Calcolo Numerico

Ahmad Shatti

2020-2021

Indice capitoli

1 Analisi dell'errore **3**
- 1.1 Rappresentazione in base . . . 3
- 1.2 Insieme dei numeri di macchina . . . 4
- 1.3 Aritmetica di macchina . . . 5
- 1.4 Operazioni di macchina . . . 6
- 1.5 Errore nel calcolo di una funzione razionale . . . 6
  - 1.5.1 Errore nel calcolo di una somma . . . 7
  - 1.5.2 Errore nel calcolo di una funzione . . . 8
- 1.6 Tecniche per l'analisi dell'errore . . . 8
- 1.7 Errore analitico . . . 10

2 I problemi dell'algebra lineare numerica **11**
- 2.1 Norme vettoriali . . . 11
- 2.2 Norme matriciali . . . 12
- 2.3 Condizionamento della risoluzione di sistemi lineari . . . 13
- 2.4 Teoremi di localizzazione per autovalori . . . 14

3 Metodi diretti per la risoluzione di sistemi lineari **17**
- 3.1 Sistemi triangolari . . . 17
- 3.2 Fattorizzazione LU . . . 17
- 3.3 Matrici elementari di Gauss . . . 19
- 3.4 Metodo di eliminazione gaussiana . . . 20
- 3.5 Tecniche di Pivoting . . . 21

4 Metodi iterativi per la risoluzione di sistemi lineari **22**
- 4.1 Generalità sui metodi iterativi . . . 22
- 4.2 Metodi di Jacobi e Gauss-Seidel . . . 25
- 4.3 Criteri di arresto . . . 25
- 4.4 Convergenza dei metodi di Jacobi e Gauss-Seidel . . . 26

5 Il metodo delle potenze **28**
- 5.1 Generalità sul metodo delle potenze . . . 28
- 5.2 Varianti del metodo delle potenze . . . 29
- 5.3 Algoritmi di ranking per motori di ricerca . . . 30
  - 5.3.1 Random Surfer Model . . . 30

6 Approssimazione degli zeri di una funzione **32**
- 6.1 Il metodo di bisezione . . . 32
- 6.2 Metodi di iterazione funzionale . . . 32
- 6.3 Ordine di convergenza . . . 34
- 6.4 Metodo delle tangenti . . . 34

Indice teoremi

1.1.1 Teorema (**Teorema di rappresentazione in base**) . . . 3
1.3.1 Teorema . . . 5
1.3.2 Teorema . . . 6
1.5.1 Teorema . . . 7
1.6.1 Teorema . . . 8

2.1.1 Teorema (**Equivalenza delle norme**) . . . 11
2.2.1 Teorema (**Compatibilità delle norme**) . . . 12
2.2.2 Teorema . . . 12
2.3.1 Teorema . . . 13
2.4.1 Teorema (**Teorema di Gershgorin**) . . . 14
2.4.2 Teorema (**Teorema di Hirsch**) . . . 14
2.4.3 Teorema (**Secondo teorema di Gershgorin**) . . . 15
2.4.4 Teorema . . . 16

3.2.1 Teorema (**Esistenza ed unicità della fattorizzazione LU**) . . . 18
3.5.1 Teorema . . . 21

4.1.1 Teorema (**Fa la cosa giusta**) . . . 22
4.1.2 Teorema (**Condizione sufficiente per la convergenza del metodo**) . . . 24
4.1.3 Teorema (**Condizione necessaria per la convergenza del metodo**) . . . 24
4.1.4 Teorema . . . 24
4.1.5 Teorema (**Condizione necessaria e sufficiente per la convergenza del metodo**) . . . 24
4.4.1 Teorema . . . 26

6.2.1 Teorema . . . 33
6.2.2 Teorema (**Teorema del punto fisso**) . . . 33
6.2.3 Teorema . . . 33
6.4.1 Teorema . . . 35
6.4.2 Teorema (**Teorema di convergenza in largo**) . . . 36

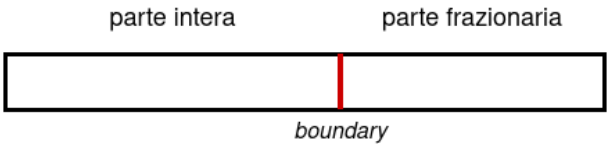
Chapter 1

Analisi dell'errore

1.1 Rappresentazione in base

Una macchina avendo memoria finita può rappresentare solo un numero finito di configurazioni, quindi si pone il problema di come associare ai numeri reali, che sono infiniti, dei numeri di macchina. Questo per garantire una buona rappresentazione di tutti numeri reali all'interno di un certo intervallo e anche di poter controllare gli errori che si generano nell'evolvere di un algoritmo. Sostanzialmente ci sono due modi:

- **Virgola fissa**, dato un certo numero reale lo si scrive all'interno di un registro. Il registro sarà diviso in due parti: *parte intera* e *parte frazionaria*, e il numero viene scomposto di conseguenza. Questo metodo si chiama *virgola fissa* perché si stabilisce dove è il *boundary* cioè in un numero a 32 bit si può scegliere 20 bit per la parte intera e 12 per la parte frazionaria. I sistemi che usano questa rappresentazione sono sistemi molto semplici come microprocessori o microcontrollori. Questa rappresentazione dei numeri è particolarmente vantaggiosa per operazioni come somma, sottrazione, moltiplicazione e divisione. Gli svantaggi sono: l'intervallo dei numeri rappresentabili è piccolo e vi è una bassa precisione dovuta alla parte frazionaria;



- **Virgola mobile** o *floating point*, rappresenta il modo più comune per fare l'*encoding* dei numeri reali. Si basa sul seguente teorema:

Teorema 1.1.1 (**Teorema di rappresentazione in base**). *Dato $x \in \mathbb{R}$ e $x \neq 0$, esistono e sono univocamente determinati:*

1. *un intero $p \in \mathbb{Z}$ detto esponente della rappresentazione*
2. *una successione di numeri naturali $\{d_i\}_{i \geq 1}$ con $d_1 \neq 0$, $0 \leq d_i \leq \beta - 1$ e d_i non definitivamente uguali a $\beta - 1$ dette cifre della rappresentazione*

tali per cui si ha:

$$x = sign(x)\beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}$$

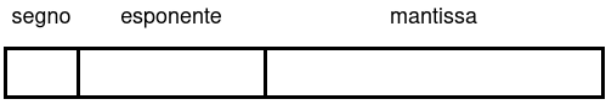
La quantità $\sum_{i=1}^{\infty} d_i \beta^{-i}$ è detta mantissa ed è compresa tra 0 ed 1 esclusi gli estremi.

La rappresentazione di un numero reale x si dice *rappresentazione normalizzata in virgola mobile* in quanto l'esponente p specifica di quanto la virgola si deve spostare. Esempi con $\beta = 10$:

$$123 = +10^3(0.123) = +10^3(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3})$$
$$0.00123 = +10^{-2}(0.123) = +10^{-2}(1 \cdot 10^{-1} + 2 \cdot 10^{-2} + 3 \cdot 10^{-3})$$

$$75.2 = +10^2(0.752) = +10^2(7 \cdot 10^{-1} + 5 \cdot 10^{-2} + 2 \cdot 10^{-3})$$

dove nel primo e secondo esempio i d_i sono $1, 2$ e 3 ed essendo in base 10 i numeri possono essere tra 0 e 9. La condizione $d_1 \neq 0$ serve per garantire l'unicità della rappresentazione, altrimenti 0.0123 lo si potrebbe scrivere come $10^{-1}(0.123)$ o ancora $10^0(0.0123)$ o $10^5(0.00000123)$ e così via. Anche il vincolo "d_i *non definitivamente uguale a* $\beta - 1$" serve per garantire l'unicità della rappresentazione infatti in base 10 il valore $0.\overline{9} = 1$. Nella rappresentazione floating point, il registro anziché essere diviso in parte intera e frazionaria, sarà composto da segno, esponente e mantissa:



I computer che implementano lo standard IEEE 754 prevedono la memorizzazione su registri lunghi 32 bit ripartiti come $1 + 8 + 23$ per la *singola precisione* e 64 bit come $1 + 11 + 52$ per la *doppia precisione*.

1.2 Insieme dei numeri di macchina

I registri di un calcolatore consentono la memorizzazione di un numero finito di cifre, l'**insieme dei numeri di macchina**, cioè l'insieme dei numeri reali esattamente rappresentabili in macchina ha cardinalità finita. Questo insieme dipende da $\mathcal{F}(\beta, t, m, M) = \{0\} \cup \{x \in \mathbb{R} : \ x = sign(x)\beta^p \sum_{i=1}^{t} d_i \beta^{-i}\}$ con $d_1 \neq 0, 0 \leq d_i \leq \beta - 1$ e $-m \leq p \leq M$, dove:

- β, base di rappresentazione che generalmente nei calcolatore è 2;
- t, il numero di cifre che servono per rappresentare la mantissa;
- m e M per definire il *range* dell'esponente.

Ovviamente se si aumenta t allora si aumenta la precisione di approssimazione di un numero. Lo zero viene messo a parte perché il teorema parla di ogni $x \neq 0$: lo zero viene rappresentato con una configurazione di bit uguali a zero. Il terzo vincolo: "d_i *non definitivamente uguale a* $\beta - 1$" sparisce perché il numero di cifre rappresentabili è limitato e quindi non ha senso di parlare di un numero periodico. Si osserva che:

- si definisce con Ω il più grande numero di macchina rappresentabile. Per esempio con $\mathcal{F}(10, 3, m, M)$ si ha $\Omega = 10^M(0.999)$ perché su tre cifre in base dieci questa è la mantissa più grande che si riesce a generare. In generale Ω vale $\Omega = \beta^M(1 - \beta^{-t})$
- per analogia ω è il numero *positivo* più piccolo rappresentabile e vale $\omega = \beta^{-m-1}$
- l'insieme dei numeri di macchina $\mathcal{F}(\beta, t, m, M)$ ha cardinalità $\#\mathcal{F} = 1 + 2(m + M + 1)(\beta^t - \beta^{t-1})$ dove:
  - 1 per la rappresentazione dello zero;
  - 2 per il segno: numeri positivi e negativi (l'insieme dei numeri di macchina è simmetrico: se x è rappresentabile allora lo è anche $-x$);
  - $m + M + 1$ è il numero di possibili esponenti;
  - $\beta^t - \beta^{t-1}$ è il numero di possibili mantisse.

  Per esempio $\#\mathcal{F}(2, 3, 1, 1) = 1 + 2(3)(2^3 - 2^2) = 1 + 6 \cdot 4 = 25$.

Come precedentemente detto i numeri floating point per la doppia precisione sono rappresentati in 64 bit (8 byte) e sono allocati nel seguente modo: 1 bit per il segno, 11 bit per l'esponente e i rimanenti 52 bit per la mantissa. Con base due si ha quindi $\mathcal{F}(2, 53, -1021, 1024)$ (53 e non 52 perché dato che la prima cifra deve essere diversa da zero allora sarà sicuramente uguale ad 1 in base 2). Le configurazioni possibili per l'esponente con 11 bit sono $2^{11} = 2048$, ma il numero di configurazioni possibili tra -1021 e 1024 sono $1021 + 1024 + 1 = 2046$ ($+1$ per lo zero). Quindi avanzano due configurazioni che vengono utilizzate per rappresentare delle situazioni particolari:

- con la configurazione di 11 bit tutti uguali a *zero* per l'esponente si rappresentano i **numeri denormalizzati** che sono del tipo $2^{-1022}(0d_2...d_{53})_2$
- con la configurazione di 11 bit tutti uguali a *uno* per l'esponente permette di identificare i **numeri speciali** cioè
  - $\pm\infty$, mantissa con tutti i bit uguali a zero, si verifica quando si ha un errore di *overflow*: il numero x che si vuole rappresentare non rientra nell'intervallo $[-\Omega, \Omega]$
  - NaN (*Not a Number*) rappresenta tutte le situazioni anomale come la divisione per zero, $\frac{\infty}{\infty}$, $\infty - \infty$, ...

1.3 Aritmetica di macchina

Quando il valore $x \in \mathbb{R}$ (con $x = sign(x)\beta^p \sum_{i=1}^{\infty} d_i\beta^{-i}$) si rappresenta su $\mathcal{F}$ significa che l'esponente rientra nel range $-m \leq p \leq M$ e x ha una rappresentazione finita a t cifre: $x = sign(x)\beta^p \sum_{i=1}^{t} d_i\beta^{-i}$ quindi x è un numero di macchina. L'altro scenario è quando x non si rappresenta su $\mathcal{F}$ ma con esponente corretto nel range $-m \leq p \leq M$, in questo caso si associa ad x un numero di macchina:

- tramite *troncamento* $trn(x)$, si prende la rappresentazione di x e la si tronca alla t-esima cifra;
- tramite *arrotondamento* $arr(x)$, è quello che generalmente ha più senso: si considera la cifra $(t+1)$-esima ed a seconda del suo valore lo si arrotonda alla cifra t-esima. Quindi è

$$arr(x) = \begin{cases} trn(x) & se\ d_{t+1} < \beta/2 \\ trn(x) + \beta^{p-t} & se\ d_{t+1} \geq \beta/2 \end{cases}$$

Rappresentare un numero reale x in macchina significa approssimare x con un numero $\hat{x} \in \mathcal{F}(\beta, t, m, M)$ commettendo un **errore relativo** ϵ_x di rappresentazione:

$$\epsilon_x = \frac{\hat{x} - x}{x} = \frac{\eta_x}{x} \qquad x \neq 0$$

La quantità $\eta_x = \hat{x} - x$ è detta **errore assoluto** di rappresentazione. Valgono i seguenti teoremi:

Teorema 1.3.1. *Se non si verifica overflow o underflow allora* $|trn(x) - x| < \beta^{p-t}$ *e* $|arr(x) - x| < \frac{1}{2}\beta^{p-t}$.

