Università di Pisa

Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Corso 2° anno - 6 CFU

Calcolo Numerico

Professore:
Prof. Luca Germignani

Autore:
Matteo Giuntoni
Filippo Ghirardini

Anno Accademico 2023/2024

Contents

Calcolo Numerico

Realizzato da: Giuntoni Matteo e Filippo Ghirardini

A.A. 2023-2024

---

1 Aritmetica di Macchina

Per una macchina la scrittura $(x+y)+z$ è diverso da $x+(y+z)$. Vediamo dunque che ci sono alcuni punti focali da considerare per far si che una macchina funzioni correttamente:

- Trovare uno standard per come **memorizzare** i numeri.
- Trovare uno standard per come **manipolare** i numeri.

Da questi due punti possiamo ricondurci ad un solo problema, come andare a **rappresentare** i numeri.

1.1 Teorema di rappresentazione

Teorema 1.1.1. Dato $x \in \mathbb{R}, x \neq 0$[1] e una base di numerazione $B, B \in \mathbb{N}, B > 1$ esistono e sono univocamente determinati:

1. Un intero $p \in \mathbb{Z}$ detto **esponente** della rappresentazione
2. Una successione di numeri naturali $\{d_i\}_{i=1}^{+\infty}$ con $d_i \neq 0, 0 \leq d_i \leq B-1$ e d_i non definitivamente uguali a $B-1$, dette cifre della rappresentazione tali per cui x si scrive in modo **unico** nella seguente forma:
$$x = sign(x)B^p \sum_{i=1}^{+\infty} d_i B^{-1}. \tag{1}$$
dove la sommatoria viene chiamata **mantissa**

Esempio 1.1.1 (Esempio in base 10)**.** Poniamo come numero da rappresentare 0.1 in base 10 è

$$0.1 = +10^0(0.1)$$

Andiamo ora ad analizzare il significato di questo teorema. Esso descrive quella che viene chiamata **rappresentazione in virgola mobile**, in quanto l'esponente p on è determinato in modo da avere la parte intera nulla. Le cose da considerare in questo teorema sono:

- La condizione $d_i \neq 0$ e d_i non definitivamente uguale a $B-1$ sono introdotte per garantire l'unicità delle rappresentazioni. Ad esempio:

  $$B = 10 \text{ abbiamo } 1 = +10^1(1 \cdot 10^{-1}) = +10^2(0 \cdot 10^{-1} + 1 \cdot 10^{-1})$$

  Quindi due rappresentazioni diverse per lo stesso numero, però considerando le condizioni scritte sopra la seconda non risulta accettabile perché la prima cifra è nulla.
  Questa clausola ci garantisce anche l'unicità delle rappresentazioni nei numeri **periodici**:

  $$0.\overline{9} = 10^0(0.99\ldots9)$$

  Non è ammissibile in quanto è definitivamente uguale a $B-1$.

- Questa rappresentazione si estende anche all'insieme dei numeri complessi del tipo $z = a + ib$, utilizzando una rappresentazione come coppie di numeri reali del tipo (a, b).

Possiamo dedurre che visto che stiamo lavorano con registri di memoria di un calcolatore con memoria a numero finito, anche la quantità di cifre rappresentabili saranno a numero finito esso viene chiamato **insieme dei numeri di macchina**.

Dal teorema di rappresentazione in base di un numero reale può avvenire assegnando delle posizioni di memoria per il segno, per l'esponente e per le cifre della rappresentazione.

[1] Lo zero viene utilizzato dalla macchina per alcune operazioni come il confronto, quindi deve averlo ma lo rappresenta in un modo particolare

Definizione 1.1.1 (Insieme dei numeri di macchina)**.** *Si definisce l'insieme dei numeri di macchina in rappresentazione floating point con t cifre, base B e range $-m, M$ l'insieme dei numeri reali.*

$$\mathbb{F}(B,t,m,M) = \{0\} \cup \{s \in \mathbb{R} : x = \pm B^p \sum_{i=1}^{t} d_i B^{-1}, d_1 \neq 0, -m \leq p \leq M\}$$

Si osserva in questa definizione che:

- L'insieme $\mathbb{F}$ ha cardinalità **finita**: $N = 2B^{t-1}(B-1)(M+m+1)+1$.
- L'insieme dei numeri di macchina $\mathbb{F}(B,t,m,M)$ è **simmetrico** rispetto all'origine.
- Possiamo definire $\Omega = B^M \sum_{i=1}^{t}(B-1)b^{-1}$ come il **più grande** numero macchina e $\omega = +B^{-m}B^{-1}$ come invece il **più piccolo** positivo.
- Posto un $x = B^P \sum_{i=1}^{t} d_i B^{-1}$ possiamo definire il suo **successivo** numero di macchina come $y = B^p(\sum_{i=1}^{t-1} d_i B^{-1} + (d_t+1)B^{-t})$.
  Da qui vediamo che la distanza $y - x = B^p - t$ porta i numeri ad essere **non equidistanti** fra di loro, quindi la distanza aumenta con l'avvicinarci a Ω.
  Questo ci fa comodo perché ci interessa l'**errore relativo**, quindi su numeri piccoli ci serve un errore piccolo mentre su numeri grandi posso fare errori grandi.

