

Elaborato di  
**Calcolo Numerico**  
Anno Accademico 2016/2017

Gabriele Puliti - 5300140 - *gabriele.puliti@stud.unifi.it*  
Luca Passaretta - -

August 30, 2017

# Capitoli

|          |                          |          |
|----------|--------------------------|----------|
| <b>1</b> | <b>Capitolo 1</b>        | <b>3</b> |
| 1.1      | Esercizio 1 . . . . .    | 3        |
| 1.2      | Esercizio 2 . . . . .    | 3        |
| 1.3      | Esercizio 3 . . . . .    | 4        |
| 1.4      | Esercizio 4 . . . . .    | 4        |
| 1.5      | Esercizio 5 . . . . .    | 5        |
| 1.6      | Esercizio 6 . . . . .    | 6        |
| 1.7      | Esercizio 7 . . . . .    | 7        |
| 1.8      | Esercizio 8 . . . . .    | 7        |
| 1.9      | Esercizio 9 . . . . .    | 8        |
| 1.10     | Esercizio 10 . . . . .   | 8        |
| 1.11     | Esercizio 11 . . . . .   | 8        |
| 1.12     | Esercizio 12 . . . . .   | 9        |
| 1.13     | Esercizio 13 . . . . .   | 9        |
| <b>2</b> | <b>Grafici</b>           | <b>i</b> |
| 2.1      | Esercizio 1.4 . . . . .  | i        |
| 2.2      | Esercizio 1.13 . . . . . | i        |

# 1 Capitolo 1

## 1.1 Esercizio 1

Sapendo che il metodo iterativo è convergente a  $x^*$  allora per definizione si ha:

$$\lim_{k \rightarrow +\infty} x_k = x^*$$

inoltre per definizione di  $\Phi$  si calcola il limite:

$$\lim_{k \rightarrow +\infty} \Phi(x_k) = \lim_{k \rightarrow +\infty} x_{k+1} = x^*$$

infine ipotizzando che la funzione  $\Phi$  sia uniformemente continua è possibile calcolare il limite:

$$\lim_{k \rightarrow +\infty} \Phi(x_k) = \Phi\left(\lim_{k \rightarrow +\infty} x_k\right) = \Phi(x^*)$$

dai due limiti si ha la tesi:

$$\Phi(x^*) = x^*$$

## 1.2 Esercizio 2

Dal momento che le variabili intere di 2 byte in Fortran vengono gestite in Modulo e Segno, la variabile **numero**, inizializzata con

```
integer*2 numero
```

varia tra  $-32768 \leq \text{numero} \leq 32767$  ( $-2^{15} \leq \text{numero} \leq 2^{15} - 1$ ). Durante la terza iterazione del primo ciclo for si arriva al valore massimo rappresentabile tramite gli interi a 2 byte, alla quarta iterazione si avrà quindi la somma del **numero** in modulo e segno:

$$\begin{aligned} (32767)_{10} + (1)_{10} &= (0111111111111111)_{2,MS} + (0000000000000001)_{2,MS} = \\ &= (1000000000000000)_{2,MS} = (-32768)_{10} \end{aligned}$$

Nel secondo ciclo for durante la quinta iterazione, al **numero** viene sottratto 1:

$$\begin{aligned} (-32768)_{10} - (1)_{10} &= (1000000000000000)_{2,MS} - (0000000000000001)_{2,MS} = \\ &= (0111111111111111)_{2,MS} = (32767)_{10} \end{aligned}$$

Da cui si spiega l'output del codice.

### 1.3 Esercizio 3

Per definizione si ha che la precisione di macchina  $u$  per arrotondamento e' data da:

$$u = \frac{1}{2}b^{1-m}$$

Se  $b = 8, m = 5$  si ha:

$$u = \frac{1}{2} \cdot 8^{1-5} = \frac{1}{2} \cdot 8^{-4} = 1,2 \cdot 10^{-4}$$

### 1.4 Esercizio 4

Il codice seguente:

```
1 format long e;  
2  
3 h=zeros(12,1);  
4 f=zeros(12,1);  
5  
6 for i=1:12  
7     h(i)= power(10,-i);  
8 end  
9  
10 for j=1:12  
11     f(j)=lim(0,v(j));  
12 end  
13  
14 f  
15  
16 function p=lim(x,y)  
17     p=(exp(x+y) - exp(x))/y;  
18 end
```

restituisce questo risultato (assumendo che  $f(x) = e^x$  e  $x_0 = 0$ ):

