

# Esperimentazioni di Fisica Computazionale

## Relazione Finale

13 maggio 2024



Riccardo Bonomelli  
Matricola 885801

Università degli Studi di Milano-Bicocca

Dipartimento di Fisica

Anno Accademico 2023/2024



## Abstract e struttura della relazione

L'analisi numerica è la branca della matematica che si occupa di studiare procedure risolutive di un problema per mezzo di algoritmi computazionali. L'utilità di tale disciplina consiste nel fornire metodi risolutivi, aventi struttura algoritmica, in grado di determinare soluzioni, spesso approssimate, a problemi complessi, inaccessibili analiticamente.

La presente relazione si propone di studiare diversi aspetti legati al calcolo numerico: il comportamento di un calcolatore nelle operazioni di calcolo e alcuni metodi numerici per la risoluzione di integrali, di equazioni differenziali ordinarie, per la generazione di sequenze casuali e per la ricerca di zeri di funzione. Le tecniche discusse sono state poi applicate allo studio di alcuni noti e meno noti sistemi dinamici (anche meccanici), al fine di verificare o dedurre alcune loro proprietà. In particolare, si intende discutere la risoluzione e lo studio di 20 esercizi al calcolatore legati ai temi in esame.

Al fine di rendere il lavoro un lavoro chiuso, ogni blocco di esercizi è preceduto da una breve prefazione teorica, in cui sono stati sintetizzati i concetti essenziali per la comprensione delle procedure e delle tecniche utilizzate senza particolari altri riferimenti. Le prefazioni sono state utilizzate anche per raggruppare alcuni risultati ricorrenti ricavati, che sono poi stati richiamati durante lo svolgimento degli esercizi al fine di evitare un numero eccessivo di ripetizioni.

Il linguaggio di programmazione con il quale sono stati scritti i codici che hanno permesso i risultati che seguono è C++. Per la parte grafica e per alcuni aspetti legati all'analisi dati (principalmente fit) si sono utilizzate le librerie di ROOT (CERN).



# Indice

<b>Condizionamento e stabilità</b>	<b>1</b>
Esercizio 1 . . . . .	5
Esercizio 2 . . . . .	13
<b>Integrazione deterministica</b>	<b>25</b>
Esercizio 3 . . . . .	34
Esercizio 4 . . . . .	44
Esercizio 5 . . . . .	52
Esercizio 6 . . . . .	59
Esercizio 7 . . . . .	72
<b>Numeri casuali e densità di probabilità</b>	<b>75</b>
Esercizio 8 . . . . .	79
Esercizio 9 . . . . .	92
Esercizio 10 . . . . .	107
Esercizio 11 . . . . .	118
Esercizio 12 . . . . .	126
Esercizio 13 . . . . .	137
<b>Metodi ad un passo per le ODE</b>	<b>147</b>
Esercizio 14 . . . . .	152
Esercizio 15 . . . . .	166
Esercizio 16 . . . . .	199
Esercizio 17 . . . . .	210
Esercizio 18 . . . . .	225
<b>Zeri di funzione</b>	<b>242</b>
Esercizio 19 . . . . .	245
Esercizio 20 . . . . .	253



## Condizionamento e stabilità

Il risultato di un calcolo numerico, ossia di una successione di operazioni al calcolatore, è sempre determinato a meno di una certa *incertezza* o *errore*. I tipi di errore risultanti da un calcolo numerico sono sostanzialmente due.

### Errori di arrotondamento

Le operazioni su un calcolatore vengono, molto spesso, eseguite su numeri reali. Tuttavia, in generale, i numeri reali possiedono rappresentazione decimale illimitata: si pensi, ad esempio, al caso dei numeri irrazionali

$$\pi = 3.141592\ldots \quad \text{o} \quad \sqrt{2} = 1.414213\ldots$$

Per tale ragione, il calcolatore, disponendo di uno spazio in memoria limitato, dovrà necessariamente troncare la serie decimale in qualche punto, in funzione della quantità di informazione che si vuole estrarre da ogni variabile, ossia della *precisione* selezionata. La precisione di un numero reale può essere semplice, doppia o quadrupla: essa stabilisce il numero di cifre significative allocate in memoria e può essere impostata dall'utente in funzione delle necessità specifiche del calcolo. In particolare, un numero reale è rappresentato nella memoria di un calcolatore secondo il principio della virgola mobile, che corrisponde, di fatto, all'analogo della notazione scientifica in base 2. La rappresentazione di  $x \in \mathbb{R}$  in virgola mobile si compone di tre costituenti fondamentali:

$$x = s \cdot m \cdot 2^e$$

dove  $s$  è il segno, per il quale si alloca generalmente sempre 1 bit,  $m$  è la mantissa, ossia la parte decimale e, infine,  $e$  è l'esponente, che determina l'ordine di grandezza del numero reale. Al fine di poter rappresentare numeri molto vicini allo zero, l'esponente viene sempre scritto come

$$e = n - \text{offset}$$

da cui seguono i problemi di overflow e underflow, ossia l'esistenza di intervalli reali non rappresentabili: un intorno sferico dello zero e due intervalli illimitati sul semiasse reale positivo e negativo. Di seguito è riportata una tabella contenente il numero di bit allocati per le tre diverse precisioni.

Precisione	$s$	$m$	$e$	Bit totali
semplice	1	23	8	32
doppia	1	52	11	64
quadrupla	1	112	15	128

Al diverso spazio in memoria destinato per la mantissa e per l'esponente corrisponderà una diversa quantità di informazione immagazzinata sul numero reale rappresentato. Di seguito una tabella riassuntiva.

Precisione	$m$	$e$
semplice	7 cifre	(−38, 38)
doppia	16 cifre	(−308, 308)
quadrupla	34 cifre	(−4932, 4932)

Gli errori di arrotondamento sono dunque errori derivanti dal troncamento della serie decimale di un numero reale in un punto, e seguono dalle modalità con cui il calcolatore rappresenta in memoria i numeri. In particolare, gli errori di arrotondamento si manifestano durante un calcolo numerico in modo non trascurabile quando si svolgono due tipi di operazione:

- differenze di numeri vicini tra loro
- somme di numeri aventi diverso ordine di grandezza

In entrambi i casi è facile verificare che esiste una buona probabilità di perdere gran parte dell'informazione relativa al calcolo, che può ripercuotersi nelle operazioni successive. I modi per evitare di incorrere in problemi di arrotondamento sono, sostanzialmente, due: aumentare la precisione con cui si allocano i numeri reali, oppure definire il problema in modo diverso ma matematicamente equivalente, al fine di limitare le operazioni problematiche di cui sopra.

### Errori di troncamento

Gli errori di troncamento consistono negli errori derivanti dall'approssimazione di una successione infinita di operazioni ad una finita. I casi tipici che si avranno modo di studiare sono l'approssimazione di un integrale alla sua successione delle somme parziali o l'approssimazione di una derivata al suo rapporto incrementale finito. Il concetto alla base è quindi analogo a quello degli errori di arrotondamento: così come non è possibile immagazzinare un numero di cifre infinito in memoria, non sarà possibile, per il calcolatore, svolgere una successione infinita di steps per risolvere un certo problema. Detta  $\tilde{x}$  la soluzione numerica di un problema a seguito di un troncamento e  $x_{\text{true}}$  il valor vero della soluzione, avremo allora che

$$x_{\text{true}} = \tilde{x} \pm \varepsilon$$

ossia la soluzione al problema sarà determinata a meno di un certo errore  $\varepsilon$ . Uno dei punti centrali dell'analisi numerica consiste nel saper quantificare o definire l'andamento di  $\varepsilon$ .

Esistono, infine, due concetti chiave per definire precisamente un problema e un algoritmo nel loro contesto.

**Definizione 0.1.** Si consideri un problema numerico  $P$ , ossia una relazione funzionale tra i dati e i risultati della forma

$$P : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$\vec{x} \mapsto \vec{P} = \vec{P}(\vec{x})$$

dove  $\vec{x}$  è il vettore dei dati e  $\vec{P}$  quello dei risultati. Diremo che  $P$  è un problema *ben condizionato* se (in una norma opportuna)

$$\left\| \vec{P}(\vec{x} + \vec{\epsilon}) - \vec{P}(\vec{x}) \right\| \rightarrow 0 \quad \text{per} \quad \|\vec{\epsilon}\| \rightarrow 0$$

ossia se a piccole variazioni dei dati corrispondono piccole variazioni dei risultati. Altrimenti, diremo che il problema è *mal condizionato*.

Si noti che nella definizione si è assunto un generico problema  $P$  in astratto, ossia la definizione è indipendente dal calcolatore e dunque dagli errori di calcolo di cui si è discusso in precedenza. Il condizionamento è quindi legato esclusivamente al problema numerico in esame, e non ha alcuna relazione né con l'architettura di un calcolatore, né con l'algoritmo utilizzato per risolvere il problema. Quando un problema è mal condizionato, l'unica strada da seguire risulta quindi essere quella aumentare la precisione e lavorare in un range tale per cui gli effetti del mal condizionamento siano ridotti. Un problema, infatti, può essere mal condizionato per alcuni dati ma ben condizionato per altri. In generale, per problemi mal condizionati è opportuna una riformulazione equivalente ben condizionata, ove possibile. Si consideri ora il caso particolare della definizione 0.1 in una dimensione. Esiste un modo per quantificare e prevedere il condizionamento di un problema.

**Definizione 0.2.** Si consideri un problema numerico  $P$ . Sia  $\tilde{x}$  il dato  $x$  perturbato di una quantità infinitesima. Chiameremo *errore relativo su  $x$*  la quantità

$$\varepsilon(x) := \left| \frac{\tilde{x} - x}{x} \right|$$

Chiameremo *errore relativo su  $P$*  la quantità

$$\varepsilon[P(x)] := \left| \frac{P(\tilde{x}) - P(x)}{P(x)} \right|$$

Chiameremo *numero di condizionamento* la quantità

$$k := \frac{\varepsilon[P(x)]}{\varepsilon(x)} = \left| \frac{xP'(x)}{P(x)} \right|$$

Visto come è stato definito il numero di condizionamento, la proposizione che segue sarà allora evidente all'intuizione.

**Proposizione 0.3.** *Si consideri un problema numerico  $P$ . Sia  $k$  il suo numero di condizionamento. Allora*

- se  $k$  grande  $\implies P$  è mal condizionato
- se  $k$  piccolo  $\implies P$  è ben condizionato

Il numero di condizionamento  $k$ , dunque, quantifica e caratterizza la sensibilità della soluzione di un problema a piccoli cambiamenti dei dati, rendendo possibile prevedere i casi in cui trattare un problema con una certa attenzione numerica. Ad ogni problema è possibile associare uno o più algoritmi risolutivi.

**Definizione 0.4.** Si consideri un problema numerico  $P$  e un certo algoritmo  $A$  che lo risolva. Sia  $\tilde{x}$  il dato  $x$  perturbato di una quantità infinitesima. Diremo che  $A$  è un algoritmo *stabile* per  $P$  se

$$\left| \frac{A(\tilde{x}) - P(\tilde{x})}{P(\tilde{x})} \right| \sim n$$

dove  $n$  è il numero di operazioni che compongono  $A$ . Altrimenti, diremo che l'algoritmo  $A$  è *instabile* per il problema  $P$ .

Un algoritmo, dunque, si dice stabile per un problema quando l'inevitabile propagazione dell'errore dovuto all'errore iniziale di arrotondamento e all'aritmetica finita del calcolatore è proporzionale al numero di operazioni elementari che compongono l'algoritmo stesso, ossia se è dell'ordine di grandezza della precisione macchina. La stabilità di un algoritmo è allora sempre definita in relazione ad un certo problema: un algoritmo può essere stabile per un problema, ma non esserlo per un altro. La stabilità valuta la reazione fornita da un algoritmo ad una perturbazione dei dati iniziali: un algoritmo stabile per un certo problema non amplifica più dell'inevitabile gli errori durante il calcolo, a differenza di un algoritmo instabile. Negli esercizi che seguono si avrà modo di osservare come i problemi di stabilità algoritmica possano portare alla produzione di risultati del tutto insensati.

## Esercizio 1

Si vuole stimare numericamente il valore della serie numerica

$$\sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.64493406685$$

Il valore esatto può essere ricavato per via analitica utilizzando opportunamente l'identità di Parseval.

Anzitutto, si noti che una serie numerica è definita come il limite della sua successione delle somme parziali, ossia

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

Alla luce del fatto che il calcolatore può compiere solo un numero finito di operazioni, risulta allora possibile ottenere una stima del valore della serie numerica valutando la successione  $S(N)$  in  $N = \tilde{N} \in \mathbb{N}$  dove  $\tilde{N}$  è un numero grande. Si noti che, con questa operazione, si sta commettendo un errore di troncamento: si sta infatti approssimando una somma infinita di termini in una somma finita. Ci si aspetta che, in generale, la stima della serie numerica migliori in precisione all'aumentare di  $N$  oppure, equivalentemente, che l'errore dovuto al troncamento dei termini della serie diminuisca all'aumentare di  $N$ . Vedremo che, sotto certe condizioni, questa supposizione verrà smentita. Vogliamo quindi verificare come si comporta il calcolatore nel calcolo di  $S(N)$ . In particolare, si vogliono confrontare i risultati ottenuti in precisione singola e in precisione doppia nei casi in cui si effettua la stima in somma diretta ( $1 \rightarrow N$ ) e nei casi in cui la si effettua in somma inversa ( $N \rightarrow 1$ ). Sia nel caso della somma diretta, che in quello della somma inversa, a seguito di un'analisi qualitativa dei dati ottenuti, si è svolto uno studio più accurato plottando la dispersione dal valore vero, definita come

$$\Delta(N) := \left| S(N) - \frac{\pi^2}{6} \right|$$

considerando un range di valori sufficientemente ampio da apprezzarne l'andamento. Evidentemente, affinché l'algoritmo  $S(N)$  possa fornire una stima non distorta del valore della serie numerica, dovrà valere la condizione di convergenza

$$\lim_{N \rightarrow +\infty} \Delta(N) = 0 \tag{1}$$

### Precisione singola

Si sono stampati a schermo i valori di  $S(N)$  al variare di  $N$  in precisione singola, prima in somma diretta  $S_d$  e poi in somma inversa  $S_i$ . Si è scelto il range

$$100 \leq N < 10000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

convenzionalmente, al fine di svolgere una prima analisi qualitativa dei risultati analizzando un range di valori sufficientemente ampio. La tabella che segue mostra quanto ottenuto.

$N$	$S_d(N)$	$S_i(N)$	$N$	$S_d(N)$	$S_i(N)$
100	1.63498	1.63498	5100	1.64473	1.64474
600	1.64327	1.64327	5600	1.64473	1.64476
1100	1.64403	1.64403	6100	1.64473	1.64477
1600	1.64431	1.64431	6600	1.64473	1.64478
2100	1.64446	1.64446	7100	1.64473	1.64479
2600	1.64455	1.64455	7600	1.64473	1.64480
3100	1.64461	1.64461	8100	1.64473	1.64481
3600	1.64467	1.64466	8600	1.64473	1.64482
4100	1.64473	1.64469	9100	1.64473	1.64482
4600	1.64473	1.64472	9600	1.64473	1.64483

Anzitutto, è possibile notare che

$$\exists \tilde{N} \in \mathbb{N} \quad \text{tale che} \quad S_d(N) = 1.64473 = \text{cost.} \quad \forall N > \tilde{N}$$

I risultati in somma diretta in singola precisione, dunque, a differenza dei risultati in somma inversa, cessano di migliorare in precisione per un valore di  $\tilde{N} \approx 4000$ , stabilizzandosi ad un valore costante. Di seguito è riportato l'andamento qualitativo per punti di  $\Delta(N)$  per un range di  $N$  maggiore.

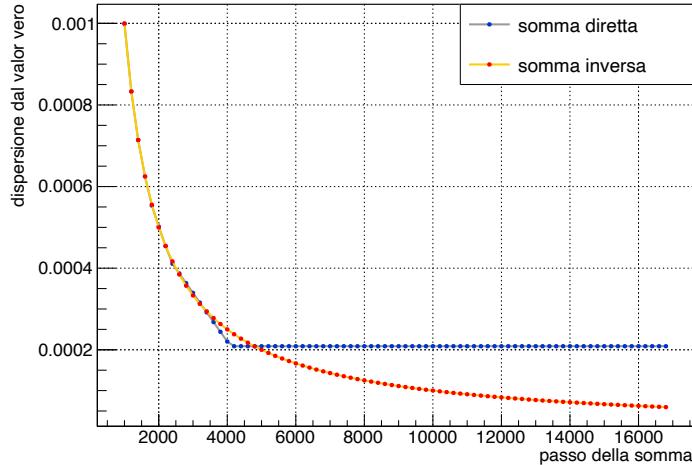


Figura 1: confronto  $\Delta(N)$  in singola precisione

L'andamento di  $\Delta(N)$  evidenzia quanto osservato in precedenza: la condizione di convergenza (1) appare qualitativamente verificata solo per i risultati in somma inversa. In particolare, restringendo il range di  $N$  e campionando ad ogni passo, si verifica che il primo passo della somma a partire dal quale la stima cessa di migliorare in precisione è  $\tilde{N} = 4094$ . Più precisamente, si ha

$$\Delta_d(N) \approx 0.00021 \quad \forall N \geq \tilde{N}$$

Il fenomeno osservato è correlato al fatto che la singola precisione è dell'ordine di  $10^{-8}$ , come si può verificare dalle tabelle introduttive. Dunque, lavorando in precisione singola, i numeri reali sono rappresentati in memoria di lavoro

conservando l'informazione di 7 o 8 cifre significative, a seconda dell'architettura specifica del calcolatore utilizzato. D'altra parte, è immediato notare che

$$\frac{1}{\tilde{N}^2} \approx 10^{-8}$$

Considerando che il termine generale della successione delle somme parziali è una successione decrescente del suo argomento, si avrà quindi

$$\frac{1}{n^2} \leq 10^{-8} \quad \forall n \geq \tilde{N}$$

Ma allora, per valori più grandi di  $\tilde{N}$ , il calcolatore approssimerà (nella somma) il termine generale della successione delle somme parziali a 0, in quanto si sta effettuando una somma di numeri di ordine di grandezza molto diverso, di cui uno è di ordine di grandezza inferiore rispetto all'ordine della singola precisione: a  $N > \tilde{N}$  fissato, le somme successive al passo critico  $\tilde{N}$  saranno, dunque, somme di termini nulli. Più formalmente, si avrà

$$\begin{aligned} \sum_{n=1}^N \frac{1}{n^2} &= \sum_{n=1}^{\tilde{N}} \frac{1}{n^2} + \sum_{n=\tilde{N}+1}^N \frac{1}{n^2} \approx \\ &\approx \sum_{n=1}^{\tilde{N}} \frac{1}{n^2} + 0 = 1.64473 \end{aligned}$$

per il calcolatore. Si noti, infatti, che i valori 1.64473 e  $10^{-8}$  differiscono di ben 8 ordini di grandezza. La somma tra numeri di ordine così diverso porta ad evidenti errori di arrotondamento successivi dati dalla precisione poco elevata che si sta utilizzando, e quindi dalla limitatezza di bit disponibili per la rappresentazione dei numeri reali nella memoria di lavoro del calcolatore. Ciò che accade è quindi una perdita di significatività, nel calcolo della somma, dei termini generali aventi ordine di grandezza più piccolo o uguale della precisione singola di  $10^{-8}$ . Questo problema, invece, non sussiste nel caso delle somme inverse. Sia  $\tilde{N}$  il passo critico per i risultati in somma diretta. L'algoritmo di somma inversa a  $N > \tilde{N}$  fissato consiste nell'eseguire una somma del tipo

$$\sum_{n=N}^1 \frac{1}{n^2} = \frac{1}{N^2} + \dots + \frac{1}{\tilde{N}^2} + \dots + \frac{1}{9} + \frac{1}{4} + 1$$

Come è possibile notare, in questo caso, i termini di ordine minore dell'ordine di grandezza della singola precisione vengono sommati immediatamente. In queste prime somme non avremo problemi di perdita di informazione, in quanto gli ordini di grandezza dei termini non differiscono significativamente tra loro. Mano a mano che le somme vengono eseguite aggiungendo termini sempre maggiori, chiaramente, parte dell'informazione andrà persa per effetto dell'arrotondamento, ma in modo molto meno marcato rispetto alla somma diretta, in quanto i numeri più grandi, la cui informazione tende a dominare, verranno aggiunti alla fine. In somma diretta, dopo sole tre somme, si ha un valore di  $\sim 1.36$ , le cui cifre significative tenderanno presto a dominare sui termini più piccoli che verranno aggiunti. Si noti, inoltre, che dalla figura 1 si ha un altro fatto rilevante: per i primi valori di  $N$ , i due grafici mostrano un andamento

coincidente. La ragione di quanto si osserva è di fatto analoga a quella data per il fenomeno osservato a  $N$  grande. In somma inversa, infatti, non vi sono problemi di perdita di significatività, in quanto inizialmente vengono sommati termini piccoli, ma con ordine di grandezza confrontabile. In somma diretta, invece, seppur vengano sommati inizialmente numeri più grandi, non si verificano errori di arrotondamento sempre per il fatto che gli ordini di grandezza non differiscono ancora in modo significativo per valori di  $N$  sufficientemente piccoli. I fatti esplicitati spiegano sia il diverso andamento asintotico dei due grafici di  $\Delta(N)$  in somma diretta e in somma inversa, sia il medesimo andamento per valori di  $N$  piccoli. Tuttavia, come è possibile notare dal plot, esistono dei valori di  $N$  tali che la dispersione dal valore vero della somma diretta risulti più piccola della dispersione in somma inversa. In particolare, si ha che

$$\Delta_d(N) < \Delta_i(N) \quad \forall N \in \mathbb{N} \quad \text{tale che} \quad 3200 < N < 4800$$

Si sono quindi campionati valori della dispersione in un intorno dell'intervallo in esame, addensando il passo della somma per uno studio più preciso. Il grafico che segue mostra il risultato ottenuto.

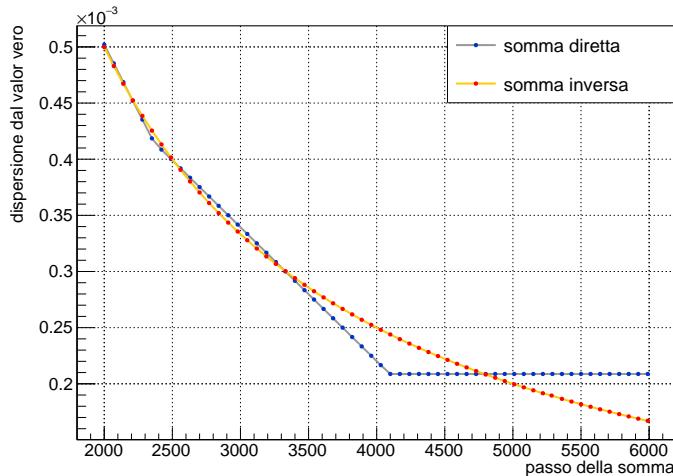


Figura 2: confronto  $\Delta(N)$  in singola precisione, zoom

Osserviamo che il comportamento nell'intervallo in esame è conseguenza di un fenomeno più generale: il grafico di  $\Delta_d(N)$  assume un andamento rettilineo definito a tratti in  $(2000, \tilde{N}) \cap \mathbb{N}$ . In particolare, si ha

$$\Delta_d(N) = \begin{cases} mN + q & \text{se } 2000 < N \leq 2300 \\ rN + p & \text{se } 2300 < N < \tilde{N} \end{cases} \quad \text{con} \quad m < r < 0$$

L'unica condizione che può verificarsi in grado di spiegare questo comportamento consiste nel fatto che, nei due intervalli esplicitati, il calcolatore sommi sempre la medesima quantità  $a = \text{cost.} > 0$ . Infatti, in tal caso avremo

$$S_d(N) = \sum_{n=1}^N \frac{1}{n^2} \approx \sum_{n=1}^N a = aN \quad \text{per} \quad 2000 < N < \tilde{N}$$

da cui segue che la discrepanza si scriverà come

$$\Delta_d(N) = \left| S_d(N) - \frac{\pi^2}{6} \right| \approx \left| aN - \frac{\pi^2}{6} \right|$$

ossia come l'equazione di una retta nell'intervallo in esame. Il fenomeno ha la stessa natura di quello già osservato in precedenza: a causa della diversità di ordine di grandezza delle coppie di numeri che vengono sommate di volta in volta, nei due diversi intervalli indicati, il calcolatore approssimerà di fatto numeri piccoli successivi allo stesso numero piccolo nel calcolo delle somme a causa dell'arrotondamento. La conseguenza è una somma di termini costanti, prima nell'intervallo  $2000 < N \leq 2300$ , poi nell'intervallo  $2300 < N < \tilde{N}$ . In definitiva, per valori del passo  $N$  grandi, la somma inversa fornisce una stima più precisa della serie numerica rispetto alla somma diretta, che risulta invece affetta da problemi di approssimazione successivi dovuti alla somma di numeri di ordine di grandezza molto diverso tra loro.

### Precisione doppia

Si sono stampati a schermo i valori di  $S(N)$  al variare di  $N$  in precisione doppia, prima in somma diretta e poi in somma inversa. Esattamente come in precedenza, si è scelto di considerare  $N \in \mathbb{N}$  tale che

$$100 \leq N < 10000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

convenzionalmente, al fine di svolgere una prima analisi qualitativa dei risultati analizzando un range di valori sufficientemente ampio, ma anche al fine di operare un confronto con i risultati ottenuti in precisione singola a parità di condizioni. La tabella che segue mostra i risultati ottenuti.

$N$	$S_d(N)$	$S_i(N)$	$N$	$S_d(N)$	$S_i(N)$
100	1.63498	1.63498	5100	1.64474	1.64474
600	1.64327	1.64327	5600	1.64476	1.64476
1100	1.64403	1.64403	6100	1.64477	1.64477
1600	1.64431	1.64431	6600	1.64478	1.64478
2100	1.64446	1.64446	7100	1.64479	1.64479
2600	1.64455	1.64455	7600	1.64480	1.64480
3100	1.64461	1.64461	8100	1.64481	1.64481
3600	1.64466	1.64466	8600	1.64482	1.64482
4100	1.64469	1.64469	9100	1.64482	1.64482
4600	1.64472	1.64472	9600	1.64483	1.64483

Anzitutto, è possibile notare che, a differenza dello studio precedente in precisione singola, in questo caso non esiste un valore di  $N$  tale che la somma diretta smetta di migliorare in precisione. Sia in somma diretta, che in somma inversa, i risultati continuano a migliorare asintoticamente, all'aumentare di  $N$ . Inoltre, i due andamenti risultano, dai valori stampati, completamente sovrapponibili. Per verificare tale fatto si sono quindi plottati i valori della dispersione  $\Delta(N)$  sia in somma diretta che in somma inversa. Per la somma diretta si è ottenuto quanto segue.

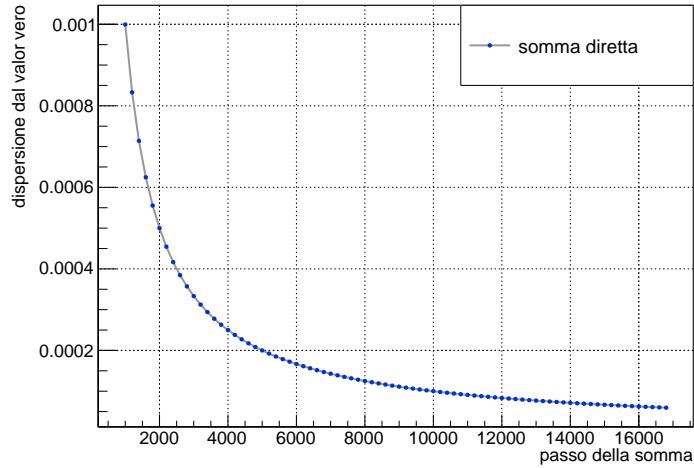


Figura 3:  $\Delta(N)$  in doppia precisione: somma diretta

Per la somma inversa, invece, come ci aspettiamo dai dati raccolti sintetizzati nella tabella precedente, l'andamento è il seguente.

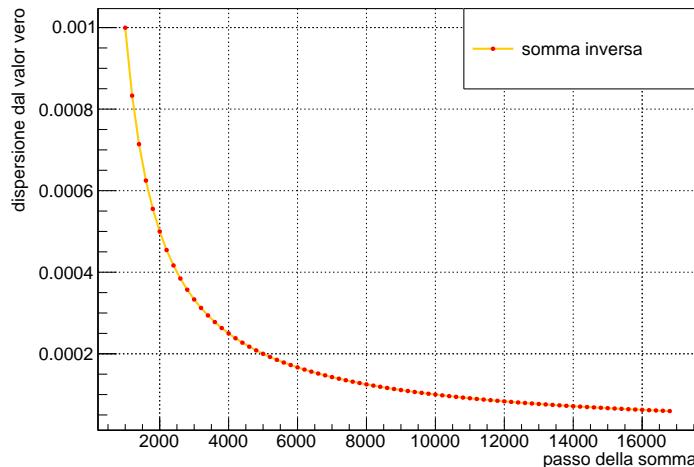


Figura 4:  $\Delta(N)$  in doppia precisione: somma inversa

Nel caso del calcolo di entrambe le somme in doppia precisione, dunque, il valore della successione delle somme parziali tende al valore vero della serie numerica all'aumentare del passo  $N$  della somma, verificando qualitativamente la condizione di convergenza (1). Inoltre, sovrapponendo i grafici in doppia precisione con il grafico in precisione singola in somma inversa ottenuto in precedenza è possibile verificare una completa corrispondenza dei risultati ottenuti. Il diverso risultato ottenuto in somma diretta, in questo caso, è dovuto al fatto che la doppia precisione dedica un numero maggiore di bit in memoria per la mantissa. La conseguenza è che, in doppia precisione, la somma di numeri che differiscono di un ordine di grandezza pari a  $10^{-8}$  verrà eseguita correttamente, senza perdita rilevante di informazione data dall'arrotondamento. Infatti, seppur anche in doppia precisione l'arrotondamento sia necessario per rappresentare un numero

illimitato di cifre, in questo caso i numeri dell'ordine di  $10^{-8}$  non saranno più di fatto termini nulli nelle somme per  $N > \tilde{N}$ , ma saranno, invece, quantità piccole non nulle la cui informazione verrà in gran parte conservata per il range di  $N$  selezionato. In definitiva, la condizione di convergenza (1) risulta verificata per  $S_i(N)$  in singola precisione e per  $S_d(N)$  e  $S_i(N)$  in doppia precisione per le ragioni spiegate. Segue che la miglior stima della serie numerica in esame è data da tali successioni valutate in  $N$  grande. La stima data dalla successione delle somme parziali  $S_d(N)$  in singola precisione risulta, invece, definitivamente distorta. Chiaramente, il fatto di non osservare gli effetti dell'arrotondamento nel caso della somma diretta in precisione doppia è strettamente legato al range di  $N$  in cui si valuta la successione delle somme parziali. Gli effetti della cattiva positura dell'algoritmo in somma diretta, infatti, sono osservabili anche in precisione doppia, a partire da valori di  $N$  tali che

$$\frac{1}{N^2} \approx 10^{-16} \quad \iff \quad N \approx 10^8$$

per ragioni analoghe a quelle già discusse. A partire da questo valore, ci si aspetta che i risultati in somma diretta cessino di migliorare in precisione per ragioni di arrotondamento. Ci si aspetta, inoltre, che comportamenti inaspettati come somme di quantità costanti possano verificarsi anche prima del raggiungimento del valore critico determinato.

In definitiva, possiamo concludere che l'algoritmo di calcolo delle somme sia instabile per il problema numerico di stima diretta ( $1 \rightarrow N$ ) di alcune serie numeriche, soprattutto per valori di  $N$  grandi. Generalizzando in modo poco formale, possiamo affermare che tale algoritmo presenta problemi dati da errori di arrotondamento per la stima di tutte quelle serie convergenti il cui termine generale è monotono decrescente, in quanto in tutti questi casi il calcolo in somma diretta porterà a sommare per primi i termini più grandi, le cui cifre significative domineranno sulle cifre dei termini sommati in seguito. Evidentemente, questo costituisce un problema non trascurabile, in quanto le serie convergenti presentano spesso termini generali decrescenti. Come si ha avuto modo di osservare, si sono trovati due modi diversi per aggirare il problema, non necessariamente esclusivi tra loro:

- riformulare il problema numerico in modo matematicamente equivalente con la somma inversa (equivalente per proprietà associativa della somma), ma tale che la nuova formulazione non presentasse il calcolo di numeri di ordine di grandezza maggiore all'inizio della somma
- aumentare la precisione con cui rappresentare i numeri reali, in modo che anche in somma diretta la quantità di informazione immagazzinata fosse tale da evitare arrotondamenti troppo precoci nella serie decimale

Tuttavia, come si ha avuto modo di notare, lavorare in precisione doppia non risolve il problema, ma lo rimanda soltanto a valori del passo maggiori. Gli errori di arrotondamento derivanti dall'algoritmo di calcolo delle somme dirette sono strutturali e seguono direttamente dalla modalità con la quale il calcolatore rappresenta in memoria i numeri reali. All'aumento della precisione, ove possibile, è sempre preferibile affiancare una riformulazione matematica equivalente del problema numerico. Vista in altro modo, il problema affrontato permette

di fare un'altra considerazione rilevante. Sia  $\oplus$  l'operazione di somma tra numeri reali in aritmetica finita (ossia al calcolatore). Non è difficile intuire che la proprietà commutativa tipica della somma è conservata, ossia vale sempre

$$x \oplus y = y \oplus x$$

per ogni coppia di reali  $x, y$  rappresentati in virgola mobile. Ciò che l'esercizio permette di verificare è il fatto che la proprietà associativa per la somma, invece, in generale viene meno in aritmetica finita. In altre parole

$$x \oplus (y \oplus z) \neq (x \oplus y) \oplus z$$

per alcune terne  $x, y, z$  rappresentate in virgola mobile. Si ha già avuto modo di spiegare che la ragione di questo fatto consiste negli inevitabili errori di approssimazione dei numeri reali in aritmetica finita, che può produrre instabilità in alcuni algoritmi risolutivi apparentemente innocui, se applicati a problemi con caratteristiche precise. Quanto discusso pone l'accento sul fatto che sia sempre bene, prima di risolvere numericamente un problema, interrogarsi sulla formulazione ottimale del problema stesso.

## Esercizio 2

Si vuole stimare numericamente la potenza  $n$ -esima  $\phi_1^n$ , dove  $\phi_1 := \frac{\sqrt{5}-1}{2}$  e  $n \in \mathbb{N}$ . In particolare, si vuole utilizzare la formula ricorsiva  $\chi$  data da

$$\begin{cases} \chi_0 = 1 \\ \chi_1 = \phi_1 \\ \chi_{n+1} = \chi_{n-1} - \chi_n \end{cases} \quad (2)$$

Il numero reale  $\phi_1$ , infatti, soddisfa la relazione di ricorrenza

$$\phi^{n+1} = \phi^{n-1} - \phi^n \quad (3)$$

Portando i termini ad uno dei due membri e raccogliendo si ha

$$\begin{aligned} \phi^{n-1} (\phi^2 + \phi - 1) &= 0 \iff \\ \phi = 0 \quad \vee \quad \phi = \phi_1 \quad \vee \quad \phi = \phi_2 := -\frac{\sqrt{5} + 1}{2} \end{aligned}$$

Sviluppando la (2) avremo allora

$$\begin{aligned} \chi_0 &= 1 = \phi_1^0 \\ \chi_1 &= \phi_1 = \phi_1^1 \\ \chi_2 &= 1 - \phi_1 = \phi_1^2 \\ \chi_3 &= \phi_1 - \phi_1^2 = \phi_1^3 \\ &\vdots \\ \chi_n &= \phi_1^n \end{aligned}$$

dove la seconda uguaglianza è stata ottenuta, in tutti i casi non banali, sostituendo opportunamente la (3) espressa in forme diverse. Si è allora ottenuto che la formula ricorsiva  $\chi$  permette di calcolare il valore di una qualunque potenza  $n$ -esima di  $\phi_1$ .

L'algoritmo definito dalla (2) risulta quindi, in astratto, perfettamente funzionale per la risoluzione del problema numerico in esame. Tuttavia, si noti che  $\phi_1$  non rappresenta l'unica soluzione della (3): anche  $\phi_2$  rappresenta una soluzione non nulla della formula di ricorrenza. A causa dell'inevitabile errore di arrotondamento dato dalla rappresentazione in virgola mobile si avrà, dunque, una contaminazione della seconda radice nella rappresentazione decimale della prima, ossia il calcolatore rappresenterà

$$\tilde{\chi}_0 = 1 \quad \text{e} \quad \tilde{\chi}_1 = \phi_1 + \varepsilon \phi_2$$

dove  $\varepsilon$  denota la precisione con cui si è deciso di rappresentare i numeri reali. Sviluppando di nuovo la (2) alla luce della contaminazione avremo

$$\begin{aligned} \tilde{\chi}_2 &= 1 - \phi_1 - \varepsilon \phi_2 = \phi_1^2 + \varepsilon \phi_2^2 - \varepsilon \\ \tilde{\chi}_3 &= \phi_1 + \varepsilon \phi_2 - \phi_1^2 - \varepsilon \phi_2^2 + \varepsilon = \phi_1^3 + \varepsilon \phi_2^3 + \varepsilon \\ &\vdots \\ \tilde{\chi}_n &\sim \phi_1^n + \varepsilon \phi_2^n \end{aligned}$$

trascurando i termini di ordine inferiore. Si noti ora che possiamo scrivere

$$|\phi_1^n| = e^{n \log |\phi_1|}$$

e chiaramente si ha

$$0 < |\phi_1| < 1 \implies \log |\phi_1| < 0$$

ossia  $|\phi_1^n|$  si presenta nella forma di un esponenziale decrescente del suo argomento. Con passaggi del tutto analoghi avremo poi

$$|\phi_2^n| = e^{n \log |\phi_2|} \quad (4)$$

per la quale, invece, varrà

$$|\phi_2| > 1 \implies \log |\phi_2| > 0$$

ossia  $|\phi_2^n|$  è un esponenziale crescente del suo argomento. Segue quindi che la dispersione della stima  $\tilde{\chi}_n$  dal valore vero  $\phi_1^n$  si scriverà come

$$\Delta(n) := |\tilde{\chi}_n - \phi_1^n| \sim \varepsilon |\phi_2^n|$$

e tenendo conto della (4) si otterrà infine

$$\Delta(n) \sim \varepsilon e^{n \log |\phi_2|} \quad \text{con} \quad \log |\phi_2| > 0 \quad (5)$$

L'andamento della dispersione dal valore vero che ci si aspetta di osservare sarà quindi un andamento esponenziale crescente, ossia ci si aspetta che l'applicazione dell'algoritmo dato dalla formula ricorsiva  $\chi$  porti ad un aumento dell'errore esponenziale al crescere della potenza  $n$ -esima calcolata. L'unico modo per controllare l'errore consisterebbe, allora, nell'operare sul coefficiente  $\varepsilon$ , ossia l'unico parametro non fissato della (4). Nei fatti, questo equivale a dire che l'unico modo per controllare la crescita inevitabile dell'errore numerico consiste nell'aumentare la precisione con cui si opera al calcolatore.

### Verifica dell'instabilità

Si è quindi verificata l'instabilità dell'algoritmo per il problema in esame verificando l'andamento esponenziale dato dalla (5) in tre diversi casi: lavorando in precisione singola, doppia e doppia estesa. Il significato di quest'ultimo termine verrà chiarito più avanti nella trattazione. A tal proposito, si è fissato un valore massimo  $N = 70$  di potenza  $n$ -esima, per poi calcolare

$$\tilde{\phi}_1^n = \tilde{\chi}_n \quad \forall n = 0, \dots, N$$

stimando quindi la potenza per mezzo della formula ricorsiva  $\chi$ . Si è poi calcolato, per ogni  $n$ , quello che si è assunto essere il valore vero (o quantomeno una stima non distorta) della potenza  $n$ -esima grazie ad una funzione dedicata, disponibile nelle librerie matematiche standard. L'implementazione della funzione di calcolo di una qualunque potenza  $n$ -esima può variare a seconda del compilatore e del sistema operativo. Un approccio comune è quello di calcolare

$$a^b = e^{b \log a} \quad \text{con} \quad a > 0 \quad \text{e} \quad b \in \mathbb{R}$$

e quindi di ricondurre il problema del calcolo di una potenza reale alla valutazione della composizione di due funzioni note in un punto. Ovviamente, al calcolatore, le valutazioni di funzioni elementari in punti presentano anch'esse errori di approssimazione dati, per lo più, dal troncamento ad un ordine finito del loro sviluppo in serie di Taylor: il modo in cui vengono spesso rappresentate in un calcolatore funzioni reali di variabile reale. Ad ogni modo, il problema in esame consiste nel calcolo di potenze naturali: l'approccio più semplice consiste nel prodotto del numero in esame con se stesso  $n$  volte. Anche in questo caso, chiaramente, verranno commessi errori di arrotondamento, ma al netto degli errori di calcolo inevitabili risulta comunque ragionevole assumere tale stima come il valore vero con il quale eseguire il confronto. Si sono quindi calcolate le dispersioni dal valore vero come

$$\Delta(n) = |\tilde{\chi}_n - \phi_1^n| \quad \forall n = 0, \dots, N$$

Questi passaggi sono stati eseguiti per tutte le tre diverse precisioni di rappresentazione di numeri reali selezionate. In particolare, al fine di svolgere una verifica quantitativa della (5), si sono plottati e interpolati i dati raccolti  $(n, \Delta(n))$  con una generica funzione esponenziale della forma  $f(n) = ae^{bn}$ , con  $a$  e  $b$  parametri liberi. Per la precisione singola si sono ottenuti i seguenti risultati.

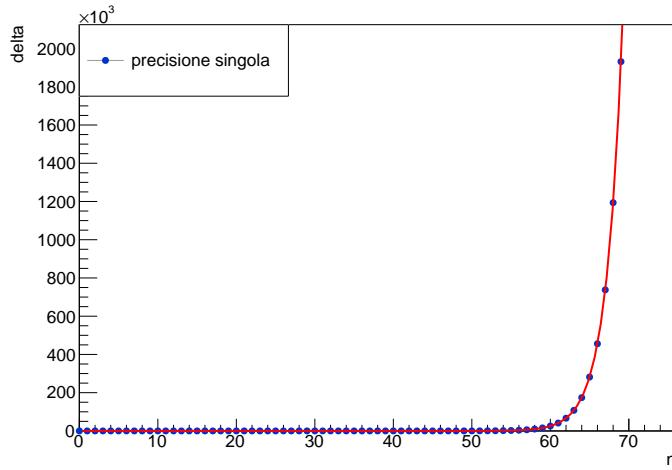


Figura 5:  $(n, \Delta(n))$  in precisione singola: fit

I parametri stimati risultano

$$a = 7.34 \cdot 10^{-9} \quad \text{e} \quad b = \log |\phi_2| = 0.481$$

Tenendo conto del fatto che vale  $\log |\phi_2| \approx 0.481211$ , risulta possibile concludere la consistenza tra il parametro  $b$  stimato e il valore atteso. Inoltre, si noti che  $a \sim \varepsilon$ . In precisione singola, infatti, si ha  $\varepsilon \approx 10^{-8}$ . La lieve discordanza di una potenza del 10 dal parametro atteso è dovuta al fatto che la relazione (5) definisce solo un andamento, ossia vale in modo approssimato avendo trascurato i termini di ordine inferiore. Dai parametri stimati è quindi possibile concludere la verifica della (5) in singola precisione. Si noti che gli algoritmi di minimizzazione per lo svolgimento di fit operano con meno difficoltà in presenza di forme

analitiche della funzione interpolante meno complicate. Una delle forme analitiche più facili da gestire, anche con un grande numero di dati, è la relazione lineare  $f(x) = a + bx$ . Per tale ragione, ove possibile, è sempre bene operare manipolazioni algebriche sulla funzione interpolante (e quindi sui dati) al fine di ricondursi sempre ad un caso in cui la probabilità di fallimento del processo di minimizzazione sia minore. Nel caso in esame è possibile notare che

$$\Delta(n) \sim \varepsilon e^{n \log |\phi_2|} \iff \log \Delta(n) \sim \log \varepsilon + n \log |\phi_2|$$

ossia il calcolo del logaritmo naturale permette di trasformare la forma esponenziale in una forma affine. Si sono quindi calcolate le coppie  $(n, \log \Delta(n))$  per ogni  $n$ , per poi plottare e interpolare i dati secondo la relazione lineare  $y = q + mn$ , ottenendo quanto segue.

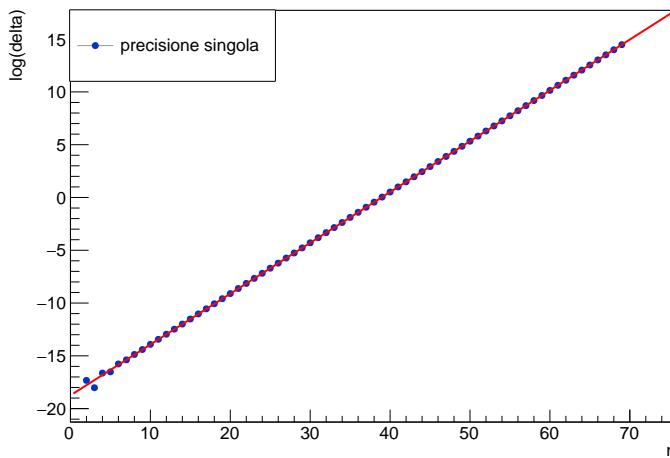


Figura 6:  $(n, \log \Delta(n))$  in precisione singola: fit

I parametri stimati risultano

$$q = \log \varepsilon = -18.7 \quad \text{e} \quad m = \log |\phi_2| = 0.482$$

Si nota facilmente che i valori ottenuti risultano del tutto sovrapponibili a quelli stimati con il fit esponenziale. Tuttavia, in questo caso è stato necessario rimuovere i primi due punti, in quanto il calcolo dei logaritmi ha prodotto infiniti in loro corrispondenza. Il fit lineare, inoltre, mostra una discordanza visiva dalla previsione dei 4 punti successivi (con potenze minori), che il fit esponenziale non mostra. Tutti questi risultati si spiegano con il fatto che la contaminazione della seconda radice  $\phi_2$  nella soluzione si osserva all'aumentare di  $n$ , ossia con il procedere delle iterazioni date dalla formula ricorsiva  $\chi$ . Il fit lineare permette dunque di concludere che la contaminazione della seconda radice in precisione singola si apprezza in modo significativo e secondo le previsioni a partire dalla potenza naturale  $n = 6$ . Si sono quindi ripetute le medesime procedure in precisione doppia, calcolando i punti  $(n, \Delta(n))$  per ogni  $n$  fino a  $N$ . Si sono poi interpolati i dati secondo la medesima funzione esponenziale utilizzata in precedenza, ottenendo quanto segue.

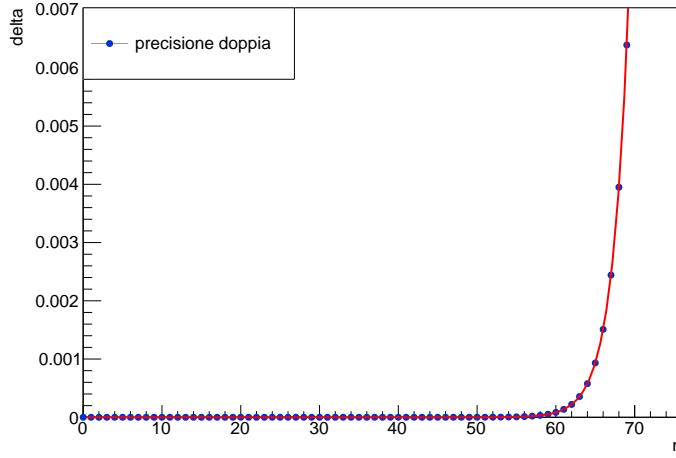


Figura 7:  $(n, \Delta(n))$  in precisione doppia: fit

I parametri stimati risultano

$$a = 2.4 \cdot 10^{-17} \quad \text{e} \quad b = \log |\phi_2| = 0.481$$

Tenendo conto del fatto che, in questo caso, lavorando in precisione doppia si ha  $\varepsilon \approx 10^{-16}$  e vista la consistenza del parametro  $b$ , risulta possibile concludere la complessiva concordanza dei dati raccolti con la relazione (5) in doppia precisione. Si sono quindi calcolati, anche in questo caso, i logaritmi delle dispersioni al fine di eseguire un fit lineare estraendo più informazioni dai dati come in precedenza. Di seguito sono riportati i risultati ottenuti.

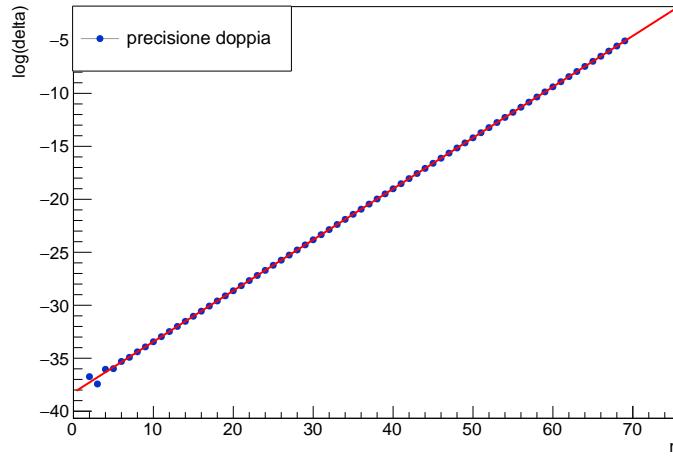


Figura 8:  $(n, \log \Delta(n))$  in precisione doppia: fit

I parametri stimati risultano

$$q = \log \varepsilon = -38.2 \quad \text{e} \quad m = \log |\phi_2| = 0.481$$

Le stime dei parametri sono dunque compatibili con la relazione (5). Anche in questo caso è stato necessario rimuovere i primi due punti, in quanto i logaritmi delle loro dispersioni hanno prodotto infiniti. Inoltre, anche in precisione doppia, si nota che i successivi 4 punti si discostano significativamente dalla funzione attesa, per le ragioni già discusse. Il primo valore di potenza tale che il contributo di contaminazione di  $\phi_2$  possa produrre un andamento consistente con la (5) è  $n = 6$ . Il calcolatore e il programma utilizzati per la stesura dei codici non hanno permesso direttamente l'utilizzo di tipi di variabili reali in precisione quadrupla. Tuttavia, hanno permesso l'utilizzo di un tipo di variabile reale contenente più informazione rispetto alla precisione doppia. Svolgendo diversi test di stampa e utilizzando una funzione apposita si ha avuto modo di verificare che la variabile *long double* utilizzata permette la corretta rappresentazione di 19 cifre decimali, rispetto alle 16 cifre della precisione doppia. Da qui in avanti, per semplicità di notazione, chiameremo questo nuovo tipo di precisione con il termine *precisione doppia estesa*. Il non utilizzo della precisione quadrupla non altera in alcun modo il fine del presente studio: i passaggi svolti in questa sezione possono essere ripetuti con qualunque precisione, anche quadrupla. Lo scopo della trattazione è proprio quello di verificare come l'esplosione esponenziale dell'errore sia del tutto indipendente dalla precisione selezionata, a meno di variazioni dei parametri della funzione esponenziale. Si sono allora ripetuti i medesimi passaggi in precisione doppia estesa, calcolando i punti  $(n, \Delta(n))$  per ogni  $n$  fino a  $N$ , per poi interpolare i dati con una generica funzione esponenziale della stessa forma delle precedenti. Di seguito sono riportati i risultati ottenuti.

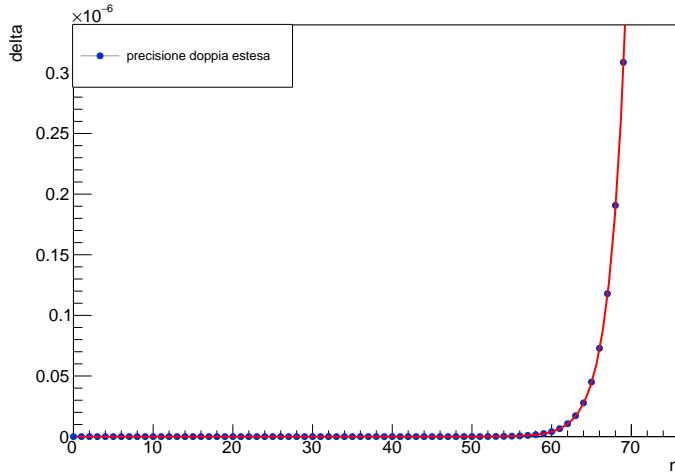


Figura 9:  $(n, \Delta(n))$  in precisione doppia estesa: fit

I parametri stimati risultano

$$a = 1.17 \cdot 10^{-21} \quad \text{e} \quad b = \log |\phi_2| = 0.481$$

Dalla compatibilità dei valori ottenuti con quelli attesi segue la verifica della relazione (5) in precisione doppia estesa. Si sono quindi calcolati, anche in questo caso, i logaritmi delle dispersioni al fine di eseguire un fit lineare ed estrarre più informazioni dai dati come in precedenza, ottenendo i seguenti risultati.

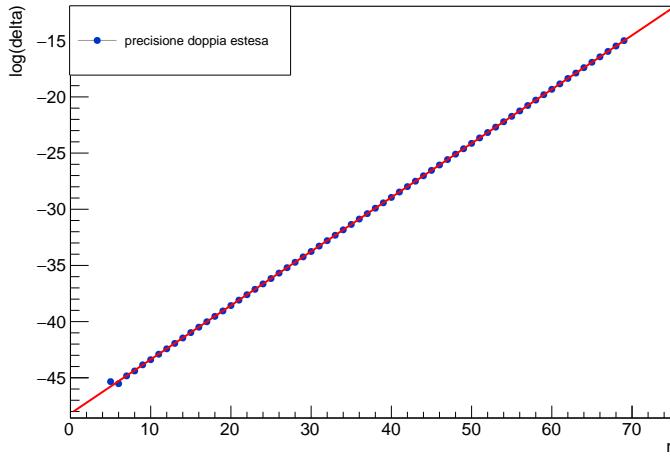


Figura 10:  $(n, \log \Delta(n))$  in precisione doppia estesa: fit

I parametri stimati risultano

$$q = \log \varepsilon = -48.2 \quad \text{e} \quad m = \log |\phi_2| = 0.481$$

evidentemente compatibili con la (5). Si noti che, in questo caso, è stato necessario rimuovere i primi 5 punti, in quanto il calcolo dei logaritmi ha prodotto infiniti in loro corrispondenza. A differenza dei casi in precisione singola e doppia, dunque, la precisione doppia estesa consente di tenere una stima sufficientemente precisa della potenza di  $\phi_1$  per valori di  $n$  fino a 5, producendo quindi uno scarto nullo dal valor vero. Come è possibile notare dal fit lineare, anche i successivi 2 punti risultano particolarmente distanti dalla funzione attesa rispetto ai rimanenti. Per considerazioni analoghe a quelle precedenti è allora possibile concludere che il primo valore di potenza tale che il contributo di contaminazione di  $\phi_2$  possa generare un andamento consistente con la (5) è  $n = 7$ . Idealmente, ci si aspetta che in precisione quadrupla questo valore possa aumentare per le ragioni già discusse.

Per tutti i valori di precisione  $\varepsilon$  selezionati si ha quindi avuto modo di verificare che l'errore dato dall'utilizzo dell'algoritmo  $\chi$  assume un andamento esponenziale all'aumentare della potenza  $n$  di  $\phi_1$  secondo la relazione (5). Questo fatto non dipende dalla struttura del problema in esame, ma dall'utilizzo dell'algoritmo ricorsivo (2) applicato al problema numerico di calcolo di una potenza  $n$ -esima. Il fatto che il problema non sia mal condizionato è evidente dal calcolo della potenza con l'utilizzo della funzione dedicata che si è assunta come valore vero. D'altra parte, se non fosse stato possibile calcolare  $\phi_1^n$  con errore trascurabile, non sarebbe nemmeno stato possibile il calcolo di  $\Delta(n)$ . Come si può intuire osservando le scale verticali dei grafici dell'andamento di  $\Delta(n)$  nei tre casi, il solo modo per diminuire l'errore di calcolo risulta essere quello di aumentare la precisione  $\varepsilon$  con cui si rappresentano i numeri reali. Mettendo a confronto i risultati in singola e doppia precisione, ad esempio, si ha il grafico che segue.

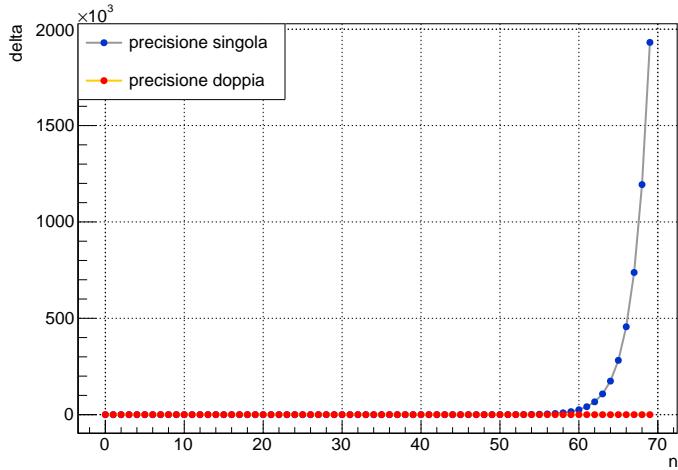


Figura 11: confronto  $\Delta(n)$  in precisione singola e doppia

L'esplosione esponenziale dell'errore di calcolo in precisione doppia appare, dunque, trascurabile rispetto all'esplosione in precisione singola. Similmente, l'esplosione in precisione doppia estesa apparirà trascurabile rispetto a quella in precisione doppia. Nonostante questo, in tutti i casi, l'errore sulla stima della potenza  $n$ -esima data da  $\chi$  avrà sempre un inevitabile andamento esponenziale dato dalla (5). Per tutte queste ragioni, dall'analisi del problema è possibile concludere che l'algoritmo (2) è instabile per il problema numerico di calcolo di una potenza  $n$ -esima di  $\phi_1$ .

### Simulazione della contaminazione

I risultati ottenuti fino a questo punto confermano l'instabilità dell'algoritmo numerico  $\chi$  per il problema in esame. La relazione (5), che rende conto quantitativamente dell'instabilità, è stata ricavata notando che, vista l'esistenza di una seconda radice non nulla nella (3), il calcolatore, per effetto dell'arrotondamento, rappresenterà in memoria la variabile reale  $\chi_1$  come

$$\tilde{\chi}_1 = \phi_1 + \varepsilon \phi_2 \quad (6)$$

dove  $\varepsilon$  rappresenta la precisione con cui si sta operando. Si vuole quindi verificare direttamente che il calcolatore rappresenti la variabile  $\chi_1$  come combinazione lineare delle due radici  $\phi_1$  e  $\phi_2$ . Per farlo, risulta possibile "simulare" questa prima contaminazione che genera la deviazione a catena dal valor vero nelle iterazioni successive. In particolare, basterà fissare una precisione  $\varepsilon$  corrispondente alla generazione di uno dei comportamenti studiati in precedenza, per poi rappresentare tutte le variabili reali con una precisione molto maggiore rispetto alla precisione  $\varepsilon$  fissata. Dichiarando manualmente la variabile (6) sarà allora possibile simulare il comportamento del calcolatore, visto che la precisione molto maggiore delle variabili consentirà di lavorare in un ambiente dove gli errori di calcolo saranno trascurabili rispetto agli errori indotti dalla contaminazione artificiale di  $\phi_2$  al primo step. Si è quindi fissato un valore di  $\varepsilon = 10^{-8}$  corrispondente alla precisione singola, dichiarando tutte le variabili reali e gli oggetti

dedicati alla loro manipolazione in precisione doppia estesa, molto maggiore della singola. Si è poi modificato l'algoritmo assegnando i valori iniziali

$$\tilde{\chi}_0 = 1 \quad \text{e} \quad \tilde{\chi}_1 = \phi_1 + \varepsilon\phi_2$$

Ci si aspetta allora che questo processo di simulazione possa portare ad ottenere risultati del tutto confrontabili rispetto ai risultati ottenuti nella sezione precedente in precisione singola. Al fine di verificarlo si è fissato, anche questa volta, un valore massimo di potenza pari a  $N = 70$ . Si sono poi calcolate e plottate le coppie  $(n, \tilde{\chi}_n)$  per ogni  $n$  fino a  $N$ , ottenendo il seguente andamento.

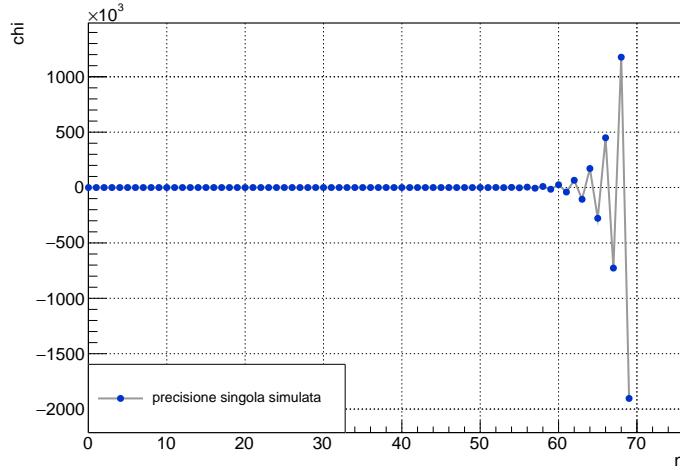


Figura 12:  $(n, \tilde{\chi}_n)$  in precisione singola simulata

Come si nota, l'andamento di  $\tilde{\chi}_n$  appare definitivamente oscillante in  $n$ . Questo accade in quanto, al passo  $n$ , sappiamo valere

$$\tilde{\chi}_n \sim \phi_1^n + \varepsilon\phi_2^n$$

D'altra parte, si ha  $\phi_1 > 0$  e  $\phi_2 < 0$ , da cui segue che

$$\phi_1^n > 0 \quad \forall n \in \mathbb{N}$$

Invece, per il secondo termine avremo

$$\begin{cases} \phi_2^n > 0 & \text{se } n \text{ pari} \\ \phi_2^n < 0 & \text{se } n \text{ dispari} \end{cases}$$

Si è già osservato in precedenza che  $0 < |\phi_1| < 1$  e  $|\phi_2| > 1$ . Ma allora, quando il contributo di  $\phi_2^n$  inizierà ad essere rilevante in modulo, la somma dei due termini in  $\tilde{\chi}_n$  inizierà a produrre valori negativi per  $n$  dispari, positivi per  $n$  pari. Questo primo fatto qualitativo dà le prime indicazioni circa l'instabilità dell'algoritmo: ai fini del problema numerico in esame, siccome ogni potenza di  $\phi_1$  è positiva, l'esistenza di stime negative è un risultato privo di significato. La funzione  $\tilde{\chi}_n$ , per le ragioni appena discusse, è allora una funzione oscillante definitivamente, ossia a partire da un certo valore di  $n$ . Stampando a schermo

i valori di  $\tilde{\chi}_n$  o ingrandendo il plot è possibile osservare quanto ci si aspetta: la stima decresce fino ad un certo valore di  $n$ , ossia fintanto che il contributo di  $\phi_1^n$  prevale sul secondo termine, ma inizia ad oscillare dal momento in cui il secondo termine dato dalla contaminazione assume lo stesso ordine di grandezza del primo. Risulta quindi interessante trovare il valore di  $n = \bar{n}$  tale che

$$\tilde{\chi}_n = (-1)^n |\tilde{\chi}_n| \quad \forall n > \bar{n}$$

ossia tale che, da quel punto in poi, la funzione inizi ad oscillare. A tale scopo si noti che, in corrispondenza di  $n = \bar{n}$ , la funzione discreta  $|\tilde{\chi}_n|$  presenterà un minimo per le ragioni discusse. Tenendo conto del fatto che il logaritmo naturale è una funzione monotona crescente del suo argomento è evidente che il comportamento della funzione  $|\tilde{\chi}_n|$  sarà il medesimo del comportamento di

$$\log |\tilde{\chi}_n| \sim \log |\phi_1^n + \varepsilon \phi_2^n|$$

Per ragioni di ottimizzazione di visualizzazione dei dati, si sono allora calcolate e plottate le coppie  $(n, \log |\tilde{\chi}_n|)$  per ogni  $n$  fino a  $N$ , ottenendo quanto segue.

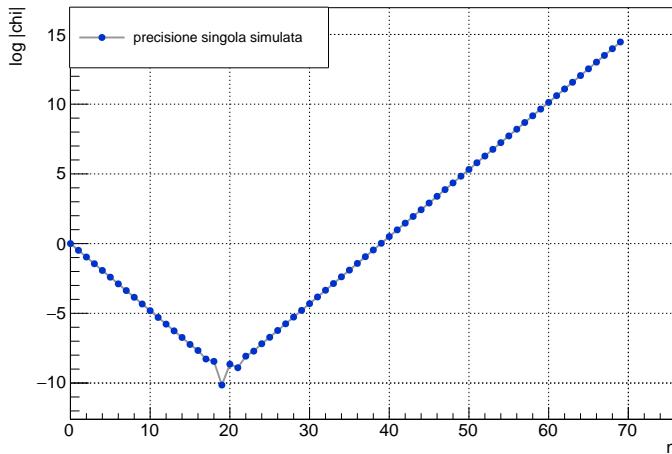


Figura 13:  $(n, \log |\tilde{\chi}_n|)$  in precisione singola simulata

Come si nota dalla figura, il punto di minimo della funzione discreta  $\log |\tilde{\chi}_n|$ , che coincide con il punto di minimo di  $|\tilde{\chi}_n|$ , è il valore di potenza  $\bar{n} = 19$ . In effetti, con un semplice conto è immediato verificare che

$$|\phi_1^{\bar{n}}| \approx 0.00011 \quad \text{e} \quad \varepsilon |\phi_2^{\bar{n}}| \approx 0.00009$$

ossia i due termini di  $\tilde{\chi}_n$  assumono, in modulo, un ordine di grandezza del tutto confrontabile. A partire da  $\bar{n} + 1$  in poi, la stima  $\tilde{\chi}_n$  assumerà allora un andamento oscillante, dato dalla prevalenza del secondo termine in modulo rispetto al primo. Chiaramente, le considerazioni fatte fino a questo punto, derivanti da uno studio più dettagliato dell'andamento della stima  $\tilde{\chi}_n$ , possono essere replicate in modo del tutto analogo anche negli studi non simulati dell'algoritmo del paragrafo precedente, ottenendo andamenti qualitativi del tutto analoghi a quelli appena ottenuti. Al fine di verificare quantitativamente la corrispondenza

dei risultati ottenuti in precisione singola simulata rispetto a quelli ottenuti in precisione singola non simulata, si sono plottate le coppie  $(n, \Delta(n))$ , per poi interpolare i dati secondo  $f(n) = ae^{bn}$ , ottenendo quanto segue.

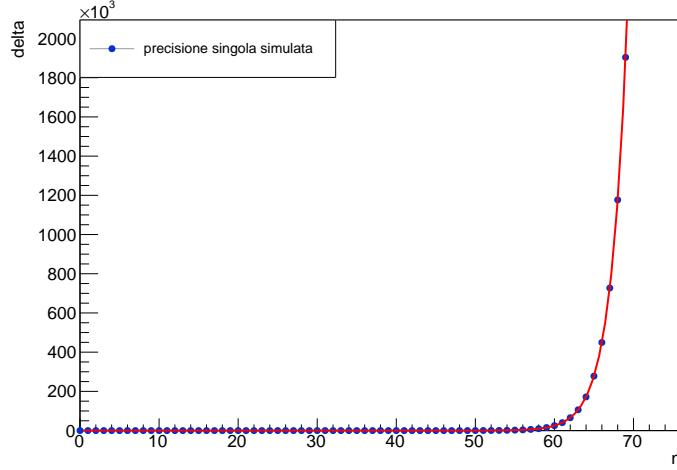


Figura 14:  $(n, \Delta(n))$  in precisione singola simulata: fit

I parametri stimati risultano

$$a = 7.24 \cdot 10^{-9} \quad \text{e} \quad b = \log |\phi_2| = 0.481$$

Ricordando il valore di  $\varepsilon$  fissato manualmente, entrambe le stime risultano compatibili con la relazione (5). Inoltre, i risultati ottenuti sono completamente sovrapponibili alle stime ricavate in precisione singola nel paragrafo precedente. Anche in questo caso, si sono calcolati i logaritmi delle dispersioni, operando poi un fit lineare della forma  $y = q + mn$ , ottenendo quanto segue.

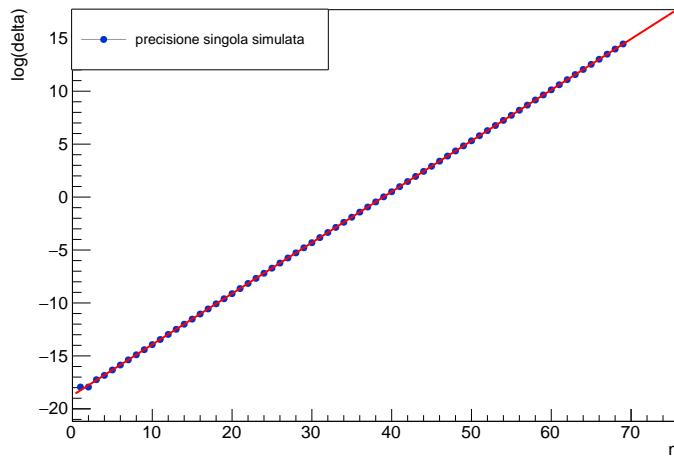


Figura 15:  $(n, \log \Delta(n))$  in precisione singola simulata: fit

I parametri stimati risultano

$$q = \log \varepsilon = -18.7 \quad \text{e} \quad m = \log |\phi_2| = 0.481$$

I valori ottenuti, anche qui, sono del tutto confrontabili con le stime prodotte dal programma di minimizzazione nel caso dello studio dell'errore in precisione singola non simulata. Tuttavia, come si può notare da un rapido confronto tra la figura 15 e la figura 6, nel caso della simulazione è stato necessario rimuovere il solo punto in corrispondenza di  $n = 0$  in quanto, in sua corrispondenza, il calcolo del logaritmo ha prodotto un infinito. Inoltre, nella simulazione, l'andamento di  $\log \Delta(n)$  sembra convergere all'andamento rettilineo atteso più rapidamente rispetto al caso non simulato: a partire da  $n = 3$ , rispetto a  $n = 6$  del caso reale. Ovviamente, questa piccola discordanza tra la simulazione e il caso non simulato si spiega ricordando l'artificiosità dell'operazione svolta in questa sezione, tenendo anche conto del fatto che, pur allocando le variabili con precisione molto maggiore rispetto alla singola, la simulazione non può essere svolta, nei fatti, con precisione infinita. Anche in questo caso avremo allora errori di arrotondamento successivi dati dall'aritmetica finita del calcolatore. Dalla corrispondenza del modello con i dati raccolti in questa simulazione è allora possibile concludere il fatto che, come ipotizzato, il calcolatore rappresenti effettivamente in memoria la variabile reale  $\tilde{\chi}_1$  secondo l'assegnazione (6).

In definitiva, l'algoritmo dato dalla (2) è instabile per il problema numerico di calcolo di una potenza  $n$ -esima di  $\phi_1$  per definizione 0.4, visto l'andamento esponenziale dell'errore. L'algoritmo inizia a produrre stime completamente insensate, come si ha avuto modo di verificare, a partire da valori di potenza molto piccoli. In un ipotetico calcolatore perfetto, in grado di rappresentare numeri reali con infinite cifre decimali, l'algoritmo  $\chi$  rappresenterebbe una strada del tutto sensata per la risoluzione del problema, ma l'aritmetica finita di macchina fa in modo che l'unica strada percorribile per l'utilizzo di questo algoritmo nel problema in esame sia quella di lavorare con il massimo della precisione possibile.

## Integrazione deterministica

Siamo interessati al calcolo di un integrale definito della forma

$$I := \int_a^b f(x) dx$$

con  $a < \infty$  e  $b < \infty$ , ossia alla stima numerica di un integrale di Riemann unidimensionale su un intervallo di integrazione limitato. Un modo possibile per affrontare il problema consiste nell'utilizzo dei *metodi deterministicici* di integrazione. Il nome è evocativo e deriva dal fatto che il loro funzionamento si basa su una precisa partizione dell'intervallo di integrazione  $[a, b]$ , in contrapposizione alle tecniche che sfruttano le sequenze casuali, come si avrà modo di studiare più avanti. In particolare, in questa sezione verranno analizzati e discussi due approcci deterministicici strutturalmente diversi tra loro:

- metodi di Newton-Cotes
- metodo di quadratura gaussiana

I primi approcciano il problema calcolando  $f$  in diversi punti, tutti equidistanti all'interno dell'intervallo di integrazione  $[a, b]$ . Il metodo di quadratura gaussiana, invece, consiste nell'utilizzo di una serie di nodi e pesi specifici, selezionati in modo da garantire, sotto certe condizioni, una stima precisa anche con un numero relativamente basso di valutazioni della funzione  $f$ . Si noti che considerare un intervallo di integrazione limitato non porta a perdite di generalità, in quanto i metodi che si analizzeranno per integrali definiti possono essere utilizzati anche per il calcolo di integrali impropri, facendo uso di opportuni cambi di variabile, al fine di mappare intervalli illimitati in opportuni intervalli limitati. Evidentemente, gli integrali definiti su un intervallo limitato a seguito di un cambio di variabile conterranno punti singolari all'interno del dominio di integrazione. Vedremo che, questo caso, sarà gestibile combinando i metodi standard con le formule aperte di integrazione, facenti parte dei metodi di Newton-Cotes. Sia dunque  $I$  l'integrale che si vuole stimare e  $\tilde{I}$  una qualunque stima numerica dell'integrale in esame. Anche in questo caso, chiameremo dispersione dal valore vero la quantità

$$\Delta(N) := |\tilde{I}(N) - I|$$

Chiaramente, la dispersione dipenderà dal numero  $N$  di punti utilizzati per il calcolo dell'integrale. Affinché  $\tilde{I}$  possa rappresentare una buona stima numerica di  $I$  dovrà valere la solita condizione di convergenza

$$\lim_{N \rightarrow +\infty} \Delta(N) = 0 \tag{7}$$

Più precisamente, vedremo che  $\Delta(N)$  dipende dal numero  $N$  di punti in cui si divide l'intervallo di integrazione nel caso dei metodi di Newton-Cotes. Dipende, invece, dal numero  $N$  di zeri dei polinomi ortogonali considerati nel caso del metodo di quadratura gaussiana. Discutiamo, quindi, i metodi di integrazione che si sono utilizzati nella risoluzione degli esercizi, al fine di dare alcune considerazioni generali che saranno poi richiamate in seguito.

## Metodi di Newton-Cotes

Sia  $N$  il numero di punti in cui si vuole calcolare  $f$  all'interno dell'intervallo di integrazione  $[a, b] = [x_1, x_N]$ . Dall'equidistanza dei punti segue che l'ampiezza di ogni sotto-intervallo in cui verrà diviso  $[x_1, x_N]$  sarà

$$h := \frac{b - a}{N - 1}$$

In tal modo, saremo in grado di esprimere l'accuratezza della stima in funzione del numero di punti, ossia di descrivere l'evoluzione di  $\Delta$  in funzione di  $N$ .

### 1) Metodo del trapezio

Il metodo del trapezio esteso consiste nel sommare aree di trapezi successivi che vengono costruiti su ogni sotto-intervallo. La formula estesa ha la forma

$$\int_{x_1}^{x_N} f(x) dx = h \left[ \frac{1}{2}f_1 + f_2 + \dots + f_{N-1} + \frac{1}{2}f_N \right] + O\left(\frac{1}{N^2}\right)$$

da cui segue direttamente che

$$\Delta_T(N) = O\left(\frac{1}{N^2}\right) \iff \lim_{N \rightarrow +\infty} N^2 \Delta_T(N) = k \in \mathbb{R}$$

Si avrà, quindi, che

$$\Delta_T(N) \approx k \frac{1}{N^2} \quad \text{con } N \text{ grande} \quad (8)$$

Al fine di osservare più facilmente l'andamento dei risultati è possibile plottare la dispersione dal valore vero in funzione di  $n := \frac{1}{N}$ . Posto questo cambio di variabile, si ha un ramo di parabola centrato nell'origine della forma

$$\Delta_T(n) \approx k n^2 \quad (9)$$

La relazione è ulteriormente meglio visualizzabile calcolando il logaritmo della dispersione dal valore vero. In questo modo, si ottiene una retta della forma

$$\log \Delta_T(n) \approx \log k + 2 \log n \quad (10)$$

La (10) risulta utile anche per effettuare fit di dati raccolti al calcolatore. Le rette, infatti, rispetto ad altre forme funzionali, risultano più facilmente trattabili numericamente dai programmi di minimizzazione: un chiaro esempio sono gli algoritmi di interpolazione dei dati.

### 2) Metodo di Simpson

Il metodo di Simpson esteso consiste nel sommare aree di figure piane non poligonali successive che vengono costruite su ogni sotto-intervallo. In questo metodo, la funzione integranda  $f$  viene approssimata, su ogni sotto-intervallo, ad archi di parabola, ossia ad una funzione quadratica. In particolare, si ha

$$\int_{x_1}^{x_N} f(x) dx = \frac{h}{3} [f_1 + 4f_2 + 2f_3 + \dots + 2f_{N-2} + 4f_{N-1} + f_N] + O\left(\frac{1}{N^4}\right)$$

Si noti che, a differenza del metodo del trapezio, il metodo di Simpson fornisce una stima consistente solo nel caso di un numero  $N$  dispari di punti. Per un numero  $N$  pari di punti, infatti, non verrebbe sommato l'addendo  $4f_{N-1}$ , determinando una sottostima dell'integrale in esame. Dalla formula estesa segue direttamente che

$$\Delta_S(N) = O\left(\frac{1}{N^4}\right) \iff \lim_{N \rightarrow +\infty} N^4 \Delta_S(N) = k \in \mathbb{R}$$

Si avrà, quindi, la relazione asintotica

$$\Delta_S(N) \approx k \frac{1}{N^4} \quad \text{con } N \text{ grande} \quad (11)$$

Anche in questo caso, è possibile plottare la dispersione in funzione di  $n := \frac{1}{N}$  al fine di visualizzare meglio i dati raccolti. Si avrà, quindi, la quartica

$$\Delta_S(n) \approx k n^4 \quad (12)$$

Calcolando i logaritmi, questa volta si otterrà una relazione lineare della forma

$$\log \Delta_S(n) \approx \log k + 4 \log n \quad (13)$$

che consentirà di visualizzare meglio gli andamenti ed eseguire fit più facilmente.