Esempio 1.1.2. Facciamo ora un esempio in cui andiamo a rappresentare il numero successivo di $x = B^p \sum_{t}^{i=1} d_i B^{-1}$. Esso si può scrivere come $y = B^p\left(\sum_{i=1}^{t-1} d_i B^{-1} + (d_t+1)B^{-t}\right)$.
Mentre si può scrivere la distanza fra questi due valori come $y - x = B^p - t$.

E' stato fissato uno standard IEEE 754 negli anni 70/80 che dice che, visto ci sono macchine che hanno metodi di rappresentazione diversi, bisogna fissare un standard, ovvero $B = 2$ con registri a 32 o 64 bit.

Questa rappresentazione ha uno svantaggio che può sembrare minimo ma non lo è, lo 0 si rappresenta due volte con $-0, +0$. Per ovviare a questo problema si è andato ad abbandonare questa rappresentazione in esponenti ma si rappresentato i numeri nel seguente modo: $p_1 2^0 + p_1 2^1 + \cdots + p_1 1 s^1 0$ che rappresentano numeri da 0 a $2^1 1 - 1$ quindi 2047 numeri, mentre lo 0 si può scrivere come:

- O tenendo tutti i valori a 0
- Oppure tendendo tutti i valori a 1

In entrambi i casi abbiamo un range di valori che va da $[-1022, 1024]$. A questo punto ho 2^{P-1022} numeri che la macchina rappresenta come $\pm 2^{P-1022}(0.1d_1 \dots d_{52})$.
Impostando questo standard abbiamo $\Omega = 2^{1024}(01\dots1)_2$ e $\omega = 2^{-1022}(101)_2$.

Osservazione 1.1.1. Quando $p = 0$ abbiamo i numeri che si trovano nella porzione della retta dei numeri che è compresa fra $-\omega$ ω e possiamo qui avere anche tutti 0 e quindi si introduce il caso dei numeri denormalizzati.

Se abbiamo l'esponente uguale a tutti 1, la convenzione è che tutte le cifre della mantissa sono tutti uguali a 0/1 questo numero indica il $\pm\infty$ altrimenti sta a significare NaN (not a number). Questi valori ci permettono di gestire forme indeterminate.

1.2 Errore di rappresentazione

Quando si va a rappresentare un numero reale non nullo $x \in \mathbb{R}$ e con $x \neq 0$ si può andare a commettere degli errori di rappresentazione detto anche **errore relativo di approssimazione**, e di definisce come, prendendo un $\tilde{x} \in \mathbb{F}(B,t,m,M)$

$$\epsilon_x = \frac{\tilde{x} - x}{x} = \frac{\eta x}{x}, x \neq 0$$

Definiamo $|\epsilon_x| = \left|\frac{\tilde{x}-x}{x}\right| \leq \frac{B^{P-t}}{|x|} \leq \frac{B^{P-t}}{B^{P-1}} = B^{1-t} = u$ la u è definita come **precisione di macchina**.

Andiamo inoltre a definire le condizioni di underflow e overflow. Dato un $x \in \mathbb{R}, x \neq 0$ abbiamo che:

1. Se $|x| < \omega$ o $|x| > \Omega$ overflow. In questo caso si va ad associare il $+\infty$.
2. Se invece $\omega \leq |x| \leq \Omega$ abbiamo underflow. In questo caso allora prendiamo una $x = B^p \sum_{i=1}^{\infty} d_i B^{-1} \longrightarrow B^p \sum_{i=1}^{t} d_i B^{-1} = \tilde{x}$ che è una approssimazione

1.3 Operazioni di macchina

Consideriamo ora due numeri $x, y \in \mathbb{F}$ e chiediamoci perché la macchina non possa fare l'operazione $x + y$. La risposta è che i risultati da questa operazione di ritornano fra i numeri di macchina. Per ovviare a questo problema dovremo usare le Operazioni di macchina che si identificano come $\oplus \ominus \otimes \oslash$. Nel nostro caso l'addizione di macchina $x \oplus y = troncamento(x+y) = (x+y)(1+\epsilon_1)$ con $|\epsilon_1| \leq u$ con e_1 detto errore locale dell'operazione.

Esempio 1.3.1. Supponiamo di dover calcolare in macchina la funzione $f(x) = \frac{x-1}{x}$. In macchina questa funzione corrisponderebbe a $g(\tilde{x}) = (\tilde{x} \ominus 1) \oslash \tilde{x}$. Abbiamo quindi:

$$g(\tilde{x}) = \frac{(x(1+\epsilon_x)-1)(1+\epsilon_1)}{x(1+\epsilon_x)} \cdot (1+\epsilon_1) = \frac{(x(1+\epsilon_x)-1)(1+\epsilon_1+\epsilon_2)}{x(1+\epsilon_x)} = \frac{(x(1+\epsilon_x)-1)(1+\epsilon_1+\epsilon_2-\epsilon_x)}{x}$$

$$g(\tilde{x}) = (\tilde{x} \oplus 1) \oslash \tilde{x} = \frac{(x(1+\epsilon_x)-1)(1+\epsilon_1+\epsilon_2-\epsilon_x)}{x}$$

$$\frac{g(\tilde{x}-f(x))}{f(x)} = \frac{((x-1)/x) + (\epsilon_1+\epsilon_2)((x-1)/x) + \epsilon_2/x - ((x-1)/x)}{(x-1)/x} = \epsilon_1 + \epsilon_2 - \frac{\epsilon_x}{x-1}$$

