

Università degli studi Milano Bicocca - Dipartimento di Fisica

# Esperimentazioni di Fisica Computazionale

S. Franceschina

June 14, 2025

## Abstract

La presente relazione contiene gli esercizi svolti durante il corso di Fisica Computazionale.

Gli esercizi sono stati svolti in Julia e sono raccolti in una cartella git, di cui si riporta il link:

[https://github.com/bunchi3/lab\\_computazionale1.git](https://github.com/bunchi3/lab_computazionale1.git)

## Contents

<b>1</b>	<b>Analisi dell'errore</b>	<b>4</b>
1.1	Esercizio 1.0.1 . . . . .	4
1.1.1	Soluzione . . . . .	4
1.2	Esercizio 1.2.1 . . . . .	5
1.2.1	Soluzione . . . . .	5
1.3	Esercizio 1.4.1 . . . . .	6
1.3.1	Soluzione . . . . .	7
1.4	Esercizio 1.4.2 . . . . .	8
1.4.1	Soluzione . . . . .	9
<b>2</b>	<b>Sistemi lineari</b>	<b>11</b>
2.1	Esercizi 2.1.1, 2.1.2, 2.1.3, 2.1.4 . . . . .	11
2.1.1	Soluzione . . . . .	12
2.2	Esercizi 2.3.1, 2.3.2, 2.3.4 . . . . .	12
2.2.1	Soluzione . . . . .	13
2.3	Esercizio 2.3.3 . . . . .	13
2.3.1	Soluzione . . . . .	14
2.4	Esercizi 2.4.1, 2.4.2 . . . . .	14
2.4.1	Soluzione . . . . .	15
2.5	Esercizio 2.5.1 . . . . .	15
2.5.1	Soluzione . . . . .	15
2.5.2	Conclusioni . . . . .	16
2.6	Esercizio 2.6.1 . . . . .	17
2.6.1	Soluzione . . . . .	17
2.7	Esercizio 2.6.2 . . . . .	17
2.7.1	Soluzione . . . . .	17

2.8	Esercizio 2.6.3 . . . . .	18
2.8.1	Soluzione . . . . .	19
2.8.2	Conclusioni . . . . .	19
<b>3</b>	<b>Radici di equazioni non lineari</b>	<b>20</b>
3.1	Esercizi 3.2.1, 3.2.2 . . . . .	20
3.1.1	Soluzione . . . . .	20
3.2	Esercizi 3.3.1, 3.3.2 . . . . .	22
3.2.1	Soluzione . . . . .	22
3.3	Esercizio 3.3.3 . . . . .	26
3.3.1	Soluzione . . . . .	26
3.4	Esercizio 3.3.4 . . . . .	27
3.5	Soluzione . . . . .	28
3.6	Esercizi 3.4.1, 3.4.2 . . . . .	28
3.6.1	Soluzione . . . . .	29
3.7	Esercizio 3.4.3 . . . . .	29
3.7.1	Soluzione . . . . .	30
<b>4</b>	<b>Interpolazioni</b>	<b>30</b>
4.1	Esercizi 4.2.1, 4.2.2 . . . . .	30
4.1.1	Soluzione . . . . .	31
4.2	Esercizi 4.4.1, 4.4.2 . . . . .	31
4.2.1	Soluzione . . . . .	32
4.3	Esercizio 4.4.3 . . . . .	33
4.3.1	Soluzione . . . . .	33
4.4	Esercizio 4.4.4 . . . . .	34
4.4.1	Soluzione . . . . .	34
4.5	Esercizio 4.6.1 . . . . .	34
4.5.1	Soluzione . . . . .	34
<b>5</b>	<b>Integrazione numerica</b>	<b>36</b>
5.1	Esercizi 5.1.1, 5.1.2 . . . . .	36
5.1.1	Soluzione . . . . .	37
5.2	Esercizi 5.1.3, 5.1.4 . . . . .	39
5.2.1	Soluzione . . . . .	39
5.3	Esercizio 5.3.1 . . . . .	40
5.3.1	Soluzione . . . . .	42
5.4	Esercizio 5.4.1 . . . . .	43
5.4.1	Soluzione . . . . .	44
5.5	Esercizio 5.4.2 . . . . .	45
5.5.1	Soluzione . . . . .	46
5.6	Esercizio 5.4.3 . . . . .	47
5.6.1	Soluzione . . . . .	47
<b>6</b>	<b>Equazioni differenziali ordinarie</b>	<b>47</b>
6.1	Esercizi 6.2.1, 6.2.2 . . . . .	47
6.1.1	Soluzione . . . . .	48
6.2	Esercizio 6.2.3 . . . . .	50
6.2.1	Soluzione . . . . .	50
6.3	Esercizio 6.3.1 . . . . .	51
6.3.1	Soluzione . . . . .	52

6.4	Esercizio 6.3.2 . . . . .	52
6.4.1	Soluzione . . . . .	53
6.5	Esercizio 6.3.3 . . . . .	53
6.5.1	Soluzione . . . . .	54
6.6	Esercizio 6.3.5 . . . . .	55
6.6.1	Soluzione . . . . .	55
6.7	Esercizio 6.3.6 . . . . .	56
6.7.1	Soluzione . . . . .	57
6.8	Esercizio 6.3.7 . . . . .	58
6.8.1	Soluzione . . . . .	58
<b>7</b>	<b>Appendice</b>	<b>59</b>
7.1	Dati . . . . .	59

# 1 Analisi dell'errore

## 1.1 Esercizio 1.0.1

Considera la funzione  $f(x) = e^x$  nell'intervallo  $x \in [0, 1]$ . Scrivi un programma che calcoli la serie approssimante:

$$g_N(x) = \sum_{n=0}^N \frac{x^n}{n!}. \quad (1)$$

1. Verifica che l'errore assoluto  $\Delta = |f(x) - g_N(x)|$  scala approssimativamente come  $x^{N+1}/(N+1)!$  per  $N = 1, 2, 3, 4$ .
2. L'errore  $\Delta$ , nell'intervallo dato di  $x$ , differisce da  $x^{N+1}/(N+1)!$ . Perché accade questo e per quali valori di  $x$ ?

### 1.1.1 Soluzione

Al fine dell'esercizio vengono rappresentati nel grafico 1 le funzioni  $\Delta$  e  $\frac{x^{N+1}}{(N+1)!}$ , con  $N = 1, 2, 3, 4$ , al variare di  $x$  nell'intervallo  $[0, 1]$ .

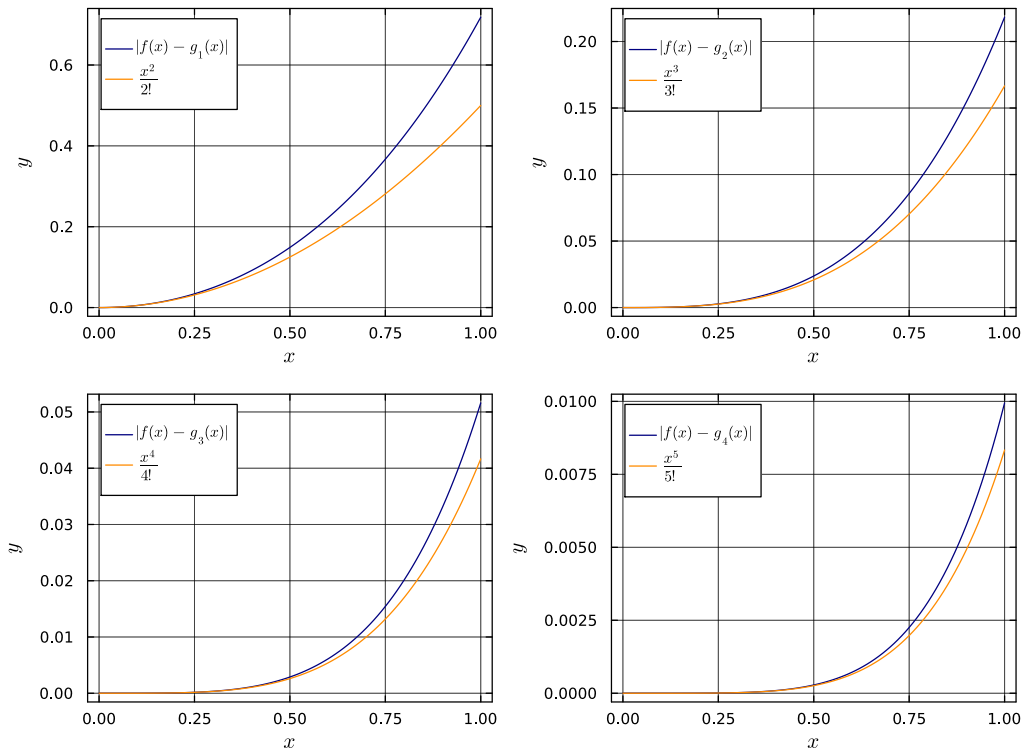


Figure 1: Confronto tra  $\Delta$  e  $\frac{x^{N+1}}{(N+1)!}$  per  $N = 1, 2, 3, 4$ .

La prima considerazione che possiamo fare è che all'aumentare di  $N$  la funzione  $\Delta$  assume valori sempre più vicini allo zero. Questo significa che la distanza tra il valore della funzione  $f(x)$  presa in esame e la sua espansione di Taylor troncata all'ordine  $N$  diminuisce. In effetti ci aspettiamo che la funzione  $\Delta$  sia esattamente zero nel caso in cui  $N \rightarrow \infty$ . Inoltre possiamo notare che la funzione  $\Delta$ , in ognuno dei grafici, è tanto più prossima allo zero quanto più ci si avvicina all'origine, poichè l'espansione in serie richiede  $x \rightarrow 0$ .

La seconda considerazione è che le funzioni  $\Delta$  e  $\frac{x^{N+1}}{(N+1)!}$  si avvicinano tra loro all'aumentare di

$N$ . Questo risponde alla richiesta dell'esercizio, cioè che l'errore scali come un polinomio di ordine  $N + 1$ .

## 1.2 Esercizio 1.2.1

Calcolare la seguente somma

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} = \lim_{N \rightarrow \infty} S(N) \quad \text{con} \quad S(N) = \sum_{n=1}^N \frac{1}{n^2} \quad (2)$$

1. Calcolarla in single precision utilizzando l'ordinamento normale,  $n = 1, 2, 3, \dots, N$ .
2. Calcolarla in single precision utilizzando l'ordinamento inverso,  $n = N, \dots, 2, 1$ .
3. Studiare la convergenza di entrambe le implementazioni in funzione di  $N$  tracciando il grafico di  $|S(N) - \pi^2/6|$ .
4. Ripetere i punti da 1 a 3 utilizzando double precision.

### 1.2.1 Soluzione

Procediamo prendendo in esame i due casi, single precision e double precision.

**Single precision:** Si riportano in figura 2 gli errori di troncamento della serie  $\sum_{n=1}^{\infty} \frac{1}{n^2}$ , laddove  $N$  è il troncamento. La somme sono state eseguite con variabili di tipo Float32 (single precision).

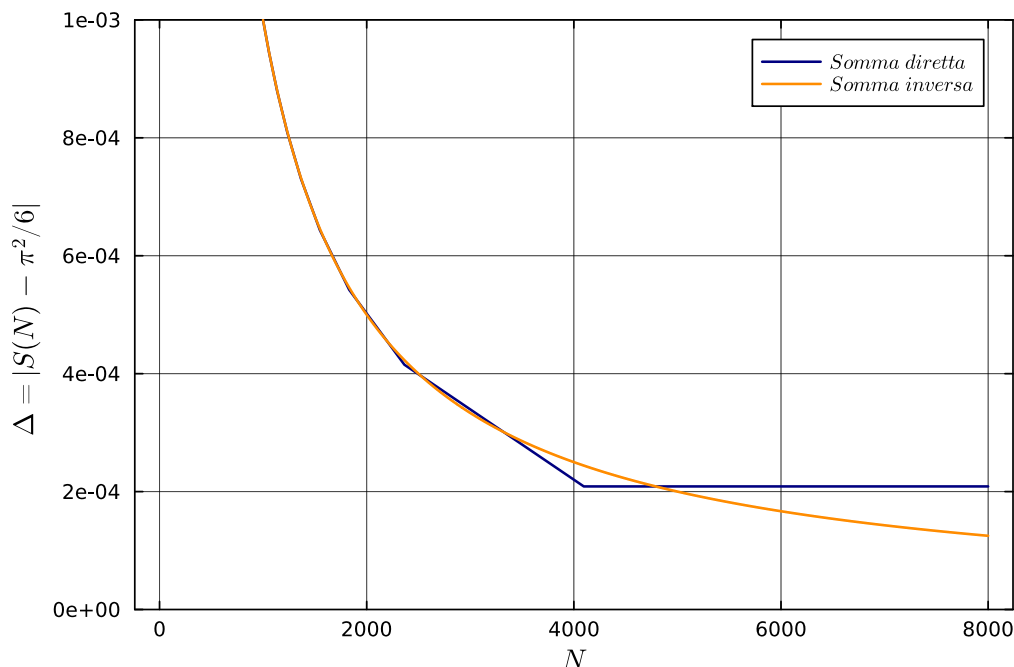


Figure 2:  $S(N) = \sum_{n=1}^N \frac{1}{n^2}$  in single precision

In figura 2 notiamo due comportamenti degni di nota. Il primo riguarda la curva blu, che arresta la sua discesa poco dopo  $N = 4000$ . Il secondo riguarda il discostarsi delle due curve, già a partire da  $N \simeq 2000$ , a causa dell'andamento a linea spezzata della curva blu. La spiegazione del primo fenomeno è la seguente: nel caso della somma con ordinamento diretto,

all'aumento dell'indice di somma, gli addendi sono sempre più piccoli. In particolare la somma comincia da 1, e sappiamo che dovrà raggiungere  $\frac{\pi^2}{6} \simeq 1.64$  e perciò per tutto il processo la somma parziale resterà nell'intervallo  $[1, 2)$ . In tale intervallo la distanza tra due floating point numbers è  $\epsilon_{mach} = 2^{-23}$ . Osserviamo il grafico 2: la curva blu comincia ad essere costante a partire da  $N = 4096 = 2^{12}$ . Tale numero corrisponde all'addendo  $\frac{1}{N^2} = 2^{-24}$ , che è appena più piccolo di  $\epsilon_{mach}$ , ovvero della distanza tra due floating point numbers in  $[1, 2)$ , e quindi è come sommare 0.

La spiegazione del secondo fenomeno è simile: in questo caso alla somma viene aggiunto un addendo che è più piccolo del precedente, ma non così tanto da essere più piccolo di  $\epsilon_{mach}$ . In questo modo la somma viene migliorata, ma il nuovo addendo è arrotondato rispetto al suo valore vero. Il successivo addendo, pur essendo differente in teoria, a causa dell'arrotondamento risulta essere uguale al precedente. Ciò fa in modo che venga sommato sempre lo stesso numero, e così l'andamento della curva blu è lineare.

I due fenomeni non si verificano per la somma con ordinamento inverso perchè il primo numero ad essere sommato è molto piccolo. Questo fa in modo che i successivi floating point number siano molto vicini tra loro, è così sommare l'addendo successivo fa cadere la somma parziale vicina al floating point number che la approssima.

Si può mostrare matematicamente che gli errori  $\Delta = |S(N) - \pi^2/6|$  scalano come  $O(\frac{1}{N})$ . Per verificare questa affermazione possiamo interpolare la curva degli errori con un metodo che vedremo nella sezione 2. Sulla base delle considerazioni di questa sezione, possiamo aspettarci che sia più conveniente interpolare la curva con ordinamento inverso, perchè i valori che la compongono sono meno affetti da errori di tipo numerico. Utilizzando il modello linearizzato  $\log(\Delta) = -c \log(N)$  con  $c = 1$  si ottiene per l'ordinamento inverso  $c_{fit} = 1.00003$  e per quello diretto  $c_{fit} = 0.986$ . Come atteso il caso di ordinamento diretto restituisce un risultato peggiore di quello inverso.

**Double precision:** Si riporta in figura 3 un ingrandimento del grafico degli errori di troncamento della serie  $\sum_{n=1}^{\infty} \frac{1}{n^2}$ , laddove  $N$  è il troncamento. La somme sono state eseguite con variabili di tipo Float64 (double precision).

Nel caso di figura 3 possiamo notare che le due curve sono sovrapposte, nonostante l'ingrandimento a valori di  $N \in [5000, 10000]$ . Per osservare un fenomeno simile a quello di figura 2 dovremmo raggiungere valori di  $N = 2^{26}$ . Infatti l'approssimazione non migliora quando vengono sommati numeri dell'ordine di  $\epsilon_{mach}$  per double precision, cioè  $\frac{1}{N^2} \simeq 2^{-52}$ .

### 1.3 Esercizio 1.4.1

(a) In statistica, definiamo la varianza di un campione di valori  $x_1, \dots, x_n$  come

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (3)$$

Scrivi una funzione che prenda in input un vettore  $x$  di lunghezza arbitraria e restituisca  $\sigma^2$  calcolata con la formula sopra. Dovresti testare la funzione con  $x = [1, 1, \dots, 1]$  e con alcuni vettori casuali.

(b) La formula 3 ha lo svantaggio di scorrere due volte il vettore dei dati. Considera la formula 4 a un ciclo:

$$\sigma^2 = \frac{1}{n-1} \left( u - \frac{1}{n} v^2 \right), \quad u = \sum_{i=1}^n x_i^2, \quad v = \sum_{i=1}^n x_i. \quad (4)$$

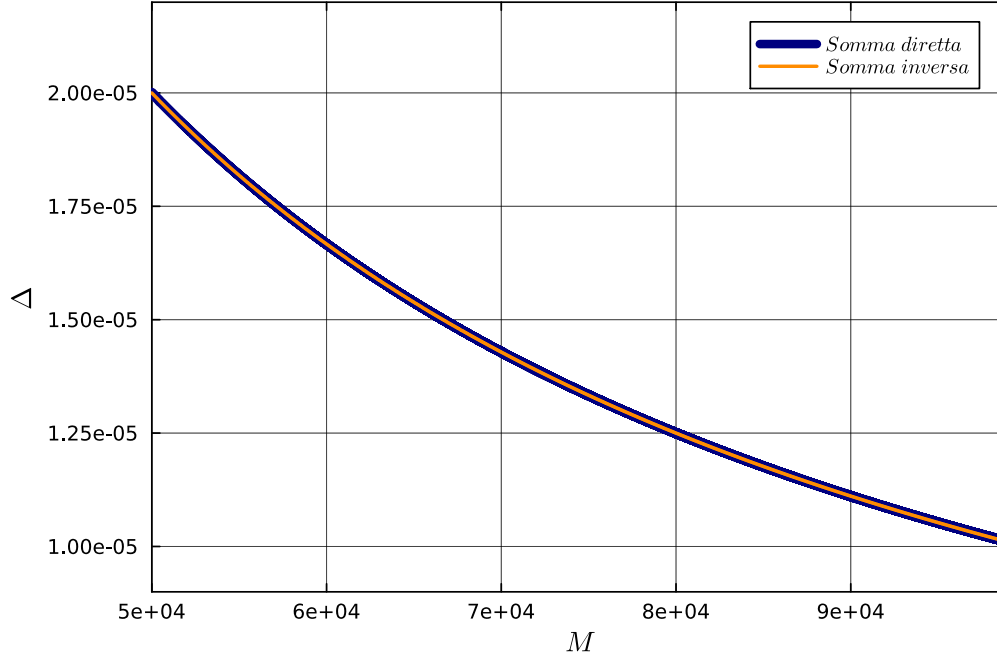


Figure 3: Ingrandimento di  $S(N) = \sum_{n=1}^N \frac{1}{n^2}$  in double precision

Prova entrambe le formule per i seguenti dataset, ciascuno dei quali ha varianza esattamente uguale a 1. Esegui i calcoli sia in single che in double precision.

$$\begin{aligned} x_1 &= [1 \cdot 10^3, 1 + 10^3, 2 + 10^3] & x_2 &= [1 \cdot 10^6, 1 + 10^6, 2 + 10^6] \\ x_3 &= [1 \cdot 10^7, 1 + 10^7, 2 + 10^7] & x_4 &= [1 \cdot 10^8, 1 + 10^8, 2 + 10^8] \end{aligned} \quad (5)$$

### 1.3.1 Soluzione

(a) Si è testato il codice che implementa la formula 3 su tre vettori contenenti elementi di tipo Float64. Il primo contiene solo numeri 1, il secondo contiene numeri generati a partire da una distribuzione uniforme, il terzo numeri generati a partire da una distribuzione normale. I valori sono riportati in tabella 1.

Table 1: Risultati del test della formula 3 su tre vettori di tipo Float64

Vettore	Lunghezza	Varianza attesa	Varianza (output)
Solo uno	$10^3$	0.000	0.000
Uniforme	$10^8$	0.083	0.083
Normale	$10^8$	1.000	1.000

I risultati di tabella 1 mostrano che l'implementazione del codice è corretta.

(b) Si è testato il codice che implementa la formula 4 su quattro vettori contenenti elementi di tipo Float32, Float64 e Long Double. I risultati sono riportati in tabella 2.

I risultati di tabella 2 mostrano che il calcolo della varianza con la formula 4 è più affetto dalla mancanza di precisione della rappresentazione dei numeri durante il calcolo. Infatti, mano a mano che la precisione dei numeri aumenta, il risultato della formula 4 si avvicina a quello atteso.

C'è da notare che anche la formula 3 è affetta dalla stessa problematica, ma in maniera meno evidente. Basti osservare il risultato del metodo 1 per il vettore  $x_4$  in Float32, che è 0.0, contro

Table 2: Valori di varianza. Metodo 1: equazione 3. Metodo 2: equazione 4.

Set	$\sigma^2$ attesa	Float32		Float64		Long Double	
		Metodo 1	Metodo 2	Metodo 1	Metodo 2	Metodo 1	Metodo 2
$x_1$	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$x_2$	1.0	1.0	-131072.0	1.0	1.0	1.0	1.0
$x_3$	1.0	1.0	0.0	1.0	1.0	1.0	1.0
$x_4$	1.0	0.0	0.0	1.0	0.0	1.0	1.0

l'atteso 1.0.

Parte del problema risiede nella cancellazione numerica. Infatti, per valori degli elementi del vettore molto grandi, la differenza tra il quadrato della somma e la somma dei quadrati è molto piccola e incorre in problema di cancellazione. Per verificare questa affermazione, si riportano in tabella 3 i valori della differenza  $u - \frac{v^2}{n}$  per la formula 4.

Table 3: Differenza  $u - \frac{v^2}{n}$  per ciascun dataset.

Dataset	Single	Double	Long Double
$x_1$	2.0	2.0	2.0
$x_2$	-262144.0	2.0	2.0
$x_3$	0.0	2.0	2.0
$x_4$	0.0	0.0	2.0

Leggendo la tabella 3 possiamo notare che per il dataset  $x_1$  la differenza è esattamente 2.0, che è il risultato atteso. Per il dataset  $x_2$  la differenza è molto grande,  $-262144.0$ , il che indica un significativo problema di cancellazione numerica. Per il dataset  $x_3$  e  $x_4$ , la differenza è 0.0, indice del fatto che i singoli valori  $u - \frac{v^2}{n}$  sono vicini tra loro meno di  $\epsilon_{mach}$ , e perciò la loro differenza è zero.

Per concludere, possiamo dire che la formula 4 è più efficiente in termini di tempo di calcolo, ma è più affetta da errori numerici rispetto alla formula 3.

## 1.4 Esercizio 1.4.2

Sia  $f(x) = \frac{e^x - 1}{x}$ .

(a) Trova il numero di condizionamento  $\kappa_f(x)$ . Qual è il massimo di  $\kappa_f(x)$  nell'intervallo  $-1 \leq x \leq 1$ ?

(b) Usa l'algoritmo "naive"

$$f(x) = \frac{e^x - 1}{x} \quad (6)$$

per calcolare  $f(x)$  per  $x = 10^{-3}, 10^{-4}, 10^{-5}, \dots, 10^{-16}$ .

(c) Crea un secondo algoritmo utilizzando i primi  $n$  termini della serie di McLaurin, cioè

$$p(x) = 1 + \frac{1}{2!}x + \frac{1}{3!}x^2 + \dots + \frac{1}{(n+1)!}x^n. \quad (7)$$

Valutalo sugli stessi valori di  $x$  del punto (b). Per farlo devi scegliere un valore per  $n$ . Verifica la stabilità del risultato al variare di  $n$ . Avresti potuto indovinare un buon valore di  $n$  fin dall'inizio?

(d) Confronta i risultati delle due implementazioni in funzione di  $x$ . Quale algoritmo pensi sia più accurato, e perché?



### 1.4.1 Soluzione

(a) Il numero di condizionamento di una funzione  $f$  è definito come:

$$\kappa_f(x) = \left| \frac{x f'(x)}{f(x)} \right|. \quad (8)$$

Per la funzione  $f(x) = \frac{e^x - 1}{x}$ , calcoliamo la derivata:

$$f'(x) = \frac{e^x x - (e^x - 1)}{x^2} = \frac{e^x(x - 1) + 1}{x^2}. \quad (9)$$

Quindi il numero di condizionamento diventa:

$$\kappa_f(x) = \left| \frac{x \left( \frac{e^x(x-1)+1}{x^2} \right)}{\frac{e^x-1}{x}} \right| = \left| \frac{e^x(x-1)+1}{(e^x-1)x} \right| = \left| \frac{e^x}{e^x-1} - 1 \right|. \quad (10)$$

La derivata di  $k_f(x)$  non si annulla mai nell'intervallo  $-1 \leq x \leq 1$ , e  $\kappa_f(0) = 0$ . Dato che  $\kappa_f(x)$  è monotona crescente per  $x \geq 0$  e monotona decrescente per  $x < 0$ , il massimo si ha in  $x = 1$ , dove assume il valore  $\kappa_f(1) = 0.58198$ .

(b)-(c): Si riportano in figura 4 i risultati dei due algoritmi. La dicitura  $f(x)$  si riferisce alla funzione calcolata con l'algoritmo "naive", mentre  $p_n(x)$  si riferisce alla funzione calcolata con la serie di Maclaurin, fino al termine  $n$ .

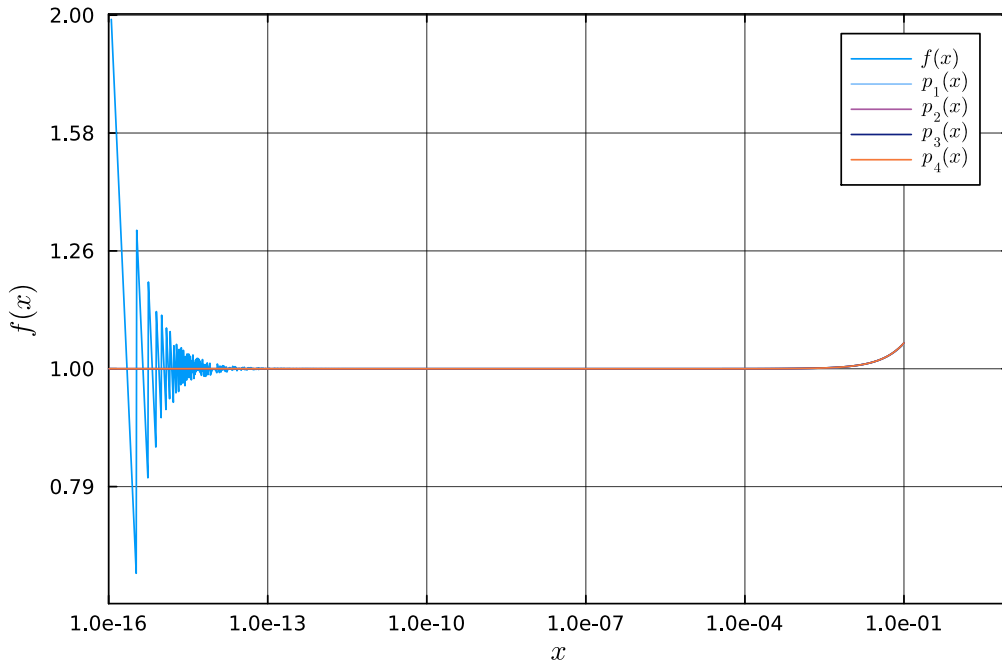


Figure 4: Confronto tra  $f(x)$  e  $p_n(x)$  al variare di  $x$ .

Per uno studio più accurato, si riportano nelle figure 5 e 6 ingrandimenti delle curve di figura 4.

Per quanto riguarda il comportamento della funzione  $f(x)$  calcolata con l'algoritmo 6, possiamo notare che per valori di  $x$  piccoli la funzione assume valori non corretti, nonostante  $\kappa_f(x)$  abbia un massimo non molto grande e peraltro non assunto nell'intorno di zero. La spiegazione di questo fenomeno è che l'algoritmo 6, nonostante non abbia  $\kappa_f(x)$  complessivo molto grande,

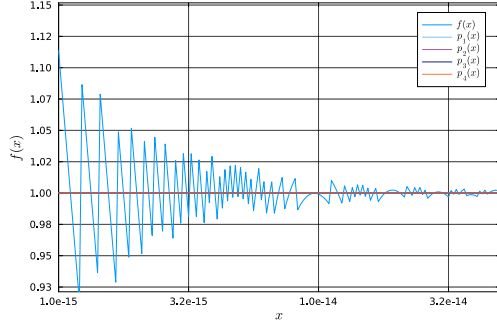


Figure 5: Ingrandimento.  $x \in [10^{-15}, 10^{-14}]$ .

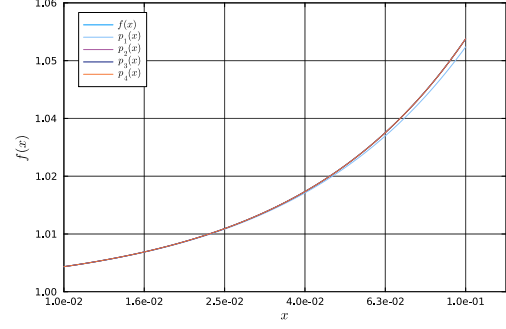


Figure 6: Ingrandimento.  $x \in [10^{-2}, 10^{-1}]$ .