### 3) Metodo di Romberg

La formula del trapezio estesa può essere implementata in modo ricorsivo aumentando esponenzialmente il numero dei punti in cui si divide l'intervallo di integrazione, senza ricominciare il calcolo ogni volta. Sia dunque  $J \geq 1$ . Vogliamo dividere l'intervallo di integrazione in  $2^J$  sotto-intervalli. Si avrà allora

$$h = \frac{b-a}{2^J} \quad \text{con } N = 2^J + 1$$

Indicheremo la stima numerica ricorsiva con il metodo del trapezio esteso con  $T(J)$ . Evidentemente, infatti, tale stima, dipendendo dal numero di punti  $N$ , dipenderà anche da  $J$  con le definizioni poste. Il metodo di Romberg consiste nel generalizzare la formula ricorsiva del trapezio al fine di aumentarne l'accuratezza numerica. In particolare, indicheremo con  $R$  una stima numerica ottenuta con il metodo di Romberg. Sia dunque

$$R(J, 0) := T(J)$$

Posto il passo zero, il metodo di Romberg è definito in modo ricorsivo secondo

$$R(J, K) := \frac{4^K R(J, K-1) - R(J-1, K-1)}{4^K - 1}$$

Dalla formula è evidente che il modo più naturale di rappresentare i risultati del metodo di Romberg sia mediante la matrice  $R^*$  di  $J+1$  righe e  $K+1$  colonne

$$R^* := \begin{pmatrix} R(0, 0) & 0 & 0 & \dots & 0 \\ R(1, 0) & R(1, 1) & 0 & \dots & 0 \\ R(2, 0) & R(2, 1) & R(2, 2) & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ R(J, 0) & R(J, 1) & R(J, 2) & \dots & R(J, K) \end{pmatrix} \in \text{Mat}_{J+1, K+1}(\mathbb{R})$$

Si dimostra che, nel metodo di Romberg, la dispersione si scrive come

$$\Delta_{R_K}(N) = O\left(\frac{1}{N^{2K+2}}\right) \iff \lim_{N \rightarrow +\infty} N^{2K+2} \Delta_{R_K}(N) = k \in \mathbb{R}$$

da cui segue che

$$\Delta_{R_K}(N) \approx k \frac{1}{N^{2K+2}} \quad \text{con } N \text{ grande} \quad (14)$$

Scrivendo la dispersione in funzione di  $n := \frac{1}{N}$  come nei casi precedenti, si ha

$$\Delta_{R_K}(n) \approx k n^{2K+2} \quad (15)$$

Calcolando i logaritmi si ottiene la solita relazione lineare della forma

$$\log \Delta_{R_K}(n) \approx \log k + (2K+2) \log n \quad (16)$$

Si noti che, per costruzione, le colonne corrispondenti a  $K = 0$  e a  $K = 1$  coincidono, rispettivamente, con i risultati del metodo del trapezio esteso e con quelli del metodo di Simpson esteso. Segue che, negli studi degli esercizi successivi, i casi rilevanti da studiare saranno quelli per  $K \geq 2$ . Chiaramente, considerando che la precisione aumenta di molto all'aumentare di  $K$  e all'aumentare di  $N$ , la miglior stima di  $I$  utilizzando questo metodo si ottiene dall'elemento  $(J+1, K+1)$ -esimo della matrice triangolare inferiore  $R^*$ .

#### 4) Formule aperte

Le formule aperte di integrazione sono formule approssimate utilizzate per stimare il valore di  $I$  sull'intervallo aperto  $(a, b)$ . Queste formule possono essere utili nel caso si voglia calcolare il valore di un integrale nell'intorno di un punto nel quale la funzione integranda presenta una singolarità. I metodi discussi fino a qui, infatti, prevedono sempre di valutare  $f$  negli estremi dell'intervallo di integrazione, non permettendo la corretta gestione di questi casi particolari. Se  $N - 2$  è il numero di punti in cui si vuole calcolare la funzione all'interno dell'intervallo  $(a, b)$ , vale sempre la definizione del passo

$$h := \frac{b - a}{N - 1}$$

L'aspetto negativo da tenere presente consiste nel fatto che tali formule non sono facilmente generalizzabili per costruire formule estese come nei casi precedenti e, inoltre, possiedono un'accuratezza limitata. Per tale ragione, analizziamo solo due casi particolari.

$N = 3$

Si supponga di dividere l'intervallo di integrazione in 2 sotto-intervalli. Si considerino, quindi,  $N = 3$  punti tali che  $x_0 = a$  e  $x_2 = b$ . Sotto queste ipotesi vale la formula aperta

$$\int_{x_0}^{x_2} f(x) dx = 2hf(x_1) + O(h^3)$$

Come è possibile notare, la funzione non viene calcolata negli estremi dell'intervallo di integrazione.

$$N = 6$$

Si supponga di dividere l'intervallo di integrazione in 5 sotto-intervalli. Si considerino, quindi,  $N = 6$  punti tali che  $x_0 = a$  e  $x_5 = b$ . Sotto queste ipotesi vale la formula aperta

$$\int_{x_0}^{x_5} f(x) dx = \frac{h}{24} [55f(x_1) + 5f(x_2) + 5f(x_3) + 55f(x_4)] + O(h^5)$$

Anche in questo caso, la funzione non viene mai calcolata negli estremi di integrazione, come ci aspettiamo da una formula aperta ben posta.

### Metodo di quadratura gaussiana

Siamo interessati al calcolo di un integrale definito di Riemann del tipo

$$\bar{I} := \int_a^b W(x)f(x) dx \quad (17)$$

Il metodo di quadratura di Gauss fornisce la seguente approssimazione

$$\bar{I} \approx \sum_{j=1}^N w_j f(x_j) \quad (18)$$

dove  $W$  è detta funzione peso,  $w_j$  sono coefficienti indipendenti dalla funzione  $f$  e  $x_j$  sono gli  $N$  zeri di un set di polinomi ortonormali di grado  $N$ . Inoltre, l'intervallo  $(a, b)$  è l'intervallo in cui sono contenuti gli zeri dei polinomi considerati. L'aspetto cruciale di questo metodo consiste nel fatto che è possibile dimostrare che l'approssimazione di quadratura gaussiana è esatta per polinomi  $f$  di grado  $2N - 1$ . Chiaramente, dato un set di polinomi ortonormali rispetto alla funzione peso  $W$ , se si è interessati al calcolo di un integrale del tipo

$$I := \int_a^b g(x) dx$$

utilizzando il metodo delle quadrature gaussiane, sarà sufficiente notare che

$$I = \int_a^b W(x) \frac{g(x)}{W(x)} dx \quad \text{con} \quad W \neq 0$$

da cui segue, dalla (18), che

$$I \approx \sum_{j=1}^N w_j \frac{g}{W}(x_j) \quad (19)$$

In tal modo è possibile ricondurre ogni integrale definito che si vuole calcolare ad un integrale della forma di  $\bar{I}$  così come definito dal metodo di Gauss anche se, come si avrà modo di verificare, il metodo di Gauss risulta particolarmente

efficiente quando l'integrandi si presenta già nella forma (17). Si supponga, dunque, di essere interessati alla stima numerica di  $I$  utilizzando il metodo delle quadrature gaussiane, con  $a < \infty$  e  $b < \infty$ . In altre parole, si supponga che  $I$  sia un integrale definito e non un integrale improprio. Esistono diversi polinomi ortonormali utilizzabili per la stima numerica dell'integrale in esame con il metodo di Gauss. Ognuno di questi ammette  $N$  zeri in un certo intervallo di  $\mathbb{R}$ , limitato o illimitato. Segue che, se  $U \subset \mathbb{R}$  è l'intervallo in cui sono presenti gli zeri dei polinomi, per il calcolo di  $I$  sarà necessario ricondurre l'integrale ad un integrale equivalente definito in  $U$ , dove è possibile applicare il metodo di Gauss utilizzando i pesi e gli zeri tabulati dei vari polinomi ortonormali. In altre parole, risulta necessario trovare una mappa biettiva

$$\Phi : (a, b) \rightarrow U$$

ossia un opportuno cambio di variabile per l'integrale, al fine di garantire il corretto funzionamento dato dall'algoritmo di Gauss. Determiniamo quindi le trasformazioni in questo caso particolare per alcuni polinomi utilizzati.

### 1) Polinomi di Legendre

Si supponga che l'intervallo in cui sono presenti gli zeri dei polinomi ortonormali considerati sia un chiuso e limitato della forma

$$U = [c, d] \subset \mathbb{R}$$

con  $c \neq a$  e  $d \neq b$ , come nel caso dei polinomi di Legendre. In tal caso, risulta necessario trovare una mappa  $\Phi$  che mandi  $x \in (a, b)$  nell'intervallo  $[c, d]$ . In particolare, si è supposto che tale mappa fosse affine, ossia della forma

$$y = mx + q \quad \text{con} \quad m, q \in \mathbb{R}$$

Si è quindi imposto il passaggio della funzione per  $c$  e per  $d$  nei punti in cui, rispettivamente,  $x = a$  e  $x = b$ , ottenendo il seguente sistema

$$\begin{cases} c = ma + q \\ d = mb + q \end{cases}$$

Sottraendo le due equazioni si ha

$$m = \frac{c - d}{a - b}$$

Sostituendo  $m$  ricavato in una delle due equazioni si ha

$$q = \frac{ad - cb}{a - b}$$

Ma allora, la mappa cercata è ora completamente determinata e ha la forma

$$\Phi : (a, b) \rightarrow [c, d]$$

$$x \mapsto y = \frac{c - d}{a - b}x + \frac{ad - cb}{a - b}$$

Abbiamo quindi ottenuto un cambio di coordinate  $\Phi$  che permette di trasformare qualunque integrale unidimensionale definito in  $(a, b)$  in un integrale equivalente definito in  $[c, d]$ . In particolare, è immediato notare che la variabile di integrazione  $x$  in funzione della nuova variabile  $y$  si scriverà come

$$x = \frac{a-b}{c-d} \left( y - \frac{ad-cb}{a-b} \right)$$

da cui segue che il determinante della matrice jacobiana, che nel caso unidimensionale si determina differenziando  $\Phi$ , sarà dato da

$$dx = \frac{a-b}{c-d} dy$$

Ma allora, se  $[c, d]$  è l'intervallo in cui sono presenti gli  $N$  zeri dei polinomi ortonormali, si ha l'integrale equivalente

$$I = \int_c^d g \left[ \frac{a-b}{c-d} \left( y - \frac{ad-cb}{a-b} \right) \right] \frac{a-b}{c-d} dy \quad (20)$$

I polinomi di Legendre sono definiti nell'intervallo  $[-1, 1]$ . Segue che, per la stima numerica di  $I$ , varrà la (20) con

$$c = -1 \quad \text{e} \quad d = 1$$

Si noti, inoltre, che i polinomi di Legendre sono ortogonali rispetto alla funzione

$$W(x) = 1$$

Segue che, in questo caso particolare, la (18) e la (19) coincidono. Per il calcolo di  $I$ , dunque, non sarà necessario combinare il cambio di variabile con la presenza della funzione peso: basterà implementare direttamente la (20).

## 2) Polinomi di Laguerre

Esistono polinomi ortonormali tali che l'intervallo in cui sono presenti gli zeri sia un aperto di  $\mathbb{R}$  semi-illimitato, come nel caso dei polinomi di Laguerre. In tal caso, il cambio di coordinate operato precedentemente non vale più, ed è quindi necessario determinare un'altra trasformazione o un altro metodo per ricondursi all'intervallo desiderato. I polinomi di Laguerre sono definiti nell'intervallo

$$U = (0, +\infty) \subset \mathbb{R}$$

Si considerino, dunque, gli integrali impropri

$$I_a := \int_a^{+\infty} g(x) dx \quad \text{e} \quad I_b := \int_b^{+\infty} g(x) dx$$

Per la proprietà di additività dell'integrale vale

$$I = I_a - I_b \quad (21)$$

Tuttavia, si noti che, posto che l'integrale  $I$  che vogliamo stimare converga, la (21) vale solo se è verificata la convergenza di  $I_a$  e  $I_b$ , ossia se vale

$$I_a < \infty \quad \text{e} \quad I_b < \infty \quad (22)$$

Se così non fosse, otterremmo infiniti o forme di indecisione del tipo  $\infty - \infty$  e la (21) perderebbe di significato. Condizione necessaria affinchè valga la (22) è

$$\lim_{x \rightarrow +\infty} g(x) = 0 \quad (23)$$

Seppur la (23) sia solo una condizione necessaria, questa ci permette di escludere la funzionalità del metodo nel caso in cui la funzione integranda diverga all'infinito. A questo punto, sotto l'ipotesi esplicitata, per l'utilizzo dei polinomi di Laguerre è sufficiente ricondurre gli integrali  $I_a$  e  $I_b$  ad integrali sull'intervallo aperto  $(0, +\infty)$ . La trasformazione  $\Phi$  in questo caso risulta quindi essere una banale traslazione. Si consideri, infatti, un integrale improprio della forma

$$I^* := \int_{\gamma}^{+\infty} g(x) dx \quad \text{con} \quad \gamma \in \mathbb{R}$$

Si consideri, quindi, il cambio di variabile affine

$$y = x - \gamma$$

Notiamo che per questa trasformazione vale

$$\lim_{x \rightarrow \gamma} y(x) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} y(x) = +\infty$$

ossia la mappa presenta le caratteristiche cercate per ricondursi all'intervallo  $U$ . La trasformazione ha, dunque, la forma

$$\Phi : (\gamma, +\infty) \rightarrow (0, +\infty)$$

$$x \mapsto y = x - \gamma$$

Valgono allora le relazioni

$$x = y + \gamma \quad \text{e} \quad dx = dy$$

Applicando il cambio di variabile si avrà quindi

$$I^* = \int_0^{+\infty} g(y + \gamma) dy \quad (24)$$

In definitiva, combinando la (21) e la (24), per l'utilizzo del metodo di Gauss con i polinomi di Laguerre si dovrà calcolare

$$\begin{aligned} I &= \int_0^{+\infty} g(x + a) dx - \int_0^{+\infty} g(x + b) dx = \\ &= \int_0^{+\infty} [g(x + a) - g(x + b)] dx \end{aligned} \quad (25)$$

Si noti che, in questo caso, al posto di cercare una trasformazione generale  $\Phi$ , si sono utilizzate le proprietà dell'integrale di Riemann per semplificare i conti, a patto di aggiungere l'ipotesi di convergenza (22). L'implementazione di un cambio di variabile generale, in questo caso, infatti, necessiterebbe di un numero elevato di calcoli successivi. Questo fatto porterebbe alla possibilità di incorrere in errori di arrotondamento, che trascinati nelle successive operazioni al

calcolatore rischierebbero di generare una perdita significativa di informazione, minando all'ottenimento di una buona stima di  $I$ . Negli esercizi, i casi che non verificheranno l'ipotesi (22) verranno analizzati e discussi di volta in volta. Si noti che i polinomi di Laguerre per  $\alpha = 0$  sono ortogonali rispetto alla funzione

$$W(x) = e^{-x}$$

Segue che, in questo caso, l'implementazione del metodo dovrà combinare la (25) e la (19) per garantire il corretto funzionamento. L'utilizzo dei polinomi di Laguerre risulta quindi particolarmente utile e vantaggioso quando la funzione integranda assume la forma

$$g(x) = e^{-x} f(x)$$

in quanto consente la possibilità di ottenere una stima più precisa e, eventualmente, esatta quando  $f$  è un polinomio di grado  $2N - 1$ .

### 3) Polinomi di Hermite

I polinomi di Hermite sono definiti nell'intervallo

$$U = (-\infty, +\infty) = \mathbb{R}$$

In questo caso, allora, determinare trasformazioni generali risulta particolarmente svantaggioso: sarà più comodo specializzare la trasformazione di volta in volta sfruttando eventuali simmetrie della funzione integranda. Per amore di completezza, notiamo soltanto che una strada possibile consiste nel mandare

$$(a, b) \rightarrow \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

per mezzo di una dilatazione combinata ad una traslazione, per poi mandare

$$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \rightarrow (-\infty, +\infty)$$

sfruttando la mappa  $y = \tan(x)$ . Non è difficile ricavare che l'integrale, a seguito di questi due cambi di variabile, risulta particolarmente complicato a livello di operazioni successive che devono essere eseguite al calcolatore, da cui segue la poca efficienza di questa strada. I polinomi di Hermite sono ortogonali rispetto alla funzione peso

$$W(x) = e^{-x^2}$$

da cui segue che il loro utilizzo nel metodo di Gauss sarà particolarmente utile in presenza di funzioni integrande del tipo

$$g(x) = e^{-x^2} f(x)$$

permettendo eventualmente l'esattezza per polinomi  $f$  di grado  $2N - 1$ .

Quanto ottenuto vale solo per integrali definiti: nei casi in cui l'intervallo di integrazione risulterà illimitato verrà cercato un opportuno cambio di variabile a seconda dell'integrale in esame.

### Esercizio 3

Si vuole stimare numericamente il valore dell'integrale

$$I_3 := \int_0^5 x^7 e^{-x} dx$$

utilizzando i metodi di Newton-Cotes e il metodo di quadratura gaussiana, al fine di verificare la loro efficienza e la loro stabilità nel problema in esame.

Analiticamente (integrando 7 volte per parti) o con l'aiuto di un calcolatore avanzato, si ricava che il valore esatto è

$$I_3 = 5040 - \frac{648240}{e^5} \approx 672.193237312836809276648$$

Siamo ora in grado di calcolare le dispersioni delle stime dal valore vero.

#### Trapezio

Si è calcolato  $I_3$  utilizzando il metodo del trapezio esteso, variando il numero di punti  $N$  in input. Ci si aspetta che, in generale, la stima migliori all'aumentare del numero di punti in cui si divide l'intervallo di integrazione. Equivalentemente, ci si aspetta che l'errore decresca al crescere di  $N$  secondo la (8). In particolare, al fine di visualizzare meglio i dati raccolti, si è deciso di plottare i risultati secondo la relazione equivalente (9). Si sono quindi calcolate e plottate le dispersioni al variare di  $1/N$  nel range

$$300 \leq N < 1500 \quad \text{con} \quad N_{i+1} = N_i + 50$$

ottenendo il grafico che segue.

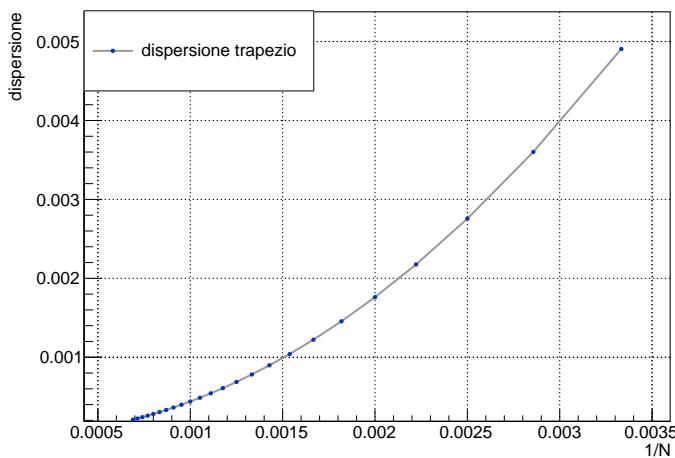


Figura 16: dispersione metodo del trapezio

Come è possibile notare, l'andamento qualitativo è parabolico, come ci si aspetta per valori di  $N$  grandi, ossia in regime asintotico. Si sono quindi interpolati i

dati raccolti al fine di verificare la bontà dell'adattamento della curva ai dati. In particolare, si è deciso di svolgere il fit passando ai logaritmi, provando quindi a verificare la relazione equivalente (10). Il passaggio ai logaritmi permette, infatti, di interpolare i dati utilizzando una più semplice funzione lineare della forma  $y = p + mx$ , facilitando l'algoritmo di minimizzazione utilizzato. Di seguito sono riportati i risultati ottenuti.

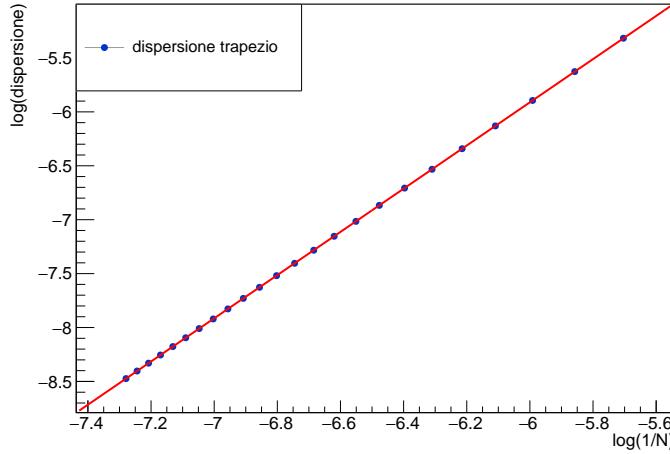


Figura 17: log dispersione metodo del trapezio: fit

I parametri stimati risultano

$$p = 6.11 = \log k \quad \text{e} \quad m = 2$$

Come è possibile notare, la stima del parametro  $m$  è consistente con il coefficiente angolare della relazione trovata analiticamente. Questo fatto garantisce la bontà del fit effettuato, rendendo possibile concludere che, nella formula di integrazione del trapezio esteso, l'errore scali come  $\frac{1}{N^2}$  per  $N$  grande, come ci si aspetta. Si noti che la verifica delle leggi che governano l'errore al variare di  $N$  implicano anche la verifica della convergenza della stima al valore vero. In particolare, il grafico 16 rappresenta una buona verifica visiva della condizione di convergenza (7), ricordando che sull'asse verticale è rappresentato l'inverso del numero di punti.

### Simpson

Si è poi calcolato  $I_3$  utilizzando il metodo di Simpson esteso, variando il numero di punti in input. In questo caso, siamo allora interessati a verificare che valga la relazione (11). Per ragioni di comodità di visualizzazione dei dati, anche in questo caso, si è deciso di provare a visualizzare la relazione equivalente (12). Si sono quindi calcolate e plottate le dispersioni al variare di  $1/N$  nel range

$$501 \leq N < 5000 \quad \text{con} \quad N_{i+1} = N_i + 100$$

selezionando solo gli  $N$  dispari, al fine di garantire il corretto funzionamento dell'algoritmo dato dal metodo di Simpson, ottenendo quanto segue.

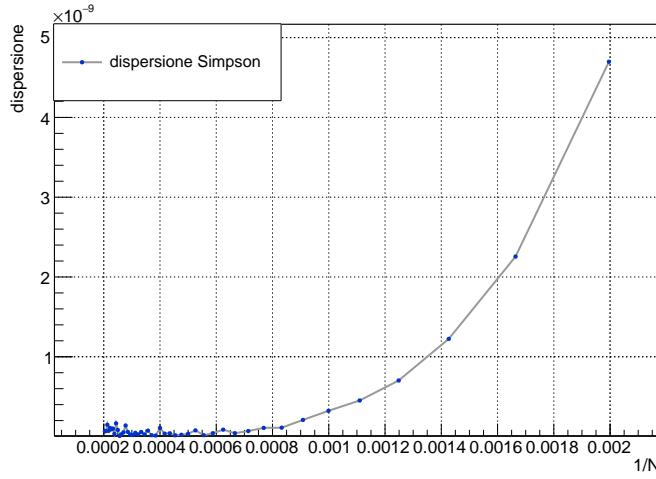


Figura 18: dispersione metodo di Simpson

Come è possibile notare dal grafico, l'andamento non è quello di una quartica come ci si aspetta. In particolare, il grafico presenta un andamento inaspettato per valori di  $\frac{1}{N}$  piccoli. Si sono quindi provati a piazzare i logaritmi come in precedenza, al fine di verificare qualitativamente l'andamento lineare dato dalla (13). Il grafico dei punti è il seguente.

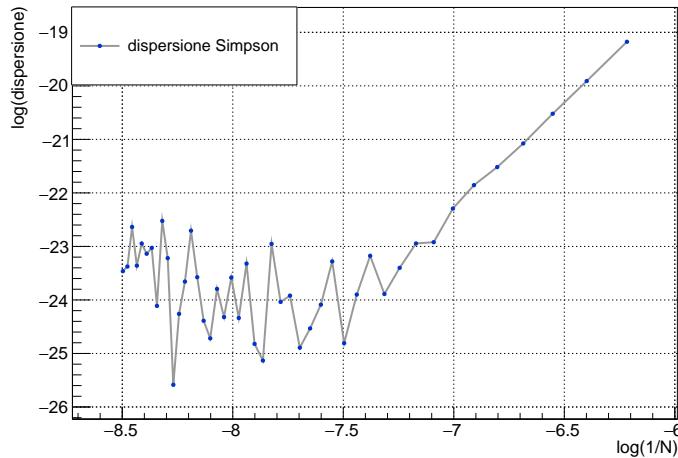


Figura 19: log dispersione metodo di Simpson

Risulta immediato osservare che l'andamento rettilineo è verificato solo per valori sufficientemente piccoli di  $N$ . In particolare,  $\exists \tilde{N} \in \mathbb{N}$  tale che la dispersione appaia una funzione oscillante secondo una legge non banale  $\forall N > \tilde{N}$ . La ragione di questo comportamento inatteso è dato dal problema di perdita di significatività generato dalla differenza di numeri vicini tra loro. Infatti il passo del campionamento e visualizzando il plot risultante è possibile dedurre

qualitativamente che la funzione cessa di avere un comportamento rettilineo per

$$\Delta_S < \tilde{\Delta}_S \quad \text{tale che} \quad -21 < \log \tilde{\Delta}_S < -22$$

Segue che il valore critico della dispersione è

$$\tilde{\Delta}_S \approx 10^{-10}$$

Il calcolo della dispersione, dunque, a partire da valori di  $N$  corrispondenti a  $\tilde{\Delta}_S$ , appare come differenza di numeri molto vicini tra loro. Ciò che accade è quindi una perdita di informazione, per valori di  $N > \tilde{N}$ , a causa dell'effetto dell'arrotondamento combinato alla limitatezza della precisione doppia utilizzata per il calcolo. Il motivo per il quale, utilizzando il metodo del Trapezio, non si è potuto osservare questo fenomeno è dovuto al fatto che, in quel caso, la dispersione scala molto più lentamente rispetto alla dispersione del metodo di Simpson. Per tale ragione, lavorando a parità di precisione, il valore di  $\tilde{N}$  a partire dal quale si osserverà il fenomeno, nel metodo del trapezio, sarà molto più elevato del valore di  $\tilde{N}$  nel metodo di Simpson. A dimostrazione di questo fatto, è immediato verificare che, lavorando in precisione doppia estesa, il valore critico  $\tilde{N}$  aumenta di molto, in quanto lo spazio di memoria allocato per il salvataggio di un numero reale diventa maggiore, aumentando di conseguenza il numero di cifre immagazzinate. Alla luce di quanto detto, al fine di verificare la proporzionalità tra la dispersione e  $\frac{1}{N^4}$  è necessario considerare un range di valori sufficientemente grande da lavorare in regime asintotico, ma sufficientemente piccolo da evitare i problemi di differenza di numeri troppo vicini tra loro. Si è quindi scelto il range di valori

$$250 \leq N < 550 \quad \text{con} \quad N_{i+1} = N_i + 5$$

selezionando solo gli  $N$  dispari. Si è poi plottata la deviazione in funzione di  $1/N$ , ottenendo quanto segue.

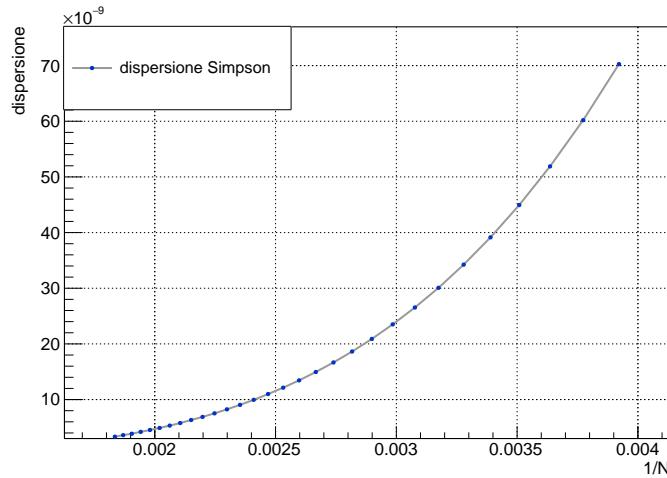


Figura 20: dispersione metodo di Simpson per  $N < \tilde{N}$

Quella ottenuta risulta, ragionevolmente, una quartica in  $n$ . Al fine di verificarlo quantitativamente si sono calcolati, come nel caso precedente, i logaritmi delle

dispersioni in funzione dei logaritmi dell'inverso del numero di punti. Si sono quindi fittati i dati con una funzione lineare della forma  $y = p + mx$ , ottenendo quando segue.

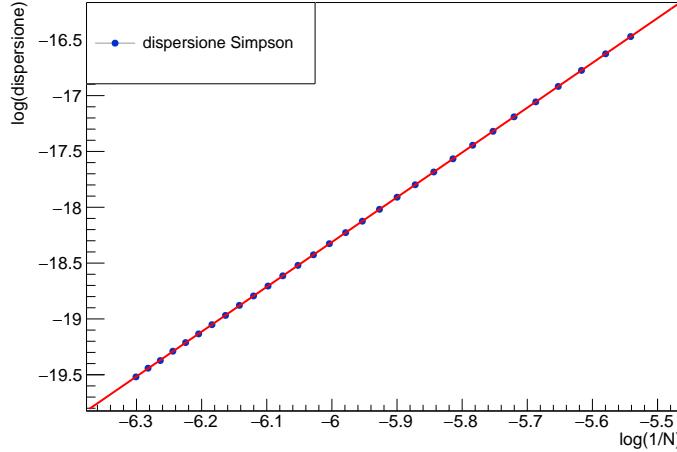


Figura 21: log dispersione metodo di Simpson per  $N < \tilde{N}$ : fit

I parametri stimati risultano

$$p = 5.76 = \log k \quad \text{e} \quad m = 4.01 \approx 4$$

Vista l'accordanza tra la stima di  $m$  e il parametro noto, la legge (13) che governa la dispersione in funzione del numero di punti può dirsi, dunque, verificata. Si può ipotizzare che la non esattezza della stima di  $m$  ottenuta sia dovuta al fatto di aver considerato un range di  $N$  ristretto. D'altra parte, lavorando in doppia precisione, è stato considerato un range tale da assumere il regime asintotico ma, allo stesso tempo, tale da non incorrere in problemi di arrotondamento dati dalla differenza di numeri vicini.

### Romberg

Si è poi calcolato  $I_3$  utilizzando il metodo di Romberg. In particolare, si è calcolata la matrice di Romberg per un valore di

$$J_{\max} = 20$$

al fine di ottenere un sufficiente numero di punti per lo studio dell'andamento della dispersione. Si è poi calcolato, per ogni valore di  $J$ , il numero  $N$  di punti in cui è stato diviso l'intervallo di integrazione dato da

$$N = 2^J + 1$$

Al fine di studiare l'andamento della dispersione del metodo si sono quindi studiati i risultati per  $K = 2$  e per  $K = 3$ , ossia per i primi due valori di  $K$  non banali, corrispondenti alla terza e alla quarta colonna della matrice di Romberg.

## **K = 2**

Si vuole verificare la relazione (14) per  $K = 2$ . In questo caso, si avrà allora

$$2K + 2 = 6$$

Si è deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con più dettaglio variazioni grandi e piccole dei risultati. Di seguito si è mostrato quanto ottenuto.

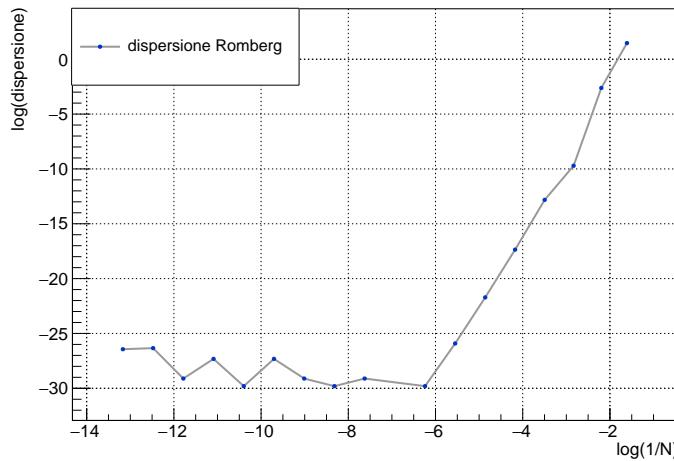


Figura 22: log dispersione metodo di Romberg per  $K = 2$

Anzitutto, è possibile notare che i soliti problemi dovuti alla differenza di numeri molto vicini tra loro genera un andamento non significativo della dispersione a partire da un certo valore  $\tilde{N}$ . Un andamento inatteso non rettilineo di nota anche per valori di  $N$  piccoli. Questo è chiaramente dovuto al fatto che l'andamento della dispersione discusso vale solo in regime asintotico. Per valori di  $N$  non sufficientemente grandi avremo, infatti, dei termini aggiuntivi nella (16) con contributo non più trascurabile. Come già accennato in precedenza, l'unico modo per ovviare al problema di perdita di significatività della stima a partire da un certo numero di punti (che produce l'andamento oscillante) consiste nell'aumentare la precisione con cui vengono rappresentati i numeri reali. Tutti i risultati ottenuti in questa sezione sono stati ricavati in precisione doppia, ma al fine di ottenere andamenti più definiti o stime più precise dai fit risulta opportuno lavorare in precisione doppia estesa o, qualora fosse possibile, in precisione quadrupla, soprattutto se si utilizzano metodi il cui errore scala con una certa rapidità. In questo caso, al fine di svolgere un fit dei dati con la funzione lineare  $y = p + mx$ , si è quindi selezionato un range di punti in cui l'andamento fosse qualitativamente rettilineo. In particolare, basandosi sui risultati ottenuti in figura 22, si sono selezionati i 5 punti tali che

$$-6.5 < \log(n) < -3$$

ottenendo i risultati che seguono.

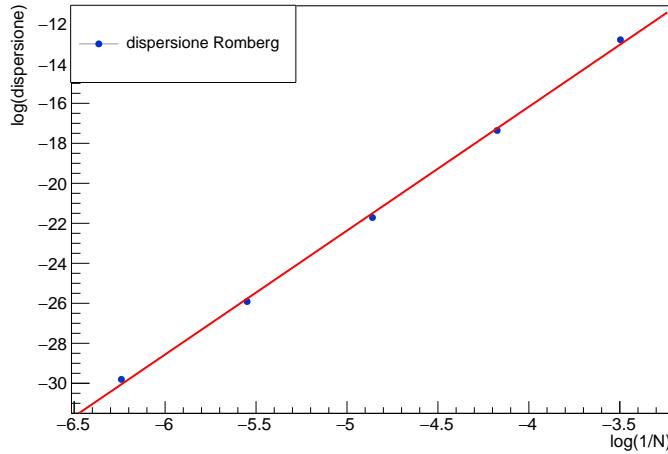


Figura 23: log dispersione metodo di Romberg per  $K = 2$ : fit

I parametri stimati risultano essere

$$p = 8.61 = \log k \quad \text{e} \quad m = 6.2 \approx 6$$

Il valore di  $m$  stimato permette di concludere la verifica della (16) per  $K = 2$ . La non esattezza dipende, chiaramente, dai pochi punti utilizzati per il fit.

### **K = 3**

Si vuole verificare la relazione (14) per  $K = 3$ . In questo caso, vale

$$2K + 2 = 8$$

Si è deciso di eseguire il plot calcolando solo i logaritmi, ottenendo quanto segue.

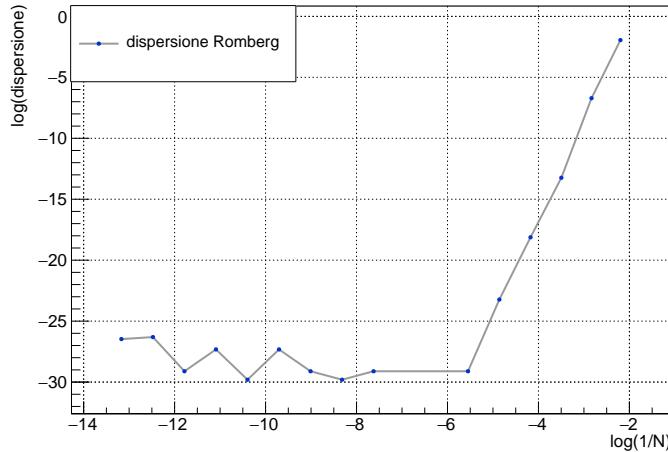


Figura 24: log dispersione metodo di Romberg per  $K = 3$

In questo caso, è immediato notare che i problemi derivanti dalla differenza di numeri molto vicini rende più complesso individuare un intervallo in cui avere sufficienti dati per eseguire un fit, lavorando in precisione doppia. Questo accade in quanto all'aumentare di  $K$  la dispersione decresce molto più rapidamente rispetto alle colonne della matrice di Romberg precedenti. Segue che la stima convergerà più rapidamente al valore vero, e gli effetti della differenza tra numeri troppo vicini si manifesteranno per valori di  $N$  più piccoli nel calcolo della dispersione. Nonostante questo, è possibile notare che i valori di  $\Delta_{R_3}$  tali che

$$-6 < \log(n) < -3$$

risultano avere qualitativamente un andamento rettilineo. Si è quindi eseguito un fit lineare secondo  $y = p + mx$  considerando solo tali valori, ottenendo i seguenti risultati.

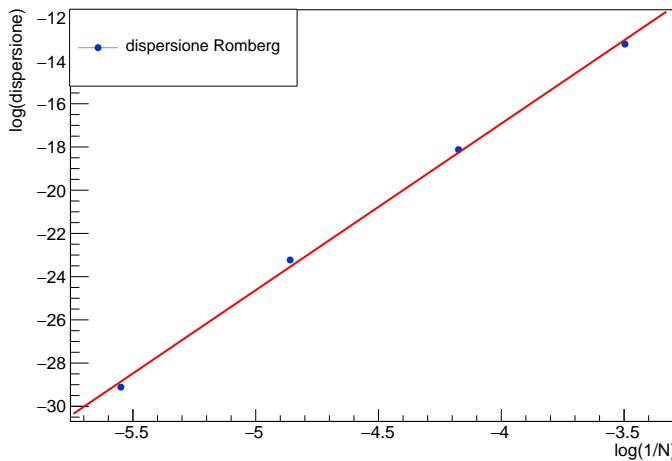


Figura 25: log dispersione metodo di Romberg per  $K = 3$ : fit

Si è ottenuta la seguente stima dei parametri.

$$p = 13.9 = \log k \quad \text{e} \quad m = 7.71 \approx 8$$

Il valore di  $m$  stimato permette di concludere la verifica della legge (16) che governa la dispersione del metodo per  $K = 3$ . La non esattezza della stima ottenuta dipende, chiaramente, dal range ristretto e quindi dai pochi punti utilizzati per il fit. Come è possibile osservare, la stima di  $m$ , in questo caso, risulta peggiore rispetto alla stima dello stesso parametro per  $K = 2$ . Questo è consistente con il fatto che si è dovuto considerare un numero minore di dati in un range più ristretto rispetto al caso precedente. In generale, per poter verificare l'andamento della dispersione del metodo di Romberg per  $K > 3$  risulta necessario l'utilizzo della precisione quadrupla per evitare di incorrere in andamenti insensati per  $N$  così piccoli da non permettere di assumere il regime asintotico.

### Gauss

Si è calcolato  $I_3$  utilizzando il metodo di quadratura gaussiano. In particolare, si sono utilizzati i polinomi di Legendre e quelli di Laguerre per  $\alpha = 0$  con

$N = 2, 4, 8$  punti. Per il calcolo si sono utilizzati i valori tabulati degli zeri e dei pesi dei polinomi in esame, corrispondenti al numero di punti scelto.

### Legendre

L'integrale  $I_3$  è definito sul compatto  $U := [0, 5]$ . Si è quindi utilizzata la relazione (20) al fine di mappare  $U$  nell'intervallo  $[-1, 1]$  di definizione dei polinomi di Legendre. La tabella che segue mostra i risultati ottenuti al variare del numero  $N$  di zeri dei polinomi.

$N$	$\tilde{I}_3$	$\Delta_{G_L}$
2	720	48
4	672.69	0.49
8	672.193246	$9.0 \cdot 10^{-6}$

Come è possibile notare, la stima utilizzando i polinomi di Legendre migliora all'aumentare del numero di zeri fino ad arrivare, per 8 punti, ad una stima che differisce dal valore vero per un valore di  $9.0 \cdot 10^{-6}$ . Quanto ottenuto è consistente con quanto ci si aspetta, infatti, nel compatto  $U$ , la funzione integranda può essere riscritta come un polinomio di Taylor di un certo grado  $M$ , sufficientemente grande da "ricoprire" la funzione su tutto l'intervallo. Sappiamo che la stima è esatta per polinomi di grado  $2N - 1$ : all'aumentare di  $N$ , si ha che  $2N - 1$  si avvicina sempre più al valore di  $M$ . Segue che la precisione della stima aumenterà con  $N$ , fino al raggiungimento di un risultato molto preciso per un numero di zeri dei polinomi di Legendre sufficientemente grande.

### Laguerre

Si noti che la funzione integranda assume la forma

$$g(x) = W(x)P_7(x)$$

dove  $W(x) = e^{-x}$  è il peso dei polinomi di Laguerre per  $\alpha = 0$  e  $P_7(x) = x^7$  è un polinomio di grado 7. Si ha quindi un integrale della forma (17), ossia tale che sia possibile isolare il peso dei polinomi di Laguerre nella funzione integranda. Si è quindi calcolato  $\tilde{I}_3$  utilizzando la relazione (25). In questo caso, infatti, è immediato verificare che vale la condizione necessaria di convergenza (23). In particolare, la presenza dell'esponenziale decrescente rende integrabile la funzione  $g$  all'infinito. La tabella che segue mostra i risultati ottenuti al variare del numero  $N$  di zeri dei polinomi.

$N$	$\tilde{I}_3$	$\Delta_{G_G}$
2	-3130	3802
4	672.19323731284	$2.0 \cdot 10^{-13}$
8	672.19323731284	$3.0 \cdot 10^{-12}$

Ricordiamo che il metodo delle quadrature gaussiane fornisce una stima esatta quando  $P$  è un polinomio di grado  $2N - 1$ . In questo caso,  $P_7$  è un polinomio di grado 7. Segue che la stima numerica sarà esatta  $\forall N \geq \tilde{N}$  tale che

$$2\tilde{N} - 1 = 7 \iff \tilde{N} = 4$$

Si noti, tuttavia, che in tabella  $\Delta_{G_G}$  appare diversa da 0 anche per  $N = 4, 8$ . D'altra parte, osservando l'ordine di grandezza della dispersione, è evidente che questo fatto sia solo dovuto al raggiungimento del limite della doppia precisione utilizzata. Alla luce di questo, allora, i risultati ottenuti sono consistenti: la stima è esatta per  $N = 4$  e per  $N = 8$ , come ci si aspetta dai risultati teorici. Appare, invece, una stima molto distante dal valore vero per  $N = 2$ . Questi risultati mostrano anche che il metodo di quadratura gaussiana può fornire stime che differiscono sensibilmente tra loro per piccole variazioni di  $N$ . Questo fatto implica che è necessario prestare una certa attenzione quando si applica il metodo di Gauss per stimare un integrale: senza curarsi della forma analitica della funzione e della proprietà di esattezza è possibile ottenere risultati privi di significato.

Nel caso dell'integrale  $I_3$ , allora, i risultati ottenuti mostrano che la strada meno costosa computazionalmente per ottenere una stima sufficientemente precisa sia quella dell'applicazione del metodo di Gauss con i polinomi di Laguerre, come ci si aspetta vista la presenza esplicita della funzione peso dei polinomi di Laguerre nell'integrandà.

## Esercizio 4

Si vuole stimare numericamente il valore dell'integrale

$$I_4 := \int_3^8 \cosh(x) dx$$

utilizzando i metodi di Newton-Cotes e il metodo di quadratura gaussiana, al fine di verificare la loro efficienza e la loro stabilità nel problema in esame.

Analiticamente (esplicitando  $\cosh$  come combinazione di esponenziali) o con l'aiuto di un calcolatore avanzato, si ricava che il valore esatto è

$$I_4 = \sinh(8) - \sinh(3) \approx 1480.46095086214028421690$$

Siamo ora in grado di calcolare le dispersioni delle stime dal valore vero.

### Trapezio

Si è calcolato  $I_4$  utilizzando il metodo del trapezio esteso, variando il numero di punti in input. Ci si aspetta che, in generale, la stima migliori all'aumentare del numero di punti in cui si divide l'intervallo di integrazione. Equivalentemente, ci si aspetta che l'errore decresca al crescere del numero di punti  $N$  secondo la (8). In particolare, al fine di visualizzare meglio i dati raccolti, si sono plot-tati i risultati secondo la relazione equivalente (9). Si sono quindi calcolate le dispersioni al variare di  $1/N$  nel range

$$300 \leq N < 1500 \quad \text{con} \quad N_{i+1} = N_i + 50$$

ottenendo il grafico che segue.

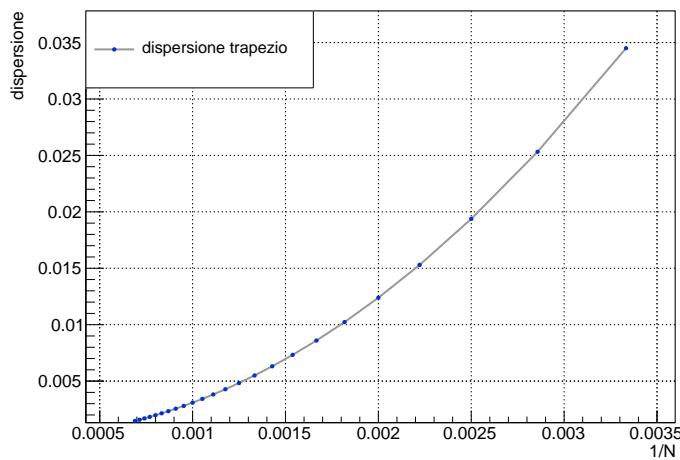


Figura 26: dispersione metodo del trapezio

Si noti che l'andamento qualitativo è parabolico, come ci si aspetta per valori di  $N$  grandi, ossia in regime asintotico. Si è quindi deciso di interpolare i dati raccolti, al fine di verificare la bontà dell'adattamento della curva ai dati. In

tal senso, si è deciso di svolgere il fit passando ai logaritmi, provando quindi a verificare la relazione equivalente (10). Si sono quindi fittati i dati con una funzione lineare della forma  $y = p + mx$ . Di seguito sono riportati i risultati restituiti dall'algoritmo di minimizzazione.

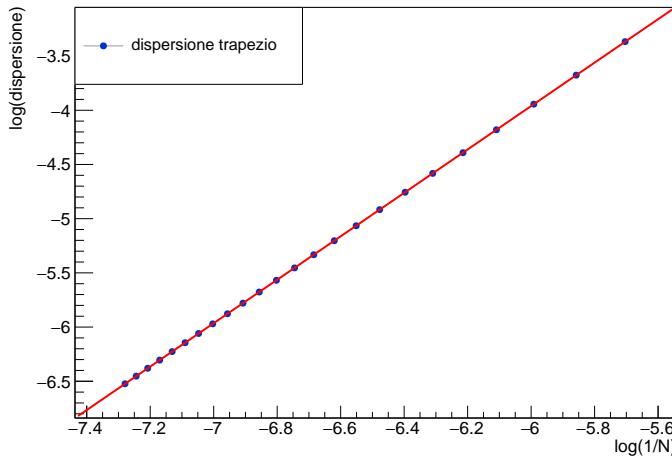


Figura 27: log dispersione metodo del trapezio: fit

I parametri stimati risultano

$$p = 8.06 = \log k \quad \text{e} \quad m = 2$$

Come è possibile notare, la stima del parametro  $m$  è consistente con il coefficiente angolare della relazione trovata analiticamente. Questo fatto garantisce la bontà del fit effettuato, rendendo possibile concludere che, nella formula di integrazione del trapezio esteso, l'errore scali come  $\frac{1}{N^2}$  per  $N$  grande, come ci si aspetta. Come nell'esercizio precedente, la velocità con cui scala l'errore nel metodo del trapezio non è tale da incorrere in andamenti non significativi dati dalla precisione finita di macchina per il range di  $N$  selezionato. Ci aspettiamo che, in generale, per i metodi che seguiranno possano presentarsi problemi di arrotondamento.

### Simpson

Si è poi calcolato  $I_4$  utilizzando il metodo di Simpson esteso, variando il numero di punti in input. In questo caso, siamo allora interessati a verificare che valga la relazione (11). Per ragioni di comodità di visualizzazione dei dati, si è deciso di provare a visualizzare la relazione equivalente (13), passando direttamente ai logaritmi. Si sono quindi calcolati e plottati i logaritmi delle dispersioni al variare di  $\log(1/N)$  nel range

$$401 \leq N < 3000 \quad \text{con} \quad N_{i+1} = N_i + 30$$

selezionando solo gli  $N$  dispari, al fine di garantire il corretto funzionamento dell'algoritmo dato dal metodo di Simpson. Il grafico che segue mostra quanto ottenuto.

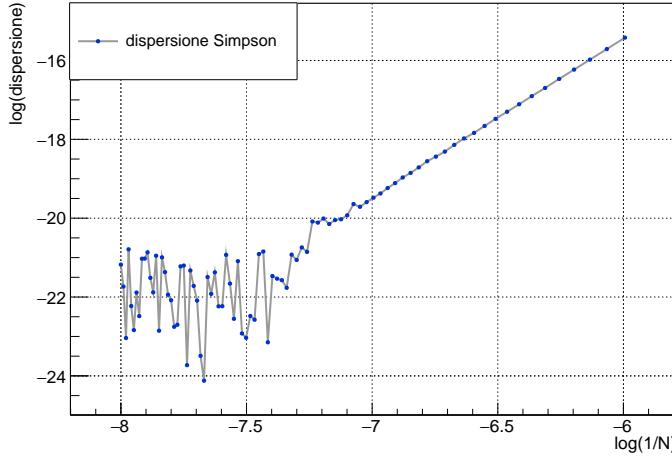


Figura 28: log dispersione metodo di Simpson

Come ci si può aspettare,  $\exists \tilde{N} \in \mathbb{N}$  tale che la dispersione assuma un andamento oscillante secondo una legge non banale  $\forall N > \tilde{N}$  per le solite ragioni dovute alla differenza di valori molto vicini tra loro. Si è quindi ristretto l'intervallo di punti considerato, infittendo il passo di campionamento dei dati in una regione in cui l'andamento risultasse, qualitativamente, rettilineo. In particolare, si sono ripetuti i medesimi passaggi nel range

$$301 \leq N < 700 \quad \text{con} \quad N_{i+1} = N_i + 10$$

In tale regione, si sono interpolati i dati raccolti con la funzione  $y = p + mx$ , ottenendo quanto segue.

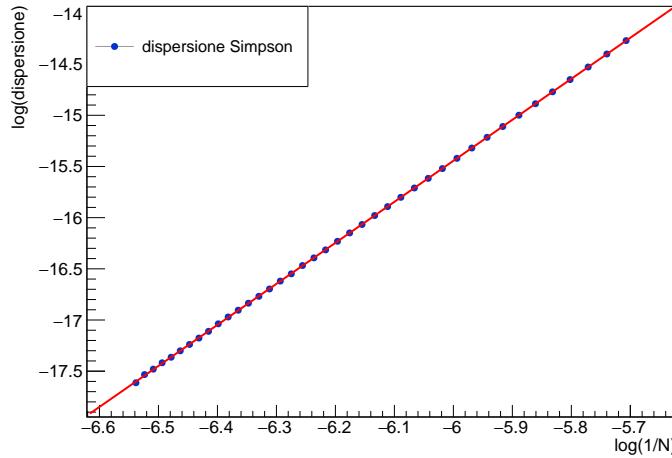


Figura 29: log dispersione metodo di Simpson per  $N < \tilde{N}$ : fit

I parametri stimati risultano

$$p = 8.6 = \log k \quad \text{e} \quad m = 4.01 \approx 4$$

Vista la compatibilità tra la stima di  $m$  e il parametro noto, la legge (13) che governa la dispersione in funzione del numero di punti può dirsi, dunque, verificata. La non esattezza della stima ottenuta è dovuta al fatto di aver considerato un range di  $N$  ristretto.

### Romberg

Si è calcolato  $I_4$  utilizzando il metodo di Romberg. In particolare, si è calcolata la matrice di Romberg per un valore di

$$J_{max} = 20$$

Si è poi calcolato, per ogni valore di  $J$ , il numero  $N$  di punti in cui è stato diviso l'intervallo di integrazione, dato da

$$N = 2^J + 1$$

Al fine di studiare l'andamento della dispersione del metodo, si sono quindi studiati i risultati per  $K = 2$  e per  $K = 3$ , ossia per i primi valori di  $K$  non coincidenti al metodo del trapezio o al metodo di Simpson.

### **K = 2**

Si vuole verificare la relazione (14) per  $K = 2$ . In questo caso avremo

$$2K + 2 = 6$$

Si è deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con massimo dettaglio l'eventuale presenza di andamenti oscillanti non significativi. Di seguito sono mostrati i risultati ottenuti.

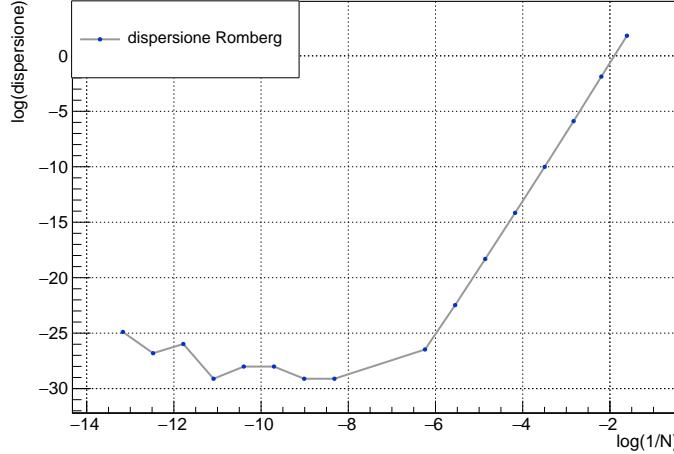


Figura 30: log dispersione metodo di Romberg per  $K = 2$

Anzitutto, è possibile notare che i problemi dati dalla differenza di numeri molto vicini tra loro genera un andamento non significativo della dispersione a partire da un certo valore  $\tilde{N}$ . Al fine di svolgere un fit lineare dei dati con la funzione

$y = p + mx$ , si è quindi selezionato un range di punti in cui l'andamento fosse qualitativamente rettilineo. In particolare, tenendo presente la figura 30, si sono selezionati gli ultimi 8 punti tali che

$$-6.5 < \log(n) < -1$$

ottenendo i seguenti risultati.

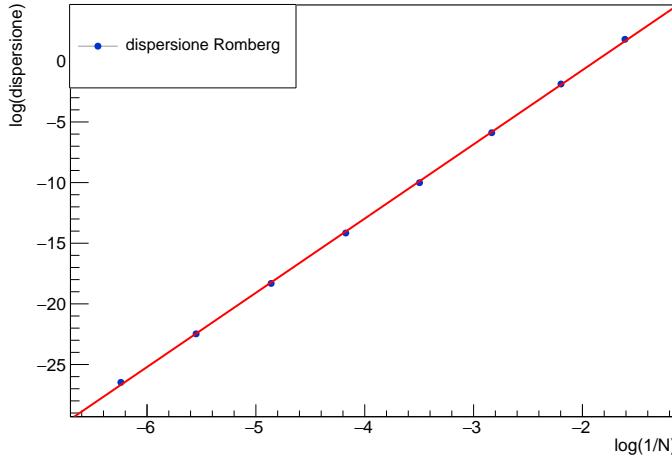


Figura 31: log dispersione metodo di Romberg per  $K = 2$ : fit

I parametri stimati risultano essere i seguenti.

$$p = 11.5 = \log k \quad \text{e} \quad m = 6.12 \approx 6$$

Il valore di  $m$  stimato permette di concludere che la legge (16) che governa la dispersione del metodo per  $K = 2$  è verificata. La non esattezza della stima ottenuta dipende, chiaramente, dal range ristretto e, di conseguenza, dai pochi punti utilizzati per il fit. D'altra parte, a differenza degli altri metodi di Newton-Cotes, nel metodo di Romberg non è possibile scegliere a piacere un range di  $N$  in cui produrre più dati con andamento rettilineo a causa della struttura stessa del metodo. L'unica strada per lo studio dell'andamento dell'errore, in questo caso, consiste nel selezionare opportunamente dati sensati. In alternativa, come si ha già avuto modo di osservare, risulta necessario lavorare con una precisione maggiore della precisione doppia.

### K = 3

Si vuole verificare la relazione (14) per  $K = 3$ . In questo avremo allora

$$2K + 2 = 8$$

Si è quindi deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con massimo dettaglio l'eventuale presenza di andamenti oscillanti non significativi. Di seguito sono mostrati i risultati ottenuti.

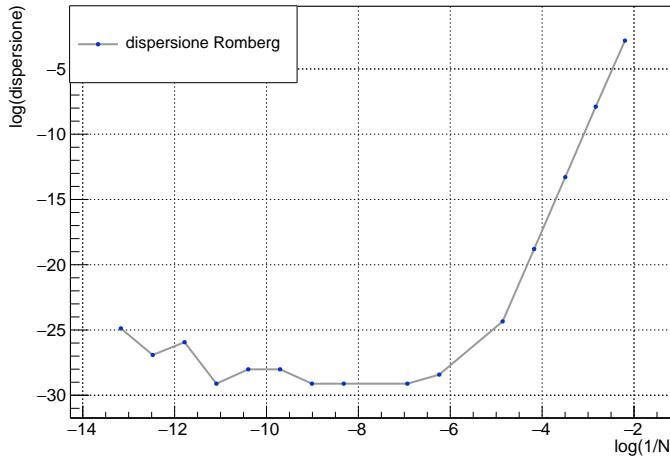


Figura 32: log dispersione metodo di Romberg per  $K = 3$

I problemi di rappresentazione dei numeri reali, vista la migliore precisione per  $K = 3$ , rendono più complesso individuare un range in cui la funzione appaia lineare. Nonostante questo, è possibile notare che i valori di  $\Delta_{R_3}$  tali che

$$-5 < \log(n) < -2$$

risultano avere qualitativamente un andamento rettilineo. Si è quindi eseguito un fit lineare secondo  $y = p + mx$  considerando solo tali punti, ottenendo i seguenti risultati.

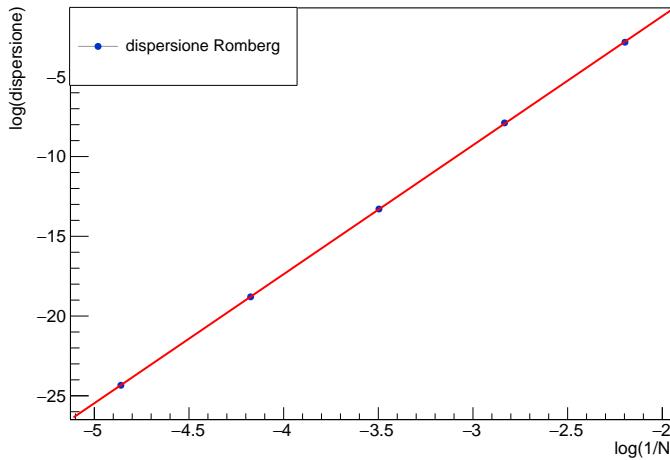


Figura 33: log dispersione metodo di Romberg per  $K = 3$ : fit

I parametri stimati risultano essere i seguenti.

$$p = 15 = \log k \quad \text{e} \quad m = 8.09 \approx 8$$

Il valore di  $m$  stimato permette di concludere la verifica della legge (16) per  $K = 3$ . La non esattezza della stima ottenuta, anche in questo caso, dipende dal range ristretto e dai pochi punti utilizzati per il fit.

### Gauss

Si è calcolato  $I_4$  utilizzando il metodo di quadratura gaussiana. In particolare, si sono utilizzati i polinomi di Legendre per  $N = 2, 4, 8$  punti e quelli di Laguerre con  $\alpha = 0$  per  $N = 2, 4$  punti. Per il calcolo, si sono utilizzati i valori tabulati degli zeri e dei pesi dei polinomi in esame, corrispondenti al numero di punti scelto.

### Legendre

L'integrale  $I_4$  è definito sul compatto  $U := [3, 8]$ . Si è quindi utilizzata la relazione (20), al fine di mappare  $U$  nell'intervallo  $[-1, 1]$  di definizione dei polinomi di Legendre. La tabella che segue mostra i risultati ottenuti al variare del numero  $N$  di zeri dei polinomi.

$N$	$\tilde{I}_4$	$\Delta_{G_L}$
2	1368	113
4	1480.304	0.157
8	1480.4609508604	$1.7 \cdot 10^{-9}$

Come è possibile notare, la stima utilizzando i polinomi di Legendre migliora all'aumentare del numero di zeri, fino ad arrivare, per 8 punti, ad una stima che differisce dal valore vero per un valore di  $1.7 \cdot 10^{-9}$ . Questo è consistente con quanto ci si aspetta, infatti, nel compatto  $U$ , la funzione integranda può essere riscritta come un polinomio di Taylor di un certo grado  $M$ . Sappiamo che la stima è esatta per polinomi di grado  $2N - 1$ . All'aumentare di  $N$ , si ha che  $2N - 1$  si avvicina sempre più al valore di  $M$ . Segue che la precisione della stima aumenta con il numero di punti, fino al raggiungimento di una precisione molto elevata per un numero di zeri dei polinomi sufficientemente grande.

### Laguerre

Anzitutto, si noti che, in questo caso, la condizione necessaria di convergenza (23) non è verificata. Infatti

$$\lim_{x \rightarrow +\infty} \cosh(x) = +\infty$$

Segue che il metodo spiegato nella sezione dedicata, implementando direttamente il coseno iperbolico come funzione integranda, non funziona. Come prova di questo fatto, si sono calcolati i valori di  $\tilde{I}_4$  e di  $\Delta_{G_L}$  utilizzando il metodo implementato, senza tener conto delle ipotesi di utilizzo della relazione (25). Si sono ottenuti i seguenti risultati.

$N$	$\tilde{I}_4$	$\Delta_{G_G}$
2	-204348	205829
4	-116044644	116046124

I risultati ottenuti risultano, chiaramente, insensati. Dalla condizione (23) non verificata segue, infatti, che il coseno iperbolico non è integrabile all'infinito. Si è quindi notato che la funzione integranda è definita come

$$\cosh(x) := \frac{e^x + e^{-x}}{2}$$

da cui segue che l'integrale in esame può essere riscritto come

$$I_4 = \frac{1}{2} \int_{-3}^8 e^x dx + \frac{1}{2} \int_{-3}^8 e^{-x} dx$$

Chiamiamo ora  $I_4^a$  il primo integrale e  $I_4^b$  il secondo. Risulta immediato osservare che  $I_4^b$  verifica la condizione di convergenza (23). In particolare, è integrabile all'infinito e appare come una funzione  $g$  della forma

$$g(x) = W(x)P_0(x)$$

dove  $W(x) = e^{-x}$  è il peso dei polinomi di Laguerre per  $\alpha = 0$  e  $P_0(x) = 1$  è un polinomio di grado 0. Segue che è possibile applicare la (25) a  $I_4^b$ , ottenendo l'esattezza della stima a partire da  $N = 1$ . Risulta quindi necessario manipolare solo  $I_4^a$ , al fine di poter utilizzare il metodo implementato anche in questo caso. Si consideri, dunque, la mappa biettiva

$$y = -x$$

Sotto questo cambio di variabile si ha

$$I_4^a = \frac{1}{2} \int_{-8}^{-3} e^{-x} dx$$

Ma allora, in questo modo abbiamo ottenuto un integrale che presenta caratteristiche analoghe all'integrale  $I_4^b$ , a cui possiamo applicare il metodo implementato. Si sono, quindi, calcolati singolarmente  $I_4^a$  e  $I_4^b$  utilizzando la (25), per poi sommare i due risultati. La tabella che segue mostra i risultati ottenuti.

$N$	$\tilde{I}_4$	$\Delta_{G_G}$
2	1480.46095086214	$2.3 \cdot 10^{-13}$
4	1480.46095086214	$6.8 \cdot 10^{-13}$

Tenendo conto della precisione macchina, i risultati ottenuti risultano, ragionevolmente, esatti per entrambi i valori di  $N$ . Quanto ottenuto è consistente con il fatto che, per polinomi di grado 0, come per la funzione costante  $P_0(x) = 1$ , l'esattezza della stima si ha  $\forall N \geq \tilde{N}$  tale che

$$2\tilde{N} - 1 = 0 \iff \tilde{N} = \frac{1}{2}$$

D'altra parte, il numero di zeri di un polinomio è un numero naturale. Segue che l'esattezza della stima si ha per tutti i valori di  $N$  a partire da  $\tilde{N} = 1$ . Somma di stime esatte restituisce una stima esatta, nei limiti della precisione doppia utilizzata nel calcolo.

Anche nel caso del calcolo di  $I_4$ , allora, la strada più furba e meno costosa computazionalmente per ottenere un'ottima stima risulta quella data dalla quadratura gaussiana con i polinomi di Laguerre, a patto di non applicare ciecamente il metodo numerico all'integrale in esame.

## Esercizio 5

Si vuole stimare numericamente il valore dell'integrale

$$I_5 := \int_{-1}^8 (x^2 + x \sin(4x)) dx$$

utilizzando i metodi di Newton-Cotes e il metodo di quadratura gaussiana, al fine di verificare la loro efficienza e la loro stabilità nel problema in esame.

Analiticamente (sfruttando la linearità dell'integrale) o con l'aiuto di un calcolatore avanzato, si ricava che il valore esatto è

$$\begin{aligned} I_5 &= \frac{1}{16} (2736 - 4 \cos(4) - 32 \cos(32) + \sin(4) + \sin(32)) \approx \\ &\approx 169.482128195823742577934 \end{aligned}$$

Siamo ora in grado di calcolare le dispersioni delle stime dal valore vero.

### Trapezio

Si è calcolato  $I_5$  utilizzando il metodo del trapezio esteso, variando il numero di punti in input. Ci si aspetta che, in generale, la stima migliori all'aumentare del numero di punti in cui si divide l'intervallo di integrazione. Equivalentemente, ci si aspetta che l'errore decresca al crescere del numero di punti  $N$  secondo la (8). In particolare, al fine di visualizzare meglio i dati raccolti, si sono plottati i dati secondo la relazione equivalente (9). Si sono quindi calcolate le dispersioni al variare di  $1/N$  nel range

$$300 \leq N < 1500 \quad \text{con} \quad N_{i+1} = N_i + 50$$

ottenendo il grafico che segue.

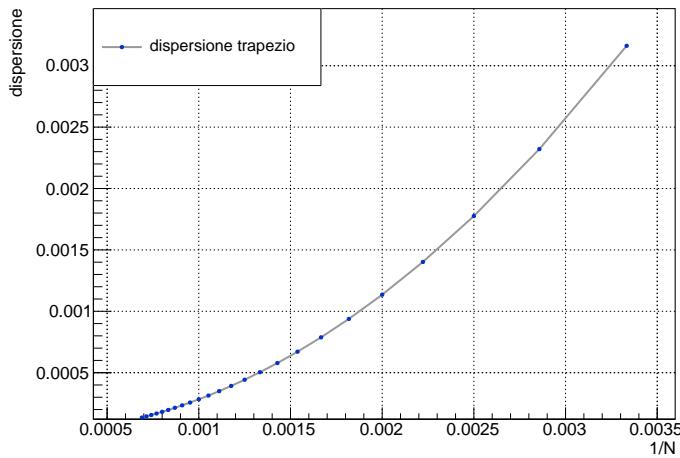


Figura 34: dispersione metodo del trapezio

Si noti che l'andamento qualitativo è parabolico, come ci si aspetta per valori di  $N$  grandi, ossia in regime asintotico. Si è quindi deciso di interpolare i dati

raccolti, al fine di verificare la bontà dell'adattamento della curva ai risultati. In tal senso, si è deciso di svolgere il fit passando ai logaritmi, provando quindi a verificare la relazione equivalente (10). Si sono dunque fittati i dati con una funzione lineare della forma  $y = p + mx$ , ottenendo quanto segue.

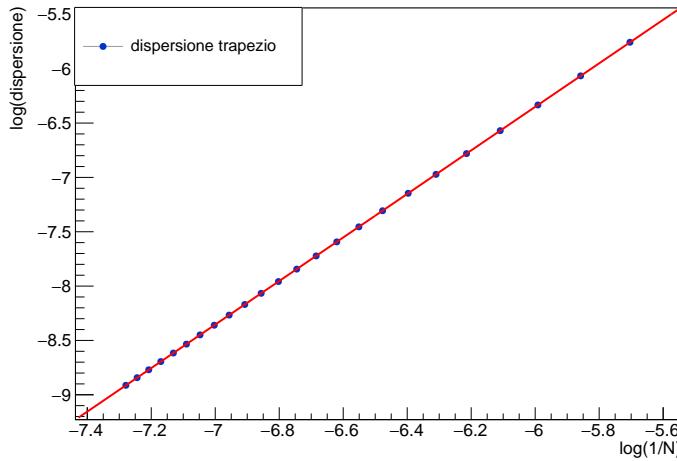


Figura 35: log dispersione metodo del trapezio: fit

I parametri stimati risultano

$$p = 5.67 = \log k \quad \text{e} \quad m = 2$$

Come è possibile notare, la stima del parametro  $m$  è consistente con il coefficiente angolare della relazione trovata analiticamente. Questo fatto garantisce la bontà del fit effettuato, rendendo possibile concludere che, nella formula di integrazione del trapezio esteso, l'errore scali come  $\frac{1}{N^2}$  per  $N$  grande, come ci si aspetta. Come nei casi precedenti, la velocità con cui scala l'errore nel metodo del trapezio non è tale da consentire il raggiungimento della precisione macchina. Per tale ragione, non è stato necessario scartare valori di dispersione per eseguire il fit. Ci aspettiamo che, in generale, per i metodi che seguiranno possano presentarsi problemi di arrotondamento.

### Simpson

Si è poi calcolato  $I_5$  utilizzando il metodo di Simpson esteso, variando il numero di punti in input. In questo caso, siamo allora interessati a verificare che valga la relazione (11). Per ragioni di comodità di visualizzazione dei dati, si è deciso di provare a visualizzare la relazione equivalente (13), passando direttamente ai logaritmi. Si sono quindi calcolati e plottati i logaritmi delle dispersioni al variare di  $\log(1/N)$  nel range

$$401 \leq N < 7000 \quad \text{con} \quad N_{i+1} = N_i + 100$$

selezionando solo gli  $N$  dispari, al fine di garantire il corretto funzionamento del metodo. Il grafico che segue mostra quanto ottenuto.

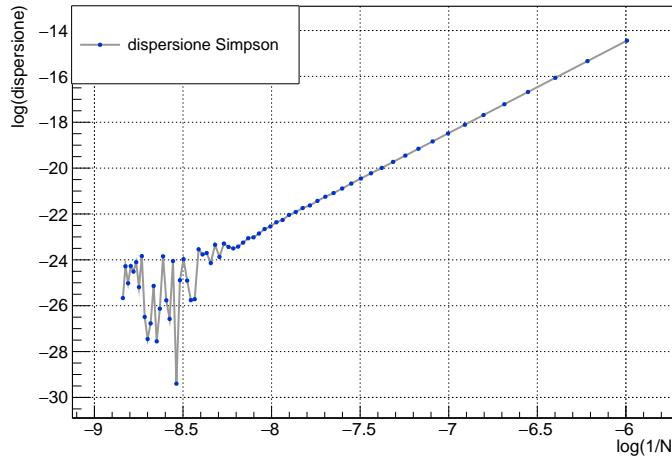


Figura 36: log dispersione metodo di Simpson

Anche in questo caso,  $\exists \tilde{N} \in \mathbb{N}$  tale che la dispersione assuma un andamento oscillante secondo una legge non banale  $\forall N > \tilde{N}$  per le solite questioni dovute alla differenza di valori molto vicini tra loro. Si è quindi ristretto l'intervallo di punti considerato fino ad una regione in cui l'andamento risultasse, qualitativamente, rettilineo. In particolare, si è considerato il range

$$401 \leq N < 2000 \quad \text{con} \quad N_{i+1} = N_i + 40$$

In tale regione, si sono interpolati i dati raccolti con la funzione  $y = p + mx$ .

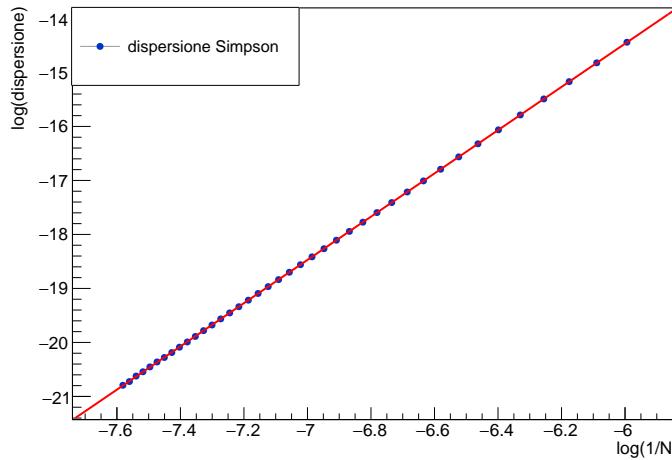


Figura 37: log dispersione metodo di Simpson per  $N < \tilde{N}$ : fit

I parametri stimati risultano i seguenti.

$$p = 9.56 = \log k \quad \text{e} \quad m = 4$$

Vista la compatibilità tra la stima di  $m$  e il parametro noto, la legge (13) che governa la dispersione in funzione del numero di punti può dirsi, dunque, verificata. Si noti che, in questo caso, si è ottenuta una stima esatta del parametro  $m$ . Questo è dovuto al fatto di aver considerato un range di valori di  $N$  più grande rispetto ai casi precedenti, in quanto i problemi dati dalla rappresentazione finita di macchina sono comparsi per un valore di  $N$  maggiore. In questo caso, infatti, la forma analitica della funzione integranda porta ad una convergenza meno rapida della stima al valore vero, permettendo di incorrere in problemi di differenze di numeri vicini a partire da un numero di punti maggiore.

### Romberg

Si è calcolato  $I_5$  utilizzando il metodo di Romberg. In particolare, si è calcolata la matrice di Romberg per un valore di

$$J_{max} = 20$$

Si è poi calcolato, per ogni valore di  $J$ , il numero  $N$  di punti in cui è stato diviso l'intervallo di integrazione, dato da

$$N = 2^J + 1$$

Al fine di studiare l'andamento della dispersione del metodo si sono quindi studiati i risultati per  $K = 2$  e per  $K = 3$ .

#### **K = 2**

Si vuole verificare la relazione (14) per  $K = 2$ . In questo caso vale allora

$$2K + 2 = 6$$

Si è deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con più dettaglio i risultati ottenuti. Si è ottenuto quanto segue.

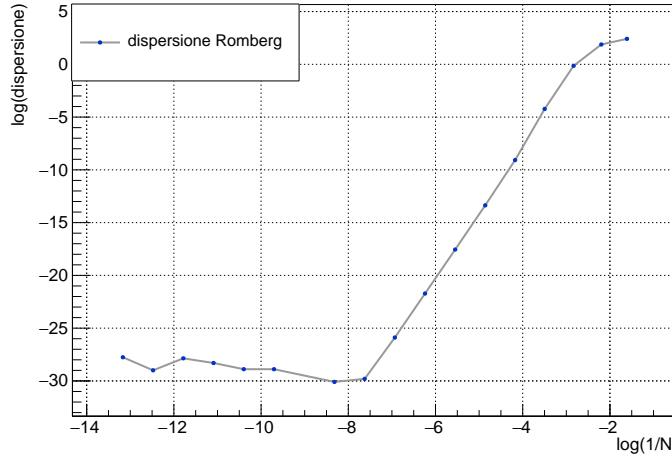


Figura 38: log dispersione metodo di Romberg per  $K = 2$

Anzitutto, è possibile notare che i problemi dovuti alla limitatezza della precisione macchina generano un andamento non significativo della dispersione a partire da un certo valore di  $\tilde{N}$ . Al fine di svolgere un fit lineare dei dati con la funzione  $y = p + mx$ , si è quindi selezionato un range di punti in cui l'andamento fosse qualitativamente rettilineo. In particolare, facendo riferimento alla figura 38, si sono selezionati i 6 punti tali che

$$-8 < \log(n) < -4$$

Per valori di  $\log(n)$  maggiori, infatti, il grafico mostra chiaramente che non è possibile assumere di lavorare in regime asintotico. Di seguito sono mostrati i risultati ottenuti.

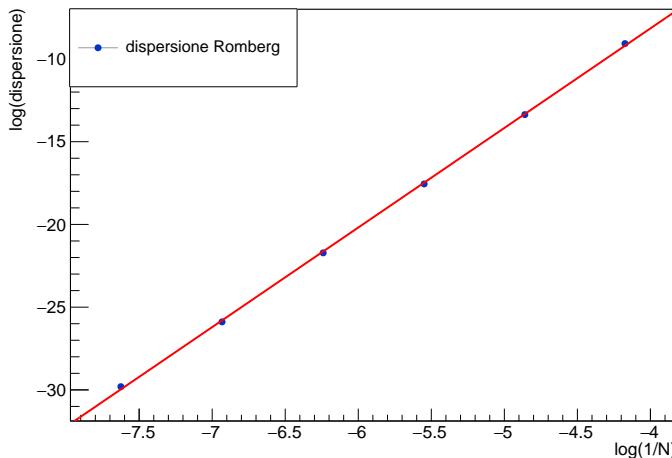


Figura 39: log dispersione metodo di Romberg per  $K = 2$ : fit

I parametri stimati risultano i seguenti.

$$p = 15.9 = \log k \quad \text{e} \quad m = 6.02 \approx 6$$

Il valore di  $m$  stimato permette di concludere che la legge (16) che governa la dispersione del metodo per  $K = 2$  è verificata. La non esattezza della stima ottenuta dipende, chiaramente, dal range ristretto e, di conseguenza, dai pochi punti utilizzati per il fit.

