L'analisi degli errori

Dario A. Bini, Beatrice Meini,
Dipartimento di Matematica, Università di Pisa
a.a. 2016-2017

24 settembre 2020

Sommario

Questo modulo didattico contiene risultati relativi allo studio della propagazione degli errori e al loro controllo nello svolgimento di elaborazioni numeriche.

1 Introduzione

Nella risoluzione di problemi del mondo reale è frequente incontrare errori a vari livelli, molto spesso anche a nostra insaputa. Gli errori hanno varia natura e sono generalmente causati dalla "finitezza" delle risorse a nostra disposizione quali strumenti di misura e risorse di calcolo.

Ad esempio, le misure fatte con gli strumenti della tecnologia, quali misure di velocità, temperature, pressioni, non possono essere esatte. Infatti gli strumenti fisici forniscono approssimazioni del valore reale, molto accurate ma pur sempre approssimazioni. La finitezza delle risorse di calcolo è un'altra sorgente importante di errori.

Un esempio significativo a questo proposito è dato dalla rappresentazione dei numeri. Si pensi ad esempio di memorizzare il numero π in un computer mediante le cifre di una sua rappresentazione in qualche base. Ovviamente non possiamo memorizzare un numero infinito di cifre, visto che il numero di celle di memoria, seppur grande, è finito. Siamo quindi costretti a fare un troncamento di π introducendo conseguentemente un errore.

La stessa situazione si presenta anche in casi apparentemente innocui e per questo più insidiosi. Ad esempio, quando digitiamo in un qualche sistema di calcolo `x=0.1` intendendo la rappresentazione in base 10, quindi 1/10, il computer apparentemente memorizza nella sua memoria il valore 1/10 che viene associato alla variabile `x`. Se poi richiediamo al computer di mostrarci `x`, ci apparirà sullo schermo il valore `0.1`. Ad esempio, usando il linguaggio *Octave*[1] e scrivendo

```
x=0.1;
disp(x)
```

[1] Un manuale di Octave si trova sul Web in versione html e in versione pdf

si ottiene

```
0.10000
```

anche usando il formato a più cifre col comando `format long` si otterrebbe

```
0.100000000000000
```

Tutto sembra regolare e tranquillo, perché stupirsi?. Però, poiché la rappresentazione dei numeri fatta all'interno del nostro computer è in base 2 (ormai è così nella quasi totalità dei computer), e poichè il numero 1/10 in base 2 ha una rappresentazione periodica, il numero effettivamente memorizzato nella variabile `x` non è 1/10 bensì una sua approssimazione, molto precisa ma pur sempre un'approssimazione, ottenuta troncando lo sviluppo periodico della rappresentazione in base 2.

In altre situazioni è il problema stesso che vorremmo risolvere che non può essere risolto in modo esatto e quindi richiede una approssimazione. Si pensi ad esempio agli zeri di un polinomio di grado maggiore o uguale a 5. Sappiamo dalla teoria di Galois che non esiste alcuna espressione formale che ci permette di rappresentare questi zeri, nel caso generale, attraverso le sole operazioni aritmetiche ed estrazioni di radici. Se vogliamo avere informazioni sugli zeri dobbiamo necessariamente approssimarli.

In certi casi, come nella risoluzione di sistemi lineari, anche se la soluzione si può esprimere attraverso un numero finito di operazioni aritmetiche, è spesso più conveniente per ragioni di costo computazionale calcolarla in modo approssimato.

La presenza degli errori nella rappresentazione dei dati come pure gli errori sviluppati nello svolgimento dei calcoli a causa del troncamento dei risultati parziali può alterare in modo drammatico il risultato finale del calcolo. Nel sito Some disasters caused by numerical errors si può trovare una lista di catastrofi causata da un errato controllo della propagazione degli errori. Tra questi, l'esplosione dell'Ariane 5, il fallimento dei missili Patriot, l'affondamento della piattaforma Sleipner.

Diventa quindi di fondamentale importanza sviluppare una strumentazione adeguata di concetti e proprietà che ci permetta di controllare e dominare gli errori e la loro propagazione.

Nel seguito, se $\tilde{x}$ è una approssimazione di x denotiamo con $\tilde{x} - x$ l'*errore assoluto* e, se $x \neq 0$, denotiamo con $(\tilde{x} - x)/x$ l'*errore relativo*. Si osserva che nell'errore relativo si rapporta l'errore assoluto al valore del dato x per cui il suo valore va letto come una "percentuale" di errore. Ad esempio, un errore relativo tale che $|(\tilde{x} - x)/x| = 1$ significa un errore del 100%, cioè una approssimazione molto scadente.

Naturalmente siamo interessati agli errori che derivano da procedimenti matematici e quindi non è compito nostro trattare gli errori provenienti da misure fisiche. La sorgente di errore più importante per noi è quella causata dalla rappresentazione in base dei numeri fatta con un numero finito di cifre. Ci occupiamo di questo nei prossimi paragrafi.

2 Rappresentazione in base

Sia $B \geq 2$ un numero intero, vale il seguente risultato di rappresentazione:

Teorema 1 (di rappresentazione in base) *Per ogni numero reale $x \neq 0$ esistono unici un intero p ed una successione $\{d_i\}_{i\geq 1}$ con le seguenti proprietà*

1) $0 \leq d_i \leq B-1$,

2) $d_1 \neq 0$,

3) per ogni $k > 0$ *esiste un* $j \geq k$ *tale che* $d_j \neq B-1$

per cui

$$x = \text{segno}(x) B^p \sum_{i=1}^{\infty} d_i B^{-i} \tag{1}$$

La proprietà espressa dal precedente risultato assume una forma più familiare se scegliamo $B = 10$ e se conveniamo di allineare gli elementi della successione in una notazione posizionale come

$$x = \pm 0.d_1 d_2 d_3 \cdots \times 10^p$$

In questo modo il numero π può essere rappresentato da

$$\pi = 0.3141592 \cdots \times 10^1$$

dove il segno + è stato omesso.

L'intero B è detto *base della rappresentazione*. Gli interi d_i per $i = 1, 2, \ldots$, sono detti le *cifre della rappresentazione* mentre p è chiamato *esponente*. Il numero $\sum_{i=1}^{\infty} d_i B^{-i}$ viene chiamato *mantissa*.

La condizione 2 è detta di *normalizzazione* ed ha un duplice scopo. Da una parte serve a garantire l'unicità della rappresentazione visto che permette di evitare rappresentazioni equivalenti quali

$$0.3141592 \times 10^1, \quad 0.0003141592 \times 10^4, \quad 3141.592 \times 10^{-3}.$$

Dall'altra permette di memorizzare in modo più efficiente un numero reale. Ad esempio nella rappresentazione 0.0003141592×10^4 sono impiegate molte più cifre rispetto a 0.3141592×10^1. Infatti, se si usa la rappresentazione normalizzata, l'informazione contenuta nelle tre cifre nulle dopo il punto decimale è codificata in modo più compatto nell'esponente che occupa una sola cifra.

La condizione 3 stabilisce che non sono ammesse configurazioni in cui da un certo punto in poi tutte le cifre sono uguali a $B-1$. Ad esempio il numero $13/100$ può essere scritto come 0.13 oppure come $0.1299999 \cdots$. La seconda rappresentazione viene vietata per garantire l'unicità

3 Numeri floating point

Un computer, essendo una macchina finita, può memorizzare una quantità finita di cifre per cui non sono fisicamente rappresentabili configurazioni dotate di un numero infinito di cifre. Diventa allora necessario, nella definizione dei numeri utilizzabili in un computer, limitarsi a rappresentazioni dotate di un numero finito di cifre. Diamo allora la seguente

Definizione 1 Dati gli interi $B \geq 2$, $t \geq 1$, $m, M > 0$, l'insieme

$$\mathcal{F}(t, B, m, M) = \{0\} \cup \{\pm B^p \sum_{i=1}^{t} d_i B^{-i},\ \ d_1 \neq 0,\ 0 \leq d_i \leq B-1,\ -m \leq p \leq M\}$$

è detto insieme dei *numeri di macchina* o anche dei *numeri in virgola mobile* o dei *numeri floating point*.

Si osservi che lo zero non è rappresentabile nella forma $\pm B^p \sum_{i=1}^{t} d_i B^{-i}$ essendo $d_1 \neq 0$. Per cui viene inserito di ufficio nell'insieme $\mathcal{F}$ dei numeri di macchina.

Sia $x \neq 0$ un numero reale rappresentato come in (1) e si consideri

$$\tilde{x} = \text{segno}(x) B^p \sum_{i=1}^{t} d_i B^{-i}.$$

Se $-m \leq p \leq M$ il numero x viene ben rappresentato in $\mathcal{F}$ dal numero $\tilde{x}$, e in questo caso la quantità $\epsilon_x = (\tilde{x} - x)/x$, cioè *l'errore relativo di rappresentazione*, è tale che

$$\left|\frac{\tilde{x} - x}{x}\right| < B^{1-t}, \qquad \left|\frac{\tilde{x} - x}{\tilde{x}}\right| < B^{1-t}. \tag{2}$$

Infatti, per definizione di $\tilde{x}$ risulta

$$|\tilde{x} - x| = B^p \sum_{i=t+1}^{+\infty} d_i B^{-i} = B^{p-t-1} \sum_{i=0}^{+\infty} d_{t+1+i} B^{-i} < \frac{B^{p-t-1}(B-1)}{1 - B^{-1}} = B^{p-t},$$

dove la diseguaglianza stretta è ottenuta maggiorando tutte le cifre con $B-1$, configurazione che non può essere mai raggiunta per le ipotesi fatte. Inoltre, poiché $d_1 \neq 0$ risulta $|x| \geq B^p \times B^{-1}$, $|\tilde{x}| \geq B^p \times B^{-1}$. Ciò implica

$$|(\tilde{x} - x)/x| < B^{1-t}, \quad |(\tilde{x} - x)/\tilde{x}| < B^{1-t}.$$

Se invece $p < -m$ oppure $p > M$ allora il numero non è rappresentabile in $\mathcal{F}(t, B, m, M)$. Nel primo caso si dice che si è incontrata una situazione di *UNDERFLOW*. Nel secondo caso si dice che si è incontrata una situazione di *OVERFLOW*. Nel caso di underflow il numero ha un valore assoluto troppo

piccolo per essere rappresentato in $\mathcal{F}$. In certi sistemi numerici esso viene rappresentato da zero. Questo fatto è deprecabile poichè se $\tilde{x} = 0$, l'errore relativo di rappresentazione vale $|\tilde{x} - x|/|x| = 1$ cioè è un errore del 100%.

Si osservi che la (2) dice che, indipendentemente dal valore di $x \neq 0$, purché non si verifichino condizioni di overflow o di underflow, l'errore relativo di rappresentazione è limitato superiormente dalla costante B^{1-t}. Cioè il sistema di numeri floating point fornisce una limitazione *uniforme* all'errore relativo di rappresentazione.

Il numero B^{1-t} viene chiamato *precisione di macchina* e rappresenta il massimo livello di precisione di un sistema floating point. Nel seguito denoteremo con $u = B^{1-t}$ la precisione di macchina.

Il numero $\tilde{x}$ rappresenta x con la precisione data da t cifre in base B. Infatti si dice che $\tilde{x}$ ha t *cifre significative* poiché tutte le t cifre di $\tilde{x}$ concorrono nel dare la massima informazione su x. Questo fatto ci porta ad estendere il concetto di numero di cifre significative nel modo seguente. Se in generale y è una approssimazione di $x \in \mathbb{R}$ tale che $|(y-x)/x| < B^{1-c}$ si dice che y ha c *cifre significative* in base B. Ciò non implica che le prime c cifre della rappresentazione in base B di x e di y coincidono. Si considerino ad esempio con $B = 10$ e $c = 5$, i valori $x = 0.12000$ e $y = 0.11999$ in cui tutte e 5 le cifre sono significative ma non tutte coincidono. In ogni caso, se y approssima x con errore relativo $\epsilon = (y - x)/x$ possiamo dire che y ha $1 + \log_B |\epsilon|^{-1}$ cifre significative.

Si osservi ancora che due sistemi floating point in base B_1 e B_2 rispettivamente con t_1 e t_2 cifre hanno precisione di macchina rispettivamente $B_1^{1-t_1}$ e $B_2^{1-t_2}$. Per cui il primo è più preciso del secondo se $B_1^{1-t_1} < B_2^{1-t_2}$. Da cui $(1 - t_1) \log B_1 < (1 - t_2) \log B_2$, cioè

$$t_1 > (t_2 - 1)\frac{\log B_2}{\log B_1} + 1.$$

Ad esempio un sistema in base 2 con 53 cifre, come quello più diffuso sui computer in commercio, ha la stessa precisione di macchina di un sistema in base 4 con 27 cifre essendo $2^{1-53} = 4^{1-27}$. Inoltre, poiché $10^{-16} < 2^{1-53} < 2.2205 \times 10^{-16}$, la precisione di questo sistema fornisce almeno 16 cifre significative in base 10.

Il tipo di approssimazione di x con $\tilde{x}$ lo abbiamo ottenuto mediante *troncamento* della rappresentazione (1). Un'altra possibilità di rappresentazione consiste nel considerare l'*arrotondamento* di x, cioè il numero di macchina più vicino a x. In questo caso si può verificare che l'errore relativo di rappresentazione è minore o uguale a $\frac{1}{2}B^{1-t}$. Nel seguito trattiamo solo il caso in cui si considera il troncamento, il caso di arrotondamento può essere trattato in modo analogo.

3.1 Rappresentazione fisica

Generalmente le rappresentazioni floating point implementate sui computer sono fatte in base $B = 2$. In particolare così è lo standard IEEE che ritroviamo sui pc con processori INTEL, e AMD. Ad esempio, una rappresentazione in base 2 con 24 cifre viene realizzata nel modo seguente ed è chiamata precisione semplice.

- Il numero viene memorizzato su 32 bit (bit è la contrazione di binary digit, cifra binaria).
- Il primo bit più a sinistra memorizza il segno della mantissa. Se il bit è 0 allora la mantissa è intesa positiva; se il bit è 1 allora la mantissa è intesa negativa.
- Gli 8 bit successivi, procedendo da sinistra a destra, racchiudono l'informazione sull'esponente p. Più precisamente, se c è il numero intero corrispondente alla rappresentazione binaria degli 8 bit, allora l'esponente p viene inteso come $p = c - 127$. Ad esempio se gli 8 bit sono `10000011` che corrisponde al numero $1 + 2 + 2^7 = 129$, allora con questa convenzione l'esponente è $p = 2$.

  In questa rappresentazione lo standard IEEE fa due eccezioni: la configurazione di 8 bit nulli e quella di 8 bit uguali a 1 vengono usate per individuare situazioni di eccezione quali `NaN` (Not-a-Number), o `+Infty` e `-Infty`. Queste situazioni si incontrano quando viene chiesto al sistema di calcolo di eseguire operazioni vietate quali ad esempio $0/0$, $\sqrt{-1}$, o, rispettivamente $x/0$ con $x \neq 0$.
- I rimanenti 23 bit contengono le cifre $d_2, d_3, \ldots, d_{24}$. Si osservi che non importa memorizzare d_1 poiché, essendo $d_1 \neq 0$ e $d_1 < 2$ è necessariamente $d_1 = 1$.
- Lo zero viene rappresentato, rompendo la convenzione, mediante la configurazione di tutti bit nulli.

Si osservi che lo zero non sarebbe rappresentabile con la convenzione adottata che richiede $d_1 \neq 0$. Identificando lo zero con la configurazione di tutti bit nulli viene sacrificato il numero $B^{-1} \times B^{-m}$ che corrisponderebbe appunto a tale configurazione. Lo standard IEEE include anche altre codifiche di eccezione quali i *denormal numbers* che permettono di memorizzare numeri più piccoli in valore assoluto di 2^{-127}. Per valutare i limiti effettivi dell'esponente provate con Matlab o Octave a scrivere

```
a = 1023 ; 2^a
```

e ottenete

```
8.9885e+307
```

mentre ponendo `a = 1024;` ottenete `Inf`. Similmente, scegliendo `a = -1075;` si ottiene `0`, mentre con `a = -1074;` si ottiene `4.9407e-324`.

La precisione di macchina in Matlab e in Octave è indicata con `eps`. Infatti se digitiamo `log2(eps)` si ottiene il valore `-52`.

Lo standard IEEE prevede anche una rappresentazione su 64 bit chiamata precisione doppia e una rappresentazione su 128 bit chiamata precisione quadrupla. Le principali caratteristiche di queste rappresentazioni sono riportate nella tabella 1.

Nome	Bit esponente	Bit mantissa	Min/Max esponente	unit roundoff	Cifre base 10	Massimo esponente decimale
Precisione semplice	8	24	-127/128	1.2×10^{-7}	7.92	38.23
Precisione doppia	11	53	-1023/1024	2.2×10^{-16}	16.65	307.95
Precisione quadrupla	15	113	-16383/16384	1.9×10^{-34}	34.7	4931.77

Tabella 1: Rappresentazioni floating point in base 2: ripartizione dei bit tra esponente e mantissa e massimi valori ottenibili per l'esponente, la precisione, l'equivalente delle cifre decimali e l'equivalente dell'esponente in base 10. Nello standard IEEE effettivamente implementato alcune configurazioni di bit sono dedicate a codificare situazioni di eccezione quali `NaN`, `+Inf`, `-Inf` e altro, inoltre sono codificati i *denormal numbers* che permettono di memorizzare numeri di valore assoluto più piccolo del minimo altrimenti consentito. Per questo motivo i limiti dell'esponente sono leggermente diversi.

Una rappresentazione basata su 80 bit e adottata nei coprocessori matematici della serie Intel 8087 e successivamente nel Motorola 68881, è la rappresentazione estesa che usa 64 bit per la mantissa, 15 bit per l'esponente e un bit per il segno. Il numero di cifre della rappresentazione in questo caso è $t = 65$, e corrisponde a poco più di 19 cifre decimali. Il massimo e minimo esponente consentiti forniscono una copertura dei numeri positivi nel segmento $[3.65 \times 10^{-4951}, 1.19 \times 10^{4932}]$.

Sistemi di rappresentazione numerica a precisione variabile sono stati progettati per venire incontro a certe applicazioni, come quelle incontrate nei problemi di critto-analisi, in cui sono richieste alte precisioni di calcolo. Vale la pena citare uno dei pacchetti tra i più efficienti e di libero uso, coperto dalla licenza GNU, che è il GMP - GNU MultiPrecision Arithmetic Library, dove sono implementati gli algoritmi più sofisticati e veloci per la moltiplicazione di numeri dotati di molte cifre, quali il metodo di Karatsuba e l'algoritmo di Schoenhage-Strassen.

3.2 Aritmetica di macchina

Si osservi che se $a, b \in \mathcal{F}(t, B, m, M)$ non è detto che $c = a \,\mathrm{op}\, b$ appartenga ad $\mathcal{F}$, dove "op" è una delle quattro operazioni aritmetiche. Quindi per poter operare sui numeri di $\mathcal{F}$ dobbiamo introdurre una aritmetica approssimata nel seguente modo

$$\hat{c} = a\,[\mathrm{op}]\,b, \quad a\,[\mathrm{op}]\,b = \mathrm{tronc}(a \,\mathrm{op}\, b)$$

dove $\mathrm{tronc}(x)$ indica il troncamento del numero reale x a t cifre. In questo modo, se nello svolgere l'operazione aritmetica non si verificano situazioni di underflow

o di overflow, e $\hat{c} \neq 0$, allora per la (2) l'errore relativo $\delta = (\hat{c} - c)/c$ è tale che $|\delta| < u$. Analogamente per $\eta = (\hat{c} - c)/\hat{c}$ vale $|\eta| < u$ dove $u = B^{1-t}$ è la precisione di macchina. Possiamo quindi scrivere che

$$\hat{c} = (a \,[\mathrm{op}]\, b) = c(1+\delta) = c/(1+\eta), \qquad |\delta|, |\eta| < u. \tag{3}$$

L'errore relativo commesso δ (rispetto a c) o η (rispetto a $\hat{c}$) nel mantenere il risultato dell'operazione dentro l'insieme $\mathcal{F}$ viene chiamato *errore locale* generato dall'operazione floating point. Ciò definisce una aritmetica approssimata nell'insieme $\mathcal{F}$ che purtroppo non gode di molte delle proprietà formali algebriche. In particolare per l'aritmetica floating point non vale l'associatività delle operazioni e la distributività del prodotto rispetto alla somma. Inoltre ogni operazione aritmetica è potenzialmente sorgente di errori. Diventa quindi particolarmente importante *capire se e quando questi errori generati possono o meno alterare il risultato di un calcolo in modo significativo.*

4 Errori nel calcolo di una funzione

Supponiamo di avere assegnata una funzione $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$. Il nostro desiderio è quello di calcolare il valore di $f(x)$ per un valore assegnato di $x \in \Omega \subset \mathbb{R}^n$. Purtroppo dobbiamo accontentarci di calcolare $f(\tilde{x})$ dove $\tilde{x} \in \mathcal{F}^n \cap \Omega$ è una n-upla di numeri di macchina tali che $\tilde{x}_i = x_i(1+\epsilon_i)$, dove ϵ_i sono gli errori di rappresentazione tali che $|\epsilon_i| < u$. Operando in questo modo, già prima di iniziare i calcoli, abbiamo a che fare con l'errore relativo

$$\epsilon_{\mathrm{in}} = \frac{f(\tilde{x}) - f(x)}{f(x)}$$

definito se $f(x) \neq 0$. Tale errore viene chiamato *errore inerente* ed è l'errore dovuto agli errori di rappresentazione. Cioè esso è indotto nella funzione dal fatto che il valore della variabile indipendente $x \in \mathbb{R}^n$ viene alterato in $\tilde{x} \in \mathcal{F}^n$.

4.1 Funzioni razionali

Supponiamo ora che la funzione $f(x) : \mathbb{R} \to \mathbb{R}$ sia razionale, cioè una funzione data dal quoziente di due polinomi, dove Ω è l'insieme dei valori in cui il denominatore non si annulla. Le funzioni razionali sono le sole che si possono calcolare con un numero finito di operazioni aritmetiche. Vogliamo calcolare il valore di $f(\tilde{x})$ eseguendo una opportuna sequenza di operazioni aritmetiche che per semplicità definiremo *algoritmo di calcolo* o più semplicemente algoritmo. Nella realizzazione in aritmetica floating point di un algoritmo ogni operazione aritmetica potenzialmente introduce un errore locale limitato superiormente in valore assoluto dalla precisione di macchina.

In questo modo il valore che otteniamo alla fine dei calcoli in generale non coinciderà con quello di $f(\tilde{x})$ ma sarà qualcosa di diverso in generale che indichiamo con $\varphi(\tilde{x})$. Definiamo quindi l'*errore algoritmico* come

$$\epsilon_{\text{alg}} = \frac{\varphi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}.$$

L'errore algoritmico è quindi generato dall'accumularsi degli errori locali relativi a ciascuna operazione aritmetica eseguita in floating point.

Chiamiamo invece `errore totale` la quantità

$$\epsilon_{\text{tot}} = \frac{\varphi(\tilde{x}) - f(x)}{f(x)}$$

che esprime di quanto il valore effettivamente calcolato $\varphi(\tilde{x})$ si discosta dal valore $f(x)$ che avremmo voluto calcolare.

Si può dimostrare facilmente che vale

$$\epsilon_{\text{tot}} = \epsilon_{\text{in}} + \epsilon_{\text{alg}} + \epsilon_{\text{in}}\epsilon_{\text{alg}} \doteq \epsilon_{\text{in}} + \epsilon_{\text{alg}}, \tag{4}$$

dove col segno $\doteq$ indichiamo l'uguaglianza delle parti lineari negli errori. Di fatto, con l'operazione $\doteq$ manteniamo solamente la parte lineare nell'errore trascurando tutti i termini di ordine quadratico o superiore. Questo tipo di analisi, detta *al primo ordine* è significativa in concreto poiché nella pratica gli errori sono piccoli e i loro prodotti o le loro potenze intere con esponente maggiore di 1 diventano trascurabili.

La dimostrazione di (4) è un semplice conto. Vale infatti

$$\epsilon_{\text{tot}} = \frac{\varphi(\tilde{x})}{f(x)} - 1 = \frac{\varphi(\tilde{x})}{f(\tilde{x})}\frac{f(\tilde{x})}{f(x)} - 1 = (\epsilon_{\text{alg}} + 1)(\epsilon_{\text{in}} + 1) - 1 = \epsilon_{\text{alg}} + \epsilon_{\text{in}} + \epsilon_{\text{alg}}\epsilon_{\text{in}}.$$

Cioè, in una analisi al primo ordine, l'errore totale può essere scisso nella somma dell'errore inerente e di quello algoritmico. Per cui basta studiare separatamente questi due tipi di errori per avere il valore dell'errore totale. È quindi importante disporre di strumenti per studiare l'errore inerente ϵ_{in} e l'errore algoritmico ϵ_{alg}.

4.2 Funzioni non razionali

Nel caso di una funzione non razionale $g(x) : \Omega \subset \mathbb{R} \to \mathbb{R}$, vale ancora la definizione di errore inerente, mentre non possiamo definire un errore algoritmico poiché $g(x)$ non può essere calcolata in un numero finito di operazioni aritmetiche. Per poter calcolare $g(x)$ dobbiamo selezionare una funzione razionale $f(x)$ che ben approssimi $g(x)$. Per questo introduciamo *l'errore analitico* definito da

$$\epsilon_{\text{an}} = \frac{f(x) - g(x)}{g(x)}$$

che esprime di quanto si discosta la funzione razionale $f(x)$ dalla $g(x)$.

Un esempio tipico è il calcolo di e^x, con $x > 0$, mediante la formula

$$e^x = \sum_{i=0}^{+\infty} \frac{x^i}{i!}$$

per cui si può scegliere la funzione razionale

$$f(x) = \sum_{i=0}^{n} \frac{x^i}{i!}$$

dove n è sufficientemente grande in modo che il resto $x^{n+1}e^{\xi}/(n+1)! < x^{n+1}e^x/(n+1)!$, con $0 < \xi < x$, sia minore di ue^x. Questo si ottiene se n è tale che $x^{n+1}/(n+1)! < u$. Infatti, sotto quest'ultima condizione, aggiungere ulteriori addendi alla quantità $\sum_{i=0}^{n} x^i/i!$ non cambia le prime t cifre della rappresentazione in base.

La scelta di $g(x)$ può essere fatta in vari modi ad esempio troncando degli sviluppi in serie, come si è fatto nel calcolo dell'esponenziale, oppure mediante tecniche di interpolazione, approssimanti di Padé ed altro ancora.

Considerando l'errore totale come $(\varphi(\tilde{x}) - g(x))/g(x)$, si può dimostrare la seguente proprietà

$$\epsilon_{\text{tot}} = \epsilon_{\text{in}}+\epsilon_{\text{alg}}+\epsilon_{\text{an}}(\tilde{x})+\epsilon_{\text{in}}\epsilon_{\text{alg}}+\epsilon_{\text{in}}\epsilon_{\text{an}}(\tilde{x})+\epsilon_{\text{alg}}\epsilon_{\text{an}}(\tilde{x})+\epsilon_{\text{alg}}\epsilon_{\text{an}}(\tilde{x})\epsilon_{\text{in}} \doteq \epsilon_{\text{in}}+\epsilon_{\text{alg}}+\epsilon_{\text{an}}(\tilde{x}).$$

Infatti si ha

$$\begin{aligned}
\epsilon_{\text{tot}} &= \frac{\varphi(\tilde{x})}{g(x)} - 1 = \frac{\varphi(\tilde{x})}{f(\tilde{x})}\frac{f(\tilde{x})}{g(\tilde{x})}\frac{g(\tilde{x})}{g(x)} - 1 \\
&= (\epsilon_{\text{alg}} + 1)(\epsilon_{\text{an}}(\tilde{x}) + 1)(\epsilon_{\text{in}} + 1) - 1 \\
&= \epsilon_{\text{alg}} + \epsilon_{\text{an}}(\tilde{x}) + \epsilon_{\text{in}} + \epsilon_{\text{alg}}\epsilon_{\text{in}} + \epsilon_{\text{alg}}\epsilon_{\text{an}}(\tilde{x}) + \epsilon_{\text{an}}(\tilde{x})\epsilon_{\text{in}} + \epsilon_{\text{alg}}\epsilon_{\text{an}}(\tilde{x})\epsilon_{\text{in}}.
\end{aligned}$$

Quindi, ai fini dell'analisi degli errori possiamo studiare separatamente gli errori di ϵ_{in}, ϵ_{alg}, ϵ_{an}. Per quanto riguarda l'errore analitico possiamo usare gli strumenti dell'analisi e della teoria dell'approssimazione di funzioni. Lo studio dell'errore inerente e algoritmico viene riportato di seguito.

4.3 Analisi dell'errore inerente

Se $n = 1$, cioè se $f(x) : \mathbb{R} \to \mathbb{R}$, allora assumendo che $f(x)$ sia definita e derivabile almeno due volte con continuità nel segmento di estremi x e $\tilde{x}$, uno sviluppo in serie di Taylor di $f(x)$ fornisce

$$f(\tilde{x}) = f(x) + (\tilde{x} - x)f'(x) + \frac{1}{2}(\tilde{x} - x)^2 f''(\xi), \quad |\xi - x| < |\xi - \tilde{x}|.$$

Da cui, considerando l'errore di rappresentazione $\delta_x = (\tilde{x} - x)/x$, si ricava

$$\epsilon_{\text{in}} = \delta_x \frac{xf'(x)}{f(x)} + \delta_x^2 \frac{x^2 f''(\xi)}{f(x)} \doteq \delta_x \frac{xf'(x)}{f(x)} \tag{5}$$

La quantità $\frac{xf'(x)}{f(x)}$ che compare nella (5) viene detta *coefficiente di amplificazione* e ci dice di quanto l'errore relativo δ_x presente nei dati viene amplificato (o ridotto) nel valore di $f(x)$ in una analisi al primo ordine. Ad esempio, se $f(x) = x^p$, con p intero, un semplice calcolo mostra che il coefficiente di amplificazione di $f(x)$ è p, cioè un errore relativo δ_x presente nella x viene amplificato

di p volte nella $f(x)$. Mentre per la funzione $x^{1/p}$ il coefficiente di amplificazione è $1/p$, cioè l'errore relativo nella x viene ridotto di p volte nella $f(x)$. Se $f(x) = \log x$ allora il coefficiente di amplificazione è $1/\log x$. Per cui, se $x > e$ o se $x < 1/e$ c'è una riduzione di errore.

Si osservi che la definizione di errore inerente non richiede che $f(x)$ sia razionale, basta solo che sia definita e sufficientemente regolare sull'intervallo di estremi x e $\tilde{x}$.

Nel caso di funzioni $f : \mathbb{R}^n \to \mathbb{R}$ vale una formula analoga per l'errore inerente. Infatti, posto $x = (x_i)$ e $\delta_{x_i} = (\tilde{x}_i - x_i)/x_i$ per $i = 1, \ldots, n$, risulta

$$\epsilon_{\text{in}} \doteq \sum_{i=1}^{n} \delta_{x_i} C_i, \qquad C_i = \frac{x_i \frac{\partial f(x)}{\partial x_i}}{f(x)}. \tag{6}$$

Le quantità C_i sono i coefficienti di amplificazione rispetto alla variabile x_i, per $i = 1, \ldots, n$, dove il simbolo $\partial f(x)/\partial x_i$ denota la derivata di $f(x)$ rispetto alla variabile x_i.

Legato all'errore inerente è il concetto di *condizionamento* di un problema. Si dice che un problema è ben condizionato se una "piccola" variazione relativa dei valori di input produce una "piccola" variazione relativa dei valori di output. Si dice *mal condizionato* se una piccola variazione relativa in input produce una "grande" variazione relativa nell'output. In altri termini, il condizionamento di un problema, quale il calcolo di una funzione, è ben o mal condizionato a seconda della grandezza in modulo dei coefficienti di amplificazione.

4.4 Analisi dell'errore algoritmico

Per l'errore inerente abbiamo introdotto lo strumento dei coefficienti di amplificazione che ci permette, mediante uno studio analitico, di calcolare tale errore in modo agevole. Per studiare l'errore algoritmico occorre faticare un po' di più.

Per studiare l'errore algoritmico consideriamo la più semplice funzione possibile $f(x_1, x_2) = x_1 \,\text{op}\, x_2$, dove op è una delle quattro operazioni aritmetiche, col più semplice algoritmo possibile: quello che esegue una singola operazione aritmetica

$$s = x_1 \,\text{op}\, x_2.$$

Nell'esecuzione di questo semplice algoritmo in aritmetica floating point l'unico errore generato dall'aritmetica approssimata è l'errore locale δ generato dal troncamento del risultato dell'operazione aritmetica. Infatti il valore $\tilde{s}$ effettivamente calcolato è dato da

$$\tilde{s} = (\tilde{x}_1 \,\text{op}\, \tilde{x}_2)(1 + \delta),$$

dove $|\delta| < u$ è l'errore locale, $\tilde{x}_1 = x_1(1 + \epsilon_{x_1})$ e $\tilde{x}_2 = x_2(1 + \epsilon_{x_2})$ sono i valori approssimati degli operandi. Gli errori ϵ_{x_1} e ϵ_{x_2} possono essere gli errori di rappresentazione di x_1 e di x_2, se essi sono dati in input, oppure possono essere gli errori accumulati nelle operazioni precedentemente svolte per calcolare x_1 e

Operazione	C_1	C_2
moltiplicazione	1	1
divisione	1	-1
addizione	$\frac{x_1}{x_1+x_2}$	$\frac{x_2}{x_1+x_2}$
sottrazione	$\frac{x_1}{x_1-x_2}$	$-\frac{x_2}{x_1-x_2}$

Tabella 2: Coefficienti di amplificazione delle operazioni aritmetiche

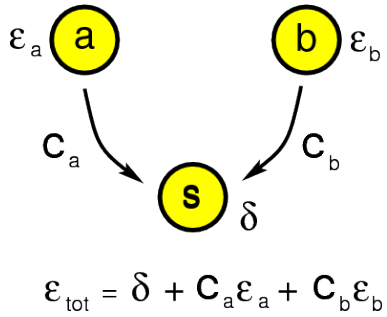


Figura 1: Errore totale in una singola operazione aritmetica

x_2. Quest'ultimo caso è quello che si avrebbe considerando questa singola operazione aritmetica come singola parte di un algoritmo più complesso costituito da più operazioni.

Dalla (4) è evidente che l'errore totale in s, al primo ordine, è dato dalla somma dell'errore algoritmico cioè δ e dell'errore inerente, cioè $C_1\epsilon_{x_1} + C_2\epsilon_{x_2}$, dove C_1 e C_2 sono i coefficienti di amplificazione relativamente a x_1 ed a x_2 della funzione $f(x_1, x_2) = x_1 \,\mathrm{op}\, x_2$, cioè

$$\tilde{s} \doteq (x_1 \,\mathrm{op}\, x_2)(1 + \delta + C_1\epsilon_{x_1} + C_2\epsilon_{x_2}).$$

Diventa quindi determinante studiare i coefficienti di amplificazione delle quattro funzioni $x_1 + x_2$, $x_1 - x_2$, $x_1 \times x_2$, x_1/x_2.

Un semplice calcolo ci permette di ottenere i coefficienti di amplificazione C_1 e C_2 relativi alle due variabili x_1 e x_2 che sono riportati nella tabella 2.

La figura 1 mostra in modo grafico il flusso delle operazioni e degli errori nell'esecuzione di una singola operazione applicata agli operandi a e b.

Si osserva che le operazioni di moltiplicazione e di divisione non amplificano eventuali errori presenti negli operandi. Anche l'addizione tra numeri di segno concorde ha dei coefficienti di amplificazione più piccoli di 1 in valore assoluto.caso L'unica operazione pericolosa che può amplificare in modo incontrollato gli errori presenti nei dati è la sottrazione di numeri concordi o l'addizione di numeri discordi in segno. Infatti in questi casi i rapporti $x_1/(x_1+x_2)$ e $x_2/(x_1+x_2)$ possono assumere valori arbitrariamente grandi in valore assoluto.

Questo fenomeno di amplificazione degli errori che si manifesta nel caso di somma di numeri di segno opposto o sottrazione di numeri di segno concorde viene chiamato *cancellazione numerica*. Il termine è motivato dal fatto che due numeri dello stesso segno che nella sottrazione danno un risultato piccolo in valore assoluto hanno necessariamente molte cifre in comune che si cancellano nell'operazione di sottrazione.

Esempio. Siano $a = 0.12345678$ e $b = 0.12345675$ le rappresentazioni in base 10 di due numeri. Siano $\tilde{a} = 0.12345679$, $\tilde{b} = 0.12345674$ i due valori perturbati in cui $\epsilon_a = 0.81 \times 10^{-7}$, $\epsilon_b = -0.81 \times 10^{-7}$. Risulta $c = a - b = 0.3 \times 10^{-7}$, mentre $\tilde{c} = \tilde{a} - \tilde{b} = 0.5 \times 10^{-7}$, dove la sottrazione è svolta in modo esatto. L'errore che compare nel risultato è $(\tilde{c} - c)/c = 0.4$. Si è passati da un errore relativo in a e in b dell'ordine di 1 su 10 milioni ad un errore relativo nel risultato del 40%. Nello svolgere la sottrazione si verifica la cancellazione di cifre

$$\begin{array}{ll} 0.12345679 & - \\ 0.12345674 & = \\ \hline 0.00000005 & \end{array}$$

È importante ribadire che il fenomeno della cancellazione consiste nella amplificazione degli errori presenti negli operandi e non è dovuta all'errore locale dell'addizione o sottrazione eseguita.

Affinchè un algoritmo non amplifichi troppo l'errore è importante fare in modo che non si presentino delle cancellazioni nei singoli passi dell'algoritmo stesso. Questa semplice regola pratica può essere applicata senza difficoltà in molte situazioni. Abbiamo quindi la seguente ricetta da seguire

Evitare di addizionare numeri di segno opposto o, equivalentemente, di sottrarre numeri dello stesso segno

Esempio. Nel calcolare le radici di un polinomio di secondo grado $ax^2 + bx + c$ viene generalmente usata la formula

$$\frac{-b \pm \sqrt{\Delta}}{2a}, \quad \Delta = b^2 - 4ac.$$

In almeno una delle due operazioni in cui il segno $\pm$ è coinvolto si verifica una cancellazione numerica. Ad esempio, se è $b > 0$, allora il calcolo di $x_1 = -b + \sqrt{\Delta}$ comporta una somma di numeri di segno opposto, mentre il calcolo di $x_2 = -b - \sqrt{\Delta}$ è sicuro. La cancellazione nel calcolo di x_1 può essere evitata utilizzando il fatto che $x_1 x_2 = c/a$ per cui possiamo scrivere $x_1 = c/(a * x_2)$. Quest'ultima formula non comporta cancellazioni numeriche.

Esempio. Nell'approssimare il valore di e^x per un valore assegnato di x possiamo usare lo sviluppo in serie

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

e sommare finchè il risultato non cambia più cioè finché il resto dello sviluppo in serie diventa più piccolo in valore assoluto della precisione di macchina per e^x. Se $x > 0$ non si verifica cancellazione, ma se $x < 0$ abbiamo una somma a segni alterni in cui, se $x << -1$ il risultato finale è molto più piccolo rispetto agli addendi. Quindi la formula genera cancellazione. In tal caso possiamo rimuovere il problema scrivendo $e^x = 1/e^{-x}$ e approssimando lo sviluppo in serie di e^{-x} che si riconduce ad una somma di termini positivi.

Esempio. Un altro esempio significativo riguarda il calcolo della somma

$$\sum_{i=1}^{2n} \frac{(-1)^{i-1}}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

Se applichiamo la formula così com'è andiamo a sommare e sottrarre quantità positive e quindi si incorre nella cancellazione. Se invece riscriviamo l'espressione come

$$\sum_{i=1}^{n} \frac{1}{2i-1} - \frac{1}{2i} = \sum_{i=1}^{n} \frac{1}{2i(2i-1)}$$

andiamo ad eseguire solo somme di numeri positivi evitando cancellazione.

Un modo per valutare l'errore algoritmico generato da un algoritmo di calcolo consiste nell'applicare a ciascuna operazione aritmetica l'analisi descritta sopra. Possiamo vedere questo nel caso del calcolo della funzione $a^2 - b^2$ mediante i due seguenti schemi di calcolo

Schema 1
$s_1 = a \times a$
$s_2 = b \times b$
$s_3 = s_1 - s_2$

Schema 2
$s_1 = a + b$
$s_2 = a - b$
$s_3 = s_1 \times s_2$

Il primo algoritmo esegue due moltiplicazioni e una addizione, il secondo una addizione, una sottrazione e una moltiplicazione. I due algoritmi sono rappresentati dai grafi in figura 2.

Denotando con ϵ_i l'errore algoritmico sulla variabile s_i, per il primo algoritmo si ha: $\epsilon_1 = \delta_1$, $\epsilon_2 = \delta_2$, da cui

$$\epsilon_3 \doteq \delta_3 + \frac{a^2}{a^2 - b^2}\delta_1 - \frac{b^2}{a^2 - b^2}\delta_2$$

che ci fornisce la maggiorazione al primo ordine

$$|\epsilon_3| < u(1 + \frac{a^2 + b^2}{|a^2 - b^2|}).$$

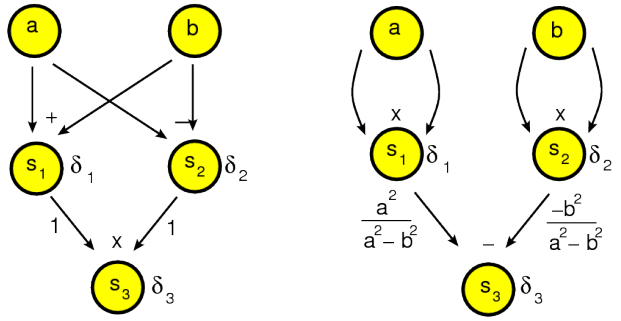


Figura 2: Due algoritmi per il calcolo di $a^2 - b^2$

Nel secondo algoritmo si ha $\epsilon_1 = \delta_1$, $\epsilon_2 = \delta_2$ e quindi

$$\epsilon_3 \doteq \delta_3 + \delta_1 + \delta_2$$

da cui la maggiorazione al primo ordine

$$|\epsilon_3| < 3u.$$

Un modo formalmente diverso, ma sostanzialmente equivalente di condurre una analisi dell'errore consiste nell'applicare la relazione (3) ad ogni operazione. Ad esempio nel caso del secondo algoritmo, la formula $s_3 = (a - b)(a + b)$ si trasforma in

$$\tilde{s}_3 = [(a - b)(1 + \delta_1)][(a + b)(1 + \delta_2)](1 + \delta_3) \doteq (a^2 - b^2)(1 + \delta_1 + \delta_2 + \delta_3).$$

5 Analisi all'indietro (backward analysis)

L'analisi dell'errore che abbiamo descritto ha l'obiettivo di arrivare a dare maggiorazioni al valore assoluto dell'errore algoritmico ottenuto alla fine dei calcoli ed è chiamata *analisi in avanti*. Generalmente una analisi di questo tipo è piuttosto tecnica e laboriosa. Una possibilità diversa che in molti casi semplifica lo studio dell'errore è l'*analisi all'indietro* introdotta da J. H. Wilkinson. La descriviamo prima nel caso di una singola operazione.

Consideriamo la somma $z \in \mathbb{R}$ di due numeri di macchina $x, y \in \mathcal{F}$, cioè $z = x + y$. Se l'operazione viene eseguita in aritmetica floating point si otterrà un valore $\tilde{z} \in \mathcal{F}$ tale che $\tilde{z} = (x + y)(1 + \delta)$ dove $|\delta| < u$. Questa espressione la possiamo scrivere in questa forma

$$\tilde{z} = \hat{x} + \hat{y}, \quad \hat{x} = x(1 + \delta_x) \in \mathbb{R}, \quad \hat{y} = y(1 + \delta_y) \in \mathbb{R}.$$

Cioè il risultato *effettivamente calcolato in aritmetica floating* point lo posso vedere come il risultato *calcolato in modo esatto* a partire però dai valori $\hat{x}$ e $\hat{y}$

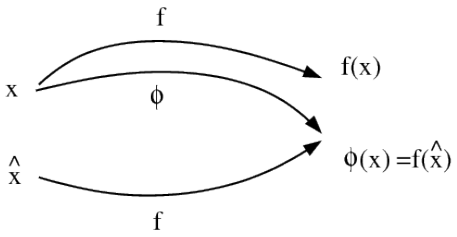


Figura 3: Analisi all'indietro dell'errore

che rispetto ai valori originali x, y hanno un errore relativo rispettivamente δ_x e δ_y.

In generale, sia $f(x_1, \ldots, x_n)$ una funzione razionale e si denoti $\varphi(x_1, \ldots, x_n)$ la funzione definita su $\mathcal{F}^n$ i cui valori sono ottenuti calcolando $f(x_1, \ldots, x_n)$ con l'aritmetica di macchina. Nell'analisi all'indietro dell'errore si cercano delle perturbazioni $\delta_1, \ldots, \delta_n$ tali che denotando con $\hat{x}_i = x_i(1 + \delta_i)$ per $i = 1, \ldots, n$ risulti

$$\varphi(x_1, \ldots, x_n) = f(\hat{x}_1, \ldots, \hat{x}_n).$$

Cioè si cerca di esprimere il valore di una funzione effettivamente calcolato in aritmetica di macchina come il valore della funzione originale $f(x_1, \ldots, x_n)$ calcolato però in un punto $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ leggermente spostato. Lo scopo è quello di dare delle maggiorazioni al valore assoluto delle perturbazioni δ_i. In questo modo l'errore algoritmico viene visto formalmente come un errore inerente, cioè causato da una perturbazione dell'input. A questo punto, se vogliamo dare maggiorazioni all'errore algoritmico conoscendo limitazioni superiori al valore assoluto delle perturbazioni, possiamo applicare i coefficienti di perturbazione e usare l'espressione (5).

La figura 3 mostra graficamente l'idea alla base della analisi all'indietro dell'errore.

Occorre dire che in generale non è sempre possibile svolgere una analisi all'indietro. Questo accade tipicamente quando il numero di variabili in gioco è inferiore al numero di operazioni da svolgere.

6 Esempi

L'errore totale generato nel calcolo del prodotto di n numeri si analizza facilmente sia mediante una analisi in avanti che mediante una analisi all'indietro. Sia $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i$. Il coefficiente di amplificazione rispetto alla variabile x_i è 1. Quindi l'errore inerente è dato al primo ordine da

$$\epsilon_{\text{in}} \doteq \sum_{i=1}^{n} \epsilon_{x_i},$$

per cui, se $|\epsilon_{x_i}| < u$ allora $|\epsilon_{\text{in}}| < nu$.

Per quanto riguarda l'errore algoritmico, calcolando il prodotto mediante la formula

$$(\cdots((x_1 \times x_2) \times x_3) \times \cdots) \times x_n$$

si ha

$$\varphi = \prod_{i=1}^{n} x_i \prod_{j=1}^{n-1}(1+\delta_j)$$

dove δ_i è l'errore locale dell'i-esima moltiplicazione. Per cui l'errore algoritmico al primo ordine è maggiorato in valore assoluto da $(n-1)u$.

L'analisi all'indietro si effettua scrivendo la relazione precedente come $\prod_{i=1}^{n} \hat{x}_i$ con $\hat{x}_i = x_i(1+\delta_i)$ per $i = 1, \ldots, n-1$, $\hat{x}_n = x_n$.

Una situazione più problematica si incontra nello studio degli errori di una somma di n termini. Infatti, già nell'analisi dell'errore inerente si incontrano coefficienti di amplificazione dati da

$$x_i / \sum_{j=1}^{n} x_j$$

che possono avere valore assoluto arbitrariamente elevato a meno che la somma non abbia tutti addendi dello stesso segno. In questo caso la somma dei valori assoluti dei coefficienti di amplificazione fa 1 per cui l'errore inerente è maggiorato in modulo al primo ordine da u.

Invece per l'errore algoritmico occorre prima specificare in che modo la somma di n addendi viene calcolata. Due tra i numerosi modi diversi, legati alla proprietà associativa dell'addizione, sono dati dal metodo di somma in sequenza e dal metodo di somma in parallelo. Il primo procede secondo lo schema

$s_0 = x_1$
$s_i = s_{i-1} + x_{i+1}$, per $i = 1, \ldots, n-1$

Il secondo procede in base allo schema che per semplicità riportiamo nel caso di $n = 2^p$, p intero positivo.

$s_i^{(0)} = x_i$ per $i = 1, \ldots, n$
$s_i^{(k)} = s_{2i-1}^{(k-1)} + s_{2i}^{(k-1)}$ per $i = 1, \ldots, n/2^k$, $k = 1, 2, \ldots, p-1$.

Se n non fosse potenza intera di 2 basta porre $x_i = 0$ per $i = n+1, \ldots, 2^p$ dove 2^p è la più piccola potenza intera di 2 maggiore o uguale a n.

I grafi relativi al flusso delle operazioni sono riportati nelle figure 4, 5.

Di seguito si riportano i codici *Octave* dei due metodi di somma dove la somma in parallelo viene implementata in modo ricorsivo.

```
function s=somma(x)
% Algoritmo di somma in sequenza
```

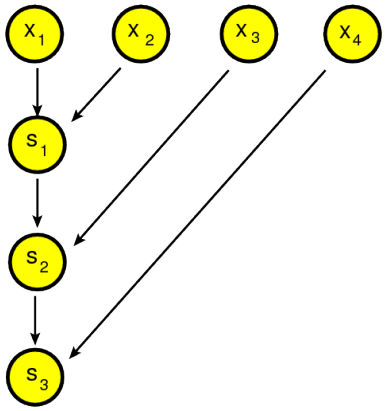


Figura 4: Somma sequenziale

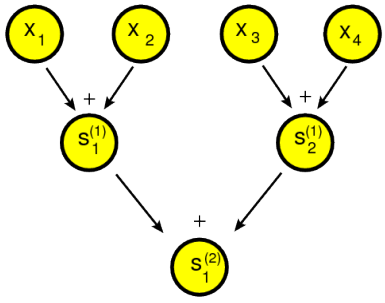


Figura 5: Somma in parallelo

Errori algoritmici	sequenziale	parallelo
crescente	$(n-1)u$	$\lceil \log_2 n \rceil u$
decrescente	$\frac{n}{2}u$	$\lceil \log_2 n \rceil u$

```
  n=length(x);
  s=x(1);
  for i=2:n
    s=s+x(i);
  endfor
endfunction

function s=somma_p(x)
% Algoritmo di somma in parallelo
% implementazione ricorsiva
  n=length(x);
  if n==2
  s=x(1)+x(2);
  elseif n==1
    s=x(1)
  else
    if mod(n,2)==0              % n pari
      y=x(1:2:n)+x(2:2:n);
      s=somma_p(y);
    else                        % n dispari
      y=x(1:2:n-1)+x(2:2:n-1);
      s=somma_p(y)+x(n);
    endif
  endif
endfunction
```

Si può dimostrare che nel caso di coefficienti non negativi l'errore algoritmico generato dai due algoritmi con i due ordinamenti diversi è maggiorato al primo ordine dalle seguenti quantità

Una analisi all'indietro del metodo di somma sequenziale fornisce il seguente risultato

$$\text{float}(\texttt{somma(x)}) = \sum_{i=1}^{n} \tilde{x}_i, \quad \tilde{x}_i = x_i(1+\epsilon_i^{(n)}), \quad |\epsilon_i^{(n)}| < u(n-i+1)$$

Per dimostrare questo si procede per induzione su n. Se $n = 2$ allora $\text{float}(x_1+x_2) = (x_1+x_2)(1+\delta_1) = \tilde{x}_1+\tilde{x}_2$, dove $\tilde{x}_1 = x_1(1+\delta_1)$, $\tilde{x}_2 = x_2(1+\delta_1)$ con $|\delta_1| < u$, e quindi la proprietà è valida. Assumendo valida la proprietà per $n-1$ si considera il caso n. Vale $s_n = s_{n-1} + x_n$, con $s_n = \sum_{i=1}^{n} x_i$. Per cui, il valore effettivamente calcolato $\tilde{s}_n$ è tale che $\tilde{s}_n = (\tilde{s}_{n-1} + x_n)(1+\delta_{n-1})$ con δ_{n-1} errore locale dell'addizione. Ne segue $\tilde{s}_n = \sum_{i=1}^{n-1} x_i(1+\epsilon_i^{(n-1)})(1+\delta_{n-1})$.

Da cui $\epsilon_i^{(n)} \doteq \epsilon_i^{(n-1)} + \delta_{n-1}$ per $i = 1, \ldots, n-1$, $\epsilon_n^{(n)} = \delta_{n-1}$. Quindi $|\epsilon_i^{(n)}| < |\epsilon_i^{(n-1)}| + u < n - i + 1$. La dimostrazione è completa.

Una analisi all'indietro del metodo di somma parallela fornisce il seguente risultato

$$\text{float}(\texttt{somma\_p(x)}) = \sum_{i=1}^{n} \tilde{x}_i, \quad \tilde{x}_i = x_i(1+\epsilon_i), \quad |\epsilon_i| < u\lceil \log_2 n \rceil.$$

Per semplicità dimostriamo questo fatto nel caso in cui $n = 2^q$ con q intero positivo. Nel caso generale basta aggiungere addendi nulli fino ad arrivare ad un numero di addendi uguale alla prima potenza di 2 maggiore o uguale ad n. Riscriviamo l'algoritmo nel seguente modo:

$$\begin{aligned} &s_i^{(0)} = x_i, \quad i = 1, \ldots, n \\ &s_i^{(k+1)} = s_{2i-1}^{(k)} + s_{2i}^{(k)}, \quad i = 1, \ldots, n/2^k, \quad k = 0, 1, \ldots, p-1 \end{aligned}$$

Dimostriamo per induzione su k che $\tilde{s}_i^{(k)} = s_i^{(k)}(1+\epsilon_i^{(k)})$, $|\epsilon_i^{(k)}| < ku$. Per $k = 0$ la relazione è chiaramente verificata. Per il passo induttivo si ha

$$\tilde{s}_i^{(k+1)} = (\tilde{s}_{2i-1}^{(k)} + \tilde{s}_{2i}^{(k)})(1+\delta_i^{(k)}), \quad |\delta_i^{(k)}| < u.$$

Da cui

$$\tilde{s}_i^{(k+1)} \doteq s_{2i-1}^{(k)}(1+\epsilon_{2i-1}^{(k)} + \delta_i^{(k)}) + s_{2i}^{(k)}(1+\epsilon_{2i}^{(k)} + \delta_i^{(k)}).$$

Dall'ipotesi induttiva si deduce che $|\epsilon_{2i-1}^{(k+1)}| = |\epsilon_{2i-1}^{(k)} + \delta_i^{(k)}| < (k+1)u$ e, similmente, $|\epsilon_{2i}^{(k+1)}| = |\epsilon_{2i}^{(k)} + \delta_i^{(k)}| < (k+1)u$. Questo completa la dimostrazione.

Un esempio significativo di analisi all'indietro riguarda il calcolo del determinante di una matrice tridiagonale $n \times n$

$$A_n = \begin{bmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{bmatrix}$$

Infatti, denotando con $x_n = \det A_n$, calcolando il determinante con regola di Laplace dello sviluppo per righe si ha

$$\begin{aligned} &x_n = a_n x_{n-1} - c_n b_{n-1} x_{n-2} \\ &x_1 = a_1, \\ &x_0 = 1. \end{aligned}$$

I valori $\tilde{x}_i$ effettivamente calcolati verificano la relazione

$$\begin{aligned} &\tilde{x}_n = a_n \tilde{x}_{n-1}(1+\alpha_n)(1+\beta_n) - c_n b_{n-1} \tilde{x}_{n-2}(1+\beta_n)(1+\gamma_n)(1+\delta_n) \\ &\tilde{x}_1 = a_1, \\ &\tilde{x}_0 = 1. \end{aligned}$$

dove $\alpha_n, \beta_n, \gamma_n$ e δ_n sono gli errori locali generati nelle quattro operazioni aritmetiche e sono maggiorati in valore assoluto dalla precisione di macchina u. Per cui, definendo $\tilde{a}_n = a_n(1+\alpha_n)(1+\beta_n)$ e $\tilde{b}_{n-1} = b_{n-1}(1+\delta_n)$, $\tilde{c}_n = c_n(1+\beta_n)(1+\gamma_n)$, si ottiene

$$\begin{aligned}
&\tilde{x}_n = \tilde{a}_n\tilde{x}_{n-1} - \tilde{c}_n\tilde{b}_{n-1}\tilde{x}_{n-1}\\
&\tilde{x}_1 = a_1,\\
&\tilde{x}_0 = 1.
\end{aligned}$$

cioè i valori effettivamente calcolati sono i determinanti delle matrici tridiagonali definiti da $\tilde{a}_i$, $\tilde{b}_i$ e $\tilde{c}_i$. Inoltre, in una analisi al primo ordine, le perturbazioni relative indotte nelle variabili di input a_n, b_{n-1} e c_n sono limitati rispettivamente da $2u, u$ e $2u$. Si ha quindi un algoritmo stabile all'indietro.

7 Esercizi

In questo paragrafo abbiamo raccolto esercizi relativi all'analisi degli errori alcuni dei quali riportano la risoluzione. Spesso useremo il simbolo $\dot{\leq}$ per denotare la diseguaglianza a meno di termini di ordine u^2 o superiore.

Un modo sistematico per trattare l'analisi all'indietro, che conviene usare quando una analisi più diretta diventa difficoltosa, è operare nel seguente modo. Supponiamo per semplicità di dover valutare una funzione di 3 variabili $f(x, y, z)$ denotiamo $\varphi(x, y, z)$ la funzione effettivamente calcolata in aritmetica floating point ed esprimiamola in termini degli errori locali sostituendo ogni operazione aritmetica del tipo $a \,\mathrm{op}\, \mathrm{b}$ con l'operazione di macchina $(a \,\mathrm{op}\, \mathrm{b})(1 + \epsilon)$ dove ϵ è l'errore locale tale che $|\epsilon| < u$. Denotiamo poi $\hat{x} = x(1 + \delta_x)$, $\hat{y} = y(1 + \delta_y)$, $\hat{z} = z(1 + \delta_z)$ i valori perturbati rispettivamente di x, y, z. Per ricavare le tre perturbazioni incognite δ_x, δ_y, δ_z si scrive il sistema lineare ottenuto uguagliando le parti lineari negli errori di $f(\hat{x}, \hat{y}, \hat{z})$ e di $\varphi(x, y, z)$. Se il sistema lineare ha soluzione allora l'analisi all'indietro può essere completata risolvendo il sistema.

Un esempio di questo modo di procedere è dato nell'esercizio 1 che viene affrontato nei due modi descritti.

Per ottenere stime dell'errore algoritmico, dopo aver svolto l'analisi all'indietro, basta esprimere l'errore algoritmico in funzione delle perturbazioni calcolate e dei corrispondenti coefficienti di amplificazione. Questo è mostrato nell'esercizio 22.

Per condurre l'analisi in avanti è possibile rappresentare l'algoritmo in termini del suo grafo associato e applicare ad ogni nodo la proprietà che l'errore presente nella variabile associata al nodo è dato al primo ordine dalla somma dell'errore locale e dell'errore proveniente dagli operandi moltiplicato per i coefficienti di amplificazione.

Un altro modo equivalente di procedere è sostituire ad ogni operazione del tipo $a \,\mathrm{op}\, \mathrm{b}$ l'operazione di macchina $(a \,\mathrm{op}\, \mathrm{b})(1+\epsilon)$, raccogliere la parte che contiene gli errori locali e dividerla per il valore della funzione. Occorre poi maggiorare il valore assoluto di questa quantità usando la diseguaglianza triangolare.

Esercizio 1 Dati numeri di macchina x_1, x_2, x_3, costruire un algoritmo numericamente stabile all'indietro per il calcolo di

$$f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3x_1.$$

Detto $\varphi(x_1, x_2, x_3)$ il risultato fornito dall'algoritmo in aritmetica floating point, si dimostri che esistono δ_1, δ_2, δ_3 tali che $\varphi(x_1, x_2, x_3) = f(\hat{x}_1, \hat{x}_2, \hat{x}_3)$, $\hat{x}_i = x_i(1 + \delta_i)$, per $i = 1, 2, 3$. Dare maggiorazioni a $|\delta_i|$, $i = 1, 2, 3$.

Soluzione. L'algoritmo, che si basa sulla formula

$$f(x_1, x_2, x_3) = (x_1 + x_3)x_2 + (x_1x_3),$$

consiste nell'effettuare le seguenti operazioni:

$$\begin{aligned} s_1 &= x_3 \cdot x_1, \\ s_2 &= x_1 + x_3, \\ s_3 &= s_2 \cdot x_2, \\ s_4 &= s_1 + s_3, \end{aligned}$$

dove $f(x_1, x_2, x_3) = s_4$. Se $\tilde{s}_i$, $i = 1, \ldots, 4$, sono i valori effettivamente calcolati, otteniamo che

$$\begin{aligned} \tilde{s}_1 &= (x_3x_1)(1 + \epsilon_1), \\ \tilde{s}_2 &= (x_1 + x_3)(1 + \epsilon_2), \\ \tilde{s}_3 &= \tilde{s}_2x_2(1 + \epsilon_3), \\ \tilde{s}_4 &= (\tilde{s}_1 + \tilde{s}_3)(1 + \epsilon_4), \end{aligned}$$

dove $|\epsilon_i| < u$, $i = 1, \ldots, 4$ sono gli errori locali generati nelle singole operazioni aritmetiche floating point per il calcolo di s_i. Dunque si ha

$$\varphi(x_1, x_2, x_3) = \big(x_2(x_1 + x_3)(1 + \epsilon_2)(1 + \epsilon_3) + x_3x_1(1 + \epsilon_1)\big)(1 + \epsilon_4) = x_2(x_1 + x_3)(1 + \epsilon_2)(1 + \epsilon_3)(1 + \epsilon_4) + x_3x_1(1 + \epsilon_1)(1 + \epsilon_4).$$

Se usiamo l'approccio sistematico abbiamo che, posto $\hat{x}_1 = x_1(1 + \delta_1)$, $\hat{x}_2 = x_2(1 + \delta_2)$, $\hat{x}_3 = x_3(1 + \delta_3)$, vale

$$f(\hat{x}_1, \hat{x}_2, \hat{x}_3) \doteq x_1x_2(1 + \delta_1 + \delta_2) + x_2x_3(1 + \delta_2 + \delta_3) + x_3x_1(1 + \delta_3 + \delta_1).$$

per cui, dalla relazione ottenuta prendendo le parti lineari negli errori in $\varphi(x_1, x_2, x_3) = f(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ si ottiene il sistema

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} \epsilon_2 + \epsilon_3 + \epsilon_4 \\ \epsilon_2 + \epsilon_3 + \epsilon_4 \\ \epsilon_1 + \epsilon_4 \end{bmatrix}$$

La soluzione del sistema è data da $\delta_3 = (\epsilon_1 + \epsilon_4)/2$, $\delta_2 = \epsilon_2 + \epsilon_3 + (\epsilon_4 - \epsilon_1)/2$, $\delta_1 = (\epsilon_1 + \epsilon_4)/2$.

Invece seguendo un approccio più diretto, poiché

$$(1 + \epsilon_1)(1 + \epsilon_4) \doteq (1 + \epsilon_1 + \epsilon_4) \doteq \left(1 + \frac{\epsilon_1 + \epsilon_4}{2}\right)\left(1 + \frac{\epsilon_1 + \epsilon_4}{2}\right),$$

vale

$$x_3x_1(1+\epsilon_1)(1+\epsilon_4) = \hat{x}_3\hat{x}_1$$

dove $\hat{x}_1 = x_1(1+\delta_1)$, $\hat{x}_3 = x_3(1+\delta_3)$, dove $\delta_1 = \delta_3 \doteq \frac{\epsilon_1+\epsilon_4}{2}$. D'altra parte

$$\begin{array}{l}(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4) \doteq (1+\epsilon_2+\epsilon_3+\epsilon_4) = \\ (1+\epsilon_2+\epsilon_3+\epsilon_4)\left(1+\frac{\epsilon_1+\epsilon_4}{2}\right)^{-1}\left(1+\frac{\epsilon_1+\epsilon_4}{2}\right) \doteq \\ (1+\epsilon_2+\epsilon_3+\epsilon_4)\left(1-\frac{\epsilon_1+\epsilon_4}{2}\right)\left(1+\frac{\epsilon_1+\epsilon_4}{2}\right) \doteq \\ (1-\frac{\epsilon_1}{2}+\epsilon_2+\epsilon_3+\frac{\epsilon_4}{2})\left(1+\frac{\epsilon_1+\epsilon_4}{2}\right).\end{array}$$

Dunque

$$x_2(x_1+x_3)(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4) = \hat{x}_2(\hat{x}_1+\hat{x}_3)$$

dove $\hat{x}_2 = x_2(1+\delta_2)$ e $\delta_2 \doteq -\frac{\epsilon_1}{2}+\epsilon_2+\epsilon_3+\frac{\epsilon_4}{2}$. Quindi $\varphi(x_1,x_2,x_3) = f(\hat{x}_1,\hat{x}_2,\hat{x}_3)$ dove $\hat{x}_i = x_i(1+\delta_i)$ e $|\delta_1| \dot{\leq} \frac{|\epsilon_1|+|\epsilon_4|}{2} < u$, $|\delta_2| \dot{\leq} \frac{|\epsilon_1|}{2}+|\epsilon_2|+|\epsilon_3|+\frac{|\epsilon_4|}{2} < 3u$, $|\delta_3| \dot{\leq} \frac{|\epsilon_1|+|\epsilon_4|}{2} < u$. □

Esercizio 2 Dati numeri di macchina x_1, x_2, costruire un algoritmo che calcoli

$$f(x_1,x_2) = x_1^2/x_2 + x_2/x_1$$

con tre operazioni aritmetiche. Si provi la stabilità all'indietro dimostrando che esistono δ_1, δ_2 tali che $\varphi(x_1,x_2) = f(\tilde{x}_1,\tilde{x}_2)$, $\tilde{x}_i = x_i(1+\delta_i)$, per $i = 1,2$, dove $\varphi(x_1,x_2)$ è il risultato fornito dall'algoritmo in aritmetica floating point. Dare maggiorazioni a $|\delta_i|$, $i = 1,2$.

Soluzione. L'algoritmo si basa sulla formula

$$f(x_1,x_2) = x_1/(x_2/x_1) + (x_2/x_1)$$

ed è dato da

$$\begin{aligned} s_1 &= x_2/x_1, \\ s_2 &= x_1/s_1, \\ s_3 &= s_2+s_1, \end{aligned}$$

dove $s_3 = f(x_1,x_2)$. Operando in aritmetica floating point si ha

$$\begin{aligned} \tilde{s}_1 &= (1+\epsilon)x_2/x_1, \\ \tilde{s}_2 &= (1+\theta)x_1/\tilde{s}_1, \\ \tilde{s}_3 &= (\tilde{s}_2+\tilde{s}_1)(1+\eta). \end{aligned}$$

Dove abbiamo indicato con ϵ, θ, η gli errori locali generati dalle singole operazioni aritmetiche. Risulta quindi

$$\tilde{s}_3 = \frac{(1+\eta)(1+\theta)x_1}{(x_2/x_1)(1+\epsilon)} + (x_2/x_1)(1+\epsilon)(1+\eta)$$

Vale allora

$$\tilde{s}_3 = \tilde{x}_1/(\tilde{x}_2/\tilde{x}_1) + (\tilde{x}_2/\tilde{x}_1)$$

dove si è posto $\tilde{x}_1 = x_1(1+\eta)^2(1+\theta)$, $\tilde{x}_2 = x_2(1+\eta)^3(1+\theta)(1+\epsilon)$. Per cui, essendo gli errori locali maggiorati in modulo dalla precisione di macchina u risulta $|\delta_1| \dot{\leq} 3u$, $|\delta_2| \dot{\leq} 5u$. □

Esercizio 3 Dati numeri di macchina $x_1, x_2 \neq 0$ costruire un algoritmo che calcoli

$$f(x_1, x_2) = x_1^2/x_2 + x_2^2/x_1$$

con 4 operazioni aritmetiche. Svolgere l'analisi in avanti dell'errore e dare maggiorazioni al valore assoluto dell'errore algoritmico. Si studi in particolare il caso in cui $x_1 x_2 > 0$.

Soluzione. L'algoritmo si basa sulla formula

$$f(x_1, x_2) = x_1(x_1/x_2) + x_2/(x_1/x_2)$$

e consiste nell'effettuare le seguenti operazioni:

$$\begin{aligned} s_1 &= x_1/x_2, \\ s_2 &= x_1 s_1, \\ s_3 &= x_2/s_1, \\ s_4 &= s_2 + s_3. \end{aligned}$$

Indichiamo con ϵ_i l'errore algoritmico per il calcolo di s_i e δ_i l'errore locale dovuto alla operazione aritmetica svolta nel calcolo di s_i. Vale allora $\epsilon_1 = \delta_1$ con $|\delta_1| < u$.

Vale inoltre

$$\begin{aligned} \epsilon_2 &\doteq \delta_2 + \epsilon_1, \\ \epsilon_3 &\doteq \delta_3 - \epsilon_1, \\ \epsilon_4 &\doteq \delta_4 + \tfrac{s_2}{s_2+s_3}\epsilon_2 + \tfrac{s_3}{s_2+s_3}\epsilon_3, \end{aligned}$$

con $|\delta_i| < u$, $i = 2, 3, 4$. Dunque

$$\epsilon_4 \doteq \delta_4 + \frac{x_1^2/x_2}{x_1^2/x_2 + x_2^2/x_1}(\delta_1 + \delta_2) + \frac{x_2^2/x_1}{x_1^2/x_2 + x_2^2/x_1}(\delta_3 - \delta_1)$$

da cui

$$|\epsilon_4| \dot{\leq} u\left(1 + 2\frac{|x_1^2/x_2|}{|x_1^2/x_2 + x_2^2/x_1|} + 2\frac{|x_2^2/x_1|}{|x_1^2/x_2 + x_2^2/x_1|}\right).$$

Nel caso $x_1 x_2 > 0$ si ottiene $|\epsilon_4| \dot{\leq} 3u$.

Esercizio 4 Sia $f(x, y) = xy + (x^2)/y = x(y + x/y)$. Si effettui l'analisi in avanti dei due algoritmi per il calcolo di $f(x, y)$ definiti dalle espressioni precedenti. Si confrontino le prestazioni dei due metodi in termini di complessità e di stabilità numerica. Si valuti anche l'errore totale.

Soluzione. In una analisi al primo ordine l'errore totale è, la somma dell'errore inerente e di quello algoritmico. L'errore inerente è dato da $\epsilon_{in} \doteq c_1\sigma_1 + c_2\sigma_2$ dove $|\sigma_i| < u$ per $i = 1, 2$ e $c_1 = \frac{x}{f(x,y)}\frac{\partial f}{\partial x} = \frac{2x+y^2}{y^2+x}$, $c_2 = \frac{y}{f(x,y)}\frac{\partial f}{\partial y} = \frac{y^2-x}{y^2+x}$. Dunque

$$\epsilon_{in} \dot{\leq} u\left(\frac{|2x+y^2|}{|y^2+x|} + \frac{|y^2-x|}{|y^2+x|}\right).$$

Per l'analisi dell'errore algoritmico supponiamo che x e y siano numeri di macchina.

Il primo algoritmo consiste nell'effettuare le seguenti operazioni:

$$\begin{aligned} s_1 &= xy, \\ s_2 &= x^2, \\ s_3 &= s_2/y, \\ s_4 &= s_1 + s_3. \end{aligned}$$

Il numero di operazioni è 4. Indichiamo con ϵ_i l'errore algoritmico per il calcolo di s_i e δ_i l'errore locale generato nel calcolo di s_i. Vale $\epsilon_1 = \delta_1$ e $\epsilon_2 = \delta_2$, con $|\delta_i| < u$ per $i = 1, 2$; inoltre

$$\begin{aligned} \epsilon_3 &\doteq \delta_3 + \epsilon_2 \\ \epsilon_4 &\doteq \delta_4 + \tfrac{s_1}{s_1+s_3}\epsilon_1 + \tfrac{s_3}{s_1+s_3}\epsilon_3 \end{aligned}$$

con $|\delta_i| < u$, $i = 3, 4$. Dunque, svolgendo i conti,

$$\epsilon_4 \doteq \delta_4 + \frac{y^2}{x+y^2}\delta_1 + \frac{x}{x+y^2}(\delta_3 + \delta_2),$$

da cui

$$|\epsilon_4| \dot{\leq} u\left(1 + \frac{y^2+2|x|}{|x+y^2|}\right).$$

Nel caso $x > 0$ si ottiene $|\epsilon_4| \dot{\leq} 3u$.

Il secondo algoritmo consiste nell'effettuare le seguenti operazioni:

$$\begin{aligned} s_1 &= x/y, \\ s_2 &= s_1 + y, \\ s_3 &= xs_2. \end{aligned}$$

Il numero di operazioni è 3. Indichiamo con ϵ_i l'errore algoritmico per il calcolo di s_i e con δ_i l'errore locale generato dall'operazione aritmetica svolta nel calcolare s_i. Vale dunque $\epsilon_1 = \delta_1$ con $|\delta_1| < u$; inoltre

$$\begin{aligned} \epsilon_2 &\doteq \delta_2 + \tfrac{s_1}{s_1+y}\epsilon_1 \\ \epsilon_3 &\doteq \delta_3 + \epsilon_2 \end{aligned}$$

con $|\delta_i| < u$, $i = 2, 3$.

Dunque, svolgendo i conti,

$$\epsilon_3 \doteq \delta_3 + \delta_2 + \frac{x}{x+y^2}\delta_1,$$

da cui

$$|\epsilon_3| \stackrel{\cdot}{\leq} u\left(2+\frac{|x|}{|x+y^2|}\right).$$

Nel caso $x>0$ si ottiene $|\epsilon_3| \stackrel{\cdot}{\leq} 3u$. □

Esercizio 5 Sia $g(x,a,b)=ax+b/x$ e si indichino con $\tilde{g}_1,\ \tilde{g}_2$ i valori ottenuti calcolando $g(x,a,b)$ con l'algoritmo

$$\begin{aligned} s_1 &= a\cdot x,\\ s_2 &= b/x,\\ g_1 &= s_1+s_2, \end{aligned}$$

e con l'algoritmo

$$\begin{aligned} t_1 &= x\cdot x,\\ t_2 &= a\cdot t_1,\\ t_3 &= t_2+b,\\ g_2 &= t_3/x, \end{aligned}$$

Dati $a,b,x\in\mathcal{F}$ dire se esistono valori $\hat{a},\hat{b},\tilde{a},\tilde{b}\in\mathbb{R}$ tali che $\tilde{g}_1=g(x,\hat{a},\hat{b})$, $\tilde{g}_2=g(x,\tilde{a},\tilde{b})$. Nel caso stimare il valore assoluto degli errori relativi $\delta_a,\delta_b,\epsilon_a,\epsilon_b$ tali che $\hat{a}=a(1+\delta_a)$, $\hat{b}=b(1+\delta_b)$, $\tilde{a}=a(1+\epsilon_a)$, $\tilde{b}=b(1+\epsilon_b)$.

Soluzione. Con il primo algoritmo, se $\tilde{s}_1$ e $\tilde{s}_2$ sono i valori effettivamente calcolati, allora $\tilde{s}_1=(ax)(1+\epsilon_1)$ e $\tilde{s}_2=(b/x)(1+\epsilon_2)$ con $|\epsilon_1|,|\epsilon_2|<u$ errori locali. Dunque il valore della funzione effettivamente calcolato sarà

$$\tilde{g}_1=(\tilde{s}_1+\tilde{s}_2)(1+\epsilon_3)=(ax)(1+\epsilon_1)(1+\epsilon_3)+(b/x)(1+\epsilon_2)(1+\epsilon_3),$$

con $|\epsilon_3|<u$. Quindi $\tilde{g}_1=g(x,\hat{a},\hat{b})$ dove $\hat{a}=a(1+\delta_a)$, $\hat{b}=b(1+\delta_b)$, e $|\delta_a|\doteq|\epsilon_1+\epsilon_3|<2u$, $|\delta_b|\doteq|\epsilon_2+\epsilon_3|<2u$.

Con il secondo algoritmo, se $\tilde{t}_i$ sono i valori effettivamente calcolati, allora $\tilde{t}_1=x^2(1+\epsilon_1)$, $\tilde{t}_2=(a\tilde{t}_1)(1+\epsilon_2)$, $\tilde{t}_3=(\tilde{t}_2+b)(1+\epsilon_3)$ dove gli errori locali sono tali che $|\epsilon_1|,|\epsilon_2|,|\epsilon_3|<u$. Dunque il valore della funzione effettivamente calcolato sarà

$$\tilde{g}_2=(\tilde{t}_3/x)(1+\epsilon_4)=(ax(1+\epsilon_1)(1+\epsilon_2)+b/x)(1+\epsilon_3)(1+\epsilon_4),$$

con $|\epsilon_4|<u$. Quindi $\tilde{g}_2=g(x,\hat{a},\hat{b})$ dove $\hat{a}=a(1+\delta_a)$, $\hat{b}=b(1+\delta_b)$, e $|\delta_a|\doteq|\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4|<4u$, $|\delta_b|\doteq|\epsilon_3+\epsilon_4|<2u$. □

Esercizio 6 Si descriva un algoritmo per il calcolo di $\sum_{i=0}^{2^k-1}x^i$, dati x e k, basato sulla seguente identità

$$\sum_{i=0}^{2^k-1}x^i=(1+x)(1+x^2)(1+x^{2^2})\cdots(1+x^{2^{k-1}})$$

che impieghi al più $3k$ operazioni aritmetiche. Si scriva una function nella sintassi di Octave che lo implementa. Si dia una maggiorazione al primo ordine del valore assoluto dell'errore algoritmico nel caso in cui $0 < x < 1$.

Esercizio 7 Per calcolare la funzione $f(x) = x^3 - 1$, per valori di $x \in \mathcal{F}$, si consideri l'algoritmo $\mathcal{A}_1$ che calcola nell'ordine

$$\begin{aligned} s_1 &= x \cdot x, \\ s_2 &= x \cdot s_1, \\ s_3 &= s_2 - 1, \end{aligned}$$

dove $f(x) = s_3$, e l'algoritmo $\mathcal{A}_2$ che si basa sull'identità $f(x) = (x-1)((x+1)^2 - x)$ e che calcola nell'ordine

$$\begin{aligned} t_1 &= x + 1, \\ t_2 &= t_1 \cdot t_1, \\ t_3 &= t_2 - x, \\ t_4 &= x - 1, \\ t_5 &= t_3 \cdot t_4, \end{aligned}$$

dove $f(x) = t_5$.

a) Mediante un'analisi in avanti si diano maggiorazioni al primo ordine α_1 e α_2 ai valori assoluti degli errori algoritmici generati rispettivamente da $\mathcal{A}_1$ e $\mathcal{A}_2$.

b) Si dimostri che α_2 è limitato superiormente da una costante e che α_1 può assumere valori arbitrariamente grandi.