Dimostrazione:

- *se* $x = \Omega$ *allora* $trn(x) = arr(x) = x$ *e quindi* $trn(x) - x = arr(x) - x = 0$ *dunque è dimostrato* $0 = |trn(x) - x| < \beta^{p-t}$ *e* $0 = |arr(x) - x| < \frac{1}{2}\beta^{p-t}$
- *se* $0 < x < \Omega$ *con* $x = \beta^p \sum_{i=1}^{\infty} d_i\beta^{-i}$ *e dati* a *e* b *due numeri di macchina consecutivi* $a \leq x < b$*:*

$$a = \beta^p \sum_{i=1}^{t} d_i\beta^{-i} \qquad b = \beta^p(\sum_{i=1}^{t} d_i\beta^{-i} + \beta^{-t})$$

Risulta che $b - a = \beta^{p-t}$ *quindi nel caso del troncamento* $trn(x) = a$ *allora*

$$|trn(x) - x| = |a - x| < |a - b| = \beta^{p-t}$$

Nel caso dell'arrotondamento:

$$arr(x) = \begin{cases} a & se\ x < \frac{a+b}{2} \\ b & se\ x \geq \frac{a+b}{2} \end{cases}$$

allora $|arr(x) - x| \leq \frac{b-a}{2} = \frac{1}{2}\beta^{p-t}$ *e l'uguaglianza vale se e solo se* $x = \frac{a+b}{2}$

Teorema 1.3.2. *Se non si verificano situazioni di overflow o di underflow e sia* $x \in \mathbb{R}, \omega \leq |x| \leq \Omega$ *e* $x \neq 0$ *allora* $|\epsilon_x| = \left|\frac{\hat{x}-x}{x}\right| \leq u$ *dove:*

$$u = \begin{cases} \beta^{1-t} & se\ \hat{x} = trn(x) \\ \frac{1}{2}\beta^{1-t} & se\ \hat{x} = arr(x) \end{cases}$$

la quantità u *è detta* ***precisione di macchina****.*

Dimostrazione: dato $x > \beta^{p-1}$ *si ha* $|trn(x) - x| < \beta^{p-t}$ *segue che* $|\epsilon_x| = \left|\frac{trn(x)-x}{x}\right| < \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{1-t}$*.*
Analogamente per l'arrotondamento $|arr(x) - x| \leq \frac{1}{2}\beta^{p-t}$ *si ha* $|\epsilon_x| = \left|\frac{arr(x)-x}{x}\right| < \frac{1/2\beta^{p-t}}{\beta^{p-1}} = \frac{1}{2}\beta^{1-t}$*.*

La precisione di macchina per l'aritmetica IEEE, nel caso dell'arrotondamento è $u = \frac{1}{2}\beta^{1-t}$ che nel caso di una macchina $\beta = 2$ e $t = 53$ e quindi $u = 2^{-53} \simeq 10^{-16}$. A volte capiterà di indicare $\hat{x}$ come $float(x)$ che è la rappresentazione di x in macchina. Sapendo che $|\epsilon_x| < u$ si può scrivere $float(x) = \hat{x} = x(1 + \epsilon_x)$.

1.4 Operazioni di macchina

La somma di due numeri di macchina $x, y \in \mathcal{F}(\beta, t, m, M)$ è un numero che può non appartenere all'insieme $\mathcal{F}(\beta, t, m, M)$. Per esempio se $\beta = 10$ e $t = 2$ la somma di due numeri di macchina $x = 0.123\ 10^3$ e $y = 0.456\ 10^0$ è $x + y = 0.123456\ 10^3 \notin \mathcal{F}(10, 2, m, M)$ perché ha bisogno di 6 cifre per essere rappresentato e non $t = 3$. Si presenta dunque il problema di approssimare nel modo più conveniente possibile, il risultato di un'operazione aritmetica fra due numeri di macchina con un numero di macchina. Occorre perciò individuare le "operazioni aritmetiche" su $\mathcal{F}(\beta, t, m, M)$ ossia definire un'**aritmetica di macchina** che meglio approssimi l'aritmetica esatta dei numeri reali. Indicando con $\circledcirc$ l'operazione di macchina che approssima l'operazione esatta $\circ$, per tutti i numeri di macchina x e y per cui l'operazione non dia luogo a situazioni di underflow o di overflow, deve valere:

$$x \circledcirc y = float(x \circ y) = (x \circ y)(1 + \epsilon) \quad \textit{con } |\epsilon| < u$$

Nell'esempio precedente $x \oplus y = float(x + y) = (0.123)10^3$. L'errore relativo ϵ è detto **errore locale dell'operazione** e vale:

$$\epsilon = \frac{(x \circledcirc y) - (x \circ y)}{(x \circ y)}$$

Non tutte le proprietà algebriche delle operazioni nel campo reale sono soddisfatte dalle operazioni di macchina. Ad esempio in generale la somma e la moltiplicazione non sono associative, non vale la distributività della somma, la legge di cancellazione e la semplificazione tra due numeri.

1.5 Errore nel calcolo di una funzione razionale

Si introducono le seguenti definizioni:

- **errore inerente**, è la quantità:

  $$\epsilon_{in} = \frac{f(\hat{x}) - f(x)}{f(x)} \quad \textit{con } f(x) \neq 0$$

  è un errore dovuto alla rappresentazione del dato in macchina. È indipendente dall'algoritmo utilizzato per il calcolo di $f(x)$ e quindi per la risoluzione del problema matematico associato. Se un problema (calcolo di $f(x)$) ha un errore inerente elevato per un certo dato x si dice che il problema è *mal condizionato* intorno ad x;

- **errore algoritmico**, è la quantità

  $$\epsilon_{alg} = \frac{g(\hat{x}) - f(\hat{x})}{f(\hat{x})} \quad \textit{con } f(\hat{x}) \neq 0$$

  la funzione g dipende dall'algoritmo utilizzato per calcolare $f(x)$. In generale differenti algoritmi conducono a differenti errori algoritmici. Se l'errore algoritmico nel calcolo di $g(x)$ è basso si dice che l'algoritmo è *stabile*;

- **errore totale**, è la quantità

$$\epsilon_{tot} = \frac{g(\hat{x}) - f(x)}{f(x)} \quad \textit{con } f(x) \neq 0$$

misura la differenza relativa tra l'output atteso e l'output effettivamente calcolato.

Teorema 1.5.1. *L'errore totale vale:* $\epsilon_{tot} = \epsilon_{in} + \epsilon_{alg}$

Dimostrazione:

$$\epsilon_{tot} = \frac{g(\hat{x}) - f(x)}{f(x)} = \frac{g(\hat{x}) - f(\hat{x}) + f(\hat{x}) - f(x)}{f(x)} = \frac{g(\hat{x}) - f(\hat{x})}{f(x)} + \underbrace{\frac{f(\hat{x}) - f(x)}{f(x)}}_{\epsilon_{in}} =$$

$$\frac{\boldsymbol{g(\hat{x}) - f(\hat{x})}}{f(x)} \frac{f(\hat{x})}{\boldsymbol{f(\hat{x})}} + \epsilon_{in} = \boldsymbol{\epsilon_{alg}} \underbrace{\left(\frac{f(\hat{x})}{f(x)}\right)}_{\epsilon_{in}+1} + \epsilon_{in} = \epsilon_{alg}\epsilon_{in} + \epsilon_{alg} + \epsilon_{in} \doteq \epsilon_{alg} + \epsilon_{in}$$

dove con $\doteq$ *si intende che l'uguaglianza vale considerando le sole componenti lineari negli errori e trascurando le componenti di ordine superiore al primo (sinteticamente riferita come* ***analisi al primo ordine*** *dell'errore).*

Il teorema esprime il fatto che nel calcolo di una funzione razionale le due fonti di generazione degli errori individuate precedentemente forniscono contributi separati che possono essere analizzati indipendentemente. L'obiettivo dell'analisi numerica è pertanto quello di individuare algoritmi numericamente stabili per problemi ben condizionati.

1.5.1 Errore nel calcolo di una somma

Mettendo insieme tutti i tasselli: in caso in cui non vi sono situazioni di overflow e di underflow se si vuole calcolare la somma di due numeri reali a e b si deve fare:

1. rappresentare a in macchina e quindi si avrà $\hat{a}$ che è la rappresentazione in macchina di a. Stessa cosa per b, quindi si avrà $\hat{b}$;
2. ora si deve fare l'operazione di macchina: $\hat{a} \oplus \hat{b}$. Si studia ora l'errore che si genera:
   (a) come detto precedentemente $\hat{a} \oplus \hat{b} = float(\hat{a} + \hat{b}) = (\hat{a} + \hat{b})(1 + \epsilon)$ dove ϵ è l'*errore locale della somma*;
   (b) a questo punto riportiamo il valore di $\hat{a}$ e $\hat{b}$ cioè $\hat{a} = float(a) = a(1 + \epsilon_a)$ e $\hat{b} = float(b) = b(1 + \epsilon_b)$ dove $\epsilon_a = \frac{\hat{a}-a}{a}$ e $\epsilon_b = \frac{\hat{b}-b}{b}$ sono *errori di rappresentazione*;
   (c) ora riportiamo il calcolo in $\hat{a}\oplus b = (\hat{a}+\hat{b})(1+\epsilon)$ otteniamo che $\hat{a}\oplus\hat{b} = [a(1+\epsilon_a)+b(1+\epsilon_b)](1+\epsilon)$;
   (d) a questo punto si effettua la moltiplicazione: $\hat{a} \oplus \hat{b} = a(1 + \epsilon_a)(1 + \epsilon) + b(1 + \epsilon_b)(1 + \epsilon)$. Ora si effettua la moltiplicazione svolgendo un *analisi al primo ordine* che permette di trascurare le quantità superiore al primo (come $\epsilon_a \cdot \epsilon$ o $\epsilon_b \cdot \epsilon$). Dunque $\hat{a}\oplus\hat{b} = a(1+\epsilon_a)(1+\epsilon)+b(1+\epsilon_b)(1+\epsilon) \doteq a + a\epsilon_a + a\epsilon + b + b\epsilon_b + b\epsilon = a + b + a\epsilon_a + b\epsilon_b + (a + b)\epsilon$;
   (e) l'*errore totale* per definizione è uguale a $\frac{\hat{a}\oplus\hat{b}-(a+b)}{a+b}$ ora sostituendo $\hat{a} \oplus \hat{b}$ con il valore trovato nel punto precedente abbiamo:

$$\frac{a + b + a\epsilon_a + b\epsilon_b + (a + b)\epsilon - (a + b)}{a + b}$$

   semplificando , l'errore totale è $= \frac{a}{a+b}\epsilon_a + \frac{b}{a+b}\epsilon_b + \epsilon$. Come visto precedentemente dall'errore totale si possono individuare due componenti:
   - $\frac{a}{a+b}\epsilon_a + \frac{b}{a+b}\epsilon_b$ che dipende dall'errore di rappresentazione dei dati. Nel caso in cui a e b siano numeri di macchina allora $\epsilon_a = \epsilon_b = 0$. Se invece a e b sono dei numeri reali che non sono di macchina e quindi hanno un certo errore allora l'errore di rappresentazione si amplificherà a seconda dei coefficienti $\frac{a}{a+b}$ e $\frac{b}{a+b}$. Questa prima componente è l'*errore inerente*;

- $+\epsilon$ questo errore è dovuto all'operazione cioè è l'*errore algoritmico*.

La parte di errore inerente è una cosa insita nel problema che si sta calcolando infatti viene chiamato anche *errore inevitabile*, mentre, si può agire sull'errore algoritmico cambiando algoritmo.

1.5.2 Errore nel calcolo di una funzione

Si vuole calcolare $f(x) = x^2 - 1$ che lo si può calcolare come $x \to x^2 \to x^2 - 1$ o come $(x-1)(x+1)$ e quindi $x \to x-1$ poi $x \to x+1$ e infine fare la moltiplicazione tra le due quantità $x-1$ e $x+1$. Si vuole confrontare dal punto di vista dell'errore quale tra le due è la strategia migliore:

1. dato $x \in \mathbb{R}$ lo si rappresenta in macchina come $float(x) = \hat{x} = x(1+\epsilon_x)$;
2. come detto $f(x)$ può essere calcolata in due modi:

   I. $g_1(\hat{x}) = \hat{x} \otimes \hat{x} \ominus 1$ dove prima si svolge il prodotto $float(\hat{x} \otimes \hat{x}) = (\hat{x} \cdot \hat{x})(1+\epsilon_1)$ e poi la sottrazione. La sottrazione è una operazione un po' delicata: si definisce $\hat{z}$ come $\hat{z} = float(\hat{x} \otimes \hat{x})$ e ora si vuole calcolare $float(\hat{z} \ominus 1)$ che è uguale ad $float(\hat{z} \ominus 1) = (\hat{z}-1)(1+\epsilon_2)$. Ora si riscrive il tutto: $g_1(\hat{x}) = [(\hat{x} \cdot \hat{x})(1+\epsilon_1)-1](1+\epsilon_2)$ dove ϵ_i è l'errore locale dell'operazione i-esima. A questo punto si riscrive anche $\hat{x}$ e quindi $g_1(\hat{x}) = \{[x(1+\epsilon_x) \cdot x(1+\epsilon_x)](1+\epsilon_1) - 1\}(1+\epsilon_2)$. Ora si effettuano le varie moltiplicazioni trascurando i vari ϵ_x^2, $\epsilon_x\epsilon_1, \epsilon_x\epsilon_2$ e $\epsilon_1\epsilon_2$ per l'analisi del primo ordine ottenendo:

   $$g_1(\hat{x}) \doteq (x^2-1) + x^2(\epsilon_1 + 2\epsilon_x) + \epsilon_2(x^2-1)$$

   L'errore totale è $\frac{g_1(\hat{x})-f(x)}{f(x)}$ che nel nostro caso è:

   $$\frac{(x^2-1) + x^2(\epsilon_1+2\epsilon_x) + \epsilon_2(x^2-1) - (x^2-1)}{x^2-1} = \frac{(x^2-1)\epsilon_2 + x^2\epsilon_1 + 2x^2\epsilon_x}{x^2-1} = \frac{2x^2}{x^2-1}\epsilon_x + \frac{x^2}{x^2-1}\epsilon_1 + \epsilon_2$$

   II. $g_2(\hat{x}) = (\hat{x} \ominus 1) \otimes (\hat{x} \oplus 1)$ dove $(\hat{x} \ominus 1)$ viene riscritto come $[x(1+\epsilon_x)-1](1+\delta_1)$ mentre $(\hat{x} \oplus 1)$ come $[x(1+\epsilon_x)+1](1+\delta_2)$. Ora si effettua la moltiplicazione come $\{[x(1+\epsilon_x)-1](1+\delta_1) \cdot [x(1+\epsilon_x)+1](1+\delta_2)\}(1+\delta_3)$. Ora si svolgono le moltiplicazioni effettuando l'analisi del primo ordine ottenendo:

   $$g_2(\hat{x}) \doteq (x^2-1) + (x^2-1)\delta_1 + (x^2-1)\delta_2 + (x^2-1)\delta_3 + 2x^2\epsilon_x$$

   L'errore totale è $\frac{g_2(\hat{x})-f(x)}{f(x)}$ che nel nostro caso è

   $$\frac{(x^2-1) + (x^2-1)\delta_1 + (x^2-1)\delta_2 + (x^2-1)\delta_3 + 2x^2\epsilon_x - (x^2-1)}{x^2-1} = \frac{(x^2-1)\delta_1 + (x^2-1)\delta_2 + (x^2-1)\delta_3 + 2x^2\epsilon_x}{x^2-1} = \frac{2x^2}{x^2-1}\epsilon_x + \delta_1 + \delta_2 + \delta_3$$

3. ora si effettua il confronto tra l'errore totale di $g_1(x)$ e $g_2(x)$: si nota che hanno una componente in comune $\frac{2x^2}{x^2-1}\epsilon_x$ che corrisponde all'errore inerente e invece differiscono per l'errore algoritmico. Se si dovesse fare un confronto quale tra i due algoritmi è preferibile? Dipende da x: per $x \pm 1$ la quantità $\frac{x^2}{x^2-1}\epsilon_1$ tende all'infinito. Quindi il primo algoritmo non è *stabile* per $x \to \pm 1$ mentre il secondo algoritmo è sempre stabile cioè non dipende dai dati in input.