è affetto da errori di cancellazione numerica a causa della forma in cui è scritta la funzione, ma non a causa del fatto che il problema sia intrinsecamente instabile. Di fatto possiamo intuire che esista una possibile riscrittura della funzione che non sia affetta da cancellazione numerica, e che sia più stabile, proprio perchè sappiamo che il massimo di  $\kappa_f(x)$  non è molto grande. Tale riscrittura è quella che utilizza la serie di Mc Laurin, implementata nell'algoritmo  $p_n(x)$ .

Possiamo notare che l'algoritmo 7, per valori di  $x$  piccoli, restituisce il valore corretto, dato che non è affetto da cancellazione numerica. Al contrario, nella regione ingradita di figura 6, l'algoritmo 7 comincia a restituire valori errati perchè ci allontaniamo dalla regione di convergenza della serie di Mc Laurin.

Osservando la figura 6, notiamo che, per valori di  $x$  fissati, l'algoritmo 7 è stabile, cioè le curve che descrivono i valori di  $p_n(x)$  sono sempre più vicine tra loro all'aumentare di  $n$ , fino a sovrapporsi. Per stimare un buon valore di  $n$  da utilizzare per calcolare  $p_n(x)$ , possiamo osservare che la serie di Mc Laurin è tanto meno stabile quanto più ci si allontana da zero. Nel caso di studio valutiamo la distanza tra i valori di  $p_n(x)$  al variare di  $n$ , fissato  $x = 0.1$ . Si è rappresentato in figura 7 l'andamento della distanza tra i valori di  $p_n(0.1)$  e  $p_{n+1}(0.1)$  al variare di  $n$ . In formula:

$$\Delta_n = |p_n(0.1) - p_{n+1}(0.1)|. \quad (11)$$

Ai fini di questo esercizio possiamo dire che  $n = 2$  è già un buon valore perchè le distanze tra i polinomi successivi non sono apprezzabili nel range  $[1, 1.06]$ , caratteristico della figura 6. È chiaro che volendo raggiungere precisione massima occorrerà scegliere  $n = 10$ , dato che  $\Delta_n$  corrispondente è più piccolo di  $\epsilon_{mach}$ .

**(d):** Come confronto tra i due algoritmi si riporta in figura 7 il grafico dell'errore tra i due algoritmi, cioè

$$\Delta(x) = |f(x) - p_n(x)|. \quad (12)$$

La figura 8 mostra una evidente differenza tra i due algoritmi nella regione di instabilità numerica di dell'algoritmo 6, cioè per valori di  $x$  piccoli. E' interessante notare che la distanza tra i due algoritmi diminuisce all'aumentare di  $x$ , ma torna ad aumentare a partire da  $x = 10^{-5}$ , laddove l'algoritmo 7 esce dalla regione di convergenza della serie di Maclaurin.

In conclusione, possiamo affermare che l'algoritmo 7 è più accurato per valori di  $x$  piccoli, mentre l'algoritmo 6 è più accurato per valori di  $x$  grandi.

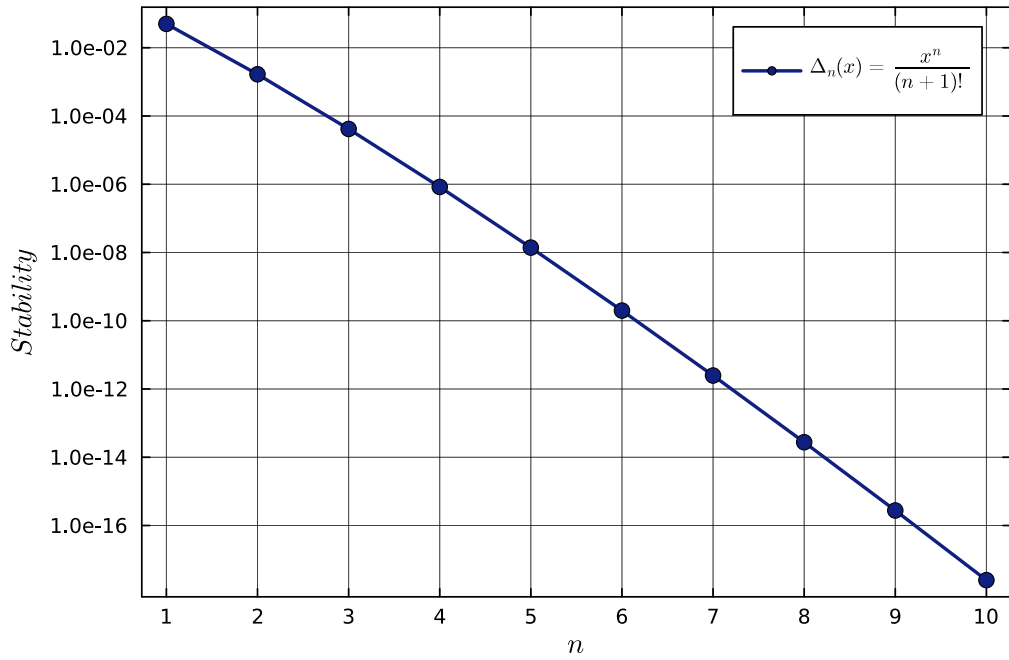


Figure 7: Distanza tra  $p_n(0.1)$  e  $p_{n+1}(0.1)$  al variare di  $n$ .

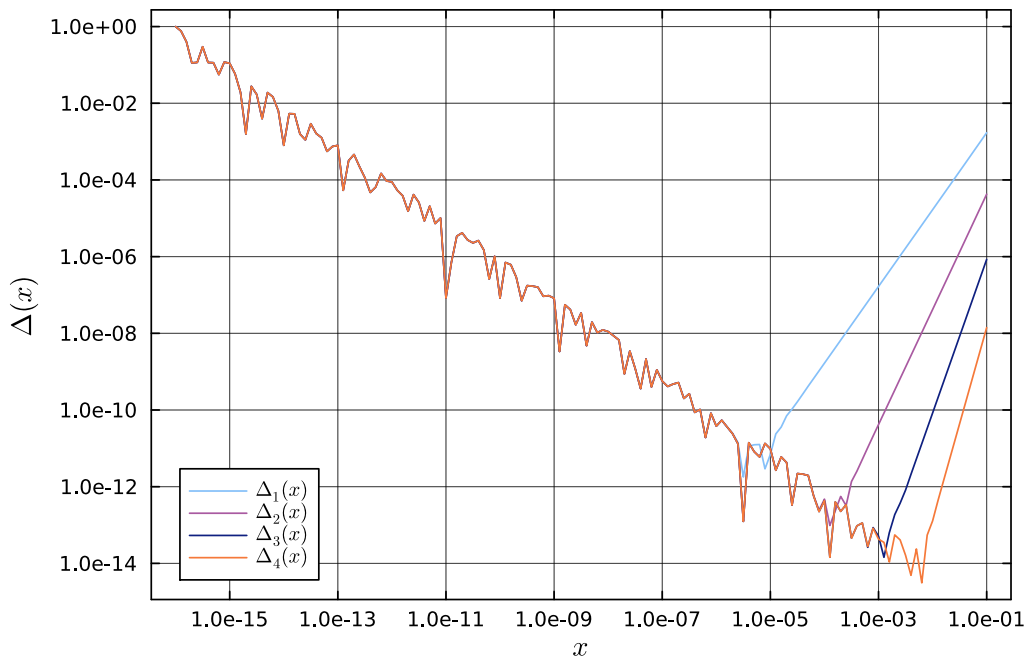


Figure 8: Distanza tra  $f(x)$  e  $p_n(x)$ .

## 2 Sistemi lineari

### 2.1 Esercizi 2.1.1, 2.1.2, 2.1.3, 2.1.4

Scrivi una funzione che esegua la sostituzione in avanti su una matrice triangolare inferiore  $n \times n$ .

2. Testa il tuo codice sui seguenti sistemi lineari:

$$(a) \begin{bmatrix} -2 & 0 & 0 \\ 1 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} -4 \\ 2 \\ 1 \end{bmatrix}$$

$$(b) \begin{bmatrix} 4 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 \\ -1 & 4 & 4 & 0 \\ 2 & -5 & 5 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} -4 \\ 1 \\ -3 \\ 5 \end{bmatrix}$$

Puoi verificare la soluzione risolvendo il sistema a mano e/o calcolando  $\mathbf{Lx} - \mathbf{b}$ .

3. Scrivi una funzione che esegua la sostituzione all'indietro su una matrice triangolare superiore  $n \times n$ .

4. Testa il tuo codice sui seguenti sistemi lineari:

$$(a) \begin{bmatrix} 3 & 1 & 0 \\ 0 & -1 & -2 \\ 0 & 0 & 3 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ 6 \end{bmatrix}$$

$$(b) \begin{bmatrix} 3 & 1 & 0 & 6 \\ 0 & -1 & -2 & 7 \\ 0 & 0 & 3 & 4 \\ 0 & 0 & 0 & 5 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 4 \\ 1 \\ 1 \\ 5 \end{bmatrix}$$

Puoi verificare la soluzione risolvendo il sistema a mano e/o calcolando  $\mathbf{Ux} - \mathbf{b}$ .

### 2.1.1 Soluzione

Per risolvere i sistemi lineari proposti, si sono implementate le funzioni di sostituzione in avanti e all'indietro. Di seguito sono riportati i risultati delle soluzioni.

#### Sostituzione in avanti:

- Per il sistema (a) si ha  $\mathbf{x} = [-4.0 \ 2.0 \ 1.0]$ .
- Per il sistema (b) si ha  $\mathbf{x} = [-4.0 \ 1.0 \ -3.0 \ 5.0]$ .

Risolvendo i sistemi a mano, i valori sono:

- Per il sistema (a) si ha  $\mathbf{x} = [-4 \ 2 \ 1]$ .
- Per il sistema (b) si ha  $\mathbf{x} = [-4 \ 1 \ -3 \ 5]$ .

#### Sostituzione all'indietro:

- Per il sistema (a) si ha  $\mathbf{x} = [1.0 \ 1.0 \ 6.0]$ .
- Per il sistema (b) si ha  $\mathbf{x} = [4.0 \ 1.0 \ 1.0 \ 5.0]$ .

Risolvendo i sistemi a mano, i valori sono:

- Per il sistema (a) si ha  $\mathbf{x} = [1 \ 1 \ 6]$ .
- Per il sistema (b) si ha  $\mathbf{x} = [4 \ 1 \ 1 \ 5]$ .

I risultati confermano la correttezza delle implementazioni delle funzioni di sostituzione in avanti e all'indietro.

## 2.2 Esercizi 2.3.1, 2.3.2, 2.3.4

1. Scrivi una funzione che esegua la fattorizzazione LU di una matrice  $n \times n$ .
2. Per ciascuna matrice, calcola la fattorizzazione LU e verifica la correttezza.

$$(A) \begin{bmatrix} 2 & 3 & 4 \\ 4 & 5 & 10 \\ 4 & 8 & 2 \end{bmatrix} \quad (B) \begin{bmatrix} 6 & -2 & -4 & 4 \\ 3 & -3 & -6 & 1 \\ -12 & 8 & 21 & -8 \\ -6 & 0 & -10 & 7 \end{bmatrix} \quad (C) \begin{bmatrix} 1 & 4 & 5 & -5 \\ -1 & 0 & -1 & -5 \\ 1 & 3 & -1 & 2 \\ 1 & -1 & 5 & -1 \end{bmatrix}$$

4. Calcola il determinante delle matrici dell'Esercizio 2 tramite la fattorizzazione LU.

### 2.2.1 Soluzione

(**Punti 1 e 2:** ) Si sono implementati gli algoritmi di fattorizzazione LU, e si sono testati sui tre sistemi proposti. Le scomposizioni ottenute sono le seguenti:

$$\mathbf{L}_A = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 2.0 & 1.0 & 0.0 \\ 2.0 & -2.0 & 1.0 \end{bmatrix} \quad \mathbf{U}_A = \begin{bmatrix} 2.0 & 3.0 & 4.0 \\ 0.0 & -1.0 & 2.0 \\ 0.0 & 0.0 & -2.0 \end{bmatrix}$$

$$\mathbf{L}_B = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.0 & 0.0 \\ -2.0 & -2.0 & 1.0 & 0.0 \\ -1.0 & 1.0 & -2.0 & 1.0 \end{bmatrix} \quad \mathbf{U}_B = \begin{bmatrix} 6.0 & -2.0 & -4.0 & 4.0 \\ 0.0 & -2.0 & -4.0 & -1.0 \\ 0.0 & 0.0 & 5.0 & -2.0 \\ 0.0 & 0.0 & 0.0 & 8.0 \end{bmatrix}$$

$$\mathbf{L}_C = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ -1.0 & 1.0 & 0.0 & 0.0 \\ 1.0 & -0.25 & 1.0 & 0.0 \\ 1.0 & -1.25 & -1.0 & 1.0 \end{bmatrix} \quad \mathbf{U}_C = \begin{bmatrix} 1.0 & 4.0 & 5.0 & -5.0 \\ 0.0 & 4.0 & 4.0 & -10.0 \\ 0.0 & 0.0 & -2.0 & 6.0 \\ 0.0 & 0.0 & 0.0 & 2.0 \end{bmatrix}$$

Per verificare la correttezza delle scomposizioni, si è calcolato il prodotto  $\mathbf{LU}$  per ciascuna matrice e lo si è sottratto alla matrice originale. Si sono ottenute matrici di zeri, in tutti i casi, confermando la correttezza dell'algoritmo.

(**Punto 4:** ) Si è calcolato il determinante delle matrici dell'esercizio 2 utilizzando la loro fattorizzazione LU. Ricordando le proprietà del determinante e che la matrice  $\mathbf{L}$  è triangolare inferiore con elementi diagonali unitari, si ha che:

$$\det(\mathbf{M}) = \det(\mathbf{LU}) = \det(\mathbf{L}) \cdot \det(\mathbf{U}) = \det(\mathbf{U}). \quad (13)$$

Per il calcolo del determinante delle matrici proposte si è eseguito il prodotto degli elementi della diagonale principale della matrice  $\mathbf{U}$ . A titolo di confronto si è calcolato il determinante delle matrici con la funzione `det` di Julia. I risultati sono riportati in tabella 4.

Table 4: Determinanti delle matrici dell'esercizio 2.

Matrice	Determinante (LU)	Determinante (Julia)
A	4.0	4.0
B	-480.0	-480.0
C	80.0	80.0

Si ottengono risultati identici.

### 2.3 Esercizio 2.3.3

Le matrici

$$\mathbf{T}(x, y) = \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{R}(\theta) = \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

sono usate per rappresentare traslazioni e rotazioni di punti del piano in computer grafica. Per quanto segue, sia

$$\mathbf{A} = \mathbf{T}(3, -1) \mathbf{R}(\pi/5) \mathbf{T}(-3, 1), \quad \mathbf{z} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}.$$

- (a) Calcolare  $\mathbf{b} = \mathbf{Az}$ .
- (b) Trovare la fattorizzazione LU di  $\mathbf{A}$ .
- (c) Usare i fattori e le sostituzioni triangolari per risolvere  $\mathbf{Ax} = \mathbf{b}$  e calcolare  $\mathbf{x} - \mathbf{z}$ .

### 2.3.1 Soluzione

- (a) Si calcola il prodotto tra le matrici  $\mathbf{A}$  e  $\mathbf{z}$ , ottenendo:

$$\mathbf{b} = \mathbf{Az} = \begin{bmatrix} 3.95 \\ 2.01 \\ 1.0 \end{bmatrix}. \quad (14)$$

- (b) Si calcola la fattorizzazione LU della matrice  $\mathbf{A}$ , ottenendo:

$$\mathbf{L} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ -0.8 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 0.81 & 0.59 & 1.16 \\ 0.0 & 1.24 & 2.42 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}. \quad (15)$$

- (c) Si risolve il sistema  $\mathbf{Ax} = \mathbf{b}$  utilizzando le sostituzioni triangolari. Si ottiene, con le dovute approssimazioni:

$$\mathbf{x} = \begin{bmatrix} 2.0 \\ 2.0 \\ 1.0 \end{bmatrix}, \quad \mathbf{x} - \mathbf{z} = \begin{bmatrix} -2.22 \cdot 10^{-16} \\ -4.44 \cdot 10^{-16} \\ 0.0 \end{bmatrix}. \quad (16)$$

## 2.4 Esercizi 2.4.1, 2.4.2

1. Scrivi un programma che esegua la decomposizione di Cholesky di una matrice  $n \times n$ .
2. Per ciascuna matrice, utilizza la decomposizione di Cholesky per determinare se è definita positiva.

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 4 & 5 \\ -1 & 5 & 10 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 5 \\ 1 & 5 & 10 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 5 \\ 1 & 5 & 1 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 6 & 2 & 1 & 0 \\ 2 & 6 & 2 & 1 \\ 1 & 2 & 5 & 2 \\ 0 & 1 & 2 & 4 \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} 4 & 1 & 2 & 7 \\ 1 & 1 & 3 & 1 \\ 2 & 3 & 5 & 3 \\ 7 & 1 & 3 & 1 \end{bmatrix}$$

### 2.4.1 Soluzione

Si è implementato l'algoritmo di decomposizione di Cholesky, e si sono testate le matrici proposte. L'algoritmo è giunto a conclusione per le matrici **A**, **B** e **D**, ma ha restituito un messaggio di errore per le matrici **C** e **E**, indicando che non sono definite positive. Si riportano i risultati delle decomposizioni giunte a termine, con i dovuti arrotondamenti.

$$\mathbf{R}_A = \begin{bmatrix} 1.0 & 0.0 & -1.0 \\ 0.0 & 2.0 & 2.5 \\ 0.0 & 0.0 & 1.7 \end{bmatrix} \quad \mathbf{R}_B = \begin{bmatrix} 1.0 & 0.0 & 1.0 \\ 0.0 & 2.0 & 2.5 \\ 0.0 & 0.0 & 1.7 \end{bmatrix} \quad \mathbf{R}_D = \begin{bmatrix} 2.45 & 0.82 & 0.41 & 0.00 \\ 0.00 & 2.29 & 0.73 & 0.41 \\ 0.00 & 0.00 & 2.06 & 0.82 \\ 0.00 & 0.00 & 0.00 & 1.84 \end{bmatrix}$$

## 2.5 Esercizio 2.5.1

Consideriamo la matrice:

$$\mathbf{A} = \begin{bmatrix} -\epsilon & 1 \\ 1 & -1 \end{bmatrix}$$

Se  $\epsilon = 0$ , la fattorizzazione LU senza pivoting parziale fallisce per **A**. Ma se  $\epsilon \neq 0$ , possiamo procedere senza pivoting, almeno in linea di principio.

- (a) Costruisci  $\mathbf{b} = \mathbf{Ax}$  prendendo  $\epsilon = 10^{-12}$  per la matrice **A** e  $\mathbf{x} = [1, 1]$  come soluzione.
- (b) Fattorizza la matrice usando la decomposizione LU senza pivoting e risolvi numericamente per  $\mathbf{x}$ . Quanto è accurato il risultato?
- (c) Verifica la fattorizzazione LU calcolando  $\mathbf{A} - \mathbf{LU}$ . Funziona?
- (d) Ripeti per  $\epsilon = 10^{-20}$ . Quanto è accurato ora il risultato?
- (e) Calcola il numero di condizionamento della matrice **A**; puoi scegliere la norma che preferisci. Il sistema  $\mathbf{Ax} = \mathbf{b}$  è mal condizionato?
- (f) Calcola le matrici **L** e **U** della fattorizzazione LU di  $\mathbf{A} = \mathbf{LU}$  a mano.
- (g) Calcola i numeri di condizionamento di **L** e **U**.

### 2.5.1 Soluzione

Per poter operare un confronto tra i risultati ottenuti con  $\epsilon = 10^{-12}$  e  $\epsilon = 10^{-20}$ , si è scelto di trattare il punto (d) insieme ai punti (a), (b) e (c). Il pedice 1 è legato al caso  $\epsilon = 10^{-12}$ , il pedice 2 al caso  $\epsilon = 10^{-20}$ .

- (a) Si è calcolato il prodotto tra la matrice **A** e il vettore  $\mathbf{x}$ , ottenendo, a meno di arrotondamenti, i vettori:

$$\mathbf{b}_1 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}. \quad (17)$$

- (b) Si è calcolata la fattorizzazione LU della matrice **A**, ottenendo:

$$\mathbf{L}_1 = \begin{bmatrix} 1.0 & 0.0 \\ -1.0 \cdot 10^{12} & 1.0 \end{bmatrix}, \quad \mathbf{U}_1 = \begin{bmatrix} -1.0 \cdot 10^{-12} & 1.0 \\ 0.0 & 1.0 \cdot 10^{12} \end{bmatrix}, \quad (18)$$

$$\mathbf{L}_2 = \begin{bmatrix} 1.0 & 0.0 \\ -1.0 \cdot 10^{20} & 1.0 \end{bmatrix}, \quad \mathbf{U}_2 = \begin{bmatrix} -1.0 \cdot 10^{-20} & 1.0 \\ 0.0 & 1.0 \cdot 10^{20} \end{bmatrix}. \quad (19)$$

Si è risolto il sistema  $\mathbf{Ax} = \mathbf{b}$  utilizzando le sostituzioni triangolari, ottenendo:

$$\mathbf{x}_1 = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.0 \\ 1.0 \end{bmatrix}. \quad (20)$$

(c) Si è calcolato il prodotto tra le matrici  $\mathbf{L}$  e  $\mathbf{U}$ , in seguito lo si è sottratto ad  $\mathbf{A}$  ottenendo:

$$\mathbf{A}_1 - (\mathbf{LU})_1 = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}, \quad \mathbf{A}_2 - (\mathbf{LU})_2 = \begin{bmatrix} 0.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}. \quad (21)$$

È evidente che il risultato per  $\epsilon = 10^{-12}$  è accurato, al contrario del risultato ottenuto per  $\epsilon = 10^{-20}$ , in cui sia la soluzione che la verifica della fattorizzazione LU sono affette da errori numerici.

(e) Si è calcolato  $\kappa(A)$ , utilizzando sia la norma 1 che la norma infinito, ottenendo lo stesso numero a causa della natura simmetrica della matrice  $\mathbf{A}$ . La formula generale per il numero di condizionamento della matrice  $\mathbf{A}$  presa in esame è  $\kappa(A) = \frac{4}{1-\epsilon}$ .

$$\kappa_1(A) = \kappa_\infty(A) = 4.0000000000004, \quad \kappa_1(A) = \kappa_\infty(A) = 4.0. \quad (22)$$

Come si può notare, in entrambi i casi il numero di condizionamento non è molto grande, quindi il sistema  $\mathbf{Ax} = \mathbf{b}$  non è mal condizionato.

(f) Si sono calcolate a mano le matrici  $\mathbf{L}$  e  $\mathbf{U}$  della fattorizzazione LU di  $\mathbf{A}$  per ciascun valore di  $\epsilon$ . I calcoli dettagliati sono disponibili nella cartella git, nel percorso `lab\computazionale1\esercizi\pen_and_paper`. Ad ogni modo le matrici ottenute sono già state riportate al punto 2.5.1.

(g) Si sono calcolati i numeri di condizionamento di  $\mathbf{L}$  e  $\mathbf{U}$ , utilizzando sia la norma 1 che la norma infinito, ancora una volta ottenendo lo stesso numero.

Le formule generali per il numero di condizionamento delle matrici  $\mathbf{L}$  e  $\mathbf{U}$  sono:

$$\kappa_1(L) = \kappa_\infty(L) = \left(1 + \frac{1}{\epsilon}\right)^2, \quad \kappa_1(U) = \kappa_\infty(U) = \frac{1}{\epsilon^2}. \quad (23)$$

In tabella 5 sono riassunti i risultati ottenuti.

Table 5: Valori dei parametri  $L$  e  $U$  per diverse tolleranze  $\epsilon$

	$\epsilon = 10^{-12}$	$\epsilon = 10^{-20}$
$L$	$1.0 \cdot 10^{24}$	$1.0 \cdot 10^{40}$
$U$	$1.0 \cdot 10^{24}$	$1.0 \cdot 10^{40}$

## 2.5.2 Conclusioni

L'esercizio ha mostrato come il numero di condizionamento di una matrice non sia un indicatore sufficiente per valutare l'esattezza della risoluzione del sistema lineare associato. Infatti, sebbene il numero di condizionamento della matrice  $\mathbf{A}$  sia relativamente basso, la risoluzione del sistema richiede il calcolo delle matrici  $\mathbf{L}$  e  $\mathbf{U}$ , il cui numero di condizionamento potrebbe risultare elevato. È questo il caso con  $\epsilon = 10^{-20}$ , in cui  $\kappa(A) = 4.0$  ma  $\kappa(L) = 1.0 \cdot 10^{40}$  e  $\kappa(U) = 1.0 \cdot 10^{40}$ . Dato che l'algoritmo utilizza le matrici  $\mathbf{L}$  e  $\mathbf{U}$  per risolvere il sistema è evidente che l'errore numerico aumenta a causa delle stesse.



## 2.6 Esercizio 2.6.1

- (a) Scrivi un programma che prenda in input una matrice  $\mathbf{A}$  e un vettore  $\mathbf{b}$  e risolva il problema dei minimi quadrati  $\arg\min \|\mathbf{b} - \mathbf{Ax}\|_2$  utilizzando la decomposizione di Cholesky.
- (b) Per verificare il funzionamento del tuo codice, calcola la soluzione ai minimi quadrati quando

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 0 & 1 \\ -2 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ -5 \\ 6 \end{bmatrix}$$

### 2.6.1 Soluzione

Si è implementato l'algoritmo per il metodo dei minimi quadrati e lo si è testato sulle matrici proposte, ottenendo:

$$\mathbf{x} = \begin{bmatrix} -2.0 \\ -1.0 \end{bmatrix}. \quad (24)$$

## 2.7 Esercizio 2.6.2

Keplero ha scoperto che il periodo orbitale  $\tau$  di un pianeta dipende dalla sua distanza media  $R$  dal Sole secondo la legge  $\tau = cR^\alpha$ , dove  $\alpha$  è un semplice numero razionale.

- (a) Esegui un fit lineare ai minimi quadrati utilizzando la tabella 12 per determinare il valore più probabile e semplice di  $\alpha$ .
- (b) Realizza un grafico dei dati e del risultato del fit.

### 2.7.1 Soluzione

(a) Il metodo dei minimi quadrati che è stato sviluppato durante il corso è adatto a risolvere problemi lineari, in cui le costanti che si vogliono determinare sono coefficienti lineari di funzioni dei dati. Non è il caso del problema proposto, dato che  $\alpha$  si trova ad esponente. Per risolvere la questione si è scelto di linearizzare il problema tramite il logaritmo naturale. In formule:

$$\ln(\tau) = \ln(c) + \alpha \ln(R). \quad (25)$$

Da qui in poi si è implementato il metodo dei minimi quadrati, applicando la scomposizione di Cholesky alla matrice  $n \times n$  ottenuta dal prodotto della matrice di design e la sua trasposta. Per chiarezza si riporta la matrice di design trasposta:

$$\mathbf{A}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \ln(57.59) & \ln(108.11) & \cdots & \ln(4499.9) \end{bmatrix}. \quad (26)$$

La matrice di regressione è stata costruita in modo che la prima colonna fosse composta da 1, per poter calcolare il termine noto della retta, e la seconda colonna fosse composta dai logaritmi naturali delle distanze dei pianeti dal Sole.

La legge di Keplero prevista è la seguente:

$$\tau = cR^{3/2}, \quad \text{con} \quad \alpha = \frac{3}{2} \quad c = \frac{2\pi}{\sqrt{G(M+m)}} \quad (27)$$

dove  $G$  è la costante di gravitazione universale,  $M$  è la massa del Sole,  $m$  è la massa del pianeta. La regressione lineare che si vuole effettuare considera  $c$  costante, nonostante dipenda dalla

massa del pianeta. Consideriamo Giove, il pianeta più massivo di quelli presi in esame ( $m \approx 1.90 \cdot 10^{27}$  kg). Si ha:

$$c_1 = \frac{2\pi}{\sqrt{G(M+m)}} = 5.47 \cdot 10^{-10} \frac{\text{sec}}{m^{3/2}} = 0.203 \frac{\text{days}}{Mkm^{3/2}} \quad (28)$$

$$c_2 = \frac{2\pi}{\sqrt{GM}} = 5.47 \cdot 10^{-10} \frac{\text{sec}}{m^{3/2}} = 0.203 \frac{\text{days}}{Mkm^{3/2}} \quad (29)$$

$$\frac{c_1}{c_2} \approx 1.0 \quad (30)$$

Possiamo quindi affermare che la costante  $c$  è indipendente dalla massa del pianeta, entro la precisione necessaria per il nostro scopo.

Si sono quindi stimati il valore di  $\alpha$  e della costante  $c$ , tramite la regressione, ottenendo:

$$c = 0.206 \frac{\text{days}}{Mkm^{3/2}} \quad \alpha = 1.499 \quad (31)$$

I risultati sono in accordo con quanto atteso.

(b) Si riporta in figura 9 il grafico dei dati e relativo fit.

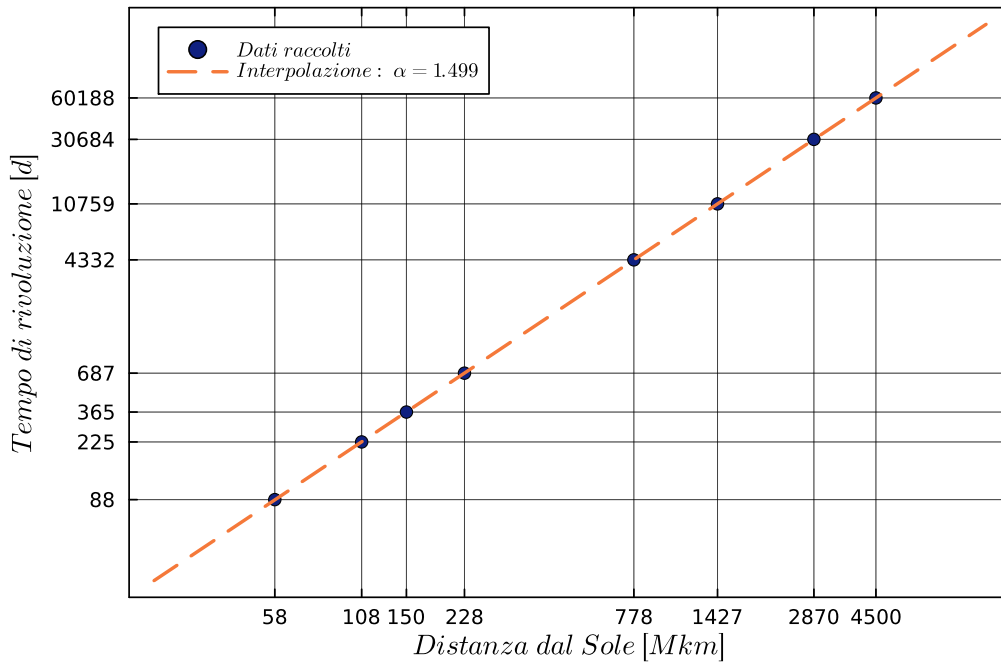


Figure 9: Fit lineare ai minimi quadrati dei dati dei pianeti del Sistema Solare.