Soluzione. Consideriamo l'algoritmo $\mathcal{A}_1$. Indichiamo con ϵ_i l'errore algoritmico per il calcolo di s_i e δ_i l'errore locale dovuto all'operazione aritmetica svolta nel calcolo di s_i. Vale allora

$$\begin{aligned} &\epsilon_1 = \delta_1, \\ &\epsilon_2 \doteq \delta_2 + \epsilon_1, \\ &\epsilon_3 \doteq \delta_3 + \tfrac{s_2}{s_3}\epsilon_2, \end{aligned}$$

con $|\delta_i| < u$, $i = 1, 2, 3$, da cui otteniamo

$$\epsilon_3 \doteq \delta_3 + \frac{x^3}{x^3 - 1}(\delta_1 + \delta_2).$$

Passando ai moduli abbiamo

$$|\epsilon_3| \stackrel{\cdot}{\leq} \left(1 + 2\frac{|x^3|}{|x^3 - 1|}\right) u = \alpha_1(x)u,$$

quindi $\alpha_1(x)$ può essere arbitrariamente grande quando x si avvicina a 1.

Consideriamo ora l'algoritmo $\mathcal{A}_2$. Indichiamo con τ_i l'errore algoritmico per il calcolo di t_i e σ_i l'errore locale dovuto alla operazione aritmetica svolta nel calcolo di t_i. Vale allora

$$\begin{aligned}
&\tau_1 = \sigma_1,\\
&\tau_2 \doteq \sigma_2 + \tau_1 + \tau_1,\\
&\tau_3 \doteq \sigma_3 + \tfrac{t_2}{t_3}\tau_2,\\
&\tau_4 = \sigma_4,\\
&\tau_5 \doteq \sigma_5 + \tau_3 + \tau_4,
\end{aligned}$$

con $|\sigma_i| < u$, $i = 1, \ldots, 5$, da cui otteniamo

$$\tau_5 \doteq \sigma_3 + \sigma_4 + \sigma_5 + \frac{(x+1)^2}{x^2+x+1}(2\sigma_1 + \sigma_2).$$

Passando ai moduli abbiamo

$$|\tau_5| \stackrel{\cdot}{\leq} 3\left(1 + \frac{(x+1)^2}{x^2+x+1}\right)u = \alpha_2(x)u,$$

e la funzione α_2 è limitata superiormente da una costante.

Osservazione: Se x fosse un numero di macchina, per cui l'errore inerente sarebbe nullo, ci sarebbe un chiaro vantaggio del secondo algoritmo sul primo poiché l'errore algoritmico ha una maggiorazione migliore rispetto al primo. Se peró x non fosse un numero di macchina, allora non ci sarebbe vantaggio di un algoritmo rispetto all'altro. Infatti l'errore inerente sarebbe $\epsilon_{in} = \epsilon\frac{x(3x^2)}{x^3-1}$, dove ϵ è l'errore di rappresentazione. Dunque l'errore inerente è arbitrariamente grande quando x si avvicina a 1 e dominerebbe comunque sull'errore algoritmico.

Esercizio 8 Per calcolare l'espressione $f(a,b,c,d) = (a^2+bc)/(c+d)$ si consideri l'algoritmo definito da:

$$\begin{aligned}
s_1 &= a \cdot a,\\
s_2 &= b \cdot c,\\
s_3 &= c + d,\\
s_4 &= s_1 + s_2,\\
s_5 &= s_4/s_3,
\end{aligned}$$

dove $f(a,b,c,d) = s_5$. Si provi che tale algoritmo è numericamente stabile all'indietro nei punti in cui $f(a,b,c,d)$ è definita, dimostrando che il valore $\varphi(a,b,c,d)$ effettivamente calcolato in aritmetica floating point è uguale a $f(\tilde{a},\tilde{b},\tilde{c},\tilde{d})$ per opportuni valori $\tilde{a},\tilde{b},\tilde{c},\tilde{d}$. Si diano limitazioni superiori alle perturbazioni $|\tilde{a}-a|, |\tilde{b}-b|, |\tilde{c}-c|, |\tilde{d}-d|$.

Soluzione. Indichiamo con $\tilde{s}_i$ i valori effettivamente calcolati di s_i. Allora

$$\begin{aligned}
\tilde{s}_1 &= a \cdot a(1+\epsilon_1),\\
\tilde{s}_2 &= b \cdot c(1+\epsilon_2),\\
\tilde{s}_3 &= (c+d)(1+\epsilon_3),\\
\tilde{s}_4 &= (\tilde{s}_1 + \tilde{s}_2)(1+\epsilon_4),\\
s_5 &= \tilde{s}_4/\tilde{s}_3(1+\epsilon_5),
\end{aligned}$$

con $|\epsilon_i| < u$. Allora, trascurando i termini quadratici negli ϵ_i, vale

$$\tilde{s}_5 \doteq \frac{a^2(1+\epsilon_1+\epsilon_4+\epsilon_5)+bc(1+\epsilon_2+\epsilon_4+\epsilon_5)}{(c+d)(1+\epsilon_3)} \doteq \frac{\hat{a}^2-\hat{b}\hat{c}}{\hat{c}\hat{d}}$$

dove

$$\begin{aligned}
&\hat{a} = a(1+\delta_a), \quad \delta_a = (\epsilon_1+\epsilon_4+\epsilon_5)/2,\\
&\hat{b} = b(1+\delta_b), \quad \delta_b = \epsilon_2+\epsilon_4+\epsilon_5-\epsilon_3,\\
&\hat{c} = c(1+\delta_c), \quad \delta_c = \epsilon_3,\\
&\hat{d} = d(1+\delta_d), \quad \delta_d = \epsilon_3.
\end{aligned}$$

Esercizio 9 Per calcolare la funzione $f(a,b) = a^2+b^2/a$ si consideri l'algoritmo che svolge i seguenti passi:

$$\begin{aligned}
s_1 &= a \cdot a,\\
s_2 &= b \cdot b,\\
s_3 &= s_2/a,\\
s_4 &= s_1 + s_3.
\end{aligned}$$

Si dimostri la stabilità all'indietro dell'algoritmo e si diano maggiorazioni al primo ordine per $|\epsilon_a|, |\epsilon_b|$ tali che $\varphi(a,b) = f(a(1+\epsilon_a), b(1+\epsilon_b))$, dove $\varphi(a,b)$ è il valore ottenuto eseguendo l'algoritmo in aritmetica floating point con numeri di macchina a, b. Si ricavi una maggiorazione al primo ordine del valore assoluto dell'errore algoritmico nel caso in cui $2a^3 > b^2 > 0$.

Esercizio 10 Si descriva un algoritmo per calcolare l'espressione

$$f(a,b,c) = \frac{ac+bc}{a-b}$$

che sia stabile all'indietro e si diano maggiorazioni ai valori assoluti delle perturbazioni $\delta_a, \delta_b, \delta_c$ per cui $\varphi(a,b,c) = f(a(1+\delta_a), b(1+\delta_b), c(1+\delta_c))$, dove $\varphi(a,b,c)$ è il valore calcolato eseguendo l'algoritmo in aritmetica floating point con numeri di macchina a, b, c, $a \neq b$.

Soluzione. Si usa l'espressione $f(a,b,c) = c(a+b)/(a-b)$. Calcolando in aritmetica floating point e denotando con ϵ_i, $i = 1,2,3,4$ rispettivamente gli errori locali generati nell'addizione, sottrazione, divisione e moltiplicazione si ha

$$\varphi(a,b,c) = (c(a+b)/(a-b))(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4) \doteq f(a,b,c)(1+\epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4).$$

Ponendo quindi $\hat{a} = a$, $\hat{b} = b$, $\hat{c} = c(1+\delta_c)$, con $\delta_c = \epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4$, risulta $\varphi(a,b,c) \doteq f(\hat{a},\hat{b},\hat{c})$. Vale inoltre $|\delta_c| < 4u$. Si osservi inoltre che, per definizione, indipendentemente dall'analisi all'indietro svolta, risulta $\epsilon_{\text{alg}} := \epsilon_1+\epsilon_2+\epsilon_3+\epsilon_4$.

Esercizio 11 Si consideri la successione $\{x_k\}$ definita da

$$x_{k+1} = a_k(x_k^2) + x_{k-1}/b_k, \quad k = 1, 2, \ldots,$$

dove $x_0 = x_1 = 1$ e a_k, b_k sono numeri di macchina. Siano $\tilde{x}_k$ i valori ottenuti applicando la formula in aritmetica floating point con precisione u. Si dimostri che esistono ϵ_k, δ_k, $k = 1, 2, \ldots$, tali che

$$\tilde{x}_{k+1} = \hat{a}_k(\tilde{x}_k^2) + \tilde{x}_{k-1}/\hat{b}_k, \quad k = 1, 2, \ldots,$$

con $\tilde{x}_0 = \tilde{x}_1 = 1$, $\hat{a}_k = a_k(1 + \delta_a^{(k)})$, $\hat{b}_k = b_k(1 + \delta_b^{(k)})$. Si diano maggiorazioni al primo ordine in funzione di u a $|\delta_a^{(k)}|$ e $|\delta_b^{(k)}|$.

Soluzione. Si calcola il passo k-esimo nel modo seguente

$$\begin{aligned} s_1 &= x_k \cdot x_k \\ s_2 &= a_k \cdot s_1 \\ s_3 &= x_{k-1}/b_k \\ x_{k+1} &= s_2 + s_3 \end{aligned}$$

Operando in aritmetica floating point e denotando $\tilde{x}_k$ e $\tilde{x}_{k-1}$ le quantità calcolate in aritmetica floating point nei passi precedenti, si ottiene

$$\begin{aligned} \tilde{s}_1 &= \tilde{x}_k \cdot \tilde{x}_k(1 + \epsilon_1^{(k)}) \\ \tilde{s}_2 &= a_k \cdot \tilde{s}_1(1 + \epsilon_2^{(k)}) \\ \tilde{s}_3 &= (\tilde{x}_{k-1}/b_k)(1 + \epsilon_3^{(k)}) \\ \tilde{x}_{k+1} &= (\tilde{s}_2 + \tilde{s}_3)(1 + \epsilon_4^{(k)}) \end{aligned}$$

dove $\epsilon_i^{(k)}$, $i = 1, 2, 3, 4$ sono gli errori locali generati nelle corrispondenti operazioni aritmetiche. Si ottiene quindi

$$\tilde{x}_{k+1} = \left[a_k\tilde{x}_k^2(1 + \epsilon_1^{(k)})(1 + \epsilon_2^{(k)}) + (\tilde{x}_{k-1}/b_k)(1 + \epsilon_3^{(k)})\right](1 + \epsilon_4^{(k)})$$

Ponendo quindi $\hat{a}_k = a_k(1 + \epsilon_1^{(k)})(1 + \epsilon_2^{(k)})(1 + \epsilon_4^{(k)}) =: a_k(1 + \delta_a^{(k)})$ e $\hat{b}_k = b_k/((1 + \epsilon_3^{(k)})(1 + \epsilon_4^{(k)})) =: b_k(1 + \delta_b^{(k)})$, risulta

$$\tilde{x}_{k+1} = \hat{a}_k\tilde{x}_k^2 + \tilde{x}_{k-1}/\hat{b}_k$$

inoltre $\delta_a^{(k)} \doteq \epsilon_1^{(k)} + \epsilon_2^{(k)} + \epsilon_4^{(k)}$, $\delta_b^{(k)} \doteq -\epsilon_3^{(k)} - \epsilon_4^{(k)}$, per cui $|\delta_a^{(k)}| \dot{\leq} 3u$, $|\delta_b^{(k)}| \dot{\leq} 2u$. □

Esercizio 12 Per calcolare la funzione $f(x_1, ..., x_n) = \sum_{k=1}^{n} \prod_{j=1}^{k} x_j$ si consideri l'algoritmo seguente

$$s_1 = x_1, \quad s_i = (1 + s_{i-1})x_i, \quad i = 2, \ldots, n.$$

a) Si dimostri che l'esecuzione dell'algoritmo in aritmetica floating point genera dei numeri di macchina $\tilde{s}_i$ tali che $\tilde{s}_i = (1 + \tilde{s}_{i-1})\tilde{x}_i$, per $i = 2, \ldots, n$ dove $\tilde{x}_i = x_i(1 + \epsilon_i)$, $|\epsilon_i| \dot{\leq} 2u$, $\tilde{s}_1 = s_1 = x_1$ e u è la precisione dell'aritmetica.

b) Nell'ipotesi che i dati x_i siano compresi tra 0 e 1, si maggiorino i coefficienti di amplificazione di $f(x_1, ..., x_n)$ e si usi il risultato del punto a) per dare una maggiorazione al primo ordine dell'errore algoritmico.

Soluzione. a) Dimostriamo la proprietà per induzione. Se $n = 2$, vale

$$\tilde{s}_2 = ((1 + s_1)(1 + \alpha_1))x_2(1 + \alpha_2)$$

dove α_1 e α_2 sono gli errori locali generati dalle singole operazioni aritmetiche, $|\alpha_1| < u$, $|\alpha_2| < u$. Dunque $\tilde{s}_2 = (1 + \tilde{s}_1)\tilde{x}_2$, dove $\tilde{x}_2 = x_2(1 + \epsilon_2)$ e $|\epsilon_2| \doteq |\alpha_1 + \alpha_2| < 2u$ e $\tilde{s}_1 = s_1 = x_1$.

Supponiamo che sia $\tilde{s}_i = (1 + \tilde{s}_{i-1})\tilde{x}_i$, con $\tilde{x}_i = x_i(1 + \epsilon_i)$ e $|\epsilon_i| \dot{\leq} 2u$, per $i = 2, \ldots, n - 1$. Allora

$$\tilde{s}_n = ((1 + \tilde{s}_{n-1})(1 + \alpha_{1,n}))x_n(1 + \alpha_{2,n})$$

dove $\alpha_{1,n}$ e $\alpha_{2,n}$ sono gli errori locali generati dalle singole operazioni aritmetiche, $|\alpha_{1,n}| < u$, $|\alpha_{2,n}| < u$. Da cui $\tilde{s}_n = (1 + \tilde{s}_{n-1})\tilde{x}_n$, dove $\tilde{x}_n = x_n(1 + \epsilon_n)$ e $|\epsilon_n| \doteq |\alpha_{1,n} + \alpha_{2,n}| < 2u$.

b) Poiché il valore effettivamente calcolato è $f(\tilde{x}_1, \ldots, \tilde{x}_n)$, l'errore algoritmico è dato da $\epsilon_{\text{alg}} \doteq \sum_{i=1}^{n} c_i\epsilon_i$ dove $c_i = \frac{x_i}{f(x_1,\ldots,x_n)}\frac{\partial f}{\partial x_i}$. Dalla definizione di $f(x_1, \ldots, x_n)$ segue che, poiché $x_i > 0$ per $i = 1, \ldots, n$, $0 < \frac{\frac{\partial f}{\partial x_i}}{f(x_1,\ldots,x_n)} < 1$ per ogni i. Dunque, poiché $0 < x_i < 1$ e $|\epsilon_i| \dot{\leq} 2u$, si ottiene $\epsilon_{\text{alg}} \dot{\leq} 2nu$. □

Esercizio 13 Dati un intero $n > 1$ e tre vettori $a = (a_i)$, $c = (c_i) \in \mathbb{R}^n$, $b = (b_i) \in \mathbb{R}^{n-1}$ si consideri il sistema $Tx = c$ dove $T = (t_{i,j})$ è la matrice triangolare inferiore tale che $t_{i,i} = a_i$, per $i = 1, \ldots, n$, $t_{i+1,i} = b_i$, $i = 1, \ldots, n-1$, e $t_{i,j} = 0$ altrove. Descrivere un algoritmo per la risoluzione del sistema che impieghi al più $3n$ operazioni aritmetiche. Fare l'analisi all'indietro dell'errore e scrivere una function nella sintassi di Octave che implementi l'algoritmo.

Esercizio 14 Siano $u, v \in \mathbb{R}^n$ con $u_1 = v_1$ e si definisca la matrice $n \times n$ $A_n = (a_{i,j})$ tale che $a_{i,i+1} = 1$, $i = 1, \ldots, n-1$, $a_{i,i} = u_i$, $a_{i,1} = v_i$, $i = 1, \ldots, n$. Si denoti $d_n = \det A_n$.

a) Scrivere una relazione che lega d_n con d_{n-1} e ricavarne un algoritmo per il calcolo di d_n che impieghi non più di $2n$ operazioni aritmetiche.

b) Scrivere una function con la sintassi di Octave che implementi l'algoritmo del punto a).

c) Dimostrare la stabilità all'indietro dell'algoritmo.

Soluzione. a) Sviluppando il determinante della matrice A_n, con $n \geq 2$, rispetto all'ultima riga, si ottiene che $d_n = u_n d_{n-1} + (-1)^{n+1} v_n$. Dunque d_n può essere calcolato mediante l'algoritmo

$$\begin{aligned} &d_1 = u_1 \\ &d_i = u_i d_{i-1} + (-1)^{i+1} v_i, \quad i = 2, \ldots, n \end{aligned}$$

che impiega $2(n-1)$ operazioni aritmetiche.

c) Sia $\tilde{d}_i$ il valore effettivamente calcolato di d_i. Vale

$$\begin{aligned} &\tilde{d}_1 = d_1 = u_1 \\ &\tilde{d}_i = \big((u_i \tilde{d}_{i-1})(1+\alpha_i) + (-1)^{i+1} v_i\big)(1+\beta_i), \quad i = 2, \ldots, n \end{aligned}$$

dove α_i, β_i, con $|\alpha_i|, |\beta_i| < u$, sono gli errori locali. Dunque

$$\tilde{d}_i = \tilde{u}_i \tilde{d}_{i-1} + (-1)^{i+1} \tilde{v}_i, \quad i = 2, \ldots, n$$

dove $\tilde{u}_i = u_i(1+\alpha_i)(1+\beta_i)$, $\tilde{v}_i = v_i(1+\alpha_i)$. Dunque il valore effettivamente calcolato è il determinante della matrice definita di vettori $\tilde{u}$, $\tilde{v}$, con elementi $\tilde{u}_i$ e $\tilde{v}_i$, rispettivamente. □

Esercizio 15 Sono dati tre vettori $b = (b_i), u = (u_i), v = (v_i) \in \mathbb{R}^n$, con $b_1 = u_1 = v_1$. Si consideri la matrice $A_n = (a_{i,j})$ di dimensione $n \times n$ con $a_{i,i} = b_i$, $i = 1, \ldots, n$, $a_{1,i} = u_i$, $a_{i,1} = v_i$, $i = 2, \ldots, n$. Si ponga $d_n = \det A_n$.

a) Si scriva la relazione che lega d_n e d_{n-1}

b) Si implementi tale relazione in una function nella sintassi di Octave che prende come input i tre vettori b, u, v e dà in output il vettore $d = (d_i) \in \mathbb{R}^n$

c) Si dica se tale formula è stabile all'indietro.

Esercizio 16 Scrivere una function nella sintassi di Octave che data una matrice $n \times n$, reale A con elementi diagonali nulli, calcoli la somma dei determinanti di tutte le sottomatrici principali 2×2 di A. Svolgere un'analisi dell'errore dell'algoritmo implementato dando una limitazione superiore all'errore algoritmico nell'ipotesi che A abbia elementi compresi tra 0 e 1.

Esercizio 17 La funzione $f(x) = \frac{1}{x(1-x)}$ può essere scritta come $f(x) = \frac{1}{x} - \frac{1}{1-x}$. Si analizzi l'errore algoritmico nel calcolo di $f(x)$ con i metodi ottenuti dalle due diverse rappresentazioni. Dire quale dei due metodi è numericamente più stabile.

Esercizio 18 Dato un intero $n > 0$ e assegnati i numeri reali positivi a_i, b_i, c_i per $i = 0, 1, \ldots, n$, si definisca $x_{i+1} = a_i + c_i b_i / x_i$, per $i = 0, 1, \ldots n$, dove $x_0 > 0$ è assegnato. Siano $\tilde{x}_i$ i valori ottenuti calcolando gli x_i con aritmetica *floating point* con precisione u. Dire se esistono perturbazioni $\alpha_i, \beta_i, \gamma_i$ tali che posto $\tilde{a}_i = a_i(1+\alpha_i)$, $\tilde{b}_i = b_i(1+\beta_i)$, $\tilde{c}_i = c_i(1+\gamma_i)$, risulta $\tilde{x}_{i+1} = \tilde{a}_i + \tilde{c}_i \tilde{b}_i / \tilde{x}_i$; in tal caso si diano maggiorazioni a $|\alpha_i|, |\beta_i|, |\gamma_i|$ in funzione di u.
Svolgere una analisi analoga nel caso del calcolo di $x_{i+1} = a_i + c_i b_i / (a_i x_i)$.

Esercizio 19 Sia $n > 0$ intero e sia $s_n = s_n(a_1, \ldots, a_n, b_1, \ldots, b_n)$ tale che

$$s_k = a_k s_{k-1} + b_k / s_{k-1} + b_k, \quad k = 1, \ldots, n, \quad s_0 = 1.$$

Determinare un algoritmo stabile all'indietro per il calcolo di s_n. Denotando $\widetilde{s}_k$ i valori calcolati dall'algoritmo in aritmetica *floating point*, dare maggiorazioni a $|\alpha_k|$ e $|\beta_k|$ per cui $\widetilde{s}_n = s_n(\widetilde{a}_1, \ldots, \widetilde{a}_n, \widetilde{b}_1, \ldots, \widetilde{b}_n)$, dove $\widetilde{a}_k = a_k(1 + \alpha_k)$, $\widetilde{b}_k = b_k(1 + \beta_k)$.

Esercizio 20 Siano $f(x), g(x) : \mathbb{R} \to \mathbb{R}$ funzioni razionali e $\varphi(\xi)$, $\gamma(\eta)$ i risultati forniti da due algoritmi per il loro calcolo applicati in aritmetica *floating point* con precisione u a partire da numeri di macchina ξ, η. Si assuma che in assenza di *overflow* o *underflow*, esistano $\delta_1, \delta_2 \in \mathbb{R}$ tali che $\varphi(\xi) = f(\xi(1 + \delta_1))$, $\gamma(\eta) = g(\eta(1 + \delta_2))$, dove $|\delta_1| \leq u\theta$ e $|\delta_1| \leq u\theta$, $\theta > 0$. Dimostrare che esiste $\epsilon \in \mathbb{R}$ tale che $\gamma(\varphi(\xi)) = g(f(\xi(1 + \epsilon)))$, in assenza di *overflow* o *underflow*, e dare maggiorazioni a $|\epsilon|$ al primo ordine in funzione di u.

Esercizio 21 Descrivere un algoritmo numericamente stabile all'indietro per il calcolo di $f(x, y, z) = x/(yz) + y/(xz)$ e farne l'analisi all'indietro dell'errore. Dire se l'algoritmo trovato può generare errori di cancellazione numerica per $z > 0$.

Soluzione. Vale $f(x, y, z) = (x/y + y/x)/z$ per cui si può usare il seguente algoritmo

$$\begin{aligned} s_1 &= x/y \\ s_2 &= y/x \\ s_3 &= s_1 + s_2 \\ f &= s_3/z \end{aligned}$$

Eseguendo l'algoritmo in aritmetica floating point si ha

$$\begin{aligned} \tilde{s}_1 &= (x/y)(1 + \epsilon_1) \\ \tilde{s}_2 &= (y/x)(1 + \epsilon_2) \\ \tilde{s}_3 &= (\tilde{s}_1 + \tilde{s}_2)(1 + \epsilon_3) \\ \tilde{f} &= (\tilde{s}_3/z)(1 + \epsilon_4) \end{aligned}$$

dove ϵ_i, $i = 1, 2, 3, 4$ sono gli errori locali generati nelle corrispondenti operazioni aritmetiche. Si ottiene allora

$$\tilde{f} = \frac{x}{yz}(1+\epsilon_1)(1+\epsilon_3)(1+\epsilon_4) + \frac{y}{xz}(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4) \doteq \frac{x}{yz}(1+\epsilon_1+\epsilon_3+\epsilon_4) + \frac{y}{xz}(1+\epsilon_2+\epsilon_3+\epsilon_4).$$

Cerchiamo ora perturbazioni $\delta_x, \delta_y, \delta_z$ tali che i valori $\hat{x} = x(1 + \delta_x)$, $\hat{y} = y(1 + \delta_y)$, $\hat{z} = z(1 + \delta_z)$, soddisfino la condizione $\tilde{f} = f(\hat{x}, \hat{y}, \hat{z})$. Imponendo questa condizione sulla parte lineare degli errori si ottiene

$$\begin{aligned} \delta_x - \delta_y - \delta_z &= \epsilon_1 + \epsilon_3 + \epsilon_4 \\ \delta_y - \delta_x - \delta_z &= \epsilon_2 + \epsilon_3 + \epsilon_4. \end{aligned}$$

Sommando e sottraendo le due espressioni si ottiene il sistema equivalente

$$\begin{aligned}\delta_z &= -(\epsilon_3 + \epsilon_4 + (\epsilon_1 + \epsilon_2)/2)\\ \delta_x - \delta_y &= \epsilon_1/2 - \epsilon_2/2\end{aligned}$$

che è risolto da $\delta_x = \epsilon_1/2$, $\delta_y = \epsilon_2/2$, $\delta_z = -(\epsilon_3 + \epsilon_4 + (\epsilon_1 + \epsilon_2)/2)$, per cui $|\delta_x| < u/2$, $|\delta_y| < u/2$, $|\delta_z| < 3u$. □

Esercizio 22 Descrivere un algoritmo numericamente stabile all'indietro per il calcolo di $f(x, y) = x^2/y + x$ e farne l'analisi all'indietro dell'errore. Utilizzare tale analisi per determinare limitazioni superiori al valore assoluto dell'errore algoritmico nel caso $xy > 0$.

Soluzione. Usando l'espressione $f(x, y) = x(1 + x/y)$ si può calcolare f col seguente algoritmo

$$\begin{aligned}s_1 &= x/y,\\ s_2 &= s_1 + 1,\\ f &= x \cdot s_2.\end{aligned}$$

Operando in aritmetica floating point si ottiene

$$\begin{aligned}\tilde{s}_1 &= (x/y)(1 + \epsilon_1),\\ \tilde{s}_2 &= (\tilde{s}_1 + 1)(1 + \epsilon_2),\\ \tilde{f} &= x \cdot \tilde{s}_2(1 + \epsilon_3),\end{aligned}$$

dove $\epsilon_1, \epsilon_2, \epsilon_3$ sono gli errori locali generati dalle tre operazioni aritmetiche. Vale quindi

$$\tilde{f} = x(1 + (x/y)(1 + \epsilon_1))(1 + \epsilon_3)(1 + \epsilon_2)$$

Per cui, posto $\hat{x} = x(1 + \epsilon_3)(1 + \epsilon_2)$ $\hat{y} = y(1 + \epsilon_3)(1 + \epsilon_2)/(1 + \epsilon_1)$ vale

$$\tilde{f} = f(\hat{x}, \hat{y}).$$

Inoltre vale $\hat{x} = x(1 + \delta_1)$, $\hat{y} = y(1 + \delta_2)$, $\delta_1 \doteq \epsilon_2 + \epsilon_3$, $\delta_2 \doteq \epsilon_3 - \epsilon_1 - \epsilon_2$, per cui $|\delta_1| \dot{\leq} 2u$, $|\delta_2| \dot{\leq} 3u$.

Per quanto riguarda l'errore algoritmico vale

$$\epsilon_{\text{alg}} \doteq C_x\delta_1 + C_y\delta_2$$

dove $C_x = (2x^2/y + x)/(x^2/y + x)$, $C_y = -(x^2/y)/(x^2/y + x)$ sono i coefficienti di amplificazione di $f(x, y)$. Risulta allora

$$|\epsilon_{\text{alg}}| \dot{\leq} u(4|x^2/y| + 2|x| + 3|x^2/y|)/|x^2/y + x|$$

Se $xy > 0$ si ottiene

$$|\epsilon_{\text{alg}}| \dot{\leq} u|7x^2/y + 2x|/|x^2/y + x| = u(7 - 5/(x/y + 1)) < 7u.$$

Un'altra possibilità è usare l'algoritmo

$$\begin{aligned} s_1 &= x \cdot x, \\ s_2 &= s_1/y, \\ f &= s_2 + x. \end{aligned}$$

Procedendo in modo analogo, il valore effettivamente calcolato in aritmetica floating point è dato da

$$\tilde{f} = \left[(x^2/y)(1+\epsilon_1)(1+\epsilon_2) + x\right](1+\epsilon_3)$$

Ponendo allora $\hat{x} = x(1+\epsilon_3)$ e $\hat{y} = y(1+\epsilon_3)/((1+\epsilon_1)(1+\epsilon_2))$ si ha $\tilde{f} = f(\hat{x}, \hat{y})$. Vale quindi $\hat{x} = x(1+\delta_1)$, $\hat{y} = y(1+\delta_2)$, $|\delta_1| < u$, $|\delta_2| \stackrel{\cdot}{\leq} 3u$. La limitazione all'errore algoritmico ottenuta in questo modo è quindi migliore e, se $xy > 0$ vale

$$|\epsilon_{\text{alg}}| \stackrel{\cdot}{\leq} u(2|x^2/y| + |x| + 3|x^2/y|)/|x^2/y + x| \leq u(1 + 4|x^2/y|/|x^2/y + x|) < 5u$$

□

Esercizio 23 Descrivere un algoritmo per il calcolo di

$$c + \sum_{i=1,n} a_i/(b_i - x)$$

numericamente stabile all'indietro e farne l'analisi all'indietro dell'errore.

Esercizio 24 Siano a, b, c tre numeri di macchina. Per il calcolo di $ab - bc$ si considerino le seguenti espressioni

$$b(a-c), \quad ab - bc, \quad a(b+c) - c(a+b).$$

Dare limitazioni superiori al valore assoluto dell'errore algoritmico nei tre casi e confrontare.

Esercizio 25 Si considerino le funzioni $f_k = f_k(x_1, \ldots, x_k, y_1, \ldots, y_k)$, $g_k = g_k(x_1, \ldots, x_k, y_1, \ldots, y_k)$, definite da

$$\begin{aligned} f_{k+1} &= x_{k+1}\, f_k\, g_k, \\ g_{k+1} &= (f_k^2 - g_k^2)/y_{k+1} \end{aligned} \qquad k = 0, 1, \ldots, n-1, \tag{7}$$

dove $f_0 = g_0 = 1$. Dare un algoritmo per il calcolo della coppia (f_n, g_n) numericamente stabile all'indietro e farne l'analisi all'indietro dell'errore.

Soluzione. È sufficiente riscrivere la seconda delle due espressioni come $g_{k+1} = (f_k^2 - g_k^2)/y_{k+1} = (f_k - g_k)(f_k + g_k)/y_{k+1}$. Il valore calcolato in aritmetica floating point è

$$\tilde{g}_{k+1} = [(f_k - g_k)(f_k + g_k)/y_{k+1}]\,(1+\epsilon_k)(1+\mu_k)(1+\eta_k),$$

dove $\epsilon_k, \mu_k, \eta_k, \nu_k$ sono gli errori locali generati rispettivamente dalla sottrazione, addizione moltiplicazione e divisione. Si ha quindi

$$\tilde{g}_{k+1} = (f_k - g_k)(f_k + g_k)/\hat{y}_{k+1}$$

con $\hat{y}_{k+1} = y_{k+1}/((1+\epsilon_k)(1+\mu_k)(1+\eta_k)(1+\nu_k)) = y_k(1+\delta_k)$, $\delta_k \doteq -\epsilon_k - \mu_k - \eta_k - \nu_k$ per cui $|\delta_k| \dot{\leq} 4u$. □

Esercizio 26 Dare un algoritmo stabile all'indietro per il calcolo di $f(a,b,c,x) = -ax^{-1} + b + ax + cx^2$ e farne l'analisi all'indietro dell'errore.

Soluzione. Vale $f(a,b,c,x) = a(x-1)(x+1)/x + b + cx^2$ da cui si ricava l'algoritmo sintetizzato dalla formula

$$f(a,b,c,x) = ((cx^2 + b) + a(x-1)(x+1)/x)$$

che applicato in aritmetica floating point produce

$$\begin{aligned}\tilde{f} =& cx^2(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4)\\ &+b(1+\epsilon_3)(1+\epsilon_4)\\ &+a(x-1)(x+1)/x(1+\epsilon_4)(1+\theta_1)(1+\theta_2)(1+\theta_3)(1+\theta_4)(1+\theta_5),\end{aligned}$$

dove $\epsilon_1, \ldots, \epsilon_4$ sono gli errori locali commessi nell'esecuzione delle tre operazioni aritmetiche nel calcolo di $cx^2 + b$ e nell'operazione di addizione della quantità così ottenuta con l'addendo successivo. Mentre $\theta_1, \ldots, \theta_5$ sono gli errori locali commessi nell'esecuzione delle 5 operazioni aritmetiche nel calcolo del secondo addendo $a(x-1)(x+1)/x$. L'analisi all'indietro si completa ponendo

$$\begin{aligned}\hat{c} &= c(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3)(1+\epsilon_4) =: c(1+\delta_c),\\ \hat{b} &= b(1+\epsilon_3)(1+\epsilon_4) =: b(1+\delta_b),\\ \hat{a} &= a(1+\epsilon_4)(1+\theta_1)(1+\theta_2)(1+\theta_3)(1+\theta_4)(1+\theta_5)(1+\epsilon_4) =: a(1+\delta_a)\end{aligned}$$

Vale quindi $|\delta_a| \dot{\leq} 7u$, $|\delta_b| \dot{\leq} 2u$, $|\delta_c| \dot{\leq} 4u$. □

Esercizio 27 Si considerino le successioni $\{x_k\}$ e $\{y_k\}$ definite da

$$x_{k+1} = (x_k + y_k)/2, \qquad y_{k+1} = \sqrt{x_k y_k}$$

a partire da $x_0, y_0 > 0$. Si dia una maggiorazione a $\theta_k = \max(|\xi_k|, |\eta_k|)$ dove ξ_k e η_k sono gli errori algoritmici nel calcolo rispettivamente di x_k e y_k in aritmetica *floating point* con precisione u, dove la radice quadrata viene calcolata da una funzione "di macchina" SQRT tale che $\mathrm{SQRT}(w) = \sqrt{w}(1+\epsilon)$, $|\epsilon| < u$. Se $u = 2^{-52}$ qual è il massimo valore di n per cui ci sono almeno 32 bit corretti nei valori effettivamente calcolati di x_n e y_n?

Esercizio 28 Descrivere un algoritmo stabile all'indietro per il calcolo di $f(x,y,z) = (xy+yz+zx)/(xyz)$ e farne l'analisi all'indietro dell'errore.

Soluzione Vale $f(x,y,z) = 1/z + 1/x + 1/y$ per cui un algoritmo di calcolo è dato da:

$$\begin{aligned} s_1 &= 1/z \\ s_2 &= 1/x \\ s_3 &= 1/y \\ s_4 &= s_1 + s_2 \\ f &= s_4 + s_3 \end{aligned}$$

Una sua esecuzione in aritmetica floating point genera quantità date da

$$\begin{aligned} \tilde{s}_1 &= (1/z)(1+\epsilon_1) \\ \tilde{s}_2 &= (1/x)(1+\epsilon_2) \\ \tilde{s}_3 &= (1/y)(1+\epsilon_3) \\ \tilde{s}_4 &= (\tilde{s}_1 + \tilde{s}_2)(1+\epsilon_4) \\ \tilde{f} &= (\tilde{s}_4 + \tilde{s}_3)(1+\epsilon_5) \end{aligned}$$

dove ϵ_i, $i = 1, \ldots, 5$ sono gli errori locali generati rispettivamente nelle corrispondenti operazioni aritmetiche. Vale quindi

$$\tilde{f} = (1+\epsilon_5)(1+\epsilon_4)(1+\epsilon_1)\frac{1}{z} + (1+\epsilon_5)(1+\epsilon_4)(1+\epsilon_2)\frac{1}{x} + (1+\epsilon_5)(1+\epsilon_3)\frac{1}{y}$$

da cui $\tilde{f} = f(\hat{x}, \hat{y}, \hat{z})$ con $\hat{x} = x/((1+\epsilon_5)(1+\epsilon_4)(1+\epsilon_2)) =: x(1+\delta_1)$, $\hat{y} = y/((1+\epsilon_5)(1+\epsilon_3)) =: y(1+\delta_2)$ $\hat{z} = z/((1+\epsilon_5)(1+\epsilon_4)(1+\epsilon_1)) =: z(1+\delta_3)$, dove $\delta_1 \doteq -\epsilon_5 - \epsilon_4 - \epsilon_2$, $\delta_2 \doteq -\epsilon_5 - \epsilon_3$, $\delta_3 \doteq -\epsilon_5 - \epsilon_4 - \epsilon_1$. Per cui $|\delta_1| \dot{\leq} 3u$, $|\delta_2| \dot{\leq} 2u$, $|\delta_3| \dot{\leq} 3u$. □

Esercizio 29 Dare limitazioni superiori al valore assoluto degli errori algoritmici commessi nel calcolo di $f(x) = x^2 + x + 1$ mediante le due espressioni

$$f(x) = x(x+1) + 1 = (x+1)^2 - x$$

in aritmetica *floating point* con precisione di macchina u. Confrontare le due limitazioni così ottenute per $x > 0$.

Soluzione. Detto $\tilde{f}$ il valore effettivamente calcolato nel primo algoritmo si ha

$$\tilde{f} = (x(x+1)(1+\nu)(1+\eta) + 1)(1+\mu)$$

dove ν, η, μ sono gli errori locali generati rispettivamente dalla prima addizione, dalla moltiplicazione e dalla seconda addizione. Per cui l'errore algoritmico diventa

$$\epsilon_{alg_1} \doteq \frac{x(x+1)(\eta+\nu+\mu)+\mu}{x^2+x+1}$$

da cui $|\epsilon_{alg_1|} \dot{\leq} u\frac{3|x(x+1)|+1}{|x^2+x+1|}$

Detto $\hat{f}$ il valore effettivamente calcolato nel secondo algoritmo si ha

$$\tilde{f} = ((x+1)^2(1+\mu)^2(1+\nu)-x)(1+\eta)$$

dove μ è l'errore locale della prima addizzione, ν è l'errore locale dell'elevamento a quadrato, mentre η è l'errore locale della sottrazione. Per cui l'errore algoritmico diventa

$$\epsilon_{alg_2} \doteq \frac{(x+1)^2(\eta+\nu+2\mu)+\eta x}{x^2+x+1}$$

da cui $|\epsilon_{alg_2}| \dot{\leq} u\frac{4(x+1)^2+|x|}{|x^2+x+1|}$.

Il primo algoritmo è più conveniente del secondo se $3|x(x+1)|+1 < 4(x+1)^2+|x|$, Cioè se $x < -3$ oppure $x > -3/7$.

L'analisi dell'errore algoritmico poteva essere svolta equivalentementw costruendo il grafo di calcolo associato ai due algoritmi. □

Esercizio 30 Si considerino le funzioni $f_k = f_k(x_1, \ldots, x_k, y_1, \ldots, y_k)$, $g_k = g_k(x_1, \ldots, x_k, y_1, \ldots, y_k)$, definite da

$$\begin{aligned} f_{k+1} &= f_k\, g_k/x_{k+1}, \\ g_{k+1} &= y_{k+1}\,(f_k^2+g_k^2-2g_k\, f_k), \end{aligned} \qquad k = 0, 1, \ldots, n-1, \tag{8}$$

dove $f_0 = g_0 = 1$. Dare un algoritmo numericamente stabile all'indietro per il calcolo della coppia (f_n, g_n) e farne l'analisi all'indietro dell'errore.

Soluzione. I valori effettivamente calcolati mediante le espressioni $f_{k+1} = (f_k g_k)/x_{k+1}$ e $g_{k+1} = y_{k+1}(f_k - g_k)^2$, generano numeri di macchina $\tilde{f}_k$ e $\tilde{g}_k$ tali che

$$\begin{aligned} \tilde{f}_{k+1} &= \tilde{f}_k\tilde{g}_k(1+\alpha_k)(1+\beta_k)/x_{k+1} \\ \tilde{g}_{k+1} &= y_{k+1}(\tilde{f}_k-\tilde{g}_k)^2(1+\gamma_k)(1+\eta_k)(1+\nu_k) \end{aligned}$$

dove α_k e β_k sono gli errori locali generati dalla moltiplicazione e dalla divisione nel calcolo di $\tilde{f}_{k+1}$, mentre γ_k, η_k, ν_k sono gli errori locali generati rispettivamente dalla addizione, l'elevamento a quadrato e dalla moltiplicazione per y_{k+1}. Per cui i valori assoluti di questi errori locali sono maggiorati dalla precisione di macchina u. Quindi, ponendo $\hat{x}_{k+1} = x_{k+1}/((1+\alpha_k)(1+\beta_k))$ e $\hat{y}_{k+1} = y_{k+1}(1+\gamma_k)(1+\eta_k)(1+\nu_k)$ si ottiene la relazione

$$\begin{aligned} \tilde{f}_{k+1} &= \tilde{f}_k\tilde{g}_k/\hat{x}_{k+1} \\ \tilde{g}_{k+1} &= \hat{y}_{k+1}(\tilde{f}_k-\tilde{g}_k)^2 \end{aligned}$$

che dimostra la stabilità all'indietro dell'algoritmo. Inoltre vale $\hat{x}_{k+1} \doteq x_{k+1}(1-\alpha_k-\beta_k) =: x_{k+1}(1+\delta_k)$, $|\delta_k| < 2u$, $\hat{y}_{k+1} \doteq y_{k+1}(1+\gamma_k)(1+\eta_k)(1+\nu_k) =: y_{k+1}(1+\delta'_k)$, $|\delta'_k| < 3u$. □

Esercizio 31 Dato un intero pari $n = 2m > 0$ e numeri reali a_i, f_i, $i = 1, \ldots, n$, b_i, $i = 1, \ldots, n-1$, si consideri il sistema lineare $Hx = f$, dove $H = (h_{i,j})$ è la matrice $n \times n$ tale che gli elementi non nulli sono tutti e soli gli $h_{i,i} = a_i$, per $i = 1, \ldots, n$, gli $h_{2i-1,2i} = b_{2i-1}$, per $i = 1, \ldots, m$ e gli $h_{2i+1,2i} = b_{2i}$, per $i = 1, m-1$.

a) Si descriva un algoritmo per risolvere il sistema $Hx = f$ che impieghi non più di $3n$ operazioni aritmetiche, e se ne faccia l'analisi all'indietro dell'errore.

Esercizio 32 Dato un intero $n > 2$ e i vettori $u = (u_i), f = (f_i) \in \mathbb{R}^n$, $v = (v_i) \in \mathbb{R}^{n-1}$, sia $A = (a_{i,j})$ la matrice bidiagonale inferiore con elementi diagonali $a_{i,i} = u_i$, $i = 1, \ldots, n$ ed elementi sottodiagonali $a_{i+1,i} = v_i$, $i = 1, \ldots, n-1$.

Si descriva un algoritmo per risolvere il sistema $Ax = f$ che impieghi non più di $3n$ operazioni aritmetiche e si svolga una analisi all'indietro dell'errore.

Esercizio 33 Dato un intero $n > 2$ e i vettori $u = (u_i), f = (f_i) \in \mathbb{R}^n$, $v = (v_i) \in \mathbb{R}^{n-1}$, sia $A = (a_{i,j})$ la matrice definita da $a_{i,i} = u_i$, $i = 1, \ldots, n$, $a_{i,1} = v_{i-1}$, $i = 2, \ldots, n$, $a_{i,j} = 0$ altrove. Descrivere un algoritmo per risolvere il sistema $Ax = f$ che impieghi non più di $3n$ operazioni aritmetiche, studiarne la stabilità all'indietro.

Esercizio 34 Si consideri la funzione $f(m,n) = \sum_{i=m}^{n} \frac{(-1)^i}{i}$ se $m \leq n$, mentre $f(m,n) = 0$ se $m > n$. Descrivere un algoritmo per il calcolo di $f(m,n)$, per $m, n > 0$, che abbia un errore algoritmico limitato superiormente da $\gamma(n-m)u$ dove γ è una costante positiva e u è la precisione di macchina.

Esercizio 35 Dati un intero $n > 1$ e tre vettori $a = (a_i)$, $c = (c_i) \in \mathbb{R}^n$, $b = (b_i) \in \mathbb{R}^{n-1}$ si consideri il sistema $Tx = c$ dove $T = (t_{i,j})$ è la matrice triangolare inferiore tale che $t_{i,i} = a_i$, per $i = 1, \ldots, n$, $t_{i+1,i} = b_i$, $i = 1, \ldots, n-1$, e $t_{i,j} = 0$ altrove. Descrivere un algoritmo per la risoluzione del sistema che impieghi al più $3n$ operazioni aritmetiche. Fare l'analisi all'indietro dell'errore.

Esercizio 36 Siano $u, v \in \mathbb{R}^n$ con $u_1 = v_1$ e si definisca la matrice $n \times n$ $A_n = (a_{i,j})$ tale che $a_{i,i+1} = 1$, $i = 1, \ldots, n-1$, $a_{i,i} = u_i$, $a_{i,1} = v_i$, $i = 1, \ldots, n$. Si denoti $d_n = \det A_n$.

a) Scrivere una relazione che lega d_n con d_{n-1} e ricavarne un algoritmo per il calcolo di d_n che impieghi non più di $2n$ operazioni aritmetiche.

b) Scrivere una function con la sintassi di Octave che implementi l'algoritmo del punto a).

c) Dimostrare la stabilità all'indietro dell'algoritmo.

Esercizio 37 Sono dati tre vettori $b = (b_i), u = (u_i), v = (v_i) \in \mathbb{R}^n$, con $b_1 = u_1 = v_1$. Si consideri la matrice $A_n = (a_{i,j})$ di dimensione $n \times n$ con

$a_{i,i} = b_i$, $i = 1, \ldots, n$, $a_{1,i} = u_i$, $a_{i,1} = v_i$, $i = 2, \ldots, n$. Si ponga $d_n = \det A_n$.
a) Si scriva la relazione che lega d_n e d_{n-1}
b) Si implementi tale relazione in una function nella sintassi di Octave che prende come input i tre vettori b, u, v e dà in output il vettore $d = (d_i) \in \mathbb{R}^n$.
c) Si dica se tale formula è stabile all'indietro.

Esercizio 38 Siano $a = (a_i), x = (x_i), y = (y_i) \in \mathbb{R}^n$, $b = (b_i), c = (c_i) \in \mathbb{R}^{n-1}$, A la matrice tridiagonale $n \times n$ con elementi diagonali a_i, per $i = 1, \ldots, n$, sottodiagonali e sopradiagonali rispettivamente b_i, c_i, $i = 1, \ldots, n-1$, tali che $y = Ax$.

a) Scrivere le formule che legano le componenti di y a quelle di x.

b) Dimostrare che il calcolo di y dati x, a, b, c mediante queste formule è numericamente stabile all'indietro.

c) Scrivere una function nella sintassi di octave che presi in input i vettori x, a, b, c fornisce in output il vettore y.

Esercizio 39 Dati i vettori $d = (d_i), u = (u_i), v = (v_i) \in \mathbb{R}^n$ si definisca la matrice $n \times n$ diagonale D con elementi diagonali $d_1, \ldots, d_n$ e si ponga $A = D + uu^T$. Si descriva un algoritmo per il calcolo di $v^T Av$ che impieghi al più $5n$ operazioni aritmetiche. Farne l'analisi all'indietro dell'errore e dare una *function* nella sintassi di Octave che lo implementi.

Riferimenti bibliografici

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

[2] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia 2002.

I teoremi di Gerschgorin

Dario A. Bini, Beatrice Meini
Università di Pisa

10 ottobre 2019

Sommario

Questo modulo didattico contiene risultati relativi ai teoremi di Gerschgorin che permettono di localizzare nel piano complesso gli autovalori di una matrice.

1 Introduzione

In certe situazioni è utile disporre di criteri facilmente applicabili che forniscano localizzazioni nel campo complesso degli autovalori di una matrice assegnata. Un esempio significativo a questo riguardo è dato dallo studio della stabilità di un sistema dinamico governato da un sistema di equazioni differenziali del tipo

$$\begin{cases} y'(t) = Ay(t), & t > 0 \\ y(0) = y_0 \end{cases}$$

dove $y(t)$ è una funzione da $\mathbb{R}$ a valori in $\mathbb{R}^n$ derivabile con continuità ed A è una matrice $n \times n$. Si può dimostrare che, se la matrice A ha autovalori con parte reale minore o uguale a zero, tutte le soluzioni di questo problema, ottenute al variare delle condizioni iniziali $y(0) = y_0$, sono limitate per ogni valore di t. Cioè, come si usa dire, il sistema dinamico rappresentato dal sistema di equazioni differenziali è stabile.

Dim. La soluzione si può scrivere come

$$y(t) = \exp(tA)y_0$$

dove si definisce

$$\exp(tA) = I + tA + \frac{t^2A^2}{2!} + \frac{t^3A^3}{3!} + \cdots,$$

e la serie è convergente per ogni valore di t. Consideriamo per semplicità il caso in cui la matrice A sia diagonalizzabile, cioè esiste S non singolare tale che $A = SDS^{-1}$, dove D è diagonale; allora $\exp(tA) = S\exp(tD)S^{-1}$ per cui gli autovalori di tA sono $e^{t\alpha_i}$, con α_i autovalori di A. Se il generico autovalore α lo scriviamo come $\alpha = \beta + \mathtt{i}\gamma$, con $\beta, \gamma \in \mathbb{R}$ dove $\mathtt{i}$ è l'unità immaginaria,

allora $e^{t\alpha} = e^{t\beta}e^{\mathtt{i}t\gamma} = e^{t\beta}(\cos(t\gamma) + \mathtt{i}\sin(t\gamma))$. Mentre il secondo fattore è limitato avendo modulo 1, il primo fattore $e^{t\beta}$ è limitato se e solo se $\beta \leq 0$. La dimostrazione nel caso generale può essere fatta considerando la forma normale di Jordan di A e trattando separatamente il caso di un singolo blocco di Jordan. □

Per valutare la stabilità di un sistema dinamico basta allora controllare che le parti reali degli autovalori siano minori o uguali a zero. In altre situazioni occorre controllare che gli autovalori di una matrice abbiano tutti modulo minore o uguale a 1. In altri casi, quando ad esempio si deve verificare se una matrice reale simmetrica è definita positiva, ci basta controllare che i suoi autovalori siano tutti positivi.

Certamente una possibilità per accertarsi che queste condizioni siano verificate consiste nel calcolare tutti gli autovalori e controllare se sono valide le proprietà richieste. Ma ciò comporta un costo troppo elevato, infatti calcolare gli autovalori di una matrice ha un costo computazionale non trascurabile.

I teoremi di Gerschgorin forniscono una valida alternativa al calcolo di tutti gli autovalori di una matrice A assegnata. Infatti essi permettono, con minimo sforzo computazionale, di determinare un insieme di dischi nel piano complesso la cui unione contiene tutti gli autovalori di A.

2 I teoremi di Gerschgorin

Nel seguito $A = (a_{i,j})$ è una matrice $n \times n$ ad elementi reali o complessi.

Teorema 1 (Primo teorema di Gerschgorin) *Gli autovalori di A appartengono all'insieme*

$$\cup_{i=1}^{n} K_i, \quad K_i = \{z \in \mathbb{C} : \ |z - a_{i,i}| \leq \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|\}.$$

Inoltre se $v = (v_i)$ è un autovettore corrispondente all'autovalore λ di A, cioè $Av = \lambda v$, allora $\lambda \in K_h$ dove h è tale che $|v_h| = \max_i |v_i|$. In altri termini un autovalore λ appartiene a quei cerchi che corrispondono alle componenti di modulo massimo di un autovettore corrispondente.

Dim. Siano λ e v rispettivamente autovalore e autovettore di A, cioè $Av = \lambda v$ e $v \neq 0$. Leggendo questa relazione nell'i-esima componente si ottiene:

$$\sum_{j=1}^{n} a_{i,j}v_j = \lambda v_i$$

da cui

$$(a_{i,i} - \lambda)v_i = -\sum_{j=1,\ j\neq i}^{n} a_{i,j}v_j.$$

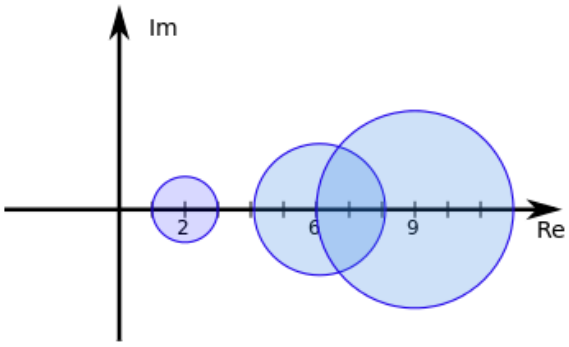


Figura 1: Cerchi di Gerschgorin della matrice A

Prendendo i moduli di entrambi i membri e usando la diseguaglianza triangolare si ha

$$|a_{i,i} - \lambda| \cdot |v_i| \leq \sum_{j=1,\ j\neq i}^{n} |a_{i,j}| \cdot |v_j|.$$

Scegliendo l'indice $i = h$ per cui $|v_h| \geq |v_j|$ per $j = 1, \ldots, n$, si ottiene $|v_h| \neq 0$ e, dividendo per $|v_h|$ si ha

$$|a_{h,h} - \lambda| \leq \sum_{j=1,\ j\neq h}^{n} |a_{h,j}| \frac{|v_j|}{|v_h|} \leq \sum_{j=1,\ j\neq h}^{n} |a_{h,j}|,$$

dove l'ultima diseguaglianza segue dal fatto che $|v_j|/|v_h| \leq 1$. Si conclude allora che $\lambda \in K_h$. □

Gli insiemi K_i per $i = 1, \ldots, n$ sono cerchi nel piano complesso detti cerchi di Gerschgorin di A. Ciascuno di essi ha per centro l'elemento diagonale corrispondente e per raggio la somma dei moduli degli elementi non diagonali che stanno sulla stessa riga.

Come esempio di applicazione si consideri la matrice

$$A = \begin{bmatrix} 2 & 1 & 0 \\ -1 & 6 & -1 \\ 1 & -2 & 9 \end{bmatrix}$$

i cui cerchi hanno centro e raggio rispettivamente c_i e r_i per $i = 1, 2, 3$, dove $(c_1, r_1) = (2, 1)$, $(c_2, r_2) = (6, 2)$, $(c_3, r_3) = (9, 3)$. Il primo teorema di Gerschgorin ci permette di dire subito che questa matrice ha tutti autovalori con parte reale positiva e di modulo maggiore o uguale a 1. Ciò risulta evidente dalla figura 2 dove sono riportati in forma grafica i cerchi di Gerschgorin di A nel piano complesso.

Si osservi che poiché le matrici A e A^T hanno gli stessi autovalori, applicando il primo teorema di Gerschgorin ad A^T si ottiene che gli autovalori di A

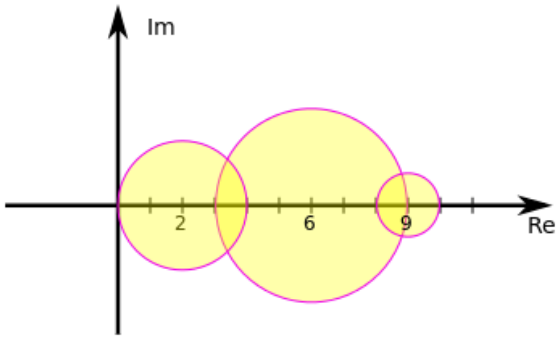


Figura 2: Cerchi di Gerschgorin della matrice A^T

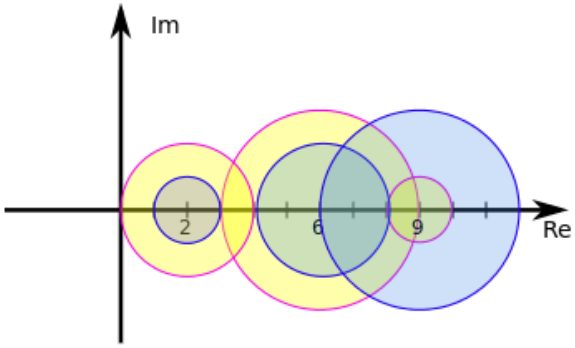


Figura 3: Cerchi di Gerschgorin delle matrici A e A^T

appartengono all'unione

$$\cup_{i=1}^{n} H_i, \quad H_i = \{z \in \mathbb{C} : \ |z - a_{i,i}| \leq \sum_{j=1,\ j\neq i}^{n} |a_{j,i}|\}.$$

La figura 2 riporta graficamente i cerchi di Gerschgorin della matrice A^T.

Possiamo quindi dire che gli autovalori di A appartengono a

$$(\cup_{i=1}^{n} K_i) \cap (\cup_{i=1}^{n} H_i)\,.$$

Attenzione che questa intersezione di unioni non coincide con l'unione delle intersezioni dei singoli cerchi. Nelle due figure 3 e 4 si riportano i cerchi di Gerschgorin relativi alle matrici A e A^T e l'intersezione dell'unione dei cerchi relativi rispettivamente ad A e ad A^T.

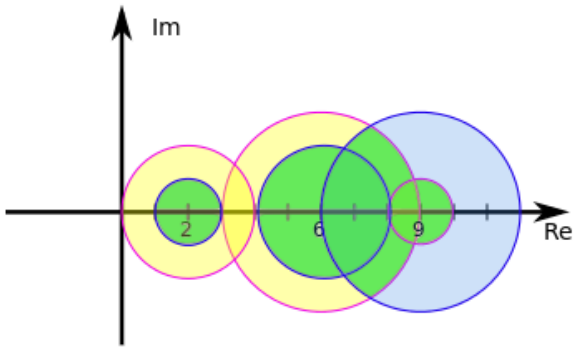


Figura 4: Intersezione in verde dell'unione dei cerchi di Gerschgorin di A e dell'unione di quelli di A^T

In generale non è vero che ogni disco contiene un solo autovalore. Però sotto opportune condizioni si può dire molto di più in questo senso.

Teorema 2 (Secondo teorema di Gerschgorin) *Si assuma che l'unione dei cerchi di Gerschgorin sia formata da due sottoinsiemi disgiunti M_1, e M_2, cioè $\cup_i K_i = M_1 \cup M_2$, $M_1 \cap M_2 = \emptyset$, dove M_1 è costituito da n_1 cerchi, e M_2 è costituito da n_2 cerchi con $n_1 + n_2 = n$. Allora in M_1 sono contenuti n_1 autovalori e in M_2 sono contenuti n_2 autovalori.*

Dim. Supponiamo per semplicità che M_1 sia costituita dai primi n_1 cerchi. La dimostrazione di questo teorema si basa su un ragionamento "per continuità". Infatti si considera una matrice dipendente da un parametro $A(t) = D+t(A-D)$ dove D è la matrice diagonale che ha elementi diagonali uguali a quelli di A. è evidente che $A(0) = D$ e $A(1) = A$. Cioè $A(t)$ descrive i punti del segmento nello spazio delle matrici che unisce D con A. Dimostriamo prima che gli autovalori di $A(t)$ dipendono in modo continuo da t. Per fare questo diamo per buono un risultato classico sui polinomi che afferma che gli zeri di un polinomio sono funzioni continue dei coefficienti. Poiché gli autovalori di una matrice sono gli zeri del polinomio caratteristico $\det(A(t) - \lambda I)$, e poiché i coefficienti del polinomio caratteristico di una matrice dipendono in modo polinomiale, e quindi continuo, dagli elementi di una matrice, possiamo affermare che gli autovalori di $A(t)$ come zeri del polinomio caratteristico dipendono in modo continuo da t. Ora osserviamo che i cerchi di Gerschgorin $K_i(t)$ di $A(t)$ hanno centro $a_{i,i}$ e raggio tr_i, dove r_i è il raggio del cerchio di Gerschgorin K_i. Essi sono quindi contenuti dentro i K_i per ogni $0 \leq t \leq 1$. Quindi i primi n_1 cerchi della matrice $A(t)$ saranno disgiunti dai rimanenti cerchi per ogni $t \in [0, 1]$. Non è quindi possibile che al variare di t in $[0, 1]$ qualche autovalore passi dall'insieme costituito dai primi n_1 cerchi all'insieme costituito dai rimanenti n_2 cerchi. Cioè

il numero di autovalori di $A(t)$ contenuti in M_1 rimane costante. Per $t = 0$ la matrice A coincide con D che è diagonale, e i suoi autovalori coincidono con i centri dei cerchi. Quindi M_1 contiene n_1 autovalori, tanti quanti sono i centri dei cerchi che lo costituiscono. □

In particolare se K_i è un cerchio disgiunto da tutti gli altri allora K_i contiene un solo autovalore. Se i cerchi sono a due a due disgiunti allora ciascuno di essi contiene un solo autovalore.

Una conseguenza utile di questo risultato è la seguente. Se la matrice A è reale e se K_i è un cerchio disgiunto dagli altri allora K_i contiene un autovalore reale. Infatti, poiché gli autovalori non reali di una matrice reale compaiono a coppie complesse coniugate, se $\lambda \in K_i$ non fosse reale allora anche il coniugato $\bar{\lambda}$ sarebbe autovalore e apparterrebbe ancora a K_i. è infatti il simmetrico di λ rispetto all'asse reale. Quindi K_i conterrebbe λ e $\bar{\lambda}$, cioè due autovalori che contraddirebbe il secondo teorema di Gerschgorin.

Una applicazione di questa proprietà alla matrice A dell'esempio mostrato sopra, ci dice che A ha un autovalore reale nell'intervallo $[1, 3]$. Infatti il primo cerchio di Gerschgorin è disgiunto dagli altri.

Definiamo una matrice $A = (a_{i,j})$ *fortemente dominante diagonale* se $|a_{i,i}| > \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|$ per $i = 1, \ldots, n$. È chiaro per il primo teorema di Gerschgorin che una matrice fortemente dominante diagonale è non singolare. Infatti, poichè il raggio di ogni cerchio di Gerschgorin è strettamente minore della distanza del centro dall'origine del piano complesso, nessun cerchio di Gerschgorin interseca l'origine e quindi zero non può essere autovalore di A.

Si consideri la seguente matrice tridiagonale

$$A = \begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & 2 & -1 \\ 0 & & -1 & 2 \end{bmatrix} \tag{1}$$

Si può verificare che tutti i cerchi di Gerschgorin, dal secondo al penultimo hanno centro 2 e raggio 2, mentre il primo e l'ultimo hanno centro 2 e raggio 1. Ci si può chiedere se zero può essere autovalore di A. Il primo teorema di Gerschgorin non ci può aiutare in questo senso, infatti zero appartiene all'unione dei cerchi di Gerschgorin. Più precisamente appartiene alla frontiera dell'unione dei cerchi.

Il terzo teorema di Gerschgorin permette di dimostrare che in una situazione come quella dell'esempio non è possibile che zero sia un autovalore della matrice A. L'ipotesi aggiuntiva che occorre mettere è che A sia una matrice *irriducibile*. Ricordiamo che una matrice A si dice riducibile se esiste una matrice di permutazione P tale che PAP^T ha la forma

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

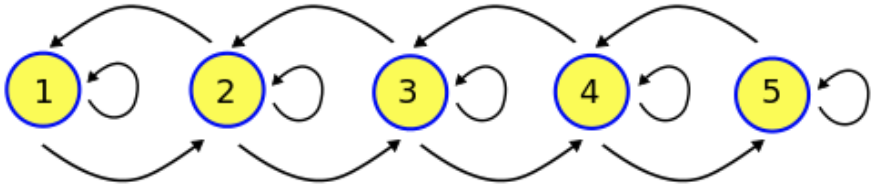


Figura 5: Grafo diretto associato alla matrice tridiagonale A

dove A_{11} e A_{22} sono matrici quadrate, cioè se esiste una permutazione di righe e colonne che porta A in forma triangolare a blocchi. Una matrice si dice irriducibile se non è riducibile.

La irriducibilità di una matrice può essere facilmente verificata utilizzando il concetto di *grafo diretto associato ad* A. Definiamo in modo informale questo concetto. Data una matrice $n \times n$ $A = (a_{i,j})$ consideriamo il grafo diretto $G[A]$ formato da n nodi, che numeriamo da 1 a n, nel quale un arco orientato unisce il nodo i al nodo j se $a_{i,j} \neq 0$. Ad esempio, la matrice tridiagonale A considerata poco fa ha il grafo diretto associato che si riporta per $n = 5$ in figura 5.

Un grafo diretto si dice *fortemente connesso* se per ogni coppia di nodi (i, j) esiste una successione di archi orientati che connette il nodo i col nodo j. In altri termini un grafo orientato è fortemente connesso se è possibile transitare per tutti i nodi percorrendo un cammino di archi orientati. Ad esempio il grafo in figura, associato alla matrice tridiagonale A è fortemente connesso.

Vale il seguente risultato.

Teorema 3 *Una matrice è irriducibile se e solo se il suo grafo associato è fortemente connesso.*

Dim. Si osserva innanzitutto che se P è una matrice di permutazione allora i grafi associati alle matrici A e $B = PAP^T$ differiscono unicamente per la numerazione dei nodi. Infatti, poiché $a_{i,j} = b_{\sigma_i,\sigma_j}$, dove σ_i è la permutazione associata alla matrice P, si ha che $a_{i,j} \neq 0$ se e solo se $b_{\sigma_i,\sigma_j} \neq 0$. Quindi un arco orientato unisce il nodo i col nodo j del grafo associato ad A se e solo se un arco orientato unisce il nodo σ_i col nodo σ_j nel grafo associato a B.

Ora, se la matrice A è riducibile allora esiste una matrice di permutazione P tale che $B = PAP^T$ è triangolare a blocchi cioè è del tipo

$$B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ 0 & B_{2,2} \end{bmatrix}$$

con $B_{1,1}$ matrice $m \times m$. Quindi il grafo associato a B non ha archi che uniscono i nodi $i > m$ ai nodi $j \leq m$. Quindi non è fortemente connesso.

Viceversa, se il grafo associato ad A non è fortemente connesso si trova la matrice di permutazione che porta A nella forma triangolare a blocchi nel modo seguente. Sia (p, q) una coppia di nodi per cui a partire da p non si possa raggiungere q percorrendo archi orientati nel grafo. Allora costruiamo

l'insieme $\mathcal{P}$ dei nodi raggiungibili da p e l'insieme $\mathcal{Q}$ dei nodi non raggiungibili da p. Certamente $q \in \mathcal{Q}$ per cui $\mathcal{Q}$ non è vuoto. Inoltre non possono esserci archi orientati che connettono nodi di $\mathcal{P}$ con nodi di $\mathcal{Q}$. Infatti, in tal caso percorrendo uno di questi archi potremmo connettere il nodo p con un nodo di $\mathcal{Q}$ che è assurdo. Allora basta ordinare le righe e le colonne di A in modo che in testa ci siano gli indici di $\mathcal{Q}$ e in coda i nodi di $\mathcal{P}$. In questo modo il blocco in basso a sinistra con indice di riga in $\mathcal{P}$ e indice di colonna in $\mathcal{Q}$ sarà costituito tutto da elementi nulli. □

Siamo ora pronti per enunciare il terzo teorema di Gerschgorin.

Teorema 4 (Terzo teorema di Gerschgorin) *Supponiamo che λ sia un autovalore di A con la seguente proprietà: se λ appartiene a K_i allora appartiene al bordo di K_i. In formula: $\lambda \in K_i \;\Rightarrow\; \lambda \in \partial K_i$, dove il simbolo ∂ denota il bordo. Se la matrice è irriducibile allora λ appartiene a tutti i cerchi di Gerschgorin e quindi appartiene all'intersezione delle frontiere dei cerchi.*

Dim. Ripercorriamo la dimostrazione del primo teorema di Gerschgorin. Se $Ax = \lambda x$, $x \neq 0$ e x_k è la componente di modulo massimo di x, allora vale

$$|a_{k,k} - \lambda| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \left|\frac{x_j}{x_k}\right| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}|,$$

quindi λ appartiene al k-esimo cerchio. Per ipotesi λ deve stare sulla frontiera del cerchio, quindi nella relazione precedente vale l'uguaglianza, cioè

$$|a_{k,k} - \lambda| = \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \left|\frac{x_j}{x_k}\right| = \sum_{j=1,\, j\neq k}^{n} |a_{k,j}|.$$

Ciò è possibile solo se $\left|\frac{x_j}{x_k}\right| = 1$ in corrispondenza di quegli indici j per cui $a_{k,j} \neq 0$. Poichè la matrice A è irriducibile il grafo associato è fortemente connesso, quindi esiste una successione di nodi $k_1 = k, k_2, k_3, \ldots, k_m$ tale che $m \geq n$, $\{k_1, k_2, \ldots, k_m\} = \{1, 2, \ldots, n\}$, per cui nel grafo c'è un arco orientato che unisce k_i con k_{i+1} e quindi $a_{k_i,k_{i+1}} \neq 0$ per $i = 1, \ldots, m-1$. Dal fatto che $a_{k_1,k_2} = a_{k,k_2} \neq 0$, ne segue che $\left|\frac{x_k}{x_{k_2}}\right| = 1$ e quindi anche x_{k_2} è una componente di x di modulo massimo per cui λ appartiene al k_2-esimo cerchio di Gerschgorin. Ripetendo lo stesso ragionamento successivamente con $k = k_2, k_3, \ldots, k_{m-1}$ si deduce che λ appartiene a tutti i cerchi di Gerschgorin e quindi a tutte le loro frontiere. □

Come applicazione di questo risultato si dimostra facilmente che la matrice tridiagonale A in (1) è non singolare. Infatti essa è irriducibile, inoltre, se $\lambda = 0$ fosse autovalore allora starebbe sulla frontiera di tutti i cerchi a cui appartiene, cioè $K_2, \ldots, K_{n-1}$. Quindi per il teorema dovrebbe appartenere alle frontiere di tutti i cerchi. Ma ciò è assurdo poiché $\lambda = 0$ non appartiene al primo e all'ultimo cerchio.

La matrice tridiagonale A di questo esempio è un caso particolare di matrice *irriducibilmente dominante diagonale*, cioè tale che

- A è irriducibile
- $|a_{i,i}| \geq \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|$
- esiste un indice k per cui $|a_{k,k}| > \sum_{j=1,\ j\neq k}^{n} |a_{k,j}|$

L'esempio mostrato della matrice tridiagonale irriducibilmente dominante diagonale è abbastanza speciale (infatti si può facilmente dimostrare per induzione che $\det A = n+1$). Però esistono ampie classi di problemi provenienti dalle applicazioni in cui intervengono matrici caratterizzate dall'essere irriducibilmente dominanti diagonali, per le quali il terzo teorema di Gerschgorin garantisce la non singolarità.

3 Estensione dei teoremi di Gerschgorin

I teoremi di Gerschgorin hanno delle generalizzazioni interessanti. Una di queste è data dal teorema di Brauer che coinvolge gli *ovali di Cassini*

Dati numeri complessi a, b e un numero reale $r \geq 0$, si definisce ovale di Cassini associato alla terna (a, b, r) il seguente insieme

$$C(a, b, r) = \{z \in \mathbb{C} : \quad |(z-a)(z-b)| \leq r\}$$

Teorema 5 (Brauer) *Siano $r_i = \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|$ i raggi dei cerchi di Gerschgorin. Gli autovalori di A sono contenuti nell'insieme*

$$\cup_{i>j} C(a_{i,i}, a_{j,j}, r_i r_j).$$

Un'altra estensione proposta da Richard Varga si basa sulla seguente osservazione. Sia $d = (d_i)$ un vettore di n componenti positive. Se D è la matrice diagonale con elementi diagonali $d_1, d_2, \ldots, d_n$, allora $A_d = D^{-1}AD$ ha gli stessi autovalori di A. Per cui applicando il primo teorema di Gerschgorin si arriva al seguente risultato.

Teorema 6 *Tutti gli autovalori di A sono contenuti nell'insieme*

$$\Omega = \cap_{d>0} \cup_{i=1}^{n} K_i(d), \quad K_i(d) = \{z \in \mathbb{C} : \ |z - a_{i,i}| \leq \frac{1}{d_i} \sum_{j=1,\ j\neq i}^{n} |a_{i,j}| d_j\}.$$

dove $d = (d_i)$, $d_i > 0$.

Richard Varga ha dimostrato che per ogni elemento ξ della frontiera di Ω esiste una matrice $B = (b_{i,j})$ tale che $|b_{i,j}| = |a_{i,j}|$ che ha ξ come autovalore. Ciò mostra che l'insieme Ω fornisce una inclusione stretta degli autovalori.

4 Esempi d'uso dei teoremi di Gerschgorin

Sia $p(x) = x^n + \sum_{i=0}^{n-1} a_i x^i$ un polinomio con coefficienti $a_0, a_1, \ldots, a_{n-1}$ e si consideri la matrice *companion*

$$C = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & \ddots & \ddots & \\ 0 & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{n-1} \end{bmatrix}.$$

È facile dimostrare per induzione su n che $\det(xI - C) = p(x)$ per cui gli zeri del polinomio $p(x)$ sono gli autovalori di C.

Il teorema di Gerschgorin applicato a C e a C^T permette allora di localizzare gli zeri di $p(x)$. Se $a_0 \neq 0$ allora lo stesso risultato applicato al polinomio "ribaltato" $q(x) = (1/a_0)x^n p(x^{-1}) = x^n + (1/a_0)\sum_{i=0}^{n-1} a_{n-i-1}x^i$ fornisce inclusioni per i reciproci degli zeri.

Un'altra localizzazione degli zeri di polinomi deriva dalla seguente osservazione. Siano $x_1, \ldots, x_n$ approssimazioni degli zeri $\lambda_1, \ldots, \lambda_n$ del polinomio $p(x)$ introdotto sopra e si costruisca la matrice $A = D - ue^T$ dove $D = \mathrm{diag}(x_1, \ldots, x_n)$ è matrice diagonale con elementi diagonali $x_1, \ldots, x_n$, e dove $e^T = (1, \ldots, 1)$, $u = (u_i)$,

$$u_i = p(x_i) / \prod_{j=1,\ j\neq i}^{n} (x_i - x_j).$$

Allora si può dimostrare che $\det(xI - A) = p(x)$. Il teorema di Gerschgorin applicato ad A fornisce delle stime dell'errore con cui i valori x_i approssimano gli zeri. In particolare, i cerchi di Gerschgorin K_i hanno centro $x_i - u_i$ e raggio $(n-1)|u_i|$ per $i = 1, \ldots, n$.

In certe situazioni sono di aiuto trasformazioni per similitudine date da matrici diagonali. Infatti tali trasformazioni non alterano gli autovalori e permettono di scalare i raggi dei cerchi di Gerschgorin. Si consideri ad esempio la matrice tridiagonale

$$A = \begin{bmatrix} 1 & 1 & & \\ -1 & 100 & 1 & \\ & -1 & 2 & -1 \\ & & 1 & 101 \end{bmatrix}.$$

I suoi cerchi di Gerschgorin hanno due componenti connesse formate da due cerchi ciascuna. Ma la matrice ottenuta moltiplicando per $\epsilon > 0$ la prima e la terza riga e dividendo per ϵ la prima la terza colonna, data da

$$D_\epsilon A D_\epsilon^{-1} = \begin{bmatrix} 1 & \epsilon & & \\ -\epsilon^{-1} & 100 & \epsilon^{-1} & \\ & -\epsilon & 2 & -\epsilon \\ & & \epsilon^{-1} & 101 \end{bmatrix}, \quad D_\epsilon = \mathrm{diag}(\epsilon, 1, \epsilon, 1),$$

ha il primo cerchio disgiunto dagli altri se $1/48 \leq \epsilon < 1/3$ e quindi si deduce dal secondo teorema di Gerschgorin che A ha due autovalori reali rispettivamente

negli intervalli $[1-\epsilon, 1+\epsilon]$ e $[2-2\epsilon, 2+2\epsilon]$. In modo analogo possiamo operare con gli altri due cerchi e trovare delle scalature di righe e colonne che permettono di ottenere cerchi disgiunti centrati in 100 e 101. □

5 Note storiche

Il primo teorema di Gerschgorin è stato pubblicato nel 1931 nell'articolo Über die Abgrenzung der Eigenwerte einer Matrix dal matematico bielorusso di origine ebraica Semyon Aranovich Gerschgorin (24 Agosto 1901 - 30 Maggio 1933). Fra il 1946 e il 1948 Brauer, un allievo di Schur, pubblica una raccolta sistematica di teoremi di limitazione degli autovalori, fra i quali quelli noti come secondo e terzo teorema di Gerschgorin e anche la generalizzazione data in termini degli ovali di Cassini.

Il primo teorema di Gerschgorin è stato preceduto da diversi risultati simili. Nel 1881 Levy dimostra che una matrice con elementi diagonali negativi ed elementi rimanenti positivi in cui le somme degli elementi sulle righe sono negative ha determinante non nullo. Hadamard estende questo risultato al caso complesso nel 1898. Successivamente anche Minkovsky dà una estensione di questo risultato.

Risultati di localizzazione più recenti si trovano su numerosi articoli pubblicati su varie riviste. Per una ricerca a riguardo si veda il data base MathSciNet dell'American Mathematical Society (AMS).

6 Esercizi

Esercizio 1 Localizzare gli autovalori della matrice

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 5 & 1 & 0 \\ 1 & 0 & 8 & 1 \\ 1 & 0 & 0 & 11 \end{bmatrix}$$

usando i cerchi di Gerschgorin di A e di A^T.

Esercizio 2 Siano $C = (c_{i,j})$, $D = (d_{i,j})$ matrici $n \times n$ con $n \geq 2$, tali che $c_{i,j} = 1$ se $i - j = 1 \mod n$, $c_{i,j} = 0$ altrimenti, $d_{i,j} = i$ per $i = j$, $d_{i,j} = 0$ altrimenti. Sia $A(t) = D + tC$, $t \geq 0$.
a) Utilizzando i teoremi di Gerschgorin si diano condizioni sufficienti su t affinché la matrice $A(t)$ abbia autovalori reali.
b) Per $\epsilon \neq 0$ sia $F_\epsilon = (f_{i,j})$ tale che $f_{i,j} = \epsilon c_{i,j}$ per $i \geq j$, $f_{i,j} = \epsilon^{1-n} c_{i,j}$ per $i < j$. Si dimostri che $A(t)$ è simile alla matrice $A_\epsilon(t) = D + tF_\epsilon$ per ogni $\epsilon \neq 0$.
c) Applicando i teoremi di Gerschgorin a $A_\epsilon(t)$ e alla sua trasposta si diano condizioni sufficienti affinché $A(t)$ abbia autovalori reali.