Studiare il condizionamento di un problema significa studiare l'errore inerente: $\epsilon_{in} = \frac{2x^2}{x^2-1}\epsilon_x$ vi sono dei valori di x per cui la funzione non può essere limitata: $lim_{x \to \pm 1} \frac{2x^2}{x^2-1} = +\infty$ allora il problema del calcolo di $f(x)$ è mal condizionato per $x \to \pm 1$.

1.6 Tecniche per l'analisi dell'errore

Teorema 1.6.1. *Se $f \in C^2([a,b])$ (cioè è continua e derivabile due volte in $[a,b]$) allora l'errore inerente approssimato al primo ordine lo si può scrivere come*

$$\epsilon_{in} \doteq \frac{f'(x)}{f(x)} x\epsilon_x = c_x\epsilon_x$$

*la quantità $c_x = \frac{f'(x)}{f(x)}x$ è detta **coefficiente di amplificazione** e misura quanto l'errore inerente sul dato si amplifica nel calcolo della funzione.*

Dimostrazione: dalla relazione dell'errore inerente

$$\epsilon_{in} = \frac{f(\hat{x}) - f(x)}{f(x)} = \frac{f(\hat{x}) - f(x)}{\hat{x} - x}\frac{\hat{x} - x}{f(x)}\frac{x}{x} = \frac{f(\hat{x}) - f(x)}{\hat{x} - x}\frac{x}{f(x)}\underbrace{\frac{\hat{x} - x}{x}}_{\epsilon_x}$$

dove ϵ_x è l'errore di rappresentazione; si ricava che se è f ha derivata prima e seconda continua su $[a, b]$ allora vale lo sviluppo di Taylor:

$$f(\hat{x}) = f(x) + f'(x)(\hat{x} - x) + f''(\xi_x)\frac{(\hat{x} - x)^2}{2!} \qquad |\xi_x - x| \leq |\hat{x} - x|$$

da cui si ottiene:

$$\epsilon_{in} = \frac{f(\hat{x}) - f(x)}{\hat{x} - x}\frac{x}{f(x)}\epsilon_x = \frac{f'(x)(\hat{x} - x) + f''(\xi_x)\frac{(\hat{x}-x)^2}{2!}}{\hat{x} - x}\frac{x}{f(x)}\epsilon_x = \left[f'(x) + f''(\xi_x)\frac{(\hat{x} - x)}{2!}\right]\frac{x}{f(x)}\epsilon_x = \frac{f'(x)}{f(x)}x\epsilon_x + f''(\xi_x)\frac{(\hat{x} - x)}{2!}\frac{x}{f(x)}\epsilon_x$$

ora il secondo componente può essere riscritta come $\frac{f''(\xi_x)x}{2!f(x)}(\frac{\hat{x}-x}{x})x\epsilon_x = \frac{f''(\xi_x)}{2!f(x)}x^2\epsilon_x^2$ e quindi per l'analisi del primo ordine può essere trascurata.

Esempio: riprendiamo $f(x) = x^2 - 1$ allora $c_x = \frac{x}{x^2-1}2x = \frac{2x^2}{x^2-1}$ che corrisponde con il calcolo fatto precedentemente. Più generalmente se $f : \Omega \to \mathbb{R}$ è definita su un insieme aperto di $\mathbb{R}^n$, differenziabile due volte su Ω ed il segmento di estremi $\hat{x}$ e x è contenuto in Ω allora vale:

$$\epsilon_{in} \doteq \sum_{i=1}^{n} c_{x_i}\epsilon_{x_i} \quad \textit{con } c_{x_i} = \frac{x_i}{f(x)}\frac{\partial f}{\partial x_i}$$

Per le operazioni aritmetiche si ottiene:

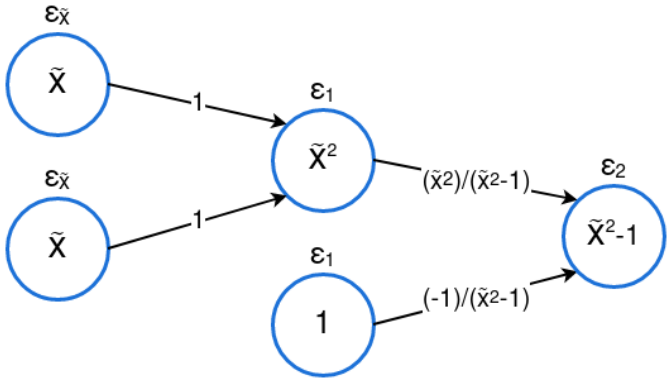
$$f(x, y) = x + y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x + y}, c_y = \frac{y}{x + y}$$

$$f(x, y) = x - y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x - y}, c_y = \frac{-y}{x - y}$$

$$f(x, y) = x \cdot y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1, c_y = 1$$

$$f(x, y) = x/y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1, c_y = -1$$

Il calcolo dell'errore algoritmico totale può essere reso intuitivo con l'aiuto di un grafo che permette di analizzare la precedenza degli operatori. I nodi corrispondono ai risultati intermedi generati dall'algoritmo mentre in corrispondenza a ciascun arco è riportato il relativo coefficiente di correlazione. In figura è riportato il grafo corrispondente all'operazione $g_1(\hat{x}) = \hat{x}^2 - 1$



ogni volta che si svolge una operazione si introduce l'errore locale dell'operazione: quindi $\epsilon_{\hat{x}} = 0$ per entrambi i casi perché $\hat{x}$ è un numero di macchina; ϵ_1 del nodo in alto è l'errore locale del quadrato mentre ϵ_1 del nodo in basso è uguale a 0 perché 1 è un numero di macchina; infine ϵ_2 è l'errore locale della sottrazione. Negli archi si inserisce il coefficiente di correlazione: partendo da sinistra verso destra vi è $\hat{x} \cdot \hat{x}$ che seguendo la formula si ricava $c_x = 1$ e $c_y = 1$; poi vi è la sottrazione di $\hat{x}^2$ e 1 allora $c_x = \frac{x^2}{x^2-1}$ e $c_y = \frac{-1}{x^2-1}$. L'errore algoritmico si ottiene percorrendo il grado dall'ultimo nodo verso i nodi iniziali:

$$\epsilon_{alg} = \epsilon_2 - \frac{1}{\hat{x}^2 - 1} \underbrace{\epsilon_1}_{=0} + \frac{\hat{x}^2}{\hat{x}^2 - 1}(\epsilon_1 + 1 \underbrace{\epsilon_{\hat{x}}}_{=0} + 1 \underbrace{\epsilon_{\hat{x}}}_{=0}) = \epsilon_2 + \frac{\hat{x}^2}{\hat{x}^2 - 1}\epsilon_1$$

1.7 Errore analitico

L'errore che si ha approssimando una certa funzione $f(x)$ non razionale con una funzione $h(x)$ razionale si chiama **errore analitico** ed è definito come:

$$\epsilon_{an} = \frac{h(x) - f(x)}{f(x)} \quad \textit{con } h(x) \neq 0$$

Nel calcolo della funzione $h(x)$ si genera poi un errore inerente ed algoritmico in accordo a quanto visto sopra:

$$\epsilon_{tot} = \frac{g(\hat{x}) - f(x)}{f(x)} \doteq \epsilon_{an} + \epsilon_{in} + \epsilon_{alg}$$

Chapter 2

I problemi dell'algebra lineare numerica

2.1 Norme vettoriali

Si dice **norma vettoriale** su $\mathbb{F}^n$ una funzione $f : \mathbb{F}^n \to \mathbb{R}$ che soddisfa le seguenti proprietà:

1. $\forall v \in \mathbb{F}^n,\ f(v) \geq 0$ ed inoltre $f(v) = 0 \Leftrightarrow v = 0$ (la norma è una funzione non negativa)
2. $\forall v \in \mathbb{F}^n, \forall \alpha \in \mathbb{F},\ \ f(\alpha v) = |\alpha| f(v)$
3. $\forall v, w \in \mathbb{F}^n,\ f(v + w) \leq f(v) + f(w)$ (disuguaglianza triangolare)

Se f è una norma vettoriale su $\mathbb{F}^n$ indicheremo per comodità di notazione $f(v) = \|v\|$. Si osservi che una norma vettoriale su $\mathbb{F}^n$ induce una *distanza* $d : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{R}$ tra elementi (punti) di $\mathbb{F}^n$ definita come $\forall v, w \in \mathbb{R}^n,\ d(v, w) = \|v - w\|$. Le proprietà della norma si traducono in analoghe proprietà della distanza indotta:

1. $\forall v, w \in \mathbb{F}^n, d(v, w) \geq 0$ e $d(v, w) = 0 \Leftrightarrow v = w$ (non negatività)
2. $\forall v, w \in \mathbb{F}^n, d(v, w) = d(w, v)$ (simmetria)
3. $\forall v, w, z \in \mathbb{F}^n, d(v, w) \leq d(v, z) + d(z, w)$ (diseguaglianza triangolare)

Le norme che useremo sono:

- **norma infinito** nel caso reale $\mathbb{R}^n$ e complesso $\mathbb{C}^n$ si ha $f(v) = \|v\|_\infty = max_{1 \leq i \leq n} |v_i|$
- **norma 1**, nel caso reale $\mathbb{R}^n$ e complesso $\mathbb{C}^n$ si ha $f(v) = \|v\|_1 = \sum_{i=1}^n |v_i|$
- **norma 2 o euclidea**:
  - nel caso reale $\mathbb{R}^n$ si ha $f(v) = \|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2} = \sqrt{v^T v}$
  - nel caso complesso $\mathbb{C}^n$ si ha $f(v) = \|v\|_2 = \sqrt{\sum_{i=1}^n |v_i|^2} = \sqrt{v^H v}$

Sebbene l'utilizzo di norme differenti forniscono valori differenti, le proprietà qualitative risultano preservate: le norme sono topologicamente equivalenti. Vale infatti il seguente teorema:

Teorema 2.1.1 (**Equivalenza delle norme**). *Siano f, g due norme su $\mathbb{F}^n$ allora esistono costanti $\alpha, \beta \in \mathbb{R}$ e > 0 tali che*

$$\forall v \in \mathbb{F}^n\ \alpha g(v) \leq f(v) \leq \beta g(v)$$

Esempio:

$$x^T = \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \text{ si ha } \|x\|_2 = \sqrt{3},\ \|x\|_1 = 3,\ \|x\|_\infty = 1$$

$$\text{dati } \alpha = 1, \beta = \sqrt{n} \text{ vale } \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty \text{ infatti } 1 \leq \sqrt{3} \leq \sqrt{3}$$

Le norme vettoriali possono essere estese alle matrici.

2.2 Norme matriciali

Si dice **norma matriciale** su $\mathbb{F}^{n\times n}$ una funzione $f : \mathbb{F}^{n\times n} \to \mathbb{R}$ che soddisfa le seguenti proprietà:

1. $\forall A \in \mathbb{F}^{n\times n},\ f(A) \geq 0$ ed inoltre $f(A) = 0 \Leftrightarrow A = 0$
2. $\forall A \in \mathbb{F}^{n\times n},\ \forall \alpha \in \mathbb{F},\ f(\alpha A) = |\alpha| f(A)$
3. $\forall A, B \in \mathbb{F}^{n\times n},\ f(A + B) \leq f(A) + f(B)$
4. $\forall A, B \in \mathbb{F}^{n\times n},\ f(A \cdot B) \leq f(A) \cdot f(B)$

Se f è una norma vettoriale su $\mathbb{F}^{n\times n}$ indicheremo per comodità di notazione $f(A) = \|A\|$. Analogamente a prima una norma matriciale su $\mathbb{F}^{n\times n}$ induce una *distanza* $d : \mathbb{F}^{n\times n} \times \mathbb{F}^{n\times n} \to \mathbb{R}$ tra elementi di $\mathbb{F}^{n\times n}$ definita come $\forall A, B \in \mathbb{F}^{n\times n}\ d(A, B) = \|A - B\|$. Dalla proprietà 1 e 4 si ricava che $\|I_n\| \geq 1$ con I_n matrice identità. Data $\|.\|$ una norma vettoriale su $\mathbb{F}^n$ si dice **norma matriciale indotta** o **compatibile** con la norma vettoriale la funzione $f : \mathbb{F}^{n\times n} \to \mathbb{R}$ definita da :

$$\forall A \in \mathbb{F}^{n\times n},\ f(A) = \max_{\{v\in\mathbb{F}^n:\, \|v\|=1\}} \|Av\|$$

Per comodità di notazione scriveremo $f(A) = \|A\|$ utilizzando lo stesso simbolo della norma vettoriale che induce la norma matriciale. Si osserva che:

- la definizione è ben posta infatti la funzione $v \to \|Av\|$ è continua per cui dal teorema di Weierstrass esiste il massimo. Inoltre la funzione f così definita verifica le proprietà 1-4 delle norme matriciali;
- per una norma matriciale indotta da una norma vettoriale vale $\|I_n\| = 1$;
- esistono norme matriciali che non sono indotte da una norma vettoriale. Ad esempio si consideri la funzione *norma di Frobenius* definita come:

$$\forall A = (a_{ij}) \in \mathbb{F}^{n\times n},\ \|A\| = \sqrt{\sum_{i\leq i,j\leq n} |a_{ij}|^2}$$

Teorema 2.2.1 (**Compatibilità delle norme**). *Sia $\|.\|$ una norma vettoriale su $\mathbb{F}^n$ e sia $\|.\|$ la norma matriciale indotta vale allora che $\forall A \in \mathbb{F}^{n\times n},\ \forall v \in \mathbb{F}^n,\ \|Av\| \leq \|A\|\|v\|$.*

Dimostrazione: se $v = 0$ allora $\|Av\| = \|0\| = 0$ per proprietà 1 delle norme vettoriali. Allora $0 = \|Av\| \leq \|A\|\|v\| = 0$, quindi la relazione vale. Se invece $v \neq 0$ allora:

$$\frac{1}{\|v\|}\|Av\| \leq \|A\| = \max_{\|z\|=1} \|Az\| \quad \textit{cioè} \quad \|A\frac{v}{\|v\|}\| \leq \|A\| = \max_{\|z\|=1} \|Az\|$$

La valutazione delle norme matriciali indotte dalla norma euclidea, dalla norma 1 e dalla norma infinito in accordo alla definizione risulta computazionalmente non praticabile. Vengono pertanto fornite le seguenti caratterizzazioni che seguono dalla definizione e mediante l'individuazione del punto di massimo forniscono lo strumento per il calcolo effettivo delle norme.