## 2.8 Esercizio 2.6.3

In questo esercizio si vuole trovare un'approssimazione della funzione periodica  $g(t) = e^{\sin(t-1)}$  su un periodo,  $0 < t \leq 2\pi$ . Come dati, si definiscono

$$t_i = \frac{2\pi i}{60}, \quad y_i = g(t_i), \quad i = 1, \dots, 60. \quad (32)$$

(a) Trova i coefficienti del fit ai minimi quadrati

$$y(t) \approx c_1 + c_2 t + \dots + c_7 t^6. \quad (33)$$

Sovrapponi un grafico dei valori dei dati come punti con una curva che mostra il fit.  
(b) Trova i coefficienti del fit ai minimi quadrati

$$y \approx d_1 + d_2 \cos(t) + d_3 \sin(t) + d_4 \cos(2t) + d_5 \sin(2t). \quad (34)$$

A differenza del punto (a), questa funzione di fitting è essa stessa periodica. Sovrapponi un grafico dei valori dei dati come punti con una curva che mostra il fit.

### 2.8.1 Soluzione

Dopo avere generato i dati richiesti, si è eseguito il metodo dei minimi quadrati con le seguenti matrici di regressione:

$$\mathbf{D}_{\text{pol}} = \begin{bmatrix} 1 & t_1 & \dots & t_1^6 \\ 1 & t_2 & \dots & t_2^6 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{60} & \dots & t_{60}^6 \end{bmatrix}, \quad \mathbf{D}_{\text{Four}} = \begin{bmatrix} 1 & \cos(t_1) & \sin(t_1) & \cos(2t_1) & \sin(2t_1) \\ 1 & \cos(t_2) & \sin(t_2) & \cos(2t_2) & \sin(2t_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos(t_{60}) & \sin(t_{60}) & \cos(2t_{60}) & \sin(2t_{60}) \end{bmatrix}, \quad (35)$$

ottenendo i coefficienti delle formule 33 e 34:

- Coefficienti del polinomio:  $[0.712; -1.351; 2.225; -0.538; -0.074; 0.033; -0.003]$
- Coefficienti della funzione di Fourier:  $[1.266; -0.951; 0.611; 0.113; -0.247]$

Si riportano i grafici dei dati e dei fit in figura 10.

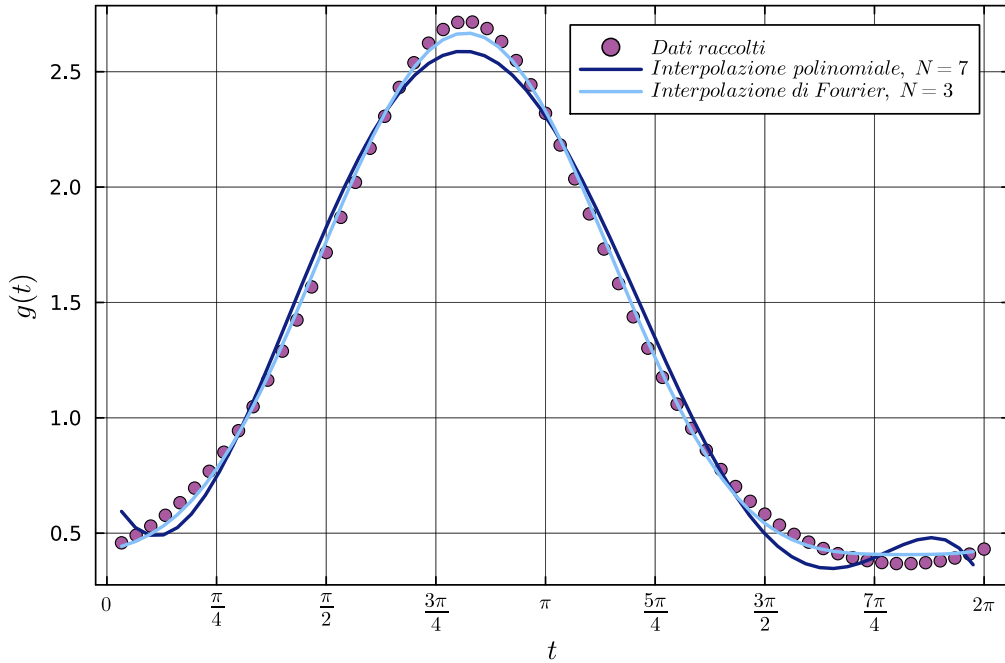


Figure 10: Fit polinomiale e di Fourier ai minimi quadrati della funzione  $g(t) = e^{\sin(t-1)}$ .

### 2.8.2 Conclusioni

In figura 10 si osserva che il fit meglio riuscito è quello di Fourier, nonostante richieda meno parametri. Il motivo è che la funzione  $g(t)$  è periodica, e quindi il fit con il modello 34 riesce a

catturare meglio la periodicità della funzione rispetto al fit eseguito con 33.

In un contesto sperimentale la funzione  $g(t)$  non è nota a priori, anzi lo scopo dell'esperimento è solitamente quello di risalire a tale funzione. Il fatto che il fit secondo il modello di Fourier riesca meglio è indice della natura periodica del fenomeno osservato.

## 3 Radici di equazioni non lineari

### 3.1 Esercizi 3.2.1, 3.2.2

1. Implementa il metodo di bisezione e testalo con una funzione semplice, ad esempio  $f(x) = x$ .
2. (a) Usa il metodo di bisezione per trovare le soluzioni di  $x^3 - 7x^2 + 14x - 6 = 0$  su  $[0, 1]$ .  
(b) Studia la convergenza del metodo di bisezione rispetto alla radice trovata in (a). (Puoi prendere come approssimazione per la radice  $r = m_{n_{\max}}$ , dove  $n_{\max}$  è l'indice corrispondente all'ultima iterazione di bisezione.)

Qual è il suo ordine di convergenza  $q$ ? Puoi stimare la costante d'errore asintotica  $C$ ?

Suggerimento: Studia  $d_n = -\log_{10} |x_n - r|$  come funzione di  $n$ .

#### 3.1.1 Soluzione

(1) Si è implementato il metodo di bisezione, e si è testato con la funzione  $f(x) = x$  per due intervalli:  $[-1.0, 1.0]$  e  $[-5.0, 1.0]$ . Il motivo della scelta di questi due intervalli è che ci aspettiamo che il metodo restituisca  $x = 0$  come soluzione e nel caso di intervallo simmetrico il numero di iterazioni è pari ad 1. Con un intervallo asimmetrico il test è più efficiente perchè il numero di iterazioni aumenta.

Il metodo si interrompe quando la distanza tra gli estremi risulta minore di  $\max(|a|, |b|) \cdot 10^{-16}$ , dove  $a$  e  $b$  sono gli estremi dell'intervallo. Il metodo ha prodotto i seguenti risultati:

- Intervallo  $[-1.0, 1.0]$ :  $x = 0.0$  in 1 iterazione.
- Intervallo  $[-5.0, 1.0]$ :  $x = 1.1 \cdot 10^{-162}$  in 539 iterazioni.

(2a) Si è implementato il metodo di bisezione per la funzione  $f(x) = x^3 - 7x^2 + 14x - 6$ . Il metodo ha restituito un valore della radice  $r = 0.5857864376269051$  con  $f(r) = 0.0$ , in 50 iterazioni. In figura 11 si riporta il grafico della funzione e delle radici successive trovate.

(2b) Per studiare la convergenza del metodo di bisezione bisogna fare alcune considerazioni preliminari.

È noto che per il metodo di bisezione vale la relazione  $|m_n - r| < 2^{-(n+1)}(b_0 - a_0)$ , dove  $m_n$  è la stima della radice al passo  $n$ ,  $r$  è la radice cercata,  $a_0$  e  $b_0$  sono gli estremi dell'intervallo iniziale.

Ad ogni iterazione l'intervallo viene dimezzato, in modo che al passo successivo la stima della radice si trovi in una delle due metà dell'intervallo nella quale si trovava la stima precedente.

In formule:  $|m_{n+1} - r| < \frac{|m_n - r|}{2}$ .

Unendo le due relazioni presentate sopra si ricava  $\frac{|m_{n+1} - r|}{|m_n - r|} < \frac{1}{2}$ , e passando al limite:

$$\lim_{n \rightarrow \infty} \frac{|m_{n+1} - r|}{|m_n - r|} = \frac{1}{2}, \quad (36)$$

da cui si ottiene che  $q$ , ordine di convergenza, è 1, mentre  $C$ , la costante di errore asintotico, vale  $\frac{1}{2}$ .

Verifichiamolo numericamente, ricordando le relazioni:

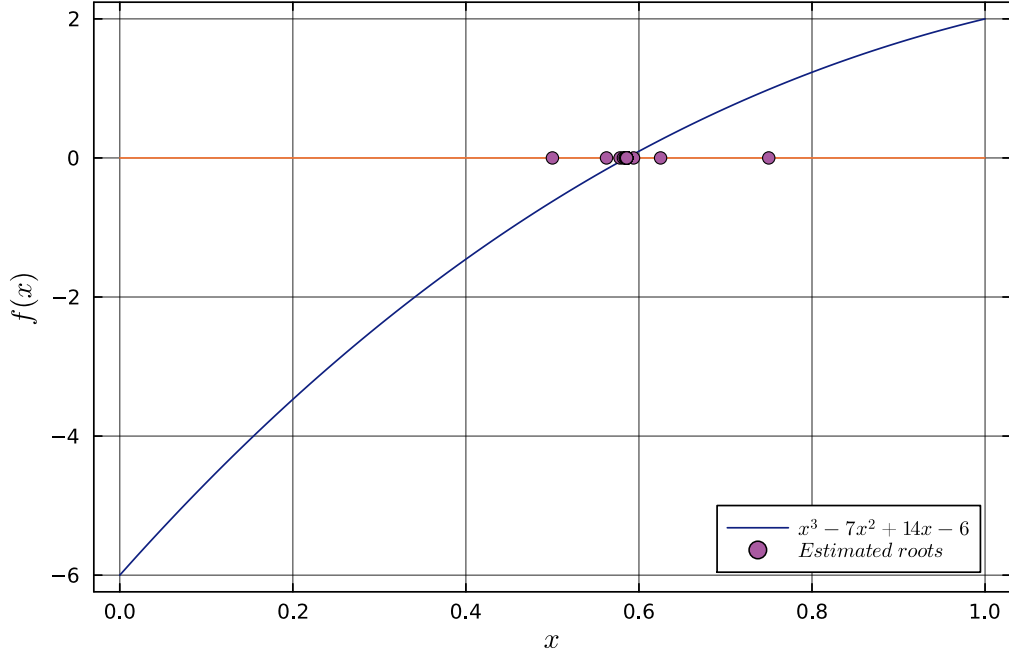


Figure 11: Grafico della funzione  $f(x) = x^3 - 7x^2 + 14x - 6$  e della radice trovata con il metodo di bisezione.

Per verificare i valori teorici di  $q$  e  $C$  è necessario tenere a mente le seguenti relazioni. La prima si ottiene applicando il logaritmo decimale alla 36:

$$d_{n+1} \approx q d_n - \log_{10} |C| \quad \Rightarrow \quad \frac{d_{n+1}}{d_n} \approx q - \frac{\log_{10} |C|}{d_n} \xrightarrow{n \rightarrow \infty} q \quad (37)$$

dove  $d_n = -\log_{10} |m_n - r|$ .

La seconda si ottiene iterando la relazione  $|x_{k+1} - r| = C|x_k - r|$ :

$$|x_{k+1} - r| = C^{k+1} |x_0 - r| \quad (38)$$

**Stima di  $q$  e  $C$**  Per stimare  $q$  si è calcolato il rapporto tra i valori di  $d_n$  e  $d_{n+1}$  e lo si è messo in grafico in funzione di  $n$ , ottenendo la figura 12, in cui è presente un ingrandimento per i valori più alti di  $n$ .

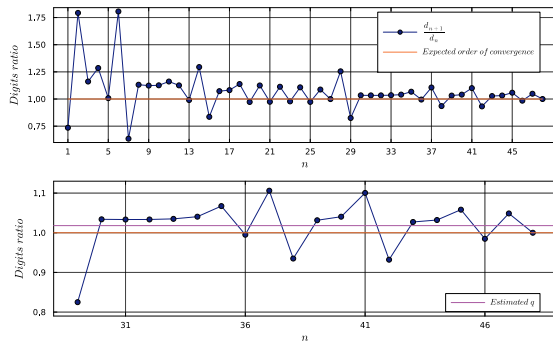


Figure 12: Stima di  $q$

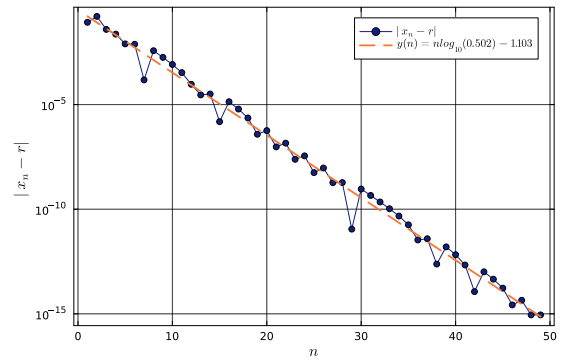


Figure 13: Stima di  $C$

Il rapporto tende a 1, come atteso. Si vuole ora quantificare tale valore. La relazione 37 vale per  $n \rightarrow \infty$ . Ciò significa che la stima di  $q$  dovrebbe essere calcolata per l' $n$

più grande possibile. Ciononostante si è deciso di stimare il valore di  $q$  tramite l'interpolazione con funzione costante perchè il valore di  $q$  fluttua attorno a 1, ed è possibile che per  $n$  maggiori di quelli disponibili il valore cercato continui a fluttuare attorno a 1. Per l'interpolazione si sono considerati gli ultimi 20 punti della serie, riportati nell'ingrandimento di 12. Il fit su tali punti restituisce un valore di  $q = 1.02$ .

Per stimare  $C$  si è eseguito il fit con metodo dei minimi quadrati sul modello:

$$\log|x_{k+1} - r| = (k + 1)\log(C) + \log|x_0 - r| \rightarrow y(x) = x\log(C) + A, \quad (39)$$

avendo considerato  $r$  come l'ultimo valore di  $m_n$  calcolato. Si riporta il grafico del fit in figura 13.

Il fit ha restituito i seguenti valori:

$$C = 0.5, \quad A = -0.1. \quad (40)$$

Si osserva che il valore di  $C$  è in accordo con quanto atteso, mentre il valore di  $A$  non è significativo.

**Commento:** È possibile stimare i valori di  $q$  e  $C$  anche in modo diverso, e con un solo fit. Interpolando la relazione 37 con un'iperbole, e ignorando il limite, si riescono a ottenere i valori di  $q$  e  $C$  in un colpo solo. I risultati ottenuti hanno una precisione minore di quelli ottenuti con i metodi precedenti, pertanto non sono riportati. Si rimanda alla cartella `git lab\computazionale1\esercizi\3_2_2.ipynb` per i dettagli.

## 3.2 Esercizi 3.3.1, 3.3.2

1. Scrivi un programma che implementi il metodo di Newton.
  2. Per ciascuna delle seguenti funzioni, esegui i seguenti passi.
    - (a) Riscrivi l'equazione nella forma standard per la ricerca degli zeri,  $f(x) = 0$ , e calcola  $f'(x)$ .
    - (b) Traccia il grafico di  $f$  nell'intervallo indicato e determina quante radici sono presenti nell'intervallo.
    - (c) Usa il metodo di Newton per trovare ciascuna radice.
    - (d) Studia l'errore nella sequenza di Newton per determinare numericamente se la convergenza è approssimativamente quadratica per la radice trovata.
- Suggerimento: Studia  $d_{n+1}/d_n$  con  $d_n = -\log_{10} |x_n - r|$  per verificare il tasso di convergenza  $q$ . Una volta stabilito, prova a stimare la costante d'errore asintotica  $C$ .

- $x^2 = e^{-x}$ , su  $[-2, 2]$
- $2x = \tan x$ , su  $[-0.2, 1.4]$
- $e^{x+1} = 2 + x$ , su  $[-2, 2]$

### 3.2.1 Soluzione

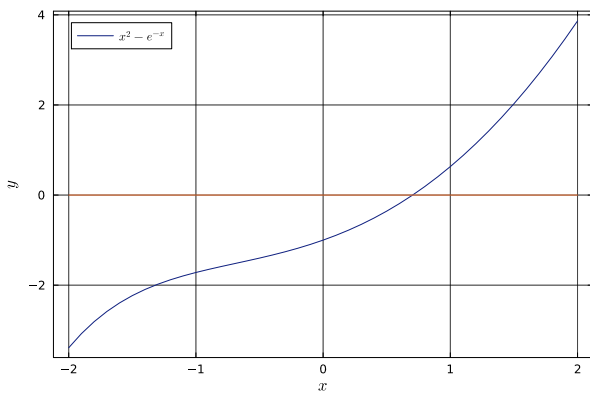
(1) Si è implementato il metodo di Newton, e si è testato con la funzione  $f(x) = e^x - 1$  con valore iniziale  $x_0 = 10$ . Il metodo ha restituito il valore  $-6.68 \cdot 10^{-17}$  in 16 iterazioni. Valore atteso:  $x = 0$ .

(2a) Le funzioni in forma di ricerca degli zeri sono:

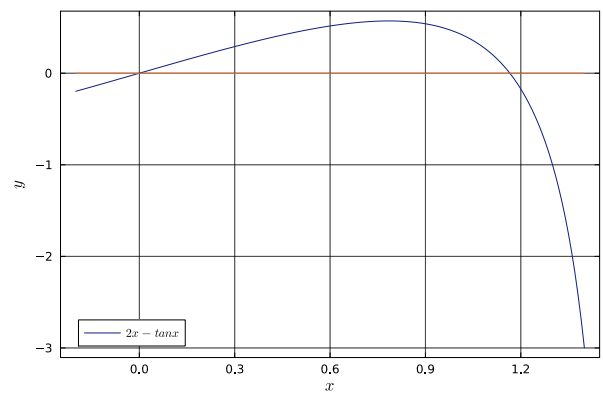
- $f_1(x) = x^2 - e^{-x}$ , con  $f'_1(x) = 2x + e^{-x}$ , su  $[-2, 2]$ .
- $f_2(x) = 2x - \tan x$ , con  $f'_2(x) = 2 - \tan^2 x$ , su  $[-0.2, 1.4]$ .
- $f_3(x) = e^{x+1} - 2 - x$ , con  $f'_3(x) = e^{x+1} - 1$ , su  $[-2, 2]$ .

(2b) I grafici delle funzioni sono riportati in figura 14, da cui si evince che:

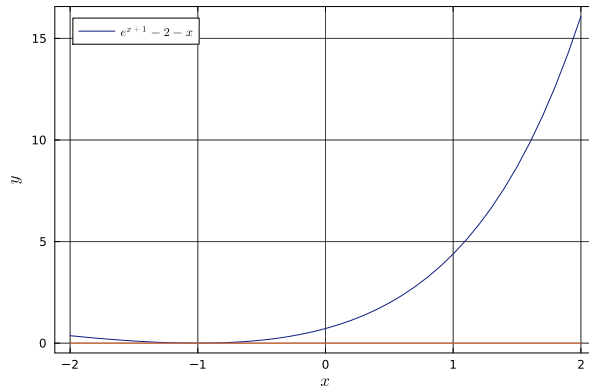
- $f_1$  ha una radice nell'intervallo  $[-2, 2]$ .
- $f_2$  ha due radici nell'intervallo  $[-0.2, 1.4]$ .
- $f_3$  ha una radice nell'intervallo  $[-2, 2]$ .



(a)



(b)



(c)

Figure 14: Grafici delle funzioni  $f_1$ ,  $f_2$  e  $f_3$ .

(2c) L'algoritmo di Newton è stato applicato alle tre funzioni, con i seguenti risultati:

- $f_1$ : radice  $x = 0.703$ , in 8 iterazioni. Punto di partenza  $x_0 = -2$ .
- $f_2$ : radici  $x_1 = 0.000$ ,  $x_2 = 1.166$ , entrambe in 7 iterazioni. Punti di partenza  $x_{0,1} = 0.65$ ,  $x_{0,2} = 1.4$ .
- $f_3$ : radice  $x = -1.000$  in 29 iterazioni. Punto di partenza  $x_0 = 2$ .

**(2d) - Prima funzione** Scopo dello studio dell'errore è verificare se i valori di  $q$  e  $C$  sono in accordo con quanto atteso. In particolare ci aspettiamo che il metodo di Newton per questo particolare zero abbia  $q = 2$  e  $C = \frac{1}{2} \left| \frac{f''(r)}{f'(r)} \right| = 0.396$ . Per determinare i valori di  $q$ , qui e nelle altre funzioni, si fa riferimento alla formula 37, e si calcola il rapporto tra i valori di  $d_n$  e  $d_{n+1}$ ; per determinare  $C$ , qui e nelle altre funzioni, si fa riferimento alla relazione  $\lim_{n \rightarrow \infty} \frac{|x_{n+1}-r|}{|x_n-r|^q} = \lim_{n \rightarrow \infty} \frac{|\epsilon_{n+1}|}{|\epsilon_n|^q} = C$ . Si riportano i due grafici in figura 15.

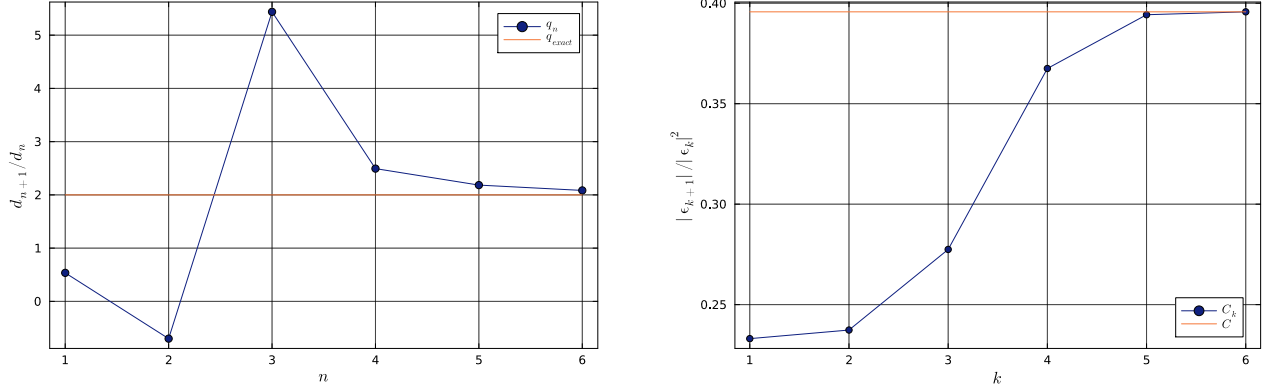


Figure 15: Studio del valore di  $q$  e  $C$  per  $f_1$ .

Per le stime di  $q$  e  $C$  si è scelto di considerare l'ultimo punto della successione corrispondente, a causa del numero limitato di iterazioni. Si ottiene:

$$q = 2.084, \quad C = 0.396. \quad (41)$$

Si osserva che i valori ottenuti sono in accordo con quanto atteso, e quindi si può concludere che il metodo di Newton converge in modo quadratico per la radice trovata.

## (2d) - Seconda funzione

**Prima radice,  $x = 0.0$ :** È importante notare che la funzione  $f_2$  in zero ha un andamento lineare, e perciò la sua derivata seconda è nulla. Questo implica che i valori di  $q$  e  $C$  non sono determinati dalle formule precedenti, perchè tali formule sono ricavate a partire da un'espansione di Taylor al secondo ordine. Per ovviare al problema si è scelto di ricavare le formule già utilizzate ma a ordine 3. I calcoli sono riportati nella cartella git `lab\computazionale1\esercizi\pen_and_paper`.

Di seguito sono riportati i valori per  $q$  e  $C$  con le dovute correzioni:

$$q = 3, \quad C = \frac{f'''(r)}{3f'(r)} = 0.667. \quad (42)$$

In figura 16 si riportano i grafici delle successioni  $q_n$  e  $C_n$  per la prima radice di  $f_2$ .

Si ottengono i seguenti valori:

$$q = 3.065, \quad C = 0.667. \quad (43)$$

Ci si può chiedere cosa succede se, non conoscendo l'andamento della funzione, si utilizzano le formule precedenti. In particolare cosa succede se si fissa  $q = 2$  e si stima  $C$  come l'ultimo elemento della successione  $C_n = \frac{|x_{n+1}-r|}{|x_n-r|^2}$ . L'aspettativa è che  $C$  risulti nullo, dato che la derivata seconda è nulla. Il grafico corrispondente può essere trovato nella cartella git `lab_computazionale1\esercizi\3_3_2.ipynb`, e mostra che il valore di  $C$  è effettivamente nullo.



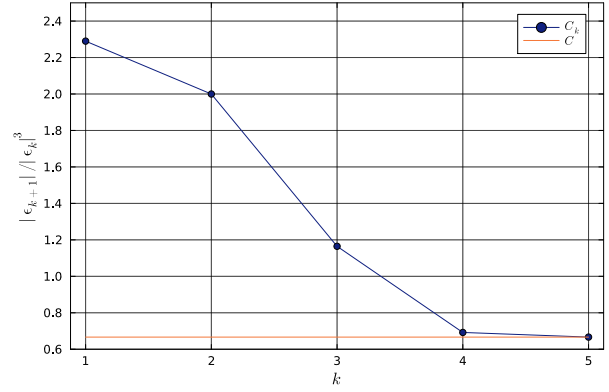
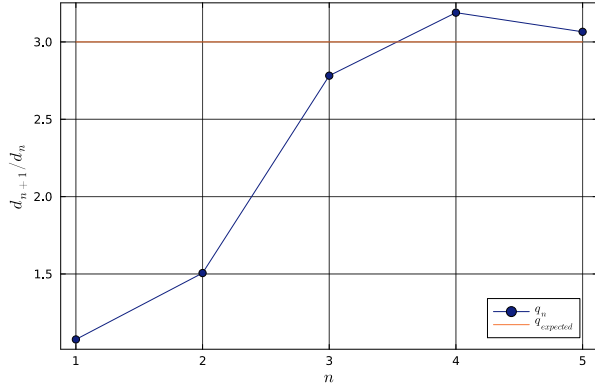


Figure 16: Studio del valore di  $q$  e  $C$  per  $f_2$ , prima radice.

**Seconda radice,  $x = 1.166$ :** la seconda radice di  $f_2$  non presenta problemi, così le formule sono quelle viste inizialmente. I valori attesi sono  $q = 2$  e  $C = 3.383$ . Si riportano i grafici delle successioni  $q_n$  e  $C_n$  in figura 17. Si ottengono i seguenti valori:

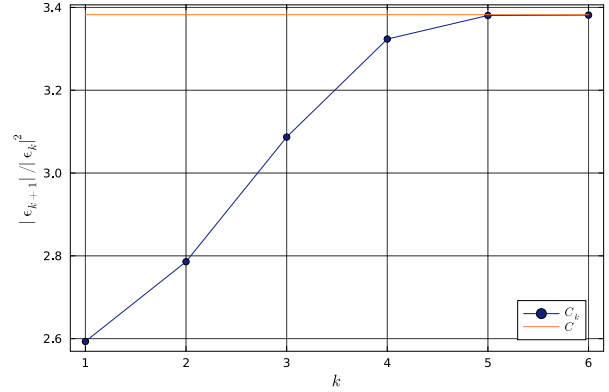
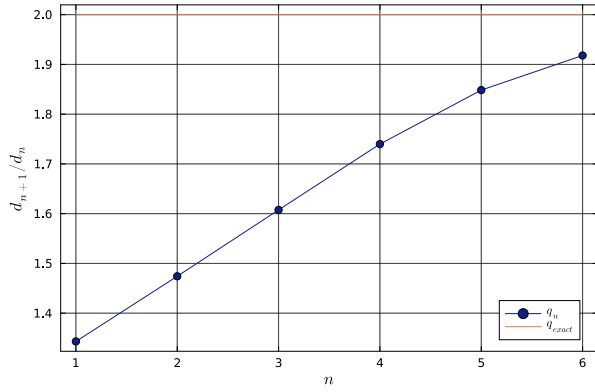


Figure 17: Studio del valore di  $q$  e  $C$  per  $f_2$ , seconda radice.

$$q = 2.000, \quad C = \frac{f''(r)}{2f'(r)} = 3.382. \quad (44)$$

In accordo con quanto atteso, il metodo di Newton converge in modo quadratico per la radice trovata.

**(2d) - Terza funzione** Osservando il grafico della funzione  $f_3$  si nota che la radice è un punto di minimo, e quindi ha una molteplicità  $m$ . D'altra parte, possiamo intuire che la radice di  $f_3$  ha una molteplicità anche dal numero di iterazioni, molto elevato rispetto alle altre funzioni. Come è noto dalla teoria, il metodo di Newton in caso di radici multiple converge linearmente, e nella formula  $\epsilon_{n+1} = C(\epsilon_n)^q$  ci aspettiamo  $q = 1$ . Inoltre in caso di radici multiple la costante d'errore asintotica diviene  $(1 - \frac{1}{m})$ .

Il primo passo è verificare il valore di  $q$ . Si riporta in figura 18 il grafico della successione  $q_n$ , da cui possiamo confermare stimare  $q = 1.067$ .

Si riportano in figura 18 i grafici della successione  $C_n$ , una volta senza considerare la molteplicità della radice, e una volta considerando la molteplicità della radice.