Esercizio 3 Sia $\alpha \in \mathbb{R}$, $n \geq 4$ un intero e $A = (a_{i,j})$ matrice $n \times n$ definita da $a_{i,i} = i\alpha$, per $i = 1, \ldots, n$, $a_{i+1,i} = -1$, $a_{i,i+1} = 1$, per $i = 1, \ldots, n-1$ e $a_{i,j} = 0$ altrove.

a) Determinare i valori di α per cui i cerchi di Gerschgorin di A sono a due a due disgiunti. Dimostrare che per tali valori di α la matrice A ha autovalori reali.

b) Posto $\alpha = 4$ si calcolino cerchi di inclusione per ciascun autovalore di A col raggio più piccolo possibile. Suggerimento: si applichino i teoremi di Gerschgorin alla matrice $D^{-1}AD$ dove D è una matrice diagonale opportunamente scelta.

c) Valutare nel modo più accurato possibile il valore $\beta > 0$ per cui la matrice A ha autovalori reali per ogni α tale che $|\alpha| \geq \beta$.

Soluzione

La matrice ha la forma seguente

$$A = \begin{bmatrix} \alpha & 1 & & & \\ -1 & 2\alpha & 1 & & \\ & -1 & 3\alpha & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & -1 & n\alpha \end{bmatrix}.$$

Se $\alpha = 0$ la matrice è antisimmetrica e quindi ha autovalori immaginari o nulli. Possiamo assumere per semplicità $\alpha > 0$, infatti se $\alpha < 0$ possiamo considerare la matrice $-A^T$ che rientra nel caso precedente.

L'i-esimo cerchio di Gerschgorin di una matrice $A = (a_{i,j})$ di ordine n ha centro $c_i = a_{i,i}$ e raggio $r_i = \sum_{j=1,\ j\neq i}^{n} |a_{i,j}|$. Nel nostro caso vale

$$\begin{array}{lll} c_1 = \alpha, & r_1 = 1, & \\ c_i = i\alpha, & r_i = 2, & i = 2, \ldots, n-1, \\ c_n = n\alpha, & r_n = 1. & \end{array}$$

a) I cerchi hanno centro reale e il bordo del cerchio i-esimo interseca l'asse reale nei punti $i\alpha + 2$ e $i\alpha - 2$, $i = 2, \ldots, n-1$, mentre i bordi del primo e dell'ultimo cerchio intersecano l'asse reale in $\alpha - 1$, $\alpha + 1$ e rispettivamente in $n\alpha - 1$, $n\alpha + 1$. Quindi i cerchi sono disgiunti se e solo se

$$\begin{array}{ll} \alpha + 1 < 2\alpha - 2, & \\ i\alpha + 2 < (i+1)\alpha - 2, & i = 2, \ldots, n-2, \\ (n-1)\alpha + 2 < n\alpha - 1. & \end{array}$$

Dalla prima e dall'ultima relazione si ricava $\alpha > 3$. Dalle rimanenti si ricava $\alpha > 4$. La condizione richiesta è dunque $\alpha > 4$.

Per il secondo teorema di Gerschgorin, se i cerchi sono a due a due disgiunti allora ciascun cerchio contiene un solo autovalore della matrice. Poiché nel nostro caso la matrice è reale, i suoi eventuali autovalori non reali compaiono a

coppie complesse coniugate. Per cui, essendo i centri dei cerchi di Gerschgorin reali, se un cerchio contenesse l'autovalore non reale λ, esso conterrebbe anche l'autovalore coniugato $\bar{\lambda}$. Ciò è assurdo poiché ogni cerchio deve contenere un solo autovalore.

b) Per $\alpha = 4$ il primo cerchio di centro 4 e raggio 1 è disgiunto dagli altri cerchi. Quindi contiene un solo autovalore. È possibile ottenere un raggio di inclusione più piccolo operando nel seguente modo. La matrice ottenuta da A moltiplicando la prima riga per un numero $0 < \epsilon < 1$ e la prima colonna per ϵ^{-1} ha gli stessi autovalori di A. Il raggio del primo cerchio di Gerschgorin diventa ϵ, il raggio del secondo cerchio diventa $1 + \epsilon^{-1}$. Determiniamo per quali valori di ϵ il primo cerchio è ancora disgiunto dai rimanenti. Poiché i raggi degli altri cerchi non sono cambiati si ha la condizione $4 + \epsilon < 8 - \epsilon^{-1} - 1$ che fornisce $\epsilon^2 - 3\epsilon + 1 < 0$ che è verificata per $(3 - \sqrt{5})/2 < \epsilon < (3 + \sqrt{5})/2$. Per cui, finché la diseguaglianza è verificata, il cerchio di centro 4 e raggio ϵ contiene un solo autovalore che è reale. Quindi, anche per $\epsilon = (3 - \sqrt{5})/2$ il cerchio di centro 4 e raggio ϵ contiene un solo autovalore. Analogamente si dimostra che il cerchio di centro n e raggio $(3 - \sqrt{5})/2$ contiene un solo autovalore.

Se $2 \leq i \leq n - 1$, moltiplicando la i-esima riga per ϵ e la i-esima colonna per ϵ^{-1} si ottiene una nuova matrice con gli stessi autovalori di A. I cerchi di Gerschgorin della nuova matrice sono gli stessi della matrice A ad esclusione dei tre cerchi di indice $i-1, i, i+1$ che hanno raggi rispettivamente $1+\epsilon^{-1}, 2\epsilon, 1+\epsilon^{-1}$. L'$i$-esimo cerchio è quindi disgiunto dai due contigui se $4(i-1) + 1 + \epsilon^{-1} < 4i - 2\epsilon < 4i + 2\epsilon < 4(i+1) - 1 - \epsilon^{-1}$,cioè

$$\epsilon^{-1} - 3 < 2\epsilon < 3 - \epsilon^{-1}.$$

La doppia diseguaglianza è verificata per $1/2 < \epsilon < 1$. Per cui il cerchio di centro $4i$ e raggio ϵ contiene un solo autovalore che è reale per ogni $1/2 < \epsilon < 1$ e quindi anche per $\epsilon = 1/2$.

c) Ripetendo il ragionamento precedente per un valore generico di α si ottiene per il primo cerchio la condizione $\epsilon^2 + \epsilon(1-\alpha) + 1 < 0$ che ha soluzioni se $\alpha > 3$ o $\alpha < -1$. Analogo risultato vale per l'ultimo cerchio. Se $2 \leq i \leq n - 1$ si ottengono le condizioni

$$1 - \alpha + \epsilon^{-1} < 2\epsilon < \alpha - 1 - \epsilon^{-1}$$

che hanno soluzioni se $\alpha > \sqrt{8} + 1$ oppure $\alpha < -\sqrt{8} + 1$. □

Esercizio 4 Sia $A = (a_{i,j})$ matrice reale tale che $a_{i,j} \leq 0$ se $i \neq j$. Si dimostri che se esiste un vettore $v = (v_i)$, $v_i > 0$ per ogni i, tale che posto $w = (w_i) = Av$ risulta $w_i > 0$ per ogni i, allora tutti gli autovalori di A hanno parte reale positiva.

Si dimostri che vale la stessa proprietà se A è irriducibile e la condizione $w_i > 0$ viene sostituita da $w_i \geq 0$ per ogni i, ed esiste un indice k tale che $w_k > 0$.

Esercizio 5 Sia A_n matrice $n \times n$, $n = 2m$, con $m < 100$, tale che $A_n = \mathrm{diag}(u) + \alpha H$, $\alpha \geq 0$, dove $u = (1, 100, 2, 200, \ldots, m, 100m)$, $H = (h_{i,j})$, $h_{i,i+1} = -h_{i+1,i} = 1$, $h_{i,j} = 0$ altrove. Determinare un numero θ tale che per ogni $\alpha \leq \theta$ la matrice A_n ha autovalori reali per ogni n.

Esercizio 6 Sia A la matrice $n \times n$ con elementi $a_{1,i} = -a_{i,1} = 1$, $i = 2, \ldots, n$, $a_{1,1} = n^2$, $a_{i,i} = i$, $i = 2, \ldots, n$. Dare delle inclusioni più accurate possibili agli autovalori di A, in particolare dire se gli autovalori della matrice sono reali e distinti.

Esercizio 7 Sia A la matrice tridiagonale $n \times n$, con $n \geq 3$, data da

$$A = \begin{bmatrix} 0 & a & & & & \\ a & 0 & a & & & \\ & 1/2 & 0 & 1/2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1/2 & \ddots & 1/2 \\ & & & & 1/2 & 0 \end{bmatrix}$$

dove $a \geq 0$. Si trovino condizioni su a affiché la matrice abbia raggio spettrale minore di 1.

Esercizio 8 Sia $\underline{i}$ l'unità immaginaria tale che $\underline{i}^2 = -1$, sia A la matrice $2n \times 2n$

$$A = \left[\begin{array}{c|c} \underline{i}D_1 & B \\ \hline C & -\underline{i}D_2 \end{array}\right]$$

dove D_1 e D_2 sono matrici diagonali di dimensioni $n \times n$ con autovalori reali nell'intevallo $[a, b]$, con $1 < a \leq b$. Inoltre B e C sono matrici complesse tali che $\|B\|_\infty = \|C\|_\infty = 1$.
a) Dimostrare che A è invertibile, ha autovalori con parte reale di modulo al più 1 e nessuno di essi è reale.
b) Dimostrare che A ha autovalori con parte reale di modulo al più $a - \sqrt{a^2 - 1}$ e con parte immaginaria di modulo almeno $\sqrt{a^2 - 1}$ e dare in funzione di a e b limitazioni superiori al raggio spettrale di A il più accurate possibile.

Soluzione. a) Poiché B e C hanno norma infinito 1 la somma dei moduli dei loro elementi su ogni riga è al più 1. Quindi i cerchi di Gerschgorin di A hanno centro $\underline{i}x$ con $x \in [\underline{i}a, \underline{i}b] \cup [-\underline{i}b, -\underline{i}a]$, e raggio al più 1. Questi cerchi non intersecano l'asse reale. Applicando il primo teorema di Gerschgorin si ottiene il punto a).

b) Coniugando la matrice A con $\mathrm{diag}(I, \epsilon I)$ si ottiene

$$A_\epsilon = \begin{bmatrix} \underline{i}D_1 & \epsilon B \\ \epsilon^{-1}C & -\underline{i}D_2 \end{bmatrix}.$$

Imponiamo che i cerchi con centro con parte immaginaria positiva siano disgiunti da quelli con centro con parte immaginaria negativa. Questo avviene se $a - \epsilon > -a + \epsilon^{-1}$. La condizione diventa quindi

$$\epsilon^2 - 2a\epsilon + 1 < 0$$

che è verificata se $a - \sqrt{a^2-1} < \epsilon < a + \sqrt{a^2-1} = 1/(a - \sqrt{a^2-1})$. Per il secondo teorema di Gerschgorin, per tutti questi valori di ϵ l'unione dei primi n cerchi contiene n autovalori e l'unione dei restanti cerchi contiene n autovalori di A. I raggi dei primi n cerchi sono al più ϵ. Per cui scegliendo $\epsilon = a - \sqrt{a^2-1}$ si deduce che l'unione dei primi n cerchi di raggio al più $a - \sqrt{a^2-1}$ contiene n autovalori e scegliendo $\epsilon = 1/(a - \sqrt{a^2-1})$ si deduce che l'unione del secondo gruppo di n cerchi che hanno raggio al più $a - \sqrt{a^2-1}$ contiene n autovalori. Questo implica il punto b). Inoltre, poichè i centri hanno modulo al più b dai teoremi di Gerschgorin segue che il raggio spettrale è al più $b + a - \sqrt{a^2-1}$.

□

Esercizio 9 Sia A una matrice $n \times n$ ad elementi complessi. Dimostrare le proprietà seguenti
a) Se A è irriducibile e se $Ax = \lambda x$, $\lambda \in \mathbb{C}$, $x = (x_i) \in \mathbb{C}^n$, e $x_k \neq 0$ è l'unica componente di x di modulo massimo, allora λ appartiene alla parte interna del cerchio di Gerschgorin di centro $a_{k,k}$ e raggio $\sum_{j=1,\, j\neq k}^{n} |a_{k,j}|$.
b) Se $Ax = \lambda x$, $Ay = \mu y$ con $\lambda, \mu \in \mathbb{C}$, $x, y \in \mathbb{C}^n$, e se inoltre esiste un indice k tale che $x_k \neq 0$ e $y_k \neq 0$ sono componenti di modulo massimo rispettivamente di x e di y, allora $|\lambda - \mu| \leq 2 \sum_{j=1,\, j\neq k}^{n} |a_{k,j}|$. Sia A irriducibile; dimostrare che la diseguaglianza diventa stretta se x_k è l'unica componente di modulo massimo di x, oppure se y_k è l'unica componente di modulo massimo di y.
c) Se A è reale e tutti i cerchi di Gerschgorin di A hanno parti interne a due a due disgiunte, allora A ha tutti autovalori reali.

Soluzione

a) Seguendo la dimostrazione del primo teorema di Gerschgorin si ha $|a_{k,k} - \lambda| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \frac{|x_j|}{|x_k|}$. Poiché A è irriducibile allora esiste almeno un indice h per cui $a_{k,h} \neq 0$. Poichè $|x_k|$ è l'unica componente di modulo massimo vale $|x_h|/|x_k| < 1$ quindi $|a_{k,k} - \lambda| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \frac{|x_j|}{|x_k|} < \sum_{j=1,\, j\neq k}^{n} |a_{k,j}|$.

b) Dalla dimostrazione del primo teorema di Gerschgorin segue che sia λ che μ appartengono allo stesso cerchio di Gerschgorin per cui $|\lambda - \mu|$ è minore o uguale al diametro del cerchio.

Se inoltre x_k è l'unica componente di modulo massimo di x allora $|a_{k,k} - \lambda| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \frac{|x_j|}{|x_k|}$. Essendo A irriducibile esiste h tale che $a_{k,h} \neq 0$, e poiché $|x_h|/|x_k| < 1$, vale $|a_{k,k} - \lambda| \leq \sum_{j=1,\, j\neq k}^{n} |a_{k,j}| \frac{|x_j|}{|x_k|} < \sum_{j=1,\, j\neq k}^{n} |a_{k,j}|$. Si procede analogamente nel caso in cui y_k sia la componente di modulo massimo di y.

c) Si usa un argomento di continuità. Precisamente si considera la matrice

$A(t) = D + t(A - D)$ dove $D = \mathrm{diag}(a_{1,1}, \ldots, a_{n,n})$. Se $t \in [0, 1)$ la matrice $A(t)$ ha tutti i cerchi di Gerschgorin a due a due disgiunti. Per cui dalla teoria sappiamo che $A(t)$ ha tutti autovalori $\lambda_i(t)$ reali. Per la continuità degli autovalori si ha che $\lambda_i(1) = \lim_{t\to 1} \lambda_i(t)$ e $\lambda_i(1)$ sono gli autovalori di A. Poiché $\lambda_i(t)$ è reale per ogni $t \in [0, 1)$, ed essendo $\lambda_i(t)$ funzione continua di t allora il limite $\lambda_i(1) = \lim_{t\to 1} \lambda_i(t)$ è reale. □

Esercizio 10 a) Sia $n \geq 2$ intero e $m = 2$. Dimostrare che se una matrice $n \times n$ H ha un autovalore λ di molteplicità geometrica m allora λ appartiene ad almeno m cerchi di Gerschgorin di H. Cioè esistono due indici $h \neq k$ tali che $\lambda \in C_k \cap C_h$, dove C_i denota il cerchio di Gerschgorin relativo all'i-esima riga di H.
b) Dimostrare la proprietà del punto a) per m generico. Dire se la proprietà vale per la molteplicità algebrica, dare eventualmente un controesempio.

Soluzione

a) Sia $m = 2$. Per l'autovalore λ esistono allora due autovettori u, v linearmente indipendenti. Sia $|u_k| = \max_i |u_i|$, $|v_h| = \max_i |v_i|$. Se $k \neq h$ allora procedendo come nella dimostrazione del primo teorema di Gerschgorin si deduce che λ appartiene a due cerchi: il k-esimo e l'h-esimo. Se invece $h = k$, posto $t = v_h/u_h$ si ha che $w = v - tu$ è non nullo poiché u e v sono linearmente indipendenti, e inoltre è autovettore essendo combinazione lineare non nulla di due autovettori corrispondenti allo stesso autovalore. Inoltre $w_h = 0$ per cui il $\max_i |w_i|$ viene preso su un indice $\widehat{k}$ diverso da h. Quindi, procedendo ancora come nella dimostrazione del primo teorema di Gerschgorin, si deduce che $\lambda \in C_h \cap C_{\widehat{k}}$.

b) Se $v^{(1)}, \ldots, v^{(m)}$ sono gli autovettori, considero la matrice V le cui righe sono $v^{(i)T}$, $i = 1, \ldots, m$. Sia $V = LU\Pi$ la fattorizzazione di V ottenuta applicando l'eliminazione gaussiana col massimo pivot sulle colonne, dove Π è matrice di permutazione. In questo modo le righe di $V\Pi$ sono ancora autovettori di H essendo combinazioni lineari di autovettori, inoltre, poiché gli elementi diagonali di V sono quelli di massimo modulo sulle rispettive righe, si ha che ciascuna riga di $V\Pi^T$ è autovettore con elemento di modulo massimo preso su indici diversi. Il primo teorema di Gerschgorin completa la dimostrazione.

La matrice companion associata al polinomio $(x + a)^2$, cioè

$$\begin{bmatrix} 0 & -a^2 \\ 1 & -2a \end{bmatrix}$$

è tale che $\lambda = -a$ è autovalore di moltelicità algebrica 2 e geometrica 1. Se $a = 1/2$, solo il secondo cerchio contiene λ. □

Esercizio 11 a) Sia $n \geq 2$ intero e $b \in \mathbb{R}$. Sia A la matrice tridiagonale $(2n) \times (2n)$ con elementi diagonali $a_{i,i} = ib$, $a_{n+i,n+i} = i(b-1)$, per $i = 1, \ldots, n$,

inoltre $a_{i,i+1} = b^2$, $a_{i+1,i} = -b^2$, per $i = 1, \ldots, 2n-1$. Si dimostri che se $0 < b \leq 1/2$ allora A ha n autovalori con parte reale positiva e n autovalori con parte reale negativa.
b) Si diano condizioni sufficienti su b affinché A abbia n autovalori reali positivi.

Soluzione

a) Si applicano i teoremi di Gerschgorin. Si denoti C_i il cerchio di Gerschgorin costruito sulla i-esima riga di A per cui C_1 ha centro b e raggio b^2, C_i ha centro ib e raggio $2b^2$ per $i = 2, \ldots, n$ mentre C_i ha centro $(i-n)(b-1)$ e raggio $2b^2$ per $i = n+1, \ldots, 2n-1$ infine C_{2n} ha centro $n(b-1)$ e raggio b^2.

Se $0 < b \leq 1/2$ risulta $b - b^2 > 0$ e $ib - 2b^2 > 0$ per $i = 2, \ldots, n$ per cui i primi n cerchi sono contenuti nel semipiano destro aperto di $\mathbb{C}$ costituito dai numeri complessi con parte reale positiva. Inoltre, poiché $(b-1)(i-n) + 2b^2 \leq 0$ per $i = n+1, \ldots, 2n-1$ e $(b-1)n + b^2 < 0$, i cerchi $C_{n+1}, \ldots, C_{2n}$ sono contenuti nel semipiano sinistro chiuso di $\mathbb{C}$ costituito dai numeri complessi con parte reale ≤ 0. Di questi, C_{n+1} è l'unico cerchio che contiene 0 per $b = 1/2$, che inoltre sta sul bordo.

Per il secondo teorema di Gerschgorin, poiché $\cup_{i=1}^{n} C_i$ è disgiunta da $\cup_{i=n+1}^{2n} C_i$, la matrice A ha n autovalori con parte reale positiva e n autovalori con parte reale ≤ 0. Se per $b = 1/2$ esistesse un autovalore con parte reale nulla questo starebbe nella frontiera di C_{n+1} e per il terzo teorema di Gerschgorin, visto che la matrice è irriducibile, dovrebbe appartenere alle frontiere di tutti i cerchi. Il che è assurdo.

b) È sufficiente che i primi cerchi siano disgiunti tra loro e che gli altri cerchi siano contenuti nel semipiano sinistro. Infatti la tesi segue dal secondo teorema di Gerschgorin. La condizione $C_i \cap C_{i+1} = \emptyset$ è $ib + 2b^2 < (i+1)b - 2b^2$ per $i = 2, \ldots, n$, mentre per $i = 1$ è $b + b^2 < 2b - 2b^2$. Si hanno quindi le condizioni

$$3b^2 - b < 0$$
$$4b^2 - b < 0,$$

che sono verificate per $0 < b < 1/4$. Per questi valori di b i rimanenti cerchi sono ancora contenuti nel semipiano sinistro per cui i rimanenti autovalori hanno parte reale negativa. Per motivi di continuità, gli autovalori rimangono reali anche per $b = 1/4$ e n di questi sono positivi.

Riferimenti bibliografici

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

[2] A. Brauer, Limits for the characteristic roots of matrices II, Duke Math. J. 14 (1947) 21-26.

[3] S. Gerschgorin, Über die Abgrenzung der Eigenwerte einer Matrix, Izv. Ahad. Nauk. SSSR, Ser. Mat. 7 (1931), 749-754.

[4] G. H. Golub, C. F. Van Loan. Matrix Computations. Johns Hopkins University Press, Baltimore, 1996.

[5] R. S. Varga, Gershgorin and His Circles, Springer Verlag 2004.

La forma normale di Schur

Dario A. Bini, Università di Pisa

24 ottobre 2019

Sommario

Questo modulo didattico contiene risultati relativi alla forma normale di Schur, alle sue proprietà e alle sue applicazioni.

1 Introduzione

Tra le diverse forme canoniche di una matrice disponibili sul mercato la forma di Schur è particolarmente utile poiché si ottiene con una trasformazione per similitudine data da una matrice unitaria. Ricordiamo che una matrice $U \in \mathbb{C}^{n\times n}$ è detta *unitaria* se $U^H U = UU^H = I$, dove I indica la matrice identica $n \times n$ e A^H indica la matrice trasposta hermitiana di A, cioè la matrice che si ottiene trasponendo A e coniugando gli elementi complessi. Una matrice reale e unitaria è detta *ortogonale* .

Enunciamo questo risultato e ne diamo una dimostrazione costruttiva che non utilizza la forma normale di Jordan. La dimostrazione che vedremo può essere opportunamente trasformata in un algoritmo di calcolo.

È importante osservare che il calcolo numerico della forma normale di Jordan è un compito numericamente intrattabile. Infatti tale forma non è stabile sotto perturbazioni della matrice pur arbitrariamente piccole purché non nulle. A questo riguardo si consideri un singolo blocco di Jordan di dimensione n, ad esempio con autovalore nullo

$$J = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}.$$

È evidente che la matrice J ha $\lambda = 0$ come unico autovalore di molteplicità algebrica n e di molteplicità geometrica 1. Si modifichi ora J alterando l'elemento di posto $(n, 1)$ con una perturbazione $\epsilon > 0$ ottenendo

$$J_\epsilon = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ \epsilon & & & 0 \end{bmatrix}.$$

Non è difficile verificare che questa matrice ha polinomio caratteristico $\lambda^n - \epsilon$ e quindi ha n autovalori distinti dati da $\lambda_i = \epsilon^{1/n}\omega_n^i$, $i = 1, \ldots, n$, dove $\omega_n = \cos(2\pi/n) + \mathtt{i}\sin(2\pi/n)$, dove $\mathtt{i}$ è l'unità immaginaria tale che $\mathtt{i}^2 = -1$. Quindi la sua forma normale di Jordan è data da una matrice *diagonale* .

La forma normale di Schur non presenta questo inconveniente.

2 Forma normale di Schur

Vale il seguente

Teorema 1 (Forma normale di Schur) *Per ogni matrice $A \in \mathbb{C}^{n\times n}$ esistono una matrice triangolare superiore T e una matrice unitaria U tali che*

$$U^H A U = T.$$

Dim. Si procede per induzione sulla dimensione n della matrice A. Se $n = 1$ la matrice A è un numero complesso e la decomposizione di Schur vale con $T = A$ e $U = 1$. In generale, supponiamo che la fattorizzazione valga per dimensione $n-1$. Consideriamo $A \in \mathbb{C}^{n\times n}$ e denotiamo con λ e x rispettivamente un autovalore e un autovettore di A, cioè $Ax = \lambda x$. Supponiamo che x sia normalizzato in modo che $x^H x = 1$ e consideriamo una base ortonormale formata dai vettori $y_2, \ldots, y_n$ dello spazio ortogonale a x. Costruiamo la matrice Q le cui colonne sono $x, y_2, \ldots, y_n$. Questa matrice è unitaria, cioè $Q^H Q = I$ e risulta $Qe^{(1)} = x$, $e^{(1)} = Q^H x$ dove $e^{(1)} = (1, 0, \ldots, 0)^T$. Inoltre vale

$$Q^H A Q = \begin{bmatrix} \lambda & u^T \\ 0 & A_{n-1} \end{bmatrix},$$

dove $A_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$. Per verificare quest'ultima relazione basta considerare $Q^H A Q e^{(1)}$ che è la prima colonna di $Q^H A Q$. Vale

$$Q^H A Q e^{(1)} = Q^H A x = Q^H \lambda x = \lambda Q^H x = \lambda e^{(1)}$$

come richiesto. Inoltre per l'ipotesi induttiva si ha $A_{n-1} = U_{n-1} T_{n-1} U_{n-1}^H$, da cui posto

$$U_n = Q \begin{bmatrix} 1 & 0 \\ & U_{n-1} \end{bmatrix},$$

si ottiene

$$U_n^H A U_n = \begin{bmatrix} \lambda & u^T U_{n-1}^H \\ 0 & T_{n-1} \end{bmatrix}$$

che è una matrice triangolare. □

Si osservi che per trasformare la dimostrazione del teorema precedente in algoritmo di calcolo è sufficiente disporre di due "scatole nere": la prima che data in input una matrice A ci fornisce in output un suo autovalore λ e il corrispondente autovettore x; la seconda che dato in input un vettore x ci fornisce in output una base ortonormale del sottospazio ortogonale a x, o, equivalentemente, fornisce una matrice Q tale che $Qx = e^{(1)}$. Vedremo in un altro articolo come la seconda scatola nera sia di facile costruzione usando le matrici elementari di Householder.

Un'altra osservazione utile è che la forma di Schur non è unica. Infatti scegliendo diversi ordinamenti degli autovalori arriviamo a forme normali in generale diverse tra loro.

Seguendo l'impostazione della dimostrazione del teorema precedente è possibile dimostrare un risultato specifico per le matrici reali. Per questo abbiamo bisogno della seguente definizione

Definizione 1 *Una matrice $T \in \mathbb{R}^{n\times n}$ si dice* quasi triangolare *se si può scrivere nella forma*

$$T = \begin{bmatrix} T_{1,1} & \dots & T_{1,n} \\ & \ddots & \vdots \\ & & T_{m,m} \end{bmatrix}$$

dove $T_{i,i}$ per $i = 1, \dots, m$ possono essere matrici 2×2 con coppie di autovalori complessi coniugati, oppure matrici 1×1, cioè numeri reali.

Gli autovalori di una matrice quasitriangolare sono gli autovalori delle sottomatrici $T_{i,i}$ per $i = 1, \dots, m$.

Ad esempio, la matrice

$$\begin{bmatrix} 2 & 1 & 1 & 1 & 1 \\ -1 & 2 & 1 & 1 & 1 \\ 0 & 0 & 5 & 2 & 3 \\ 0 & 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix}$$

è quasi triangolare. I suoi autovalori sono dati dagli autovalori di $\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$ che sono $2 + \mathtt{i}$ e $2 - \mathtt{i}$, dagli autovalori di $\begin{bmatrix} 3 & -1 \\ 1 & 3 \end{bmatrix}$ che sono $3 + \mathtt{i}$, $3 - \mathtt{i}$, e da 5.

Teorema 2 (Forma reale di Schur) *Per ogni matrice $A \in \mathbb{R}^{n\times n}$ esistono una matrice quasi triangolare superiore $T \in \mathbb{R}^{n\times n}$ e una matrice ortogonale $Q \in \mathbb{R}^{n\times n}$ tali che*

$$Q^T A Q = T.$$

3 Applicazioni

La forma di Schur ha delle conseguenze interessanti, alcune di facile dimostrazione. Una prima conseguenza riguarda le matrici hermitiane, cioè quelle matrici A tali che $A = A^H$.

Infatti, se $U^H AU = T$ è la forma di Schur della matrice hermitiana A, allora la proprietà $A = A^H$ implica

$$T^H = U^H A^H U = U^H AU = T.$$

Cioè T è hermitiana e quindi, essendo triangolare, è diagonale. Inoltre gli elementi diagonali di T sono tali che $t_{i,i} = \bar{t}_{i,i}$ e quindi sono reali. Cioè si ottiene che le matrici hermitiane sono diagonalizzabili con una trasformazione per similitudine unitaria.

Se A è anti-hermitiana, cioè se $A = -A^H$ allora procedendo come sopra si ottiene che anche T è anti-hermitiana. Ne segue che T è una matrice diagonale tale che $t_{i,i} = -\bar{t}_{i,i}$. Cioè una matrice anti-hermitiana è diagonalizzabile da una trasformazione ortogonale e i suoi autovalori sono numeri immaginari.

Un risultato più generale è espresso dal seguente

Teorema 3 *Una matrice $A \in \mathbb{C}^{n\times n}$ ha forma normale di Schur diagonale se e solo se $A^H A = AA^H$.*

Dim. Se T è la forma di Schur di A, allora $A^H A = AA^H$ se e solo se $T^H T = TT^H$. Ciò segue dal fatto che

$$A^H A = U^H T^H UU^H TU = U^H T^H TU, \quad AA^H = U^H TUU^H T^H U = U^H TT^H U.$$

Se T è matrice diagonale allora $T^H T = TT^H$ e quindi $A^H A = AA^H$. Viceversa, se $A^H A = AA^H$ allora $T^H T = TT^H$. Dimostriamo che la condizione $T^H T = TT^H$ implica che T è matrice diagonale. Per questo procediamo per induzione su n. Se $n = 1$ non c'è nulla da dimostrare. Assumiamo vera la tesi per matrici di dimensione $n-1$ e consideriamo il caso di dimensione n. Leggendo la relazione $T^H T = TT^H$ sull'elemento di posto $(1, 1)$ otteniamo

$$\bar{t}_{1,1} t_{1,1} = \sum_{j=1}^{n} t_{1,j} \bar{t}_{1,j}$$

cioè $|t_{1,1}|^2 = |t_{1,1}|^2 + |t_{1,2}|^2 + \cdots + |t_{1,n}|^2$ da cui $|t_{1,2}|^2 + \cdots + |t_{1,n}|^2 = 0$. Data la non negatività degli addendi questo implica che $t_{1,j} = 0$ per $j = 2, \ldots, n$. Quindi la matrice T ha la forma

$$T = \begin{bmatrix} t_{1,1} & 0 \\ 0 & T_{n-1} \end{bmatrix}.$$

La condizione $T^H T = TT^H$ implica allora $T_{n-1}^H T_{n-1} = T_{n-1} T_{n-1}^H$, e, per l'ipotesi induttiva segue che T_{n-1} è matrice diagonale. □

La classe di matrici che verificano la condizione $A^H A = AA^H$ è detta classe delle matrici *normali*

Una matrice unitaria è in particolare una matrice normale e quindi ha una forma di Schur diagonale. Inoltre da $A^H A = I$ segue che $|t_{i,i}|^2 = 1$. Cioè le matrici unitarie sono diagonalizzabili da una trasformazione ortogonale e hanno autovalori di modulo 1.

4 Esercizi

1. Determinare una forma di Schur della matrice $A = uv^T$ dove $u, v \in \mathbb{R}^n$.

2. Usando la forma normale di Schur determinare l'insieme dei possibili autovalori delle matrici A ad elementi complessi $n \times n$ che risolvono l'equazione $A^2 - 5A^H + 6I = 0$.

3. Sia α un numero reale e si consideri la classe $\mathcal{A}_n(\alpha)$ delle matrici $n \times n$ a elementi complessi tali che $\alpha A + A^H A + A^H = I$. Utilizzando la forma normale di Schur si descriva l'insieme $\Lambda(\alpha) = \{\lambda \in \mathbb{C} : \quad \exists A \in \mathcal{A}_n(\alpha),\ \det(A - \lambda I) = 0\}$.

4. Si ricavi la forma normale di Schur di una matrice A dalla forma normale di Jordan $A = SJS^{-1}$.

5. Siano $a, b \in \mathbb{R}$ e
$$C = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}. \tag{1}$$
Si determini una matrice Q unitaria tale che che $Q^H C Q = D$, dove D diagonale con elementi diagonali $a \pm \mathbf{i}b$ dove $\mathbf{i}^2 = -1$. Si dimostri che tutte le matrici della forma (1) commutano.

   Soluzione. Calcolando gli zeri del polinomio caratteristico, troviamo che gli autovalori di C sono $\lambda_{1,2} = a \pm \mathbf{i}b$. Inoltre gli autovettori corrispondenti sono $x_1 = \begin{bmatrix} 1 \\ \mathbf{i} \end{bmatrix}$, $x_2 = \begin{bmatrix} 1 \\ -\mathbf{i} \end{bmatrix}$, che sono tra loro ortogonali. Dunque, se definiamo la matrice $Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ \mathbf{i} & -\mathbf{i} \end{bmatrix}$, questa è una matrice unitaria tale che
$$Q^H C Q = \begin{bmatrix} a + \mathbf{i}b & 0 \\ 0 & a - \mathbf{i}b \end{bmatrix}.$$
   In particolare, tutte le matrici della forma (1) sono diagonalizzabili mediante la stessa trasformazione, quindi commutano.

6. Dimostrare la forma normale reale di Schur del teorema 2.

   Soluzione. Si procede per induzione sulla dimensione n della matrice A con la stesso approccio usato per dimostrare il teorema di Schur nel caso

generale. Se $n = 1$ non c'è nulla da dimostrare. Se $n = 2$ e gli autovalori sono reali, allora posto v un autovettore reale tale che $v^T v = 1$, la matrice Q le cui colonne sono v e u con u reale, ortogonale a v e tale che $u^T u = 1$, risulta $Q^T Q = I$ e $Q^T A Q = T$ triangolare superiore. Quindi T è in forma reale di Schur. Se invece con $n = 2$ la matrice ha una coppia di autovalori complessi coniugati allora è già nella forma reale di Schur. Per il passo induttivo, sia A matrice reale $n \times n$. Se λ è un autovalore reale si procede come nella dimostrazione del teorema di Schur nel caso generale e si riconduce il problema a dimensione $n - 1$. Se invece $\lambda = a + \mathtt{i}b$ e $\bar{\lambda} = a - \mathtt{i}b$, $a, b \in \mathbb{R}$, sono una coppia di autovalori complessi coniugati corrispondenti agli autovettori $v = u + \mathtt{i}w$, $\bar{v} = u - \mathtt{i}w$, con $u, w \in \mathbb{R}^n$ si osserva che

$$A \begin{bmatrix} u & w \end{bmatrix} = \begin{bmatrix} u & w \end{bmatrix} \begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

Cioè il sottospazio di $\mathbb{R}^n$ generato da u e w è invariante per A. I vettori u e w sono linearmente indipendenti perché, se non lo fossero, il vettore u sarebbe autovettore, dunque A avrebbe un autovettore reale corrispondente a un autovalore con parte immaginaria non nulla, che è un assurdo. Allora si costruisce una base ortonormale di questo sotto spazio, formata da due vettori $u^{(1)}$ e $u^{(2)}$ scegliendo combinazioni lineari di u e w, ad esempio

$$\begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} = \begin{bmatrix} u, w \end{bmatrix} S$$

con S opportuna matrice 2×2. In questo modo si ha

$$A \begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} = \begin{bmatrix} u^{(1)}, u^{(2)} \end{bmatrix} T_1$$

con $T_1 = S^{-1} \left[\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix} \right] S$ matrice 2×2. I vettori $u^{(1)}$ e $u^{(2)}$ si completano con una base ortonormale dello spazio ortogonale a $\mathrm{span}(u^{(1)}, u^{(2)})$. La matrice U che ha per colonne questi vettori è tale che $U^T A U$ è triangolare superiore a blocchi col primo blocco sulla diagonale principale che coincide con T_1. Mentre il blocco di posto $(2, 2)$ ha dimensione $n - 2$ quindi per l'ipotesi induttiva ha forma normale reale di Schur.

7. Usando la forma normale reale di Schur si dimostri che se A è una matrice $n \times n$ reale allora A è normale se e solo se esiste una matrice reale ortogonale Q tale che $Q^T A Q$ è una matrice diagonale a blocchi, reale, con blocchi di dimensione minore o uguale a 2 in cui i blocchi di dimensione 2 hanno la forma (1).

   Soluzione. Se la matrice $Q^T A Q$ è diagonale a blocchi con blocchi T_i di dimensione 1 o 2 e struttura definita in (1) allora poiché $A^T A = A A^T$ se e solo se $T_i^T T_i = T_i T_i^T$, basta dimostrare che $T_i^T T_i = T_i T_i^T$. Se la dimensione di T_i è 1 questo è banalmente vero. Se la dimensione è 2 e i blocchi hanno la struttura (1) allora la proprietà segue dall'esercizio 5. Per dimostrare l'implicazione opposta si procede per induzione come nella dimostrazione del teorema 3. Per $n = 1$ non c'è nulla da dimostrare. Per

$n = 2$ se la matrice ha autovalori reali allora il risultato segue dal teorema 3. Se invece $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ è reale con autovalori complessi coniugati, dalla relazione $A^T A - AA^T = 0$ segue $b^2 = c^2$, $(b-c)(d-a) = 0$ da cui $b = \pm c$. Se fosse $b = c$ la matrice sarebbe simmetrica e avrebbe autovalori reali. Allora deve essere $b = -c$ e conseguentemente $a = d$. Si ottiene quindi la struttura in (1). Per il passo induttivo si considera la forma normale di Schur reale T di A. Siano $T_{i,j}$ i blocchi di T. Leggendo l'uguaglianza $T^T T = TT^T$ nel blocco di posto $(1,1)$ si ha

$$T_{1,1}^T T_{1,1} = T_{1,1} T_{1,1}^T + \sum_{i \geq 2} T_{1,i} T_{1,i}^T$$

da cui

$$T_{1,1}^T T_{1,1} - T_{1,1} T_{1,1}^T = \sum_{i \geq 2} T_{1,i} T_{1,i}^T$$

La matrice a sinistra ha traccia nulla, e quindi anche la matrice a destra. Poiché ciascun addendo è semidefinito positivo deve avere traccia non negativa. Poiché la somma delle tracce è nulla allora ciascun addendo ha traccia nulla. Ma una matrice reale simmetrica semidefinita positiva con traccia nulla deve essere identicamente nulla. Ne segue che $T_{1,i} = 0$ per ogni i. Si osserva infatti che se $T_{1,i} \neq 0$ esisterebbe $x \neq 0$ tale che $T_{1,i}^T x \neq 0$, da cui $x^T T_{1,i} T_{1,i}^T x > 0$ per cui $T_{1,i} T_{1,i}^T$ non può esser nulla.

L'uguaglianza a zero della prima riga a blocchi di T, con eventuale esclusione di $T_{1,1}$ riconduce il problema a dimensione inferiore per cui si può applicare l'ipotesi induttiva.

Riferimenti bibliografici

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

Norme di vettori e matrici

Dario A. Bini, Università di Pisa

7 luglio 2020

Sommario

Questo modulo didattico contiene risultati e proprietà relativi alle norme di vettori e di matrici.

1 Introduzione

Nello studio dei metodi di risoluzione di sistemi lineari è utile disporre del concetto di *norma* per valutare attraverso un numero reale non negativo la grandezza di un vettore o di una matrice.

1.1 Norme di vettori

Diamo la seguente

Definizione 1 *Una applicazione* $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ *viene detta* norma vettoriale *se soddisfa alle seguenti proprietà*

- $\|x\| \geq 0$, *per ogni* $x \in \mathbb{R}$; $\|x\| = 0$ *se e solo se* $x = 0$;
- $\|\alpha x\| = |\alpha| \, \|x\|$, *per ogni* $\alpha \in \mathbb{C}$, $x \in \mathbb{C}^n$;
- $\|x + y\| \leq \|x\| + \|y\|$, *per ogni* $x, y \in \mathbb{C}^n$ *(diseguaglianza triangolare)*

Esempi importanti di norme sono:

- $\|x\|_1 = \sum_{i=1}^n |x_i|$ (norma 1)
- $\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{\frac{1}{2}}$ (norma euclidea, o norma 2)
- $\|x\|_\infty = \max_i |x_i|$ (norma infinito, o norma del massimo)

Queste norme sono casi speciali della *norma di Hölder*

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}, \quad p \geq 1,$$

dove $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$.

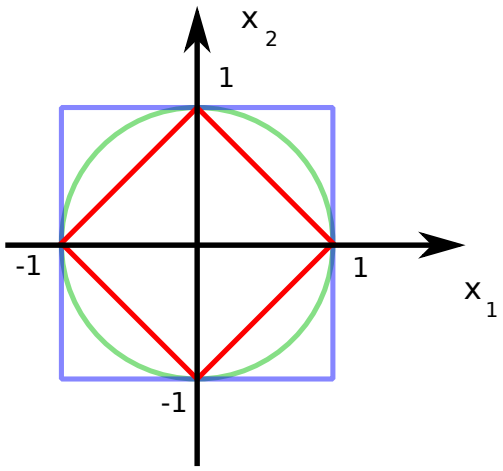


Figura 1: Palle unitarie in norma 1 (rosso), norma 2 (verde) e norma infinito (blu).

Nella figura 1 si riportano le sfere unitarie, cioè gli insiemi $S = \{x \in \mathbb{C}^n : \|x\| = 1\}$ nel caso delle norme 1,2 e ∞ per $n = 2$. In blu la palla per la norma infinito, in verde quella della norma 2 e in rosso quella della norma 1.

Non è difficile dimostrare che per una norma l'insieme $\{x \in \mathbb{C}^n : \quad \|x\| \leq 1\}$ è un insieme convesso.

Dim. Siano infatti x, y, tali che $\|x\|, \|y\| \leq 1$ e si consideri un generico punto z del segmento di estremi x e y. Vale allora $z = \alpha x + (1 - \alpha)y$ con $0 \leq \alpha \leq 1$. Dalla diseguaglianza triangolare segue

$$\|z\| \leq \alpha\|x\| + (1 - \alpha)\|y\| \leq \alpha + 1 - \alpha = 1$$

e quindi $\|z\| \leq 1$. □

Questo fatto ci permette di dire che per $0 < p < 1$ l'espressione $\|x\|_p$ non può essere una norma essendo l'insieme $\{x \in \mathbb{C}^n : \quad \|x\| \leq 1\}$ non convesso. Nella figura 2 si riporta l'insieme S_p per valori di $p < 1$.

Si ricorda che un *prodotto scalare* su $\mathbb{C}^n$ è una applicazione da $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ che alla coppia (x, y) associa il numero $\langle x, y\rangle$ tale che

- $\langle x, y\rangle = \overline{\langle y, x\rangle}$.
- $\langle x, \alpha y\rangle = \alpha\langle x, y\rangle$, per ogni $\alpha \in \mathbb{C}$
- $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
- $\langle x, x\rangle \geq 0$
- $\langle x, x\rangle = 0$ se e solo se $x = 0$

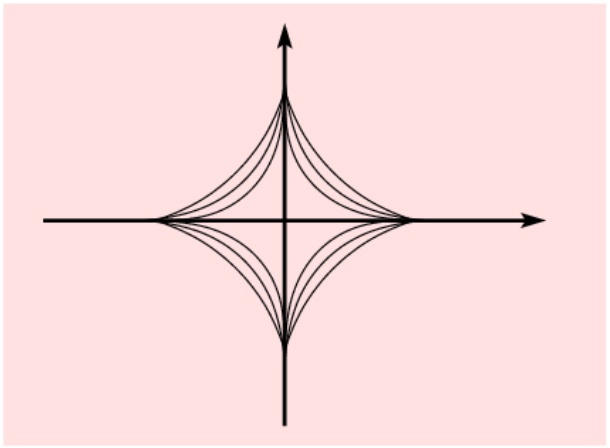


Figura 2: Insiemi S_p per valori di p minori di 1

Dalle proprietà del prodotto scalare discende

- $\langle \alpha x, y \rangle = \overline{\alpha} \langle x, y \rangle$, per ogni $\alpha \in \mathbb{C}$
- $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$

Un esempio di prodotto scalare su $\mathbb{R}^n$ è dato da $\langle x, y \rangle = x^T y$; esso è detto *prodotto scalare euclideo*. Un esempio di prodotto scalare su $\mathbb{C}^n$ è dato da $\langle x, y \rangle = x^H y$; esso è detto *prodotto scalare hermitiano*.

Si osserva che, dato un prodotto scalare $\langle x, y \rangle$ su $\mathbb{C}^n$, allora l'applicazione $x \to \langle x, x \rangle^{1/2}$ è una norma. Questa norma viene detta *norma indotta dal prodotto scalare*. Ad esempio, la norma 2 è la norma indotta dal prodotto scalare hermitiano $\langle x, y \rangle = x^H y$.

Un prodotto scalare soddisfa la diseguaglianza di Cauchy-Schwarz nota anche come diseguaglianza di Cauchy-Bunyakowski-Schwarz

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle, \tag{1}$$

di cui si riporta la breve dimostrazione

dim. Se $y = 0$ la diseguaglianza è soddisfatta. Se $y \neq 0$ si pone $t = \langle y, x \rangle / \langle y, y \rangle$ per cui

$$0 \leq \langle x - ty, x - ty \rangle = \langle x, x \rangle - |\langle x, y \rangle|^2 / \langle y, y \rangle$$

da cui, moltiplicando per $\langle y, y \rangle$, si ottiene la diseguaglianza di Cauchy-Schwarz □

La diseguaglianza di Cauchy-Schwarz implica che $\langle x, y \rangle \leq 1$ per ogni coppia di vettori x e y tali che $\|x\| = \|y\| = 1$, dove $\| \cdot \|$ è la norma indotta dal

prodotto scalare. Nel caso di $\mathbb{R}^n$ questo ci permette di definire l'angolo $\theta \in [0, \pi]$ formato da due vettori x e y mediante l'espressione $\cos\theta = \langle x, y\rangle/(\|x\|\,\|y\|)$. Nel caso di $\mathbb{C}^n$ si può definire l'angolo $\theta \in [0, \pi/2]$ mediante la relazione $\cos\theta = |\langle x, y\rangle|/(\|x\|\,\|y\|)$. Questo ci permette di definire l'ortogonalità di due vettori x, y quando $\langle x, y\rangle = 0$.

Ci si può chiedere se la norma 1 o la norma 2 siano o meno indotte da un prodotto scalare. Per questo vale il seguente risultato.

Teorema 1 *Una norma $\|\cdot\|$ è indotta da un prodotto scalare se e solo se vale la legge del parallelogramma*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2). \tag{2}$$

Inoltre, su $\mathbb{R}^n$ il prodotto scalare che induce la norma $\|\cdot\|$ è dato da

$$\langle x, y\rangle = \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2),$$

mentre su $\mathbb{C}^n$ il prodotto scalare è dato da

$$\langle x, y\rangle = \frac{1}{4}(\|x + y\|^2 - \|x - y\|^2 + \mathtt{i}(\|x + \mathtt{i}y\|^2 - \|x - \mathtt{i}y\|^2)).$$

La condizione (2) data nel teorema precedente, che abbiamo chiamato legge del parallelogramma, esprime il fatto che la somma dei quadrati delle diagonali di un parallelogramma è uguale alla somma dei quadrati dei quattro lati. Si può verificare che la norma infinito non soddisfa la legge del parallelogramma. Ad esempio, in $\mathbb{R}^2$, scegliendo $x = (1, 0)$ e $y = (0, 1)$ vale $\|x+y\|_\infty^2 = 1$, $\|x-y\|_\infty^2 = 1$, mentre $2(\|x\|_\infty^2 + \|y\|_\infty^2) = 4$. Gli stessi vettori forniscono un controesempio per dimostrare che anche la norma 1 non è indotta da un prodotto scalare.

La diseguaglianza triangolare si può esprimere in modo equivalente nella seguente forma

$$|\|x\| - \|y\|| \le \|x - y\|, \quad x, y \in \mathbb{C}^n. \tag{3}$$

Infatti, consideriamo $a, b \in \mathbb{C}^n$. Allora dalla diseguaglianza triangolare applicata a a e b, si ha $\|a + b\| \le \|a\| + \|b\|$ da cui $\|a + b\| - \|a\| \le \|b\|$. Quindi ponendo $x = a + b$ e $y = a$ si ha $\|x\| - \|y\| \le \|x - y\|$. Scambiando x con y si deduce che $\|y\| - \|x\| \le \|y - x\| = \|x - y\|$ da cui segue la diseguaglianza (3).

Una conseguenza immediata della diseguaglianza triangolare espressa nella forma (3) è che ogni norma è una funzione *uniformemente continua* cioè vale

$$\forall\, \epsilon > 0\ \exists\, \delta > 0 :\ \forall x, y \in \mathbb{C}^n\ |x_i - y_i| \le \delta\ \Rightarrow\ |\,\|x\| - \|y\|\,| \le \epsilon.$$

L'uniformità della continuità sta nel fatto che il valore di δ non dipende dalla scelta di x e y in $\mathbb{C}^n$. Questa utile proprietà si dimostra nel modo seguente

Dim. Si consideri la base canonica di $\mathbb{C}^n$ formata dai vettori $e^{(j)} = (e_i^{(j)})$, $j = 1, \ldots, n$ tali che $e_i^{(j)} = \delta_{i,j}$, dove $\delta_{i,j}$ è il delta di Kronecker. In questo modo $x = \sum_{i=1}^n x_i e^{(i)}$ e $y = \sum_{i=1}^n y_i e^{(i)}$. Vale allora

$$\|x - y\| = \|\sum_{i=1}^n (x_i - y_i) e^{(i)}\| \leq \sum_{i=1}^n |x_i - y_i| \, \|e^{(i)}\|,$$

da cui, se $|x_i - y_i| \leq \delta$ ne segue che

$$\|x - y\| \leq \delta\gamma, \quad \gamma = \sum_{i=1}^n \|e^{(i)}\|,$$

dove γ è indipendente da x e da y. Quindi dalla (3) si deduce che

$$| \, \|x\| - \|y\| \, | \leq \delta\gamma.$$

Allora per completare la dimostrazione basta scegliere $\delta < \epsilon/\gamma$. □

La continuità delle norme vettoriali implica la seguente proprietà di equivalenza delle norme su $\mathbb{C}^n$.

Teorema 2 (Proprietà di equivalenza delle norme) *Per ogni coppia di norme $\|\cdot\|'$ e $\|\cdot\|''$ su $\mathbb{C}^n$ esistono costanti positive α, β tali che per ogni $x \in \mathbb{C}^n$ vale*

$$\alpha\|x\|' \leq \|x\|'' \leq \beta\|x\|'.$$

Dim. Se $x = 0$ la proprietà è verificata. Supponiamo allora $x \neq 0$. Possiamo scegliere una delle due norme a piacere ad esempio la norma infinito. Infatti se dimostriamo la doppia diseguaglianza per la coppia $\|\cdot\|', \|\cdot\|_\infty$ e per la coppia $\|\cdot\|'', \|\cdot\|_\infty$, componendo le due doppie diseguaglianze otteniamo le costanti che legano $\|\cdot\|'$ e $\|\cdot\|''$. Scegliamo allora $\|\cdot\|' = \|\cdot\|_\infty$ e osserviamo che, per la definizione di $\|\cdot\|_\infty$, l'insieme $S_\infty = \{x \in \mathbb{C}^n : \quad \|x\|_\infty = 1\}$ è limitato. Inoltre, essendo S_∞ l'immagine inversa dell'insieme chiuso $\{1\}$ tramite $\|\cdot\|_\infty$ ed essendo la norma infinito continua, ne segue che S_∞ è anch'esso chiuso. Sappiamo allora che essendo S_∞ un sottoinsieme chiuso e limitato di $\mathbb{C}^n$ esso è compatto. Quindi la funzione continua $\|\cdot\|''$ ha massimo e minimo su S_∞. Siano β e α i valori di questo massimo e minimo. Allora, poiché vale $y = \frac{1}{\|x\|_\infty} x \in S_\infty$, ne segue che

$$\alpha \leq \|y\|'' \leq \beta$$

da cui

$$\alpha\|x\|_\infty \leq \|x\|'' \leq \beta\|x\|_\infty.$$

□

Se $x = (x_i) \in \mathbb{C}^n$, definiamo $y = (y_i)$ dove $y_i = |x_i|$. Se $\|\cdot\|$ è una delle tre norme 1,2 e infinito vale $\|x\| = \|y\|$. In generale questo non è vero. Si provi a trovare degli esempi di questo fatto. Una norma che verifica questa proprietà è detta *norma assoluta*.

Si possono agevolmente determinare le costanti α e β che legano le norme 1,2 e infinito. Vale infatti il seguente

Teorema 3 *Per ogni $x \in \mathbb{C}^n$ si ha*

$$\begin{array}{l} \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty \\ \|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2 \\ \|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty \end{array} \tag{4}$$

Dim. Poiché $|x_i| \le \|x\|_\infty$ si ha $\|x\|_1 = \sum_{i=1}^n |x_i| \le n\|x\|_\infty$ e $\|x\|_1 \ge \|x\|_\infty$. Da cui segue la prima coppia di diseguaglianze. Dall'identità $\left(\sum_{i=1}^n |x_i|\right)^2 = \sum_{i=1}^n |x_i|^2 + 2\sum_{i>j} |x_i x_j|$ segue $\|x\|_1^2 \ge \|x\|_2^2$ e quindi $\|x\|_1 \ge \|x\|_2$. Per dimostrare la diseguaglianza $\|x\|_1 \le \sqrt{n}\|x\|_2$ si deve ricorrere alla diseguaglianza di Cauchy-Schwarz soddisfatta dal prodotto scalare hermitiano. Infatti scegliendo $y = (y_i)$ con $|y_i| = 1$ e $y_i \bar{x}_i = |x_i|$, dalla diseguaglianza di Cauchy-Schwarz (1) si ottiene

$$(\sum_{i=1}^n |x_i|)^2 = |x^H y|^2 \le \langle x, x\rangle \langle y, y\rangle = n \sum_{i=1}^n |x_i|^2$$

da cui $\|x\|_1 \le \sqrt{n}\|x\|_2$. Per quanto riguarda la terza coppia di diseguaglianze si osserva che la seconda si ottiene dalla formula $\|x\|_2^2 = \sum_{i=1}^n |x_i|^2$ maggiorando ciascun $|x_i|$ con $\|x\|_\infty$. Per la prima diseguaglianza basta osservare che $\sum_{i=1}^n |x_i|^2 \ge |x_j|^2$ per ogni j. □

Una proprietà interessante della norma 2 è che è invariante sotto trasformazioni unitarie. Infatti se $y = Ux$ ed U è unitaria, cioè $U^H U = I$, allora

$$\|y\|_2^2 = y^H y = (Ux)^H (Ux) = x^H U^H U x = x^H x = \|x\|_2^2.$$

1.2 Alcuni Esercizi

1. Se $x \to \|x\|$ è norma allora $x \to \alpha\|x\|$ è norma per ogni $\alpha > 0$.
2. Se $x \to \|x\|$ è norma e S è matrice invertibile allora $x \to \|Sx\|$ è norma.
3. Se S è matrice $m \times n$ di rango massimo con $m \ge n$, e se $\|\cdot\|$ è una norma su $\mathbb{R}^m$, è vero che $x \to \|Sx\|$ è norma su $\mathbb{R}^n$?
4. Se $\|x\|'$ e $\|x\|''$ sono norme allora $\alpha\|x\|' + \beta\|x\|''$ è norma per $\alpha, \beta > 0$.
5. Dire se le seguenti applicazioni sono norme
   - $x \to \min_i |x_i|$
   - $x \to \max_i |x_i| + \min_i |x_i|$
   - $(x_1, x_2) \to |x_1 - x_2| + |x_1|$
   - $(x_1, x_2) \to |x_1| + \min\{|x_1|, |x_2|\}$
6. Sia $x \in \mathbb{R}^{2n}$ e si partizioni x in due sottovettori $x^{(1)} = (x_1, \ldots, x_n)$ e $x^{(2)} = (x_{n+1}, \ldots, x_{2n})$. Siano $\|\cdot\|'$ una norma su $\mathbb{R}^n$ e $\|\cdot\|''$ una norma su $\mathbb{R}^2$. dire se l'applicazione $x \to \|(\|x^{(1)}\|', \|x^{(2)}\|')\|''$ è una norma.

7. Sia A matrice hermitiana definita positiva. Dimostrare che $(x, y) \to y^H Ax$ è un prodotto scalare.

8. Dimostrare che una matrice A è unitaria se e solo se per ogni $x \in \mathbb{C}^n$ vale $\|x\|_2 = \|Ax\|_2$.

2 Norme di matrici

L'insieme delle matrici $n \times n$ è uno spazio vettoriale di dimensioni n^2. In altri termini una matrice A $n \times n$ può essere vista come un vettore di n^2 componenti. Per cui in linea teorica potremmo usare le definizioni e le proprietà delle norme di vettori per matrici in generale. Però è conveniente usare una definizione leggermente più forte che impone qualche utile proprietà.

Definizione 2 (Norma di matrice) *Si dice* norma di matrice *una applicazione* $\|\cdot\|$ *da* $\mathbb{C}^{n\times n}$ *in* $\mathbb{R}$ *tale che*

- $\|A\| \geq 0, \qquad \|A\| = 0$ *se e solo se* $A = 0$
- $\|\alpha A\| = |\alpha| \|A\|$ *per* $\alpha \in \mathbb{C}$
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|AB\| \leq \|A\| \, \|B\|$

Le prime tre proprietà sono le analoghe di quelle date nel caso dei vettori. La quarta, più specifica al contesto delle matrici, è detta *proprietà submoltiplicativa*. Un esempio di norma di matrice è la *norma di Frobenius* definita da

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|^2 \right)^{1/2}.$$

Si osserva che la norma di Frobenius non è altro che la norma euclidea applicata al vettore

$$vec(A) = (a_{1,1}, a_{2,1}, \ldots, a_{2,n}, a_{1,2}, \ldots, a_{n,n})^T$$

che si ottiene sovrapponendo una sull'altra le colonne di A. La norma di Frobenius può anche essere scritta come $\|A\|_F = \text{traccia}(A^H A)^{1/2}$ ed è la norma indotta dal prodotto scalare $\langle A, B \rangle = \text{traccia}(A^H B)$.

Tra le norme di matrici giocano un ruolo importante le *norme di matrici indotte* da una norma vettoriale, dette anche *norme operatore*.

Sia $\|\cdot\|$ una norma su $\mathbb{C}^n$ e A una matrice $n \times n$. Poiché la sfera unitaria $S = \{x \in \mathbb{C}^n : \qquad \|x\| = 1\}$ è un chiuso e limitato quindi un compatto, ed essendo la norma una funzione continua, esiste il

$$\max_{x \in S} \|Ax\|.$$

È facile dimostrare che l'applicazione che associa ad A questo massimo è una norma che chiamiamo norma di matrice indotta dalla norma vettoriale $\|\cdot\|$ e denotiamo col simbolo $\|A\|$. Poniamo quindi

$$\|A\| := \max_{\|x\|=1} \|Ax\|.$$

La dimostrazione delle quattro proprietà è una semplice verifica e non viene riportata.

Se guardiamo ad una matrice come un operatore lineare tra due spazi vettoriali, la norma indotta di una matrice ci dice qual è il massimo allungamento che l'operatore produce nel trasformare i vettori. L'allungamento viene misurato nella norma vettoriale assegnata. Per questo motivo questa norma viene chiamata anche norma operatore.

Una proprietà che segue direttamente dalla definizione e che è molto utile nelle elaborazioni successive è la seguente

$$\|Ax\| \leq \|A\| \, \|x\|. \tag{5}$$

Un'altra conseguenza interessante della definizione di norma di matrice indotta è $\|I\| = 1$. Questa proprietà non è verificata dalla norma di Frobenius per cui $\|I\| = \sqrt{n}$. Quindi la norma di Frobenius non è una norma indotta.

Ci possiamo chiedere come sono fatte le norme di matrice indotte dalle norme vettoriali 1,2 e infinito. Vale per questo il seguente risultato che ci permette di valutare queste norme in modo agevole in almeno due casi su 3.

Teorema 4 *Per le norme di matrice indotte dalla norma 1,2 e infinito vale*

$$\begin{array}{l} \|A\|_1 = \max_j \sum_{i=1}^n |a_{i,j}| \\ \|A\|_2 = \left(\rho(A^H A)\right)^{1/2} \\ \|A\|_\infty = \max_i \sum_{j=1}^n |a_{i,j}| \end{array} \tag{6}$$

dove $\rho(A)$ denota il raggio spettrale di una matrice A, cioè il massimo dei moduli dei suoi autovalori.

Dim. La metodologia dimostrativa è la stessa nel caso delle tre norme. In un primo passo si dimostra che le relazioni (6) valgono con il segno $\leq$. In un secondo passo si dimostra che esiste un vettore x di norma unitaria per cui $\|Ax\|$ coincide con il membro destro in (6) nei tre casi distinti. Partiamo con la norma 1 e dimostriamo che $\|A\|_1 \leq \max_j \sum_{i=1}^n |a_{i,j}|$. Dato $x \in \mathbb{C}^n$ tale che $\|x\|_1 = 1$ vale

$$\|Ax\|_1 = \sum_{i=1}^n \left| \sum_{j=1}^n a_{i,j} x_j \right| \leq \sum_{i=1}^n \sum_{j=1}^n |a_{i,j}| \, |x_j| = \sum_{j=1}^n |x_j| \sum_{i=1}^n |a_{i,j}|.$$

Maggiorando $\sum_{i=1}^n |a_{i,j}|$ con $\max_j \sum_{i=1}^n |a_{i,j}|$ si ottiene

$$\|Ax\|_1 \leq \max_j \sum_{i=1}^n |a_{i,j}| \sum_{j=1}^n |x_j| = \max_j \sum_{i=1}^n |a_{i,j}|.$$

Inoltre se $\max_j \sum_{i=1}^n |a_{i,j}|$ è preso sulla colonna k-esima, il vettore $x = e^{(k)}$ che ha componenti nulle tranne la k-esima che vale 1, è tale che $\|x\|_1 = 1$ e

$$\|Ax\|_1 = \sum_{i=1}^n |a_{i,k}| = \max_j \sum_{i=1}^n |a_{i,j}|.$$

La dimostrazione procede in modo analogo con la norma infinito. Infatti, se x è un vettore tale che $\|x\|_\infty = 1$, vale

$$\|Ax\|_\infty = \max_i \left| \sum_{j=1}^n a_{i,j} x_j \right| \leq \max_i \sum_{j=1}^n |a_{i,j}|\,|x_j|.$$

Maggiorando $|x_j|$ con 1, si ottiene

$$\|Ax\|_\infty \leq \max_i \sum_{j=1}^n |a_{i,j}|.$$

Supponiamo che il massimo sia raggiunto sull'indice k. Allora scegliendo il vettore x in modo che $x_j = \bar{a}_{k,j}/|a_{k,j}|$ se $a_{k,j} \neq 0$ e $x_j = 1$ altrimenti, risulta $\|x\|_\infty = 1$ e la k-esima componente di Ax è uguale a $\sum_{j=1}^n |a_{k,j}| = \max_i \sum_{j=1}^n |a_{i,j}|$.

Per la norma 2 si procede in modo analogo. Se x è tale che $\|x\|_2 = 1$ allora $\|Ax\|_2^2 = x^H A^H A x$, e, poiché $A^H A$ è hermitiana, esiste una matrice unitaria U tale che $U^H A^H A U = D$, con D matrice diagonale e $d_{i,i} \geq 0$. Risulta allora, con $y = U^H x$,

$$\|Ax\|_2^2 = x^H (UDU^H) x = y^H D y = \sum_{i=1}^n |y_i|^2 \lambda_i \leq \rho(A^H A) \sum_{i=1}^n |y_i|^2 = \rho(A^H A),$$

dove l'uguaglianza è raggiunta quando x è autovettore di $A^H A$ corrispondente a $\rho(A^H A)$. □

Si può notare come la norma 1 e la norma infinito siano facilmente calcolabili. Mentre la norma 2, richiedendo il calcolo degli autovalori di una matrice è calcolabile con maggiori difficoltà computazionali.

Si è già osservato che dalla definizione di norma 2 e di norma di Frobenius segue $\|A\|_2 \leq \|A\|_F$. Infatti $\|A\|_F^2 = \text{traccia}(A^H A)$ è la somma degli autovalori della matrice semidefinita positiva $A^H A$, mentre $\|A\|_2^2 = \rho(A^H A)$ è il massimo degli autovalori di $A^H A$.

Dalle relazioni (4) si possono ricavare agevolmente le costanti che legano attraverso diseguaglianze le norme di matrice indotte dalla norma 1,2 e infinito.

È interessante osservare che se U e V sono matrici unitarie e $B = UAV$ allora

$$B^H B = (UAV)^H (UAV) = V^H A^H U^H U A V = V^H A^H A V$$

cioè $A^H A$ e $B^H B$ sono (unitariamente) simili ed hanno quindi gli stessi autovalori per cui $\rho(A^H A) = \rho(B^H B)$ e $\text{traccia}(A^H A) = \text{traccia}(B^H B)$.

Una proprietà che lega la norma 2 e la norma di Frobenius è che $\|A\|_2 \leq \|A\|_F$. Ciò vale poiché

$$\|A\|_2^2 = \rho(A^H A) \leq \text{traccia}(A^H A) = \|A\|_F^2,$$

infatti la traccia è la somma degli autovalori, che per $A^H A$ sono tutti maggiori o uguali a zero essendo $A^H A$ semidefinita positiva. Questo ci permette di dimostrare agevolmente la proprietà submoltiplicativa della norma di Frobenius. Infatti vale

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2 \leq \|A\|_F \|x\|_2$$

Questo implica che

$$\|AB\|_F^2 = \sum_{j=1}^{n} \|ABe^{(j)}\|_2^2 \leq \sum_{j=1}^{n} \|A\|_2^2 \|Be^{(j)}\|_2^2 \leq \|A\|_F^2 \sum_{j=1}^{n} \|Be^{(j)}\|_2^2 = \|A\|_F^2 \|B\|_F^2.$$

2.1 Norme e raggio spettrale

Ci sono delle relazioni molto strette tra le norme di matrici indotte e il raggio spettrale. La prima osservazione interessante è che se λ è autovalore di A, cioè $Ax = \lambda x$ per $x \neq 0$, allora per la (5)

$$\|Ax\| \leq \|A\| \, \|x\|$$

da cui $|\lambda| \, \|x\| \leq \|A\| \, \|x\|$. Ne segue che $\|A\| \geq |\lambda|$ per ogni autovalore λ per cui

$$\|A\| \geq \rho(A). \tag{7}$$

Ci possiamo chiedere allora se $\rho(A)$ può essere una norma per ogni matrice A. Per questo vale il seguente risultato

Teorema 5 *Per ogni matrice A e per ogni $\epsilon > 0$ esiste una norma di matrice indotta $\|\cdot\|$ tale che*

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon.$$

Inoltre, se gli autovalori di A di modulo uguale al raggio spettrale stanno in blocchi di Jordan di dimensione 1, allora esiste una norma indotta che applicata ad A coincide con $\rho(A)$.

Dim. Se S è una matrice invertibile allora si verifica facilmente che, data una norma $\|\cdot\|$ vettoriale, l'applicazione $x \to \|Sx\|$ è una norma. Definiamo allora $\|x\|_S := \|Sx\|$. Verifichiamo che la norma di matrice indotta dalla norma $\|\cdot\|_S$ è $\|A\|_S = \|SAS^{-1}\|$. Infatti

$$\|A\|_S := \max_{\|x\|_S = 1} \|Ax\|_S = \max_{\|Sx\| = 1} \|SAx\| = \max_{\|y\| = 1} \|SAS^{-1}y\| = \|SAS^{-1}\|,$$

con $y = Sx$. Ora portiamo A in forma di Jordan, $J = W^{-1}AW$, e fissato $\epsilon > 0$, consideriamo la matrice diagonale D_ϵ i cui elementi diagonali sono $1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1}$. Osserviamo che la matrice $\widehat{J} = D_\epsilon^{-1} J D_\epsilon$ è diagonale a blocchi con blocchi che differiscono dai blocchi di Jordan per il fatto che gli elementi sopra diagonali sono uguali a ϵ. Per cui la norma infinito di $\widehat{J}$ è data da $\rho(A)+\epsilon$ se esiste un blocco di Jordan di dimensione maggiore di 1 con autovalore di modulo uguale al raggio spettrale. Mentre, se tutti gli autovalori di modulo uguale al raggio spettrale stanno in blocchi di Jordan di dimensione 1 e se ϵ è scelto abbastanza piccolo in modo che $|\lambda| + \epsilon \leq \rho(A)$ per ogni altro autovalore λ, allora $\|\widehat{J}\|_\infty = \rho(A)$. Ne segue che il teorema vale con $\|A\| = \|A\|_S$ dove $S = D_\epsilon^{-1} W^{-1}$. Inoltre $\|A\| = \rho(A)$ se tutti gli autovalori di modulo $\rho(A)$ stanno in blocchi di dimensione 1. □

Un altro risultato che lega il raggio spettrale con le norme di matrici è dato dal seguente.

Teorema 6 *Per ogni norma di matrice $\|\cdot\|$ e per ogni matrice A vale*

$$\lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \rho(A)$$

Dim. La dimostrazione consiste di due parti. Una prima parte in cui si fa vedere che se il limite esiste allora non dipende dalla norma. Una seconda parte in cui si sceglie una norma speciale per cui il limite è proprio $\rho(A)$. Per la prima parte supponiamo di avere due norme $\|\cdot\|$ e $\|\cdot\|'$ e di sapere che esiste il $\lim_{k\to\infty} \|A^k\|^{1/k} = \ell$. Per il teorema di equivalenza delle norme si ha che esistono costanti α e β per cui vale la relazione

$$\alpha\|X\| \leq \|X\|' \leq \beta\|X\|$$

qualunque sia la matrice X. Scegliendo allora $X = A^k$ si ha

$$\alpha\|A^k\| \leq \|A^k\|' \leq \beta\|A^k\|$$

e, prendendo la radice k-esima delle tre espressioni si ottiene

$$\alpha^{1/k}\|A^k\|^{1/k} \leq (\|A^k\|')^{1/k} \leq \beta^{1/k}\|A^k\|^{1/k}$$

Prendendo il limite per $k \to \infty$, $\alpha^{1/k}$ e $\beta^{1/k}$ convergono ad 1 per cui la parte di sinistra e la parte di destra convergono allo stesso limite ℓ. Per il teorema dei carabinieri la parte centrale converge anch'essa ad ℓ.

Rimane ora da dimostrare che per una particolare norma vale $\ell = \rho(A)$. Per questo ricorriamo alla forma di Jordan $J = S^{-1}AS$ di A e, come abbiamo già fatto, consideriamo la norma infinito di J come norma particolare di A. Denotiamo quindi $\|A\| = \|J\|_\infty$. Valutiamo quindi $\|A^k\| = \|J^k\|_\infty$. Poiché la matrice J^k è diagonale a blocchi con le potenze k-esime dei blocchi di Jordan di A sulla diagonale principale, la norma infinito di J^k è la massima norma infinito delle potenze k-esime dei blocchi di Jordan di A. Sia $\widehat{J}$ un generico blocco di Jordan corrispondente all'autovalore λ. Se il blocco ha dimensione 1

la sua norma è $\|\widehat{J}^k\|_\infty = |\lambda|^k$, se il blocco ha dimensione m maggiore di 1 allora vale:

$$J^k = \begin{bmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \dots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \lambda^k & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda^{k-1} \\ & & & \lambda^k \end{bmatrix}$$

Ciò si vede applicando la formula del binomio di Newton al blocco di Jordan $\widehat{J}$ scritto nella forma $\widehat{J} = \lambda I + H$ dove $H = (h_{i,j})$ è la matrice che ha tutti elementi nulli tranne $h_{i,i+1} = 1$ per $i = 1, \dots, m-1$, essendo $H^k = 0$ per $k \geq m$. Quindi, se $\lambda = 0$ vale $J^k = 0$ per $k \geq m$. Se $\lambda \neq 0$ allora prendendo la norma infinito di $\widehat{J}^k$ si ottiene

$$\|\widehat{J}^k\| = \sum_{i=0}^{m-1} |\lambda|^{k-i} \binom{k}{i} = |\lambda|^k \sum_{i=0}^{m-1} |\lambda|^{-i} \binom{k}{i}$$

Ora, prendendo la radice k-esima dell'espressione precedente e prendendo il limite per $k \to \infty$ si ottiene

$$\lim_{k\to\infty} \left(\|\widehat{J}^k\|_\infty\right)^{1/k} = |\lambda| \lim_{k\to\infty} (\sum_{i=0}^{m-1} |\lambda|^{-i} \binom{k}{i})^{1/k} = \lambda$$

infatti la parte sotto radice k-esima, come funzione di k, è un polinomio di grado $m-1$ per cui $\lim_{k\to\infty}(\sum_{i=0}^{m-1} |\lambda|^{-i}\binom{k}{i})^{1/k} = 1$. □

Se A è tale che $\|A\| < 1$ per qualche norma di matrice indotta, allora, per la (7), A ha autovalori in modulo più piccoli di 1 e conseguentemente $I - A$ è invertibile. Una diseguaglianza utile che fornisce una stima di $\|(I-A)^{-1}\|$ è la seguente:

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}.$$

Infatti, da $(I-A)^{-1}(I-A) = I$ si ricava che

$$(I-A)^{-1} = (I-A)^{-1}A + I.$$

Prendendo le norme e usando la diseguaglianza triangolare si ottiene

$$\|(I-A)^{-1}\| \leq 1 + \|A\| \, \|(I-A)^{-1}\|$$

da cui la tesi.

Si osservi che sostituendo $-A$ ad A si ottiene $\|(I+A)^{-1}\| \leq 1/(1-\|A\|)$.

3 Numero di condizionamento

Lo studio del condizionamento di un sistema lineare $Ax = b$ ha come obiettivo capire in quali condizioni una piccola perturbazione introdotta nel vettore

dei termini noti o negli elementi della matrice A, si ripercuota più o meno amplificato nella soluzione. Uno strumento per fare questo sono i coefficienti di amplificazione. Ricordiamo che nel calcolo di una funzione $f(t_1, \ldots, t_n)$ di n variabili il coefficiente di amplificazione rispetto a t_i è dato da $t_i \frac{\partial f}{\partial t_i}/f$. Nel nostro caso le funzioni di cui calcolare i coefficienti di amplificazione sono $x_i(b_1, \ldots, b_n, a_{1,1}, \ldots, a_{n,n})$ per $i = 1, \ldots, n$ di $n^2 + n$ variabili dove $x = A^{-1}b$. usare questo strumento nel caso dei sistemi lineari porta ad una complicazione formale molto pesante. È allora più conveniente, anziché valutare le perturbazioni componente a componente, dare una valutazione globale usando le norme. Anche se le stime che otteniamo in questo modo sono meno precise, esse hanno il vantaggio di essere di uso più facile e immediato.

Siano $\delta_b \in \mathbb{C}^n$ e $\delta_A \in \mathbb{C}^{n\times n}$ perturbazioni che introduciamo rispettivamente in b e in A.

Assumendo $b \neq 0$ e $A \neq 0$, definiamo ora

$$\epsilon_b = \|\delta_b\|/\|b\|, \qquad \epsilon_A = \|\delta_A\|/\|A\|$$

le perturbazioni relative espresse in norma. Denotiamo con $x + \delta_x$ la soluzione del sistema perturbato, cioè tale che

$$(A + \delta_A)(x + \delta_x) = b + \delta_b$$

e definiamo $\epsilon_x = \|\delta_x\|/\|x\|$ la variazione relativa nel risultato conseguente alle due perturbazioni introdotte. Si osserva che se $\delta_A = 0$, dalle relazioni $Ax = b$ e $A(x + \delta_x) = b + \delta_b$ si ricava

$$A\delta_x = \delta_b$$

e, assumendo $\det A \neq 0$, si ricava $\delta_x = A^{-1}\delta_b$ da cui $\|\delta_x\| \leq \|A^{-1}\|\,\|\delta_b\|$. D'altra parte, poiché $Ax = b$ ne segue $\|b\| \leq \|A\|\,\|x\|$ da cui

$$\frac{\|\delta_x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\delta_b\|}{\|b\|}.$$

Cioè la perturbazione relativa in norma ϵ_b introdotta nel termine noto ha causato una variazione relativa in norma ϵ_x nella soluzione limitata da $\epsilon_x \leq \|A\|\,\|A^{-1}\|\epsilon_b$.

Il numero $\mu(A) = \|A\|\,\|A^{-1}\|$ è detto *numero di condizionamento* di A ed esprime la massima amplificazione che può subire l'errore introdotto nel termine noto.

Una maggiorazione analoga anche se un po' più complessa vale nel caso in cui $\delta_A \neq 0$.

Teorema 7 *Se* $\det A \neq 0$ *e* $\|A^{-1}\|\,\|\delta_A\| < 1$ *allora* $\det(A + \delta_A) \neq 0$*, inoltre*

$$\epsilon_x \leq \frac{\|A\|\,\|A^{-1}\|}{1 - \epsilon_A\|A\|\,\|A^{-1}\|}(\epsilon_b + \epsilon_A).$$

Dim. Dalle relazioni $Ax = b$ e $(A+\delta_A)(x+\delta_x) = b+\delta_b$, sottraendo membro a membro si ottiene $(A + \delta_A)\delta_x = -\delta_A x + \delta_b$. Poiché $A + \delta_A = A(I + A^{-1}\delta_A)$ e $\|A^{-1}\delta_A\| \leq \|A^{-1}\|\,\|\delta_A\| < 1$, la matrice $I + A^{-1}\delta_A$ risulta invertibile e vale

$$\delta_x = (I + A^{-1}\delta_A)^{-1}A^{-1}(-\delta_A x + \delta_b).$$

Poiché $\|(I + A^{-1}\delta_A)^{-1}\| \leq 1/(1 - \|A^{-1}\delta_A\|) \leq 1/(1 - \|A^{-1}\|\,\|\delta_A\|)$, risulta

$$\|\delta_x\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|\delta_A\|}(\|\delta_A\|\,\|x\| + \|\delta_b\|)$$

che assieme alla relazione $\|b\| \leq \|A\|\,\|x\|$ conduce al risultato □

Si osservi che in questo caso il coefficiente di amplificazione ha una forma più complessa. Però nel caso in cui ϵ_A è molto più piccolo di $\|A\|\,\|A^{-1}\|$, allora il coefficiente di amplificazione è ben approssimato dal numero di condizionamento $\mu(A)$ di A.