### K = 3

Si vuole verificare la relazione (14) per  $K = 3$ . In questo caso avremo

$$2K + 2 = 8$$

Come al solito, si è deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con massimo dettaglio l'eventuale presenza di andamenti oscillanti non significativi. Ci aspettiamo di osservare un numero minore di punti che presentano un andamento rettilineo. Di seguito sono mostrati i risultati ottenuti.

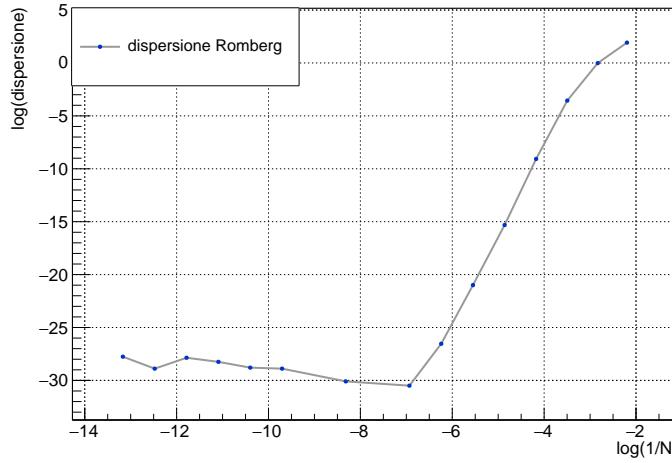


Figura 40: log dispersione metodo di Romberg per  $K = 3$

I problemi dati dalla differenza di numeri vicini, vista la migliore precisione per  $K = 3$ , rendono più complesso individuare un range in cui la funzione appaia lineare. Nonostante questo, è possibile notare che i valori di  $\Delta_{R_3}$  tali che

$$-6.5 < \log(n) < -4$$

risultano avere qualitativamente un andamento rettilineo. Si è quindi eseguito un fit lineare secondo  $y = p + mx$  considerando solo tali punti, ottenendo i seguenti risultati.

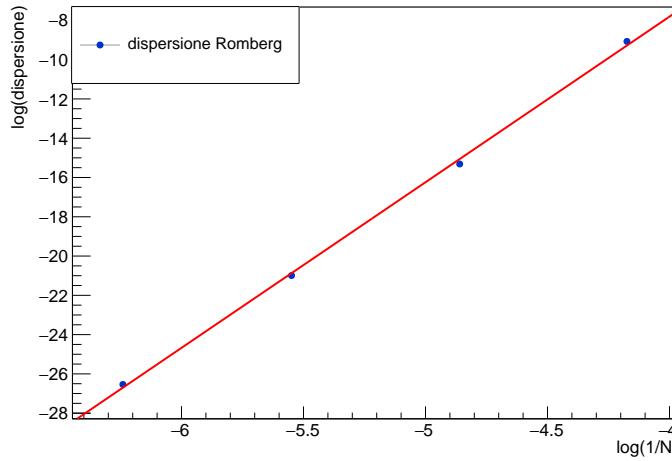


Figura 41: log dispersione metodo di Romberg per  $K = 3$ : fit

I parametri ottenuti risultano essere i seguenti.

$$p = 25.9 = \log k \quad \text{e} \quad m = 8.43 \approx 8$$

Il valore di  $m$  stimato permette di concludere, anche in questo caso, la verifica della legge (16) per  $K = 3$ . La non esattezza della stima ottenuta dipende dal range ristretto e dai pochi punti utilizzati per il fit.

### Gauss

Si è calcolato  $I_5$  utilizzando il metodo di quadratura gaussiana. In particolare, si sono utilizzati i polinomi di Legendre per  $N = 2, 4, 8, 16, 48$  punti. Per il calcolo, si sono utilizzati i valori tabulati degli zeri e dei pesi dei polinomi in esame, corrispondenti al numero di punti scelto.

### Legendre

L'integrale  $I_5$  è definito sul compatto  $U := [-1, 8]$ . Si è quindi utilizzata la relazione (20) al fine di mappare  $U$  nell'intervallo  $[-1, 1]$  di definizione dei polinomi di Legendre. La tabella che segue mostra i risultati ottenuti al variare del numero  $N$  di zeri dei polinomi.

$N$	$\tilde{I}_5$	$\Delta_{G_L}$
2	151	19
4	181	11
8	149	20
16	169.48209	$3.9 \cdot 10^{-5}$
48	169.482128195824	0

Come è possibile notare, la dispersione di  $\tilde{I}_5$  dal valore vero utilizzando i polinomi di Legendre rimane sostanzialmente stabile sull'ordine di una potenza del 10 per  $N = 2, 4, 8$  punti. Questo accade in quanto lo sviluppo polinomiale in Taylor della funzione integranda sul compatto  $U$  di integrazione non ha un grado tale da consentire un miglioramento della stima per i primi tre punti considerati. Per  $N = 16$  punti, invece, la stima appare sensibilmente migliorata di 6 ordini di grandezza. Evidentemente, in questo caso, il grado dello sviluppo dell'integranda su  $U$  si avvicina al valore  $2N - 1$ . Infine, la stima appare sostanzialmente confrontabile con il valore vero per  $N = 48$  punti. Si noti che il valore di dispersione nullo ottenuto non è da interpretarsi come stima esatta, ma è solo il risultato di "sporcizia" di macchina a seguito del raggiungimento del limite della precisione doppia utilizzata. Come è possibile notare, in questo caso, la convergenza di  $\tilde{I}_5$  al valore vero appare più lenta. Risulta quindi necessario un numero maggiore di punti al fine di ottenere una buona stima con il metodo di quadratura gaussiana. Questo è dovuto al fatto che la funzione integranda è una composizione di un polinomio e di una funzione trigonometrica. La presenza di una funzione non polinomiale rende meno efficiente il metodo, richiedendo un numero di punti maggiore per garantire la precisione della stima.

Ad ogni modo, nonostante l'efficienza ridotta rispetto agli esercizi precedenti, anche nel caso di  $I_5$  il metodo di Gauss con i polinomi di Legendre risulta essere la strada meno costosa a livello computazionale per ottenere una stima sufficientemente precisa. Nei metodi di Newton-Cotes, infatti, il numero di punti in cui è necessario valutare la funzione per il raggiungimento della precisione macchina è ben al di sopra dei 48 punti richiesti dal metodo di Gauss.

## Esercizio 6

Si vuole stimare numericamente il valore dell'integrale

$$I_6 := \int_3^{+\infty} x^5 e^{-x^2} dx$$

utilizzando i metodi di Newton-Cotes e il metodo di quadratura gaussiana, al fine di verificare la loro efficienza e la loro stabilità nel problema in esame.

Analiticamente (facendo uso di un opportuno cambio di variabile) o con l'aiuto di un calcolatore avanzato, si ricava che il valore esatto è

$$I_6 = \frac{101}{2e^9} \approx 0.00623219510637731724963065$$

Siamo ora in grado di calcolare le dispersioni delle stime dal valore vero.

Si noti che, in questo caso, l'integrale da stimare risulta essere un integrale improprio. Questo fatto sarà particolarmente utile per la stima numerica con il metodo di quadratura gaussiana. D'altra parte, il metodo del trapezio, di Simpson e di Romberg sono costruiti per il calcolo di integrali definiti su un compatto. Risulta quindi necessario determinare una trasformazione  $\Phi$  che permetta di mandare l'intervallo illimitato  $[3, +\infty)$  in un intervallo limitato di  $\mathbb{R}$ . Si consideri, dunque, il cambio di variabile

$$z = e^{-x^2}$$

evidentemente biunivoco all'interno dell'intervallo di integrazione. Ma allora, agli estremi del dominio vale la mappatura

$$z(3) = e^{-9} \quad \text{e} \quad \lim_{x \rightarrow +\infty} z(x) = 0$$

Differenziando il cambio di variabile e invertendo la trasformazione si ha

$$dz = -2xe^{-x^2} dx \quad \text{e} \quad x^2 = -\log(z)$$

da cui segue immediatamente che l'integrale può essere riscritto come

$$I_6 = \frac{1}{2} \int_0^{e^{-9}} \log^2(z) dz \tag{26}$$

Si noti che, nonostante  $I_6$  sia ora scritto come un integrale equivalente definito sull'intervallo limitato  $U := (0, e^{-9}]$ , risulta comunque impossibile applicare direttamente i tre metodi in questione. Infatti, vale

$$\lim_{z \rightarrow 0} \log^2(z) = +\infty$$

ossia la funzione integranda  $g(z) := \log^2(z)$  presenta una singolarità in  $z = 0$ . Il metodo del trapezio, di Simpson e di Romberg calcolano la funzione integranda in ogni punto in cui si decide di dividere l'intervallo di integrazione. In particolare, il primo punto  $x_1$  coincide, per costruzione, all'estremo sinistro del compatto su cui l'integrale è definito. Risulta evidente, dunque, che il calcolo

di  $f_1 = f(x_1)$  produrrebbe un infinito nei casi in cui l'integrandà diverga nel punto, rendendo i metodi in esame inapplicabili. Per stimare numericamente  $I_6$  con i tre metodi in esame, allora, risulta necessario combinare le formule aperte di integrazione con i metodi di Newton-Cotes. Si consideri, dunque, l'intervallo di integrazione  $U$ . Sia poi  $0 < \epsilon \ll e^{-9}$  tale che

$$U = U_\epsilon \cup \tilde{U}$$

dove  $U_\epsilon := (0, \epsilon)$  e  $\tilde{U} := [\epsilon, e^{-9}]$ . Le formule aperte di integrazione verranno applicate nell'intervallo  $U_\epsilon$ , nell'intorno destro di  $z = 0$ , ossia del punto singolare. Le formule chiuse, invece, saranno applicate sul compatto  $\tilde{U}$ , ora privo di punti singolari. Si noti poi che l'integrandà di (26) ammette primitiva esprimibile per mezzo di funzioni elementari, infatti

$$F(x) = \frac{1}{2} \int \log^2(x) dx = \frac{1}{2}x(\log^2(x) - 2\log(x) + 2) + c \quad \forall c \in \mathbb{R}$$

da cui segue che, all'interno di  $\tilde{U}$ , si avrà

$$I_6^{\tilde{U}} = F(e^{-9}) - F(\epsilon)$$

per teorema fondamentale del calcolo integrale. La conoscenza esplicita del valore vero della frazione di  $I_6$  all'interno dell'intervallo non singolare  $\tilde{U}$  permetterà di svolgere uno studio quantitativo sull'andamento dell'errore per i tre metodi di Newton-Cotes come nei precedenti casi, indipendentemente dall'utilizzo delle formule aperte di integrazione.

### Formule aperte

L'idea originaria per lo studio dell'integrale in esame, dunque, prevedeva di studiare separatamente l'andamento dell'errore in  $\tilde{U}$  e in  $U_\epsilon$ . In particolare, si è pensato di dedicare una sezione al solo studio del corretto andamento dell'errore per le formule aperte di integrazione. Nonostante non sia possibile disporre di formule estese, infatti, si è notato che

$$\begin{cases} h = \epsilon/2 & \text{se } N = 3 \\ h = \epsilon/5 & \text{se } N = 6 \end{cases}$$

D'altra parte, risolvendo un semplice integrale improprio, si ha che

$$I_6^{U_\epsilon} = F(\epsilon)$$

In altre parole, non è possibile variare  $h$  a parità di integrale come per le formule estese, ma è possibile disporre analiticamente del valore vero di  $I_6^{U_\epsilon}$  al variare dell'estremo  $\epsilon$  di integrazione. Siccome  $\epsilon$  è legato al passo  $h$  come sopra, segue che è possibile pensare di verificare il corretto andamento dell'errore calcolando le dispersioni di tanti integrali diversi al variare di  $\epsilon$ . In sostanza, se non è possibile variare l'ampiezza dei sotto-intervalli aumentando  $N$ , l'unico altro modo consiste nel variare gli estremi di integrazione. Seppur questa strada sembra sensata, svolgendo vari test si è notato che la stima dei coefficienti per l'errore restituisce risultati spesso non consistenti con quelli attesi. Ad esempio, l'applicazione di questo metodo a  $I_6^{U_\epsilon}$  produce un andamento al primo ordine in  $h$  per

entrambe le formule aperte. Di seguito sono mostrati gli andamenti ottenuti per  $N = 3$  e  $N = 6$  a confronto, al variare dell'estremo  $\epsilon$  nel range

$$10^{-9} \leq \epsilon < 0.5 \cdot 10^{-6} \quad \text{con} \quad \epsilon_{i+1} = \epsilon_i + 0.5 \cdot 10^{-9}$$

In particolare, il plot è stato ottenuto calcolando i logaritmi delle dispersioni e dei passi corrispondenti, al fine di visualizzare immediatamente i risultati con il massimo dettaglio.

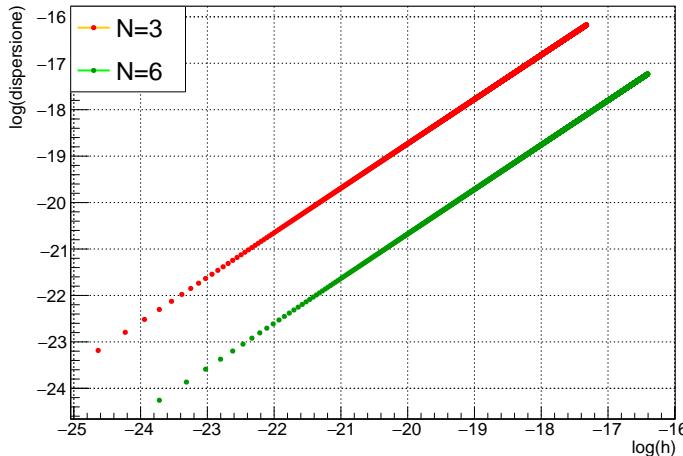


Figura 42: confronto dispersione formule aperte per  $N = 3$  e  $N = 6$

Come si nota dal grafico, la formula aperta per  $N = 6$  risulta sempre più precisa della formula aperta per  $N = 3$ , in quanto restituisce un risultato di dispersione dal valore vero sempre minore per ogni passo: questo fatto è consistente con quanto ci si aspetta. Tuttavia, come è possibile notare, gli andamenti definiscono due rette tra loro parallele, ossia caratterizzate dallo stesso coefficiente angolare. Inoltre, osservando la scala verticale e orizzontale di figura 42, non è difficile convincersi che tale coefficiente angolare abbia proprio valore unitario. Si potrebbe pensare ad un errore di implementazione delle formule aperte, ma la convergenza al valore atteso dei plot mostrati suggerisce che questa ipotesi risulti poco ragionevole. Si è anche pensato che il problema consistesse nell'applicazione delle formule aperte all'interno di un intervallo contenente una singolarità. Per tale ragione, si sono svolti diversi test con funzioni elementari in intervalli privi di punti singolari, ottenendo sempre risultati parzialmente o del tutto inconsistenti rispetto a quelli attesi. D'altra parte, le formule aperte sono pensate proprio per i casi di integrazione di funzioni con qualche singolarità nel dominio: sarebbe del tutto irragionevole pensare che il problema possa consistere nella presenza di punti di questo tipo. Per qualche ragione, allora, possiamo pensare che sia possibile studiare l'errore solo lavorando a parità di integrale, ossia solo se si dispone di formule estese.

### Trapezio

Al fine di verificare il corretto andamento dell'errore nell'intervallo non singolare con il metodo del trapezio si è fissato, arbitrariamente, l'estremo  $\epsilon = 10^{-6}$ ,

per poi procedere in modo usuale all'interno dell'intervallo  $\tilde{U}$ . Si sono quindi calcolate le dispersioni al variare di  $1/M$  nel range

$$300 \leq M < 1500 \quad \text{con} \quad M_{i+1} = M_i + 50$$

dove  $M$  è il numero di punti in cui si è diviso l'intervallo di integrazione. Di seguito si è mostrato quanto ottenuto.

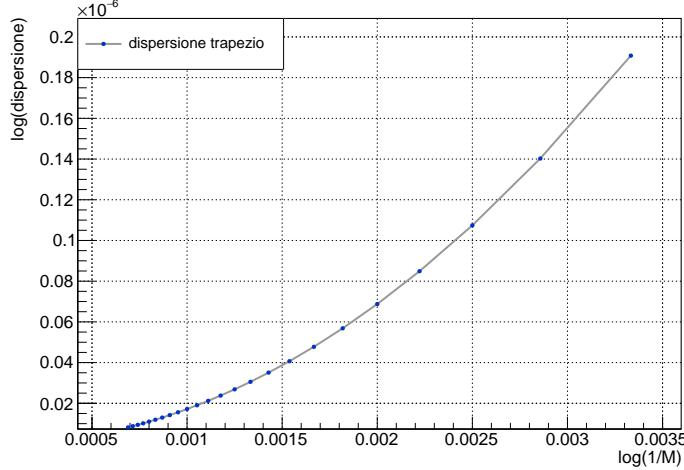


Figura 43: dispersione metodo del trapezio in  $\tilde{U}$

Come è possibile notare, la dispersione assume qualitativamente andamento parabolico, coerentemente con la (9). Si è quindi deciso di interpolare i dati raccolti passando ai logaritmi, provando quindi a verificare la relazione equivalente (10). In tal modo sarà anche possibile visualizzare con dettaglio eventuali andamenti non significativi. Si sono quindi interpolati i dati con una funzione lineare della forma  $y = p + mx$ . Di seguito sono mostrati i risultati ottenuti.

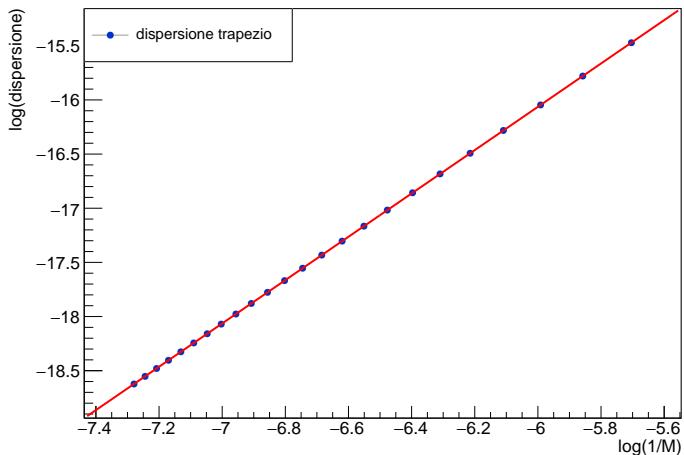


Figura 44: log dispersione metodo del trapezio in  $\tilde{U}$ : fit

Si è ottenuta la seguente stima dei parametri.

$$p = -4.06 = \log k \quad \text{e} \quad m = 2$$

Non è difficile notare che la stima del parametro  $m$  è consistente con il coefficiente angolare della relazione trovata analiticamente. Questo fatto garantisce la bontà del fit effettuato, rendendo possibile concludere che, nella formula di integrazione del trapezio esteso, l'errore scala come  $\frac{1}{M^2}$  per  $M$  grande, come ci si aspetta.

A questo punto, siamo interessati a studiare globalmente l'andamento della dispersione per l'integrale in esame al variare dell'estremo di integrazione. Si è allora deciso di studiare la dispersione aggiungendo il calcolo dell'integrale in  $U_\epsilon$  per mezzo delle formule aperte. In particolare, si è fissato  $M = 10000$  per la stima in  $\tilde{U}$  con il metodo del trapezio, variando  $\epsilon$  nel range

$$10^{-11} \leq \epsilon < 1.5 \cdot 10^{-8} \quad \text{con} \quad \epsilon_{i+1} = \epsilon_i + 0.5 \cdot 10^{-11}$$

ottenendo i seguenti risultati.

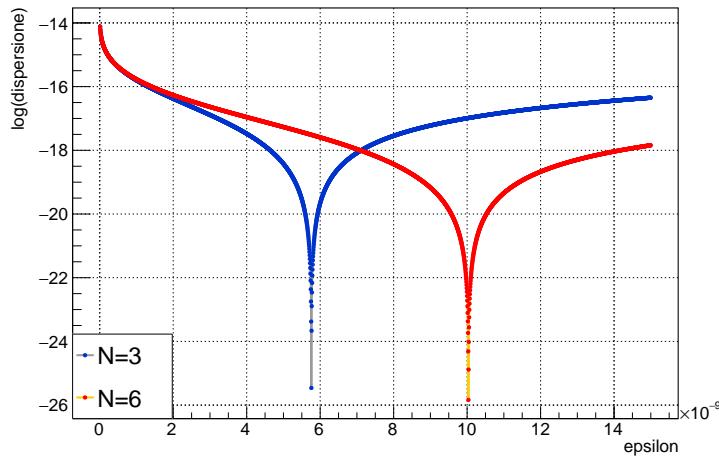


Figura 45: confronto log dispersione con trapezio in  $\tilde{U}$  e formule aperte in  $U_\epsilon$

Come si nota dalla figura, esiste un valore critico dell'estremo  $\epsilon$  tale che la stima di  $I_6$  raggiunga un massimo di precisione, ossia un minimo di dispersione dal valore vero. Questo fatto risulta vero per entrambe le formule aperte. In particolare, il valore di dispersione minimo si raggiunge per un valore di  $\epsilon \approx 10^{-8}$  con la più precisa formula aperta per  $N = 6$  punti. La ragione di questo andamento trova una spiegazione piuttosto intuitiva: al diminuire di  $\epsilon$ , se  $M$  è fissato, le formule aperte miglioreranno in precisione per quanto verificato in precedenza, ma la stima con il metodo in esame nell'intervallo non singolare peggiorerà, in quanto dall'equipartizione dell'intervallo seguirà un valore più grande del passo di integrazione. D'altra parte, all'aumentare di  $\epsilon$ , se  $M$  è fissato avremo un aumento in precisione della stima in  $\tilde{U}$ , ma un peggioramento della stessa all'interno dell'intervallo singolare  $U_\epsilon$  in quanto avremo un valore di  $h$  più grande. I punti di minimo degli andamenti in figura 45 rappresentano, dunque, i valori

dell'estremo di integrazione in cui i due effetti descritti si bilanciano, ottimizzando la stima dell'integrale. Ovviamente, il valore critico di  $\epsilon$  dipenderà dalla forma analitica della funzione integranda, ossia dalla rilevanza del contributo di area sottesa alla curva nell'intervallo singolare rispetto al contributo di area nell'intervallo non singolare. Inoltre, il valore critico dipenderà, come è naturale aspettarci, anche dal numero  $M$  di punti con cui si divide l'intervallo non singolare. Al fine di verificare questo fatto, si è deciso di svolgere la medesima operazione di studio globale della dispersione per il valore critico  $\epsilon = 10^{-8}$  fissato, al variare di  $M$  nel range

$$1000 \leq M < 18000 \quad \text{con} \quad M_{i+1} = M_i + 100$$

ottenendo i seguenti andamenti.

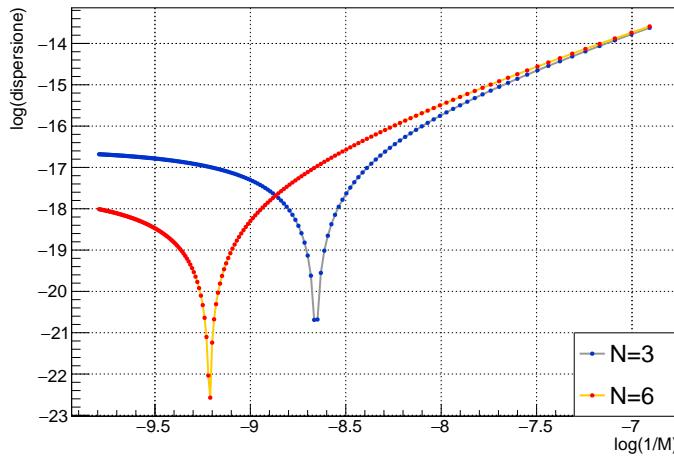


Figura 46: confronto log dispersione con trapezio in  $\tilde{U}$  e formule aperte in  $U_\epsilon$

Come si nota, il grafico ottenuto mostra un andamento simile al precedente. In altre parole, permette di verificare che la dispersione dal valore vero dipenda, a tutti gli effetti, anche dal numero di punti in cui si divide l'intervallo non singolare. Il fatto che la dispersione dipenda sia da  $\epsilon$  che da  $M$  rende lo studio particolarmente complicato in assenza del valore vero.

### Simpson

Al fine di verificare il corretto andamento dell'errore nell'intervallo non singolare con il metodo di Simpson si è fissato, arbitrariamente, l'estremo  $\epsilon = 10^{-6}$ , per poi procedere in modo usuale all'interno dell'intervallo  $\tilde{U}$ . Si sono quindi calcolate le dispersioni al variare di  $1/M$  nel range

$$401 \leq M < 2000 \quad \text{con} \quad M_{i+1} = M_i + 40$$

selezionando solo gli  $M$  dispari. In particolare, si è deciso di procedere immediatamente con il calcolo dei logaritmi e con un fit dei dati raccolti secondo la relazione affine  $y = p + mx$ , ottenendo quanto segue.

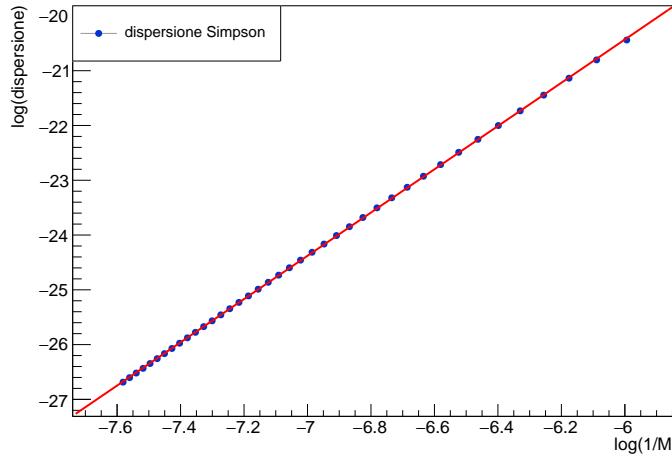


Figura 47: log dispersione metodo di Simpson in  $\tilde{U}$ : fit

Si è ottenuta la seguente stima dei parametri.

$$p = 3.26 = \log k \quad \text{e} \quad m = 3.95 \approx 4$$

La consistenza della stima del parametro  $m$  con il valore del coefficiente angolare atteso permette di concludere la verifica quantitativa del fatto che, nel metodo di Simpson, l'errore scala come  $\frac{1}{M^4}$  per  $M$  grande, come ci si aspetta.

Anche in questo caso, si è deciso di procedere per uno studio globale della dispersione di  $I_6$ , combinando il metodo in esame nell'intervallo non singolare con le formule aperte di integrazione nell'intervallo singolare. Si è considerato  $M = 10001$ , variando  $\epsilon$  nello stesso range utilizzato per lo studio precedente, ottenendo quanto segue.

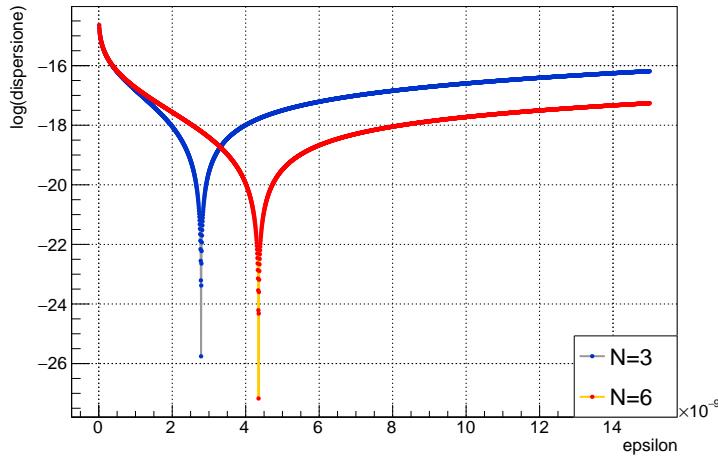


Figura 48: confronto log dispersione con Simpson in  $\tilde{U}$  e formule aperte in  $U_\epsilon$

Come è possibile notare, i risultati ottenuti sono del tutto analoghi a quelli analizzati in precedenza. In questo caso, la miglior stima per il valore di  $M$  selezionato si ha per un valore di  $\epsilon \approx 4.4 \cdot 10^{-9}$  con la più precisa formula aperta per  $N = 6$ .

### Romberg

Al fine di verificare il corretto andamento dell'errore nell'intervallo non singolare con il metodo di Romberg si è fissato, come nei punti precedenti, l'estremo  $\epsilon = 10^{-6}$ , per poi procedere in modo usuale all'interno dell'intervallo  $\tilde{U}$ . In particolare, si è calcolata la matrice di Romberg per un valore di

$$J_{max} = 20$$

Si è poi calcolato, per ogni valore di  $J$ , il numero  $M$  di punti in cui è stato diviso l'intervallo di integrazione, dato da

$$M = 2^J + 1$$

Al fine di studiare l'andamento della dispersione si sono quindi studiati i risultati per  $K = 2$ . In questo caso, si avrà allora

$$2K + 2 = 6$$

Si è deciso di eseguire il plot calcolando solo i logaritmi, al fine di visualizzare subito con più dettaglio variazioni grandi e piccole dei risultati. Di seguito si è mostrato quanto ottenuto.

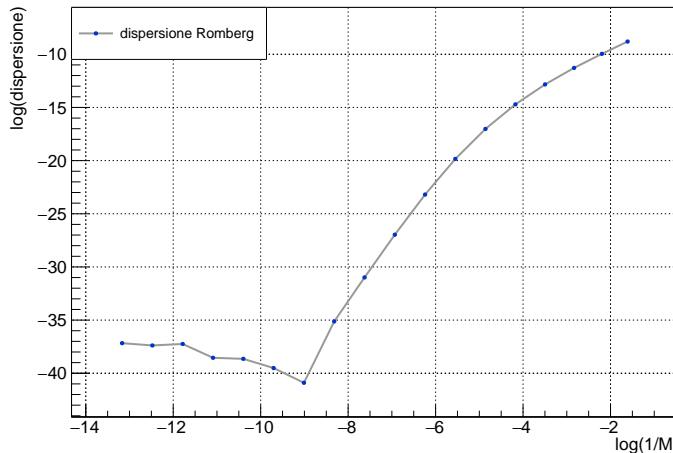


Figura 49: log dispersione metodo di Romberg per  $K = 2$

Anzitutto, è possibile notare che i problemi di arrotondamento dati dalla differenza di numeri molto vicini generano un andamento non significativo della dispersione a partire da un certo valore di  $M$ . Al fine di svolgere un fit lineare dei dati con la funzione  $y = p + mx$ , si è quindi selezionato un range di punti in cui l'andamento fosse qualitativamente rettilineo. In particolare, facendo riferimento alla figura 49, si sono selezionati i 5 punti tali che

$$-8.5 < \log(1/M) < -5.5$$

ottenendo quanto segue.

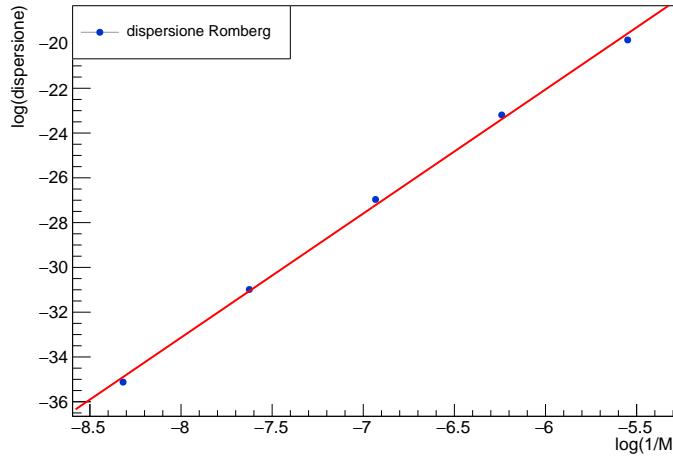


Figura 50: log dispersione metodo di Romberg per  $K = 2$ : fit

I parametri stimati risultano i seguenti.

$$p = 11.2 = \log k \quad \text{e} \quad m = 5.54 \approx 6$$

Il valore di  $m$  stimato permette di concludere la verifica della legge (16) che governa la dispersione del metodo per  $K = 2$ . La non esattezza della stima ottenuta dipende, chiaramente, dal range ristretto e, di conseguenza, dai pochi punti utilizzati per il fit.

Anche in questo caso, si è deciso di procedere per uno studio globale della dispersione di  $I_6$ , considerando  $J = K = 16$  e variando  $\epsilon$  nello stesso range utilizzato per lo studio precedente, ottenendo quanto segue.

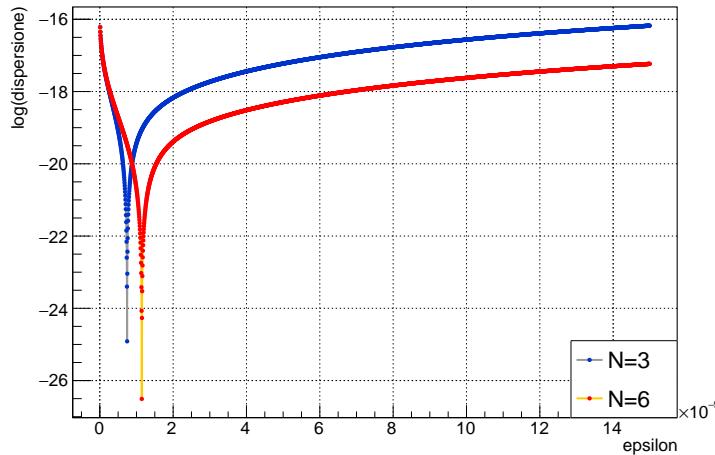


Figura 51: confronto log dispersione con Romberg in  $\tilde{U}$  e formule aperte in  $U_\epsilon$

Non è difficile accorgersi che i risultati ottenuti risultano analoghi a quelli ottenuti in precedenza. In questo caso, la miglior stima per il valore di  $M$  selezionato si ha per un valore di  $\epsilon \approx 1.2 \cdot 10^{-9}$  con la più precisa formula aperta per  $N = 6$ .

### Gauss

Si è calcolato  $I_6$  utilizzando il metodo di quadratura gaussiana. In particolare, si sono utilizzati i polinomi di Hermite per  $N = 2, 4, 8, 100$  punti e quelli di Laguerre per  $N = 2, 4, 8$  punti. Per il calcolo si sono utilizzati i valori tabulati degli zeri e dei pesi dei polinomi in esame, corrispondenti al numero di punti scelto.

### Hermite

Anzitutto, si noti che l'integrale in esame è definito sull'intervallo illimitato  $U := (3, +\infty)$ . Risulta allora necessario determinare un cambio di variabile in grado di mappare l'intervallo in esame su tutto l'asse reale  $(-\infty, +\infty)$ , ossia sull'intervallo di definizione dei polinomi di Hermite. Inoltre, tale trasformazione deve garantire di poter scrivere l'integranda come una certa funzione moltiplicata per il peso  $W$  dei polinomi in esame, al fine di garantire la massima efficienza del metodo. Si consideri allora la traslazione

$$y = x - 3$$

Agli estremi del dominio valgono le trasformazioni

$$y(3) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} y(x) = +\infty$$

La vecchia coordinata in funzione di quella nuova e il nuovo differenziale si scriveranno come

$$x = y + 3 \quad \text{e} \quad dx = dy$$

Da cui segue immediatamente che l'integrale in esame equivale all'integrale

$$I_6 = \int_0^{+\infty} (x+3)^5 e^{-(x+3)^2} dx$$

Si consideri, dunque, la funzione integranda

$$g(x) := (x+3)^5 e^{-(x+3)^2}$$

La mappa in esame non presenta alcuna simmetria rispetto agli assi coordinati, ma per definizione di valore assoluto vale sempre

$$g(|x|) = \begin{cases} g(x) & \text{se } x \geq 0 \\ g(-x) & \text{se } x < 0 \end{cases}$$

ossia  $g(|x|)$  è una funzione pari, simmetrica rispetto all'asse delle ordinate. Grazie alla precedente ridefinizione di  $I_6$  sul semiasse reale positivo possiamo quindi estendere l'integrale su tutto l'asse reale sfruttando la simmetria come

$$I_6 = \frac{1}{2} \int_{-\infty}^{+\infty} (|x|+3)^5 e^{-(|x|+3)^2} dx$$

Sviluppando il quadrato di binomio e sfruttando le proprietà delle potenze si ha

$$I_6 = \frac{1}{2} \int_{-\infty}^{+\infty} \frac{(|x|+3)^5}{\exp(6|x|+9)} e^{-x^2} dx$$

Avendo isolato la funzione peso siamo allora riusciti a riscrivere l'integrale nella forma desiderata per l'applicazione del metodo di Gauss. Si è quindi calcolato  $I_6$  utilizzando il metodo di Gauss con i polinomi di Hermite per  $N = 2, 4, 8, 100$  punti, ottenendo i risultati sintetizzati nella tabella che segue.

$N$	$\tilde{I}_6$	$\Delta_{G_H}$
2	0.00110	0.00513
4	0.00232	0.00391
8	0.00369	0.00254
100	0.00597	0.00026

Come si nota dai risultati ottenuti, la stima migliora progressivamente all'aumentare del numero di punti, ma molto lentamente. Inoltre, seppur l'errore assoluto non sia elevato è importante notare che il valore vero dell'integrale sia anch'esso molto piccolo. L'errore relativo percentuale per  $N = 100$  risulta essere

$$\varepsilon_r^{100} = \frac{\Delta_{G_H}^{100}}{I_6} \approx 4.2\%$$

tutt'altro che trascurabile. Evidentemente, visti i valori maggiori di errore assoluto per  $N = 2, 4, 8$  punti, in quei casi la stima di  $I_6$  può dirsi decisamente poco precisa. Chiaramente, il risultato è dovuto alla particolare complessità della funzione integranda, che di fatto genera una convergenza molto lenta della stima al valore vero. D'altra parte, si ha già avuto modo di notare che più l'integranda differisce da un polinomio, più il metodo di quadratura gaussiana fatica a convergere con rapidità.

### Laguerre

Al fine di garantire la massima efficienza, l'utilizzo dei polinomi di Laguerre necessita della presenza esplicita nell'integranda della funzione peso  $W$  rispetto alla quale vale l'ortogonalità. Si consideri, dunque, il cambio di variabile

$$y = x^2$$

evidentemente biettivo per tutti i valori di  $x$  positivi. Agli estremi del dominio, rispetto all'integrale originale, valgono le trasformazioni

$$y(3) = 9 \quad \text{e} \quad \lim_{x \rightarrow +\infty} y(x) = +\infty$$

La vecchia coordinata in funzione di quella nuova e il nuovo differenziale si scriveranno come

$$x = \sqrt{y} \quad \text{e} \quad dy = 2xdx$$

dove si è scelto solo il ramo positivo poiché  $x > 0$ . Dalle relazioni determinate segue immediatamente che l'integrale in esame equivale all'integrale

$$I_6 = \frac{1}{2} \int_9^{+\infty} x^2 e^{-x} dx$$

nel quale appare ora in modo esplicito il peso dei polinomi di Laguerre nell'integrandra. L'ultimo passaggio da compiere sarà allora quello di mandare  $(9, +\infty)$  nell'intero semiasse reale positivo, in cui sono presenti gli zeri dei polinomi utilizzati. Si consideri allora la traslazione

$$z = x - 9$$

Agli estremi del dominio si avrà

$$z(9) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} z(x) = +\infty$$

Per le vecchie variabili di integrazione avremo poi

$$x = z + 9 \quad \text{e} \quad dx = dz$$

Da cui segue immediatamente che l'integrale in esame equivale all'integrale

$$I_6 = \frac{1}{2} \int_0^{+\infty} (x+9)^2 e^{-(x+9)} dx$$

Sviluppando il quadrato di binomio e sfruttando le proprietà delle potenze si ha

$$I_6 = \frac{1}{2e^9} \int_0^{+\infty} (x+9)^2 e^{-x} dx$$

Siamo quindi riusciti a riformulare in modo equivalente l'integrale, nella forma corretta per l'applicazione del metodo di Gauss. Si è dunque calcolato  $I_6$  utilizzando i polinomi di Laguerre per  $N = 2, 4, 8$  punti, ottenendo i risultati sintetizzati nella tabella che segue.

$N$	$\tilde{I}_6$	$\Delta_{G_G}$
2	0.00623219510637732	$8.7 \cdot 10^{-19}$
4	0.00623219510637732	$8.7 \cdot 10^{-19}$
8	0.00623219510637732	$8.7 \cdot 10^{-19}$

Come si nota, il valore stimato risulta estremamente preciso per ogni  $N$ . Quanto ottenuto è dovuto al fatto che l'integrandra assume la forma

$$g(x) = P_2(x)W(x)$$

dove  $P_2$  è un polinomio di grado 2. Per la proprietà fondante del metodo di Gauss si ha allora un risultato esatto per ogni  $N \geq \tilde{N}$  tale che

$$2\tilde{N} - 1 = 2 \iff \tilde{N} = 2$$

poiché  $N$  è un intero. Ma allora, il risultato ottenuto, a meno dell'errore macchina, sarà da interpretarsi come una verifica dell'esattezza della stima. Anche il fatto di aver ottenuto stime identiche per ogni  $N$  trova quindi giustificazione: qualunque sia il valore del numero di punti, il metodo, per come è scritta l'integrandra, restituirà sempre la stessa stima esatta. L'utilizzo dei polinomi di Laguerre risulta allora, in questo caso, la strada migliore per la stima di  $I_6$ . Tuttavia, questo non è sempre vero, anche davanti ad uno stesso integrale. Si noti, infatti, che l'integrale in esame, sotto la sola traslazione

$$y = x - 3$$

può essere riscritto come

$$I_6 = \int_0^{+\infty} (x+3)^5 e^{-x^2-5x-9} e^{-x} dx$$

Anche in questa forma si ha una scrittura consistente con l'utilizzo dei polinomi di Laguerre, ma svolgendo il calcolo numerico si ottengono i seguenti risultati.

$N$	$\tilde{I}_6$	$\Delta_{G_G}$
2	0.00237	0.00386
4	0.005415	0.000817
8	0.006276	$4.3 \cdot 10^{-5}$

Come si nota, in questo caso, seppur si sia ricondotto  $I_6$  nella forma opportuna per l'utilizzo dei polinomi di Laguerre, la stima appare decisamente poco precisa per tutti i punti selezionati rispetto ai risultati della tabella precedente. Evidentemente, questo è dovuto al fatto che la funzione integranda appare ora scritta come la funzione peso  $W$  moltiplicata per una funzione molto diversa da un polinomio. Per quanto spiegato in precedenza, allora, il metodo faticherà molto di più a convergere rispetto al banale caso polinomiale dato dal precedente cambio di variabile.

Abbiamo allora ottenuto un risultato importante: per un problema numerico di calcolo di un integrale con la quadratura di Gauss è fondamentale procedere lucidamente con le manipolazioni algebriche e analitiche al fine di ottenere una riformulazione ottimale. Non è allora del tutto corretto affermare che un set di polinomi ortogonali sia migliore di un altro "in toto", ma è sempre necessario relazionare i polinomi utilizzati con la forma in cui si è ricondotto l'integrale in esame. In sostanza, la ragione di quanto detto consiste nel fatto che

$$\int_a^b f(x) dx = \int_a^b g(x) dx \Rightarrow f(x) = g(x)$$

da cui segue che è sempre bene chiedersi se la trasformazione determinata per riscrivere un integrale sia quella ottimale per il set di polinomi ortogonali considerato. Un altro fatto rilevante consiste nel notare che se l'integrale originale presenta esplicitamente una certa funzione peso  $W$  per un set di polinomi, non è sempre vero che l'utilizzo di tale set sia la scelta migliore. Ne è un chiaro esempio il confronto tra i risultati ottenuti con Hermite e i primi risultati ottenuti con Laguerre: nel secondo caso, la stima risulta addirittura esatta per ogni  $N$ , a meno della precisione doppia utilizzata per il calcolo.

## Esercizio 7

Si vuole stimare numericamente il valore dell'integrale

$$I_7 := \int_0^{+\infty} x^{14} e^{-x^2} dx$$

utilizzando il metodo di quadratura gaussiana con i polinomi di Hermite e di Laguerre per  $\alpha = 0$ .

Analiticamente (facendo uso di un opportuno cambio di variabile) o con l'aiuto di un calcolatore avanzato, si ricava che il valore esatto è

$$I_7 = \frac{1}{256}(135135\sqrt{\pi}) \approx 935.627152898894173238039$$

Siamo ora in grado di calcolare le dispersioni delle stime dal valore vero.

### Hermite

Anzitutto, si noti che la funzione integranda si scrive come

$$f(x) = P_{14}(x)e^{-x^2}$$

ossia presenta la forma di una funzione polinomiale moltiplicata per la funzione peso  $W$  dei polinomi di Hermite. I polinomi di Hermite sono definiti in  $(-\infty, +\infty)$ : l'unico passaggio analitico da compiere risulta, dunque, quello di scrivere un integrale equivalente a  $I_7$  che si estenda su tutto l'asse reale, nel quale sono presenti gli zeri dei polinomi in esame. Si noti, dunque, che la funzione integranda  $f$  è tale che

$$f(-x) = f(x) \quad \forall x \in \mathbb{R}$$

ossia risulta una funzione pari definita su tutto  $\mathbb{R}$ . Dalla simmetria rispetto all'asse delle ordinate segue che

$$I_7 = \frac{1}{2} \int_{-\infty}^{+\infty} x^{14} e^{-x^2} dx$$

La riscrittura dell'integrale come un integrale improprio su tutto  $\mathbb{R}$  di una funzione integranda contenete il peso dei polinomi di Hermite consente ora l'applicazione diretta della (18). Si è quindi calcolato  $I_7$  utilizzando il metodo di Gauss con i polinomi di Hermite per  $N = 2, 4, 8$  punti, ottenendo i risultati sintetizzati nella tabella che segue.

$N$	$\tilde{I}_7$	$\Delta_{G_H}$
2	0.007	935.620
4	91	845
8	935.627152898894	$1.14 \cdot 10^{-13}$

Dai risultati ottenuti è evidente che il metodo di Gauss, in questo caso, restituisca risultati incompatibili con il valor vero per un numero di punti pari a  $N = 2$  e  $N = 4$ . A meno della doppia precisione utilizzata restituisce, invece, il valore

esatto per  $N = 8$  punti. Quanto ottenuto è consistente con quanto ci si aspetta: il metodo di Gauss garantisce l'esattezza della stima quando la funzione  $f$  della (17) assume la forma di un polinomio di grado minore o pari a  $2N - 1$ , dove  $N$  è il numero di punti utilizzato. Nel caso in esame, il polinomio che moltiplica il peso di Hermite è di grado 14. Segue che l'esattezza della stima si ha  $\forall N \geq \tilde{N}$  tale che

$$2\tilde{N} - 1 = 14 \iff \tilde{N} = 8$$

poiché  $N$  è un intero. I risultati incompatibili ottenuti per  $N = 2, 4$  punti trovano quindi spiegazione con il fatto che, per tali valori, il grado del polinomio della funzione integranda risulta ancora troppo distante dal numero di punti utilizzato. Il fatto che la stima migliori di una piccola quantità per  $N = 4$  rispetto a  $N = 2$  è comunque consistente con quanto appena spiegato. Infine, l'esattezza della stima risulta anch'essa in linea con la proprietà cardine del metodo di quadratura gaussiano:  $N = 8$  punti è il primo intero che consente di ottenere, come si è appena mostrato, una stima esatta.

### Laguerre

I polinomi di Laguerre per  $\alpha = 0$  sono definiti sul semiasse reale  $(0, +\infty)$ . In questo caso, dunque, l'intervallo di integrazione di  $I_7$  si presenta già nella forma corretta per l'applicazione del metodo con i polinomi in esame. Risulta dunque necessario determinare un cambio di variabile, al fine di poter riscrivere la funzione integranda come il prodotto di una certa funzione  $g$  con il peso  $W(x) = e^{-x}$  dei polinomi di Laguerre, mantenendo inalterato l'intervallo di integrazione. Si consideri, dunque, il cambio di variabile

$$z = x^2$$

ben posto in quanto invertibile e differenziabile con continuità nell'intervallo di integrazione  $(0, +\infty)$ . Evidentemente, valgono le seguenti trasformazioni agli estremi del dominio:

$$z(0) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} z(x) = +\infty$$

In altre parole, il dominio di integrazione rimane inalterato, come da richiesta al fine di operare con i polinomi in esame. Differenziando il cambio di variabile e invertendo la trasformazione si ha

$$dz = 2x dx \quad \text{e} \quad x = \sqrt{z}$$

Si noti che si è scelto il solo ramo positivo in quanto  $x \in (0, +\infty)$ . Operando le sostituzioni esplicitate, segue immediatamente che l'integrale può essere riscritto nella forma desiderata come

$$I_7 = \frac{1}{2} \int_0^{+\infty} \sqrt{z^{13}} e^{-z} dz \tag{27}$$

La riscrittura dell'integranda come una funzione contenete il peso dei polinomi di Laguerre consente ora l'applicazione diretta della (18). Si è quindi calcolato  $I_7$  utilizzando il metodo di Gauss con i polinomi di Laguerre per  $N = 2, 4, 8, 24, 64$  punti, ottenendo i seguenti risultati.

$N$	$\bar{I}_7$	$\Delta_{G_G}$
2	214	721
4	936.106	0.479
8	935.627230	0.000143
24	935.62715291	$1.15 \cdot 10^{-8}$
64	935.6271528989	$1.57 \cdot 10^{-12}$

Come si nota dai risultati ottenuti, la stima dell'integrale appare migliorare progressivamente all'aumentare del numero di punti utilizzato, come ci si aspetta. Nel caso in esame, la funzione che moltiplica il peso dei polinomi di Laguerre a seguito del cambio di variabile non è più una funzione polinomiale, ma si presenta nella forma di una radice di ordine pari del tipo

$$g(x) = \sqrt{x^{13}}$$

Segue che l'esattezza della stima utilizzando la proprietà centrale del metodo non è più possibile. Il fatto che il risultato ottenuto per  $N = 64$  punti risulti essere dello stesso ordine di grandezza dei risultati che in precedenza si sono interpretati come esatti dipende unicamente dalla limitatezza della precisione doppia utilizzata per il calcolo: in questo caso, la forma analitica di  $g$  è tale da consentire il raggiungimento della precisione macchina per  $N = 64$  punti.

In definitiva, l'utilizzo dei polinomi di Hermite risulta essere, in questo caso, più conveniente da un punto di vista computazionale rispetto all'utilizzo dei polinomi di Laguerre. Il cambio di coordinate è, infatti, quasi immediato nel primo caso. Ma l'aspetto più importante consiste nel fatto che il numero minimo di somme, e quindi di iterazioni, che deve essere effettuato per ottenere una stima esatta con i polinomi di Hermite è  $N = 8$ . La stima con Laguerre, invece, a causa della forma analitica non polinomiale di  $g$ , non consente l'esattezza del risultato e risulta ancora lontana dalla precisione macchina per  $N = 24$  punti.

## Numeri casuali e densità di probabilità

I numeri casuali svolgono un ruolo chiave in diversi ambiti della matematica e della fisica. Basti pensare che, ad esempio, il risultato di ogni osservazione sperimentale può essere pensato come un numero casuale distribuito in un certo modo intorno al valore vero della grandezza misurata. La probabilità e la statistica si occupano dello studio e della formalizzazione di questi oggetti.

Dall'esperienza comune risulta naturale pensare che esistano fenomeni con la proprietà di accadere o meno, per le ragioni più disparate.

**Definizione 0.5.** Chiameremo *evento casuale* o *aleatorio* un qualunque evento o fenomeno che può accadere o non accadere.

Se a questo evento è possibile associare univocamente un numero reale, come nel caso del risultato di una misura sperimentale, tale evento prende il nome di *numero casuale*.

**Definizione 0.6.** Chiameremo *spazio delle fasi* di una misura l'insieme  $\Omega$  di tutti i possibili risultati della misura stessa.

Un esempio classico è quello del lancio di una moneta non truccata, per il quale lo spazio delle fasi è l'insieme discreto e finito  $\Omega = \{\text{testa, croce}\}$ .

**Definizione 0.7.** Chiameremo *campione* l'insieme

$$\{x_i\}_{i=1,\dots,N}$$

delle  $N$  misure effettuate. Chiameremo *popolazione* il limite di tutte le possibili misure che possono essere effettuate, ossia

$$\lim_{N \rightarrow +\infty} \{x_i\}_{i=1,\dots,N}$$

Evidentemente, dalle definizioni date segue che il campione è sempre un sottoinsieme della popolazione. Più il numero di dati è elevato, più si hanno informazioni circa il fenomeno oggetto dello studio. In altre parole, l'aumento della dimensione  $N$  di un campione è sempre associato ad un aumento in *precisione* della misura. Nell'esempio del lancio di una moneta, un possibile campione può essere rappresentato dall'insieme  $\{T, T, C, T, C, C, C\}$ . Ogni evento di natura casuale ha una certa probabilità di accadimento, in funzione del fenomeno che si sta studiando. In particolare, la nozione di probabilità può essere formalizzata in modo costruttivo come segue.

**Definizione 0.8** (assiomatica di Kolmogorov). Sia  $E$  un evento casuale e  $\Omega$  lo spazio delle fasi associato. Chiameremo *probabilità* di  $E$  un numero  $P(E) \in \mathbb{R}$  tale che

- $P(E) \geq 0 \quad \forall E \subseteq \Omega$
- $P(\Omega) = 1$
- $P(E_i \cup E_j) = P(E_i) + P(E_j) \quad \forall i, j \text{ t.c. } E_i \cap E_j = \emptyset \text{ con } i \neq j$

Evidentemente, la definizione assiomatica di probabilità ha scarsa utilità operativa. Per operare con le probabilità, solitamente, vengono utilizzate altre due definizioni: quella *a priori* e quella *a posteriori*, a seconda del fatto che si disponga della popolazione o solo di un campione della grandezza in esame. La definizione di Kolmogorov, tuttavia, è essenziale per la verifica della consistenza di diverse definizioni di probabilità. Ogni volta che viene definita una probabilità, infatti, per dirsi tale dovrà rispettare le tre condizioni della 0.8.

A seconda della cardinalità di  $\Omega$  risulta possibile costruire alcuni oggetti utili che consentono di trattare i numeri casuali con gli strumenti dell'analisi matematica standard. Accenniamo quindi a due costruzioni di grande importanza per gli esercizi che seguono.

### Distribuzioni continue e discrete

Le misure fisiche o le stime numeriche sono spesso numeri casuali definiti in uno spazio delle fasi con cardinalità del continuo. Siamo allora interessati alla costruzione di uno strumento che permetta di descrivere la distribuzione di questi numeri e la loro probabilità di accadimento limitandoci, per semplicità, al solo caso unidimensionale.

**Definizione 0.9.** Sia  $X$  una variabile casuale continua e  $\Omega \subseteq \mathbb{R}$  lo spazio delle fasi. Chiameremo *funzione densità di probabilità (pdf)* di  $X$  l'applicazione  $f : \Omega \rightarrow \mathbb{R}$  non negativa e integrabile secondo Lebesgue tale che

$$P(X \in A \subseteq \Omega) = \int_A f(x) dx \quad (28)$$

Dalla definizione 0.9 si ha che un modo analogo e più esplicito di definizione di una pdf unidimensionale sarà

$$f(x) := \frac{dP}{dx}$$

Spesso la funzione  $P(x) \equiv F(x)$  è detta *funzione cumulativa* della densità di probabilità  $f$ . La pdf permette di descrivere l'andamento di una variabile casuale continua: dato un campione di misure, infatti,  $f$  si sovrappone all'andamento dell'istogramma normalizzato delle frequenze relative di accadimento. Si noti che la probabilità (28) verifica il primo assioma di Kolmogorov per definizione, e il terzo per linearità dell'integrale di Lebesgue, in questo caso coincidente con quello di Riemann. Affinché la probabilità appena introdotta possa dirsi ben definita basterà allora assicurarsi che

$$\int_{\Omega} f(x) dx = 1 \quad (29)$$

ossia che la funzione densità di probabilità  $f$  sia *normalizzata* su tutto il suo insieme di definizione.

Evidentemente, possono verificarsi casi in cui una misura sia rappresentata da numero casuale definito in un insieme discreto, ossia un insieme, con cardinalità finita o infinita, composto soltanto da punti isolati. Anche in questo caso è possibile definire una densità come segue.

**Definizione 0.10.** Sia  $X$  una variabile casuale discreta e  $\Omega \subseteq \mathbb{R}$  lo spazio delle fasi. Chiameremo *distribuzione discreta di probabilità* di  $X$  l'applicazione  $f : \Omega \rightarrow \mathbb{R}$  non negativa tale che

$$P(X \in A \subseteq \Omega) = \sum_{x \in A} f(x) \quad (30)$$

Anche in questo caso, affinché la definizione di densità discreta possa dirsi una definizione consistente di probabilità è necessario che siano verificati i tre assiomi di Kolmogorov.

Esistono diverse quantità che possono essere definite per lo studio delle caratteristiche di una pdf o di una distribuzione discreta, come i momenti e i momenti centrali. Inoltre, esiste un teorema di capitale importanza per lo studio dei fenomeni casuali: il teorema centrale del limite. Tutti questi fatti verranno discussi, analizzati e applicati nel dettaglio negli esercizi che seguono.

### Sequenze pseudo-casuali e metodo Monte Carlo

Alla luce dell'utilità dei numeri casuali in diversi campi di studio risulta importante saper simulare la loro generazione al fine, ad esempio, di simulare una misura sperimentale al calcolatore. Seppur il calcolatore sia una macchina deterministica è infatti possibile generare una sequenza

$$\{x_i\}_{i=1,\dots,N}$$

di  $N$  numeri distribuiti secondo le più diverse distribuzioni di probabilità  $f$ . A tale scopo esistono diverse funzioni nelle librerie di programmazione standard in grado, ad esempio, di generare numeri distribuiti uniformemente. Un esempio di implementazione per un generatore di numeri uniformi è il *generatore lineare congruenziale*, definito ricorsivamente come

$$x_{n+1} = (ax_n + c) \mod m$$

dove  $a$ ,  $c$  e  $m$  sono opportuni coefficienti interi e positivi. Il dato iniziale  $x_0$  definisce il *seme* della sequenza, e deve essere cambiato ad ogni esecuzione per ottenere sequenze diverse. Ovviamente, la natura deterministica di un calcolatore non permette di generare veri numeri casuali: i numeri sono sempre generati da funzioni, e quindi da algoritmi che svolgono calcoli definiti. Un buon generatore di sequenze casuali è in grado di nascondere, qualitativamente, la struttura funzionale dietro alla generazione. Ogni generatore è quindi caratterizzato da un *periodo*, ossia da un numero minimo di valori tale che la sequenza non si ripeta uguale a se stessa. Quanto più un generatore è buono, tanto più il periodo è elevato. Per tutte queste ragioni, le sequenze casuali prodotte da un algoritmo computazionale sono dette *sequenze pseudo-casuali*. In seguito, a partire dalla generazione uniforme, verranno discusse diverse tecniche per ottenere numeri distribuiti secondo densità  $f$  a piacere.

Oltre ai casi di simulazione di misure sperimentali, rilevanti per la verifica di modelli teorici, le sequenze pseudo-casuali trovano applicazione in analisi numerica. In particolare, è possibile sfruttare il loro utilizzo per il calcolo di

integrali di funzioni reali. I metodi numerici che sfruttano la generazione di sequenze casuali sono detti *metodi Monte Carlo*. Si consideri, infatti, il problema del calcolo di un integrale monodimensionale

$$I := \int_a^b g(x) dx$$

con  $a < \infty$  e  $b < \infty$ . Ricordando che l'integrale di Riemann è definito come il limite comune di due successioni delle somme parziali è evidente che un'approssimazione di  $I$  sarà data da

$$I \approx \frac{b-a}{N} \sum_{i=1}^N g(x_i) \quad \text{con} \quad f(x) = U(a, b) \quad (31)$$

dove  $U(a, b)$  rappresenta la distribuzione uniforme nell'intervallo di integrazione. Il metodo numerico dato dalla (31) va sotto il nome di *Monte Carlo sampling*. In generale, i metodi Monte Carlo, sfruttando una generazione casuale, possono essere meno precisi dei metodi di Newton-Cotes a parità di punti generati. Tuttavia, i metodi che sfruttano la generazione di sequenze casuali presentano due notevoli vantaggi: anzitutto, sono molto più semplici da utilizzare in presenza di domini di integrazione complicati: l'equipartizione deterministica di intervalli non standard, infatti, può presentare notevoli difficoltà. Il vantaggio principale, tuttavia, consiste nel fatto che questi metodi possono essere estesi al calcolo di integrali multidimensionali di funzioni in più variabili senza impattare in modo considerevole sulla complessità computazionale dell'algoritmo, e quindi sul tempo di esecuzione. Si avrà poi modo di osservare che lo svantaggio della poca precisione della stima Monte Carlo potrà essere aggirato effettuando un campionamento ad importanza, ossia generando numeri con distribuzioni piccate in corrispondenza dei massimi della funzione integranda. L'analisi di tutti questi fatti sarà oggetto degli esercizi che seguono.

## Esercizio 8

Si vuole stimare numericamente il volume della sfera unitaria  $M$  dimensionale utilizzando il metodo deterministico del midpoint e il metodo Monte Carlo, per poi operare un confronto in efficienza.

L'ipersfera unitaria centrata nell'origine di un sistema cartesiano è definita come

$$S_M := \left\{ (x_1, \dots, x_M) \in \mathbb{R}^M \mid \sum_{i=1}^M x_i^2 = 1 \right\} \quad (32)$$

e rappresenta una varietà  $M - 1$  dimensionale immersa in uno spazio  $M$  dimensionale. Posto  $\vec{x} := (x_1, \dots, x_M)$  e ricordando la definizione di norma euclidea, la (32) può essere riscritta in modo più compatto come

$$S_M = \left\{ \vec{x} \in \mathbb{R}^M \mid \|\vec{x}\| = 1 \right\}$$

L'ipersfera  $S_M$  rappresenta quindi l'insieme di frontiera della palla  $M$  dimensionale, definita banalmente come

$$P_M := \left\{ \vec{x} \in \mathbb{R}^M \mid \|\vec{x}\| \leq 1 \right\}$$

Siamo allora interessati al calcolo della misura di Lebesgue dell'insieme  $P_M$ , ossia al calcolo dell'integrale multidimensionale

$$\begin{aligned} V_M := \mathcal{L}(P_M) &= \int_{P_M} dx_1 \dots dx_M = \\ &= \int_{P_{M-1}} dx_1 \dots dx_{M-1} \sqrt{1 - \sum_{i=1}^{M-1} x_i^2} \end{aligned} \quad (33)$$

esprimendo una variabile della (32) in funzione delle altre  $M - 1$ . Si dimostra che analiticamente vale

$$V_M = \frac{\pi^{M/2}}{\Gamma(\frac{M}{2} + 1)} = \begin{cases} \frac{\pi^{M/2}}{\frac{M!}{2}} & \text{se } M \text{ pari} \\ \frac{2^{\frac{M+1}{2}} \pi^{\frac{M-1}{2}}}{M!!} & \text{se } M \text{ dispari} \end{cases} \quad (34)$$

dove  $\Gamma$  è la funzione Gamma di Eulero e il doppio fattoriale di  $M$  è definito come il prodotto tra tutti i naturali pari che lo precedono se  $M$  è pari, tra tutti i naturali dispari se  $M$  è dispari.

### Stabilità dei valori veri

Anzitutto, notando che la (34) può essere scritta in due forme equivalenti, si è deciso di verificare quale dei due metodi fosse più opportuno utilizzare, da un punto di vista computazionale, come metodo di stima del valore vero del volume di  $S_M$ . Si sono quindi implementate entrambe le forme del risultato analitico, definendo separatamente le funzioni per il calcolo del fattoriale e del doppio fattoriale per la seconda forma, che indicheremo con  $V_M^2$ . Per la prima forma, che indicheremo con  $V_M^1$ , il calcolo è stato eseguito staticamente utilizzando la

funzione Gamma implementata nelle librerie standard. Si è quindi calcolato il volume nel range di dimensioni

$$1 \leq M < 49 \quad \text{con} \quad M_{i+1} = M_i + 1$$

Si sono poi plottate le coppie  $(M, V_M^1)$  e  $(M, V_M^2)$  sovrapposte, ottenendo il seguente risultato.

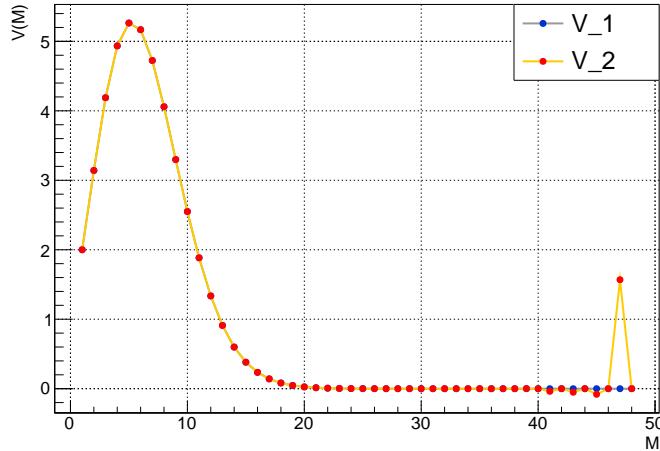


Figura 52: plot  $(M, V_M^1)$  e  $(M, V_M^2)$  a confronto

Come è possibile notare,  $V_M^2$  presenta un comportamento inatteso a partire da un certo valore di  $M$ , scostandosi da  $V_M^1$ , che invece sembra mantenere una certa regolarità. Ipotizzando che l'andamento corretto fosse rappresentato da  $V_M^1$ , si è allora deciso di svolgere un fit dei dati  $(M, V_M^1)$  secondo la funzione  $y = \frac{a^{M/2}}{\Gamma(\frac{M}{2} + 1)}$  con  $a$  parametro libero, ottenendo quanto segue.

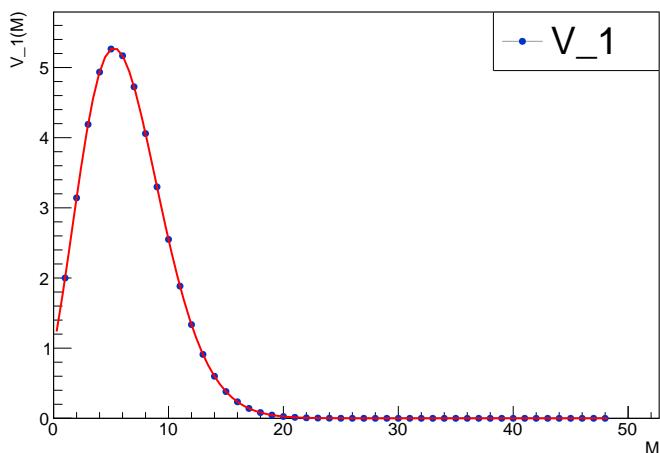


Figura 53: grafico  $(M, V_M^1)$ : fit

Si è ottenuta la stima del parametro

$$a = 3.14 \approx \pi$$

La compatibilità del parametro stimato con il valore atteso conferma il corretto andamento della stima  $V_M^1$ . Stampando a schermo i valori di  $|V_M^1 - V_M^2|$  al variare di  $M$  si osserva che la seconda stima inizia a fluttuare restituendo valori insensati per  $M$  tale che  $M > 34$ . Svolgendo diverse prove di controllo si è notato che le funzioni per il calcolo del fattoriale e del doppio fattoriale iniziano a restituire valori inconsistenti a partire dal valore di  $M$  critico individuato. Questo fatto può essere spiegato ricordando che i numeri interi, a causa della limitatezza della memoria del calcolatore, hanno un range massimo di rappresentabilità all'interno della memoria di lavoro. Le funzioni per il calcolo del fattoriale e del doppio fattoriale producono interi che crescono più rapidamente di un esponenziale, raggiungendo velocemente l'intero massimo rappresentabile dal calcolatore. Infatti, sostituendo il tipo di dato *long int* con il tipo *int*, che presenta una limitatezza maggiore di rappresentazione, operando la stessa operazione di stampa a schermo si osserva che la stima  $V_M^2$  perde di significatività a partire da un valore più piccolo, di  $M = 20$ , in linea con le considerazioni fatte. Possiamo allora affermare che il metodo dato da  $V_M^2$  per la stima del volume dell'ipersfera unitaria risulta (definitivamente) instabile per il problema in esame. Si è quindi assunta, da qui fino alla fine dell'esercizio, la stima  $V_M^1$  come valore vero del volume di  $S_M$ .

### Estensione multidimensionale e algoritmi

Il problema numerico in esame consiste nella stima dell'integrale (33). Si consideri, dunque, l'applicazione

$$\Phi_{M-1}(x_1, \dots, x_{M-1}) := \sqrt{1 - \sum_{i=1}^{M-1} x_i^2}$$

Al fine di rendere il problema più comodo a livello computazionale è possibile definire la funzione reale di  $M-1$  variabili reali

$$S_{M-1}(x_1, \dots, x_{M-1}) := \begin{cases} \Phi_{M-1}(\vec{x}) & \text{se } \|\vec{x}\| \leq 1 \\ 0 & \text{altrimenti} \end{cases} \quad (35)$$

che rappresenta una porzione di ipersfera estesa su tutto  $\mathbb{R}^{M-1}$ . Dall'estensione appena mostrata seguirà banalmente che l'integrale (33) oggetto dello studio si potrà scrivere come

$$V_M = 2^M \int_0^1 \cdots \int_0^1 S_{M-1}(x_1, \dots, x_{M-1}) dx_1 \dots dx_{M-1} \quad (36)$$

per teorema di Fubini combinato alla simmetria della sfera  $M$  dimensionale. Il problema numerico si riduce, allora, a meno di un fattore di riscalamento  $2^M$ , al calcolo dell'integrale  $M-1$  dimensionale di (35) sul dominio  $M-1$  dimensionale rappresentato dall'iperquadrato

$$Q_{M-1} := \underbrace{[0, 1] \times \dots \times [0, 1]}_{M-1 \text{ volte}} = \prod_{i=1}^{M-1} [0, 1]$$

Risulta dunque necessario estendere i metodi numerici di integrazione dal caso unidimensionale al caso del calcolo di integrali in più dimensioni.

Il metodo di integrazione del *midpoint*, nel caso monodimensionale, consiste nel dividere il dominio di integrazione  $(a, b)$  in  $N$  intervalli della stessa misura. Definendo il passo in modo usuale come

$$h := \frac{b - a}{N}$$

vengono generati deterministicamente  $N$  punti all'interno di  $(a, b)$  secondo

$$x_i = a + ih + \frac{h}{2}$$

ossia in corrispondenza del punto medio di ogni sotto-intervallo in cui è stato diviso il dominio. Il metodo consiste quindi nel valutare la funzione integranda  $f$  in ogni punto generato, e ricordando la solita definizione di integrale  $I$  come limite comune di due successioni, si avrà che la somma di tutti i rettangoli

$$I \approx \frac{b - a}{N} \sum_{i=1}^N f(x_i)$$

consentirà una stima del valore dell'integrale, tanto migliore quanto più è grande il numero  $N$  di punti. Si dimostra che l'errore commesso è del tutto equivalente all'errore del metodo del trapezio. Si noti che il metodo del midpoint, a differenza di gran parte dei metodi Newton-Cotes, consente una stima senza il calcolo del valore della funzione agli estremi del dominio. L'idea dell'estensione del midpoint nel caso dell'integrale in esame sarà allora quella di dividere l'iperquadrato  $Q_{M-1}$  in  $N^{M-1}$  iperquadратi, per poi generare lo stesso numero di  $(M-1)$ -uple centrate in ognuno di essi. In tal modo, ricordando la definizione di integrale multidimensionale, si avrà che

$$V_{\text{mid}} \approx 2^M \frac{b_1 - a_1}{N_1} \dots \frac{b_{M-1} - a_{M-1}}{N_{M-1}} \sum_{n_1=1}^{N_1} \dots \sum_{n_{M-1}=1}^{N_{M-1}} S_{M-1}(x_{n_1}, \dots, x_{n_{M-1}})$$

Tenendo conto del fatto che  $(b_1 - a_1) \dots (b_{M-1} - a_{M-1}) = 1$  nel caso in esame, ossia che la misura di un iperquadrato di lato unitario è unitaria, possiamo riscrivere l'approssimazione numerica in modo più compatto come

$$V_{\text{mid}} \approx \frac{2^M}{N^{M-1}} \sum_{n_1=1}^{N_1} \dots \sum_{n_{M-1}=1}^{N_{M-1}} S_{M-1}(x_{n_1}, \dots, x_{n_{M-1}}) \quad (37)$$

Si noti che l'implementazione della (37) da un punto di vista computazionale consiste in  $M-1$  cicli di iterazione annidati.

L'estensione al caso multidimensionale del metodo Monte Carlo, invece, consiste nel generare  $(M-1)$ -uple casuali  $\vec{x}_i$  con distribuzione uniforme all'interno dell'iperquadrato complessivo di integrazione  $Q_{M-1}$ . Segue che una stima del valore dell'integrale sarà data da

$$V_{\text{MC}} \approx 2^M \frac{V^{(M-1)}}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} S_{M-1}(\vec{x}_i)$$

dove  $V^{(M-1)} := (b_1 - a_1) \dots (b_{M-1} - a_{M-1}) = 1$  per le ragioni già discusse, e  $N_{\text{tot}}$  è il numero totale di punti generati in  $Q_{M-1}$ . Possiamo quindi scrivere in modo più compatto l'approssimazione anche in questo caso come

$$V_{\text{MC}} \approx \frac{2^M}{N_{\text{tot}}} \sum_{i=1}^{N_{\text{tot}}} S_{M-1}(\vec{x}_i) \quad (38)$$

In questo caso, la stima data dall'algoritmo (38) presenta una sola sommatoria: risulta allora naturale pensare che la sua implementazione, nei casi multidimensionali, sia molto più vantaggiosa rispetto all'utilizzo del midpoint o di altri metodi deterministici in termini di costo computazionale. Si noti, invece, che nel caso particolare in cui  $M - 1 = 1$  il metodo del midpoint e quello di Monte Carlo, di fatto, coincidono in termini di tempi di calcolo.

Si sono quindi implementati i due metodi discussi, dati dalla (37) e dalla (38). L'implementazione del Monte Carlo multidimensionale risulta piuttosto semplice e naturale: l'idea è quella di fare un ciclo sul numero di punti totale  $N_{\text{tot}}$  e generare, per ogni punto, le  $M - 1$  coordinate corrispondenti in  $Q_{M-1}$  con un numero casuale distribuito uniformemente per ogni direzione coordinata. Nei fatti, questa procedura consiste in due cicli iterativi annidati: uno sui punti e uno sulla dimensione. L'implementazione del midpoint multidimensionale presenta, invece, diverse difficoltà. La difficoltà principale si ha quando la dimensione  $M$  dell'ipersfera non è fissata. In tal caso la (37) richiede la presenza di  $M - 1$  cicli iterativi annidati a causa delle somme consecutive. Esiste un modo per risolvere il problema senza perdere di generalità: utilizzare una funzione ricorsiva che permette di eseguire un numero non fissato di cicli. Al fine di evitare di complicare eccessivamente il codice è possibile individuare un'altra soluzione, al prezzo di fissare uno dei parametri in gioco. Si supponga, infatti, di fissare il numero di punti per direzione coordinata come

$$N := N_1 = N_2 = \dots = N_{M-1} = 10$$

Il numero di punti necessario per generare una griglia regolare  $M - 1$  dimensionale nell'iperquadrato  $Q_{M-1}$  sarà allora

$$N_{\text{tot}} = 10^{M-1} \quad (39)$$

Sia dunque  $n \in \mathbb{N}$  tale che  $0 \leq n \leq N_{\text{tot}}$  una variabile indicizzata che conta i punti generati. Siccome  $n$  è un intero positivo potrà sempre essere scritto come

$$n = n_1 \dots n_L \quad \text{con} \quad n_i \in \{0, 1, 2, \dots, 9\} \quad \forall i = 1, \dots, L$$

dove  $n_i$  denotano le cifre ordinate consecutive che rappresentano il numero naturale  $n$ . Ad esempio, se  $n = 317$  si ha  $n_1 = 3$ ,  $n_2 = 1$  e  $n_3 = 7$ . Si consideri, dunque, la mappa

$$\begin{aligned} \Phi : \mathbb{N} &\rightarrow \mathbb{N}^L \\ n &\longmapsto \vec{n} := (n_1, \dots, n_L) \end{aligned}$$

ossia la mappa che associa, ad ogni naturale, un vettore le cui componenti sono le cifre che costituiscono il numero nell'ordine in cui sono scritte. Nell'esempio

riportato, l'applicazione trasformerà  $n = 317$  nel vettore  $(3, 1, 7)$ . Evidentemente, la dimensione  $L$  dello spazio in cui vive il vettore dipenderà dalla grandezza del numero naturale in esame e coincide esattamente con il numero di cifre che compongono  $n$ . Non è difficile capire che, per come sono stati generati i numeri, vale sempre  $L \leq M - 1$ . Consideriamo, allora, anche la mappa

$$\begin{aligned}\Psi : \mathbb{N}^L &\rightarrow \mathbb{N}^{M-1} \\ \vec{n} &\longmapsto (0, \dots, 0, n_1, \dots, n_L)\end{aligned}$$

ossia la mappa che associa, ad ogni vettore delle cifre che compongono  $n$ , un vettore a cui è stato aggiunto, nelle prime posizioni, un numero sufficiente di zeri tale da raggiungere la dimensione fissata  $M - 1$ . A questo punto, non è difficile mostrare che la mappa composta  $\Psi \circ \Phi$  è biunivoca. In particolare, l'applicazione in esame assocerà univocamente, ad ogni punto che compone la griglia, un vettore  $(n_1, \dots, n_{M-1})$  con  $n_i \in \mathbb{N}$  tale che  $0 \leq n_i \leq 9$ . Ma allora, la mappa costruita risulta essere in grado di costruire un set di punti  $M - 1$  dimensionali tutti equispaziati tra loro. Al fine di generare punti nell'iperquadrato  $Q_{M-1}$  e di centrarli in ogni sottoinsieme di quest'ultimo, come richiede il metodo del midpoint, sarà allora necessario compiere un'ultima operazione di riscalamento e di traslazione. Il passo è definito, nel caso in esame, come  $h = \frac{1}{N}$ . La mappa in grado di svolgere queste ultime due operazioni sarà allora

$$\begin{aligned}\Upsilon : \mathbb{N}^{M-1} &\rightarrow Q_{M-1} \\ \vec{n} &\longmapsto h(n_1, \dots, n_{M-1}) + \left( \frac{h}{2}, \dots, \frac{h}{2} \right)\end{aligned}$$

Siamo quindi riusciti a costruire formalmente una mappa

$$\text{coord} := \Upsilon \circ \Psi \circ \Phi \tag{40}$$

evidentemente biunivoca in grado di associare, ad ogni punto della griglia indicizzato da un naturale, il corrispondente vettore  $M - 1$  dimensionale delle coordinate centrato in un opportuno sotto-iperquadrato di  $Q_{M-1}$ . Grazie alla mappa determinata sarà allora possibile costruire un algoritmo con un'implementazione simile al caso Monte Carlo. Infatti,  $\text{coord}(n)$  permette di associare un set di coordinate al variare dell'indice naturale  $n$ . Da un punto di vista operativo, un modo per implementare la mappa (40) consiste nel trasformare  $n$  in una stringa di cifre, aggiungere manualmente il numero necessario di zeri in testa alla stringa fino alla dimensione  $M - 1$  e, infine, convertire la stringa di cifre nel corrispondente vettore di interi. Si noti che il metodo di generazione della griglia dato dalla mappa  $\text{coord}(n)$  funziona solo se il numero di punti per direzione coordinata del dominio è fissato a  $N = 10$ . Infatti, per un numero diverso di punti, la mappa costruita perde la biumivocità che garantisce il corretto funzionamento. La verifica di questo fatto può essere effettuata pensando, ad esempio, ad un semplice caso di dominio bidimensionale. Il fatto di fissare il valore di  $N$ , tuttavia, non è una condizione che pesa particolarmente ai fini dell'esercizio in esame. Per operare un confronto a parità di punti, ad esempio, basterà variare il numero totale di punti generato in Monte Carlo, dove la struttura dell'algoritmo permette la massima flessibilità in termini di variazione dei parametri. In qualche senso, per garantire un'implementazione ragionevole del midpoint multidimensionale al variare della dimensione dell'ipersfera, si è dovuto accettare

di fissare un altro parametro, meno rilevante ai fini dei risultati che si vogliono verificare. Questa asimmetria di semplicità nell'implementazione tra midpoint e Monte Carlo permette, fin da qui, di capire che l'utilizzo delle sequenze pseudo-casuali risulta più comoda ai fini dell'integrazione multidimensionale.