Come si può notare da 18, non avendo considerato la molteplicità della radice, il valore di  $C$  si discosta molto da quello calcolato con la solita formula  $C = \frac{f''(r)}{2f'(r)}$ . Invece, nel secondo

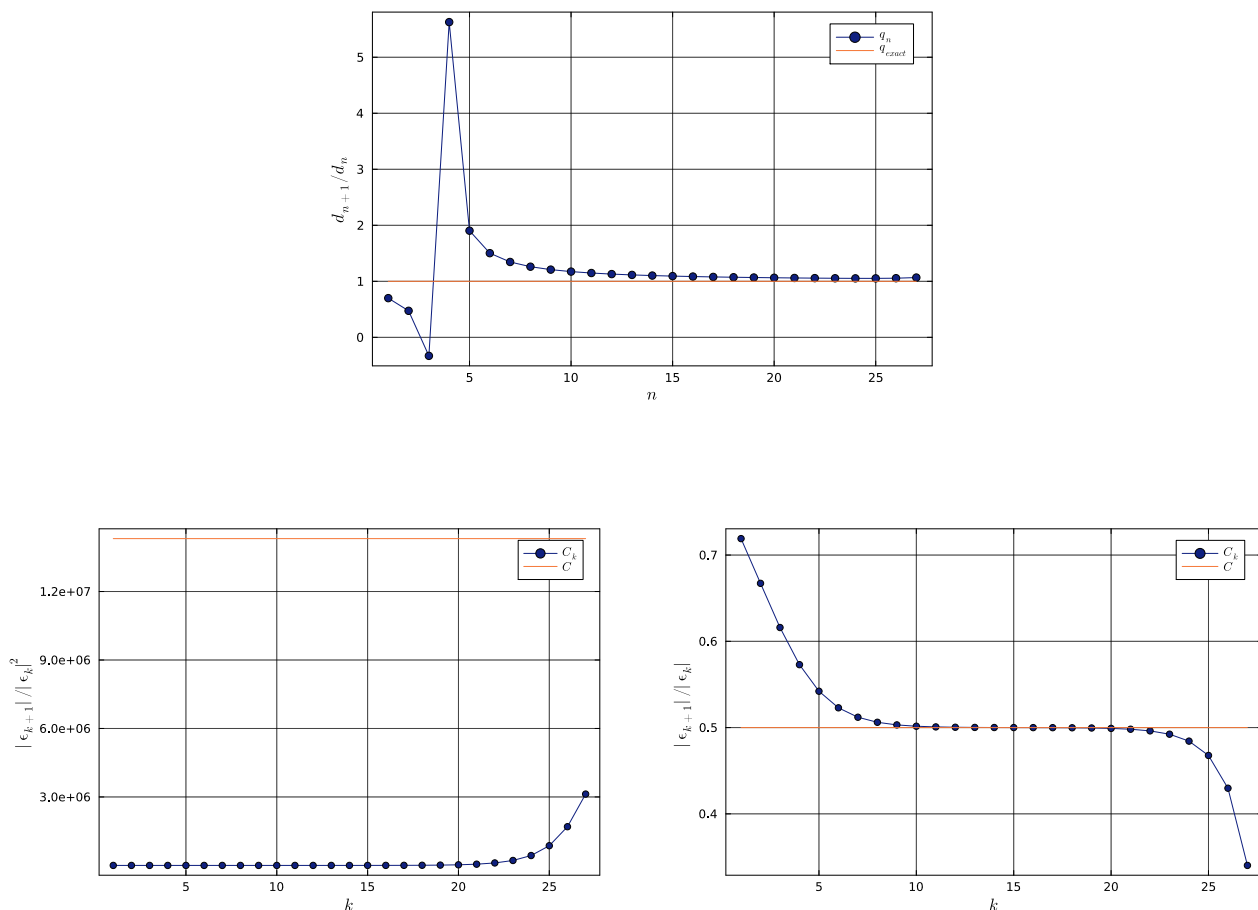


Figure 18: Studio del valore di  $q$  (sopra) e  $C$  (sotto) per  $f_3$ .

grafico della stessa figura si osserva che i valori successivi di  $C_n$  si stabilizzano attorno al valore  $C = 0.5$ , per poi calare nuovamente a causa di errori di cancellazione numerica.

Scegliendo  $C = C_{15}$  come stima del plateau si ottengono i seguenti valori:

$$C = 0.500, \quad m = 2.000, \quad (45)$$

in accordo con quanto atteso.

### 3.3 Esercizio 3.3.3

Traccia la funzione  $f(x) = x^{-2} - \sin x$  nell'intervallo  $x \in [0.5, 10]$ . Per ciascun valore iniziale  $x_1 = 1, x_1 = 2, \dots, x_1 = 7$ , applica il metodo di Newton a  $f$  e realizza una tabella che mostri  $x_1$  e la radice trovata dal metodo. In quale caso l'iterazione converge a una radice diversa da quella più vicina al valore iniziale? Usa il grafico per spiegare perché ciò è accaduto.

#### 3.3.1 Soluzione

Si riporta in figura 19 il grafico della funzione  $f(x) = x^{-2} - \sin x$  nell'intervallo  $[0.5, 10]$ , con i punti di partenza del metodo di Newton.

Applicando il metodo di Newton con i valori iniziali, non sempre si ottengono le radici più vicine al valore di partenza. Si riportano i risultati in tabella 6.

Le due colonne evidenziate in arancione sono i punti iniziali che non convergono alla radice più vicina. Osservando il grafico della funzione è facile osservare che i punti iniziali  $x_2$  e  $x_5$  sono

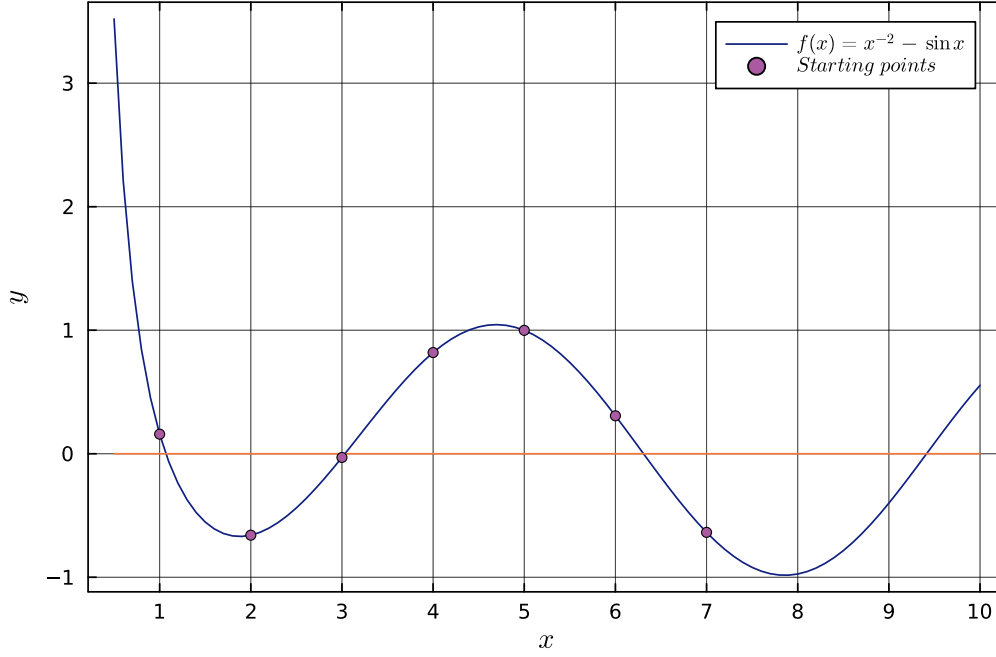


Figure 19: Grafico della funzione  $f(x) = x^{-2} - \sin x$  nell'intervallo  $[0.5, 10]$ .

Table 6: Valori relativi ai punti  $x_1, \dots, x_7$

	$x_1 = 1$	$x_2 = 2$	$x_3 = 3$	$x_4 = 4$	$x_5 = 5$	$x_6 = 6$	$x_7 = 7$
$x_{\text{start}}$	1.0	2.0	3.0	4.0	5.0	6.0	7.0
$x_{\text{root}}$	1.1	6.3	3.0	3.0	9.4	6.3	6.3
$f'(x)$	-2.5	0.2	0.9	0.6	-0.3	-0.9	-0.8

vicini a un minimo e a un massimo locali, rispettivamente. In effetti la derivata prima in questi punti è vicina a zero, come riportato in tabella. Non è esattamente zero, altrimenti il caso sarebbe quello della seconda funzione all'esercizio 3.3.2, ma è comunque piccola: ciò significa che il metodo di Newton converge quadraticamente, ma non necessariamente alla radice più vicina.

### 3.4 Esercizio 3.3.4

Le proprietà più facilmente osservabili dell'orbita di un corpo celeste attorno al Sole sono il periodo  $\tau$  e l'eccentricità ellittica  $\epsilon$ . (Un cerchio ha  $\epsilon = 0$ .) Da questi parametri è possibile determinare, in ogni istante  $t$ , la *vera anomalia*  $\theta(t)$ . Questa è l'angolo tra la direzione del perielio (il punto dell'orbita più vicino al Sole) e la posizione attuale del corpo, visto dal fuoco principale dell'ellisse (dove si trova il Sole). Questo si ottiene tramite

$$\tan \frac{\theta}{2} = \sqrt{\frac{1+\epsilon}{1-\epsilon}} \tan \frac{\psi}{2} \quad (46)$$

dove l'*anomalia eccentrica*  $\psi(t)$  soddisfa l'equazione di Keplero:

$$\psi - \epsilon \sin \psi - \frac{2\pi t}{\tau} = 0. \quad (47)$$

L'equazione 47 deve essere risolta numericamente per trovare  $\psi(t)$ , e poi la 46 può essere risolta analiticamente per ottenere  $\theta(t)$ . L'asteroide Eros ha  $\tau = 1.7610$  anni e  $\epsilon = 0.2230$ . Utilizzando

il metodo di Newton per la 47, realizza un grafico di  $\theta(t)$  per 100 valori di  $t$  compresi tra 0 e  $\tau$ , cioè per un'orbita completa.

(Nota: usa  $\text{mod}(\theta, 2\pi)$  per riportare l'angolo tra 0 e  $2\pi$  se vuoi che il risultato sia una funzione continua.)

### 3.5 Soluzione

Per prima cosa si è risolta l'equazione 47 con il metodo di Newton. Per capire meglio il procedimento si riporta l'equazione 47 nel grafico sinistro di 20, dove si osserva che la funzione ha un solo zero nell'intervallo  $[0, 2\pi]$ . Il metodo di Newton è stato applicato al variare del tempo  $t$ , cioè è stato applicato alla stessa funzione ma traslata in verticale, come si può vedere nella figura.

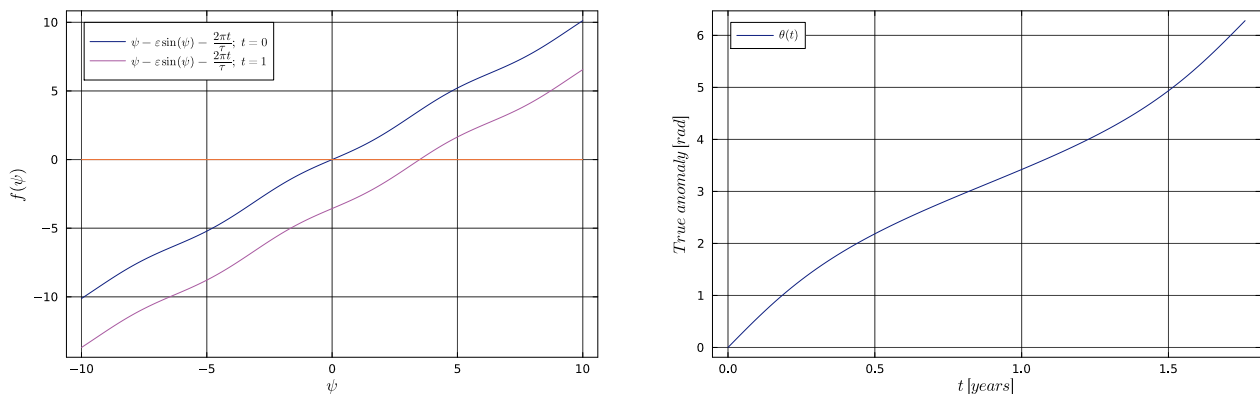


Figure 20: Grafico dell'equazione 47 (sinistra) e  $\theta(t)$  (destra).

Ottenuti i valori di  $\psi$  in corrispondenza dei 100 valori di  $t$  compresi tra 0 e  $\tau$ , si è calcolata la vera anomalia  $\theta$  tramite la relazione 46. Si riporta il grafico di  $\theta(t)$  in figura sinistra di 20.

### 3.6 Esercizi 3.4.1, 3.4.2

1. Scrivi un programma che implementi il metodo della secante.
2. Per ciascuna delle seguenti funzioni, esegui i seguenti passi.
  - (a) Riscrivi l'equazione nella forma standard per la ricerca degli zeri,  $f(x) = 0$ .
  - (b) Traccia il grafico di  $f$  nell'intervallo indicato e determina quante radici sono presenti nell'intervallo.
  - (c) Per ciascuna radice determina un intervallo che la racchiuda. Poi usa il metodo della secante, partendo dagli estremi dell'intervallo, per trovare ciascuna radice.
  - (d) Per una delle radici, usa gli errori nella sequenza della secante per determinare numericamente se la convergenza è apparentemente compresa tra lineare e quadratica.
1.  $x^2 = e^{-x}$ , su  $[-2, 2]$
2.  $2x = \tan x$ , su  $[-0.2, 1.4]$
3.  $e^{x+1} = 2 + x$ , su  $[-2, 2]$

### 3.6.1 Soluzione

(1) Si è implementato il metodo della secante e lo si è testato con la funzione  $x^2 - 1$ , ottenendo come radice  $x = 1.0$  in 10 iterazioni.

(2a)-(2b) Per questi esercizi si rimanda alle sezioni 3.2.1 e 3.2.1.

(2c) L'algoritmo della secante è stato applicato alle tre funzioni, con i seguenti risultati:

- $f_1$ : radice  $x = 0.703$ , in 9 iterazioni. Intervallo iniziale  $[1, 2]$ .
- $f_2$ : radici  $x_1 = 0.000$ ,  $x_2 = 1.166$ , in 6 e 11 iterazioni, rispettivamente. Intervalli iniziali:  $[-0.1, 0.2]$  e  $[0.9, 1.3]$ .
- $f_3$ : radice  $x = -1.000$  in 42 iterazioni. Intervallo iniziale  $[-2, 0]$ .

In generale si può notare che il numero di iterazioni nel metodo della secante è maggiore rispetto a quello del metodo di Newton, nonostante gli intervalli iniziali nel caso della secante siano stati scelti in modo da essere molto vicini alla radice. Ciò è dovuto al fatto che il metodo della secante converge con ordine del rapporto aureo, mentre il metodo di Newton converge con ordine quadratico per radici semplici.

(2d) Si è scelto di studiare l'errore della secante per la radice di  $f_1$ . Si riporta in figura 21 il grafico della successione  $d_k = |x_k - x_{k-1}|$ . Come stima di  $q$  si è scelto di considerare l'ultimo punto della successione, e si ottiene  $q = 1.692$ , sufficientemente in accordo con  $q_{atteso} = 1.618$ .

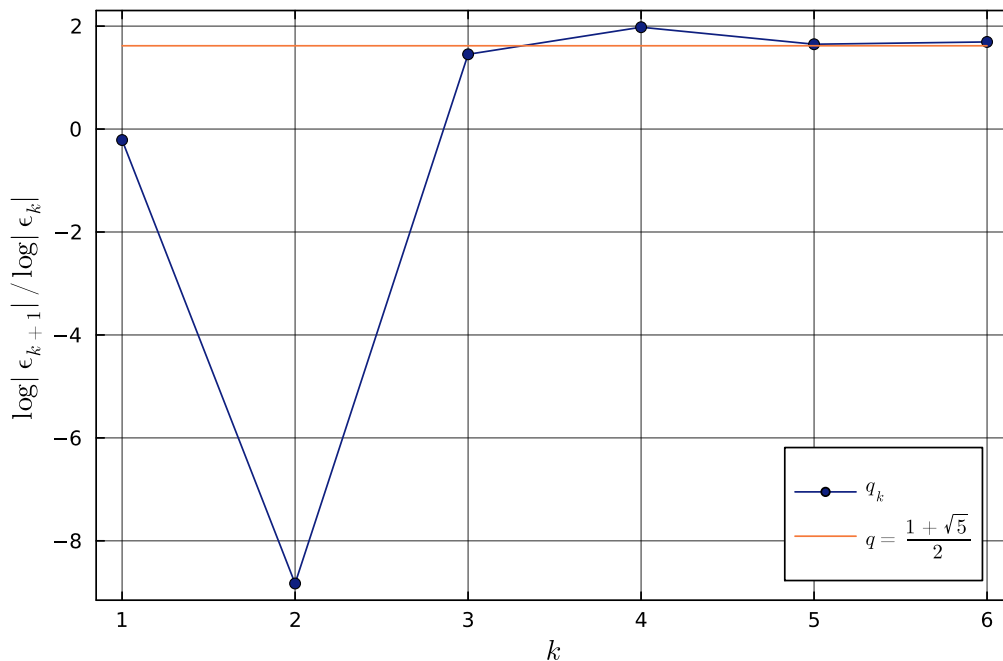


Figure 21: Studio del valore di  $q$  per la radice di  $f_1$ .

### 3.7 Esercizio 3.4.3

Utilizza un grafico per localizzare approssimativamente tutte le radici di  $f(x) = x^{-2} - \sin(x)$  nell'intervallo  $[0.5, 10]$ . Poi trova una coppia di punti iniziali per ciascuna radice tale che il metodo della secante converga a quella radice.

### 3.7.1 Soluzione

Si riporta in figura 22 il grafico della funzione  $f(x) = x^{-2} - \sin(x)$  nell'intervallo  $[0.5, 10]$ , da cui si evince che la funzione ha 4 radici nell'intervallo. In figura sono riportate anche le radici stimate.

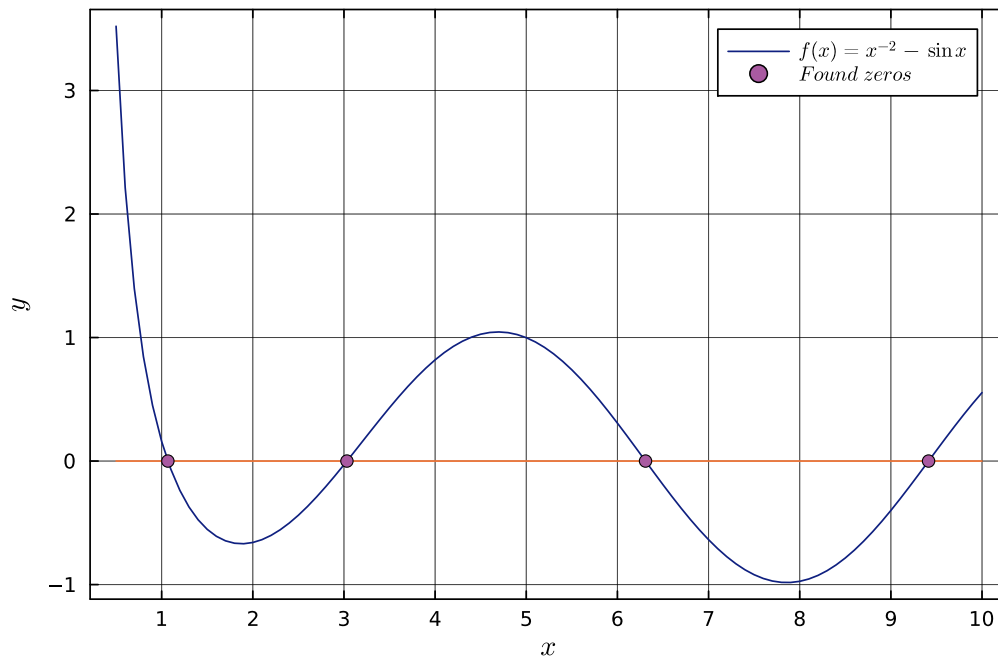


Figure 22: Grafico della funzione  $f(x) = x^{-2} - \sin(x)$  nell'intervallo  $[0.5, 10]$ .

Si riportano in tabella 7 le radici stimate, i punti iniziali scelti e il numero di iterazioni necessarie per la convergenza. Possiamo osservare che le radici trovate sono in accordo con

Radice	Radice stimata	Punti iniziali	Numero di iterazioni
$r_1$	1.068	$[1.0, 2.0]$	11
$r_2$	3.033	$[2.0, 4.0]$	8
$r_3$	6.308	$[6.0, 7.0]$	7
$r_4$	9.413	$[9.0, 10.0]$	7

Table 7: Radici stimate e punti iniziali per il metodo della secante.

quelle stimate nell'esercizio 3.3 con metodo di Newton.

## 4 Interpolazioni

### 4.1 Esercizi 4.2.1, 4.2.2

1. Scrivi un programma che implementi la formula baricentrica dell'interpolazione di Lagrange per nodi in posizioni generiche.
2. In ciascun caso, interpola la funzione data usando  $n$  nodi equispaziati nell'intervallo indicato. Rappresenta graficamente ciascuna funzione interpolante insieme alla funzione esatta.
  - (a)  $f(x) = \ln(x)$ ,  $n = 2, 3, 4$ ,  $x \in [1, 10]$
  - (b)  $f(x) = \tanh(x)$ ,  $n = 2, 3, 4$ ,  $x \in [-3, 2]$
  - (c)  $f(x) = \cosh(x)$ ,  $n = 2, 3, 4$ ,  $x \in [-1, 3]$

(d)  $f(x) = |x|$ ,  $n = 3, 5, 7$ ,  $x \in [-2, 1]$

#### 4.1.1 Soluzione

(1) Si è implementata la formula baricentrica dell'interpolazione di Lagrange per nodi in posizioni generiche, e la si è testata nell'esercizio 4.2.2.

(2) In questo esercizio, nonostante sia data la possibilità di utilizzare la formula baricentrica semplificata dai nodi equispaziati, si è scelto di utilizzare la formula baricentrica per nodi generici, come test dell'implementazione dell'esercizio precedente. In esercizi successivi sarà richiesto di implementare la formula baricentrica per nodi di Chebyshev, e in tale contesto si è sviluppata una funzione Julia che chiede la natura dei nodi e restituisce la funzione baricentrica corrispondente, anche per il caso equispaziato.

Si riportano i grafici delle funzioni interpolanti in figura 23.

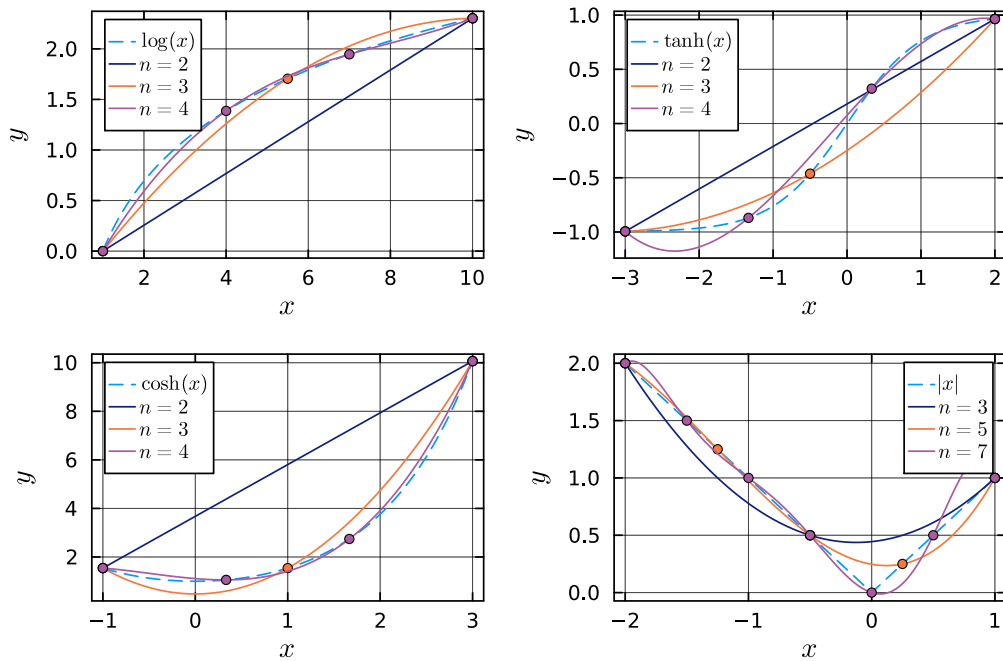


Figure 23: Grafici delle funzioni interpolanti per i vari casi.

Come si può notare, all'aumentare del numero di nodi l'interpolazione migliora. Il grafico della funzione  $|x|$  però richiede un numero di nodi maggiore per essere interpolato con una certa accuratezza, ciò è dovuto alla non derivabilità della funzione in  $x = 0$ . Il problema è che si sta cercando di interpolare una funzione non derivabile con un polinomio, che notoriamente è derivabile in ogni punto.

## 4.2 Esercizi 4.4.1, 4.4.2

1. Scrivi un programma che implementi la formula baricentrica dell'interpolazione di Lagrange per il caso particolare dei nodi di Chebyshev, utilizzando i risultati analitici per i pesi. Verifica la correttezza della tua implementazione confrontandola con la routine per nodi generici.

2. (a) Per ciascun caso sotto riportato, calcola il polinomio interpolante usando  $n$  nodi di Chebyshev di seconda specie in  $[-1, 1]$  per  $n = 4, 8, 12, \dots, 60$ . Per ogni valore di  $n$ , calcola

l'errore in norma  $\infty$  (ossia,  $\|f - p\|_\infty = \max_{x \in [-1,1]} |p(x) - f(x)|$  valutato su almeno 4000 valori di  $x$ ). Utilizzando una scala log-lineare, rappresenta graficamente l'errore in funzione di  $n$  e determina una buona approssimazione della costante  $K$ .

- $f(x) = 1/(25x^2 + 1)$
- $f(x) = \tanh(5x + 2)$
- $f(x) = \cosh(\sin x)$
- $f(x) = \sin(\cosh x)$

(b) Realizza un grafico analogo utilizzando  $n$  punti equidistanti per l'interpolazione.

#### 4.2.1 Soluzione

(1) Si è implementata la formula baricentrica dell'interpolazione di Lagrange per il caso particolare dei nodi di Chebyshev, e la si è testata nell'esercizio 4.4.2, confrontandola con il caso di nodi equidistanti.

(2a) Il calcolo dell'errore in norma  $\infty$  è stato effettuato con un passo di campionamento sulle  $x$  pari a 0.0005, in modo da ottenere 4000 punti tra  $-1$  e  $1$ . Si riportano in figura 24 a sinistra i grafici dell'errore in norma  $\infty$  in funzione di  $n$ , per i casi delle varie funzioni.

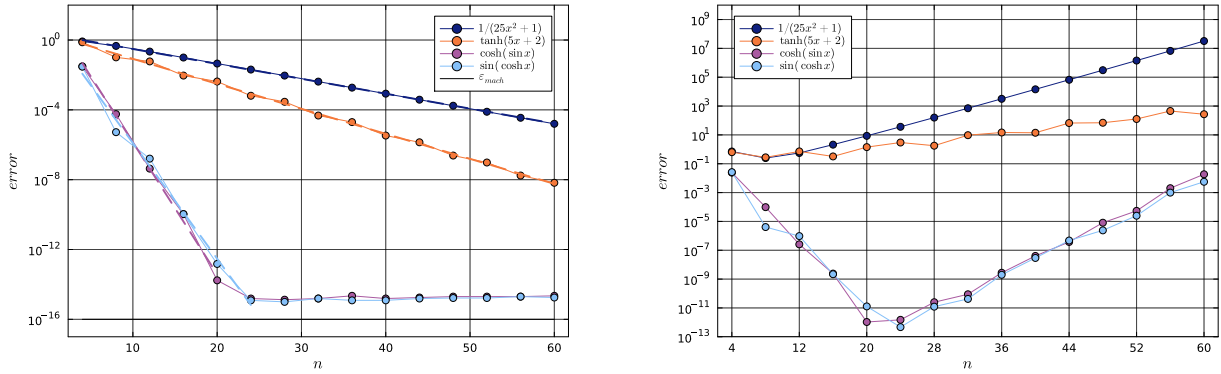


Figure 24: Errori per nodi di Chebyshev (sinistra) e nodi equidistanti (destra).

La prima osservazione che si può fare è che l'errore in norma  $\infty$  decresce all'aumentare di  $n$ , come ci si aspetta data la scelta dei nodi di Chebyshev. La seconda osservazione è che l'errore decresce con un andamento esponenziale, fino ad arrestare la sua discesa intorno al valore di  $\epsilon_{mach} = 10^{-16}$ .

La relazione che ha consentito di stimare la costante  $K$  è la seguente:

$$|p(x) - f(x)| = C \cdot K^{-n}, \quad \log_{10} |p(x) - f(x)| = \log_{10}(C) + n(-\log_{10}(K)), \quad (48)$$

dove si è scelto di linearizzare i dati con il logaritmo per poter stimare la costante  $K$  tramite il coefficiente angolare della retta di regressione. I fit sono stati eseguiti scartando i punti corrispondenti a errori minori di  $10^{-16}$ , e in figura 24 sono rappresentati dalle linee tratteggiate. Si riportano i valori di  $K$  ottenuti per le varie funzioni:

- $f_1(x) = 1/(25x^2 + 1)$ :  $K = 1.217$ .
- $f_2(x) = \tanh(5x + 2)$ :  $K = 1.391$ .



- $f_3(x) = \cosh(\sin x)$ :  $K = 5.704$ .
- $f_4(x) = \sin(\cosh x)$ :  $K = 4.598$ .

(2b) Si è ripetuto lo stesso procedimento con i nodi equidistanti, e si riportano in figura 24 a destra i grafici dell'errore in norma  $\infty$  in funzione di  $n$ , per i vari casi. Come atteso, l'errore in norma  $\infty$  non decresce globalmente all'aumentare di  $n$ . Per le prime due funzioni l'errore è crescente fin dall'inizio, mentre per le ultime due funzioni l'errore decresce fino a un certo punto, per tornare a crescere successivamente. È un comportamento noto come *fenomeno di Runge*, ed è dovuto alla scelta dei nodi equidistanti.

### 4.3 Esercizio 4.4.3

I punti di Chebyshev possono essere utilizzati anche quando l'intervallo di interpolazione è  $[a, b]$  invece di  $[-1, 1]$ , tramite un opportuno cambiamento di variabile. Traccia il polinomio interpolante di  $f(z) = \cosh(\sin z)$  su  $[0, 2\pi]$  utilizzando  $n = 40$  nodi di Chebyshev.