Esempio 1.3.2. Supponiamo ora di calcolare la funzione $f(x) = \frac{x-1}{x}$ in un altro modo, $g_2(\tilde{x}) = \frac{g_2(\tilde{x})-f(x)}{f(x)}$ ed andiamo a fare l'analisi dell'errore.

$$\frac{g_1(\tilde{x}) - f(x)}{f(x)} \doteq \epsilon_1 + \epsilon_2 + \frac{\epsilon_1}{(x-1)} \text{ Questo è il risultato di un analisi al primo ordine}$$

$$g_2(\tilde{x}) = 1 \ominus \frac{1}{\tilde{x}}(1+\delta_1) = 1 \ominus \frac{1}{x}(1+\delta_1)(1-\epsilon_x) = [1 - \frac{1}{x}(1-\delta_1)(1-\epsilon_1)](1+\delta_2)$$

$$\doteq (1+\delta_1) - \frac{1}{x}(1+\delta_1+\delta_2+\epsilon_x) \doteq (1-\frac{1}{x}) + \delta_2(1-\frac{1}{x}) - \frac{\delta_1}{x} + \frac{\epsilon_x}{x} \doteq \delta_2 - \frac{\delta_1}{x-1} + \frac{\epsilon_x}{x-1}$$

$$\frac{g_2(\tilde{x}) - f(x)}{f(x)} = \delta_2 - \frac{\delta_1}{x-1} + \frac{\epsilon_x}{x-1}$$

Questo è il risultato finale dove $\delta_2 - \frac{\delta_1}{x-1}$ viene definita come parte stabilità mentre $\delta_2 - \frac{\delta_1}{x-1}$ viene chiamato condizionamento,il risultato finale viene definito invece numero stabile.

2 Calcolo degli errori

Supponiamo di avere una funzione $f : [a, b] \to \mathbb{R}$ e $f \neq 0$, per andare a calcolare questa funzione come già visto usiamo un algoritmo che esprime tale valore come risutlato di una sequanza di operazioni aritmentiche. Questa rappresentazione come abbiamo già potuto verificare con esempi produce degli errori di approssimazione. Questi errori possono essere suddivisini in 3 tipologie.

Definizione 2.0.1 (Errore inerente o inevitabile)**.** *Si dice errore inerente o inevitabile generato nel calcolo di $f(x) \neq 0$ la quantità:*

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

Definizione 2.0.2 (Errore algoritmico)**.** *Si dice errore algoritmico generato nel calcolo di $f(x) \neq 0$ la quantità:*

$$\epsilon_{alg} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$$

Definizione 2.0.3 (Errore totale)**.** *Si dice errore algoritmico totale nel calcolo di $f(x) \neq 0$ mediante l'algoritmo specificato da g la quantità:*

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

Osservazione 2.0.1. Vediamo che se $|\epsilon_{in}|$ è grande il problema si definisce **problema mal condizionato**. Mentre se $|\epsilon_{alg}$ è grande l'algoritmo si dice che **algoritmo è numericamente instabile**.

3 Matrici

3.1 Norme vettoriali

Sono uno strumento che ci permette di definire una distanza tra due vettori.

Definizione 3.1.1 (Norma vettoriale)**.** *È una funzione del tipo* $f : \mathbb{R}^n \to \mathbb{R}$ *che deve soddisfare tre proprietà:*

1. $f(x) \geq 0 \wedge f(x) = 0 \Leftrightarrow x = 0$
2. $f(\alpha x) = |\alpha| f(x) \quad \forall \alpha \in \mathbb{R} \quad \forall x \in \mathbb{R}^n$
3. ***Disuguaglianza triangolare****:* $f(x + y) \leq f(x) + f(y) \quad \forall x, y \in \mathbb{R}^n$

e la indichiamo come

$$f(x) = \|x\|$$

Detto questo possiamo definire una distanza come:

$$dist(x, y) = \|x - y\| \tag{2}$$

Le tre proprietà ci danno alcune informazioni sulla distanza:

1. La distanza deve essere non negativa e valere 0 solo se i due vettori coincidono
2. La distanza tra x e y deve essere uguale a quella tra y e x
3. $\|x - y\| = \|(x - a) + (a - y)\| \leq \|x - a\| + \|a - y\|$

Definizione 3.1.2 (Distanza euclidea - Norma 2)**.**

$$f(x) = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} > \|x\|_2 \tag{3}$$

Definizione 3.1.3 (Norma infinito)**.**

$$f(x) = max|x_1| = \|x\|_\infty \tag{4}$$

Definizione 3.1.4 (Norma 1)**.**

$$f(x) = \sum_{i=1}^{n} |x_i| = \|x\|_1 \tag{5}$$

Esempio 3.1.1. Prendiamo due vettori

$$\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} \quad \begin{Bmatrix} 2 \\ -2 \end{Bmatrix}$$

e calcoliamo le varie norme:

$$\|x\|_2 = \sqrt{2} \quad \|x\|_\infty = 1 \quad \|x\|_1 = 2$$

Definizione 3.1.5 (Norma matriciale)**.** *È una funzione del tipo* $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ *che deve soddisfare tre proprietà:*

1. $f(A) \geq 0 \wedge f(A) = 0 \Leftrightarrow A = 0$
2. $f(\alpha A) = |\alpha| f(x) \quad \forall \alpha \in \mathbb{R} \quad \forall A \in \mathbb{R}^{n \times n}$
3. ***Disuguaglianza triangolare****:* $f(A + B) \leq f(A) + f(B) \quad \forall A, B \in \mathbb{R}^{n \times n}$
4. $f(A \cdot B) \leq f(A) f(B)$

e la indichiamo come

$$f(A) = \|A\|$$

4 Condizionamento

Studiare il condizionamento di un problema in forma $Ax = b$ significa chiedersi di quanto cambia la soluzione perturbando di poco A e B.