Se la matrice A è hermitiana e definita positiva, ordinando i suoi autovalori come

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n,$$

risulta $\|A\|_2 = \lambda_n$, $\|A^{-1}\|_2 = \lambda_1^{-1}$. Per cui

$$\mu_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\lambda_n}{\lambda_1}.$$

Cioè il numero di condizionamento in norma 2 è dato dal rapporto tra il massimo e il minimo autovalore si A.

Se A è hermitiana ma non è definita positiva, ordinando i suoi autovalori in modo che $|\lambda_i| \leq |\lambda_{i+1}|$ per $i = 1, \ldots, n-1$, vale $\|A\|_2\|A^{-1}\|_2 = \frac{|\lambda_n|}{|\lambda_1|}$, cioè il rapporto tra l'autovalore di massimo modulo e quello di minimo modulo di A.

Se A è normale allora dal fatto che la sua forma di Schur $D = Q^H AQ$ è diagonale segue che il numero condizionamento di A in norma 2 è ancora dato dal rapporto tra l'autovalore di massimo modulo e quello di minimo modulo.

Se A non è normale la situazione è più complicata. Prendiamo ad esempio la matrice

$$A = \begin{bmatrix} 1 & -a & & & \\ & 1 & -a & & \\ & & \ddots & \ddots & \\ & & & 1 & -a \\ & & & & 1 \end{bmatrix}$$

e valutiamo il numero di condizionamento in norma infinito. Vale

$$A^{-1} = \begin{bmatrix} 1 & a & a^2 & \ldots & a^{n-2} & a^{n-1} \\ & 1 & a & a^2 & \ddots & a^{n-2} \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & 1 & a & a^2 \\ & & & & 1 & a \\ & & & & & 1 \end{bmatrix}$$

da cui $\|A\|_\infty = 1 + |a|$, mentre, assumendo per semplicità $|a| \neq 1$ si ha $\|A^{-1}\|_\infty = \sum_{i=0}^{n-1} |a|^i = \frac{|a|^n - 1}{|a| - 1}$. Si osserva allora che, se $|a| > 1$ il numero

di condizionamento cresce esponenzialmente con n su base $|a|$. Ad esempio, se $a = 10$ e $n = 100$ il numero di condizionamento è dell'ordine di 10^{100}. Questa è la situazione tipica di un blocco di Jordan relativo ad un autovalore λ tale che $|\lambda| < 1$.

4 Esercizi

1. Sia $S \in \mathbb{R}^{m\times n}$, con $m \geq n$, e sia $\|\cdot\|$ una norma su $\mathbb{R}^m$. Si dica sotto quali ipotesi su S l'applicazione $x \longrightarrow \|Sx\|$ è una norma su $\mathbb{R}^n$.

2. Si dica, motivando la risposta, quali delle seguenti funzioni è una norma su $\mathbb{R}^n$:

   (a) $x \longrightarrow |x_1| + |x_n|$

   (b) $x \longrightarrow \max_i |x_i| + \min_i |x_i|$

   (c) $x \longrightarrow |x_1| + \max_i |x_i|$

   (d) su $\mathbb{R}^2$, $x \longrightarrow |x_1 - x_2| + |x_1|$

3. Dire se $\frac{1}{\sqrt{n}}\|\cdot\|_F$ è norma indotta.

4. Dimostrare che l'applicazione $A \to \max_{i,j} |a_{i,j}|$ soddisfa ai primi tre assiomi delle norme ma non è submultiplicativa.

5. Dimostrare che $\|A\|_F \leq \sqrt{r}\|A\|_2$ dove r è il rango di A.

6. Dimostrare che $\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}$.

7. Dimostrare che per una matrice unitaria il numero di condizionamento in norma 2 è 1.

8. Stimare il numero di condizionamento in norma 2 e in norma infinito di un blocco di Jordan $n \times n$ di autovalore λ.

9. Stimare il numero di condizionamento in norma 2 e in norma infinito di una matrice elementare, cioè del tipo $I - uv^T$ dove $u, v \in \mathbb{C}^n$. (Si veda l'articolo sulle matrici elementari)

10. Se A è matrice triangolare non singolare si dimostri che $\|A\|_2\|A^{-1}\|_2 \geq \max_i |a_{i,i}|/\min_i |a_{i,i}|$.

11. Si dimostri che una matrice non singolare A ha numero di condizionamento 1 in norma 2 se e solo se $A^H A = \alpha I$.

12. Si determinino costanti α e β tali che $\alpha\mu_\infty(A) \leq \mu_2(A) \leq \beta\mu_\infty(A)$.

13. Si dimostri che se A è hermitiana allora $\mu_2(A) \leq \mu(A)$, dove $\mu(A)$ è il numero di condizionamento rispetto a una qualsiasi norma indotta.

14. Si dica qual è il massimo numero di condizionamento in norma infinito di una matrice triangolare inferiore con elementi diagonali uguali a 1 ed elementi non diagonali di modulo minore o uguale a 1.

15. Se x è la soluzione del sistema $Ax = b$ e y è una sua approssimazione, posto $r = Ay - b$ il *residuo* di y, si dimostri che vale

$$\frac{\|x - y\|}{\|x\|} \leq \mu(A) \frac{\|r\|}{\|b\|}.$$

16. Siano $\|\cdot\|'$ e $\|\cdot\|''$ due norme su $\mathbb{R}^n$. Si consideri l'applicazione che alla matrice $n \times n$ reale A associa il numero reale $f(A) = \max_{\|x\|'=1} \|Ax\|''$. Dire quali proprietà della norma di matrici soddisfa questa applicazione. Si dimostri che scegliendo $\|\cdot\|' = \|\cdot\|_1$ e $\|\cdot\|'' = \|\cdot\|_\infty$ si ha $f(A) = \max_{i,j} |a_{i,j}|$.

17. Sia $A = \begin{bmatrix} 3 & -1 \\ 1 & 1 \end{bmatrix}$. Si dica se esiste una norma matriciale indotta $\|\cdot\|$ tale che:

    (a) $\|A\| = 4$

    (b) $\|A\| = 2 + 10^{-20}$

    (c) $\|A\| = 2$.

Riferimenti bibliografici

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

Fattorizzazioni LU e QR

Dario A. Bini, Università di Pisa

27 settembre 2019

Sommario

Questo modulo didattico contiene risultati e proprietà relativi alle fattorizzazioni LU e QR di una matrice. Si danno condizioni di esistenza e unicità della fattorizzazione LU e si mostra l'utilità di queste fattorizzazioni nella risoluzione di sistemi lineari.

Si consideri il sistema lineare $Ax = b$, dove A è una matrice $n \times n$ non singolare e $b \in \mathbb{R}^n$ è il vettore dei termini noti. Se la matrice A è triangolare inferiore, cioè se $a_{i,j} = 0$ per $i < j$, allora il sistema può essere facilmente risolto mediante il metodo di *sostituzione in avanti* definito dalle seguenti formule

$$\begin{aligned} &x_1 = b_1/a_{1,1}, \\ &x_i = (b_i - \textstyle\sum_{j=1}^{i-1} a_{i,j}x_j)/a_{i,i}, \quad i = 2, \ldots, n. \end{aligned}$$

In base a queste formule x_1 si ricava dalla prima equazione e si sostituisce nelle altre, x_2 si ricava dalla seconda equazione e si sostituisce nelle successive e via di seguito finché si ricava x_n dall'ultima equazione. Queste formule richiedono n^2 operazioni aritmetiche. Inoltre è facile dimostare che queste formule sono numericamente stabili all'indietro.

Un discorso analogo vale se la matrice è triangolare superiore, cioè se $a_{i,j} = 0$ se $i > j$. In questo caso le formule di risoluzione sono

$$\begin{aligned} &x_n = b_n/a_{n,n}, \\ &x_{n-i} = (b_{n-i} - \textstyle\sum_{j=i+1}^{n} a_{n-i,j}x_j)/a_{n-i,n-i}, \quad i = 1, \ldots, n-1, \end{aligned}$$

che definiscono il metodo di *sostituzione all'indietro*. Anche queste formule richiedono un numero di operazioni aritmetiche pari a n^2 e sono numericamente stabili all'indietro.

Un'altra situazione favorevole si incontra quando A è una matrice unitaria. Infatti in questo caso l'inversa di A coincide con A^H per cui la soluzione del sistema si ottiene come $x = A^H b$ e può essere calcolata con $2n^2 - n$ operazioni aritmetiche e il calcolo è numericamente stabile all'indietro.

Nel caso in cui A è una matrice arbitraria non singolare si può cercare di fattorizzare A nel prodotto di due o più matrici per le quali la risoluzione del sistema originale sia più facile. Infatti, se $A = BC$ è una fattorizzazione di A, dove B e C sono matrici $n \times n$, allora il sistema $Ax = b$ può essere riscritto come una coppia di sistemi da risolvere in successione:

$$\begin{cases} By = b \\ Cx = y \end{cases}$$

Infatti la risoluzione del primo sistema ci fornisce il vettore y che viene successivamente usato come termine noto nel secondo sistema la cui soluzione x coincide con la soluzione del sistema originale.

Se le matrici B e C sono triangolari o unitarie allora si può trarre vantaggio da quanto detto sopra e risolvere i due sistemi con un costo computazionale proporzionale a n^2. Il costo complessivo della risoluzione del sistema originale è allora la somma dei costi della risoluzione dei due sistemi più il costo del calcolo della fattorizzazione $A = BC$.

Le fattorizzazioni di matrici più studiate in letteratura sono le seguenti:

- fattorizzazione $A = LU$, dove L è matrice triangolare inferiore con elementi diagonali uguali a 1, U è matrice triangolare superiore;
- fattorizzazione $A = PLU$, dove L ed U sono come sopra mentre P è matrice di permutazione;
- fattorizzazione $A = P_1 L U P_2$, dove L e U sono come sopra mentre P_1, P_2 sono matrici di permutazione;
- fattorizzazione $A = QR$, dove Q è unitaria mentre R è triangolare superiore.

Forniamo ora una condizione di esistenza e unicità della fattorizzazione LU. Per questo è utile dare prima alcune definizioni.

Se $\Omega \subset \{1, 2, \ldots, n\}$, la matrice di elementi $a_{i,j}$ con $i, j \in \Omega$ è detta *sottomatrice principale* di A. Una sottomatrice principale ha elementi diagonali che sono anche elementi diagonali di A. Se $\Omega = \{1, 2, \ldots, k\}$ la sottomatrice di elementi con indici in Ω viene detta *sottomatrice principale di testa* di A.

Teorema 1 *Se tutte le sottomatrici principali di testa $k \times k$ di A sono non singolari per $k = 1, \ldots, n-1$ allora esiste ed è unica la fattorizzazione LU di A.*

Dim.

Si procede per induzione su n. Per $n = 1$ non c'e' nulla da dimostrare poiché $L = (1)$, $A = U = (a_{1,1})$. Assumiamo vera la tesi per dimensione $n-1$ e la dimostriamo per dimensione n. Cerchiamo quindi una fattorizzazione $A = LU$ che scriviamo nella seguente forma dopo aver partizionato opportunamente le matrici in blocchi:

$$\left[\begin{array}{c|c} A_{n-1} & b \\ \hline c^T & a_{n,n} \end{array}\right] = \left[\begin{array}{c|c} L_{n-1} & \\ \hline x^T & 1 \end{array}\right] \left[\begin{array}{c|c} U_{n-1} & y \\ \hline 0 & u_{n,n} \end{array}\right].$$

Uguagliando tra loro i quattro blocchi in entrambi i membri della espressione precedente si ottiene

$$A_{n-1} = L_{n-1}U_{n-1}, \quad b = L_{n-1}y,$$
$$c^T = x^T U_{n-1}, \qquad a_{n,n} = x^T y + u_{n,n}.$$

Poiché A_{n-1} ha sottomatrici principali di testa non singolari, per l'ipotesi induttiva esistono uniche matrici L_{n-1} e U_{n-1} tali che $A_{n-1} = L_{n-1}U_{n-1}$, dove L_{n-1} è triangolare inferiore con elementi diagonali uguali a 1, U_{n-1} è triangolare superiore. Inoltre, poiché L_{n-1} è triangolare con elementi diagonali uguali a 1, il suo determinante è uguale a 1 e quindi L_{n-1} è non singolare. Allora esiste unico il vettore $y = L_{n-1}^{-1}b$. Poiché $A_{n-1} = L_{n-1}U_{n-1}$ e A_{n-1} è non singolare per ipotesi, anche U_{n-1} risulta non singolare, quindi esiste unico $x^T = c^T U_{n-1}^{-1}$. Infine $u_{n,n}$ è dato in modo univoco dalla relazione $u_{n,n} = a_{n,n} - x^T y$. □

La condizione di non singolarità delle sottomatrici principali di testa di A data nel teorema precedente non è necessaria per l'esistenza della fattorizzazione LU come mostra il semplice esempio

$$\begin{bmatrix} 0 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

È facile dimostrare che se A è invertibile allora la condizione data nel teorema 1 è anche necessaria per l'esistenza della fattorizzazione LU. Si lascia questo per esercizio. Si può inoltre dimostrare che una fattorizzazione PLU esiste sempre qualunque sia la matrice A. Cioè permutando le righe di una qualsiasi matrice A in modo opportuno ci si può ricondurre ad una matrice che ammette una fattorizzazione LU.

Vedremo in un altro articolo come si possono costruire e analizzare algoritmi per il calcolo della fattorizzazione LU di una matrice A.

Per quanto riguarda la fattorizzazione QR daremo un metodo che calcola tale fattorizzazione qualunque sia la matrice A. Quello che è evidente è che la fattorizzazione QR non è unica. Ciò si vede con questo semplice ragionamento. Se $A = QR$ è una tale fattorizzazione e se D è una qualsiasi matrice diagonale con elementi di modulo 1 sulla diagonale principale, allora D è sia unitaria che triangolare superiore per cui $A = QDD^{-1}R = \widehat{Q}\widehat{R}$ con $\widehat{Q} = QD$, $\widehat{R} = D^{-1}R$ è ancora una fattorizzazione QR di A. Infatti $\widehat{Q}$ unitaria come prodotto di matrici unitarie e $\widehat{R}$ triangolare superiore come prodotto di matrici triangolari superiori.

Se A è invertibile allora si può dimostrare che la fattorizzazione QR è unica a meno di trasformazioni ottenute mediante matrici diagonali unitarie. Infatti, se $A = Q_1R_1 = Q_2R_2$, dall'ipotesi di nonsingolarità segue che $Q_2^H Q_1 = R_2R_1^{-1}$. Quindi la matrice triangolare superiore $R_2R_1^{-1}$ è unitaria ed è facile verificare che le uniche matrici triangolari unitarie sono quelle diagonali.

Esercizi

1. Si dimostri che se A è fortemente dominante diagonale allora esiste ed è unica la fattorizzazione LU di A.

2. Si dimostri che se A è hermitiana e definita positiva allora esiste ed è unica la fattorizzazione LU di A.

3. Si dimostri che se A è hermitiana e definita positiva allora esiste la fattorizzazione $A = LDL^T$, dove D è matrice diagonale con elementi diagonali positivi ed L è triangolare inferiore con elementi diagonali uguali a 1.

Riferimenti bibliografici

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

Matrici elementari e fattorizzazioni

Dario A. Bini, Università di Pisa

1 febbraio 2020

Sommario

Questo modulo didattico introduce ed analizza la classe delle matrici elementari. Tale classe verrà usata per costruire algoritmi di fattorizzazione di matrici e risolvere sistemi lineari.

1 Introduzione

Si introduce la classe delle matrici elementari che hanno proprietà computazionali interessanti e permettono di costruire algoritmi efficienti per calcolare le principali fattorizzazioni di matrici. Poi si considerano due sottoclassi particolari di matrici elementari: le matrici elementari di Gauss che sono triangolari inferiori, le matrici elementari di Householder che sono unitarie e hermitiane. Successivamente mostriamo come le matrici elementari possono essere usate per calcolare una generica fattorizzazione del tipo $A = SU$ dove S è un prodotto di matrici elementari e U è triangolare superiore. La specializzazione nella scelta delle matrici elementari alle matrici di Gauss e di Householder conduce ai metodi di Gauss e di Householder per la fattorizzazione LU e QR di una matrice.

2 Matrici elementari

Siano $u, v \in \mathbb{C}^n$. Per ragioni di chiarezza osserviamo che, con le nostre notazioni, l'espressione $v^H u = \sum_{i=1}^n \bar{v}_i u_i$ fornisce un numero complesso; mentre l'espressione uv^H, prodotto righe per colonne di un vettore colonna e di un vettore riga, fornisce una matrice $n \times n$ di elementi $u_i \bar{v}_j$. Ciò premesso, siamo pronti per dare la definizione di matrice elementare

Definizione 1 Una matrice del tipo

$$M = I - \sigma u v^H,$$

dove I è la matrice identica $n \times n$, σ un numero complesso e $u, v \in \mathbb{C}^n$, è detta *matrice elementare*.

Si osservi che nella definizione data c'è ridondanza di parametri, infatti il parametro scalare σ potrebbe essere inglobato in uno dei due vettori, inoltre uno dei due vettori potrebbe essere normalizzato a piacimento. Questa ridondanza ci permette una maggior facilità nell'analizzare le proprietà computazionali di questa classe di matrici.

Si osservi ancora che se $x \in \mathbb{C}^n$ allora $Mx = x - \sigma u(v^H x)$, per cui tutti i vettori dello spazio ortogonale a v, cioè tali che $v^H x = 0$, vengono trasformati da M in se stessi. Se d'altro canto $x = u$ allora $Mu = (1 - \sigma v^H u)u$, cioè $1 - \sigma v^H u$ è autovalore di M corrispondente all'autovettore u. In particolare se $\sigma v^H u = 1$ allora M è singolare. Viceversa, se M è singolare, esiste $x \neq 0$ tale che $Mx = 0$, cioè deve essere $x = \sigma(v^H x)u$, $v^H x \neq 0$. Per cui a meno di una costante moltiplicativa vale $x = u$, e $\sigma v^H u = 1$.

Studiamo ora alcune proprietà computazionali della classe di matrici elementari. Per semplicità assumiamo che $\sigma u v^H \neq 0$. Questa condizione esclude il caso in cui $M = I$ che non ha bisogno di ulteriore analisi.

Poiché M lascia l'intero sottospazio ortogonale a v invariato, e ha u come autovettore, così fa l'inversa di M se $\det M \neq 0$. Quindi viene naturale cercare l'inversa di M nella classe delle matrici elementari dello stesso tipo di M. Proviamo allora a cercare un numero complesso τ tale che $M^{-1} = (I - \tau u v^H)$. Cioè imponiamo la condizione $(I - \tau u v^H)M = I$, dove naturalmente supponiamo che $\sigma v^H u \neq 1$, condizione necessaria e sufficiente di invertibilità. Sviluppando i calcoli si ottiene

$$(\tau\sigma(v^H u) - \tau - \sigma)uv^H = 0$$

che è verificata se e solo se

$$\tau(1 - \sigma v^H u) = -\sigma.$$

Se $\sigma v^H u = 1$ abbiamo già osservato che la matrice non è invertibile, infatti l'equazione di sopra non ha soluzione. Se $\sigma v^H u \neq 1$ allora l'equazione ha soluzione

$$\tau = \frac{-\sigma}{1 - \sigma v^H u}.$$

Si può sintetizzare questo risultato col seguente

Teorema 1 *La matrice $M = I - \sigma u v^H$ è non singolare se e solo se $\sigma v^H u \neq 1$ e la sua inversa è data da*

$$I - \tau u v^H, \qquad \tau = \frac{-\sigma}{1 - \sigma v^H u}.$$

Il calcolo di τ richiede l'esecuzione di $n+1$ moltiplicazioni, n addizioni e una divisione. Un'altra proprietà interessante è riportata nel seguente

Teorema 2 *Data $M = I - \sigma u v^H$ matrice elementare e dato un vettore b, il calcolo del prodotto $y = Mb$ costa $2n + 1$ moltiplicazioni e $2n - 1$ addizioni. La risoluzione del sistema $Mx = b$ costa $3n + 2$ moltiplicazioni $3n - 1$ addizioni e una divisione.*

Una proprietà importante delle matrici elementari è che sono in grado di trasformare un qualsiasi vettore non nullo in un qualsiasi altro vettore non nullo come è precisato nel seguente

Teorema 3 *Per ogni $x, y \in \mathbb{C}^n \backslash 0$ esiste una matrice elementare $M = I - \sigma u v^H$ non singolare tale che $Mx = y$.*

Dim. Si procede in modo costruttivo. La condizione $(I - \sigma u v^H)x = y$ si riscrive come $\sigma u(v^H x) = x - y$. Basta quindi scegliere v non ortogonale a x e porre $\sigma u = (x - y)/(v^H x)$. Per avere la non singolarità di M basta imporre la condizione $\sigma v^H u \neq 1$, cioè $v^H(x - y)/(v^H x) \neq 1$, che è equivalente a $v^H y \neq 0$. Basta allora scegliere v in modo che non sia ortogonale né a x né a y. □

3 Matrici elementari di Gauss

Una matrice elementare di Gauss è ottenuta ponendo $v = e^{(1)}$, primo versore della base canonica, $\sigma = 1$ e u tale che $u_1 = 0$. Cioè per una matrice elementare di Gauss M vale

$$M = I - uv^H = \begin{bmatrix} 1 & & & 0 \\ -u_2 & 1 & & \\ \vdots & 0 & \ddots & \\ -u_n & & & 1 \end{bmatrix}$$

È facile verificare che $M^{-1} = I + uv^H$ cioè

$$M^{-1} = \begin{bmatrix} 1 & & & 0 \\ u_2 & 1 & & \\ \vdots & 0 & \ddots & \\ u_n & & & 1 \end{bmatrix}.$$

Quindi invertire una matrice elementare di Gauss non richiede alcuna operazione aritmetica. Basta cambiare segno agli elementi della prima colonna escluso l'elemento diagonale.

Si osservi che se $x = (x_i)$ è tale che $x_1 \neq 0$ allora esiste una matrice elementare di Gauss M tale che $Mx = x_1 e^{(1)}$. Infatti basta porre $u_i = x_i/x_1$, cioè

$$M = \begin{bmatrix} 1 & & & 0 \\ -x_2/x_1 & 1 & & \\ \vdots & 0 & 1 & \\ -x_n/x_1 & & & 1 \end{bmatrix}.$$

Si osservi che le matrici M e M^{-1} hanno entrambe norma infinito uguale a $1 + \max_i |x_i|/|x_1|$ quindi il numero di condizionamento in norma infinito di M

è $(1 + \max_i |x_i|/|x_1|)^2$. Se poi x è tale che $|x_1| = \max_i |x_i|$ allora il numero di condizionamento di M in norma infinito è al più 4. È quindi indipendente dalla dimensione n.

Un'altra classe di matrici elementari con numero di condizionamento indipendente da n è quella delle matrici di Householder.

4 Matrici elementari di Householder

Una matrice elementare di Householder è una matrice elementare hermitiana e unitaria. Vale quindi $M = I - \beta uu^H$ con $\beta = 0$ oppure $\beta = 2/(u^H u)$ e $u \neq 0$. Infatti, risulta

$$(I - \beta uu^H)(I - \beta uu^H)^H = (I - \beta uu^H)^2 = I - 2\beta uu^H + \beta^2(u^H u)uu^H = I.$$

Evidentemente l'inversa di una matrice di Householder M è M stessa. Inoltre in norma 2 il condizionamento di M è 1 poiché le matrici unitarie hanno norma 2 unitaria.

Non è difficile costruire una matrice di Householder che trasformi un vettore x non nullo in un vettore del tipo $\alpha e^{(1)}$. Infatti, si osserva subito che, essendo M unitaria essa trasforma i vettori lasciando invariata la norma 2, risulta quindi $|\alpha| = \|x\|_2$. Inoltre, poiché M è hermitiana, il valore di $x^H Mx$ è reale qualunque sia x. Questo implica che $x^H \alpha e^{(1)} = \bar{x}_1 \alpha$ è reale. Ciò permette di determinare il valore di α. Infatti di α conosciamo il modulo, quindi α è determinato a meno di un fattore complesso di modulo 1. Vale cioè $\alpha = \theta\|x\|_2$, con $|\theta| = 1$. Si verifica facilmente che se poniamo

$$\theta = \begin{cases} \pm x_1/|x_1| & \text{se } x_1 \neq 0, \\ \pm 1 & \text{se } x_1 = 0, \end{cases}$$

nel caso $x_1 \neq 0$ risulta

$$\bar{x}_1 \alpha = \pm \bar{x}_1 x_1/|x_1| \, \|x\|_2 = \pm |x_1| \, \|x\|_2 \in \mathbb{R},$$

se invece $x_1 = 0$ si ha $\bar{x}_1 \alpha = 0 \in \mathbb{R}$.

In teoria tutte e due i segni vanno bene, ma per ragioni di stabilità numerica vedremo tra poco che la scelta obbligata è il segno meno.

A questo punto siamo pronti per determinare il vettore u e conseguentemente lo scalare $\beta = 2/(u^H u)$. Infatti, dalla condizione $Mx = \alpha e^{(1)}$ si deduce che $(I - \beta uu^H)x = \alpha e^{(1)}$ e quindi

$$\beta(u^H x)u = x - \alpha e^{(1)}.$$

Ciò permette di determinare il vettore u a meno della sua lunghezza. Ma la lunghezza di u non è rilevante poichè questa informazione viene inglobata nel parametro β. Infatti basta porre

$$u = x - \alpha e^{(1)}$$

e ricavare β dalla condizione $\beta = 2/(u^H u)$. Infatti con questa scelta di β si verifica facilmente che $\beta(u^H x) = 1$. Si osserva che il vettore u ha tutte le componenti uguali a quelle di x tranne la prima che, nel caso $x_1 \neq 0$ è

$$u_1 = x_1 - \alpha = x_1 \mp \theta \|x\|_2 = x_1(1 \mp \|x\|_2/|x_1|).$$

È evidente che nella determinazione di θ conviene optare per il segno meno in modo che nella formula precedente non ci sia cancellazione numerica. Chiaramente, nel caso $x_1 = 0$ risulta $u_1 = -\alpha = \mp\|x\|_2$. In questo caso il segno non è influente ai fini della stabilità numerica. Per semplicità scegliamo ancora il segno meno. In questo modo si ha

$$u_1 = \begin{cases} x_1(1 + \frac{1}{|x_1|}\|x\|_2) & \text{se } x_1 \neq 0, \\ \|x\|_2 & \text{se } x_1 = 0, \end{cases}$$

e vale

$$\beta = 1/(\|x\|_2^2 + |x_1|\,\|x\|_2). \tag{1}$$

Il costo computazionale del calcolo di u e di β è dominato dal calcolo di $\|x\|_2$, cioè n moltiplicazioni e $n-1$ addizioni più una estrazione di radice. Si osservi inoltre che da (1) segue che $\beta(u^H x) = 1$ come richiesto.

5 Fattorizzazione mediante matrici elementari

Mostriamo come sia possibile utilizzare le matrici elementari per realizzare metodi per fattorizzare una matrice A nel prodotto $A = SU$ dove

$$S = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$$

è un prodotto di matrici elementari ed U è una matrice triangolare superiore. La disponibilità di tale fattorizzazione ci permette di risolvere agevolmente il sistema $Ax = b$ attraverso la risoluzione dei due sistemi $Sy = b$ e $Ux = y$. Il secondo, essendo con matrice triangolare superiore, si risolve con n^2 operazioni aritmetiche. Il primo permette di esprimere la soluzione mediante la formula

$$y = M_{n-1} \cdots M_1 b$$

e quindi, se le matrici elementari $M_1, \ldots, M_{n-1}$ sono disponibili, permette di calcolare la soluzione in circa $3n^2$ moltiplicazioni e altrettante addizioni grazie al teorema 2. Cioè, data la fattorizzazione di A, il sistema $Ax = b$ è risolubile in $O(n^2)$ operazioni aritmetiche. Quindi tutto è ricondotto al calcolo della fattorizzazione $A = SU$. Vediamo ora come si realizza.

Posto $A_1 = A$ andiamo a generare una successione di matrici $A_k = (a_{i,j}^{(k)})$, $k = 1, 2 \ldots, n$ tali che $A_{k+1} = M_k A_k$ dove M_k è una matrice elementare e dove A_k ha le prime $k-1$ colonne in forma triangolare, cioè

$$A_k = \left[\begin{array}{ccc|ccc} a^{(k)}_{1,1} & \dots & a^{(k)}_{1,k-1} & a^{(k)}_{1,k} & \dots & a^{(k)}_{1,n} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & a^{(k)}_{k-1,k-1} & a^{(k)}_{k-1,k} & \dots & a^{(k)}_{k,n} \\ \hline 0 & \dots & 0 & a^{(k)}_{k,k} & \dots & a^{(k)}_{k+1,n} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a^{(k)}_{n,k} & \dots & a^{(k)}_{n,n} \end{array}\right] =: \left[\begin{array}{c|c} T_k & V_k \\ \hline 0 & W_k \end{array}\right] \tag{2}$$

dove T_k ha dimensione $(k-1)\times(k-1)$, W_k ha dimensione $(n-k+1)\times(n-k+1)$ e V_k ha dimensione $(k-1)\times(n-k+1)$.

Vediamo il primo passo. Consideriamo la matrice $A = A_1$, denotiamo $a^{(1)}$ la sua prima colonna. Per il teorema 3 esiste una matrice elementare M_1 che trasforma $a^{(1)}$ in un vettore proporzionale al primo vettore $e^{(1)}$ della base canonica di $\mathbb{C}^n$, cioè M_1 è tale che $M_1 a^{(1)} = a^{(2)}_{1,1} e^{(1)}$. Vale allora

$$M_1 A_1 = \left[\begin{array}{c|c} a^{(2)}_{1,1} & V_2 \\ \hline 0 & W_2 \end{array}\right] =: A_2$$

dove W_2 è matrice $(n-1)\times(n-1)$ mentre V_2 ha dimensioni $1\times(n-1)$. Cioè mediante la moltiplicazione per M_1 abbiamo iniziato il primo passo del processo di triangolarizzazione di A. Adesso ripetiamo lo stesso procedimento sulla matrice W_2 costruendo una matrice elementare che trasformi la prima colonna di W_2 in un vettore proporzionale al primo versore della base canonica di $\mathbb{C}^{n-1}$. Descriviamo questo processo nella sua generalità.

Supponiamo di avere la matrice A_k e di voler costruire la matrice M_k tale che $A_{k+1} = M_k A_k$ abbia le prime k colonne in forma triangolare. Per questo si consideri una matrice elementare $\widehat{M}_k$ di dimensione $(n-k+1)\times(n-k+1)$ che trasformi la prima colonna di W_k in un vettore proporzionale al primo versore della base canonica di $\mathbb{C}^{n-k+1}$. Tale matrice esiste per il teorema 3. Vale allora

$$\widehat{M}_k W_k = \left[\begin{array}{c|c} a^{(k+1)}_{k,k} & z^T \\ \hline 0 & W_{k+1} \end{array}\right]$$

con W_{k+1} di dimensioni $(n-k)\times(n-k)$.

Quindi ponendo

$$M_k = \left[\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & \widehat{M}_k \end{array}\right] \tag{3}$$

dove I_{k-1} è una matrice identica di ordine $k-1$, si ha

$$M_k A_k = \left[\begin{array}{c|c} T_k & V_k \\ \hline 0 & \widehat{M}_k W_k \end{array}\right] = \left[\begin{array}{c|cc} T_k & \multicolumn{2}{c}{V_k} \\ \hline 0 & a^{(k+1)}_{k,k} & z^T \\ & 0 & W_{k+1} \end{array}\right] := A_{k+1}$$

dove A_{k+1} ha la stessa struttura in (2) con k sostituito da $k+1$.

Inoltre la matrice M_k è ancora una matrice elementare essendo $M_k = I - \sigma u v^H$ dove

$$u = \begin{bmatrix} 0 \\ \hat{u} \end{bmatrix}, \qquad v = \begin{bmatrix} 0 \\ \hat{v} \end{bmatrix}.$$

con $\widehat{M}_k = I_{n-k+1} - \sigma \hat{u}\hat{v}^H$.

In questo modo si è realizzato il generico passo del processo di triangolarizzazione di A.

Dopo $n-1$ passi si ottiene la fattorizzazione $A_n = M_{n-1} \cdots M_1 A$, dove A_n è triangolare superiore, da cui si ottiene

$$A = M_1^{-1} \cdots M_{n-1}^{-1} A_n.$$

Si osservi che per risolvere il sistema lineare $Ax = b$, la tecnica di fattorizzazione appena introdotta può essere utilizzata in due modi diversi ma equivalenti.

- Calcolare e memorizzare ogni singola matrice elementare M_k assieme alla matrice A_n triangolare superiore e alla fine calcolare $y = M_{n-1} \cdots M_1 b$ mediante prodotti successivi, e poi risolvere il sistema triangolare $A_n x = y$.
- Costruire la successione di sistemi equivalenti $A_k x = b^{(k)}$, dove $b^{(k+1)} = M_k b^{(k)}$ e $b^{(1)} = b$, e alla fine risolvere il sistema $A_n x = b^{(n)}$.

La differenza tra i due approcci riguarda solo la tempistica delle operazioni.

Nel secondo caso si applica una strategia usa e getta nel calcolo delle matrici M_k che comporta un ingombro di memoria più basso rispetto al primo approccio.

Computazionalmente i due metodi sono equivalenti visto che entrambi calcolano, anche se in tempi diversi, gli $n-1$ prodotti matrice-vettore $y = b^{(n)} = M_{n-1} \cdots M_1 b$.

6 Metodi di Gauss e di Householder

Se ad ogni passo del metodo descritto sopra si sceglie come M_k una matrice del tipo (3) dove $\widehat{M}_k = I - \hat{\beta}_k \hat{u}^{(k)} \hat{u}^{(k)H}$ è matrice di Householder, allora anche M_k è matrice di Householder essendo $M_k = I - \beta_k u^{(k)} u^{(k)H}$ con

$$u^{(k)} = \begin{bmatrix} 0 \\ \hat{u}^{(k)} \end{bmatrix}$$

e $\beta_k = \hat{\beta}_k$. La fattorizzazione che si ottiene in questo modo è una fattorizzazione QR dove Q è un prodotto di matrici di Householder e quindi è unitaria.

Poiché esiste sempre una matrice di Householder che trasforma un arbitrario vettore in un vettore proporzionale al primo versore della base canonica, la costruzione mostrata può essere sempre portata a termine senza interruzioni.

Questo dimostra in modo costruttivo che la fattorizzazione QR di una matrice esiste sempre.

Scegliendo invece ad ogni passo come matrice elementare la matrice M_k descritta in (3) dove $\widehat{M}_k$ è una matrice di Gauss si arriva alla fattorizzazione LU.

Si osserva che la matrice $\widehat{M}_k$ ha elementi

$$\widehat{M}_k = \begin{bmatrix} 1 & & & 0 \\ -a_{2,k}^{(k)}/a_{k,k}^{(k)} & 1 & & \\ \vdots & & \ddots & \\ -a_{n,k}^{(k)}/a_{k,k}^{(k)} & & & 1 \end{bmatrix}. \tag{4}$$

Diversamente dalle matrici di Householder, l'esistenza della matrice elementare di Gauss è legata alla condizione $a_{k,k}^{(k)} \neq 0$. Infatti se tale condizione non fosse verificata la matrice di Gauss (4) che realizza un passo del processo di triangolarizzazione non esisterebbe in generale. L'elemento $a_{k,k}^{(k)}$ viene chiamato *elemento pivot*.

È possibile verificare agevolmente che se tutte le sottomatrici principali di testa di A di dimensione $k \times k$ sono non singolari per $k = 1, \ldots, n-1$, per cui esiste ed è unica la fattorizzazione LU, allora tutti gli elementi pivot sono diversi da zero e quindi il processo di triangolarizzazione può essere portato a termine senza interruzioni.

Infatti, supponiamo per assurdo che il metodo di fattorizzazione LU appena visto si interrompa al passo k-esimo poiché $a_{k,k}^{(k)} = 0$.

Allora, dalla relazione $A_k = M_{k-1} \cdots M_1 A$, poiché $L_k = M_{k-1} \cdots M_1$è triangolare inferiore si ha che la sottomatrice principale di testa di A_k di dimensione $k \times k$ è uguale alla sottomatrice principale di testa di L_k per la sottomatrice principale di testa di A che è non singolare per ipotesi.

Questo implica che la sottomatrice principale di testa $k \times k$ di A_k è non singolare. Ma questo è assurdo essendo tale matrice triangolare superiore con elementi diagonali $a_{i,i}^{(k)}$ ed essendo $a_{k,k}^{(k)} = 0$.

I metodi per il calcolo della fattorizzazione LU e QR che si ottengono nel modo descritto sono detti rispettivamente metodo di Gauss, o metodo di eliminazione Gaussiana, e metodo di Householder. Una loro analisi computazionale viene svolta nel prossimo articolo Aspetti computazionali dei metodi di Gauss e Householder.

7 Il complemento di Schur

Si partizioni la matrice A nel seguente modo

$$A = \left[\begin{array}{c|c} A_{1,1} & A_{1,2} \\ \hline A_{2,1} & A_{2,2} \end{array}\right]$$

dove $A_{1,1}$ è $k \times k$ e non singolare. La matrice

$$S = A_{2,2} - A_{2,1} A_{1,1}^{-1} A_{1,2}$$

è definita il *complemento di Schur* di $A_{2,2}$ in A. Il complemento di Schur è legato alla fattorizzazione LU di A. Infatti si verifica facilmente che vale la fattorizzazione a blocchi

$$A = \begin{bmatrix} I & 0 \\ A_{2,1}A_{1,1}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & S \end{bmatrix}. \tag{5}$$

Inoltre S coincide con la matrice W_k che compare in (2).

Il complemento di Schur ha proprietà interessanti. Ad esempio, da (5) segue che $\det A = \det A_{1,1} \det S$. Quindi se A è invertibile anche S lo è. Inoltre S^{-1} coincide con la sottomatrice principale di A^{-1} formata dagli indici $i, j > k$.

Riferimenti bibliografici

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

Aspetti computazionali dei metodi di Gauss e di Householder

Dario A. Bini, Università di Pisa

18 agosto 2019

Sommario

Questo modulo didattico contiene una discussione sugli aspetti computazionali dei metodi di Gauss e di Householder per risolvere sistemi lineari. Particolare attenzione viene rivolta a questioni di complessità e di stabilità numerica.

1 Introduzione

In questo articolo esaminiamo alcuni aspetti computazionali e implementativi relativi ai metodi di Gauss e di Householder per il calcolo delle fattorizzazioni LU e QR di una matrice A e per la risoluzione di un sistema lineare $Ax = b$. Analizzeremo il costo computazionale e la stabilità numerica di questi metodi. In particolare, per il metodo di Gauss vedremo che il costo è dell'ordine di $\frac{2}{3}n^3$ operazioni aritmetiche, e, da un'analisi all'indietro dell'errore vedremo che non ci sono garanzie di stabilità numerica per tale metodo. Per questo introdurremo le strategie di massimo pivot che permettono di rendere il metodo di eliminazione Gaussiana numericamente affidabile.

Per quanto riguarda il metodo di Householder, un'analisi computazionale ci mostra che il costo computazionale è sensibilmente più alto, cioè dell'ordine di $\frac{4}{3}n^3$ operazioni aritmetiche, e che il metodo non incontra problemi di stabilità numerica anche se applicato senza strategie di massimo pivot. Discuteremo infine su come applicare tali metodi per il calcolo del determinante, dell'inversa e del rango di una matrice.

2 Metodo di eliminazione Gaussiana

Nel calcolo della fattorizzazione LU di una matrice A si genera una successione di matrici A_k tali che $A_{k+1} = M_k A_k$, dove $M_k = (m_{i,j}^{(k)})$ con

$$m_{i,j}^{(k)} = \begin{cases} 1 & \text{per } i = j, \\ -\frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} & \text{per } j = k, i > k \\ 0 & \text{altrove} \end{cases} \tag{1}$$

e dove

$$A_k = \left[\begin{array}{ccc|ccc} a_{1,1}^{(k)} & \dots & a_{1,k-1}^{(k)} & a_{1,k}^{(k)} & \dots & a_{1,n}^{(k)} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & a_{k-1,k-1}^{(k)} & a_{k-1,k}^{(k)} & \dots & a_{k,n}^{(k)} \\ \hline 0 & \dots & 0 & a_{k,k}^{(k)} & \dots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a_{n,k}^{(k)} & \dots & a_{n,n}^{(k)} \end{array}\right] \tag{2}$$

La relazione $A_{k+1} = M_k A_k$ trascritta in componenti si riduce a

$$a_{i,j}^{(k+1)} = \begin{cases} a_{i,j}^{(k)} & \text{se } j < k, \text{ oppure } i \leq k \\ 0 & \text{se } j = k \text{ e } i > k \\ a_{i,j}^{(k)} + m_{i,k}^{(k)} a_{k,j}^{(k)} & \text{se } i > k,\ j > k \end{cases} \tag{3}$$

L'algoritmo per il calcolo della fattorizzazione LU è ricondotto ad applicare le (1) e (3). Se utilizziamo la stessa variabile $\mathtt{A} = (\mathtt{a_{i,j}})$ per memorizzare tutti i valori di $A^{(k)}$ e la stessa variabile $\mathtt{M} = (\mathtt{m_{i,j}})$ per memorizzare i valori degli $m_{i,k}^{(k)}$, l'algoritmo viene sintetizzato dalle relazioni seguenti

Algoritmo 1

1. *Per* $\mathtt{k} = 1, \dots, \mathtt{n} - 1$
2. *per* $\mathtt{i} = \mathtt{k} + 1, \dots, \mathtt{n}$
3. $\mathtt{m_{i,k}} = -\mathtt{a_{i,k}}/\mathtt{a_{k,k}}$
4. *per* $\mathtt{j} = \mathtt{k} + 1, \dots, \mathtt{n}$
5. $\mathtt{a_{i,j}} = \mathtt{a_{i,j}} + \mathtt{m_{i,k}}\mathtt{a_{k,j}}$
6. *Fine ciclo* $\mathtt{j}$
7. *Fine ciclo* $\mathtt{i}$
8. *Fine ciclo* $\mathtt{k}$

Nel caso della risoluzione di un sistema lineare $Ax = b$ l'algoritmo può essere modificato aggiungendo il calcolo di $b^{(k+1)} = M_k b^{(k)}$ mediante l'istruzione

$$\mathtt{b_i} = \mathtt{b_i} + \mathtt{m_{i,k}}\mathtt{b_k}$$

che va inserita subito dopo aver calcolato $\mathtt{m_{i,k}}$. In questo modo la soluzione del sistema si ottiene risolvendo il sistema triangolare $A_k x = b^{(k)}$.

Alla fine dell'applicazione dell'algoritmo 1 la matrice A_{n-1} contenuta nella variabile $\mathtt{A}$ conterrà gli elementi della matrice U della fattorizzazione $A = LU$. Non è difficile verificare che la matrice L ha elementi dati da

$$\ell_{i,k} = -m_{i,k} = a_{i,k}^{(k)}/a_{k,k}^{(k)}.$$

Ciò segue dal fatto che

$$L = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$$

e che M_k differisce dalla matrice identica solo nella colonna k-esima in cui gli elementi di indice maggiore di k sono $a_{i,k}^{(k)}/a_{k,k}^{(k)}$.

2.1 Costo computazionale

Al passo k-esimo dell'algoritmo descritto sopra viene richiesto il calcolo delle quantità $\mathtt{a_{i,j}}$ per $\mathtt{i,j} = \mathtt{k+1},\ldots,\mathtt{n}$ che comporta l'esecuzione di $2(n-k)^2$ operazioni aritmetiche. Altre $n-k$ divisioni sono richieste per il calcolo di $\mathtt{m_{i,k}}$ per $\mathtt{i} = \mathtt{k+1},\ldots,\mathtt{n}$. In totale si arriva ad un costo computazionale di

$$C_n = \sum_{k=1}^{n-1} \big(2(n-k)^2 + (n-k)\big) = \sum_{k=1}^{n-1} (2k^2 + k)$$

operazioni aritmetiche. Usando le formule

$$\sum_{i=1}^{m} i^2 = \frac{m^3}{3} + \frac{m^2}{2} + \frac{m}{6}, \quad \sum_{i=1}^{m} i = \frac{m(m+1)}{2},$$

si arriva al costo

$$C_n \doteq \frac{2}{3} n^3$$

dove $\doteq$ indica l'uguaglianza a meno di termini di ordine inferiore ad n^3.

2.2 Stabilità numerica e strategie del massimo pivot

Applicando gli strumenti standard dell'analisi degli errori è possibile condurre una analisi all'indietro del metodo di fattorizzazione LU descritto dall'algoritmo 1 si arriva a dimostrare il seguente risultato

Teorema 1 *Sia A una matrice $n \times n$ con sottomatrici principali di testa fino all'ordine $n-1$ non singolari per cui esiste unica la fattorizzazione $A = LU$. Se $\widetilde{L}$ e $\widetilde{U}$ sono le matrici effettivamente calcolate applicando l'algoritmo 1 in aritmetica floating point con precisione di macchina u, allora vale*

$$A + \Delta_A = \widetilde{L}\widetilde{U}$$

con

$$|\Delta_A| \leq 2nu(|A| + |\tilde{L}|\,|\tilde{U}|) + O(u^2)$$

dove si è indicato con $|A|$ la matrice i cui elementi sono $|a_{i,j}|$ e dove la diseguaglianza tra matrici va intesa elemento per elemento. Inoltre, se $\tilde{y}$ è la soluzione

del sistema $\widetilde{L}y = b$ *effettivamente calcolata in aritmetica floating point mediante sostituzione in avanti e* $\tilde{x}$ *è il vettore effettivamente calcolato risolvendo il sistema triangolare* $\widetilde{U}x = \tilde{y}$ *in aritmetica floating point mediante sostituzione all'indietro, vale*

$$(A + \widehat{\Delta}_A)\tilde{x} = b$$

con

$$|\widehat{\Delta}_A| \leq 4nu(|A| + |\widetilde{L}|\,|\widetilde{U}|) + O(u^2).$$

Questo risultato ci dice che i valori effettivamente calcolati $\widetilde{L}$ e $\widetilde{U}$ sono i fattori L ed U esatti di una matrice ottenuta perturbando A.

L'entità della perturbazione, nella componente proporzionale ad u, è lineare in n ma dipende anche dalla grandezza in modulo degli elementi di L e di U, di cui $\widetilde{L}$ e $\widetilde{U}$ sono i valori effettivamente calcolati, sulla quale non abbiamo alcun controllo a priori. Per cui se con certi dati gli elementi $a_{i,j}^{(k)}$ delle matrici A_k o gli elementi $m_{i,k} = -a_{i,k}^{(k)}/a_{k,k}^{(k)}$ prendono valori molto elevati in modulo rispetto ai valori iniziali di A, allora non dobbiamo aspettarci un buon comportamento numerico del metodo di eliminazione gaussiana. Questo vale sia per il calcolo della fattorizzazione che per la risoluzione del sistema lineare.

Ciò accade ad esempio se per qualche k un elemento pivot $a_{k,k}^{(k)}$ prende valori piccoli in modulo rispetto agli altri elementi $a_{i,k}^{(k)}$. In questo caso i rapporti $m_{i,k} = -a_{i,k}^{(k)}/a_{k,k}^{(k)}$ potrebbero prendere valori molto grandi in modulo.

Un modo per rimuovere questa eventualità consiste nell'effettuare una permutazione di righe alla matrice A_k prima di proseguire col passo di triangolarizzazione. Più precisamente, si sceglie un indice h per cui $|a_{h,k}| \geq |a_{i,k}|$ per ogni altro indice $i \geq k$. Successivamente si permuta la riga h-esima con la riga k-esima, ottenendo una nuova matrice che ha sempre la stessa struttura triangolare a blocchi (2) ma dove l'elemento pivot ha modulo maggiore o uguale a quello degli elementi sulla stessa colonna sotto la diagonale. Se indichiamo con P_k la matrice di permutazione che effettua lo scambio delle componenti k e h, allora il generico passo del metodo così modificato diventa

$$A_{k+1} = M_k P_k A_k.$$

Ora si osserva che se $M_i = I - v^{(i)}e^{(i)T}$, con $i < k$, allora

$$P_k M_i = \widetilde{M}_i P_k,$$

dove $\widetilde{M}_i = I - \tilde{v}^{(i)}e^{(i)T}$ e $\tilde{v}^{(i)} = P_k v^{(i)}$. Per cui, se k è il primo intero in cui si effettua la permutazione delle righe, dalla relazione $A_k = M_{k-1}\cdots M_1 A$ si deduce che

$$A_{k+1} = M_k P_k A_k = M_k P_k M_{k-1}\cdots M_1 A = M_k \widetilde{M}_{k-1}\cdots \widetilde{M}_1 P_k A.$$

In modo analogo si vede induttivamente che, se la permutazione viene applicata ad ogni passo del metodo, vale

$$A_{k+1} = M_k \widetilde{M}_{k-1}\cdots \widetilde{M}_1 P_k P_{k-1}\cdots P_1 A.$$

Cioè, alla fine del procedimento si ottiene

$$A_n = M_{n-1}\widetilde{M}_{n-2}\cdots\widetilde{M}_1(P_{n-1}\cdots P_1)A$$

da cui la fattorizzazione LU della matrice PA:

$$PA = LU, \qquad P = P_{n-1}\cdots P_1,$$

con P matrice di permutazione, o equivalentemente la fattorizzazione del tipo PLU: $A = P^T LU$.

Questa strategia, detta del *massimo pivot parziale* ha il vantaggio di contenere la potenziale crescita degli $|m_{i,k}|$ risultando $|m_{i,k}| \leq 1$. Dalla pratica del calcolo questa strategia fornisce un sostanziale miglioramento della stabilità numerica dell'eliminazione gaussiana anche se, come vedremo tra poco, non dà garanzie assolute di stabilità.

Un altro vantaggio di questa strategia è che permette di portare avanti il procedimento anche se nel calcolo si dovesse incontrare un pivot nullo. Infatti, il processo così come lo abbiamo descritto, può subire un arresto solo nel caso in cui al passo k-esimo risulta $a_{i,k}^{(k)} = 0$ per $i = k, \ldots, n$. Ma in questo caso non è più necessario svolgere il k-esimo passo di triangolarizzazione essendo la k-esima colonna della matrice in forma già triangolare. Basta quindi saltare il k-esimo passo e continuare il processo dal passo $k+1$. Chiaramente ciò accade solo se la matrice è singolare. A causa degli errori locali generati dall'aritmetica floating point le quantità che in teoria devono risultare nulle, nel calcolo vengono rappresentate da valori di modulo piccolo ma in generale non nullo. È quindi impossibile numericamente stabilire se una quantità calcolata in aritmetica floating point che risulta piccola in modulo rappresenti un valore nullo o un valore piccolo.

Come si è già detto, il fatto che con la strategia del massimo pivot parziale risulti $|m_{i,k}| \leq 1$, migliora in pratica le prestazioni numeriche dell'algoritmo di eliminazione gaussiana. Però questo non garantisce nulla sulla limitatezza uniforme del fattore U. Infatti, dalla relazione (1) si deduce che con la strategia del massimo pivot parziale vale

$$|a_{i,j}^{(k+1)}| \leq |a_{i,j}^{(k)}| + |m_{i,k}|\,|a_{k,j}^{(k)}| \leq |a_{i,j}^{(k)}| + |a_{k,j}^{(k)}|.$$

Quindi, se denotiamo con

$$g_k(A) = \max_{i,j} |a_{i,j}^{(k)}| / \max_{i,j} |a_{i,j}| \tag{4}$$

si ha che

$$g_{k+1}(A) \leq 2g_k(A).$$

Ciò conduce alla limitazione

$$g_n(A) = \max_{i,j} |u_{i,j}| / \max_{i,j} |a_{i,j}| \leq 2^{n-1}.$$

Purtroppo questa limitazione superiore ha una crescita esponenziale in n che non promette nulla di buono. Inoltre la maggiorazione data non è lasca come

uno potrebbe sperare, infatti viene raggiunta dal seguente esempio che si riporta nel caso di $n = 5$:

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

Si vede immediatamente che dopo un passo del metodo di eliminazione gaussiana in cui non si effettuano permutazioni poiché l'elemento pivot ha sempre modulo massimo, la matrice A_1 risulta

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & -1 & 1 & 0 & 2 \\ 0 & -1 & -1 & 1 & 2 \\ 0 & -1 & -1 & -1 & 2 \end{bmatrix}$$

Come si può vedere, gli elementi dell'ultima colonna, tranne il primo sono raddoppiati. Procedendo ulteriormente si ha

$$A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & -1 & 1 & 4 \\ 0 & 0 & -1 & -1 & 4 \end{bmatrix}$$

e alla fine l'elemento di posto n, n vale $a_{n,n}^{(n)} = 2^{n-1}$.

Per fortuna casi come questo non si presentano nelle applicazioni importanti e questo esempio è più un artificio matematico che non un esempio tratto da un problema reale.

Esiste però un rimedio efficace che permette di raggiungere una maggior stabilità numerica. Si tratta della strategia del *massimo pivot totale*. Questa strategia consiste nell'operare nel seguente modo:

- al passo k-esimo si selezionano due indici $r, s \geq k$ tali che $|a_{r,s}^{(k)}| \geq |a_{i,j}^{(k)}|$ per ogni $i, j \geq k$;
- si scambiano le righe k ed r e le colonne k ed s di $A^{(k)}$;
- si prosegue con l'eliminazione gaussiana sulla matrice ottenuta.

Procedendo in questo modo si mantiene sempre un elemento pivot che ha modulo maggiore o uguale ai moduli degli elementi di indice $i, j \geq k$. Si può dimostrare che la crescita della quantità $g_k(A)$ definita in (4) ottenuta con la strategia del massimo pivot totale è limitata da

$$g_k(A) \leq \sqrt{n \prod_{j=2}^{n} j^{1/(j-1)}}. \tag{5}$$

Tale funzione ha una crescita molto più contenuta rispetto all'esponenziale. Ad esempio, per $n = 100$ la maggiorazione in (5) è all'incirca 3500.

Si conoscono rarissimi esempi di matrici $n \times n$, con $n = 18, 20, 25$ in cui il fattore $g_n(A)$ raggiunge valori di poco superiori a n. Nel resto dei casi noti il valore di $g_n(A)$ non supera n. Lo studio della crescita di $g_n(A)$ è un problema aperto non facile da risolvere.

Nel caso di una matrice A singolare, la strategia del massimo pivot totale si arresta quando la matrice A_k ha elementi nulli per $i, j \geq k$. In questo caso possiamo dire che il rango di A è $k-1$. Non solo, ma la fattorizzazione $P_1 A P_2 = LU$ che si trova in questo modo fornisce anche una base per il nucleo (spazio nullo) di A e per l'immagine di A. Questo fatto può essere utile in un calcolo in cui non vengono generati errori nelle operazioni. Infatti, in un calcolo in aritmetica floating point, a causa degli errori locali non siamo in grado di dire se un valore calcolato corrisponde a una quantità nulla o ad una di modulo piccolo.

Si osserva che il metodo di fattorizzazione LU può essere applicato anche a matrici di interi generando matrici A_k razionali. Per matrici razionali è possibile mantenere una aritmetica senza errori locali rappresentando ciascun razionale come coppia di interi ed operando sui razionali con una aritmetica intera.

In questo modello di calcolo si incontra però un altro problema. Durante lo svolgimento di ciascuna operazione il numero di cifre con cui vengono rappresentati gli interi che danno il numeratore e il denominatore dei numeri razionali aumenta proporzionalmente al numero di operazioni aritmetiche. Ciò può rendere il calcolo estremamente lento. Si pensi ad esempio alla somma di due numeri del tipo $1/a + 1/b = (a + b)/ab$. Nel caso peggiore il denominatore non si semplifica col numeratore e l'intero ab ha un numero di cifre pari a circa la somma delle cifre di a e quelle di b.

Esiste una variante dell'eliminazione gaussiana nota come *metodo di Bareiss* che applicata ad una matrice di elementi interi mantiene il calcolo sugli interi. Questo metodo ha l'inconveniente di produrre interi il cui numero di cifre cresce proporzionalmente alla dimensione della matrice.

Con la variante di Bareiss, o con una aritmetica razionale esatta, è possibile calcolare il rango di una matrice seppur ad un costo computazionale che può essere molto elevato.

Esiste un approccio poco costoso per calcolare il rango di una matrice di interi che ha il prezzo di fornire un risultato non garantito al 100% ma comunque altamente affidabile. Si basa sugli algoritmi che usano la casualità quali gli algoritmi Montecarlo e gli algoritmi LasVegas. Un algoritmo Montecarlo usa al suo interno dei numeri generati in modo pseudocasuale e dà un risultato che, in relazione al numero casuale estratto può essere corretto o meno. La probabilità di errore si conosce ed è generalmente piccola. Un metodo LasVegas procede come un metodo Montecarlo, ma certifica il risultato. Cioè l'output di un metodo LasVegas è il risultato esatto oppure "fallimento".

Ad esempio, si consideri questo metodo per calcolare il rango di una matrice di interi. Si sceglie un numero primo p a caso e si applica l'eliminazione gaussiana con la strategia del pivot totale (per eliminare eventuali pivot nulli) con

aritmetica modulo p. Se il processo arriva a termine in k passi si otterrà una fattorizzazione del tipo

$$P_1 A P_2 = L A_k \qquad \mod p$$

dove A_k è triangolare superiore e, o $k = n$ con $\det A_n \neq 0$, oppure A_k ha elementi nulli per $i, j \geq k$. Ne segue che, nel primo caso A è non singolare modulo p e quindi è non singolare. Nel secondo caso il rango di A modulo p è $k-1$. Cioè tutte le sottomatrici di A di dimensione maggiore o uguale a k sono singolari modulo p cioè i loro determinanti sono multipli interi di p, mentre c'è una sottomatrice di dimensione $k-1$ che ha determinante diverso da zero e non multiplo di p. Ne segue che il rango di A è almeno $k-1$.

Se guardiamo a questa proprietà in termini probabilistici, potremmo scommettere che scegliendo un primo p a caso la probabilità che il rango di A modulo p sia più piccolo del rango r di A sia molto bassa. Infatti affinché ciò accada occorre che gli $\binom{n}{r}$ determinanti di tutte le sottomatrici $r \times r$ di A siano multipli del numero primo p che abbiamo scelto a caso. Questa sembra una eventualità rara.

Se poi vogliamo essere più tranquilli, possiamo scegliere a caso $m \geq 2$ numeri primi p_i per $i = 1, \ldots, m$ e ripetere il calcolo modulo p_i, $i = 1, \ldots, m$ ottenendo ranghi $r_i = k_i - 1$, $i = 1, \ldots, m$. In questo modo si deduce che il rango di A è almeno $r = \max_i r_i$. La probabilità che il rango sia maggiore di r è che gli $\binom{n}{r}$ determinanti delle sottomatrici $r \times r$ di A siano multipli interi di tutti i p_i, $i = 1, \ldots, m$ che abbiamo scelto a caso. Si presume che tale probabilità sia bassa.

Negli algoritmi Montecarlo e LasVegas la probabilità di fallimento deve essere calcolata a priori per avere chiara l'affidabilità del metodo. Non è ora nostro compito fare questa analisi relativamente all'esempio mostrato.

Un'ultima considerazione sulle strategie di pivot è che il costo della strategia del massimo pivot parziale richiede $n-k+1$ confronti per passo che comportano globalmente una aggiunta di circa $n^2/2$ confronti al costo computazionale. Ciò non altera il costo globale che rimane $\frac{2}{3}n^3$, a meno di termini di ordine inferiore.

Diversa è la situazione della strategia del massimo pivot totale. Infatti, in questo caso, al generico passo k occorre svolgere $(n-k+1)^2$ confronti per individuare l'elemento di massimo modulo in una sottomatrice $(n-k+1) \times (n-k+1)$. Globalmente questo comporta $\frac{1}{3}n^3$ confronti da aggiungere al costo computazionale. Equiparando il costo di un confronto a quello di una operazione aritmetica si arriva ad un costo globale di n^3 operazioni e confronti per l'eliminazione gaussiana con la strategia del massimo pivot totale.

2.3 Calcolo dell'inversa e del determinante

La conoscenza di una fattorizzazione LU o PLU di una matrice A permette di calcolare agevolmente sia il determinante di A che l'inversa se $\det A \neq 0$. Infatti, dalla relazione $PA = LU$ segue che $\det A = \pm \det U$ dove il segno dipende dalla parità o disparità della permutazione associata a P, cioè la parità o disparità del

numero di scambi di righe svolti durante l'esecuzione della strategia di pivot. Inoltre vale $\det U = \prod_{i=1}^{n} u_{i,i}$, per cui il calcolo di $\det A$ costa ancora $\frac{2}{3}n^3$ operazioni.

Per calcolare l'inversa di A, occorre risolvere gli n sistemi lineari $LUx = e^{(k)}$, per $k = 1, \ldots, n$, dove al solito $e^{(k)}$ indica il k-esimo vettore della base canonica. Ciò è ricondotto alla risoluzione dei due sistemi

$$\begin{aligned} Ly &= e^{(k)} \\ Ux &= y \end{aligned}$$

Il primo sistema è di fatto ricondotto a risolvere un sistema triangolare inferiore $(n-k+1) \times (n-k+1)$ e comporta quindi $(n-k+1)^2$ operazioni aritmetiche a meno di termini di ordine inferiore. Il secondo richiede n^2 operazioni. Sommando per $k = 1, \ldots, n$ si arriva ad un costo aggiuntivo di $n^3/3 + n^3$ operazioni aritmetiche che aggiunte al costo della fattorizzazione LU comporta $2n^3$ operazioni aritmetiche.

È naturale chiedersi se il costo di $2n^3$ operazioni è necessario per invertire una matrice $n \times n$ o se è possibile costruire un algoritmo di inversione di costo più basso.

Nel 1969 Volker Strassen ha dimostrato che se esiste un algoritmo per moltiplicare matrici con al più γn^θ operazioni, dove $\gamma > 0$ e $2 < \theta \leq 3$, allora esiste un algoritmo per invertire una matrice con al più $\gamma' n^\theta$ operazioni e viceversa. Inoltre Strassen ha dato un algoritmo per moltiplicare matrici $n \times n$ con al più γn^θ con $\theta = \log_2 7 = 2.8073...$. All'inizio degli anni '80 sono stati individuati altri metodi per moltiplicare matrici con costo asintoticamente più basso. L'algoritmo di Coppersmith e Winograd impiega γn^θ operazioni con $\theta = 2.376...$ La determinazione del valore minimo dell'esponente θ della complessità del prodotto di matrici e dell'inversione è un problema tuttora aperto.

2.4 Implementazione in Octave

Il calcolo della fattorizzazione LU mediante eliminazione gaussiana si realizza facilmente usando il linguaggio di programmazione *Octave*. Il seguente è una prima stesura di una *function* che calcola i fattori L ed U senza strategie di pivot. Poiché non ci sono controlli sulla nullità del pivot, la *function* si arresta se nello svolgimento dei calcoli si incontra un pivot nullo.

```
function [L,U]=flu(A)
% FLU calcola la fattorizzazione LU della matrice A
  n=size(A)(1);
  L=eye(n);
  for k=1:n-1
     for i=k+1:n
        m=A(i,k)/A(k,k);
        L(i,k)=m;
        for j=k+1:n
           A(i,j)=A(i,j)-m*A(k,j);
```

```
        end
      end
    end
    U=triu(A);
```

Una versione più efficiente in ambiente Octave si ottiene riscrivendo in forma vettoriale la parte relativa ai due cicli `for` pilotati dagli indici `i,j`; cioè in modo che le operazioni su vettori e matrici vengono svolte simultaneamente in termini di vettori e non singolarmente sulle componenti. Ad esempio, anziché scrivere

```
for i=k1:k2
   x(i)=y(i)*z(i+1);
end
```

conviene scrivere

```
x(k1:k2)=y(k1:k2).*z(k1+1:k2+1);
```

dove si ricorda che l'operatore .* usato nella precedente istruzione significa il prodotto componente a componente dei due vettori. Cioè il prefisso dato dal punto trasforma un operatore aritmetico tra scalari in un operatore aritmetico tra vettori o matrici dove la sua azione va intesa componente a componente.

Il tempo di esecuzione richiesto da un codice Octave aumenta notevolmente in presenza di cicli `for`. Un programma in cui i cicli `for` che svolgono operazioni sulle singole componenti di vettori o matrici vengono sostituiti da operazioni che coinvolgono in blocco vettori o matrici diventa molto più veloce.

Nel caso dell'algoritmo di fattorizzazione LU si ha il seguente codice più efficiente

```
function [L,U]=vflu(A)
% VFLU calcola la fattorizzazione LU di A in modo vettoriale
  n=size(A)(1);
  L=eye(n);
  for k=1:n-1
      L(k+1:n,k)=A(k+1:n,k)/A(k,k);
      A(k+1:n,k+1:n)=A(k+1:n,k+1:n)-L(k+1:n,k)*A(k,k+1:n);
  end
  U=triu(A);
```

Occorre ricordare che Octave ha la sua implementazione efficiente del calcolo della fattorizzazione LU di una matrice che è realizzata dalla function `lu`.

3 Il metodo di Householder

Nel calcolo della fattorizzazione QR mediante matrici di Householder si genera una successione di matrici A_k tali che $A_{k+1} = M_k A_k$ dove $M_k = I - \beta_k u^{(k)} u^{(k)H}$

è matrice di Householder tale che

$$u_i^{(k)} = \begin{cases} 0 & \text{se } i < k \\ a_{k,k}^{(k)}(1 + \frac{(\sum_{i=k}^n |a_{i,k}^{(k)}|^2)^{1/2}}{|a_{k,k}^{(k)}|}) & \text{se } i = k \\ a_{i,k}^{(k)} & \text{se } i > k \end{cases} \tag{6}$$
$$\beta^{(k)} = 2/\sum_{i=k}^n |u_i^{(k)}|^2$$

e dove A_k ha la forma (2).

Le relazioni che forniscono i valori di $a_{i,j}^{(k+1)}$ sono le seguenti

$$a_{i,j}^{(k+1)} = \begin{cases} a_{i,j}^{(k)} & \text{se } j < k, \text{ se } i \leq k \\ 0 & \text{se } j = k, \text{ e } i > k \\ a_{i,j}^{(k)} - \beta^{(k)} u_i^{(k)} \sum_{r=k}^n \bar{u}_r^{(k)} a_{r,j}^{(k)} & \text{se } i \geq k, j > k \end{cases} \tag{7}$$

Nel caso della risoluzione di un sistema lineare $Ax = b$ l'algoritmo può essere modificato aggiungendo il calcolo di $b^{(k+1)} = M_k b^{(k)}$ mediante

$$b_i^{(k+1)} = b_i^{(k)} - \beta_k u_i^{(k)} \sum_{r=k}^n \bar{u}_r^{(k)} b_r^{(k)}, \quad i = k, \ldots, n. \tag{8}$$

3.1 Costo computazionale

Per quanto riguarda il costo computazionale del k-esimo passo del metodo di Householder si hanno di $2(n-k+1)$ operazioni, a meno di costanti additive, per il calcolo di $u_1^{(k)}$ e per il calcolo di $\beta^{(k)}$. Inoltre l'aggiornamento di $a_{i,j}^{(k+1)}$ richiede $4(n-k+1)^2$ operazioni aritmetiche a meno di termini di ordine inferiore. Sommando su k si arriva a un costo dominato da $\frac{4}{3}n^3$ operazioni. Questo costo non comprende il calcolo degli elementi del fattore $Q = M_1 M_2 \cdots M_{n-1}$, ma prevede di memorizzare la matrice Q in modo implicito attraverso le matrici M_k, $k = 1, \ldots, n-1$.

Confrontando col metodo di eliminazione gaussiana per il calcolo della fattorizzazione LU abbiamo quindi un costo doppio.

3.2 Stabilità numerica

Il maggior costo computazionale viene bilanciato da una migliore stabilità numerica del metodo di Householder che differentemente dall'eliminazione gaussiana non richiede l'applicazione di strategie di massimo pivot. Vale infatti il seguente risultato.

Teorema 2 *Si consideri il sistema $Ax = b$ e sia $\widetilde{R}$ la matrice triangolare superiore ottenuta applicando il metodo di Householder dato dalle formule (6) e (7) per il calcolo della fattorizzazione $A = QR$, in aritmetica floating point con precisione di macchina u. Sia inoltre $\tilde{x}$ la soluzione ottenuta risolvendo in aritmetica floating point il sistema $\widetilde{R}x = \tilde{b}^{(n)}$ dove $\tilde{b}^{(n)}$ è il vettore effettivamente*

calcolato in aritmetica floating point mediante le (8). Allora il vettore $\tilde{x}$ risolve il sistema perturbato $(A + \Delta_A)\widetilde{x} = b + \delta_b$ dove

$$\|\Delta_A\|_F \leq u(\gamma n^2 \|A\|_F + n\|\widetilde{R}\|_F) + O(u^2), \qquad \|\delta_b\|_2 \leq \gamma n^2 u \|b\|_2 + O(u^2),$$

dove γ è una costante positiva.

Dalla limitazione superiore data alla norma di Frobenius di Δ_A risulta che la stabilità del metodo di Householder è legata alla norma di Frobenius di A, che è una costante che non dipende dall'algoritmo, e dalla norma di Frobenius di $\widetilde{R}$. Quest'ultima matrice è la matrice effettivamente calcolata al posto di R in aritmetica floating point e quindi la sua norma differisce da quella di R per un termine proporzionale a u. Inoltre dalla relazione $A = QR$ e dalle proprietà delle norme di Frobenius si deduce che $\|R\|_F = \|A\|_F$ per cui la limitazione data nel teorema si può riscrivere come

$$\|\Delta_A\|_F \leq \gamma u(n^2 + n)\|A\|_F + O(u^2).$$

Cioè l'entità della perturbazione su A dipende in modo quadratico da n. Ciò implica la possibilità di risolvere numericamente sistemi lineari con il metodo di Householder in modo efficace anche per dimensioni grandi di n. Occorre dire che nella pratica del calcolo la crescita dell'errore algoritmico in funzione di n avviene in misura sostanzialmente inferiore a quella data nelle maggiorazioni dei teoremi 1 e 2.

4 Matrici speciali

Una matrice $A = (a_{i,j})$ si dice matrice a banda di ampiezza $2q + 1$ se $a_{i,j} = 0$ per $|i - j| > q$. Ad esempio, una matrice a banda di ampiezza 3, detta *matrice tridiagonale* ha la forma

$$A = \begin{bmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_n \\ & & & c_n & a_n \end{bmatrix}.$$

È facile dimostrare che se esiste la fattorizzazione LU di A allora i due fattori triangolari $L = (\ell_{i,j})$ ed $U = (u_{i,j})$ sono anch'essi a banda, cioè vale $\ell_{i,j} = u_{i,j} = 0$ per $|i - j| > q$. Inoltre, le matrici A_k generate nel processo di eleiminazione gaussiana sono anch'esse matrici a banda di ampiezza $2k + 1$.

In questo caso la formule (1) si semplifica in $m_{i,k} = -a_{i,k}^{(k)}/a_{k,k}^{(k)}$, $i = k + 1, \ldots, \min(k + q, n)$, mentre la (3) si semplifica in

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} + m_{i,k} a_{i,j}^{(k)}, \qquad i, j > k, \quad |i - j| \leq q.$$

In questo modo il costo computazionale del calcolo della fattorizzazione LU è minore o uguale a $(2q^2 + q)n$.

Una analoga proprietà vale per matrici per cui $a_{i,j} = 0$ se $j - i > q_1$ o se $i - j > q_2$ per $1 \leq q_1, q_2 < n$. Infatti anche questa struttura a banda si conserva nelle matrici A_k e nei fattori L ed U.

Nel caso del metodo di Householder la struttura si conserva nel fattore U dove però l'ampiezza di banda diventa $q_1 + q_2$. Anche per il metodo di Householder applicato a una matrice a banda il costo computazionale si riduce.

Riferimenti bibliografici

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

Metodi iterativi per sistemi lineari

Dario A. Bini, Università di Pisa

30 agosto 2020

Sommario

Questo modulo didattico contiene risultati relativi ai metodi iterativi per risolvere sistemi di equazioni lineari.

I metodi basati sulle fattorizzazioni LU e QR per risolvere un sistema di n equazioni lineari in n incognite $Ax = b$ forniscono la soluzione in un numero di operazioni aritmetiche dell'ordine di n^3 . In certi problemi provenienti dalle applicazioni il valore di n è molto elevato. Ad esempio, nel problema del calcolo del vettore di PageRank nei motori di ricerca sul Web, il valore di n è circa 40 miliardi. In questo caso il costo computazionale di n^3 operazioni diventa proibitivo. Anche usando il calcolatore più veloce esistente attualmente dovrebbero passare molti millenni prima che l'eliminazione gaussiana o il metodo di Householder fornisca la soluzione del sistema. Quindi per trattare problemi di dimensioni così elevate occorre escogitare qualcosa di diverso.

Una caratteristica particolare delle matrici di grosse dimensioni che si incontrano nelle applicazioni del web è che la maggior parte dei loro elementi sono nulli e il numero di elementi diversi da zero è dell'ordine di grandezza di n. Questa proprietà è nota come *sparsità*. Per queste matrici "sparse" è relativamente poco costoso calcolare il prodotto matrice - vettore, visto che il numero di operazioni aritmetiche da eseguire è circa il doppio del numero di elementi diversi da zero.

La figura 1 riporta una tipica matrice sparsa che descrive l'insieme dei collegamenti di $n = 1000$ pagine sul Web. Gli elementi diversi da 0, riportati graficamente con un asterisco, sono poche migliaia. Un elemento non nullo in posizione (i, j) denota un link dalla pagina i alla pagina j per $i, j = 1, \ldots, n$.

In questo articolo descriviamo dei metodi che generano successioni di vettori $x^{(k)}$ che convergono alla soluzione cercata. Per generare il generico vettore $x^{(k)}$ questi metodi richiedono il calcolo del prodotto di una matrice per un vettore e questa è l'operazione computazionale più costosa del metodo. Se la successione converge abbastanza velocemente alla soluzione allora è sufficiente calcolare pochi elementi della successione per raggiungere una buona approssimazione. Inoltre in molti casi, come ad esempio nelle applicazioni grafiche, non è necessario conoscere molte cifre della soluzione per cui il numero di elementi della successione da calcolare diventa trascurabile rispetto alla dimensione.

Questa classe di metodi che approssimano la soluzione generando successioni di vettori è nota come classe dei *metodi iterativi*.

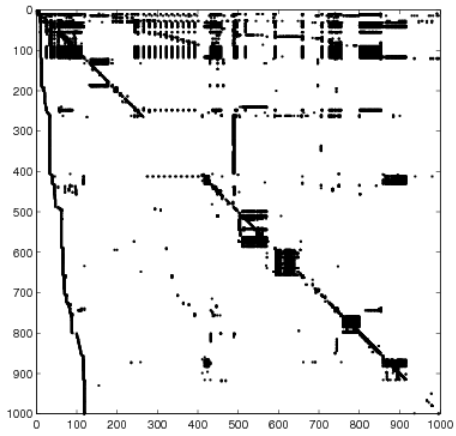


Figura 1: Matrice sparsa 1000x1000 che descrive le connessioni di un insieme di 1000 pagine del Web. L'elemento di posto (i, j) è uguale a 1 se esiste un link dalla pagina i alla pagina j. Gli asterischi denotano elementi non nulli

1 Metodi stazionari

Dato il sistema lineare $Ax = b$ dove A è una matrice $n \times n$ e b vettore di n componenti, si consideri un generico partizionamento additivo di A:

$$A = M - N, \qquad \det M \neq 0\,.$$

Possiamo allora riscrivere equivalentemente il sistema come $Mx = Nx + b$ che conduce alla seguente scrittura equivalente del sistema originale

$$x = M^{-1}Nx + M^{-1}b =: Px + q\,. \tag{1}$$

La formulazione del problema in (1) è data nella forma di *problema di punto fisso* e produce in modo naturale il seguente metodo per generare una successione di vettori una volta scelto $x^{(0)} \in \mathbb{C}^n$:

$$x^{(k+1)} = Px^{(k)} + q\,. \tag{2}$$

Chiamiamo *metodo iterativo stazionario* l'espressione (2) o, più formalmente, l'insieme delle successioni generate da (2) al variare di $x^{(0)} \in \mathbb{C}^n$. Il metodo iterativo (2) è detto *stazionario* poichè nell'espressione (2) la matrice P e il vettore q sono indipendenti da k. La matrice P viene detta *matrice di iterazione* del metodo.

Osserviamo che se la successione generata dalla (2) fissato $x^{(0)}$ ha un limite x^* allora x^* è soluzione del sistema lineare $Ax = b$. Infatti il limite del primo

membro della (2) è x^*, e per la continuità delle applicazioni lineari, il limite del secondo membro è $Px^* + q$. Per cui $x^* = Px^* + q$, cioè $Ax^* = b$.

Questo fatto implica che ci basta dimostrare la convergenza della successione $x^{(k)}$ per concludere sull'efficacia del metodo iterativo.

1.1 Convergenza

Il seguente risultato fornisce condizioni sufficienti di convergenza facilmente verificabili.

Teorema 1 *Se esiste una norma di matrice indotta $\|\cdot\|$ tale che $\|P\| < 1$ allora la matrice A è invertibile per cui esiste unica la soluzione x del sistema $Ax = b$. Inoltre per ogni vettore iniziale $x^{(0)} \in \mathbb{C}^n$ la successione generata da (2) converge alla soluzione x.*

Dim. Se fosse $\det A = 0$ esisterebbe v vettore non nullo tale che $Av = 0$ cioè $Mv = Nv$ quindi $v = M^{-1}Nv$. Cioè la matrice $P = M^{-1}N$ avrebbe raggio spettrale ρ maggiore o uguale a 1 per cui per ogni norma indotta si avrebbe $\|P\| \geq \rho(P) \geq 1$ che contraddice le ipotesi. Sia allora x la soluzione del sistema $Ax = b$, definiamo il vettore $e^{(k)} = x^{(k)} - x$ che ci rappresenta l'errore di approssimazione al passo k e osserviamo che sottraendo da entrambi i membri della (2) i membri corrispondenti della relazione $x = Px + q$, si ottiene

$$e^{(k+1)} = Pe^{(k)}.$$

Applicando la relazione precedente in modo induttivo si ottiene

$$e^{(k)} = Pe^{(k-1)} = P^2e^{(k-2)} = \cdots = P^ke^{(0)}. \tag{3}$$

Scegliamo una norma vettoriale arbitraria $\|\cdot\|$ e consideriamo la norma di matrice indotta. Allora dalle proprietà delle norme si ha che per ogni $x^{(0)}$, e quindi per ogni $e^{(0)}$ vale

$$\|e^{(k)}\| \leq \|P^k\|\,\|e^{(0)}\| \leq \|P\|^k\|e^{(0)}\|$$

da cui $\lim_k \|e^{(k)}\| = 0$ per ogni $x^{(0)}$. □

Definiamo il metodo iterativo (2) *convergente* se per ogni scelta del vettore $x^{(0)}$ la successione $x^{(k)}$ converge ad una soluzione del sistema. Il teorema 1 dà quindi una condizione sufficiente di convergenza per il metodo iterativo. Inoltre la norma di P ci fornisce una maggiorazione della riduzione dell'errore in ciascun passo del metodo iterativo. Viene quindi naturale confrontare la velocità di convergenza di due metodi iterativi in base alla norma delle rispettive matrici di iterazione. Questo confronto non è rigoroso poiché dipende dalla norma scelta e inoltre si basa su maggiorazioni dell'errore che non sappiamo quanto siano accurate.