Teorema 2.2.2. *Sia $A = (a_{ij}) \in \mathbb{F}^{n\times n}$ si ha:*

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \qquad \|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \qquad \|A\|_2 = \sqrt{\rho(A^H A)}$$

*Il numero $\rho(M)$ è detto **raggio spettrale** di una matrice $M \in \mathbb{F}^{n\times n}$ definito come il modulo dell'autovalore di modulo massimo di M: $\rho(M) = max_{1\leq i\leq n}|\lambda_i|$ con λ_i autovalore di M.*

Esempio:

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

$$\|A\|_\infty = max\{1+2,\ |-1|+3\} = max\{3, 4\} = 4$$

$$\|A\|_1 = max\{1+|-1|,\ 2+3\} = max\{2, 5\} = 5$$

$$A^TA = \begin{bmatrix}1 & -1\\ 2 & 3\end{bmatrix}\begin{bmatrix}1 & 2\\ -1 & 3\end{bmatrix} = \begin{bmatrix}2 & -1\\ -1 & 13\end{bmatrix} \Rightarrow p(\lambda) = (2-\lambda)(13-\lambda) - 1 = \lambda^2 - 15\lambda + 25 = 0$$

$$max\{\tfrac{15+\sqrt{125}}{2}, \tfrac{15-\sqrt{125}}{2}\} = \tfrac{15+\sqrt{125}}{2} \quad \textit{allora} \quad \|A\|_2 = \sqrt{\tfrac{15+\sqrt{125}}{2}}$$

Da ricordare: le matrici simmetriche hanno autovalori reali e se M è una matrice triangolare superiore o inferiore, gli elementi sulla diagonale principale sono gli autovalori.

2.3 Condizionamento della risoluzione di sistemi lineari

Data $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ matrice invertibile e dato $b \in \mathbb{R}^n$ si cerca un vettore $x \in \mathbb{R}^n$ tale che $Ax = b$. La matrice A è detta *matrice dei coefficienti* del sistema lineare; il vettore b è detto *vettore dei termini noti* e il vettore x è detto *vettore delle incognite*. Studiare il **condizionamento** significa studiare quanto è sensibile la soluzione x al variare di A e b. Se si deve risolvere il sistema $Ax = b$ dove però A e b hanno un certo errore ci si trova a risolvere il sistema $\hat{A}\hat{x} = \hat{b}$. Valutare il condizionamento significa andare a valutare la distanza tra $\hat{x}$ e x. Esempio:

$$\begin{cases}1.01x + 1.02y = 2.03\\ 0.99x + y = 1.99\end{cases} \quad \textit{che in termini matriciali equivale a scrivere } A = \begin{bmatrix}1.01 & 1.02\\ 0.99 & 1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}2.03\\ 1.99\end{bmatrix}$$

$$\textit{e la soluzione è } \begin{bmatrix}x\\ y\end{bmatrix} = \begin{bmatrix}1\\ 1\end{bmatrix}$$

Cosa succede se perturbiamo di un centesimo il coefficiente 0.99? Cioè si introduce un errore di un centesimo su uno dei dati e quindi la matrice perturbata di dati equivale a:

$$\hat{A} = A + \begin{bmatrix}0 & 0\\ \frac{1}{100} & 0\end{bmatrix} \quad \textit{e la soluzione approssimata è} \quad \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix} = \begin{bmatrix}-0.02....\\ 2.01.....\end{bmatrix}$$

Abbiamo introdotto una certa perturbazione ma ci aspettavamo che il problema fosse ben condizionato e quindi il risultato doveva essere affetto da un errore dell'ordine della perturbazione introdotta cioè andavano bene delle soluzioni come 1.01 o 0.997 invece troviamo $\hat{x} = -0.02$ e $\hat{y} = 2.01$. Questo vuol dire che il problema è mal condizionato in maniera severa: piccole perturbazioni sui dati provocano un errore molto elevato sul risultato. La quantità:

$$\frac{\|\hat{x} - x\|}{\|x\|}$$

misura il condizionamento di un problema. Per semplicità di analisi assumiamo di perturbare il solo termine noto $b \neq 0$ e di non avere alcun errore su A.

Teorema 2.3.1. *Dato A non singolare, $b \neq 0$ e scelta una norma che induce una norma matriciale allora vale che:*

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\hat{b} - b\|}{\|b\|}$$

*la quantità $u(A) = \|A\|\|A^{-1}\|$ si chiama **numero di condizionamento** ed è l'equivalente del coefficiente di amplificazione del capitolo precedente.*

<u>Dimostrazione</u>: $\hat{x} - x = A^{-1}\hat{b} - A^{-1}b = A^{-1}(\hat{b} - b)$ a questo punto $\|\hat{x} - x\| = \|A^{-1}(\hat{b} - b)\| \leq \|A^{-1}\|\|\hat{b} - b\|$ per la compatibilità tra norma vettoriale e norma matriciale indotta. Studiamo ora il denominatore: $\|b\| = \|Ax\| \leq \|A\|\|x\|$ per la compatibilità tra norme, quindi si ricava che $\|x\| \geq \frac{\|b\|}{\|A\|}$ dove sicuramente $\|A\| \neq 0$ perché A è non singolare. Mettendo tutto insieme:

$$\frac{\|\hat{x} - x\|}{\|x\|} \leq \frac{1}{\|x\|}\|A^{-1}\|\|\hat{b} - b\| \leq \|A\|\|A^{-1}\|\frac{\|\hat{b} - b\|}{\|b\|}$$

Il numero di condizionamento è una quantità sempre maggiore di 1. Per una norma matriciale indotta da una norma vettoriale vale: $1 = \|I_n\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = u(A)$.

2.4 Teoremi di localizzazione per autovalori

I teoremi di localizzazione individuano regioni del piano complesso dove gli autovalori sono confinati. Il più classico è il seguente teorema:

Teorema 2.4.1 (**Teorema di Gershgorin**). *Data una matrice $A \in \mathbb{C}^{n\times n}$, definiamo il cerchio nel piano complesso come:*

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

*questo è detto **cerchio di Gershgorin** di centro a_{ii} e raggio $r_i = \sum_{j=1,j\neq i} |a_{ij}|$ relativo alla riga i. Per $i = 1, ..., n$ se λ è autovalore di A allora $\lambda \in \bigcup_{i=1}^{n} K_i$.*

Dimostrazione: se λ autovalore allora $\exists x \in \mathbb{C}^n, x \neq 0$ tale che $Ax = \lambda x$. Questa relazione implica che

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i, \;\; 1 \leq i \leq n \;\; \text{da cui} \;\; \sum_{j=1, j\neq i}^{n} a_{ij}x_j = (\lambda - a_{ii})x_i, \;\; 1 \leq i \leq n$$

Sia p l'indice tale per cui $|x_p| = \|x\|_\infty$. Poiché $x \neq 0$ si ha $|x_p| > 0$. Sia $i = p$ allora

$$(\lambda - a_{pp})x_p = \sum_{j=1, j\neq p}^{n} a_{pj}x_j \;\; \text{da cui passando ai valori assoluti}$$

$$|(\lambda - a_{pp})x_p| = |\lambda - a_{pp}||x_p| = \left| \sum_{j=1, j\neq p}^{n} a_{pj}x_j \right| \leq \sum_{j=1, j\neq p}^{n} |a_{pj}||x_j|$$

e quindi dividendo ambo i membri per $|x_p|$

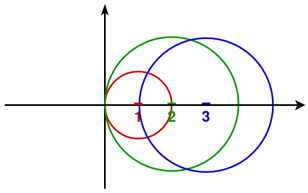
$$|\lambda - a_{pp}| \leq \sum_{j=1, j\neq p}^{n} |a_{pj}| \frac{|x_j|}{|x_p|} \leq \sum_{j=1, j\neq p}^{n} |a_{pj}|$$

Questa relazione implica che $\lambda \in K_p$

Esempio:

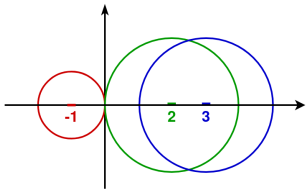
$$A_1 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 1 \\ 2 & 0 & 3 \end{bmatrix}$$

$$K_1 = \{z \in \mathbb{C} : |z - 1| \leq 1\}$$
$$K_2 = \{z \in \mathbb{C} : |z - 2| \leq 2\}$$
$$K_3 = \{z \in \mathbb{C} : |z - 3| \leq 2\}$$



$$A_2 = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 2 & 1 \\ 2 & 0 & 3 \end{bmatrix}$$

$$K_1 = \{z \in \mathbb{C} : |z + 1| \leq 1\}$$
$$K_2 = \{z \in \mathbb{C} : |z - 2| \leq 2\}$$
$$K_3 = \{z \in \mathbb{C} : |z - 3| \leq 2\}$$



Gli autovalori saranno all'interno dell'unione dei cerchi disegnati sul grafico. Un risultato di inclusione generalmente più debole è fornito dal seguente teorema:

Teorema 2.4.2 (**Teorema di Hirsch**). *Data una matrice $A \in \mathbb{C}^{n\times n}$ e sia $\|.\|$ una norma matriciale indotta allora vale*

$$\forall \lambda \text{ autovalore di } A \Rightarrow |\lambda| \leq \|A\|$$

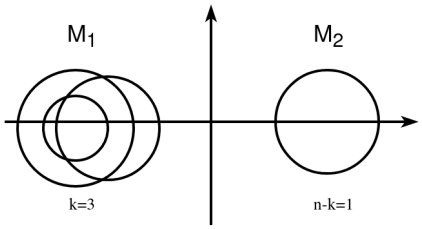
Dimostrazione: se λ è autovalore allora $\exists x \in \mathbb{C}^n, x \neq 0$ tale che $Ax = \lambda x$. Questa relazione implica che:

$$\|\lambda x\| = |\lambda| \, \|x\| = \|Ax\| \leq \|A\|\|x\| \text{ ora dividendo per } \|x\| \text{ si ottiene } |\lambda| \leq \|A\|$$

Da ricordare: se tra le radici di un polinomio vi è una radice complessa allora c'è anche la sua radice coniugata.

Teorema 2.4.3 (**Secondo teorema di Gershgorin**). *Se l'unione M_1 di k cerchi di Gershgorin è disgiunta dall'unione dei rimanenti $n-k$ cerchi M_2 allora k autovalori appartengono a M_1 e $n-k$ a M_2.*

Esempio: supponiamo di avere una matrice A che è 4×4:



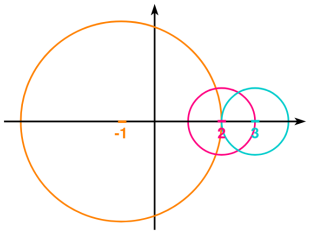
se $A \in \mathbb{R}^{n \times n}$ si può dire che in M_2 c'è un solo autovalore reale mentre in M_1 ci sono tre autovalori dove almeno uno di loro è reale. Si definisce H_j i cerchi per colonna come:

$$H_j = \{z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{i=1, i \neq j}^{n} |a_{ij}|\}$$

Riprendiamo A_2 dell'esempio precedente e notiamo che i cerchi per riga sono completamente diversi dai cerchi per colonna (essi coincidono se la matrice è simmetrica)

$$A_2 = \begin{bmatrix} -1 & -1 & 0 \\ -1 & 2 & 1 \\ 2 & 0 & 3 \end{bmatrix}$$

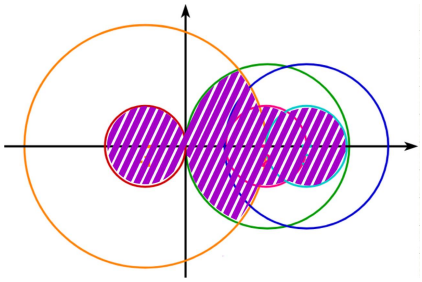
$$H_1 = \{z \in \mathbb{C} : |z + 1| \leq 3\}$$
$$H_2 = \{z \in \mathbb{C} : |z - 2| \leq 1\}$$
$$H_3 = \{z \in \mathbb{C} : |z - 3| \leq 1\}$$



Siccome gli autovalori di A e di A^T coincidono allora:

$$\lambda \in \Big(\bigcup_{i=1}^{n} K_i\Big) \cap \Big(\bigcup_{j=1}^{n} H_j\Big)$$

nel nostro caso:



Una matrice $A \in \mathbb{C}^{n \times n}$ si dice a **predominanza diagonale per righe** se vale

$$|a_{ii}| > \sum_{j=1, j \neq i}^{n} |a_{ij}|, \quad \forall i = 1, ..., n$$

Esempio:

$A = \begin{bmatrix} 1 & 0 & \frac{1}{2} \\ 1 & -2 & \frac{2}{3} \\ -1 & 1 & 3 \end{bmatrix}$ *è a predominanza diagonale per righe:* $1 > 0 + \frac{1}{2}$; $|-2| > 1 + \frac{2}{3}$ *e* $3 > |-1| + 1$

Tuttavia non è a predominanza per colonna perché per la prima colonna $1 \not> 1 + |-1|$.

Teorema 2.4.4. *Se A è a predominanza diagonale (per riga o per colonna) allora A è non singolare.*

Dimostrazione: A è non singolare se e solo se tutti gli autovalori sono diversi da zero. Allora basta verificare che il $\lambda = 0$ non appartiene all'unione $\bigcup_{i=1}^{n} K_i$. Poiché A è a predominanza diagonale $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$ dove $|a_{ii}| = |0 - a_{ii}|$. Ora riscrivendo la prima relazione si ottiene:

$$|0 - a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|, \ \forall i = 1, ..., n$$

$|0 - a_{ii}|$ *significa che* $0 \notin K_i, \ \forall i = 1, ..., n$ *quindi* $0 \notin \bigcup_{i=1}^{n} K_i$.

Chapter 3

Metodi diretti per la risoluzione di sistemi lineari

3.1 Sistemi triangolari

Sistema lineare $Ax = b$ con $A \in \mathbb{R}^{n\times n}, \mathbb{C}^{n\times n}$, la soluzione x esiste ed è unica se:

- $rank(A) = n$ ($dimImm(A) = n$);
- $Ker(A) = \{\emptyset\}$ ($dimKer(A) = 0$);
- $det(A) \neq 0$;
- *tutti gli autovalori sono diversi da zero*.