4.1 Metodi diretti

Dato un sistema lineare $Ax = b$ le soluzioni possono essere trovate tramite

$$x_i = f_i(a_{11}, \dots, a_{1n}, b1, \dots, b_n) \quad i : 1 \dots n$$

Si può partire studiando la forma di A per cercare dei sistemi risolvibili facilmente.

4.1.1 Matrice diagonale

Il primo caso è quando A è una matrice **diagonale**:

$$a_{ij} = 0 \Longleftarrow i \neq j$$

Il determinante di una matrice diagonale è $det(A) = \prod_{i=1}^{n} a_i$ e il determinante è diverso da 0 solo se tutti gli elementi della diagonale lo sono. Possiamo riscrivere la matrice in un sistema di equazioni lineari:

$$Ax = b \Leftrightarrow \begin{cases} a_1 x_1 = b_1 \Leftrightarrow x1 = \frac{b1}{a1} \\ \vdots \\ a_n x_n = b_n \Leftrightarrow x_n = \frac{b_n}{a_n} \end{cases}$$

Questo sistema lo risolviamo in $O(n)$. Dato però che ridurre una matrice normale in una diagonale è un processo complesso, non conviene farlo.

4.1.2 Matrice triangolare

Una matrice è **triangolare inferiore** quando $a_{ij} = 0$ per $j > i$, mentre è **triangolare superiore** quando $a_{ij} = 0$ per $i > j$.
Per calcolare il determinante di una matrice triangolare superiore con Laplace facciamo:

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix}$$

Per risolvere $Ax = b$ con una matrice triangolare, vediamo che il sistema associato ci porta ad avere prima un equazione con tutte le incognite fino all'ultima con una sola incognita. È intuitivo che per calcolarne la soluzione è sufficiente partire dall'ultima ed eseguire una **sostituzione all'indietro**.

```
function[x]=backward_substitution(a,b)
   n=length(b);
   x = zeros(n,1);
   for k=n:-1:1
      s=0;
      for j=k+1:n
         s=s+a(k,j)*x(j);
      end
      x(k)=(b(k)-s)/a(k,k);
   end
```

Dato che l'operazione moltiplicativa in macchina richiede leggermente più lunga di quella della somma, è sufficiente considerare le prime per calcolare la complessità di questo programma.
La complessità al caso pessimo è quindi $\frac{n \cdot (n+1)}{2} = O(n^2) = \frac{n^2}{2} + O(n)$.

Esempio 4.1.1 (Matrice bidiagonale superiore)**.** Supponiamo di avere la seguente matrice bidiagonale superiore:

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & a_{n-1n} \\ 0 & 0 & 0 & a_{nn} \end{bmatrix} = \begin{bmatrix} a_1 & b_1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & b_{n-1} \\ 0 & 0 & 0 & a_n \end{bmatrix}$$

In questo caso il codice si può cambiare sostituendo il secondo ciclo con l'operazione

```
s=a(k,k+1)*x(k+1);
```

e la complessità diventa $O(n)$.

5 Metodi iterativi

Una classe di metodi che permette di risolvere sistemi lineari è quella dei metodi iterativi. L'idea è di costruire una successione di vettori partendo dalla matrice che convergano alla soluzione del sistema lineare.

$$\{X_k\}_{k\in\mathbb{N}}$$

Osservazione 5.0.1. Non potendo generare infiniti termini, ad un certo punto ci sarà bisogno di fermarsi quando si pensa di essere sufficientemente vicini alla soluzione.

Diciamo che

$$\lim_{k\to+\infty} x_k \Leftrightarrow \lim_{k\to+\infty} ||x_k - x||_\infty = 0 \tag{6}$$

Questo è vero perché:

$$\forall j = 1,\dots,n \quad 0 \le |x_j^{(k)} - x_j| \le ||x^{(k)} - x||_\infty$$

Abbiamo una matrice invertibile A. Supponiamo di scomporre la matrice come:

$$A = M - N \tag{7}$$

con l'assunzione che anche M sia invertibile ($det(M) = 0$). A questo punto possiamo dire che:

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

Dovendo trovare la soluzione di $x = Px + q$, possiamo scegliere un vettore iniziale $x_0 \in \mathbb{R}^n$ e costruirci la successione di vettori

$$x_{k+1} = Px_k + q \tag{8}$$

Se questa successione converge ho trovato la soluzione del sistema lineare.