Il seguente risultato dà una condizione *necessaria e sufficiente* di convergenza in termini del raggio spettrale $\rho(P)$ della matrice di iterazione P.

Teorema 2 *Il metodo iterativo è convergente e* $\det A \neq 0$ *se e solo se* $\rho(P) < 1$.

Dim. Se $\rho(P) < 1$ allora esiste un $\epsilon > 0$ tale che $\rho(P) + \epsilon < 1$. Sappiamo dalle proprietà delle norme che esiste una norma indotta $\|\cdot\|$ tale che $\|P\| \leq \rho(P) + \epsilon < 1$. Allora per il teorema 1 il metodo iterativo è convergente e $\det A \neq 0$. Viceversa, supponiamo che $\det A \neq 0$ e che il metodo sia convergente. In questo caso la soluzione x del sistema esiste ed è unica. L'arbitrarietà della scelta di $x^{(0)}$ implica l'arbitrarietà del vettore $e^{(0)} = x^{(0)} - x$. Per cui, per le ipotesi, la successione degli errori $e^{(k)}$ converge a zero qualunque sia il vettore $e^{(0)}$. Scegliendo allora $e^{(0)}$ uguale ad un autovettore di P corrispondente all'autovalore λ, cioè tale che $Pe^{(0)} = \lambda e^{(0)}$, dalla (3) si ha

$$e^{(k)} = P^k e^{(0)} = \lambda^k e^{(0)}.$$

Poichè $\lim_k e^{(k)} = 0$, ne segue che $\lim_k \lambda^k e^{(0)} = 0$ da cui $\lim_k \lambda^k = 0$. Ciò implica che $|\lambda| < 1$. □

Si osservi che la condizione $\det A \neq 0$ non può essere rimossa dalle ipotesi del teorema precedente. Si consideri infatti il caso di un sistema singolare omogeneo $Ax = 0$ in cui $A = M - N$ dove $M = I$ e N ha autovalori λ_i e autovettori linearmente indipendenti $v^{(i)}$, $i = 1, \ldots, n$ cioè $Nv^{(i)} = \lambda_i v^{(i)}$, dove $\lambda_1 = 1$ e $|\lambda_i| < 1$ per $i = 2, \ldots, n$. Posto $x^{(0)} = \sum_{i=1}^n \xi_i v^{(i)}$ vale $x^{(k)} = \sum_{i=1}^n \xi_i \lambda_i^k v^{(i)}$. Per cui si ha $\lim_k x^{(k)} = \xi_1 v^{(1)}$, cioè si ha convergenza della successione qualunque sia $x^{(0)}$ anche se il limite dipende dalla scelta di $x^{(0)}$. Cioè il metodo iterativo è convergente ma il raggio spettrale di $P = M^{-1}N = N$ è 1.

È interessante osservare che dal teorema precedente segue che se un metodo iterativo è convergente allora esiste una norma indotta per cui $\|P\| < 1$. Cioè vale anche il "solo se" nel teorema 1. Per dimostrare questo basta osservare che essendo $\rho(P) < 1$, esiste un $\epsilon > 0$ per cui $\rho(P) + \epsilon < 1$ e, per le proprietà delle norme indotte, esiste una norma indotta per cui $\|P\| \leq \rho(P) + \epsilon < 1$.

Il raggio spettrale di P, oltre a darci una condizione necessaria e sufficiente di convergenza, esprime in una forma che preciseremo meglio tra poco, la riduzione media asintotica per passo dell'errore. Per cui è corretto confrontare la velocità di convergenza di due metodi iterativi in base al raggio spettrale delle rispettive matrici di iterazione.

Osserviamo che data una norma $\|\cdot\|$, il quoziente $\|e^{(k)}\|/\|e^{(k-1)}\|$ esprime la riduzione dell'errore al k-esimo passo del metodo iterativo misurato nella norma vettoriale scelta. Consideriamo allora la media geometrica di queste riduzioni dell'errore fatta sui primi k passi.

$$\theta_k(e^{(0)}) = \left(\frac{\|e^{(1)}\|}{\|e^{(0)}\|} \cdot \frac{\|e^{(2)}\|}{\|e^{(1)}\|} \cdots \frac{\|e^{(k)}\|}{\|e^{(k-1)}\|} \right)^{\frac{1}{k}} . \tag{4}$$

Semplificando numeratori e denominatori in (4) e tenendo presente che $e^{(k)} = P^k e^{(0)}$ si ottiene

$$\theta_k(e^{(0)}) = \left(\frac{\|P^k e^{(0)}\|}{\|e^{(0)}\|} \right)^{\frac{1}{k}} \leq \|P^k\|^{\frac{1}{k}}$$

dove la diseguaglianza non è lasca poiché l'uguaglianza si ottiene per un particolare vettore $e^{(0)}$, quello che realizza il $\max \|P^k e^{(0)}\|/\|e^{(0)}\| = \|P^k\|$.

Definiamo la *riduzione asintotica media per passo* di un metodo iterativo con errore iniziale $e^{(0)}$ il valore $\theta(e^{(0)}) = \lim_k \theta_k(e^{(0)})$. Prendendo il limite su k si ottiene

$$\lim_k \theta_k(e^{(0)}) \leq \lim_k \|P^k\|^{\frac{1}{k}} = \rho(P)\,.$$

L'uguaglianza è raggiunta per $e^{(0)}$ uguale ad un autovettore di P corrispondente ad un autovalore λ di P tale che $|\lambda| = \rho(P)$.

Si può concludere col seguente risultato

Teorema 3 *La riduzione asintotica media per passo $\theta(e^{(0)})$ dell'errore di un metodo iterativo applicato con errore iniziale $e^{(0)}$ è minore o uguale al raggio spettrale della matrice di iterazione P. Inoltre, se $e^{(0)}$ è proporzionale a un autovettore corrispondente ad un autovalore di modulo massimo, allora $\theta(e^{(0)})$ coincide con il raggio spettrale di P.*

Un esempio significativo di metodo iterativo è dato dal *metodo di Richardson* definito dalla relazione

$$x^{(k+1)} = x^{(k)} - \alpha(Ax^{(k)} - b)$$

dove α è un opportuno parametro. Se la matrice è definita positiva, la scelta $\alpha = 1/\|A\|$, con $\|\cdot\|$ norma indotta, garantisce la convergenza del metodo.

2 I metodi di Jacobi e di Gauss-Seidel

Si decomponga la matrice A nel modo seguente

$$A = D - B - C$$

dove

- D è la matrice diagonale con elementi diagonali $d_{i,i} = a_{i,i}$;
- B è la matrice strettamente triangolare inferiore con elementi $b_{i,j} = -a_{i,j}$ per $i > j$;
- C è la matrice strettamente triangolare superiore con elementi $c_{i,j} = -a_{i,j}$ per $i < j$.

Il metodo iterativo ottenuto col partizionamento additivo $A = M - N$ con $M = D$ e $N = B + C$ è detto *metodo di Jacobi*. Il metodo che si ottiene ponendo $M = D - B$ e $N = C$ viene detto *metodo di Gauss-Seidel*.

Chiaramente, per la loro applicabilità questi metodi richiedono che $\det M \neq 0$ e quindi $a_{i,i} \neq 0$ per $i = 1, \ldots, n$.

Le matrici di iterazione $P = M^{-1}N$ del metodo di Jacobi e del metodo di Gauss-Seidel, denotate rispettivamente con J e G, sono date da

$$J = D^{-1}(B + C), \qquad G = (D - B)^{-1}C\,.$$

Vale il seguente teorema che dà condizioni sufficienti di convergenza facilmente verificabili.

Teorema 4 *Se vale una delle seguenti condizioni allora* $\rho(J) < 1$ *e* $\rho(G) < 1$*:*

1. *A è fortemente dominante diagonale;*
2. *A^T è fortemente dominante diagonale;*
3. *A è irriducibilmente dominante diagonale;*
4. *A^T è irriducibilmente dominante diagonale.*

Dim. Si osserva innanzitutto che sotto le condizioni del teorema risulta $a_{i,i} \neq 0$ per $i = 1, \ldots, n$. Si supponga per assurdo che la matrice J abbia un autovalore λ di modulo maggiore o uguale a 1. Allora dalla condizione $\det(\lambda I - J) = 0$ si deduce che $\det(\lambda I - D^{-1}(B + C)) = 0$, cioè

$$\det(\lambda D - B - C) = 0\,.$$

Ma la matrice $H = \lambda D - B - C$ si ottiene dalla matrice A moltiplicando i suoi elementi diagonali per λ. Quindi se A è fortemente dominante diagonale, a maggior ragione H è fortemente dominante diagonale essendo $|\lambda| \geq 1$. Analogamente se A è irriducibilmente dominante diagonale così è H. Quindi, per il primo o per il terzo teorema di Gerschgorin (a seconda delle ipotesi su A) risulta $\det H \neq 0$ che contraddice l'ipotesi fatta. Discorso analogo vale se A^T è fortemente o irriducibilmente dominante diagonale. Per quanto riguarda il metodo di Gauss-Seidel, l'esistenza di un autovalore λ di G di modulo maggiore o uguale a 1 implica che $\det(\lambda I - G) = 0$ cioè $\det(\lambda I - (D - B)^{-1}C) = 0$. Si deduce quindi che

$$\det(D - B - \lambda^{-1}C) = 0\,.$$

Ma la matrice $H = D - B - \lambda^{-1}C$ differisce dalla matrice A per il fatto che gli elementi della parte triangolare superiore sono moltiplicati per λ^{-1} cioè per una quantità di modulo minore o uguale a 1. Per cui se A è fortemente o irriducibilmente dominante diagonale lo è a maggior ragione la matrice H. Quindi, anche in questo caso, per il primo o per il terzo teorema di Gerschgorin (a seconda dell'ipotesi su A) risulta $\det H \neq 0$ che contraddice l'ipotesi fatta. □

3 Aspetti computazionali

L'iterazione del metodo di Jacobi si lascia scrivere come

$$x^{(k+1)} = D^{-1}((B+C)x^{(k)} + b)$$

che in componenti diventa

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}(b_i - \sum_{j=1,\ j\neq i}^{n} a_{i,j}x_j^{(k)})\,. \tag{5}$$

L'iterazione del metodo di Gauss-Seidel si lascia scrivere come

$$x^{(k+1)} = (D-B)^{-1}(Cx^{(k)} + b)\,.$$

Se guardiamo a questa relazione come a un sistema lineare con matrice triangolare inferiore, cioè $(D-B)x^{(k+1)} = Cx^{(k)} + b$, allora la sua risoluzione mediante il metodo di sostituzione in avanti fornisce la formula che ci dà il metodo di Gauss-Seidel in componenti:

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)})\,. \tag{6}$$

La stessa relazione si ottiene in modo equivalente scrivendo l'espressione $(D-B)x^{(k+1)} = Cx^{(k)} + b$ nella forma $Dx^{(k+1)} = Cx^{(k)} + Bx^{(k+1)} + b$ e quindi

$$x^{(k+1)} = D^{-1}(Cx^{(k)} + Bx^{(k+1)} + b).$$

Un confronto tra la (5) e la (6) mostra che i due metodi impiegano lo stesso numero di operazioni aritmetiche che è di circa n^2 moltiplicazioni e n^2 addizioni. Il metodo di Jacobi esegue queste operazioni intervenendo sulle componenti di $x^{(k)}$ mentre il metodo di Gauss-Seidel usa anche le componenti già aggiornate di $x^{(k+1)}$. Questo fatto porta a pensare che il metodo di Gauss-Seidel, utilizzando informazione più aggiornata rispetto al metodo di Jacobi, debba avere migliori proprietà di convergenza. Generalmente è così anche se si possono costruire degli esempi in cui il metodo di Jacobi converge più velocemente del metodo di Gauss-Seidel. Esistono particolari classi di matrici per le quali si può mettere in relazione esplicita $\rho(J)$ con $\rho(G)$. Questo lo vedremo tra poco.

Si osserva ancora che dalla (6) segue che l'aggiornamento della i-esima componente di $x^{(k)}$ può essere effettuato solo dopo aver aggiornato i valori delle componenti di indice minore di i. Per questo motivo il metodo di Gauss-Seidel è chiamato anche metodo degli *spostamenti successivi* mentre il metodo di Jacobi viene detto metodo degli *spostamenti simultanei*. Questa differente caratteristica è cruciale dal punto di vista computazionale quando si dispone di un ambiente di calcolo parallelo in cui diversi processori matematici possono svolgere simultaneamente operazioni aritmetiche sui dati. Per il metodo di Jacobi la disponibilità di n processori permette di aggiornare simultaneamente tutte le

componenti di $x^{(k)}$ nel tempo richiesto dall'aggiornamento di una singola componente. Per il metodo di Gauss-Seidel questo non è possibile e la disponibilità di più processori aritmetici non può essere pienamente sfruttata.

Un'altra osservazione utile riguarda il fatto che, se la matrice A è sparsa, cioè il numero dei suoi elementi non nulli è dell'ordine di n, allora anche B e C sono sparse, per cui il calcolo di un passo di questi due metodi costa un numero di operazioni proporzionale a n e non a n^2 come sarebbe per una matrice densa. Ciò costituisce un grosso vantaggio per tutti quei problemi in cui la proprietà di sparsità non si accompagna ad esempio ad una struttura a banda, per cui applicando l'eliminazione gaussiana o il metodo di Householder si ottengono matrici A_k (complementi di Schur) che perdono rapidamente la proprietà di sparsità. Questo fenomeno è chiamato *fill-in*. A causa del fill-in il costo computazionale dei metodi basati sulla fattorizzazione quali l'eliminazione gaussiana o il metodo di Householder diventa dell'ordine di n^3 per cui non si riesce a trarre vantaggio dal fatto che la matrice è sparsa. Occorre sottolineare che in letteratura sono state sviluppate tecniche di riordinamento di righe e colonne di una matrice sparsa che in alcuni casi permettono di contenere il fenomeno del fill-in e quindi raggiungere un costo computazionale inferiore a quello che si avrebbe applicando un metodo di risoluzione basato su fattorizzazioni al caso di una matrice densa.

4 Confronto tra i metodi di Jacobi e Gauss-Seidel

Come già osservato, ci sono delle classi di matrici per cui si può dare una relazione esplicita tra i raggi spettrali di J e di G.

Teorema 5 *(di Stein-Rosenberg). Se la matrice A ha elementi diagonali non nulli e J ha elementi non negativi allora vale una sola delle seguenti proprietà*

- $\rho(J) = \rho(G) = 0$,
- $0 < \rho(G) < \rho(J) < 1$,
- $\rho(J) = \rho(G) = 1$,
- $1 < \rho(J) < \rho(G)$.

Quindi alla luce del teorema 3, nelle ipotesi del teorema 5 la convergenza del metodo di Gauss-Seidel è più veloce di quella del metodo di Jacobi.

Per matrici tridiagonali si riesce a quantificare la maggior velocità di convergenza del metodo di Gauss-Seidel. Vale infatti il seguente

Teorema 6 *Se A è una matrice tridiagonale con elementi diagonali non nulli, allora per ogni autovalore λ di J esiste un autovalore μ di G tale che $\mu = \lambda^2$. Per ogni autovalore non nullo μ di G esiste un autovalore λ di J tale che $\mu = \lambda^2$. In particolare vale*

$$\rho(G) = \rho(J)^2 .$$

Dim.

Sia λ autovalore di J e μ autovalore di G allora vale $\det(\lambda I - J) = 0$ e $\det(\mu I - G) = 0$. Queste due condizioni possono essere riscritte rispettivamente come

$$\begin{aligned} &\det(\lambda D - B - C) = 0\,, \\ &\det(\mu D - \mu B - C) = 0\,. \end{aligned} \tag{7}$$

Ora si consideri un parametro $\alpha \neq 0$ e si definisca la matrice diagonale $D_\alpha = \mathrm{diag}(1, \alpha, \alpha^2, \ldots, \alpha^{n-1})$ e si osservi che $D_\alpha D D_\alpha^{-1} = D$, poiché D è diagonale, mentre $D_\alpha B D_\alpha^{-1} = \alpha B$ e $D_\alpha C D_\alpha^{-1} = \alpha^{-1} C$, essendo B bidiagonale inferiore e C bidiagonale superiore. Quindi si ottiene

$$D_\alpha(\lambda D - B - C)D_\alpha^{-1} = \lambda D - \alpha B - \alpha^{-1} C = \alpha^{-1}(\lambda\alpha D - \alpha^2 B - C)\,.$$

Per cui la prima delle due condizioni in (7) si può riscrivere come

$$\det(\lambda\alpha D - \alpha^2 B - C) = 0\,. \tag{8}$$

Confrontando la (8) con la seconda delle (7) si vede allora che scegliendo $\alpha = \lambda$ e $\mu = \lambda^2$ ne segue che se λ è autovalore di J non nullo allora $\mu = \lambda^2$ è autovalore di G. Se μ è autovalore non nullo di G allora i λ tali che $\lambda^2 = \mu$ sono autovalori di J. D'altro canto, se λ è autovalore nullo di J allora $\mu = 0$ è comunque autovalore di G essendo G singolare. Quindi la restrizione $\lambda \neq 0$ può essere tolta dall'enunciato. □

Dal teorema precedente, alla luce del teorema 3, segue che mediamente il metodo di Gauss-Seidel richiede la metà dei passi richiesti dal metodo di Jacobi per ridurre l'errore di una quantità prefissata.

5 Metodi a blocchi

Se la matrice A di dimensione $mn \times mn$ è partizionata in n^2 blocchi $m \times m$

$$A = (A_{i,j}) = \begin{bmatrix} A_{1,1} & A_{1,2} & \ldots & A_{1,n} \\ A_{2,1} & A_{2,2} & \ldots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \ldots & A_{n,n} \end{bmatrix}$$

possiamo considerare una decomposizione additiva $A = D - B - C$ dove D è la matrice diagonale a blocchi con blocchi diagonali uguali a $A_{i,i}$, $i = 1, \ldots, n$, $B = (B_{i,j})$ la matrice triangolare inferiore a blocchi tale che $B_{i,j} = -A_{i,j}$ se $i > j$ mentre $B_{i,j} = 0$ altrimenti, e $C = (C_{i,j})$ la matrice triangolare superiore a blocchi tale che $C_{i,j} = -A_{i,j}$ per $i < j$ e $C_{i,j} = 0$ altrimenti. In questo modo, se i blocchi diagonali $A_{i,i}$ di A sono non singolari, possiamo considerare i metodi iterativi definiti da $M = D$, $N = B + C$, e da $M = D - B$, $N = C$. Il primo metodo è detto metodo di *Jacobi a blocchi* mentre il secondo è detto metodo di *Gauss-Seidel a blocchi*.

Per i metodi di Jacobi a blocchi e di Gauss-Seidel a blocchi vale un analogo del teorema 6.

Teorema 7 *Sia $A = (A_{i,j})$ una matrice tridiagonale a blocchi, cioè tale che $A_{i,j} = 0$ se $|i - j| \geq 2$, con blocchi diagonali $A_{i,i}$ non singolari. Siano J_B e G_B le matrici di iterazione dei metodi di Jacobi a blocchi e di Gauss-Seidel a blocchi. Allora per ogni autovalore di λ di J_B esiste un autovalore μ di G_B tale che $\mu = \lambda^2$. Per ogni autovalore non nullo μ di G_B esiste un autovalore λ di J_B tale che $\mu = \lambda^2$. In particolare vale*

$$\rho(G) = \rho(J)^2 .$$

La dimostrazione del teorema 7 si svolge esattamente nello stesso modo della dimostrazione del teorema 6, per cui non viene riportata.

6 Metodi iterativi non stazionari (cenno)

Esistono metodi iterativi non stazionari in cui la successione $x^{(k)}$ non si può scrivere nella forma (2). Una classe molto studiata di questi metodi è quella basata sulle successioni dei *sottospazi di Krylov*. Dato un vettore iniziale $x^{(0)}$ si costruisce lo spazio di Krylov $\mathcal{S}_k$ di dimensione al più k che è generato dai vettori $x^{(0)}, Ax^{(0)}, \ldots, A^{k-1}x^{(0)}$. In questa classe di metodi iterativi il vettore $x^{(k)}$ viene scelto nello spazio $\mathcal{S}_k$ con opportuni criteri che determinano il tipo di metodo e le proprietà computazionali e di convergenza.

Fanno parte di questa classe i metodi del gradiente quali il metodo della *discesa più ripida* e il metodo del *gradiente coniugato*.

7 Esercizi

Esercizio 1 Sia $n > 2$ un intero e si denoti con $e = (1, \ldots, 1)^T \in \mathbb{R}^{n-1}$, $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{n-1}$.

a) Costruire la matrice di Householder P tale che $Pe = \theta e_1$, dove $\theta \in \mathbb{R}$

b) Sia $A = (a_{i,j})$ matrice reale $n \times n$ tale che $a_{1,i} = a_{i,1} = 1$ per $i = 2, \ldots, n$ e $a_{i,j} = 0$ altrove. Si costruisca una matrice ortogonale Q tale che $B = QAQ^T$ abbia tutti elementi nulli tranne $b_{2,1} = b_{1,2}$. Si determini il valore di $b_{2,1}$.

c) Dare condizioni su α affinché il metodo di Jacobi applicato al sistema $(I - \alpha A)x = b$ sia convergente.

d) Dare condizioni su α affinché il metodo di Gauss-Seidel applicato al sistema $(I - \alpha A)x = b$ sia convergente. Si confrontino le velocità di convergenza dei due metodi.

Soluzione

Una matrice reale di Householder P è tale che $P = I - \beta uu^T$ dove $u \in \mathbb{R}^n$, $u \neq 0$ e $\beta = 2/\|u\|_2^2$.

a) Dalla relazione $Pe = \theta e_1$ e dalla ortogonalità di P si deduce che $\|e\|_2 = |\theta|$, cioè $\theta = \pm\sqrt{n-1}$. Dalla espressione $P = I - \beta uu^T$ si deduce che $\beta uu^T e = e - \theta e_1$. Per evitare cancellazione numerica si sceglie allora $\theta = -\sqrt{n-1}$, per cui $u = e + \sqrt{n-1}e_1$, $\beta = 2/\|u\|_2^2 = 1/(\sqrt{n-1} + n - 1)$.

b) Se P è la matrice di Householder tale che $Pe = \theta e_1$, la matrice $Q = \begin{bmatrix} 1 & 0 \\ 0 & P \end{bmatrix}$ risulta ortogonale e inoltre

$$QAQ^T = \begin{bmatrix} 0 & e^T P \\ Pe & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\sqrt{n-1} & 0 & \dots & 0 \\ -\sqrt{n-1} & & & & \\ \vdots & & & 0 & \\ 0 & & & & \end{bmatrix}$$

c) La matrice di iterazione del metodo di Jacobi è αA. Il suo raggio spettrale è α per il raggio spettrale di A. Quest'ultimo coincide col raggio spettrale della matrice $\begin{bmatrix} 0 & -\sqrt{n-1} \\ -\sqrt{n-1} & 0 \end{bmatrix}$ che è $\sqrt{n-1}$. Quindi il metodo di Jacobi è convergente se $\alpha < 1/\sqrt{n-1}$.

d) La matrice di iterazione del metodo di Gauss-Seidel è

$$\begin{bmatrix} 1 & 0 \\ \alpha e & I \end{bmatrix}^{-1} \begin{bmatrix} 0 & \alpha e^T \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \alpha e^T \\ 0 & -\alpha^2 ee^T \end{bmatrix}$$

Il suo raggio spettrale coincide con il raggio spettrale di $\alpha^2 ee^T$. Questa matrice ha $n-1$ autovalori nulli corrispondenti agli autovettori nello spazio ortogonale a e e un autovalore pari a $\alpha^2(n-1)$ corrispondente all'autovettore e. Il suo raggio spettrale è quindi $\alpha^2(n-1)$ e la condizione di convergenza del metodo di Gauss-Seidel è quindi $|\alpha| < 1/\sqrt{n-1}$. Cioè la stessa del metodo di Jacobi. Però il raggio spettrale della matrice di iterazione del metodo di Gauss-Seidel è il quadrato di quello della matrice del metodo di Jacobi. Per cui la velocità di convergenza del metodo di Gauss-Seidel è doppia di quella del metodo di Jacobi. □

Esercizio 2 È dato il sistema lineare $Ax = b$ dove A è una matrice $n \times n$ reale simmetrica definita positiva di autovalori $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Si consideri il metodo iterativo definito da

$$x^{(i+1)} = x^{(i)} + \alpha(b - Ax^{(i)}),$$

con α parametro reale.

a) Dare condizioni su α in termini degli autovalori di A necessarie e sufficienti per la convergenza del metodo.

b) Determinare in funzione degli autovalori di A il valore di α che massimizza la velocità di convergenza.

c) Determinare condizioni su α in funzione degli elementi di A sufficienti per la convergenza del metodo.

d) Se A è una matrice $n \times n$ reale arbitraria determinare condizioni sugli autovalori di A affinché esista un α che rende il metodo iterativo convergente.

Soluzione

Il metodo iterativo si può scrivere come $X^{(i+1)} = Px^{(i)} + q$ con $P = I - \alpha A$ e $q = \alpha b$. Inoltre x è soluzione del sistema $Ax = b$ se e solo se $x = Px + q$. Condizione necessaria e sufficiente per la convergenza di un metodo iterativo è che il raggio spettrale della matrice di iterazione sia minore di 1.

a) La matrice di iterazione del metodo è $P = I - \alpha A$, i suoi autovalori sono quindi $1 - \alpha\lambda_i$, $i = 1, \ldots, n$. Il raggio spettrale di P è il massimo dei $|1 - \alpha\lambda_i|$, $i = 1, \ldots, n$. Quindi la condizione cercata su α è $-1 < 1 - \lambda_i\alpha < 1$ cioè $\alpha < 2/\lambda_i$. Dato l'ordinamento degli autovalori si ottiene $\alpha < 2/\lambda_n$.

b) Poiché il raggio spettrale è uguale alla riduzione asintotica media per passo dell'errore, per massimizzare la velocità di convergenza basta minimizzare il raggio spettrale. Basta cioè trovare il

$$\min_{\alpha} \max_{i} |1 - \alpha\lambda_i|$$

cioè il minimo dell'inviluppo superiore delle funzioni $f_i(\alpha) = |1 - \alpha\lambda_i|$. Come si vede dalla figura tale inviluppo è dato da

$$f(x) = \begin{cases} 1 - \alpha_1 x & \text{per } x < \alpha_0 \\ 1 - \alpha_n & \text{per } x \geq \alpha_0 \end{cases}$$

con $\alpha_0 = 2/(\alpha_1 + \alpha_n)$ e il suo minimo, preso in α_0 vale $(\lambda_n - \lambda_1)/(\lambda_1 + \lambda_n)$

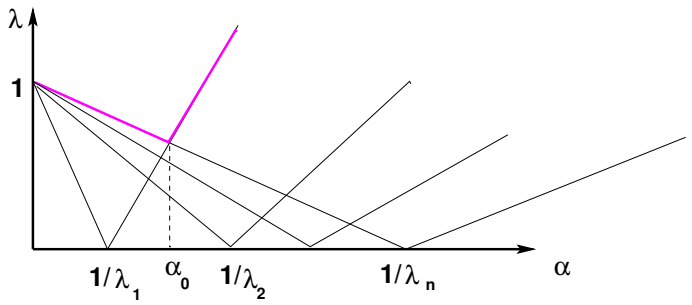


Figura 2: Inviluppo superiore delle funzioni $|1 - \alpha\lambda_i|$

c) Dal punto a) si ha che per la convergenza occorre e basta che $\alpha < 2/\lambda_n$. Poiché per ogni norma matriciale indotta $\|\cdot\|$ vale $\lambda_n = |\lambda_n| \leq \|A\|$, ne segue che la condizione $\alpha < 2/\|A\|$ è sufficiente per la convergenza. Ad esempio, con la norma infinito basta che $\alpha < 2/\max_i \sum_{j=1}^{n} |a_{i,j}|$.

d) Gli autovalori di A devono essere tali che esista un α per cui $|1 - \lambda_i\alpha| < 1$, cioè $-1 < 1 - \lambda_i\alpha < 1$. Le due diseguaglianze valgono se e solo se $0 < \lambda_i\alpha < 2$. Per cui gli autovalori devono essere tutti dello stesso segno e non nulli, inoltre $0 < \alpha < 2/\lambda_i$ oppure $2/\lambda_i < \alpha < 0$. □

Esercizio 3 Siano rispettivamente J e G le matrici di iterazione dei metodi di Jacobi e di Gauss-Seidel applicati al sistema $Hx = b$ dove $H = (h_{i,j})$ è la matrice tridiagonale $n \times n$ tale che $h_{i,i} = 3$ per $i = 1, \ldots, n$ e $h_{i+1,i} = h_{i,i+1} = 1$ per $i = 1, \ldots, n-1$.

a) Dimostrare che per $n \geq 2$ vale $\rho(J) < 2/3$, $\rho(G) < 4/9$, dove $\rho(\cdot)$ indica il raggio spettrale.

b) Sia $n = 2m$ pari, e si partizioni H in blocchi 2×2 in modo che H possa essere vista come matrice $m \times m$ i cui elementi sono blocchi 2×2. In questo modo H risulta tridiagonale a blocchi con blocchi diagonali $A = \left[\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right]$, blocchi sopradiagonali $B = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, blocchi sottodiagonali B^T. Si consideri il metodo iterativo di Jacobi a blocchi dato dal partizionamento $H = M - N$ con M matrice diagonale a blocchi 2×2 con blocchi diagonali uguali ad A. Si dimostri che il raggio spettrale della matrice di iterazione $M^{-1}N$ è minore di $1/2$.

c) Sia $\widetilde{H}$ la matrice ottenuta da H ponendo uguale a 1 l'elemento di posto $(1, 1)$. Dimostrare che il metodo di Jacobi applicato al sistema $\widetilde{H}x = b$ è convergente e dare una limitazione superiore più accurata possibile del raggio spettrale della matrice di iterazione.

Soluzione

La matrice di iterazione del metodo di Jacobi è $-(1/3)H$ dove $H = (h_{i,j})$ è la matrice tridiagonale con elementi $h_{i+1,i} = h_{i,i+1} = 1$ e nulli altrove. La norma infinito di H è 2 quindi per il raggio spettrale ρ vale $\rho(-(1/3)H) \leq 2/3$. Poiché la matrice del sistema A è tridiagonale, sappiamo che il raggio spettrale della matrice di iterazione del metodo di Gauss-Seidel è il quadrato di quello della matrice del metodo di Jacobi. Esso è quindi maggiorato da $(2/3)^2 = 4/9$.

La matrice di iterazione del metodo di Jacobi a blocchi è data dalla matrice tridiagonale a blocchi con blocchi diagonali nulli, blocchi sopra diagonali uguali a $\begin{bmatrix} -1/8 & 0 \\ 3/8 & 0 \end{bmatrix}$, e blocchi sottodiagonali $A^{-1}B^T = \begin{bmatrix} 0 & 3/8 \\ 0 & -1/8 \end{bmatrix}$. I cerchi di Gerschgorin per questa matrice hanno centri 0 e tutti quelli relativi alle righe dalla terza alla terzultima hanno raggio $3/8 + 1/8 = 1/2$. Il primo e l'ultimo cerchio hanno raggio $1/8$, il secondo e il penultimo hanno raggio $3/8$. Quindi il raggio spettrale è minore o uguale a $1/2$. Se la matrice fosse irriducibile, il terzo teorema di Gerschgorin garantirebbe che non ci sono autovalori di modulo $1/2$.

Verifichiamo se la matrice è o meno irriducibile. Dalla sua struttura si vede che nel grafo diretto associato ci sono archi che connettono ogni nodo dispari col nodo dispari successivo e col nodo pari precedente, cioè il nodo $2i + 1$ si collega col nodo $2i + 3$ e col nodo $2i$, quando i valori $2i + 3$ e $2i$ sono compresi tra 1 e n, mentre ogni nodo pari si connette al dispari successivo e al pari precedente, cioè il nodo $2i$ si collega ai nodi $2i + 1$ e a $2i - 1$ quando questi sono compresi tra 1 e n. Per cui ogni nodo dispari si collega a tutti i nodi dispari successivi, ogni nodo pari si collega a tutti i pari precedenti, da un nodo pari si può passare al dispari successivo e da un nodo dispari al pari precedente. Tutti i nodi sono quindi raggiungibili da ogni altro nodo tranne il primo e l'ultimo che non sono raggiungibili. Infatti la prima e l'ultima colonna della matrice sono nulle. Quindi manca la riducibiità della matrice. In effetti il grafo associato

è formato da tre componenti connesse, una formata dal singolo nodo 1, una formata dal singolo nodo n e una formata dai rimanenti nodi $2, \ldots, n-1$. Se la matrice viene perturbata ponendo uguale ad $\epsilon \neq 0$ gli elementi di posto $(2,1)$ e $(n-1,n)$, la nuova matrice così ottenuta è irriducibile poiché al grafo associato sono stati aggiunti due archi che collegano la componente connessa formata dai nodi $2, \ldots, n-1$ al primo e ultimo nodo. Il terzo teorma di Gerschgorin si può allora applicare e si ha che se $\epsilon \leq 1/8$ tutti i raggi dei cerchi di Gerschgorin valgono al più $1/2$ mentre il primo e l'ultimo sono minori di $1/2$. Per cui gli autovalori hanno modulo minore di $1/2$ per ogni ϵ di modulo al più $1/8$. Facendo tendere ϵ a zero, per la continuità degli autovalori si conclude che anche la matrice originale ha raggio spettrale di modulo strettamente minore di $1/2$.

La matrice J di iterazione del metodo di Jacobi applicato al sistema con $\widetilde{H}$ è tridiagonale con elementi principali nulli, elementi sotto diagonali uguali a $-1/3$, elementi sopra diagonali uguali a $-1/3$ ad eccezione dell'elemento in posizione $(1,2)$ che vale -1. Vale cioè

$$\begin{bmatrix} 0 & -1 & & & & \\ -1/3 & 0 & -1/3 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & & -1/3 \\ & & & & -1/3 & 0 \end{bmatrix}$$

Il secondo teorema di Gerschgorin garantisce che il raggio spettrale di questa matrice è minore di 1 e quindi il metodo di Jacobi è convergente. Una stima del raggio spettrale si può ottenere considerando la matrice ottenuta moltiplicando la prima riga per $\epsilon > 0$ e dividendo la prima colonna per ϵ. Questa matrice ha gli stessi autovalori di J però il primo cerchio di Gerschgorin ha raggio $r_1 = \epsilon$ mentre il secondo ha raggio $r_2\epsilon^{-1}/3 + 1/3$. La condizione $r_1 = r_2$ fornisce il valore ottimale $\epsilon = (1 + \sqrt{13})/6$ che costituisce una limitazione superiore al raggio spettrale. □

Esercizio 4 Sia $A = (a_{i,j})$ la matrice $n \times n$, $n \geq 2$, tale che $a_{i,i} = 1$ per $i = 1, \ldots, n$, $a_{1,i} = a_{i,1} = \alpha \neq 0$ per $i = 2, \ldots, n$, $a_{i,j} = 0$ altrimenti. Si denotino con J e G le matrici di iterazione rispettivamente dei metodi di Jacobi e di Gauss-Seidel applicati al sistema lineare $Ax = b$.

a) Determinare J e G e dimostrare che J ha rango 2 mentre G ha rango 1.

b) Determinare il raggio spettrale di J e di G e dare condizioni necessarie e sufficienti di convergenza per i metodi di Jacobi e di Gauss-Seidel

c) Determinare una matrice di Householder P tale che la matrice $B = PAP^T$ abbia la forma

$$B = \begin{bmatrix} 1 & \theta & & & \\ \theta & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

e si confronti il raggio spettrale delle matrici di iterazione dei metodi di Jacobi e di Gauss-Seidel relativi a B con quelli relativi ad A.

Soluzione

a) La matrice A ha la forma

$$A = \begin{bmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & & \\ \vdots & & \ddots & \\ \alpha & & & 1 \end{bmatrix}$$

Quindi per definizione vale

$$J = -\begin{bmatrix} 0 & \alpha & \dots & \alpha \\ \alpha & 0 & & \\ \vdots & & \ddots & \\ \alpha & & & 0 \end{bmatrix}, \quad G = -\begin{bmatrix} 1 & 0 & \dots & 0 \\ \alpha & 1 & & \\ \vdots & & \ddots & \\ \alpha & & & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \alpha & \dots & \alpha \\ 0 & 0 & & \\ \vdots & & \ddots & \\ 0 & & & 0 \end{bmatrix} = uv^T$$

dove $u = [-1, \alpha, \dots, \alpha]^T$ e $v^T = [0, \alpha, \dots, \alpha]$. Chiaramente $G = uv^T$ ha rango 1. Inoltre J si lascia scrivere come $J = e_1 v^T + v e_1^T$, dove $e_1 = [1, 0, \dots, 0]^T$ e quindi ha rango 2.

b) La matrice $G = uv^T$ ha $n-1$ autovalori nulli e un autovalore uguale a $v^T u = \alpha^2(n-1)$. Quindi il suo raggio spettrale è $\alpha^2(n-1)$. Poiché v è ortogonale a e_1, rappresentando la matrice J in una base in cui il primo vettore è e_1 e il secondo vettore è $\hat{v} = (1/\theta)v$, $\theta = \alpha\sqrt{n-1}$, la matrice J si trasforma in $\theta e_1 e_2^T + \theta e_2 e_1^T$ ed ha $n-2$ autovalori nulli e due autovalori uguali a $\pm\theta$. Quindi il raggio spettrale di J è $|\alpha|\sqrt{n-1}$. Si può concludere che i metodi di Jacobi e di Gauss-seidel sono convergenti se e solo se $|\alpha| < 1/\sqrt{n-1}$. Inoltre vale $\rho(G) = \rho(J)^2$, quindi il metodo di Gauss-Seidel, se convergente, ha velocità di convergenza doppia rispetto al metodo di Jacobi

c) È sufficiente scegliere una matrice di Householder Q di ordine $n-1$, tale che $Qe = \sqrt{n-1}e_1$, dove $e, e_1 \in \mathbb{R}^{n-1}$ sono rispettivamente il vettore di tutti uni e il primo versore della base canonica. Ponendo $P = \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$, P risulta di Householder e vale $PAP^T = B$ con $\theta = \alpha\sqrt{n-1}$. Denotando con $\hat{J}$ e $\hat{G}$ le matrici di iterazione dei metodi di Jacobi e di Gauss-Seidel applicati a un sistema con matrice B, risulta $\rho(\hat{J}) = |\theta| = |\alpha|\sqrt{n-1}$, $\rho(\hat{G}) = |\theta^2| = \alpha^2(n-1)$. I raggi spettrali sono gli stessi del caso precedente. □

Esercizio 5 Sia $A = (a_{i,j})$ la matrice $n \times n$ con elementi $a_{i+1,i} = 1, a_{i,n} = \alpha$ per $i = 1, \dots, n-1$, $a_{i,j} = 0$ altrove.

a) Dare condizioni sufficienti su α affinché gli autovalori di A abbiano modulo minore di 1.

b) Dare condizioni sufficienti su α affinché il metodo di Jacobi applicato al sistema lineare $(I + A)x = b$ sia convergente.

c) Dare condizioni sufficienti su α affinché il metodo di gauss-seidel applicato al sistema lineare $(I + A)x = b$ sia convergente.

d) Per $\alpha = 1/n$ dire di quanto viene ridotto l'errore iniziale dopo n passi del metodo di Gauss-seidel e del metodo di Jacobi.

Soluzione

□

Esercizio 6 Dato un numero reale α e un intero $n > 2$ sia $M = (m_{i,j})$ la matrice $n \times n$ triangolare superiore con elementi $m_{i,j} = 1$ per $i \leq j$. Sia inoltre $N = \alpha uv^T$ dove $u = (u_i), v = (v_i) \in \mathbb{R}^n$ sono tali che $u_i = i$ per $i = 1, \ldots, n$ e $v_i = (-1)^{i+1}$ per $i = 1, \ldots, n-1$ e $v_n = 0$. Si consideri il sistema lineare $Ax = b$, con $b \in \mathbb{R}^n$, dove $A = M - N$, e il metodo iterativo $x^{(k+1)} = M^{-1}(Nx^{(k)} + b)$.
a) Dire per quali valori di α il metodo iterativo è convergente.

b) Sia $\alpha = 3/4$ e $n = 1000$. Calcolare il raggio spettrale $\rho(P)$ della matrice $P = M^{-1}N$ e dimostrare che esiste una norma indotta $\|\cdot\|$ tale che $\|P\| = \rho(P)$. Dire quanti passi occorrono per ridurre l'errore iniziale di un fattore 10^{-6} utilizzando la norma individuata.

c) Sia $\alpha = 3/4$ e $n = 1001$. Calcolare il raggio spettrale $\rho(P)$ della matrice $P = M^{-1}N$ e dire se esiste una norma indotta $\|\cdot\|$ tale che $\|P\| = \rho(P)$. Dire quanti passi occorrono per ridurre l'errore iniziale di un fattore 10^{-6} utilizzando la norma infinito.

Soluzione

□

Esercizio 7 Dati i vettori $u = (u_i), v = (v_i), w = (w_i), z = (z_i), b = (b_i) \in \mathbb{R}^n$ si consideri il sistema di $2n$ equazioni e $2n$ incognite $Ax = b$ dove

$$A = \begin{bmatrix} I & \alpha uv^T \\ \alpha wz^T & I \end{bmatrix}.$$

a) Valutare i raggi spettrali $\rho(J)$ e $\rho(G)$ delle matrici di iterazione J e G rispettivamente dei metodi di Jacobi e di Gauss-Seidel applicati al sistema $Ax = b$.

In particolare imostrare che $\rho(G) = \rho(J)^2$ e dare condizioni sufficienti su α per la convergenza.

b) Sia $\alpha = 1/(2\sqrt{n})$, e $u_i = 1$, $v_i = i$, $w_i = 1/i$, $z_i = (-1)^i$ per $i = 1, \ldots, n$. Dire quante iterazioni del metodo di Jacobi e del metodo di Gauss-Seidel sono sufficienti per ridurre l'errore iniziale di un fattore 10^{-6} se $n = 1000$ e se $n = 1001$.

c) Dire se esiste una norma indotta $\|\cdot\|$ tale che $\|J\| = \rho(J)$. Dire se esiste una norma indotta $\|\cdot\|$ tale che $\|G\| = \rho(G)$.

Soluzione

□

Esercizio 8 La matrice A reale $n \times n$ ha m autovalori uguali ad $a \in \mathbb{R}$ e $n-m$ autovalori uguali a $b \in \mathbb{R}$, dove $1 \leq m < n$.

a) Dire sotto quali condizioni su a e b esiste un $\alpha \in \mathbb{R}$ tale che le successioni definite da $x^{(k+1)} = G_\alpha(x^{(k)})$ con $G_\alpha(x) = x + \alpha(f - Ax)$, convergono alla soluzione del sistema $ax = f$, con $f \in \mathbb{R}^n$ per ogni $x^{(0)} \in \mathbb{R}^n$. Nell'ipotesi di esistenza determinare il valore di α che massimizza la velocità di convergenza.

b) Dire sotto quali condizioni su a e b esistono $\alpha, \beta \in \mathbb{R}$ tali che le successioni generate da $x^{(k+1)} = G_\alpha(G_\beta(x^{(k)}))$ convergono alla soluzione del sistema per ogni $x^{(0)} \in \mathbb{R}^n$. In caso di convergenza determinare i valori ottimali di α e β.

c) Con i valori di α e β ottenuti al punto b) studiare la velocità di convergenza delle successioni al punto b) se A ha m autovalori in $[a-\epsilon, a+\epsilon]$ e $n-m$ autovalori in $[b-\epsilon, b+\epsilon]$, dove $|a|, |b| > 1$ e $0 < \epsilon < |a-b|/4$.

Soluzione

□

Esercizio 9 Sia $n \geq 3$ intero e $u = (u_i), v = (v_i) \in \mathbb{R}^{n-1}$, $u, v \neq 0$. Si consideri la matrice $B(u,v) = (b_{i,j}) \in \mathbb{R}^{n \times n}$ tale che $b_{1,i+1} = v_i$, $b_{i+1,1} = u_i$ per $i = 1, \ldots, n-1$, $b_{i,j} = 0$ altrove.

a) Si dimostri che esiste una matrice di Householder P tale che $PB(u,v)P^T = B(\alpha e_1, \widehat{v})$ dove $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{n-1}$ e $\widehat{v} \in \mathbb{R}^{n-1}$. Si deduca che B ha $n-2$ autovalori nulli e due autovalori di modulo $|v^T u|^{1/2}$.

b) Si determinino i raggi spettrali delle matrici J e G rispettivamente dei metodi di Jacobi e di Gauss Seidel applicati al sistema $Ax = b$ dove $A = I - B$. Si diano condizioni sui vettori u e v per la convergenza dei due metodi e si confrontino le velocità di convergenza.

c) Posto $u_i = i$, per $i = 1, \ldots, n-1$, e $v_i = 1$ se $i = 0 \bmod 4$, o se $i = 1 \bmod 4$, $v_i = -1$ altrimenti, si dica se si ha convergenza dei metodi di Jacobi e di Gauss Seidel per $n = 4000$ e $n = 4001$. Nel caso di convergenza si dica per i due metodi quante iterazioni sono sufficienti a ridurre l'errore iniziale di un fattore almeno 10^{10} in norma infinito.

Soluzione

□

Esercizio 10 Sia A la matrice tridiagonale $n \times n$, dove n è pari, tale che $a_{i,i} = (-1)^i \alpha$, per $i = 1, \ldots, n$, $a_{i,i+1} = a_{i+1,i} = 1$ per $i = 1, \ldots, n-1$, dove $\alpha \in \mathbb{R}$. Per la risoluzione del sistema $Ax = b$ si consideri il metodo iterativo $Mx_{k+1} = Nx_k + b$, ottenuto dal partizionamento $A = M - N$, dove M è la matrice diagonale a blocchi 2×2 con blocchi diagonali $\begin{bmatrix} -\alpha & 1 \\ 1 & \alpha \end{bmatrix}$.

a) Dimostrare che la successione $\{x_k\}$ è ben definita per ogni $\alpha \in \mathbb{R}$.

b) Dare condizioni sufficienti su α per la convergenza del metodo e dare una limitazione superiore al raggio spettrale della matrice di iterazione.

c) Confrontare il metodo col metodo di Jacobi, più precisamente mostrare che

esistono valori di α per cui il metodo è convergente mentre il metodo di Jacobi non lo è.

Soluzione

□

Esercizio 11 Sia A una matrice reale simmetrica $n \times n$ con autovalori λ_i, $i = 1, \ldots, n$ e autovettori ortonormali u_i, $i = 1, \ldots, n$. Per la risoluzione del sistema $Ax = b$ si consideri il metodo iterativo $x_{k+1} = M^{-1}(Nx_k + b)$, dove $A = M - N$ e $\det M \neq 0$.
a) Si diano condizioni necessarie e sufficienti sui λ_i affinché il metodo ottenuto con $M = I$ sia convergente.
b) Si diano condizioni necessarie e sufficienti sui λ_i affinché esista un numero reale α tale che il metodo ottenuto con $M = \alpha I$ sia convergente.
c) Supponendo $0 < \lambda_i \leq \lambda_{i+1} < 1$ per $i = 1, 2, \ldots, n-1$ e assumendo di conoscere λ_1 e u_1, si determinino i valori di α per cui il metodo iterativo ottenuto con $M = I + \alpha u_1 u_1^T$ sia convergente. Si determini un valore di α che massimizzi la velocità di convergenza del metodo.
d) Supponendo $\lambda_i \leq \lambda_{i+1} < 1$ per $i = 1, 2, \ldots, n-1$ si diano condizioni necessarie e sufficienti sugli autovalori di A affinché esista un α tale che il metodo ottenuto con $M = I + \alpha u_1 u_1^T$ sia convergente.

Soluzione

□

Esercizio 12 Per n intero positivo, siano $u, v, b \in \mathbb{R}^n$ e D una matrice $n \times n$ diagonale tale che $\det D \neq 0$. Per risolvere il sistema lineare $Ax = b$ con $A = D + uv^T$ si consideri il metodo iterativo $x_{k+1} = M^{-1}(Nx_k + b)$ dove $A = M - N$ e $M = \alpha D$, con $\alpha \in \mathbb{R}$, $\alpha \neq 0$.
a) Dire sotto quali condizioni su u, v, D esiste un α per cui il metodo iterativo è convergente e determinare il raggio spettrale di $P = M^{-1}N$.
b) Determinare il valore di α che minimizza il raggio spettrale di P.
c) Si dimostri che per $\alpha = 1$ l'errore di approssimazione $e_1 = x_1 - x$ ottenuto dopo un passo a partire da un qualunque x_0 è proporzionale a $D^{-1}u$. Se $v = D^{-1}u$ si usi questa proprietà per determinare x_0 tale che $e_1 = 0$.

Soluzione

□

Esercizio 13 Si consideri il sistema lineare $Ax = b$ dove la matrice $n \times n$ $A = (a_{i,j})$ è tale che $a_{i,j} = d_i$ se $i = j$, $a_{1,j} = v_j$, se $j > 1$, $a_{i,1} = u_i$ se $i > 1$ e $a_{i,j} = 0$ altrimenti.
a) Si consideri il metodo iterativo dato dal partizionamento $A = M - N$ con $M = (m_{i,j})$, $m_{1,1} = \alpha$, $m_{i,j} = a_{i,j}$ per $i \geq j$ e $(i, j) \neq (1, 1)$. Dare condizioni sui valori di u_i e v_i affinché esista un α per cui il metodo è convergente.
b) Determinare il valore ottimale di α che massimizza la velocità di convergenza.

c) Se $\alpha = 0$ e $\sum_{i=2}^{n} u_i v_i = 0$ determinare il numero di passi del metodo iterativo sufficienti a ridurre l'errore iniziale di un fattore 10^{-10}.

Soluzione

□

Esercizio 14 Per $n > 2$ intero, si consideri la matrice $n \times n$,

$$C = (c_{i,j}) = \begin{bmatrix} 0 & & & -1 \\ 1 & 0 & & -1 \\ & \ddots & \ddots & \vdots \\ & & 1 & -1 \end{bmatrix}$$

tale che $c_{i+1,i} = 1$, $i = 1, \ldots, n-1$, $c_{i,n} = -1$, $i = 1, \ldots, n$, $c_{i,j} = 0$ altrove. Sia inoltre $A = \alpha I - C$, con $\alpha > 0$ parametro reale.
a) Si dimostri che gli autovalori di C sono le radici $(n+1)$-esime dell'unità diverse da 1.
b) Si studi la convergenza del metodo iterativo dato dal partizionamento $A = M - N$, $M = \alpha I$, $N = C$, per la risoluzione del sistema $Ax = b$, al variare di α.
c) Si studi la convergenza del metodo di Gauss Seidel applicato al sistema $Ax = b$ al variare di α e si confrontino i risultati con quelli del punto b).

Soluzione

□

Esercizio 15 Sia $m \geq 2$ un numero intero e si definiscano le matrici $m \times m$ $B = (b_{i,j})$, $E = (e_{i,j})$ tali che $b_{i,i} = \alpha$, $b_{i,i+1} = -1$, $i = 1, \ldots, m-1$, $b_{i,j} = 0$ altrimenti; $e_{m,1} = -1$, $e_{i,j} = 0$ altrimenti. Si consideri la matrice $2m \times 2m$ definita da

$$A = \begin{bmatrix} B & E \\ E^T & B^T \end{bmatrix}$$

e il sistema lineare $Ax = b$. Dire per quali valori di α reale i metodi iterativi basati sul partizionamento $A = M - N$ sono convergenti alla soluzione del sistema e si valuti il raggio spettrale delle relative matrici di iterazione
a) $M = \alpha I$; b) $M = \begin{bmatrix} \alpha I & 0 \\ E^T & B^T \end{bmatrix}$; c) $M = \begin{bmatrix} \alpha I & E \\ E^T & \alpha I \end{bmatrix}$.
Per $\alpha = 2$ si dica quale dei metodi è più conveniente per ridurre l'errore iniziale di un fattore 10^{-10} e si valuti il numero di operazioni necessario a tale scopo.

Soluzione

□

Esercizio 16 Sia $A \in \mathbb{C}^{n \times n}$ partizionata in 9 blocchi $A_{i,j}$, $i, j = 1, 2, 3$, dove i blocchi $A_{i,i}$, $i = 1, 2, 3$, sono quadrati e $A_{1,2}, A_{2,3}, A_{3,1}$ sono nulli. Si decomponga A in $A = D - B - C$ dove

$$D = \begin{bmatrix} A_{1,1} & 0 & 0 \\ 0 & A_{2,2} & 0 \\ 0 & 0 & A_{3,3} \end{bmatrix}, \; B = -\begin{bmatrix} 0 & 0 & 0 \\ A_{2,1} & 0 & 0 \\ 0 & A_{3,2} & 0 \end{bmatrix}, \; C = -\begin{bmatrix} 0 & 0 & A_{1,3} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

a) Si dimostri che se $\det(\lambda D - B - C) = 0$ e $\lambda \neq 0$, allora $\mu = \lambda^3$ è tale che $\det(\mu(D-B) - C) = 0$ e viceversa.
b) Si dimostri che il metodo iterativo $x_{k+1} = M^{-1}(Nx_k + b)$ per risolvere il sistema lineare $Ax = b$ ottenuto con $M = D$, $N = B + C$ (Jacobi a blocchi) è convergente se e solo se lo è il metodo ottenuto con $M = D - B$, $N = C$ (Gauss-Seidel a blocchi). Si confrontino le velocità asintotiche di convergenza dei due metodi.

Soluzione

□

Esercizio 17 Sia $E \in \mathbb{R}^{n\times n}$, $E = -E^T$ e i l'unità immaginaria tale che $i^2 = -1$.
a) Si dimostri che se $n = 2m$, $m > 0$ intero, allora gli autovalori di E sono $\{\pm i\mu_j,\ \mu_j \in \mathbb{R},\ j = 1, \ldots, m\}$; se $n = 2m + 1$ allora gli autovalori sono $\{0\} \cup \{\pm i\mu_j,\ \mu_j \in \mathbb{R},\ j = 1, \ldots, m\}$, dove $m > 0$ è un intero.
b) Sia inoltre $\rho(E) = 2$, $A = I - E$, $M = \alpha I$, $\alpha \in \mathbb{R}$, $N = M - A$. Si dimostri che il metodo iterativo per risolvere il sistema lineare $Ax = b$, dato dal partizionamento $A = M - N$, è convergente se e solo se $\alpha > 5/2$ e per $\alpha = 5$ raggiunge la massima velocità di convergenza
c) Si determini in funzione di $\rho(E)$ il valore ottimale di α che massimizza la velocità di convergenza del metodo iterativo.

Soluzione

□

Esercizio 18 Si consideri il sistema lineare $Ax = b$, con A matrice reale $n \times n$.

a) Si dimostri che se esistono matrici diagonali D_1 e D_2 tali che D_1AD_2 è fortemente dominante diagonale (per righe o per colonne), oppure è irriducibile e dominante diagonale (per righe o per colonne) allora i metodi di Jacobi e Gauss-Seidel applicati al sistema $Ax = b$ sono convergenti.

b) Sia $a_{i,j} < 0$ per $i \neq j$ e $a_{i,j} > 0$ per $i = j$. Si dimostri che se esiste w con componenti positive tale che Aw oppure $A^T w$ hanno componenti positive (oppure non negative e A è irriducibile) allora i metodi di Jacobi e Gauss-Seidel applicati a $Ax = b$ sono convergenti.

c) Si dica per quali valori di $\alpha \leq 1$ i metodi di Jacobi e Gauss Seidel applicati alla matrice $n \times n$ tridiagonale $A = (a_{i,j})$, $a_{1,1} = \alpha$, $a_{i,i} = 2$, $a_{i,i-1} = a_{i-1,i} = -1$, $i = 2, \ldots, n$, sono convergenti.

Soluzione

□

Esercizio 19 Dato il polinomio $p(x) = x^n - \sum_{i=0}^{n-1} x^i a_i$, dove n è un intero positivo, si associ a $p(x)$ la matrice $n \times n$ $B = (b_{i,j})$, tale che $b_{i+1,i} = 1$, $i = 1, \ldots, n-1$, $b_{1,j} = a_{n-j}$, $j = 1, \ldots, n$, $b_{i,j} = 0$ altrimenti.
a) Si dimostri che $\det(xI - B) = p(x)$.

b) Sia $n = 2m \geq 4$, $p(x) = (x^m - 1)^2$, $A = \alpha I - B$. Si studi al variare di $\alpha \in \mathbb{R}$ la convergenza dei metodi di Jacobi e di Gauss-Seidel applicati al sistema $Ax = b$ e si confrontino le velocità di convergenza dei due metodi.
c) Per $\alpha > 1$ e $0 < \theta < 1$, si valuti in funzione di α e θ il numero k di iterazioni per cui $||e^{(k)}||_\infty \leq \theta ||e^{(0)}||_\infty$, dove $e^{(k)}$ è l'errore generato dopo k passi del metodo di Gauss-Seidel. Si tratti poi il caso particolare in cui $e_m^{(0)} = e_n^{(0)} = 0$.

Soluzione

□

Esercizio 20 Dati $a, b, c, d \in \mathbb{R}^n$ si dimostri che i vettori ortogonali a b e d stanno nel nucleo della matrice $V = ab^T + cd^T$ e che gli autovalori di

$$\begin{bmatrix} b^T a & b^T c \\ d^T a & d^T c \end{bmatrix}$$

sono autovalori di V. Si consideri poi il sistema lineare $Ax = f$, $x, f \in \mathbb{R}^n$, $A = (a_{i,j}) \in \mathbb{R}^{n \times n}$, dove $a_{i,i}, a_{n,i}, a_{1,i} \neq 0$, $i = 1, \ldots, n$, $a_{i,j} = 0$ altrimenti. Si studi la convergenza dei metodi di Jacobi e di Gauss-Seidel applicati a tale sistema e se ne confrontino le velocità di convergenza.

Soluzione

□

Esercizio 21 Sia $A_n = (a_{i,j})$ la matrice tridiagonale $n \times n$ con elementi $a_{i,i} = 2$, $i = 1, \ldots, n$, $a_{i+1,i} = a_{i,i+1} = -1$, $i = 1, \ldots, n-1$.

a) Verificare che A è invertibile e che la prima ed ultima colonna dell'inversa di A sono rispettivamente $\frac{1}{n+1}(n, n-1, \ldots, 2, 1)^T$ e $\frac{1}{n+1}(1, 2, \ldots, n-1, n)^T$.

b) Si supponga che $n = 2m$, si partizioni A_n in 4 blocchi $m \times m$ e si scriva la matrice di iterazione del metodo di Jacobi a blocchi applicato al sistema $Ax = b$. Si analizzi la velocità di convergenza dei metodi di Jacobi e Gauss-Seidel a blocchi applicati al sistema $Ax = b$.

c) Cosa si può dire se $n = 3m$ e A_n è partizionata in 9 blocchi $m \times m$? Cosa si può dire nel caso generale in cui $n = km$ e A_n è partizionata in k^2 blocchi $m \times m$?

Soluzione

□

Esercizio 22 Sia $A = (a_{i,j})$ una matrice simmetrica $n \times n$ con autovalori $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = 1$. Per risolvere il sistema $Ax = b$ si consideri la classe di metodi iterativi definiti da $x^{(k+1)} = x^{(k)} - \theta H(Ax^{(k)} - b)$, dove H è una matrice $n \times n$. Si studi la convergenza nei due casi $H = I$ e $H = 2I - A$ al variare del parametro θ. Si determini in funzione dei λ_i il valore ottimo di θ nei due casi. Si dica quali dei due metodi è più conveniente in termini computazionali (cioè in base al costo computazionale per passo e in base alla velocità di convergenza) nel caso in cui θ abbia il valore ottimo.

Soluzione

□

Esercizio 23 Sia $A = (a_{i,j})$ la matrice $n \times n$ di elementi $a_{i,i} = i$, per $i = 1, \ldots, n$, $a_{i,n} = -1$, $a_{i+1,i} = -1$ per $i = 1, \ldots, n-1$, $a_{i,j} = 0$ altrove. Si scrivano le matrici di iterazione J e G dei metodi di Jacobi e di Gauss Seidel applicati a un sistema lineare con matrice A e si dimostri che tali metodi sono convergenti. Si diano maggiorazioni dei raggi spettrali di J e di G e si individui il metodo che ha la migliore stima di velocità di convergenza.
Sia A_α la matrice ottenuta da A moltiplicando gli elementi diagonali di A per α. Dire per quali valori di α il metodo di Gauss-Seidel applicato ad un sistema con matrice A_α è convergente.

Soluzione

□

Esercizio 24 Sia A una matrice $n \times n$ non singolare di elementi reali e si consideri il partizionamento additivo $A = M - N_1 - N_2$, con M, N_1, N_2 matrici $n \times n$ ad elementi reali, $\det M \neq 0$. Per risolvere il sistema lineare $Ax = b$, $b \in \mathbb{R}^n$, si consideri l'iterazione $x^{(k+1)} = M^{-1}(N_1 x^{(k)} + N_2 x^{(k-1)} + b)$, $k = 1, 2, \ldots$, dove $x^{(0)}, x^{(1)}$ sono scelti in modo arbitrario.
a) Si riscriva l'iterazione nella forma

$$z^{(k+1)} = \mathcal{P} z^{(k)} + q, \quad z^{(k)} = \begin{bmatrix} x^{(k)} \\ x^{(k-1)} \end{bmatrix}$$

esplicitando $\mathcal{P} \in \mathbb{R}^{2n \times 2n}$ e $q \in \mathbb{R}^{2n}$.
b) Si dimostri che gli autovalori di $\mathcal{P}$ sono gli zeri di $\det(\lambda^2 M - \lambda N_1 - N_2)$.
c) Sia M la matrice diagonale con elementi principali $a_{i,i}$, N_1 la matrice strettamente triangolare inferiore con elementi $-a_{i,j}$ per $i > j$. Si dimostri che se A è fortemente dominante diagonale o irriducibilmente dominante diagonale allora la successione $\{x_k\}$ converge per ogni scelta di x_0, x_1.
d) Con le specifiche di M, N_1 e N_2 del punto c), se A è matrice tridiagonale si confronti il raggio spettrale di $\mathcal{P}$ con quello della matrice di iterazione del metodo di Jacobi applicato al sistema lineare $Ax = b$.

Soluzione

□

Esercizio 25 Dati i vettori $a = (a_i), f = (f_i) \in \mathbb{R}^n$ e $b = (b_i), c = (c_i) \in \mathbb{R}^{n-1}$ si consideri la matrice tridiagonale T di dimensione $n \times n$ che ha elementi diagonali $a_i \neq 0$, sottodiagonali b_i e sopradiagonali c_i. Si scrivano le relazioni che legano due iterate successive del metodo di Jacobi applicato al sistema $Tx = f$. Si dimostri che le formule sono stabili all'indietro. Si scriva una function nella sintassi di Octave che, presi in input $a, u, f \in \mathbb{R}^n$, $b, c \in \mathbb{R}^{n-1}$ fornisce in output l'approssimazione $v \in \mathbb{R}^n$ ottenuta applicando un passo del metodo di Jacobi al vettore u.

Soluzione

□

Esercizio 26 Dati i vettori $a = (a_i), f = (f_i) \in \mathbb{R}^n$ e $b = (b_i), c = (c_i) \in \mathbb{R}^{n-1}$ si consideri la matrice tridiagonale T di dimensione $n \times n$ che ha elementi diagonali $a_i \neq 0$, sottodiagonali b_i e sopradiagonali c_i. Si scrivano le relazioni che legano due iterate successive del metodo di Jacobi applicato al sistema $Tx = f$. Si dimostri che le formule sono stabili all'indietro. Si scriva una function nella sintassi di Octave che, presi in input $a, u, f \in \mathbb{R}^n$, $b, c \in \mathbb{R}^{n-1}$ fornisce in output l'approssimazione $v \in \mathbb{R}^n$ ottenuta applicando un passo del metodo di Jacobi al vettore u.

Soluzione

□

Esercizio 27 Si consideri il sistema lineare $Ax = f$ con $f \in \mathbb{R}^n$ e $A = (a_{i,j})$ tale che $a_{i+1,i} = -\alpha$ per $i = 1, \ldots, n-1$, $a_{1,n} = -\beta$, $a_{i,i} = 1$ per $i = 1, \ldots, n$, $a_{i,j} = 0$ altrove.

a) Scrivere le matrici di iterazione dei metodi di Jacobi e Gauss-Seidel applicati a tale sistema, e determinare i loro raggi spettrali. Dare condizioni su α e β necessarie e sufficienti per la convergenza dei due metodi.

b) Confrontare le velocità di convergenza dei due metodi in termini asintotici nel numero di iterazioni.

c) Dimostrare che dopo un passo del metodo di Gauss-Seidel l'errore di approssimazione è proporzionale al vettore $(1, \alpha, \alpha^2, \ldots, \alpha^{n-1})^T$.
Usare questo fatto per ricavare direttamente la soluzione del sistema dal vettore ottenuto applicando un solo passo del metodo ad un qualsiasi vettore iniziale.

Soluzione La matrice A ha la forma

$$A = \begin{bmatrix} 1 & \ldots & 0 & -\beta \\ -\alpha & 1 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & \ldots & -\alpha & 1 \end{bmatrix}$$

Le matrici di Jacobi e di Gauss Seidel hanno la forma

$$J = \begin{bmatrix} 0 & \ldots & 0 & \beta \\ \alpha & 0 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & \ldots & \alpha & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & \ldots & 0 & 0 \\ -\alpha & 1 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & \ldots & -\alpha & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & \ldots & 0 & \beta \\ 0 & & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ 0 & \ldots & 0 & 0 \end{bmatrix}$$

per cui

$$G = \begin{bmatrix} 1 & \ldots & 0 & 0 \\ \alpha & 1 & \ldots & 0 \\ \vdots & \ddots & \ldots & \vdots \\ \alpha^{n-1} & \ldots & \alpha & 1 \end{bmatrix} \beta e_1 e_n^T = \beta u e_n^T, \quad u = (1, \alpha, \alpha^2, \ldots, \alpha^{n-1})^T$$

Gli autovalori λ di J sono tali che $\det(\lambda - J) = 0$. Calcolando il determinante con la regola di Laplace sulla prima riga si ottiene $\det(\lambda I - J) = \lambda^n - \beta\alpha^{n-1}$ da cui gli autovalori di J sono le radici n-esime di $\beta\alpha^{n-1}$ e il raggio spettrale è quindi $\rho(J) = |\beta\alpha^{n-1}|^{1/n}$.

La matrice $G = ue_n^T$ ha l'autovalore 0 di molteplicità $n-1$ corrispondente agli autovettori ortogonali a e_n, e l'autovalore $\mu = \beta e_n^T u = \beta\alpha^{n-1}$ corrispondente all'autovettore u. pertanto $\rho(G) = |\beta\alpha^{n-1}|$. Si osserva quindi che $\rho(G) = \rho(J)^n$. Condizione necessaria e sufficiente per la convergenza di entrambi i metodi è $|\beta\alpha^{n-1}| < 1$. Poiché il raggio spettrale dà la riduzione asintotica media per passo, dalla relazione $\rho(G) = \rho(J)^n$ si deduce che il metodo di Gauss Seidel impiega asintoticamente un numero di iterazioni n volte inferiore a quello richiesto dal metodo di Jacobi.

Se x^* è la soluzione del sistema e se $\epsilon^{(0)} = x^{(0)} - x^*$ è l'errore iniziale allora dalla teoria sappiamo che dopo un passo del metodo di Gauss-Seidel l'errore è $\epsilon^{(1)} = Ge^{(0)} = ue_n^T\epsilon^{(0)} = \epsilon_n^{(0)}u =: \xi u$.

Vale quindi $x^{(1)} - x^* = \xi u$ per calcolare x^* conoscendo $x^{(1)}$ e u, basta calcolare ξ. Poiché $Ax^* = f$ si ha $\xi Au = Ax^{(1)} - f$ da cui, considerando ad esempio la prima componente, $\xi = (x_1^{(1)} - \beta x_n^{(1)} - f_1)/(1 - \beta\alpha^{n-1})$. □

Esercizio 28 Siano A, D_1, D_2 matrici reali $n \times n$, D_1 e D_2 matrici diagonali non singolari. Si consideri il sistema lineare $Ax = f$ con $f \in \mathbb{R}^n$ e il sistema equivalente $By = g$ con $B = D_1AD_2$, $x = D_2y$, $g = D_1f$.

a) Si metta a confronto il metodo di Jacobi applicato al sistema $Ax = f$ e al sistema $By = g$. Si svolga la stessa analisi per il metodo di Gauss-Seidel.

b) Si scrivano le matrici di iterazione dei metodi di Jacobi e di Gauss Seidel applicati al sistema $Ax = b$ dove $A = uv^T$ con $u, v \in \mathbb{R}^n$ vettori con componenti non nulle e si calcolino i loro autovalori (si osservi che $uv^T = D_uED_v$, con E matrice di elementi uguali a 1, D_u e D_v matrici diagonali con elementi diagonali rispettivamente u_i, v_i).

Soluzione

Riferimenti bibliografici

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

Zeri di funzioni

Dario A. Bini, B. Meini, Università di Pisa

1 dicembre 2021

Sommario

Questo modulo didattico contiene risultati relativi ai metodi per approssimare numericamente gli zeri di una funzione continua.

Un problema interessante dal punto di vista computazionale consiste nel calcolare le soluzioni di una equazione o di un sistema di equazioni non lineari. Nelle applicazioni questo problema si incontra in diverse forme. Ad esempio, nella progettazione di robot industriali per l'assemblaggio di oggetti costituiti da più parti, la configurazione che il robot deve assumere per poter svolgere le sue funzioni dipende dalla soluzione di un sistema di equazioni.

Infatti, date le coordinate del punto dello spazio che l'estremità del braccio di un robot deve raggiungere, occorre determinare i parametri che definiscono la configurazione del robot, tipo gli angoli tra i segmenti che costituiscono il braccio e lunghezze di tali segmenti, in modo che l'estremità del braccio occupi il punto che ha quelle coordinate. Questi parametri e le coordinate del punto di arrivo sono legati da un sistema di equazioni tipicamente non lineare.

Ad esempio, nel disegno in figura 1 è rappresentato un robot semplificato formato da due bracci. Il primo, di lunghezza u fissata, può ruotare nel piano attorno all'origine. Il secondo, di lunghezza $v < u$ fissata, può ruotare sempre nel piano attorno all'estremità libera del primo braccio. All'estremità del secondo braccio, di coordinate (a, b) è collocata la pinza del robot. È evidente che

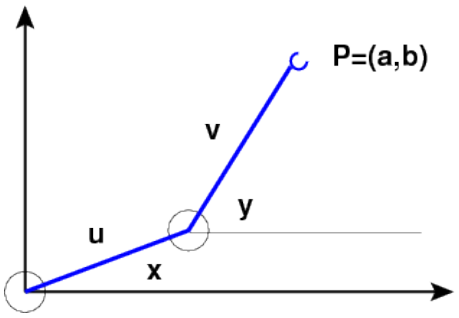


Figura 1: Robot costituito da due bracci articolati

la pinza del robot può occupare tutti i punti della corona circolare di centro l'origine e raggi $u-v$ e $u+v$. Se chiamiamo con x e y rispettivamente gli angoli che il primo e il secondo braccio formano con una retta orizzontale, valgono le relazioni

$$\begin{cases} a = u\cos x + v\cos y \\ b = u\sin x + v\sin y \end{cases}$$

Se, dati a e b vogliamo determinare gli angoli x e y per fare raggiungere alla pinza del robot la posizione (a, b) dobbiamo risolvere il sistema non lineare di due equazioni e due incognite descritto sopra.

Purtroppo solo in rari casi siamo in grado di dare delle formule esplicite che ci permettono di rappresentare le soluzioni di una equazione o di un sistema. Si pensi ad esempio al caso dei familiari polinomi $p(x) = \sum_{i=0}^{n} a_i x^i$ in una variabile x. Se il grado n è 1 o 2 sappiamo esprimere le radici di $p(x)$ con delle formule esplicite.

Nel caso di equazioni cubiche e di equazioni quartiche esistono formule esplicite pubblicate da Girolamo Cardano nel 1545. La soluzione nel caso $n = 3$ fu comunicata a Cardano da Niccolò Tartaglia, mentre la soluzione nel caso $n = 4$ fu trovata da Scipione del Ferro, studente di Cardano.

Nel caso di polinomi di grado maggiore o uguale a 5 non esistono formule che permettano di esprimere le radici in termini di radicali e operazioni aritmetiche. Questo segue dalla Teoria di Galois.

Quindi dal punto di vista computazionale occorre individuare dei metodi che permettano di *approssimare* le soluzioni di una equazione o di un sistema di equazioni attraverso la generazione di successioni che ad esse convergano.

In questo articolo ci occupiamo del progetto ed analisi di metodi per generare tali successioni, assieme all'introduzione di strumenti generali per la loro analisi.

1 Zeri di funzioni continue da $\mathbb{R}$ in $\mathbb{R}$

Ci occupiamo qui del caso più semplice che si può incontrare: il caso in cui è data una funzione $f(x) : [a, b] \to \mathbb{R}$ continua sull'intervallo $[a, b]$, tale che $f(a)f(b) < 0$. Con queste ipotesi, per un noto teorema dell'analisi, esiste almeno un punto $\alpha \in [a, b]$ per cui $f(\alpha) = 0$. Vogliamo allora generare una successione $\{x_k\}_k$ che converga ad α.

Uno dei metodi più semplici per fare ciò è il metodo della *bisezione* detto anche *metodo dicotomico*. Esso si basa su una strategia antica e universale che viene generalmente applicata con successo in molte situazioni, dall'arte militare (dai tempi di Giulio Cesare) alla politica e all'informatica: la strategia del *divide et impera* che in letteratura anglosassone diventa *divide-and-conquer*.

Il metodo funziona nel modo seguente:

- si pone $a_0 = a$, $b_0 = b$
- per $k = 0, 1, \ldots$, si calcola il punto medio $c_k = (a_k + b_k)/2$ del segmento $[a_k, b_k]$, assieme al valore di $f(c_k)$

- se $f(c_k) = 0$ abbiamo trovato una soluzione $\alpha = c_k$ e abbiamo finito; altrimenti, se $f(c_k)f(a_k) < 0$, deduciamo che la funzione si annulla in almeno un punto α del sottointervallo di sinistra $[a_k, c_k]$; se invece $f(c_k)f(a_k) > 0$ si deduce che la funzione si annulla in un punto α_k del sottointervallo di destra $[c_k, b_k]$;

- nel primo caso ripetiamo il procedimento applicandolo all'intervallo $[a_k, c_k]$, cioè poniamo $a_{k+1} = a_k$, $b_{k+1} = c_k$, nel secondo caso lo applichiamo all'intervallo $[c_k, b_k]$, cioè poniamo $a_{k+1} = c_k$, $b_{k+1} = b_k$, finché non abbiamo ottenuto un intervallo di ampiezza sufficientemente piccola, cioè finché $b_{k+1} - a_{k+1} < \epsilon$.

È evidente che con questa strategia l'ampiezza dell'intervallo corrente si riduce della metà ad ogni passo che viene fatto. In questo modo, dopo k passi abbiamo un intervallo di ampiezza $b_k - a_k = \frac{1}{2^k}(b - a)$. Quindi la quantità $\epsilon_k = b_k - a_k$, maggiorazione dell'errore di approssimazione di α, converge a zero in modo esponenziale.

Apparentemente la convergenza esponenziale a zero di ϵ_k denota una "velocità" di convergenza alta. In effetti non è proprio così visto che più tardi introdurremo dei metodi in cui l'errore di approssimazione converge a zero in modo *doppiamente esponenziale* cioè è limitato superiormente da

$$\beta^{p^k} \quad \text{con } 0 < \beta < 1, \quad p \geq 2.$$

Dal punto di vista computazionale il metodo di bisezione conviene implementarlo nel modo descritto nel codice *Octave* del listato 1.

In particolare, poiché il calcolo viene svolto in aritmetica floating point, la condizione di arresto viene data sull'errore relativo mediante la diseguaglianza $b - a \leq \epsilon \min\{|a|, |b|\}$. Inoltre viene messo un limite massimo al numero di passi di bisezione. Un'altra caratteristica dell'implementazione è che vengono usate solo due variabili `a`, `b` per generare gli intervalli di inclusione: di fatto viene mantenuto solo l'ultimo intervallo generato e la storia passata viene dimenticata. Poiché il segno della funzione $f(x)$ nell'estremo sinistro degli intervalli generati è sempre lo stesso, così come il segno della funzione nell'estremo destro, è sufficiente calcolarsi questo segno una volta per tutte usando la variabile `fa` e calcolare ad ogni passo solamente il valore di $f(c)$ da confrontarsi con `fa`. In questo modo il costo di ogni iterazione si riduce al calcolo del punto c e del valore di $f(c)$. Il codice presuppone che sia stata definita una function `f(x)`.

In una esecuzione in aritmetica floating point del metodo di bisezione i valori effettivamente calcolati di $f(x)$ sono affetti da errore. Se supponiamo che il valore effettivamente calcolato sia compreso tra $f(x) - \delta$ e $f(x) + \delta$, dove $\delta > 0$ è una quantità nota, allora il problema del calcolo degli zeri di $f(x)$ si trasforma in quello del calcolare le intersezioni tra la striscia di piano formata dai punti $\{(x, y) \in \mathbb{R}^2 : \quad f(x) - \delta \leq y \leq f(x) + \delta\}$ e l'asse delle x, come mostrato in figura 2.