La soluzione x esiste ma non è unica quando $Imm[A|b] = Imm[A]$. In generale per risolvere il sistema si usa il *metodo di Gauss* che trasforma la matrice in una matrice *a scalini* o **triangolare superiore**.

$$\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ & \ddots & \vdots & \vdots \\ & & a_{nn} & b_n \end{array}\right]$$

Trasformare la matrice in questo modo è conveniente perché ora si può effettuare il **metodo di sostituzione all'indietro** come segue: dall'ultima equazione si ricava:

$$a_{nn}x_n = b_n \quad \Rightarrow \quad x_n = \frac{b_n}{a_{nn}}$$

dopodiché si continua con la penultima equazione e si ricavano tutte le incognite. In generale vale la formula:

$$x_k = \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}, \quad k = n, n-1, ..., 1$$

La matrice $A^{n\times n}$ è invertibile se la sua matrice triangolare superiore ha n pivot.

3.2 Fattorizzazione LU

Una matrice $A \in \mathbb{R}^{n\times n}$ si dice **fattorizzabile nella forma LU** se esistono $U \in \mathbb{R}^{n\times n}$ matrice triangolare superiore ed $L \in \mathbb{R}^{n\times n}$ matrice triangolare inferiore con elementi uguali ad 1 sulla diagonale principale tali che $A = L \cdot U$. Se $A \in \mathbb{R}^{n\times n}$ invertibile è fattorizzabile nella forma LU allora segue che U è pure invertibile e dunque il sistema lineare $Ax = b$ può essere risolto mediante la sequenza di sistemi triangolari

$$\begin{cases} Ly = b \\ Ux = y \end{cases}$$

Esempio: si rappresenta il seguente sistema in forma matriciale e si utilizza il metodo di Gauss:

$$\begin{cases} 2x_1 - x_2 + x_3 = 1 \\ -2x_1 + 2x_2 + x_3 = -1 \\ 4x_1 - 4x_2 + x_3 = 5 \end{cases}$$

$$\left[\begin{array}{ccc|c} 2 & -1 & 1 & 1 \\ -2 & 2 & 1 & -1 \\ 4 & -4 & 1 & 5 \end{array}\right] \begin{array}{l} r_2 \to r_2 - (\mathbf{-1})r_1 \\ r_3 \to r_3 - (\mathbf{2})r_1 \end{array} \left[\begin{array}{ccc|c} 2 & -1 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & -2 & -1 & 3 \end{array}\right] \quad r_3 \to r_3 - (\mathbf{-2})r_2 \quad \left[\begin{array}{ccc|c} 2 & -1 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 3 & 3 \end{array}\right]$$

Implicitamente, utilizzando il metodo di Gauss, abbiamo calcolato la fattorizzazione LU della matrice:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \mathbf{-1} & 1 & 0 \\ \mathbf{2} & \mathbf{-2} & 1 \end{bmatrix} \quad U = \left[\begin{array}{ccc|c} 2 & -1 & 1 & 1 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 3 & 3 \end{array}\right] \quad \Rightarrow \quad L \cdot U = \begin{bmatrix} 2 & -1 & 1 \\ -2 & 2 & 1 \\ 4 & -4 & 1 \end{bmatrix}$$

e anche il sistema $Ly = b$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 5 \end{bmatrix}$$

infatti si può notare che il vettore y corrisponde ai termini noti trovati durante la riduzione a scalini della matrice: $y_1 = 1, y_2 = 0, y_3 = 3$.

Teorema 3.2.1 (**Esistenza ed unicità della fattorizzazione LU**). *Sia $A \in \mathbb{R}^{n \times n}$. Se le sottomatrici principali di testa di A di ordine k con $k = 1, ..., n-1$ sono non singolari allora esiste unica la fattorizzazione LU di A.*

Dimostrazione: per induzione sulla dimensione n della matrice.

Caso base: $n = 1$ allora $A = [a_{11}]$ non ha sottomatrici principali dunque il teorema è verificato $L = [1]$ $U = [a_{11}]$ ed è unica per la struttura di L.

Passo induttivo: supponiamo che il teorema sia vero per matrici di ordine $m \leq n-1$ e dimostriamo che vale per A di ordine n. Se A è fattorizzabile LU allora esistono L ed U tali che $A = LU$ con L triangolare inferiore con 1 sulla diagonale e U triangolare superiore. Si partiziona a blocchi la matrice A e le matrici incognite L ed U

$$A = \left[\begin{array}{c|c} A_{n-1} & z \\ \hline x^T & a_{nn} \end{array}\right] = \left[\begin{array}{c|c} \hat{L} & 0 \\ \hline w^T & 1 \end{array}\right] \left[\begin{array}{c|c} \hat{U} & y \\ \hline 0^T & \beta \end{array}\right]$$

dove:

- *A_{n-1} è la sottomatrice principale di testa di ordine $n-1$, z vettore e x^T è l'ultima riga di A;*
- *$\hat{L}$ è una matrice triangolare inferiore di dimensione $n-1$ e w^T è l'ultima riga di L;*
- *$\hat{U}$ è una matrice triangolare superiore di dimensione $n-1$, y vettore incognita e β scalare.*

Allora si impone:

$$\begin{cases} A_{n-1} = \hat{L}\hat{U} + 0 \cdot 0^T = \hat{L}\hat{U} \\ z = \hat{L}y + 0 \cdot \beta = \hat{L}y \\ x^T = w^T\hat{U} + 1 \cdot 0^T = w^T\hat{U} \\ a_{nn} = w^T y + 1 \cdot \beta \end{cases} \quad \Rightarrow \quad \begin{cases} A_{n-1} = \hat{L}\hat{U} \\ z = \hat{L}y \\ x^T = w^T\hat{U} \\ a_{nn} = w^T y + \beta \end{cases}$$

Per la prima equazione si può notare che la matrice A_{n-1} è invertibile e le sottomatrici principali di testa di A_{n-1} e di A coincidono. Questo implica che per le ipotesi del teorema A_{n-1} ha le sottomatrici principali di testa di ordine k con $k = 1, ..., n-2$ non singolari. Allora A_{n-1} soddisfa l'ipotesi induttiva e quindi è fattorizzabile in modo unico LU. Il sistema lineare $z = \hat{L}y$ ammette un'unica soluzione poiché $det\hat{L} = 1$. Il terzo sistema lo si può riscrivere come: $x = \hat{U}^T w$ allora la soluzione w esiste ed è unica se e solo se $det\hat{U} \neq 0$ ma noi sappiamo che $det\hat{U} = detA_{n-1}$ dove $detA_{n-1} \neq 0$ per le ipotesi del teorema. Infine l'ultima equazione equivale a $\beta = a_{nn} - w^T y$ ed è unico.

Da ricordare: il determinante di matrici triangolari è il prodotto degli elementi diagonali. Data una generica matrice A le sue sottomatrici principali di testa sono:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \dots & \dots & a_{nn} \end{bmatrix} \qquad A_1 = a_{11} \quad A_2 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \dots \quad A_k = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{bmatrix}$$

Quindi data la seguente matrice:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \textit{si deve controllare:} \quad detA_1 = det[1] = 1 \quad detA_2 = det \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix} = -3$$

e dato che $detA_1$ e $detA_2$ sono diversi da zero allora si può affermare che la matrice A ammette unica fattorizzazione LU: esistono L e U tali che $A = L \cdot U$.

3.3 Matrici elementari di Gauss

Una matrice $E \in \mathbb{R}^{n \times n}$ si dice **matrice elementare di Gauss** se esiste un $k \in \mathbb{N}$ e un $v \in \mathbb{R}^n$ con $v_1 = ...v_k = 0$ tale che

$$E = I_n - ve_k^T$$

con e_k vettore k-esimo della base canonica. Esempio: con $k = 2$ allora v avrà le prime due componenti nulle e e_k^T ha un 1 in seconda posizione e 0 altrimenti

$$E_2 = I_n - \begin{bmatrix} 0 \\ 0 \\ v_3 \\ \vdots \\ v_n \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & -v_3 & 1 & & \\ & \vdots & & \ddots & \\ & -v_n & & & 1 \end{bmatrix}$$

Proprietà:

1. le matrici elementari di Gauss sono triangolari inferiori con 1 sulla diagonale principale;
2. le matrici elementari di Gauss sono invertibili, E^{-1} è ancora una matrice elementare di Gauss e in particolare vale $E^{-1} = I_n + ve_k^T$. Infatti:

$$(I_n - ve_k^T)(I_n + ve_k^T) = I_n - ve_k^T + ve_k^T - ve_k^T ve_k^T = I_n - \underbrace{(e_k^T v)ve_k^T}_{e_k^T v = v_k = 0} = I_n$$

3. dato $x \in \mathbb{R}^n$ con $x_k \neq 0$, esiste una matrice elementare di Gauss tale che $Ex = \begin{bmatrix} x_1, \dots, x_k, 0, \dots, 0 \end{bmatrix}^T$, cioè restituisce un vettore che è rimasto invariato nelle prime k componenti e 0 altrimenti. Infatti basta porre $E - I_n - ve_k^T$ con $x_j - v_j x_k = 0 \Leftrightarrow v_j = x_j / x_k, \quad k + 1 \leq j \leq n$;
4. Se $E_k = I_n - ve_k^T$ e $E_h = I_n - we_h^T$ sono matrici elementari di Gauss con $h > k$ allora $e_k^T w = 0$ e dunque $E_k \cdot E_h = I_n - ve_k^T - we_h^T$. Ciò implica che la matrice prodotto risulta costruita semplicemente apponendo nella corretta posizione i vettori v e w dei fattori;
5. Il prodotto Ey di una matrice elementare di Gauss costa $O(n - k)$ operazioni moltiplicative. Si ha infatti

$$z = Ey = (I_n - v_k^T)y = y - ve_k^T y = y - v\underbrace{(e_k^T y)}_{= y_k} = y - y_k v$$

quindi $z_i = y_i$ per $i \leq k$ e $z_i = y_i - y_k v_i$ per $i = k + 1...n$ dove in quest'ultima ogni componente richiede una sottrazione e una moltiplicazione allora calcolare Ey costa $n - k$ sottrazioni e $n - k$ moltiplicazioni.

3.4 Metodo di eliminazione gaussiana

Il seguente processo detto **metodo di eliminazione gaussiana** utilizza le proprietà delle matrici elementari di Gauss sotto opportune ipotesi di una matrice $A = A_0$ in forma triangolare superiore. Indichiamo con $a_1^{(k)}, ..., a_n^{(k)}$ i vettori colonna della matrice $A_k = (a_{ik}^{(k)}), 1 \leq i, j \leq n, 0 \leq k \leq n-1$. Se assumiamo che $a_{11}^{(0)} \neq 0$ allora per la proprietà 3 possiamo determinare E_1 tale da aversi:

$$E_1 a_1^{(0)} = [a_{11}^{(0)}, 0, ..., 0]^T$$

Risulta:

$$E_1 = I_n - [0, a_{21}^{(0)}/a_{11}^{(0)}, ..., a_{n1}^{(0)}/a_{11}^{(0)}]^T e_1^T$$

I termini $m_{21}^{(0)} = a_{21}^{(0)}/a_{11}^{(0)}, ..., m_{n1}^{(0)} = a_{n1}^{(0)}/a_{11}^{(0)}$ sono detti **moltiplicatori di Gauss** $m_{ik}^{(k-1)} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ mentre il termine $a_{11}^{(0)}$ è il pivot. Poniamo dunque:

$$A_1 = E_1 A_0 \quad b_1 = E_1 b_0$$

Il processo prosegue operando sulla matrice A_1: se assumiamo che $a_{22}^{(1)} \neq 0$ allora per la proprietà 3 possiamo determinare E_2 tale da aversi:

$$E_2 a_2^{(1)} = [a_{12}^{(1)}, a_{22}^{(1)}, 0, ..., 0]^T$$

Risulta:

$$E_2 = I_n - [0, 0, a_{32}^{(1)}/a_{22}^{(1)}, ..., a_{n2}^{(1)}/a_{22}^{(0)}]^T e_2^T$$

Si osserva che $E_2 a_1^{(1)} = a_1^{(1)}$. Poniamo dunque:

$$A_2 = E_2 A_1 \quad b_2 = E_2 b_1$$

In questo modo assumendo che valga $a_{jj}^{(j-1)} \neq 0,\ 1 \leq j \leq n-1$ è possibile determinare una sequenza di matrici elementari di Gauss $E_1, ..., E_{n-1}$ tali da avere:

$$E_{n-1}E_{n-2}...E_1 A_0 = E_{n-1}E_{n-2}....E_1 A = A_{n-1} = U$$

Le relazioni $A_k = E_k A_{k-1},\ b_k = E_k b_{k-1}$ espresse in termini di componenti si scrivono:

$$\begin{cases} a_{ij}^{(k)} = a_{ij}^{(k-1)} & se\ i \leq k\ o\ j \leq k-1 \\ a_{ik}^{(k)} = 0 & se\ i > k \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}^{(k-1)} a_{kj}^{(k-1)} & se\ i > k\ e\ j > k \end{cases} \qquad \begin{cases} b_i^{(k)} = b_i^{(k-1)} & se\ i \leq k \\ b_i^{(k)} = b_i^{(k-1)} - m_{ik}^{(k-1)} b_k^{k-1)} & se\ i > k \end{cases}$$

Studiamo il costo computazionale di questa procedura: al passo k-esimo si calcola $m_{ik}^{(k-1)}$ che costa $n-k$ divisioni e si deve aggiornare $(n-k)^2$ elementi con la formula $a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}^{(k-1)} a_{kj}^{(k-1)}$ cioè il costo è pari a $2(n-1)^2$. In totale il costo computazionale è:

$$\sum_{k=1}^{n-1} (n-k) + 2(n-k)^2$$

che con un cambio di variabile $n - k = i$ si ottiene:

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

Dalla proprietà 4 segue che il prodotto delle matrici elementari $L = E_1^{-1}E_2^{-1}, ..., E_{n-1}^{-1}$ è determinato apponendo nel corretto ordine i moltiplicatori di Gauss

$$L = \begin{bmatrix} 1 & & & & \\ \frac{a_{21}^{(0)}}{a_{11}^{(0)}} & 1 & & & \\ \frac{a_{31}^{(0)}}{a_{11}^{(0)}} & \frac{a_{32}^{(1)}}{a_{22}^{(1)}} & \ddots & & \\ \vdots & \vdots & \vdots & \ddots & \\ \frac{a_{n1}^{(0)}}{a_{11}^{(0)}} & \frac{a_{n2}^{(1)}}{a_{22}^{(1)}} & \cdots & \frac{a_{n,n-1}^{(n-1)}}{a_{n-1,n-1}^{(n-1)}} & 1 \end{bmatrix}$$

3.5 Tecniche di Pivoting

Teorema 3.5.1. *Data una matrice $A \in \mathbb{R}^{n\times n}$ allora le sottomatrici principali di testa di ordine k non singolare per $k = 1, ..., n-1$ se e soltanto se $a_{kk}^{(k-1)} \neq 0$ per $k = 1, 2, ..., n-1$.*

Esempio: la matrice A sottostante ha 0 in posizione a_{11}, quindi non lo si può utilizzare come pivot.