Osservazione 5.0.2. Nell'implementazione pratica non userò mai l'equazione 8 ma invece

$$Mx_{k+1} = Nx_k + b \tag{9}$$

5.1 Convergenza

Esempio 5.1.1. Prendiamo

$$\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} \quad b = \begin{bmatrix}0\\0\end{bmatrix} \quad Ax = b \Leftrightarrow x = \begin{bmatrix}0\\0\end{bmatrix}$$

Partiamo definendo le due matrici per la scomposizione:

$$M = \begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix} \quad N = \begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix}$$

e calcolando la matrice P

$$P = M^{-1}N = \begin{bmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}0 & -1\\ -1 & 0\end{bmatrix} = \begin{bmatrix}0 & -\frac{1}{2}\\ -\frac{1}{2} & 0\end{bmatrix}$$

e il vettore q Iniziamo il metodo iterativo:

$$x^{k+1} = Px^{(k)} = \begin{bmatrix}0 & -\frac{1}{2}\\ -\frac{1}{2} & 0\end{bmatrix}x^{(k)}$$
$$x^{(1)} = \begin{bmatrix}0 & -\frac{1}{2}\\ -\frac{1}{2} & 0\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}-\frac{1}{2}y\\ -\frac{1}{2}x\end{bmatrix}$$
$$x^{(2)} = \begin{bmatrix}0 & -\frac{1}{2}\\ -\frac{1}{2} & 0\end{bmatrix}\begin{bmatrix}-\frac{1}{2}y\\ -\frac{1}{2}x\end{bmatrix} = \begin{bmatrix}\frac{1}{4}x\\ \frac{1}{4}y\end{bmatrix}$$

e notiamo che la successione tende a

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

Se prendiamo invece

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad N = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

con

$$P = M^{-1}N = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & -2 & -2 & 0 \end{bmatrix}$$

abbiamo che la successione **diverge**.

Definizione 5.1.1. *Un metodo iterativo è convergente se*

$$\forall x^{(0)} \in \mathbb{R}^n \quad x_k \to x \tag{10}$$

Ovvero se per ogni vettore di partenza scelto, il metodo converge.

Teorema 5.1.1. Dato $x^{(k+1)} = Px^{(k)} + q$, se

$$\exists ||\cdot|| \text{ tc } ||P|| < 1 \tag{11}$$

allora il metodo è convergente.

Esempio 5.1.2. Data una matrice

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

definiamo le matrici per la scomposizione

$$M = \begin{bmatrix} 3 & & \\ & \ddots & \\ & & 3 \end{bmatrix} \quad N = \begin{bmatrix} 0 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix}$$

Per capire se il metodo è convergente dobbiamo calcolare la norma infinito di P:

$$M^{-1}N = \begin{bmatrix} 0 & \frac{1}{3} & & \\ \frac{1}{3} & \ddots & \ddots & \\ & \ddots & \ddots & \frac{1}{3} \\ & & \frac{1}{3} & 0 \end{bmatrix} \quad ||P||_\infty = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} < 1$$

Il problema di questo metodo iterativo è che ha complessità $O(n)$ per ogni iterazione e non è quindi competitivo con l'eliminazione Gaussiana.

Esempio 5.1.3. Data una matrice

$$A = \begin{bmatrix} T & I & & \\ -I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & -I & T \end{bmatrix} \quad T = \begin{bmatrix} 5 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 5 \end{bmatrix}$$

In questo caso i metodi iterativi sono vantaggiosi poiché anche se la matrice è predominante diagonale (e quindi permette Gauss senza problemi), a causa dell'effetto del fill-in lo rende sconveniente.

Esempio 5.1.4. Data una matrice

$$A = \begin{bmatrix} n & -1 & \dots & -1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \dots & -1 & n \end{bmatrix}$$

calcoliamo P

$$N = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \quad P = \begin{bmatrix} 0 & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{n} \\ \frac{1}{n} & \dots & \frac{1}{n} & 0 \end{bmatrix}$$

La sua norma vale

$$||P||_\infty = \frac{n-1}{n}$$

e quindi converge.

Note 5.1.1. Se la norma vale 1 non posso dire nulla sulla convergenza.

Teorema 5.1.2.

$$x^{(k+1)} = Px^{(k)} + q \text{ è convergente } \Leftrightarrow \phi(P) < 1 \tag{12}$$

Esempio 5.1.5. Data una matrice

$$A = \begin{bmatrix} 1 & & & \alpha \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

troviamo le matrici per la scomposizione

$$M = I \quad N = \begin{bmatrix} 0 & & & -\alpha \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

Per quali α il metodo è convergente?
Troviamo la fattorizzazione LU della matrice per il calcolo del determinante e otteniamo così il polinomio caratteristico:

$$x^n + \alpha$$

Dobbiamo poi trovare il modulo degli autovalori per poterli confrontare:

$$\begin{aligned} x^n &= -\alpha \\ |\lambda|^n = |-\alpha| \Rightarrow |\lambda| &= \sqrt[n]{|\alpha|} \\ \sqrt[n]{|\alpha|} < 1 \Leftrightarrow |\alpha| &< 1 \end{aligned}$$

5.2 Metodo di Jacobi

In questa tecnica prendiamo M come la matrice diagonale principale:

$$M = diag(A) \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \quad M = \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix}$$

Chiaramente questo metodo è applicabile sono quando la diagonale non contiene zeri, in quanto altrimenti non sarebbe invertibile.
Possiamo ottenere la matrice N facendo $M - A$:

$$N = \begin{bmatrix} 0 & -a_{12} & \dots & \dots & -a_{1n} \\ -a_{21} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -a_{n-1n} \\ -a_{n1} & & & -a_{nn-1} & 0 \end{bmatrix}$$

e il vettore al tempo k:

$$x_e^{(k+1)} = \frac{1}{a_{ll}} \left[b_l - \sum_{j=1, j\neq l}^{n} a_{lj} x_j^{(k)} \right] \quad l = 1, \dots, n \tag{13}$$

che avrà complessità $O(nnz(A))$ essendo molto parallelizzabile.