### Maledizione della dimensionalità

Al fine di ottenere alcuni primi risultati circa il comportamento dei due metodi si sono quindi calcolate le stime  $V_{\text{mid}}$  e  $V_{MC}$  al variare della dimensione  $M$  dell'ipersfera nel range

$$1 \leq M_i \leq 9 \quad \text{con} \quad M_{i+1} = M_i + 1$$

Ogni funzione dedicata al calcolo del volume è stata dotata di un'apposita funzione, inclusa nelle librerie standard, per il calcolo del tempo di esecuzione espresso in secondi. Per il metodo del midpoint il numero di punti è fissato, mentre per il metodo Monte Carlo si è considerato un numero totale di punti pari a  $N_{\text{tot}} = 10^5$ . La tabella che segue mostra quanto ottenuto.

<b><math>M</math></b>	<b>midpoint</b>		<b>Monte Carlo</b>		<b>esatto</b>
	$V_{\text{mid}}$	$t_{\text{mid}}$	$V_{MC}$	$t_{MC}$	$V_M$
1	1.9975	$1.1 \cdot 10^{-5}$	2	0.018426	2.00000
2	3.15241	$10^{-5}$	3.14047	0.07008	3.14159
3	4.19824	$9 \cdot 10^{-5}$	4.19286	0.080568	4.18879
4	4.94357	0.001003	4.91125	0.092323	4.93480
5	5.27027	0.01153	5.25911	0.101763	5.26379
6	5.17474	0.126084	5.20505	0.111869	5.16771
7	4.73259	1.38952	4.67994	0.11891	4.72477
8	4.06299	13.8036	4.10798	0.129858	4.05871
9	3.27951	151.355	3.18449	0.138184	3.29851

Come è possibile notare, nel caso del metodo del midpoint, i tempi di calcolo aumentano molto più velocemente rispetto ai tempi di calcolo in Monte Carlo. D'altra parte, il metodo Monte Carlo fornisce, all'aumentare della dimensione  $M$  dell'ipersfera, una stima del volume sempre meno precisa, a differenza del metodo deterministico del midpoint.

A tal proposito si noti che, dato un set di dati qualunque  $\{x_i\}_{i=1,\dots,N}$  con  $N$  fissato distribuiti in uno spazio  $M$  dimensionale, all'aumentare della dimensione  $M$  dello spazio il set di dati perde di significatività e di correlazione. Intuitivamente questo fatto è banale: se il numero di dati è fisso, all'aumentare della dimensione questi divengono più sparsi, in quanto vengono mano a mano distribuiti in un volume sempre maggiore. Questo fenomeno è indipendente dal modo in cui i dati vengono generati ed è anche conosciuto come *maledizione della dimensionalità / curse of dimensionality*. Se gli  $N$  dati vengono utilizzati per il calcolo di una stima di un integrale è evidente che, all'aumentare della dimensione dello spazio su cui si integra, la precisione del calcolo peggiorerà progressivamente se il numero di punti in cui viene valutata l'integranda rimane costante. Al fine di verificare la consistenza dei risultati ottenuti alla luce delle considerazioni fatte si è quindi deciso di calcolare le dispersioni dal valore vero al variare di  $M$ , sia nel caso del midpoint, che in quello Monte Carlo. Il grafico che segue mostra i due andamenti a confronto.

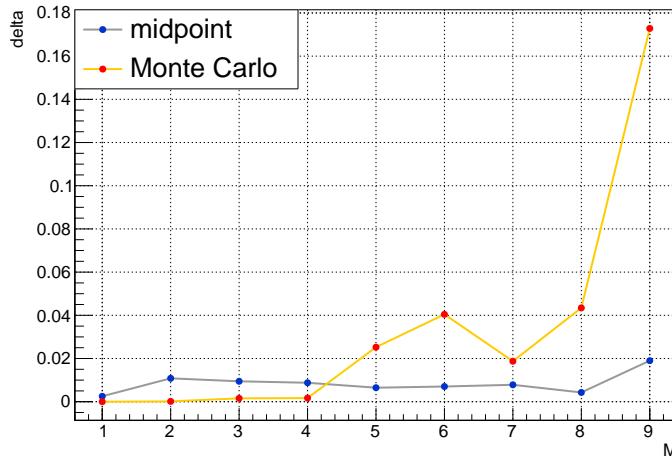


Figura 54: plot  $\Delta_{\text{mid}}(M)$  e  $\Delta_{\text{MC}}(M)$  a confronto

Il grafico mostra che i risultati ottenuti sono perfettamente in linea con quanto ci si aspetta. La precisione in midpoint rimane sostanzialmente costante in quanto il metodo stesso, per costruzione, genera una griglia di

$$N_{\text{tot}} = 10^{M-1}$$

punti, ossia i punti totali aumentano esponenzialmente all'aumentare della dimensione dello spazio. L'effetto della maledizione della dimensionalità viene così nascosto, al prezzo di un aumento esponenziale del numero di punti nella griglia, e quindi dei tempi di calcolo. L'errore commesso con il calcolo Monte Carlo, invece, diviene rapidamente elevato, in quanto si è fissato un numero  $N_{\text{tot}}$  costante di punti generati in  $Q_{M-1}$ , che all'aumentare della dimensione del dominio perdono di significatività. Il metodo del midpoint è vincolato, per costruzione, alla generazione di una griglia regolare in uno spazio  $M - 1$  dimensionale. Per tale ragione, seppur sia garantita una stima con precisione relativamente elevata, il tempo di esecuzione richiesto esploderà all'aumentare di  $M$ . Il metodo Monte Carlo, invece, sfruttando la generazione casuale, non ha alcun vincolo circa il numero di punti generato in  $Q_{M-1}$ . Risulta allora interessante studiare per quali valori della dimensione l'utilizzo di questo metodo permetta comunque una stima accettabile entro una certa precisione per  $N_{\text{tot}} = 10^5$  punti. A tale scopo si sono calcolati e plottati tre diversi set di coppie  $(M, \Delta_{\text{MC}})$  mantenendo costante il numero totale di punti generati. Si è scelto di calcolare tre diversi insiemi di dati in quanto la generazione di numeri pseudo-casuali può produrre, per un solo set di dati, risultati non attendibili. La medesima operazione può essere svolta per qualunque range di  $M$  e per ogni valore del numero totale di punti generato. L'analisi svolta in questa sezione è solo un esempio per illustrare un metodo operativo di analisi dell'errore commesso in Monte Carlo. Inoltre, di seguito si farà esplicito riferimento al solo errore assoluto. Viene da sé che per ottenere risultati più informativi e di pratica utilità sarà utile il calcolo dell'errore relativo percentuale, che per ragioni di semplicità verrà omesso. Di seguito sono mostrati i tre andamenti ottenuti a confronto.

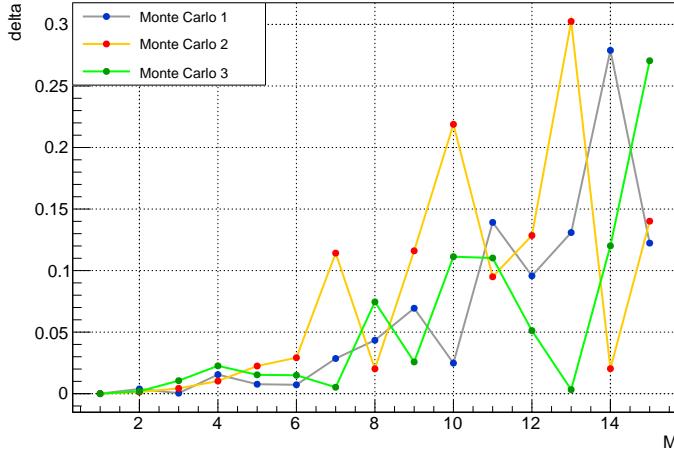


Figura 55: plot di 3 set di dati  $\Delta_{MC}(M)$  a confronto

Come è possibile notare, le dispersioni dal valore vero aumentano globalmente in tutti i casi. Le fluttuazioni che si osservano sono semplicemente date dall'utilizzo delle sequenze casuali. Dal grafico risulta possibile concludere che il metodo Monte Carlo restituisce una stima che, fino alla dimensione  $M = 15$  dell'ipersfera, dista dal valore vero di un valore più piccolo di 0.3. Nel caso del metodo del midpoint, invece, come si nota dalla figura 54, la stima è sempre definita a meno di un errore massimo di 0.02, per ogni valore di  $M$ . Tentativi di calcolo con il midpoint per dimensioni superiori a  $M = 9$  risultano, di fatto, irrealizzabili a causa del veloce aumento del costo computazionale dell'algoritmo. Il metodo Monte Carlo, invece, non dà particolari problemi anche per dimensioni molto elevate. Risulta allora interessante svolgere un'analisi più accurata circa l'andamento dei tempi di calcolo. Si sono quindi calcolate e plottate le coppie  $(M, t_{\text{mid}})$  e  $(M, t_{MC})$  al variare di  $M$ , ottenendo quanto segue.

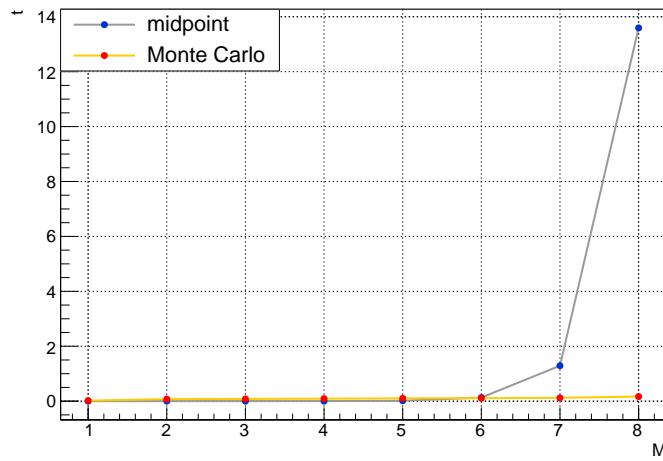


Figura 56: plot  $t_{\text{mid}}(M)$  e  $t_{MC}(M)$  a confronto

Come è possibile notare, i tempi di calcolo nel caso del metodo del midpoint esplodono a partire da un valore di  $M = 6$ . I tempi di calcolo Monte Carlo, invece, risultano sostanzialmente trascurabili in relazione all'esplosione di cui sopra. Vista la rapidità della crescita, dal grafico è possibile ipotizzare che l'andamento descritto dai tempi del midpoint sia un esponenziale crescente. Al fine di svolgere un'analisi più accurata su questo aspetto si sono allora interpolati i dati midpoint al variare di  $M$  da 2 fino a 8 con la generica funzione esponenziale  $t(M) = ae^{bM}$ . Non è stato possibile analizzare un range di valori più ampio per le ragioni già discusse. Di seguito sono riportati i risultati ottenuti.

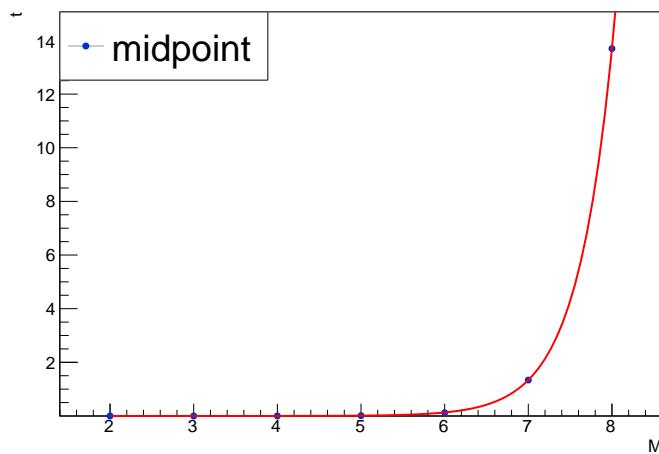


Figura 57: grafico  $t_{\text{mid}}(M)$ : fit

I parametri stimati risultano essere

$$a = 1.33 \cdot 10^{-7} \quad \text{e} \quad b = 2.31$$

In questo caso non è possibile svolgere alcun confronto con parametri noti per la verifica della correttezza del fit, ma dal grafico è possibile notare che la funzione interpolante appare descrivere correttamente i dati raccolti. L'esplosione esponenziale risulta in linea con le considerazioni fatte: il metodo del midpoint, infatti, per costruzione genera un numero totale di punti che varia esponenzialmente con la dimensione dell'ipersfera. Si è quindi deciso di svolgere la medesima operazione nel caso della stima Monte Carlo. In particolare, svolgendo un semplice fit delle coppie di dati  $(M, t_{\text{MC}})$  generate, si è notato un andamento qualitativamente rettilineo. Tenendo conto dell'andamento trascurabile dei tempi Monte Carlo rispetto a quelli del midpoint di figura 56 e tenendo conto della gerarchia degli infiniti, l'ipotesi di andamento lineare appare del tutto ragionevole. Si è allora interpolato il tempo di calcolo in funzione della dimensione con una funzione lineare della forma  $t(M) = q + pM$ . Il numero di punti generati è stato lasciato fisso al valore di  $N_{\text{tot}} = 10^5$ , per coerenza con i risultati ottenuti fino a questo punto. Si è considerato il range di  $M$  da 2 fino a 18, sfruttando la flessibilità del metodo nella scelta della dimensione. Di seguito sono riportati i risultati ottenuti.

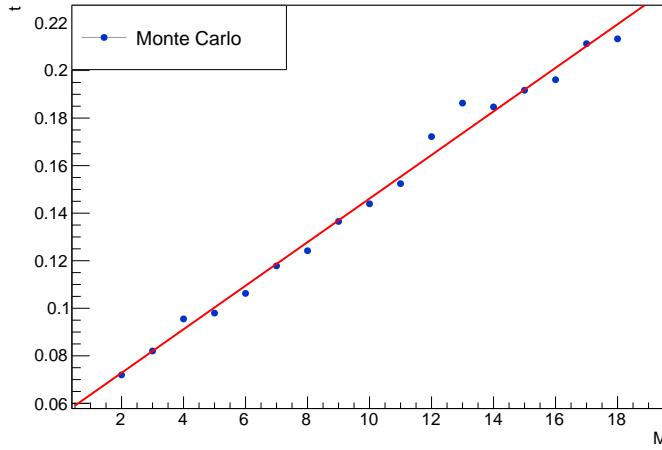


Figura 58: grafico  $t_{MC}(M)$  per  $N_{tot} = 10^5$ : fit

I parametri stimati risultano essere

$$q = 0.055 \quad \text{e} \quad p = 0.0092$$

A meno delle solite fluttuazioni date dall'utilizzo delle sequenze casuali, la funzione interpolante descrive correttamente i dati raccolti. Siamo allora riusciti a verificare in modo più quantitativo che l'andamento dei tempi di calcolo per il midpoint corre all'infinito molto più rapidamente rispetto all'andamento dei tempi Monte Carlo. Tuttavia, i risultati ottenuti presentano una certa disparità di condizioni: nel caso del metodo del midpoint il numero di punti totale in  $Q_{M-1}$  varia esponenzialmente con la dimensione. Nel caso Monte Carlo, invece, rimane fissato ad un valore costante per ogni  $M$ . Si sono quindi calcolate e plottate ancora una volta le coppie  $(M, t_{mid})$  e  $(M, t_{MC})$ , variando il numero dei punti totali secondo la (39) anche in Monte Carlo, ottenendo quanto segue.

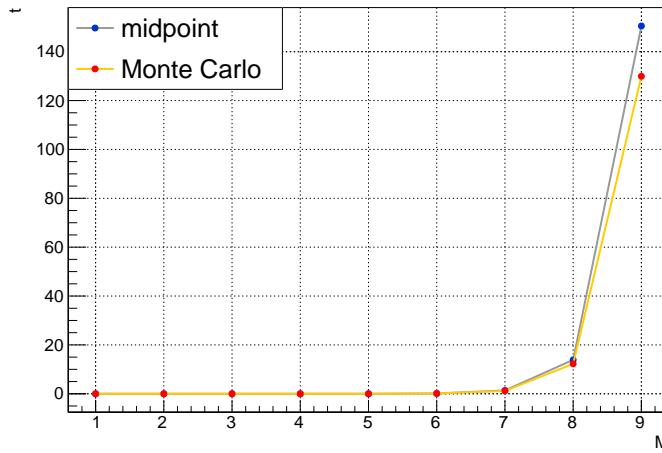


Figura 59: plot  $t_{mid}(M)$  e  $t_{MC}(M)$  con  $N_{tot} = 10^{M-1}$  a confronto

Come è possibile notare, imponendo un numero totale di punti generati nel dominio pari al numero totale del metodo del midpoint, anche il metodo Monte Carlo diviene soggetto ad un aumento esponenziale dei tempi di calcolo. Al fine di verificarlo con più precisione si è deciso, anche in questo caso, di interpolare i dati Monte Carlo al variare di  $M$  da 2 fino a 8 con la generica funzione esponenziale  $t(M) = ae^{bM}$ , ottenendo i seguenti risultati.

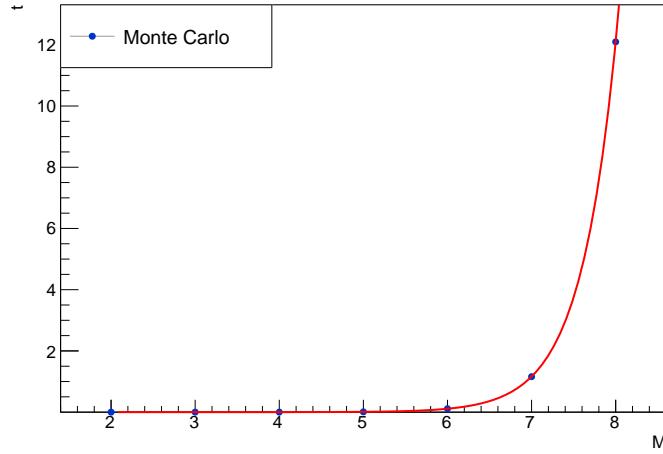


Figura 60: grafico  $t_{\text{MC}}(M)$  con  $N_{\text{tot}} = 10^{M-1}$ : fit

I parametri stimati risultano essere

$$a = 8.69 \cdot 10^{-8} \quad \text{e} \quad b = 2.34$$

Anche in questo caso non è possibile alcuna verifica per mezzo di confronti con valori attesi, ma la funzione interpolante appare descrivere correttamente i dati raccolti. Anche il metodo Monte Carlo, dunque, lavorando a parità di punti generati, risulta essere soggetto ad un andamento esponenziale dei tempi di calcolo. D'altra parte, risulta abbastanza intuitivo pensare che il costo computazionale di un algoritmo abbia andamento esponenziale in  $M$  se viene generato un numero di punti esponenziale nello stesso  $M$ .

Evidentemente, generare un numero di punti esponenziale nella dimensione risulta del tutto insensato nel caso del metodo Monte Carlo, in quanto la sua costruzione risulta vantaggiosa proprio nei casi di integrazione multidimensionale, in cui si vuole evitare la generazione esponenziale. Tuttavia, i risultati ottenuti permettono di concludere che, così come nel caso del midpoint, anche il metodo Monte Carlo è soggetto alla maledizione della dimensionalità. Il vantaggio del calcolo Monte Carlo per dimensioni elevate rispetto ai metodi deterministici risiede solo nella costruzione molto diversa dei due metodi. Nel metodo del midpoint è necessario generare una griglia regolare  $M - 1$  dimensionale per essere certi della convergenza della stima prodotta dall'algoritmo al valore vero (supponendo una statistica infinita). L'utilizzo delle sequenze pseudo-casuali, invece, permette di superare la rigidità dei metodi deterministici, garantendo la possibilità di inchiodare il numero di punti generati nel dominio di integrazione qualunque sia il valore di  $M$ . Così facendo si guadagna sui tempi di calcolo, ma

si perde in precisione a causa dell'effetto della maledizione della dimensionalità. L'utilizzo di un metodo piuttosto che un altro, allora, come in tutti i casi, dipenderà dalle esigenze specifiche del problema numerico in esame. Nel caso del calcolo di  $V_M$  l'utilizzo dei metodi Monte Carlo appare significativamente migliore per dimensioni elevate, a partire da  $M = 10$ , dove il tempo richiesto dal midpoint supera diversi minuti. Seppur la stima Monte Carlo non abbia la stessa precisione del midpoint, infatti, siamo almeno certi di riuscire a produrne una senza disporre di un supercomputer, entro una certa precisione che sappiamo quantificare ( $\sim 1/\sqrt{N_{\text{tot}}}$ ). In questo senso, il metodo Monte Carlo risulta cruciale per affrontare il problema della maledizione della dimensionalità, sotto il quale i metodi deterministici falliscono.

## Esercizio 9

Si vuole verificare numericamente il teorema centrale del limite per variabili casuali distribuite secondo due diverse distribuzioni densità di probabilità.

**Teorema 0.11** (centrale del limite). *Sia  $\{x_1, \dots, x_N\}$  un insieme di  $N$  variabili casuali statisticamente indipendenti e identicamente distribuite secondo  $f = f(x)$  tale che*

$$\langle x \rangle = \mu \quad e \quad \text{Var}(x) = \sigma^2$$

*Si consideri la media campionaria della sequenza*

$$\bar{x} := \frac{1}{N} \sum_{i=1}^N x_i$$

*Sia  $g = g(\bar{x})$  la distribuzione densità di probabilità di  $\bar{x}$ . Allora*

$$\lim_{N \rightarrow +\infty} g(\bar{x}) = G \left( \bar{x} \mid \mu, \frac{\sigma}{\sqrt{N}} \right)$$

*dove*

$$G(x \mid \mu, \sigma) := \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right)$$

*è la distribuzione normale (o gaussiana) con media  $\mu$  e deviazione standard  $\sigma$ .*

Il teorema 0.11 ci assicura che, qualunque sia la distribuzione  $f$  della variabile casuale  $x$ , la media campionaria delle  $x_i$  misure sarà distribuita, asintoticamente, come una gaussiana centrata nella media di  $f$  e con medesima deviazione standard, riscalata di un fattore  $1/\sqrt{N}$ .

**Definizione 0.12.** Sia  $x$  una variabile casuale distribuita secondo la funzione densità di probabilità  $f : \Omega \rightarrow \mathbb{R}$ .

- Se  $f$  è una distribuzione continua, chiameremo *momento di ordine  $m$*  di  $f$  la quantità

$$E(x^m) \equiv \langle x^m \rangle := \int_{\Omega} x^m f(x) dx$$

- Se  $f$  è una distribuzione discreta, chiameremo *momento di ordine  $m$*  di  $f$  la quantità

$$E(x^m) \equiv \langle x^m \rangle := \sum_{x \in \Omega} x^m f(x)$$

L'operatore  $E$  che compare nella definizione 0.12 è chiamato *valore di aspettazione* di  $x^m$  rispetto alla densità  $f$ . In entrambi i casi, il momento di ordine 1 è detto *media* della distribuzione. A partire dalla media di  $f$  è possibile definire anche quelli che sono chiamati *momenti centrali di ordine  $m$* . I momenti, e soprattutto i momenti centrali di una distribuzione, sono valori particolarmente importanti in quanto forniscono indicazioni circa alcune caratteristiche della distribuzione in esame, come la sua simmetria o la presenza di picchi accentuati. Esistono, infatti, distribuzioni particolarmente complesse da trattare analiticamente: il calcolo di questi numeri permette, anche in questi casi, di estrarre informazioni circa il comportamento qualitativo di una densità. In particolare,

il momento centrale di ordine 2 è detto *varianza* della distribuzione e dà informazioni circa la sua larghezza.

Discussiamo allora alcune relazioni notevoli che sussistono tra i momenti di una distribuzione. Si dimostra che, nel caso gaussiano del teorema centrale del limite, per il momento di ordine 2 vale la relazione asintotica

$$\begin{aligned}\langle \bar{x}^2 \rangle &= O\left(\frac{1}{N}\right) \quad \text{per} \quad N \rightarrow +\infty \quad \iff \\ \langle \bar{x}^2 \rangle &\approx k \frac{1}{N} \quad \text{con} \quad N \text{ grande e } k \in \mathbb{R}\end{aligned}\tag{41}$$

Posto  $n := \frac{1}{N}$  la relazione diviene lineare e passante per l'origine. In particolare

$$\langle \bar{x}^2 \rangle \approx kn \quad \text{con} \quad n \text{ piccolo}\tag{42}$$

Calcolando i logaritmi ad entrambi i membri assume la forma

$$\log \langle \bar{x}^2 \rangle \approx \log k + \log n \quad \text{con} \quad n \text{ piccolo}\tag{43}$$

Il momento di ordine 4 è relazionato al momento di ordine 2 secondo

$$\langle \bar{x}^4 \rangle = 3 \langle \bar{x}^2 \rangle^2$$

In regime asintotico vale la (42). Sostituendo e passando ai logaritmi si ottiene una relazione lineare della forma

$$\log \langle \bar{x}^4 \rangle \approx \log 3k^2 + 2 \log n \quad \text{con} \quad n \text{ piccolo}\tag{44}$$

Il momento di ordine 6, invece, è relazionato a quello di ordine 2 secondo

$$\langle \bar{x}^6 \rangle = 15 \langle \bar{x}^2 \rangle^3$$

In regime asintotico vale ancora la (42). Sostituendo e passando ai logaritmi si ottiene una relazione lineare della forma

$$\log \langle \bar{x}^6 \rangle \approx \log 15k^3 + 3 \log n \quad \text{con} \quad n \text{ piccolo}\tag{45}$$

Uno dei modi per verificare il teorema 0.11 sarà, quindi, quello di plottare e fissare i valori dei momenti al variare della dimensione  $N$  del pacchetto di dati. Si noti che le relazioni lineari ricavate possono essere ulteriormente semplificate calcolando i rapporti dei vari momenti di ordine pari ricavati. In particolare, il rapporto tra il momento di ordine 4 e quello di ordine 2 sarà

$$\frac{\langle \bar{x}^4 \rangle}{\langle \bar{x}^2 \rangle} = \frac{3 \langle \bar{x}^2 \rangle^2}{\langle \bar{x}^2 \rangle} = 3 \langle \bar{x}^2 \rangle$$

Per  $N$  sufficientemente grande vale la relazione (42). Sostituendo e passando ai logaritmi si ottiene

$$\log \left( \frac{\langle \bar{x}^4 \rangle}{\langle \bar{x}^2 \rangle} \right) \approx \log 3k + \log n \quad \text{con} \quad n \text{ piccolo}\tag{46}$$

Il rapporto tra il momento di ordine 6 e quello di ordine 4 sarà invece

$$\frac{\langle \bar{x}^6 \rangle}{\langle \bar{x}^4 \rangle} = \frac{15\langle \bar{x}^2 \rangle^3}{3\langle \bar{x}^2 \rangle^2} = 5\langle \bar{x}^2 \rangle$$

Per  $N$  abbastanza grande vale la relazione (42). Sostituendo e passando ai logaritmi si ottiene

$$\log \left( \frac{\langle \bar{x}^6 \rangle}{\langle \bar{x}^4 \rangle} \right) \approx \log 5k + \log n \quad \text{con} \quad n \text{ piccolo} \quad (47)$$

Plottare e fissare la (43), (44) e (45) o la (46) e la (47) risulta, dunque, del tutto equivalente ai fini della verifica numerica del teorema centrale del limite. Si procederà per la seconda strada vista la normalizzazione del coefficiente di  $\log n$  nel secondo caso. In particolare, in entrambi i casi si è deciso di rappresentare negli istogrammi direttamente il numero di eventi in funzione di  $x$ . Risulta possibile anche rappresentare le frequenze relative, dividendo il numero di eventi in un bin per il prodotto tra la dimensione del bin e il numero di eventi totali. In tal modo, l'istogramma risulterebbe normalizzato, rendendo consistente la rappresentazione dei dati con la normalizzazione di ogni funzione densità di probabilità che li descrive. In ogni caso, i fattori di normalizzazione non influiscono sui momenti e sui momenti centrali di una distribuzione di probabilità. Per tale ragione, si è deciso di snellire i passaggi nel codice rappresentando direttamente il numero di eventi per bin.

Siamo quindi interessati a verificare il teorema 0.11 con due diverse distribuzioni di probabilità  $f$  di partenza: la distribuzione uniforme centrata nello zero e la distribuzione discreta di Bernoulli.

### pdf uniforme

Si sono generati  $M = 10000$  pacchetti di valori  $\{x_i\}_{i=1,\dots,N}$  identicamente distribuiti secondo la distribuzione uniforme definita da

$$U(x) := \begin{cases} \frac{1}{b-a} & \text{se } x \in \Omega \\ 0 & \text{se } x \notin \Omega \end{cases} \quad \text{con} \quad \Omega = (a, b) = (-1, 1)$$

Si noti che, per la distribuzione uniforme in esame, valgono

$$\mu = \frac{a+b}{2} = 0 \quad \text{e} \quad \sigma = \frac{b-a}{\sqrt{12}} = \frac{\sqrt{3}}{3}$$

Si è scelto un valore di  $M$  sufficientemente grande al fine di ottenere un numero di medie tale da facilitare la visualizzazione dell'andamento della distribuzione. Tale valore è stato mantenuto costante durante tutta l'analisi che segue. Per amore di completezza, si è deciso di verificare il teorema 0.11 con due diverse modalità: una verifica statica ad un valore  $N = \tilde{N}$  sufficientemente grande fissato, e una verifica dinamica al variare di  $N$ , utilizzando le relazioni tra i momenti ricavate in precedenza.

## Verifica statica

Al fine di verificare qualitativamente l'approssimazione della distribuzione risultante a quella gaussiana si sono calcolate le medie campionarie

$$\bar{x}_j := \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} x_{j,i} \quad \forall j = 1, \dots, M$$

con  $\tilde{N} = 1, 2, 5, 20, 1000$ . Di seguito sono riportati gli istogrammi dei dati ottenuti.

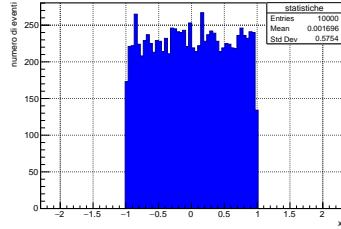


Figura 61:  $\tilde{N} = 1$  con  $f$  uniforme

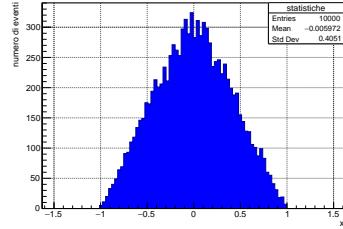


Figura 62:  $\tilde{N} = 2$  con  $f$  uniforme

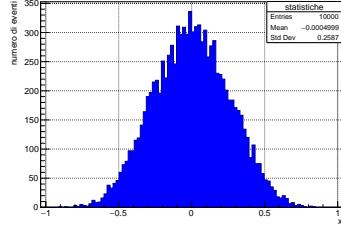


Figura 63:  $\tilde{N} = 5$  con  $f$  uniforme

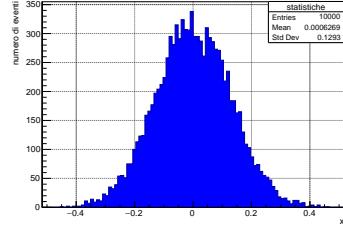


Figura 64:  $\tilde{N} = 20$  con  $f$  uniforme

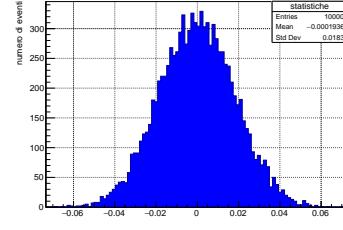


Figura 65:  $\tilde{N} = 1000$  con  $f$  uniforme

Come è possibile notare, l'andamento della media campionaria  $\bar{x}$  assume la forma a campana tipica della distribuzione gaussiana all'aumentare della dimensione  $N$  dei pacchetti di dati, come ci si aspetta. Al fine di svolgere una prima verifica quantitativa si è quindi deciso di fittare la distribuzione binnata ottenuta per  $\tilde{N} = 1000$  con una gaussiana  $G(\bar{x} | \tilde{\mu}, \tilde{\sigma})$  non normalizzata, con  $\tilde{\mu}$  e  $\tilde{\sigma}$  parametri liberi. L'algoritmo di minimizzazione ha prodotto quanto segue.

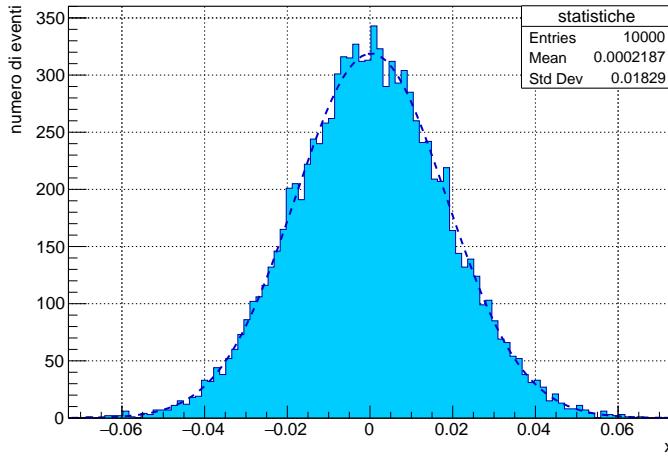


Figura 66: distribuzione per  $\tilde{N} = 1000$  con  $f$  uniforme: fit

Si è ottenuta la seguente stima dei parametri.

$$\tilde{\mu} = 0.00022 \quad \text{e} \quad \tilde{\sigma} = 0.01829$$

Si noti che si sono ottenuti parametri leggermente differenti da quelli riportati in figura 65 nonostante si sia lavorato a parità di  $\tilde{N}$  a causa del cambiamento del seme delle sequenze pseudo-casuali ad ogni esecuzione. Per teorema centrale del limite ci si aspetta che la gaussiana che descrive la distribuzione delle medie campionarie abbia la forma

$$G = G \left( \bar{x} \mid 0, \frac{\sigma}{\sqrt{\tilde{N}}} \right)$$

Dunque, la deviazione standard attesa risulta essere

$$\frac{\sigma}{\sqrt{\tilde{N}}} = \frac{\sqrt{3}}{3\sqrt{1000}} \approx 0.0183$$

Ma allora, notiamo subito che

$$\tilde{\mu} \approx 0 \quad \text{e} \quad \tilde{\sigma} \approx 0.0183$$

La non esattezza delle stime ottenute dipende, chiaramente, dal fatto che l'approssimazione alla distribuzione gaussiana vale asintoticamente. D'altra parte, computazionalmente, risulta possibile lavorare solo a valori di  $N$  finiti. I valori della media e della deviazione standard sono quindi consistenti con quelli dati dal teorema 0.11. Un'ulteriore conferma è data dal valore di probabilità associata al  $Q^2$  stimata dal programma di minimizzazione, che risulta ben al di sopra della soglia di accettabilità dell'ipotesi. Segue che la gaussiana è una forma funzionale che ben descrive la distribuzione ottenuta per  $\tilde{N} = 1000$ , rendendo possibile affermare la verifica del teorema per questo fissato valore di  $N$ .

## Verifica dinamica

Al fine di verificare l'approssimazione della distribuzione risultante a quella gaussiana si sono poi calcolati e plottati i momenti di ordine  $k = 1, \dots, 6$  delle medie campionarie come

$$\langle x^k \rangle_j := \frac{1}{N} \sum_{i=1}^N x_{j,i}^k \quad \forall j = 1, \dots, M$$

Al variare di  $N$  nel range

$$2 \leq N < 100 \quad \text{con} \quad N_{i+1} = N_i + 1$$

Si sono poi plottati i valori dei momenti in funzione della dimensione  $N$  del pacchetto di dati costruito. In particolare, al fine di svolgere una prima verifica qualitativa, si è deciso di plottare separatamente i momenti di ordine pari da quelli di ordine dispari per confrontarne l'andamento. Ci si aspetta, infatti, che i momenti di ordine dispari presentino un andamento qualitativo simile. Lo stesso per i momenti di ordine pari, a meno della velocità di decrescita. Per i momenti di ordine dispari si sono ottenuti i seguenti risultati.

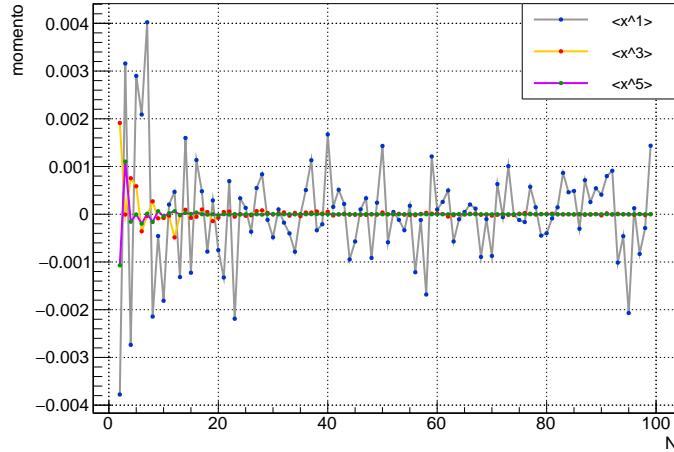


Figura 67: confronto  $\langle x^{2k-1} \rangle$  con  $f$  uniforme

Come è possibile notare, i momenti di ordine dispari della distribuzione di  $\bar{x}$  oscillano, in generale, in un intorno dello zero delle ordinate. Il fatto che, mediamente, i momenti di ordine dispari siano nulli è ben giustificato assumendo che, per il teorema 0.11,  $\bar{x}$  sia distribuita secondo

$$G\left(\bar{x} | 0, \frac{\sigma}{\sqrt{N}}\right)$$

Osservando la forma analitica di una gaussiana è evidente che, quanto  $\mu = 0$ , la funzione è pari rispetto allo zero, ossia

$$G(-\bar{x}) = G(\bar{x})$$

D'altra parte,  $x^{2k-1}$  è una funzione dispari. Segue che, per  $N$  grande vale

$$\langle x^{2k-1} \rangle = \int_{-\infty}^{+\infty} x^{2k-1} G(\bar{x}) dx = 0 \quad \forall k \in \mathbb{N} \setminus \{0\}$$

poiché il prodotto tra una funzione pari e una dispari è una funzione dispari, che, integrata sull'intervallo simmetrico  $(-\infty, +\infty)$ , produce sempre un valore nullo. Il fatto che la relazione valga in regime asintotico spiega la ragione per la quale, per valori di  $N$  piccoli, si siano ottenuti valori più distanti dallo zero. Si noti che, ciò che cambia nei momenti di ordine dispari è la rapidità con cui la funzione oscillante si stabilizza al valore nullo. In particolare, all'aumentare di  $k$  dispari, il momento di ordine  $k$  tende a stabilizzarsi a zero per valori di  $N$  minori. Per i momenti di ordine pari si sono ottenuti i risultati che seguono.

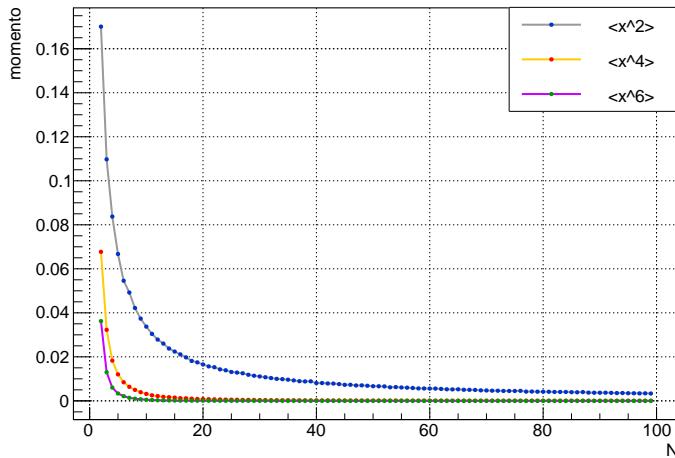


Figura 68: confronto  $\langle x^{2k} \rangle$  con  $f$  uniforme

Anzitutto, notiamo che i valori dei momenti risultano sempre positivi. Questo è consistente con il fatto che il calcolo di potenze pari produce sempre numeri più grandi di zero. Inoltre, è possibile notare che l'andamento nei tre diversi casi è qualitativamente consistente con quanto ci si aspetta. Infatti, ci si aspetta che  $\langle x^2 \rangle$  abbia un andamento descritto dalla (41), come accade in figura. Si noti poi che  $\langle x^4 \rangle$  e  $\langle x^6 \rangle$  sono relazionati a  $\langle x^2 \rangle$  tramite potenze, rispettivamente, di ordine 2 e 3. Segue che ci si aspetta di osservare un andamento simile, ma con una convergenza più rapida verso lo zero all'aumentare di  $k$ , esattamente come si osserva nel plot ottenuto. Al fine di verificare quantitativamente l'andamento gaussiano al variare di  $N$  si è quindi deciso di verificare la relazione  $\langle \bar{x} \rangle = \mu$  data direttamente dal teorema 0.11, la relazione (46) e la relazione (47). In particolare, nonostante l'aumento considerevole dei tempi di calcolo, al fine di ottenere una stima più accurata dei parametri si è deciso di calcolare i momenti, e quindi i loro rapporti, nel range

$$100 \leq N < 1000 \quad \text{con} \quad N_{i+1} = N_i + 10$$

Si sono quindi calcolati, plottati e interpolati i valori dei rapporti dei momenti di ordine pari in funzione di  $N$ . Vista la linearità delle relazioni da verificare, tutti

i fit sono stati effettuati utilizzando una funzione lineare della forma  $y = mx + q$ . Per la prima relazione si sono ottenuti i risultati che seguono

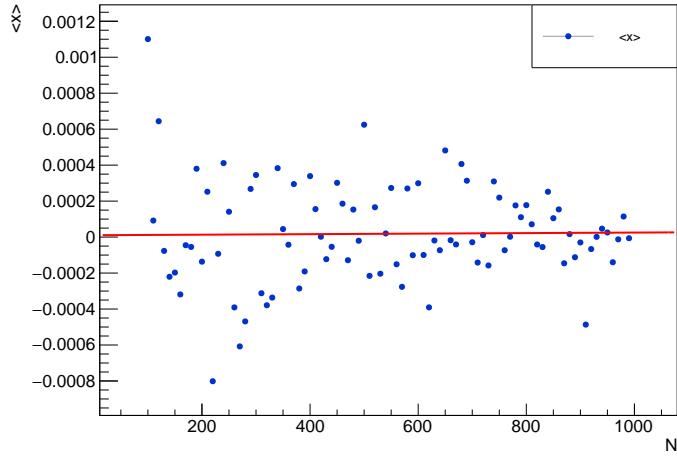


Figura 69:  $\langle \bar{x} \rangle$  con  $f$  uniforme: fit

I parametri stimati risultano essere

$$q = 1.04 \cdot 10^{-5} \approx 0 \quad \text{e} \quad m = 1.47 \cdot 10^{-8} \approx 0$$

La relazione  $\langle \bar{x} \rangle = \mu = 0$  risulta, dunque, verificata. Per la relazione (46) si è ottenuto quanto segue.

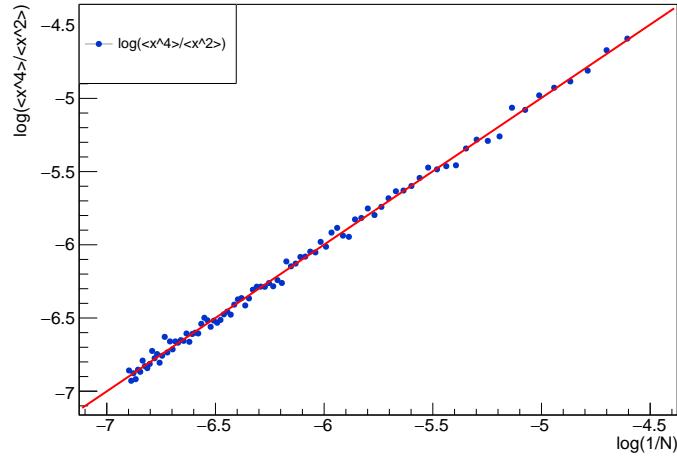


Figura 70:  $\log\left(\frac{\langle \bar{x}^4 \rangle}{\langle \bar{x}^2 \rangle}\right)$  con  $f$  uniforme: fit

I parametri stimati risultano essere

$$q = \log 3k = 0.014 \quad \text{e} \quad m = 1$$

Il valore del coefficiente angolare risulta consistente con il coefficiente della relazione in esame. L'esattezza della stima è dovuta al range di  $N$  più grande

considerato per il fit. Alla luce dei risultati, risulta possibile affermare la verifica della (46). Per la relazione (47) si è ottenuto quanto segue.

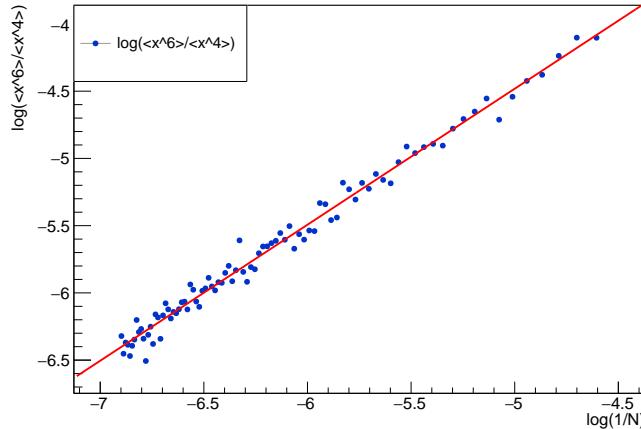


Figura 71:  $\log\left(\frac{\langle \bar{x}^6 \rangle}{\langle \bar{x}^4 \rangle}\right)$  con  $f$  uniforme: fit

I parametri stimati risultano essere

$$q = \log 5k = 0.576 \quad \text{e} \quad m = 1.01 \approx 1$$

Il valore del coefficiente angolare risulta consistente con il coefficiente della relazione in esame. Alla luce dei risultati, risulta possibile affermare la verifica della legge (47).

La verifica combinata delle relazioni  $\langle \bar{x} \rangle = \mu$ , (46) e (47) permette di concludere anche la verifica dinamica del teorema 0.11 per la pdf uniforme.

### pdf di Bernoulli

Si sono generati  $M = 10000$  pacchetti di valori  $\{x_i\}_{i=1,\dots,N}$  identicamente distribuiti secondo la distribuzione discreta di Bernoulli definita da

$$B(x | p) := p^x (1-p)^{1-x} \quad \text{con} \quad x \in \Omega = \{-1, +1\} \quad \text{e} \quad p = \frac{1}{2}$$

La distribuzione in esame è un caso particolare di distribuzione binomiale per un singolo lancio. Si noti che, per una Bernoulli, valgono

$$\mu = -\frac{1}{2} + \frac{1}{2} = 0 \quad \text{e} \quad \sigma = \sqrt{\frac{1}{2} \sum_{x \in \Omega} k^2} = 1$$

Si è scelto un valore di  $M$  sufficientemente grande al fine di ottenere un numero sufficiente di medie per facilitare la visualizzazione dell'andamento della distribuzione. Anche in questo caso, si è deciso di verificare il teorema 0.11 con due diverse modalità: una verifica statica ad un valore  $N = \tilde{N}$  sufficientemente grande fissato, e una verifica dinamica al variare di  $N$ , utilizzando le relazioni ricavate in precedenza.

## Verifica statica

Al fine di verificare qualitativamente l'approssimazione della distribuzione risultante a quella gaussiana, si sono calcolate le medie campionarie

$$\bar{x}_j := \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} x_{j,i} \quad \forall j = 1, \dots, M$$

con  $\tilde{N} = 1, 2, 5, 20, 1000$ . Di seguito sono riportati gli istogrammi dei dati ottenuti.

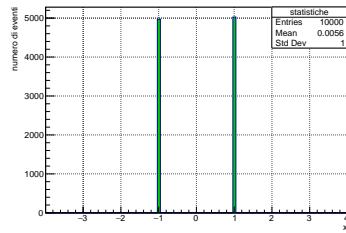


Figura 72:  $\tilde{N} = 1$  con  $f$  Bernoulli

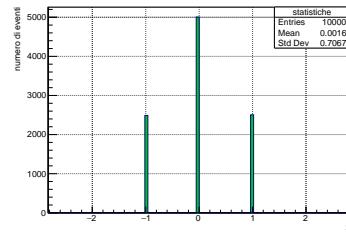


Figura 73:  $\tilde{N} = 2$  con  $f$  Bernoulli

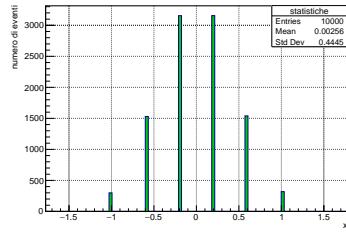


Figura 74:  $\tilde{N} = 5$  con  $f$  Bernoulli

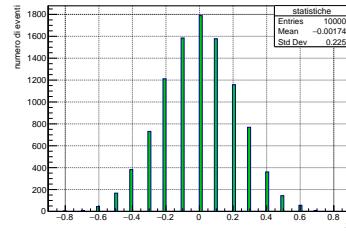


Figura 75:  $\tilde{N} = 20$  con  $f$  Bernoulli

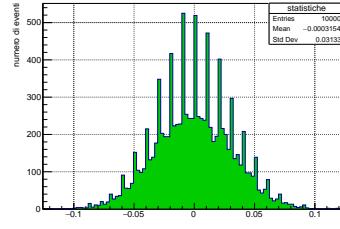


Figura 76:  $\tilde{N} = 1000$  con  $f$  Bernoulli

Come è possibile notare, l'andamento della media campionaria  $\bar{x}$  assume una forma a campana tipica della distribuzione gaussiana all'aumentare della dimensione  $N$  dei pacchetti, come ci si aspetta. Al fine di svolgere una prima verifica quantitativa sulla forma della distribuzione, si è deciso di fittare la distribuzione

binnata ottenuta per  $\tilde{N} = 10000$  con una gaussiana  $G(\bar{x} | \tilde{\mu}, \tilde{\sigma})$  non normalizzata, con  $\tilde{\mu}$  e  $\tilde{\sigma}$  parametri liberi. L'algoritmo di minimizzazione ha prodotto quanto segue.

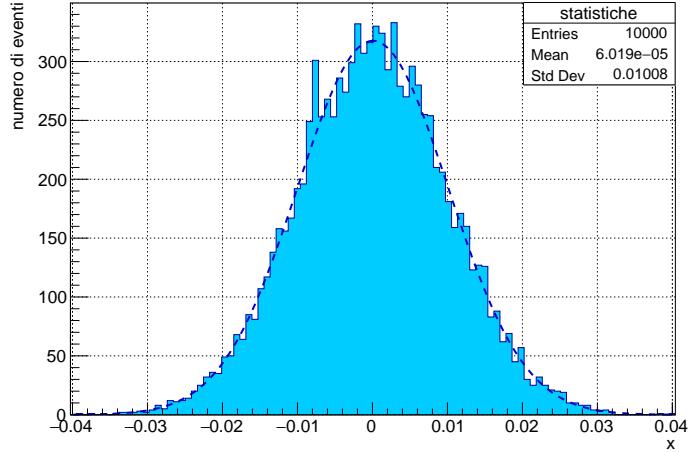


Figura 77: distribuzione per  $\tilde{N} = 10000$  con  $f$  Bernoulli: fit

I parametri stimati sono i seguenti.

$$\tilde{\mu} = 5.75 \cdot 10^{-5} \quad \text{e} \quad \tilde{\sigma} = 0.0101$$

Per il teorema 0.11 ci si aspetta che la gaussiana che descrive la distribuzione delle medie campionarie abbia la forma

$$G = G\left(\bar{x} | 0, \frac{\sigma}{\sqrt{\tilde{N}}}\right)$$

Dunque, la deviazione standard attesa risulta essere

$$\frac{\sigma}{\sqrt{\tilde{N}}} = \frac{1}{\sqrt{10000}} \approx 0.01$$

Ma allora, notiamo subito che

$$\tilde{\mu} \approx 0 \quad \text{e} \quad \tilde{\sigma} \approx 0.01$$

La non esattezza delle stime ottenute dipende, chiaramente, dal fatto che l'approssimazione alla distribuzione gaussiana vale asintoticamente. D'altra parte, computazionalmente, risulta possibile lavorare solo a valori di  $N$  finiti. I valori della media e della deviazione standard sono quindi consistenti con quelli dati dal teorema 0.11. Un'ulteriore conferma è data dal valore di probabilità associata al  $Q^2$  stimata dal programma di minimizzazione, che risulta ben al di sopra della soglia di accettabilità dell'ipotesi. Segue che la gaussiana è una forma funzionale che ben descrive la distribuzione ottenuta per  $\tilde{N} = 10000$ , rendendo possibile affermare la verifica del teorema per questo dato valore di  $N$ .

## Verifica dinamica

Al fine di verificare l'approssimazione della distribuzione risultante a quella gaussiana, si sono calcolati e plottati i momenti di ordine  $k = 1, \dots, 6$  delle medie campionarie come

$$\langle x^k \rangle_j := \frac{1}{N} \sum_{i=1}^N x_{j,i}^k \quad \forall j = 1, \dots, M$$

Al variare di  $N$  nel range

$$2 \leq N < 100 \quad \text{con} \quad N_{i+1} = N_i + 1$$

Si sono poi plottati i valori dei momenti in funzione della dimensione  $N$  del pacchetto di dati. Anche in questo caso, si è deciso di plottare separatamente i momenti di ordine pari da quelli di ordine dispari per confrontarne l'andamento. Per i momenti di ordine dispari si sono ottenuti i seguenti risultati.

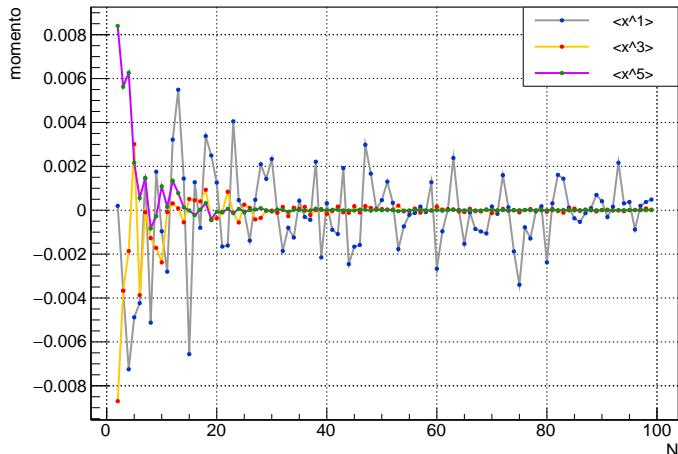


Figura 78: confronto  $\langle x^{2k-1} \rangle$  con  $f$  Bernoulli

Come è possibile notare, i momenti di ordine dispari della distribuzione di  $\bar{x}$  oscillano, in generale, in un intorno dello zero delle ordinate. Anche in questo caso, il fatto è ben giustificabile notando che

$$G(-\bar{x}) = G(\bar{x})$$

poiché la media della distribuzione di Bernoulli è, in questo caso, nulla. D'altra parte,  $x^{2k-1}$  è una funzione dispari. Segue che, per  $N$  grande vale

$$\langle x^{2k-1} \rangle = \sum_{x \in \Omega} x^{2k-1} G(x) = 0 \quad \forall k \in \mathbb{N} \setminus \{0\}$$

Il fatto che la relazione valga in regime asintotico spiega la ragione per la quale, per valori di  $N$  piccoli, si siano ottenuti valori più distanti dallo zero. Anche in questo caso, ciò che cambia nei momenti di ordine dispari è la rapidità con cui la funzione oscillante si stabilizza al valore nullo. Per i momenti di ordine pari si sono ottenuti i risultati che seguono.

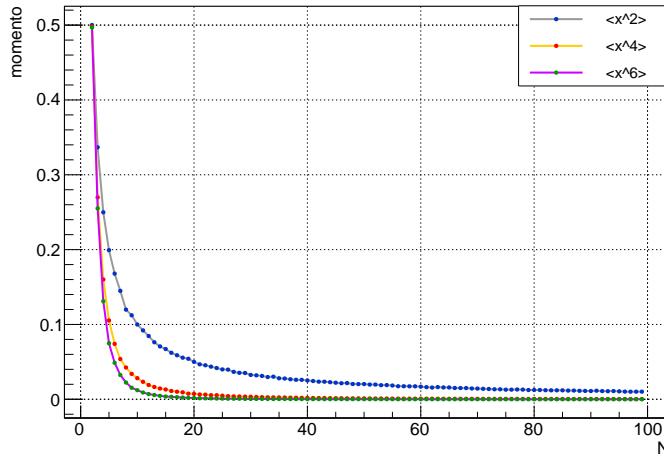


Figura 79: confronto  $\langle x^{2k} \rangle$  con  $f$  Bernoulli

Anche in questo caso, i risultati sono analoghi a quelli ottenuti in precedenza e qualitativamente consistenti alle relazioni che devono valere in regime asintotico. Al fine di verificare quantitativamente l'andamento gaussiano al variare di  $N$  si è quindi deciso di verificare la relazione  $\langle \bar{x} \rangle = \mu$  data direttamente dal teorema 0.11, la relazione (46) e la relazione (47). In particolare, nonostante l'aumento considerevole dei tempi di calcolo, al fine di ottenere una stima più accurata dei parametri si è deciso di calcolare i momenti, e quindi i loro rapporti, nel range

$$50 \leq N < 500 \quad \text{con} \quad N_{i+1} = N_i + 10$$

Per la prima relazione si sono ottenuti i risultati che seguono

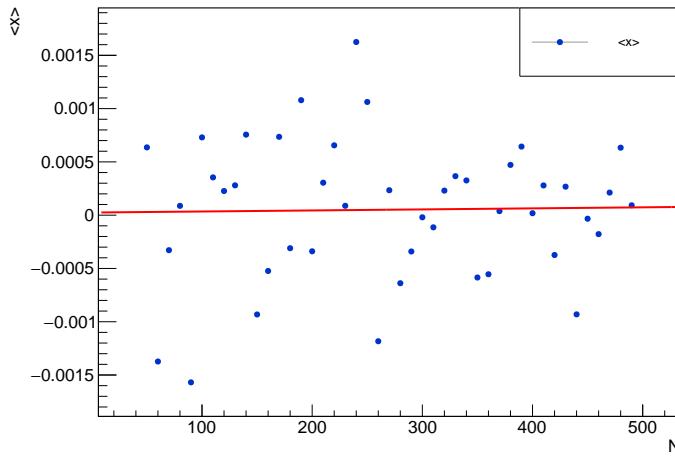


Figura 80:  $\langle \bar{x} \rangle$  con  $f$  Bernoulli: fit

I parametri stimati sono i seguenti.

$$q = 2.48 \cdot 10^{-5} \approx 0 \quad \text{e} \quad m = 9.76 \cdot 10^{-8} \approx 0$$

Vista la vicinanza di entrambe le stime al valore nullo, la relazione  $\langle \bar{x} \rangle = \mu = 0$  può dirsi, dunque, verificata. Per la relazione (46) si è ottenuto quanto segue.

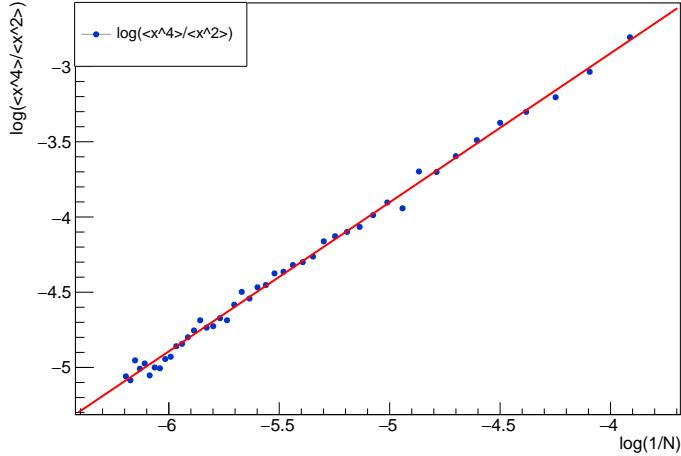


Figura 81:  $\log\left(\frac{\langle \bar{x}^4 \rangle}{\langle \bar{x}^2 \rangle}\right)$  con  $f$  Bernoulli: fit

Con i seguenti parametri stimati

$$q = \log 3k = 1.05 \quad \text{e} \quad m = 0.99 \approx 1$$

Il valore del coefficiente angolare risulta consistente con il coefficiente della relazione in esame. Alla luce dei risultati, risulta possibile affermare la verifica della (46). Per la relazione (47) si è ottenuto quanto segue.

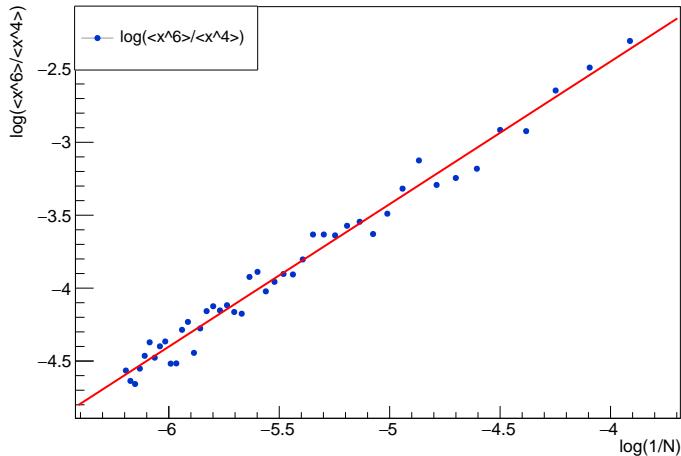


Figura 82:  $\log\left(\frac{\langle \bar{x}^6 \rangle}{\langle \bar{x}^4 \rangle}\right)$  con  $f$  Bernoulli: fit

Con i seguenti parametri stimati

$$q = \log 5k = 1.46 \quad \text{e} \quad m = 0.978 \approx 1$$

Il valore del coefficiente angolare risulta consistente con il coefficiente della relazione in esame. Alla luce dei risultati, risulta possibile affermare la verifica della legge (47).

La verifica combinata delle relazioni  $\langle \bar{x} \rangle = \mu$ , (46) e (47) permette di concludere anche la verifica dinamica del teorema 0.11 per la pdf di Bernoulli.

## Esercizio 10

Si vogliono stimare numericamente gli integrali  $I_3$ ,  $I_4$  e  $I_5$  utilizzando il metodo Monte Carlo sampling e il metodo hit or miss. In particolare, si vuole studiare la deviazione al variare del numero  $N$  di punti generato in entrambi i casi.

Si consideri il problema di calcolo del solito integrale definito  $I$ . Abbiamo già visto che il metodo Monte Carlo consiste nel generare  $N$  numeri pseudo-casuali  $\{x_i\}_{i=1,\dots,N}$  nell'intervallo di integrazione  $(a, b)$  con distribuzione uniforme. In particolare, la stima è data dalla relazione

$$I \approx \frac{b-a}{N} \sum_{i=1}^N f(x_i) \quad \text{con } N \text{ grande} \quad (48)$$

Ricordando la definizione di media campionaria, lo stimatore dato dalla (48) può essere riscritto in modo più compatto come

$$I \approx (b-a)\langle f \rangle$$

Il metodo hit or miss, invece, consiste nel generare un'area  $R$  all'interno della quale è contenuto l'integrale  $I$  che si vuole calcolare. In particolare, vengono generati  $N$  punti  $\{(x_i, y_i)\}_{i=1,\dots,N}$  con distribuzione uniforme all'interno di  $R$ , per poi contare il numero di punti  $N_h$  che sono caduti all'interno dell'area  $I$  che identifica l'integrale di cui si vuole avere una stima. La stima che si vuole calcolare sarà allora data dalla relazione

$$I \approx R \frac{N_h}{N} \quad \text{con } N \text{ grande} \quad (49)$$

Si noti che il metodo Monte Carlo sampling e il metodo hit or miss sono, a ben vedere, del tutto equivalenti. Si consideri, infatti, la funzione di due variabili reali

$$g(x, y) = \begin{cases} 0 & \text{se } (x, y) \notin I \\ 1 & \text{se } (x, y) \in I \end{cases} \quad \text{con } (x, y) \in R$$

Generalizzando il metodo Monte Carlo sampling in due dimensioni, vale

$$\iint_R g(x, y) dx dy \approx R \frac{1}{N} \sum_{i=1}^N g(x_i, y_i) = R \frac{N_h}{N}$$

che coincide esattamente con la stima (49) data dal metodo hit or miss. Dalla coincidenza dei due metodi mostrata segue che l'andamento dell'errore sarà lo stesso in entrambi i casi. I metodi mostrati si fondano, quindi, sul calcolo della media

$$I \approx V^{(n)} \langle f \rangle$$

Dove  $V^{(n)}$  rappresenta, in generale, il volume  $n$  dimensionale in cui sono stati generati i punti. L'unica sorgente di errore risulta essere associata alla media  $\langle f \rangle$ . L'errore associato alla stima di  $I$  sarà quindi dati dall'errore sulla media

$$\text{err}(I) = V^{(n)} \text{err} \langle f \rangle = V^{(n)} \frac{\sigma_f}{\sqrt{N}}$$

Ricordando che per la varianza vale la relazione notevole

$$\sigma_f^2 = \langle f^2 \rangle - \langle f \rangle^2$$

possiamo allora scrivere

$$\text{err}(I) = \frac{1}{\sqrt{N}} V^{(n)} \sqrt{\langle f^2 \rangle - \langle f \rangle^2}$$

Un modo pratico utile per lo studio dell'errore consiste, come al solito, nello studio della dispersione dal valore vero, definita nel modo usuale come

$$\Delta(N) := |I - \tilde{I}(N)|$$

Dove  $\tilde{I}(N)$  rappresenta una stima di  $I$  per  $N$  punti con i metodi introdotti. Infatti, vale banalmente la relazione

$$\text{err}(I) \propto \Delta(N) \iff \text{err}(I) = \tilde{k} \Delta(N)$$

Arrangiando i termini e le costanti reali si ottiene

$$\Delta(N) = k \frac{1}{\sqrt{N}} \quad \text{con} \quad N \text{ grande} \quad \text{e} \quad k \in \mathbb{R} \quad (50)$$

Con l'usuale tecnica di passaggio ai logaritmi si avrà la relazione lineare più facilmente verificabile

$$\log \Delta(N) = \log k - \frac{1}{2} \log N \quad (51)$$

Al fine di verificare il corretto andamento dell'errore nei due metodi sarà quindi sufficiente stimare il coefficiente angolare della retta (51) per mezzo di un fit, per poi confrontare la stima con il valore atteso.

Di seguito sono riportati gli studi sulla dispersione dal valore vero per i due metodi descritti applicati agli integrali  $I_3$ ,  $I_4$  e  $I_5$ , così come definiti negli esercizi precedenti.

### Monte Carlo

Si sono calcolate le dispersioni  $\Delta_3$ ,  $\Delta_4$  e  $\Delta_5$  con il metodo Monte Carlo sampling al variare del numero  $N$  di punti generati casualmente. In particolare, si è considerato il range

$$1000 \leq N < 100000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

assicurandosi di lavorare in regime asintotico e di definire l'andamento in un intervallo sufficientemente ampio di punti. In particolare, al fine di verificare la relazione (51), si sono plottati i logaritmi delle dispersioni al variare dei logaritmi del numero di punti, per poi interpolare i dati con una relazione lineare della forma  $y = p + mx$ . Analizziamo quindi i tre diversi integrali singolarmente nel dettaglio.

### I<sub>3</sub>

Per l'integrale  $I_3$  si sono ottenuti i risultati che seguono.

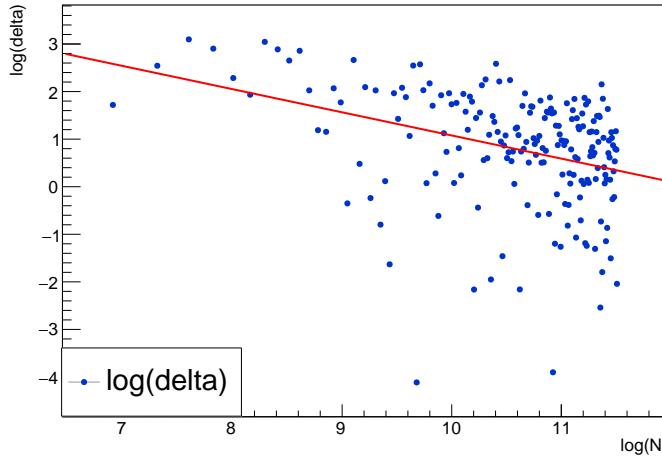


Figura 83:  $\log \Delta_3$  Monte Carlo: fit

I parametri stimati risultano

$$p = 5.94 = \log k \quad \text{e} \quad m = -0.487 \approx -0.5$$

Come è possibile notare, seppur il fit abbia stimato un coefficiente angolare compatibile con quello della relazione (51), la stima di  $m$  risulta distorta di un fattore considerevole. Tale fatto è consistente con quanto si nota dal grafico: seppur mediamente l'andamento cercato sia verificato, i punti risultano sparsi in un range delle ordinate sufficientemente ampio. Al fine di migliorare la stima di  $m$  risulta quindi opportuno addensare maggiormente i punti intorno all'andamento rettilineo cercato. Per farlo, si è deciso di stimare  $M$  volte l'integrale allo stesso  $N$  fissato, per poi calcolare

$$\langle \Delta_3 \rangle_M \quad \forall N = N_{min}, \dots, N_{max}$$

Al posto di associare ad ogni  $N$  la dispersione dal valore vero di una singola stima si è quindi associato, ad ogni  $N$ , il valore medio di  $M$  diverse dispersioni dell'integrale. La media campionaria, infatti, è uno stimatore non distorto del valore di aspettazione di una variabile casuale. Come tale, permette di migliorare una data stima all'aumentare della dimensione del campione, ossia all'aumentare di  $M$ . In altre parole, ci si aspetta che la stima dell'integrale possa aumentare in precisione per ogni  $N$ , definendo un andamento meno dispersivo intorno alla retta cercata. Evidentemente, l'andamento sarà meglio definito all'aumentare del numero  $M$  di elementi che vengono mediati ad  $N$  fissato. Si sono quindi svolte le medesime operazioni per  $M = 1, 20, 100$ . Si sono poi sovrapposti i diversi plot al fine di verificare l'addensamento dell'andamento intorno ad una retta all'aumentare del numero  $M$  di elementi mediati. Di seguito sono riportati i risultati ottenuti.

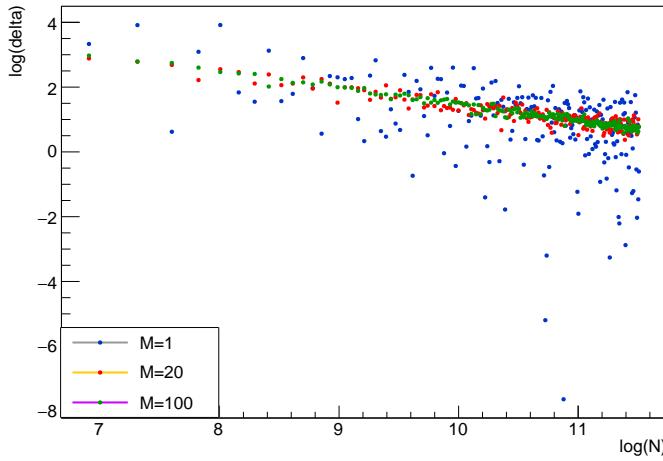


Figura 84:  $\log \Delta_3$  Monte Carlo: confronto per  $M = 1, 20, 100$

Risulta evidente che, all'aumentare della dimensione  $M$  del pacchetto di dati, l'andamento della dispersione in funzione di  $N$  divenga, qualitativamente, più rettilineo. Al fine di verificarlo anche quantitativamente, si è svolto un fit con i soli dati ottenuti per  $M = 100$ , ottenendo quanto segue.

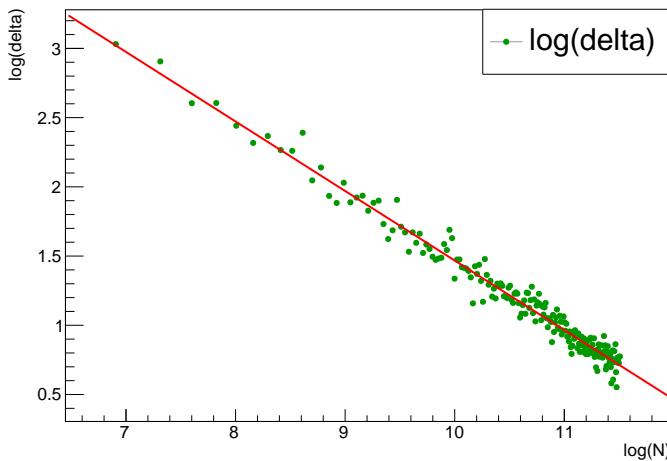


Figura 85:  $\log \Delta_3$  Monte Carlo per  $M = 100$ : fit

I parametri stimati risultano

$$p = 6.49 = \log k \quad \text{e} \quad m = -0.502 \approx -0.5$$

Come previsto, il parametro  $m$  risulta più compatibile con il coefficiente angolare della relazione (51), coerentemente con quanto si osserva qualitativamente. La media campionaria di  $M$  diverse dispersioni ha quindi permesso di verificare con più precisione la relazione che governa la dispersione del metodo al variare del numero di punti.

#### I<sub>4</sub>

Si è notato che  $I_3$ ,  $I_4$  e  $I_5$  presentano caratteristiche di regolarità analoghe. Per gli integrali rimanenti si è quindi deciso di plottare direttamente il confronto al variare di  $M$ , per poi verificare la relazione in modo più accurato per  $M = 100$ , in quanto ci si aspetta un comportamento qualitativo simile a quello osservato nella stima precedente. Di seguito sono riportati i risultati ottenuti.

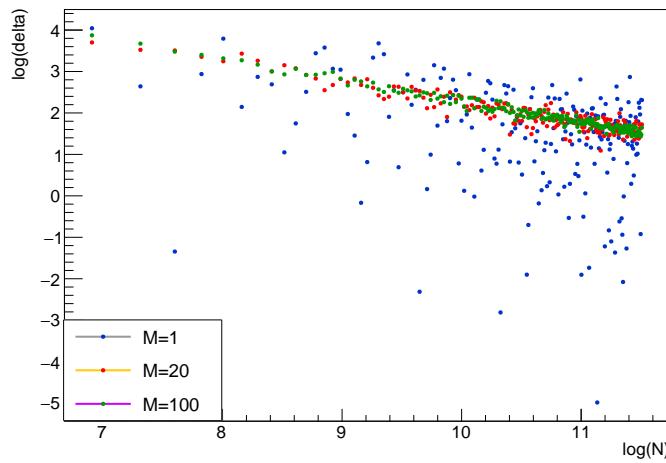


Figura 86:  $\log \Delta_4$  Monte Carlo: confronto per  $M = 1, 20, 100$

Anche in questo caso, i punti si addensano maggiormente intorno ad una retta all'aumentare della dimensione  $M$  del pacchetto di dati su cui si effettua la media, come ci si aspetta. Si sono quindi interpolati i dati per  $M = 100$ , ottenendo quanto segue.

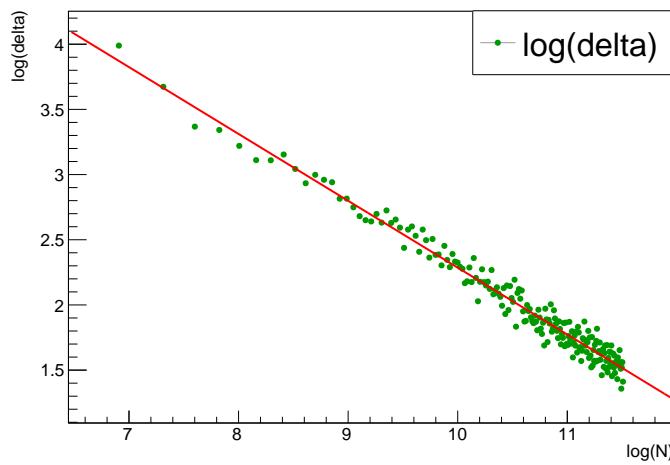


Figura 87:  $\log \Delta_4$  Monte Carlo per  $M = 100$ : fit

I parametri stimati risultano

$$p = 7.41 = \log k \quad \text{e} \quad m = -0.513 \approx -0.5$$

La compatibilità tra la stima di  $m$  ottenuta e il coefficiente angolare atteso permette di concludere la verifica numerica della (51).

### **I<sub>5</sub>**

Di seguito sono riportati i risultati ottenuti per  $M = 1, 20, 100$ .