| h          | $\Psi_h(0)$           |
|------------|-----------------------|
| $10^{-1}$  | 1.051709180756477e+00 |
| $10^{-2}$  | 1.005016708416795e+00 |
| $10^{-3}$  | 1.000500166708385e+00 |
| $10^{-4}$  | 1.000050001667141e+00 |
| $10^{-5}$  | 1.000005000006965e+00 |
| $10^{-6}$  | 1.000000499962184e+00 |
| $10^{-7}$  | 1.000000049433680e+00 |
| $10^{-8}$  | 9.999999939225290e-01 |
| $10^{-9}$  | 1.000000082740371e+00 |
| $10^{-10}$ | 1.000000082740371e+00 |
| $10^{-11}$ | 1.000000082740371e+00 |
| $10^{-12}$ | 1.000088900582341e+00 |

si può notare che al diminuire del valore  $h$ , la funzione  $\Psi_h(0)$  approssima sempre meglio il valore  $f'(0)$ , come si può vedere dal plot 1.

### 1.5 Esercizio 5

Per dimostrare le due uguaglianze è necessario sviluppare in serie di Taylor  $f(x)$  fino al secondo ordine:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2 f''(x_0)}{2} + O((x - x_0)^2)$$

Da cui possiamo sostituire con i valori di  $x = x + h$  e  $x = x - h$ :

$$f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2)$$

$$f(x_0 - h) = f(x_0) - hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2)$$

Andando a sostituire questi valori si ottiene, nel primo caso:

$$\begin{aligned} & \frac{f(x_0 + h) - f(x_0 - h)}{2h} = \\ &= \frac{(f(x_0) + hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2)) - (f(x_0) - hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2))}{2h} = \\ &= \frac{2hf'(x_0) + O(h^2)}{2h} = f'(x_0) + O(h^2) \end{aligned}$$

nel secondo caso:

$$\begin{aligned}
 & \frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h))}{h^2} = \\
 = & \frac{f(x_0) + hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2) - 2f(x_0) + f(x_0) - hf'(x_0) + \frac{h^2 f''(x_0)}{2} + O(h^2)}{h^2} = \\
 = & \frac{h^2 f''(x_0) + O(h^2)}{h^2} = f''(x_0) + O(h^2)
 \end{aligned}$$

Abbiamo quindi dimostrato che:

$$\frac{f(x_0 + h) - f(x_0 - h))}{2h} = f'(x_0) + O(h^2)$$

$$\frac{f(x_0 + h) - 2f(x_0) + f(x_0 - h))}{h^2} = f''(x_0) + O(h^2)$$

## 1.6 Esercizio 6

Il codice MatLab, indicando con  $x=x_n$  e  $r=\epsilon$ :

```

1 format longEng
2 format compact
3
4 conv=sqrt(2);
5 x=[2,1.5];
6 r=[x(1)-conv,x(2)-conv];
7
8 for i= 2:7
9     x(i+1) = (x(i)*x(i-1)+2)/(x(i)+x(i-1))
10 end
11
12 for i=3:8
13     r(i)=x(i)-conv;
14 end
15
16 x
17 r

```

restituisce i valori:

| n | $x_n$                  | $\epsilon$             |
|---|------------------------|------------------------|
| 0 | 2.000000000000000e+000 | 585.786437626905e-003  |
| 1 | 1.500000000000000e+000 | 85.7864376269049e-003  |
| 2 | 1.42857142857143e+000  | 14.3578661983335e-003  |
| 3 | 1.41463414634146e+000  | 420.583968367971e-006  |
| 4 | 1.41421568627451e+000  | 2.12390141496321e-006  |
| 5 | 1.41421356268887e+000  | 315.774073555986e-012  |
| 6 | 1.41421356237310e+000  | 0.000000000000000e+000 |
| 7 | 1.41421356237310e+000  | 0.000000000000000e+000 |

I valori indicano che per valori di  $n$  superiori a 5 l'errore, indicato con  $\epsilon$ , è dell'ordine di  $10^{-12}$ .

## 1.7 Esercizio 7

Sapendo che la rappresentazione del numero è stata fatta usando l'arrotondamento allora la precisione di macchina si calcola:

$$u = \frac{b^{1-m}}{2}$$

che sappiamo essere pari a:

$$u \approx 4.66 \cdot 10^{-10}$$

dato che stiamo cercando il numero di cifre binarie allora si deve avere  $b = 2$ , è possibile quindi ricavarsi  $m$ :

$$m = 1 - \log_2(4.66 \cdot 10^{-10}) \approx 31.99$$

possiamo quindi affermare che servono 32 cifre dedicate alla mantissa per rappresentare il numero con precisione macchina  $4.66 \cdot 10^{-10}$ .