Esempio 5.2.1. Data la matrice

$$A = \begin{bmatrix} 1 & & & -\alpha \\ & \ddots & & \vdots \\ & & \ddots & -\alpha \\ -\beta & & & 1 \end{bmatrix}$$

la sua matrice J è

$$J = \begin{bmatrix} & & & \alpha \\ & & & \vdots \\ & & & \alpha \\ \beta & \dots & \beta & 0 \end{bmatrix}$$

Utilizziamo la fattorizzazione LU per calcolare il polinomio caratteristico:

$$det(\lambda I - J) = det \begin{bmatrix} \lambda & & & -\alpha \\ & \ddots & & \vdots \\ & & \ddots & -\alpha \\ -\beta & \dots & 0 & \lambda \end{bmatrix} = \begin{bmatrix} & I & & 0 \\ -\frac{\beta}{\lambda} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda & & & -\alpha \\ & \ddots & & \vdots \\ & & \lambda & -\alpha \\ & & & u \end{bmatrix}$$

dove $u = \lambda - \frac{\alpha\beta}{\lambda}$ e il raggio spettrale vale

$$\phi(J) < 1 \Leftrightarrow \sqrt{|\alpha\beta|} < 1$$

Dimostrare che per i valori per cui Jacobi è convergente, la matrice è invertibile.
Sappiamo che A non è invertibile se e solo se:

$$\exists x \neq 0 | Ax = 0 \Leftrightarrow (M - N)x = 0 \Leftrightarrow Mx = Nx \Leftrightarrow x = M^{-1}Nx \Leftrightarrow x = Px$$

Questo ci fa capire che che possiamo scriverlo come $Px = \lambda x$e che il metodo quindi non è convergente in quanto ha autovalore λcon autovettore x. Quindi se è convergente allora è invertibile.

5.3 Metodo di Gauss-Seidel

Gauss ragiona sul fatto che ad ogni iterazione avrò calcolato tutte le componenti precedenti fino a $l-1$ al tempo $k+1$. Riscrive quindi l'iterazione di Jacobi nella seguente forma:

$$x_e^{(k+1)} = \frac{1}{a_{ll}} \left[b_e - \sum_{j=1}^{l-1} a_{lj} x_j^{(k+1)} - \sum_{j=l+1}^{n} a_{ej} x_j^{(k)} \right] \tag{14}$$

La differenza è che questo metodo, rispetto a quello di Jacobi, perde il parallelismo ma converge più velocemente.

Teorema 5.3.1. Se A è predominante diagonale, allora:

- I metodi di Jacobi e Gauss-Seidel sono applicabili
- I metodi di Jacobi e Gauss-Seidel sono convergenti

Dimotrazione 5.3.1. Dimostriamo i due punti del teorema:

- Il primo punto è semplice in quanto se la matrice è predominante diagonale per forza di cose avremo almeno un valore diverso da 0 (perché vale la disuguaglianza stretta)
- Per dimostrare che sono convergenti vogliamo mostrare che:

$$A\text{predominante diagonale } \Rightarrow \phi(J), \phi(GS) < 1 \Rightarrow \text{Convergente}$$

Stiamo usando il fatto che il raggio spettrale sia **sufficiente** a garantire la convergenza.

$$\begin{aligned} det(\lambda I - P) & \\ &= det(\lambda I - M^{-1}N) \\ &= det(\lambda M^{-1}M - M^{-1}N) \\ &= det(M^{-1}(\lambda M - N)) \\ &= det(M^{-1}) \cdot det(\lambda M - N) \end{aligned}$$

$$det(\lambda I - P) = 0 \Leftrightarrow det(\lambda M - N) = 0$$

Quindi se λ è autovalore di P allora

$$det(\lambda M - N) = 0 \tag{15}$$

Assumiamo che esista un autovalore λ di P con $|\lambda| \geq 1$:

$$\lambda M - N = \lambda \begin{bmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{bmatrix} - \begin{bmatrix} 0 & a_{12} & \dots & . \\ & \ddots & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & \lambda a_{nn} \end{bmatrix}$$

dimostriamo ora che la matrice è predominante diagonale

$$|\lambda_{ll}| > \sum_{j=1, j\neq l}^{n} |a_{lj}| \quad |\lambda||a_{ll}| \geq |a_{ll}| > \sum_{j=1, j\neq l}^{n} |a_{lj}|$$

ma se è predominante diagonale allora la matrice è per forza invertibile e il determinante non è 0. Quindi abbiamo dimostrato per assurdo che l'uguaglianza 15 non è vera.