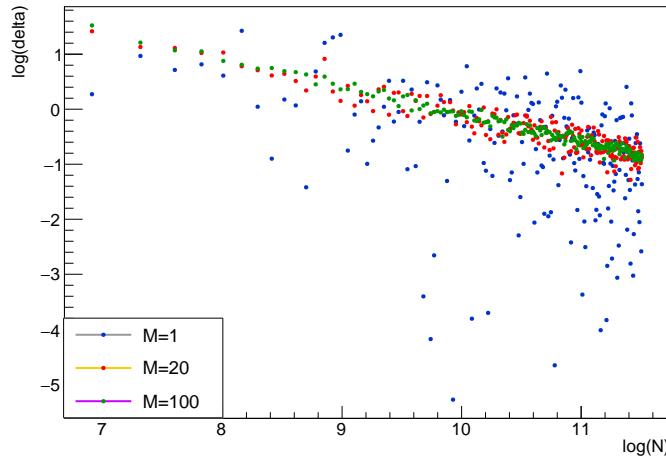


Figura 88:  $\log \Delta_5$  Monte Carlo: confronto per  $M = 1, 20, 100$

Anche in questo caso, i punti si addensano maggiormente in un andamento rettilineo all'aumentare di  $M$ , come ci si aspetta. Si sono quindi interpolati i dati per  $M = 100$ , ottenendo quanto segue.

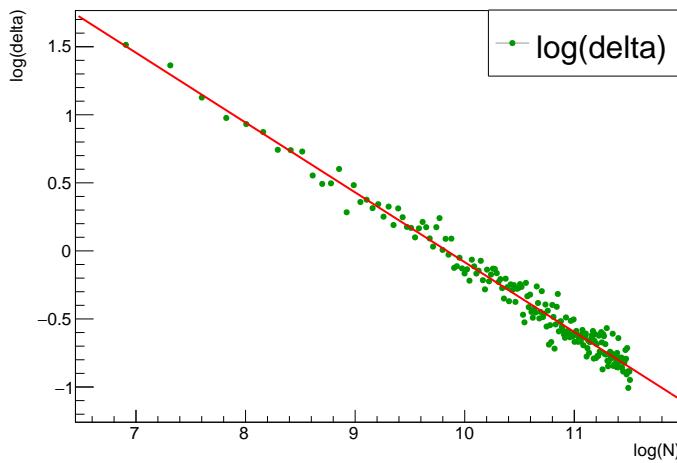


Figura 89:  $\log \Delta_5$  Monte Carlo per  $M = 100$ : fit

I parametri stimati risultano

$$p = 5.05 = \log k \quad \text{e} \quad m = -0.513 \approx -0.5$$

La compatibilità tra la stima di  $m$  ottenuta e il coefficiente angolare della relazione (51) permette di concludere la verifica della legge che governa l'errore della stima al variare del numero di punti.

### Hit or miss

Si sono calcolate le dispersioni dal valore vero  $\Delta_3$ ,  $\Delta_4$  e  $\Delta_5$  con il metodo hit or miss, al variare del numero  $N$  di punti generato casualmente. In particolare, si è considerato il range

$$1000 \leq N < 100000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

assicurandosi di lavorare in regime asintotico e di definire l'andamento in un intervallo sufficientemente ampio di punti. Al fine di verificare la relazione (51) si sono piazzati i logaritmi delle dispersioni al variare dei logaritmi del numero di punti, per poi interpolare i dati con la relazione lineare  $y = p + mx$ .

### I<sub>3</sub>

Di seguito sono riportati i risultati ottenuti.

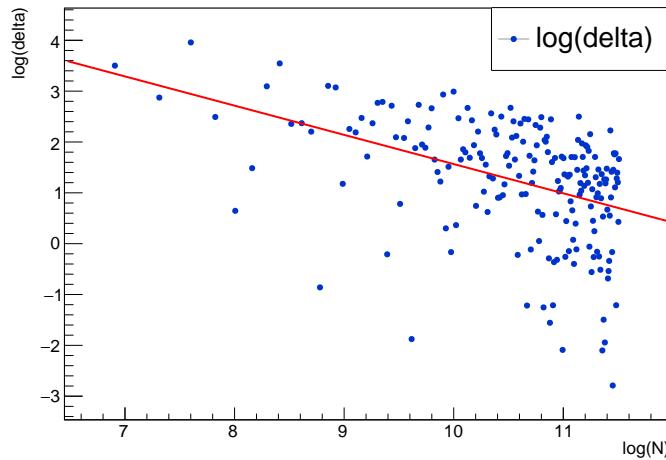


Figura 90:  $\log \Delta_3$  hit or miss: fit

I parametri stimati risultano

$$p = 7.31 = \log k \quad \text{e} \quad m = -0.575 \approx -0.5$$

Anche nel caso del metodo hit or miss, seppur il fit abbia stimato un coefficiente angolare compatibile con quello della relazione (51), la stima di  $m$  risulta distorta di un fattore non trascurabile. Anche in questo caso, dunque, al fine di

migliorare la stima del coefficiente angolare  $m$ , si sono calcolate le medie delle dispersioni dal valore vero come

$$\langle \Delta_3 \rangle_M \quad \forall N = N_{min}, \dots, N_{max}$$

fissando  $M$  dimensione del pacchetto di stime della dispersione. Si sono quindi svolte le medesime operazioni per  $M = 1, 20, 100$ . Si sono poi sovrapposti i diversi plot al fine di verificare l'addensamento dell'andamento intorno ad una retta, come segue.

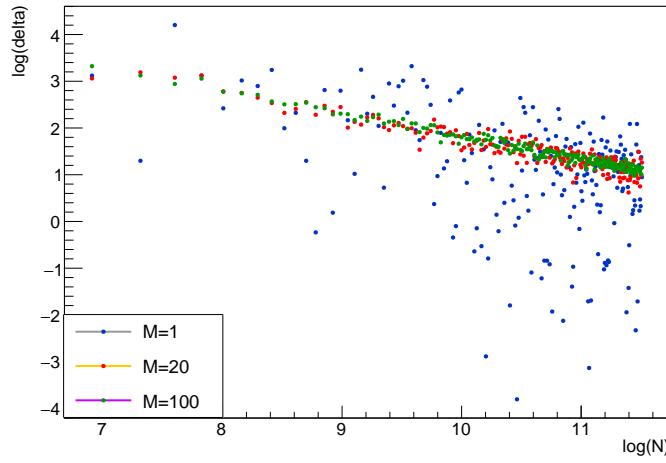


Figura 91:  $\log \Delta_3$  hit or miss: confronto per  $M = 1, 20, 100$

Si sono quindi fittati i dati per  $M = 100$  ottenendo quanto segue.

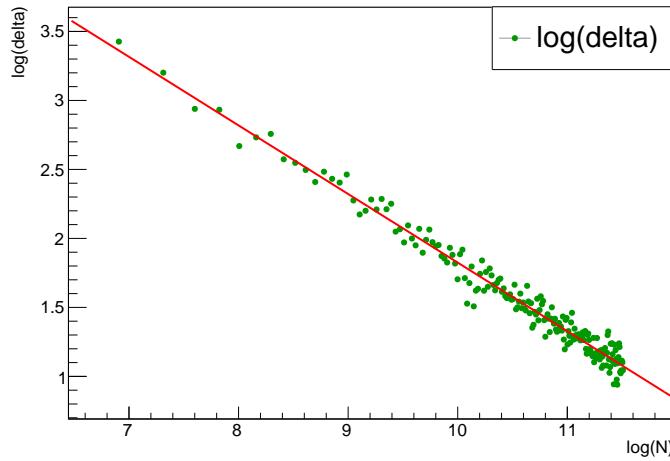


Figura 92:  $\log \Delta_3$  hit or miss per  $M = 100$ : fit

I parametri stimati risultano

$$p = 6.8 = \log k \quad \text{e} \quad m = -0.497 \approx -0.5$$

Come ci si aspetta, il coefficiente  $m$  risulta molto più vicino al valor vero nel caso di  $M = 100$  rispetto al caso  $M = 1$  studiato precedentemente. La compatibilità tra la stima di  $m$  ottenuta e il coefficiente angolare della relazione (51) permette di concludere la verifica della legge che governa l'errore della stima al variare del numero di punti.

#### I<sub>4</sub>

Anche in questo caso si è deciso, per gli integrali rimanenti, di plottare direttamente il confronto al variare di  $M$ . Per  $I_4$  si è ottenuto quanto segue.

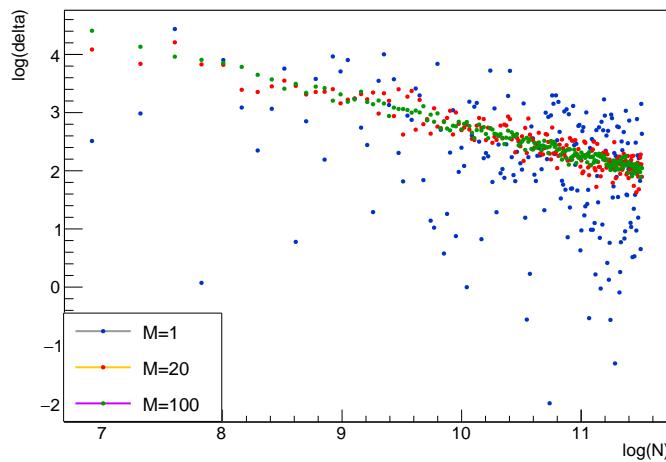


Figura 93:  $\log \Delta_4$  hit or miss: confronto per  $M = 1, 20, 100$

Si sono quindi interpolati i dati per  $M = 100$ , ottenendo i seguenti risultati.

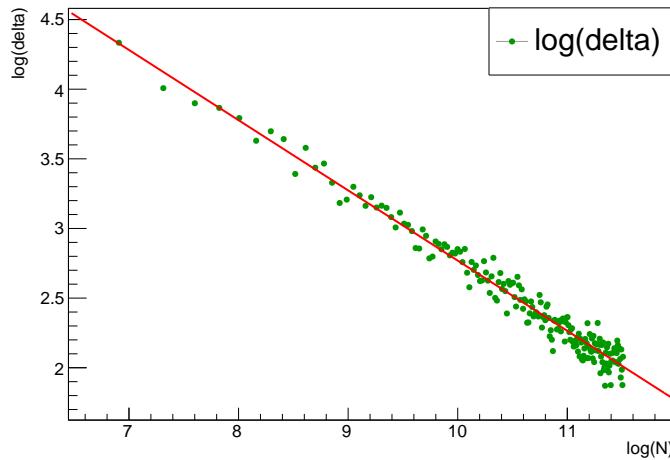


Figura 94:  $\log \Delta_4$  hit or miss per  $M = 100$ : fit

I parametri stimati risultano

$$p = 7.81 = \log k \quad \text{e} \quad m = -0.504 \approx -0.5$$

La stima di  $m$  ottenuta permette di concludere la verifica della relazione (51), vista la compatibilità con il coefficiente angolare noto della relazione.

### **$I_5$**

Di seguito sono riportati i risultati ottenuti dal calcolo delle dispersioni per  $M = 1, 20, 100$  nel caso dell'integrale  $I_5$ .

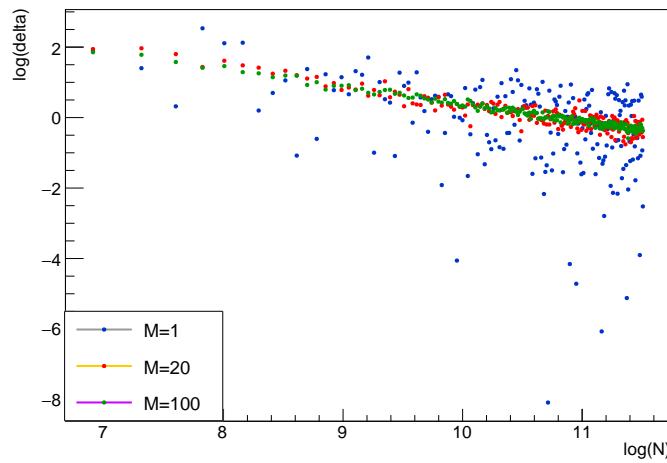


Figura 95:  $\log \Delta_5$  hit or miss: confronto per  $M = 1, 20, 100$

Si sono quindi interpolati i dati per  $M = 100$ , ottenendo quanto segue.

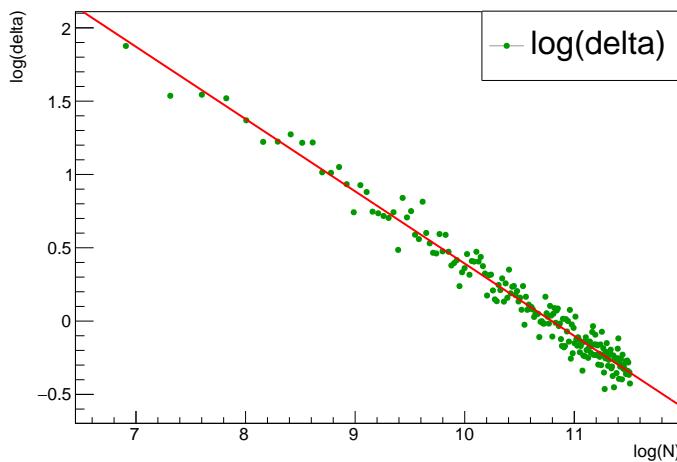


Figura 96:  $\log \Delta_5$  hit or miss per  $M = 100$ : fit

I parametri stimati risultano

$$p = 5.32 = \log k \quad \text{e} \quad m = -0.493 \approx -0.5$$

Anche in questo caso, la compatibilità tra la stima di  $m$  ottenuta e il coefficiente angolare della relazione (51) permette di concludere la verifica della legge che governa l'errore della stima al variare del numero di punti.

In definitiva, sia nel caso del metodo Monte Carlo sampling, sia per il metodo hit or miss si sono ottenuti risultati, sostanzialmente, sovrapponibili in termini di andamento asintotico dell'errore. Tale verifica è consistente con l'equivalenza dei due metodi mostrata in precedenza. L'applicazione del metodo hit or miss è di particolare utilità nel caso di funzioni integrande regolari su un intervallo di integrazione altrettanto regolare. Il metodo Monte Carlo, invece, come si ha già avuto modo di mostrare, è particolarmente efficace per integrali in più dimensioni o in caso di domini di integrazione complicati. Ad ogni modo, l'utilizzo di un metodo piuttosto che un altro, in questo caso, è del tutto equivalente in termini di precisione della stima.

## Esercizio 11

Si vogliono generare numeri pseudo-casuali distribuiti secondo diverse funzioni densità di probabilità  $f_i(x)$  all'interno dell'intervallo  $\Omega_i \subset \mathbb{R}$  utilizzando il metodo esatto della funzione inversa.

**Proposizione 0.13** (metodo della funzione inversa). *Sia  $U$  una variabile casuale continua distribuita uniformemente in  $[0, 1]$ . Sia  $F$  la distribuzione cumulativa della distribuzione  $f$ . Allora, la variabile casuale*

$$X := F^{-1}(U)$$

*è distribuita secondo la funzione densità di probabilità  $f$ .*

Il metodo della funzione inversa permette, dunque, di generare un campione di  $N$  numeri casuali  $\{x_i\}_{i=1,\dots,N}$  distribuiti come  $f$ , a patto di essere in grado di determinare analiticamente una primitiva di tale funzione e di essere, poi, in grado di invertirla. Evidentemente, seppur la 0.13 permetta di generare numeri casuali in modo esatto, il metodo della funzione inversa è applicabile solo quando la funzione  $f$  è sufficientemente semplice, tale da permettere l'integrazione per via analitica, ossia tale che sia garantita la possibilità di riuscire ad esprimere la primitiva per mezzo di una combinazione di funzioni elementari. Inoltre, la cumulativa  $F$  deve risultare biettiva (o, al più, iniettiva) all'interno dell'intervallo  $\Omega$  nel quale si vogliono generare le sequenze pseudo-casuali. La biettività, infatti, è condizione necessaria e sufficiente per l'invertibilità di una funzione. Si noti, infatti, che seppur la cumulativa  $F$  sia, per definizione, una funzione monotona crescente del suo argomento, questo non garantisce l'iniettività della mappa nell'intervallo considerato: può infatti accadere che la cumulativa possa presentare tratti costanti in qualche sottoinsieme di  $\Omega$ .

Ricordiamo poi che una funzione densità di probabilità  $f$  è sempre, per definizione, normalizzata all'interno dell'intervallo  $\Omega$  su cui è definita. Per il secondo assioma della definizione costruttiva di probabilità di Kolmogorov deve infatti valere

$$\int_{\Omega} f(x) dx = 1 \quad (52)$$

A seguito della verifica della positività, il passaggio centrale che verrà eseguito in ognuno dei casi sarà, dunque, quello di imporre la condizione (52) al fine di permettere la corretta generazione nell'intervallo  $\Omega$  selezionato.

### Esponenziale in $\Omega$ limitato

Si vogliono generare numeri random distribuiti secondo la funzione

$$f(x) = ke^{-x} \quad \text{con } x \in \Omega = (0, 2)$$

Anzitutto, si noti che, siccome l'esponenziale è sempre positivo, se  $k > 0$  anche la candidata densità in esame sarà sempre positiva qualunque sia il valore del numero casuale sull'asse reale. In particolare, dunque, sarà vero che

$$f(x) > 0 \quad \forall x \in \Omega \quad \text{con } k > 0$$

Per valori positivi di  $k$ , risulta quindi verificato il primo assioma della 0.8, poiché integrali di funzioni positive restituiscono valori positivi. Si è poi determinata la costante  $k \in \mathbb{R}$  di normalizzazione imponendo la (52) come

$$\int_{\Omega} f(x) dx = 1 \iff k \int_0^2 e^{-x} dx = 1 \iff k \frac{e^2 - 1}{e^2} = 1$$

Da cui segue che la costante di normalizzazione assume la forma

$$k = \frac{e^2}{e^2 - 1} > 0$$

Si è quindi calcolata la funzione cumulativa integrando la mappa in esame da 0 fino ad un generico estremo  $x$ , ottenendo

$$F(x) = k \int_0^x e^{-t} dt = k(1 - e^{-x})$$

che risulta chiaramente biunivoca in  $\Omega = (0, 2)$ . In particolare, ricavando  $x$  in funzione di  $F(x)$  si è determinata l'inversa della cumulativa, che assume la forma

$$F^{-1}(x) = -\log\left(1 - \frac{x}{k}\right) \quad (53)$$

Si è quindi utilizzato il metodo della funzione inversa dato dalla 0.13 con la funzione (53). In particolare, si sono generati  $N = 1000000$  numeri pseudo-casuali, al fine di visualizzare nel dettaglio l'andamento. Di seguito è riportato l'istogramma dei dati ottenuto.

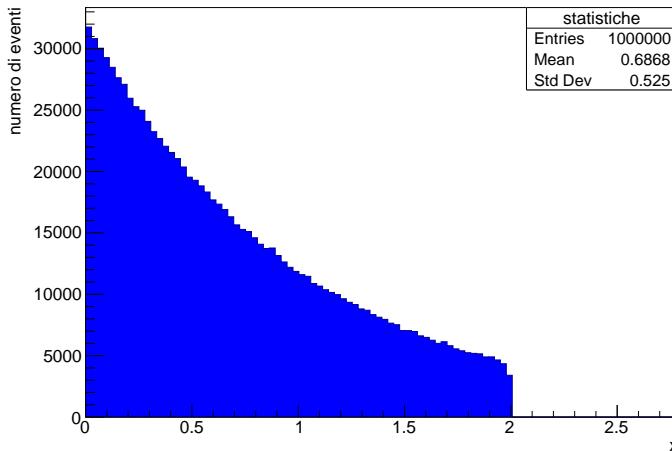


Figura 97: distribuzione esponenziale in  $(0, 2)$  per  $N = 1000000$

Come è possibile notare, l'istogramma non normalizzato ottenuto mostra qualitativamente un andamento esponenziale decrescente nell'intervallo  $(0, 2)$ , come ci si aspetta. Al fine di verificare quantitativamente la corretta generazione si è deciso di svolgere un fit della distribuzione binnata interpolando i dati con la funzione  $f(x) = ae^{bx}$ . Di seguito sono riportati i risultati ottenuti.

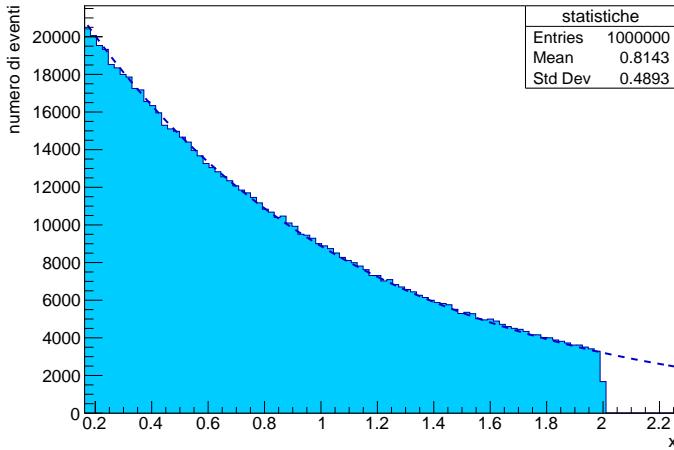


Figura 98: distribuzione esponenziale in  $(0, 2)$  per  $N = 1000000$ : fit

I parametri stimati dal programma di fit risultano

$$a = 2.46 \cdot 10^4 \quad \text{e} \quad b = -1.02 \approx 1$$

Come è possibile notare, a meno della normalizzazione dell'istogramma data dal parametro  $a$ , il parametro  $b$  risulta compatibile con il valore della costante che ci aspettiamo. La forma analitica della distribuzione risulta, dunque, consistente con la forma attesa. La correttezza della costante di normalizzazione  $k$  può essere verificata visivamente notando che il metodo della funzione inversa implementato ha prodotto sequenze pseudo-casuali esattamente nel range  $(0, 2)$ , ossia il range all'interno del quale si è normalizzata la funzione  $f$ . Un altro modo per la verifica della corretta generazione può essere quello dello studio dei momenti della distribuzione, che per una pdf esponenziale si calcolano con facilità. Tuttavia, i momenti di una distribuzione sono, in generale, meno informativi rispetto ad un fit completo dei dati ottenuti. Per tale regione, si è preferito non percorrere questa seconda strada. Si noti che non è stato necessario verificare il terzo assioma della probabilità per garantire la corretta generazione. La ragione risiede nel fatto che l'assioma di linearità è già automaticamente dato grazie alle proprietà dell'operatore di integrazione. Anche negli studi che seguono, dunque, la verifica del terzo assioma non sarà necessaria.

### Esponenziale in $\Omega$ illimitato

Si vogliono generare numeri random distribuiti secondo la funzione

$$f(x) = ke^{-x} \quad \text{con} \quad x \in \Omega = (1, +\infty)$$

Anzitutto, si è notato che, anche in questo caso, la candidata densità di probabilità è definita come un esponenziale. Per tale ragione, se  $k > 0$ , la mappa in esame sarà positiva su tutto l'asse reale. In particolare, sarà vero che

$$f(x) > 0 \quad \forall x \in \Omega \quad \text{con} \quad k > 0$$

Per valori positivi di  $k$ , risulta quindi verificato il primo assioma della 0.8. Al fine di verificare anche il secondo assioma, si è poi determinata la costante  $k \in \mathbb{R}$  di normalizzazione imponendo la (52) come

$$\int_{\Omega} f(x) dx = 1 \iff k \int_1^{+\infty} e^{-x} dx = 1 \iff ke^{-1} = 1$$

Da cui segue che la costante di normalizzazione assume la forma

$$k = e > 0$$

Si è quindi calcolata la funzione cumulativa

$$F(x) = k \int_1^x e^{-t} dt = k \left( \frac{1}{e} - \frac{1}{e^x} \right)$$

che risulta chiaramente biunivoca in  $\Omega = (1, +\infty)$ . In particolare, ricavando  $x$  in funzione di  $F(x)$  si è trovata l'inversa di  $F$ , che assume la forma

$$F^{-1}(x) = -\log \left( \frac{1}{e} - \frac{x}{k} \right) \quad (54)$$

Si è quindi utilizzato il metodo della funzione inversa dato dalla 0.13 con la funzione (54). In particolare, si sono generati  $N = 1000000$  numeri pseudo-casuali, al fine di visualizzare con più dettaglio l'andamento risultante. Di seguito è riportato l'istogramma ottenuto.

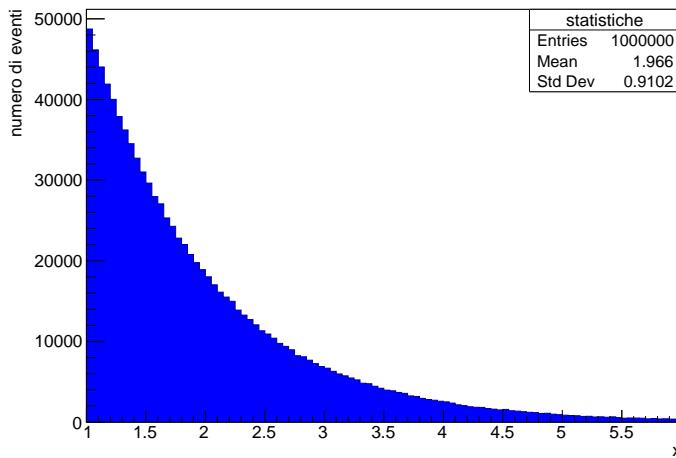


Figura 99: distribuzione esponenziale in  $(1, +\infty)$  per  $N = 1000000$

Come è possibile notare, l'istogramma non normalizzato ottenuto mostra qualitativamente un andamento esponenziale decrescente nell'intervallo  $(1, +\infty)$ , come ci si aspetta. Al fine di verificare quantitativamente la corretta generazione esponenziale si è deciso di svolgere un fit della distribuzione binnata interpolando i dati con la funzione  $f(x) = ae^{bx}$ . Di seguito sono mostrati i risultati ottenuti.

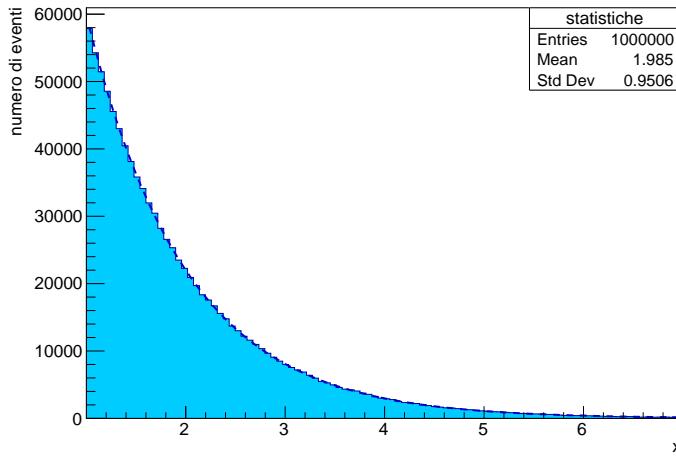


Figura 100: distribuzione esponenziale in  $(1, +\infty)$  per  $N = 1000000$ : fit

I parametri stimati dal programma di fit risultano

$$a = 1.63 \cdot 10^5 \quad \text{e} \quad b = -1$$

Come è possibile notare, a meno della normalizzazione dell'istogramma data dal parametro  $a$ , il parametro  $b$  risulta compatibile con il valore della costante che ci si aspetta. In questo caso, la stima di  $b$  risulta esatta. Questo è dovuto al fatto che, a differenza del caso precedente, si sono generati numeri random con pdf esponenziale in un range illimitato. L'interpolazione dei dati, in questo caso, a parità di eventi  $N$  generati sarà, dunque, più precisa, in quanto il programma di fit è in grado di vedere l'andamento della distribuzione in un range maggiore. La forma analitica della distribuzione risulta, dunque, consistente con la forma attesa. La correttezza della costante di normalizzazione  $k$  può essere verificata visivamente notando che il metodo della funzione inversa implementato ha prodotto sequenze pseudo-casuali nel range  $(1, +\infty)$ , ossia il range all'interno del quale si è normalizzata la funzione  $f$ .

### Combinazione di funzioni in $\Omega$ illimitato

Si vogliono generare numeri random distribuiti secondo la funzione

$$f(x) = kxe^{-x^2} \quad \text{con} \quad x \in \Omega = (0, +\infty)$$

Anzitutto, come al solito, risulta necessario verificare il segno della candidata distribuzione. Escludendo il valore inammissibile  $k = 0$ , studiando una banale disequazione parametrica si ha

$$\begin{cases} f(x) > 0 & \forall x \in (0, +\infty) \quad \text{se} \quad k > 0 \\ f(x) > 0 & \forall x \in (-\infty, 0) \quad \text{se} \quad k < 0 \end{cases}$$

Per valori positivi di  $k$ , risulta quindi verificato il primo assioma della definizione di probabilità 0.8 all'interno dell'intervallo  $\Omega$  in esame. Si è quindi determinata

la costante  $k \in \mathbb{R}$  di normalizzazione imponendo la (52), ossia calcolando il solito integrale

$$\int_{\Omega} f(x) dx = 1 \iff k \int_0^{+\infty} xe^{-x^2} dx = 1 \iff k \frac{1}{2} = 1$$

La sostituzione che ha permesso di ricondurre l'integrale in esame ad un integrale notevole è mostrata di seguito. Dal calcolo svolto, segue che la costante di normalizzazione è positiva e assume la forma

$$k = 2 > 0$$

Ma allora, per il valore di  $k$  determinato, la candidata densità verifica il primo assioma di Kolmogorov. Si è quindi calcolata la funzione cumulativa

$$F(x) = k \int_0^x te^{-t^2} dt$$

Posto il cambio di variabile

$$u = t^2 \implies du = 2tdt$$

ben definito in quanto sufficientemente regolare e invertibile nell'intervallo in esame, si ha l'integrale immediato

$$F(x) = \frac{k}{2} \int_0^{x^2} e^{-u} du = \frac{k}{2} \left( 1 - e^{-x^2} \right)$$

Si noti che la mappa  $F$  determinata non risulta iniettiva su tutto l'asse reale. Tuttavia, non è difficile verificare l'iniettività nell'intervallo  $(0, +\infty)$ , ossia all'interno dell'intervallo  $\Omega$  di nostro interesse per la generazione di sequenze pseudo-casuali. Si è quindi calcolata la funzione inversa della cumulativa ricavando  $x$  in funzione di  $F(x)$ , ottenendo la coppia di equazioni

$$F_{\pm}^{-1}(x) = \pm \sqrt{-\log \left( 1 - \frac{2x}{k} \right)}$$

Siccome deve valere  $x \in (0, +\infty)$  si è selezionato il solo ramo positivo. L'inversa della cumulativa nell'intervallo in esame avrà allora la forma

$$F^{-1}(x) = \sqrt{-\log \left( 1 - \frac{2x}{k} \right)} \tag{55}$$

Si è quindi utilizzato il metodo della funzione inversa dato dalla 0.13 utilizzando la funzione (55). In particolare, si sono generati  $N = 1000000$  numeri pseudo-casuali al fine di definire in dettaglio, da un punto di vista visivo, l'andamento della distribuzione. Si noti, inoltre, che considerare un numero elevato di misure non offre solo vantaggi grafici di visualizzazione, ma permette anche di ottenere stime dei parametri più precise dall'interpolazione dei dati. Di seguito è riportato l'istogramma ottenuto.

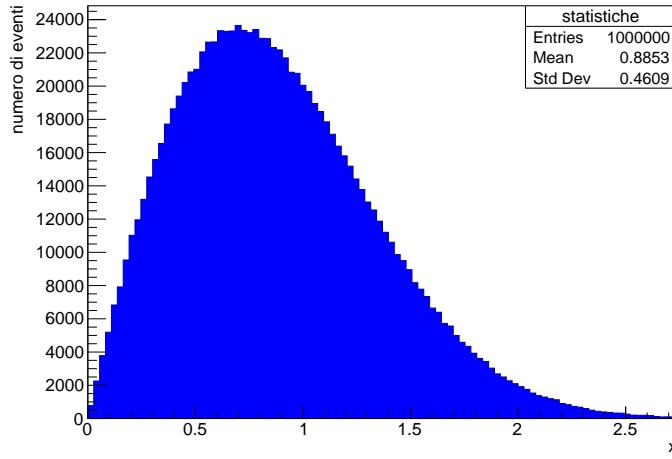


Figura 101: distribuzione  $f$  in  $(0, +\infty)$  per  $N = 1000000$

Plotando la mappa  $f$  su un calcolatore grafico è possibile notare che l'istogramma non normalizzato ottenuto mostra qualitativamente l'andamento atteso nell'intervallo  $(0, +\infty)$ . Al fine di verificare quantitativamente la corretta generazione si è deciso di svolgere un fit della distribuzione binnata interpolando i dati con la funzione  $f(x) = axe^{bx^2}$  ottenendo quanto segue.

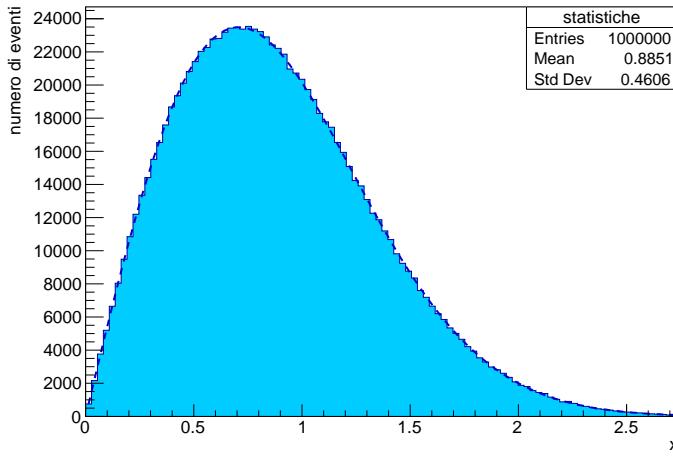


Figura 102: distribuzione  $f$  in  $(0, +\infty)$  per  $N = 1000000$ : fit

I parametri stimati dal programma di fit risultano

$$a = 5.48 \cdot 10^4 \quad \text{e} \quad b = -1$$

Come è possibile notare, a meno della normalizzazione dell'istogramma data dal parametro  $a$ , il parametro  $b$  risulta compatibile con il valore atteso. Anche in questo caso, la stima di  $b$  risulta esatta per via del fatto che si sono generate

sequenze pseudo-casuali in un intervallo illimitato, garantendo maggiore precisione all'algoritmo di minimizzazione. La forma analitica della distribuzione risulta, dunque, consistente con la forma attesa. La correttezza della costante di normalizzazione  $k$  può essere verificata visivamente notando che il metodo della funzione inversa implementato ha prodotto sequenze pseudo-casuali esattamente nel range  $(0, +\infty)$ , ossia il range all'interno del quale si è normalizzata la funzione  $f$ .

## Esercizio 12

Si vogliono generare numeri pseudo-casuali distribuiti secondo una distribuzione gaussiana con  $\mu = 0$  e  $\sigma = 1/\sqrt{2}$ , ossia secondo la densità di probabilità

$$f(x) := \frac{1}{\sqrt{\pi}} e^{-x^2} \quad \text{con } x \in (-\infty, +\infty)$$

utilizzando il metodo esatto e il metodo accept-reject, per poi verificare che la distribuzione generata sia la medesima.

Si noti che, all'interno dello spazio delle fasi  $\Omega = (-\infty, +\infty)$ , la distribuzione risulta già normalizzata, infatti

$$\int_{\Omega} f(x) dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-x^2} dx = \frac{1}{\sqrt{\pi}} \sqrt{\pi} = 1$$

dove l'integrale di  $e^{-x^2}$  su tutto l'asse reale è noto e può essere ottenuto con il teorema dei residui promuovendo  $f$  a funzione di variabile complessa, o per mezzo dell'analisi reale estendendo  $f$  in due dimensioni in coordinate polari.

### Box-Muller

L'unico metodo esatto di generazione di sequenze casuali incontrato fino a qui è il metodo della funzione inversa, definito nella proposizione 0.13. Nel caso della distribuzione gaussiana, tuttavia, non è possibile esprimere la primitiva per mezzo di funzioni elementari. Vale, infatti

$$\frac{1}{\sqrt{\pi}} \int e^{-x^2} dx = \frac{1}{2} \operatorname{erf}(x) + c \quad \forall c \in \mathbb{R}$$

dove la *funzione degli errori* è una funzione integrale data da

$$\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Vista l'impossibilità di integrazione per via analitica, il metodo della funzione inversa non può essere applicato direttamente. Tuttavia, si è già notato che, nonostante la gaussiana non sia integrabile analiticamente in una dimensione risulta, invece, integrabile passando in due dimensioni. Si consideri, dunque, la gaussiana bidimensionale definita come

$$g(x, y) := \frac{1}{\pi} e^{-(x^2+y^2)} \quad \text{con } (x, y) \in (-\infty, +\infty) \times (-\infty, +\infty)$$

Passando in coordinate polari, date dalla trasformazione

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix} \quad \text{con } r \in [0, +\infty) \text{ e } \theta \in [0, 2\pi)$$

e tenendo conto del determinante della matrice Jacobiana del cambio di coordinate, sarà banalmente vero che

$$\iint_{\mathbb{R}^2} g(x, y) dx dy = \int_0^{2\pi} \frac{1}{2\pi} d\theta \int_0^{+\infty} 2r e^{-r^2} dr = 1 \quad (56)$$

poiché entrambi gli integrali hanno valore unitario. In altre parole, la gaussiana bidimensionale definita come sopra risulta positiva e normalizzata, come ci si aspetta da ogni funzione densità di probabilità. Per ottenere la cumulativa corrispondente basterà risolvere l'integrale

$$G(r, \theta) = \frac{1}{2\pi} \int_0^\theta d\phi \int_0^r 2\rho e^{-\rho^2} d\rho = \frac{1}{2\pi} \theta (1 - e^{-r^2})$$

Si noti, dunque, che la cumulativa ottenuta può essere vista come il prodotto di due cumulative unidimensionali indipendenti, ossia

$$G(r, \theta) = H(\theta)W(r)$$

dove

$$H(\theta) := \frac{\theta}{2\pi} \quad \text{e} \quad W(r) := 1 - e^{-r^2}$$

Le funzioni densità di probabilità corrispondenti, per quanto visto, saranno

$$h(\theta) = \frac{1}{2\pi} \quad \text{e} \quad w(r) = 2re^{-r^2}$$

ben definite secondo Kolmogorov in quanto sempre positive nell'intervallo di definizione di  $\theta$  e  $r$  e normalizzate per la (56). Ma allora, per generare numeri secondo la gaussiana bidimensionale  $g$ , il metodo della funzione inversa ci assicura che basterà invertire singolarmente le due cumulative  $H$  e  $W$  che compongono la cumulativa bidimensionale  $G$ .

- Per la funzione  $H$  avremo

$$\alpha := \frac{\theta}{2\pi} \implies \theta = 2\pi\alpha$$

- Per la funzione  $W$  avremo

$$\beta := 1 - e^{-r^2} \implies r = \sqrt{-\log(1 - \beta)}$$

Si noti che si è scelto il solo ramo positivo in quanto  $r \geq 0$  per definizione delle coordinate introdotte. Si noti, inoltre, che  $\alpha$  e  $\beta$  sono due diverse variabili casuali con distribuzione uniforme nell'intervallo  $(0, 1)$  per proposizione 0.13. Ricordando ora le trasformazioni corrispondenti alle coordinate polari e sostituendo le relazioni ottenute, le coppie ordinate  $(x, y) \in \mathbb{R}^2$  tali che

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \sqrt{-\log(1 - \beta)} \cos(2\pi\alpha) \\ \sqrt{-\log(1 - \beta)} \sin(2\pi\alpha) \end{pmatrix} \quad \text{con} \quad \alpha, \beta \in U(0, 1) \quad (57)$$

saranno distribuite secondo la gaussiana bidimensionale  $g$  per il metodo della funzione inversa. Vista l'indipendenza statistica di  $x$  e  $y$  e nota la simmetria della gaussiana bidimensionale, per ottenere una gaussiana monodimensionale con questo metodo basterà generare numeri random secondo una delle due componenti della (57). Il procedimento descritto per generare pacchetti gaussiani in modo esatto va sotto il nome di *trasformazione di Box-Muller* e consiste, di fatto, come si ha avuto modo di mostrare, in una generalizzazione multidimensionale del metodo della funzione inversa.

Al fine di verificare la corretta generazione di punti casuali distribuiti come una gaussiana bidimensionale con il metodo di Box-Muller, si sono generati  $N = 500000$  punti  $(x_i, y_i)$  calcolati secondo la relazione (57). L'istogramma bidimensionale che segue mostra quanto ottenuto.

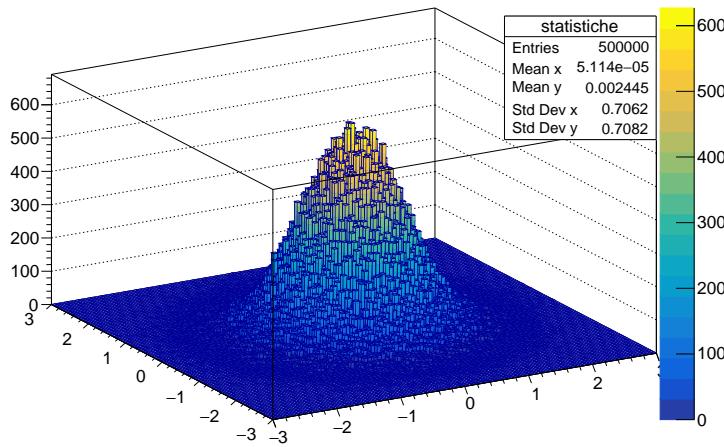


Figura 103: gaussiana bidimensionale con funzione inversa per  $N = 500000$

Come è possibile notare, l'istogramma ottenuto presenta caratteristiche qualitative consistenti con la distribuzione normale bidimensionale in esame: appare simmetrica, piccata rispetto allo zero e con medie nulle rispetto ai due assi coordinati. Si è dunque generata la sequenza di numeri pseudo-casuali  $\{x_i\}_{i=1,\dots,N}$  con  $N = 1000000$  secondo la relazione

$$x = \sqrt{-\log(1 - \beta)} \cos(2\pi\alpha)$$

con  $\alpha$  e  $\beta$  uniformi in  $(0, 1)$ , ottenendo il seguente istogramma non normalizzato.

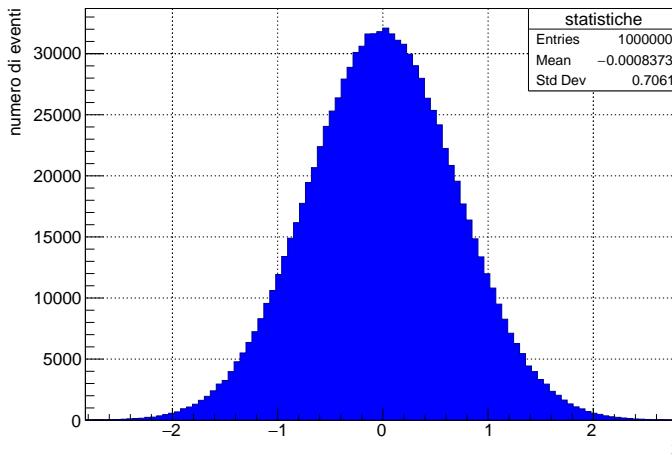


Figura 104: restrizione lungo  $x$  con funzione inversa per  $N = 1000000$

L'istogramma ottenuto mostra che i numeri casuali generati secondo la prima componente della relazione (57) risultano, qualitativamente, distribuiti secondo una gaussiana. Inoltre, i valori ottenuti nel box statistiche permettono, fin da qui, di verificare la corretta generazione. Ad ogni modo, al fine di operare una verifica quantitativa più completa, si è deciso di svolgere un fit della distribuzione binnata interpolando i dati secondo una gaussiana con media  $\mu$ , deviazione standard  $\sigma$  e costante di normalizzazione  $k$  parametri liberi. Di seguito sono riportati i risultati ottenuti.

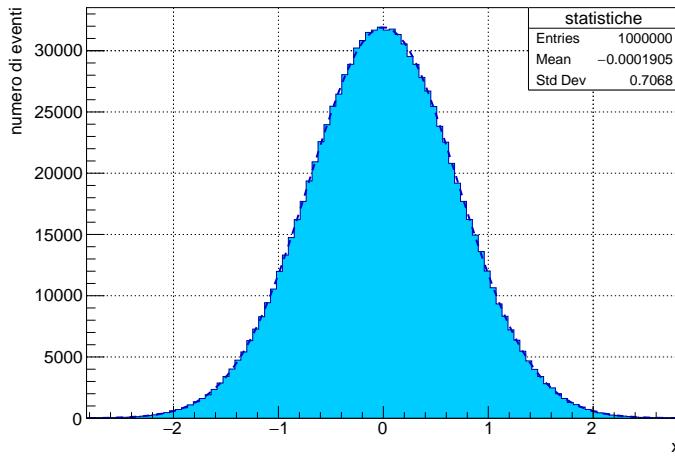


Figura 105: restrizione lungo  $x$  con funzione inversa per  $N = 1000000$ : fit

Si è ottenuta la seguente stima dei parametri.

$$k = 3.19 \cdot 10^4 \quad \text{e} \quad \mu = -0.0002 \quad \text{e} \quad \sigma = 0.707$$

Notiamo che valori attesi dei momenti e dei momenti centrali notevoli della gaussiana monodimensionale in esame sono

$$\mu_t = 0 \quad \text{e} \quad \sigma_t = \frac{1}{\sqrt{2}} \approx 0.707$$

Dall'evidente compatibilità dei valori stimati con quelli attesi segue la verifica della corretta generazione di sequenze distribuite secondo la gaussiana  $f$  grazie al metodo di Box-Muller con restrizione lungo l'asse coordinato delle  $x$ . Si è poi deciso di verificare la corretta generazione di dati secondo una distribuzione normale con restrizione della gaussiana bidimensionale lungo l'asse coordinato delle  $y$ . Ci si aspetta, per ragioni di simmetria, di ottenere una distribuzione equivalente a quella appena ottenuta. Inoltre, ci aspettiamo di ottenere una stima dei momenti compatibile con quella presente nel box statistiche di figura 103. Si è dunque generata la sequenza di numeri pseudo-casuali  $\{y_i\}_{i=1,\dots,N}$  con  $N = 1000000$  secondo la relazione

$$y = \sqrt{-\log(1-\beta)} \sin(2\pi\alpha)$$

con  $\alpha$  e  $\beta$  aventi distribuzione uniforme in  $(0, 1)$ , ottenendo il seguente istogramma non normalizzato.

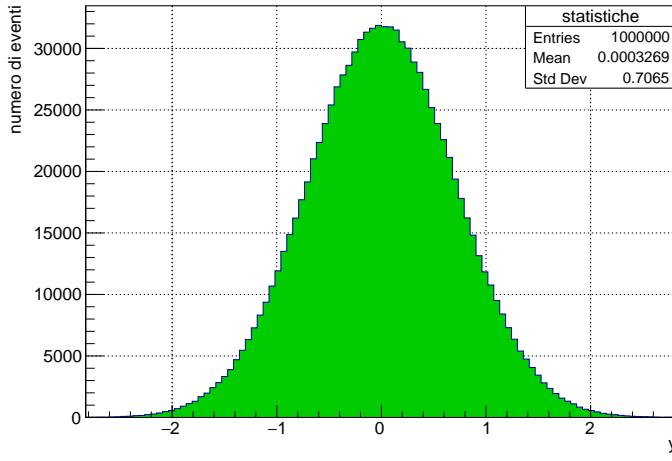


Figura 106: restrizione lungo  $y$  con funzione inversa per  $N = 1000000$

L’istogramma ottenuto mostra che i punti risultano, qualitativamente, distribuiti secondo una gaussiana. Al fine di svolgere una verifica quantitativa quanto più completa possibile si è svolto, anche in questo caso, un fit della distribuzione binnata interpolando i dati secondo una generica funzione gaussiana con media  $\mu$ , deviazione standard  $\sigma$  e costante di normalizzazione  $k$  parametri liberi, ottenendo quanto segue.

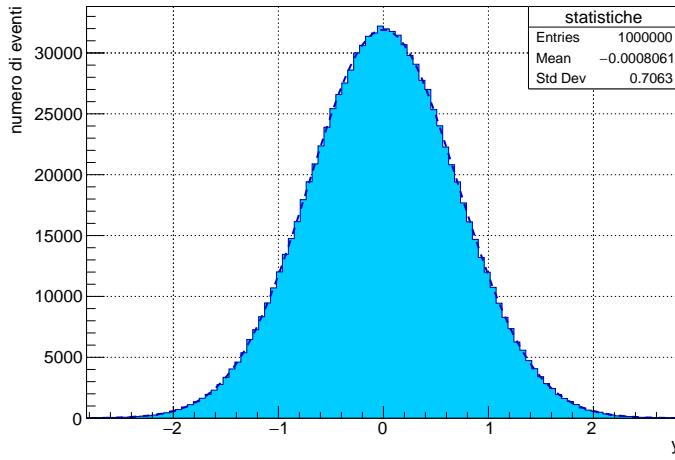


Figura 107: restrizione lungo  $y$  con funzione inversa per  $N = 1000000$ : fit

Si è ottenuta la seguente stima dei parametri.

$$k = 3.19 \cdot 10^4 \quad \text{e} \quad \mu = -0.0008 \quad \text{e} \quad \sigma = 0.706$$

I parametri stimati risultano chiaramente compatibili con i momenti attesi della distribuzione  $f$ . Per transitività, risultano compatibili anche con i parametri

stimati dall'andamento dato dalla restrizione lungo  $x$  della distribuzione bidimensionale. Inoltre, si nota facilmente che le medie e le deviazioni stimate singolarmente dai fit delle due restrizioni lungo  $x$  e  $y$  risultano compatibili con i valori dei momenti della distribuzione bidimensionale riportati in figura 103. Per tale ragione, risulta possibile concludere, più in generale, anche la compatibilità tra le due distribuzioni monodimensionali e la distribuzione bidimensionale ottenuta. Siamo quindi riusciti a mostrare computazionalmente che la gaussiana bidimensionale  $g$  definita in precedenza risulta essere, a tutti gli effetti, un'estensione consistente della gaussiana monodimensionale  $f$  di partenza, ossia una distribuzione che conserva i momenti e i momenti centrali notevoli della gaussiana monodimensionale in esame.

### Accept-reject

Il metodo di generazione di numeri pseudo-casuali accept-reject consiste nel generare un set di numeri random  $\{x_i\}_{i=1,\dots,M}$  distribuiti secondo una densità di probabilità  $g$ , detta *proposal density*, tale che

$$g(x) \geq f(x) \quad \forall x \in \Omega$$

Vengono poi generati, con distribuzione uniforme in  $(0, g(x))$ , le corrispondenti ordinate pseudo-casuali  $\{y_i\}_{i=1,\dots,M}$ . Se è verificata la condizione

$$y_i < f(x_i) \tag{58}$$

l'estrazione viene accettata, altrimenti viene reiterato il procedimento per  $x_{i+1}$ . La sequenza di valori  $\{x_i\}_{i=1,\dots,N < M}$  che sopravvive al controllo (58) sarà distribuita secondo la funzione densità di probabilità  $f$ . Il metodo accept-reject risulta, chiaramente, meno efficiente rispetto al metodo della funzione inversa o rispetto a varianti affini in più dimensioni come Box-Muller, in quanto consiste nello scartare punti che non rispettano una data condizione. Tuttavia, può risultare un'alternativa utile nel caso in cui la densità  $f$  non sia sufficientemente semplice da rispettare le ipotesi di integrabilità e invertibilità del metodo della funzione inversa. Infatti, il metodo accept-reject può essere sempre utilizzato per generare sequenze casuali ponendo, ad esempio,  $g = U(\Omega)$ . La generazione di sequenze distribuite uniformemente è, infatti, banale e può essere effettuata con diversi metodi elementari. D'altra parte, l'efficienza del metodo dipende dal rapporto tra numero di punti che viene accettato e il numero di punti che viene scartato. Al fine di migliorare l'efficienza risulta, dunque, necessario individuare una proposal density  $g$  che sia quanto più possibile vicina alla densità  $f$  con cui vogliamo generare i punti. In tal modo, infatti, l'area in cui i punti possono essere generati diminuisce e, di conseguenza, diminuisce la probabilità di generare una coppia  $(x_i, y_i)$  che non verifichi la (58). Come al solito, dunque, più informazioni si hanno circa il risultato che si vuole ottenere, più sarà possibile aumentare l'efficienza di un metodo. In questo caso, la conoscenza dell'andamento di  $f$  permetterà di costruire una dominante  $g$  ottimale.

Al fine di generare numeri distribuiti come la gaussiana  $f$ , si è considerata la proposal density

$$g(x) := \begin{cases} A & \text{se } x \in (0, 1] \\ Axe^{1-x^2} & \text{se } x \in (1, +\infty) \end{cases} \quad \text{con} \quad A := \frac{1}{\sqrt{\pi}} \tag{59}$$

Analiticamente, o anche visivamente plotando le due funzioni, vale la condizione

$$g(x) \geq f(x) \quad \forall x \in (0, +\infty)$$

di applicabilità del metodo accept-reject. Anzitutto, è possibile notare che

$$\begin{aligned} \int_0^{+\infty} g(x) dx &= A \int_0^1 dx + A \int_1^{+\infty} xe^{1-x^2} dx = \\ &= A + \frac{A}{2} = \frac{3}{2\sqrt{\pi}} \neq 1 \end{aligned}$$

ossia la proposal density non è normalizzata all'interno dell'intervallo  $(0, +\infty)$ . Al fine di rendere  $g$  una densità di probabilità ben definita si è quindi determinato  $k \in \mathbb{R}$  tale che

$$\begin{aligned} k \int_0^{+\infty} \frac{g(x)}{A} dx = 1 &\iff k \int_0^1 dx + k \int_1^{+\infty} xe^{1-x^2} dx = 1 \\ &\iff k + \frac{k}{2} = 1 \iff k = \frac{2}{3} \end{aligned}$$

dove il secondo integrale è stato risolto ponendo il cambio di variabile  $u = x^2$ . Segue che la funzione densità di probabilità associata alla funzione maggiorante  $g$  si scriverà come

$$g_{\text{pdf}}(x) = \begin{cases} k & \text{se } x \in (0, 1] \\ kxe^{1-x^2} & \text{se } x \in (1, +\infty) \end{cases} \quad \text{con } k = \frac{2}{3} \quad (60)$$

Al fine di applicare il metodo accept-reject risulta quindi necessario generare numeri pseudo-casuali distribuiti secondo  $g_{\text{pdf}}(x)$ . A tale scopo, vista la semplicità analitica della distribuzione, si è deciso di utilizzare il metodo della funzione inversa. Si è quindi calcolata la funzione cumulativa della densità  $g_{\text{pdf}}$  come

$$\begin{aligned} G_{\text{pdf}}(x) &= \begin{cases} \int_0^x k dt & \text{se } x \in (0, 1] \\ \int_1^x kte^{1-t^2} dt + \int_0^1 k dt & \text{se } x \in (1, +\infty) \end{cases} = \\ &= \begin{cases} kx & \text{se } x \in (0, 1] \\ \frac{k}{2} \left(1 - e^{1-x^2}\right) + k & \text{se } x \in (1, +\infty) \end{cases} \end{aligned}$$

Dove il termine  $\int_0^1 k dt$  nel secondo tratto della funzione è stato aggiunto per ragioni di continuità della cumulativa nel punto  $x = 1$ . Risulta banale verificare che la mappa trovata presenta le caratteristiche attese da una funzione di ripartizione, infatti

$$\lim_{x \rightarrow 0^+} G_{\text{pdf}}(x) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} G_{\text{pdf}}(x) = \frac{3}{2}k = 1$$

ossia, nell'ordine: nessun evento accade prima di  $x = 0$  e la probabilità che un evento sia accaduto in  $(0, +\infty)$  è unitaria, come ci si aspetta. Inoltre,  $G_{\text{pdf}}(x)$  è una funzione monotona crescente del suo argomento. Si è quindi invertita la cumulativa facendo attenzione a determinare le due nuove condizioni della funzione definita a tratti. Svolgendo i conti si ha

$$G_{\text{pdf}}^{-1}(x) = \begin{cases} \frac{x}{k} & \text{se } x \in (0, k] \\ \sqrt{1 - \log(3 - \frac{2x}{k})} & \text{se } x \in (k, +\infty) \end{cases} \quad \text{con } k = \frac{2}{3}$$

Al fine di verificare qualitativamente la corretta generazione di sequenze pseudo-casuali secondo la proposal density (60), si è generato il set  $\{x_i\}_{i=1,\dots,N}$  con  $N = 1000000$  utilizzando il metodo della funzione inversa espresso dalla 0.13, ottenendo il seguente istogramma non normalizzato.

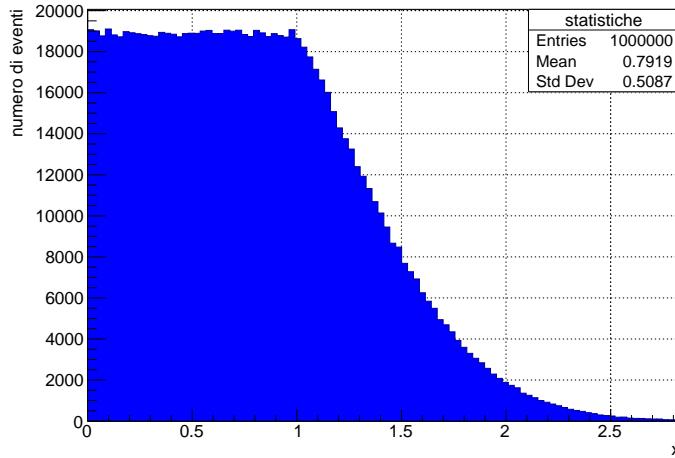


Figura 108: distribuzione  $g$  con funzione inversa per  $N = 1000000$

Plotando la mappa  $g_{\text{pdf}}(x)$  su un calcolatore grafico risulta evidente che, a meno della normalizzazione, la distribuzione generata presenta caratteristiche qualitative consistenti con quelle attese. Appurata la corretta generazione della proposal density, si è quindi implementato il metodo accept-reject, generando sequenze secondo  $g_{\text{pdf}}(x)$  e valutando la condizione (58) utilizzando la funzione di partenza  $g(x)$ : si noti, infatti,  $g_{\text{pdf}}$  non è una maggiorante ottimale di  $f$ . In particolare, si sono generati  $\{x_i\}_{i=1,\dots,N}$  con  $N = 500000$  valori con il metodo accept-reject, ottenendo il seguente istogramma non normalizzato.

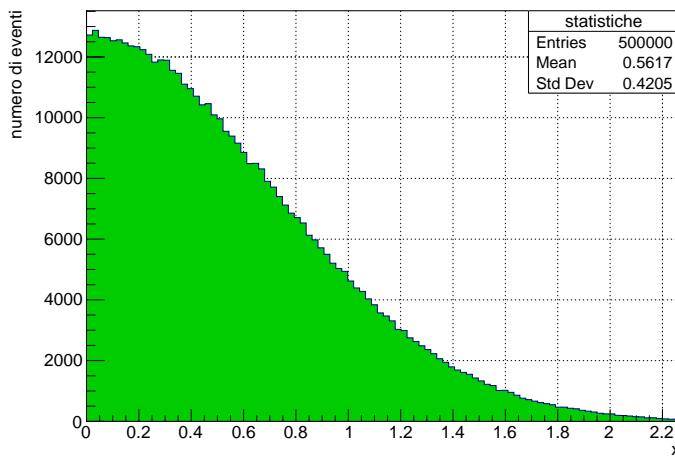


Figura 109: distribuzione  $f$  con accept-reject per  $N = 500000$ :  $x > 0$

Come è possibile notare, l'istogramma dei dati appare distribuito come una semi-gaussiana per  $x > 0$ . Quanto ottenuto è chiaramente consistente con il metodo implementato e con le funzioni  $g$  e  $G_{\text{pdf}}^{-1}$  utilizzate, che risultano definite solo per valori di  $x$  positivi. Al fine di ottenere la seconda metà dei dati, per generare una gaussiana  $f$  completa, risulta possibile procedere in due modi diversi:

- ridefinire la funzione  $g$  per  $x < 0$  facendo una simmetria rispetto all'asse  $y$  e procedere con passaggi analoghi
- generare un altro set di dati con le stesse funzioni, per poi moltiplicare ogni valore ottenuto per  $-1$

Per semplicità, si è scelta la seconda strada, generando una nuova sequenza  $\{x_i\}_{i=1,\dots,N}$  con  $N = 500000$  con il metodo accept-reject. Si è poi determinato il nuovo set  $\{y_i\}_{i=1,\dots,N}$  come

$$y_i = -x_i \quad \forall i = 1, \dots, N$$

Per ragioni di simmetria di una gaussiana rispetto all'origine ci si aspetta che, questo secondo set di valori, sia banalmente distribuito come una semi-gaussiana per  $x < 0$ . Si sono quindi rappresentati i dati ottenuti in un istogramma non normalizzato, ottenendo quanto segue.

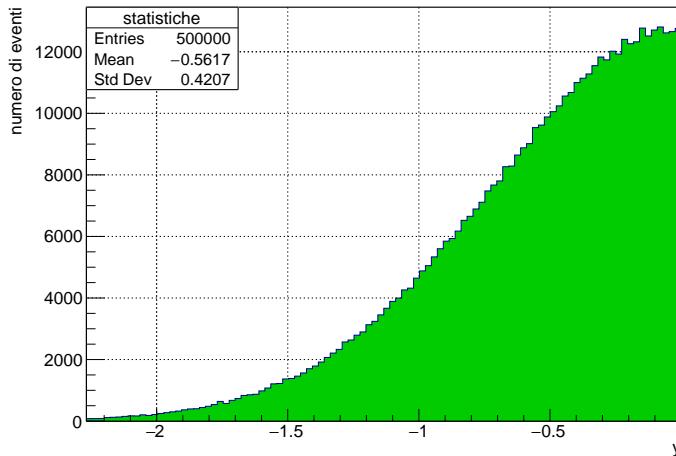


Figura 110: distribuzione  $f$  con accept-reject per  $N = 500000$ :  $x < 0$

Come è possibile notare, i dati risultano distribuiti secondo una semi-gaussiana per  $x < 0$ , come ci si aspetta. Inoltre, dai momenti delle due distribuzioni visibili nelle figure, si ha una verifica quantitativa della simmetria rispetto allo zero dei due set di dati. Al fine di generare una gaussiana completa della forma di  $f$  così come definita inizialmente con il metodo accept-reject basterà, dunque, visualizzare in un istogramma il set di valori

$$\{x_i\} \cup \{y_i\} \quad \forall i = 1, \dots, N$$

Nel nostro caso, abbiamo scelto un numero di punti pari a  $N = 500000$ . Di seguito è riportato l'istogramma ottenuto.

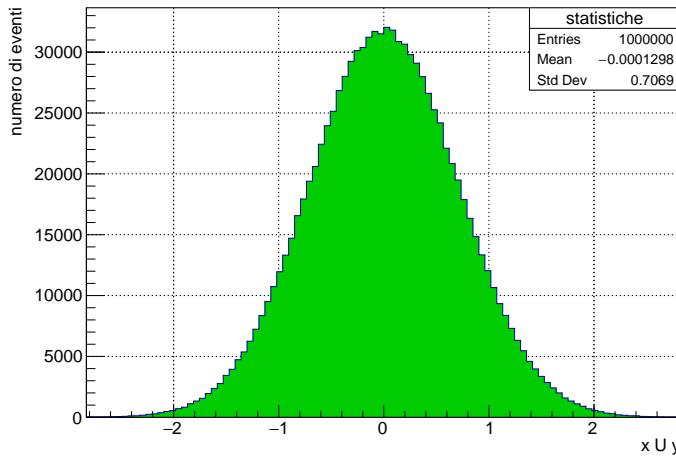


Figura 111: distribuzione  $f$  con accept-reject per  $N = 1000000$

L'istogramma risulta, qualitativamente, gaussiano e centrato nell'origine come ci si aspetta. Al fine di svolgere una verifica quantitativa completa, si è svolto un fit della distribuzione binnata interpolando i dati secondo la distribuzione gaussiana con media  $\mu$ , deviazione standard  $\sigma$  e costante di normalizzazione  $k$  parametri liberi, ottenendo quanto segue.

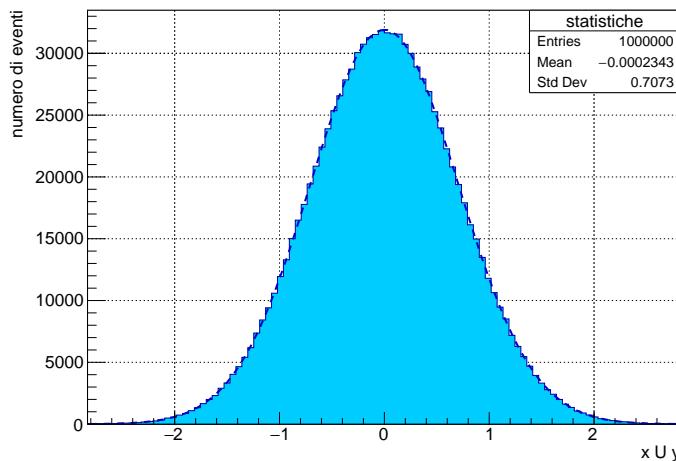


Figura 112: distribuzione  $f$  con accept-reject per  $N = 1000000$ : fit

Si è ottenuta la seguente stima dei parametri.

$$k = 3.19 \cdot 10^4 \quad \text{e} \quad \mu = -0.0002 \quad \text{e} \quad \sigma = 0.707$$

Dalla stima dei parametri ottenuta dal fit, è immediato notare che i dati generati utilizzando il metodo accept-reject sono distribuiti secondo una gaussiana della forma di  $f$ , come desiderato. Inoltre, dal parametro  $k$  di normalizzazione è evidente che la distribuzione ottenuta con questo metodo sia la medesima della

distribuzione ottenuta con il metodo della funzione inversa. A parità di numero di eventi  $N$  generati, infatti, la costante  $k$  assume, in entrambi i casi, un valore compatibile, se non uguale.

In definitiva, abbiamo scoperto che è sempre possibile generare numeri secondo una distribuzione  $f$  a piacere, anche se non sono verificate le ipotesi del metodo della funzione inversa. Il metodo accept-reject, infatti, al costo di una perdita di efficienza data dai punti scartati che non rispettano la condizione (58), risulta efficace a tale scopo. Evidentemente, questo metodo è l'unico percorribile in tutti quei casi in cui  $f$  risulti sufficientemente complicata da non permettere l'integrazione per via analitica, anche se abbiamo scoperto che, nel caso gaussiano, il metodo della funzione inversa può essere esteso notando che una gaussiana bidimensionale ammette primitiva elementare. Ovviamente, la trasformazione di Box-Muller sarà sempre più efficiente del metodo accept-reject, qualunque sia la proposal density. Questo è dovuto al fatto che, nei metodi esatti, ogni punto generato viene utilizzato.

### Esercizio 13

Si vogliono stimare numericamente gli integrali

$$I_c := \int_0^{\pi/2} x \cos(x) dx \quad \text{e} \quad I_s := \int_0^\pi x \sin(x) dx$$

utilizzando il metodo dell'importance sampling. In particolare, si vuole studiare la deviazione al variare del numero  $N$  di punti, confrontandola con la deviazione prodotta dal metodo Monte Carlo.

L'importance sampling (campionamento ad importanza) è una variante del metodo di integrazione numerica Monte Carlo che permette, a parità di numero  $N$  di punti casuali generati, di diminuire l'errore sul calcolo. Si consideri, infatti, il problema di calcolo dell'integrale definito

$$I := \int_a^b f(x) dx$$

tale che l'integrandra  $f$  ammetta un massimo in  $\tilde{x} \in (a, b)$ . Si ha già avuto modo di mostrare che, per il metodo Monte Carlo, l'errore sul calcolo di  $I$  è dato da

$$\text{err}(I) = V^{(n)} \text{err}\langle f \rangle \propto \frac{\sigma_f}{\sqrt{N}}$$

ossia  $\text{err}(I)$  presenta una dipendenza diretta dalla deviazione standard di  $f$  e dal numero di punti  $N$  con cui si decide di eseguire la stima. Al fine di ridurre l'errore sulla stima, dunque, è possibile agire aumentando  $N$ , ma anche riducendo  $\sigma_f$ : l'importance sampling ha proprio questo secondo obiettivo. Se la funzione  $f$  ammette un massimo in  $\tilde{x}$ , infatti, il contributo maggiore dell'integrale sarà dato, numericamente, dalle somme parziali nell'intorno  $B(\tilde{x}, \epsilon) \subset (a, b)$  per qualche  $\epsilon > 0$ . Ma allora, generare numeri distribuiti uniformemente in  $(a, b)$  risulta, in questo caso, sconveniente. Risulta, invece, conveniente generare  $\{x_i\}_{i=1,\dots,N}$  secondo una funzione densità di probabilità  $g$  tale che

$$\exists \bar{x} \in (a, b) \quad \text{t.c.} \quad \max_{x \in (a,b)} g(x) = g(\bar{x}) \quad \text{e} \quad \bar{x} \approx \tilde{x} \quad (61)$$

ossia tale che  $g$  ammetta un massimo nell'intervallo di integrazione vicino al massimo della funzione integranda  $f$ . Questo metodo consiste, dunque, nel generare numeri pseudo-casuali non più distribuiti uniformemente in  $(a, b)$ , ma distribuiti secondo un densità di probabilità  $g$  avente un picco in corrispondenza del picco di  $f$ , al fine di dare maggiore peso ai contributi più rilevanti di area sottesa alla curva. Analiticamente, dovremo porre il cambio di variabile

$$x = G^{-1}(z) \quad \text{con} \quad z \in U(0, 1)$$

dove  $G^{-1}$  è l'inversa della cumulativa della densità di probabilità  $g$ . In tal modo, infatti, per il metodo della funzione inversa dato dalla proposizione 0.13, avremo che  $x$  sarà una variabile casuale distribuita secondo  $g$ . Invertendo il cambio di variabile vale banalmente la relazione

$$z = G(x)$$

Si ha quindi un nuovo differenziale della forma

$$dz = \frac{d}{dx} G(x) dx = \left( \frac{d}{dx} \int_a^x g(t) dt \right) dx = g(x) dx$$

per teorema fondamentale del calcolo integrale. Si noti, quindi, che per gli estremi di integrazione vale la mappatura

$$x = a \implies z = 0 \quad \text{e} \quad x = b \implies z = 1$$

poiché  $G$  è la cumulativa di  $g$ , e per costruzione assume valore nullo all'estremo sinistro e valore unitario in quello destro. Sostituendo quanto ricavato in  $I$  avremo un'espressione della forma

$$I = \int_0^1 \frac{f(G^{-1}(z))}{g(G^{-1}(z))} dz \approx \frac{1}{N} \sum_{i=1}^N \frac{f(G^{-1}(z_i))}{g(G^{-1}(z_i))} \quad (62)$$

dove l'approssimazione numerica scritta vale banalmente per la (48), in questo caso con  $b - a = 1$ . Per la costruzione fatta,  $\{z_i\}_{i=1,\dots,N}$  è una sequenza di  $N$  numeri pseudo-casuali distribuiti uniformemente nell'intervallo  $(0, 1)$ .

Studiamo quindi i due integrali in esame, confrontando l'usuale metodo Monte Carlo con la sua variante di campionamento ad importanza.

### Studio di $I_c$

Anzitutto, integrando per parti, si ha che il valore vero dell'integrale è dato da

$$I_c = \frac{1}{2}(\pi - 2)$$

Per la stima numerica dell'integrale  $I_c$  si è utilizzata la densità di probabilità

$$g(x) = k \cos(x)$$

Si noti che i punti stazionari di  $f$  in  $(0, \pi/2)$  sono dati da  $x$  tali che

$$\frac{d}{dx} f(x) = 0 \iff \cos(x) - x \sin(x) = 0$$

Numericamente, si ricava che lo zero della funzione derivata nell'intervallo in esame è  $\tilde{x} \approx 0.86$ , che coincide con un punto di massimo. Si calcola facilmente, infatti, che la derivata seconda di  $f$  nel punto è negativa. La densità  $g$ , invece, nell'intervallo in esame, ammette massimo per  $\bar{x} = 0$  per ogni valore di  $k > 0$ . La misura di Lebesgue dell'intervallo di integrazione in  $\mathbb{R}$  è  $\mathcal{L}(a, b) = \frac{\pi}{2} \approx 1.57$ . Ma allora, in questo caso, rispetto alla misura dell'intervallo, il massimo di  $g$  si discosta significativamente dal massimo di  $f$ . In altre parole, si ha

$$\tilde{x} \not\simeq \bar{x}$$

Ci si aspetta, dunque, che l'importance sampling utilizzando la densità  $g$  possa non risultare più efficiente rispetto all'utilizzo del semplice metodo Monte Carlo, in quanto la condizione (61) non risulta verificata. Affinché la densità  $g$  possa

dirsi ben definita, si è normalizzata la funzione nell'intervallo, ricavando  $k \in \mathbb{R}$  tale che

$$k \int_0^{\pi/2} \cos(x) dx = 1 \iff k = 1$$

Segue che la funzione densità di probabilità in esame si scriverà come

$$g(x) = \cos(x) \quad \text{con} \quad x \in \left(0, \frac{\pi}{2}\right)$$

evidentemente positiva, come ci si aspetta. La sua cumulativa avrà la forma

$$G(x) = \int_0^x \cos(t) dt = \sin(x)$$

Ricavando  $x$  in funzione di  $G$  si ottiene la funzione inversa della cumulativa

$$G^{-1}(x) = \arcsin(x)$$

Si è quindi applicato il cambio di variabile dato dal metodo dell'importance sampling dato dalla relazione (62), ottenendo

$$I_c = \int_0^1 \frac{\arcsin(x) \cos(\arcsin(x))}{\cos(\arcsin(x))} dx = \int_0^1 \arcsin(x) dx$$

Si è poi applicato il metodo numerico espresso dalla seconda parte della (62). In particolare, si è scelto il range di punti

$$1000 \leq N < 50000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

Si sono quindi plottati i logaritmi delle dispersioni in funzione dei logaritmi del numero  $N$  di punti. In tal modo, come sappiamo, la dispersione si scriverà come una retta della forma (51). Si sono dunque fissati i dati con una generica funzione affine del tipo  $y = p + mx$ , ottenendo quanto segue.

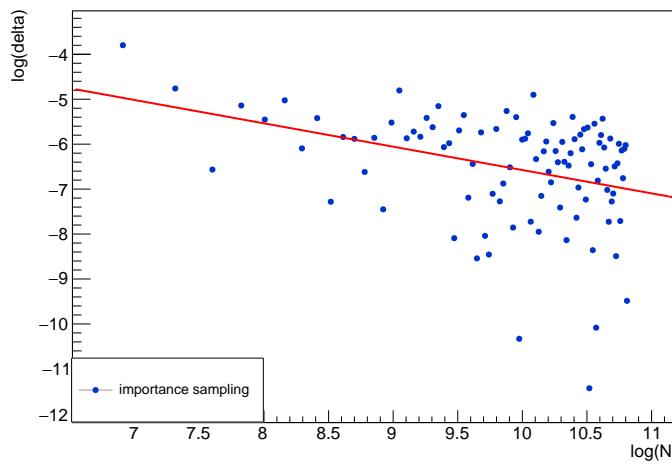


Figura 113:  $\log \Delta_c$  importance sampling: fit

Si è ottenuta la stima dei parametri che segue.

$$p = -1.38 = \log k \quad \text{e} \quad m = -0.519 \approx -0.5$$

Come è possibile notare, il coefficiente angolare stimato risulta compatibile con il valore atteso dato dalla relazione (51). Tuttavia, coerentemente con quanto si ha già avuto modo di discutere negli esercizi precedenti, i punti risultano sparsi in un range considerevole dell'asse delle ordinate. Sappiamo che risulta possibile rettificare l'andamento, consentendo una stima migliore del coefficiente angolare, calcolando  $M$  volte lo stesso integrale ad  $N$  fissato, per poi determinare

$$\langle \Delta_c \rangle_M \quad \forall N = N_{min}, \dots, N_{max}$$

Al posto di associare ad ogni  $N$  la dispersione dal valor vero di una singola stima, si è quindi associato, ad ogni  $N$ , il valor medio di  $M$  diverse dispersioni dell'integrale. Per le proprietà dello stimatore di media campionaria, infatti, avremo, ad ogni  $N$ , una stima più precisa del valore di  $I_c$ . Si è ripetuta la medesima interpolazione dei dati considerando un valore di  $M = 200$  e mantenendo lo stesso range di  $N$ , ottenendo quanto segue.

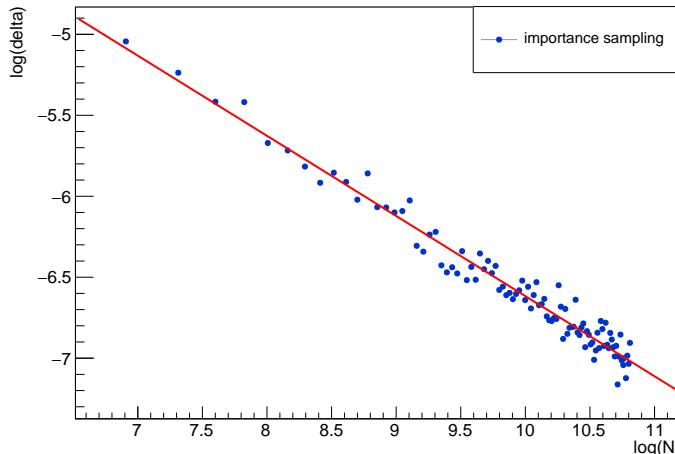


Figura 114:  $\log \Delta_c$  importance sampling per  $M = 200$ : fit

Si è ricavata la più precisa stima dei parametri

$$p = -1.67 = \log k \quad \text{e} \quad m = -0.495 \approx -0.5$$

Come ipotizzato, la procedura di rettificazione per mezzo del calcolo delle medie ha consentito di ottenere una stima di  $m$  più compatibile con il valore atteso. Al fine di procedere per un'analisi più profonda, si è deciso di confrontare il metodo dell'importance sampling con il classico metodo di integrazione Monte Carlo dato dalla (48) plottando, nello stesso grafico, i due logaritmi delle dispersioni in funzione dei logaritmi del numero di punti. In particolare, l'operazione è stata svolta per  $M = 1$  e nello stesso range di  $N$  definito precedentemente. Di seguito sono mostrati i risultati ottenuti.