## 1.8 Esercizio 8

Sapendo che la mantissa in decimale è calcolabile tramite la funzione:

- $m = 1 - \log_{10}u$  (troncamento)
- $m = 1 - \log_{10}(2 \cdot u)$  (arrotondamento)

e che la precisione di macchina assuma un valore accettabilmente piccolo in modo tale che il  $\log_{10}u \approx 1$ , cioè  $0 \leq u < 1$ , allora è possibile scrivere:

- $m = 1 - \log_{10}u \approx -\log_{10}u$  (troncamento)
- $m = 1 - \log_{10}2u = 1 - \log_{10}2 - \log_{10}u \approx -\log_{10}u$  (arrotondamento)

## 1.9 Esercizio 9

```
1 x=0;  
2 delta = 1/10;  
3 while x<1,x=x+delta , end
```

Il valore di delta è uguale a  $\frac{1}{10}$ , la rappresentazione binaria di questo numero però non è esatta. Si tratta di una rappresentazione periodica e quindi, in decimale, sarà circa 0.0999, prendendo quindi  $\delta \approx 0.9$  vedremo che per  $i = 10$   $x \approx 0.999$ , mentre per  $i = 11$   $x \approx 1.0989$ . Per questo motivo la condizione  $x=1$  non si avvererà mai e il programma non terminerà.

## 1.10 Esercizio 10

All'interno della radice può presentarsi un problema di overflow dato che la somma dei due quadrati potrebbe essere molto grande, tanto grande da poter superare il limite massimo rappresentabile dalla macchina:

$$realmax = (1 - b^{-m}) \cdot b^{b^s - \nu}$$

Per risolvere questo problema è necessario prendere il massimo valore tra le due variabili:

$$m = \max\{|x|, |y|\}$$

E moltiplicare e dividere per questo valore:

$$\sqrt{x^2 + y^2} = m \cdot \frac{\sqrt{x^2 + y^2}}{m} = m \cdot \sqrt{\frac{x^2 + y^2}{m^2}} = m \cdot \sqrt{\left(\frac{x}{m}\right)^2 + \left(\frac{y}{m}\right)^2}$$

In questo modo si eviterà il problema di overflow, il problema è ben condizionato dato che potenza e radice sono ben condizionate e grazie alla modifica proposta.

## 1.11 Esercizio 11

Le due espressioni in aritmetica finita vengono scritte tenendo conto dell'errore di approssimazione sul valore reale:

- $fl(fl(fl(x) + fl(y)) + fl(z)) = ((x(1 + \varepsilon_x) + y(1 + \varepsilon_y))(1 + \varepsilon_a) + z(1 + \varepsilon_z))(1 + \varepsilon_b)$
- $fl(fl(x) + fl(fl(y) + fl(z))) = (x(1 + \varepsilon_x) + (y(1 + \varepsilon_y) + z(1 + \varepsilon_z))(1 + \varepsilon_a))(1 + \varepsilon_b)$



Indicando con  $\varepsilon_x, \varepsilon_y, \varepsilon_z$  i relativi errori di  $x, y, z$  e con  $\varepsilon_a, \varepsilon_b$  gli errori delle somme.

Per calcolare l'errore relativo delle due espressioni consideriamo  $\varepsilon_m = \max\{\varepsilon_x, \varepsilon_y, \varepsilon_z, \varepsilon_a, \varepsilon_b\}$ , dalla definizione di errore relativo si ha quindi:

•

$$\begin{aligned}\varepsilon_1 &= \frac{((x(1+\varepsilon_x) + y(1+\varepsilon_y))(1+\varepsilon_a) + z(1+\varepsilon_z))(1+\varepsilon_b) - (x+y+z)}{x+y+z} \approx \\ &\approx \frac{x(1+\varepsilon_x + \varepsilon_a + \varepsilon_b) + y(1+\varepsilon_y + \varepsilon_a + \varepsilon_b) + z(1+\varepsilon_z + \varepsilon_b) - x - y - z}{x+y+z} \leq \\ &\leq \frac{3 \cdot x \cdot \varepsilon_m + 3 \cdot y \cdot \varepsilon_m + 2 \cdot z \cdot \varepsilon_m}{x+y+z} \leq \frac{3 \cdot \varepsilon_m \cdot (x+y+z)}{x+y+z} = 3 \cdot \varepsilon_m\end{aligned}$$