#### 4.3.1 Soluzione

Il cambio di variabile utilizzato per passare da  $[-1, 1]$  a  $[a, b]$  è il seguente:

$$z = \psi(x) = a + (b - a)\frac{x + 1}{2}, \quad (49)$$

dove  $x$  è il nodo di Chebyshev.

L'algoritmo implementato tiene conto di questo cambio di variabile. In particolare si sono calcolati i nodi di Chebyshev, li si è trasformati con 49 e poi si è calcolato il polinomio nei nuovi punti con i soliti pesi. Si riporta in figura 25 il grafico ottenuto con  $n = 40$  nodi di Chebyshev, per la funzione  $f(z) = \cosh(\sin z)$  su  $[0, 2\pi]$ .

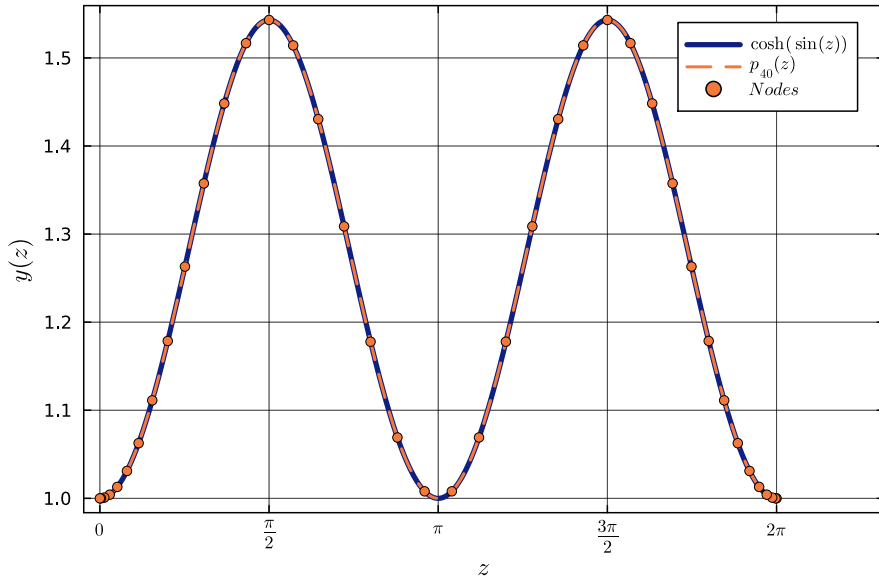


Figure 25:  $f(z) = \cosh(\sin z)$  su  $[0, 2\pi]$  con  $n = 40$  nodi di Chebyshev.

## 4.4 Esercizio 4.4.4

Siano  $x_1, \dots, x_n$  i punti di Chebyshev standard. Questi vengono mappati nella variabile  $z$  come  $z_i = \phi(x_i)$  per ogni  $i$ , dove  $\phi$  è una trasformazione sulla retta reale. Supponiamo che  $f(z)$  sia una funzione data, il cui dominio è l'intera retta reale. Allora i valori della funzione  $y_i = f(z_i)$  possono essere associati ai nodi di Chebyshev  $x_i$ , portando a un polinomio interpolante  $p(x)$ . Questo implica a sua volta una funzione interpolante sulla retta reale, definita come

$$q(z) = p(\phi^{-1}(z)) = p(x). \quad (50)$$

Implementa questa idea per tracciare un interpolante di  $f(z) = (z^2 - 2z + 2)^{-1}$  usando  $n = 30$ . Il tuo grafico deve mostrare  $q(z)$  valutato in 1000 punti equispaziati in  $[-6, 6]$ , con marcatori sui valori nodali (quelli che ricadono nella finestra  $[-6, 6]$ ).

Suggerimento: se preferisci evitare di gestire potenziali infiniti considera l'uso dei nodi di Chebyshev di prima specie.

### 4.4.1 Soluzione

Il cambio di variabile utilizzato è il seguente:

$$z = \phi(x) = \frac{2x}{1 - x^2}, \quad (51)$$

La funzione Julia implementata tiene conto di questo cambio di variabile. In particolare calcola i nodi di Chebyshev del secondo tipo nell'intervallo  $[-1, 1]$ , che vengono denotati come  $x_i$ , e li trasformati con 51 per ottenere i nodi di Chebyshev  $z_i$ , ora distribuiti sull'intera retta reale. I punti per i quali passa il polinomio vengono indicati con  $y_i = f(z_i)$ .

La funzione restituisce il polinomio interpolante  $q(z)$ . Per calcolarlo si utilizza la trasformazione inversa di 51 per portare i punti  $z$  in cui verrà calcolato il polinomio  $q$  nell'intervallo  $[-1, 1]$ . A questo punto avviene l'interpolazione con i nodi di Chebyshev  $x_i$  e i valori  $y_i$ , e così il polinomio restituito interpola su tutto l'asse reale avendo calcolato l'interpolazione su un intervallo finito. Si riporta in figura 26 il grafico della funzione  $f(z) = (z^2 - 2z + 2)^{-1}$  e della sua interpolazione con  $n = 30$  nodi di Chebyshev. Per motivi di rappresentazione grafica l'intervallo è stato limitato a  $[-6, 6]$ , si noti infatti che il numero di nodi in grafico è pari a venti: i dieci restanti sono al di fuori dell'intervallo.

## 4.5 Esercizio 4.6.1

Ognuna delle seguenti funzioni è 2-periodica. Scrivi una funzione che esegua l'interpolazione trigonometrica su  $[-1, 1]$  e traccia la funzione insieme ai suoi interpolanti trigonometrici per  $n = 3, 6, 9$ . Poi, per  $n = 2, 3, \dots, 30$ , calcola l'errore in norma  $\infty$  dell'interpolante trigonometrico campionando in almeno 1000 punti, e realizza un grafico di convergenza in scala semi-logaritmica.

(a)  $f(x) = e^{\sin(2\pi x)}$

(b)  $f(x) = \log[2 + \sin(3\pi x)]$

(c)  $f(x) = \cos^{12}[\pi(x - 0.2)]$

### 4.5.1 Soluzione

Ai fini dell'esercizio si è sviluppata una funzione Julia che calcola l'interpolazione trigonometrica su un intervallo  $[-1, 1]$  per una funzione  $f(x)$ , e restituisce il polinomio interpolante. Dato che le funzioni sono 2-periodiche le si è interpolate correttamente sull'intervallo  $[-1, 1]$ . Per poter

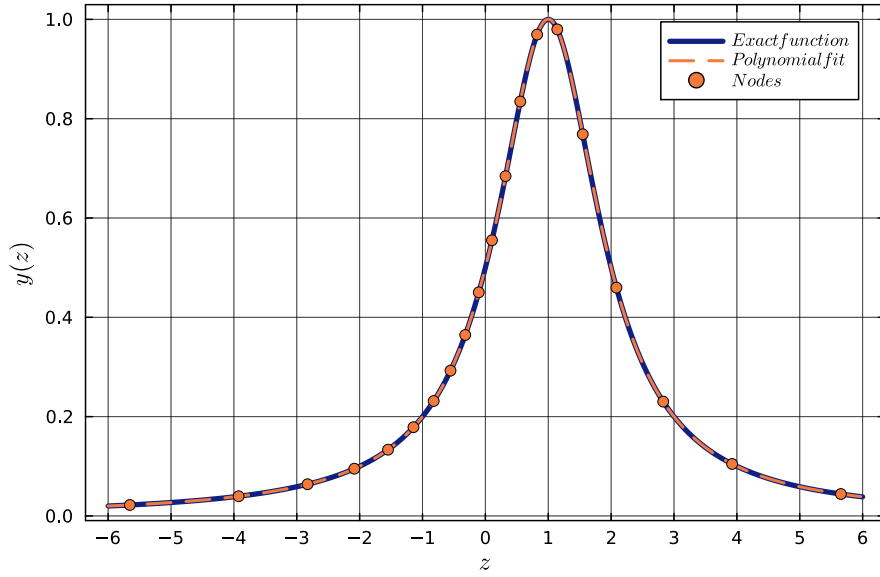


Figure 26:  $f(z) = (z^2 - 2z + 2)^{-1}$  e relativa interpolazione con 30 nodi di Chebyshev.

interpolare funzioni T-periodiche bisogna inserire nell'algoritmo un cambio di variabile, proprio come per l'esercizio 4.4.3.

(a) Si riporta in figura 27 il grafico della funzione  $f(x) = e^{\sin(2\pi x)}$  e dei suoi interpolanti trigonometrici per  $n = 3, 6, 9$ . A fianco si riporta anche il grafico dell'errore in norma  $\infty$  dell'interpolante trigonometrica al variare del numero di nodi, avendo campionato in 1000 punti l'intervallo.

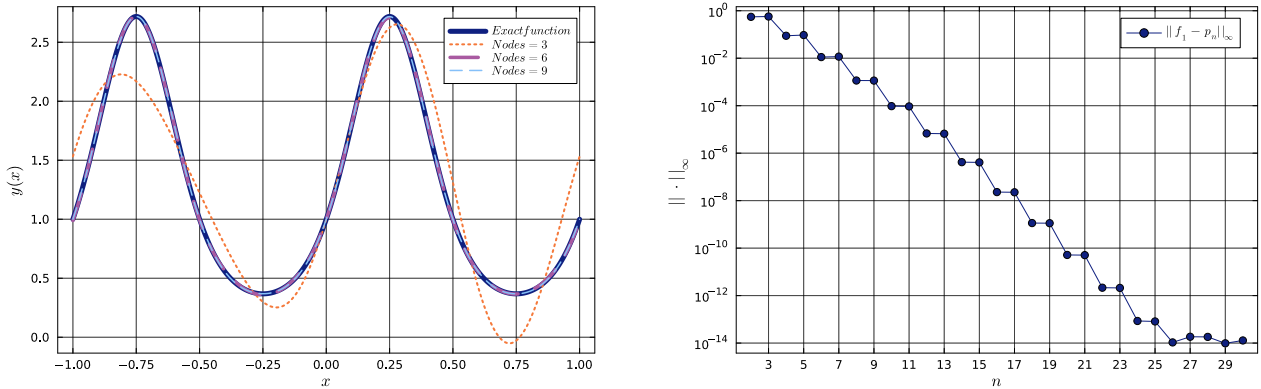


Figure 27: Studio delle interpolazioni trigonometriche per la funzione  $f(x) = e^{\sin(2\pi x)}$ .

(b) Si riporta in figura 28 il grafico della funzione  $f(x) = \log[2 + \sin(3\pi x)]$  e dei suoi interpolanti trigonometrici per  $n = 3, 6, 9$ . A fianco si riporta anche il grafico dell'errore in norma  $\infty$  dell'interpolante trigonometrica al variare del numero di nodi, avendo campionato in 1000 punti l'intervallo.

(c) Si riporta in figura 29 il grafico della funzione  $f(x) = \cos^{12}[\pi(x - 0.2)]$  e dei suoi interpolanti trigonometrici per  $n = 3, 6, 9$ . A fianco si riporta anche il grafico dell'errore in norma  $\infty$  dell'interpolante trigonometrica al variare del numero di nodi, avendo campionato in 1000 punti l'intervallo.

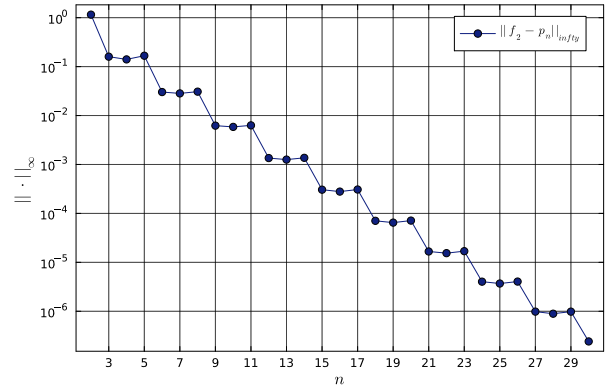
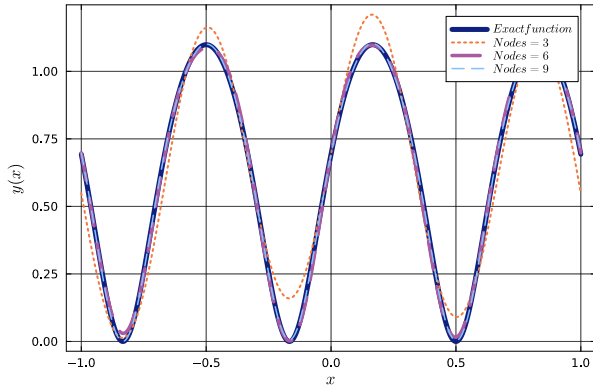


Figure 28: Studio delle interpolazioni trigonometriche per la funzione  $f(x) = \log[2 + \sin(3\pi x)]$ .

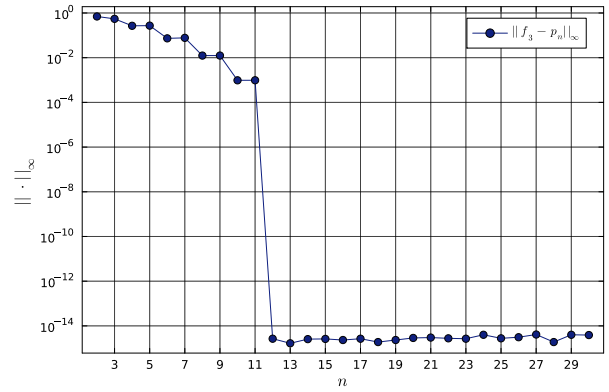
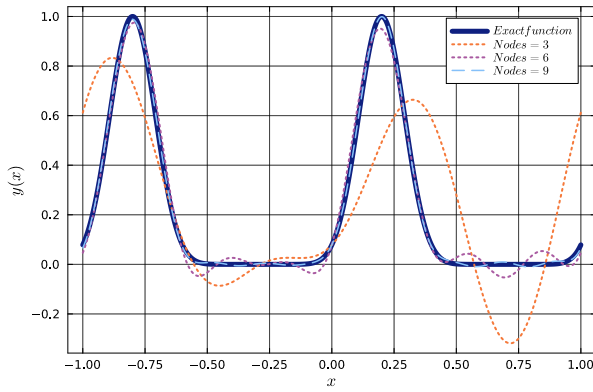


Figure 29: Studio delle interpolazioni trigonometriche per la funzione  $f(x) = \cos^{12}[\pi(x - 0.2)]$ .

**Osservazioni** Si osserva che l'errore in norma  $\infty$  diminuisce all'aumentare del numero di nodi in tutti e tre i casi, confermando la convergenza dell'interpolazione trigonometrica. In generale l'errore non diminuisce mai oltre il valore di  $10^{-14}$  per motivazioni legate alla rappresentazione numerica. È interessante notare che in tutti e tre i casi l'errore ha un andamento a gradino. Questo fenomeno mostra che l'aggiunta di nodi non sempre migliora l'accuratezza dell'interpolazione, anche se globalmente l'errore diminuisce.

Curioso infine è il comportamento dell'errore per la terza funzione, in cui per  $n = 12$  si osserva una diminuzione drastica dell'errore. Sulle ragioni di questo fatto si è avanzata una spiegazione. Osservando il grafico 29 si nota che all'aumento di  $n$  le zone che sono peggio approssimate sono quelle di appiattimento della funzione, e cioè sono quelle che contribuiscono all'errore, calcolato come il massimo della distanza tra la funzione e il polinomio interpolante, su tutto l'intervallo. La figura 30 è un ingrandimento di figura 29 nella regione interessata e si osserva che per  $n = 12$  il polinomio interpolante si appiattisce definitivamente, con la conseguente riduzione dell'errore.

## 5 Integrazione numerica

### 5.1 Esercizi 5.1.1, 5.1.2

1. Scrivi un programma che implementi la regola del trapezio composta per un intervallo generico  $[a, b]$ .
2. Per ciascuno dei seguenti integrali, utilizza la regola del trapezio per stimare il valore dell'integrale con  $m = 10 \cdot 2^k$  nodi, per  $k = 1, 2, \dots, 10$ . Per ogni caso, rappresenta in scala

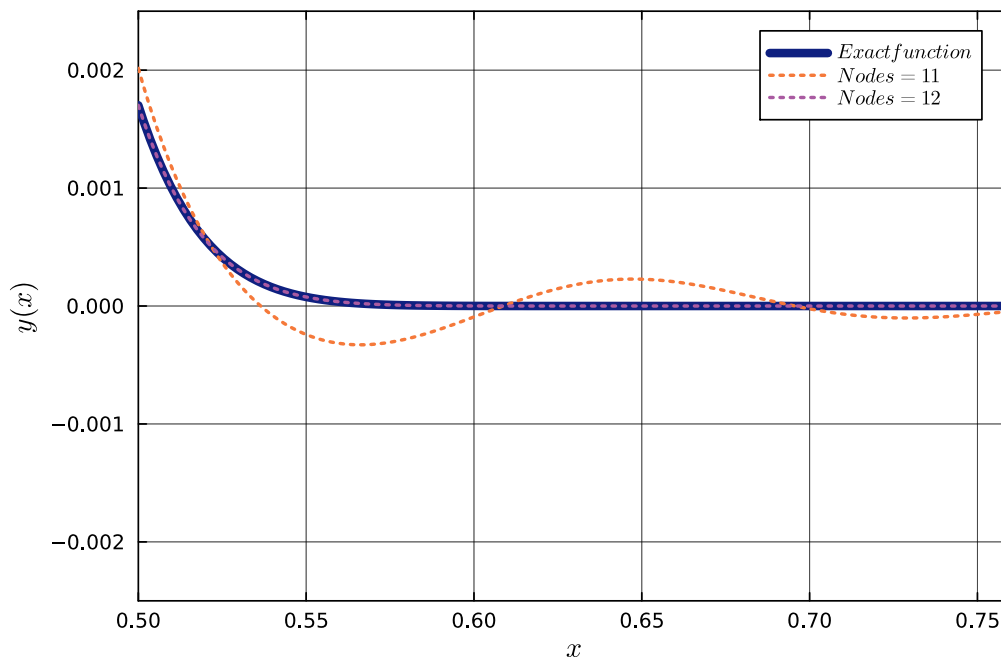


Figure 30: Ingrandimento della figura 29 nella regione di appiattimento della funzione.

log-log l'errore in funzione di  $m$  e verifica se la convergenza è di secondo ordine.

(a)  $\int_0^1 x \log(1+x) dx = \frac{1}{4}$

(b)  $\int_0^1 x^2 \tan^{-1} x dx = \frac{\pi - 2 + 2 \log 2}{12}$

(c)  $\int_0^{\pi/2} e^x \cos x dx = \frac{e^{\pi/2} - 1}{2}$

(d)  $\int_0^1 \frac{\tan^{-1}(\sqrt{2+x^2})}{(1+x^2)\sqrt{2+x^2}} dx = \frac{5\pi^2}{96}$

(e)  $\int_0^1 \sqrt{x} \log(x) dx = -\frac{4}{9}$

*Suggerimento:* Sebbene l'integranda abbia limite zero per  $x \rightarrow 0$ , non può essere valutata direttamente in  $x = 0$ . Una possibilità è iniziare l'integrale da  $x = \epsilon_{\text{mach}}$ . In alternativa, puoi definire una funzione wrapper  $g(x)$  che restituisce  $g(0) = 0$  e  $g(x) = \sqrt{x} \log(x)$  altrimenti.

(f)  $\int_0^1 \sqrt{1-x^2} dx = \frac{\pi}{4}$

### 5.1.1 Soluzione

(1) Si è implementata la regola del trapezio composta per un intervallo generico  $[a, b]$ , e la si è testata nell'esercizio 5.1.2.

(2) Si è calcolato l'integrale con la regola del trapezio per tutti i casi, al variare del numero di nodi. Oltre alla rappresentazione grafica dell'errore, si sono rappresentate le derivate seconde dell'integranda, in modo da poter verificare la convergenza di ordine secondo della regola del trapezio. La motivazione di tale scelta è che l'errore della regola del trapezio scala al secondo ordine con il passo di campionamento, ma ha per coefficiente la derivata seconda dell'integranda, calcolata in un certo  $\xi$  appartenente all'intervallo in considerazione. In formule:

$$R_T(h) = -\frac{(b-a)^3}{12} h^2 f''(\xi)$$

Si riportano nelle figure 31, 32, 33, 34, 35, 36 gli studi dei vari casi.

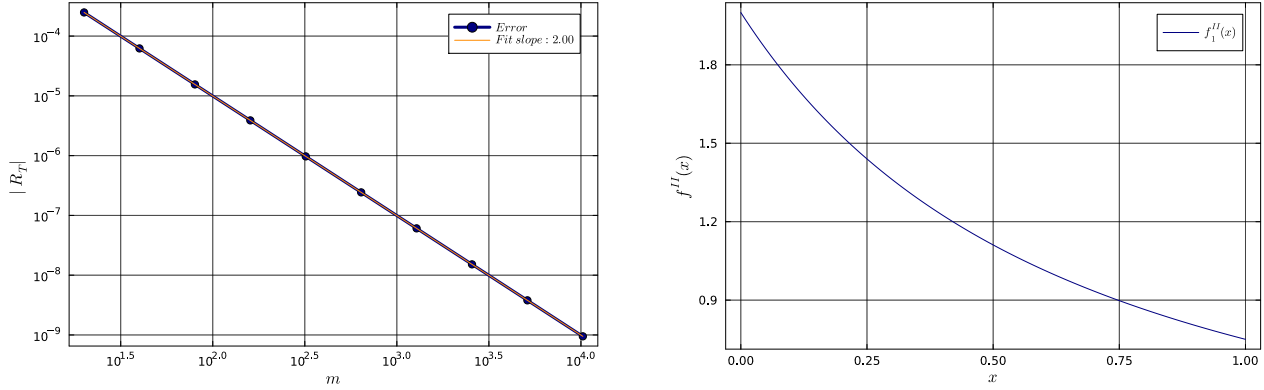


Figure 31: Studio della regola del trapezio per l'integrale  $\int_0^1 x \log(1+x) dx$ .

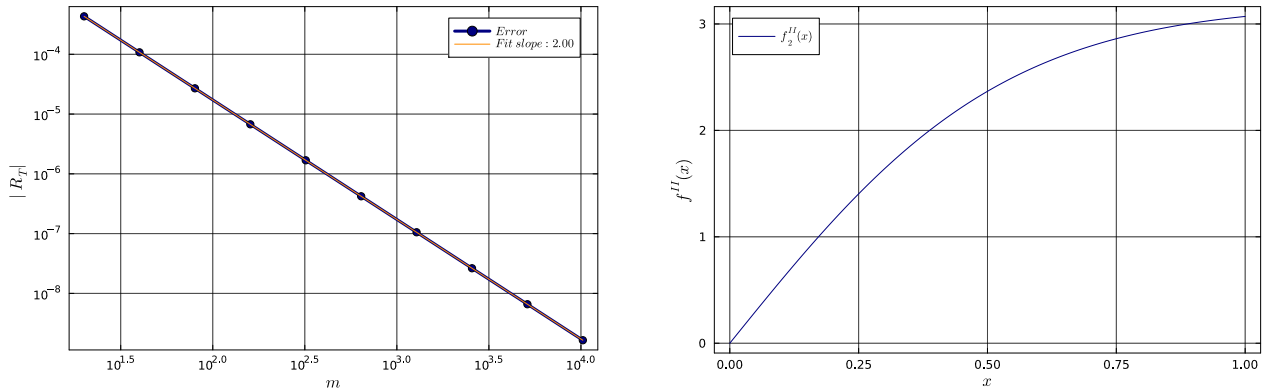


Figure 32: Studio della regola del trapezio per l'integrale  $\int_0^1 x^2 \tan^{-1} x dx$ .

Nei grafici viene riportata la pendenza dell'interpolante dell'errore e nei primi quattro casi l'errore decresce con un andamento quadratico nel passo  $h$ , come atteso. Nel quinto e sesto caso l'errore decresce con un ordine inferiore. Come già accennato, questo è dovuto al fatto che l'integranda ha derivata seconda non ben definita nel dominio di integrazione.

**Commento:** Qui e nei prossimi esercizi l'interpolazione è stata eseguita per l'errore  $R_T$  (o  $R_S$ ) in funzione del passo  $h$ , e non in funzione del numero di nodi  $m$ , per mostrare l'andamento quadratico (o quartico) dell'errore. I grafici sono stati realizzati in funzione del numero di nodi come richiesto e quindi la retta di regressione è stata calcolata in funzione del numero di nodi  $m$  e non del passo  $h$ . Si ricorda che il passo  $h$  è legato al numero di nodi  $m$  dalle relazioni:

$$h = \frac{b-a}{m}, \quad h = \frac{b-a}{2m}, \quad (52)$$

la prima per la regola del trapezio e la seconda per la regola di Simpson.

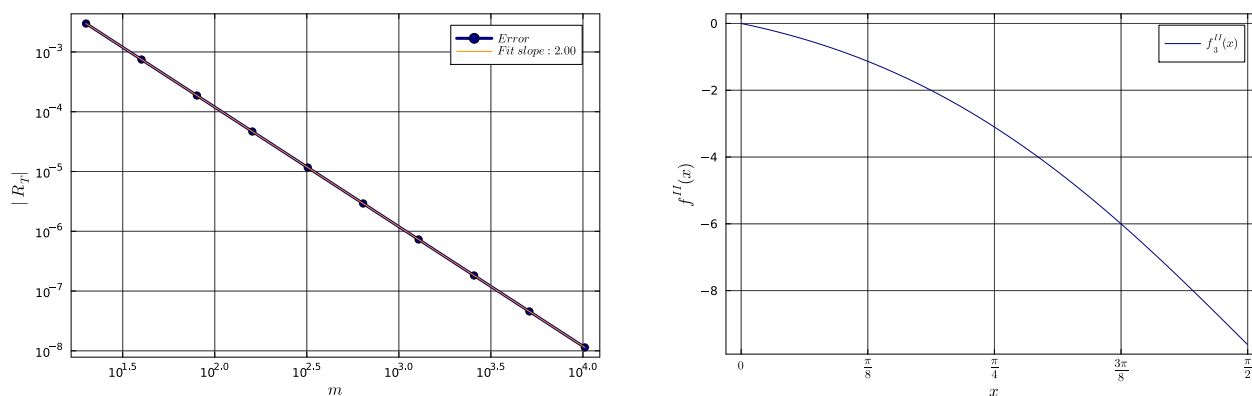


Figure 33: Studio della regola del trapezio per l'integrale  $\int_0^{\pi/2} e^x \cos x \, dx$ .

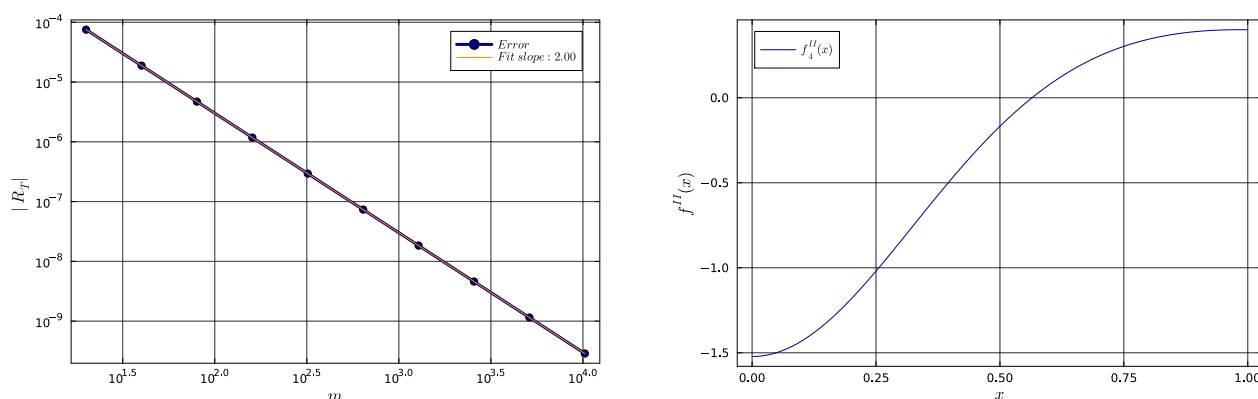


Figure 34: Studio della regola del trapezio per l'integrale  $\int_0^1 \frac{\tan^{-1}(\sqrt{2+x^2})}{(1+x^2)\sqrt{2+x^2}} \, dx$ .

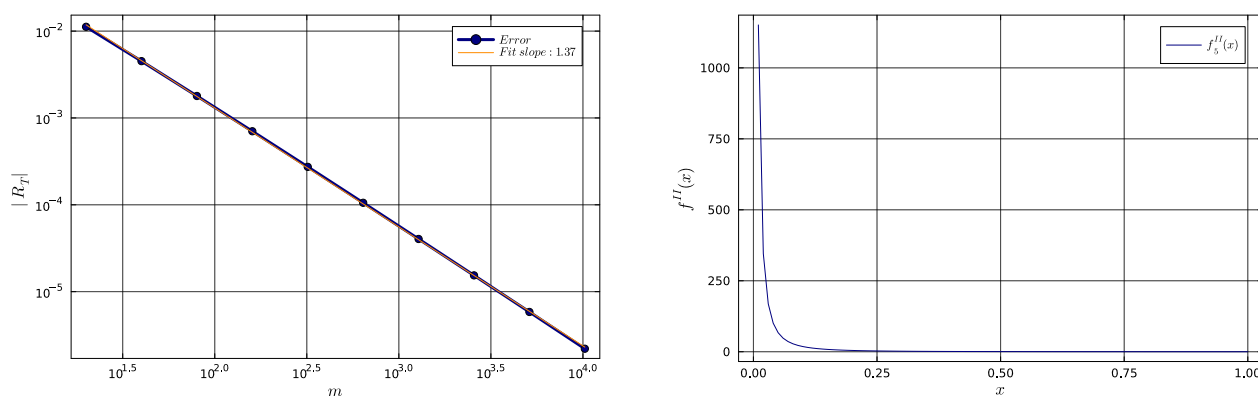


Figure 35: Studio della regola del trapezio per l'integrale  $\int_0^1 \sqrt{x} \log(x) \, dx$ .