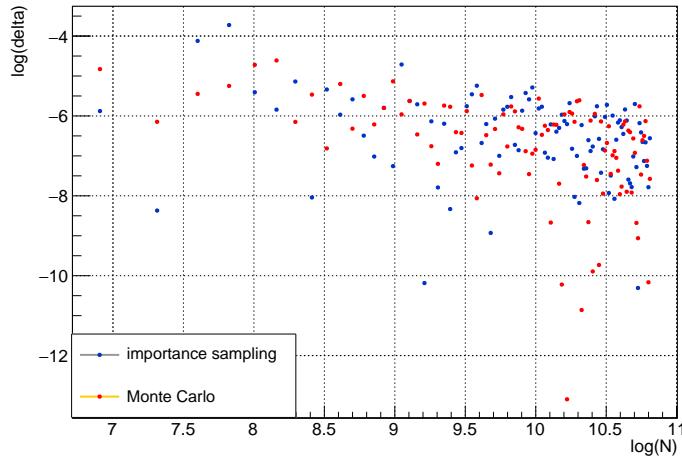


Figura 115: confronto  $\log \Delta_c$  per  $M = 1$

Come si nota dal plot dei punti, il calcolo di  $I_c$  con l'importance sampling si discosta poco in precisione dal calcolo utilizzando il classico metodo Monte Carlo: per  $M = 1$  i due andamenti risultano, qualitativamente, sovrapponibili. Nonostante l'esplosione dei tempi di calcolo, si è quindi deciso di svolgere la medesima operazione per  $M = 200$ , ottenendo i seguenti andamenti.

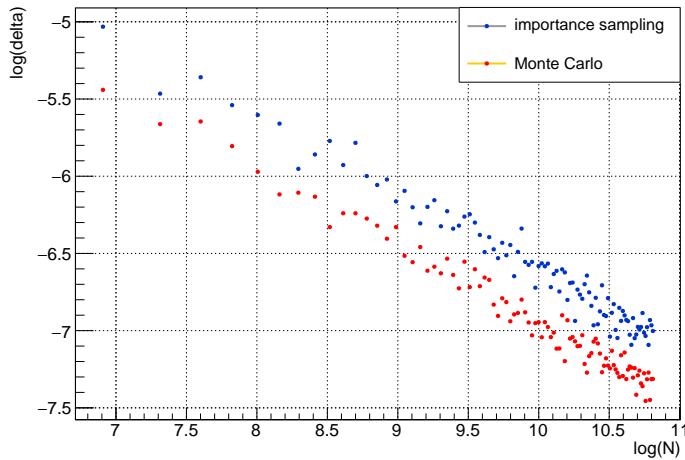


Figura 116: confronto  $\log \Delta_c$  per  $M = 200$

Dalla rettificazione dei dati risulta ora molto più evidente come, in questo caso, il metodo dell'importance sampling risulti meno efficiente del metodo Monte Carlo: la dispersione nel caso Monte Carlo, infatti, risulta, a parità di numero di punti  $N$  generati, sempre inferiore rispetto alla dispersione dell'importance sampling. Quantitativamente è possibile verificare questo fatto interpolando i dati ottenuti con il metodo Monte Carlo con la mappa affine  $y = q + wx$ .

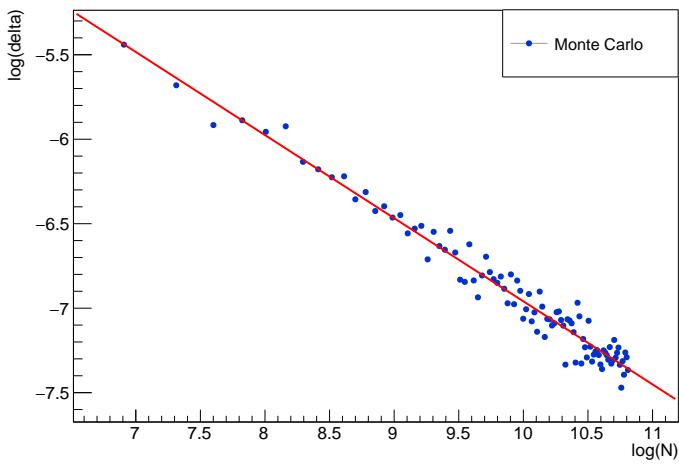


Figura 117:  $\log \Delta_c$  Monte Carlo per  $M = 200$ : fit

I parametri stimati risultano essere

$$q = -2.04 = \log k \quad \text{e} \quad w = -0.492 \approx -0.5$$

Confrontando la retta appena stimata con quella già ricavata dal fit per  $M = 200$  con il metodo dell'importance sampling, si ha che

$$q < p \quad \text{e} \quad m \approx w$$

da cui segue la verifica quantitativa della maggior efficienza del metodo Monte Carlo classico se si utilizza la densità  $g$  per il campionamento ad importanza. D'altra parte, è facile verificare che nell'intervallo di integrazione  $(0, \pi/2)$  la densità  $g$  e l'integrandà  $f$  assumono il seguente andamento.

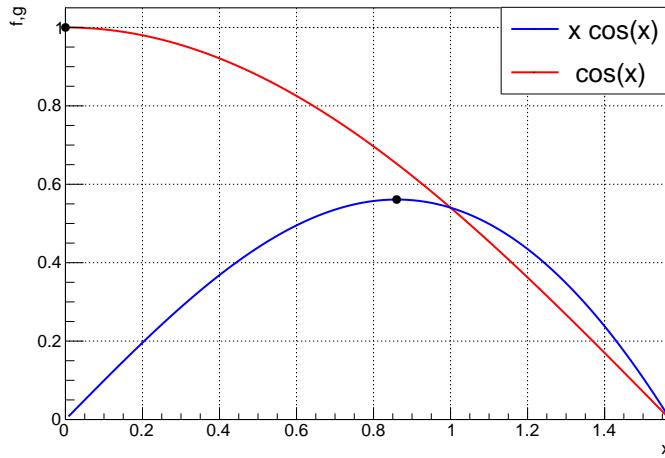


Figura 118: grafico delle funzioni  $f$  e  $g$  con rispettivi punti di massimo

Anche graficamente è possibile notare quanto si è già ottenuto per via analitica: i due punti di massimo si discostano significativamente l'uno dall'altro.

La dispersione maggiore ottenuta con il campionamento ad importanza è allora consistente con le considerazioni fatte inizialmente circa la discordanza tra il punto di massimo di  $f$  e il punto di massimo della densità  $g$ .

Ciò che accade, non venendo rispettata l'ipotesi di efficienza (61) è che, con il metodo dell'importance sampling, usando  $g$  come densità, vengono generati, a parità di  $N$ , più numeri pseudo-casuali in corrispondenza di un intorno destro dello zero rispetto al metodo Monte Carlo. Vengono quindi sommate aree di rettangoli in una zona dove il contributo di area rispetto all'area totale data da  $I_c$  è meno rilevante, inducendo un errore maggiore a parità di punti, coerentemente con la costruzione del metodo esposta inizialmente.

### Studio di $I_s$

Anzitutto, integrando per parti, si ha che il valore vero dell'integrale è dato da

$$I_s = \pi$$

Per la stima numerica dell'integrale  $I_s$  si è utilizzata la densità di probabilità

$$g(x) = k \sin(x)$$

Si noti che i punti stazionari di  $f$  in  $(0, \pi)$  sono dati da  $x$  tali che

$$\frac{d}{dx} f(x) = 0 \iff \sin(x) + x \cos(x) = 0$$

Numericamente, si ricava che lo zero della funzione derivata nell'intervallo in esame risulta  $\tilde{x} \approx 2.029$ , che coincide con un punto di massimo. Si calcola facilmente, infatti, che la derivata seconda di  $f$  nel punto è negativa. La densità  $g$ , invece, nell'intervallo in esame ammette massimo per  $\bar{x} = \frac{\pi}{2} \approx 1.57$  per ogni valore di  $k > 0$ . La misura di Lebesgue dell'intervallo di integrazione in  $\mathbb{R}$  è  $\mathcal{L}(a, b) = \pi \approx 3.14$ . Ma allora, in questo caso, rispetto alla misura dell'intervallo, il massimo di  $g$  non si discosta significativamente dal massimo di  $f$ . In altre parole, può dirsi verificata la condizione (61) in quanto si ha

$$\tilde{x} \approx \bar{x}$$

Ci si aspetta, dunque, che l'applicazione dell'importance sampling utilizzando la densità  $g$  possa risultare, in questo caso, più efficiente rispetto all'utilizzo del semplice metodo Monte Carlo. Affinché la densità  $g$  possa dirsi ben definita, si è normalizzata la funzione nell'intervallo, ricavando  $k \in \mathbb{R}$  tale che

$$k \int_0^\pi \sin(x) dx = 1 \iff 2k = 1 \iff k = \frac{1}{2}$$

Segue che la funzione densità di probabilità in esame si scriverà come

$$g(x) = \frac{1}{2} \sin(x) \quad \text{con} \quad x \in (0, \pi)$$

evidentemente positiva, come ci si aspetta. La sua cumulativa avrà la forma

$$G(x) = \frac{1}{2} \int_0^x \sin(t) dt = \frac{1}{2}(1 - \cos(x))$$

Ricavando  $x$  in funzione di  $G$  si ottiene la funzione inversa della cumulativa

$$G^{-1}(x) = \arccos(1 - 2x)$$

Si è quindi applicato il cambio di variabile dato dal metodo dell'importance sampling espresso dalla relazione (62), ottenendo

$$I_s = \int_0^1 \frac{\arccos(1 - 2x) \sin(\arccos(1 - 2x))}{2^{-1} \sin(\arccos(1 - 2x))} dx = 2 \int_0^1 \arccos(1 - 2x) dx$$

Si è poi applicato il metodo numerico espresso dalla seconda parte della (62). In particolare, si è scelto il range di punti

$$1000 \leq N < 50000 \quad \text{con} \quad N_{i+1} = N_i + 500$$

In questo caso, al fine di ottenere immediatamente un fit lineare più preciso, si è deciso di calcolare direttamente, per ogni  $N$ , le medie

$$\langle \Delta_s \rangle_M \quad \forall N = N_{min}, \dots, N_{max}$$

per  $M = 200$ , come nel caso precedente, per poi fissare i dati con la relazione lineare  $y = p + mx$ . Di seguito sono riportati i risultati ottenuti.

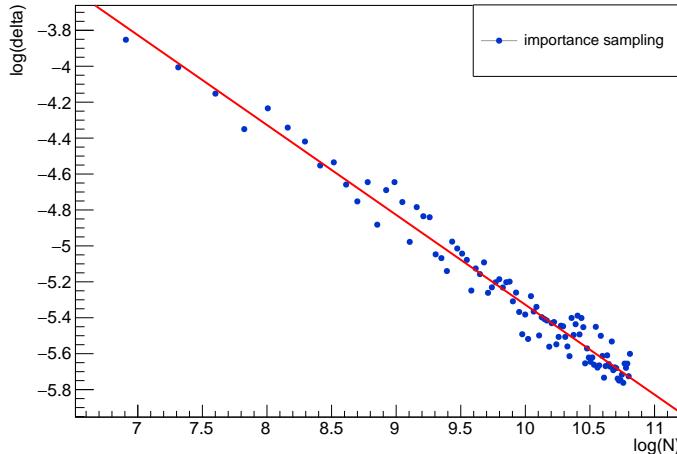


Figura 119:  $\log \Delta_s$  importance sampling per  $M = 200$ : fit

I parametri stimati risultano essere

$$p = -0.325 = \log k \quad \text{e} \quad m = -0.5$$

Come in precedenza, la procedura di rettificazione per mezzo del calcolo delle medie ha consentito di ottenere una stima di  $m$  molto compatibile con il valore atteso: in questo caso esatta, consentendo la verifica della solita relazione (51). Si è poi deciso di confrontare il metodo dell'importance sampling con il classico metodo di integrazione Monte Carlo dato dalla (48) plottando, nello stesso grafico, i due logaritmi delle dispersioni direttamente per  $M = 200$  e nello stesso range di  $N$  definito precedentemente, al fine di visualizzare subito un andamento preciso. Si sono ottenuti i seguenti risultati.

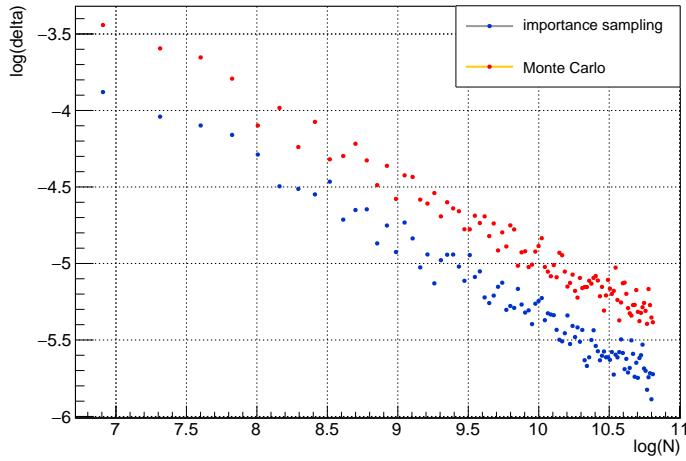


Figura 120: confronto  $\log \Delta_s$  per  $M = 200$

In questo caso, è evidente che il metodo dell'importance sampling risulti più efficiente del metodo Monte Carlo: la dispersione nel caso di Monte Carlo, infatti, risulta, a parità di numero di punti  $N$  generati, sempre maggiore rispetto rispetto alla dispersione dell'importance sampling. Quantitativamente è possibile verificare questo fatto interpolando i dati ottenuti con il metodo Monte Carlo con la funzione lineare  $y = q + wx$ .

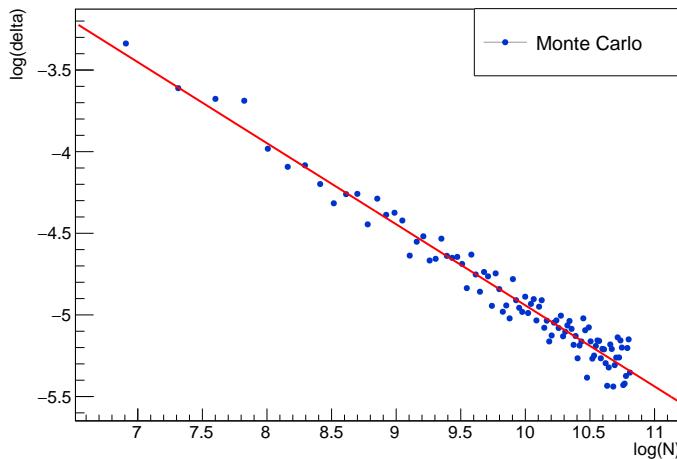


Figura 121:  $\log \Delta_s$  Monte Carlo per  $M = 200$ : fit

I parametri stimati risultano essere

$$q = 0.025 = \log k \quad \text{e} \quad w = -0.497 \approx -0.5$$

Confrontando la retta appena stimata con quella già ricavata dal fit per  $M = 200$  con il metodo dell'importance sampling si ha che

$$q > p \quad \text{e} \quad m \approx w$$

A parità di coefficiente angolare, si ha che l'intercetta della retta che descrive la dispersione del metodo Monte Carlo risulta maggiore dell'intercetta della retta che descrive la dispersione del metodo dell'importance sampling. Si ha quindi avuto modo di verificare, anche quantitativamente, la maggior efficienza del metodo dell'importance sampling se si utilizza la densità  $g$  come distribuzione. Quanto ottenuto è consistente con le considerazioni fatte inizialmente circa la vicinanza del punto di massimo di  $f$  dal punto di massimo della densità  $g$ . Si noti, infatti, che nell'intervallo di integrazione  $(0, \pi)$  le due funzioni presentano il seguente andamento.

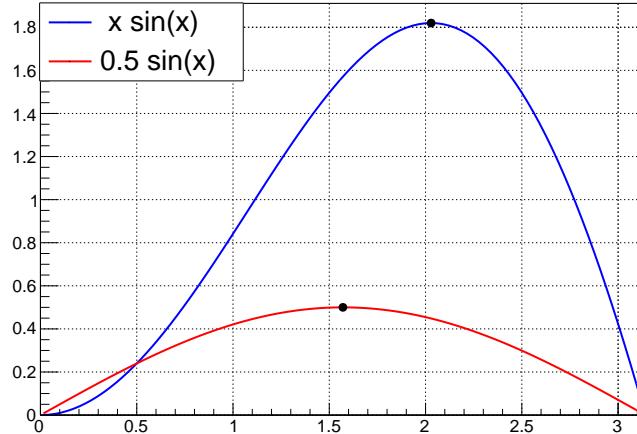


Figura 122: grafico delle funzioni  $f$  e  $g$  con rispettivi punti di massimo

Anche graficamente notiamo quanto si è già ricavato analiticamente: i due punti di massimo non si discostano significativamente l'uno dall'altro, da cui segue la verifica della condizione (61) di efficienza del metodo dell'importance sampling.

A differenza del caso precedente, per il calcolo di  $I_s$  con la densità  $g$ , si ha che vengono generati, a parità di  $N$ , più punti in un intorno del massimo dell'integrandà  $f$  rispetto al classico metodo Monte Carlo. Verranno quindi sommate più aree di rettangoli in una zona dove il contributo dell'integrale è più rilevante, inducendo una diminuzione di  $\sigma_f$  e dunque una maggiore efficienza del metodo dell'importance sampling a parità di punti generati.

## Metodi ad un passo per le ODE

Siamo interessati allo studio di diversi metodi numerici per la risoluzione di equazioni differenziali ordinarie e di problemi di Cauchy.

**Definizione 0.14.** Chiameremo *equazione differenziale ordinaria* (ODE) di *ordine n* un'equazione la cui incognita è una funzione di una variabile reale  $x = x(t)$  contenente le sue derivate fino alla derivata  $n$ -esima, ossia

$$\Phi(t, x, \dot{x}, \ddot{x}, \dots, x^{(n)}) = 0$$

dove  $\dot{x} := \frac{dx}{dt}$ ,  $\ddot{x} := \frac{d^2x}{dt^2}$ , ...,  $x^{(n)} := \frac{d^n x}{dt^n}$ . Inoltre, chiameremo

- *scalare* una ODE tale che

$$x : (t_0, t_1) \rightarrow \mathbb{R}$$

- *vettoriale* una ODE tale che

$$x : (t_0, t_1) \rightarrow \mathbb{R}^m$$

Se all'equazione differenziale è associato un sistema di  $n$  condizioni iniziali, chiameremo *problema di Cauchy* il sistema composto dalla ODE e dai *dati iniziali*.

**Definizione 0.15.** Chiameremo *sistema dinamico* (SD) in  $\mathbb{R}^n$  un sistema di  $n$  equazioni differenziali ordinarie al primo ordine della forma

$$\begin{cases} \dot{x}_1 = f_1(t, x_1, x_2, \dots, x_n) \\ \dot{x}_2 = f_2(t, x_1, x_2, \dots, x_n) \\ \vdots \\ \dot{x}_n = f_n(t, x_1, x_2, \dots, x_n) \end{cases}$$

Esistono diverse classi di equazioni differenziali che possono essere integrate. Tuttavia, in generale, una ODE di ordine superiore al primo non è risolubile analiticamente a meno di alcuni casi particolari, tipicamente appartenenti alla classe delle equazioni differenziali lineari di ordine  $n$ . Tale problema è in parte aggirabile grazie alla seguente proposizione.

**Proposizione 0.16** (Equivalenza ODE - SD). *Ogni equazione differenziale di ordine  $n$  è equivalente ad un sistema dinamico in  $\mathbb{R}^n$  come segue*

$$x^{(n)} = \Psi(t, x, \dot{x}, \ddot{x}, \dots, x^{(n-1)}) \iff \begin{cases} \dot{x} = x_1 \\ \dot{x}_1 = x_2 \\ \vdots \\ \dot{x}_{n-1} = x_n \\ \dot{x}_n = \Psi(t, x, x_1, x_2, \dots, x_n) \end{cases}$$

Le sostituzioni poste a partire dalla derivata prima della soluzione permettono dunque di ricondurre lo studio di una qualsiasi equazione differenziale di

ordine  $n$  allo studio di un sistema dinamico in  $\mathbb{R}^n$  della forma esplicitata. Chiaramente, non vale il viceversa della proposizione enunciata. Si noti, inoltre, che il lato destro del sistema dinamico di proposizione 0.16 può essere visto come un campo vettoriale che associa, ad ogni punto dello spazio delle fasi, la tangente alla soluzione del sistema. Questo fatto risulta particolarmente rilevante in quanto il calcolo e il disegno delle tangenti fornisce una verifica qualitativa del corretto andamento della soluzione di una ODE ricavata, ad esempio, computazionalmente per via numerica. Un caso fisico rilevante di applicazione dei concetti introdotti consiste nella risoluzione dell'equazione di Newton al fine di determinare l'equazione del moto di una particella dotata di massa. A tal proposito, chiameremo *equazione di tipo Newton* ogni ODE della forma

$$\ddot{x} = f(t, x, \dot{x})$$

Dalla proposizione 0.16 è immediato notare che ogni equazione di tipo Newton è equivalente al sistema dinamico in  $\mathbb{R}^2$

$$\begin{cases} \dot{x} = z \\ \dot{z} = f(t, x, z) \end{cases}$$

Le curve  $t \mapsto (x(t), z(t))$  soluzioni di tale sistema costituiscono le *curve di fase* e caratterizzano univocamente il moto della particella. Riuscire a risolvere una generica ODE al primo ordine, dunque, garantisce la possibilità di risolvere qualsiasi ODE di ordine  $n$  grazie alla proposizione 0.16. Siamo allora interessati allo studio di un problema di Cauchy della forma

$$\begin{cases} \frac{dx}{dt} = f(t, x) \\ x(t_0) = x_0 \end{cases} \quad (63)$$

caratterizzato da una generica ODE al primo ordine e da un generico dato iniziale. In generale, le ODE al primo ordine si risolvono analiticamente separando le variabili solo se la funzione  $f$  può essere scritta come

$$f(t, x) = g(x)h(t)$$

Oppure, utilizzando una nota formula di integrazione quando

$$f(t, x) = a(t)x(t) + b(t)$$

Se la forma analitica di  $f$  risulta troppo complessa, l'integrazione non è possibile, e risulta necessario procedere per via numerica. Analizziamo, quindi, tre diversi metodi numerici di risoluzione di problemi di Cauchy del tipo (63), in ordine crescente di precisione. Tutti i metodi analizzati consistono nel discretizzare la derivata prima, trasformandola nel suo rapporto incrementale. L'idea è quella di calcolare il punto  $x_{n+1}$  a partire dal punto precedente  $x_n$ , variando in modo costante il tempo  $t_n$  di un passo  $h$ . Ma allora, in tutti i casi si avrà

$$t_n = t_0 + nh \quad (64)$$

Un metodo numerico di risoluzione di una ODE che calcola  $x_{n+1}$  soltanto a partire dal punto precedente  $x_n$  è detto *metodo ad un passo*.

### Metodo di Eulero

Posto il dato iniziale, il metodo di Eulero ricostruisce iterativamente la soluzione della (63) al passo  $n + 1$  come

$$x_{n+1} = x_n + h f(t_n, x_n) + O(h^2) \quad (65)$$

Come si nota, l'errore sul singolo passo si scrive come

$$\Delta_E^n(h) = O(h^2)$$

Tuttavia, fissando un tempo  $t_{\tilde{n}}$  stiamo compiendo un numero  $n$  di steps. Dalla (64) si ha la relazione di proporzionalità

$$n \propto \frac{1}{h}$$

Combinandola con l'errore sul singolo passo, avremo che l'errore sul metodo di Eulero ad un tempo fissato sarà dato da

$$\Delta_E(h) = O(h) \iff \lim_{h \rightarrow 0} \frac{\Delta_E(h)}{h} = k \in \mathbb{R}$$

Approssimando la relazione al finito e calcolando i logaritmi si ricava la relazione lineare che descrive l'errore per il metodo di Eulero, che assume la forma

$$\log \Delta_E(h) \approx \log k + \log h \quad \text{con } h \text{ piccolo} \quad (66)$$

La formula di Eulero risulta, dunque, una formula al primo ordine in  $h$  e come tale porta, in generale, ad una ricostruzione poco precisa della soluzione al problema ai dati iniziali, mostrando una divergenza dalla funzione attesa già per valori di tempo piccoli. Il metodo di Eulero può essere utilizzato anche per risolvere numericamente sistemi dinamici in  $\mathbb{R}^m$ , applicando la (65) ad ogni ODE al primo ordine che costituisce il sistema. In tal caso, otterremo

$$x_{n+1}^i = x_n^i + h f^i(t_n, x_n^1, x_n^2, \dots, x_n^i) \quad \forall i = 1, \dots, m \quad (67)$$

Nel caso di un'equazione di tipo Newton valgono

$$m = 2 \quad \text{e} \quad f^1(t, x, z) = z$$

e il metodo da implementare si ridurrà banalmente al sistema accoppiato

$$\begin{cases} x_{n+1} = x_n + h z_n \\ z_{n+1} = z_n + h f(t_n, x_n, z_n) \end{cases}$$

dove i valori al passo zero sono definiti dai due dati di Cauchy associati.

### Metodo di Runge-Kutta 2

Posto il dato iniziale, il metodo di Runge-Kutta 2 ricostruisce iterativamente la soluzione della (63) al passo  $n + 1$  come

$$\begin{cases} x_{n+1} = x_n + k_2 + O(h^3) \\ k_1 = h f(t_n, x_n) \\ k_2 = h f(t_n + \frac{h}{2}, x_n + \frac{k_1}{2}) \end{cases} \quad (68)$$

Come si nota, l'errore sul singolo passo si scrive come

$$\Delta_{RK2}^n(h) = O(h^3)$$

Per ragioni esattamente analoghe a quelle precedenti, ad un tempo fissato, il metodo di Runge-Kutta 2 risulta un metodo al secondo ordine in  $h$ . Calcolando i logaritmi e troncando al finito si ottiene una relazione lineare della forma

$$\log \Delta_{RK2}(h) \approx \log k + 2 \log h \quad \text{con } h \text{ piccolo} \quad (69)$$

Il metodo di Runge-Kutta 2 può essere utilizzato anche per risolvere numericamente sistemi dinamici in  $\mathbb{R}^m$ , applicando la (68) ad ogni ODE al primo ordine che costituisce il sistema. In tal caso, otterremo

$$\begin{cases} x_{n+1}^i = x_n^i + k_2^i \\ k_1^i = hf^i(t_n, x_n^1, x_n^2, \dots, x_n^i) \\ k_2^i = hf^i\left(t_n + \frac{h}{2}, x_n^1 + \frac{k_1^i}{2}, \dots, x_n^i + \frac{k_1^i}{2}\right) \end{cases} \quad \forall i = 1, \dots, m \quad (70)$$

Nel caso di un'equazione di tipo Newton valgono

$$m = 2 \quad \text{e} \quad f^1(t, x, z) = z$$

e il metodo da implementare sarà il caso particolare

$$\begin{cases} x_{n+1} = x_n + k_2^x \\ z_{n+1} = z_n + k_2^z \\ k_1^x = hz_n \\ k_2^x = h\left(z_n + \frac{k_1^z}{2}\right) \\ k_1^z = hf(t_n, x_n, z_n) \\ k_2^z = hf\left(t_n + \frac{h}{2}, x_n + \frac{k_1^x}{2}, z_n + \frac{k_1^z}{2}\right) \end{cases}$$

dove i valori al passo zero sono definiti dai due dati di Cauchy associati.

#### Metodo di Runge-Kutta 4

Posto il dato iniziale, il metodo di Runge-Kutta 4 ricostruisce iterativamente la soluzione della (63) al passo  $n + 1$  come

$$\begin{cases} x_{n+1} = x_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) + O(h^5) \\ k_1 = hf(t_n, x_n) \\ k_2 = hf\left(t_n + \frac{h}{2}, x_n + \frac{k_1}{2}\right) \\ k_3 = hf\left(t_n + \frac{h}{2}, x_n + \frac{k_2}{2}\right) \\ k_4 = hf(t_n + h, x_n + k_3) \end{cases} \quad (71)$$

Come si nota, l'errore sul singolo passo si scrive come

$$\Delta_{RK2}^n(h) = O(h^5)$$

Per le solite ragioni, ad un tempo fissato, il metodo di Runge-Kutta 4 è un metodo al quarto ordine in  $h$ . Calcolando i logaritmi e troncando al finito si ottiene una relazione lineare della forma

$$\log \Delta_{RK4}(h) \approx \log k + 4 \log h \quad \text{con } h \text{ piccolo} \quad (72)$$

Il metodo di Runge-Kutta 4 fornisce un'approssimazione decisamente migliore della soluzione numerica rispetto agli altri due metodi discussi. Per tale ragione, risulta essere il metodo più comunemente utilizzato in questo contesto. Il metodo di Runge-Kutta 4 può essere utilizzato anche per risolvere numericamente sistemi dinamici in  $\mathbb{R}^m$ , applicando la (71) ad ogni ODE al primo ordine che costituisce il sistema. In tal caso, otterremo

$$\begin{cases} x_{n+1}^i = x_n^i + \frac{1}{6}(k_1^i + 2k_2^i + 2k_3^i + k_4^i) \\ k_1^i = hf^i(t_n, x_n^1, \dots, x_n^i) \\ k_2^i = hf^i\left(t_n + \frac{h}{2}, x_n^1 + \frac{k_1^i}{2}, \dots, x_n^i + \frac{k_1^i}{2}\right) \\ k_3^i = hf^i\left(t_n + \frac{h}{2}, x_n^1 + \frac{k_1^i}{2}, \dots, x_n^i + \frac{k_2^i}{2}\right) \\ k_4^i = hf^i\left(t_n + h, x_n^1 + k_3^i, \dots, x_n^i + k_3^i\right) \end{cases} \quad \forall i = 1, \dots, m \quad (73)$$

Nel caso di un'equazione di tipo Newton valgono

$$m = 2 \quad \text{e} \quad f^1(t, x, z) = z$$

e il metodo da implementare sarà il caso particolare

$$\begin{cases} x_{n+1} = x_n + \frac{1}{6}(k_1^x + 2k_2^x + 2k_3^x + k_4^x) \\ z_{n+1} = z_n + \frac{1}{6}(k_1^z + 2k_2^z + 2k_3^z + k_4^z) \\ k_1^x = hz_n \\ k_2^x = h\left(z_n + \frac{k_1^z}{2}\right) \\ k_3^x = h\left(z_n + \frac{k_2^z}{2}\right) \\ k_4^x = h(z_n + k_3^z) \\ k_1^z = hf(t_n, x_n, z_n) \\ k_2^z = hf\left(t_n + \frac{h}{2}, x_n + \frac{k_1^x}{2}, z_n + \frac{k_1^z}{2}\right) \\ k_3^z = hf\left(t_n + \frac{h}{2}, x_n + \frac{k_2^x}{2}, z_n + \frac{k_2^z}{2}\right) \\ k_4^z = hf\left(t_n + h, x_n + k_3^x, z_n + k_3^z\right) \end{cases}$$

dove i valori al passo zero sono definiti dai due dati di Cauchy associati.

## Esercizio 14

Si vuole studiare numericamente la soluzione al problema di Cauchy

$$\ddot{\theta} = -\theta \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 1 \end{cases}$$

utilizzando il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4. Si vuole quindi simulare il moto di un oscillatore armonico classico monodimensionale con pulsazione  $\omega = 1$  e massa unitaria della particella. Oppure, equivalentemente, di un pendolo matematico con approssimazione alle piccole oscillazioni, caratterizzato dalle medesime costanti fisiche.

Si noti che il caso dell'oscillatore armonico è uno dei pochi casi fisici per i quali l'equazione di Newton può essere integrata, permettendo di ottenere la soluzione esatta. Si è dunque ricavata la soluzione al problema di Cauchy notando che l'ODE in esame è lineare, al secondo ordine, omogenea e a coefficienti costanti. Le soluzioni del polinomio caratteristico sono  $\lambda \in \mathbb{C}$  tali che

$$\lambda^2 + 1 = 0 \quad \iff \quad \lambda = \pm i$$

Segue che lo spazio vettoriale delle soluzioni avrà la forma

$$V_\theta = \text{span}\{\cos(t), \sin(t)\}$$

ossia la famiglia a due parametri soluzione della ODE si scriverà come

$$\theta(t) = A \cos(t) + B \sin(t) \quad \forall A, B \in \mathbb{R}$$

Imponendo il passaggio per i dati di Cauchy si ricava la specializzazione dei parametri reali  $A$  e  $B$ , e dunque l'unica soluzione al problema in un intorno (almeno) dei dati iniziali, infatti

$$\begin{cases} \theta(0) = A \cos(0) + B \sin(0) = 0 \\ \dot{\theta}(0) = -A \sin(0) + B \cos(0) = 1 \end{cases} \iff \begin{cases} A = 0 \\ B = 1 \end{cases}$$

Ma allora, la soluzione al problema di Cauchy in esame è del tipo

$$\theta(t) = \sin(t)$$

Si noti che la soluzione analitica è ottenibile anche imponendo i dati iniziali alla nota soluzione dell'oscillatore armonico

$$\theta(t) = \theta_0 \sin(\omega t + \phi_0)$$

che equivale alla soluzione appena ricavata applicando la formula di addizione del seno e ridefinendo opportunamente le due nuove variabili fisiche  $\theta_0$  e  $\phi_0$  in funzione di  $A$  e  $B$ .

Vista la conoscenza, in questo caso particolare, della soluzione esatta, si vogliono confrontare i tre diversi metodi ad un tempo fissato, studiando la convergenza della soluzione numerica alla soluzione analitica, al fine di verificare il

corretto andamento dell'errore nei tre diversi casi. La verifica del corretto andamento dell'errore in questo caso particolare consentirà, infatti, di dare un certo grado di fiducia ai tre metodi implementati, rendendo legittima la loro applicazione anche negli esercizi successivi, nei quali l'integrazione analitica non sarà possibile. Si noti, dunque, che l'equazione in esame è di tipo Newton, pertanto, per quanto mostrato precedentemente, questa equivale ad un sistema dinamico in  $\mathbb{R}^2$  ai dati iniziali della forma

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\theta \end{cases} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ z(0) = 1 \end{cases}$$

al quale risulta ora possibile applicare le relazioni ricorsive dei tre metodi di integrazione numerica. Anzitutto, prima di studiare quantitativamente l'andamento dell'errore per i vari metodi in esame, si è deciso di svolgere una verifica qualitativa circa il corretto andamento delle soluzioni numeriche, confrontando i risultati ottenuti nei tre diversi casi. In particolare, si è fissato un valore del passo pari a  $h = 0.15$ , generando i punti  $(t_i, \theta_i)_{i=1, \dots, N}$  con  $N = 100$  secondo i tre metodi in esame. Si è poi plottata la funzione attesa sovrapposta ai tre set di punti, al fine di confrontarne visivamente la convergenza. Ci si aspetta che la soluzione numerica diverga più rapidamente da quella esatta al diminuire dell'ordine in  $h$  dei metodi. Di seguito sono riportati i risultati ottenuti.

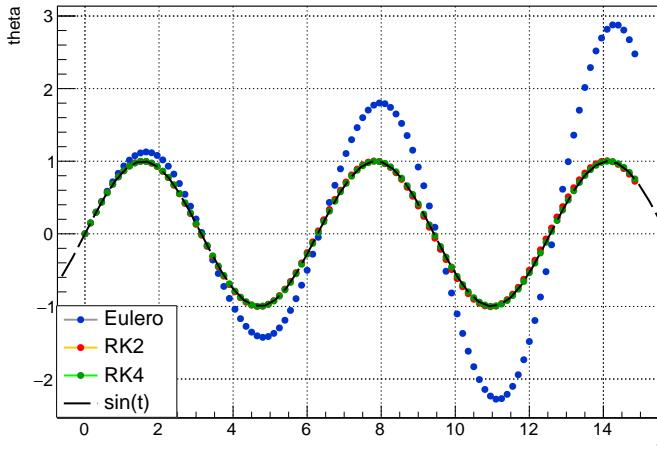


Figura 123: confronto E, RK2 e RK4 per  $h = 0.15$  e  $N = 100$

Come ci si aspetta, a parità di passo  $h$ , il metodo di Eulero ricostruisce una soluzione che diverge molto più rapidamente dalla soluzione attesa rispetto agli altri due metodi. I metodi di Runge-Kutta risultano, invece, qualitativamente sovrapponibili rispetto alla scala di grandezze in figura per i valori del passo e del numero di punti selezionati: sia tra loro, sia con la soluzione analitica. Si è quindi deciso di confrontare soltanto il metodo al secondo e al quarto ordine, aumentando il passo di integrazione a  $h = 0.3$  e diminuendo il numero di punti generati a  $N = 50$ , al fine di ottenere un intervallo di visualizzazione dei dati di misura identica rispetto alla misura dell'intervallo precedente. Di seguito sono riportati i risultati ottenuti a confronto.

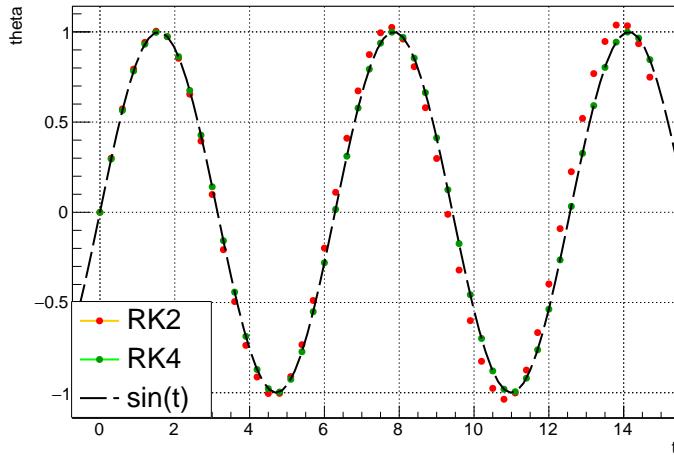


Figura 124: confronto RK2 e RK4 per  $h = 0.3$  e  $N = 50$

Ad un passo raddoppiato è possibile osservare qualitativamente quanto ci si aspetta: Runge-Kutta 2 risulta, all'aumentare del tempo, meno preciso rispetto al corrispondente metodo al quarto ordine, discostandosi sensibilmente dalla soluzione attesa. Il metodo di Runge-Kutta 4, invece, risulta essere qualitativamente sovrapponibile alla soluzione esatta anche ad un valore del passo di  $h = 0.3$ , coerentemente con il fatto che consiste in un metodo al quarto ordine in  $h$ . Appurata la consistenza delle caratteristiche qualitative delle soluzioni numeriche date dai tre metodi a confronto, siamo ora interessati ad uno studio più accurato a metodo fissato.

### Eulero

Fissato il metodo numerico siamo interessati, anzitutto, a svolgere una prima analisi qualitativa circa l'andamento della soluzione numerica al variare di  $h$ . Si è quindi deciso di plottare tre set di soluzioni per tre diversi valori del passo. In particolare, al fine di rendere efficace l'analisi visiva dei risultati, si vuole fare in modo che il range sull'asse dei tempi sia il medesimo per le tre diverse soluzioni, nonostante il diverso valore del passo di integrazione. Il passo, nei metodi analizzati, è legato al numero di punti dalla (64), da cui segue che

$$N = \frac{t_N - t_0}{h} \quad (74)$$

Si è scelto un intervallo di misura pari ad un periodo della soluzione esatta, ossia  $t_N = 2\pi$ . Si sono poi scelti e fissati i passi  $h_1 = 0.15$ ,  $h_2 = 0.1$  e  $h_3 = 0.05$ . Si sono quindi costruite le soluzioni numeriche con il metodo di Eulero con un numero di punti dato, per ogni  $h$ , dalla relazione (74). Per ragioni ovvie, ci si aspetta che la soluzione numerica converga alla soluzione esatta al diminuire del passo di integrazione. Allo stesso tempo, ci aspettiamo che ogni soluzione numerica si discosti sempre più dalla soluzione esatta all'aumentare del tempo a causa della propagazione dell'errore nei calcoli successivi. Di seguito sono riportati i tre diversi andamenti a confronto con la soluzione esatta.

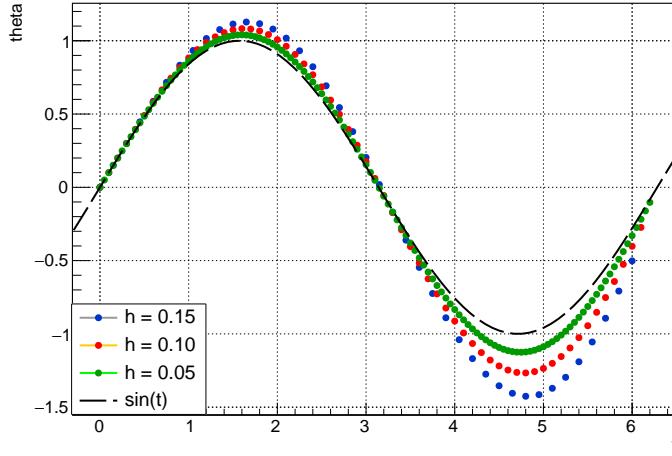


Figura 125: confronto E per  $h_1 = 0.15$ ,  $h_2 = 0.1$  e  $h_3 = 0.05$

Come ci si aspetta, la soluzione numerica mostra una crescente approssimazione alla soluzione analitica al diminuire di  $h$ . Allo stesso tempo, in tutti i casi, mostra una crescente discordanza dalla funzione attesa all'aumentare del tempo. Si vuole quindi verificare quantitativamente l'andamento dell'errore in funzione del passo di integrazione, dato dalla relazione (66). A tale scopo, si sono calcolate le dispersioni dal valore vero come

$$\Delta(h) = |\theta_N(h) - \sin(t_N)|$$

ad un tempo fissato  $t_N = 4$ . Al fine di assicurarsi che ogni vettore dei dati contenesse la posizione  $\theta_N$  corrispondente al tempo fissato  $t_N$ , si è deciso di non variare direttamente il passo  $h$  in un certo range, ma di variare il numero di punti generato  $N$  tale che fosse verificata la relazione (74). Tale relazione, infatti, lega il passo di integrazione  $h$  al numero totale di punti  $N$ : ad una variazione del numero dei punti corrisponderà una variazione inversamente proporzionale del passo. Formalmente, variando  $N$ , la (74) ci assicura che

$$t_N \in \vec{t} := (t_i)_{i=1,\dots,N} \quad \forall N = N_{min}, \dots, N_{max}$$

In particolare, ogni vettore dei tempi conterrà il tempo fissato  $t_N$  nella posizione corrispondente alla sua ultima componente. La ragione di questo passaggio indiretto consiste nel fatto che operare cicli iterativi su variabili reali può produrre perdite rilevanti di informazione a causa della rappresentazione in virgola mobile. Si sono quindi generate soluzioni numeriche  $(t_i, \theta_i)_{i=1,\dots,N}$  al problema di Cauchy con il metodo di Eulero al variare di  $N$  nel range

$$100 \leq N < 1000 \quad \text{con} \quad N_{i+1} = N_i + 10$$

salvando l'ultimo elemento del vettore  $\vec{\theta}$  delle posizioni. Ad ogni  $N$ , si è calcolato il passo  $h$  grazie alla relazione (74), per poi calcolare i logaritmi di  $\Delta(h)$  e del passo corrispondente. Si sono quindi interpolati i dati raccolti secondo una relazione lineare della forma  $y = p + mx$ , ottenendo quanto segue.

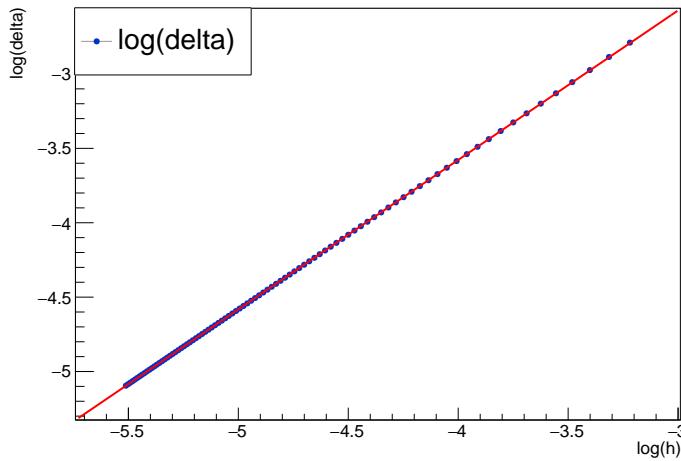


Figura 126:  $\log \Delta(h)$  con Eulero per  $t_N = 4$ : fit

I parametri stimati risultano

$$p = 0.442 = \log k \quad \text{e} \quad m = 1$$

La stima del coefficiente angolare della retta utilizzata per l'interpolazione dei dati risulta, dunque, consistente con la relazione (66) che governa la dispersione al variare del passo di integrazione. La verifica, in questo caso particolare, del fatto che il metodo di integrazione numerica implementato risulti al primo ordine in  $h$ , consente di garantire un certo grado di fiducia sulla corretta implementazione dell'algoritmo.

### Runge-Kutta 2

Al fine di studiare l'andamento dell'errore per il metodo di Runge-Kutta 2 si è deciso di svolgere immediatamente una verifica quantitativa dei risultati. In questo caso, infatti, il discostarsi della soluzione numerica rispetto a quella analitica, a parità di variazione di  $h$  risulta, visivamente, meno percettibile rispetto al metodo di Eulero, coerentemente con il fatto che il metodo di Runge-Kutta 2 risulta essere un metodo al secondo ordine in  $h$ . Anche in questo caso, dunque, si sono calcolati i logaritmi delle dispersioni in funzione dei logaritmi dei passi a tempo  $t_N = 4$  fissato, generando soluzioni numeriche  $(t_i, \theta_i)_{i=1,\dots,N}$  al variare del numero di punti  $N$  nel range

$$100 \leq N < 1000 \quad \text{con} \quad N_{i+1} = N_i + 10$$

sufficientemente grande da garantire il regime asintotico. Come in precedenza, il valore del passo è stato calcolato, per ogni  $N$ , secondo la relazione (74) a causa dei soliti problemi di perdita di informazione dati dalla rappresentazione finita dei numeri reali in un calcolatore. Cicli iterativi su interi positivi sono sempre preferibili in quanto non soggetti a problemi di approssimazione dati dalla rappresentazione in virgola mobile. Si sono quindi interpolati i dati raccolti secondo una funzione lineare della forma  $y = p + mx$ . Di seguito sono riportati i risultati ottenuti.

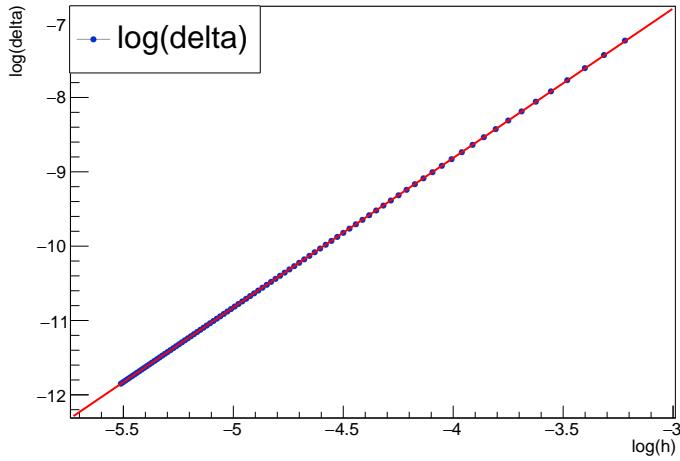


Figura 127:  $\log \Delta(h)$  con Runge-Kutta 2 per  $t_N = 4$ : fit

I parametri stimati risultano

$$p = -0.773 = \log k \quad \text{e} \quad m = 2.01 \approx 2$$

La stima del coefficiente angolare della retta utilizzata per l'interpolazione dei dati risulta, dunque, compatibile con il coefficiente della relazione (69) che governa la dispersione a variare del passo di integrazione. La verifica, in questo caso particolare, del fatto che il metodo di integrazione numerica implementato risulti al secondo ordine in  $h$ , consente di garantire un certo grado di fiducia sulla corretta implementazione.

#### Runge-Kutta 4

Anche in questo caso, al fine di studiare l'andamento dell'errore per il metodo di Runge-Kutta 4, si è deciso di svolgere immediatamente una verifica quantitativa dei risultati, tenendo conto della maggiore difficoltà ad avere verifiche visive adeguate a parità di variazione del passo. In questo caso, infatti, ci aspettiamo che il metodo risulti al quarto ordine in  $h$ , ossia che sia ancora più complesso trovare valori di  $h$  tali che sia possibile apprezzare graficamente la convergenza della soluzione numerica a quella attesa. Con una procedura del tutto analoga a quella adottata per gli altri due metodi numerici, si sono calcolati i logaritmi delle dispersioni in funzione dei logaritmi dei passi a tempo  $t_N = 4$  fissato, generando soluzioni numeriche  $(t_i, \theta_i)_{i=1, \dots, N}$  al variare di  $N$  nel range

$$100 \leq N < 1000 \quad \text{con} \quad N_{i+1} = N_i + 10$$

sufficientemente grande da ottenere un andamento ben definito e tale da assumere il regime asintotico. In particolare, anche in questo caso, il valore del passo è stato calcolato, per ogni  $N$ , secondo la relazione (74) per ragioni legate agli errori di approssimazione dei numeri reali. Si sono quindi interpolati i dati raccolti secondo una funzione lineare della forma  $y = p + mx$ . Di seguito sono riportati i risultati ottenuti.

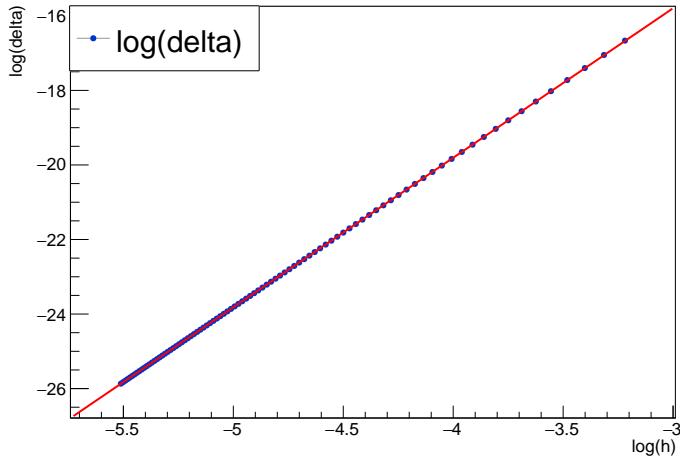


Figura 128:  $\log \Delta(h)$  con Runge-Kutta 4 per  $t_N = 4$ : fit

I parametri stimati risultano

$$p = -3.76 = \log k \quad \text{e} \quad m = 4.01 \approx 4$$

La stima del coefficiente angolare della retta utilizzata per l'interpolazione dei dati risulta, dunque, compatibile con il coefficiente della relazione (72) che governa la dispersione a variare del passo di integrazione. La verifica del fatto che il metodo di integrazione numerica implementato risulti al quarto ordine in  $h$  consente di garantire, anche in questo caso, un certo grado di fiducia sulla corretta implementazione.

### Analisi del sistema meccanico

Fino ad ora si sono studiati i metodi numerici di risoluzione al problema di Cauchy in esame. Concentriamoci, ora, sull'analisi del problema da un punto di vista meccanico. Siamo interessati, dunque, alla verifica o alla deduzione delle proprietà di un oscillatore armonico sfruttando la potenza di calcolo che abbiamo adesso a disposizione. Assumendo questa volta la corretta implementazione del codice, dalla convergenza dei fit effettuati segue indirettamente la verifica, anche da un punto di vista numerico, del fatto che la soluzione dell'oscillatore armonico classico in esame abbia la forma

$$\theta(t) = \sin(t)$$

Osserviamo ora che il sistema dinamico equivalente alla ODE di tipo Newton in esame risulta essere un sistema *autonomo e posizionale*, ossia tale che

$$\frac{df}{dt} = 0 \quad \text{e} \quad \frac{df}{d\dot{\theta}} = 0$$

dove  $f(t, \theta, \dot{\theta}) = -\theta$  è la forza di tipo elastico che agisce sul punto materiale. Siccome la forza in gioco è monodimensionale e non dipende né dal tempo né dalla velocità, segue che quest'ultima ammetterà potenziale. Scegliendo come

zero del potenziale elastico il punto  $\theta = 0$  corrispondente alla posizione a riposo della molla, il potenziale si scriverà come

$$U(\theta) = - \int_0^\theta f(\tilde{\theta}) d\tilde{\theta} = \frac{1}{2} \theta^2$$

per definizione di energia potenziale di un sistema conservativo. Ricordiamo, dunque, che vale il seguente teorema di capitale importanza per la meccanica classica.

**Teorema 0.17** (conservazione dell'energia meccanica). *Si consideri un sistema dinamico di tipo Newton, autonomo e posizionale della forma*

$$\begin{cases} \dot{x} = z \\ \dot{z} = -\frac{dU}{dx} \end{cases} \quad (75)$$

Allora, l'energia meccanica del sistema

$$E(x, z) := \frac{1}{2} z^2 + U(x)$$

è una costante del moto lungo le soluzioni di (75). Equivalentemente

$$\frac{dE}{dt} = 0 \quad \forall (x(t), z(t)) \in S$$

dove  $S$  è lo spazio delle curve soluzioni di (75).

Dalla conservazione dell'energia meccanica per un sistema la cui risultante ammette potenziale segue che, noti i dati iniziali  $x_0$  e  $z_0$ , si ha

$$E_0 := E(x_0, z_0) = E(x(t), z(t)) \quad \forall t \in \mathbb{R}$$

Il primo fatto fisico rilevante da verificare, dunque, sarà la conservazione dell'energia meccanica lungo le soluzioni del sistema: ci si aspetta che questa non cambi nel tempo, ossia che definisca una funzione costante per ogni  $t$ . Valutando l'energia meccanica nei dati di Cauchy si ottiene

$$E_0 = \frac{1}{2}$$

Ma allora, ci aspettiamo che la funzione energia evolva nel tempo conservando sempre lo stesso valore iniziale di  $E_0 = 1/2$ . Si è quindi costruita la soluzione al problema in esame  $(t_i, \theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  con  $N = 100$ , utilizzando il metodo più preciso di Runge-Kutta 4 con un valore del passo di  $h = 0.01$ , sufficientemente piccolo da consentire una buona approssimazione alla soluzione esatta. Si è poi calcolata l'energia meccanica ad ogni punto come

$$E_i = \frac{1}{2} \dot{\theta}_i^2 + \frac{1}{2} \theta_i^2 \quad \forall i = 1, \dots, N$$

Si sono quindi plottati i punti  $(t_i, E_i)_{i=1,\dots,N}$ , interpolando i dati con una funzione lineare della forma  $y = p + mx$ , al fine di svolgere una verifica quantitativa dell'andamento atteso. Di seguito sono riportati i risultati ottenuti.

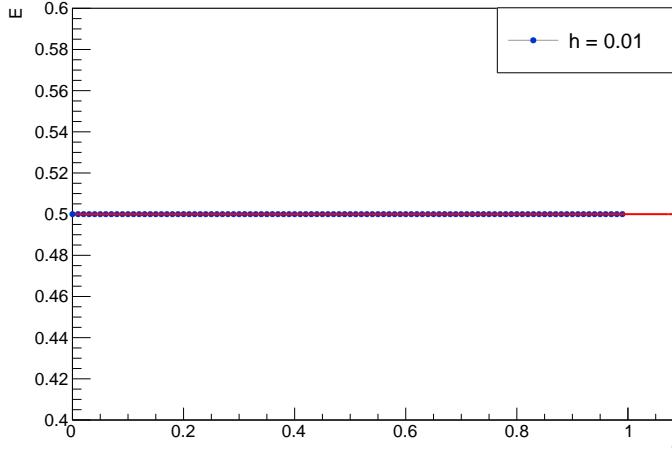


Figura 129:  $E(t)$  con Runge-Kutta 4 per  $h = 0.01$  e  $N = 100$ : fit

I parametri stimati risultano

$$p = 0.5 \quad e \quad m = -6.75 \cdot 10^{-13} \approx 0$$

Come è possibile notare, i parametri stimati risultano consistenti con il valore iniziale di energia  $E_0$ , fornendo addirittura una stima esatta dell'intercetta. Visto l'ordine di grandezza del coefficiente angolare è possibile concludere che il risultato non nullo ottenuto sia semplicemente dovuto alla limitatezza della precisione doppia utilizzata per il calcolo congiuntamente al fatto che, come ogni metodo numerico, anche il metodo di Runge-Kutta 4, per quanto preciso, fornisce soltanto risultati approssimati.

Dalla conservazione dell'energia segue che, posti i dati di Cauchy, la velocità di un punto materiale la cui evoluzione è descritta da un sistema della forma (75) è nota in funzione della posizione a meno di un segno come

$$z = \pm \sqrt{2(E_0 - U(x))} \quad (76)$$

Lo studio dell'andamento qualitativo della (76) è uno dei punti centrali della teoria dei sistemi dinamici applicata al caso dei sistemi meccanici conservativi. Le curve date da questa relazione, al variare dell'energia (e quindi dei dati iniziali), compongono il *diagramma di fase* del sistema meccanico. Il loro andamento permette di studiare l'equilibrio del sistema, la limitatezza del moto e la sua periodicità. Nel caso dell'oscillatore armonico in esame la (76) diventa

$$\dot{\theta}(\theta) = \pm \sqrt{2 \left( E_0 - \frac{1}{2} \theta^2 \right)} \quad \text{con} \quad E_0 = \frac{1}{2}$$

ossia, più esplicitamente

$$\dot{\theta}(\theta) = \pm \sqrt{1 - \theta^2} \quad \iff \quad \dot{\theta}^2 + \theta^2 = 1$$

Ci si aspetta, dunque, che la curva di fase per  $E = E_0$  sia la circonferenza unitaria centrata nell'origine. Al fine di verificare il corretto andamento qualitativo della curva di fase si è quindi costruita la soluzione  $(\theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  con  $N = 251$  utilizzando il metodo più preciso di Runge-Kutta 4, con un valore del passo di  $h = 0.05$ , sufficientemente piccolo da consentire una buona approssimazione alla soluzione esatta. Il numero  $N$  di punti generato è stato scelto, a seguito di diversi tentativi al calcolatore, al fine di compiere due giri completi della curva di fase. In particolare, si è ottenuto quanto segue.

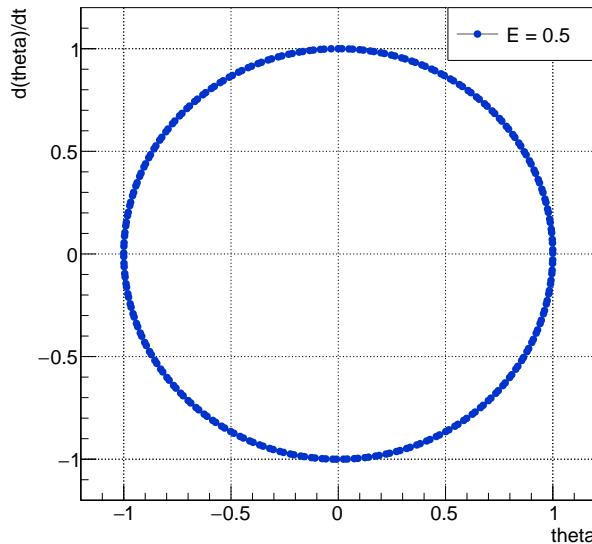


Figura 130:  $\dot{\theta}(\theta)$  con Runge-Kutta 4 per  $h = 0.05$  e  $N = 251$

Come è possibile notare, la curva di fase ottenuta assume la forma di una circonferenza unitaria centrata nell'origine, come ci si aspetta. Al fine di svolgere una verifica quantitativa dei risultati ottenuti vogliamo effettuare un fit dei dati raccolti. Tuttavia, si è notato che il programma di minimizzazione utilizzato svolge fit di dati solo secondo funzioni reali di variabile reale. Siccome la circonferenza, nelle usuali coordinate cartesiane, non è una funzione, l'interpolazione è possibile solo seguendo tre strade:

- la prima strada consiste nel dividere il set di dati  $(\theta_i, \dot{\theta}_i)$  in due sottoinsiemi: il primo contenente tutti i punti ad ordinata positiva e il secondo contenente i punti ad ordinata negativa. Si procede poi fittando singolarmente i due sottoinsiemi con le due semicirconferenze trovate in precedenza, che assumono la forma di funzioni ben definite
- la seconda strada consiste nell'elevare al quadrato le ordinate di ogni punto al fine di non avere problemi di segno, interpolando quindi i dati secondo una parabola
- in alternativa, si può pensare di scrivere la circonferenza come funzione delle due variabili polari, per poi svolgere due fit rettilinei separati

In questo caso, si è scelta la seconda strada, in quanto si ha avuto modo di verificare che l'interpolazione di dati secondo funzioni contenenti radici di ordine pari presenta diverse difficoltà a livello computazionale. Inoltre, la seconda via permette di svolgere un singolo fit senza scorrelare il set di dati raccolti. Si è quindi svolto un fit delle coppie ordinate  $(\theta_i, \dot{\theta}_i^2)_{i=1,\dots,N}$  secondo la funzione parabolica  $y = p + mx^2$ , ottenendo i seguenti risultati.

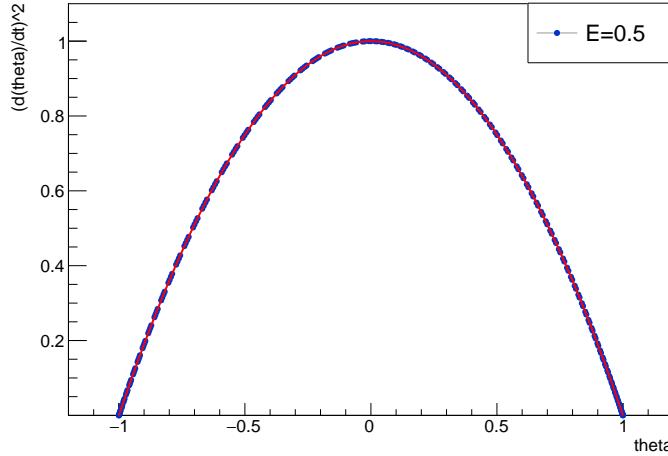


Figura 131:  $\dot{\theta}^2(\theta)$  con Runge-Kutta 4 per  $h = 0.05$  e  $N = 251$ : fit

I parametri stimati risultano

$$p = 1 \quad \text{e} \quad m = -1$$

Le stime ottenute risultano consistenti con la relazione analitica che descrive la curva di fase per il valore di energia  $E_0$ . Anche da quest'ultima verifica numerica, dunque, è possibile concludere che il moto di un oscillatore armonico per  $\omega = 1$  risulta limitato, in quanto la curva di fase è una curva limitata in un compatto reale. Inoltre, il moto è chiaramente periodico, in quanto si evolve sempre identico a se stesso da un punto di elongazione minima  $\theta = -1$  ad un punto di elongazione massima  $\theta = 1$ .

Sfruttando la potenza di calcolo di un calcolatore esiste un'altra strada per verificare qualitativamente il corretto andamento di una curva di fase. Come si è accennato nella parte introduttiva, ad ogni sistema dinamico è possibile associare univocamente un campo vettoriale, dato dal lato destro del sistema stesso. Nel caso in esame si ha la mappa

$$\begin{pmatrix} \theta \\ z \end{pmatrix} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \mapsto \begin{pmatrix} z \\ -\theta \end{pmatrix} = \begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix}$$

Come si nota dalle seconde uguaglianze il campo associa, ad ogni punto del piano delle fasi, la tangente alla curva di fase corrispondente. Il fatto rilevante consiste nel notare che disponiamo analiticamente del lato destro del sistema dinamico. Ma allora, per avere un'idea qualitativa dell'andamento delle curve di fase di un sistema dinamico autonomo, basterà valutare il lato destro del

sistema in diversi punti del piano delle fasi ottenendo, in questo modo, i vettori tangenti alle curve nel punto. Si è allora costruita una griglia regolare di  $N = 20$  punti per lato del tipo  $\{(\theta_i, \dot{\theta}_j)\}_{i,j=1,\dots,N^2}$  nel quadrato  $[-5, 5] \times [-5, 5]$  centrato nell'origine di un sistema cartesiano. Si è poi calcolato il lato destro del sistema ottenendo punti della forma  $\{(\dot{\theta}_i, \ddot{\theta}_j)\}_{i,j=1,\dots,N^2}$  corrispondenti alle componenti dei vettori che definiscono il campo. Posto

$$\vec{V} := \begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix}$$

si è poi normalizzata ogni coppia calcolando

$$\vec{V}_n := \frac{\vec{V}}{\gamma \|\vec{V}\|} \quad \text{con} \quad \gamma \in \mathbb{R}$$

al fine di non creare problemi di visualizzazione dati dalla sovrapposizione dei vettori tangenti nel piano delle fasi. Il parametro  $\gamma$  serve unicamente a riscalare tutti i vettori di una certa quantità, per le stesse ragioni di sovrapposizione grafica: il suo valore è stato fissato dopo diversi tentativi di visualizzazione ottimale del campo vettoriale. In tal modo, chiaramente, si perde l'informazione sull'intensità del campo (che può comunque essere recuperata assegnando dei colori), ma l'interesse in questo caso è unicamente rivolto verso l'andamento qualitativo delle curve di fase, ossia verso la direzione e il verso dei vettori tangenti. Fissando  $\gamma = 2.1$  si è ottenuto il grafico che segue.

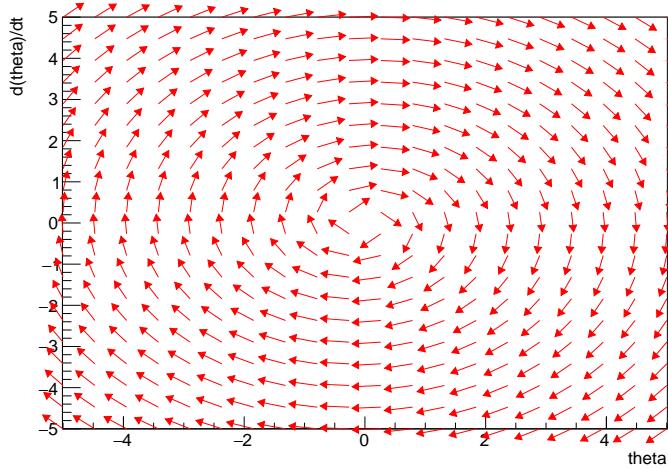


Figura 132: andamento qualitativo delle tangenti alle curve di fase

Dal grafico è possibile estrarre numerose informazioni circa il sistema dinamico in esame. Anzitutto, i moti sono periodici e limitati per ogni valore di energia, e dunque dei dati iniziali. Questo è consistente se si pensa al fatto che l'andamento del potenziale è parabolico, ossia i moti permessi risultano definiti su un compatto per ogni valore di energia. In secondo luogo, la topologia delle tangenti alle curve di fase mostra che  $(\theta, \dot{\theta}) = (0, 0)$  è un punto di equilibrio stabile per il sistema: fatto non sorprendente se si pensa alla corrispondenza fisica del problema che stiamo analizzando.

Arrivati a questo punto della trattazione siamo riusciti a verificare, in diversi modi sia qualitativi che quantitativi, che il moto di un oscillatore armonico è periodico. L'ultimo punto rilevante consiste, dunque, nella stima del periodo del moto. Detti  $x_m$  il punto di elongazione minima e  $x_M$  il punto di elongazione massima, reinterpretando la relazione (76) come

$$\frac{dx}{dt} = \pm \sqrt{2(E_0 - U(x))}$$

si ha che l'intervallo di tempo infinitesimo corrispondente allo spostamento infinitesimo del punto si scriverà come

$$dt = \frac{1}{\sqrt{2(E_0 - U(x))}} dx$$

da cui segue che il periodo per un dato valore di energia  $E_0$  fissato sarà

$$T_{E_0} := 2 \int_{x_m}^{x_M} \frac{1}{\sqrt{2(E_0 - U(x))}} dx \quad (77)$$

In particolare, nel caso dell'oscillatore armonico in esame, si avrà

$$T_{1/2}^\omega = 2 \int_{-1}^1 \frac{1}{\sqrt{1 - \theta^2}} d\theta = 2\pi$$

coerentemente con il periodo dell'equazione sinusoidale del moto ricavata inizialmente per via analitica. Si è quindi deciso di verificare numericamente il risultato ottenuto calcolando l'integrale  $T_{1/2}^\omega$  utilizzando un metodo di Newton-Cotes. L'integrale in esame, infatti, risulta monodimensionale: per quanto si ha avuto modo di verificare nelle sezioni precedenti, l'utilizzo dei metodi deterministici è da privilegiare, in questi casi, rispetto all'utilizzo dei metodi Monte Carlo, più utili nel caso di integrali multidimensionali. Anzitutto, si è notato che la funzione integranda risulta definita soltanto nell'aperto  $(-1, 1)$ . In particolare, per  $\theta = \pm 1$  presenta due singolarità, infatti

$$\lim_{\theta \rightarrow \pm 1} \frac{1}{\sqrt{1 - \theta^2}} = +\infty$$

ossia le rette  $\theta = \pm 1$  risultano asintoti verticali per la funzione. Per il calcolo di  $T_{1/2}^\omega$  risulta allora necessario l'utilizzo delle formule aperte negli intorni dei punti singolari. Si è quindi diviso l'intervallo di integrazione in tre sotto-intervalli come

$$(-1, 1) = (-1, -1 + \epsilon) \cup [-1 + \epsilon, 1 - \epsilon] \cup (1 - \epsilon, 1)$$

Il periodo si scriverà, allora, come

$$T_{1/2}^\omega = \int_{-1}^{-1+\epsilon} \frac{2}{\sqrt{1 - \theta^2}} d\theta + \int_{-1+\epsilon}^{1-\epsilon} \frac{2}{\sqrt{1 - \theta^2}} d\theta + \int_{1-\epsilon}^1 \frac{2}{\sqrt{1 - \theta^2}} d\theta$$

Si è poi applicata la più precisa formula aperta per  $N = 6$  punti negli intervalli contenenti le singolarità, e la formula di Romberg per  $J = K = 10$  nell'intervallo centrale senza punti singolari. Al fine di verificare quale fosse il valore di  $\epsilon$  che

consentisse una stima migliore del periodo si sono calcolate le dispersioni dal valore vero che, come al solito, si scrivono come

$$\Delta(\epsilon) = \left| \tilde{T}_{1/2}(\epsilon) - 2\pi \right|$$

nel range di valori

$$0.00005 \leq \epsilon < 0.001 \quad \text{con} \quad \epsilon_{i+1} = \epsilon_i + 0.000015$$

dove  $\tilde{T}_{1/2}$  rappresenta la stima numerica del periodo del moto in funzione di  $\epsilon$  con i metodi appena discussi. Si sono quindi plottati i valori della dispersione dal valore vero in funzione dei corrispondenti valori dell'estremo  $\epsilon$  di integrazione, ottenendo il seguente andamento.

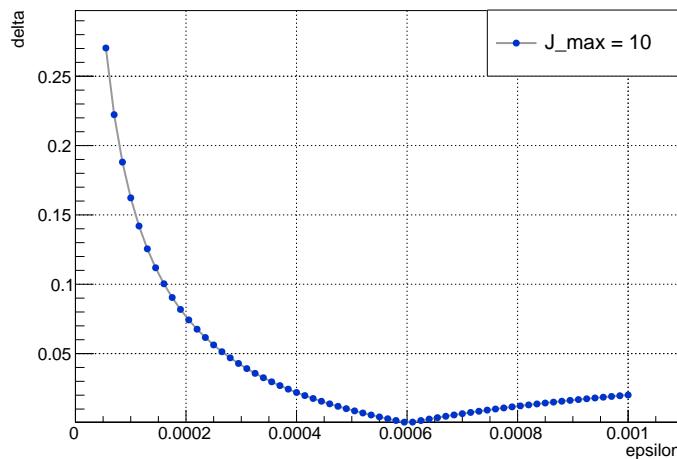


Figura 133:  $\Delta(\epsilon)$  con formula aperta per  $N = 6$  e Romberg per  $J_{max} = 10$

Come è possibile notare dal plot, la stima del periodo con la tecnica utilizzata raggiunge un valore prossimo al valore vero (con dispersione nulla) per un un valore di  $\epsilon = 0.0006$ . Visti i risultati ottenuti, risulta possibile concludere anche la verifica numerica della stima del periodo di un oscillatore armonico di pulsazione unitaria.

## Esercizio 15

Si vuole studiare numericamente la soluzione al problema di Cauchy

$$\ddot{\theta} = -\sin(\theta) + g(t, \dot{\theta}) \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 1 \end{cases}$$

utilizzando il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4. Si vuole quindi simulare numericamente il moto di un pendolo matematico senza l'approssimazione alle piccole oscillazioni, con costante normalizzata  $g/l = 1$  e massa della particella classica unitaria. In particolare, si vogliono studiare tre diversi casi fisici dello stesso problema: il moto senza attrito, il moto con un attrito dipendente dalla velocità e il moto con attrito con l'aggiunta di una forzante esterna periodica dipendente dal tempo. La specializzazione della funzione  $g$ , di volta in volta, renderà conto del caso fisico analizzato.

Non è difficile verificare che, qualunque sia la forma della funzione  $g$ , il problema di Cauchy in esame non ammette soluzione analitica esprimibile per mezzo di una combinazione di funzioni elementari. In assenza della soluzione esatta, per l'analisi della consistenza e della precisione delle soluzioni numeriche sarà allora necessario trovare altre strade.

### Pendolo in vuoto

Se il pendolo semplice non è soggetto a forze diverse da quella data dal campo di forze gravitazionale terrestre si ha

$$g(t, \dot{\theta}) = 0$$

e il problema di Cauchy in esame si riduce ad un problema della forma

$$\ddot{\theta} = -\sin(\theta) \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 1 \end{cases}$$

Dall'equivalenza 0.16 segue che il problema in esame equivale allo studio del sistema dinamico in  $\mathbb{R}^2$  di tipo Newton

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\sin(\theta) \end{cases} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ z(0) = 1 \end{cases}$$

al quale risulta ora possibile applicare le relazioni ricorsive date dai tre metodi. Come prima cosa, al fine di operare un confronto visivo dell'andamento della soluzione al variare del metodo numerico, si è deciso di calcolare e plottare le coppie  $(t_i, \theta_i)_{i=1,\dots,N}$  secondo i tre metodi, fissando un passo  $h = 0.15$  e con un numero di punti di integrazione pari a  $N = 100$ . Visto l'andamento dell'errore in funzione del passo che si ha avuto modo di verificare nel caso particolare dell'esercizio precedente, ci si aspetta che la soluzione ricostruita con il metodo di Eulero diverga significativamente dalle soluzioni ricostruite con gli altri due metodi. Ci si aspetta, inoltre, che il metodo al secondo e al quarto ordine possano ricostruire soluzioni visivamente sovrapponibili per il passo fissato. Di seguito sono riportati i risultati ottenuti a confronto.

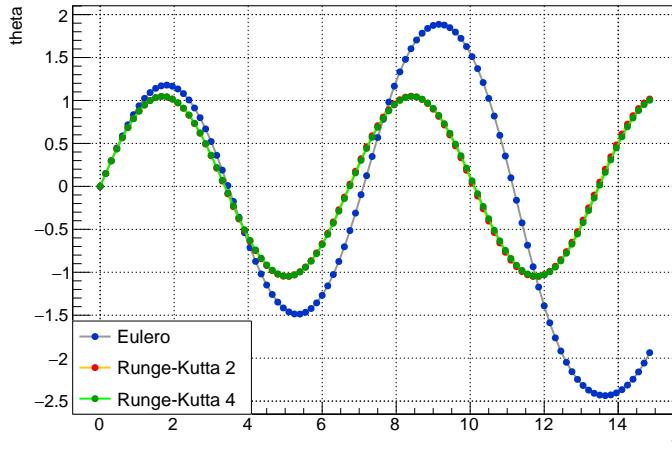


Figura 134: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.15$  e  $N = 100$

Come è possibile notare, il metodo di Eulero si discosta in modo significativo dal metodo di Runge-Kutta 2 e 4 già a partire da tempi piccoli, coerentemente con il fatto che risulta essere un metodo al primo ordine in  $h$ . Nella scala del grafico riportato, invece, i metodi Runge-Kutta appaiono, di fatto, sovrapponibili. Si è allora deciso di confrontare soltanto il metodo al secondo e al quarto ordine, aumentando il passo di integrazione a  $h = 0.3$  e diminuendo il numero di punti generati a  $N = 50$ , al fine di ottenere un intervallo di visualizzazione dei dati di misura analoga a quella precedente, ottenendo quanto segue.

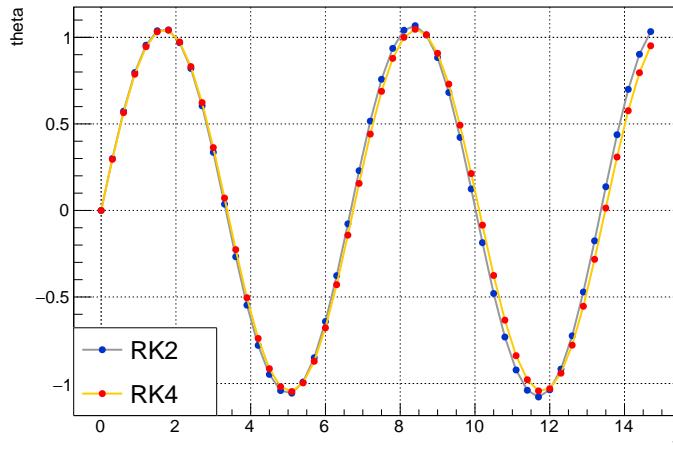


Figura 135: confronto  $\theta(t)$  con RK2 e RK4 per  $h = 0.3$  e  $N = 50$

Ad un passo raddoppiato è possibile osservare qualitativamente quanto ci si aspetta: Runge-Kutta 2 risulta discostarsi, all'aumentare del tempo, sempre di più rispetto al corrispondente metodo al quarto ordine. Si è quindi eseguita la medesima operazione calcolando e plottando  $(t_i, \dot{\theta}_i)_{i=1,\dots,N}$  secondo i tre metodi,

fissando un valore del passo pari a  $h = 0.15$  e con  $N = 100$ . I grafici della velocità in funzione del tempo risultano avere l'andamento che segue.