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

Allora si scambiano la prima e seconda riga di A e per farlo si costruisce una matrice P ottenuta dalla matrice I_n scambiando la prima e seconda riga e poi si effettua il prodotto PA

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \Rightarrow \quad PA = \begin{bmatrix} -3 & 2 & 1 \\ 0 & 1 & 2 \\ -1 & 0 & 0 \end{bmatrix}$$

Ovviamente se si vuole risolvere il sistema $Ax = b$ allora si deve applicare la stessa permutazione anche a b, quindi si risolverà il sistema $PAx = Pb$. Osservazione: se A è invertibile allora è sempre possibile trovare una matrice di permutazione P tale che $A = PLU$. Questa tecnica di pivot non viene fatta soltanto quando l'elemento è nullo: la scelta dell'elemento pivotale è suggerita da valutazioni di stabilità numerica. Se indichiamo con $\hat{L}$ il fattore L effettivamente calcolato ed analogamente $\hat{U}$ il fattore U calcolato allora

$$\hat{L}\hat{U} = A + E, \quad \|E\| = O(u)\|L\|\|U\|$$

Per controllare gli elementi di L, ad ogni passo si sceglie l'elemento di modulo massimo sulla colonna corrente. Se

$$|a_{jk}^{(k-1)}| = \max_{k\leq i\leq n} |a_{ik}^{(k-1)}|$$

allora si scambia la riga k con la riga j al passo k. Con questa tecnica, detta **massimo pivot**, l'algoritmo diventa più stabile. Ovvero i moltiplicatori $m_{ik}^{(k-1)}$ sono tutti più piccoli di 1 e quindi gli elementi di L hanno modulo minore o uguale a 1.

Chapter 4

Metodi iterativi per la risoluzione di sistemi lineari

4.1 Generalità sui metodi iterativi

Per sistemi lineari $Ax = b$ dove la matrice dei coefficienti $A \in \mathbb{R}^{n \times n}$ è sparsa o di elevate dimensioni ($n > 10^6$) allora il metodo di Gauss è poco competitivo. In questi casi si ricorre ai **metodi iterativi** che a partire da un vettore di partenza $x^{(0)}$ si genera una successione di vettori $\{x^{(k)}\}$ tale che $\{x^{(k)}\}$ converge alla soluzione del sistema lineare $Ax = b$ cioè

$$\lim_{k \to +\infty} \|x^{(k)} - x\| = 0$$

Per il teorema di equivalenza delle norme posso scegliere una qualsiasi norma. In pratica la costruzione della successione termina dopo un numero finito di passi determinato in base alla verifica di opportuni *criteri di arresto*. La qualità e l'efficienza di un metodo iterativo è pertanto determinata dalle proprietà di convergenza della successione generata. Una tecnica generale per derivare un metodo iterativo si basa sulla decomposizione additiva $A = M - N$ con M matrice invertibile. Si ha allora

$$Ax = b \quad \Leftrightarrow \quad (M - N)x = b \quad \Leftrightarrow \quad Mx - Nx = b \quad \Leftrightarrow \quad Mx = Nx + b \quad \Leftrightarrow \quad x = M^{-1}Nx + M^{-1}b$$

Posto dunque $P = M^{-1}N$ detta **matrice di iterazione** e $q = M^{-1}b$ si ottiene che

$$Ax = b \quad \Leftrightarrow \quad x = Px + q$$

ora dato un vettore iniziale $x^{(0)} \in \mathbb{R}^n$ si calcola $x^{(k+1)} = Px^{(k)} + q, \ k > 0$. Esempio: data una matrice P, un vettore q e un vettore $x^{(0)}$ come segue:

$$P = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{3} \end{bmatrix} \quad q = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

si costruisce la successione di vettori $x^{k+1} = Px^{(k)} + 0 = Px^{(k)}$

$$x^{(1)} = Px^{(0)} = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{1}{3} \end{bmatrix}$$

$$x^{(2)} = Px^{(1)} = \begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \frac{3}{2} \\ \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{13}{12} \\ \frac{1}{9} \end{bmatrix}$$

$$\vdots$$

Si vuole dimostrare che se $\{x^{(k)}\}$ converge a x allora x è soluzione del sistema $Ax = b$

Teorema 4.1.1 (Fa la cosa giusta). *Dato $x^{(0)} \in \mathbb{R}$ sia $x^{(k+1)} = g(x^{(k)}), k \geq 0$. Se $lim_{k \to +\infty} x^{(k+1)} = x$ allora $g(x) = x$ e dunque $Ax = b$.*

Dimostrazione: consideriamo la funzione $g(x) = Px + q, \ g : \mathbb{R}^n \to \mathbb{R}^n$. La funzione g è continua infatti

$$\forall \epsilon > 0 \ \exists \delta > 0 : \ \forall x^{(k)}, x : \|x^{(k)} - x\| < \delta \text{ allora } \|g(x^{(k)}) - g(x)\| \leq \epsilon$$

per vedere la continuità si può osservare che

$$\|g(x^{(k)}) - g(x)\| = \|Px^{(k)} + q - (Px + q)\| = \|P(x^{(k)} - x)\| \leq \|P\|\|x^{(k)} - x\|$$

se si prende $\delta = \dfrac{\epsilon}{\|P\|}$ *abbiamo finito perché*

$$\|P\|\|x^{(k)} - x\| \leq \|P\|\delta = \epsilon$$

Quindi

$$x = \lim_{k\to+\infty} x^{(k+1)} = \lim_{k\to+\infty} g(x^{(k)}) = g(x)$$

l'uguaglianza $x = g(x)$ *si chiama* ***punto fisso*** *per* $g(x)$*.*

Un metodo iterativo è definito come

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ x^{(k+1)} = Px^{(k)} + q \end{cases}$$

Un metodo iterativo si dice **convergente** se tutte le successioni che ottengo variando il vettore iniziale $x^{(0)}$ sono convergenti (che sarà $x = A^{-1}b$). Esempio:

$$P = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad x^{(k+1)} = Px$$

se il metodo è convergente allora dal teorema sappiamo che $lim_{k\to\infty} x^{(k+1)} = x$ è tale che

$$x = Px \quad \Leftrightarrow \quad (I_n - P)x = 0$$

si nota che $det(I_n - P) \neq 0$ allora la soluzione è il vettore x tutto nullo. Quindi la successione $\{x^{(k)}\}$ dovrebbe tendere al vettore nullo se il metodo fosse convergente. Si controlla: $x^{(k+1)} = Px^{(k)}$ dato che $q = 0$ lo si può riscrivere come $Px^{(k)} = P(Px^{(k-1)}) = P^{k+1}x^{(0)}$ allora

$$P^{k+1} = \left(\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & -2 \end{bmatrix} \right)^{k+1} = \begin{bmatrix} \frac{1}{2^{(k+1)}} & 0 & 0 \\ 0 & \frac{1}{2^{(k+1)}} & 0 \\ 0 & 0 & (-2)^{(k+1)} \end{bmatrix}$$

se ora si scegliesse:

- $x^{(0)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad x^{(k)} = P^k x^{(0)} = \begin{bmatrix} \frac{1}{2^k} \\ 0 \\ 0 \end{bmatrix}$ e per $k \to +\infty$ è uguale al vettore nullo
- $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad \Rightarrow \quad x^{(k)} = P^k x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ (-2)^k \end{bmatrix}$ non è convergente

allora questo metodo iterativo non è convergente. Se la matrice P fosse

$$P = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

allora preso qualunque $x^{(0)} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ con $\alpha, \beta \in \mathbb{R}$ si avrebbe

$$x^{(k)} = \begin{bmatrix} \frac{1}{2^k} & 0 \\ 0 & \frac{1}{2^k} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{2^k}\alpha \\ \frac{1}{2^k}\beta \end{bmatrix}$$

che tende sempre al vettore nullo perché la successione non dipende da α e β. Allora è un metodo iterativo convergente.

Teorema 4.1.2 (**Condizione sufficiente per la convergenza del metodo**). *Il metodo*

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ x^{(k+1)} = Px^{(k)} + q \end{cases}$$

è convergente se esiste una norma matriciale indotta da una norma vettoriale $\|.\|$ su $\mathbb{R}^n$ tale per cui $\|P\| < 1$.

Dimostrazione: la quantità $e^{(k)}$ è il vettore dell'errore ed è definita come $e^{(k)} = x^{(k)} - x$. Basta dimostrare che se esiste $\|P\| < 1$ allora $lim_{k\to+\infty}\|e_k\| = 0$

$$e^{(k+1)} = x^{(k+1)} - x = Px^{(k)} + q - (Px + q) = P(x^{(k)} - x) = Pe^{(k)} = P(Pe^{(k-1)}) = P^{k+1}e^{(0)}$$

Sia $\|.\|$ la norma vettoriale tale che $\|P\| < 1$

$$\|e^{(k+1)}\| = \|P^{k+1}e^{(0)}\| \leq \|P^{k+1}\|\|e^{(0)}\| \leq \|P\|\|P^k\|\|e^{(0)}\| \leq \|P\|^{k+1}\|e^{(0)}\|$$

da cui per il teorema del confronto segue che $lim_{k\to+\infty}\|e^{(k+1)}\| = 0$

Teorema 4.1.3 (**Condizione necessaria per la convergenza del metodo**). *Se il metodo iterativo è convergente allora $\rho(P) < 1$.*

Dimostrazione: sia λ tale che $|\lambda| = \rho(P)$ e v un corrispondente autovettore di P cioè $Pv = \lambda v,\ v \neq 0$. Se il metodo è convergente allora converge qualunque io scelga $x^{(0)}$. Sia $x^{(0)} = x + v$ con $x = A^{-1}b$ soluzione del sistema lineare $Ax = b$. La successione generata dal metodo iterativo con punto iniziale $x^{(0)}$ è convergente ad x, infatti

$$e^{(k+1)} = P^{k+1}e^{(0)} = P^{k+1}v = \lambda^{k+1}v$$

da cui

$$\|e^{(k+1)}\| = \|\lambda^{k+1}v\| = |\lambda|^{k+1}\|v\|$$

e quindi

$$\lim_{k\to+\infty} |\lambda|^k = 0$$

che implica $|\lambda| < 1$

Condizione necessaria per la convergenza è che tutti gli autovalori di P sono in modulo < 1. Se $|detP| \geq 1$ allora almeno un autovalore di P è in modulo ≥ 1 allora non converge. Un'altra condizione è che se la traccia di P in modulo è $\geq n$ allora esiste un autovalore tale che $\|\lambda_i\| \geq 1$ e quindi non converge.

Teorema 4.1.4. *Sia $A \in \mathbb{R}^{n\times n}$ con $\rho(A) < 1$ allora esiste una norma matriciale indotta tale per cui $\|A\| < 1$.*

Combinando tra loro i teoremi 4.1.2, 4.1.3 e 4.1.4 si arriva infine al seguente risultato.

Teorema 4.1.5 (**Condizione necessaria e sufficiente per la convergenza del metodo**). *Condizione necessaria e sufficiente per la convergenza del metodo iterativo è che $\rho(P) < 1$.*

Dimostrazione: la condizione necessaria è già stata dimostrata con il teorema 4.1.3. Condizione sufficiente: se $\rho(P) < 1$ usando il teorema 4.1.4 si conosce che esiste una norma matriciale indotta tale che $\|P\| < 1$ quindi sono verificate le ipotesi per il teorema 4.1.2.

4.2 Metodi di Jacobi e Gauss-Seidel

Sia $A \in \mathbb{R}^{n\times n}$ matrice invertibile con elementi diagonali non nulli. Poniamo $A = D - L - U$ con $D = (d_{ij}), L = (l_{ij})$ e $U = (u_{ij})$ definite come segue:

$$d_{ij} = \begin{cases} a_{ij} & se\ i = j \\ 0 & altrimenti \end{cases}$$

$$l_{ij} = \begin{cases} -a_{ij} & se\ i > j \\ 0 & altrimenti \end{cases}$$

$$u_{ij} = \begin{cases} -a_{ij} & se\ i < j \\ 0 & altrimenti \end{cases}$$

Il **metodo iterativo di Jacobi** per la risoluzione del sistema lineare $Ax = b$ è definito dal partizionamento

$$M = D, \quad N = L + U \quad \Rightarrow \quad J = D^{-1}(L + U)$$

con J matrice di iterazione. Il **metodo iterativo di Gauss-Seidel** per la risoluzione del sistema lineare $Ax = b$ è definito dal partizionamento

$$M = D - L, \quad N = U \quad \Rightarrow \quad G = (D - L)^{-1}U$$

con G matrice di iterazione. Poiché per entrambi i metodi M risulta triangolare inferiore con elementi diagonali di A non nulli, questo garantisce l'applicabilità dei metodi. Per il metodo di Jacobi si ottiene $i = 1, 2, ..., n$

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)} \quad \Rightarrow \quad x_i^{(k+1)} = \frac{b_i - \sum_{j=1,j\neq i}^{n} a_{ij}x_j^{(k)}}{a_{ii}}$$

Per il metodo di Gauss-Seidel si ottiene per $i = 1, 2, ..., n$:

$$\sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = b_i - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \quad \Rightarrow \quad x_i^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}}{a_{ii}}$$

La differenza fondamentale tra i due metodi è che con il metodo di Jacobi si ha bisogno del vettore $x^{(k)}$ per il calcolo del vettore $x^{(k+1)}$, mentre, nel metodo di Gauss-Seidel è sufficiente disporre di un solo vettore in cui si sostituiscono le componenti via via che si calcolano.

4.3 Criteri di arresto

La risoluzione numerica del sistema lineare $Ax = b$ con un metodo iterativo richiede la determinazione di un **criterio di arresto** che consente di terminare l'elaborazione. Criteri usualmente utilizzati sono:

$$\|x^{(k+1)} - x^{(k)}\| \leq tol$$

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k+1)}\|} \leq tol$$

$$\|Ax^{(k+1)} - b\| \leq tol$$

$$\frac{\|x^{(k+1)} - b\|}{\|x^{(k+1)}\|} \leq tol$$

dove tol indica una tolleranza prefissata. In generale non abbiamo la garanzia che se uno dei criteri d'arresto è verificato allora $\|x^{(k+1)} - x\| \leq tol$, questo perché:

$$e^{(k+1)} = x^{(k+1)} - x = Pe^{(k)}$$

infatti:

$$x^{(k+1)} - x^{(k)} = \underbrace{x^{(k+1)} - x}_{e^{(k+1)}} - \underbrace{(x^{(k)} - x)}_{e^{(k)}} = Pe^{(k)} - e^{(k)} = (P - I_n)e^{(k)}$$

se il metodo è convergente allora $\rho(P) < 1$ e quindi $P - I_n$ è invertibile allora:

$$e^{(k)} = (P - I_n)^{-1}(x^{(k+1)} - x^{(k)})$$

passiamo ora alle norme:

$$\|e^{(k)}\| = \|(P - I_n)^{-1}(x^{(k+1)} - x^{(k)})\| \leq \|(P - I_n)^{-1}\|\|(x^{(k+1)} - x^{(k)})\| \leq \|(P - I_n)^{-1}\| \cdot tol$$

se la quantità $\|(P - I_n)^{-1}\|$ è grande allora anche l'errore può essere grande.