• seguendo gli stessi procedimenti del punto sopra possiamo scrivere:

$$\begin{aligned}\varepsilon_2 &= \frac{(x(1+\varepsilon_x) + (y(1+\varepsilon_y) + z(1+\varepsilon_z))(1+\varepsilon_a))(1+\varepsilon_b) - (x+y+z)}{x+y+z} = \\ &= \dots \leq \frac{2 \cdot x \cdot \varepsilon_m + 3 \cdot y \cdot \varepsilon_m + 3 \cdot z \cdot \varepsilon_m}{x+y+z} \leq \frac{3 \cdot \varepsilon_m \cdot (x+y+z)}{x+y+z} = 3 \cdot \varepsilon_m\end{aligned}$$

### 1.12 Esercizio 12

Sapendo che il numero di condizionamento del problema è dato da:

$$k = \left| f'_x \cdot \frac{x}{f(x)} \right|$$

Dato che la nostra funzione è  $f(x) = \sqrt{x}$  allora la derivata è data da  $f'(x) = \frac{1}{2\sqrt{x}}$ , sostituendo otteniamo:

$$k = \left| \frac{1}{2 \cdot \sqrt{x}} \cdot \frac{x}{\sqrt{x}} \right| = \left| \frac{1}{2} \right| = \frac{1}{2}$$

Come volevamo.

### 1.13 Esercizio 13

Il problema è che stiamo rappresentando dei numeri reali in un calcolatore quindi la loro rappresentazione comporta delle approssimazioni. Nella riga 11 abbiamo calcolato e restituito in output il valore interno al logaritmo  $|3(1 - \frac{3}{4}) + 1|$  che teoricamente è zero, ma si ottiene  $2.220446049250313e-16$ . Si può vedere che il codice MatLab:

```

1 format long;
2
3 x=linspace(2/3,2,1001);
4 y= [];
5
6 for i = 1:1001
7     y(i) = log(abs(3*(1-x(i))+1))/80 + x(i)^2 +1;
8 end
9
10 plot(x,y);
11 disp (abs(3*(1-4/3)+1))

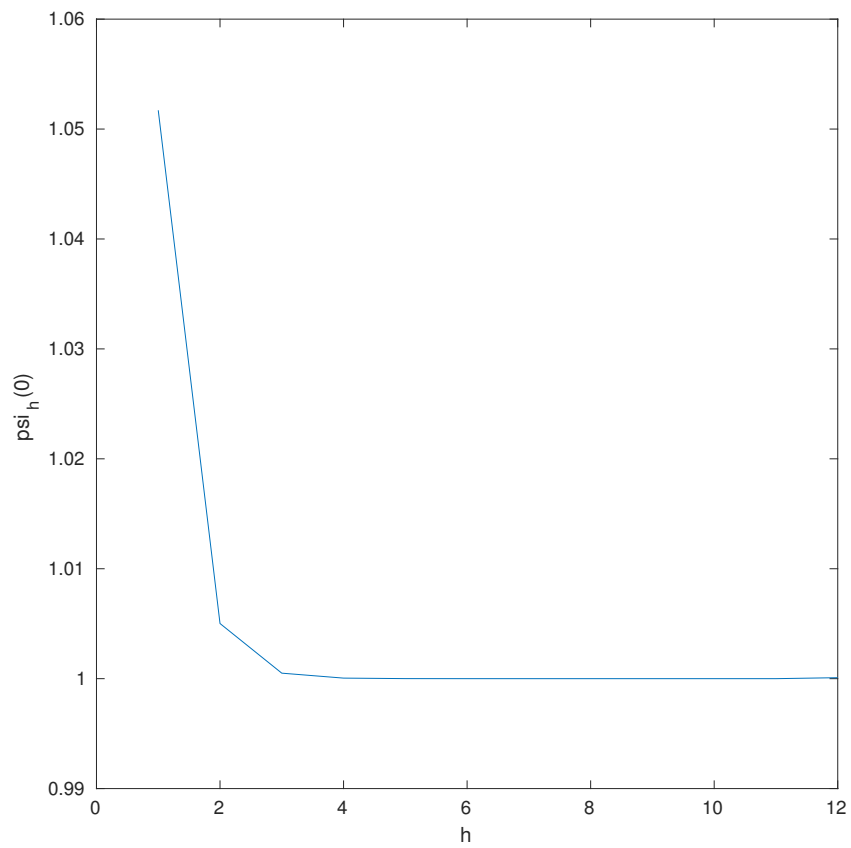
```

calcola i valori della funzione ottenendo il grafico 2 e si può notare che l'asintoto verticale in  $x = \frac{4}{3}$  viene rappresentato come minimo della funzione  $f(x)$ .

## 2 Grafici

### 2.1 Esercizio 1.4

Figure 1: Esercizio 1.4



### 2.2 Esercizio 1.13

Figure 2: Esercizio 1.13