## 5.2 Esercizi 5.1.3, 5.1.4

3. Scrivi un programma che implementi la regola di Simpson composta per un intervallo generico  $[a, b]$ .
4. Per ciascun integrale dell'Esercizio 1 sopra, applica la formula di Simpson e confronta gli errori con la convergenza di ordine quattro.

### 5.2.1 Soluzione

(3) Si è implementata la regola di Simpson composta per un intervallo generico  $[a, b]$ , e la si è testata nell'esercizio 5.1.4.

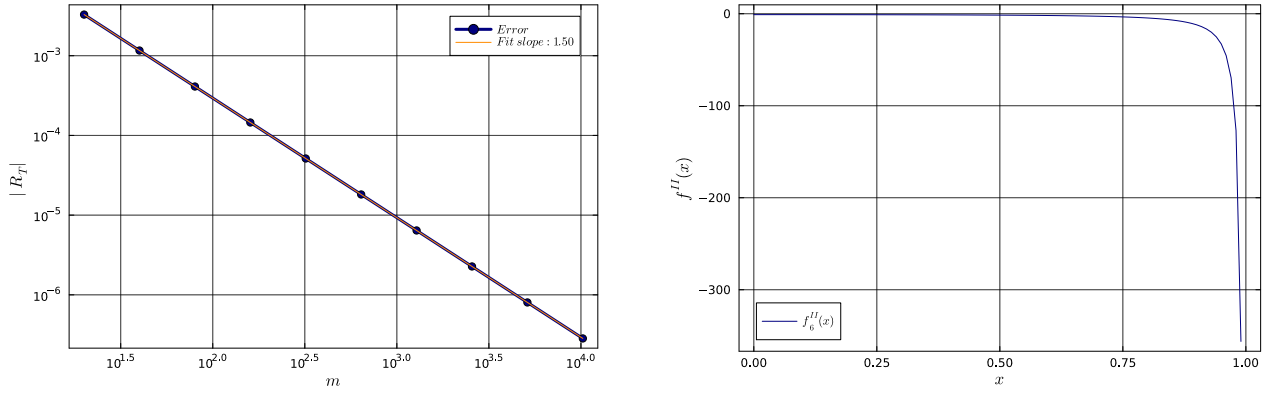


Figure 36: Studio della regola del trapezio per l'integrale  $\int_0^1 \sqrt{1-x^2} dx$ .

(4) Si è calcolato l'integrale con la regola di Simpson per tutti i casi, al variare del numero di nodi. Oltre alla rappresentazione grafica dell'errore, si sono rappresentate le derivate quarte dell'integranda, in modo da poter verificare la convergenza di ordine quarto della regola di Simpson. La motivazione di tale scelta è che l'errore della regola di Simpson scala al quarto ordine con il passo di campionamento, ma ha per coefficiente la derivata quarta dell'integranda, calcolata in un certo  $\xi$  appartenente all'intervallo in considerazione. In formule:

$$R_S(h) = -\frac{(b-a)}{180} h^4 f^{(4)}(\xi)$$

Si riportano nelle figure 37, 38, 39, 40, 41, 42 gli studi dei vari casi.

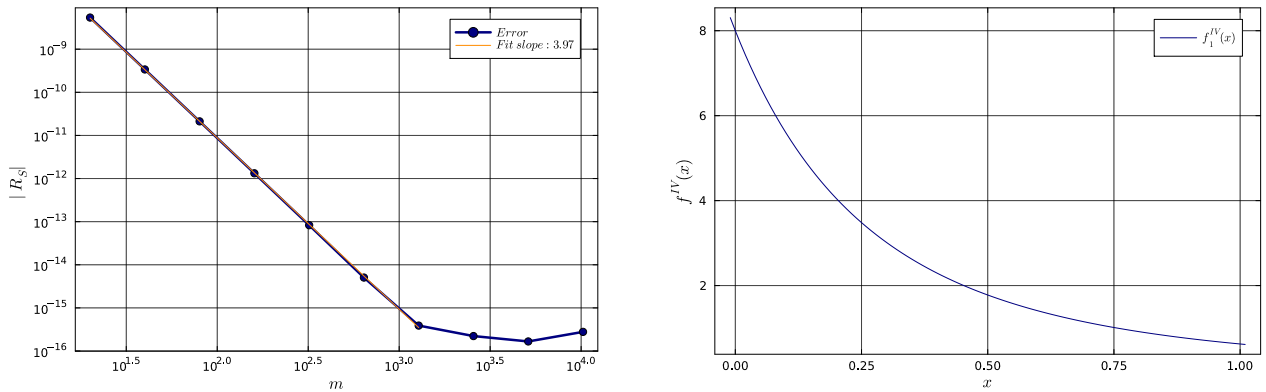


Figure 37: Studio della regola di Simpson per l'integrale  $\int_0^1 x \log(1+x) dx$ .

Si può osservare che l'errore decresce con un andamento quartico nei primi quattro casi, come atteso, ma non negli ultimi due casi, in cui la derivata quarta è non ben definita. In questi casi l'errore decresce con un ordine inferiore. Nei casi in cui l'errore decresce con un ordine quartico, si osserva che l'errore scende rapidamente fino a valore di  $\epsilon_{mach}$ , per poi stabilizzarsi per motivazioni numeriche. Chiaramente l'interpolazione dell'errore è stata eseguita solo sui punti che si trovano al di sopra di questo valore.

### 5.3 Esercizio 5.3.1

1. Scrivi un programma che calcoli i nodi e i pesi per la quadratura di Gauss-Legendre nell'intervallo  $[-1, 1]$ . Di seguito alcuni suggerimenti per l'implementazione.



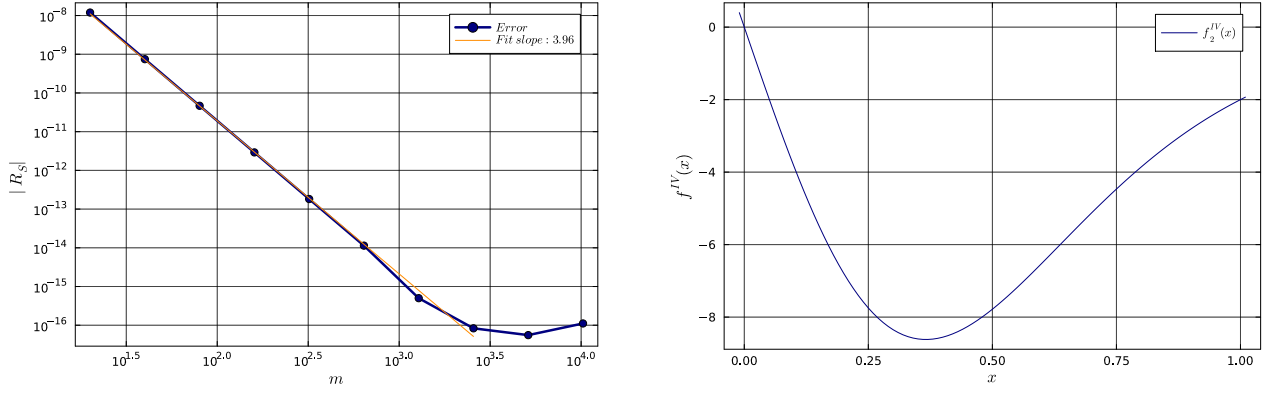


Figure 38: Studio della regola di Simpson per l'integrale  $\int_0^1 x^2 \tan^{-1} x \, dx$ .

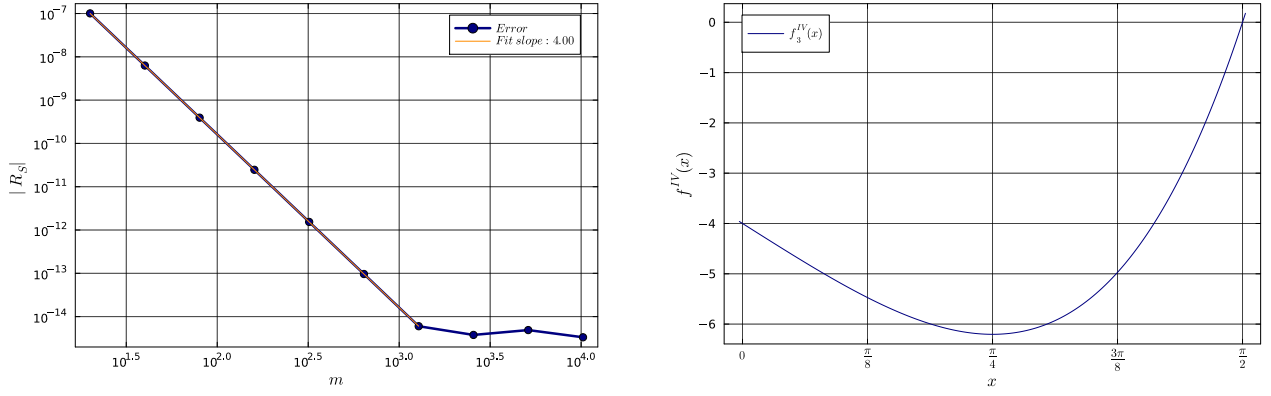


Figure 39: Studio della regola di Simpson per l'integrale  $\int_0^{\pi/2} e^x \cos x \, dx$ .

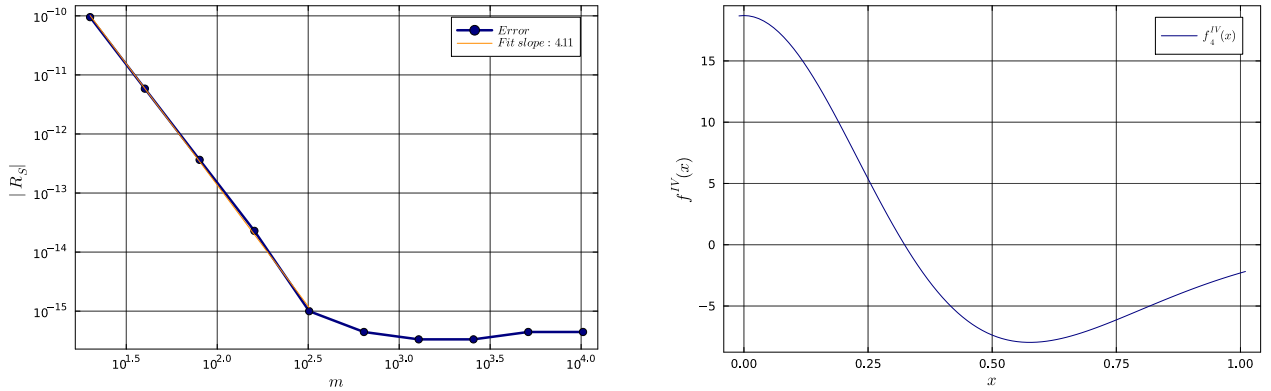


Figure 40: Studio della regola di Simpson per l'integrale  $\int_0^1 \frac{\tan^{-1}(\sqrt{2+x^2})}{(1+x^2)\sqrt{2+x^2}} \, dx$ .

(a) Per trovare le radici  $x_k$ , con  $k = 1, \dots, n$ , del polinomio di Legendre  $P_n(x)$  di ordine  $n$ , usa il metodo di Newton. Come condizione iniziale per la  $k$ -esima radice, utilizza

$$x_k^{(0)} = \cos(\phi_k), \quad \phi_k = \frac{4k-1}{4n+2}\pi. \quad (53)$$

(b) Per applicare il metodo di Newton è necessario valutare  $P_n(x_k^{(i)})$  e la derivata  $P'_n(x_k^{(i)})$ , dove  $x_k^{(i)}$  è la  $i$ -esima iterazione nella ricerca della  $k$ -esima radice  $x_k$ . Puoi ottenerle risolvendo la relazione di ricorrenza

$$(m+1)P_{m+1}(x) = (2m+1)xP_m(x) - mP_{m-1}(x), \quad (54)$$

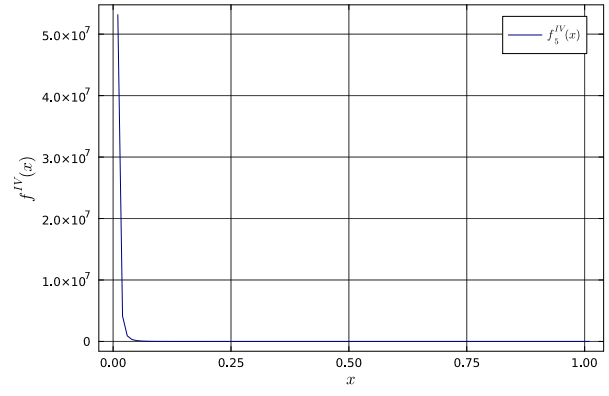
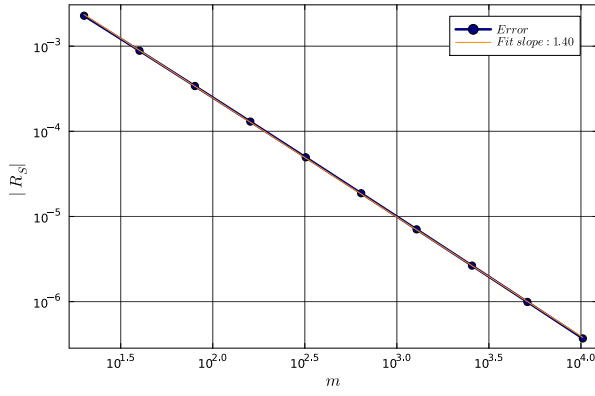


Figure 41: Studio della regola di Simpson per l'integrale  $\int_0^1 \sqrt{x} \log(x) dx$ .

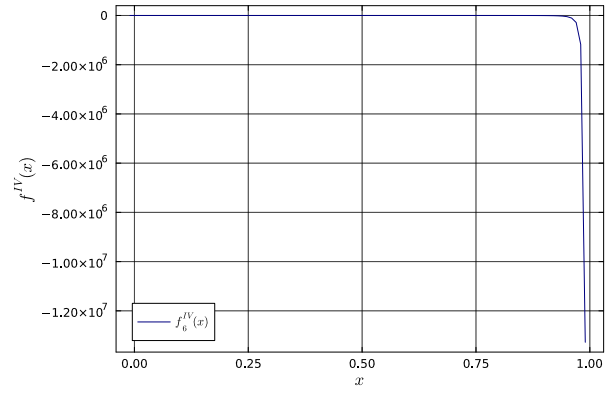
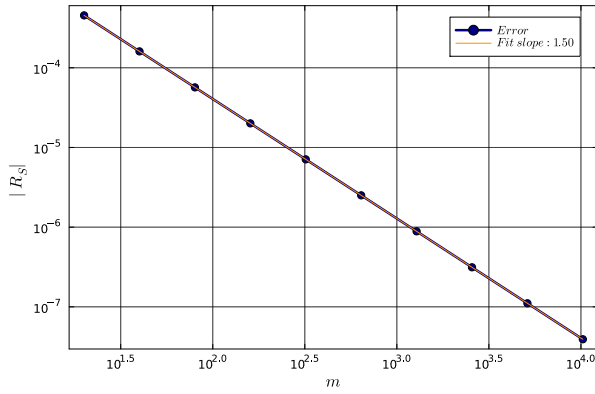


Figure 42: Studio della regola di Simpson per l'integrale  $\int_0^1 \sqrt{1-x^2} dx$ .

per  $m = 1, \dots, n-1$ , dove  $P_0(x) = 1$  e  $P_1(x) = x$ , con  $x = x_k^{(i)}$ . La derivata si calcola poi tramite la relazione

$$(x^2 - 1)P'_n(x) = n[xP_n(x) - P_{n-1}(x)]. \quad (55)$$

(c) Una volta raggiunta la radice  $x_k$  entro una certa tolleranza, il valore della derivata  $P'_n(x_k)$  permette di ottenere il peso corrispondente tramite:

$$w_k = \frac{2}{(1 - x_k^2)[P'_n(x_k)]^2}. \quad (56)$$

(d) Verifica i tuoi risultati confrontandoli con quelli tabulati.

### 5.3.1 Soluzione

Per comodità di implementazione la soluzione riporta i quattro passaggi nel seguente ordine: (b), (a), (c), (d).

**(b)** Si è sviluppato un algoritmo che calcola sia i polinomi di Legendre  $P_n(x)$  che le loro derivate  $P'_n(x)$  utilizzando la relazione di ricorrenza indicata nell'esercizio. Si riporta in figura 43 il grafico dei polinomi e delle loro derivate fino ad ordine  $n = 4$ .

**(a) (b)** Si è implementato il metodo di Newton per trovare le radici dei polinomi di Legendre  $P_n(x)$  e dei pesi associati. Una volta calcolati gli zeri di  $P_n(x)$ , si sono trovati i pesi associati

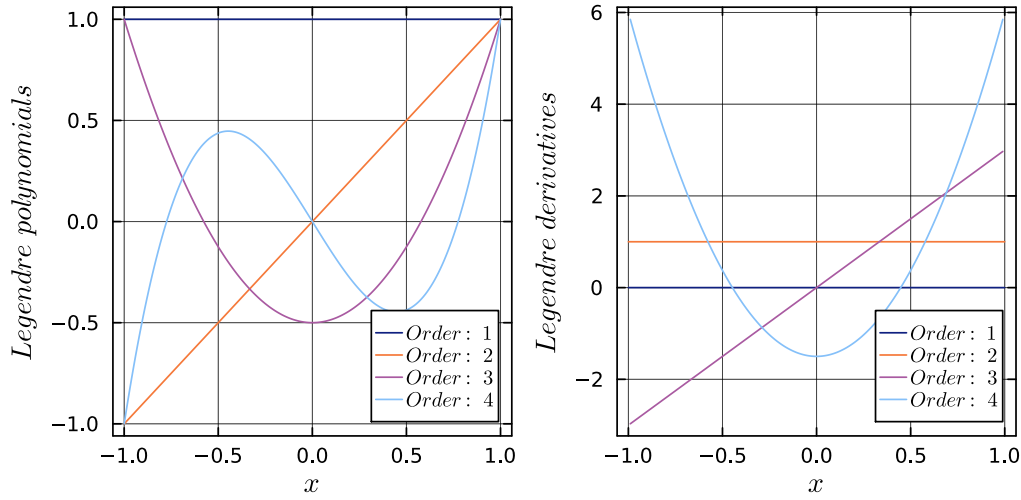


Figure 43: Polinomi di Legendre  $P_n(x)$  e le loro derivate  $P'_n(x)$  fino ad ordine  $n = 4$ .

utilizzando la formula 56. Si riporta in tabella 8 il confronto tra i nodi e pesi calcolati con i corrispettivi valori tabulati, con ordine del polinomio  $n = 4$ .

Table 8: Confronto tra zeri e pesi del polinomio di Legendre di ordine 4

$x_i$	$x_i^{\text{tab}}$	$w_i$	$w_i^{\text{tab}}$
0.86114	0.86114	0.65215	0.65215
0.33998	0.33998	0.34785	0.34785
-0.33998	-0.33998	0.34785	0.34785
-0.86114	-0.86114	0.65215	0.65215

## 5.4 Esercizio 5.4.1

Per ciascun integrale, calcola le approssimazioni utilizzando le regole di quadratura di Gauss-Legendre e Clenshaw-Curtis con  $n = 4, 6, 8, \dots, 40$ . Rappresenta graficamente gli errori  $|I_n - I|$  di entrambi i metodi in funzione di  $n$  su una scala semi-logaritmica. (Suggerimento: considera di implementare la regola di Clenshaw-Curtis solo per  $n$  pari.)

(a)  $\int_{-1}^1 e^{-4x} dx = \frac{1}{2} \sinh(4)$

(b)  $\int_{-1}^1 e^{-9x^2} dx = \frac{\sqrt{\pi}}{3} \operatorname{erf}(3)$

(c)  $\int_{-1}^1 \operatorname{sech}(x) dx = 2 \tan^{-1}[\sinh(1)]$

(d)  $\int_{-1}^1 \frac{1}{1+9x^2} dx = \frac{2}{3} \tan^{-1}(3)$

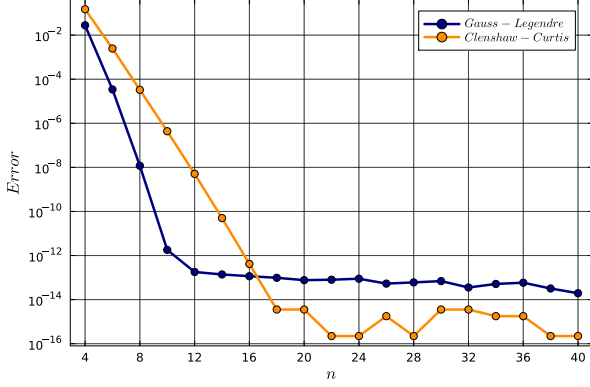
(e)  $\int_{\pi/2}^{\pi} x^2 \sin 8x dx = -\frac{3\pi^2}{32}$

### 5.4.1 Soluzione

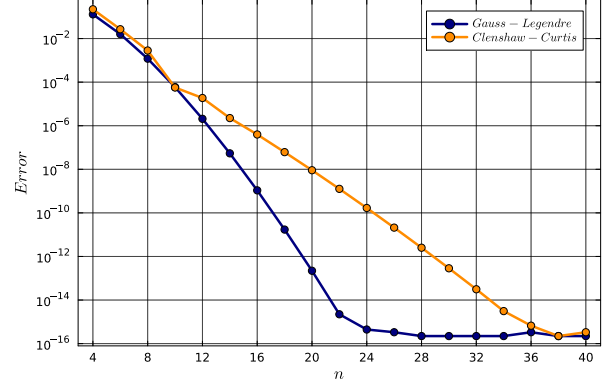
Per ciascun integrale, sono state calcolate le approssimazioni utilizzando le regole di quadratura di Gauss-Legendre e Clenshaw-Curtis con  $n = 4, 6, 8, \dots, 40$ . Gli errori  $|I_n - I|$  di entrambi i metodi sono stati rappresentati graficamente in funzione di  $n$  su una scala semi-logaritmica.

I grafici saranno presentati non nell'ordine in cui sono stati richiesti, ma dapprima gli integrali di funzioni olomorfe, poi quelli di funzioni non olomorfe.

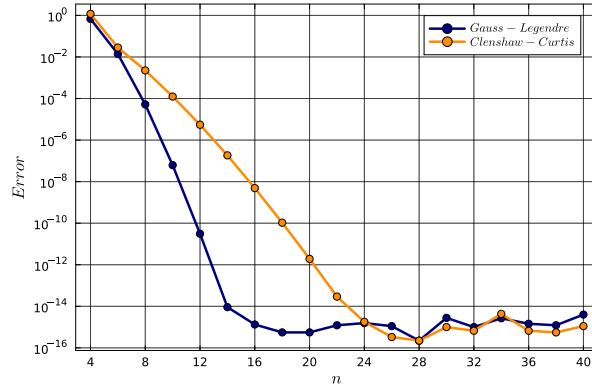
Si riportano in figura 44, gli studi degli errori per gli integrali di funzioni olomorfe.



(a)



(b)



(e)

Figure 44: Studio degli errori per gli integrali di funzioni olomorfe.

I grafici di figura 44 mostrano che gli errori decrescono in maniera molto rapida all'aumentare di  $n$ , raggiungendo l'ordine di grandezza di  $\epsilon_{mach}$  per  $n$  piccoli. Questo fenomeno è legato alla regolarità dell'integranda, che è olomorfa in tutto l'intervallo di integrazione.

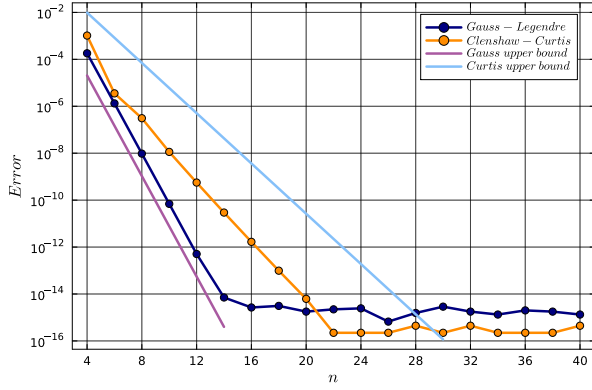
Un'altro comportamento degno di nota è che gli errori del metodo di Clenshaw-Curtis sono più grandi di quelli del metodo di Gauss-Legendre, questo è legato al fatto che il metodo di Clenshaw-Curtis integra esattamente polinomi di grado  $n - 1$  mentre il metodo di Gauss-Legendre integra esattamente polinomi di grado  $2n - 1$ .

Infine si può mostrare matematicamente che inizialmente gli errori del metodo di Clenshaw-Curtis coincidono con gli errori del metodo di Gauss-Legendre.

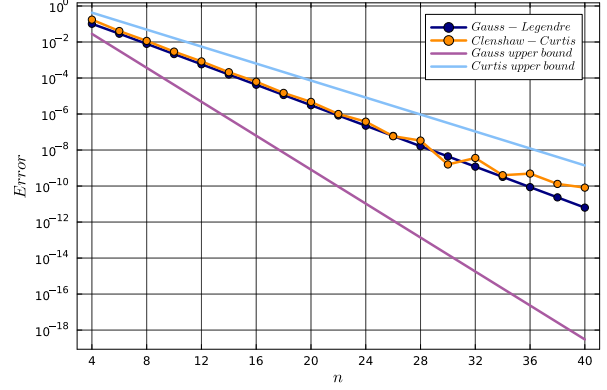
Si riportano in figura 45, gli studi degli errori per gli integrali di funzioni non olomorfe.

Nei casi di figura 45 si è potuto studiare l'errore in maniera più accurata tramite le ellissi di Bernstein. Dalla teoria è noto che la più grande ellissi nel piano complesso che non contiene singolarità è definita da:

$$E_\rho = \left\{ z \in C : z = \frac{1}{2} (\rho e^{i\theta} + \rho^{-1} e^{-i\theta}), \theta \in [0, 2\pi) \right\}. \quad (57)$$



(c)



(d)

Figure 45: Studio degli errori per gli integrali di funzioni non olomorfe.

determina il parametro  $\rho$  delle seguenti maggiorazioni:

$$|I[f] - I_n[f]| \leq \frac{64 M \rho^{1-n}}{15 \rho^2 - 1}, \quad |I[f] - I_n[f]| \leq \frac{64 M \rho^{-2n}}{15 \rho^2 - 1}. \quad (58)$$

dove la prima riguarda Clenshaw-Curtis, la seconda Gauss-Legendre.

Si è perciò calcolato il parametro  $\rho$  in entrambi i casi, sapendo che le singolarità delle integrande estese al piano complesso si trovano in  $\frac{i\pi}{2} + ik\pi$  per  $k \in N$  nel primo caso, in  $\pm i\frac{1}{3}$  nel secondo. Tramite  $\rho$  si sono stimati i limiti superiori degli errori al variare di  $n$ . Nei grafici vengono riportate le rette, ma dato che in 58 è presente anche la maggiorazione con  $M$ , il grafico si concentra sulla pendenza e non sulla traslazione verticale. Si può notare che nel caso (c) le rette ricalcano precisamente la pendenza degli errori. Nel caso (d) l'andamento di Curtis è rispettato, mentre Legendre necessita di più punti per poter entrare nel regime in cui l'errore scala come  $O(n^2)$ . D'altra parte, l'accuratezza massima raggiunta nel caso del metodo (d) è dell'ordine di  $10^{-11}$ , molto più alta rispetto agli altri casi. Infine, per i valori di  $n$  considerati, i due metodi non escono dalla regione in cui gli errori sono gli stessi. Per queste ragioni nel caso (d) la pendenza della retta nel caso di Gauss-Legendre non coincide con quella degli errori.

## 5.5 Esercizio 5.4.2

Per ciascun integrale improprio, calcola le approssimazioni utilizzando la regola di quadratura double-exponential appropriata con  $N = n/2$  e  $n = 4, 6, 8, \dots, 60$ . Rappresenta graficamente gli errori in funzione di  $n$  su una scala semi-logaritmica.

(a)  $\int_{-\infty}^{\infty} \frac{1}{1+x^2+x^4} dx = \frac{\pi}{\sqrt{3}}$

(b)  $\int_{-\infty}^{\infty} e^{-x^2} \cos(x) dx = e^{-1/4} \sqrt{\pi}$

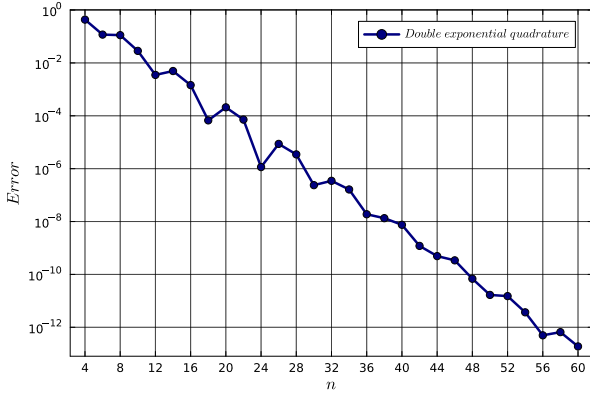
(c)  $\int_{-\infty}^{\infty} (1+x^2)^{-2/3} dx = \frac{\sqrt{\pi} \Gamma(1/6)}{\Gamma(2/3)}$

(d)  $\int_0^{\infty} \frac{1}{1+x^2} dx = \frac{\pi}{2}$

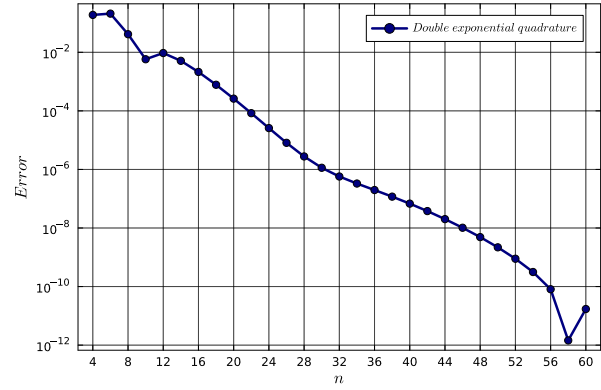
(e)  $\int_0^{\infty} \frac{e^{-x}}{\sqrt{x}} dx = \sqrt{\pi}$

### 5.5.1 Soluzione

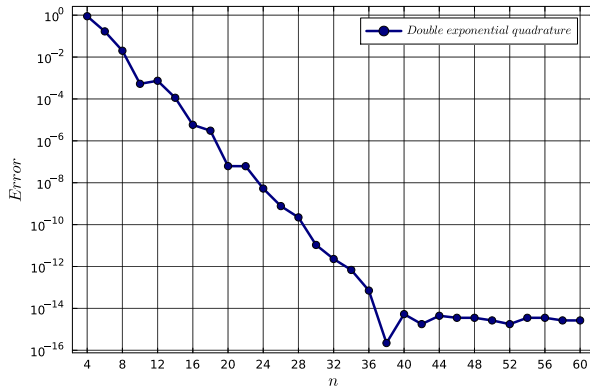
Si riportano nelle figure 46 e 47 gli studi degli errori per gli integrali impropri, con la regola di quadratura double-exponential. In figura 45 possiamo notare che gli integrali convergono



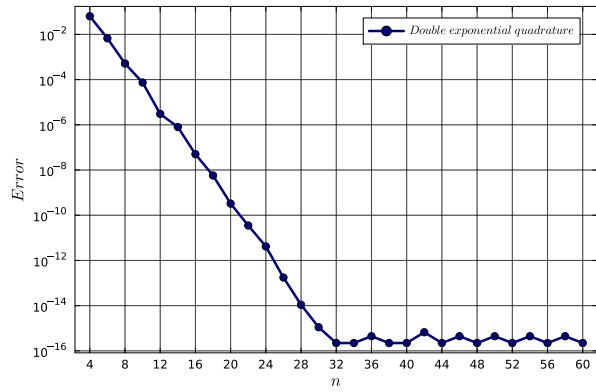
(a)



(b)



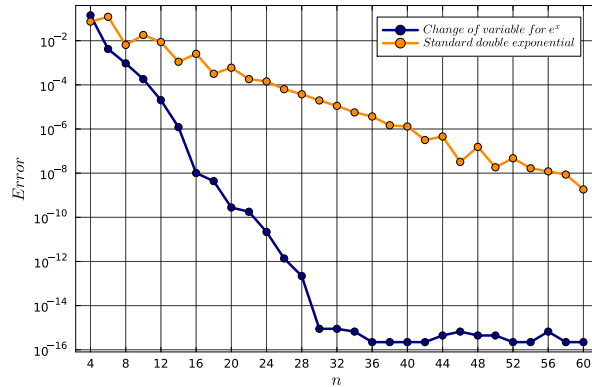
(c)



(d)

Figure 46: Studio degli errori per gli integrali impropri con la regola di quadratura double-exponential.

nonostante la presenza di singolarità nelle integrande, questo è uno dei vantaggi della regola di quadratura double-exponential.