4.4 Convergenza dei metodi di Jacobi e Gauss-Seidel

Teorema 4.4.1. *Sia $A \in \mathbb{R}^{n\times n}$, se A è a predominanza diagonale allora:*

1. *A è invertibile;*
2. *i metodi di Jacobi e Gauss-Seidel per la risoluzione di un sistema lineare $Ax = b$ sono applicabili;*
3. *i metodi di Jacobi e Gauss-Seidel per la risoluzione di un sistema lineare $Ax = b$ sono convergenti.*

Dimostrazione: dimostriamo le tre proprietà

1. *L'invertibilità di A segue dal teorema di Gershgorin, infatti vale*

$$|0 - a_{ii}| = |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \;\; 1 \leq i \leq n$$

   dunque $0 \notin \bigcup_{i=1}^{n} K_i$ cioè zero non appartiene all'unione dei cerchi quindi non può essere autovalore e allora la matrice è non singolare.

2. *Per l'applicabilità si ha*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \geq 0 \quad \Rightarrow \quad |a_{ii}| \neq 0, \;\; 1 \leq i \leq n$$

3. *Consideriamo P matrice di iterazione. Il metodo è convergente se e solo se $\rho(P) < 1$ cioè $\forall\lambda$: $|\lambda| < 1$. Se λ è autovalore di P allora $det(P - \lambda I_n) = 0$*

$$det(P - \lambda I_n) = det(M^{-1}N - \lambda I_n) = det(M^{-1}N - \lambda(M^{-1}M)) = det(-M^{-1}(\lambda M - N)) = \underbrace{det(-M^{-1})}_{\neq 0} \cdot \underbrace{det(\lambda M - N)}_{= 0}$$

   quindi per gli autovalori di P vale che la matrice $H = \lambda M - N$ è singolare. Per assurdo assumiamo che esista un autovalore di P tale che $|\lambda| \geq 1$. Nel caso di Jacobi e Gauss-Seidel, la matrice $H = \lambda M - N$, sotto le ipotesi che $|\lambda| \geq 1$ e A a predominanza diagonale, risulta a sua volta a predominanza diagonale. La prima proprietà dice che le matrici a predominanza diagonale sono invertibili ma abbiamo un assurdo perché H invece è singolare quindi tutti gli autovalori devono essere per forza < 1 allora il metodo è convergente. Si mostra che la matrice H è a predominanza diagonale

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| = \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|, \;\; 1 \leq i \leq n$$

   implica

$$|\lambda||a_{ii}| \;\; > \;\; |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + |\lambda| \sum_{j=i+1}^{n} |a_{ij}| \;\; \underbrace{\geq}_{|\lambda| \geq 1} \;\; \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|, \;\; 1 \leq i \leq n$$

e quindi per Gauss-Seidel

$$|\lambda a_{ii}| > \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

e per Jacobi

$$|\lambda a_{ii}| > \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|$$

Queste ultime due relazioni esprimono la predominanza diagonale della matrice H che implica che H è una matrice invertibile e allora abbiamo l'assurdo cioè di aver assunto che $|\lambda| \geq 1$

Chapter 5

Il metodo delle potenze

5.1 Generalità sul metodo delle potenze

Data una matrice $A \in \mathbb{R}^{n\times n}$ cerchiamo $x, y \in \mathbb{C}^n, x, y \neq 0, \lambda \in \mathbb{C}$ tale che

$$Ax = \lambda x \leftarrow \textit{autovettore destro} \qquad y^H A = Ay^H \leftarrow \textit{autovettore sinistro}$$

dove $y^H = [\overline{y}_1, ..., \overline{y}_n]$ è il vettore riga che ha come componenti il coniugato di y cioè se vi è un certo $y_k = a + ib$ allora $\overline{y}_k = a - ib$. Il **metodo delle potenze** è un metodo iterativo per approssimare l'autovalore di massimo modulo di una matrice e il corrispondente autovettore. Per funzionare ha bisogno di alcune ipotesi:

1. assumiamo che A sia diagonalizzabile

   $$X^{-1}AX = D \quad \Leftrightarrow \quad AX = XD$$

   Con D matrice diagonale. Casi particolari:

   (a) se A è una matrice simmetrica allora A è diagonalizzabile in $\mathbb{R}$

   (b) se A ammette esattamente n autovalori distinti allora A è diagonalizzabile

2. assumiamo che A ha un solo autovalore di modulo massimo cioè

   $$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \geq 0$$

Si sceglie un $v^{(0)} \in \mathbb{R}$ e si costruisce una successione di vettori: $v^{(k+1)} = Av^{(k)},\ k \geq 0$ ottenendo:

$$v^{(1)} = Av^{(0)}$$

$$v^{(2)} = Av^{(1)} = A \cdot Av^{(0)} = A^2 v^{(0)}$$

$$\vdots$$

$$v^{(k+1)} = A^{k+1} v^{(0)}$$

Sotto le due ipotesi precedenti $v^{(k)}$ punta nella direzione di $x^{(1)}$. Dimostrazione:

$$\exists y \in \mathbb{C}^n, y = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} \textit{ tale che } v^{(0)} = Xy = \psi_1 x^{(1)} + ... + \psi_n x^{(n)} \textit{ allora}$$

$$\begin{aligned} Av^{(0)} &= AXy \\ &= A(\psi_1 x^{(1)} + \cdots + \psi_n x^{(n)}) \\ &= \psi_1 Ax^{(1)} + \cdots + \psi_n Ax^{(n)} \\ &= \psi_1 \lambda_1 x^{(1)} + ... + \psi_n \lambda_n x^{(n)} \qquad \textit{(dato che } Ax^{(i)} = \lambda_i x^{(i)}\textit{)} \end{aligned}$$

quindi

$$
\begin{aligned}
&v^{(1)} = Av^{(0)} = AXy = XDy \\
&v^{(2)} = Av^{(1)} = AXDy = XDDy = XD^2y = \psi_1\lambda_1^2x^{(1)} + ... + \psi_n\lambda_n^2x^{(n)} \\
&\vdots \\
v^{(k)} &= XD^ky \\
&= \psi_1\lambda_1^kx^{(1)} + \psi_2\lambda_2^kx^{(2)} + \cdots + \psi_n\lambda_n^kx^{(n)} \\
&= \lambda_1^k(\psi_1x^{(1)} + \psi_2(\tfrac{\lambda_2}{\lambda_1})^kx^{(2)} + \cdots + \psi_n(\tfrac{\lambda_n}{\lambda_1})^kx^{(n)})
\end{aligned}
$$

avevamo detto che $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$ quindi $|\frac{\lambda_i}{\lambda_1}| < 1$ per $i = 2, ..., n$ allora $v^{(k)}$ per k sufficientemente grande converge a $\lambda_1^k\psi_1x^{(1)}$ perché le quantità $|\frac{\lambda_i}{\lambda_1}| < 1$ tendono a zero.

Tuttavia ci sono due problemi:

1. ψ_1 potrebbe essere uguale a zero allora anche la quantità $\lambda_1^k\psi_1x^{(1)}$ è uguale a zero. Stessa cosa se $\psi_1 \approx 0$ si ha una convergenza molto lenta. Di questo problema in realtà non dobbiamo preoccuparci perché prendendo un vettore a caso con all'interno elementi floating point la situazione $\psi_1 = 0$ ha probabilità zero di verificarsi;
2. cosa succede se
   (a) $|\lambda_1| > 1$ allora λ_1^k diverge e potrei avere un errore di overflow
   (b) $|\lambda_1| < 1$ allora λ_1^k va a zero e quindi $v^{(k)}$ va a finire nel vettore nullo

Per risolvere il problema si creano delle varianti:

Variante 1: dato $v^{(0)} \neq 0$ si ha $v^{(k+1)} = \frac{v^{(k+1)}}{\lambda_1}$ quindi ripetendo i calcoli precedenti

$$
\begin{aligned}
v^k &= \tfrac{1}{\lambda_1^k}A^kXy \\
&= \tfrac{1}{\lambda_1^k}XD^ky \\
&= \psi_1x^{(1)} + \psi_1(\tfrac{\lambda_2}{\lambda_1})^kx^{(2)} + \cdots + \psi_n(\tfrac{\lambda_n}{\lambda_1})^kx^{(n)}
\end{aligned}
$$

cioè la successione $v^{(k)}$ tende a $\psi_1x^{(1)}$. Però noi non conosciamo λ_1 quindi "stiamo imbrogliando";

Variante 2: dato $v^{(0)} \in \mathbb{R}^n$ si ha $v^{(k+1)} = \frac{v^{(k+1)}}{\|v^{(k+1)}\|_2}$ quindi $v^{(k)}$ punta nella direzione di $x^{(1)}$ e ha lunghezza uno. Per il calcolo degli autovalori si utilizza il **quoziente di Rayleigh**:

$$\lim_{k\to\infty} \frac{v^{(k)T}Av^{(k)}}{v^{(k)T}v^{(k)}} = \lambda_1$$

5.2 Varianti del metodo delle potenze

Il metodo delle potenze approssima un singolo autovalore, cosa fare se servono altri autovalori?

- **metodo delle potenze inverse**: il metodo delle potenze può essere modificato per permettere l'approssimazione dell'autovalore di modulo minimo. Detti $\lambda_1, ..., \lambda_n$ gli autovalori di A ed assunto che

  $$|\lambda_1| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0$$

  allora gli autovalori di A^{-1} sono

  $$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geq \cdots \geq \frac{1}{|\lambda_1|}$$

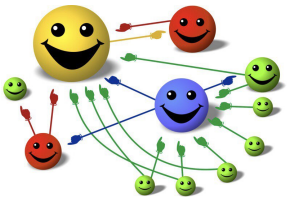
  quindi si calcola $v^{(k+1)} = A^{-1}v^{(k)}$ come $Av^{(k+1)} = v^{(k)}$ e si risolve il sistema lineare. Si potrebbe pensare anche di effettuare la fattorizzare LU di A per avere un vantaggio computazionale.

- per approssimare gli autovalori interni

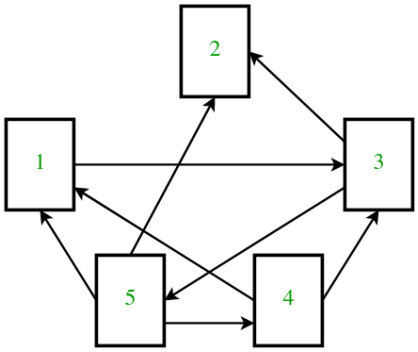
  $$|\lambda_n| \leq \ldots |\lambda_{i-1}| < |\lambda_i| < |\lambda_{i+1}| \leq \cdots \leq |\lambda_n|$$

  si applica il metodo delle potenze inverse alla matrice $A - \rho I_n$ in modo tale che $|\lambda_i - \rho|$ è il minimo tra gli autovalori $|\lambda_j - \rho|$.

5.3 Algoritmi di ranking per motori di ricerca



L'algoritmo **pagerank** di Google assegna un peso numerico ad ogni elemento (pagina) di un insieme di documenti uniti mediante collegamenti ipertestuali come il Web con lo scopo di quantificare la loro importanza relativa all'interno della raccolta. Il controllo dell'importanza di una pagina è sottratto al proprietario della pagina e affidato alla comunità. Esempio: supponiamo che il Web sia composta da cinque pagine e si costruisce la matrice di adiacenza



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

si vuole determinare il pagerank della pagina i, indicato con π_i. Ogni pagina distribuisce la propria importanza ugualmente a tutti i suoi vicini. Per esempio dalla pagina cinque ci si può muovere con una probabilità $\frac{1}{3}$ sulla pagina uno, due o quattro. Il vettore π_i staticamente viene definito come

$$\pi_j = \sum_{i \in \mathscr{I}(j)} \frac{\pi_i}{outdegree(i)}$$

con $\mathscr{I}(j)$ insieme degli archi entranti in j e $outdegree(i)$ il numero di archi uscenti dal nodo i. Esempio:

$$\pi_1 = \frac{\pi_4}{2} + \frac{\pi_5}{3} \qquad \pi_2 = \frac{\pi_3}{2} + \frac{\pi_5}{3} \qquad \pi_3 = \frac{\pi_1}{1} + \frac{\pi_4}{2} \qquad \pi_4 = \frac{\pi_5}{3} \qquad \pi_5 = \frac{\pi_3}{2}$$

Il vettore pagerank π può essere calcolato con il metodo delle potenze: si passa dalla matrice di adiacenza a una matrice P

$$G = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \quad \Rightarrow \quad P = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{bmatrix}$$

con $\sum_{j=1}^{n} P_{ij} = 1,\ i = 1, 3, 4, 5$ quindi $0 \leq P_{ij} \leq 1$ possono essere interpretati come probabilità.

5.3.1 Random Surfer Model

Per calcolare π si utilizza un modello iterativo: dato $\pi^{(0)}$ si calcola

$$\pi_j^{(k+1)} = \sum_{i \in \mathscr{I}(j)} \pi_i^{(k)} P_{ij},\ \ k = 0, 1, \ldots \qquad P_{ij} = \frac{1}{outdegree(i)}$$

che è equivalente a fare $\pi^{(k+1)} = P^T \pi^{(k)}$ e abbiamo visto che $\pi^{(k+1)}$ tende all'autovettore λ_1. Come prima per la convergenza del metodo delle potenze sono necessarie delle ipotesi: diagonalizzabilità e $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Occorre fare alcuni aggiustamenti per garantire queste proprietà: se applichiamo il teorema di Gershgorin abbiamo che poiché $P_{ii} = 0$

$$K_i = \{z \in \mathbb{C}^n : |z| \leq \textstyle\sum_{j=1}^{n} |P_{ij}|\}$$

quindi i cerchi hanno centro zero e raggio uno, oppure, raggio zero nel caso ci siano righe nulle allora $|\lambda| \leq 1$. Se la matrice P non ha righe nulle allora $\lambda_1 = 1$ perché

$$Pe = \underbrace{1}_{\lambda = 1} e$$

quindi posso forzare P ad essere stocastica (cioè a essere $Pe = e$)

$$\hat{P} = P + \frac{1}{n} d e^T \qquad d_i = \begin{cases} 1 & se\ outdegree(i) = 0 \\ 0 & altrimenti \end{cases}$$

$$\hat{P} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{bmatrix} + \frac{1}{5} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{bmatrix}$$

la matrice $\hat{P}$ è detta **stocastica per righe** e ha 1 come autovalore perché vale la relazione $\hat{P}e = 1e$. Chi assicura che l'autovalore di modulo massimo sia unico? Un teorema che si chiama **Perron-Frobenius** ci assicura che se la matrice è irriducibile (cioè il grafo è fortemente connesso) allora c'è un solo autovalore di modulo massimo. L'irriducibilità è forzata in un modo drastico, cioè si rende il grafo completo aggiungendo archi fittizi per ogni altro nodo in questo modo:

$$\overline{P} = \alpha \hat{P} + (1 - \alpha) \frac{e e^T}{n}$$

con probabilità α seguo i link, con probabilità $1 - \alpha$ vado in qualsiasi altro nodo del grafo. Ora abbiamo fatto tutti gli aggiustamenti perché

- $\overline{P}$ è stocastica: $\overline{P}e = \alpha \hat{P}e + (1 - \alpha)\frac{ee^t}{n}e = \alpha e + (1 - \alpha)\frac{e}{n}n = e$ quindi 1 è autovalore;
- $\overline{P}$ è irriducibile perché il grafo è fortemente connesso quindi 1 è l'unico autovalore di modulo massimo:
- non si ha bisogno di normalizzare perché $\lambda_1 = 1$.