5.4 Criteri di arresto

Un criterio tipico di arresto per capire quando fermare le nostre iterazioni è:

$$||x_{k+1} - x_n|| \leq \text{tolleranza} \tag{16}$$

che in codice diventa:

```
while(err > tol AND iterazoni <= it_max)
   calcola x_new partendo da x_old
   voluto err = ||x_new - x_old ||
   x_old = x_new
   iterazioni = iterazioni +1
end
```

Per capire quando fermarsi possiamo mettere in relazione la differenza al punto 16 con la soluzione come segue:

$$\begin{aligned} x_{k+1} - x_k = & \\ & = x_{k+1} - x + x - x_k \\ & = Px_k + 1 - Px - q + x - x_k \\ & = P(x_k - x) + x - x_k \\ & = (P - I)(x_k - x) \end{aligned}$$

dove la matrice $P - I$ è invertibile dato che, quando sottraiamo l'identità, gli autovalori sono quelli della prima matrice meno 1 e quindi per avere autovalori uguali a 0 dovremmo avere che un autovalore di P è 1, ma è impossibile perché dato che è convergente abbiamo che il loro valore assoluto è < 1. Ottengo quindi:

$$||x_k - x|| \leq ||(P - I)^{-1}|| \cdot ||x_{k+1} - x_k|| \tag{17}$$

Esempio 5.4.1. Data la seguente matrice

$$\begin{bmatrix} 1 & \dots & \dots & x \\ & \ddots & & \vdots \\ & & \ddots & x \\ x & \dots & \dots & 1 \end{bmatrix}$$

possiamo dire che è predominante diagonale se $|x| < 1$.

Per quali valori di x il metodo di Gauss-Seidel converge? Bisogna guardare il raggio spettrale. La matrice di iterazione per il metodo GS sarà fatta:

$$GS = \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ x & \dots & x & 1 \end{bmatrix}^{-1} \begin{bmatrix} & & -x \\ & & \vdots \\ & & -x \\ & & 0 \end{bmatrix} = \begin{bmatrix} 0 & \dots & 0 & y_1 \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & y_n \end{bmatrix}$$

Dobbiamo quindi risolvere:

$$\begin{bmatrix} -x \\ \vdots \\ -x \\ 0 \end{bmatrix} = My \quad \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ x & \dots & x & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} -x \\ \vdots \\ -x \\ 0 \end{bmatrix}$$

$$\begin{aligned} & y_n = -x \\ & y_2 = -x \\ & \dots \\ & y_{n+1} = -x \\ & x \cdot y_1 + x \cdot y_2 + \dots + x \cdot y_{n-1} + y_n = 0 \end{aligned}$$

quindi abbiamo che il raggio spettrale vale

$$\phi(GS) = (n-1)x^2$$

che deve essere minore di 1:

$$(n-1)x^2 < 1 \Leftrightarrow x^2 < \frac{1}{n-1} \Leftrightarrow -\frac{1}{\sqrt{n-1}} < x < \frac{1}{n-1}$$

6 Equazioni non lineari

Stiamo considerando equazioni del tipo $f(x) = 0$ dove la funzione f non è lineare (quindi non è una retta). Di fronte a questo tipo di equazioni, ci sono due difficoltà:

- Non c'è una teoria generale sul *numero* e sull'*esistenza* delle **soluzioni**
- Non esistono metodi diretti di risoluzione

Esempio 6.0.1. Determinare il numero di soluzioni reali dell'equazione

$$f(x) = x \log x - 1 = 0$$

Il primo passo è tracciare un grafico approssimativo di questa funzione:

- **Dominio**: $x > 0$
- **Limiti**:

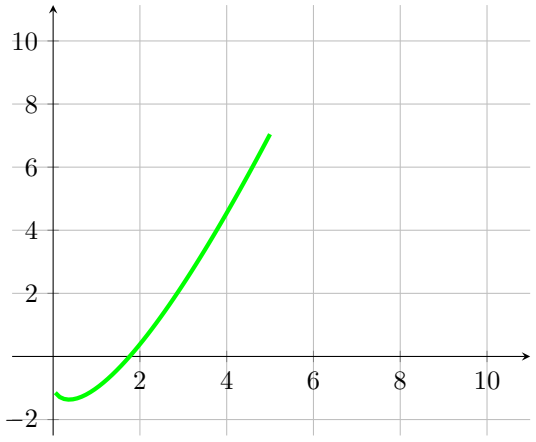
$$\lim_{x \to +\infty} x \log x - 1 = +\infty$$

$$\lim_{x \to 0^+} x \log x = \lim_{x \to 0^+} \frac{logx}{\frac{1}{x}} = \lim_{x \to 0^+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x \to 0^+} -\frac{x^2}{x} = 0 \Longrightarrow \lim_{x \to 0^+} xlogx - 1 = -1$$

- **Derivata prima**:

$$f'(x) = \log x + x \cdot \frac{1}{x} = \log x + 1$$
$$f'(x) \geq 0 \Leftrightarrow logx + 1 \geq 0 \Leftrightarrow logx \geq -1 \Leftrightarrow x \geq \frac{1}{e}$$

- **Derivata seconda**: $f''(x) = \frac{1}{x} \geq 0 \quad \forall x > 0$



Quindi possiamo dire che

$$\exists! \alpha \in \mathbb{R} | f(\alpha) = 0$$

Ci serve dare un **intervallo di localizzazione** della soluzione, ad esempio:

$$f(1) = -1$$
$$f(2) = 2 \log 2 - 1 = log4 - 1$$
$$\Rightarrow \alpha \in [1, 2]$$