(e)

Figure 47: Studio degli errori per l'integrale improprio con la regola di quadratura double-exponential.

Nella figura 47 si sono rappresentati gli errori dell'integrazione secondo la regola di quadratura double-exponential, ma in un caso con l'utilizzo del cambio di variabile apposito per funzioni con  $e^{-x}$ , nell'altro senza. Si può notare che nel primo caso l'errore ha una convergenza molto più rapida, raggiungendo l'ordine di  $\epsilon_{mach}$ .

## 5.6 Esercizio 5.4.3

Per ciascun integrale, calcola le approssimazioni utilizzando la regola di quadratura di Gauss-Legendre con  $n = 4, 6, 8, \dots, 60$  e la regola di quadratura double-exponential appropriata con  $N = n/2$ . Rappresenta graficamente gli errori in funzione di  $n$  su una scala semi-logaritmica.

$$(a) \int_0^1 \sqrt{x} \log(x) dx = -\frac{4}{9}$$

$$(b) \int_0^1 \sqrt{1-x^2} dx = \frac{\pi}{4}$$

$$(c) \int_0^1 (\log x)^2 dx = 2$$

$$(d) \int_0^{\pi/2} \log(\cos(x)) dx = -\frac{\pi}{2} \log(2)$$

$$(e) \int_0^{\pi/2} \sqrt{\tan(x)} dx = \frac{\pi}{\sqrt{2}}$$

### 5.6.1 Soluzione

Si riportano in figura 48 gli studi degli errori per gli integrali con la regola di quadratura di Gauss-Legendre, insieme agli errori della regola di quadratura double-exponential.

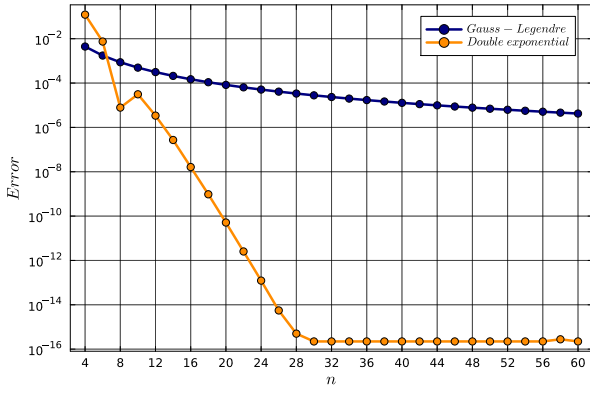
RIVEDI QUESTE CONSIDERAZIONI:

Dai grafici risulta evidente che gli errori della regola di quadratura di Gauss-Legendre decrescono molto più rapidamente rispetto a quelli della regola di quadratura double-exponential. Con questo esercizio, unitamente agli esercizi 5.4.1 e 5.4.2, giungiamo alla conclusione che la regola di Gauss-Legendre è l'algoritmo più rapido per il calcolo di integrali di funzioni regolari, almeno per la nostra implementazione. Lo svantaggio è che non riesce a integrare tutte le funzioni, per cui bisogna ricorrere alla quadratura double-exponential nel caso di funzioni con singolarità, nonostante sia molto più lento. In generale però si può considerare l'algoritmo di Clenshaw-Curtis con l'uso della "fast Fourier transform" (FFT), aumentandone l'efficienza.

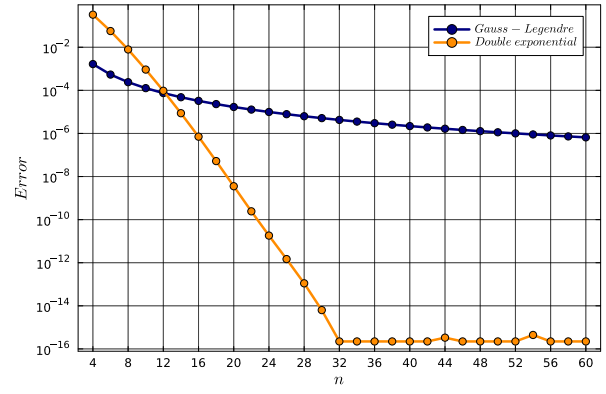
## 6 Equazioni differenziali ordinarie

### 6.1 Esercizi 6.2.1, 6.2.2

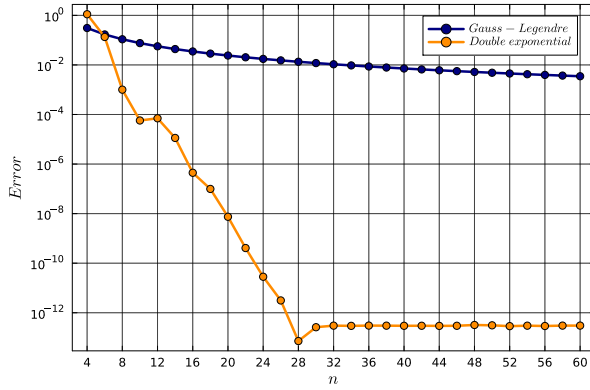
1. Scrivi un programma che implementi il metodo di Eulero per risolvere un *sistema* di equazioni differenziali.
2. Per ciascun problema ai valori iniziali (IVP), risolvi il problema usando il metodo di Eulero.
  - (i) Traccia il grafico della soluzione per  $n = 320$ .
  - (ii) Per  $n = 10 \cdot 2^k$ ,  $k = 2, 3, \dots, 10$ , calcola l'errore come  $\|u - \hat{u}\|_\infty = \max_{0 \leq i \leq n} |u_i - \hat{u}(t_i)|$  e all'istante finale,  $|u_n - \hat{u}(t_n)|$ . Realizza un grafico di convergenza log-log, includendo una retta di riferimento per la convergenza del primo ordine.



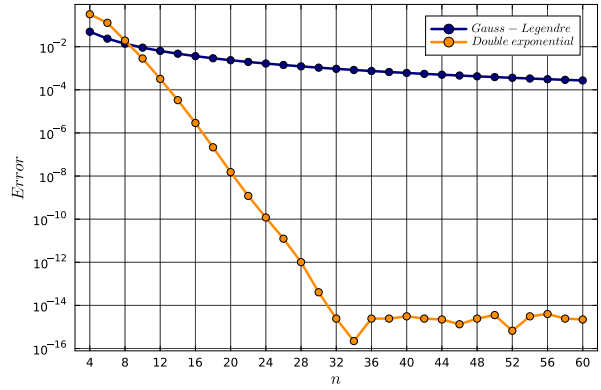
(a)



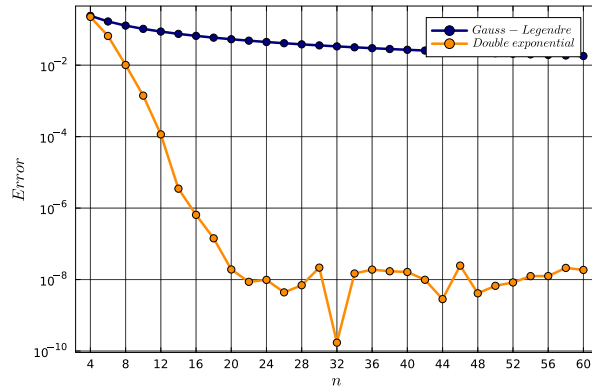
(b)



(c)



(d)



(e)

Figure 48: Studio degli errori per gli integrali con la regola di quadratura di Gauss-Legendre e double-exponential.

- (a)  $u' = -2tu$ ,  $0 \leq t \leq 2$ ,  $u(0) = 2$ ;  $\hat{u}(t) = 2e^{-t^2}$   
(b)  $u' = u + t$ ,  $0 \leq t \leq 1$ ,  $u(0) = 2$ ;  $\hat{u}(t) = -1 - t + 3e^t$   
(c)  $(1 + t^3)uu' = t^2$ ,  $0 \leq t \leq 3$ ,  $u(0) = 1$ ;  $\hat{u}(t) = [1 + (2/3) \ln(1 + t^3)]^{1/2}$

### 6.1.1 Soluzione

Si è implementato il metodo di Eulero per risolvere un sistema di equazioni differenziali. La funzione che sviluppa il metodo viene testata nei prossimi esercizi, sia nel caso di equazioni che di sistemi di equazioni.

Si riportano nelle figure 49, 50, 51 i grafici delle soluzioni, uniti ai grafici degli errori per i



tre casi richiesti. Le soluzioni sono calcolate per  $n = 320$  come richiesto, i grafici degli errori corrispondono a più soluzioni con  $n$  che varia da  $10 \cdot 2^k$  con  $k = 2, 3, \dots, 10$ .

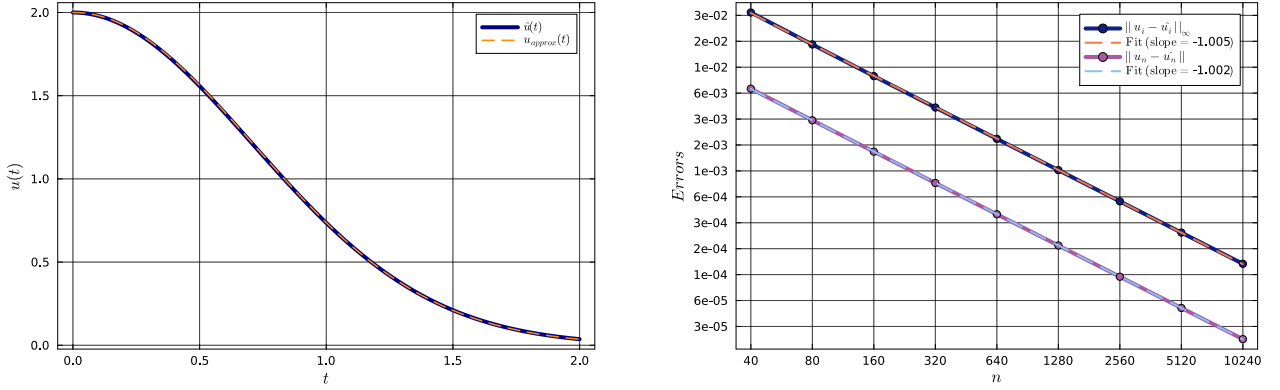


Figure 49: Soluzione e errore per il caso (a).

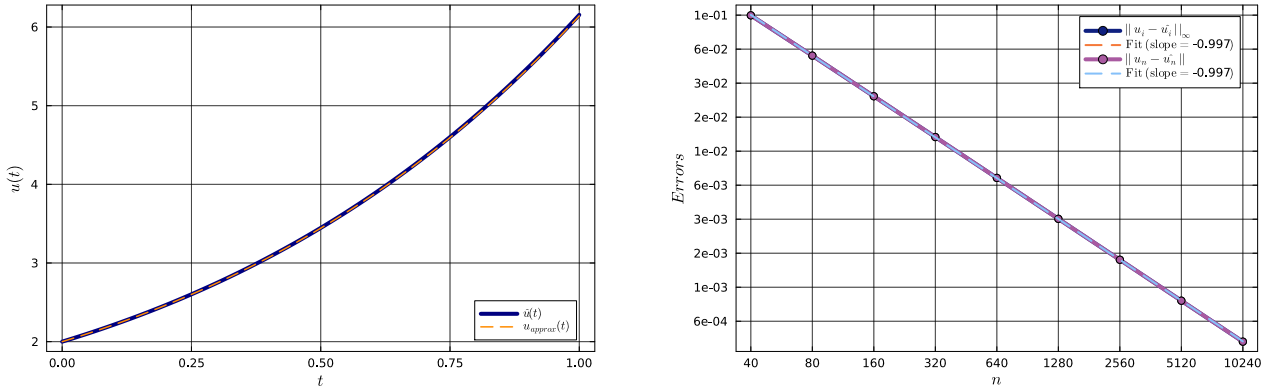


Figure 50: Soluzione e errore per il caso (b).

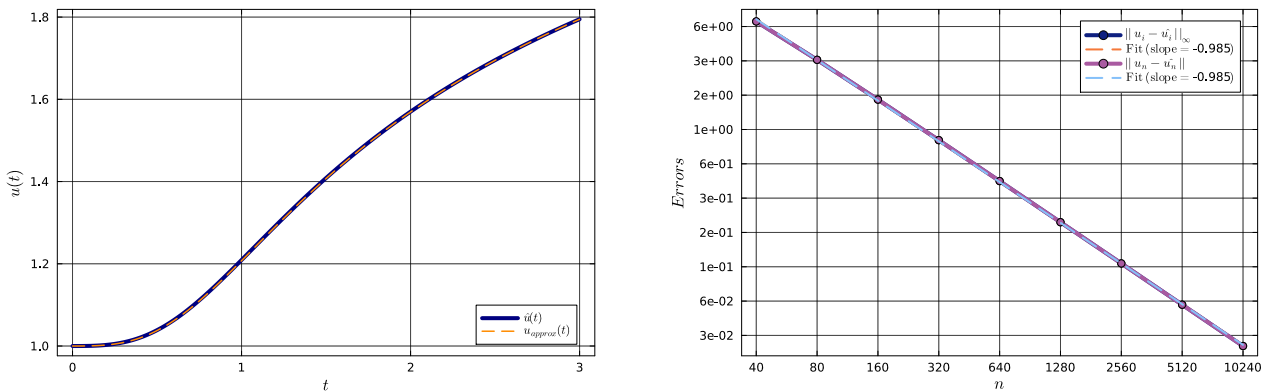


Figure 51: Soluzione e errore per il caso (c).

La prima osservazione che si può fare riguarda la convergenza del metodo di Eulero. In tutti i grafici gli errori sono stati calcolati sia come  $\|u - \hat{u}\|_\infty = \max_{0 \leq i \leq n} |u_i - \hat{u}(t_i)|$  che come  $|u_n - \hat{u}(t_n)|$ . In entrambi i casi si osserva che gli errori decrescono con un ordine di grandezza prossimo a  $O(h)$ , come atteso per il metodo di Eulero. Nei grafici infatti è stato riportato il coefficiente angolare dell'interpolazione degli errori con il metodo dei minimi quadrati. In tutti e tre i casi, per ciascuna tipologia di errore, il coefficiente angolare è prossimo a  $-1$ .

La seconda osservazione riguarda la differenza tra i due tipi di errore. Si osserva che solo nel primo grafico le due rette degli errori sono distinte, mentre negli altri due casi sono sovrapposte. Questo è lo stesso comportamento che si è studiato nella teoria, in cui perturbazioni del valore iniziale influenzano la soluzione. L'esempio fatto riguardava le due equazioni differenziali lineari  $u' = u$  e  $u' = -u$ , in cui la soluzione della prima equazione amplifica gli errori, mentre la soluzione della seconda equazione li smorza. Nel nostro caso, la soluzione dell'equazione (a) smorza gli errori avanzando nel tempo, per cui il massimo errore non si raggiunge all'istante finale. La diretta conseguenza è che le rette sono distinte. Negli altri due casi invece la soluzione amplifica gli errori, portando a rette sovrapposte.

## 6.2 Esercizio 6.2.3

Risolvi i seguenti problemi ai valori iniziali (IVP) con il metodo di Eulero utilizzando  $n = 1000$  passi. Traccia in un unico grafico la soluzione e la sua derivata prima, e in un altro grafico rappresenta l'errore in ciascuna componente in funzione del tempo.

(a)  $y'' + 9y = \sin(2t), \quad 0 < t < 2\pi, \quad y(0) = 2, \quad y'(0) = 1;$

$$\hat{y}(t) = \frac{1}{5} \sin(3t) + 2 \cos(3t) + \frac{1}{5} \sin(2t)$$

(b)  $y'' - 4y = 4t, \quad 0 < t < 1.5, \quad y(0) = 2, \quad y'(0) = -1;$

$$\hat{y}(t) = e^{2t} + e^{-2t} - t$$

(c)  $y'' + 4y' + 4y = t, \quad 0 < t < 4, \quad y(0) = 1, \quad y'(0) = \frac{3}{4};$

$$\hat{y}(t) = (3t + \frac{5}{4})e^{-2t} + \frac{t-1}{4}$$

### 6.2.1 Soluzione

Fissato  $n = 1000$  si è risolto il problema ai valori iniziali (IVP) con il metodo di Eulero. Per prima cosa si sono riscritte le equazioni differenziali di secondo ordine come un sistema di equazioni differenziali di primo ordine, in modo da poter utilizzare il metodo di Eulero:

$$\begin{cases} z_1' = z_2 \\ z_2' = -9z_1 + \sin(2t) \\ z_1(0) = 2 \\ z_2(0) = 1 \end{cases} \quad \begin{cases} z_1' = z_2 \\ z_2' = 4z_1 + 4t \\ z_1(0) = 2 \\ z_2(0) = -1 \end{cases} \quad \begin{cases} z_1' = z_2 \\ z_2' = -4z_1 - 4z_2 + t \\ z_1(0) = 1 \\ z_2(0) = \frac{3}{4} \end{cases}$$

Scritte così si è potuto utilizzare il metodo di Eulero per sistemi di equazioni differenziali. Si riportano in figura 52, 53, 54 i grafici delle soluzioni e delle loro derivate prime, insieme ai grafici degli errori per i tre casi richiesti. Gli errori rappresentano la distanza tra la soluzione calcolata e quella esatta, in ciascuna componente del sistema.

Nelle figure 52 e 53 è possibile notare che gli errori aumentano insieme al tempo. Infatti nel caso (a) la soluzione esatta è una combinazione di funzioni trigonometriche, per cui le perturbazioni iniziali si amplificano nel tempo. Nel caso (b) si ha lo stesso comportamento perchè a tempi più grandi domina l'esponenziale  $e^{2t}$ , che amplifica le perturbazioni iniziali. Al contrario in figura 54 si osserva che gli errori decrescono nel tempo, questo perchè la soluzione esatta è esponenziale decrescente, e così smorza l'errore.

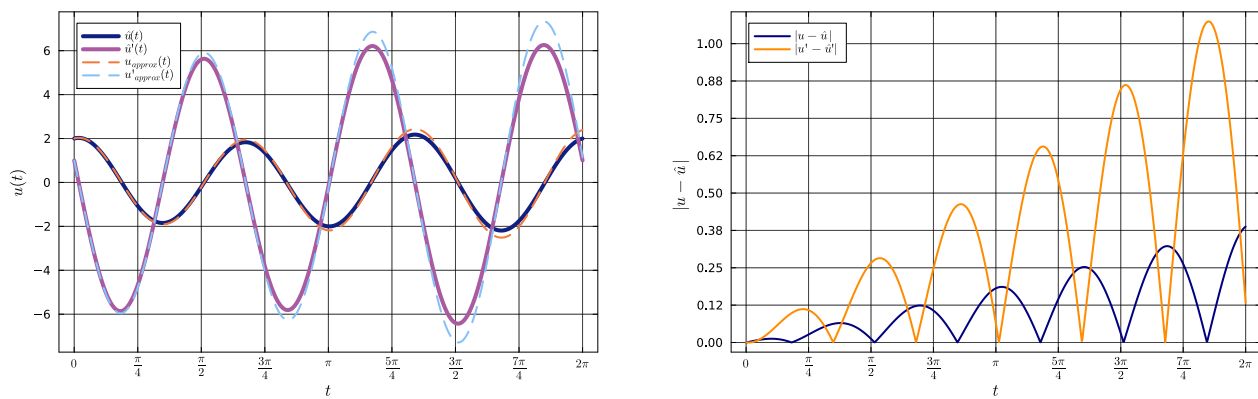


Figure 52: Soluzione e derivata prima per il caso (a). Errori.

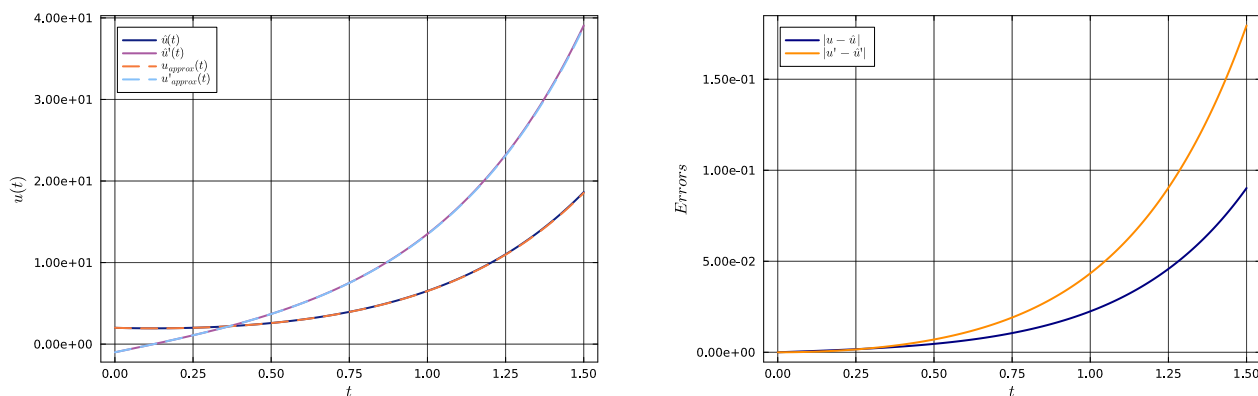


Figure 53: Soluzione e derivata prima per il caso (b). Errori.

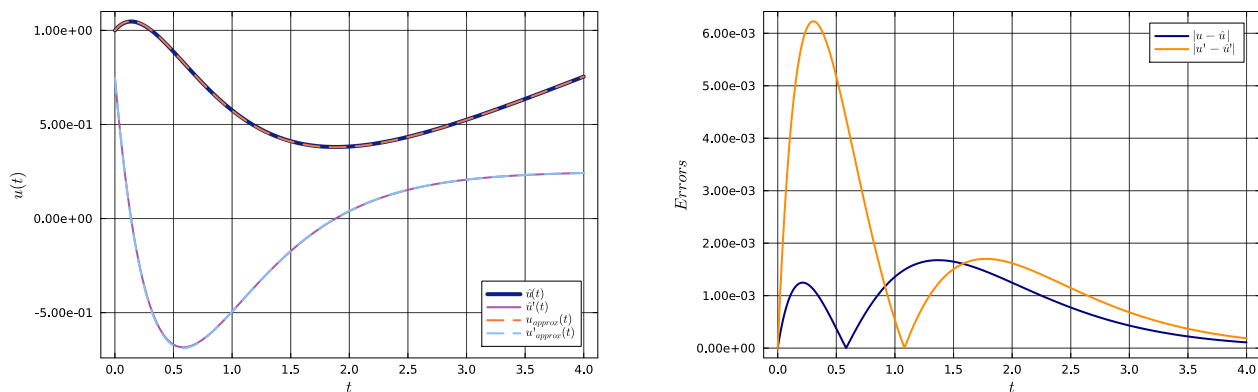


Figure 54: Soluzione e derivata prima per il caso (c). Errori.

### 6.3 Esercizio 6.3.1

- Scrivi un programma che risolva un sistema di equazioni differenziali utilizzando il metodo IE2 (Eulero implicito esplicito di ordine 2).
- Scrivi un programma che risolva un sistema di equazioni differenziali utilizzando il metodo RK4 (Runge-Kutta del quarto ordine).
- Testa la tua implementazione dei metodi IE2 e RK4 sul problema ai valori iniziali

$$u' = -2tu, \quad 0 \leq t \leq 2, \quad u(0) = 2;$$

$$\hat{u}(t) = 2e^{-t^2}$$

Risolvi per  $n = 30, 60, 90, \dots, 300$  e rappresenta graficamente l'errore massimo  $\|u - \hat{u}\|_\infty = \max_{0 \leq i \leq n} |u_i - \hat{u}(t_i)|$  in funzione del numero di valutazioni di funzione in un grafico log-log, insieme a una retta che mostri il tasso di convergenza atteso.

### 6.3.1 Soluzione

Si sono implementati i due metodi richiesti e applicati al problema ai valori iniziali (IVP) presentato. In figura 55 si riportano i grafici degli errori per i due metodi, calcolati per  $n = 30, 60, 90, \dots, 300$ . Nel grafico sono state tracciate le rette interpolanti degli errori, ottenute con

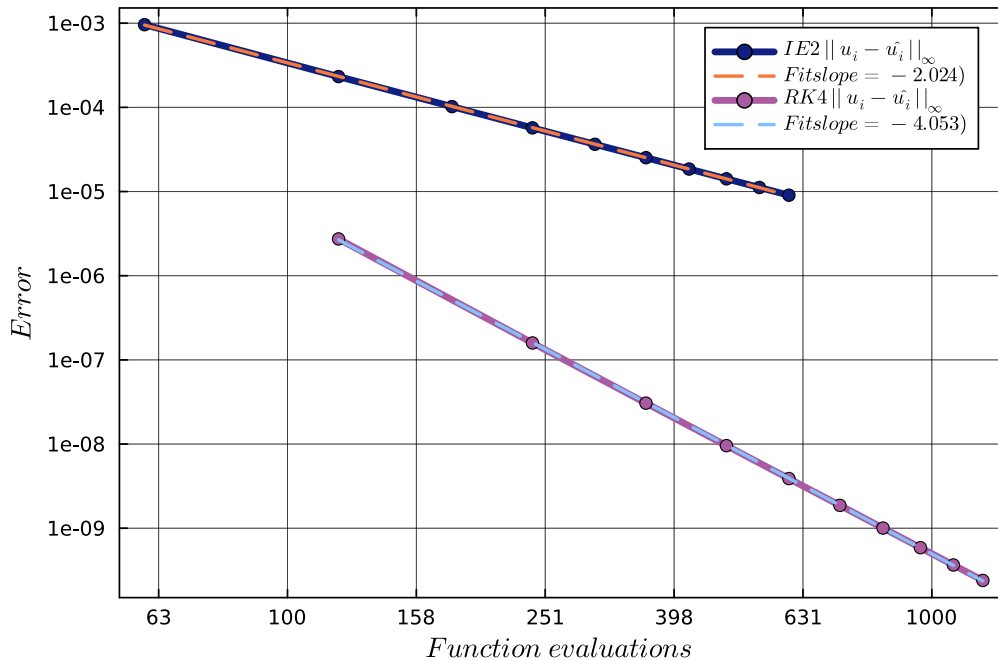


Figure 55: Confronto tra gli errori dei metodi IE2 e RK4.

il metodo dei minimi quadrati. Di tali rette sono riportate sul grafico le pendenze, in accordo con la teoria. Il metodo IE2 infatti deve avere un ordine di convergenza pari a 2, mentre il metodo RK4 deve avere un ordine di convergenza pari a 4.

I grafici sono stati realizzati in funzione del numero di valutazioni della funzione. Sapendo che IE2 richiede 2 valutazioni della funzione per ogni passo, e che RK4 richiede 4 valutazioni della funzione per ogni passo, si può spiegare il fatto che i punti del grafico si trovano incolonnati uno ogni due passi di IE2.

Infine si può notare che a parità di numero di valutazioni della funzione, il metodo RK4 ha un errore molto più piccolo rispetto al metodo IE2, evidenziandone l'efficacia.

### 6.4 Esercizio 6.3.2

In ciascuno dei seguenti casi, utilizza l'integratore RK4 per risolvere il sistema di ODE per  $0 \leq t \leq 10$  con le condizioni iniziali indicate. (A tal fine, devi trovare una procedura sensata per scegliere il numero di passi  $n$  da utilizzare.) Rappresenta i risultati come curve nel piano delle fasi, cioè con  $x$  e  $y$  come assi del grafico.

$$(a) \quad \begin{cases} x' = -4y + x(1 - x^2 - y^2) \\ y' = 4x + y(1 - x^2 - y^2) \end{cases} \\ \text{con } [x(0), y(0)] = [0.1, 0], [0, 1.9]$$

$$(b) \quad \begin{cases} x' = -4y - \frac{1}{4}x(1 - x^2 - y^2)(4 - x^2 - y^2) \\ y' = 4x - \frac{1}{4}y(1 - x^2 - y^2)(4 - x^2 - y^2) \end{cases}$$

con  $[x(0), y(0)] = [0.95, 0], \quad [0, 1.05], \quad [-2.5, 0]$

### 6.4.1 Soluzione

Si sono risolti i due sistemi di equazioni differenziali con il metodo RK4, per ciascuna delle condizioni iniziali indicate. Al fine di stimare un valore per il passo  $h$ , si è scelto di utilizzare il seguente algoritmo:

1. Si sceglie un numero di passi iniziale  $n$ .
2. Si calcola la soluzione per  $n$  passi, ottenendo la soluzione  $u$ .
3. Si calcola la soluzione per  $2n$  passi, ottenendo la soluzione  $\tilde{u}$ .
4. Si calcola l'errore come  $\|u - \tilde{u}\|_\infty$ , dove di  $\tilde{u}$  si scelgono solo i punti che corrispondono a  $u$ .
5. Se l'errore è maggiore di una tolleranza  $\epsilon$ , si raddoppia il passo e si torna al punto 2, altrimenti l'algoritmo si interrompe, restituisce  $n$  e  $h = \frac{b-a}{n}$ .