Chapter 6

Approssimazione degli zeri di una funzione

6.1 Il metodo di bisezione

Sia $f : [a, b] \to \mathbb{R}$ con $f \in C^0([a, b]$ (cioè f continua su $[a, b]$) e $f(a)f(b) < 0$, dal teorema di esistenza degli zeri segue che $\exists \alpha \in [a, b]$ tale che $f(\alpha) = 0$. Per la determinazione di un tale α il **metodo di bisezione** genera sequenze di approssimazioni $\{a_k\}, \{b_k\}$ e $\{c_k\}$ definite come segue:

```
a(1) = a; b(1) = b;
for k>=1
    c(k) = (a(k) + b(k))/2;
    if (f(a(k)) * f(c(k))<=0)
        a(k+1) = a(k);
        b(k+1) = c(k);
    else
        a(k+1) = c(k);
        b(k+1) = b(k);
    end
end
```

Il metodo di bisezione procede suddividendo ad ogni passo l'intervallo $[a, b]$ a metà e determinando in quale dei due intervalli si trova la soluzione. La successione dei k_i converge ad α. Il procedimento viene interrotto imponendo un criterio di arresto:

$$b_k - a_k \leq tol$$

che garantisce $|\alpha - c_k| \leq \frac{tol}{2}$. Poiché $\frac{b-a}{2^k} \leq tol \Leftrightarrow 2^k \cdot tol \geq b - a \Leftrightarrow 2^k \geq \frac{b-a}{tol} \Leftrightarrow k \geq log_2(\frac{b-a}{tol})$ si ha che la condizione risulta soddisfatta dopo

$$k \geq \left\lceil log_2\left(\frac{b-a}{tol}\right)\right\rceil$$

iterazioni. Questo numero può essere significativamente elevato richiedendo molte valutazioni della funzione f.

6.2 Metodi di iterazione funzionale

Siano $f, g : [a, b] \to \mathbb{R}$, le equazioni $f(x) = 0$ e $g(x) - x = 0$ si dicono **equivalenti** se $f(\alpha) = 0 \Leftrightarrow g(\alpha) = \alpha$. In tal caso la radice α dell'equazione $f(x) = 0$ è detta **punto fisso** della funzione $g(x)$. La riformulazione del problema della ricerca delle soluzioni di un'equazione come il problema della ricerca dei punti fissi di una funzione associata conduce all'introduzione dei metodi di iterazioni funzionali del tipo

$$\begin{cases} x_0 \in [a, b] \\ x_{k+1} = g(x_k), \quad k \geq 0 \end{cases}$$

si ha infatti il seguente teorema

Teorema 6.2.1. *Se $g \in C^0([a,b])$, $x_k \in [a,b],\ \forall k > 0$ e $lim_{k\to+\infty} x_k = \alpha$ allora $g(\alpha) = \alpha,\ \alpha \in [a,b]$.*

Dimostrazione: segue da $\alpha = lim_{k\to+\infty} x_{k+1} = lim_{k\to+\infty} g(x_k)$ sapendo che g è continua si può scrivere $lim_{k\to+\infty} g(x_k) = g(lim_{k\to+\infty} x_k) = g(\alpha)$.

Sia $g : [a,b] \to \mathbb{R}$ con $g(\alpha) = \alpha,\ \alpha \in (a,b)$ allora il metodo

$$\begin{cases} x_0 \in [a,b] \\ x_{k+1} = g(x_k), \quad k \geq 0 \end{cases}$$

si dice **localmente convergente** in α se $\exists \rho > 0$ tale che $\forall x_0 \in [\alpha - \rho, \alpha + \rho]$ allora

1. ogni punto della successione rimane nell'intervallo: $x_k \in [\alpha - \rho, \alpha + \rho]$
2. la successione è convergente: $lim_{k\to+\infty} x_k = \alpha$

Un classico risultato che assicura la convergenza locale è il seguente teorema:

Teorema 6.2.2 (Teorema del punto fisso). *Sia $g : [a,b] \to \mathbb{R}, g \in C^1([a,b])$ (cioè g continua e ha derivata prima continua su $[a,b]$) e $g(\alpha) = \alpha,\ \alpha \in (a,b)$. Se $\exists \rho > 0$ tale che $|g'(x)| < 1, \forall x \in [\alpha - \rho, \alpha + \rho] \subseteq [a,b]$ allora $\forall x_0 \in [\alpha - \rho, \alpha + \rho]$ vale*

1. *$x_k \in [\alpha - \rho, \alpha + \rho]$*
2. *$lim_{k\to+\infty} x_k = \alpha$*

Dimostrazione: dal teorema di Weierstrass essendo $g'(x)$ continua e $[\alpha - \rho, \alpha + \rho]$ chiuso e limitato abbiamo $\lambda = \max_{x\in[\alpha-\rho,\alpha+\rho]} |g'(x)| < 1$. Si dimostra che

$$|x_k - \alpha| \leq \lambda^k \rho,\ \forall k \geq 0$$

da cui segue la proprietà 1

$$|x_k - \alpha| \ \leq\ \lambda^k \rho \ \leq\ \rho \ \Rightarrow\ x_k \in [\alpha - \rho, \alpha + \rho]$$

e la proprietà 2 per il teorema del confronto

$$0 \ \leq\ |x_k - \alpha| \ \leq\ \lambda^k \rho \ \Rightarrow\ \lim_{k\to+\infty} x_k = \alpha$$

Passiamo a dimostrare che $|x_k - \alpha| \leq \lambda^k \rho$ per induzione su k

Caso base: per $k = 0$ si ha $|x_0 - \alpha| \leq \lambda^0 \rho = \rho$ ed è vero perché $x_0 \in [\alpha - \rho, \alpha + \rho]$.

Passo induttivo: per $k + 1$ si ha

$$|x_{k+1} - \alpha| = |g(x_k) - g(\alpha)|$$

ora per il teorema di Lagrange

$$|g(x_k) - g(\alpha)| = |g'(\eta_k)(x_k - \alpha)| = |g'(\eta_k)||(x_k - \alpha)|, \quad |\eta_k - \alpha| \leq |x_k - \alpha|$$

per ipotesi induttiva segue che $\eta_k \in [\alpha - \rho, \alpha + \rho]$ e dunque

$$|g'(\eta_k)||(x_k - \alpha)| \leq \lambda\lambda^k \rho = \lambda^{k+1} \rho$$

Dal teorema segue il seguente corollario:

Teorema 6.2.3. *Sia $g : [a,b] \to \mathbb{R}, g \in C^1([a,b])$ e $g(\alpha) = \alpha,\ \alpha \in (a,b)$. Se $|g'(\alpha) < 1|$ allora il metodo*

$$\begin{cases} x_0 \in [a,b] \\ x_{k+1} = g(x_k), \quad k \geq 0 \end{cases}$$

è localmente convergente in α.

Dimostrazione: sia $h : [a,b] \to \mathbb{R}, h(x) = |g'(x)| - 1$. *Si ha che* $h \in C^0([a,b]), h(\alpha) = |g'(\alpha)| - 1 < 0$ *e dunque per il teorema della permanenza del segno* $\exists \rho > 0$ *tale che* $\forall x \in [\alpha - \rho, \alpha + \rho], h(x) < 0$ *cioè* $|g'(x)| < 1$.

6.3 Ordine di convergenza

L'**ordine di convergenza** indica con quale velocità la successione $\{x_k\}$ è convergente alla soluzione. Viene definita come: sia $\{x_k\}$ convergente ad α punto fisso per una certa $g(x)$ e, se $x_k \neq \alpha$ e esiste un valore $p \geq 1$ tale che

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} = \gamma \quad \textit{con} \quad \begin{cases} 0 < \gamma \leq 1 & \textit{se } p = 1 \\ \gamma > 0 & \textit{se } p > 1 \end{cases}$$

allora si dice che la successione converge con ordine p. Si parla di

- **convergenza lineare** se $p = 1$ e $0 < \gamma < 1$;
- **convergenza sublineare** se $p = 1$ e $\gamma = 1$, la successione va a zero molto lentamente;
- **convergenza quadratica** se $p = 2$.

6.4 Metodo delle tangenti

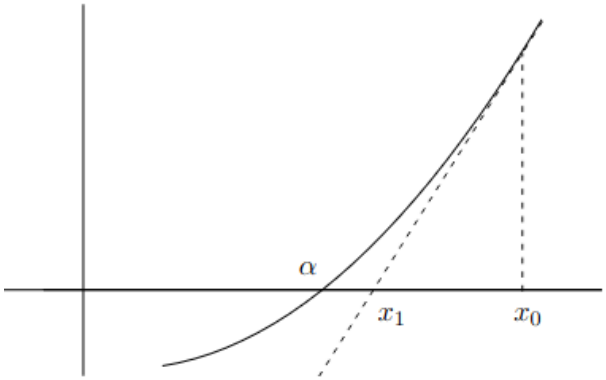
Il più noto metodo di iterazione funzionale è **metodo delle tangenti** o **di Newton**. Lo si può applicare solo se la funzione $f(x)$ è derivabile nell'intervallo $[a,b]$. Il metodo procede nel seguente modo: si sceglie nell'intervallo $[a,b]$ un punto che chiameremo x_0 e si considera la retta tangente alla curva nel punto $(x_0, f(x_0))$ di equazione

$$y = f(x_0) + f'(x_0)(x - x_0)$$

Il punto x_1 in cui la retta interseca l'asse delle x si ottiene ponendo $y = 0$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Se il punto x_0 è stato scelto opportunamente il punto x_1 appartiene ancora all'intervallo $[a,b]$ e risulta compreso fra α e x_0.



Il metodo prosegue sostituendo x_1 al posto di x_0 e calcolando un altro punto x_2 e così via. Quindi vale la formula

$$\begin{cases} x_0 \in [a,b] \\ x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)} \end{cases} \quad \textit{con } f'(x_k) \neq 0$$

Sia α tale $f(\alpha) = 0$ è detta **radice semplice** per una certa funzione $f \in C^1([a,b])$ se $\alpha \in (a,b)$ e $f'(\alpha) \neq 0$. Esempi:

$f(x) = x^2 - 4 = (x-2)(x+2)$ ha come radici $\alpha \pm 2$ che sono radici semplici. Infatti se si calcola la derivata $f'(x) = 2x$ essa non si annulla né in 2 né in -2

$f(x) = (x-2)^2$ ha come radice $\alpha = 2$. Si calcola la derivata $f'(x) = 2(x-2)$ e si nota che si annulla con $\alpha = 2$. Questo significa che α ha molteplicità 2 cioè non è semplice

Il metodo delle tangenti è localmente convergente nel caso di approssimazioni di radici semplici e la convergenza è quadratica.

Teorema 6.4.1. *Sia $f : [a,b] \to C^2([a,b])$ (cioè f è continua e ha derivata prima e seconda continua su $[a,b]$) e $f(\alpha) = 0,\ \alpha \in (a,b)$. Se α è semplice cioè $f'(\alpha) \neq 0$ allora*

1. *il metodo*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

è localmente convergente ad α;

2. *se $x_k \neq \alpha, \forall k \geq 0$ allora la convergenza è almeno quadratica cioè*

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = l \in \mathbb{R}$$

se $l \neq 0$ allora vi è una convergenza quadratica; se $l = 0$ allora vi è una convergenza più che quadratica.

<u>Dimostrazione</u>: poiché $f'(\alpha) \neq 0$ allora $f'(\alpha)$ sarà > 0 o < 0. Quindi per il teorema della permanenza del segno $\exists \eta > 0$ tale che $\forall x \in [\alpha - \eta, \alpha + \eta]$, $f'(x) \neq 0$. Ora si definisce $g(x)$ come

$$\begin{cases} g : [\alpha - \eta, \alpha + \eta] \to \mathbb{R} \\ g(x) = x - \dfrac{f(x)}{f'(x)} \end{cases}$$

che sull'intervallo $[\alpha - \eta, \alpha + \eta]$ la $g(x)$ è ben definita. Inoltre $g \in C^1([\alpha - \eta, \alpha + \eta])$ e

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

Poiché $g'(\alpha) = 0$ dal teorema 6.2.3 segue che il metodo è localmente convergente. Ora dimostriamo il secondo punto: dallo sviluppo di Taylor applicato a $f(x)$ nel punto x_k

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + f''(\xi_k)\frac{(x - x_k)^2}{2!} \qquad |\xi_k - x| \leq |x_k - x|$$

Per $x = \alpha$ abbiamo

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + f''(\hat{\xi}_k)\frac{(\alpha - x_k)^2}{2} \qquad |\hat{\xi}_k - \alpha| \leq |x_k - \alpha|$$

da cui

$$x_{k+1} - \alpha = \frac{f''(\hat{\xi}_k)}{2f'(x_k)}(x_k - \alpha)^2$$

da cui si ricava per continuità di $f'(x)$ e $f''(x)$

$$\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^2} = \lim_{k \to +\infty} \frac{|f''(\hat{\xi}_k)|}{|2f'(x_k)|} \in \mathbb{R}$$

Teorema 6.4.2 (Teorema di convergenza in largo). *Sia* $f : [a, b] \to \mathbb{R}, f \in C^2([a, b]), f(\alpha) = 0, a \in (a, b)$. *Se* $\exists \delta > 0$ *tale che* $\forall x \in (\alpha, \alpha + \delta] \subset [a, b]$ *si ha*

1. $f'(x) \neq 0$
2. $f(x)f''(x) > 0$

allora il metodo delle tangenti con $x_0 \in (\alpha, \alpha + \delta]$ *genera successioni convergenti ad* α.