Allora in questa forma alla soluzione α del problema originale viene sostituito *l'intervallo di incertezza* rappresentato in figura dal segmento verde. Se $f(x)$ è

Listing 1: Function bisezione

```
function alfa=bisezione(a,b)
% applica il metodo di bisezione per approssimare uno zero
% della funzione f(x) sull'intervallo [a,b] dove f(x) e' una
% funzione predefinita tale che f(a)f(b)<0
   maxit = 100;
   fa = f(a);
   if fa*f(b)>0
      disp('Dati inconsistenti')
      break
   end
   for k=1:maxit
      c = (a+b)/2;
      fc = f(c);
      if fc*fa<=0
         b = c;
      else
         a = c;
      end
      if b-a < eps*min(abs(a),abs(b))
         break
      end
   end
   if k==maxit
      disp('attenzione: raggiunto il numero massimo di iterazioni')
   end
   alfa=c;
```

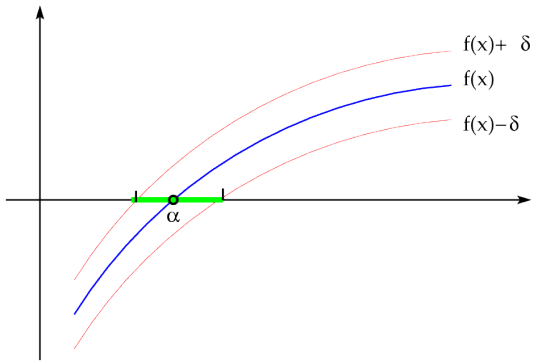


Figura 2: Problema degli zeri in caso di incertezza numerica

derivabile con continuità si riesce a dare una approssimazione dell'intervallo di incertezza sostituendo al grafico di $f(x)+\delta$ e di $f(x)-\delta$ la loro approssimazione lineare data dalle rette con coefficinete angolare $f'(\alpha)$ passanti rispettivamente per (α, δ) e $(\alpha, -\delta)$. Infatti è semplice verificare che le intersezioni di queste rette con l'asse delle x sono date da $\alpha - \delta/f'(\alpha)$ e $\alpha + \delta/f'(\alpha)$.

L'intervallo

$$[\alpha - \frac{\delta}{f'(\alpha)}, \alpha + \frac{\delta}{f'(\alpha)}]$$

fornisce l'approssimazione al primo ordine in δ dell'*intervallo di incertezza*. In queste condizioni di incertezza numerica dove la $f(x)$ non può essere nota in modo esatto, l'obiettivo del calcolo consiste nel trovare un punto dentro l'intervallo di incertezza. Il metodo di bisezione fa esattamente questo.

È interessante osservare che più piccola in valore assoluto è la derivata prima $f'(\alpha)$ e tanto più ampio è l'intervallo di incertezza. In altri termini il reciproco della derivata prima $f'(\alpha)$ fornisce una misura del condizionamento numerico del problema del calcolo dello zero α di $f(x)$. Intuitivamente, più è orizzontale il grafico della curva nel punto in cui interseca l'asse x e tanto più ampia è l'intersezione della striscia di piano individuata da $f(x) \pm \delta$ e l'asse delle x.

2 Metodi del punto fisso

Sebbene il metodo di bisezione sia efficace e robusto, non sempre è adatto per il calcolo numerico di zeri di funzioni. Infatti, il numero di passi che esso richiede può essere molto elevato e ciò diventa un grosso inconveniente quando il costo del calcolo del valore di $f(x)$ è alto. In questo caso i *metodi del punto fisso* sono più efficaci.

I metodi del punto fisso si ottengono trasformando il problema $f(x) = 0$ del calcolo di zeri di una funzione in un problema "del punto fisso" del tipo

$$x = g(x)$$

che consiste nel trovare quei numeri α che sono trasformati dalla funzione $g(x)$ in sé stessi e per questo detti *punti fissi* di $g(x)$.

La trasformazione da $f(x) = 0$ in $x = g(x)$ si può ottenere in infiniti modi diversi. Ad esempio, data una qualsiasi funzione $h(x) \neq 0$ si può porre

$$g(x) = x - f(x)/h(x).$$

La formulazione data in termini di problema di punto fisso permette di generare in modo naturale una successione di punti che, se convergente, converge ad un punto fisso di $g(x)$ purché $g(x)$ sia una funzione continua. Tale successione è semplicemente data da

$$\begin{cases} x_{k+1} = g(x_k), \quad k = 0, 1, 2, \ldots, \\ x_0 \in \mathbb{R} \end{cases} \tag{1}$$

Ci riferiamo all'espressione (1) come al *metodo di punto fisso* o *metodo di iterazione funzionale* associato a $g(x)$.

Se $\lim_k x_k = \ell$, nell'ipotesi di continuità di $g(x)$ si ha

$$\ell = \lim_k x_{k+1} = \lim_k g(x_k) = g(\lim_k x_k) = g(\ell).$$

Cioè ℓ è un punto fisso di $g(x)$.

Diventa quindi cruciale poter dare condizioni facilmente verificabili affinché la successione generata $\{x_k\}$ converga per x_0 in un intorno di α. Per questo vale il seguente risultato che enunciamo e dimostriamo in una forma adatta ad un approccio computazionale.

Teorema 1 (del punto fisso) *Sia $\mathcal{I} = [\alpha - \rho, \alpha + \rho]$ e $g(x) \in C^1(\mathcal{I})$, dove $\alpha = g(\alpha)$ e $\rho > 0$. Si denoti con $\lambda = \max_{|x-\alpha|\le\rho} |g'(x)|$. Se $\lambda < 1$ allora per ogni $x_0 \in \mathcal{I}$, posto $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, vale*

$$|x_k - \alpha| \le \lambda^k \rho,$$

per cui $x_k \in \mathcal{I}$ e $\lim_k x_k = \alpha$. Inoltre α è l'unico punto fisso di $g(x)$ in $[a, b]$.

Dim. Si procede per induzione su k. Per $k = 0$, poiché $x_0 \in \mathcal{I}$, vale $|x_0 - \alpha| \le \rho = \rho\lambda^0$. Assumiamo che $|x_k - \alpha| \le \lambda^k \rho$. Allora vale

$$x_{k+1} - \alpha = g(x_k) - g(\alpha) = g'(\xi_k)(x_k - \alpha), \qquad |\xi_k - \alpha| < |x_k - \alpha|.$$

Dove abbiamo usato il fatto che $x_{k+1} = g(x_k)$ e $\alpha = g(\alpha)$, e abbiamo usato il teorema del valor medio di Lagrange che è applicabile essendo $x_k \in \mathcal{I}$ per ipotesi induttiva e $\alpha \in \mathcal{I}$. Poiché $x_k \in \mathcal{I}$ è anche $\xi_k \in \mathcal{I}$ e si ha $|g'(\xi_k)| \le \lambda$ per cui, per l'ipotesi induttiva vale

$$|x_{k+1} - \alpha| \le |g'(\xi_k)|\,|x_k - \alpha| \le \lambda \cdot \lambda^k \rho = \lambda^{k+1}\rho.$$

Per l'unicità di α si procede per assurdo. Se $\beta \ne \alpha$ fosse un altro punto fisso in $[a, b]$, dalla relazione

$$\alpha - \beta = g(\alpha) - g(\beta) = g'(\xi)(\alpha - \beta)$$

si dedurrebbe che $g'(\xi) = 1$ il che è assurdo. □

Così come è formulato il teorema 1 non sembra di utilità pratica. Infatti, già la conoscenza degli estremi $\alpha - \rho$ e $\alpha + \rho$ dell'intervallo $\mathcal{I}$ in cui è definita $g(x)$ permette di calcolare il punto fisso α semplicemente prendendone la semisomma. Nella pratica possiamo assumere di avere a disposizione un intervallo $[a, b]$ a cui appartiene α e di sapere che su tale intervallo $|g'(x)| < 1$. Sotto queste ipotesi siamo certi che con almeno una delle due scelte $x_0 = a$, $x_0 = b$ la successione è ben definita e converge al punto fisso α. Infatti se α appartiene alla metà sinistra dell'intervallo $[a, b]$, il teorema vale con $\rho = \alpha - a$ sull'intevallo $[\alpha - \rho, \alpha + \rho]$ di cui a è estremo sinistro. Se α appartiene alla metà destra di $[a, b]$ allora

il teorema vale con $\rho = b - \alpha$ sull'intervallo $[\alpha - \rho, \alpha + \rho]$ di cui b è estremo destro. Dal punto di vista operativo basta scegliere a caso uno dei due estremi e calcolare gli elementi della successione $\{x_k\}$. Se c'è convergenza il problema è risolto. Se invece qualche x_k cade fuori dell'intervallo $[a, b]$ allora si arrestano le iterazioni e si riparte con x_0 uguale all'altro estremo.

Si noti che, a differenza dei metodi iterativi per sistemi lineari in cui la convergenza della successione generata vale per ogni scelta del punto iniziale, per i metodi del punto fisso applicati a funzioni non lineari la convergenza vale in un intorno opportuno del punto fisso α. Questo fatto si usa denotare con l'espressione *convergenza locale*. Mentre col termine *convergenza globale* ci si riferisce al fatto che le successioni generate convergono *qualunque* sia il punto iniziale x_0.

Nelle situazioni concrete in cui $g(x)$ viene calcolata con l'aritmetica floating point, il risultato fornito dal teorema 1 non vale più. È però possibile dimostrare una versione equivalente valida per l'aritmetica floating point.

Teorema 2 *Nelle ipotesi del teorema 1 sia $\tilde{x}_k$ la successione generata da*

$$\tilde{x}_{k+1} = g(\tilde{x}_k) + \delta_k$$

dove $|\delta_k| \leq \delta$ è l'errore commesso nel calcolo di $g(\tilde{x}_k)$ in aritmetica floating point e δ è una quantità nota a priori. Posto $\sigma = \delta/(1 - \lambda)$, se $\sigma < \rho$ risulta

$$|\tilde{x}_k - \alpha| \leq (\rho - \sigma)\lambda^k + \sigma$$

Dim. Si procede per induzione su k. Per $k = 0$ la diseguaglianza è soddisfatta. Dimostriamo il passo induttivo. Vale

$$\tilde{x}_{k+1} - \alpha = g(\tilde{x}_k) - g(\alpha) + \delta_k = g'(\xi_k)(\tilde{x}_k - \alpha) + \delta_k,$$

dove $|\xi_k - \alpha| < |x_k - \alpha|$. Prendendo i valori assoluti e applicando l'ipotesi induttiva si ha

$$|\tilde{x}_{k+1} - \alpha| \leq \lambda|\tilde{x}_k - \alpha| + \delta \leq \lambda((\rho - \sigma)\lambda^k + \sigma) + \delta$$

La tesi segue dal fatto che $\lambda\sigma + \delta = \lambda\delta/(1 - \lambda) + \delta = \sigma$. □

Il teorema 2 ci dice che la distanza di $\tilde{x}_k$ dal punto fisso α è limitato dalla somma di due parti. La prima converge a zero in modo esponenziale su base λ. La seconda è costante ed è data da $\sigma = \delta/(1 - \lambda)$. Questa seconda parte rappresenta l'intervallo di incertezza sotto il quale non è consentito andare. Si osservi che per la funzione $f(x) = x - g(x)$ che ha α come zero, l'intervallo di incertezza è dato al primo ordine da $[\alpha - \delta/|f'(\alpha)|, \alpha + \delta/|f'(\alpha)|]$. Cioè essendo $f'(x) = 1 - g'(x)$, se $g'(x) > 0$, l'intervallo di incertezza è contenuto in $[\alpha - \sigma, \alpha + \sigma]$.

È utile ricordare dalla dimostrazione del teorema 1 che

$$x_{k+1} - \alpha = g'(\xi_k)(x_k - \alpha). \tag{2}$$

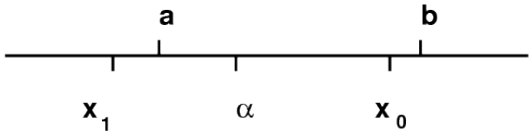


Figura 3: Punto fisso non centrato nell'intervallo

Ciò implica che se $0 < g'(x) < 1$ nell'intervallo $x \in [\alpha - \rho, \alpha + \rho]$ allora, $x_{k+1} - \alpha$ ha lo stesso segno di $x_k - \alpha$. In altri termini, se $x_0 > \alpha$ allora $x_k > \alpha$ per ogni valore di k, inoltre $\alpha < x_{k+1} < x_k$. Cioè la successione è decrescente. Analogamente se $x_0 < \alpha$ la successione $\{x_k\}$ è crescente.

Analogamente, nell'ipotesi $-1 < g'(x) < 0$, segue dalla (2) che se $x_k > \alpha$ allora $x_{k+1} < \alpha$, se $x_k < \alpha$ allora $x_{k+1} > \alpha$. Cioè la successione generata ha un comportamento alternato: le sottosuccessioni $\{x_{2k}\}$ e $\{x_{2k+1}\}$ convergono, una crescendo, l'altra decrescendo, al punto fisso α.

Questa osservazione permette di determinare il comportamento della convergenza mediante lo studio del segno della derivata di $g(x)$. Un'altra osservazione interessante è che se $[a, b]$ è un intervallo qualsiasi che contiene α, non necessariamente centrato in α, e se $g'(x) > 0$ allora per ogni $x_0 \in [a, b]$ la successione generata $\{x_k\}$ converge in modo monotono ad α. Mentre se $g'(x) < 0$ e se $[a, b]$ non è centrato in α può accadere che per un particolare $x_0 \in [a, b]$ risulti $x_1 \notin [a, b]$ anche se più vicino ad α di x_0. Per cui $g(x)$ può non esistere in x_1 o possono non essere più soddisfatte le ipotesi sulla derivata di $g(x)$. Ciò è mostrato nella figura 3.

È evidente quindi la necessità di mettere nelle ipotesi del teorema 1 la condizione $[a, b] = [\alpha - \rho, \alpha + \rho]$.

Un'altra osservazione interessante è che, come risulta dalla dimostrazione del teorema 1, la quantità ξ_k sta nell'intervallo aperto di estremi α e x_k. Ciò implica che se $x_k \to \alpha$ allora $\xi_k \to \alpha$ quindi il fattore di riduzione dell'errore $|g'(\xi_k)|$ si avvicina sempre di più al valore $|g'(\alpha)|$. Questo implica che se $g'(\alpha) = 0$ allora man mano che le iterazioni procedono, il fattore di riduzione dell'errore diventa sempre più vicino a zero. Cioè la convergenza è sempre più rapida. Tra poco formalizzeremo questa proprietà.

L'ultima osservazione che facciamo riguarda l'interpretazione grafica dei metodi del punto fisso. Il fatto che $x_{k+1} = g(x_k)$ ci permette di descrivere graficamente la costruzione della successione come riportato nelle figure 4, 5 che mostrano il caso di una funzione $g(x)$ crescente e di una decrescente in cui la successione $\{x_k\}$ è rispettivamente monotona e alternata. Nelle figure sono riportati i primi due passi dell'iterazione. Il terzo passo porterebbe ad un punto difficilmente distinguibile da α nella risoluzione dello schermo o della stampante.

Un aspetto computazionalmente interessante dei metodi del punto fisso riguarda le condizioni di arresto e la determinazione di limitazioni *a posteriori* dell'errore. È abbastanza naturale per questo considerare la quantità $|x_k - x_{k+1}|$ e arrestare le iterazioni quando $|x_k - x_{k+1}| \leq \epsilon$ per un valore di ϵ fissato. Vediamo

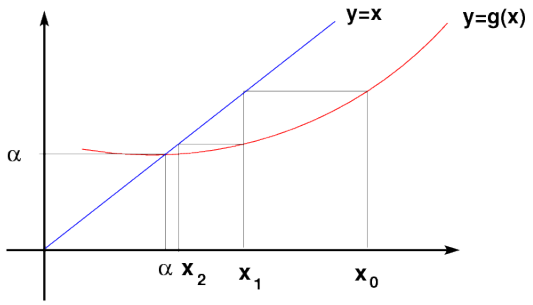


Figura 4: Convergenza monotona

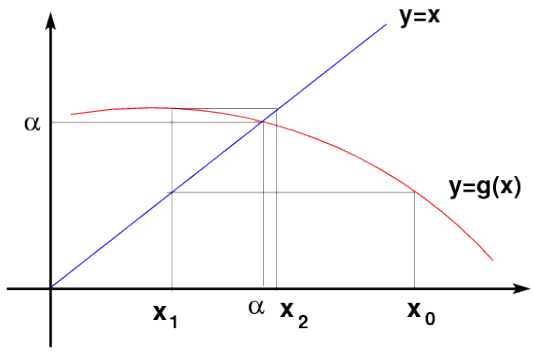


Figura 5: Convergenza alternata

ora come la quantità $|x_k - x_{k+1}|$ sia legata all'errore assoluto di approssimazione $|x_k - \alpha|$. Vale

$$x_k - x_{k+1} = x_k - \alpha - (x_{k+1} - \alpha) = (x_k - \alpha) - g'(\xi_k)(x_k - \alpha) = (1 - g'(\xi_k))(x_k - \alpha),$$

per un opportuno ξ_k tale che $|\xi_k - \alpha| < |\xi_k - \alpha|$. Da questo si ottiene

$$|x_k - \alpha| = \left| \frac{x_k - x_{k+1}}{1 - g'(\xi_k)} \right|.$$

Quindi la condizione di arresto $|x_k - x_{k+1}| \leq \epsilon$ ci fornisce la limitazione superiore all'errore

$$|x_k - \alpha| \leq \frac{1}{|1 - g'(\xi_k)|} \epsilon.$$

È interessante osservare che se $g'(x) < 0$ allora il denominatore nell'espressione precedente è maggiore di 1 e quindi la condizione di arresto ci fornisce una limitazione dell'errore significativa. Se invece $g'(x) > 0$ allora il denominatore è minore di 1 e può diventare arbitrariamente vicino a zero a seconda di quanto i valori di $g'(x)$ si avvicinano a 1. Quindi nel caso di $g'(x)$ "vicino" a 1 non solo si ha convergenza lenta ma anche la limitazione a posteriori dell'errore è poco significativa nel senso che per avere un errore di approssimazione dell'ordine della precisione di macchina occorre scegliere una condizione di arresto con un valore di ϵ molto più piccolo.

3 Velocità di convergenza

Nella scelta di un metodo iterativo è cruciale avere informazioni sulla velocità di convergenza delle successioni generate dal metodo. Per questo diamo prima alcune definizioni sulla convergenza di successioni che applicheremo alle successioni generate da metodi del punto fisso, e poi dimostriamo alcuni risultati computazionalmente utili.

Definizione 1 Sia $\{x_k\}$ una successione tale che $\lim_k x_k = \alpha$. Supponiamo esista il limite

$$\gamma = \lim_k \left| \frac{x_{k+1} - \alpha}{x_k - \alpha} \right|. \tag{3}$$

La convergenza di $\{x_k\}$ a α è detta

- lineare (o geometrica) se $0 < \gamma < 1$,
- sublineare se $\gamma = 1$,
- superlineare se $\gamma = 0$.

$x_k - \alpha$	k
$1/k$	10^{16}
$1/2^k$	54
$1/2^{2^k}$	6

Tabella 1

Nel caso di convergenza superlineare, se $p > 1$ è tale che esiste il limite

$$\lim_k \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = \sigma, \qquad 0 < \sigma < \infty,$$

si dice che la successione converge con ordine p. Se $p = 2$ si dice che la convergenza è quadratica, se $p = 3$ si dice che la convergenza è cubica.

Ad esempio, se $\gamma < 1$, la successione $x_k = \gamma^k$ ha convergenza lineare a zero. La successione $x_k = \gamma^{p^k}$ converge a zero in modo superlineare e ha ordine di convergenza p. La successione $x_k = 1/k$ converge a zero in modo sublineare. Per apprezzare meglio la differenza tra le velocità di convergenza, riportiamo nella tabella 1 il più piccolo indice k per cui il valore di $|x_k - \alpha|$ è inferiore a 10^{-16}, avendo scelto $\gamma = \sigma = 1/2$, $p = 2$.

Per successioni generate dal metodo del punto fisso è possibile determinare la velocità di convergenza semplicemenete calcolando delle derivate come viene mostrato nei seguenti teoremi.

Teorema 3 *Sia $g(x) \in C^1([a, b])$ e $\alpha \in (a, b)$ tale che $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a, b]$ tale che la successione (1) converge linearmente ad α con fattore γ definito in (3) allora $|g'(\alpha)| = \gamma$. Viceversa, se $0 < |g'(\alpha)| < 1$ allora esiste un intorno $\mathcal{I}$ di α contenuto in $[a, b]$ tale che per ogni $x_0 \in \mathcal{I}$ la successione $\{x_k\}$ generata da (1) converge ad α in modo lineare con fattore $\gamma = |g'(\alpha)|$.*

Dim. Se $\{x_k\}$ è la successione definita da $x_{k+1} = g(x_k)$ che converge linearmente al punto fisso α, allora vale $(x_{k+1} - \alpha)/(x_k - \alpha) = (g(x_k) - g(\alpha))/(x_k - \alpha)$, e per il teorema del valor medio segue

$$\frac{x_{k+1} - \alpha}{x_k - \alpha} = g'(\xi_k), \qquad \text{con } |\xi_k - \alpha| < |x_k - \alpha|.$$

Per cui, poiché $x_k \to \alpha$, anche $\xi_k \to \alpha$. Risulta allora

$$|g'(\alpha)| = \lim_k \left| \frac{x_{k+1} - \alpha}{x_k - \alpha} \right|, \tag{4}$$

e quindi $0 < |g'(\alpha)| < 1$. Viceversa, se $0 < |g'(\alpha)| < 1$, poiché g' è continua, esiste un intorno $[\alpha - \rho, \alpha + \rho]$ contenuto in $[a, b]$ in cui $|g'(x)| < 1$. Per il teorema del punto fisso, le successioni generate a partire da x_0 in questo intorno

convergono ad α e per queste successioni vale l'analisi fatta per dimostrare la prima parte per cui vale (4). Conseguentemente tutte queste successioni convergono linearmente. □

Teorema 4 *Sia $g(x) \in C^1([a,b])$ e $\alpha \in (a,b)$ tale che $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a,b]$ tale che la successione (1) converge sublinearmente ad α allora $|g'(\alpha)| = 1$. Viceversa, se $|g'(\alpha)| = 1$, esiste un intorno $\mathcal{I}$ di α contenuto in $[a,b]$ tale che per ogni $x \in \mathcal{I}$, $x \neq \alpha$ è $|g'(x)| < 1$, e $g'(x)$ non cambia segno su $\mathcal{I}$ allora tutte le successioni $\{x_k\}$ generate da (1) con $x_0 \in \mathcal{I}$ convergono ad α in modo sublineare.*

Dim. La prima parte si dimostra nello stesso modo del teorema 3. Per la seconda parte occorre dimostrare prima che le successioni generate a partire da $x_0 \in \mathcal{I}$ convergono ad α. Fatto questo la dimostrazione segue ripercorrendo la traccia data nel teorema 3. Diamo un cenno di come si dimostra la convergenza. Si osserva innanzitutto che se $x_k \in \mathcal{I}$ allora $|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)| = |g'(\xi_k)|\,|x_k - \alpha| < |x_k - \alpha|$ poiché $|g'(\xi_k)| < 1$. Quindi anche $x_{k+1} \in \mathcal{I}$. Questo permette di dimostrare che tutti i punti x_k appartengono ad $\mathcal{I}$ e che la successione $\{|x_k - \alpha|\}$ è decrescente. Se $g'(x) \geq 0$ sull'intervallo $[a,b]$ allora la successione $\{x_k\}$ è monotona e limitata quindi ha limite β tale che $\beta = g(\beta)$. Se il limite fosse $\beta \neq \alpha$ allora dalla relazione $\alpha - \beta = g(\alpha) - g(\beta) = g'(\xi)(\alpha - \beta)$ si ha un assurdo poiché $g'(\xi) < 1$. Se invece $g'(x) \leq 0$ allora si considera $G(x) = g(g(x))$ e si osserva che $x_{2k+2} = G(x_{2k})$, $x_{2k+1} = g(x_{2k})$. La funzione $G(x)$ è tale che $G(\alpha) = \alpha$ e $G'(x) = g'(g(x))g'(x) \geq 0$ Basta allora applicare il ragionamento precedente alla funzione $G(x)$ e concludere che $\lim_k x_{2k} = \alpha$ e conseguentemente $\lim_k x_{2k+1} = \lim_k g(x_{2k}) = g(\lim_k x_{2k}) = g(\alpha) = \alpha$ per ogni $x_0 \in \mathcal{I}$. □

Teorema 5 *Sia $g(x) \in C^p([a,b])$ con $p > 1$ intero e $\alpha \in (a,b)$ tale che $g(\alpha) = \alpha$. Se esiste un $x_0 \in [a,b]$ tale che la successione (1) converge superlinearmente ad α con ordine di convergenza p, allora $|g^{(k)}(\alpha)| = 0$ per $k = 1, \ldots, p-1$ e $g^{(p)}(\alpha) \neq 0$. Viceversa, se $|g^{(k)}(\alpha)| = 0$ per $k = 1, \ldots, p-1$ e $g^{(p)}(\alpha) \neq 0$ allora esiste un intorno $\mathcal{I}$ di α tale che per ogni $x_0 \in \mathcal{I}$ tutte le successioni $\{x_k\}$ generate da (1) convergono ad α in modo superlineare con ordine di convergenza p.*

Dim. Se la successione $\{x_k\}$ converge ad α con ordine p allora $\lim_k |x_{k+1} - \alpha|/|x_k - \alpha|^p = \sigma \neq 0$ finito. Per cui se $0 < q < p$ allora $\lim_k |x_{k+1} - \alpha|/|x_k - \alpha|^q = 0$. Usando questo fatto dimostriamo che $g^{(q)}(\alpha) = 0$ per $q = 1, \ldots, p-1$. Procediamo per induzione su q. La tesi è vera per $q = 1$ infatti abbiamo già dimostrato nei teoremi 3 e 4 che $g'(\alpha) = \lim_k |x_{k+1} - \alpha|/|x_k - \alpha|$, che è zero. In generale se $0 = g'(\alpha) = \cdots = g^{(q-1)}(\alpha)$, allora sviluppando $g(x)$ in serie di Taylor in un intorno di α si ottiene

$$g(x) = g(\alpha) + \frac{x-\alpha}{1!}g'(\alpha) + \cdots + \frac{(x-\alpha)^{q-1}}{(q-1)!}g^{(q-1)}(\alpha) + \frac{(x-\alpha)^q}{q!}g^{(q)}(\xi)$$

dove ξ è un punto del segmento aperto di estremi α e x. Dall'ipotesi induttiva segue

$$\frac{g(x)-g(\alpha)}{(x-\alpha)^q} = \frac{1}{q!}g^{(q)}(\xi).$$

Applicando questa espressione con $x = x_k$ dove ξ viene sostituito da ξ_k, e prendendo il limite in k, poiché ξ_k converge ad α si ottiene $g^{(q)}(\alpha) = 0$. Analogamente, con $q = p$, si deduce che

$$\frac{1}{p!}g^{(p)}(\alpha) = \lim_k \frac{x_{k+1}-\alpha}{(x_k-\alpha)^p} = \sigma \neq 0.$$

La dimostrazione della seconda implicazione è più semplice. Infatti dalla convergenza superlineare della successione $\{x_k\}$ si deduce che $g'(\alpha) = 0$ e per il teorema del punto fisso esiste un intorno $\mathcal{I} = [\alpha-\rho, \alpha+\rho]$ per cui tutte le successioni generate a partire da $x_0 \in \mathcal{I}$ convergono ad α. Il fatto che le derivate di $g(x)$ in α sono nulle fino all'ordine $p-1$, mentre $g^{(p)}(\alpha) \neq 0$, implica che nello sviluppo in serie di $g(x)$ risulta

$$g(x) = g(\alpha) + \frac{(x-\alpha)^p}{p!}g^{(p)}(\xi).$$

Ponendo $x = x_k$ si ottiene $(x_{k+1}-\alpha)/(x_k-\alpha)^p = g^{(p)}(\xi_k)/p!$ da cui, prendendo il limite in k si ottiene la tesi. □

Da questi risultati si ricava uno strumento utile per capire la velocità di convergenza dei metodi del punto fisso se applicati a funzioni sufficientemente regolari. In particolare, se siamo in grado di costruire un metodo iterativo associato ad una funzione $g(x)$ tale che $g'(\alpha) = 0$ allora disponiamo di un metodo che ha convergenza superlineare. Se poi $g(x)$ è derivabile due volte con continuità e $g''(\alpha) \neq 0$ il metodo costruito avrà convergenza quadratica. Come vedremo tra poco non sarà tanto difficile costruire un metodo che verifica la condizione $g'(\alpha) = 0$ anche se non si conosce α.

Apparentemente si potrebbe dedurre dai risultati precedenti che l'ordine di convergenza debba essere sempre un numero intero. Questo è falso come si può capire dall'esempio in cui $g(x) = x^{4/3}$. Chiaramente $\alpha = 0$ è un punto fisso di $g(x)$, inoltre $|x_{k+1}|/|x_k|^p = |x_k^{4/3}|/|x_k|^p$ per cui il limite $\lim_k |x_{k+1}|/|x_k|^p$ è finito e non nullo se e solo se $p = 4/3$. D'altro canto si vede che le ipotesi del teorema 5 non sono verificate essendo $g \in C^1$ ma $g \notin C^2$. Infatti $g'(x) = \frac{4}{3}x^{1/3}$, $g''(x) = \frac{4}{9}x^{-2/3}$.

In effetti la funzione $g(x)$ non è abbastanza regolare: per poter applicare il teorema avremmo dovuto avere o $g \in C^2$ e $g''(\alpha) \neq 0$, ma non è questo il caso, oppure, $g \in C^1$ e $g'(\alpha) \neq 0$.

Nella definizione data di convergenza lineare, sublineare e superlineare, così come nella definizione di ordine di convergenza, abbiamo assunto l'esistenza di alcuni limiti che in generale possono non esistere. Non è difficile costruire degli esempi in cui queste situazioni si presentano.

La seguente definizione può essere talvolta utile

Definizione 2 La successione $\{x_k\}$ converge ad α con *ordine almeno* p se esiste una costante β tale che

$$|x_{k+1} - \alpha| \leq \beta |x_k - \alpha|^p.$$

È facile dimostrare che una successione che converge con ordine $q \geq p$ converge anche con ordine almeno p.

È interessante osservare che se una successione x_k converge ad α in modo che l'errore relativo al passo k è limitato da

$$\epsilon_k = |x_k - \alpha|/|\alpha| \leq \beta \gamma^{p^k}$$

allora il numero di cifre significative, dato da $1 + \log_2 \epsilon_k^{-1}$ è tale che

$$1 + \log_2 \epsilon_k^{-1} \geq 1 + \log_2 \beta^{-1} + p^k \log_2 \gamma^{-1}$$

Cioè il numero di cifre significative è dato dalla somma di una parte costante, $1 + \log_2 \beta^{-1}$ e una parte che ad ogni passo aumenta di un fattore moltiplicativo p. In particolare, nel caso di $p = 2$, il numero di cifre corrispondente al secondo addendo *raddoppia* ad ogni passo.

Per un metodo del punto fisso definito da una funzione $g(x)$, diciamo che il metodo ha convergenza superlineare con ordine p se tutte le successioni generate a partire da x_0 in un opportuno intorno di α, $x_0 \neq \alpha$ convergono con ordine p. I teoremi precedenti garantiscono che se la funzione $g(x)$ è sufficientemente regolare allora si può definire l'ordine del metodo. Se invece vogliamo costruire un esempio di metodo iterativo in cui l'ordine di convergenza non è definibile, dobbiamo considerare funzioni che difettano di regolarità quale ad esempio la funzione

$$g(x) = \begin{cases} \frac{1}{2}x & \text{se } x \leq 0 \\ x^2 & \text{se } x > 0 \end{cases}$$

che ha punto fisso $\alpha = 0$. Infatti, partendo con $x_0 < 0$ la convergenza è lineare e monotona con fattore di convergenza $\gamma = 1/2$. Se invece $x_0 > 0$ la convergenza è monotona superlineare di ordine 2.

Alcuni spunti di riflessione:

- Cosa si può dire sulla convergenza delle successioni generate dalla funzione

$$g(x) = \begin{cases} -x & \text{se } x \leq 0 \\ -x^2 & \text{se } x > 0 \end{cases}$$

  per x_0 in un intorno di 0? Sono ancora applicabili le definizioni date?

- Si provi a considerare questa definizione alternativa: la successione $\{x_k\}$ ha convergenza lineare con fattore γ se $\lim_k |x_k - \alpha|^{1/k} = \gamma$. Si osservi che la quantità $(|x_k - \alpha|/|x_0 - \alpha|)^{1/k}$ dà la riduzione media dell'errore sui primi k passi del metodo, dove la media considerata è quella geometrica. Infatti vale

$$\left(\frac{|x_k-\alpha|}{|x_0-\alpha|}\right)^{1/k} = \left(\frac{|x_1-\alpha|}{|x_0-\alpha|}\cdot\frac{|x_2-\alpha|}{|x_1-\alpha|}\cdots\frac{|x_k-\alpha|}{|x_{k-1}-\alpha|}\right)^{1/k}.$$

Si verifichi che questa definizione coincide con quella data precedentemente nei casi in cui entrambe le definizioni sono applicabili. Si mostri un esempio in cui la seconda definizione è applicabile mentre la prima non lo è.

- Si definisca convergenza di ordine p se $\lim_k |x_k-\alpha|^{1/p^k} = \beta \neq 0$. Si verifichi che questa definizione coincide con quella data precedentemente nei casi in cui entrambe le definizioni sono applicabili. Si mostri un esempio in cui la seconda definizione è applicabile mentre la prima non lo è.

- Se $g(x)$ individua un metodo superlineare di ordine p, qual è l'ordine di convergenza del metodo dato dalla funzione $g_2(x) = g(g(x))$? E qual è l'ordine del metodo dato dalla funzione $g_q(x) = g(g(\cdots g(x)\cdots))$, dove la composizione viene fatta q volte? E se $g(x)$ ha ordine di convergenza 1 con fattore γ, qual è il fattore del metodo associato a $g_q(x)$?

3.1 Confronto tra metodi

Dati due metodi iterativi del punto fisso definiti da due funzioni $g_1(x)$ e $g_2(x)$ viene naturale chiedersi quale dei due sia più conveniente da usare. I due fattori principali che intervengono nella scelta sono la velocità di convergenza e il costo per passo. Confrontiamo i due metodi "alla pari". Supponiamo cioè che siano tutte e due a convergenza lineare o a convergenza superlineare. Denotiamo con c_1 e c_2 il numero delle operazioni aritmetiche per passo richieste dai due metodi.

Se i due metodi hanno convergenza lineare con fattore di convergenza γ_1 e γ_2 allora la riduzione dell'errore commesso dopo k passi sarà data rispettivamente da $\beta_1\gamma_1^k$ e $\beta_2\gamma_2^k$. I due metodi produrranno la stessa riduzione dell'errore rispettivamente con k_1 e k_2 passi se

$$\beta_1\gamma_1^{k_1} = \beta_2\gamma_2^{k_2}.$$

Prendendo i logaritmi si ha che

$$k_1\log\gamma_1 = k_2\log\gamma_2 + \log(\beta_2/\beta_1).$$

In una analisi asintotica nel numero di passi, che ha significato pratico nel caso si debba calcolare la soluzione con precisione elevata, si può trascurare il termine $\log(\beta_2/\beta_1)$, e imporre la condizione

$$k_1 = k_2\frac{\log\gamma_2}{\log\gamma_1}. \tag{5}$$

Poiché il costo globale dei due metodi applicati rispettivamente con k_1 e con k_2 iterazioni è dato rispettivamente da c_1k_1 e c_2k_2, il primo metodo risulta più

conveniente se $c_1 k_1 < c_2 k_2$. Quindi, per la (5) ciò accade se

$$\frac{c_1}{c_2} < \frac{\log \gamma_1}{\log \gamma_2}.$$

Si procede in modo analogo nell'analisi del caso di convergenza superlineare. Osserviamo prima che per la convergenza superlineare con ordine p l'errore al passo k è tale che $\epsilon_k \leq \beta \epsilon_{k-1}^p$.

Questo conduce alla diseguaglianza

$$\epsilon_k \leq \beta \beta^p \epsilon_{k-2}^{p^2} \leq \cdots \leq \beta \beta^p \beta^{p^2} \cdots \beta^{p^{k-1}} \epsilon_0^{p^k} = \eta r^{p^k}$$

dove $\eta = \beta^{-1/(p-1)}$, $r = \epsilon_0 \beta^{1/(p-1)}$, e dove si assume ϵ_0 sufficientemente piccolo in modo che $r < 1$.

Nel caso di due metodi con costanti η_1, r_1 e η_2, r_2, dove $r_1, r_2 < 1$, e di costo per passo c_1 e c_2, per capire quali dei due è asintoticamente più conveniente occorre confrontare le quantità $c_1 k_1$ e $c_2 k_2$ dove stavolta k_1 e k_2 sono legate dalla relazione

$$\eta_1 r_1^{p_1^{k_1}} = \eta_2 r_2^{p_2^{k_2}}.$$

Prendendo i logaritmi si ha

$$p_1^{k_1} \log r_1 = p_2^{k_2} \log r_2 + \log(\eta_2/\eta_1).$$

In una analisi asintotica nel numero di iterazioni possiamo trascurare la parte additiva che non dipende né da k_1 né da k_2 e quindi imporre la condizione

$$p_1^{k_1} \log r_1 = p_2^{k_2} \log r_2.$$

Cambiando segno e prendendo nuovamente i logaritmi si arriva a

$$k_1 \log p_1 = k_2 \log p_2 + \log(\log r_2^{-1} / \log r_1^{-1}),$$

che, sempre in una analisi asintotica può essere sostituita da

$$k_1 = k_2 \frac{\log p_2}{\log p_1}.$$

Per cui il primo metodo risulta asintoticamente più efficiente del secondo se $c_1 k_1 < c_2 k_2 = c_2 k_1 \log p_1 / \log p_2$, cioè se

$$\frac{c_1}{c_2} < \frac{\log p_1}{\log p_2}.$$

4 Alcuni metodi del punto fisso

Descriviamo ed analizziamo alcuni metodi del punto fisso per l'approssimazione numerica degli zeri di una funzione

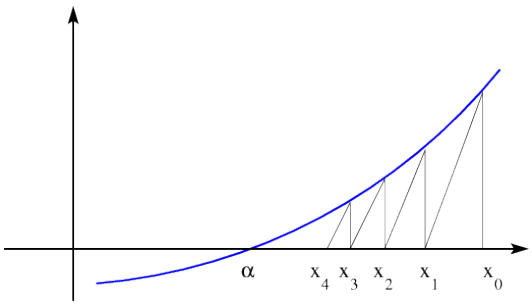


Figura 6: Metodo delle secanti

4.1 Il metodo delle secanti

Sia $f(x) \in C^1([a,b])$ e $\alpha \in [a,b]$ tale che $f(\alpha) = 0$. Il metodo definito dalla funzione

$$g(x) = x - f(x)/m$$

dove m è una opportuna costante è detto metodo delle secanti. Graficamente questo metodo consiste nel tracciare la retta passante per il punto $(x_k, f(x_k))$ di coefficiente angolare m e considerare come x_{k+1} l'ascissa del suo punto di intersezione con l'asse delle ascisse. La figura 6 mostra questa interpretazione geometrica.

Si osserva che $g'(x) = 1 - f'(x)/m$. Quindi una condizione sufficiente di convergenza è che $|1 - f'(x)/m| < 1$ in un intorno circolare di α. Questa condizione è verificata se $0 < f'(x)/m < 2$. Basta quindi scegliere m in modo che abbia lo stesso segno di $f'(x)$ e $|m| > \frac{1}{2}|f'(x)|$. In particolare, se $f'(\alpha)$ fosse nota, la scelta $m = f'(\alpha)$ darebbe una convergenza superlineare. Ma questa informazione molto raramente è disponibile.

Ad esempio, se $f(x) = x^2 - 2$, di modo che $\alpha = \sqrt{2}$ è zero di $f(x)$, la condizione da rispettare è $0 < x/m < 1$. Scegliendo ad esempio $m = 4$, la condizione viene verificata sull'intervallo $(0,4)$. Inoltre, poiché $f'(x) = 2x$, la derivata prima di $g(x) = x - f(x)/m$ vale $1 - 2x/m = 1 - x/2$ ed è positiva e minore di 1 sull'intervallo $(0,2)$. Allora per ogni x_0 compreso tra 0 e 2 la successione del punto fisso generata da $x_{k+1} = x_k - (x_k^2 - 2)/4$ converge in modo monotono ad α. La convergenza è lineare con fattore $\gamma = 1 - \sqrt{2}/2 = 0.29289...$ Scegliendo $m = 3$, la $g'(x)$ è positiva e minore di 1 in $(0, 3/2)$. Tale intervallo contiene $\sqrt{2}$ visto che $9/4 > 2$. Per cui, la successione generata da $x_{k+1} = x_k - (x_k^2 - 2)/3$ converge in modo monotono per ogni scelta di $x_0 \in (0, 3/2)$. Inoltre la convergenza è lineare con fattore di convergenza $1 - 2\sqrt{2}/3 = 0.05719...$

La tabella 2 mostra l'andamento delle successioni generate con $m = 4$, e con $m = 3$ a partire da $x_0 = 3/2$. La maggior velocità di convergenza della seconda successione a $\sqrt{2} = 1.41421356237310...$ è evidente.

k	$m = 4$	$m = 3$
1	1.43750000000000	1.41666666666667
2	1.42089843750000	1.41435185185185
3	1.41616034507751	1.41422146490626
4	1.41478281433500	1.41421401430572
5	1.41438021140058	1.41421358821949

Tabella 2: Approssimazioni di $\sqrt{2}$ ottenute col metodo delle secanti

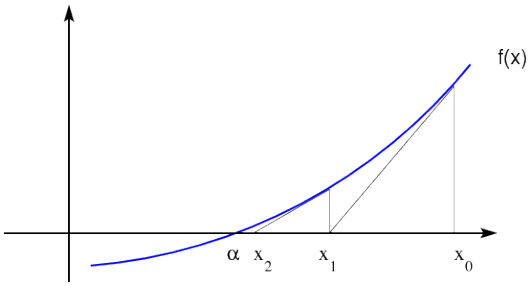


Figura 7: Metodo delle tangenti

4.2 Il metodo delle tangenti di Newton

Una naturale modifica del metodo delle secanti consiste nel variare ad ogni passo l'inclinazione m della retta passante per $(x_k, f(x_k))$ la cui intersezione con l'asse delle x ci fornisce x_{k+1}. Il modo più semplice di fare ciò si ottiene scegliendo come inclinazione quella della retta tangente al grafico di $f(x)$ nel punto $(x_k, f(x_k))$. Questo richiede che la funzione $f(x)$ sia derivabile. L'espressione che troviamo in questo modo, data da

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, \ldots, \tag{6}$$

definisce il *metodo di Newton* detto anche *metodo delle tangenti*.

La figura 7 mostra l'interpretazione geometrica del metodo di Newton.

Giusto per vedere la differenza della velocità di convergenza della successione generata, riportiamo nella tabella 3 i valori ottenuti col metodo di Newton applicato alla funzione $x^2 - 2$ a partire da $x_0 = 3/2$

Se la funzione $f(x)$ è sufficientemente regolare allora è possibile dimostrare facilmente le proprietà di convergenza del metodo di Newton. Ad esempio, supponiamo che $f(x)$ sia almeno di classe C^3 con $f'(\alpha) \neq 0$, di modo che la funzione

$$g(x) = x - \frac{f(x)}{f'(x)}$$

k	x_k
1	1.41666666666667
2	1.41421568627451
3	1.41421356237469
4	1.41421356237310

Tabella 3: Approssimazioni di $\sqrt{2}$ ottenute col metodo delle tangenti

sia di classe C^2. Allora, poiché

$$g'(x) = \frac{f(x)f''(x)}{f'(x)^2}, \quad g''(x) = \frac{f''(x)}{f'(x)} + \frac{f(x)f'''(x)}{f'(x)^2} - 2\frac{f(x)f''(x)^2}{f'(x)^3}$$

risulta $g'(\alpha) = 0$, $g''(\alpha) = f''(\alpha)/f'(\alpha)$. Per cui, per il teorema 5 il metodo di Newton ha convergenza superlineare che è di ordine 2 se $f''(\alpha) \neq 0$, mentre è di ordine almeno 2 se $f'''(\alpha) = 0$.

Possiamo però dare risultati di convergenza sotto ipotesi più deboli di regolarità.

Teorema 6 *Sia $f(x) \in C^2([a,b])$ e $\alpha \in (a,b)$ tale che $f(\alpha) = 0$. Se $f'(\alpha) \neq 0$ esiste un intorno $\mathcal{I} = [\alpha-\rho, \alpha+\rho] \subset [a,b]$ tale che per ogni $x_0 \in \mathcal{I}$ la successione (6) generata dal metodo di Newton converge ad α. Inoltre, se $f''(\alpha) \neq 0$ la convergenza è superlineare di ordine 2, se $f''(\alpha) = 0$ la convergenza è di ordine almeno 2.*

Dim. Poiché $g'(x) = f(x)f''(x)/(f'(x))^2$ risulta $g'(\alpha) = 0$. Per cui, essendo $g'(x)$ continua, esiste un intorno $\mathcal{I} = [\alpha - \rho, \alpha + \rho] \subset [a,b]$ per cui $|g'(x)| < 1$ se $x \in \mathcal{I}$. Per il teorema del punto fisso questo garantisce la convergenza delle successioni generate a partire da $x_0 \in \mathcal{I}$. Per dimostrare la convergenza quadratica si considera il rapporto

$$\frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{x_k - f(x_k)/f'(x_k) - \alpha}{(x_k - \alpha)^2}. \tag{7}$$

Sviluppando $f(x)$ in un intorno di x_k si ha che

$$0 = f(\alpha) = f(x_k) + (\alpha - x_k)f'(x_k) + \frac{(\alpha - x_k)^2}{2}f''(\xi_k), \qquad |\alpha - \xi_k| < |\alpha - x_k|,$$

da cui

$$f(x_k)/f'(x_k) = x_k - \alpha - \frac{(\alpha - x_k)^2 f''(\xi_k)}{2f'(x_k)}$$

Sostituendo nella (7) si ottiene

$$\frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{(\alpha - x_k)^2 f''(\xi_k)}{2(x_k - \alpha)^2 f'(x_k)}$$

da cui $\lim_k (x_{k+1} - \alpha)/(x_k - \alpha)^2 = f''(\alpha)/(2f'(\alpha))$, che dimostra la tesi. □

Teorema 7 *Sia $f(x) \in C^p([a,b])$ con $p > 2$ e $\alpha \in (a,b)$ tale che $f(\alpha) = 0$. Se $f'(\alpha) = \cdots = f^{(p-1)}(\alpha) = 0$, $f^{(p)}(\alpha) \neq 0$, e $f'(x) \neq 0$ per $x \neq \alpha$, allora esiste un intorno $\mathcal{I} = [\alpha - \rho, \alpha + \rho] \subset [a,b]$ in cui $f'(x) \neq 0$ per $x \in \mathcal{I}$, $x \neq \alpha$, e tale che per ogni $x_0 \in \mathcal{I}$ la successione (6) generata dal metodo di Newton converge ad α, inoltre α è l'unico zero di $f(x)$ in $\mathcal{I}$. La convergenza è lineare con fattore di convergenza $1 - 1/p$.*

Dim. Poiché $f(x) \in C^p([a,b])$ allora $f'(x) \in C^{p-1}([a,b])$ per cui si può sviluppare $f'(x)$ in serie, in un intorno di α ed essendo $f^{(i)}(\alpha) = 0$ per $i = 1, \ldots, p-1$ si ha

$$f'(x) = \frac{1}{(p-1)!}(x-\alpha)^{p-1} f^{(p)}(\xi)$$

con ξ appartenente all'intervallo aperto di estremi α e x. Questo implica che $f'(x) \neq 0$ in un intorno di α in cui $f^{(p)}(x) \neq 0$ che esiste essendo $f^{(p)}(\alpha) \neq 0$. Quindi il metodo di Newton è ben definito in questo intorno e la funzione $g(x)$ che lo definisce è data da

$$g(x) = \begin{cases} \alpha & \text{se } x = \alpha \\ x - \dfrac{f(x)}{f'(x)} & \text{se } x \neq \alpha. \end{cases}$$

Dimostriamo che $g(x)$ è di classe C^1 in un intorno di α e che $|g'(\alpha)| < 1$ in modo che la tesi discende dal teorema 1. Dimostriamo nell'ordine la continuità di $g(x)$ in α, la derivabilità in α e la continuità di $g'(x)$ in α. Essendo $f'(x) \neq 0$ per $x \neq \alpha$, la funzione $g(x)$ è continua per $x \neq \alpha$. Per la continuità di $g(x)$ su tutto l'intervallo basta dimostrare che $\lim_{x\to\alpha} g(x) = \alpha$, o equivalentemente che $\lim_{x\to\alpha} f(x)/f'(x) = 0$. Per la regola di de L'Hôpital vale $\lim_{x\to\alpha} f(x)/f'(x) = \lim_{x\to\alpha} f^{(p-1)}(x)/f^{(p)}(x) = f^{(p-1)}(\alpha)/f^{(p)}(\alpha) = 0$. Per la derivabilità di $g(x)$ in α basta dimostrare che esiste $\lim_{h\to 0}(g(\alpha+h) - g(\alpha))/h$. Per la definizione di $g(x)$ si ha $\lim_{h\to 0}(g(\alpha+h) - g(\alpha))/h = 1 - \ell$ dove $\ell = \lim_{h\to 0} f(\alpha+h)/(hf'(\alpha+h))$. Dagli sviluppi in serie di $f(\alpha+h)$ e $f'(\alpha+h)$

$$\begin{aligned} f(\alpha+h) =& f(\alpha) + hf'(\alpha) + \ldots + \frac{h^{p-1}}{(p-1)!} f^{(p-1)}(\alpha) + \frac{h^p}{p!} f^{(p)}(\xi), \\ f'(\alpha+h) =& f'(\alpha) + hf''(\alpha) + \ldots + \frac{h^{p-2}}{(p-2)!} f^{(p-1)}(\alpha) + \frac{h^{p-1}}{(p-1)!} f^{(p)}(\eta), \end{aligned} \tag{8}$$

dove ξ, η appartengono all'intervallo aperto di estremi α e $\alpha+h$, si ottiene $f(\alpha+h) = \frac{h^p}{p!} f^{(p)}(\xi)$, $f'(\alpha+h) = \frac{h^{p-1}}{(p-1)!} f^{(p)}(\eta)$. Per cui $\ell = \lim_{h\to 0} \frac{h^p/p!}{h^p/(p-1)!} \frac{f^{(p)}(\xi)}{f^{(p)}(\eta)}$. Poiché $\xi, \eta \to \alpha$ per $x \to \alpha$ e $f^{(p)}(x)$ è continua, allora $\ell = 1/p$ per cui $g(x)$ è derivabile in α e vale $g'(\alpha) = 1 - \ell = 1 - 1/p$. Per dimostrare la continuità di $g'(x)$ basta dimostrare che $\lim_{x\to\alpha} g'(x) = 1 - 1/p$. Per $x \neq \alpha$ vale $g'(x) = f(x)f''(x)/(f'(x))^2$ da cui, utilizzando gli sviluppi in serie (8) con $x = \alpha + h$ e quello di $f''(x)$

$$f''(x) = f''(\alpha) + (x-\alpha)f'''(\alpha) + \ldots + \frac{h^{p-3}}{(p-3)!} f^{(p-1)}(\alpha) + \frac{h^{p-2}}{(p-2)!} f^{(p)}(\mu),$$

con μ nell'intervallo aperto di estremi x e α, si ha $\lim_{x\to\alpha} g'(x) = 1 - 1/p$. □

Un altro utile risultato sulla convergenza delle successioni generate dal metodo di Newton riguarda la monotonia. Se la funzione $f(x)$ è crescente e convessa sull'intervallo $\mathcal{I} = [\alpha, \alpha+\rho]$, con $\rho > 0$, allora per ogni $x_0 \in \mathcal{I}$, la retta tangente al grafico della funzione giace tutta sotto il grafico per la convessità di $f(x)$. Per cui il punto x_1 deve necessariamente stare a destra di α. Non solo, ma il fatto che la $f(x)$ sia crescente implica che x_1 deve stare a sinistra di x_0. Queste considerazioni ci portano a concludere, mediante un argomento di induzione, che la successione generata dal metodo di Newton converge decrescendo ad α. Dimostriamo questa proprietà in modo analitico sotto ipotesi più generali.

Teorema 8 *Se la funzione $f(x)$ è di classe C^2 sull'intervallo $\mathcal{I} = [\alpha, \alpha+\rho]$ ed è tale che $f'(x)f''(x) > 0$ per $x \in \mathcal{I}$, allora per ogni $x_0 \in \mathcal{I}$, la successione generata dal metodo di Newton applicato ad $f(x)$ converge decrescendo ad α.*

Dim. Si supponga $f'(x) > 0$, il caso $f'(x) < 0$ si tratta in modo analogo. In questo caso è $f(x) > 0$ per $x > \alpha$, per cui $f(x)/f'(x) > 0$. Risulta quindi $x_1 < x_0$. Inoltre vale $x_1 - \alpha = g'(\xi)(x_0 - \alpha)$ con $\alpha < \xi < x_0$. Poiché $g'(\xi) = f(\xi)f''(\xi)/f'(\xi)^2$ si ha $g'(\xi) > 0$, quindi $x_1 > \alpha$. Ciò permette di dimostrare induttivamente che x_i converge decrescendo a α. □

Un risultato analogo vale su intervalli del tipo $[\alpha - \rho, \alpha]$.

5 Applicazioni del metodo di Newton

Una semplice ed efficace applicazione del metodo di Newton riguarda il calcolo del reciproco di un numero da svolgere con sole addizioni e moltiplicazioni.

Nella realizzazione di una aritmetica floating point si incontra il seguente problema: dato un numero di macchina a, vogliamo calcolare il numero di macchina che meglio approssima $1/a$. Possiamo assumere senza perdere di generalità che a sia la mantissa del numero, cioè $1/2 \le a < 1$. Si pone allora $f(x) = a - 1/x$ che si annulla in $\alpha = 1/a$ e si considera il metodo di Newton:

$$x_{k+1} = 2x_k - x_k^2 a \tag{9}$$

si verifica immediatamente che

$$(\alpha - x_{k+1})/\alpha = ((\alpha - x_k)/\alpha)^2. \tag{10}$$

Cioè l'errore relativo di approssimazione viene elevato a quadrato ad ogni passo. Se si sceglie $x_0 = 3/2$, l'errore relativo iniziale è $|3/2 - \alpha|/\alpha \le 1/2$ per cui dopo soli 6 passi, si ottiene un errore relativo limitato da $2^{-2^6} = 2^{-64}$. Tutte le 53 cifre della rappresentazione in doppia precisione sono corrette.

La tabella 4 mostra i valori che si ottengono per $a = 4/3$. Si può apprezzare la proprietà che ad ogni passo viene raddoppiato il numero di cifre significative dell'approssimazione.

k	x_k
1	1.31250000000000
2	1.33300781250000
3	1.33333325386047
4	1.33333333333333

Tabella 4: Calcolo di 4/3 col metodo di Newton senza svolgere divisioni

Un'altra utile applicazione riguarda il calcolo della radice p-esima di un numero di macchina a. Ancora possiamo supporre $a \in [1/2, 1]$. consideriamo per semplicità $p = 2$. Applicando il metodo di Newton alla funzione $x^2 - a$ per calcolare $\alpha = \sqrt{a}$, si ottiene la successione

$$x_{k+1} = \frac{x_k^2 + a}{2x_k}$$

generata dalla funzione $g(x) = (x^2+a)/2x$. Poiché $f''(\alpha) \neq 0$ il metodo converge con ordine 2.

Applicando il metodo di Newton alla funzione $x^{-2} - a^{-1}$, si ottiene la successione

$$x_{k+1} = \frac{1}{2}(3x_k - x_k^3 b), \quad b = a^{-1},$$

che, una volta calcolato $b = a^{-1}$ non comporta divisioni. Anche in questo caso l'ordine di convergenza del metodo è 2.

Per calcolare la radice p-esima possiamo applicare il metodo di Newton a una delle seguenti equazioni

$$f_q(x) = (x^p - a)x^{-q} = 0$$

con $q = 0, 1, \ldots, p$. Si scriva la funzione $g_q(x)$ corrispondente e si studi la convergenza.

5.1 Metodo di Newton nel campo complesso

Il metodo di Newton è formalmente definito anche per funzioni di variabile complessa purché derivabili, indipendentemente dal fatto che la interpretazione geometrica di metodo delle tangenti viene a mancare. Ad esempio, per un polinomio quale $f(x) = x^3 - 1$ l'espressione $g(x) = x - (x^3 - 1)/(3x^2)$, ottenuta dal metodo di Newton, ha ancora valore per valori complessi di x. La dimostrazione dei teoremi di convergenza che abbiamo dato non è più valida visto che si basa su risultati dell'analisi di funzioni di variabile reale. Però è ancora possibile dimostrare le buone proprietà di convergenza superlineare.

Sia allora $f(x) = (x - \alpha)q(x)$ un polinomio con uno zero α in generale complesso e semplice, cioè tale che il polinomio quoziente $q(x)$ non si annulla in α e inoltre $q'(\alpha) \neq 0$. Scriviamo la funzione $g(x)$ che definisce il metodo di Newton applicato a $f(x)$:

$$g(x) = x - \frac{(x-\alpha)q(x)}{q(x) + (x-\alpha)q'(x)}$$

Vale allora

$$g(x) - \alpha = (x-\alpha)^2 \frac{q'(x)}{q(x) + (x-\alpha)q'(x)} := (x-\alpha)^2 s(x)$$

Poiché $q(\alpha) \neq 0$, la funzione razionale $s(x)$ è definita in α e, per ragioni di continuità, esiste un intorno $\mathcal{U}$ di α nel piano complesso in cui $|s(x)| \leq \beta$ per β costante positiva. Per cui, per $x \in \mathcal{U}$ vale

$$|g(x) - \alpha| \leq \beta |x - \alpha|^2.$$

Questo implica che $|x_{k+1} - \alpha| \leq \beta |x_k - \alpha|^2$, dove $x_{k+1} = g(x_k)$ è la successione generata dal metodo di Newton a partire da un x_0 sufficientemente vicino ad α, e quindi la convergenza almeno quadratica a zero della successione $|x_k - \alpha|$. Se poi risulta $q'(\alpha) = 0$ allora la convergenza diventa di ordine superiore. La possibilità di definire i metodi del punto fisso anche per valori complessi di x ci porta a introdurre il concetto di *bacino di attrazione*.

5.2 Bacini di attrazione del metodo di Newton

Per capire la dinamica delle successioni generate dai metodi del punto fisso applicati ad esempio con una funzione $g(x)$ che ha dei punti fissi anche nel campo complesso, al variare di $x_0 \in \mathbb{C}$, è conveniente andare a tracciare i bacini di attrazione.

Supponiamo ad esempio di applicare il metodo di Newton al polinomio $x^3 - 1$ che ha tre radici nel campo complesso $\xi_1 = 1$, $\xi_2 = -1/2 + \mathtt{i}\sqrt{3}/2$, $\xi_3 = -1/2 - \mathtt{i}\sqrt{3}/2$, dove $\mathtt{i}$ è l'unità immaginaria tale che $\mathtt{i}^2 = -1$. Abbiamo l'iterazione

$$x_{k+1} = x_k - (x_k^3 - 1)/(3x_k^2).$$

Supponiamo di realizzare idealmente questo procedimento di colorazione dei punti del piano complesso:

- per un valore dell'intero n fissato, ad esempio $n = 400$, considero i numeri complessi $z_{k,j} = 2k/n + 2\mathtt{i}j/n$, con $k, j = -n, \ldots, n$. Questi punti ricoprono il quadrato del piano complesso centrato in 0 e di semilato 2.
- Per $k, j = -n, \ldots, n$, applichiamo il metodo di Newton con $x_0 = z_{k,j}$ (saltando $z_{0,0}$ che è nullo).
- Se la convergenza avviene verso ξ_1 coloriamo il punto di partenza di rosso; se la convergenza avviene verso ξ_2 coloriamo il punto di partenza di blu; se si ha convergenza a ξ_3 si colora di verde.
- Se non si ha convergenza si colora il punto di partenza di nero.

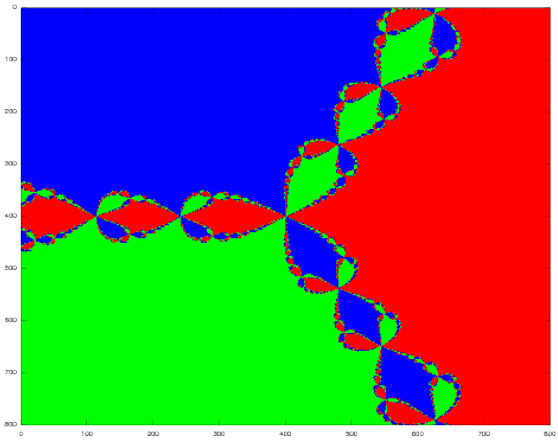


Figura 8: Bacini di attrazione del metodo di Newton applicato alla funzione $x^3 - 1$

In questo modo nel reticolo di punti scelti abbiamo evidenziato quelli che sono i bacini di attrazione verso le tre radici. Gli insiemi che si ottengono i questo modo sono molto informativi e anche suggestivi da un punto di vista estetico. L'immagine che si ottiene è riportata nella figura 8.

La colorazione la possiamo fare anche in modo più o meno intenso a seconda del numero di passi che sono stati sufficienti per entrare in un intorno fissato di ciascuna radice. È possibile usare colorazioni di fantasia, legate comunque al numero di passi impiegati per entrare in un assegnato intorno dello zero, che danno figure più suggestive. Qui sotto sono riportati alcuni di questi disegni.

5.3 Applicazioni in contesti non numerici

Il metodo di Newton può essere efficacemente usato anche in contesti non numerici. Ad esempio, consideriamo il seguente problema computazionale.

Dati i coefficienti del polinomio $a(t)$ di grado m tale che $a(0) \neq 0$, e dato un intero $n > 0$, calcolare i coefficienti del polinomio $x(t)$ di grado al più $n-1$ tale che $a(t)x(t) = 1 \mod t^n$.

Si consideri allora l'iterazione (9) riscritta sull'anello dei polinomi di grado al più $n-1$ con le operazioni modulo t^n:

$$\begin{aligned} x_{k+1}(t) &= 2x_k(t) - ax_k(t)^2 \mod t^n \\ x_0(t) &= 1/a(0) \end{aligned}$$

Si verifica facilmente che $x_0(t)a(t) - 1 = 0 \mod t$. Inoltre, come già visto in (10), vale

$$1 - x_{k+1}(t)a(t) = (1 - x_k(t)a(t))^2$$

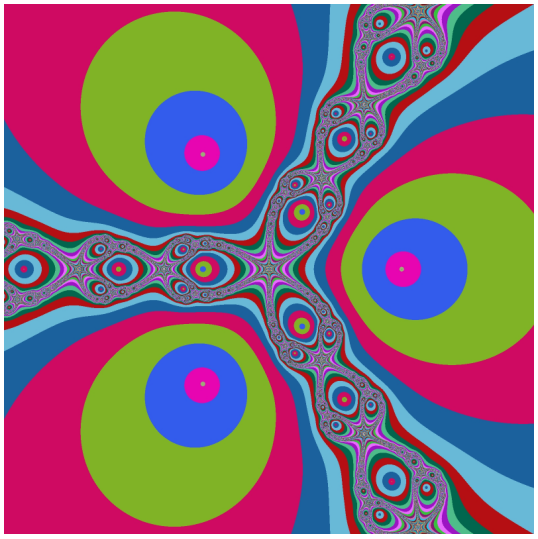


Figura 9: Bacini di attrazione per il metodo di Newton applicato alla funzione $x^3 - 1$, con colorazione di fantasia legata al numero di passi impiegati.

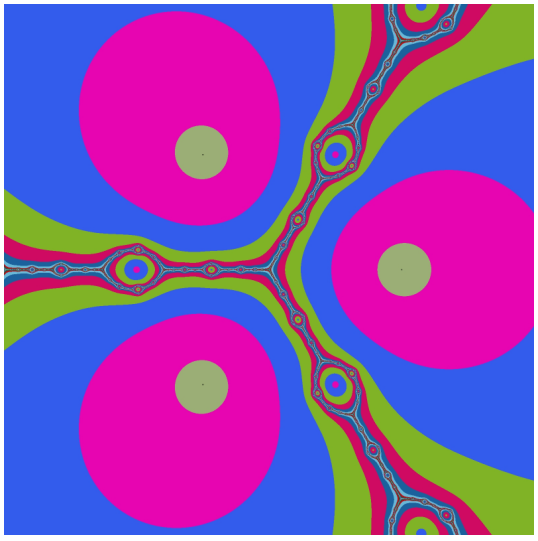


Figura 10: Bacini di attrazione per il metodo di Newton applicato alla funzione $x^2 - 1/x$, con colorazione di fantasia legata al numero di passi impiegati.

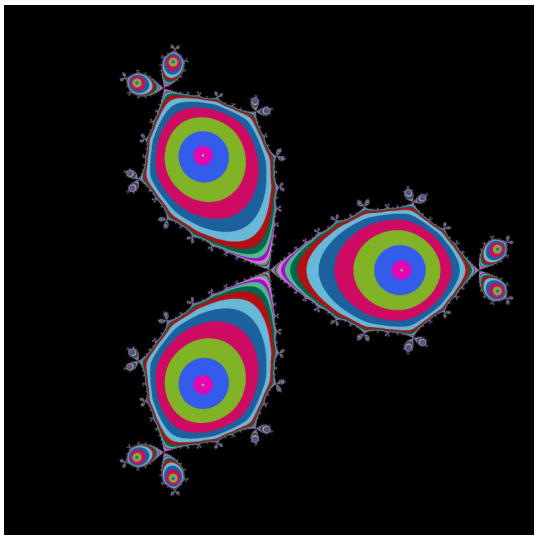


Figura 11: Bacini di attrazione per il metodo di Newton applicato alla funzione $1 - 1/x^3$, con colorazione di fantasia legata al numero di passi impiegati.

da cui si deduce induttivamente che

$$1 - x_{k+1}(t)a(t) = 0 \bmod t^{2^{k+1}}.$$

In altri termini, il polinomio $x_k(t)$ ha i primi 2^k coefficienti che coincidono con quelli della soluzione cercata. Cioè ad ogni passo il numero dei coefficienti corretti del polinomio $x_k(t)$ raddoppia. Per avere n coefficienti corretti bastano quindi $\lceil \log_2 n \rceil$ passi.

Un risultato analogo lo incontriamo nel calcolo di funzioni di polinomi modulo t^n che si possono scrivere come zeri di funzioni, quali ad esempio la radice p-esima modulo t^n:

$$x(t)^p - a(t) = 0 \bmod t^n.$$

Situazioni analoghe si incontrano quando si deve risolvere una equazione sull'anello Z_{p^n}. Ad esempio, vogliamo calcolare un intero x tale che $3x = 1 \bmod 5^{32}$. Sappiamo che $x_0 = 2$ soddisfa $3x_0 = 1 \bmod 5$. Applicando l'iterazione

$$x_{k+1} = 2x_k - 3x_k^2 \bmod 5^{2^{k+1}}$$

si ottengono dei valori tali che $3x_k = 1 \bmod 5^{2^k}$ per cui x_5 è la soluzione del nostro problema. I valori ottenuti degli x_k sono riportati nella tabella 5

6 Il caso dei polinomi

Approssimare gli zeri di polinomi in modo efficiente ha un interesse teorico in sé ma è anche importante per le applicazioni. Infatti, mediante i metodi

k	x_k
0	2
1	17
2	417
3	260417
4	101725260417
5	15522042910257975260417

Tabella 5

dell'algebra computazionale, si può sempre ricondurre il calcolo delle soluzioni di un sistema di equazioni algebriche al calcolo di tutte le radici di un opportuno polinomio. Una di queste tecniche usa lo strumento delle basi di Groebner.

Esiste una letteratura molto ampia riguardo ai metodi per l'approssimazione degli zeri reali e complessi di un polinomio. Numerosi metodi basati su tecniche del punto fisso e su altri approcci diversi sono stati introdotti ed analizzati. Riportiamo uno dei metodi della classe dei metodi di iterazione simultanea che è particolarmente efficiente ed è stato usato per la realizzazione del pacchetto MPSolve. MPSolve è un software efficiente che permette di calcolare un numero arbitrario di cifre di polinomi di grado arbitrario con coefficienti arbitrari.

Supponiamo di avere un polinomio $p(x)$ di grado n, monico, cioè in cui il coefficiente di x^n è 1, di cui conosciamo "buone" approssimazioni $x_1, \ldots, x_{n-1}$ di $n-1$ zeri. Per calcolare lo zero mancante applichiamo il metodo di Newton alla funzione razionale

$$f(x) = \frac{p(x)}{\prod_{j=1}^{n-1}(x - x_j)}.$$

Se i valori x_j, $j = 1, \ldots, n-1$ coincidessero esattamente con gli zeri $\xi_1, \ldots, \xi_{n-1}$ di $p(x)$, allora la funzione $f(x)$ sarebbe $x - \xi_n$ e quindi il metodo di Newton fornirebbe la soluzione in un solo passo qualunque sia il punto iniziale. In generale, è facile verificare che il metodo di Newton applicato a $f(x)$ è dato dalla funzione

$$g(x) = x - \frac{p(x)/p'(x)}{1 - \frac{p(x)}{p'(x)} \sum_{j=1}^{n-1} \frac{1}{x - x_j}}.$$

Il metodo di Ehrlich-Aberth usa questo fatto per approssimare simultaneamente tutti gli zeri di $p(x)$. Il metodo funziona in questo modo

- si scelgono n approssimazioni iniziali $x_j^{(0)}$, $j = 1, \ldots, n$ degli zeri di $p(x)$. Generalmente si pone

$$x_j^{(0)} = \cos(2(j + 1/2)\pi/n) + \mathtt{i} \sin(2(j + 1/2)\pi/n)$$

- per $k = 1, \ldots, k_{max}$ si calcola

$$x_i^{(k+1)} = x_i^{(k)} - \frac{p(x_i^{(k)})/p'(x_i^{(k)})}{1 - \frac{p(x_i^{(k)})}{p'(x_i^{(k)})} \sum_{j=1,\, j\neq i}^{n} \frac{1}{x_i^{(k)} - x_j^{(k)}}}, \qquad i = 1, \ldots, n.$$

- le iterazioni si arrestano se $|x_i^{(k+1)} - x_i^{(k)}| \leq \epsilon$ per $i = 1, \ldots, n$.

Si può dimostrare che se lo zero ξ_i è semplice allora la sequenza $x_i^{(k)}$ converge con ordine 3 a ξ_i dopo aver riordinato in modo opportuno le componenti $x_1^{(k)}, \ldots, x_n^{(k)}$. La convergenza a zeri multipli avviene linearmente. Non esistono risultati di convergenza globale ma non si conoscono esempi in cui la successione non converge per una scelta dei punti iniziali in un insieme di misura non nulla.

7 Esercizi

Esercizio 1 Siano $a, b \in \mathbb{R}$, $b > a$, e sia $g(x) \in C^1[a, b]$ tale che $g(\alpha) = \alpha$ con $\alpha \in [a, b]$. Inoltre vale $1/3 \leq g'(x) \leq 1/2$ per ogni $x \in [a, b]$.
a) Si dimostri che per ogni $x_0 \in [a, b]$ la successione $\{x_i\}_{i\in\mathbb{N}}$ definita da $x_{i+1} = g(x_i)$ converge ad α in modo monotono e si determini un intero k tale che $|x_k - \alpha| \leq (b - a)10^{-6}$, $\forall x_0 \in [a, b]$.
b) Per "accelerare" la convergenza della successione $\{x_i\}_{i\in\mathbb{N}}$ si consideri la successione $\{y_i\}_{i\in\mathbb{N}}$ così costruita $y_{i+1} = y_i + \theta(g(y_i) - y_i)$, dove $y_0 \in [a, b]$. Si determinino i valori di θ per cui la successione $\{y_i\}_{i\in\mathbb{N}}$ è convergente e si analizzi la velocità di convergenza.
c) È possibile determinare un valore ottimo di θ per cui ogni $\{y_i\}_{i\in\mathbb{N}}$ ha convergenza più rapida di ogni $\{x_i\}_{i\in\mathbb{N}}$ se $x_0 = y_0$?

Esercizio 2 Si considerino le successioni $\{x_i\}_{i\in\mathbb{N}}$ generate a partire da $x_0 \in \mathbb{R}$ dalla relazione $x_{i+1} = g(x_i)$ dove

$$g(x) = \begin{cases} \dfrac{x^2 + 2}{2x} & \text{se } x^2 - 2 \geq 0 \\ 1 + x - \frac{1}{2}x^2 & \text{se } x^2 - 2 < 0 \end{cases}$$

Si dica per quali valori di x_0 la successione generata è convergente e se ne determini il limite. Si determini inoltre l'ordine di convergenza e si dica per quali valori di x_0 la convergenza è monotona.

Soluzione. Cerchiamo intanto i punti fissi di $g(x)$. Siano $g_1(x) = \frac{x^2+2}{2x}$ e $g_2(x) = 1 + x - \frac{1}{2}x^2$. Si osserva che $g_1(x) = x$ se $\alpha_{1,2} = \pm\sqrt{2}$, e $g_2(x) = x$ se $\alpha_{1,2} = \pm\sqrt{2}$, dunque i punti fissi di $g(x)$ sono $\alpha_{1,2} = \pm\sqrt{2}$.

Studiamo le derivate di $g_1(x)$ e $g_2(x)$. Vale

$$g_1'(x) = \frac{1}{2} - \frac{1}{x^2}, \quad g_2'(x) = 1 - x.$$

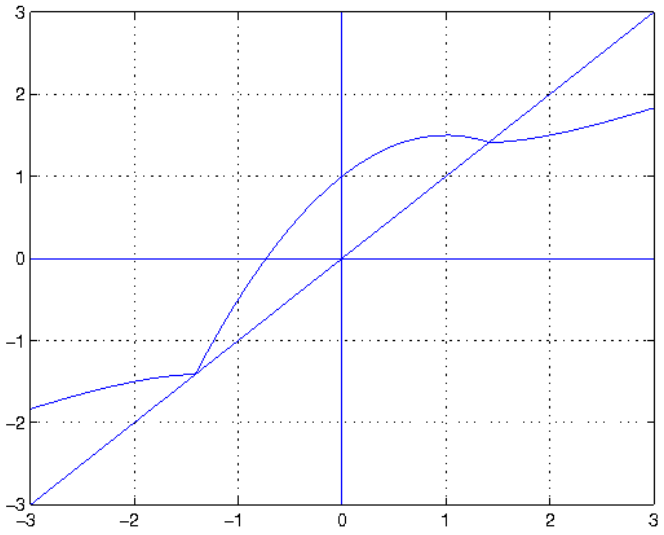


Figura 12: Funzione $g(x)$ dell'Esercizio 2

Dunque $g_1'(\alpha_{1,2}) = 0$, $g_2'(\alpha_1) = 1 - \sqrt{2}$, $g_2'(\alpha_2) = 1 + \sqrt{2}$. Si osserva inoltre che

$$0 \leq g_1'(x) < \frac{1}{2}, \text{ se } x^2 - 2 \geq 0$$

e

$$\begin{aligned} g_2(x) &> x, \quad \text{se } \alpha_2 < x < \alpha_1 \\ g_2'(x) &< 0, \quad \text{se } 1 < x \leq \sqrt{2}. \end{aligned}$$

Dunque la funzione $g(x)$ è continua in $\mathbb{R}$, ma derivabile solamente in $\mathbb{R} \setminus \{\pm\sqrt{2}\}$, e quindi non possiamo applicare i risultati di convergenza validi per funzioni di classe C^1 in un intorno del punto fisso. Il grafico della funzione è riportato in Figura 12.

Studiamo la convergenza per diversi punti iniziali x_0.

Supponiamo sia $x_0 > \alpha_1$. Allora

$$x_1 - \alpha_1 = g_1(x_0) - g_1(\alpha_1) = g_1'(\xi)(x_0 - \alpha_1)$$

per un opportuno ξ compreso tra x_0 e α_1, dunque tale che $\xi > \alpha_1$. In particolare $0 < g_1'(\xi) < \frac{1}{2}$, quindi $x_1 - \alpha_1 > 0$ e $x_1 - \alpha_1 \leq \frac{1}{2}(x_0 - \alpha_1)$. Induttivamente, possiamo dimostrare che, se $x_0 > \alpha_1$, allora

$$\alpha_1 < x_{i+1} < x_i, \quad x_i - \alpha_1 < \left(\frac{1}{2}\right)^i (x_0 - \alpha_1), \quad i \geq 0,$$

quindi $\{x_i\}$ converge in modo monotono decrescente a α_1. Riguardo la velocità di convergenza si ha,

$$\lim_i \frac{x_{i+1}-\alpha_1}{(x_i-\alpha_1)^2} = \lim_i \frac{\frac{x_i^2+2}{2x_i}-\sqrt{2}}{(x_i-\sqrt{2})^2} = \lim_i \frac{(x_i^2+2-2\sqrt{2}x_i)/(2x_i)}{(x_i-\sqrt{2})^2} = \lim_i \frac{1}{2x_i} = \frac{1}{2\sqrt{2}},$$

quindi la convergenza è quadratica.

Supponiamo ora $x_0 < \alpha_2$. Procedendo come nel caso $x_0 > \alpha_1$, si dimostra induttivamente che

$$x_i < x_{i+1} < \alpha_2, \quad \alpha_2 - x_i < \left(\frac{1}{2}\right)^i (\alpha_2 - x_0), \quad i \geq 0,$$

quindi $\{x_i\}$ converge in modo monotono crescente a α_2. Analogamente, si dimostra che la convergenza è quadratica.

Supponiamo ora $1 \leq x_0 < \alpha_1$. Allora

$$x_1 - \alpha_1 = g_2(x_0) - g_2(\alpha_1) = g_2'(\xi)(x_0 - \alpha_1)$$

per un opportuno ξ compreso tra x_0 e α_1, dunque tale che $1 < \xi < \alpha_1$. In particolare $g_2'(\xi) < 0$, quindi $x_1 - \alpha_1 > 0$. Dunque x_1 può essere visto come il valore iniziale della successione ottenuta a partire da un punto maggiore di α_1, e si applicano le considerazioni fatte nel caso $x_0 > \alpha_1$. Dunque la successione $\{x_i\}$ converge a α_1 in modo monotono decrescente per $i \geq 1$, e la convergenza è quadratica.

Supponiamo ora $\alpha_2 < x_0 < 1$. Poiché in questo intervallo aperto è $g(x) > x$, si ha che $x_{i+1} > x_i$ fintanto che $x_0 < x_i < 1$. La successione è quindi crescente finché rimane in $(x_0, 1)$. La successione non può essere contenuta tutta in $(x_0, 1)$ poiché in tal caso essendo crescente e limitata dall'alto dovrebbe avere limite nell'intervallo. Per la continuità di $g(x)$ questo limite dovrebbe essere un punto fisso di $g(x)$. Ma $g(x)$ non ha punti fissi in $[x_0, 1]$. Ci sarà quindi un valore $\bar{i}$ di i per cui $x_{\bar{i}} > 1$. Ci ritroviamo quindi in uno dei due casi precedenti per cui risulta $x_{\bar{i}+1} > \alpha_1$ quindi per $i \geq \bar{i}+1$ la successione converge in modo quadratico e decrescendo a α_1.

Si osservi che scegliendo x_0 razionale non può accadere che per un certo $\bar{i}$ sia $x_{\bar{i}} = \alpha_1$ essendo α_1 irrazionale.