Per tutte le condizioni iniziali si è scelto come tolleranza  $\epsilon = 10^{-6}$ , e come numero di punti iniziali  $n = 1000$ .

Da notare che nell'algoritmo presentato ad ogni passo bisogna calcolare solamente una nuova soluzione perchè si riutilizza quella precedente. Con l'algoritmo proposto si riesce a stimare un valore di  $h$  che garantisca un errore inferiore alla tolleranza, ma non è certamente paragonabile al metodo adattivo, che permette di scegliere il passo in maniera più efficiente, a ogni passaggio. Si riportano in figura 56, 57 i grafici delle soluzioni ottenute, rappresentate come curve nel piano delle fasi, per i due casi richiesti. Inoltre si sono rappresentate le soluzioni in tre dimensioni, con il tempo come terza coordinata.

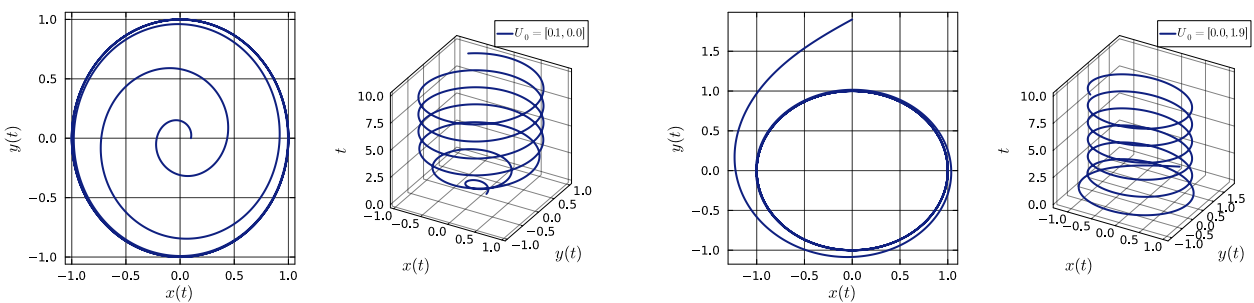


Figure 56: Soluzioni del sistema di ODE (a).

Infine si riportano in tabella 9 i valori di  $h$  ottenuti tramite il metodo precedentemente descritto.

### 6.5 Esercizio 6.3.3

Una malattia endemica in una popolazione può essere modellata tracciando la frazione di popolazione suscettibile all'infezione,  $v(t)$ , e la frazione infetta,  $w(t)$ . (Il resto della popolazione si considera guarito e immune.) Un modello tipico è il *modello SIR*:

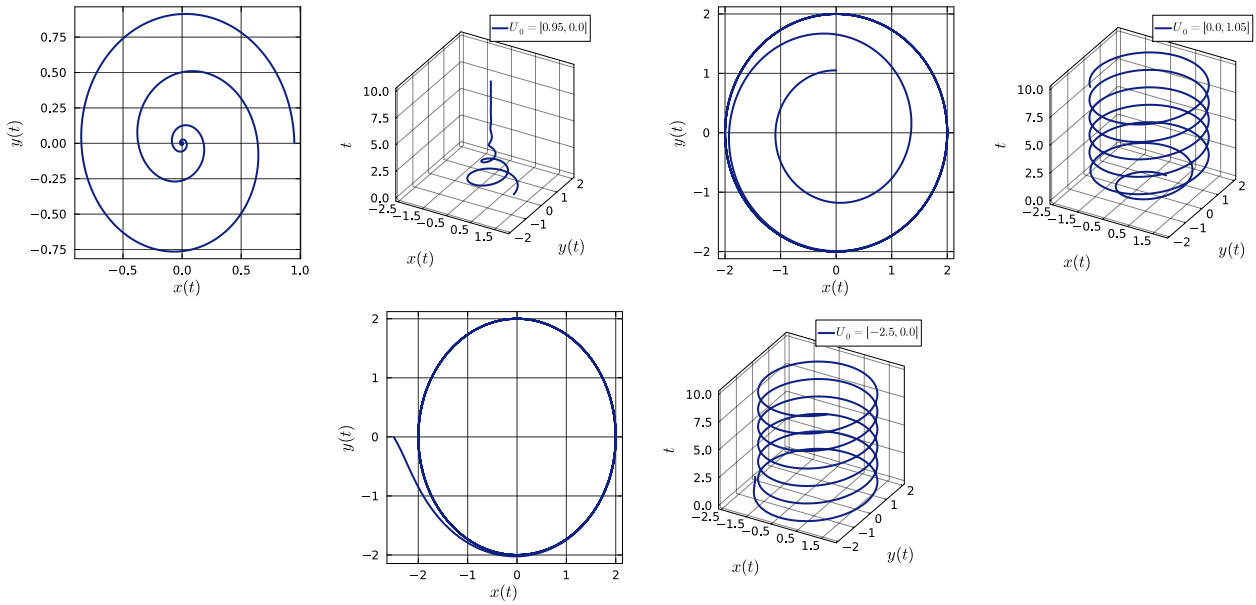


Figure 57: Soluzioni del sistema di ODE (b).

Table 9: Valori di  $h$  ottenuti per i due sistemi di ODE

Sistema	Condizioni iniziali	$h$
(a)	$[0.1, 0]$	0.0025
(a)	$[0, 1.9]$	0.0025
(b)	$[0.95, 0]$	0.0025
(b)	$[0, 1.05]$	0.00125
(b)	$[-2.5, 0]$	0.00125

$$\frac{dv}{dt} = 0.2(1 - v) - 3vw, \quad \frac{dw}{dt} = (3v - 1)w. \quad (59)$$

Partendo da  $v(0) = 0.95$  e  $w(0) = 0.05$ , utilizza il metodo RK4 per trovare i valori stazionari a lungo termine di  $v(t)$  e  $w(t)$ . Rappresenta graficamente entrambe le componenti della soluzione in funzione del tempo.

### 6.5.1 Soluzione

Si è risolto il sistema di equazioni differenziali con il metodo RK4, partendo dalle condizioni iniziali  $v(0) = 0.95$  e  $w(0) = 0.05$ . Il passo  $h$  è stato stimato allo stesso modo dell'esercizio 6.4, in particolare si è ottenuto  $h = 0.015625$  con un numero di step iniziali  $n = 100$ . Si riporta in figura 58 il grafico della soluzione, in cui si osserva che le due componenti convergono a un valore stazionario.

Per valutare i valori stazionari, una volta ottenute le soluzioni si sono calcolate le differenze tra i valori successivi di  $v$  e  $w$ , fino a quando non si è raggiunta la tolleranza di  $10^{-9}$ .

Si è poi scelto il valore di tempo più grande tra quelli ottenuti per le due soluzioni, detto  $t_{steady}$ . È importante notare che il valore di  $t_{steady}$  dipende fortemente dal metodo scelto per stimarlo e perciò potrebbe essere una sovrastima. Certamente la valutazione della distanza tra due valori successivi è una condizione molto stringente perchè non tiene conto del cambiamento globale della soluzione, inoltre suppone che la soluzione raggiunga un valore stazionario. Per quanto riguarda il caso in esame, poco importa: l'esercizio suppone che il sistema raggiunga un valore stazionario e, se anche il metodo adottato portasse a una sovrastima di  $t_{steady}$ , ciò non

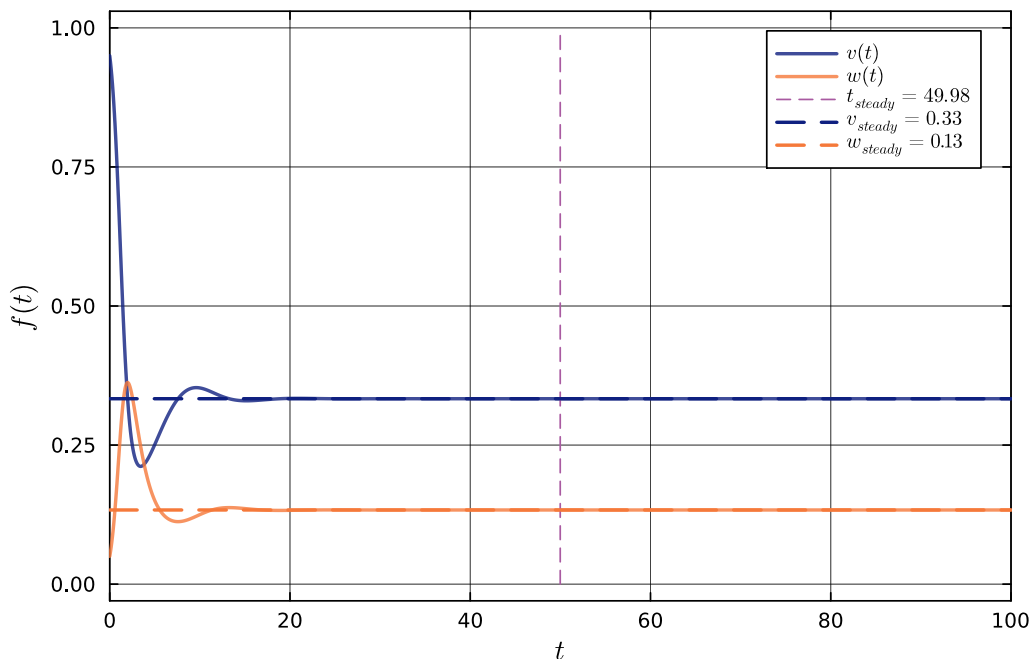


Figure 58: Soluzione del modello SIR.

influenzerebbe la valutazione degli stati stazionari. Il valore degli stessi è riportato nel grafico, insieme al valore di  $t_{steady}$ .

## 6.6 Esercizio 6.3.5

- Scrivi un programma che risolva un sistema di equazioni differenziali utilizzando il metodo adattivo BS23.
- Testa la tua implementazione risolvendo

$$u' = -2tu, \quad 0 \leq t \leq 2, \quad u(0) = 2; \quad \hat{u}(t) = 2e^{-t^2}$$

Rappresenta graficamente la soluzione in funzione del tempo per  $\delta = 10^{-8}$  in un grafico e il valore del passo in un altro. Determina inoltre il passo minimo e medio, escludendo il primo e l'ultimo passo.

- Considera  $\delta = 10^{-6}, 10^{-5}, \dots, 10^{-12}$  e studia come l'errore sulla soluzione all'istante finale,  $|u_n - \hat{u}(t_n)|$ , dipende dalla tolleranza in input  $\delta$ .

### 6.6.1 Soluzione

Si è implementato il metodo adattivo BS23 per risolvere un sistema di equazioni differenziali, e si riporta in figura 59 il grafico della soluzione ottenuta, insieme al grafico del passo  $h$  in funzione del tempo.

Si è scelto di togliere il primo e l'ultimo passo perchè il primo è arbitrario, scelto dall'utente, e l'ultimo è forzato in modo da raggiungere l'estremo  $b$ , quindi non è scelto dall'algoritmo. Così facendo si ottiene un passo minimo di  $h_{min} = 0.0043$  e un passo medio di  $h_{med} = 0.0054$ . Lo studio del passo è interessante perchè permette di capire di quanto si è guadagnato rispetto a un metodo non adattivo. Infatti se avessimo scelto passo fisso pari a  $h_{med}$  avremmo ottenuto un numero di punti maggiore perchè nel metodo adattivo il passo può anche diminuire. Nel caso in esame il guadagno non è molto significativo, ma in generale se la differenza fra il passo

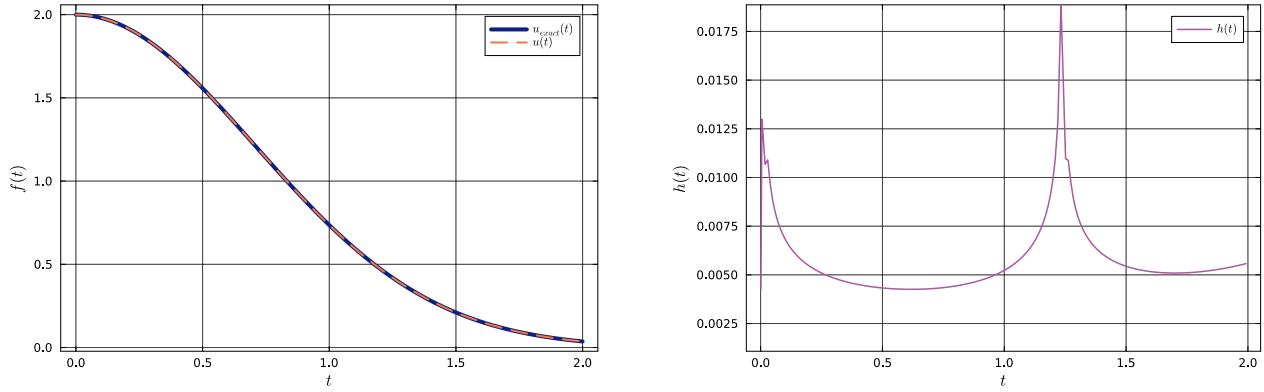


Figure 59: Soluzione e passo del metodo BS23.

minimo e il passo medio è molto grande, si possono ottenere integrazioni migliori. In figura 60 si riporta il grafico dell'errore all'istante finale,  $|u_n - \hat{u}(t_n)|$ , in funzione della tolleranza  $\delta$  scelta. Si osserva che l'errore cresce all'aumentare della tolleranza. Infatti la media dei passi  $h$  cresce all'aumentare della tolleranza, e la soluzione viene integrata da un numero minore di punti.

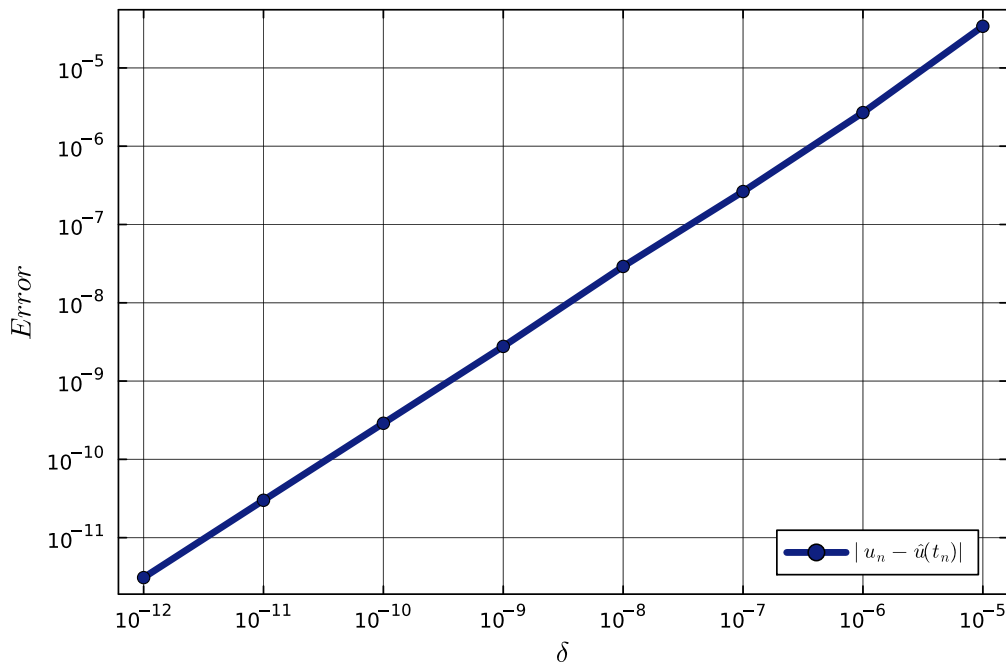


Figure 60: Errore all'istante finale in funzione della tolleranza  $\delta$ .

## 6.7 Esercizio 6.3.6

Utilizzando il metodo BS23 con tolleranza sull'errore  $\delta = 10^{-8}$ , risolvi il problema  $y'' + (1 + y')^3 y = 0$  nell'intervallo  $0 \leq t \leq 4\pi$  per le seguenti condizioni iniziali. Rappresenta graficamente  $y(t)$  e  $y'(t)$  in funzione di  $t$  e, separatamente, il passo  $h$  in funzione di  $t$ . Determina inoltre il passo minimo e medio, escludendo il primo e l'ultimo valore.

- $y(0) = 0.1, \quad y'(0) = 0$
- $y(0) = 0.5, \quad y'(0) = 0$



- $y(0) = 0.75, \quad y'(0) = 0$
- $y(0) = 0.95, \quad y'(0) = 0$

### 6.7.1 Soluzione

Il sistema di equazioni differenziali del primo ordine associato all'equazione differenziale di secondo ordine è

$$\begin{cases} z_1' = z_2 \\ z_2' = -(1 + z_2)^3 z_1 \end{cases}$$

dove  $z_1 = y$  e  $z_2 = y'$ . Nelle figure 61, 62, 63, 64 sono riportati i grafici delle soluzioni ottenute, insieme ai grafici del passo  $h$  in funzione del tempo.

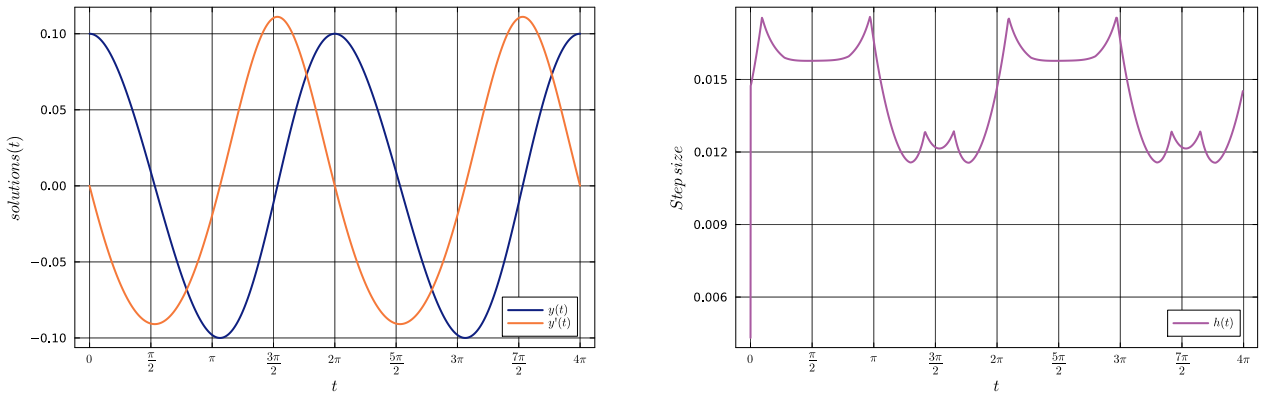


Figure 61: Soluzione e derivata prima per il caso  $y(0) = 0.1, y'(0) = 0$ . Passo del metodo BS23.

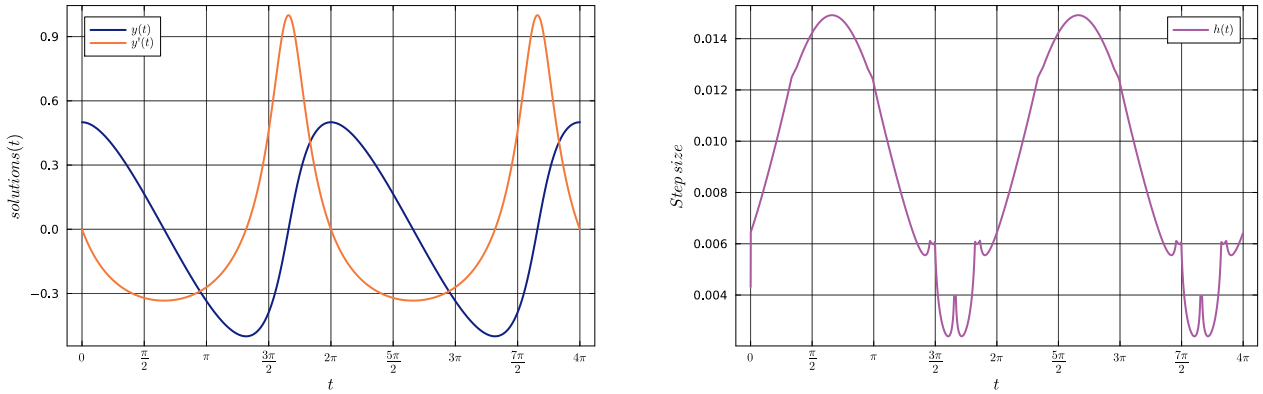


Figure 62: Soluzione e derivata prima per il caso  $y(0) = 0.5, y'(0) = 0$ . Passo del metodo BS23.

Nei vari casi si osserva che tanto più la soluzione presenta picchi accentuati, tanto più il metodo BS23 sceglie un passo piccolo. In tabella 10 sono riportati i valori di passo minimo e medio. Si rimanda all'esercizio 6.6 la discussione sull'importanza dello studio del passo.

La tabella 10 mostra il guadagno rispetto a un metodo non adattivo è molto significativo negli ultimi due casi, dove le soluzioni presentano picchi molto accentuati, perchè il metodo BS23, nonostante abbia passo medio dell'ordine di  $10^{-3}$ , raggiunge passi minimi dell'ordine di  $10^{-4}$  e  $10^{-5}$ .

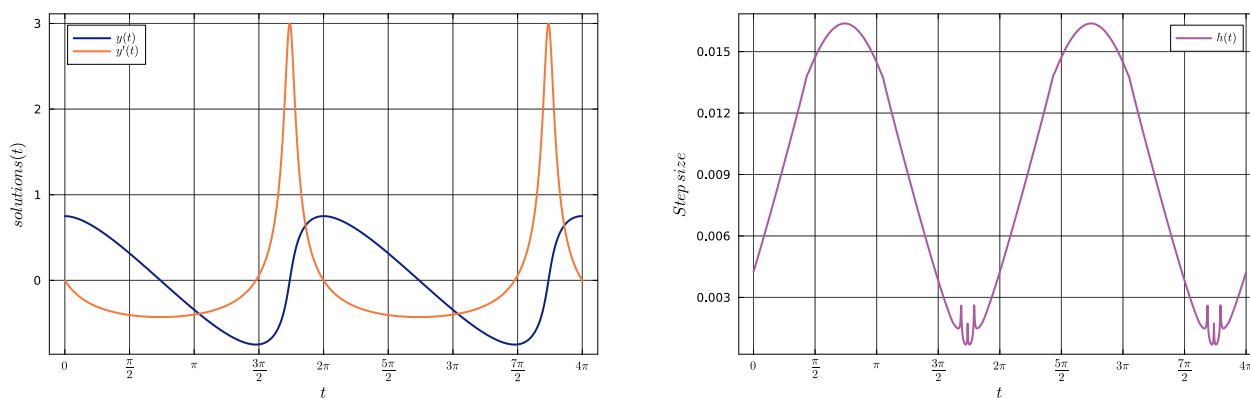


Figure 63: Soluzione e derivata prima per il caso  $y(0) = 0.75$ ,  $y'(0) = 0$ . Passo del metodo BS23.

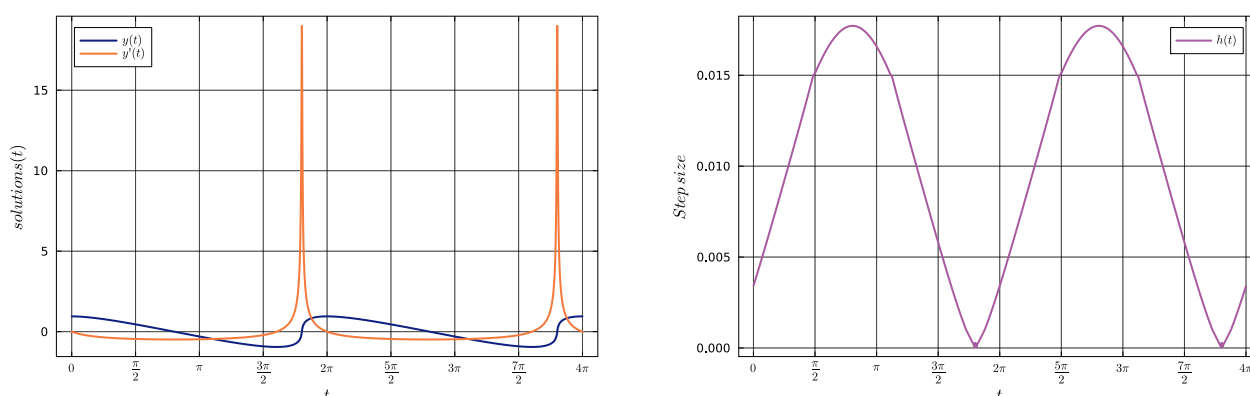


Figure 64: Soluzione e derivata prima per il caso  $y(0) = 0.95$ ,  $y'(0) = 0$ . Passo del metodo BS23.

Table 10: Valori di passo minimo e medio per i quattro casi

Caso	$h_{min}$	$h_{med}$
<b>a</b>	$4.3 \times 10^{-3}$	$1.41 \times 10^{-2}$
<b>b</b>	$2.4 \times 10^{-3}$	$6.9 \times 10^{-3}$
<b>c</b>	$6.97 \times 10^{-4}$	$4.48 \times 10^{-3}$
<b>d</b>	$5.56 \times 10^{-5}$	$2.49 \times 10^{-3}$

## 6.8 Esercizio 6.3.7

Risolvi il problema  $u' = 100u^2 - u^3$ ,  $u(0) = 0.0002$ ,  $0 \leq t \leq 100$  utilizzando il metodo BS23 e realizza grafici che mostrino sia la soluzione sia i passi temporali scelti per  $\delta = 10^{-8}$ . La soluzione effettua una rapida transizione tra due stati quasi costanti. Il passo scelto dall'algoritmo si comporta allo stesso modo in entrambi gli stati?

### 6.8.1 Soluzione

Si è risolto il problema ai valori iniziali (IVP) con il metodo BS23. Si riporta in figura 65 il grafico della soluzione ottenuta, insieme al grafico del passo  $h$  in funzione del tempo.

L'integrazione avviene con successo, ma il grafico del passo mostra un comportamento peculiare. Ci si potrebbe aspettare che, data la presenza di due zone analoghe per la soluzione, il passo scelto dall'algoritmo fosse simile in entrambe le zone, con una fase transiente in cui

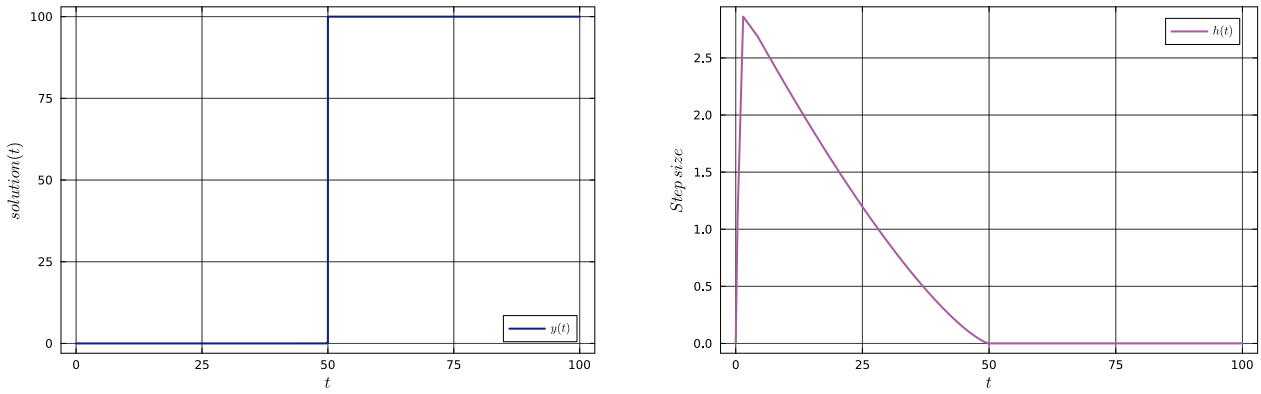


Figure 65: Soluzione e passo del metodo BS23.

il passo diminuisce rapidamente. Invece si osserva che il passo è molto più grande nella zona iniziale, dove la soluzione cresce lentamente, per poi rimanere costante nel tempo, e molto piccolo. Il problema descritto appartiene perciò alla categoria degli *stiff problems*.

Per studiare meglio il comportamento del passo, si è scelto di calcolare i passi minimo e medio, dividendo le due regioni della soluzione. La prima regione comincia da  $t = 0$  e termina a  $t = 51$ , dove il valore finale si è scelto in modo da includere la fase transiente, e la seconda regione comincia da  $t = 51$  e termina a  $t = 100$ . In tabella 11 sono riportati i valori di passo minimo e medio per le due regioni.

Giustamente si osserva il minimo nella prima regione, perchè il minimo valore di  $h$  viene

Table 11: Valori di passo minimo e medio per il problema stiff

Regione	$h_{min}$	$h_{med}$
Prima regione	$1.6 \times 10^{-6}$	0.00845
Seconda regione	0.00025	0.00025

assunto in corrispondenza del salto della soluzione, in cui il metodo adattivo cerca passi molto piccoli per ottenere una soluzione accurata. Ciò che è particolare è che nella seconda regione il passo si stabilizzi, rimanendo costante e molto piccolo.

## 7 Appendice

### 7.1 Dati

Table 12: Distanza dal Sole e periodo orbitale dei pianeti del Sistema Solare

Pianeta	Distanza dal Sole [Mkm]	Periodo orbitale [giorni]
Mercurio	57.59	87.99
Venere	108.11	224.7
Terra	149.57	365.26
Marte	227.84	686.98
Giove	778.14	4332.4
Saturno	1427	10759
Urano	2870.3	30684
Nettuno	4499.9	60188