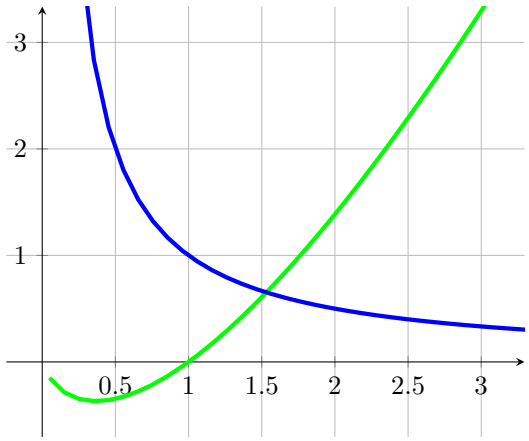
6.1 Tecnica della separazione

A partire da una funzione complessa, mi riconduco a funzioni più semplici e vedo dove si intercettano.

Esempio 6.1.1. Supponiamo di avere

$$x \log x - 1 = 0 \Leftrightarrow x \log x = 1 \Leftrightarrow \log x = \frac{1}{x}$$

Che sul grafico sono



Esempio 6.1.2. Data la seguente funzione

$$f(x) = e^x - 2x = 0 \Leftrightarrow e^x = 2x$$

È difficile usare il metodo della separazione perché l'intersezione non è facile da trovare.
Usiamo quindi la soluzione grafica:

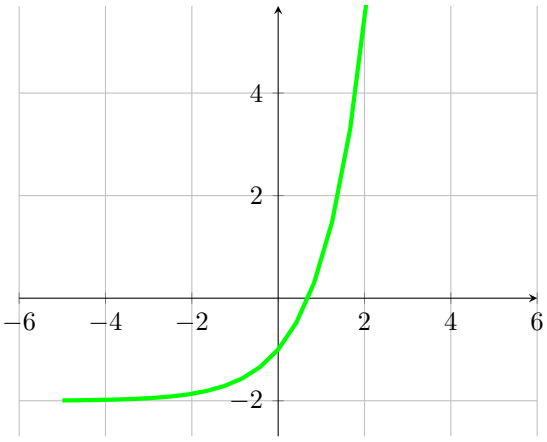
- **Dominio**: $\forall x \in \mathbb{R}$ che possiamo scrivere anche come $C^\infty(\mathbb{R})$
- **Limiti**:

$$\lim_{x \to +\infty} e^x - 2x = \lim_{x \to +\infty} e^x(1 - \frac{2x}{e^x}) = 0$$
$$\lim_{x \to -\infty} e^x - 2x = +\infty$$

- **Derivata prima**:

$$f'(x) = e^x - 2$$
$$f'(x) \geq 0 \Leftrightarrow e^x \geq 2 \Leftrightarrow x \geq \log 2$$

- **Derivata seconda**: $f''(x) = e^x$



Calcoliamo adesso il valore in $\log 2$:

$$f(\log 2) = e^{\log 2} - 2\log 2 = 2 - 2\log 2 = 2(1 - \log 2)$$

Esempio 6.1.3. Data la funzione:

$$f(x) = x^3 - 6x + 1 = 0 \tag{18}$$

In questo esempio abbiamo un polinomio e abbiamo quindi un'**equazione algebrica** di terzo grado. Quindi sappiamo il numero di soluzioni *complesse*, nel nostro caso 3. A noi però interessa il numero di soluzioni reali. Sapendo che quelle complesse devono andare sempre in coppia, potremo avere o due soluzioni complesse e una reale oppure tre soluzioni reali. Studiamo la funzione:

- **Dominio**: $\forall x \in \mathbb{R}$
- **Limiti**:

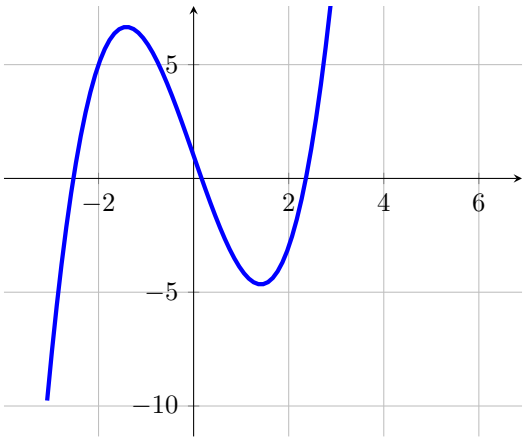
$$\lim_{x \to +\infty} x^3 - 6x + 1 = +\infty$$
$$\lim_{x \to -\infty} x^3 - 6x + 1 = -\infty$$

  Sappiamo quindi che esiste sicuramente almeno un punto in cui la funzione vale 0 per il teorema dell'esistenza degli zeri.

- **Derivata prima**:

$$f'(x) = 3x^2 - 6$$
$$f'(x) = 0 \Leftrightarrow x^2 = 2 \Leftrightarrow x = \pm\sqrt{2} \rightarrow x < -\sqrt{2} \wedge x > \sqrt{2}$$

- **Derivata seconda**: $f''(x) = 6x \rightarrow f''(x) > 0 \Leftrightarrow x > 0$



Quindi ci sono tre soluzioni reali, che possiamo localizzare come:

$$\beta \in [0, \sqrt{2}]$$
$$\gamma \in [\sqrt{2}, 3]$$
$$\alpha \in [-3, -2]$$