Esercizio 3 a) Si determini nell'insieme dei polinomi di grado al più 3 con coefficienti razionali un polinomio $g(x)$ tale che la successione generata da $x_{k+1} = g(x_k)$ a partire da x_0 in un intorno di $\sqrt{3}$ converga a $\sqrt{3}$ con massimo ordine di convergenza.

b) Nell'insieme di funzioni del tipo $ax^{-1} + b + cx$ con a, b, c razionali si determini una funzione $h(x)$ tale che la successione generata da $y_{k+1} = h(y_k)$ a partire da y_0 in un intorno di $\sqrt{3}$ converga a $\sqrt{3}$ con massimo ordine di convergenza.

c) Se $x_0 = y_0$ è tale che le successioni $\{x_i\}_i$ e $\{y_i\}_i$ convergono a $\sqrt{3}$, si dica quale delle due successioni converge più velocemente valutando $\lim_n \left|\frac{x_i - \sqrt{3}}{y_i - \sqrt{3}}\right|$.

d) Si assuma che nel punto a) i polinomi abbiano grado al più 2 e coefficienti del tipo p/q con p, q interi di modulo minore di 10. Si individui il polinomio $g(x)$ per cui la successione generata abbia massima velocità di convergenza.

Soluzione.

a) Sia $g(x) = a + bx + cx^2 + dx^3$, con $a, b, c, d \in \mathbb{Q}$. La condizione $g(\sqrt{3}) = \sqrt{3}$ diventa $a + \sqrt{3}b + 3c + 3\sqrt{3}d = \sqrt{3}$. Dunque, uguagliando le parti razionali e non in entrambi i membri, si ottiene $a + 3c = 0$, $b + 3d = 1$.

Se $g'(\sqrt{3}) = 0$, il metodo converge localmente e la convergenza è almeno quadratica. Poiché $g'(x) = b + 2cx + 3dx^2$, uguagliando i termini razionali e non in entrambi i membri dell'equazione $g'(\sqrt{3}) = 0$, si ottiene $b + 9d = 0$ e $c = 0$.

Il sistema lineare definito dalle equazioni $a + 3c = 0$, $b + 3d = 1$, $b + 9d = 0$ e $c = 0$ ha un'unica soluzione, data da $a = 0$, $b = \frac{3}{2}$, $c = 0$, $d = -\frac{1}{6}$.

Dunque esiste un unico polinomio $g(x)$ di grado al più 3 con coefficienti razionali tale che la successione converga a $\sqrt{3}$ con massimo ordine di convergenza. La convergenza è quadratica perché $g''(\sqrt{3}) = -\sqrt{3} \neq 0$.

b) Sia $h(x) = ax^{-1} + b + cx$. La condizione $h(\sqrt{3}) = \sqrt{3}$ diventa $a\frac{\sqrt{3}}{3} + b + c\sqrt{3} = \sqrt{3}$. Dunque, uguagliando le parti razionali e non in entrambi i membri, si ottiene $b = 0$, $\frac{1}{3}a + c = 1$.

Se $h'(\sqrt{3}) = 0$, il metodo converge localmente e la convergenza è almeno quadratica. Poiché $h'(x) = -ax^{-2} + c$, si ottiene $h'(\sqrt{3}) = -\frac{1}{3}a + c = 0$.

Il sistema lineare definito dalle equazioni $\frac{1}{3}a + c = 1$ e $-\frac{1}{3}a + c = 0$ ha un'unica soluzione data da $a = \frac{3}{2}$, $c = \frac{1}{2}$.

Poiché $h''(\sqrt{3}) \neq 0$, con questi coefficienti il metodo ha convergenza quadratica, ed è l'unico a velocità di convergenza massima.

c) Usando lo sviluppo in serie si ha $x_i - \sqrt{3} = (x_{i-1} - \sqrt{3})^2 \frac{1}{2} g''(\xi_i)$, $y_i - \sqrt{3} = (y_{i-1} - \sqrt{3})^2 \frac{1}{2} h''(\eta_i)$, con ξ_i e η_i opportuni valori negli intervalli aperti di estremi rispettivamente $\sqrt{3}$, x_i, e $\sqrt{3}$, y_i. Quindi

$$\frac{y_i - \sqrt{3}}{x_i - \sqrt{3}} = \left(\frac{y_{i-1} - \sqrt{3}}{x_{i-1} - \sqrt{3}} \right)^2 \frac{h''(\xi_{i-1})}{g''(\eta_{i-1})}.$$

Posto $d_i = \frac{y_i - \sqrt{3}}{x_i - \sqrt{3}}$ e $\gamma_i = \frac{h''(\xi_i)}{g''(\eta_i)}$, si ha allora $d_{i+1} = d_i^2 \gamma_i$, $d_0 = 1$. Da cui

$$d_i = \gamma_{i-1} \gamma_{i-2}^2 \cdots \gamma_0^{2^{i-1}}.$$

Poiché $|g''(\sqrt{3})| = \sqrt{3}$ mentre $|h''(\sqrt{3})| = 1/\sqrt{3}$, è $|h''(\sqrt{3})/g''(\sqrt{3})| = 1/3$. Per cui per ξ, η in un intorno opportuno di $\sqrt{3}$ vale $|h''(\sqrt{3})/g''(\sqrt{3})| \leq \lambda < 1$. Quindi per $x_0 = y_0$ in tale intorno risulta $0 \leq \gamma_i < 1$ per cui

$$\left| \frac{y_i - \sqrt{3}}{x_i - \sqrt{3}} \right| < \lambda^{1+2+\cdots+2^{i-1}}$$

cioè $\lim_i \left| \frac{y_i - \sqrt{3}}{x_i - \sqrt{3}} \right| = 0$. La successione y_i converge allora più velocemente della successione x_i.

d) Sia $g(x) = a + bx + cx^2$. La condizione $g(\sqrt{3}) = \sqrt{3}$, con $a, b, c \in \mathbb{Q}$, è equivalente a $b = 1$ e $a + 3c = 0$. Poiché $g'(\sqrt{3}) = b + 2\sqrt{3}c$, affinché sia $g'(\sqrt{3}) = 0$, dovrei avere $b = 0$ e $c = 0$, dunque non esistono coefficienti $a, b, c \in \mathbb{Q}$ tali che la successione ha convergenza locale superlineare.

La condizione di convergenza locale è $|g'(\sqrt{3})| < 1$, cioè $|1 + 2\sqrt{3}c| < 1$, e la velocità di convergenza è massima quando $|g'(\sqrt{3})|$ è minimo. Occorre quindi minimizzare $g'(\sqrt{3}) = 1 + 2\sqrt{3}c$ dove $c = p/q$ e $|p|, |q| < 10$. Per $c = -1/(2\sqrt{3})$ vale $g'(\sqrt{3}) = 0$. Poiché $g'(\sqrt{3})$ è funzione lineare di c, il valore di p/q che minimizza $|g'(\sqrt{3})|$ è quello che meglio approssima $-1/(2\sqrt{3}) = -0.28868...$, cioè $-2/7 = 0.28571...$, e vale $g'(\sqrt{3}) = 0.00296...$.

Esercizio 4 Sia $g_1(x) = \frac{1}{2}(x + \frac{a}{x})$, $g_2(x) = \frac{1}{2}x(3 - \frac{x^2}{a})$, dove $a > 0$.

a) Si verifichi che $\sqrt{a}$ è punto fisso di g_1 e g_2.

b) Si studi l'ordine di convergenza dei metodi di iterazione funzionale individuati da g_1 e da g_2.

c) Si determinino due costanti α e β tali che $G(x) = \alpha g_1(x) + \beta g_2(x)$ abbia punto fisso α e ordine di convergenza più alto possibile. Dire qual è l'ordine.

Esercizio 5 Per $a \in \mathbb{R}$ sia $f_a(x) = x \log|x + a|$, dove per $a = 0$ si definisce $f_a(0) = 0$. Si determinino gli zeri di $f_a(x)$ e si dica se il metodo di Newton è applicabile per la loro approssimazione. Studiare la convergenza locale del metodo di Newton, se applicabile, determinandone l'ordine.

Soluzione.

Vale $f_a(x) = 0$ quando $x = 0$ oppure $|x + a| = 1$, dunque gli zeri di $f_a(x)$ sono $\alpha_1 = 0$, $\alpha_2 = 1 - a$, $\alpha_3 = -1 - a$.

Sia $a \neq 0$. La funzione non è definita in $x = -a$, e per $x \neq -a$ vale

$$f_a'(x) = \log(|x + a|) + \frac{x}{x + a}, \quad f_a''(x) = \frac{x + 2a}{(x + a)^2}, \quad f_a'''(x) = -\frac{x + 3a}{x + a}.$$

In particolare la funzione è di classe C^∞ su $(-\infty, -a) \cup (-a, +\infty)$.

La funzione è di classe C^∞ in un intorno di ciascun zero α_i, $i = 1, 2, 3$, e vale $f_a'(\alpha_1) = \log(|a|)$, $f_a'(\alpha_2) = 1 - a$, $f_a'(\alpha_3) = 1 + a$. Dunque, se $a \neq \pm 1$, vale $f_a'(\alpha_i) \neq 0$, dunque il metodo di Newton è applicabile e converge localmente a α_1, α_2 e α_3, rispettivamente. Poiché $f_a''(\alpha_i) \neq 0$, $i = 1, 2, 3$, la convergenza è quadratica. La Figura 13 riporta il grafico di $f_a(x)$ nel caso $a = 2$.

Se $a = 1$, allora $\alpha_1 = \alpha_2 = 0$ e $\alpha_3 = -2$. Vale $f_a'(\alpha_1) = 0$, $f_a'(x) \neq 0$ in un intorno di α_1, e $f_a''(\alpha_1) = 2$, dunque il metodo di Newton è applicabile, converge localmente a α_1, e la convergenza è lineare con fattore di convergenza $1/2$. Poiché $f_a'(\alpha_3) = 2$ e $f_a''(\alpha_3) = 0$, il metodo converge localmente a α_3, e la convergenza è almeno quadratica; poiché $f_a'''(\alpha_3) = 1$, la convergenza è di ordine 3.

Se $a = -1$, allora $\alpha_1 = \alpha_3 = 0$ e $\alpha_2 = 2$. In modo analogo al caso $a = 1$, il metodo di Newton converge localmente a α_1, e la convergenza è lineare con fattore di convergenza $1/2$; inoltre il metodo converge localmente a α_2, e la convergenza è di ordine 3.

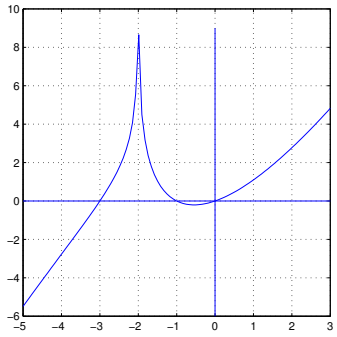


Figura 13: Esercizio 5, $a = 2$

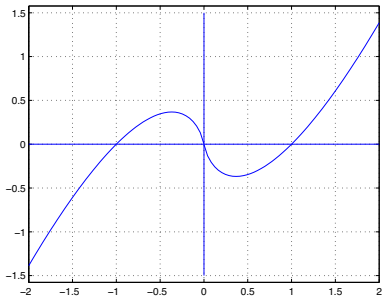


Figura 14: Esercizio 5, $a = 0$

Sia $a = 0$. Se definiamo $f(0) = 0$, allora la funzione $f(x) = x \log |x|$ è continua su $\mathbb{R}$. I suoi zeri sono $\alpha_1 = 0$, $\alpha_2 = 1$, $\alpha_3 = -1$. Se $x \neq 0$ vale $f'(x) = \log |x| + 1$, $f''(x) = 1/x$, e in particolare $f(x)$ è C^∞ in $\mathbb{R} \setminus \{0\}$. La Figura 14 riporta il grafico di $f_a(x)$ nel caso $a = 0$. La funzione è C^∞ in un intorno di α_2 e α_3, dunque il metodo di Newton è applicabile in un intorno di questi due punti. Vale $f'(\alpha_2) \neq 0$, $f''(\alpha_1) \neq 0$ e $f'(\alpha_3) \neq 0$, $f''(\alpha_3) \neq 0$, dunque converge localmente con convergenza quadratica. Consideriamo ora lo zero $\alpha_1 = 0$. Il metodo di Newton è applicabile in ogni punto diverso da 0, e la sua espressione è:

$$x_{i+1} = x_i - \frac{x_i \log |x_i|}{\log |x_i| + 1} = \frac{x_i}{\log |x_i| + 1}, \quad i \geq 0.$$

Se definiamo

$$g(x) = \begin{cases} \frac{x}{\log |x|+1} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

allora la funzione $g(x)$ è definita e continua in un intorno di 0, e possiamo definire il metodo di Newton in un intorno di 0 mediante l'espressione $x_{i+1} = g(x_i)$. La funzione $g(x)$ è anche derivabile in 0, infatti

$$\lim_{h \to 0} \frac{g(h) - g(0)}{h} = \lim_{h \to 0} \frac{1}{\log |h| + 1} = 0,$$

quindi $g'(0) = 0$. Questo implica che il metodo di Newton converge localmente. Vogliamo valutare l'ordine di convergenza. La funzione $g(x)$ non è di classe C^2 in un intorno di 0, dunque non possiamo applicare i risultati teorici di convergenza. Osserviamo che

$$\lim_{i \to \infty} \frac{|x_{i+1}|}{|x_i|} = \lim_{i \to \infty} \frac{1}{\log |x_i| + 1} = 0$$

quindi la convergenza è superlineare. Osserviamo che $\lim_{i\to\infty} \frac{|x_{i+1}|}{|x_i|^2} = \infty$ e non esiste un $p \in \mathbb{R}$ tale che

$$\lim_{i\to\infty} \frac{|x_{i+1}|}{|x_i|^p} = \gamma$$

con $0 < \gamma < \infty$, dunque per questa successione l'ordine di convergenza non è definibile. □

Esercizio 6 Sia $a > 0$ e $f(x) = x + a\sqrt[3]{x^2}$. Per $x \neq 0$ sia $g(x) = x - f(x)/f'(x)$ e, fissato $x_0 \neq 0$ si consideri la successione $\{x_k\}$ definita da $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$. Si dimostri che
a) $f(x) = 0$ se e solo se $x \in \{0, -a^3\}$; inoltre $\{x_k\}$ converge quadraticamente a $-a^3$ per x_0 in un intorno di $-a^3$;
b) $\{x_k\}$ converge a $-a^3$ per ogni $x_0 < -8/27a^3$ e la convergenza è monotona per $x_0 < -a^3$;
c) la convergenza è lineare in un intorno di 0; determinare il fattore di convergenza.

Esercizio 7 Si verifichi che i punti fissi della funzione

$$g(x) = \begin{cases} 1 + (x-1)^2 & \text{se } x \geq 1 \\ \frac{24}{25} + (x - \frac{4}{5})^2 & \text{se } x < 1 \end{cases}$$

sono $\alpha = 1$ e $\beta = 2$ e si discuta la convergenza della successione generata da $x_{k+1} = g(x_k)$ al variare di x_0. Più precisamente
a) Si dimostri che esiste un intervallo destro U di α tale che per $x_0 \in U$ la successione converge in modo monotono e quadratico; si determini il massimo intorno U per cui vale questa proprietà
b) Si dimostri che esiste un intervallo sinistro V di α tale che per $x_0 \in V$ la successione converge in modo monotono e lineare; si determini il massimo intorno V per cui vale questa proprietà
c) Determinare gli insiemi $\mathcal{A}$ e $\mathcal{B}$ per cui la successione converge rispettivamente ad α per ogni $x_0 \in \mathcal{A}$ e a β per ogni $x_0 \in \mathcal{B}$. Determinare l'insieme dei valori x_0 per cui non si ha convergenza.

Esercizio 8 Sia $g(x) = a - 1/x$ con $a \geq 0$. Determinare al variare di a i punti fissi di $g(x)$ e discutere la convergenza della successione definita da $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, al variare di $x_0 \in \mathbb{R}\backslash 0$. In particolare individuare i casi di convergenza monotona, valutare l'ordine di convergenza e descrivere l'insieme dei valori di x_0 per cui la successione non è definita.

Esercizio 9 Sia $q(x)$ un polinomio di grado al più 2 a coefficienti razionali.
a) Sia $g(x) = x - q(x)(x^2 - 2)$. Si trovino condizioni sui coefficienti di $q(x)$ per i quali per ogni x_0 in un opportuno intorno di $\sqrt{2}$ la successione definita da $x_{k+1} = g(x_k)$ sia convergente a $\sqrt{2}$ col massimo ordine di convergenza possibile. Si scriva l'espressione di $g(x)$ e si determini l'ordine di convergenza.
b) Nella classe dei metodi così ottenuti se ne individui uno che richieda non più

di 4 operazioni aritmetiche per passo (si assume che costanti razionali quali ad esempio 2/3, siano assegnate a costo zero).
c) Si determini un numero $\rho > 0$ tale che la successione generata dal metodo selezionato al punto b) sia convergente per ogni $x_0 \in [\sqrt{2}-\rho, \sqrt{2}+\rho]$, motivando la risposta.
d) Si risponda alla domanda a) nel caso di $g(x) = x - (x^2-2)/q(x)$.

Esercizio 10 Siano $g_1(x)$ e $g_2(x)$ due funzioni da $\mathbb{R}$ in $\mathbb{R}$ derivabili con continuità. Si consideri la successione generata a partire da $x_0, x_1 \in \mathbb{R}$ dalla relazione $x_{k+1} = g_1(x_k) + g_2(x_{k-1})$, $k = 0, 1, \ldots$.
a) Si dimostri che se esiste $\lim_k x_k = \alpha$ allora $\alpha = g_1(\alpha) + g_2(\alpha)$ e che per ogni $x_0, x_1 \in \mathbb{R}$ esistono $\xi_k, \eta_k \in \mathbb{R}$ tali che $e_{k+1} = g_1'(\xi_k)e_k + g_2'(\eta_k)e_{k-1}$, per $k = 1, 2, \ldots$, dove $e_k = x_k - \alpha$.
b) Riscrivendo l'espressione $e_{k+1} = g_1'(\xi_k)e_k + g_2'(\eta_k)e_{k-1}$ nella forma

$$v^{(k+1)} = \begin{bmatrix} g_1'(\xi_k) & 2g_2'(\eta_k) \\ \frac{1}{2} & 0 \end{bmatrix} v^{(k)}, \quad \text{con} \quad v^{(k)} = \begin{bmatrix} e_k \\ \frac{1}{2}e_{k-1} \end{bmatrix}$$

si dimostri che se $|g_1'(\xi)| + 2|g_2'(\eta)| \le \lambda < 1$ $\forall \xi, \eta \in \Omega = \{x : |x-\alpha| \le \rho\}$ allora $\forall\ x_0, x_1 \in \Omega$ vale $x_k \in \Omega$, $||v_k||_\infty \le \rho \max(1/2, \lambda)^{k-1}$, per $k \ge 1$, e $\lim_k x_k = \alpha$.
c) Sotto le condizioni del punto b) si provi che α è l'unico punto fisso di $g_1(x) + g_2(x)$ in Ω.
d) Dire se la condizione $|g_1'(\xi)| + |g_2'(\eta)| \le \lambda < 1$, $\forall \xi, \eta \in \Omega$, è sufficiente per la convergenza.

Esercizio 11 Sia $g(x) : [a, b] \to [a, b]$ continua tale che $g(\alpha) = \alpha$, dove $a < \alpha < b$. Sia $\{x_k\}_{k \ge 0}$ la successione definita da $x_{k+1} = g(x_k)$ a partire da $x_0 \in [a, b]$.
a) Si dimostri che se valgono le implicazioni $x \in [a, b]$, $x > \alpha \Rightarrow 0 < g(g(x)) - \alpha < x - \alpha$, e $x \in [a, b]$, $x < \alpha \Rightarrow 0 < \alpha - g(g(x)) < \alpha - x$, allora per ogni $x_0 \in [a, b]$ le sottosuccessioni x_{2k} e x_{2k+1} convergono in modo monotono ad α.
b) Si determinino i punti fissi di $g(x) = \gamma x^2/(3x-2)$ per $\gamma > 3/4$ e si dica per quali valori di γ la successione $\{x_k\}$ è localmente convergente ai punti fissi. Si determini l'ordine di convergenza. Si studi in particolare il caso $\gamma = 1$.
c) Per $\gamma = 1$ si determini un insieme di punti x_0 per cui la successione x_k converga a 0 e si dica se esiste un intorno di 1 per cui la successione converge a 1 per ogni scelta di x_0 in questo intorno.

Soluzione.
a) Fissiamo x_0 e sia $x_1 = g(x_0)$. Osserviamo che, se $h(x) = g(g(x))$, allora $x_{2(k+1)} = h(x_{2k})$ e $x_{2k+1} = h(x_{2k-1})$, per $k \ge 0$. Sia $x_0 > \alpha$. Allora vale $0 < h(x_0) - \alpha = x_2 - \alpha =< x_0 - \alpha$ e, induttivamente su k,

$$0 < x_{2(k+1)} - \alpha < x_{2k} - \alpha, \quad k \ge 0.$$

Dunque la successione $\{x_{2k}\}_k$ è monotona decrescente e limitata inferiormente da α, quindi convergente a un limite $\beta \ge \alpha$. In modo analogo, se $x_0 < \alpha$,

allora la successione $\{x_{2k}\}_k$ è monotona crescente e limitata superiormente da α, quindi convergente a un limite $\beta \leq \alpha$. Nello stesso modo si ragiona sulla successione $\{x_{2k-1}\}_k$

Dimostriamo che $\beta = \alpha$. Supponiamo di essere nel caso $x_0 > \alpha$, dunque $\beta \geq \alpha$. Essendo $h(x)$ continua, deve essere $h(\beta) = \beta$. Se fosse $\beta > \alpha$ avrei $0 < h(\beta) - \alpha < \beta - \alpha$, assurdo perché $h(\beta) = \beta$, dunque $\beta = \alpha$. Negli altri casi si ragiona in modo analogo.

b) Si osserva che la funzione $g(x)$ è definita per $x \neq 2/3$, inoltre $g(x) \neq 2/3$ per ogni $x \in \mathbb{R}$ poiché $\gamma > 3/4$. I punti fissi di $g(x)$ sono le soluzioni dell'equazione $(\gamma - 3)x^2 + 2x = 0$, dunque se $\gamma = 3$ l'unico punto fisso è $\alpha_1 = 0$, se $\gamma \neq 3$ i punti fissi sono $\alpha_1 = 0$ e $\alpha_2 = 2/(3 - \gamma)$. La funzione $g(x)$ è C^∞ in un intorno dei punti fissi, e vale

$$g'(x) = \gamma x \frac{3x - 4}{(3x - 2)^2}.$$

Vale $g'(0) = 0$ e $g''(0) \neq 0$ per ogni $\gamma \neq 0$, dunque il metodo converge localmente con convergenza quadratica in un intorno di α_1. Supponiamo ora $\gamma \neq 3$ e studiamo la convergenza in un intorno di α_2. Si può verificare che $g'(\alpha_2) = (2\gamma - 3)/\gamma$, quindi se $|2\gamma - 3|/|\gamma| < 1$ ho convergenza locale. La condizione $|2\gamma - 3|/|\gamma| < 1$ è equivalente a $1 < \gamma < 3$; in particolare per $\gamma = 1$ ho $|2\gamma - 3|/|\gamma| = 1$ e $g'(\alpha_2) = 1$, quindi se c'è convergenza locale, questa è sublineare. Se $\gamma = 3/2$ la convergenza è superlineare.

c) Il grafico di $g(x) = \frac{x^2}{3x-2}$ è riportato nella Figura 15. Si può verificare che se $x < 0$ allora $0 < g'(x) < 1$, dunque se $x_0 < 1$ la successione $\{x_i\}_i$ converge in modo monotono a 0. Se $0 < x < 2/3$ allora $g'(x) < 0$, dunque se $0 < x_0 < 2/3$ allora $x_1 < 0$ e si applicano le considerazioni fatte nel caso $x_0 < 0$.

Per la convergenza locale a 1 si applica il risultato del punto a). Infatti basta verificare che la funzione $h(x) = g(g(x))$ è tale che $0 < h'(x) < 1$ per $x \neq 1$ in un opportuno intorno di 1.

□

Esercizio 12 Si dimostri che il polinomio $p(x) = x^3 - 3x + 1$ ha tre zeri reali. Si discuta la convergenza agli zeri di $p(x)$delle successioni del tipo $x_{i+1} = g(x_i)$, $i = 0, 1, 2, \ldots$, con $x_0 \in R$ nel caso in cui

1. $g(x) = (x^3 + 1)/3$
2. $g(x) = (3x - 1)^{1/3}$
3. $g(x) = x - (x^3 - 3x + 1)\frac{x}{6x-3}$

Esercizio 13 Sia $g(x) = a \log x$ con $a \neq 0$. Determinare al variare di a il numero di punti fissi di $g(x)$ e discutere la convergenza della successione definita da $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, al variare di $x_0 \in \mathbb{R}^+$. In particolare, individuare i casi di convergenza monotona, determinare l'ordine di convergenza e l'insieme dei valori di x_0 per cui la successione non è definita.

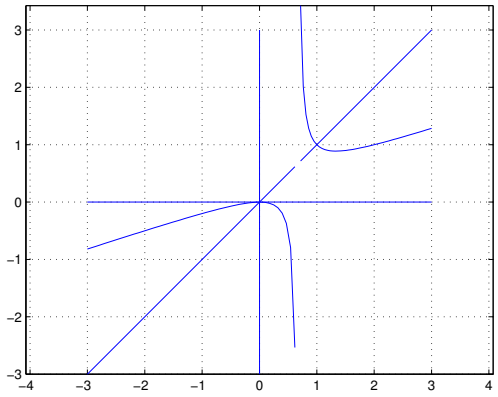


Figura 15: Esercizio 11, funzione $g(x)$, caso $\gamma = 1$

Esercizio 14 Siano $a, b, \alpha \in \mathbb{R}$, $a < \alpha < b$ e $f \in C^4[a, b]$ tale che $f(\alpha) = 0$, $f'(\alpha) \neq 0$. Si consideri il metodo iterativo definito dalla funzione $g(x) = x - t(x) - \theta t(x)^2$ con $t(x) = f(x)/f'(x)$ e θ parametro reale.
a) Si dimostri che per ogni $\theta \in \mathbb{R}$ il metodo iterativo è localmente convergente ad α con convergenza almeno quadratica.
b) Si determini il valore di θ che massimizza l'ordine di convergenza.

Esercizio 15 È data la funzione $g(x) : [a, b] \to \mathbb{R}$ e $\alpha \in [a, b]$ tale che $g(\alpha) = \alpha$, dove $a < b$ e $\alpha \neq 0$. Sia $G(x) = g(x)(\beta + \gamma g(x)/x)$ con $\beta, \gamma \in \mathbb{R}$. Si confrontino le proprietà di convergenza locale delle successioni $\{x_k\}$ e $\{y_k\}$ tali che $x_{k+1} = g(x_k)$, $y_{k+1} = G(y_k)$ con x_0, y_0 in un intorno di α. In particolare:
a) dare condizioni su $g(x)$ affinché esistano β, γ tali che $\{y_k\}$ sia localmente convergente ad α;
b) determinare i valori di β e γ per i quali $\{y_k\}$ abbia convergenza locale più elevata possibile e dire in quali casi $\{y_k\}$ converge più velocemente di $\{x_k\}$.
c) Sia $g(x) = x - f(x)/f'(x)$ dove $f(x) \in C^p$, $p \geq 2$, $0 = f(\alpha) = f'(\alpha) = \cdots = f^{p-1}(\alpha)$, $f^p(\alpha) \neq 0$. Determinare la funzione $G(x)$ che massimizza l'ordine di convergenza.

Esercizio 16 Sia $\alpha \in \mathbb{R}$ e $g_1(x) \in C^1((-\infty, \alpha])$, $g_2(x) \in C^1([\alpha, +\infty))$, tali che $g_1(\alpha) = g_2(\alpha) = \alpha$ e $g_1'(x) \leq 0$, $g_2'(x) \leq 0$, $g_1'(\alpha) \neq g_2'(\alpha)$. Si definisca $g(x) : \mathbb{R} \to \mathbb{R}$ tale che $g(x) = g_1(x)$ se $x \leq \alpha$, $g(x) = g_2(x)$ se $x \geq \alpha$, e per $x_0 \in \mathbb{R}$ si definisca $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$
a) Si discuta la convergenza e gli ordini di convergenza delle sucessioni $\{x_{2k}\}$ e $\{x_{2k+1}\}$, assumendo se necessario maggior regolarità di g_1 e di g_2, in funzione delle derivate prime in α di g_1 e di g_2.
b) Si discuta la convergenza e gli ordini di convergenza della sucessione $\{x_k\}$.

Esercizio 17 Sia $p_0(x) = x - 1/2$ e si definisca la successione di polinomi $p_k(x)$ di grado $n_k = 3^k$ mediante $p_{k+1}(x) = p_k(x)^3 + x$, $k = 0, 1, \ldots$
a) Si dimostri che nell'intervallo $[0, 1]$ ogni polinomio $p_k(x)$ ha una sola radice reale ξ_k e che vale $\xi_{k+1} < \xi_k$.
b) Si dimostri che il metodo di Newton applicato a $p_k(x)$ con punto iniziale $x_0 = 0$ genera una successione che converge in modo monotono a ξ_k con ordine 2. Si descriva il comportamento della successione ottenuta con $x_0 = \xi_{k-1}$.
c) Si determini un algoritmo per il calcolo di un passo del metodo di Newton applicato al polinomio $p_k(x)$ di costo $O(k)$.

Soluzione.
a) Dimostriamo per induzione che: $p_k'(x) > 0$ per ogni $x \in \mathbb{R}$, $p_k(0) < 0$, $p_k(\xi_{k-1}) > 0$, per $k \geq 1$. Questo implica che esiste unico $0 < \xi_k < \xi_{k-1}$ tale che $p_k(\xi_k) = 0$; in particolare ξ_k è l'unico zero in $[0, 1]$. Per $k = 1$ la tesi vale. Supponiamo valga per $k \geq 1$. È $p_{k+1}'(x) = 3p_k(x)^2 p_k'(x) + 1$, quindi $p_{k+1}'(x) > 0$; inoltre $p_{k+1}(0) = p_k(0)^3 < 0$ e $p_{k+1}(\xi_k) = \xi_k > 0$.
b) Si osserva che $p_{k+1}''(x) = 6p_k(x)p_k'(x)^2 + 3p_k(x)^2 p_k''(x)$, quindi $p_{k+1}''(x) < 0$ se $0 \leq x \leq \xi_{k+1}$. Dunque $p_{k+1}'(x)p_{k+1}''(x) < 0$ se $0 \leq x \leq \xi_{k+1}$, e questo implica la convergenza monotona se $0 \leq x_0 < \xi_{k+1}$. La convergenza è di ordine 2 perché $p_k'(\xi_k) \neq 0$ e $p_k''(\xi_k) \neq 0$.

Il metodo di Newton genera la successione

$$x_{n+1} = x_n - \frac{p_{k-1}(x_n)^3 + x_n}{3p_{k-1}(x_n)^2 p_{k-1}'(x_n) + 1} = \frac{3x_n p_{k-1}(x_n)^2 p_{k-1}'(x_n) - p_{k-1}(x_n)^3}{3p_{k-1}(x_n)^2 p_{k-1}'(x_n) + 1}$$

quindi se $x_0 = \xi_{k-1}$ allora $x_1 = 0$ e ci riconduciamo al caso $x_0 = 0$.
c) Ad ogni passo del metodo di Newton occorre calcolare $p_k(x)$ e $p_k'(x)$. Se usiamo le formula ricorsive $p_k(x) = p_{k-1}(x)^3 + x$ e $p_k'(x) = 3p_{k-1}(x)^2 p_{k-1}'(x) + 1$, il calcolo di $p_k(x)$ e di $p_k'(x)$ può essere effettuato on $o(k)$ operazioni aritmetiche.
□

Esercizio 18 Si analizzi la convergenza della successione $\{x_k\}_{k\geq 0}$ generata dal metodo di Newton applicato alla funzione $f(x) = x - 1 - 1/(x^2 - 1)$, al variare di $x_0 \in \mathbb{R}$, $x_0 \neq \pm 1$. In particolare,
a) dopo aver dimostrato che la funzione ha tre zeri reali, per ciascuno zero α di $f(x)$ si determini l'ordine di convergenza ad α;
b) per ciascuno zero α di $f(x)$ si determini l'insieme dei valori iniziali x_0 per cui la successione $\{x_k\}_{k\geq 0}$ converge ad α in modo monotono;
c) (facoltativo) per ciascuno zero α di $f(x)$ si determini l'insieme dei punti iniziali x_0 per cui la successione converge ad α.

Esercizio 19 Si analizzi la convergenza della successione $\{x_k\}_{k\geq 0}$ generata dal metodo di Newton applicato alla funzione $f(x) = (x^3 - x^2)^{1/3}$, al variare di $x_0 \in \mathbb{R}$. In particolare,
a) per ciascuno zero α di $f(x)$ si dica se c'è convergenza locale e si determini l'ordine di convergenza ad α;

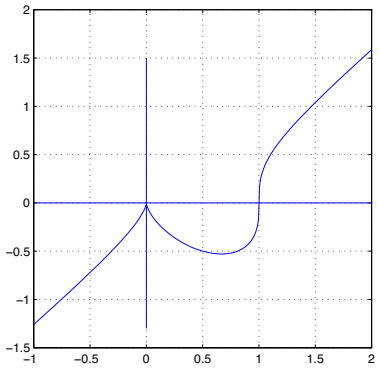


Figura 16: Esercizio 19, $f(x)$

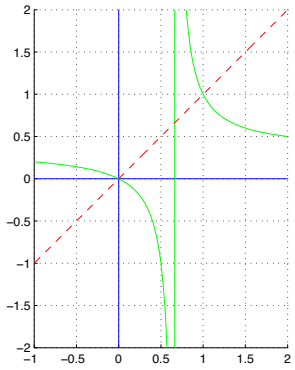


Figura 17: Esercizio 19, $g(x)$

b) per ciascuno zero α di $f(x)$ si determini un insieme dei valori iniziali x_0 per cui la successione $\{x_k\}_{k\geq 0}$ converge ad α; si determini inoltre un insieme dei punti iniziali x_0 per cui la convergenza avviene in modo alternato.

Soluzione.

a) La funzione $f(x)$ è continua su $\mathbb{R}$. I suoi zeri sono $\alpha_1 = 0$ e $\alpha_2 = 1$. Vale

$$f'(x) = \frac{1}{3}(x^3 - x^2)^{-2/3}(3x^2 - 2x),$$

dunque $\lim_{x\to 0^+} f'(x) = +\infty$, $\lim_{x\to 0^-} f'(x) = -\infty$, $\lim_{x\to 1^+} f'(x) = +\infty$, $\lim_{x\to 1^-} f'(x) = +\infty$. La funzione quindi non è derivabile in 0 e 1. Il grafico è riportato in Figura 16.

La funzione $f(x)$ è derivabile in ogni punto diverso da 0 e 1 quindi il metodo di Newton è applicabile in ogni punto diverso da questi e da $2/3$ dove la derivata è nulla.

Il metodo di Newton è:

$$x_{n+1} = x_n - 3\frac{(x_n^3 - x_n^2)^{1/3}}{(x_n^3 - x_n^2)^{-2/3}(3x_n^2 - 2x_n)} = x_n - 3\frac{x_n^3 - x_n^2}{3x_n^3 - 2x_n} = \frac{x_n}{3x_n - 2}.$$

Definisco

$$g(x) = \frac{x}{3x - 2}.$$

La funzione $g(x)$ è definita in un intorno di 0 e di 1, dunque il metodo di Newton può essere definito in un intorno di 0 e di 1 mediante la funzione $g(x)$. La funzione $g(x)$ è C^∞ in un intorno di 0 e di 1, quindi possiamo applicare i teoremi di convergenza per i metodi del punto fisso. Il grafico della funzione $g(x)$ ràppresentato nella Figura 17. Vale $g'(x) = -\frac{2}{(3x-2)^2}$, quindi $g'(\alpha_1) = -1/2$ e $g'(\alpha_2) = -2$. Dunque il metodo di Newton converge localmente in un intorno di $\alpha_1 = 0$ con convergenza lineare e non converge localmente a $\alpha_2 = 1$.

b) Poichè $g'(x) < 0$ per ogni $x < 2/3$ e $\alpha_1 = 0$, se $x_0 < 0$ vale $x_1 = x_1 - \alpha_1 = g(x_0) - g(\alpha_1) = g'(\xi)x_0 > 0$, per un opportuno $0 < \xi < x_0$. Se $0 < x_0 < 2/3$, ragionando in modo analogo, si trova che $x_1 < 0$. Induttivamente, se $x_0 < 2/3$, possiamo provare che la successione x_n è alternata fin tanto che $x_n < 2/3$.

Studiamo quindi le successioni x_{2n} e x_{2n-1}. Osserviamo che $x_{2n} = h(x_{2(n-1)})$ e $x_{2n+1} = h(x_{2n-1})$, $n \geq 1$, dove $h(x) = g(g(x)) = x/(4-3x)$. Vale $h'(x) = 4/(4-3x)^2$, dunque se $x < 0$ allora $0 < h'(x) < 1/4$, se $0 < x < 2/3$ allora $0 < h'(x) < 1$.

Supponiamo ora $x_0 < 0$. Allora $x_2 = h'(\xi)x_0$ per un opportuno $x_0 < \xi < 0$, dunque $(1/4)x_0 < x_2 < 0$; induttivamente vale $(1/4)^n x_0 < x_{2n} < 0$, dunque la successione x_{2n} converge a 0. Studiamo ora la successione x_{2n-1}. Osserviamo che se $x < 0$ allora $g(x) < 2/3$, quindi $0 < x_{2n-1} < 2/3$ per ogni n. Poiché se $0 < x < 2/3$ allora $1/4 < h'(x) < 1$, vale $0 < x_3 = h'(\eta)x_1 < x_1$, per un opportuno $0 < \eta < 2/3$. Induttivamente, vale $0 < x_{2n+1} < x_{2n-1}$ per $n \geq 1$. Quindi la successione x_{2n-1}, essendo limitata inferiormente e monotona decrescente, ha un limite $\beta \geq 0$ tale che $h(\beta) = \beta$. È necessariamente $\beta = 0$ perché se fosse $\beta > 0$, applicando il teorema di Lagrange si avrebbe $h(\beta) - h(0) = h'(\mu)\beta$ con μ nell'intervallo aperto $(0, \beta)$. Questo è assurdo essendo $0 < h'(\mu) < 1$. Il caso $0 < x_0 < 2/3$ è trattato in modo analogo.

Se $x_0 = 2/3$ il metodo di Newton non è definito.

Se $2/3 < x_0 < 1$ allora $x_1 - 1 = g(x_0) - g(1) = g'(\xi)(x_0 - 1)$ per un opportuno $x_0 < \xi < 1$, ma $g'(x) < -2$ se $2/3 < x < 1$, dunque $x_1 > 1$.

Se $x_0 > 4/3$ allora, dalla definizione di $g(x)$, segue che $x_1 = g(x_0) < 2/3$, dunque si applicano i ragionamenti fatti nel caso $x_0 < 2/3$ e la successione converge a 0.

Non esiste un insieme tale che per ogni x_0 in questo insieme la successione converge a 1. □

Esercizio 20 Dato un intero $n \geq 2$, e $a_i \in \mathbb{R}$, $a_i \geq 0$, $i = 0, \ldots, n$, $a_0 a_n \neq 0$, si ponga $f(x) = \sum_{i=1}^{n} a_i x^i - a_0$.

a) Si dimostrino le seguenti proprietà.

– esiste unico $\alpha \in \mathbb{R}$, $\alpha > 0$, tale che $f(\alpha) = 0$;

– vale $\alpha \leq a$ dove $a = \min\{(a_0/a_i)^{1/i} : \quad i = 1 \leq i \leq n, \quad a_i \neq 0\}$;

– le successioni definite da $x_{k+1} = x_k - f(x_k)/f'(a)$, $x_0 \in [0, a]$ convergono in modo monotono a α.

b) Assumendo $a_i = 1/i$, $i = 1, \ldots, n$, $a_0 = 1$ e $x_0 = 0$, determinare in funzione di n un valore di k tale che $\alpha - x_k < 10^{-15}$.

c) (Facoltativo) Enunciare proprietà analoghe a quelle del punto a) per il polinomio $f(x) = \sum_{i=0}^{n-1} a_i x^i - a_n x^n$, con $a_n \neq 0$, motivando adeguatamente.

Esercizio 21 Considerare il metodo di Newton applicato alle funzioni $x^2 - a$ e $x^{-2} - a^{-1}$ con $a > 0$ e scrivere le corrispondenti funzioni di punto fisso $g_1(x)$ e $g_2(x)$.

a) Determinare i parametri β, γ in modo che l'iterazione del punto fisso definita dalla funzione $g_3(x) = \beta g_1(x) + \gamma g_2(x)$ generi successioni localmente convergenti a $\sqrt{a}$ col massimo ordine di convergenza possibile.

b) Confrontare le tre iterazioni così ottenute in base al loro costo computazionale per approssimare $\sqrt{a}$ con errore al più ϵ. (Condurre una analisi asintotica per $\epsilon \to 0$ valutando solo il costo delle moltiplicazioni e divisioni, escluse quelle per potenze intere di 2).
c) Studiare la convergenza monotona delle successioni generate da $g_1(x), g_2(x)$ e $g_3(x)$.

Esercizio 22 Siano $a < \alpha < b$ numeri reali, $f(x) \in C^5([a,b])$, $f(\alpha) = 0$, $f'(\alpha) \neq 0$. Si consideri il metodo iterativo $x_{k+1} = g(x_k)$ definito dalla funzione $g(x) = x - t(x)/(1 - \theta t(x)s(x))$ dove θ è un parametro reale e $t(x) = f(x)/f'(x)$, $s(x) = f''(x)/f'(x)$. Dopo aver osservato che $t'(x) = 1 - s(x)t(x)$, $t(\alpha) = 0$ e $t'(\alpha) = 1$,

a) si dimostri che il metodo ha ordine di convergenza almeno 2 per ogni θ;

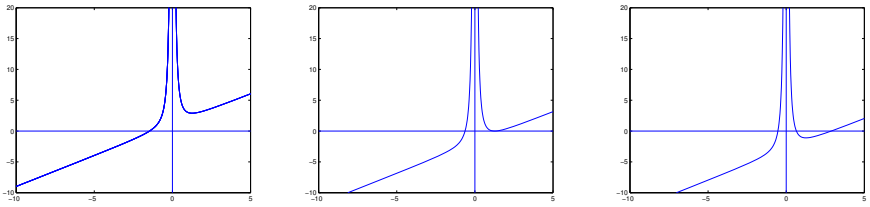
b) si determini un valore di θ per cui il metodo ha ordine almeno 3. Valutando il costo per passo in base al numero di valori di funzione e di derivate calcolati, si dimostri che il metodo così ottenuto è più conveniente in termini asintotici del metodo di Newton.

c) Si determini un valore di θ per cui il metodo ha convergenza superlineare nel caso $f'(\alpha) = 0$, $f''(\alpha) \neq 0$.

Esercizio 23 Si determini al variare del parametro $a \in \mathbb{R}$ il numero degli zeri reali della funzione $f(x) = x - x\log x - a$. Si analizzi al variare di a la convergenza del metodo di Newton applicato alla funzione $f(x)$ con particolare attenzione all'ordine di convergenza e alla convergenza monotona. Si determini l'insieme dei valori $x_0 \geq 0$ per cui la successione generata dal metodo di Newton converge.
Facoltativo: per $a = 0$ si dimostri che il metodo di Newton applicato alla funzione $f(x)/f'(x)$ ha convergenza localmente quadratica.

Esercizio 24 Siano a_i, $i = 1, \ldots, n$ numeri reali positivi e $b_1 < b_2 < \cdots < b_n$. Si dimostri che
a) la funzione razionale $f(x) = 1 + \sum_{i=1,n} a_i/(b_i - x)$ ha n zeri reali distinti $x_1 < x_2 < \cdots < x_n$ e si determinino intervalli $[\alpha_i, \beta_i]$, tali che $\alpha_i < x_i < \beta_i$ per $i = 1, \ldots, n$ e $\beta_i \leq \alpha_i$ per $i = 1, \ldots, n-1$.
b) il metodo di Newton applicato a $f(x)$ ha convergenza monotona in almeno uno dei due sottointervalli in cui ogni intervallo $[\alpha_i, \beta_i]$ viene diviso dallo zero in esso contenuto.
c) la convergenza locale è sempre almeno quadratica.

Esercizio 25 Sia $g(x) = 1/x^2 + 2x - a$ e si consideri il metodo iterativo dato da $x_{k+1} = g(x_k)$, $x_0 \in \mathbb{R}$.
a) Al variare di $a \in \mathbb{R}$ studiare il numero di punti fissi di $g(x)$ e la convergenza locale per ciascun punto fisso individuando se esistono casi di convergenza superlineare, sublineare e monotona.
b) Per $a = 2$ e per ogni punto fisso $\alpha > 0$, determinare gli insiemi dei punti $x_0 > 0$ a partire dai quali le successioni generate convergono ad α e l'insieme dei punti $x_0 > 0$ per i quali non c'è convergenza.
c) Per $a = 2$ e $x_0 > 1$, si dimostri che $x_{k+1} - 1 < 3(x_k - 1)^2$ e si determini il



(a) $a = -1$ (b) $a = a^*$ (c) $a = 3$

Figura 18: Esercizio 25, grafico di $f(x)$

numero di passi sufficienti ad approssimare il punto fisso $\alpha = 1$ con errore al più 2^{-52} se $x_0 = 13/10$.

Soluzione

a) Cerchiamo i punti fissi di $g(x)$. I punti fissi di $g(x)$ sono gli zeri della funzione $f(x) = 1/x^2 + x - a$. Tale funzione ha un minimo locale in $\bar{x} = 2^{1/3}$ e vale $f(\bar{x}) = 2^{-2/3} + 2^{1/3} - a$. Poniamo $a^* = 2^{-2/3} + 2^{1/3}$. Dunque, studiando la funzione, troviamo che:

- Se $a < a^*$, $f(x)$ ha uno zero $\alpha_1 < 0$;
- se $a = a^*$, $f(x)$ ha uno zero $\alpha_1 < 0$ e uno zero doppio $\alpha_2 = 2^{1/3}$;
- se $a > a^*$, $f(x)$ ha uno zero $\alpha_1 < 0$ e due zeri positivi, $0 < \alpha_2 < 2^{1/3} < \alpha_3$.

Il grafico della funzione $f(x)$ per alcuni valori di a è riportato in Figura 18.

La funzione $g(x)$ è C^1 in ogni intorno di ciascun punto fisso a vale $g'(x) = 2(1-1/x^3)$. In particolare $g'(\alpha_1) > 2$ perché $\alpha_1 < 0$, dunque non c'è convergenza locale in un intorno di α_1, per ogni valore di a.

Consideriamo il caso $a = a^*$ e studiamo la convergenza in un intorno di $\alpha_2 = 2^{1/3}$. Osserviamo che:

$$\begin{array}{ll} g'(x) < -1 & \text{se } 0 < x < (2/3)^{1/3} \\ -1 \le g'(x) < 0 & \text{se } (2/3)^{1/3} \le x < 1 \\ g(x) = 0 & \text{se } x = 1 \\ 0 < g'(x) \le 1 & \text{se } 1 < x \le 2^{1/3} \\ g'(x) > 1 & \text{se } x > 2^{1/3}. \end{array}$$

Vale $g'(\alpha_2) = 1$ e $0 < g'(x) < 1$ se $1 < x < \alpha_2$. Quindi, se $1 \le x_0 < \alpha_2$, la successione $\{x_k\}$ è tale che $1 < x_k < x_{k+1} < \alpha$, $k = 0, 1, \ldots$. Dunque la successione converge perché crescente e limitata, e la convergenza è sublineare. Se invece $x_0 > \alpha_2$ non ho convergenza perchè $g'(x) > 1$ se $x > \alpha_2$.

Consideriamo ora il caso $a > a^*$. Se $\alpha_2 > (2/3)^{1/3}$ allora $|g'(\alpha_2)| < 1$, c'è convergenza locale. La convergenza è superlineare se $\alpha_2 = 1$, che è vero se $a = 2$. La condizione $\alpha_2 > (2/3)^{1/3}$ è verificata quando $a < \bar{a} = (2/3)^{-2/3} + (2/3)^{1/3}$. Se $a > \bar{a}$, allora $|g'(\alpha_2)| > 1$ e non c'è convergenza locale.

b) Nel caso $a = 2$ vale $\alpha_2 = 1$, $g'(\alpha_2) = 0$, $g''(\alpha_2) \neq 0$. Quindi la convergenza è di ordine 2. Per le proprietà di segno della derivata di $g(x)$, se $\alpha_2 < x_0 \leq 2^{1/3}$, allora la convergenza è monotona decrescente. Se $(2/3)^{1/3} \leq x_0 < \alpha_2$, allora $\alpha_2 < x_1 \leq 2^{1/3}$ e la convergenza monotona dal secondo passo in poi. In α_3 non c'è convergenza locale perchè $g'(\alpha_3) > 1$.

c) Si osserva che $x_{i+1} - 1 = \frac{2x_i+1}{x_i^2}(x_i - 1)^2$. Poichè, se $x_0 > 1$ allora $x_i > 1$ per $i = 1, 2, \ldots$ e poichè $\frac{2x+1}{x^2} < 3$ se $x > 1$, allora $x_{i+1} - 1 < 3(x_i - 1)^2$. Ragionando per induzione vale $x_i - 1 < 3^{2^i-1}(x_0 - 1)^{2^i}$, dunque basta trovare i tale che $(3(x_0 - 1))^{2^i} < 2^{-52}$. Essendo $x_0 = 13/10$ si ottiene $i > \log_2(52/\log_2(10/9)) \approx 8.4183$. Dunque sono sufficienti 9 iterazioni. □

Esercizio 26 Per $\gamma > 0$ si consideri la funzione $f(x)$ definita da $f(0) = 0$, $f(x) = x - \gamma x/\log|x|$, se $x \neq 0, 1, -1$. Dopo aver determinato il numero di zeri di $f(x)$, si studi la convergenza locale del metodo di Newton applicato a $f(x)$ per ciascuno degli zeri, discutendone l'ordine di convergenza. Si studi per quali valori di x_0 la successione $\{x_k\}$ generata dal metodo di Newton converge in modo monotono a uno zero di $f(x)$; (facoltativo) dire per quali valori di $x_0 \neq \pm 1$ la successione non è definita o non converge.

Esercizio 27 Sia $g(x) : [a, b] \to \mathbb{R}$ derivabile con continuità e $\alpha \in (a, b)$ tale che $g(\alpha) = \alpha$.
a) Sapendo che esite un $x_0 \in [a, b]$ per cui la successione $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, è tale che $x_k \in (a, b)$ per $k \geq 0$ e converge ad α in modo sublineare e alternato (cioè $x_k > \alpha \quad \Rightarrow \quad x_{k+1} < \alpha$ e $x_k < \alpha \quad \Rightarrow \quad x_{k+1} > \alpha$), si dimostri che $g'(\alpha) = -1$.

b) Si dimostri che le successioni definite da $z_{k+1} = (z_k + g(z_k))/2$ convergono localmente ad α in modo superlineare e che la convergenza è almeno quadratica se $g \in C^2[a, b]$.
c) Sia $n \geq 2$ intero e sia $g \in C^2[a, b]$ con $g'(\alpha) = -1$, $g(\alpha) = \alpha$. Definiamo $G(x) = x + \sum_{i=1}^{n} \theta_i g^{[i]}(x)$, dove $\theta_i \in \mathbb{R}$, $i = 1, \ldots, n$ e $g^{[i]}(x) = g(g(\cdots g(x) \cdots))$ i-volte. Si dimostri che se $\alpha \neq 0$ e g"$(\alpha) \neq 0$ non esistono costanti $\theta_i \in \mathbb{R}$, $i = 1, \ldots, n$ tali che le successioni generate da $z_{k+1} = G(z_k)$ convergano localmente ad α con ordine maggiore di 2.

Soluzione.
a) Supponiamo sia $x_k > \alpha$. Poiché $x_{k+1} - \alpha < 0$ e $x_{k+1} - \alpha = g'(\xi_k)(x_k - \alpha)$ per un opportuno ξ_k compreso tra x_k e α, allora $g'(\xi_k) < 0$. In modo analogo si ragiona se $x_k < \alpha$. Inoltre, essendo la successione $\{x_k\}_k$ convergente a α ed essendo $g(x)$ derivabile con continuità, anche la successione $\{\xi_k\}_k$ converge ad α e si ottiene $\lim_k g'(\xi_k) = g'(\alpha) \leq 0$. Poiché la convergenza è sublineare, si ha $\lim_k \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = 1$. D'altra parte, essendo $g(x)$ derivabile con continuità, $\lim_k \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = |g'(\alpha)|$, dunque $g'(\alpha) = -1$.
b) Sia $h(x) = (x + g(x))/2$. Poiché $h \in C^2[a, b]$, se dimostriamo che $h'(\alpha) = 0$ allora la convergenza è almeno quadratica; è quadratica se $h''(\alpha) \neq 0$. Vale

$h'(x) = (1 + g'(x))/2$, dunque $h'(\alpha) = 0$. Inoltre $h''(x) = g''(x)/2$, dunque la convergenza è quadratica se $g''(\alpha) \neq 0$.
c) La condizione $G(\alpha) = \alpha$ diventa $\alpha(1 + \sum_{i=1}^{n} \theta_i) = \alpha$. Dunque se $\alpha \neq 0$ deve essere $\sum_{i=1}^{n} \theta_i = 0$. Calcoliamo $G'(x)$. Poiché $g^{[i]}(x) = g(g^{[i-1]}(x))$, $i \geq 1$, vale $g^{[i]'}(x) = g'(g^{[i-1]}(x))g^{[i-1]'}(x)$. Usando questa proprietà possiamo dimostrare per induzione che $g^{[i]'}(\alpha) = (-1)^i$. Dunque $G'(\alpha) = 1 + \sum_{i=1}^{n}(-1)^i\theta_i$ e la condizione di convergenza superlineare diventa $\sum_{i=1}^{n}(-1)^i\theta_i = -1$. Quindi esistono θ_i, $i = 1, \ldots, n$, tali che la convergenza è superlineare. In particolare, se $\alpha \neq 0$, scelti arbitrariamente $\theta_3, \ldots, \theta_n$, le costanti θ_1 e θ_2 risolvono il sistema

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} -\sum_{i=3}^{n} \theta_i \\ -1 - \sum_{i=3}^{n}(-1)^i\theta_i \end{bmatrix}.$$

Calcoliamo ora $G''(x)$. Osserviamo che $g^{[i]''}(x) = g''(g^{[i-1]}(x))g^{[i-1]'}(x)^2 + g'(g^{[i-1]}(x))g^{[i-1]''}(x)$. Usando questa ricorsione, possiamo provare per induzione che $g^{[i]''}(\alpha) = \frac{1}{2}(1-(-1)^i)g''(\alpha)$. Dunque $G''(\alpha) = \frac{1}{2}g''(\alpha)\sum_{i=1}^{n}(1-(-1)^i)\theta_i$ e, se $g''(\alpha) = 0$, la convergenza è di ordine maggiore di due. Se $g''(\alpha) \neq 0$ e $\alpha \neq 0$, poiché $\sum_{i=1}^{n} \theta_i = 0$ e $\sum_{i=1}^{n}(-1)^i\theta_i = -1$, allora la convergenza è quadratica. Se $g''(\alpha) \neq 0$, $\alpha = 0$ e $n \geq 3$, posso trovare costanti θ_i tali che $\sum_{i=1}^{n} \theta_i = 0$ e $\sum_{i=1}^{n}(-1)^i\theta_i = 0$, dunque la convergenza è di ordine maggiore di due.

□

Esercizio 28 Sia $\gamma \geq 0$; determinare il numero di punti fissi della funzione $g(x)$ definita da $g(0) = 0$, $g(x) = x(1-\gamma) + x/\log|x|$, per $x \neq 0, 1, -1$, e studiare la convergenza locale della successione $x_{k+1} = g(x_k)$ per $x_0 \in \mathbb{R}$, $x_0 \neq 1, -1$. In particolare, individuare i casi di convergenza monotona, lineare, superlineare e sublineare studiando l'ordine di convergenza.

Esercizio 29 Si studi al variare di $a \in \mathbb{R}$ il numero di punti fissi non negativi della funzione $g(x) = \sqrt{x} - a/\sqrt{x}$ e la convergenza locale delle successioni definite da $x_{k+1} = g(x_k)$ per $x_0 \geq 0$, inclusi l'ordine di convergenza e le proprietà di monotonia.

Esercizio 30 Sia $g : \mathbb{R} \to \mathbb{R}$ funzione continua ed $\alpha \in \mathbb{R}$ un suo punto fisso, cioè tale che $\alpha = g(\alpha)$. Si assuma inoltre che g sia derivabile con continuità per $x \neq \alpha$ e che esistano costanti $0 \leq a \leq 1 < b$ tali che $-b \leq g'(x) \leq -1$ per $x < \alpha$, $-a \leq g'(x) \leq 0$ per $x > \alpha$.
a) Si dimostri che α è l'unico punto fisso di g.
b) Per $x_0 \in \mathbb{R}$ si definisca $x_{k+1} = g(x_k)$, $k \geq 0$. Si diano condizioni su a, b affinché le successioni $\{x_{2k}\}$, $\{x_{2k+1}\}$ e $\{x_k\}$ siano convergenti per ogni $x_0 \in \mathbb{R}$.
c) Assumendo che $\lim_{x\to\alpha^+} g'(x) = \theta$, $\lim_{x\to\alpha^-} g'(x) = \sigma$, $\theta \neq \sigma$, sotto le condizioni trovate al punto b) si studi la velocità di convergenza delle successioni $\{x_k\}$, $\{x_{2k}\}$, $\{x_{2k+1}\}$.

Esercizio 31 Sia $n > 1$ un intero e a_i, $i = 1, 2, \ldots, n$ numeri reali positivi. Si consideri l'iterazione $x_{k+1} = g(x_k)$, con $x_0 > 0$, dove

$$g(x) = \begin{cases} x - \theta x \sum_{i=1}^{n} \log(x/a_i), & \text{se } x > 0, \\ 0, & \text{se } x = 0, \end{cases}$$

e $\theta \neq 0$.
a) Si determinino i punti fissi di $g(x)$ e si studi la convergenza locale a tali punti fissi al variare di θ.
b) Per ogni punto fisso α, dopo aver determinato il valore di θ che massimizza la velocità di convergenza ad α si dia una limitazione superiore a $|x_k - \alpha|$ nelle ipotesi $1 < a_i < 3/2$, $i = 1, \ldots, n$ e $x_0 = 1$.

Esercizio 32 Si vuole approssimare il reciproco di un numero reale $a > 0$, eseguendo solamente moltiplicazioni, addizioni e sottrazioni. Si consideri allora la classe di iterazioni del punto fisso $x_{k+1} = g(x_k)$, dove

$$g(x) = x + (1 - ax)(\beta x + \gamma a x^2), \quad \alpha, \beta, \gamma \in \mathbb{R}.$$

a) Si determinino i parametri $\beta, \gamma \in \mathbb{R}$, indipendenti da a, in modo da ottenere il metodo iterativo con l'ordine di convergenza più alto possibile.

b) Si mettano in relazione i residui $1 - ax_k$ e $1 - ax_{k+1}$. Si confronti l'efficienza del metodo ottenuto con quella del metodo dato da $\hat{g}(x) = 2x - ax^2$. In particolare, se $1/2 \leq a < 1$ e $x_0 = 1$, si valuti il numero di passi k e il numero totale di operazioni aritmetiche affinchè $|x_k - a^{-1}|/a^{-1} < 2^{-52}$.

c) Si individui una espressione generale di $g(x)$ che dia un metodo a convergenza di ordine q arbitrario e se ne valuti l'efficienza.

Esercizio 33 Sia $\varphi(x) \in C^2([a, b])$ e $\alpha \in [a, b]$ tale che $\varphi(\alpha) = 0$, dove $[a, b] \subset \mathbb{R}$ e $\varphi'(x) \neq 0$ in $[a, b]$.
a) Si dimostri che per ogni $x \in [a, b]$ esiste $\xi \in (a, b)$ tale che $\varphi(x)/\varphi'(x) = x - \alpha - (x - \alpha)^2\varphi''(\xi)/(2\varphi'(x))$.
b) Si studi la convergenza locale del metodo di Newton applicato alla funzione $f(x) = \varphi(x) \log |\varphi(x)|$ per $x \neq \alpha$, $f(\alpha) = 0$, e applicato alla funzione $f(x) = \varphi(x)(\log(1 + \varphi(x)))$.
c) Se $m(x)$ è una funzione continua per cui esiste $\gamma > 0$ tale che $|m(x) - m(y)| \leq \gamma|x - y|$, per $x, y \in [a, b]$ e inoltre $m(\alpha) = 1$, si dimostri la convergenza almeno quadratica dell'iterazione $x_{k+1} = g(x_k)$ definita da $g(x) = x - m(x)\varphi(x)/\varphi'(x)$.

Esercizio 34 Sia $f(x) = |x|^a - 2x$, dove $a \in (0, 1)$. Determinare il numero di zeri reali di $f(x)$ e, per ciascuno zero, studiare la convergenza locale del metodo di Newton per l'approssimazione di tale zero, in particolare si determini l'ordine di convergenza e l'eventuale convergenza monotona.
Determinare l'insieme dei punti iniziali per cui la successione generata dal metodo di Newton è convergente.

Esercizio 35 Per approssimare la radice p-esima del reciproco di un numero reale $a > 0$ senza eseguire divisioni si consideri la famiglia di funzioni $g(x) = x(1 + \beta R(x) + \gamma R^2(x))$, $R(x) = 1 - ax^p$, $\beta, \gamma \in \mathbb{R}$, e le successioni generate da $x_{k+1} = g(x_k)$ a partire da $x_0 \in \mathbb{R}$.

Si individui nella classe di metodi quello di massimo ordine di convergenza e quello di massima efficienza computazionale.

Come si può generalizzare la classe di algoritmi per avere ordine di convergenza arbitrario q e come si può estendere l'analisi fatta?

Esercizio 36 Si determinino al variare di $a \in \mathbb{R}$ il numero di soluzioni dell'equazione $\log(x^2 + x + 1) - x = a$ e si descriva un metodo per approssimare tali soluzioni che abbia convergenza superlineare individuando i casi di convergenza monotona ed alternata.

Esercizio 37 Al variare di $c \in \mathbb{R}$ si determini il numero di punti fissi della funzione $g(x)$ definita da

$$g(x) = x^2 - \log x + c$$

e si studi la convergenza locale delle successioni generate da $x_{k+1} = g(x_k)$ a partire da $x_0 \in \mathbb{R}$. Si confronti con il metodo di Newton applicato all'equazione $x - g(x) = 0$.

Esercizio 38 Sia $g(x) : [a, b] \to \mathbb{R}$ di classe C^2 e $\alpha \in [a, b]$ tale che $\alpha = g(\alpha)$. Si supponga di conoscere $\theta = g'(\alpha)$. Si determinino i valori di ξ_0, ξ_1 in modo che il metodo iterativo definito da $x_{k+1} = G(x_k)$ dove $G(x) = \xi_0 x + \xi_1 g(x)$ sia localmente convergente ed α ed abbia il massimo ordine di convergenza locale. Si discuta sotto quali condizioni il nuovo metodo così ottenuto è più efficiente del metodo dato dalla funzione $g(x)$.
Se $g \in C^\infty[a, b]$ e $\theta = 0$ dire se esiste un $n > 1$ per cui nella classe di metodi definiti da da $G(x) = \xi_0 x + \xi_1 g(x) + \xi_2 g(g(x)) + \cdots + \xi_n g(g(\cdots g(x) \cdots)$ esiste un metodo più efficiente del metodo originale.

Esercizio 39 Siano $\alpha \in \mathbb{R}$ e $g_1(x) \in C^1([\alpha, +\infty[)$, $g_2(x) \in C^1(]-\infty, \alpha])$ tali che $g_1(\alpha) = g_2(\alpha) = \alpha$, $-\theta < g_1'(x) \leq 0$ per $x \geq \alpha$, $-\sigma < g_2'(x) \leq 0$ per $x \leq \alpha$, dove $\sigma, \theta > 0$, $\theta\sigma = \lambda < 1$. Si definisca

$$g(x) = \begin{cases} g_1(x) & \text{se } x \geq \alpha \\ g_2(x) & \text{se } x < \alpha. \end{cases}$$

Si dimostri che
a) α è l'unico punto fisso di $g(x)$;
b) per ogni $x_0 \in \mathbb{R}$ la successione generata da $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, converge ad α.
c) Si studi il tipo di convergenza (monotona o alternata), l'ordine di convergenza e il fattore di convergenza delle successioni $\{x_k\}$, $\{x_{2k}\}$, $\{x_{2k+1}\}$; in particolare si tratti il caso in cui $g_1'(\alpha) \neq g_2'(\alpha)$.
d) Sia $\lambda_k = |x_k - \alpha|/|x_{k-1} - \alpha|$ la riduzione dell'errore al passo k, e si definisca φ_k la media geometrica delle riduzioni degli errori nei primi k passi. Si dimostri che, anche se $g_1'(\alpha) \neq g_2'(\alpha)$ il limite $\lim_k \varphi_k$ esiste ed è uguale a $\sqrt{g_1'(\alpha) g_2'(\alpha)}$.

Esercizio 40 Siano $a, b \in \mathbb{R}$ e $g(x) = ax + b/x$. Si diano condizioni sufficienti su a, b affinché $g(x)$ abbia almeno un punto fisso $\alpha \in \mathbb{R}$ ed esista un intorno

$\mathcal{I} = [\alpha - \rho, \alpha + \rho]$, $\rho > 0$, per cui la successione definita da $x_{k+1} = g(x_k)$ sia convergente per ogni $x_0 \in \mathcal{I}$.

Si diano condizioni sufficienti su a, b affinché la convergenza sia localmente monotona e localmente quadratica.

Esercizio 41 Si supponga di avere a disposizione un metodo per il calcolo del valore di e^x dato x. Per poter approssimare il logaritmo naturale $\alpha = \log(a)$ di un numero a tale che $1 < a < e$, si considerino le iterazioni del tipo $x_{k+1} = g(x_k)$, $x_0 \in [0, 1]$, dove $g(x)$ è una delle seguenti funzioni

$$g_1(x) = x + 1 - \frac{1}{a}e^x, \qquad g_2(x) = x + \frac{3}{2} - \frac{2}{a}e^x + \frac{1}{2a^2}e^{2x}$$

a) Si studi la convergenza locale delle due successioni generate in questo modo con attenzione all'ordine e alla monotonia, e si determini quale delle due è più efficiente dal punto di vista computazionale supponendo che il calcolo dell'esponenziale costi l'equivalente di 20 operazioni aritmetiche.

b) Si supponga che il valore effettivamente calcolato di e^x sia affetto da un errore minore o uguale a δ mentre le altre operazioni aritmetiche sono svolte in modo esatto. Si determini una limitazione al primo ordine in δ degli errori di approssimazione forniti dai due metodi.

c) Fra tutti i metodi definiti da $g(x) = x + p(e^x)$, dove $p(t)$ è un polinomio di grado al più m, determinare quello che garantisce il massimo ordine di convergenza e quello che dà la massima efficienza.

Esercizio 42 È assegnato un intero $n > 1$ e una funzione $f(x) \in C^\infty[a, b]$, di cui sappiamo che $f(\alpha) = 0$ per un certo $\alpha \in [a, b]$ e che $f^{(i)}(\alpha) = 1$ per $i = 1, \ldots, 2n$. Vogliamo individuare un metodo efficiente per approssimare α. Per questo consideriamo la classe di metodi iterativi definiti da

$$g(x) = x - a_0 f(x) - a_1 f'(x) - \cdots - a_n f^{(n)}(x)$$

ottenuta al variare dei parametri $a_1, \ldots, a_n$ in $\mathbb{R}$.

a) Si determini in questa classe il sottoinsieme dei metodi che convergono localmente ad α col più alto ordine di convergenza. Assumendo che il calcolo di ciascuna $f^{(i)}(x)$ abbia un costo di 10 operazioni aritmetiche, qual è il metodo di massima efficienza computazionale, in relazione all'ordine di convergenza e al costo per passo?

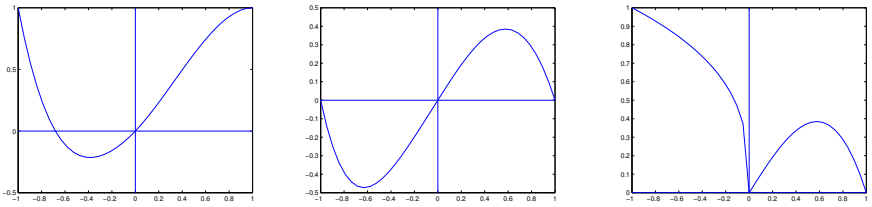
b) Dire se si possono ottenere ordini di convergenza più alti nella classe definita da

$$g(x) = x - f(x)(a_0 + a_1 f'(x) + a_2 f''(x) + \cdots + a_n f^{(n)}(x))$$

Esercizio 43 Sia $g(x) \in C^1([0, 1])$ tale che $|g'(x)| \leq 1/2$ per $x \in [0, 1]$. Sia $\alpha \in [0, 1]$ un punto fisso di $g(x)$.

a) Sapendo che $1/3 \leq \alpha \leq 2/3$ si dimostri che la successione $x_{k+1} = g(x_k)$ converge ad α per ogni scelta del punto iniziale $x_0 \in [0, 1]$.

b) Sapendo che $-\lambda \leq g'(x) \leq \lambda$, con $0 < \lambda < 1$, determinare $\alpha_1, \alpha_2 \in [0, 1]$ tali che se $\alpha_1 \leq \alpha \leq \alpha_2$ la successione x_k converge per ogni $x_0 \in [0, 1]$.



(a) Funzione a) (b) Funzione b) (c) Funzione c)

Figura 19: Esercizio 44, grafico di $f(x)$

Esercizio 44 Si studi la convergenza locale in un intorno di 0 del metodo di Newton applicato alle seguenti funzioni.

$$a)\quad f(x)=\begin{cases} -x^3+x^2+x & \text{se } x\geq 0\\ x^4+x^2+x & \text{se } x<0\end{cases} \qquad b)\quad f(x)=\begin{cases} -x^3+x & \text{se } x\geq 0\\ x^4+x & \text{se } x<0\end{cases}$$

$$c)\quad f(x)=\begin{cases} -x^3+x & \text{se } x\geq 0\\ (-x)^{1/3} & \text{se } x<0\end{cases}$$

Nel caso l'ordine di convergenza non sia definibile, si determini un valore di p per cui la convergenza è di *ordine almeno p*, oppure, qualora sia possibile, si determini l'ordine di convergenza delle due sottosuccessioni formate rispettivamente dalle componenti di indice pari e di indice dispari.
d) Sia $\{x_k\}$ la successione generata a partire da x_0 da $x_{k+1}=g(x_k)$, con $g(x):\mathbb{R}\to\mathbb{R}$ funzione continua. Se $x_i=-1/i$ per i dispari e $x_i=1/2^i$ per i pari, si dimostri che $g(x)$ non è derivabile in 0.

Soluzione. I grafici della funzione $f(x)$ sono riportati in Figura 19. Denotiamo con $f_1(x)$ e $f_2(x)$ rispettivamente l'espressione di $f(x)$ per $x\geq 0$ e per $x<0$.

a) Vale $f_1'(x)=-3x^2+2x+1$, $f_1''(x)=-6x+2$, $f_2'(x)=4x^3+2x+1$, $f_2''(x)=12x+2$. Troviamo che $f(x)$ è di classe C^2 in un intorno di zero e $f'(0)=1$, $f''(0)=2$. Dunque, applicando i risultati di convergenza del metodo di Newton, il metodo ha convergenza locale di ordine 2. Osserviamo anche che $f_1'(x)f_1''(x)>0$ se $0<x<1/3$. Quindi, per il Teorema 8, se $0<x_0<1/3$, la convergenza è monotona decrescente. Osserviamo che se $x_0<0$ è sufficientemente vicino a zero, allora $f_2(x)<0$ e $f_2''(x)>0$. Quindi, essendo $g_2'(x)=f_2(x)f_2''(x)/(f_2'(x))^2$, allora $x_1>0$ e la convergenza è monotona decrescente dal secondo passo in poi.

b) Vale $f_1'(x)=-3x^2+1$, $f_1''(x)=-6x$, $f_2'(x)=4x^3+1$, $f_2''(x)=12x$. Troviamo che $f(x)$ è di classe C^2 in un intorno di zero e $f'(0)=1$, $f''(0)=0$. Dunque, applicando i risultati di convergenza del metodo di Newton, il metodo ha convergenza locale di ordine almeno 2. Poiché f non è di classe C^3 non posso usare la teoria per determinare l'ordine di convergenza.

Il metodo di Newton applicato alla funzione $f_1(x)$ è definito dalla funzione $g_1(x)=-2x^3/(-3x^2+1)$, mentre il metodo di Newton applicato alla funzione $f_2(x)$ è definito dalla funzione $g_2(x)=3x^4/(4x^3+1)$. Si osserva che

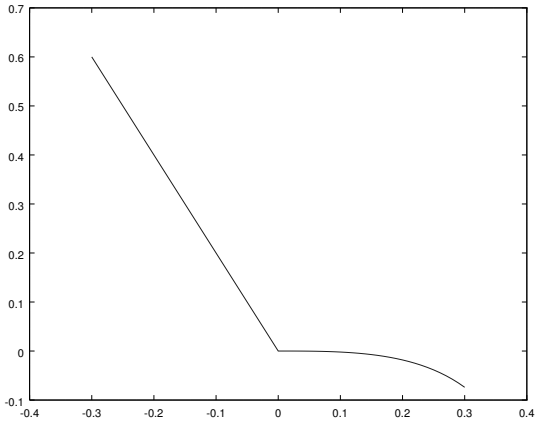
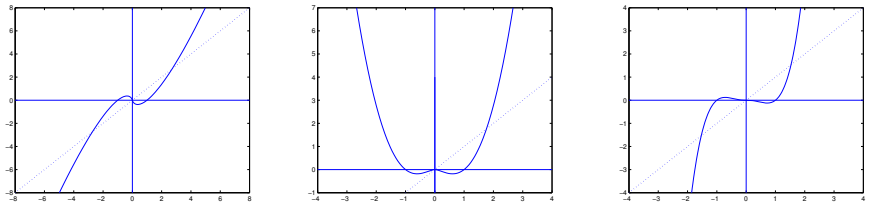


Figura 20: Esercizio 44: funzione $g(x)$ del punto c)

se $x_0 > 0$ ed è sufficientemente vicino a zero, allora $x_1 = g_1(x_0) < 0$; se $x_0 < 0$ ed è sufficientemente vicino a zero, allora $x_1 = g_2(x_0) > 0$. Dunque la successione è alternata. Vogliamo studiare l'ordine di convergenza. Si osserva che $\lim_{x\to 0^+} |g_1(x)|/|x^3| = 2$ e $\lim_{x\to 0^-} |g_2(x)|/|x^4| = 3$, quindi l'ordine di convergenza non è definibile poiché non esiste un numero $p > 1$ tale che $\lim_{i\to\infty} |x_{i+1}|/|x_i|^p = \theta \neq 0$. Possiamo però definire l'ordine di convergenza delle sottosuccessioni di indici pari e dispari. Sia $x_0 > 0$ sufficientemente vicino a 0, allora $x_{2(i+1)} = g_2(g_1(x_{2i}))$ e $x_{2i+1} = g_1(g_2(x_{2i-1}))$. Troviamo che $g_2(g_1(x)) = 48x^{12}/((3x^2-1)(32x^9+27x^6-27x^4+9x^2-1))$ e $g_1(g_2(x)) = (54x^{12})/((4x^3+1)(27x^8-16x^6-8x^3-1))$, quindi $\lim_{i\to\infty} |x_{2(i+1)}|/|x_{2i}|^{12} = 48$, $\lim_{i\to\infty} |x_{2i+1}|/|x_{2i-1}|^{12} = 54$, e le sottosuccessioni di indici pari e dispari hanno ordine di convergenza 12. In modo analogo si ragiona se $x_0 < 0$.

c) Vale $f_1'(x) = -3x^2 + 1$, $f_1''(x) = -6x$, $f_2'(x) = -1/3(-x)^{-2/3}$, dunque $f(x)$ non è di classe C^1 in un intorno di zero. Troviamo che il metodo di Newton applicato alla funzione $f_1(x)$ è definito dalla funzione $g_1(x) = -2x^3/(-3x^2+1)$, mentre il metodo di Newton applicato alla funzione $f_2(x)$ è definito dalla funzione $g_2(x) = -2x$. La funzione $g(x)$ definita come $g_1(x)$ se $x \geq 0$, $g_2(x)$ se $x < 0$, è continua ma non di classe C^1 in un intorno di zero (si veda la figura 20).

Osserviamo che se $x_0 > 0$ ed è abbastanza vicino a zero, allora $x_1 = g_1(x_0) < 0$; se $x_0 < 0$ allora $x_1 = g_2(x_0) > 0$. Studiamo quindi la convergenza locale delle sottosuccessioni di indice pari e dispari. Le funzioni $G(x) = g_2(g_1(x)) = 4x^3/(-3x^2+1)$ e $H(x) = g_1(g_2(x)) = 16x^3/(1-12x^2)$ sono di classe C^1 in un intorno di 0. Vale $G'(0) = 0$ e $H'(0)$, quindi se x_0 è sufficientemente vicino a zero, le sottosuccessioni di indici pari e dispari convergono in modo superlineare. L'ordine di convergenza è 3 poiché $\lim_{x\to 0} \frac{|G(x)|}{|x|^3} = 4 \neq 0$ e $\lim_{x\to 0} \frac{|H(x)|}{|x|^3} = 16 \neq 0$.



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$

Figura 21: Esercizio 46, grafico di $f(x)$

d) Se $g(x)$ fosse di classe C^1 in un intorno di 0 allora per il teorema di Lagrange sarebbe $x_{i+1} = x_i g'(\xi_i)$ per ξ_i compreso nell'intervallo aperto di estremi x_i e 0. Poichè $\lim x_i = 0$ si avrebbe $\lim x_{i+1}/x_i = \lim g'(\xi_i) = g'(0)$. Quindi, se i è dispari allora $x_{i+1}/x_i = -i/2^{i+1}$ mentre se i è pari $x_{i+1}/x_i = -2^i/(i+1)$. Per cui il limite di x_{i+1}/x_i non può esistere.