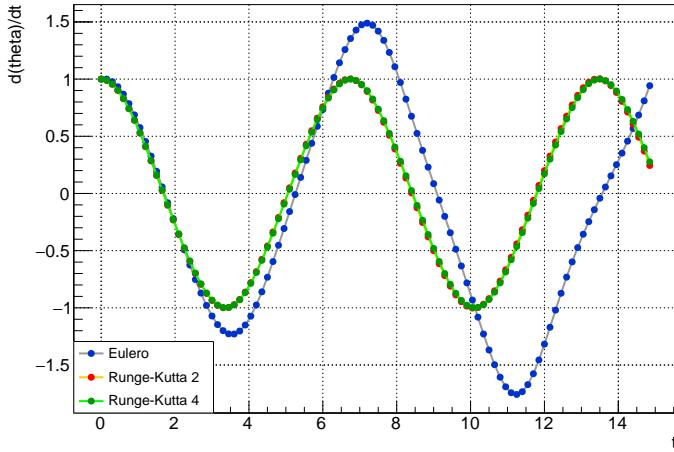


Figura 136: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.15$  e  $N = 100$

Anche nel caso delle velocità, il metodo di Eulero si discosta in modo significativo dal metodo di Runge-Kutta 2 e 4 già a partire da tempi piccoli. I due metodi di ordine 2 e 4, invece, appaiono del tutto sovrapponibili nella scala della figura riportata. Si è allora deciso di confrontare soltanto il metodo al secondo e al quarto ordine, aumentando il passo di integrazione a  $h = 0.3$  e diminuendo il numero di punti generati a  $N = 50$ , esattamente come in precedenza. Di seguito sono riportati i risultati ottenuti.

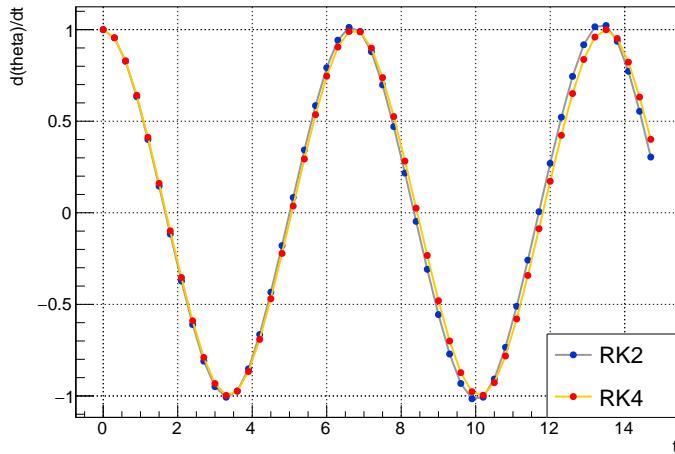


Figura 137: confronto  $\dot{\theta}(t)$  con RK2 e RK4 per  $h = 0.3$  e  $N = 50$

Coerentemente con il fatto che i due metodi differiscono di due ordini in precisione, si osserva che la curva che descrive la velocità in Runge-Kutta 2 inizia a discostarsi dalla velocità in Runge-Kutta 4 a partire da un certo valore tem-

porale. Questi primi dati confermano qualitativamente quanto ci si aspetta dai tre metodi numerici studiati. Risulta quindi interessante studiare fino a quale valore temporale la soluzione numerica al problema di Cauchy può darsi corretta entro un certo valore di precisione fissato al variare del passo  $h$  di integrazione. In questo caso, non disponendo della soluzione analitica, non è possibile svolgere questa analisi direttamente. Tuttavia, notiamo che siccome la forza  $f(\theta) = -\sin(\theta)$  dipende dalla sola coordinata generalizzata  $\theta$ , allora questa ammetterà potenziale, che possiamo scrivere esplicitamente come

$$U(\theta) = - \int_0^\theta f(\tilde{\theta}) d\tilde{\theta} = 1 - \cos(\theta)$$

una volta posto lo zero di  $U$  in corrispondenza della posizione verticale del pendolo per  $\theta = 0$ . Il sistema fisico in esame risulta allora conservativo, ossia dovrà valere il teorema di conservazione dell'energia meccanica. Sapendo che  $E(\theta, \dot{\theta})$  deve conservarsi nel tempo, risulta allora possibile utilizzare questa costante del moto per operare l'analisi di cui sopra. Anzitutto, valutando l'energia meccanica nei dati di Cauchy si ottiene

$$E_0 = \frac{1}{2}$$

Si è quindi costruita la soluzione  $(t_i, \theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  per ognuno dei tre metodi in esame, fissando tre diversi passi:  $h_1 = 0.1$ ,  $h_2 = 0.05$  e  $h_3 = 0.01$ . L'idea è quella di fissare un valore temporale  $\bar{t} = 3$ . A questo punto, il numero di punti generati per arrivare al tempo fissato sarà automaticamente dato dalla (64). Si è poi calcolata l'energia meccanica ad ogni punto come

$$E_i = \frac{1}{2} \dot{\theta}_i^2 + (1 - \cos(\theta_i)) \quad \forall i = 1, \dots, N$$

per ognuno dei tre metodi. Per il metodo di Eulero si è ottenuto quanto segue.

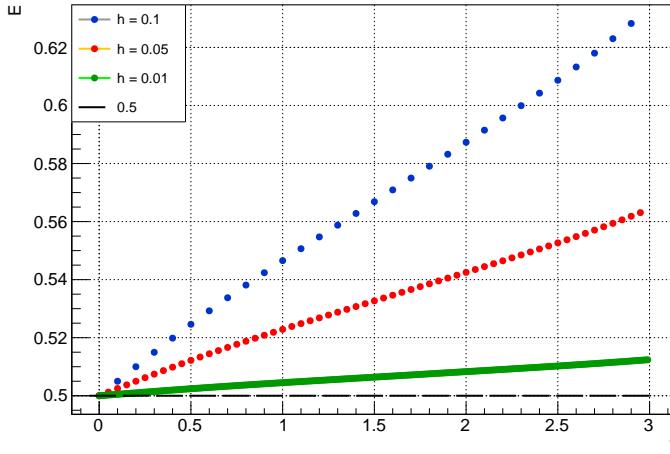


Figura 138: confronto  $E(t)$  con  $E$  per tre valori di  $h$  con  $\bar{t} = 3$

Come è possibile notare, il metodo di Eulero risulta, non sorprendentemente, poco preciso. In particolare, dal tempo iniziale fino a  $t = \bar{t}$ , Eulero riesce a

ricostruire la soluzione con un errore massimo sull'energia di  $\varepsilon_1 = 0.14$  per il passo  $h_1$ , di  $\varepsilon_2 = 0.07$  per il passo  $h_2$  e di  $\varepsilon_3 = 0.02$  per il passo  $h_3$ . I risultati utilizzando Runge-Kutta 2, invece, sono i seguenti.

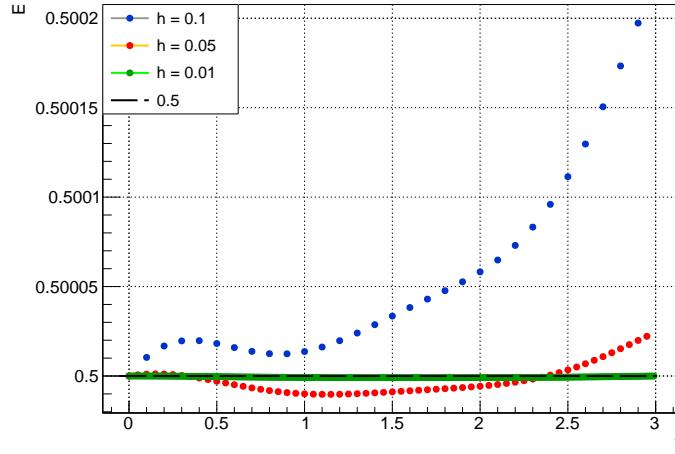


Figura 139: confronto  $E(t)$  con RK2 per tre valori di  $h$  con  $\bar{t} = 3$

Dal grafico riportato è evidente che il metodo di Runge-Kutta 2 ricostruisca la soluzione in modo decisamente più preciso rispetto al metodo di Eulero a parità di passo e di tempo finale, come ci si aspetta. In particolare, calcolando le dispersioni dal valore vero si verifica facilmente che la soluzione viene ricostruita, in questo caso, con un errore massimo sull'energia di  $\varepsilon_1 = 0.0002$  per il passo  $h_1$ , di  $\varepsilon_2 = 2.2 \cdot 10^{-5}$  per il passo  $h_2$  e di  $\varepsilon_3 = 9.1 \cdot 10^{-7}$  per il passo  $h_3$ . Per Runge-Kutta 4 si sono ottenuti i risultati che seguono.

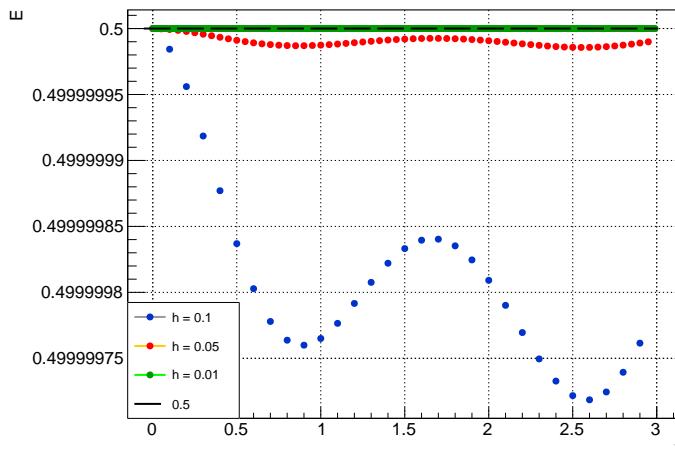


Figura 140: confronto  $E(t)$  con RK4 per tre valori di  $h$  con  $\bar{t} = 3$

Osservando la scala verticale del grafico riportato è possibile notare un nuovo sensibile miglioramento della soluzione ricostruita rispetto a quella attesa, come

ci si aspetta. Calcolando anche in questo caso le dispersioni dal valore atteso per ogni punto si è verificato che il metodo di Runge-Kutta 4 ricostruisce la soluzione fino a  $\bar{t}$  con un errore massimo sull'energia di  $\varepsilon_1 = 2.8 \cdot 10^{-7}$  per il passo  $h_1$ , di  $\varepsilon_2 = 1.4 \cdot 10^{-8}$  per il passo  $h_2$  e di  $\varepsilon_3 = 1.9 \cdot 10^{-11}$  per il passo  $h_3$ . Evidentemente, questo metodo restituisce un errore assoluto sulla sola energia meccanica. Spesso, tuttavia, siamo interessati a conoscere l'errore sulla posizione  $\theta$  o sulla velocità  $\dot{\theta}$ , piuttosto che sull'energia. In tal caso, sarà necessario propagare l'errore sulla grandezza a cui siamo interessati. Sappiamo che vale  $E = E(\theta, \dot{\theta})$  in generale. Al fine di riuscire a propagare l'errore analiticamente, dobbiamo sempre essere in grado di determinare

$$\theta = \theta(E) \quad \text{e} \quad \dot{\theta} = \dot{\theta}(E)$$

eventualmente a meno di un segno che, come vedremo, sarà irrilevante. Consideriamo, ad esempio, il problema di stima dell'errore su  $\theta$  noto l'errore su  $E$ . Con uno sviluppo in Taylor troncato al primo ordine si può dimostrare che, per funzioni di una singola variabile reale, vale la formula di propagazione

$$\text{err}(\theta) = \left| \frac{\partial}{\partial E} \theta(E) \right| \text{err}(E)$$

dove  $E$  è il valore vero dell'energia nel punto d'interesse, mentre  $\text{err}(E)$  è l'errore sull'energia associato a quel punto. Si è utilizzata la derivata parziale in quanto la funzione  $\theta$  dipende, a priori, anche da  $\dot{\theta}$ , che in questo caso stiamo supponendo fissata. Come è possibile notare, il modulo che compare nella formula di propagazione rende irrilevante il segno che compare davanti alla forma analitica di  $\theta$ . Questo garantisce che, in caso di inversioni determinate a meno di un segno, l'errore rimarrà comunque univocamente determinato. Si noti, inoltre, che dati due valori di energia  $E_1 = E_1(t_1)$  ed  $E_2 = E_2(t_2)$  tali che  $t_1 < t_2$ , una proprietà comoda che vorremmo sempre verificata è

$$\text{err}(E_1) < \text{err}(E_2)$$

Se l'errore cresce nel tempo, infatti, per essere certi di generare una soluzione precisa entro una certa precisione quantificabile fino ad un tempo fissato basterà studiare l'errore al solo tempo finale. Come si nota dai risultati ottenuti, questo fatto è vero solo in media nel problema in esame: per alcuni valori del passo l'andamento dell'errore oscilla, ma mediamente cresce con il tempo. La proprietà desiderata semplifica di molto lo studio sull'errore, in quanto in sua assenza sarebbe necessario svolgere lo stesso studio per tanti valori temporali intermedi tra il tempo iniziale e il tempo finale. Se vogliamo simulare l'evoluzione per tempi grandi, questo fatto minerebbe alla possibilità di ottenere soluzioni in tempi ragionevoli. La proprietà richiesta, semplificando molto, va sotto il nome di *stabilità* di una soluzione e fa parte delle caratteristiche di *buona positura* di un'equazione differenziale. Seppur sembri del tutto naturale aspettarci che l'errore cresca nel tempo, esistono casi di sistemi non lineari la cui soluzione risulta molto sensibile agli errori che si propagano durante il calcolo iterativo dei punti della soluzione a partire dai punti precedenti, determinando andamenti non significativi anche per tempi finali piccoli. In altre parole, in alcuni sistemi l'errore cresce esponenzialmente generando andamenti che possono anche tornare a sovrapporsi agli andamenti attesi all'aumentare del tempo (determinando

un valore di dispersione nullo), ma in modo del tutto imprevedibile. In questi casi, uno studio in precisione al solo tempo finale risulta, chiaramente, privo di significato. Sistemi di questo tipo sono detti *caotici* e avremo modo di studiarli più nel dettaglio negli esercizi che seguono.

L'esistenza di una costante del moto nel problema meccanico in esame ha quindi permesso di operare un'analisi in precisione della soluzione numerica anche nel caso di una ODE la cui soluzione analitica non è esprimibile per mezzo di funzioni elementari. Risulta piuttosto immediato concludere, visti i risultati appena ottenuti, che la conservazione dell'energia è verificata anche in questo caso. Al fine di svolgere una verifica quantitativa si è deciso di plottare i punti  $(t_i, E_i)_{i=1,\dots,N}$ , interpolando i dati con una funzione lineare della forma  $y = p + mx$ . Si è utilizzato il metodo più preciso di Runge-Kutta 4, con un passo sufficientemente piccolo di  $h = 0.05$  generando  $N = 100$  punti. Di seguito sono riportati i risultati ottenuti.

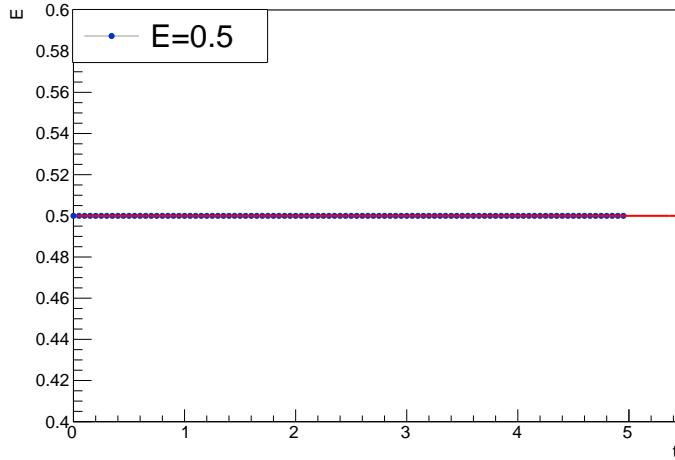


Figura 141:  $E(t)$  con Runge-Kutta 4 per  $h = 0.05$  e  $N = 100$ : fit

I parametri stimati risultano

$$p = 0.5 \quad \text{e} \quad m = -1.71 \cdot 10^{-9} \approx 0$$

Come è possibile notare, i parametri stimati risultano consistenti con il valore iniziale di energia  $E_0$ , fornendo addirittura una stima esatta dell'intercetta. Può quindi dirsi verificato in modo più quantitativo il teorema di conservazione di energia meccanica.

A parità di passo  $h$ , dunque, il metodo che meglio ricostruisce la soluzione al problema di Cauchy in esame risulta essere Runge-Kutta 4. Si noti, ad esempio da figura 135, che RK4 ricostruisce un moto che appare essere periodico e con andamento sinusoidale. Contro ogni testo di analisi, potrebbe allora sorgere il dubbio che il moto possa essere descritto dalla soluzione dell'oscillatore armonico precedente, ossia  $\theta(t) = \sin(t)$ . Si è quindi costruita una soluzione con RK4 per  $h = 0.15$  e  $N = 100$ , sovrapponendo nel plot il grafico della soluzione analitica dell'esercizio precedente, ottenendo quanto segue.

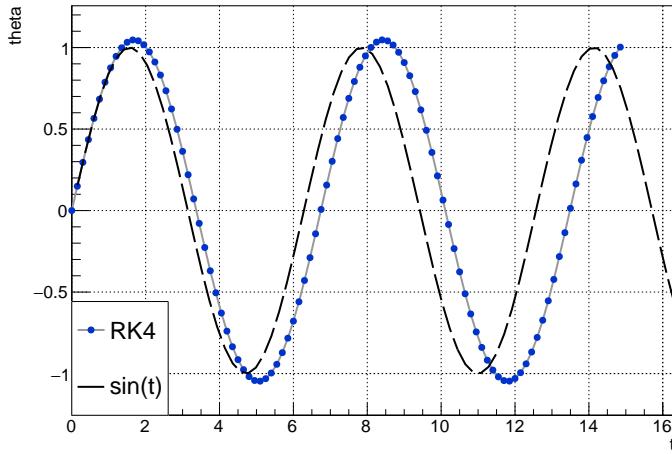


Figura 142: confronto  $\theta(t)$  con RK4 e  $\sin(t)$  per  $h = 0.15$  e  $N = 100$

Come si nota, la soluzione numerica del pendolo semplice senza approssimazione alle piccole oscillazioni si discosta sensibilmente dalla soluzione dell'oscillatore armonico precedente, ma solo a partire da un certo valore di  $t > 0$ . Per alcuni istanti iniziali, infatti, le due soluzioni appaiono visivamente sovrapponibili. Si sono quindi ripetuti i medesimi passaggi selezionando un passo più piccolo di  $h = 0.05$  e  $N = 100$ , al fine di diminuire il tempo massimo a cui giunge la soluzione, ottenendo quanto segue.

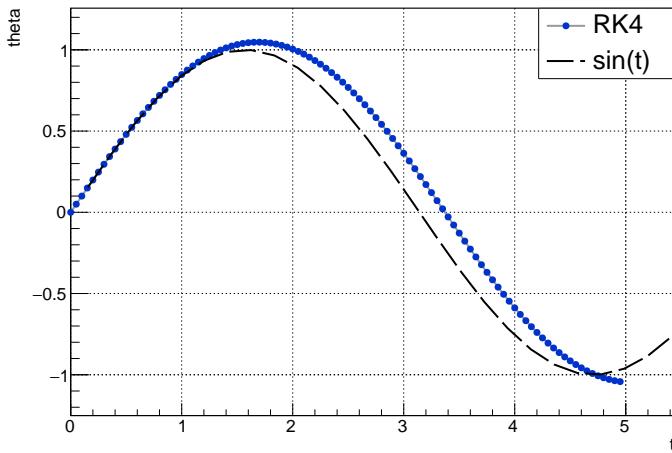


Figura 143: confronto  $\theta(t)$  con RK4 e  $\sin(t)$  per  $h = 0.05$  e  $N = 100$

Come è possibile notare, l'ingrandimento di figura 143 permette di osservare che la soluzione dell'oscillatore armonico coincide con quella numerica del pendolo matematico per  $t \in (0, 1)$ . Questo fatto può essere spiegato operando la linearizzazione del sistema dinamico che descrive il pendolo in esame intorno al punto  $\theta_0 := \theta(0)$ . In generale, come ogni sistema conservativo di tipo Newton,

il sistema dinamico in esame assume la forma del sistema di equazioni in  $\mathbb{R}^2$

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\frac{dU}{d\theta} \end{cases}$$

Sviluppando in serie di Taylor intorno a  $\theta_0$  al primo ordine si ha

$$\frac{dU}{d\theta} \approx U'(\theta_0) + U''(\theta_0)(\theta - \theta_0)$$

Tuttavia, notiamo che il potenziale si scrive come  $U(\theta) = 1 - \cos(\theta)$ . Dai dati iniziali si ha che  $\theta_0 = 0$ , da cui segue  $U'(0) = \sin(0) = 0$ . Abbiamo allora scoperto, non sorprendentemente se si pensa al caso fisico in esame, che  $\theta_0 = 0$  è un punto di equilibrio per il sistema. Scrivendo lo sviluppo avremo ora che

$$\frac{dU}{d\theta} \approx U''(\theta_0) \theta$$

Il sistema dinamico linearizzato diverrà allora

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -U''(\theta_0) \theta \end{cases} \quad \text{per } \theta \sim 0$$

la cui ODE al secondo ordine equivalente sarà

$$\ddot{\theta} = -U''(\theta_0) \theta$$

che coincide esattamente con l'equazione differenziale di un oscillatore armonico con pulsazione  $\omega = U''(\theta_0)$ . Abbiamo allora mostrato analiticamente un fatto ben noto: per angoli piccoli (piccole oscillazioni), e quindi per valori di tempo piccoli (per continuità), la soluzione dell'equazione differenziale del pendolo matematico coincide esattamente con la soluzione della ODE che descrive un oscillatore armonico. Siccome nel caso in esame si ha  $U''(\theta_0) = 1$ , allora la ODE risultante del processo di linearizzazione coincide esattamente con la ODE studiata nell'esercizio precedente, la cui soluzione è proprio  $\theta(t) = \sin(t)$ . Per tale ragione, dunque, la soluzione numerica del pendolo matematico coincide, per tempi piccoli iniziali, alla soluzione analitica di un oscillatore armonico monodimensionale. Quanto si ha avuto modo di mostrare permette di concludere che tutti i casi in cui il moto si svolge in un intorno del punto di equilibrio  $\theta = 0$  possono essere legittimamente trattati studiando un oscillatore armonico la cui soluzione, come si è mostrato nell'esercizio precedente, è nota e banale. Ovviamente, questo vale solo per soluzioni in un intorno dell'equilibrio, come suggerisce la figura 142.

Anche in questo caso è interessante studiare la curva di fase data dalla (76) per il valore di energia dato dai dati iniziali. Sostituendo il nuovo potenziale, la velocità in funzione della posizione sarà data da

$$\dot{\theta}(\theta) = \pm \sqrt{2(E_0 - (1 - \cos \theta))} \quad \text{con} \quad E_0 = \frac{1}{2}$$

ossia, più esplicitamente

$$\dot{\theta}(\theta) = \pm \sqrt{2 \cos \theta - 1}$$

Si sono allora costruite le soluzioni  $(\theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  con il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4. Tenendo conto del grafico del potenziale e della forma esplicita della curva di fase, ci si aspetta di osservare un ovale limitato centrato nello zero. Si è fissato un passo pari a  $h = 0.05$  e un valore di  $N = 270$ . Si sono quindi plottate le coppie nello stesso grafico, ottenendo i seguenti risultati.

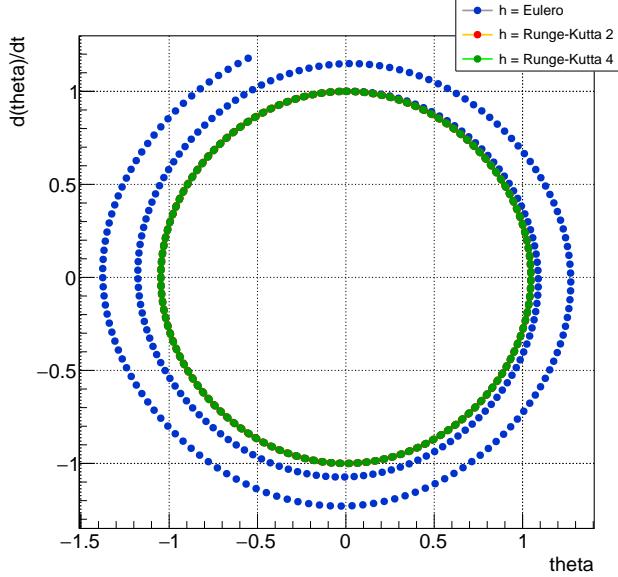


Figura 144:  $\dot{\theta}(\theta)$  con E, RK2 e RK4 per  $h = 0.05$  e  $N = 270$

Il metodo di Eulero, per  $h = 0.05$ , produce una curva di fase sensibilmente distante da quella attesa già a partire da tempi piccoli, coerentemente con i risultati ottenuti dallo studio del moto. I metodi di Runge-Kutta, invece, appaiono del tutto sovrapponibili rispetto alla scala di misure in figura e al passo selezionato. In particolare, i metodi al secondo e al quarto ordine, per il valore di  $h$  fissato, ricostruiscono una curva di fase a forma di ellisse, come ci si aspetta dalla relazione determinata grazie alla conservazione dell'energia. Possiamo quindi affermare quanto già si era intuito dallo studio della soluzione: il moto del pendolo matematico in esame è periodico e limitato, in quanto la curva di fase è costituita da un'unica componente连通的 su un compatto che si estende da un punto di elongazione minima ad uno di elongazione massimo. Al fine di verificare quantitativamente il corretto andamento della curva di fase per il valore di energia  $E_0$  è possibile notare che

$$\dot{\theta}(\theta) = \pm \sqrt{2 \cos \theta - 1} \quad \iff \quad \dot{\theta}^2(\theta) = 2 \cos \theta - 1$$

La relazione equivalente ottenuta risulta molto più gestibile a livello computazionale grazie all'assenza di segni e di radici di ordine pari. Si è quindi deciso di svolgere un fit delle coppie  $(\theta_i, \dot{\theta}_i^2)_{i=1,\dots,N}$  per ragioni di semplicità analoghe a quelle dell'esercizio precedente. Si è scelto di utilizzare i dati più precisi di Runge-Kutta 4, ponendo  $y = a \cos(\theta) + b$  come funzione interpolante, con  $a$  e  $b$  parametri liberi. Di seguito sono riportati i risultati ottenuti.

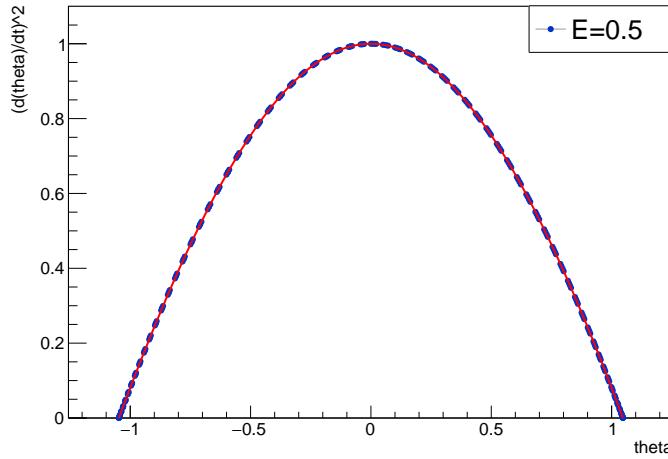


Figura 145:  $\dot{\theta}^2(\theta)$  con Runge-Kutta 4 per  $h = 0.05$  e  $N = 270$

I parametri stimati risultano essere

$$a = 2 \quad \text{e} \quad b = -1$$

Dalla consistenza delle stime ottenute con i valori attesi è possibile concludere la verifica della corretta costruzione numerica della curva di fase con RK4 per il valore del passo fissato. Anche in questo caso, allora, si ha avuto modo di verificare che l'esistenza di una costante del moto per un sistema dinamico permette di condurre uno studio più accurato e quantitativo circa la sua evoluzione, nonostante non sia nota analiticamente la sua soluzione. Risulta importante notare che l'esistenza di una costante del moto è un'eccezione e non una regola: la stragrande maggioranza dei sistemi dinamici che descrivono un fenomeno, infatti, non è caratterizzato da particolari simmetrie (e quindi costanti del moto). Vedremo che, nonostante questo, esiste comunque sempre un modo per svolgere un'analisi quantitativa circa la precisione con cui riusciamo a ricostruire numericamente la soluzione di un'equazione differenziale.

Esattamente come in precedenza, si è poi deciso di generare il campo vettoriale associato al sistema dinamico. Osservando il lato destro del sistema in esame abbiamo che, in questo caso, il campo assume la forma

$$\begin{pmatrix} \theta \\ z \end{pmatrix} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \mapsto \begin{pmatrix} z \\ -\sin(\theta) \end{pmatrix} = \begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix}$$

In particolare, sappiamo che tale mappa restituirà le tangenti alle curve di fase per ogni punto del piano delle fasi. Si è allora costruita una griglia regolare di  $N = 20$  punti per lato del tipo  $\{(\theta_i, \dot{\theta}_j)\}_{i,j=1,\dots,N^2}$  nel quadrato  $[-5, 5] \times [-5, 5]$  centrato nell'origine di un sistema cartesiano. Si è poi calcolato il lato destro del sistema ottenendo punti della forma  $\{(\dot{\theta}_j, \ddot{\theta}_j)\}_{i,j=1,\dots,N^2}$  corrispondenti alle componenti dei vettori che definiscono il campo, per poi svolgere l'usuale operazione di normalizzazione e riscalamento per ragioni di visualizzazione. Il campo vettoriale ottenuto assume l'andamento che segue.

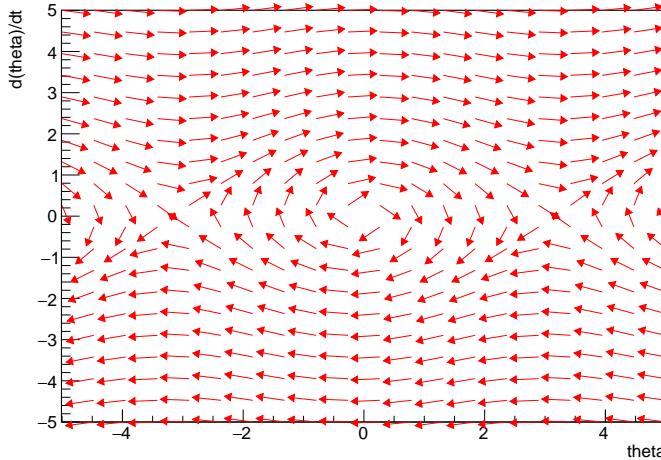


Figura 146: andamento qualitativo delle tangenti alle curve di fase

Risulta immediato notare dal grafico che, in un intorno dell'origine, le curve di fase si dispongono a formare ellissi limitate. Anche da questa verifica qualitativa è possibile concludere che i metodi di Runge-Kutta forniscono una ricostruzione della curva di fase molto più precisa rispetto al metodo di Eulero. Si noti, inoltre, che il campo vettoriale generato restituisce graficamente molte altre informazioni. Ad esempio, il comportamento delle curve di fase intorno all'origine ci fa pensare che  $\theta = 0$  sia un punto di equilibrio stabile per il sistema. Al contrario, la disposizione dei vettori in  $\theta \approx \pm 3$  mostra la presenza di due coppie di separatrici, permettendo di ipotizzare che si tratti di punti di equilibrio instabile per il sistema. D'altra parte, se si pensa al caso fisico in esame, non è difficile capire di essere in presenza di punti di equilibrio instabile per  $\theta = \pm\pi$ . In corrispondenza di tali valori della coordinata lagrangiana, infatti, il pendolo si trova in posizione verticale verso l'alto: per piccole perturbazioni il moto si allontanerà irrimediabilmente da tali punti sotto effetto della forza peso. Inoltre, seguendo le tangenti alle soluzioni nella parte alta e bassa del grafico, non è difficile convincersi che esistono valori dei dati iniziali tali che il moto risulti illimitato. Questo comportamento è molto diverso rispetto a quello che si verifica nell'oscillatore armonico, il cui moto descrive ellissi limitate per ogni valore del dato iniziale. L'interpretazione fisica di questo fatto è chiara: se al pendolo viene fornita abbastanza energia da completare almeno una rotazione, in assenza di attriti continuerà a restare in moto aumentando, ad ogni rotazione, la coordinata posizionale di un fattore  $2\pi$ . Tutti questi fatti possono essere verificati studiando qualitativamente il grafico di  $\dot{\theta}(\theta)$  al variare dell'energia e studiando la stabilità del sistema con i due teoremi di Lyapunov. Questa parte di studio verrà omessa in quanto non direttamente affine a quello che si propone di fare la presente relazione.

Dalla costruzione numerica della soluzione e dallo studio della curva di fase si ha già avuto modo di osservare la periodicità del sistema in esame. Come ultima analisi del sistema meccanico si è quindi deciso di stimare, esattamente come nell'esercizio precedente, il periodo del moto. Dalla (77) si avrà che il

periodo del moto del pendolo matematico assume la forma

$$T_{1/2} = 2 \int_{\theta_m}^{\theta_M} \frac{1}{\sqrt{2 \cos \theta - 1}} d\theta$$

I punti di inversione del moto  $\theta_m$  e  $\theta_M$  saranno quei valori angolari tali che la velocità si annulli istantaneamente, dunque tali che

$$2 \cos \theta - 1 = 0 \iff \theta = \arccos\left(\frac{1}{2}\right) = \pm \frac{\pi}{3}$$

poiché  $\theta \in (-\pi, \pi)$ . L'integrale da stimare sarà allora

$$T_{1/2} = 2 \int_{-\pi/3}^{\pi/3} \frac{1}{\sqrt{2 \cos \theta - 1}} d\theta$$

la cui soluzione non è esprimibile per mezzo di una combinazione di funzioni elementari. Risulta quindi necessario procedere computazionalmente, ricorrendo alle tecniche di calcolo numerico studiate. In particolare, tenendo conto del fatto che l'integrale in esame è monodimensionale, risulterà opportuno l'utilizzo dei metodi deterministici di integrazione. Anzitutto, si noti che l'integranda presenta due singolarità agli estremi del dominio, infatti

$$\lim_{\theta \rightarrow \pm \frac{\pi}{3}} \frac{1}{\sqrt{2 \cos \theta - 1}} = +\infty$$

ossia le rette  $\theta = \pm \pi/3$  risultano asintoti verticali per la funzione. Sarà allora necessario l'utilizzo delle formule aperte in un intorno delle singolarità. Si è quindi scritto l'intervallo di integrazione come

$$\left(-\frac{\pi}{3}, \frac{\pi}{3}\right) = \left(-\frac{\pi}{3}, -\frac{\pi}{3} + \epsilon\right) \cup \left[-\frac{\pi}{3} + \epsilon, \frac{\pi}{3} - \epsilon\right] \cup \left(\frac{\pi}{3} - \epsilon, \frac{\pi}{3}\right)$$

Il periodo si scriverà, allora, come

$$\begin{aligned} T_{1/2} &= \int_{-\pi/3}^{-\pi/3+\epsilon} \frac{2}{\sqrt{2 \cos \theta - 1}} d\theta + \int_{-\pi/3+\epsilon}^{\pi/3-\epsilon} \frac{2}{\sqrt{2 \cos \theta - 1}} d\theta + \\ &\quad + \int_{\pi/3-\epsilon}^{\pi/3} \frac{2}{\sqrt{2 \cos \theta - 1}} d\theta \end{aligned}$$

Si è poi applicata la più precisa formula aperta per  $N = 6$  punti agli intervalli contenenti le singolarità, e la formula di Romberg per  $J = K = 10$  nell'intervallo centrale senza punti singolari. Al fine di verificare quale fosse il valore di  $\epsilon$  che consentisse una stima migliore del periodo, ossia un utilizzo ottimale della formula aperta, si sono calcolate le dispersioni dal valore vero

$$\Delta(\epsilon) = \left| \tilde{T}_{1/2}(\epsilon) - T_{\text{true}} \right|$$

nel range di valori

$$0.00005 \leq \epsilon < 0.001 \quad \text{con} \quad \epsilon_{i+1} = \epsilon_i + 0.000015$$

dove  $\tilde{T}_{1/2}$  rappresenta la stima del periodo con il metodo appena discusso. Il valore vero  $T_{\text{true}}$ , invece, si è ottenuto facendo svolgere numericamente l'integrale in esame ad un programma di calcolo avanzato. Si sono quindi plottati i valori della dispersione in funzione dei corrispondenti valori di  $\epsilon$ . Di seguito sono mostrati i risultati ottenuti.

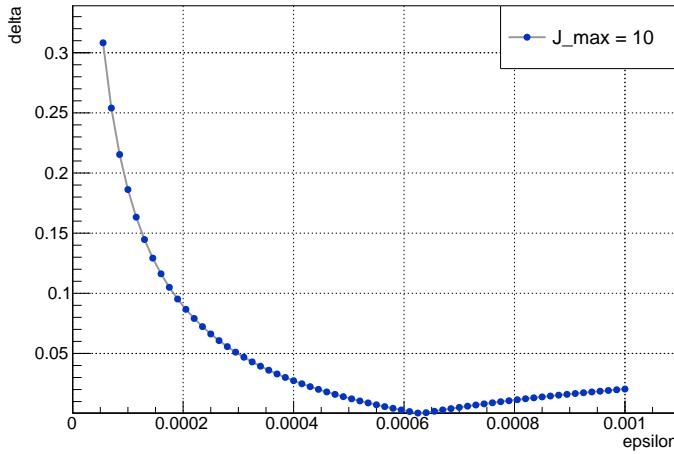


Figura 147:  $\Delta(\epsilon)$  con formula aperta per  $N = 6$  e Romberg per  $J_{max} = 10$

Come è possibile notare dal plot, la stima del periodo con la tecnica utilizzata raggiunge un valore prossimo al valore vero (con dispersione nulla) per un valore di  $\epsilon \approx 0.00064$ . Per tale valore di  $\epsilon$  si ottiene

$$T(0.00064)_{1/2} \approx 6.74 > 6.28 \approx 2\pi$$

Siamo quindi riusciti a mostrare numericamente che il periodo di un pendolo matematico è poco maggiore del periodo di un pendolo con approssimazione alle piccole oscillazioni, coerentemente con quanto si osserva in figura 142.

### Pendolo con attrito

Se il pendolo semplice è soggetto anche ad una forza di attrito di natura viscosa, come nel caso dell'attrito dell'aria, si avrà che

$$g(t, \dot{\theta}) = -\gamma \dot{\theta} \quad \text{con} \quad \gamma > 0$$

Il problema di Cauchy in esame si riduce allora ad un problema della forma

$$\ddot{\theta} = -\sin(\theta) - \gamma \dot{\theta} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 1 \end{cases}$$

Dall'equivalenza (0.16) segue che l'equazione al secondo ordine equivale allo studio del sistema dinamico in  $\mathbb{R}^2$  ai dati iniziali

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\sin(\theta) - \gamma \dot{\theta} \end{cases} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ z(0) = 1 \end{cases}$$

al quale risulta ora possibile applicare le relazioni ricorsive date dai tre metodi. Anche in questo caso, come prima cosa, si è deciso di calcolare e plottare le coppie  $(t_i, \theta_i)_{i=1,\dots,N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.15$  e con  $N = 100$ , con coefficiente di attrito  $\gamma = 0.4$ , come segue.

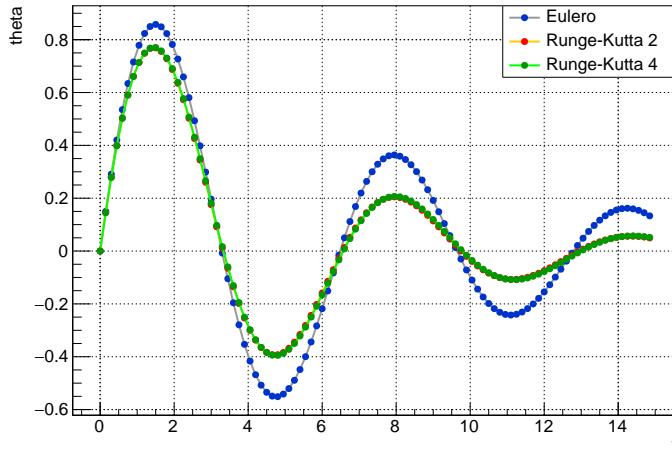


Figura 148: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.15$  e  $N = 100$ ,  $\gamma = 0.4$

Come è possibile notare, il metodo di Eulero si discosta in modo significativo dal metodo di Runge-Kutta 2 e 4 già a partire da tempi piccoli, coerentemente con il fatto che risulta essere un metodo al primo ordine in  $h$ . Nella scala del grafico riportato, invece, i metodi Runge-Kutta appaiono, di fatto, sovrapponibili. Si è allora deciso di confrontare soltanto il metodo al secondo e al quarto ordine, aumentando il passo di integrazione a  $h = 0.5$  e diminuendo il numero di punti generati a  $N = 50$  come in precedenza, ottenendo quanto segue.

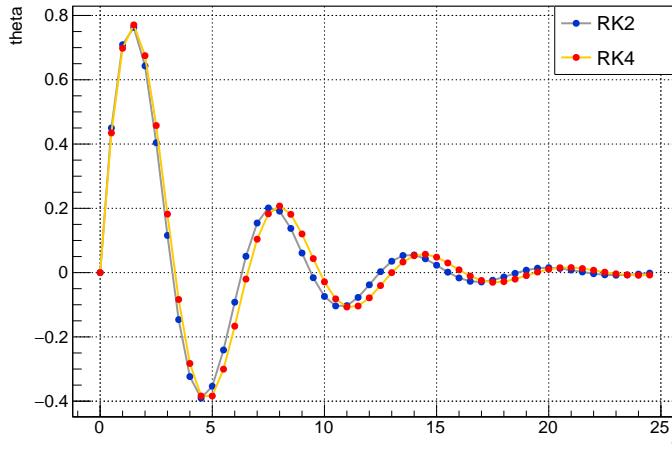


Figura 149: confronto  $\theta(t)$  con RK2 e RK4 per  $h = 0.5$  e  $N = 50$ ,  $\gamma = 0.4$

Ad un passo più grande è possibile osservare qualitativamente quanto ci si aspetta: Runge-Kutta 2 risulta discostarsi, all'aumentare del tempo, sempre di più rispetto al corrispondente metodo al quarto ordine. Si è quindi eseguita la medesima operazione calcolando e piazzando  $(t_i, \dot{\theta}_i)_{i=1, \dots, N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.15$  e con  $N = 100$ , con lo stesso va-

lore del coefficiente  $\gamma$  di attrito. I grafici della velocità in funzione del tempo risultano avere l'andamento che segue.

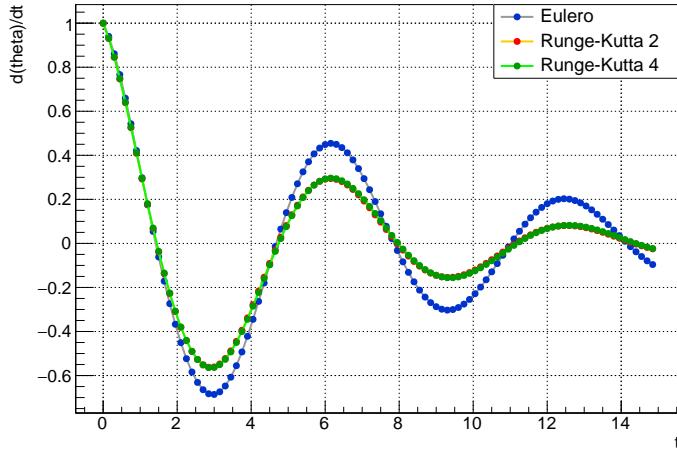


Figura 150: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.15$  e  $N = 100$ ,  $\gamma = 0.4$

Anche nel caso delle velocità, il metodo di Eulero si discosta in modo significativo dal metodo di Runge-Kutta 2 e 4 già a partire da tempi piccoli. I due metodi di ordine 2 e 4, invece, appaiono del tutto sovrapponibili nella scala della figura precedente. Si è allora deciso di confrontare soltanto il metodo al secondo e al quarto ordine, aumentando il passo di integrazione a  $h = 0.5$  e diminuendo il numero di punti generati a  $N = 50$ , esattamente come in precedenza, ottenendo quanto segue.

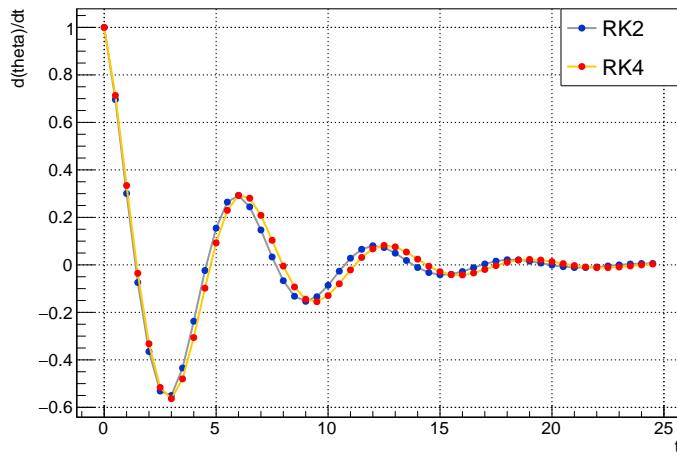


Figura 151: confronto  $\dot{\theta}(t)$  con RK2 e RK4 per  $h = 0.5$  e  $N = 50$ ,  $\gamma = 0.4$

Coerentemente con il fatto che i due metodi differiscono di due ordini in precisione, si osserva che la curva che descrive la velocità in Runge-Kutta 2 inizia a discostarsi dalla velocità in Runge-Kutta 4 a partire da un certo valore tem-

porale. Questi primi dati confermano qualitativamente quanto ci si aspetta dai tre metodi numerici studiati. Si noti ora che il sistema dinamico che descrive il pendolo matematico con attrito viscoso risulta essere autonomo, ma non più posizionale. Il termine  $g$  che è stato aggiunto nella ODE, infatti, dipende esplicitamente dalla velocità in modo lineare. Da questo fatto segue che non è più verificato il teorema di conservazione dell'energia meccanica. Considerando che la forza  $g$  non è conservativa, si potrebbe tentare di utilizzare il più generale teorema di variazione dell'energia meccanica

$$\frac{dE}{dt} = W_\gamma$$

al fine di ottenere qualche appoggio analitico per lo studio del sistema. Tuttavia, è facile verificare che il lavoro della forza non conservativa lungo la traiettoria

$$W_\gamma = \int_{\omega} \vec{F}_\gamma \cdot d\vec{s} \quad \text{con} \quad \vec{F}_\gamma = -\gamma \dot{\theta}$$

non è esprimibile direttamente mediante funzioni elementari, in quanto non è nota l'espressione analitica di  $\dot{\theta}(t)$ . Non disponendo né della soluzione analitica, né di una costante del moto, non è quindi possibile svolgere un confronto quantitativo in precisione dei tre metodi in esame con i metodi utilizzati fino a questo punto. Si potrebbe pensare di assumere la soluzione data da RK4 ad un passo molto piccolo come soluzione esatta, per poi operare un confronto come nell'esercizio precedente. Tuttavia, anche la soluzione data da RK4 è nota a meno di un errore che sappiamo quantificare, per quanto piccolo. Fissato uno dei tre metodi, in assenza di costanti del moto, è comunque possibile operare un'analisi in precisione confrontando, a tempo  $\bar{t}$  fissato, la soluzione  $\theta_h$  ricostruita ad un passo  $h$  con la soluzione  $\theta_{h/2}$  ricostruita ad un passo  $h/2$ . Evidentemente, per la relazione (64), questo equivale a confrontare una soluzione costruita per un numero di passi  $N$  con una costruita con un numero di passi  $2N$ . Fissato un metodo, si sono quindi costruite diverse coppie di soluzioni a passo dimezzato l'una rispetto all'altra fino a  $\bar{t} = 5$  nel range

$$200 \leq N < 2000 \quad \text{con} \quad N_{i+1} = N_i + 50$$

Si sono poi calcolate le distanze (in metrica euclidea)

$$\Delta(\bar{t}) = |\theta_N(\bar{t}) - \theta_{2N}(\bar{t})|$$

per ogni  $N$  nel range selezionato. In particolare, visti i risultati ottenuti nella prefazione teorica, ci si aspetta che l'errore scali come la (66), la (69) o la (72) a seconda del metodo selezionato. Siccome l'errore scala come una potenza di  $h$ , l'analisi operata, riducendo di volta in volta il passo, consentirà di determinare il valore di  $\bar{h}$  tale che la soluzione venga ricostruita, fino al tempo  $\bar{t}$ , con una precisione più piccola di una precisione  $\varepsilon$  desiderata. Cercheremo allora il primo valore del passo  $h$  di integrazione tale che

$$\Delta(\bar{t}) < \varepsilon \tag{78}$$

Ovviamente, esiste un certo grado di incertezza in questo tipo di analisi, in quanto implicitamente si sta supponendo che la soluzione ricostruita con un

passo dimezzato non si discosti in modo significativo da quella ricostruita ad un passo  $h$ : nel caso di sistemi non caotici questo fatto è sempre vero per definizione. Anzitutto, al fine di verificare il corretto andamento, si sono eseguiti i procedimenti descritti con i tre metodi, calcolando i logaritmi delle dispersioni e dei passi corrispondenti. Si sono poi interpolati i dati con una retta della forma  $y = mx + q$ , ottenendo i seguenti risultati.

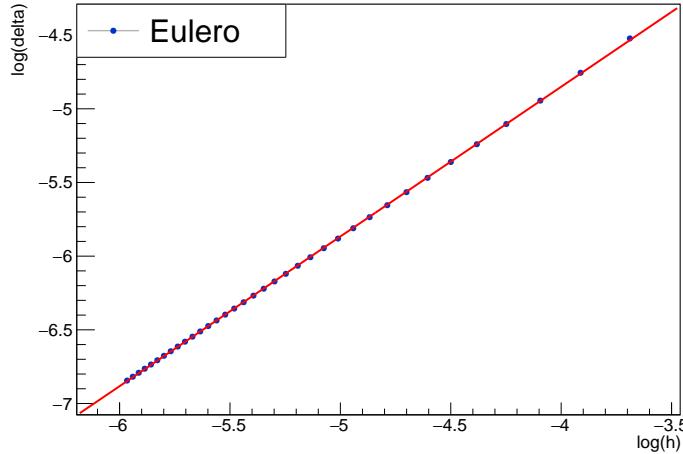


Figura 152: andamento  $\log \Delta(\bar{t})$  con Eulero: fit

Si è ottenuta la stima di parametri che segue.

$$q = -0.786 \quad \text{e} \quad m = 1.02 \approx 1$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Eulero risulta essere al primo ordine in  $h$ . Per il metodo di Runge-Kutta 2 si è ottenuto quanto segue.

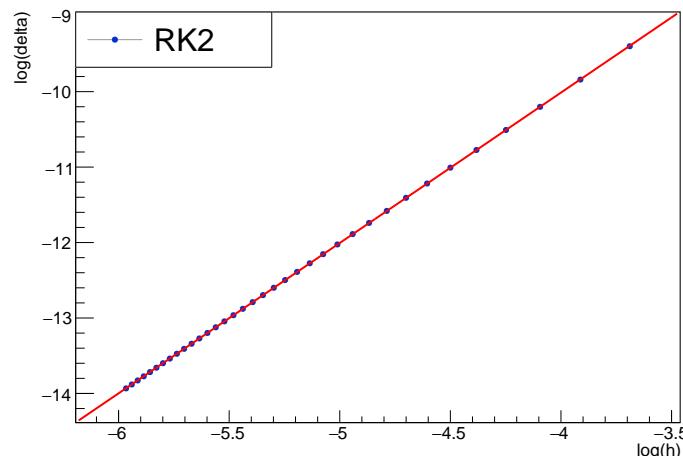


Figura 153: andamento  $\log \Delta(\bar{t})$  con RK2: fit

Si è ottenuta la stima di parametri che segue.

$$q = -2.04 \quad \text{e} \quad m = 1.99 \approx 2$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 2 risulta essere al secondo ordine in  $h$ . Anche per il metodo di Runge-Kutta 4 i risultati sono riportati di seguito.

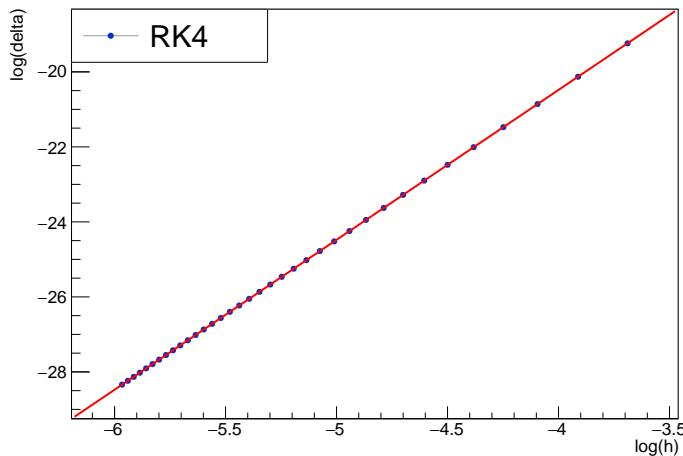


Figura 154: andamento  $\log \Delta(\bar{t})$  con RK4: fit

Si è ottenuta la stima di parametri che segue.

$$q = -4.49 \quad \text{e} \quad m = 4$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 4 risulta essere al quarto ordine in  $h$ . Verificato il corretto andamento in tutti i casi, si sono quindi ripetute le medesime operazioni in un range di  $N$  molto più grande al fine di trovare il primo valore del passo critico  $\bar{h}$  che verificasse la condizione di precisione (78). In particolare, si è scelto arbitrariamente di fissare un valore  $\varepsilon = 10^{-5}$ . Di seguito sono riportati i risultati ottenuti.

	$\bar{h}$
Eulero	$2.4 \cdot 10^{-5}$
Runge-Kutta 2	0.0086
Runge-Kutta 4	0.17

Per i valori del passo critico riportati in tabella, dunque, i metodi in esame ricostruiscono la soluzione al problema di Cauchy fino al tempo  $\bar{t} = 5$  con un errore massimo di  $\varepsilon = 10^{-5}$ . Questo tipo di analisi è sempre possibile, anche quando non si dispone di quantità conservative per il sistema, in quanto si fonda soltanto sulla convergenza della soluzione numerica al diminuire del passo. Dovremo però fare particolare attenzione davanti a sistemi dinamici caotici, per i quali l'analisi in precisione appena effettuata può, in certi casi, perdere di significato.

Valutato il problema in esame da un punto di vista numerico, siamo ora interessati allo studio del moto del pendolo con attrito viscoso al variare del coefficiente di attrito  $\gamma \in (0, 2)$  da un punto di vista strettamente fisico, operando un confronto con alcuni casi analoghi noti validi per un oscillatore armonico. In particolare, basandosi sullo studio dell'errore appena effettuato, si sono costruite tre soluzioni con RK4 fissando  $h = 0.05$  e  $N = 600$ , al fine di ottenere soluzioni numeriche sufficientemente precise fino al tempo finale in esame. Posto  $\gamma = 0.5$  si è ottenuta la soluzione che segue.

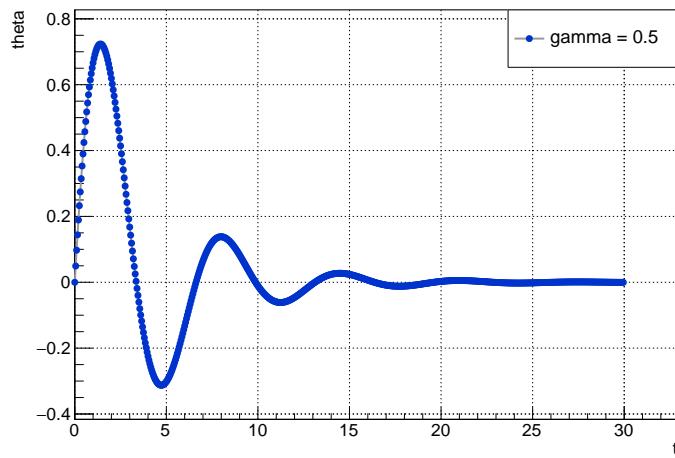


Figura 155: plot  $\theta(t)$  con RK4 per  $h = 0.05$  e  $N = 600$ ,  $\gamma = 0.5$

Come si nota, per questo valore di  $\gamma$ , il punto materiale compie alcune oscillazioni con ampiezza decrescente nel tempo fino a stabilizzarsi, per  $t \rightarrow +\infty$ , alla coordinata costante  $\theta = 0$ . Per  $\gamma = 1$  si è ottenuto quanto segue.

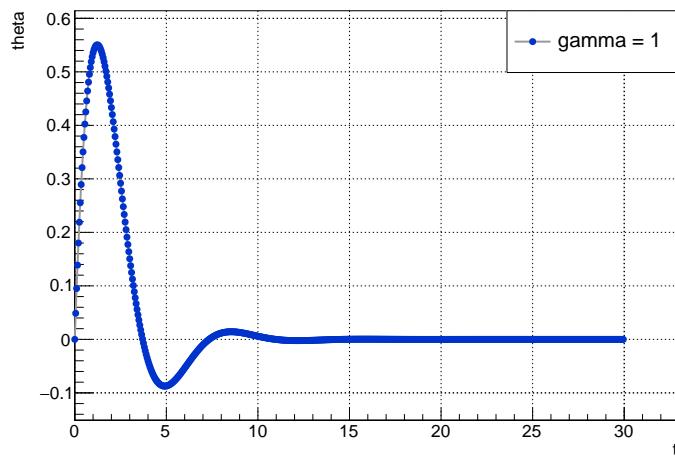


Figura 156: plot  $\theta(t)$  con RK4 per  $h = 0.05$  e  $N = 600$ ,  $\gamma = 1$

In questo caso, il coefficiente di attrito è più elevato e il punto riesce a compiere soltanto 3 oscillazioni, sempre meno marcate, prima di arrestarsi alla coordinata generalizzata nulla. Per  $\gamma = 1.8$  si è ottenuto il seguente risultato.

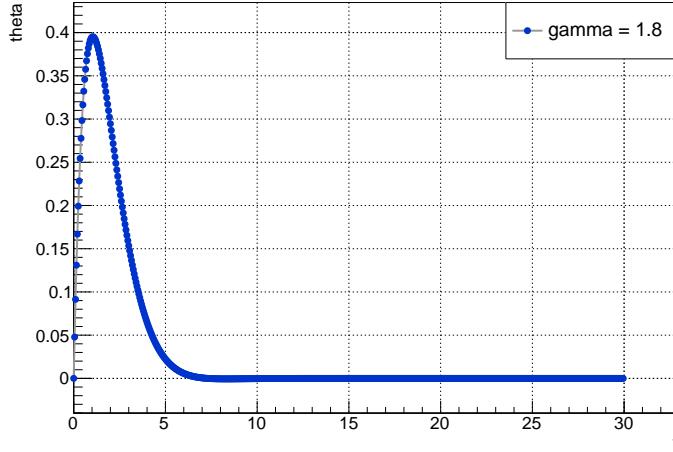


Figura 157: plot  $\theta(t)$  con RK4 per  $h = 0.05$  e  $N = 600$ ,  $\gamma = 1.8$

In questo caso, invece, il coefficiente di attrito è così elevato da non permettere al punto materiale più di un'oscillazione: dopo un solo massimo di posizione il punto si stabilizza a  $\theta = 0$  per tutti i valori temporali. I risultati ottenuti sono perfettamente in linea con l'intuizione: a differenza dei casi dell'oscillatore armonico e del pendolo matematico studiati in precedenza, nei quali il moto non decresce nel tempo vista l'assenza di forze dissipative, nel pendolo con attrito l'ampiezza delle oscillazioni decresce tanto più quanto più è grande il coefficiente  $\gamma$  di proporzionalità. L'andamento della velocità nel tempo avrà un andamento per certi versi simmetrico a quello della posizione, considerando che differisce da quest'ultima soltanto di un operatore di derivazione. Risulta interessante notare che i risultati ottenuti sono simili ai risultati che si ottengono per un oscillatore armonico smorzato. Anche in questo caso, qualitativamente, sembrano configurarsi i casi di sovra-smorzamento e sotto-smorzamento al variare di  $\gamma$ . Tuttavia, è importante ricordare che, nel caso dell'oscillatore armonico smorzato, le diverse soluzioni che portano ai casi di sovra e sotto smorzamento possono essere determinate analiticamente studiando la soluzione di una ODE lineare al secondo ordine. In questo caso, senza metodi numerici, non è possibile predire nessuno di questi comportamenti per via analitica.

Come se non bastasse, nei casi in cui la forza non dipende solo dalla posizione come quello in esame, non è possibile avere l'espressione analitica della curva di fase, in quanto la (76) deriva direttamente dalla conservazione dell'energia. Tutto ciò che rimane da fare è allora operare un confronto qualitativo delle curve di fase costruite numericamente con il campo vettoriale delle tangenti alla soluzione. Si sono quindi costruite le soluzioni  $(\theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  con il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4, fissando  $h = 0.05$  e  $N = 270$ , con una costante di attrito  $\gamma = 0.5$ . Di seguito sono riportati i risultati ottenuti a confronto nello stesso grafico.

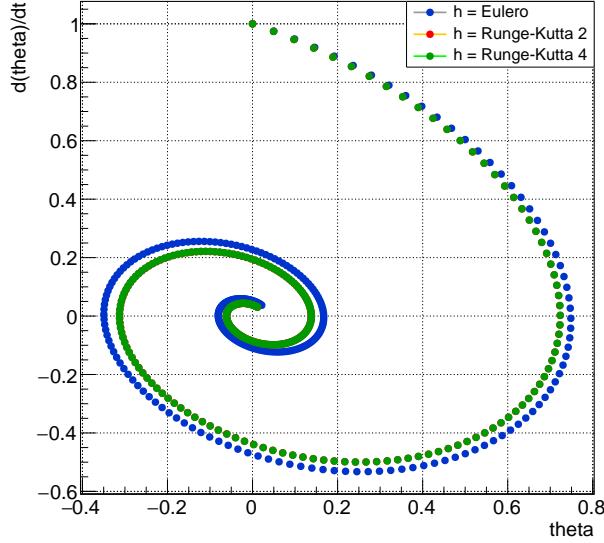


Figura 158: confronto  $\dot{\theta}(\theta)$  con E, RK2 e RK4 per  $h = 0.05$  e  $N = 270$ ,  $\gamma = 0.5$

La curva di fase ottenuta presenta una forma a spirale, avvolta intorno allo zero degli assi del piano delle fasi. L'andamento asintotico verso lo zero è chiaramente consistente con quanto abbiamo già notato: in presenza di attrito il pendolo raggiunge la posizione  $\theta = 0$  per tempi grandi. Siccome per definizione

$$\dot{\theta} = \frac{d\theta}{dt}$$

allora, asintoticamente dovrà valere la relazione

$$\lim_{t \rightarrow +\infty} \dot{\theta}(t) = 0$$

poiché  $\theta(t) = 0$  per tempi sufficientemente grandi in presenza di attrito. Quanto si osserva in figura 158, dunque, è consistente con quanto ci si aspetta. Come al solito, la soluzione data dal metodo di Eulero diverge rapidamente dalle altre due soluzioni più precise che, invece, appaiono del tutto sovrapponibili. Nel caso in esame, il campo vettoriale associato al sistema dinamico avrà la forma

$$\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \mapsto \begin{pmatrix} \dot{\theta} \\ -\sin(\theta) - \gamma\dot{\theta} \end{pmatrix}$$

e restituirà le tangenti alle curve di fase per ogni punto del piano delle fasi. Si è allora costruita una griglia regolare di  $N = 20$  punti per lato del tipo  $\{(\theta_i, \dot{\theta}_j)\}_{i,j=1,\dots,N^2}$  nel quadrato  $[-5, 5] \times [-5, 5]$  centrato nell'origine di un sistema cartesiano. Si è poi calcolato il lato destro del sistema ottenendo punti della forma  $\{(\dot{\theta}_j, \ddot{\theta}_j)\}_{i,j=1,\dots,N^2}$  corrispondenti alle componenti dei vettori che definiscono il campo, per poi svolgere la solita operazione di normalizzazione e riscalamento per ragioni di visualizzazione. Il campo vettoriale ottenuto per un coefficiente di attrito  $\gamma = 0.5$  ha il seguente andamento.

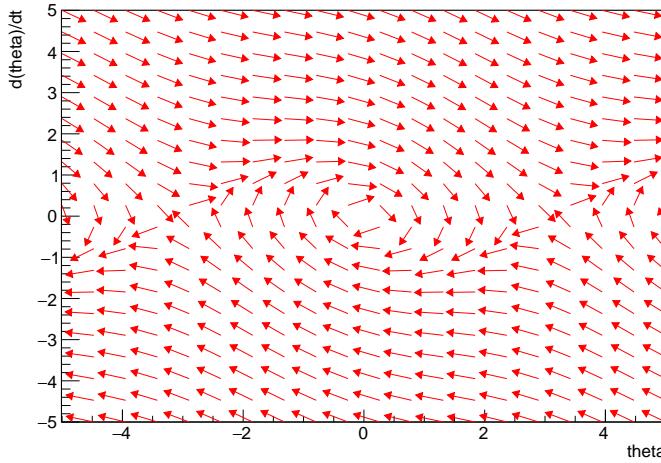


Figura 159: andamento qualitativo delle tangenti alle curve di fase,  $\gamma = 0.5$

Come è possibile notare, il grafico del campo vettoriale rende conto qualitativamente dell'andamento a spirale ottenuto. Anche in questo caso è possibile osservare due punti di equilibrio instabile e uno di equilibrio stabile centrale, coerentemente con l'intuizione. Tuttavia, rispetto al campo ottenuto nel caso del pendolo in vuoto, le tangenti alle curve di fase appaiono disporsi in modo meno simmetrico rispetto all'asse orizzontale. Si sono quindi costruite altre curve di fase con RK4 per diversi valori di  $\gamma$ , con  $h = 0.05$ . Il numero di punti  $N$  è stato deciso di volta in volta, al fine di riuscire sempre a costruire una soluzione sufficientemente estesa da raggiungere asintoticamente lo zero degli assi del piano delle fasi. Per  $\gamma = 0.4$  si è ottenuto quanto segue.

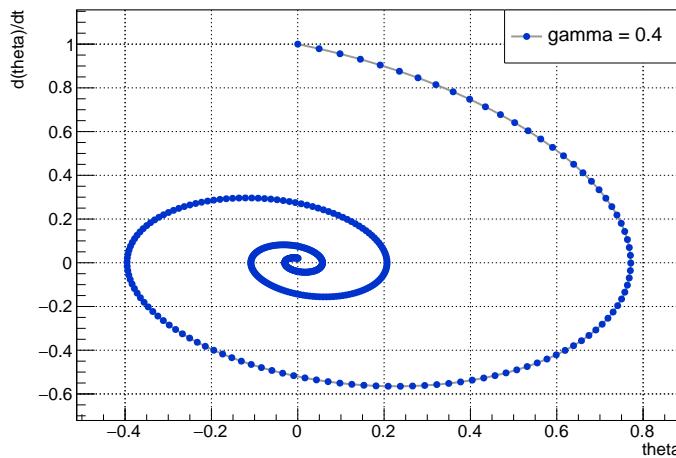


Figura 160:  $\dot{\theta}(\theta)$  con RK4 per  $\gamma = 0.4$

Per un valore di  $\gamma$  sufficientemente piccolo, il moto compie diverse oscillazioni prima di arrestarsi, come si ha già avuto modo di notare dall'andamento delle

soluzioni. Il numero di punti necessario per raggiungere il punto di equilibrio stabile risulta, infatti, sensibilmente più grande rispetto ai due casi che seguono. Per  $\gamma = 1$  la curva di fase assume l'andamento che segue.

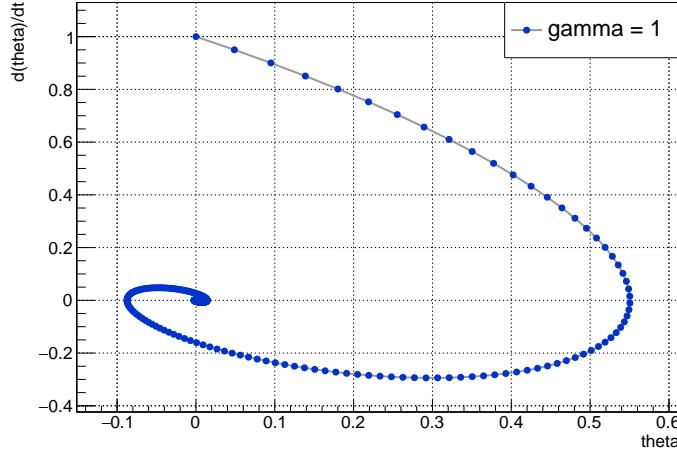


Figura 161:  $\dot{\theta}(\theta)$  con RK4 per  $\gamma = 1$

In questo caso, con  $\gamma$  più grande, il moto tende più rapidamente a raggiungere la posizione di equilibrio asintotica, come ci si aspetta. Per  $\gamma = 1.5$  la curva di fase assume l'andamento che segue.

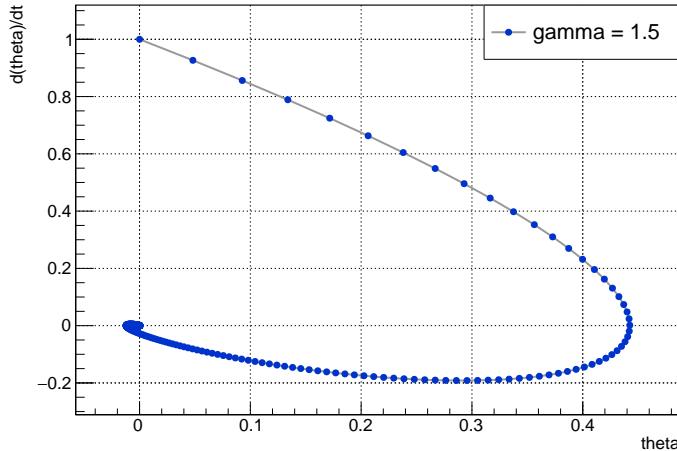


Figura 162:  $\dot{\theta}(\theta)$  con RK4 per  $\gamma = 1.5$

Ancora una volta, si ha un sensibile aumento di rapidità con la quale il moto tende ad arrestarsi per un valore del coefficiente di attrito maggiore. Si può mostrare, costruendo i campi vettoriali associati ai valori di  $\gamma$  selezionati, che le curve di fase ottenute sono consistenti con il grafico delle tangenti alle stesse in un intorno del punto di equilibrio asintotico. Infine, si noti che l'asimmetria delle tangenti in figura 159 mostra chiaramente che il campo vettoriale tende

ad attrarre ogni curva di fase nel punto stabile  $\theta = 0$ , indipendentemente dalla coppia di dati iniziali selezionata. In effetti, la forza dissipativa di attrito tenderà asintoticamente ad arrestare il moto del pendolo.

### Pendolo con attrito e forzante periodica

Se il pendolo semplice, oltre ad essere soggetto ad una forza di attrito viscoso, risulta soggetto ad una forzante periodica di pulsazione  $\omega = 2/3$  si avrà

$$g(t, \dot{\theta}) = -\gamma \dot{\theta} + A \sin(\omega t) \quad \text{con} \quad \gamma, A > 0$$

Il problema di Cauchy in esame si riduce allora ad un problema della forma

$$\ddot{\theta} = -\sin(\theta) - \gamma \dot{\theta} + A \sin(\omega t) \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 1 \end{cases}$$

Dall'equivalenza (0.16) segue che l'equazione differenziale al secondo ordine equivale allo studio del sistema dinamico in  $\mathbb{R}^2$

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\sin(\theta) - \gamma \dot{\theta} + A \sin(\omega t) \end{cases} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ z(0) = 1 \end{cases}$$

al quale risulta ora possibile applicare le relazioni ricorsive date dai tre metodi. Anzitutto, come al solito, si è deciso di calcolare e plottare le coppie  $(t_i, \theta_i)_{i=1,\dots,N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.2$  e  $N = 100$ , con coefficiente  $\gamma = 1.5$  e ampiezza della forzante  $A = 0.2$ .

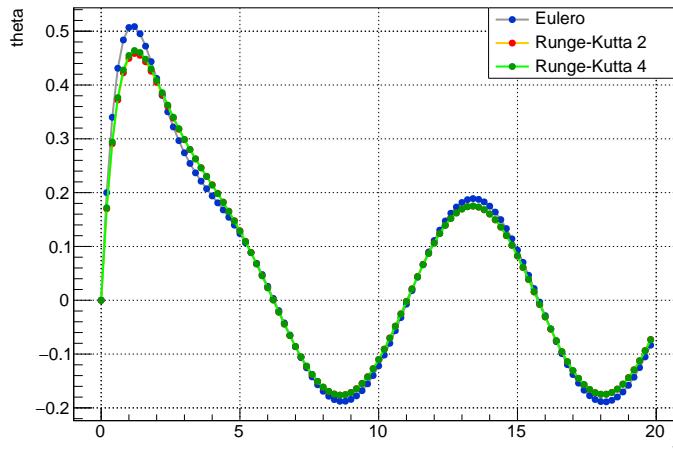


Figura 163: confronto  $\theta(t)$  per  $h = 0.2$  e  $N = 100$ ,  $\gamma = 1.5$  e  $A = 0.2$

Come è possibile notare, piuttosto sorprendentemente, a meno dell'andamento ai primi istanti temporali, il metodo di Eulero appare non discordarsi in modo significativo dai due metodi di Runge-Kutta. L'andamento qualitativo relativamente preciso si ha, infatti, per valori del passo che, nei precedenti studi, hanno sempre generato un comportamento poco preciso da parte del metodo al primo ordine. Si è allora provato ad eseguire la medesima operazione fissando  $h = 0.3$  e  $N = 150$ . Di seguito sono riportati i risultati ottenuti.

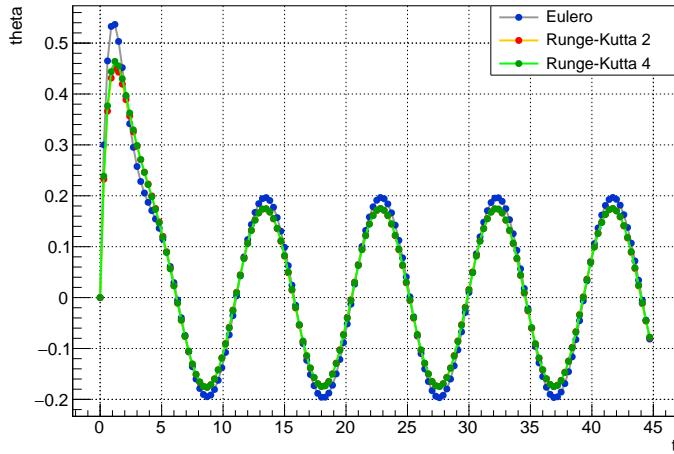


Figura 164: confronto  $\theta(t)$  per  $h = 0.3$  e  $N = 150$ ,  $\gamma = 1.5$  e  $A = 0.2$

Nonostante l'aumento del passo e del numero di punti, il metodo di Eulero continua a non distare in modo significativo dai Runge-Kutta, ma solo definitivamente. I tre metodi producono, invece, soluzioni molto diverse in un intorno dei primi valori temporali anche in questo caso. Si è quindi eseguita la medesima operazione calcolando e plottando  $(t_i, \dot{\theta}_i)_{i=1, \dots, N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.2$  e con  $N = 100$ , per gli stessi valori dei parametri  $\gamma$  e di  $A$  utilizzati in precedenza. I grafici della velocità in funzione del tempo risultano avere l'andamento che segue.

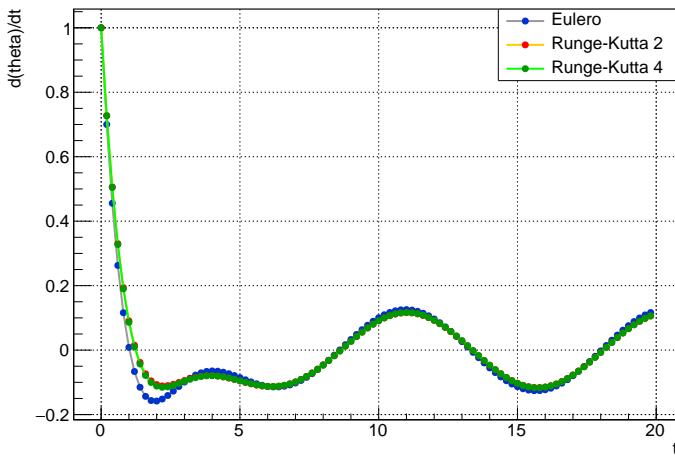


Figura 165: confronto  $\dot{\theta}(t)$  per  $h = 0.2$  e  $N = 100$ ,  $\gamma = 1.5$  e  $A = 0.2$

Anche in questo caso è possibile notare un andamento preciso del metodo di Eulero a partire da un certo tempo e un andamento più distante dai metodi di Runge-Kutta per valori di tempo iniziali. Per un passo  $h = 0.3$  e un numero di punti  $N = 150$ , invece, si è ottenuto quanto segue.

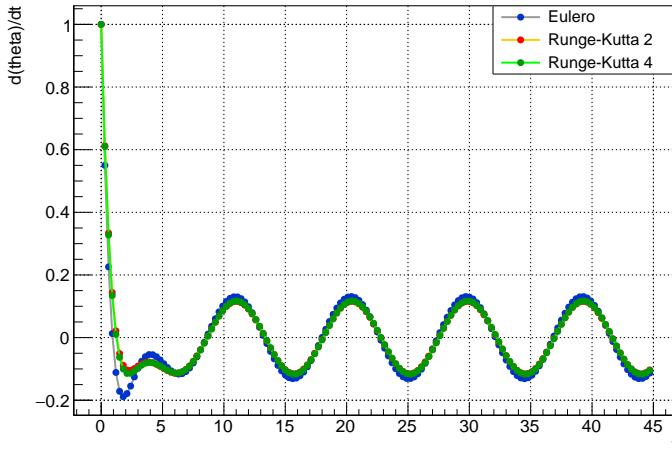


Figura 166: confronto  $\dot{\theta}(t)$  per  $h = 0.3$  e  $N = 150$ ,  $\gamma = 1.5$  e  $A = 0.2$

Visti i risultati ottenuti è possibile concludere che il metodo di Eulero, a partire da valori del passo di circa  $h = 0.1$ , risulta particolarmente preciso per tempi distanti dal tempo iniziale, ossia a partire da  $t$  tale che il moto inizi ad assumere un andamento periodico e regolare. Al contrario, i metodi risultano particolarmente sensibili in corrispondenza degli intervalli temporali in cui la soluzione varia bruscamente in modo irregolare: in un intorno destro del tempo zero. Se si vuole ottenere una soluzione precisa in questo intorno sarà allora necessario lavorare con RK4 ad un passo sufficientemente piccolo. Chiaramente, questi risultati valgono e dipendono unicamente della particolare forma analitica della ODE che descrive un pendolo smorzato e forzato, e non sono quindi generalizzabili, come si ha avuto modo di osservare negli studi precedenti.