Esercizio 45 Siano $g_i(x)$, per $i \in \mathbb{N}$, funzioni di classe C^1 sull'intervallo $\mathcal{I} = [\alpha - \rho, \alpha + \rho]$ tali che $g_i(\alpha) = \alpha$ per $i \in \mathbb{N}$. Si assuma inoltre che $|g_i'(x)| \leq \lambda_i < 1$ per $i \in \mathbb{N}$ e $x \in \mathcal{I}$. Si dimostri che
a) se $\lambda_i \leq \lambda < 1$ per $i \in \mathbb{N}$ allora per ogni $x_0 \in \mathcal{I}$ la successione generata da $x_{i+1} = g_i(x_i)$ è tale che $x_i \in \mathcal{I}$ e $\lim_i x_i = \alpha$;
b) se $\lim_i \lambda_i = \mu < 1$ allora per ogni $x_0 \in \mathcal{I}$ la successione generata da $x_{i+1} = g_i(x_i)$ è tale che $x_i \in \mathcal{I}$ e $\lim_i x_i = \alpha$.
Cosa si può dire della velocità di convergenza della successione x_i?

Esercizio 46 Sia $k \geq 1$ un intero e si consideri la funzione $f(x)$ definita da $f(0) = 0$, $f(x) = x^k \log|x|$, se $x \neq 0$. Determinare il numero dei punti fissi di $f(x)$ e studiare la convergenza locale della successione $x_{i+1} = f(x_i)$ per x_0 in un intorno di ciascun punto fisso. Cosa si può dire della convergenza locale in un intorno di 0 delle successioni generate dal metodo di Newton applicato a $f(x)$?

Soluzione. Cerchiamo i punti fissi di $f(x)$. Si osserva innanzitutto che se k è pari la funzione $f(x)$ è pari, se k è dispari allora $f(x)$ è dispari. Il grafico della funzione $f(x)$ è riportato in Figura 21. Per ogni valore di k, un punto fisso è $\alpha_1 = 0$. Se $k \geq 2$ è pari, $f(x)$ ha un punto fisso $\alpha_2 > 1$; se $k \geq 1$ è dispari, $f(x)$ ha un punto fisso $\alpha_2 > 1$ e un punto fisso $\alpha_3 = -\alpha_2$.

Studiamo la convergenza locale. Consideriamo il caso $k = 1$. Troviamo che $f'(x) = \log(|x|) + 1$ se $x \neq 0$. Quindi, poiché $\lim_{x\to 0^+} f'(x) = -\infty$ e $\lim_{x\to 0^-} f'(x) = -\infty$, la funzione non è derivabile in 0. Inoltre, per ϵ sufficientemente piccolo si ha che $|x_0| < \epsilon$ implica $|x_1| > |x_0|$ per cui non ci può essere convergenza locale in un intorno di 0. Negli altri due punti fissi vale $f'(x) = 2 > 1$ per cui non c'e' convergenza locale.

Consideriamo ora il caso $k \geq 2$. Troviamo che $f'(x) = x^{k-1}(k \log(|x|) + 1)$ se $x \neq 0$; inoltre si può verificare che $f'(x)$ esiste ed è continua in 0 e $f'(0) = 0$.

Questo implica che c'è convergenza locale superlineare in un intorno di 0. Per studiare l'ordine di convergenza applichiamo la definizione: osserviamo che

$$\lim_{i\to\infty}\frac{|x_{i+1}|}{|x_i^p|}=\lim_{i\to\infty}\frac{|x_i^k\log(|x_i|)|}{|x_i^p|}=\begin{cases}0 & \text{se } p<k\\ +\infty & \text{se } p\geq k\end{cases}$$

quindi l'ordine di convergenza non è definito, ma la convergenza è superlineare di ordine almeno $k-\epsilon$ per ogni $0<\epsilon<k$ nel senso che esiste una costante $\gamma>0$ tale che $|x_{i+1}|\leq\gamma|x_i|^{k-\epsilon}$.

Per quanto riguarda i punti fissi α_2 e α_3 (se k è dispari), poiché $f'(\alpha_2)>1$ e $f'(\alpha_3)>1$, non c'è convergenza locale in un intorno di ciascuno.

Il metodo di Newton applicato alla funzione $f(x)$, per $x\neq 0$, è definito dalla funzione

$$g(x)=x-\frac{x^k\log(|x|)}{x^{k-1}(k\log(|x|)+1)}=\frac{x((k-1)\log(|x|)+1)}{k\log(|x|)+1}.$$

Poiché $\lim_{x\to 0^+}g(x)=0$ e $\lim_{x\to 0^-}g(x)=0$, possiamo estendere $g(x)$ nel punto 0 definendo $g(0)=0$. La funzione $g(x)$ ha derivata

$$g'(x)=\frac{1}{k(k\log(|x|)+1)}-\frac{1}{(k\log(|x|)+1)^2}+\frac{k-1}{k}$$

e si può verificare che $g(x)$ è derivabile con continuità in 0 e vale $g'(0)=(k-1)/k$. Quindi il metodo di Newton ha convergenza superlineare se $k=1$, lineare se $k>1$. Nel caso $k=1$ vale

$$\lim_{i\to\infty}\frac{|x_{i+1}|}{|x_i^p|}=\lim_{i\to\infty}\frac{\frac{x_i}{\log(|x_i|)+1}}{|x_i^p|}=\frac{1}{|x_i^{p-1}|(\log(|x_i|)+1)}=\begin{cases}0 & \text{se } p=1\\ +\infty & \text{se } p>1.\end{cases}$$

Quindi la convergenza è superlineare ma l'ordine non è definito. □

Esercizio 47 Vogliamo risolvere numericamente l'equazione $x^x=a$ per $a>0$. Per questo si dia risposta alle seguenti questioni.
a) Dire quante soluzioni ha l'equazione.
b) Studiare la convergenza locale delle iterazioni del tipo $x_{k+1}=g(x_k)$ dove $g(x)$ è una delle seguenti funzioni

b1) $\log a/\log x$; b2) $a^{1/x}$; b3) $(x+\log a)/(1+\log x)$.

Per ciascuna radice si dica se c'è convergenza locale, se è lineare, superlineare o sublineare, monotona o alternata.

Esercizio 48 Sia $p>1$ intero, $g(x)\in C^p([a,b])$ e $\alpha\in(a,b)$ tale che $\alpha=g(\alpha)$. Supponiamo esista una successione $\{x_k\}$ definita da $x_{k+1}=g(x_k)$, $x_0\in[a,b]$ tale che $x_k\in[a,b]$ e $\lim_k x_k=\alpha$ con ordine di convergenza p. Si dimostri che esiste un k_0 tale che per $k\geq k_0$ la successione converge ad α con una delle seguenti modalità
a) se p è pari, la convergenza è monotona: da destra se $g^{(p)}(\alpha)>0$, da sinistra se $g^{(p)}(\alpha)<0$;
b) se p è dispari, la convergenza è monotona se $g^{(p)}(\alpha)>0$; è alternata se $g^{(p)}(\alpha)<0$.

Esercizio 49 Si definisca $g(x) = x\sin(1/x)$ se $x \neq 0$, $g(x) = 0$ se $x = 0$.
a) Dare una espressione esplicita dei punti fissi di $g(x)$.
b) Discutere la convergenza locale delle successioni generate a partire da $x_0 \neq 0$ da $x_{k+1} = g(x_k)$ in ciascun punto fisso non nullo. Nel caso di convergenza si dica se questa è monotona, lineare, superlineare o sublineare.

Esercizio 50 Siano $a, b, \alpha \in \mathbb{R}$, tali che $a < \alpha < b$. Siano $f(x) \in C^3([a,b])$, $h(x) \in C^2([a,b])$ funzioni a valori reali tali che $f(\alpha) = 0$, $f'(\alpha) = h(\alpha) \neq 0$. Si consideri il metodo di iterazione funzionale definito dalla funzione $g(x) = x - f(x)/h(x)$.

a) Dimostrare che il metodo è localmente convergente e studiarne l'ordine di convergenza.

b) Per calcolare la radice quadrata $\alpha > 0$ del numero reale $a > 0$ si consideri la funzione $f(x) = x^3 - ax$ tale che $f(\alpha) = 0$, $f'(\alpha) = 2a$, e si applichi il metodo iterativo analizzato nel punto precedente scegliendo $h(x) = 2a + \gamma x f(x)$ dove γ è un parametro reale. Determinare il parametro γ che massimizza l'ordine di convergenza del metodo e valutarne l'ordine.

c) Confrontare il metodo ottenuto (sapendo che il suo ordine è 3) con il metodo di Newton applicato all'equazione $f(x) = 0$ e dire quale dei due è computazionalmente più conveniente.

d) Studiare il punto a) nell'ipotesi in cui $f \in C^2([a,b])$ e $h \in C^1([a,b])$.

Esercizio 51 Per calcolare la radice cubica $\alpha = \sqrt[3]{a}$ del numero reale $a > 0$ si consideri l'iterazione $x_{k+1} = g(x_k)$ dove x_0 è scelto in un intorno di α e $g(x) = x - f(x)/(\mu + \nu x^q f(x))$, $f(x) = x^4 - ax$ con $\mu, \nu \in \mathbb{R}$, $q \in \mathbb{Z}$.

a) Determinare i valori di μ, ν e q che danno ordine di convergenza almeno 3 e scrivere la classe di metodi così ottenuta. Dimostrare che l'ordine di convergenza dei metodi in questa classe è 3.

b) Si ponga $\nu = 0$ e si determini μ in modo da massimizzare l'ordine di convergenza. Si confronti l'efficienza del metodo ottenuto con quella del miglior metodo nella classe di cui al punto a) e con quella del metodo di Newton applicato alla funzione $x^3 - a$.

Esercizio 52 Sia a un numero intero positivo che non sia un quadrato perfetto. Per approssimare la radice quadrata di a mediante un metodo iterativo che ad ogni passo non esegua divisioni ma solo moltiplicazione e addizioni/sottrazioni tra numeri razionali, si considerino le iterazioni del punto fisso $x_{k+1} = g(x_k)$ con $g(x)$ polinomio di grado n a coefficienti razionali.

a) Determinare il polinomio $g(x)$ di grado minimo a coefficienti razionali che garantisce convergenza locale quadratica a $\sqrt{a}$.

b) Per $a = 3$ determinare un intervallo $\mathcal{I}$ contenente $\sqrt{a}$ per cui per ogni $x_0 \in \mathcal{I}$ la successione x_k converge a $\sqrt{a}$.

c) Dare una limitazione inferiore al massimo ordine di convergenza che si può ottenere con un polinomio $g(x)$ di grado n a coefficienti razionali.

Esercizio 53 Sia $f(x) \in C^p[a,b]$, con $p \geq 3$, $\alpha \in (a,b)$, $f(\alpha) = 0$, $f'(\alpha) \neq 0$. Si assuma che la successione $x_{k+1} = x_k - f(x_k)/f'(x_k)$ generata dal metodo di

Newton a partire da $x_0 \in [a,b]$ sia ben definita, con $x_k \in [a,b]$, e converga ad α.
a) Si dimostri che se x_k converge in modo alternato ad α allora la convergenza è di ordine almeno 3.
b) Si dimostri che se $f'(\alpha)f''(\alpha) > 0$ (rispettivamente $f'(\alpha)f''(\alpha) < 0$), la convergenza è quadratica ed esiste un k_0 tale che per $k > k_0$ la successione è decrescente (rispettivamente crescente).

Esercizio 54 Sia $g(x) \in C^2([a,b])$ ed $\alpha \in [a,b]$ tale che $g(\alpha) = \alpha$. Si supponga che per un particolare $x_0 \in [a,b]$ la successione $\{x_k\}$ definita da $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, converga ad α in modo sublineare. Dire se è possibile determinare un numero reale ω tale che, posto $G(x) = (g(x)+\omega x)/(1+\omega)$ le successioni $\{y_k\}_k$ generate dalla relazione $y_{k+1} = G(y_k)$, $k = 0, 1, \ldots$, a partire da $y_0 \in [\alpha-\rho, \alpha+\rho]$ per un opportuno $\rho > 0$ convergano ad α in modo superlineare. Nei casi in cui è possibile, si studi l'ordine di convergenza del metodo definito da $G(x)$. Nel caso non sia possibile, si consideri la funzione $G(x) = (g(g(x)) + \omega_1 g(x) + \omega_2 x)/(1 + \omega_1 + \omega_2)$ e si svolga una analoga analisi.

Riferimenti bibliografici

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

[2] D.A. Bini, V. Pan *Polynomial and Matrix Computations*, Birkhäuser, 1994.

[3] Wikipedia: Equazione cubica http://it.wikipedia.org/wiki/Equazione_cubica

[4] Wikipedia: Equazione quartica http://it.wikipedia.org/wiki/Equazione_quartica

[5] Girolamo Cardano http://it.wikipedia.org/wiki/Girolamo_Cardano

[6] Wikipedia: Niccolò Tartaglia http://it.wikipedia.org/wiki/Niccol%C3%B2_Tartaglia

[7] Wikipedia: Scipione del Ferro http://it.wikipedia.org/wiki/Scipione_Dal_Ferro

[8] Teoria di Galois http://it.wikipedia.org/wiki/Teoria_di_Galois

[9] Octave http://www.gnu.org/software/octave/

[10] Wikipedia: Algebra Computazionale http://it.wikipedia.org/wiki/Sistema_di_algebra_computazionale

[11] Wikipedia: Base di Groebner http://it.wikipedia.org/wiki/Base_di_Gr%C3%B6bner

[12] Wikipedia: MPSolve http://en.wikipedia.org/wiki/MPSolve

Il problema dell'interpolazione

Dario A. Bini, Università di Pisa

26 novembre 2019

Sommario

Questo modulo didattico contiene risultati relativi al problema dell'interpolazione di funzioni

1 Introduzione

In alcune situazioni si incontra il problema di dover approssimare il valore che una certa funzione $f(x) : [a, b] \to \mathbb{R}$ assume in un punto assegnato $\xi \in [a, b]$ avendo a disposizione i valori che la funzione assume in un insieme di $n + 1$ punti $x_0, x_1, \ldots, x_n \in [a, b]$. Cioè, date le coppie (x_i, y_i), con $y_i = f(x_i)$ per $i = 0, \ldots, n$ e dato $\xi \in [a, b]$ si vuole approssimare in qualche modo $f(\xi)$. Questo problema è abbastanza naturale quando si deve tracciare il grafico continuo di una funzione che si conosce solo in alcuni punti, o quando si devono calcolare funzioni particolari, tipo le funzioni elementari quali le funzioni trigonometriche i cui valori si conoscono in alcuni punti. In altre applicazioni, in cui $f(x)$ è una funzione a valori in $\mathbb{R}^2$ caso che noi non trattiamo, viene assegnato un oggetto continuo (a tratti), tipo la firma di una persona o un carattere tipografico di una stampante laser e vogliamo memorizzarlo attraverso un numero finito di valori numerici in modo che sia poi facilmente ricostruibile in tutti i suoi punti o manipolabile.

Problemi di questo tipo possono essere trattati mediante l'interpolazione

2 L'interpolazione lineare

Siano $\varphi_0(x), \ldots, \varphi_n(x)$ funzioni assegnate definite su $[a, b]$ a valori in $\mathbb{R}$ linearmente indipendenti. Siano (x_i, y_i) per $i = 0, \ldots, n$, dei valori assegnati in modo che $x_i \in [a, b]$ e $x_i \neq x_j$ per $i \neq j$. Il *problema dell'interpolazione lineare* consiste nel determinare dei coefficienti $a_0, a_1, \ldots, a_n \in \mathbb{R}$ tali che la funzione $f(x) = \sum_{i=0}^{n} a_i \varphi_i(x)$ soddisfi le condizioni

$$f(x_i) = y_i, \quad i = 0, \ldots, n. \tag{1}$$

Le condizioni (1) sono dette *condizioni di interpolazione*. I punti $x_0, \ldots, x_n$ sono detti *nodi* dell'interpolazione.

Si possono avere vari tipi di interpolazione a seconda di come sono scelte le funzioni $\varphi_i(x)$. Ad esempio, si parla di interpolazione polinomiale quando le $\varphi_i(x)$ generano lo spazio dei polinomi di grado al più n, si parla di interpolazione trigonometrica quando le funzioni $\varphi_i(x)$ vengono scelte nell'insieme $\{\sin((i+1)x),\ \cos(ix) :\quad i = 0, 1, \ldots, m\}$. Un'altro tipo di interpolazione molto importante nelle applicazioni è l'interpolazione spline in cui le funzioni $\varphi_i(x)$ sono polinomiali a tratti relativamente ai sottointervalli $[x_i, x_{i+1}]$ per $i = 0, \ldots, n-1$, dove i nodi sono stati ordinati in modo che $a \leq x_0 < x_1 < \cdots < x_n \leq b$.

3 Interpolazione polinomiale

Una forma abbastanza naturale ed elementare di interpolazione è quella che si sottiene scegliendo $\varphi_i(x) = x^i$. In questo caso si parla di interpolazione polinomiale fatta sulla base dei monomi.

Si può osservare che in generale la condizione di interpolazione (1) si riduce al sistema lineare

$$\sum_{j=0}^{n} \varphi_j(x_i) a_j = y_i, \quad i = 0, \ldots, n. \tag{2}$$

Nel caso in cui $\varphi_i(x) = x^i$, questo sistema prende la forma

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{3}$$

La matrice del sistema, indicata con V_n, viene detta *matrice di Vandermonde*. Il sistema (3) ci dice che un problema di interpolazione polinomiale sulla base dei monomi è ricondotto alla risoluzione di un sistema con matrice di Vandermonde. Vale il seguente utile risultato

Teorema 1 *Per la matrice di Vandermonde V_n costruita sui nodi $x_0, \ldots, x_n$ vale*

$$\det V_n = \prod_{0 \leq j < i \leq n} (x_i - x_j). \tag{4}$$

Dim. Possiamo limitarci a considerare il caso in cui $x_i \neq x_j$ per $i \neq j$. Infatti se $x_i = x_j$ per qualche $i \neq j$ allora $\det V_n = 0$ avendo V_n due righe uguali, inoltre la produttoria in (4) sarebbe nulla essendo nullo almeno un fattore. Procediamo per induzione su n. Per $n = 1$ vale $\det V_1 = x_1 - x_0$ e la tesi è vera. Supponiamo la tesi valida per $n-1$ e dimostriamola per n. Per fare questo consideriamo la matrice $V_n(\lambda)$ ottenuta sostituendo l'ultima riga di V_n con $(1, \lambda, \lambda^2, \ldots, \lambda^n)$.

Calcoliamo $\det V_n(\lambda)$ sviluppandolo lungo l'ultima riga con la regola di Laplace. Si ottiene allora un polinomio in λ di grado n in cui il coefficiente di λ^n è il determinante della matrice ottenuta togliendo ultima riga e ultima colonna a $V_n(\lambda)$. Questa matrice coincide con V_{n-1}, matrice di Vandermonde costruita sui nodi $x_0, \ldots, x_{n-1}$, e per l'ipotesi induttiva il suo determinante è $\prod_{0 \le j < i \le n-1}(x_i - x_j) \neq 0$. Per cui si può scrivere $\det V_n(\lambda) = \det V_{n-1} p(\lambda)$ con $p(\lambda)$ polinomio con coefficiente di λ^n uguale a 1. Poiché $\det V_n(x_i) = 0$ per $i = 0, \ldots, n-1$, avendo $V_n(\lambda)$ in questo caso due righe uguali, risulta $p(x_i) = 0$ per $i = 0, \ldots, n-1$ e quindi $p(x) = \prod_{i=0}^{n-1}(x - x_i)$. Ne segue che

$$\begin{aligned}\det V_n &= \det V_n(x_n) = \det V_{n-1} p(x_n) \\ &= \prod_{0 \le j < i \le n-1} (x_i - x_j) \prod_{i=0}^{n-1} (x_n - x_i) = \prod_{0 \le j < i \le n} (x_i - x_j).\end{aligned}$$

□

Segue da questo risultato che se i nodi x_i sono a due a due distinti, così come abbiamo supposto, allora la matrice di Vandermonde è invertibile ed esiste unica la soluzione del problema dell'interpolazione polinomiale. Il polinomio $p(x)$ che verifica le condizioni di interpolazione viene detto *polinomio di interpolazione*. I suoi coefficienti, nella base dei monomi, possono essere calcolati semplicemente risolvendo un sistema di equazioni lineari, quindi con costo computazionale di $(2/3)n^3$ operazioni aritmetiche. È stato sviluppato in letteratura un algorimto per risolvere sistemi con matrici di Vandermonde in $O(n^2)$ operazioni aritmetiche. L'algoritmo va sotto il nome di algoritmo di Bjorck-Pereyra. Sono stati sviluppati altri algoritmi che eseguono il calcolo della soluzione in $O(n \log^2 n)$ operazioni. Per maggior dettagli si veda il libro [2].

È stato dimostrato da Walter Gautschi e Gabriele Inglese [3] che la matrice di Vandermonde con nodi reali positivi ha un numero di condizionamento esponenziale nel grado n. Ciò rende il problema dell'interpolazione polinomiale nella base dei monomi numericamente intrattabile. Per questo è conveniente cambiare approccio.

4 Interpolazione polinomiale nella base di Lagrange

Consideriamo una base di polinomi diversa da quella dei monomi. Definiamo allora

$$L_i(x) = \prod_{j=0,\, j \neq i}^{n} (x - x_j) / \prod_{j=0,\, j \neq i}^{n} (x_i - x_j), \quad i = 0, 1, \ldots, n$$

la base dei *polinomi di Lagrange*. Si osservi che, in base alla definizione, vale $L_i(x_i) = 1$ mentre $L_i(x_j) = 0$ se $i \neq j$. Scegliendo quindi $\varphi_i(x) = L_i(x)$ il sistema (2) ha come matrice la matrice identica e il polinomio di interpolazione

si lascia scrivere come

$$p(x) = \sum_{i=0}^{n} y_i L_i(x). \tag{5}$$

L'espressione (5) viene detta *polinomio di interpolazione di Lagrange*.

L'espressione (5) permette di calcolare il valore di $p(x)$ in un punto in modo efficiente. Infatti vale

$$p(x) = \prod_{i=0}^{n}(x - x_i) \sum_{i=0}^{n} \frac{y_i/\theta_i}{x - x_i}, \quad \theta_i = \prod_{j=0\, j\neq i}^{n} (x_i - x_j). \tag{6}$$

Per cui il calcolo del valore di $p(x)$ mediante la (5) costa $O(n^2)$ operazioni aritmetiche. Non solo, ma il costo del calcolo del valore di $p(x)$ in un nuovo punto $x = \eta$ dopo aver calcolato $p(x)$ costa solamente $O(n)$ operazioni, visto che i valori delle quantità θ_i sono disponibili perché già calcolati in precedenza.

Un altro vantaggio computazionale sta nel fatto che, se dovessimo aggiungere un nuovo nodo x_{n+1} e calcolare il valore del nuovo polinomio di interpolazione p_{n+1} mediante la (5) in un punto ξ, ci basterebbe calcolare i nuovi valori dei θ_i aggiornando i valori precedenti mediante $O(n)$ operazioni aritmetiche. Anche il calcolo del valore di $p_{n+1}(x)$ in un punto costerebbe $O(n)$ operazioni.

Si possono introdurre altre basi di polinomi per rappresentare il polinomio di interpolazione. Ad esempio la base $\varphi_0(x) = 1$, $\varphi_i(x) = (x - x_0) \cdots (x - x_{i-1})$ per $i = 1, \ldots, n$, detta *base di Newton*, porta al polinomio di interpolazione di Newton

$$a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0) \cdots (x - x_{n-1}),$$

dove i coefficienti a_i, detti differenze divise, sono ricavabili risolvendo il sistema (1) che in questo caso è triangolare e con una struttura molto particolare.

5 Resto dell'interpolazione

Se $f(x)$ è una funzione sufficientemente regolare è possibile dare una espressione esplicita al *resto dell'interpolazione* definito come

$$r_n(x) = f(x) - p_n(x)$$

dove $p_n(x)$ è il polinomio di interpolazione di $f(x)$ relativo ai nodi $x_0 < x_1 < \cdots < x_n$. Vale infatti il seguente

Teorema 2 *Sia $f(x) \in C^{n+1}[a, b]$ e $p(x)$ il polinomio di interpolazione di $f(x)$ relativo ai nodi $a \leq x_0 < x_1 < \cdots < x_n \leq b$. Per ogni $x \in [a, b]$ esiste $\xi \in (a, b)$ tale che*

$$r_n(x) = \prod_{i=0}^{n}(x - x_i) \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Dim. Se x coincide con uno dei nodi allora la formula è banalmente vera essendo $r_n(x) = 0$ e $\prod_{i=0}^{n}(x - x_i) = 0$. Se x è diverso da ogni x_i allora $\prod_{i=0}^{n}(x - x_i) \neq 0$ e possiamo considerare questa funzione ausiliaria

$$g(y) = r_n(y) - \prod_{i=0}^{n}(y - x_i)\frac{r_n(x)}{\prod_{i=0}^{n}(x - x_i)}$$

in cui y è la variabile mentre x è il valore che abbiamo fissato. La funzione $g(y)$ è derivabile $n + 1$ volte con continuità. Infatti $r_n(y)$ lo è come pure il secondo addendo che è un polinomio in y. Inoltre $g(y)$ si annulla in tutti i nodi x_i e anche per $y = x$. Poiché $g(y)$ si annulla in $n + 2$ punti in $[a, b]$ allora la sua derivata prima si annulla in $n + 1$ punti in (a, b), la sua derivata seconda si annulla in n punti, e procedendo in questo modo si può concludere che la derivata $(n + 1)$-esima si annulla in un punto $\xi \in (a, b)$. Vale cioè

$$0 = g^{(n+1)}(\xi) = r_n^{(n+1)}(\xi) - (n + 1)!\frac{r_n(x)}{\prod_{i=0}^{n}(x - x_i)}. \tag{7}$$

Inoltre, poiché $p_n(x)$ è un polinomio di grado al più n la sua derivata $(n+1)$-esima è nulla e vale quindi $r_n^{(n+1)}(\xi) = f^{(n+1)}(\xi)$ che assieme alla (7) dà la tesi. □

L'espressione che abbiamo dato del resto evidenzia che, per una funzione in cui la derivata $(n+1)$-esima non cambia segno, il resto cambia segno ogni volta che la variabile x oltrepassa un nodo. Dal punto di vista grafico questo comportamento è deprecabile poiché il grafico di $p(x)$ ha un andamento ondeggiante sul grafico della funzione $f(x)$ che aumenta maggiormente con l'aumentare del numero dei nodi di interpolazione. Anche se l'approssimazione numerica può essere accurata, al punto di vista estetico il risultato diventa sgradevole. Per questo motivo nei problemi di CAGD si preferisce usare delle basi diverse che non presentano questo inconveniente quali le funzioni spline che però non trattiamo in questo articolo.

Supponiamo di fissare una successione di $(n + 1)$-uple di punti in $[a, b]$ cioè $x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}$ per $n = 1, 2, \ldots$ Viene spontaneo chiedersi se la corrispondente successione di polinomi di interpolazione $\{p_n(x)\}$ ottenuta interpolando $f(x)$ nei punti $x_i^{(n)}, f(x_i^{(n)})$, $i = 0, \ldots, n$, converge puntualmente o uniformemente ad $f(x)$. Purtroppo esistono esempi di funzioni apparentemente innocue e di successioni di $(n + 1)$-uple di nodi per cui la successione $p_n(x)$ non converge neppure puntualmente a $f(x)$. Un esempio classico è dato dalla *funzione di Runge* definita da $f(x) = 1/(1 + x^2)$ sull'intervallo $[-5, 5]$. Questa è una funzione di classe $C^\infty([a, b])$, però con la successione di nodi $x_i^{(n)} = -5 + 10i/n$, $i = 0, \ldots, n$ la successione dei polinomi di interpolazione $p_n(x)$ non converge nemmeno puntualmente a $f(x)$. Se però si restringe la funzione ad un intervallo più piccolo, ad esempio $[-0.2, 0.2]$ e prendendo ancora nodi equispaziati, si ottiene la convergenza uniforme di $p_n(x)$ a $f(x)$.

Si possono dare condizioni sufficienti affinché per ogni scelta dei nodi ci sia convergenza uniforme di $p_n(x)$ a $f(x)$. Il seguente teorema, di cui non si riporta la dimostrazione, fornisce una condizione sufficiente.

Teorema 3 *Se la funzione $f(x)[a,b] \to \mathbb{R}$ è la restrizione di una funzione analitica definita sull'insieme $\Omega \subset \mathbb{C}$ tale che*

$$\Omega = \{z \in \mathbb{C} : \quad \exists\, x \in [a,b],\, |z-x| \le b-a\}$$

allora per ogni successione di nodi $x_i^{(n)}$ il polinomio di interpolazione $p_n(x)$ converge uniformemente a $f(x)$.

L'insieme Ω è assimilabile al recinto di un cane la cui catena, di lunghezza $b-a$, ha un estremo libero di scorrere lungo il segmento $[a,b]$.

È evidente che la funzione di Runge ha un polo nei punti $x = \mathtt{i}$ e $x = -\mathtt{i}$. Per cui le ipotesi del teorema del "recinto del cane" sono soddisfatte se $[a,b] = [-b,b]$ è tale che $2b < 1$.

Il prossimo risultato, che riportiamo senza dimostrazione, ci dice che data una funzione $f(x)$ possiamo sempre trovare una successione di nodi che garantisce la convergenza uniforme.

Teorema 4 *Se la funzione $f(x)$ è continua sull'intervallo $[a,b]$, allora esiste una scelta della successione di nodi per cui $p_n(x)$ converge uniformemente ad $f(x)$.*

6 Esercizi

Esercizio 1 Sia n un intero positivo e siano $x_i \in (0,1)$, $i = 0,\dots,n$ tali che $x_i \neq x_j$ per $i \neq j$. Si definiscano inoltre le funzioni $g_i(x) = x_i + x^{-i}$, $i = 0,\dots,n$.
a) Si consideri la matrice $(n+1)\times(n+1)$ $V(x) = (v_{i,j})_{i,j=0,n}$ tale che $v_{i,j} = g_j(x_i)$ per $i \neq n$ e $v_{n,j} = g_j(x)$, con $x > 0$. Si dimostri che se $\det V(\xi) = 0$ allora $\det V(\xi^{-1}) = 0$ e che $\det V(\xi) = 0$, $\xi \in (0,1)$ se e solo se $\xi \in \{x_0,\dots,x_{n-1}\}$.
b) Si dimostri che il problema di interpolazione nello spazio delle funzioni $g_i(x)$, $i = 0,\dots,n$ relativo ai nodi $x_i \in (0,1)$ ha una e una sola soluzione, cioè, per ogni $(n+1)$-upla $y_0,\dots,y_n \in \mathbb{R}$ esistono unici $a0,\dots,a_n$ tali che, posto $\varphi(x) = \sum_{j=0}^{n} a_j g_j(x)$, vale $\varphi(x_i) = y_i$, $i = 0,\dots,n$.
c) Si dia unespressione per $\varphi(x)$ e un algoritmo per il calcolo di $\varphi(x)$ in un punto ξ di costo $O(n^2)$.

Soluzione.
a) Si osserva che $V(\xi) = V(\xi^{-1})$, dunque se $\det V(\xi) = 0$ allora $\det V(\xi^{-1}) = 0$. Indichiamo con $V_n(x)$ la matrice di dimensione $(n+1)\times(n+1)$. Dimostriamo per induzione su n che $\det V_n(\xi) = 0$, $\xi \in (0,1)$ se e solo se $\xi \in \{x_0,\dots,x_{n-1}\}$. Per $n = 1$ è una verifica diretta. Supponiamo la proprietà valga per $n-1$ e la dimostriamo per n. Se $\xi \in \{x_0,\dots,x_{n-1}\}$ allora $\det V_n(\xi) = 0$ perché $V_n(\xi)$ ha due righe uguali. Proviamo il viceversa. Osserviamo che

$$V_n(x) = \begin{bmatrix} I_{n-1} & 0 \\ 0 & x^{-n} \end{bmatrix} \begin{bmatrix} V_{n-1}(x_{n-1}) & a \\ b^T & x^{2n}+1 \end{bmatrix} = D_n(x) W_n(x)$$

dove gli elementi del vettore b sono polinomi di grado minore di $2n$. Poiché per ipotesi induttiva $\det V_{n-1}(x_{n-1}) \neq 0$, sviluppando il determinante della matrice $W_n(x)$ risptto all'ultima riga, si trova che $\det W_n(x)$ è un polinomio di grado $2n$. D'altra parte $\det W_n(\xi) = 0$ se e solo se $\det V_n(\xi) = 0$, e $\det V_n(\xi) = 0$ se $\xi \in \{x_0, \ldots, x_{n-1}\}$ oppure $\xi \in \{x_0^{-1}, \ldots, x_{n-1}^{-1}\}$. Dunque gli $\xi \in \{x_0, \ldots, x_{n-1}\}$ sono gli unici possibili zeri di $\det V_n(\xi)$ nell'intervallo $(0, 1)$.
b) Si verifica che il vettore $a = (a_i)_{i=0,\ldots,n-1}$ risolve il sistema di equazioni lineari $V(x_n)a = y$, dove $y = (y_i)_{i=0,\ldots,n-1}$, e tale sistema ha un'unica soluzione perché $\det V(x_n) \neq 0$.
c) (Traccia) Osserviamo che $\varphi(x) = x^{-n}p_{2n}(x)$, dove $p_{2n}(x) = \sum_{j=0}^{n} a_j(x^{n+j} + x^{n-j}$ ha grado al più $2n$. Poiché $y_j = \varphi(x_j) = \varphi(x_j^{-1})$, per $j = 0, \ldots, n$, allora $p_{2n}(x)$ è il polinomio di grado al più $2n$ tale che $p_{2n}(x_j) = x_j^n y_j$ e $p_{2n}(x_j^{-1}) = x_j^{-n} y_j$, per $j = 0, \ldots, n$. Usando la formula (6) possiamo dare una espressione a $p_{2n}(x)$ e quindi a $\varphi(x) = x^{-n}p_{2n}(x)$, da cui possiamo derivare un metodo per il calcolo di $\varphi(x)$ con costo $O(n^2)$. □

Esercizio 2 Siano $x_1, \ldots x_{2n}$ numeri reali non nulli a due a due distinti e siano a, b due numeri reali tali che $a < x_i < b$, $i = 1, \ldots, 2n$. Si definiscano le matrici $2n \times 2n$, $W = (w_{i,j})$ e $V = (v_{i,j})$ tali che $w_{i,j} = x_i^{j-n}$, $v_{i,j} = x_i^{j-1}$, $i, j = 1, \ldots, 2n$.
a) Mettendo in relazione $\det V$ con $\det W$ e sfruttando le proprietà delle matrici di Vandermonde si dimostri che W è non singolare.
b) Assegnati i numeri reali $y_1, \ldots, y_{2n}$ dire se il problema di determinare i coefficienti α_i, $i = -n + 1, \ldots, n$ tali che la funzione $s(x) = \sum_{j=-n+1}^{n} \alpha_j x^j$ soddisfi le condizioni di interpolazione $s(x_i) = y_i$, $i = 1, \ldots, 2n$, ha soluzione e, nel caso, dire se è unica.
c) Se $x_1, \ldots, x_{2n}$ coincidono con le radici $2n$-esime dell'unità, individuare un metodo per il calcolo degli α_i che impieghi $O(n \log n)$ operazioni aritmetiche, dove n è potenza intera di 2.
d) Se gli y_i del punto b) sono i valori che una funzione $f(x) \in C^{2n}([a, b])$ assume nei punti x_i, cioè $f(x_i) = y_i$, $i = 1, \ldots, 2n$, dare una espressione esplicita per l'errore di interpolazione $r(x) = f(x) - s(x)$ per $x \in [a, b]$.

Soluzione.
a) Si osserva che $w_{i,j} = x_i^{-n+1} v_{i,j}$, $i, j = 1, \ldots, 2n$, quindi

$$W = DV, \quad D = \begin{bmatrix} x_1^{-n+1} & & 0 \\ & \ddots & \\ 0 & & x_{2n}^{-n+1} \end{bmatrix},$$

da cui $\det V = 0$ se e solo se $\det W = 0$, ma V è non singolare perché è una matrice di Vandermonde e i punti x_i sono due a due distinti.
b) Si verifica che il vettore $\alpha = (\alpha_i)_{i=-n+1,\ldots,n}$ risolve il sistema di equazioni lineari $W\alpha = y$, dove $y = (y_i)_{i=1,\ldots,2n}$, e tale sistema ha un'unica soluzione perché $\det W \neq 0$.

c) (Traccia) Il sistema $W\alpha = y$ viene risolto risolvendo il sistema $V\alpha = D^{-1}y$. Questo ultimo sistema può essere risolto in $O(n \log n)$ operazioni aritmetiche mediante FFT.
d) (Traccia) Osserviamo che $s(x) = x^{-n+1}p(x)$, dove $p(x) = \sum_{j=0}^{2n-1} \alpha_{j-n+1}x^j$. Definiamo $\tilde{f}(x) = x^{n-1}f(x)$ e osserviamo che $p(x_i) = \tilde{f}(x_i) = x_i^{n-1}y_i$, cioè $p(x)$ è il polinomio di grado al più $2n-1$ tale che interpola $\tilde{f}(x)$ nei nodi x_i, $i = 1, \ldots, 2n$. Dal Teorema 2 sappiamo che $\tilde{r}(x) = \tilde{f}(x) - p(x) = \prod_{i=1}^{2n}(x - x_i)\frac{\tilde{f}^{(2n)}(\xi)}{(2n)!}$ dove $\xi = \xi(x) \in (a, b)$, da cui ricaviamo $r(x) = x^{-n+1}\tilde{r}(x)$. □

Riferimenti bibliografici

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

[2] D.A. Bini, V. Pan *Polynomial and Matrix Computations*, Birkhäuser, 1994.

[3] Walter Gautschi, Gabriele Inglese, Lower bounds for the condition number of Vandermonde matrices, Numer. Math. 52, 3, 241-250, DOI: 10.1007/BF01398878

La trasformata discreta di Fourier e la FFT

Dario A. Bini, Università di Pisa

7 febbraio 2020

Sommario

Questo modulo didattico contiene risultati relativi alla trasformata discreta di Fourier e agli algoritmi per il suo calcolo, in particolare gli algoritmi FFT di Cooley-Tukey e di Sande-Tukey. Si dà un cenno ad alcune applicazioni.

La trasformata discreta di Fourier è una operazione che permette di rappresentare vettori nel campo complesso in una base speciale, la base di Fourier. Da questa particolare rappresentazione si possono ricavare informazioni interessanti del vettore originario che sono di grande utilità in molte applicazioni.

Una caratteristica di fondamentale importanza di questa trasformazione è che il suo costo computazionale è estremamente basso. Infatti per effettuare il cambio di base di un vettore di n componenti bastano circa $\frac{3}{2}n\log_2 n$ operazioni purché n sia una potenza intera di 2. Inoltre la trasformazione è ben condizionata e gli algoritmi veloci per il suo calcolo, chiamati Fast Fourier Transform o FFT, sono numericamente stabili. Per queste caratteristiche particolarmente favorevoli, la trasformata discreta di Fourier trova numerose impieghi in diversi campi della matematica e delle sue applicazioni.

In questo articolo introduciamo questa trasformazione nel contesto dell'interpolazione, descriviamo due algoritmi FFT per il suo calcolo e mostriamo alcune applicazioni.

1 Interpolazione alle radici n-esime dell'unità

Dato un intero positivo n, consideriamo il polinomio $x^n - 1$. Le radici di questo polinomio vengono chiamate le *radici n-esime dell'unità*. Queste radici possono essere caratterizzate in termini di una radice speciale:

$$\omega_n = \cos\frac{2\pi}{n} + \mathtt{i}\sin\frac{2\pi}{n}$$

dove con $\mathtt{i}$ abbiamo denotato l'unità immaginaria tale che $\mathtt{i}^2 = -1$. Infatti, i numeri complessi

$$\omega_n^j = \cos\frac{2\pi j}{n} + \mathtt{i}\sin\frac{2\pi j}{n}, \quad j = 0, 1, \ldots, n-1,$$

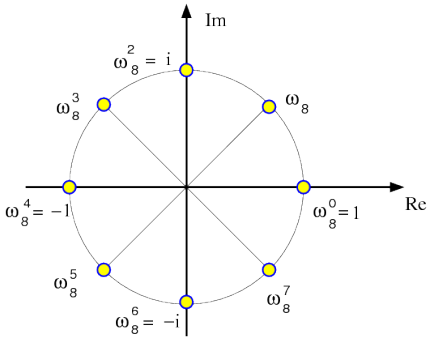


Figura 1: Radici ottave dell'unià

sono tutte e sole le radici n-esime dell'unità: esse sono n, essendo valori tutti distinti, e la loro potenza n-esima fa 1 dato che

$$(\omega_n^j)^n = (\omega_n^n)^j = 1^j = 1.$$

La figura 1 mostra graficamente nel piano complesso le radici ottave dell'unità

Ogni radice n-esima dell'unità ω tale che l'insieme $\{\omega^j : \quad j = 0, 1, \ldots, n-1\}$ costituisce l'insieme di tutte le radici n-esime viene detta *primitiva*. Si può verificare che se ζ è una radice n-esima primitiva tale che k e n siano primi tra loro, allora anche ζ^k è radice primitiva. In particolare, se n è un numero primo, ogni potenza non nulla di ω_n è radice primitiva. La radice ω_n come pure la sua coniugata sono radici primitive.

L'insieme delle radici n-esime forma un gruppo moltiplicativo essendo

$$\omega_n^p \omega_n^q = \omega_n^{p+q \bmod n}.$$

Consideriamo adesso il problema dell'interpolazione nel campo complesso e scegliamo come nodi le radici n-esime dell'unità che chiameremo *nodi di Fourier*. Poniamo cioè

$$x_j = \omega_n^j, \quad j = 0, 1, \ldots, n-1.$$

Dato allora il vettore $y = (y_0, y_1, \ldots, y_{n-1})$ vogliamo calcolare i coefficienti $(z_0, z_1, \ldots, z_{n-1})$ del polinomio $p(t) = \sum_{j=0}^{n-1} z_j t^j$ tale che $p(x_i) = y_i$ per $i = 0, 1, \ldots, n-1$. La condizione di interpolazione appena scritta conduce al sistema con matrice di Vandermonde

$$Vz = y, \quad V = (v_{i,j}), \quad v_{i,j} = x_i^j, \quad i, j = 0, \ldots, n-1,$$

dove, per la specificità dei nodi, la matrice V prende la forma $V = (\omega_n^{ij})_{i,j=0,\ldots,n-1}$. Questa matrice speciale di Vandermonde la indichiamo con Ω_n e la chiamiamo *matrice di Fourier*. Vale quindi

$$\Omega_n = (\omega_n^{ij \bmod n}).$$

Ad esempio, per $n = 7$ vale

$$\Omega_7 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \omega_7 & \omega_7^2 & \omega_7^3 & \omega_7^4 & \omega_7^5 & \omega_7^6 \\ 1 & \omega_7^2 & \omega_7^4 & \omega_7^6 & \omega_7 & \omega_7^3 & \omega_7^5 \\ 1 & \omega_7^3 & \omega_7^6 & \omega_7^2 & \omega_7^5 & \omega_7 & \omega_7^4 \\ 1 & \omega_7^4 & \omega_7 & \omega_7^5 & \omega_7^2 & \omega_7^6 & \omega_7^3 \\ 1 & \omega_7^5 & \omega_7^3 & \omega_7 & \omega_7^6 & \omega_7^4 & \omega_7^2 \\ 1 & \omega_7^6 & \omega_7^5 & \omega_7^4 & \omega_7^3 & \omega_7^2 & \omega_7 \end{bmatrix}.$$

La matrice di Fourier gode di proprietà particolarmente interessanti. Premettiamo il seguente risultato di ortogonalità delle radici n-esime dell'unità:

Lemma 1 *Le radici n-esime dell'unità sono tali che*

$$\sum_{i=0}^{n-1} \omega_n^{ki} = \begin{cases} n & \text{se } k = 0 \bmod n \\ 0 & \text{se } k \neq 0 \bmod n \end{cases}$$

Dim.
Dalla identità

$$1 - x^n = (1 - x)(1 + x + x^2 + \cdots + x^{n-1})$$

ponendo $x = \omega_n^r$, deduciamo che, se $r \neq 0 \bmod n$ allora $1-x \neq 0$ mentre $1-x^n = 0$ e quindi, poiché siamo su un campo, ne segue $1+\omega_n^r+\omega_n^{2r}+\cdots+\omega_n^{(n-1)r} = 0$. D'altro canto, se $r = 0 \bmod n$ allora l'espressione $\sum_{k=0}^{n-1} \omega_n^{rk}$ si trasforma in una somma di n addendi uguali a 1 che dà n. □

Teorema 1 *La matrice Ω_n è tale che*

- $\Omega_n = \Omega_n^T$
- $\Omega_n^H \Omega_n = nI$, *dove* Ω_n^H *è la trasposta coniugata di* Ω_n
- $\Omega_n^2 = n\Pi_n$, *dove* Π_n *è la matrice di permutazione corrispondente alla permutazione* $\pi_0 = 0$, $\pi_j = n - j$, $j = 1, \ldots, n-1$.

Dim. La matrice Ω_n è chiaramente simmetrica essendo $\omega_n^{ij} = \omega_n^{ji}$. Per dimostrare il secondo punto sull'ortogonalità delle colonne di Ω_n dobbiamo fare vedere che il prodotto tra l'i-esima riga di Ω_n^H e la j-esima colonna di Ω_n vale zero se $i \neq j$ e vale n se $i = j$, cioè

$$\sum_{k=0}^{n-1} \bar{\omega}_n^{ik} \omega_n^{jk} = \begin{cases} n & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

La sommatoria nella precedente espressione si riduce a $\sum_{k=0}^{n-1} \omega_n^{rk}$ dove $r = j - i$. Per cui la tesi segue dal Lemma 1.

Per quanto riguarda la terza espressione ci si comporta in modo analogo. Infatti l'elemento di posto (i, j) di Ω_n^2 è dato da

$$\sum_{k=0}^{n-1} \omega_n^{ik} \omega_n^{jk} = \sum_{k=0}^{n-1} \omega_n^{k(i+j)}$$

e, per il lemma 1, vale zero se $i + j \neq 0 \bmod n$, vale n se $i + j = 0 \bmod n$. □

Si osservi che la permutazione individuata dalla matrice Π se applicata alla n-upla delle radici n-esime, listate in senso antiorario a partire da 1, ci fornisce la lista delle radici n-esime listate in senso orario a partire sempre da 1.

Il risultato precedente ci permette di scrivere l'inversa di Ω_n in forme computazionalmente convenienti. Vale infatti il seguente

Corollario 1 *Per la matrice inversa di Ω_n valgono le seguenti espressioni*

$$\Omega_n^{-1} = \frac{1}{n}\Omega_n^H, \qquad \Omega_n^{-1} = \frac{1}{n}\Omega_n\Pi_n, \quad \Omega_n^{-1} = \frac{1}{n}\Pi_n\Omega_n.$$

Si osserva in particolare che, per risolvere il problema dell'interpolazione sui nodi di Fourier, non c'è bisogno di risolvere nessun sistema lineare poiché l'inversa della matrice di Vandermonde è disponibile gratuitamente. Per cui, dato il vettore y, è possibile calcolare il vettore dei coefficienti z eseguendo un prodotto matrice vettore, infatti vale

$$z = \frac{1}{n}\Omega_n^H y, \quad z = \frac{1}{n}\Omega_n\Pi_n y, \quad z = \frac{1}{n}\Pi_n\Omega_n y.$$

Assumendo di avere a disposizione le radici n-esime dell'unità, il costo del calcolo dei coefficienti del polinomio di interpolazione è di al più n^2 moltiplicazioni e $n^2 - n$ addizioni. Vediamo tra poco che si può fare molto meglio con gli algoritmi FFT.

Un'altra conseguenza molto importante delle proprietà di ortogonalità delle radici n-esime è che la matrice $F_n = \frac{1}{\sqrt{n}}\Omega_n$ è unitaria, cioè $F_n^H F_n = I$, e quindi $\|F_n\|_2 = \|F_n^H\|_2 = 1$. Conseguentemente il suo numero di condizionamento in norma 2, cioè $\mu_2 = \|F_n\|_2\|F_n^H\|_2 = 1$. Ciò implica che anche la matrice Ω_n ha condizionamento in norma 2 uguale a 1, infatti essa differisce da F_n per una costante moltiplicativa. Quindi, diversamente dal caso dei nodi reali, il problema dell'interpolazione ai nodi di Fourier è ben condizionato.

La trasformazione lineare che associa il vettore y al vettore z è detta *trasformata discreta di Fourier* e si denota $z = \mathrm{DFT}_n(y)$, la trasformazione lineare che associa z a y viene detta *trasformata discreta inversa di Fourier* e si denota $y = \mathrm{IDFT}_n(z)$. Quindi in sintesi:

$$\begin{array}{ll} \mathrm{DFT}_n(y) := z = \frac{1}{n}\Omega_n^H y & Trasformata\ discreta\ di\ Fourier \\ \mathrm{IDFT}_n(z) := y = \Omega_n z & Trasformata\ discreta\ inversa\ di\ Fourier \end{array}$$

La trasformata discreta inversa di Fourier associa ai coefficienti del polinomio i valori che esso assume nelle radici n-esime dell'unità. Quindi risolve un *problema di valutazione*. La trasformata discreta di Fourier associa ai valori che un polinomio assume nei nodi di Fourier i suoi coefficienti. Quindi risolve un *problema di interpolazione*.

Un'altra osservazione interessante è che, poiché $y = \Omega_n z$, le componenti di z sono i coefficienti di rappresentazione del vettore y nella base di $\mathbb{C}^n$ data dalle colonne di Ω_n. Questa base la chiamiamo *base di Fourier*. Possiamo quindi interpretare la DFT e la IDFT come cambiamenti di base: dalla base di Fourier alla base canonica e dalla base canonica alla base di Fourier.

In letteratura ci sono diverse definizioni di DFT e IDFT a seconda del contesto in cui viene usata. Ad esempio, in alcuni casi si preferisce definire le due trasformazioni in modo unitario normalizzando entrambe col fattore $1/\sqrt{n}$. In Octave il comando `x = fft(y);` fornisce il valore di $n\mathrm{DFT}_n(y) = \Omega_n^H y$, mentre il comando `y = ifft(x);` dà il valore di $\frac{1}{n}\mathrm{IDFT}_n(x) = \frac{1}{n}\Omega_n x$.

2 Gli algoritmi FFT

Così come è stata definita la DFT e la IDFT si possono calcolare eseguendo un prodotto tra la matrice Ω_n e un vettore. Il costo computazionale è quindi di n^2 moltiplicazioni e n^2-n addizioni. È possibile però utilizzare la speciale struttura di Ω_n per poter calcolare la DFT e la IDFT con un costo sostanzialmente più basso.

Consideriamo il caso della IDFT. Dati i valori di $z_0, z_1, \ldots, z_{n-1}$ e date le radici n-esime dell'unità vogliamo calcolare i valori di

$$y_i = \sum_{j=0}^{n-1} \omega_n^{ij} z_j, \quad i = 0, 1, \ldots, n-1\,. \tag{1}$$

Consideriamo il caso particolare in cui n è una potenza intera di 2, cioè $n = 2^q$.

L'idea che si usa per costruire un algoritmo efficiente per il calcolo di (1) si basa sulla strategia del *divide et impera*. Cioè il calcolo di una IDFT su n nodi lo riconduciamo al calcolo di due IDFT su $\frac{n}{2}$ nodi. Per far questo riscriviamo la (1) separando gli addendi che hanno indice pari da quelli che hanno indice dispari. Otteniamo allora

$$y_i = \sum_{j=0}^{\frac{n}{2}-1} \omega_n^{2ij} z_{2j} + \sum_{j=0}^{\frac{n}{2}-1} \omega_n^{i(2j+1)} z_{2j+1}, \quad i = 0, \ldots, n-1.$$

Adesso osserviamo che $\omega_n^2 = \omega_{\frac{n}{2}}$ per cui $\omega_n^{2ij} = \omega_{\frac{n}{2}}^{ij}$ e $\omega_n^{i(2j+1)} = \omega_n^i \omega_{\frac{n}{2}}^{ij}$. Possiamo allora riscrivere la precedente espressione come

$$y_i = \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j} + \omega_n^i \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j+1}, \quad i = 0, \ldots, n-1\,. \tag{2}$$

Se limitiamo il calcolo alle prime $\frac{n}{2}$ componenti la (2) ci dice che

$$(y_0, \ldots, y_{\frac{n}{2}-1})^T = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) + \mathrm{Diag}(1, \omega_n, \ldots, \omega_n^{\frac{n}{2}-1})\mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}}), \quad (3)$$

dove abbiamo indicato rispettivamente con z_{pari} e z_{dispari} la parte del vettore z costituita dalle componenti pari e la parte costituita dalle componenti dispari. Abbiamo cioè espresso la prima metà di y in termini di due trasformate discrete di ordine la metà.

La seconda metà del vettore y, cioè quella costituita dalle componenti $y_{\frac{n}{2}}, \ldots y_{n-1}$, si ottiene mettendo al posto dell'indice i nella (2) il valore $\frac{n}{2} + i$. Poiché $\omega_{\frac{n}{2}}^{\frac{n}{2}+i} = \omega_{\frac{n}{2}}^i$, e $\omega_n^{\frac{n}{2}+i} = -\omega_n^i$, si ottiene quindi

$$y_{\frac{n}{2}+i} = \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j} - \omega_n^i \sum_{j=0}^{\frac{n}{2}-1} \omega_{\frac{n}{2}}^{ij} z_{2j+1}, \quad i = 0, \ldots, \frac{n}{2} - 1.$$

Cioè in termini di $\mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}})$ e $\mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}})$ vale

$$(y_{\frac{n}{2}}, \ldots, y_{n-1})^T = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}}) - \mathrm{Diag}(1, \omega_n, \ldots, \omega_n^{\frac{n}{2}-1})\mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}}). \quad (4)$$

Mettendo insieme la (3) e la (4) possiamo rappresentare il vettore y in termini delle IDFT della parte pari e della parte dispari di z. Più precisamente si procede come segue

- calcolare $w^{(1)} = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{pari}})$ e $w^{(2)} = \mathrm{IDFT}_{\frac{n}{2}}(z_{\mathrm{dispari}})$;
- moltiplicare $w^{(2)}$ per i valori $1, \omega_n, \ldots, \omega_n^{\frac{n}{2}-1}$, cioè calcolare
$$w^{(3)} = \mathrm{Diag}(1, \omega_n, \ldots, \omega_n^{\frac{n}{2}-1})w^{(2)};$$
- infine calcolare i vettori $y^{(1)} = w^{(1)} + w^{(3)}$, $y^{(2)} = w^{(1)} - w^{(3)}$, e porre $y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}$.

Questo procedimento comporta il calcolo di due trasformate discrete inverse di ordine la metà, $\frac{n}{2}$ moltiplicazioni, $\frac{n}{2}$ addizioni e $\frac{n}{2}$ sottrazioni.

Poiché $n = 2^q$ è potenza di 2, possiamo allora ripetere questa strategia per calcolare le due trasformate di ordine $\frac{n}{2}$ mediante quattro trasformate di ordine $n/4$ e procedere ricorsivamente in questo modo finché non si arriva a trasformate di ordine 1 che non richiedono alcuna operazione. Indicando con $c(n)$ il costo computazionale per il calcolo di una IDFT di ordine n con questo metodo abbiamo la relazione

$$c(n) = 2c(\frac{n}{2}) + \frac{n}{2}M + nA$$

dove M indica "moltiplicazioni" e A indica "addizioni/sottrazioni". Poiché $c(1) = 0$ si può dimostrare induttivamente che

$$c(n) = (\frac{n}{2}M + nA)q = (\frac{n}{2}M + nA)\log_2 n.$$

L'algoritmo per il calcolo della IDFT che si ottiene in questo modo è noto come algoritmo di Cooley e Tukey.

Il calcolo della DFT si realizza utilizzando il teorema 1 e ha il costo aggiuntivo del calcolo di n divisioni per n.

2.1 Note computazionali

È possibile dare una interpretazione in termini di matrici dell'algoritmo per il calcolo della IDFT descritto sopra. Per questo introduciamo la matrice di permutazione pari-dispari P_n tale che

$$P_n z = \begin{bmatrix} z_{\text{pari}} \\ z_{\text{dispari}} \end{bmatrix}$$

e denotiamo $m = \frac{n}{2}$.

Dalla descrizione data nell'algoritmo 1 deduciamo che

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} = \begin{bmatrix} I_m & I_m \\ I_m & -I_m \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & D_m \end{bmatrix} \begin{bmatrix} \Omega_m & 0 \\ 0 & \Omega_m \end{bmatrix} P_n z$$

dove

$$D_m = \text{Diag}(1, \omega_n, \ldots, \omega_n^{m-1}).$$

Questo ci permette di dire che la matrice di Fourier Ω_n si lascia fattorizzare nel modo seguente

$$\Omega_n = \begin{bmatrix} I_m & I_m \\ I_m & -I_m \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & D_m \end{bmatrix} \begin{bmatrix} \Omega_m & 0 \\ 0 & \Omega_m \end{bmatrix} P_n. \tag{5}$$

Questa fattorizzazione assume una forma più interessante se la riscriviamo in termini del prodotto di Kronecker $\otimes$ che definiamo nel modo seguente. Siano A e B matrici di dimensioni rispettivamente $k \times k$ e $h \times h$. Definiamo $C = A \otimes B$ la matrice $hk \times hk$ tale che.

$$C = \begin{bmatrix} a_{1,1}B & a_{1,2}B & \ldots & a_{1,k}B \\ a_{2,1}B & a_{2,2}B & \ldots & a_{2,k}B \\ \vdots & \ldots & \ldots & \vdots \\ a_{k,1}B & a_{k,2}B & \ldots & a_{k,k}B \end{bmatrix}$$

Poiché $\Omega_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, la (5) prende la forma

$$\Omega_n = (\Omega_2 \otimes I_m)\text{Diag}(I, D_m)(I_2 \otimes \Omega_m)P_n, \quad D_m = \text{Diag}(1, \omega_n, \ldots, \omega_n^{m-1}). \tag{6}$$

Questa rappresentazione matriciale permette di dedurre altre interessanti proprietà. Ad esempio, poichè $\Omega_n = \Omega_n^T$ la (6) implica che

$$\Omega_n = P_n^T (I_2 \otimes \Omega_m) \mathrm{Diag}(I, D_m)(\Omega_2 \otimes I_m). \tag{7}$$

L'applicazione ricorsiva della fattorizzazione (7) conduce ad un altro algoritmo per il calcolo della IDFT noto come algoritmo di Sande e Tukey, che ha la stessa complessità dell'algoritmo di Cooley e Tukey ma svolge le operazioni aritmetiche in modo diverso. Ad esempio, mentre nell'algoritmo di Cooley e Tukey la permutazione viene fatta all'inizio, nel metodo di Sande e Tukey la permutazione viene fatta alla fine. Per questa caratteristica il primo metodo viene detto algoritmo in base 2 di decimazione in tempo, e il secondo come algoritmo in base 2 di decimazione in frequenza. Questa terminologia proviene dalle applicazioni ingegneristiche della DFT.

Non è complicato dimostrare che se $n = rs$, con r, s interi positivi, allora vale

$$\Omega_n = (\Omega_r \otimes I_s)\mathrm{Diag}(I, D_s, D_s^2, \ldots, D_s^{r-1})(I_r \otimes \Omega_s)P_n^{(r)},$$

dove $D_s = \mathrm{Diag}(1, \omega_n, \ldots, \omega_n^{s-1})$, e dove $P_n^{(r)}$ è la matrice di permutazione associata alla permutazione che mette in testa gli indici congrui a 0 modulo r seguiti dagli indici congrui a 1 modulo r, e così via fino agli indici congrui a $r-1$ modulo r.

Questa fattorizzazione permette di costruire algoritmi efficienti per il calcolo della IDFT nel caso in cui n non sia una potenza di 2 ma sia altamente fattorizzabile.

L'implementazione degli algoritmi di Cooley e Tukey e di Sande e Tukey fatta con un linguaggio di programmazione che permette la ricorsività è molto semplice. È però possibile dare una implementazione più efficiente "in place" che evita di usare esplicitamente la ricorsività. Infatti, applicando ricorsivamente la (6) si ottiene che nella parte destra della fattorizzazione finale si accumulano tutte le permutazioni di tipo pari-dispari fatte ai vari livelli dell'algoritmo. Cooley e Tukey hanno dimostrato che la composizione di tali permutazioni è data dalla permutazione *bit reversal* cioè la permutazione che associa i a j se la rappresentazione binaria di i su q bit, dove $n = 2^q$, coincide con la rappresentazione di j con i bit listati in ordine inverso. Cioè se $i = \sum_{k=0}^{q-1} 2^k i_k$ e $j = \sum_{k=0}^{q-1} 2^k j_k$, con $i_k, j_k \in \{0, 1\}$, sono le rappresentazioni binarie di i e di j allora $j_k = i_{q-k}$ per $k = 0, \ldots, q-1$. Una volta eseguita questa permutazione il resto del calcolo consiste nello svolgere semplici operazioni aritmetiche tra componenti contigue senza la necessità di dover applicare nessuna permutazione.

Maggiori informazioni a riguardo si possono trovare in questo link di Wikipedia.

Esistono algoritmi per il calcolo della DFT che non richiedono nessuna proprietà particolare dell'intero n ed hanno un costo computazionale limitato da $\alpha n \log_2 n$. Però il valore di α è molto maggiore di $3/2$ per cui il loro interesse è principalmente teorico.

Gli algoritmi di trasformata veloce sono numericamente stabili. Ad esempio, per quanto riguarda l'algoritmo di Cooley e Tukey vale il seguente risultato la cui dimostrazione è riportata in [8].

Teorema 2 *Per $n = 2^q$, q intero positivo, sia $\hat{y}$ il valore effettivamente calcolato al posto di $y = IDFT_n(z)$ con l'algoritmo di Cooley e Tukey in aritmetica floating point con precisione u dove i valori $\hat{\omega}_n^i$ effettivamente usati al posto delle radici n-esime ω_n^i sono tali che $|\hat{\omega}_n^i - \omega_n^i| \leq u$. Allora*

$$\frac{\|y - \hat{y}\|_2}{\|y\|_2} \leq \frac{\gamma u \log_2 n}{1 - \gamma u \log_2 n}$$

dove $\gamma = 1 + (\sqrt{2} + u)\frac{4}{1-4u} = 1 + 4\sqrt{2} + O(u)$.

Ne segue che l'errore algoritmico relativo, misurato in norma 2, cresce col logaritmo dell'ordine della trasformata.

2.2 FFT in Matlab

Gli algoritmi FFT per vettori di lunghezza arbitraria sono implementati in Matlab nelle function `fft` e `ifft` con una piccola differenza notazionale. Il vettore ottenuto col comando `v = fft(u)` è tale che $v_i = \sum_{j=0}^{n-1} \omega_n^{-ij} u_j$ mentre il vettore ottenuto con `u = ifft(v)` è tale che $u_i = \frac{1}{n}\sum_{j=0}^{n-1} \omega_n^{ij} v_j$. Cioè la normalizzazione $\frac{1}{n}$ è utilizzata nella definizione della trasformata inversa, e, come radice primitiva, è considerata ω_n^{-1}, cioè l'ordinamento delle radici n-esime scelto da Matlab è quello orario.

3 DFT su altri campi

Se il vettore z di cui calcolare la IDFT è reale, allora si verifica che il vettore $y = \mathrm{IDFT}_n(z)$ ha le componenti y_0 e $y_{\frac{n}{2}}$ (se n è pari) reali, mentre le rimanenti componenti sono tali che $y_j = \bar{y}_{n-j}$, $j = 1, \ldots, \frac{n}{2}$. Questa proprietà può essere usata per ridurre leggermente il costo computazionale degli algoritmi veloci descritti nel caso complesso.

La definizione di DFT e IDFT e gli algoritmi di trasformata veloce possono essere dati anche su campi finiti o, sotto condizioni aggiuntive, su anelli in cui esiste una radice n-esima dell'unità. Ad esempio l'insieme $\mathbb{Z}_{17}$ costituisce un campo e l'elemento 2 è una radice ottava primitiva dell'unità. Infatti le sue potenze (modulo 17) sono date nell'ordine da

$$2, 4, 8, 16, 15, 13, 9, 1.$$

Mentre il numero 3 è una radice 16-ma primitiva; le sue potenze (modulo 17) sono date nell'ordine da

$$3, 9, 10, 13, 5, 15, 11, 16, 14, 8, 7, 4, 12, 2, 6, 1$$

È possibile costruire la matrice di Fourier Ω_{16} su $\mathbb{Z}_{17}$ e definire la DFT e la IDFT. Infatti tutte le elaborazioni che abbiamo fatto finora si applicano poiché abbiamo usato solamente le proprietà di campo, l'esistenza di una radice n-esima e la sua primitività. L'operazione di coniugazione di una radice n-esima va vista come calcolo del reciproco.

La situazione diventa più delicata sugli anelli. Ad esempio su $\mathbb{Z}_{16}$ il numero 3 è una radice quarta dell'unità, le sue potenze sono nell'ordine

$$3, 9, 11, 1$$

e le "coniugate" cioè i reciproci sono

$$11, 9, 3, 1$$

Però la matrice di Fourier Ω_4 e la sua "coniugata", date da

$$\Omega_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 11 \\ 1 & 9 & 1 & 9 \\ 1 & 11 & 9 & 3 \end{bmatrix}, \quad \Omega_4^H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 11 & 9 & 3 \\ 1 & 9 & 1 & 9 \\ 1 & 3 & 9 & 11 \end{bmatrix}$$

sono tali che

$$\Omega_4^H \Omega_4 = \begin{bmatrix} 4 & 8 & 4 & 8 \\ 8 & 4 & 8 & 4 \\ 4 & 8 & 4 & 8 \\ 8 & 4 & 8 & 4 \end{bmatrix} \neq 4I \bmod 16.$$

Infatti il lemma 1 non è più valido. La dimostrazione del lemma non vale più visto che si richiedeva che non esistessero divisori dello zero. Inoltre, nel nostro caso il valore $n = 4$ non ha reciproco modulo 16. Per cui, anche se il prodotto fosse nI, non si potrebbe scrivere l'inversa di Ω_n come $\frac{1}{n}\Omega_n^H$.

Per potere definire la DFT e la IDFT su un anello dobbiamo garantire l'esistenza di una radice n-esima ω_n che sia *principale* cioè sia tale che $\sum_{j=0}^{n-1} \omega_n^{ij} = 0$ per $i = 1, \ldots, n$, e che n sia coprimo con la caratteristica dell'anello.

4 Applicazioni

La trasformata discreta di Fourier ha applicazioni in numerosi campi. Diamo un breve cenno su alcuni esempi di applicazioni.

4.1 Aritmetica di polinomi

Supponiamo di aver assegnato i coefficienti di due polinomi $a(t)$ e $b(t)$ di grado p e definiamo il polinomio prodotto $c(t) = a(t)b(t)$. I coefficienti di $c(t)$ sono

legati ai coefficienti di $a(t)$ e $b(t)$ dalle relazioni

$$\begin{aligned}
&c_0 = a_0 b_0 \\
&c_1 = a_0 b_1 + a_1 b_0 \\
&\cdots \\
&c_i = \textstyle\sum_{r+s=i} a_r b_s \\
&\cdots \\
&c_{2p-1} = a_{p-1} b_p + a_p b_{p-1} \\
&c_p = a_p b_p
\end{aligned}$$

Il calcolo dei coefficienti di $c(t)$ svolto con le formule precedenti richiede $O(p^2)$ operazioni aritmetiche.

È possibile calcolare i coefficienti di $c(t)$ usando la FFT con un costo asintoticamente più basso nel grado dei fattori. Procediamo nel seguente modo

Algoritmo 1.

- sia n la più piccola potenza intera di 2 maggiore del grado di $c(t)$
- si calcola $y^{(a)} = (y_i^{(a)})$, $y_i^{(a)} = a(\omega_n^i)$, $i = 0, \ldots, n-1$, con una IDFT_n;
- si calcola $y^{(b)} = (y_i^{(b)})$, $y_i^{(b)} = b(\omega_n^i)$, $i = 0, \ldots, n-1$, con una IDFT_n;
- si calcola $c(\omega_n^i) = y_i^{(a)} y_i^{(b)}$, $i = 0, \ldots, n-1$ con n moltiplicazioni
- si interpolano i valori di $c(\omega_n^i)$ ottenuti al passo precedente e si ottengono i coefficienti di $c(t)$ mediante una DFT_n

L'algoritmo richiede il calcolo di 3 trasformate veloci di Fourier n moltiplicazioni per valutare $c(\omega_n^i)$ più n moltiplicazioni per $1/n$ nel calcolo della DFT. Il costo complessivo è dominato da $\frac{9}{2} n \log_2 n$ operazioni aritmetiche.

Gli algoritmi FFT possono essere usati anche per calcolare il reciproco modulo t^n di un polinomio $p(t)$, tale che $p_0 = p(0) \neq 0$, con costo $O(n \log n)$. Infatti si può dimostrare che la successione di polinomi $x^{(k)}(t)$ definiti da

$$\begin{aligned}
&x^{(k+1)}(t) = 2x^{(k)}(t) - (x^{(k)}(t))^2 p(t) \bmod t^{2^{k+1}}, \quad k = 0, 1, \ldots, \lceil \log_2 n \rceil \\
&x^{(0)}(t) = 1/p_0
\end{aligned}$$

è tale che $x^{(k)}(t)p(t) = 1 \bmod t^{2^k}$. Infatti gli elementi di questa successione, si ottengono applicando formalmente il metodo di Newton all'equazione $x(t)^{-1} - p(t) = 0$ e la proprietà descritta sopra segue dalla convergenza quadratica del metodo di Newton. Si può verificare che utilizzando il metodo FFT per il calcolo dei prodotti di polinomi nell'iterazione di Newton, il costo complessivo per il calcolo del reciproco modulo t^n rimane dell'ordine di $n \log_2 n$. Per maggiori dettagli si veda [2].

4.2 Aritmetica degli interi

Le proprietà della DFT possono essere usate per costruire algoritmi veloci per la moltiplicazione di interi. Si considerino due interi positivi a e b ciascuno rappresentato in base B con al più p cifre. Per comodità possiamo assumere che B sia la familiare base 10; nelle implementazioni su computer B è la più realistica base 2. Per calcolare le cifre del prodotto $c = ab$ generalmente andiamo a calcolare tutti i prodotti possibili tra le cifre di a e quelle di b svolgendo p^2 moltiplicazioni più circa altrettante addizioni, inclusi i riporti, per ottenere il risultato. Ad esempio per moltiplicare 123 con 257 si procede generalmente così

```
    123 x
    257 =
---------
    861
   615
  246
---------
  31611
```

Siamo quindi abituati a svolgere più di p^2 operazioni elementari tra cifre per calcolare il prodotto di interi di p cifre. Questo ci creerebbe problemi se gli interi da moltiplicare avessero decine di cifre. Anche in un computer questo metodo può richiedere tempi di calcolo elevati se i numeri da moltiplicare hanno milioni di cifre come può capitare in certe applicazioni legate alla crittoanalisi.

L'uso della DFT permette di accelerare significativamente questo calcolo. Per vedere questo scriviamo gli interi a, b e c nella loro rappresentazione in base

$$a = \sum_{i=0}^{p-1} a_i B^i, \quad b = \sum_{i=0}^{p-1} b_i B^i, \quad c = \sum_{i=0}^{2p-1} c_i B^i$$

e associamo ad essi i polinomi

$$a(t) = \sum_{i=0}^{p-1} a_i t^i, \quad b(t) = \sum_{i=0}^{p-1} b_i t^i, \quad c(t) = \sum_{i=0}^{2p-1} c_i t^i.$$

In questo modo vale $a = a(B)$, $b = b(B)$, $c = c(B)$.

Osserviamo ora che il polinomio prodotto $\hat{c}(t) = b(t)a(t)$ assume anch'esso il valore dell'intero c nel punto B, cioè vale $\hat{c}(B) = c$. Però i suoi coefficienti $\hat{c}_i$ sono in generale diversi da quelli di $c(t)$. Infatti, i coefficienti c_i di $c(t)$ sono singole cifre mentre i coefficienti $\hat{c}_i$ di $\hat{c}(t)$ sono somma di p termini, ciascuno è il prodotto di due cifre. Il loro valore non può superare quindi $(B-1)^2 p$ per cui in base B saranno rappresentati da non più di $2 + \log_B p$ cifre.

I coefficienti di $\hat{c}(t)$ possono essere calcolati con l'algoritmo 1 eseguendo $O(p \log p)$ operazioni aritmetiche. Poiché i valori che vanno calcolati sono rappresentabili con $O(\log p)$ cifre, e la trasformata veloce di Fourier è numericamente stabile, per calcolare i coefficienti di $\hat{c}(t)$ è sufficiente eseguire i calcoli

con una aritmetica floating point dotata di $O(\log p)$ cifre. Se questa aritmetica viene implementata col metodo standard per moltiplicare numeri floating point (o interi) il costo computazionale di ogni moltiplicazione diventa $O(\log^2 p)$ operazioni elementari tra cifre. In questo modo il costo totale, espresso in termini di operazioni elementari tra cifre, per il calcolo dei coefficienti $\hat{c}_i$, dato dal numero di operazioni aritmetiche per il costo di ciascuna operazione aritmetica, diventa $O((p \log p)(\log^2 p)) = O(p \log^3 p)$.

Una volta che i coefficienti $\hat{c}_i$ sono stati calcolati possiamo recuperare le cifre c_i dalle cifre dei numeri $\hat{c}_i$ con $O(p \log p)$ addizioni. Lasciamo i dettagli di questo calcolo al lettore.

In questo modo si riesce a calcolare i coefficienti del prodotto di due interi con un numero di operazioni elementari tra cifre dell'ordine di $p \log^3 p$, quindi inferiore al costo p^2 dell'algoritmo standard di moltiplicazione. È possibile ridurre ulteriormente il costo se la moltiplicazione di numeri di $O(\log p)$ cifre, anziché calcolarla con l'algoritmo standard di costo quadratico, viene calcolata ricorsivamente con l'algoritmo che stiamo descrivendo.

Un algoritmo più efficiente, basato sulla DFT ambientata in anelli finiti, è *l'algoritmo di Schoenhage-Strassen*. La complessità di questo algoritmo è di $O(p \log p \log \log p)$ operazioni elementari tra bit. L'algoritmo è stato migliorato nel 2007 da Martin Fürer. L'algoritmo di Fürer ha una complessità di $O(p \log p \, 2^{\log^* p})$ operazioni elementari dove $\log^* p$ è uguale al numero di volte in cui $\log_2$ compare nell'espressione $\log_2 \log_2 \cdots \log_2 p$ affinché il risultato sia compreso tra 0 e 1.

Sebbene l'algoritmo di Fürer sia asintoticamente più veloce dell'algoritmo di Schönhage - Strassen, in pratica diventa più efficiente solo per valori di p estremamente elevati. Per cui nella pratica non viene utilizzato nelle implementazione dei sistemi di calcolo in multiprecisione.

Maggiori informazioni sul metodo di Fürer si trovano nell'articolo "Faster integer multiplication". Informazioni sui metodi per la moltiplicazione di interi si trovano nella pagina di Wikipedia "Multiplication Algorithm"

Recentemente è uscito un articolo di David Harvey, Joris van der Hoeven [7] in cui si dimostra che bastano $O(n \log n)$ operazioni tra bit per calcolare il prodotto di interi dotati di n bit.

4.3 Interpolazione trigonometrica

Definiamo polinomio trigonometrico una espressione del tipo

$$F(x) = \begin{cases} \frac{\alpha_0}{2} + \sum_{j=1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx) & \text{se } n = 2m-1 \\ \frac{\alpha_0}{2} + \sum_{j=1}^{m-1} (\alpha_j \cos jx + \beta_j \sin jx) + \frac{\alpha_m}{2} \cos mx & \text{se } n = 2m \end{cases}$$

dove $\alpha_j, \beta_j \in \mathbb{R}$.

Il seguente risultato ci fornisce l'espressione del polinomio trigonometrico che interpola i valori $(x_i, y_i) \in \mathbb{R}^2$, $x_i = 2i\pi/n$ per $i = 0, \ldots, n-1$.

Teorema 3 *Sia* $z = (z_0, \ldots z_{n-1}) = \mathrm{DFT}_n(y)$*, dove* $y = (y_0, \ldots, y_{n-1})$*. Il polinomio trigonometrico* $F(x)$ *definito dai coefficienti*

$$\alpha_i = 2\mathrm{re}(z_i) = \frac{2}{n}\sum_{j=0}^{n-1} y_j \cos ix_j, \quad \beta_i = -2\mathrm{im}(z_i) = \frac{2}{n}\sum_{j=0}^{n-1} y_j \sin ix_j$$

è tale che $F(x_i) = y_i$*,* $x_i = 2i\pi/n$*,* $i = 0, \ldots, n-1$*. Inoltre tale polinomio trigonometrico è unico.*

La dimostrazione di questo risultato è una semplice applicazione di note formule trigonometriche e non viene riportata.

Dal punto di vista computazionale i coefficienti del polinomio trigonometrico che interpola i valori (x_i, y_i) si calcolano con circa $\frac{2}{3}n\log_2 n$ operazioni aritmetiche se n è potenza di 2. Infatti, dati i valori di $y = (y_i)$ basta calcolare $z = \mathrm{DFT}_n(y)$ e ricavarne parte reale e immaginaria. Viceversa, i valori che un polinomio trigonometrico $F(x)$ assegnato in termini dei coefficienti α_i e β_i assume nei punti x_i si possono calcolare ancora con circa $\frac{3}{2}n\log_2 n$ operazioni aritmetiche. Infatti, dati i coefficienti α_i e β_i, basta calcolare i valori $z_i = 2(\alpha_i - \mathtt{i}\beta_i)$ e porre $y = \mathrm{IDFT}_n(z)$.

Questo fatto permette di effettuare operazioni di "filtraggio" di segnali e immagini digitali in modo efficiente.

Riferimenti bibliografici

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

[2] D.A. Bini, V. Pan *Polynomial and Matrix Computations*, Birkhäuser, 1994.

[3] CooleyTukey FFT algorithm, Wikipedia http://en.wikipedia.org/wiki/Cooley%E2%80%93Tukey_FFT_algorithm

[4] Fast Fourier transform, Wikipedia http://en.wikipedia.org/wiki/Fast_Fourier_transform

[5] Fürer's Algorithm, Wikipedia http://en.wikipedia.org/wiki/F%C3%BCrer%27s_algorithm

[6] Martin Fürer, Fast Integer Multiplication, SIAM J. Comput., 39(3), 2009, 9791005.

[7] David Harvey, Joris van der Hoeven, Integer multiplication in time O(n log n), 2019. hal-02070778. https://hal.archives-ouvertes.fr/hal-02070778/document

[8] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia 2002.

[9] Multiplication Algorithm, Wikipedia http://en.wikipedia.org/wiki/Multiplication_algorithm

[10] Schönhage-Strassen Algorithm, Wikipedia http://en.wikipedia.org/wiki/Sch%C3%B6nhage%E2%80%93Strassen_algorithm