Una volta discussa la sensibilità numerica del sistema si è deciso, anche in questo caso, di analizzare qualitativamente il comportamento fisico del problema meccanico in esame. In particolare, si sono generate diverse soluzioni con il più preciso metodo di Runge-Kutta 4, fissando un passo  $h = 0.05$  e un numero  $N$  di punti sufficientemente grande da apprezzarne l'andamento asintotico. Come prima cosa, siamo interessati a verificare come il moto  $\theta(t)$  cambi al variare dei parametri  $\gamma, A \in (0, 2)$ . A differenza del caso precedente, in cui il parametro era uno solo, lo studio dell'equazione parametrica comporta, in questo caso, un aumento notevole di difficoltà, in quanto l'aggiunta di un parametro indipendente determina l'aumento di un grado di libertà, generando molte combinazioni possibili di  $\gamma$  e di  $A$ . Costruendo diverse soluzioni al variare dei parametri, si è notato che per valori di  $\gamma$  e  $A$  tali che

$$\gamma \in (0, 2) \quad \text{e} \quad A < 0.8$$

vale spesso che

$$\exists \bar{t} \in \mathbb{R} \quad \text{tale che} \quad \theta(t) = \theta(t + T) \quad \forall t > \bar{t}$$

dove  $T$  è il periodo del moto. In altre parole, per tutte le combinazioni possibili dei parametri all'interno dei range individuati, esiste spesso un tempo tale che la

soluzione inizi ad assumere un andamento periodico e regolare per tutti i valori temporali più grandi del valore critico  $t = \bar{t}$ . Ad esempio, per  $\gamma = 0.6$  e  $A = 0.3$  si ha il seguente andamento.

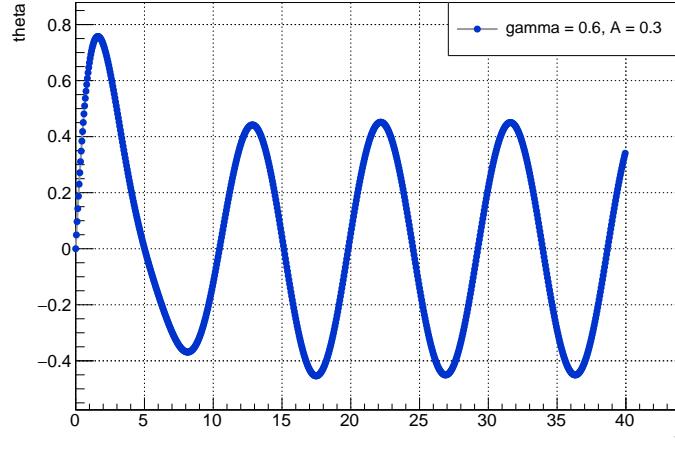


Figura 167: plot  $\theta(t)$  con RK4 per  $\gamma = 0.6$  e  $A = 0.3$

Come si nota, a meno di un intorno del tempo iniziale, la soluzione inizia ad assumere un andamento sinusoidale periodico a partire da un certo istante di tempo, coerentemente con quanto appena affermato. Per  $\gamma = 0.4$  e  $A = 0.7$  si ha, invece, quanto segue.

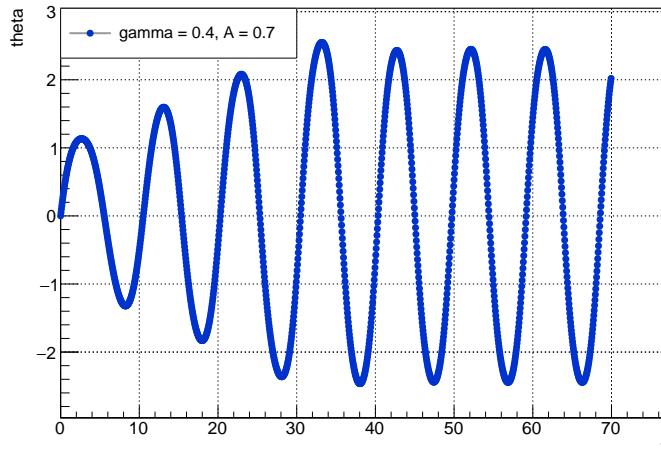


Figura 168: plot  $\theta(t)$  con RK4 per  $\gamma = 0.4$  e  $A = 0.7$

Anche in questo caso è possibile apprezzare un comportamento simile. Si noti, tuttavia, che l'ampiezza delle oscillazioni nel tratto regolare della soluzione differisce di molto rispetto all'ampiezza della soluzione precedente. Inoltre, a differenza del caso precedente, il tempo necessario per apprezzare la regolarità della soluzione è maggiore. Per  $\gamma = 1.8$  e  $A = 0.6$  si ha il seguente andamento.

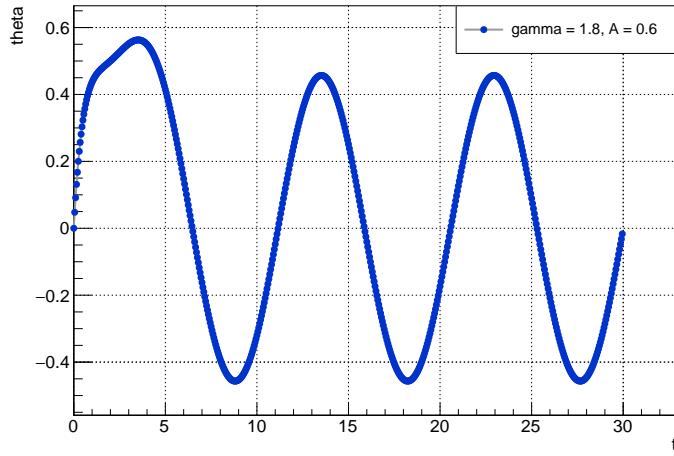


Figura 169: plot  $\theta(t)$  con RK4 per  $\gamma = 1.8$  e  $A = 0.6$

In questo caso, la soluzione inizia ad avere un comportamento regolare a partire da un tempo molto piccolo. Per ultimo, si è generata la soluzione per  $\gamma = 0.05$  e  $A = 0.05$ , ottenendo quanto segue.

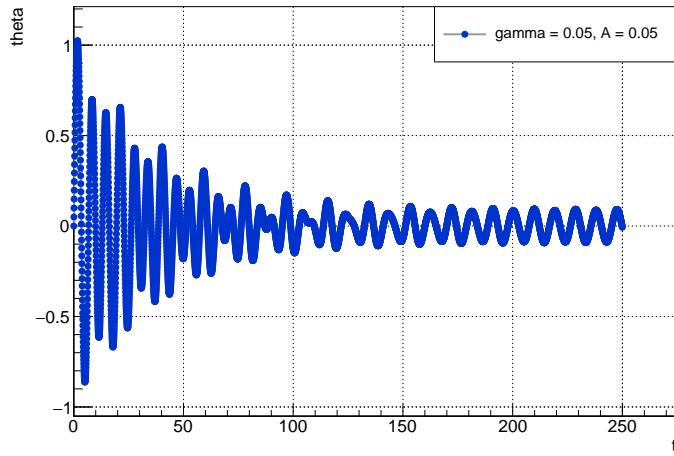


Figura 170: plot  $\theta(t)$  con RK4 per  $\gamma = 0.05$  e  $A = 0.05$

In questo caso ancora, la periodicità si manifesta a tempi molto più grandi. Come è possibile notare da tutte le soluzioni prodotte, risulta difficile trovare correlazioni tra la specializzazione dei parametri  $\gamma$  e  $A$  e le caratteristiche della soluzione, come il tempo  $\bar{t}$  o l'ampiezza delle oscillazioni in regime di periodicità. Tuttavia, è possibile affermare che la particella classica, per  $A < 0.8$ , spesso oscilla periodicamente attorno al punto di equilibrio stabile  $\theta = 0$  del corrispondente sistema senza forzante, ma solo dopo un certo intervallo di tempo in cui il moto appare irregolare e poco prevedibile, a causa del contributo della forza di attrito. La regolarità del moto per il range di ampiezze della forzante indi-

viduato, tuttavia, non è sempre verificata. Ne è un chiaro esempio la soluzione per  $\gamma = 0.05$  e  $A = 0.5$ , che assume l'andamento che segue.

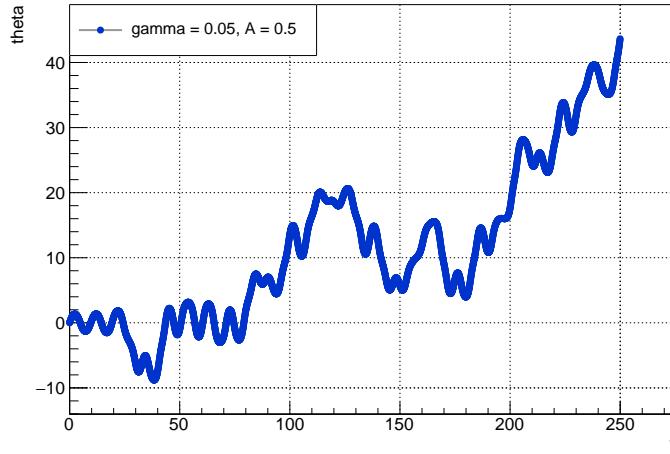


Figura 171: plot  $\theta(t)$  con RK4 per  $\gamma = 0.05$  e  $A = 0.5$

In questo caso, seppur si abbia  $A < 0.8$ , la soluzione appare del tutto irregolare e imprevedibile. Svolgendo diversi test, però, i casi in cui si ha un comportamento di questo tipo per il range di ampiezze individuato risultano rari rispetto ai casi di regolarità mostrati fino ad ora. Per valori di  $A > 0.8$ , invece, non è stato possibile trovare alcun tipo di regolarità nelle soluzioni prodotte: in quasi tutti i casi simulati, il moto risulta caotico e imprevedibile ad ogni tempo. Ad esempio, per  $\gamma = 0.05$  e  $A = 0.9$  si ha il seguente andamento.

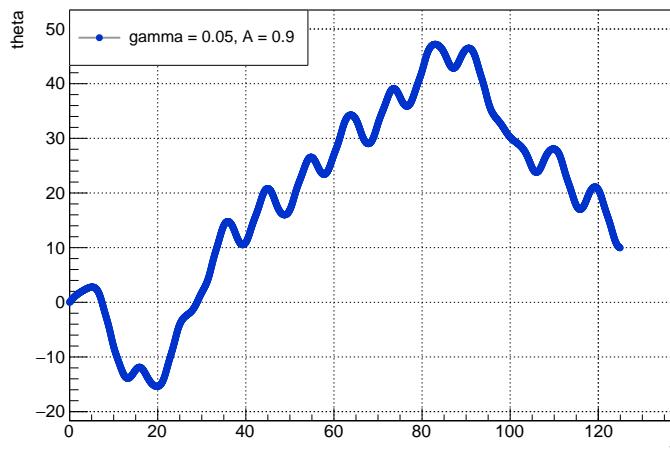


Figura 172: plot  $\theta(t)$  con RK4 per  $\gamma = 0.05$  e  $A = 0.9$

Come si nota, l'andamento appare del tutto irregolare. Si può mostrare che, anche per tempi molto grandi, non si ha mai un andamento definitivamente periodico. Per  $\gamma = 0.05$  e  $A = 1.8$  si ha quanto segue.

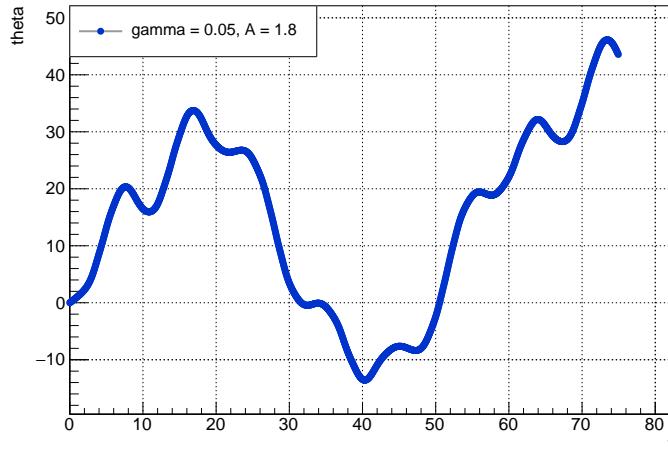


Figura 173: plot  $\theta(t)$  con RK4 per  $\gamma = 0.05$  e  $A = 1.8$

Anche in questo caso, in cui l'ampiezza della forzante supera abbondantemente il valore critico individuato, si ha un andamento sostanzialmente caotico. Tuttavia, non mancano moti in cui si ha regolarità a partire da un certo tempo  $\bar{t}$  anche nei casi in cui  $A > 0.8$ , come per  $\gamma = 0.3$  e  $A = 1.9$ .

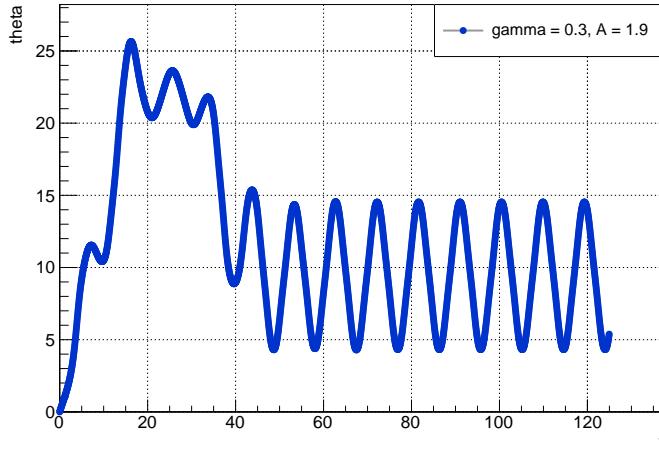


Figura 174: plot  $\theta(t)$  con RK4 per  $\gamma = 0.3$  e  $A = 1.9$

In questo caso, tuttavia, le oscillazioni periodiche non avvengono intorno a  $\theta = 0$  come nei primi casi analizzati, ma si configurano intorno ad un valore angolare positivo. In definitiva, a meno della parziale regolarità individuata per valori del parametro  $A < 0.8$ , la soluzione alla ODE in esame appare caotica nel senso di molto sensibile alla scelta dei parametri  $\gamma$  e  $A$ , rendendo di fatto impossibile trattare il fenomeno ricorrendo alle classiche considerazioni analitiche. Al fine di compiere previsioni accurate sarà allora necessario utilizzare le tecniche numeriche implementate, assicurandoci di conoscere con grande precisione i due

valori dei parametri fisici in esame. Evidentemente, tali valori dovranno essere gestiti con un tipo di variabile reale che possa consentire una precisione elevata, al fine di minimizzare gli effetti dell'errore iniziale di approssimazione.

Si noti che, nel caso in esame, il campo vettoriale associato al sistema dinamico si scriverà come

$$\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} \mapsto \begin{pmatrix} \dot{\theta} \\ -\sin(\theta) - \gamma\dot{\theta} + A\sin(\omega t) \end{pmatrix}$$

Abbiamo quindi una dipendenza diretta dal tempo nella seconda componente del vettore tangente. Non sarà più possibile, allora, disegnare il campo dei vettori tangenti alle curve di fase in modo statico, in quanto la sua geometria sarà strettamente legata all'istante di tempo in cui viene calcolato. In altre parole, la mappa individuata non risulta più biettiva nel dominio del piano delle fasi. In tutti i casi di forze dipendenti dal tempo, dunque, non è più possibile disporre nemmeno di questa verifica qualitativa. Ad ogni modo, come in precedenza, si sono costruite le soluzioni  $(\theta_i, \dot{\theta}_i)_{i=1, \dots, N}$  con il metodo di Eulero, RK2 e RK4, con  $h = 0.05$  e  $N = 270$ , fissando arbitrariamente le costanti a  $\gamma = 1.5$  e  $A = 0.2$ , ottenendo il seguente risultato.

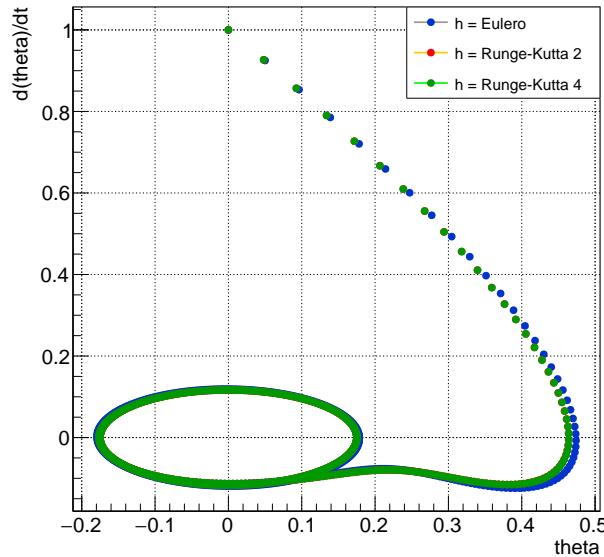


Figura 175: confronto  $\dot{\theta}(\theta)$  con E, RK2 e RK4,  $\gamma = 1.5$  e  $A = 0.2$

Come è possibile notare, i risultati in termini di efficienza numerica dei metodi appaiono del tutto confrontabili rispetto ai risultati ottenuti in figura 163 e 164, dove si sono plottate le soluzioni con i medesimi valori di  $\gamma$  e  $A$ . Il metodo di Eulero risulta approssimarsi bene ai metodi Runge-Kutta quando la soluzione diviene periodica e mostra, invece, una significativa discordanza nei tratti in cui varia bruscamente con meno regolarità. Da un punto di vista di analisi fisica, la curva di fase ottenuta è consistente con il moto osservato nelle figure precedenti: in corrispondenza dell'andamento sinusoidale della soluzione si ha una curva di

fase a ovale o ellisse, ossia un moto periodico e limitato. Al fine di verificare la divergenza del metodo di Eulero nelle soluzioni meno regolari si sono allora costruite le curve di fase per  $\gamma = 0.05$  e  $A = 0.5$ , ottenendo quanto segue.

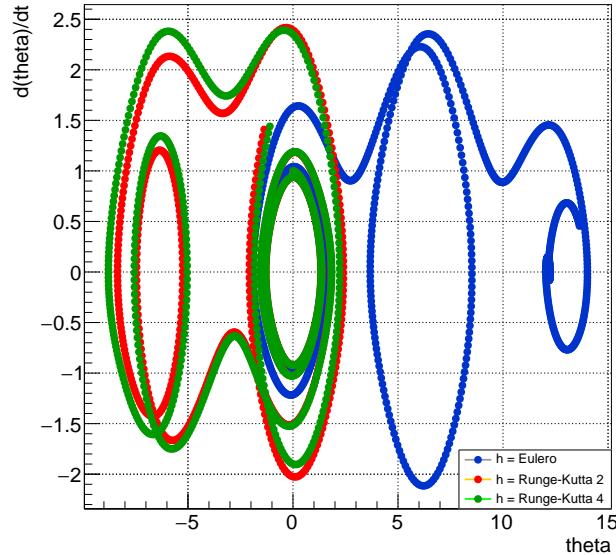


Figura 176: confronto  $\dot{\theta}(\theta)$  con E, RK2 e RK4,  $\gamma = 0.05$  e  $A = 0.5$

In questo caso è assolutamente evidente notare come il metodo di Eulero fallisca nel dare una soluzione sensata rispetto agli altri due metodi per il valore di  $h$  fissato. D'altra parte, è facile capire che tutti i metodi ad un passo ricostruiscono i punti della soluzione a partire dai punti precedenti: se viene commesso un errore non trascurabile nel calcolo di un punto e la soluzione varia bruscamente, allora l'errore si propagherà in modo molto rilevante sul punto successivo.

## Esercizio 16

Si vuole studiare numericamente la soluzione al problema di Cauchy

$$\ddot{\theta} + \gamma\dot{\theta} + (\alpha - \beta \cos(t)) \sin(\theta) = 0 \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ \dot{\theta}(0) = 0.1 \end{cases}$$

utilizzando il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4. Si vuole quindi studiare l'equazione differenziale che descrive il moto di una particella di massa unitaria soggetta ad una forza di smorzamento dipendente dalla velocità e ad una forza non lineare dipendente dal tempo e dalla posizione. In particolare, si vuole studiare il moto per  $\alpha = 0.5$  e  $\gamma = 0.03$  al variare di 3 dati valori del parametro  $\beta$ , ossia  $\beta_1 = 0.50$ ,  $\beta_2 = 0.63$  e  $\beta_3 = 0.70$ .

Non è difficile verificare che, per valori non banali di  $\alpha$ ,  $\beta$  e  $\gamma$ , l'equazione differenziale in esame non ammette soluzione analitica esprimibile per mezzo di funzioni elementari. Anzitutto, dall'equivalenza (0.16) segue che il problema in esame equivale allo studio del sistema dinamico in  $\mathbb{R}^2$

$$\begin{cases} \dot{\theta} = z \\ \dot{z} = -\gamma\dot{\theta} - (\alpha - \beta \cos(t)) \sin(\theta) \end{cases} \quad \text{con} \quad \begin{cases} \theta(0) = 0 \\ z(0) = 0.1 \end{cases}$$

al quale risulta ora possibile applicare le relazioni ricorsive date dai tre metodi.

### Studio dei moti

Vogliamo studiare il caso in cui  $\beta = \beta_1$ . Al fine di operare un confronto visivo dell'andamento della soluzione al variare del metodo numerico, si sono calcolate e plottate le coppie  $(t_i, \theta_i)_{i=1,\dots,N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.1$  e con  $N = 500$ , ottenendo i seguenti andamenti.

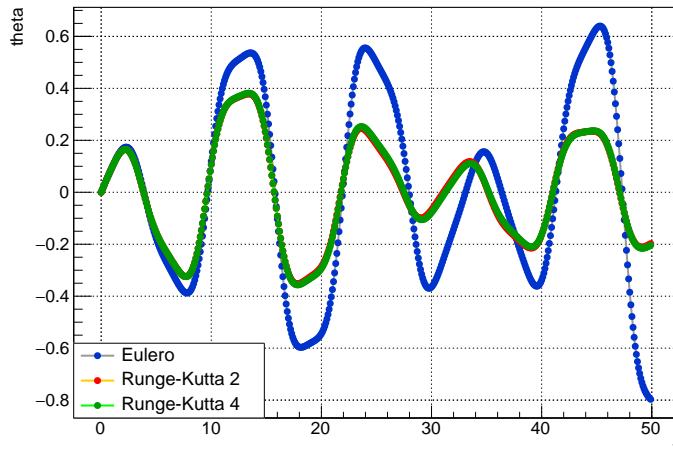


Figura 177: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_1$

Come si nota, per il valore del passo selezionato, il metodo di Eulero di discosta sensibilmente dai metodi al secondo e al quarto ordine già a partire da  $\bar{t} \approx 5$ .

Le soluzioni date da Runge-Kutta 2 e 4, invece, appaiono sovrapponibili in quasi tutte le regioni del grafico. Si è quindi diminuito il passo di integrazione a  $h = 0.01$ , aumentando il numero di punti generati a  $N = 5000$ , al fine di visualizzare la soluzione nel medesimo intervallo temporale, come segue.

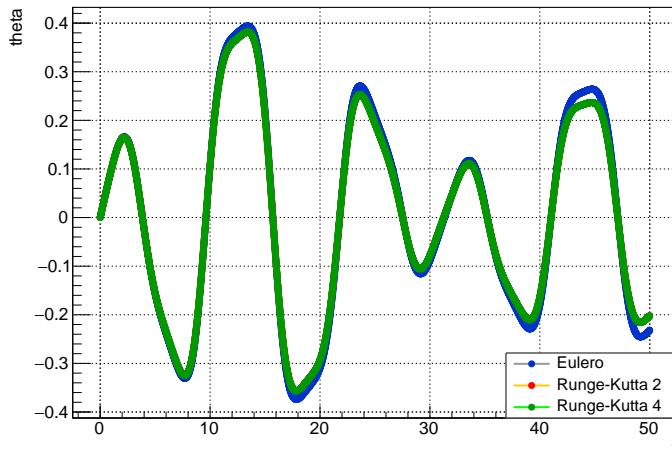


Figura 178: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_1$

Ad un valore del passo riscalato di una potenza del 10, la soluzione data da Eulero si approssima in modo decisamente migliore alle soluzioni date da Runge-Kutta, mostrando una discordanza visibile solo in corrispondenza dei massimi e dei minimi relativi di  $\theta(t)$ . Si sono quindi ripetuti gli stessi passaggi per  $\beta = \beta_2$ . In particolare, per  $h = 0.1$  e  $N = 500$  si sono ottenuti i seguenti andamenti.

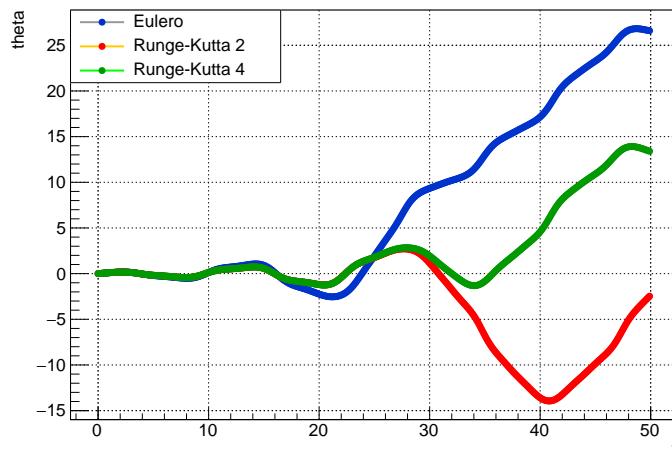


Figura 179: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_2$

In questo caso, il metodo di Eulero diverge dai due Runge-Kutta a partire da  $\bar{t} \approx 10$ , costruendo una soluzione sensibilmente diversa per ogni  $t > \bar{t}$ . Inoltre, anche la soluzione data da Runge-Kutta 2 diverge dal corrispondente metodo al

quarto ordine, ma a partire da  $\bar{t} \approx 26$ , distanziandosi in modo significativo da RK4 nel tempo. Per  $h = 0.01$  e  $N = 5000$  si è ottenuto quanto segue.

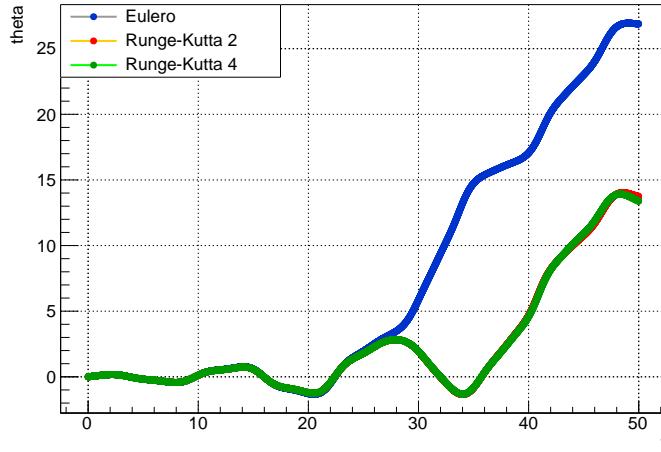


Figura 180: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_2$

Ad un passo riscalato di una potenza del 10, il tempo a partire dal quale il metodo di Eulero inizia a divergere sensibilmente aumenta a  $\bar{t} \approx 24$ . La soluzione per RK2, invece, inizia a divergere a partire degli ultimi istanti temporali in cui si è plottata la soluzione. Per  $\beta = \beta_2$ , dunque, si ha una maggiore difficoltà a costruire una soluzione precisa per tempi grandi rispetto al caso di  $\beta = \beta_1$ . Questo è dovuto al fatto che i metodi ad un passo utilizzati sono sensibili alla regolarità della soluzione che ricostruiscono. Per  $\beta = \beta_1$ , infatti, è evidente che la soluzione, rispetto al caso  $\beta = \beta_2$ , sia più regolare nel tempo, mostrando un comportamento oscillante caratterizzato dalla presenza di una pseudo-periodicità. Per  $\beta = \beta_3$  si è ottenuto quanto segue.

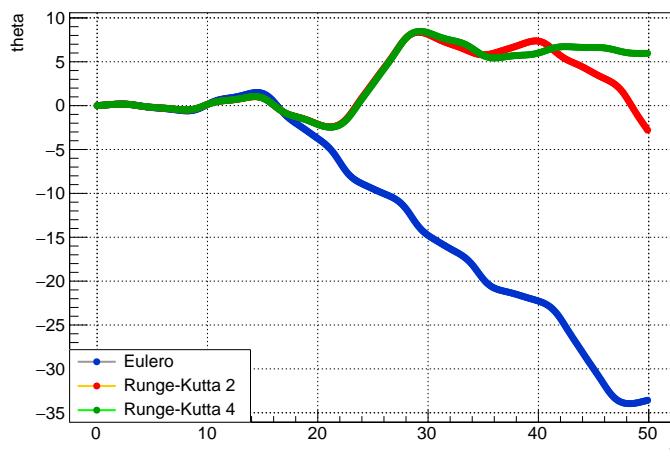


Figura 181: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_3$

La soluzione data dal metodo di Eulero diverge dalle altre due soluzioni a partire da  $\bar{t} \approx 10$ , mentre la soluzione data da Runge-Kutta 2 diverge da Runge-Kutta 4 a partire da  $\bar{t} \approx 28$ . Come nei precedenti casi, si sono allora fissati  $h = 0.01$  e  $N = 5000$ , ottenendo quanto segue.

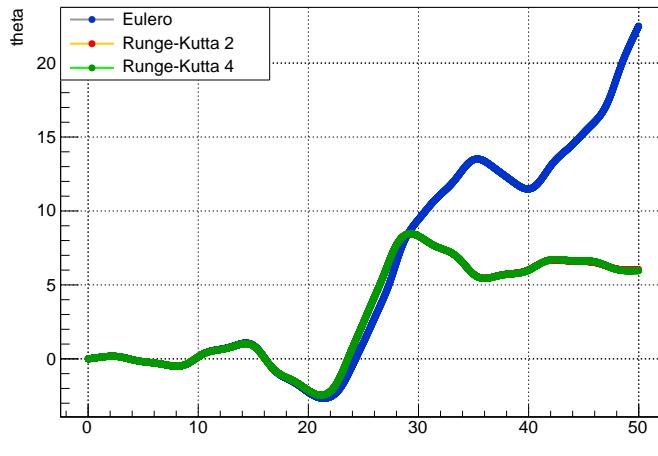


Figura 182: confronto  $\theta(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_3$

Per il valore del passo selezionato, dunque, si ha che la soluzione data dal metodo di Eulero appare discordarsi dalle altre due soluzioni a partire da  $\bar{t} \approx 20$ , mentre i metodi al secondo e al quarto ordine risultano, fino al tempo  $\bar{t} = 50$ , del tutto sovrapponibili. La spiegazione della necessità di un passo piccolo per garantire una buona precisione della soluzione è allora la medesima di quella già individuata per il confronto dei primi due casi. Visti i risultati ottenuti, si è deciso di plottare le 3 soluzioni al variare di  $\beta$  utilizzando il metodo più preciso di RK4 per  $h = 0.01$  e  $N = 5000$ , ottenendo i seguenti andamenti.

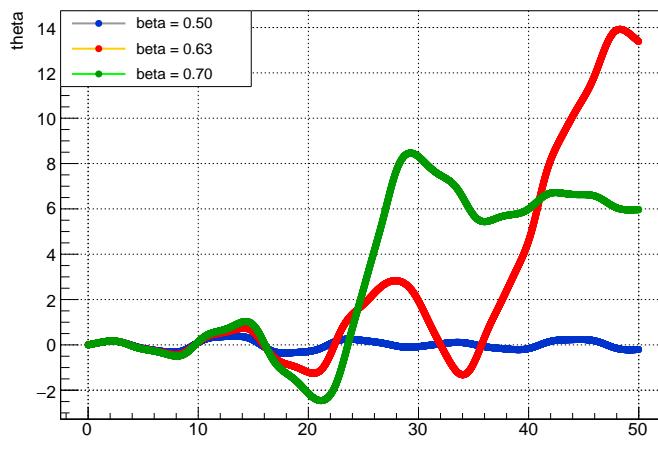


Figura 183: confronto  $\theta(t)$  con RK4 al variare di  $\beta$

Dal grafico è immediato notare che, a partire dall'istante temporale  $t \approx 10$ , le soluzioni  $\theta(t)$  si separano, definendo andamenti radicalmente differenti tra loro. Per  $\beta = \beta_1$  il moto appare limitato e caratterizzato da una certa periodicità intorno a  $\theta = 0$ , come si è già notato in precedenza. Anche per  $\beta = \beta_3$  il moto si stabilizza, ma solo definitivamente intorno a  $\theta \approx 6$ . Per  $\beta = \beta_2$ , invece, il comportamento appare molto più irregolare. Ciò che si osserva, allora, è che per piccole variazioni del parametro  $\beta$  della ODE in esame, il moto risultante si evolve in modo molto diverso e imprevedibile a priori: quello in esame è allora un primo caso rilevante di *sistema caotico*. Si noti che il problema in esame è descritto da un'equazione differenziale ordinaria e da un set di due dati iniziali: pertanto, risulta essere del tutto deterministico. Il fatto che un problema deterministico possa produrre caos può apparire, a prima vista, controintuitivo. Tuttavia, questo accade quando si è in presenza di un sistema descritto da una ODE contenente qualche tipo di non linearità, come nel nostro caso. Nel caso in esame, le condizioni iniziali sono fissate, ma l'equazione risulta sensibile alla variazione dei parametri che compaiono al suo interno determinando, di fatto, un comportamento imprevedibile. Vedremo che esistono anche casi di caoticità dati dalla sensibilità alle condizioni iniziali. Utilizzando le nozioni viste all'inizio della relazione si può affermare che, quello in esame, risulti essere un problema mal condizionato. L'unico modo per affrontarlo consiste nell'utilizzo di metodi numerici risolutivi abbastanza precisi da garantire la buona approssimazione della soluzione ad un tempo sufficientemente grande.

### Studio delle velocità

Vogliamo ora ripetere le medesime operazioni, ma studiando la velocità della particella in funzione del tempo. Si sono quindi calcolate e plottate le coppie  $(t_i, \dot{\theta}_i)_{i=1,\dots,N}$  secondo i tre metodi, fissando un valore del passo pari a  $h = 0.1$  e con  $N = 500$ . Per  $\beta = \beta_1$  si sono ottenuti i seguenti andamenti.

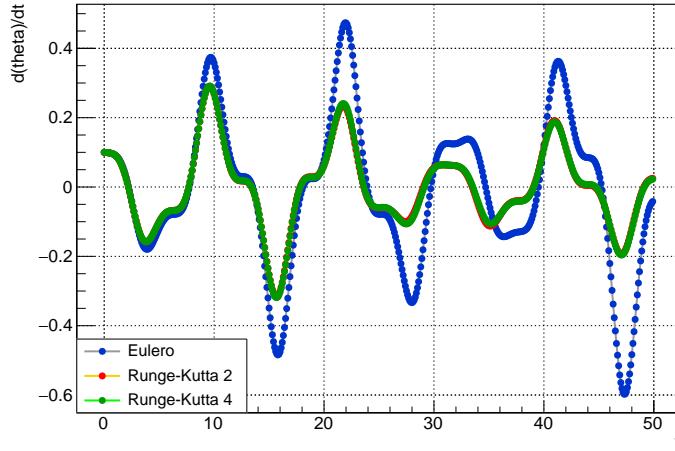


Figura 184: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_1$

La velocità data dal metodo di Eulero inizia a discostarsi dai metodi al secondo e al quarto ordine a partire da  $t \approx 5$ , esattamente come per la soluzione del

moto studiata precedentemente. Fissando i parametri dei metodi numerici ai valori di  $h = 0.01$  e  $N = 5000$  si sono ottenuti gli andamenti che seguono.

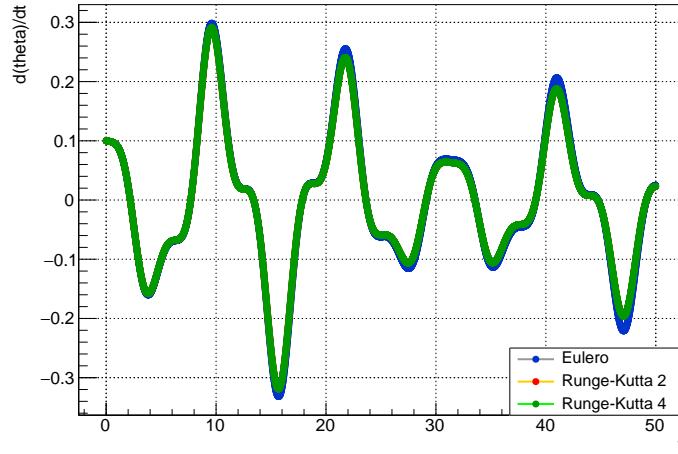


Figura 185: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_1$

Coerentemente con quanto si ha avuto modo di osservare nello studio del moto, la relativa regolarità della funzione in esame consente al metodo di Eulero una buona approssimazione per tutti i tempi rappresentati, presentando una lieve discordanza visiva solo in corrispondenza di massimi e minimi locali. Per  $\beta = \beta_2$  con  $h = 0.1$  e  $N = 500$  si è ottenuto quanto segue.

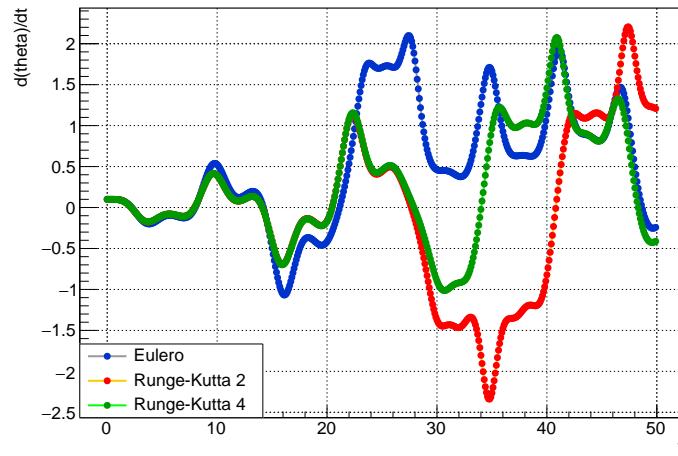


Figura 186: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_2$

Coerentemente con quanto si è ottenuto nello studio dei moti, l'irregolarità della funzione produce una difficoltà di convergenza delle soluzioni molto maggiore rispetto al caso precedente. In particolare, la velocità con Eulero si discosta dalle altre per  $\bar{t} \approx 8$ . La velocità con RK2, invece, si discosta dal corrispondente

metodo al quarto ordine per  $\bar{t} \approx 28$ . Come nei casi precedenti, si è allora ripetuta la medesima procedura per  $h = 0.01$  e  $N = 5000$ , come segue.

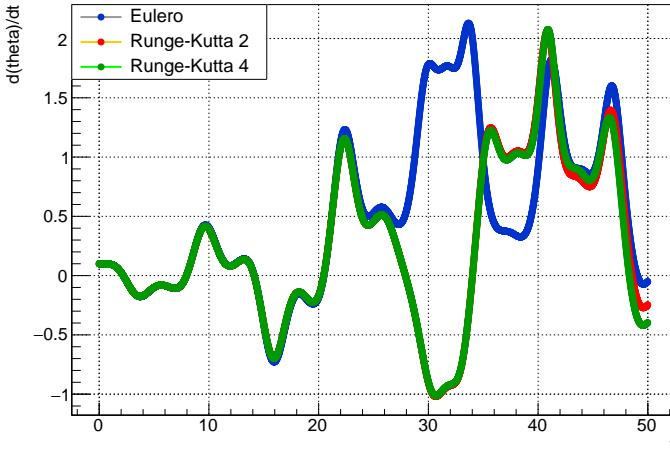


Figura 187: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_2$

Ad un passo riscalato di una potenza del 10, il metodo di Eulero produce una soluzione qualitativamente precisa fino a  $\bar{t} \approx 22$ . Il metodo di Runge-Kutta 2, invece, ricostruisce una buona soluzione fino a  $\bar{t} \approx 42$ . Per  $\beta = \beta_3$  con  $h = 0.1$  e  $N = 500$  si è ottenuto quanto segue.

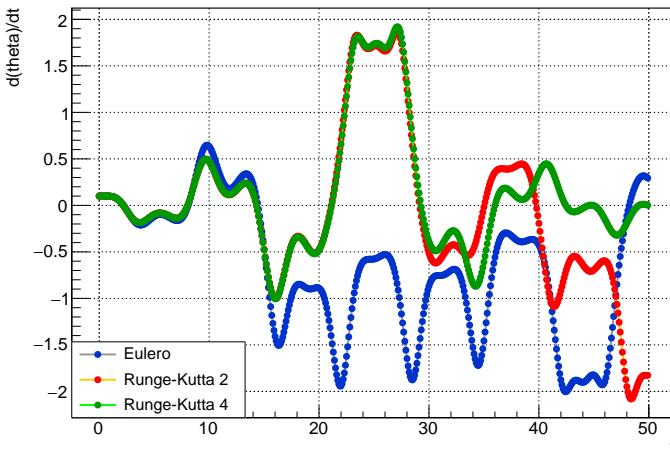


Figura 188: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.1$  e  $N = 500$ , per  $\beta = \beta_3$

In questo caso, per il solito passo iniziale fissato, la velocità data dal metodo di Eulero diverge a partire da  $\bar{t} \approx 8$ . La velocità data da RK2, invece, inizia a discostarsi dal corrispondente metodo al quarto ordine a partire da  $\bar{t} \approx 28$ . Quanto ottenuto risulta, anche in questo caso, consistente con i risultati ricavati dallo studio dei moti. Riscalando il passo e il numero di punti di integrazione del solito fattore 10 si sono ottenuti i seguenti risultati.

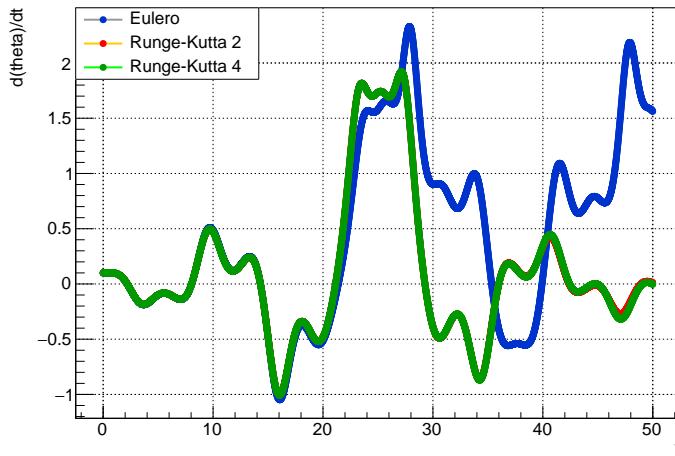


Figura 189: confronto  $\dot{\theta}(t)$  con E, RK2 e RK4 per  $h = 0.01$  e  $N = 5000$ , per  $\beta = \beta_3$

I metodi Runge-Kutta risultano sovrapponibili per tutti i tempi considerati, a meno di un'iniziale piccola divergenza per  $\bar{t} \approx 48$ . Il metodo di Eulero, invece, risulta sufficientemente preciso fino a  $\bar{t} \approx 20$ , per poi divergere caoticamente dalle soluzioni attese. Visti i risultati ottenuti si è deciso, anche in questo caso, di plottare i 3 andamenti della velocità al variare di  $\beta$  utilizzando il metodo più preciso di RK4 per  $h = 0.01$  e  $N = 5000$ , ottenendo i seguenti andamenti.

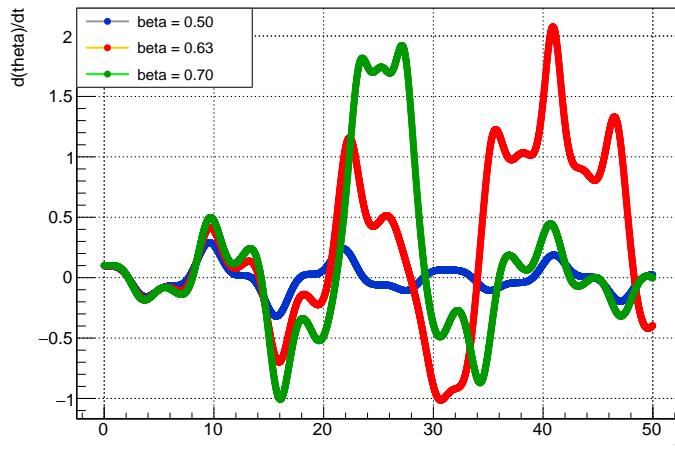


Figura 190: confronto  $\dot{\theta}(t)$  con RK4 al variare di  $\beta$

Anche i grafici delle velocità in funzione del tempo mostrano qualitativamente la caoticità del sistema per piccole variazioni del parametro  $\beta$ . A meno di alcuni istanti iniziali, le funzioni assumono un andamento del tutto diverso a partire da  $t \approx 4$ . Di fatto, la divergenza tra le varie funzioni si ha quando il termine non lineare dell'equazione differenziale che descrive il sistema inizia a diventare rilevante rispetto ai termini lineari.

## Studio delle curve di fase

Anzitutto notiamo che, in questo caso, il campo vettoriale associato al sistema dinamico assume la forma

$$\begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix} \mapsto \begin{pmatrix} \dot{\theta} \\ -\gamma\dot{\theta} - (\alpha - \beta \cos(t)) \sin(\theta) \end{pmatrix}$$

Anche qui, allora, non sarà possibile svolgere una verifica qualitativa delle curve di fase sfruttando la rappresentazione dei vettori tangenti alle curve nel piano delle fasi, in quanto la dipendenza esplicita dal tempo non permette la verifica della biunivocità dell'associazione. In altre parole, l'andamento delle tangenti dipenderà dall'istante in cui vengono calcolate. Al fine di studiare le curve di fase corrispondenti al sistema in esame si sono allora costruite le soluzioni  $(\theta_i, \dot{\theta}_i)_{i=1,\dots,N}$  con il metodo di Eulero, RK2 e RK4, con  $h = 0.1$  e  $N = 500$ . In questo caso, visti gli studi precedenti, si è deciso di non ripetere la medesima operazione due volte infittendo il passo di integrazione.

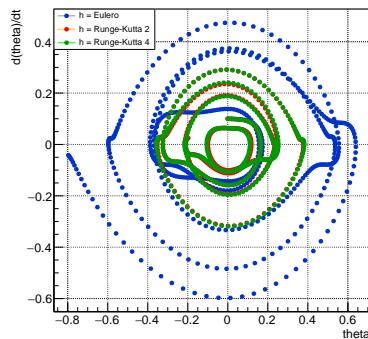


Figura 191:  $\dot{\theta}(\theta)$  per  $\beta = \beta_1$

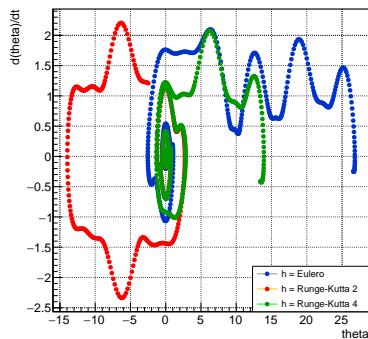


Figura 192:  $\dot{\theta}(\theta)$  per  $\beta = \beta_2$

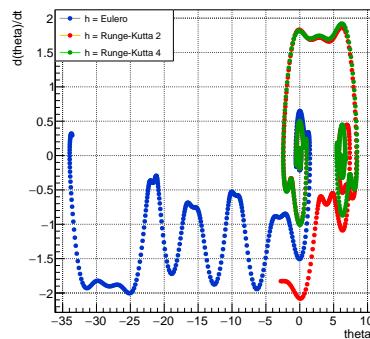


Figura 193:  $\dot{\theta}(\theta)$  per  $\beta = \beta_3$

Come è possibile notare, per  $\beta = \beta_1$  il metodo di Eulero ricostruisce un moto più ampio, ma con caratteristiche qualitative simili alle curve di fase più precise dei corrispondenti metodi al secondo e al quarto ordine. Per  $\beta = \beta_2$ , invece, l'irregolarità della soluzione porta alla produzione di curve di fase radicalmente

diverse tra loro. Per  $\beta = \beta_3$  le considerazioni sono del tutto analoghe. Quanto si osserva, dunque, risulta del tutto in linea con gli studi effettuati nelle due sezioni precedenti. Si sono quindi costruite le curve di fase  $\dot{\theta}(\theta)$  con RK4 per  $h = 0.01$  e  $N = 5000$ , ossia per un valore del passo e per un numero di punti tali da ricostruire la soluzione in modo sufficientemente preciso fino al tempo finale considerato. Per  $\beta = \beta_1$  si è ottenuto quanto segue.

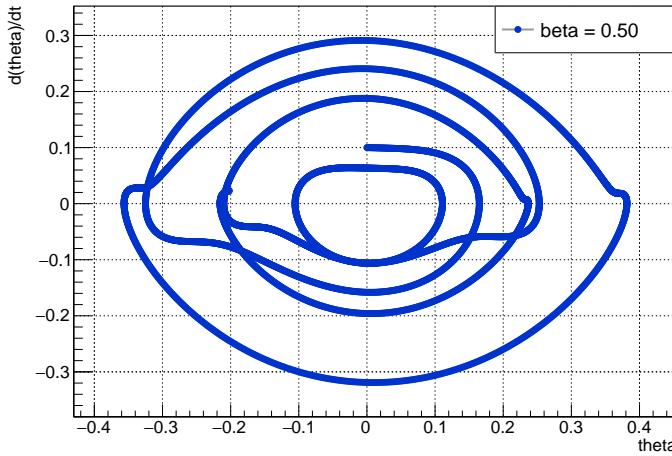


Figura 194:  $\dot{\theta}(\theta)$  con RK4 per  $\beta = \beta_1$

Come si nota, il moto ricostruito assume caratteristiche qualitative consistenti con quelle attese, restando limitato tra due valori angolari intorno al punto  $(\theta, \dot{\theta}) = (0, 0)$ . Questo fatto è consistente con quanto osservato in precedenza nello studio del moto e della velocità. Per  $\beta = \beta_2$  si è ottenuto quanto segue.

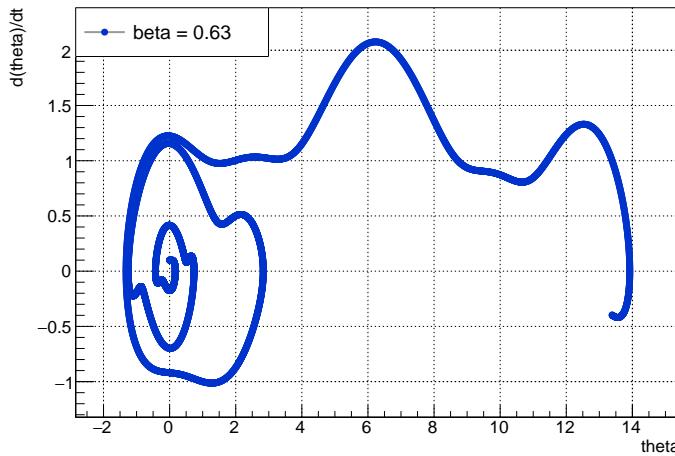


Figura 195:  $\dot{\theta}(\theta)$  con RK4 per  $\beta = \beta_2$

In questo caso, la posizione del punto materiale diverge dopo aver compiuto qualche oscillazione intorno allo zero del piano delle fasi. In particolare, il

moto appare molto più irregolare e imprevedibile rispetto al caso precedente, coerentemente con quanto ci si aspetta. Per  $\beta = \beta_3$  si è ottenuto quanto segue.

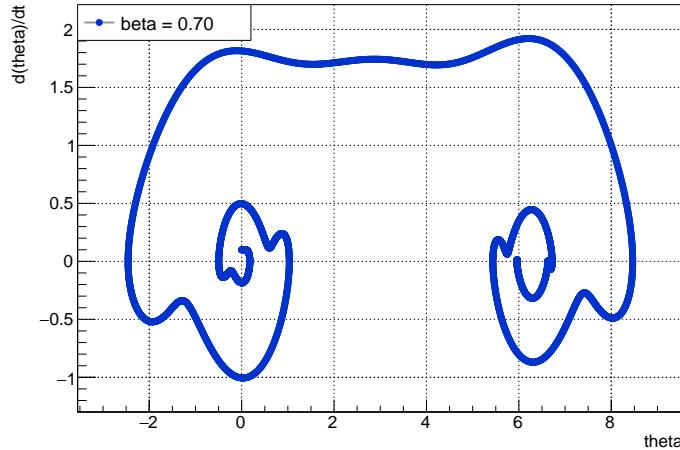


Figura 196:  $\dot{\theta}(\theta)$  con RK4 per  $\beta = \beta_3$

In questo caso, invece, il moto sembra evolvere mantenendo una particolare simmetria almeno per i primi istanti temporali. In effetti, quanto si osserva risulta consistente con i grafici di posizione e velocità in funzione del tempo prodotti in precedenza: in entrambi i casi è possibile apprezzare un andamento simmetrico che tende a stabilizzarsi asintoticamente intorno ad una cerca coordinata angolare positiva.

In definitiva, abbiamo avuto modo di osservare che esistono sistemi dinamici la cui non linearità può generare un comportamento del tutto imprevedibile al variare dei parametri che compaiono all'interno del sistema stesso. La sensibilità della soluzione al parametro  $\beta$  suggerisce che, in tutti questi casi, l'unica possibilità di compiere predizioni precise per tempi sufficientemente lunghi consista nell'utilizzo di un metodo numerico quanto più preciso possibile, congiuntamente ad un'elevata precisione con cui si devono rappresentare le variabili reali come, appunto, il valore del parametro  $\beta$  in questione.

## Esercizio 17

Si vuole studiare numericamente la soluzione al sistema dinamico in  $\mathbb{R}^3$

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = \rho x - xz - y \\ \dot{z} = xy - \beta z \end{cases}$$

con  $\sigma = 10$ ,  $\rho = 28$  e  $\beta = 8/3$ , utilizzando il metodo di Eulero, Runge-Kutta 2 e Runge-Kutta 4. Il sistema dinamico in esame è meglio noto come *attrattore di Lorenz* e rappresenta una semplificazione delle equazioni alle derivate parziali che descrivono il movimento termico di convezione di un fluido. A causa del forte troncamento, le equazioni descrivono bene il fenomeno fisico di convezione solo per  $\rho \approx 1$ . Tuttavia, esse vengono generalmente studiate fuori dal regime di sensatezza fisica ponendo, come nel nostro caso,  $\rho > 1$  in quanto rappresentano un modello a bassa dimensionalità di *sistema dinamico caotico*. L'importanza di questo sistema, infatti, trascende il modo in cui esso venne storicamente ottenuto, risiedendo essenzialmente nelle inaspettate proprietà delle sue soluzioni.

### Analisi dei metodi numerici

Anzitutto, al fine di rendere legittimi gli studi successivi sulle proprietà del sistema, si è deciso di operare un'analisi in precisione dei metodi numerici in esame. Per semplicità, si è considerato il set di condizioni iniziali

$$\begin{cases} x(0) = 1 \\ y(0) = 1 \\ z(0) = 1 \end{cases}$$

grazie alle quali risulta ora possibile l'applicazione diretta della (67), della (70) o della (73) con  $m = 3$  a seconda del metodo numerico selezionato. Si sono allora generate coppie di soluzioni ad un passo raddoppiato l'una rispetto all'altra fino al tempo  $\bar{t} = 3$  nel range

$$15000 \leq N < 30000 \quad \text{con} \quad N_{i+1} = N_i + 200$$

per poi calcolare le distanze (in metrica euclidea)

$$\Delta(\bar{t}) = |x_N(\bar{t}) - x_{2N}(\bar{t})|$$

per ogni  $N$  nel range selezionato. Si è quindi deciso di svolgere l'analisi in precisione solo per la prima componente della curva soluzione del sistema: con passaggi del tutto analoghi è possibile ottenere risultati identici per le due componenti rimanenti. In particolare, visti i risultati ottenuti nella prefazione teorica, ci si aspetta che l'errore scali come la (66), la (69) o la (72) a seconda del metodo numerico selezionato. Questo tipo di analisi è sensata solo a patto che il tempo finale  $\bar{t}$  sia sufficientemente piccolo, come verificheremo in seguito. Si sono quindi calcolati i logaritmi delle dispersioni e dei passi corrispondenti, per poi interpolare i dati con una retta della forma  $y = mx + q$ . Per il metodo di Eulero si è ottenuto quanto segue.

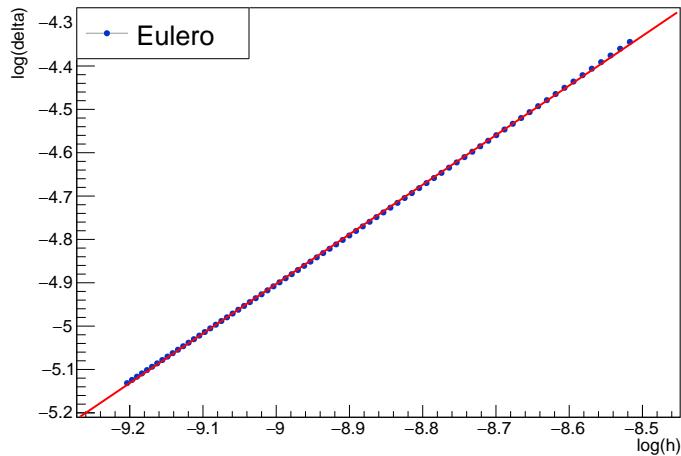


Figura 197: andamento  $\log \Delta(\bar{t})$  con Eulero: fit

Si è ottenuta la stima di parametri che segue.

$$q = 5.39 \quad \text{e} \quad m = 1.14 \approx 1$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Eulero risulta essere al primo ordine in  $h$ . Tuttavia, si noti che, diversamente da quanto si è ottenuto negli esercizi precedenti, la stima di  $m$  in questo caso risulta meno precisa. Per il metodo di Runge-Kutta 2 si è ottenuto quanto segue.

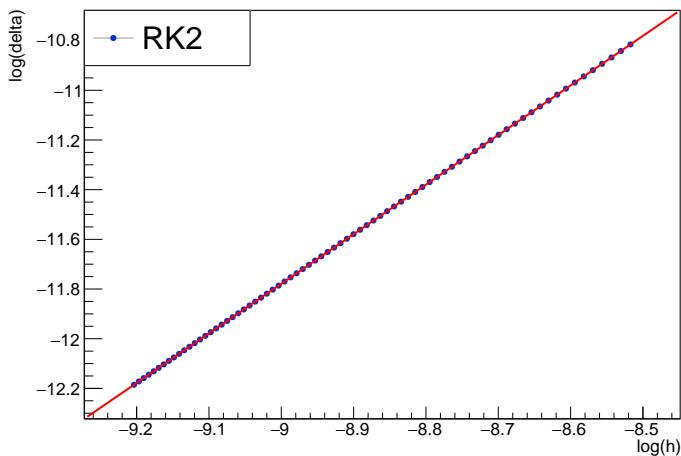


Figura 198: andamento  $\log \Delta(\bar{t})$  con RK2: fit

Si è ottenuta la stima di parametri che segue.

$$q = 6.19 \quad \text{e} \quad m = 2$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 2 risulta essere al secondo ordine in  $h$ . Per il metodo di Runge-Kutta 4 si è ottenuto quanto segue.

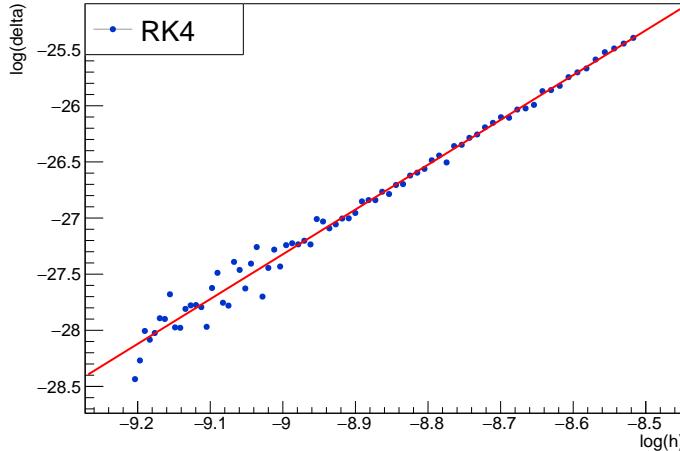


Figura 199: andamento  $\log \Delta(\bar{t})$  con RK4: fit

Si è ottenuta la stima di parametri che segue.

$$q = 8.59 \quad \text{e} \quad m = 3.99 \approx 4$$

Dal valore del coefficiente angolare stimato è possibile concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 4 risulta essere al quarto ordine in  $h$ . Si noti che i risultati ottenuti, in questo caso, presentano inaspettate fluttuazioni. Risulta possibile supporre che tale dispersione sia dovuta al fatto di aver considerato valori di  $N$  grandi, e quindi a tempo fissato valori di  $h$  molto piccoli. Vista la velocità con cui decresce l'errore in RK4, il calcolo di  $\Delta(\bar{t})$  presenterà differenze di numeri molto vicini tra loro, portando a parziali perdite di significatività dei risultati ottenuti a causa dell'errore di arrotondamento. Questa spiegazione è consistente con il fatto che, come si osserva dalla figura 199, i punti tendono a disporsi in modo meno rettilineo proprio in corrispondenza di  $\log h$  piccoli e quindi di  $h$  piccoli, tenendo conto del fatto che il logaritmo è una funzione monotona crescente del suo argomento.

Una volta verificato il corretto andamento ci si aspetta che sia sempre possibile determinare il valore di  $\bar{h}$  tale che la soluzione venga ricostruita, fino al tempo  $\bar{t}$ , con una precisione più piccola di una precisione  $\varepsilon$  desiderata. Questo fatto è sempre vero nel caso di sistemi dinamici non caotici, ossia tali che la soluzione per valori dei dati iniziali perturbati di una certa quantità infinitesima non diverga esponenzialmente dalla soluzione costruita a partire da dati iniziali non perturbati. D'altra parte, si è già anticipato che il sistema in esame è un celebre esempio di sistema caotico. Si sono allora ripetuti i medesimi passaggi di calcolo dei logaritmi delle dispersioni in funzione dei logaritmi dei passi all'interno dello stesso range di  $N$  precedente, ma fino ad un tempo finale  $\bar{t} = 15$ . A titolo di esempio, si sono costruite le soluzioni con il metodo di Eulero, ottenendo i risultati che seguono.

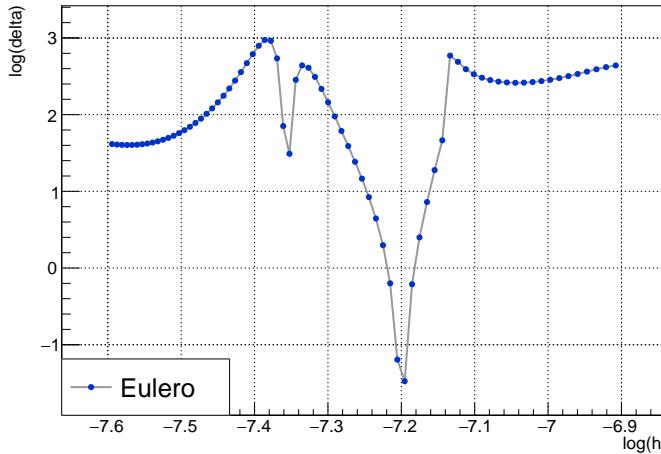


Figura 200: andamento  $\log \Delta (\bar{t} = 15)$  con Eulero

Evidentemente, l'andamento ottenuto risulta ben distante dall'andamento rettilineo atteso dato dalla (66). Si potrebbe pensare che, visto l'aumento del tempo finale, l'insensatezza dei risultati ottenuti sia dovuta al venir meno dell'ipotesi di regime asintotico sotto cui vale la (66). D'altra parte è possibile notare che i valori di  $N$  considerati sono sufficientemente grandi da garantire un valore di  $h$  piccolo, nonostante l'aumento di  $\bar{t}$ . Ciò che accade è che, nel caso di sistemi dinamici caotici, la soluzione per dati iniziali perturbati diverge caoticamente rispetto a quella originale. Questa irregolarità rende di fatto insensata l'analisi in precisione effettuata negli esercizi precedenti, o quantomeno ne determina una regione di sensatezza sull'asse dei tempi. La ragione per la quale per  $\bar{t} = 3$  si sono ottenuti risultati consistenti è data dal fatto che, per tempi piccoli, in questo caso particolare, la caoticità del sistema non è ancora rilevante. Vista l'analisi dei metodi numerici effettuata, negli studi successivi le soluzioni numeriche saranno sempre ricostruite con il metodo di Runge-Kutta 4 ad un passo molto piccolo.

### Punti di equilibrio e stabilità

Siamo ora interessati all'analisi delle proprietà delle soluzioni all'attrattore di Lorenz a partire dalla teoria dei sistemi dinamici. Come prima cosa, risulta interessante determinare i punti di equilibrio del sistema. Per definizione, i punti di equilibrio per un sistema dinamico sono tutte le sue soluzioni costanti nel tempo. Evidentemente, questo equivale a chiedere che le tangenti alle soluzioni siano i vettori nulli, ossia

$$(\dot{x}, \dot{y}, \dot{z}) = (0, 0, 0) \iff \begin{cases} \sigma(y - x) = 0 \\ \rho x - xz - y = 0 \\ xy - \beta z = 0 \end{cases}$$

Il sistema risultante non è lineare, ma per risolverlo sarà sufficiente isolare di volta in volta una variabile da un'equazione per poi sostituirla nelle rimanenti.

Con semplici passaggi si ricavano i tre punti di equilibrio

$$(x, y, z) = (0, 0, 0) \quad \text{e} \quad (x, y, z) = (\pm\sqrt{\beta(\rho-1)}, \pm\sqrt{\beta(\rho-1)}, \rho-1)$$

Sostituendo i valori numerici dei parametri nel nostro caso particolare avremo allora i punti di equilibrio

$$Q := (0, 0, 0) \quad \text{e} \quad P_{\pm} := (\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$$

Si è quindi deciso di verificare computazionalmente che i tre punti determinati fossero di equilibrio per il sistema. A tale scopo, per ogni punto di equilibrio preso come dato iniziale, si sono costruite soluzioni con il metodo di Runge-Kutta 4 fino al tempo  $\bar{t} = 15$  e con un numero di passi  $N = 7000$ , tali da consentire un valore del passo  $h$  sufficientemente piccolo da garantire una ricostruzione precisa della soluzione. Si sono quindi plottati i risultati nello spazio delle fasi  $(x, y, z)$ . Di seguito sono riportati i risultati ottenuti per i tre punti di equilibrio in esame.

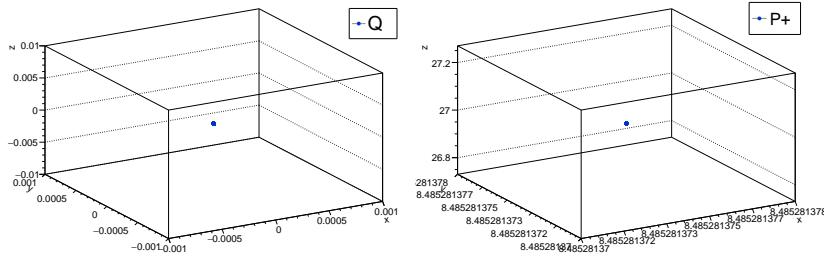


Figura 201: dati iniziali  $Q$

Figura 202: dati iniziali  $P_+$

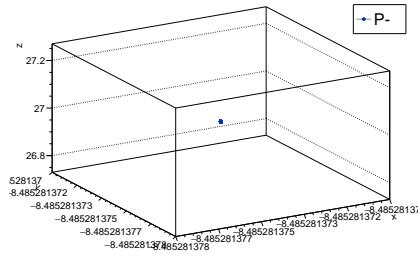


Figura 203: dati iniziali  $P_-$

Come è possibile notare, in tutti i casi in cui è stata ricostruita l'evoluzione a partire da dati iniziali coincidenti con i punti di equilibrio determinati, la curva di fase soluzione del sistema appare essere un punto isolato nello spazio delle fasi per ogni tempo. In altre parole, la soluzione risulta costante nel tempo come ci si aspetta, da cui segue la verifica computazionale dei punti di equilibrio per il sistema. Il secondo fatto rilevante da studiare consisterà, invece, nella classificazione dei punti di equilibrio, questa volta a partire dalle simulazioni numeriche. Detto  $E \in \mathbb{R}^3$  un generico punto di equilibrio, al fine di studiarne

la stabilità, un modo possibile consiste nel generare una curva di fase con dati iniziali  $E_\varepsilon$  per il sistema tali che

$$E_\varepsilon = E + \varepsilon$$

dove  $\varepsilon$  è un vettore con entrate date da numeri piccoli a piacere, ossia un vettore di norma piccola. La topologia delle curve di fase intorno al punto di equilibrio consentirà di determinare la stabilità o l'instabilità del punto considerato. Si sono allora costruite le soluzioni al sistema come in precedenza per ognuno dei punti di equilibrio in esame, ponendo

$$\varepsilon = (0.01, 0.01, 0.01)$$

al fine di simulare una perturbazione infinitesima dei dati iniziali. Di seguito sono riportati gli andamenti ottenuti.

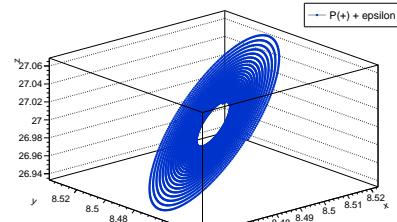
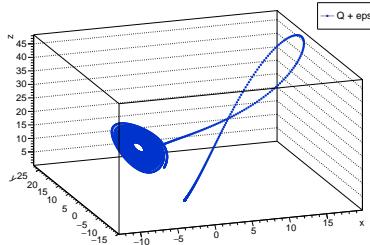


Figura 204: dati iniziali  $Q + \varepsilon$

Figura 205: dati iniziali  $P_+ + \varepsilon$

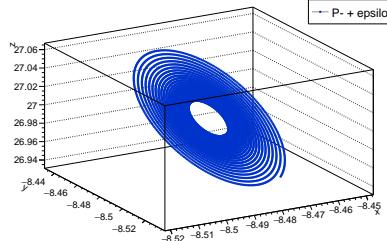


Figura 206: dati iniziali  $P_- + \varepsilon$

Come è possibile notare, la soluzione ricostruita con dati iniziali  $Q + \varepsilon$  diverge rapidamente dal punto di equilibrio  $Q$ . Questo fatto permette di concludere che si tratti di un punto di equilibrio instabile per il sistema. Anche la topologia delle curve di fase intorno a  $P_\pm$  porta alla medesima conclusione per questa coppia di punti. Infatti, seppur l'andamento delle curve appaia, in entrambi i casi, poco caotico con traiettorie regolari a spirale, l'evoluzione non rimane in una palla sufficientemente piccola centrata nei punti di equilibrio in esame. Inoltre, è possibile concludere che  $P_+$  e  $P_-$  sono punti di equilibrio caratterizzati dalla stessa natura, in quanto le curve di fase date dal dato iniziale perturbato di figura 205 e 206 risultano topologicamente equivalenti. L'instabilità dell'origine per i valori dei parametri fissati è deducibile anche analiticamente. La matrice

Jacobiana associata al sistema di Lorenz assume la forma

$$J(x, y, z) := \begin{pmatrix} \partial_x \dot{x} & \partial_y \dot{x} & \partial_z \dot{x} \\ \partial_x \dot{y} & \partial_y \dot{y} & \partial_z \dot{y} \\ \partial_x \dot{z} & \partial_y \dot{z} & \partial_z \dot{z} \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - z & -1 & -x \\ y & x & -\beta \end{pmatrix}$$

Valutandola nel punto di equilibrio  $Q$  avremo allora

$$J(Q) = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{pmatrix}$$

Applicando la formula di Laplace all'ultima riga si ha

$$\det J(Q) = \beta\sigma(\rho - 1) \neq 0 \quad \forall \sigma, \rho, \beta > 0$$

da cui segue che sono verificate le ipotesi per l'utilizzo del primo teorema di Lyapunov. Il polinomio caratteristico associato alla matrice sarà

$$\begin{aligned} P_\lambda(J) &= \det(\lambda I - J) = \\ &= (\lambda + \beta)(\lambda^2 + \lambda(\sigma + 1) - \rho\sigma + \sigma) \end{aligned}$$

Gli autovalori di  $J(Q)$  saranno dati da  $\lambda \in \mathbb{R}$  soluzioni di

$$P_\lambda(J) = 0$$

Utilizzando la legge di annullamento del prodotto e risolvendo l'equazione di secondo grado in  $\lambda$  otterremo gli autovalori

$$\lambda_1 = -\beta$$

$$\lambda_{2,3} = \frac{1}{2} \left[ -(\sigma + 1) \pm \sqrt{(\sigma + 1)^2 + 4\sigma(\rho - 1)} \right]$$

Infine, sostituendo i valori fissati dei parametri  $\sigma, \rho$  e  $\beta$  del sistema avremo che

$$\lambda_1 < 0 \quad \text{e} \quad \lambda_2 > 0 \quad \text{e} \quad \lambda_3 < 0$$

Siccome tutti gli autovalori determinati sono non nulli e ne esiste almeno uno positivo, possiamo finalmente concludere che, per il primo criterio di Lyapunov, il punto di equilibrio  $Q$  è instabile per il sistema di Lorenz, coerentemente con quanto si ha avuto modo di osservare in figura 204. Risulta possibile dimostrare, per mezzo di un procedimento del tutto analogo, che i punti di equilibrio  $P_\pm$  sono anch'essi instabili per il sistema per il valore di  $\rho$  fissato. Questo fatto è consistente con quanto si ha già avuto modo di osservare dalle figure 205 e 206: seppur l'andamento delle curve sia più regolare rispetto al caso precedente, l'evoluzione si allontana dai punti di equilibrio  $P_\pm$ , ossia i punti sono instabili per definizione. Tuttavia, si può dimostrare che  $P_\pm$  sono invece stabili per

$$1 < \rho < 24.7$$

dove  $\sigma$  e  $\beta$  sono inchiodati agli stessi valori. Questo fatto, unito alla singolare topologia delle curve di fase di figura 205 e 206, fa sicuramente pensare che la natura di  $P_\pm$  sia diversa da quella di  $Q$ . Si può allora pensare di raffinare la definizione di instabilità, individuando sottoclassi di punti di equilibrio instabile in base alla topologia delle curve di fase risultanti da un dato iniziale coincidente con il punto di equilibrio perturbato di una quantità infinitesima.

### Soluzioni generali e insiemi attrattori

Chiarita la natura dei punti di equilibrio, siamo ora interessati allo studio delle soluzioni numeriche al sistema a partire da generici dati iniziali diversi dagli equilibri. Anzitutto, notiamo che posto  $F := (\dot{x}, \dot{y}, \dot{z})$  si ha

$$\nabla \cdot F := \frac{\partial}{\partial x} \dot{x} + \frac{\partial}{\partial y} \dot{y} + \frac{\partial}{\partial z} \dot{z} = -(\sigma + \beta + 1) < 0$$

per ogni  $(x, y, z) \in \mathbb{R}^3$ , poiché  $\sigma, \beta > 0$ . Siccome la divergenza del campo vettoriale associato al sistema è negativa in ogni punto dello spazio delle fasi, il campo tenderà a contrarsi anziché espandersi. Ci si aspetta, dunque, che la soluzione tenda ad evolvere contraendo i volumi, convergendo ad una certa regione dello spazio. Per semplicità e coerenza con quanto si è svolto nella prima parte dell'esercizio, si è considerato il set di dati iniziali

$$\begin{cases} x(0) = 1 \\ y(0) = 1 \\ z(0) = 1 \end{cases}$$

Si è quindi costruita una soluzione al sistema con il metodo di Runge-Kutta 4 fino al tempo  $\bar{t} = 40$  e con un numero di passi  $N = 7000$ , tali da consentire un valore del passo  $h$  sufficientemente piccolo da garantire una ricostruzione precisa della soluzione. Si è poi plottata la soluzione nelle proiezioni  $(x, y)$ ,  $(x, z)$  e  $(y, z)$  dello spazio delle fasi. In particolare, si è anche deciso di assegnare ad ogni punto un colore in grado di dare un'indicazione qualitativa sul tempo, la cui scala è riportata a destra di ogni grafico: essa permetterà di dare un'orientazione all'evoluzione delle curve. Nel piano  $(x, y)$  si è ottenuto quanto segue.

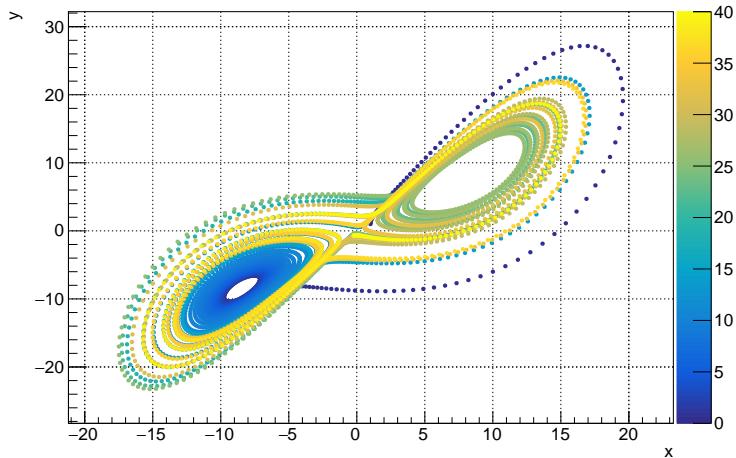


Figura 207: curva di fase nel piano  $(x, y)$

Come è possibile notare, la curva di fase risulta limitata e il suo andamento tende ad evolvere nella stessa regione dello spazio delle fasi, coerentemente con le considerazioni fatte. Nel piano  $(x, z)$  si è ottenuto quanto segue.

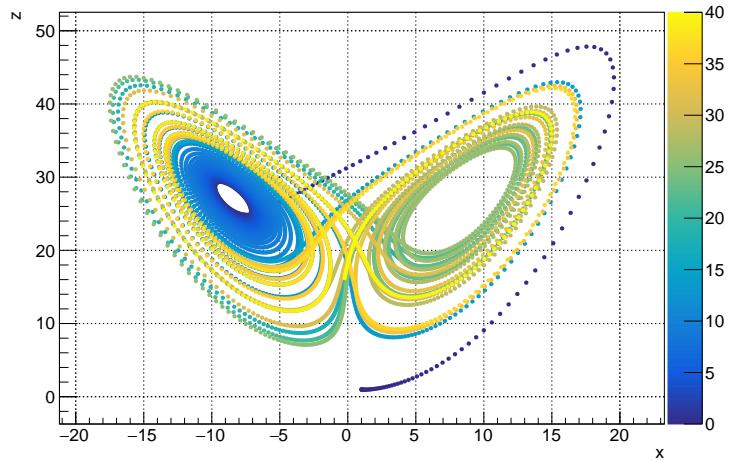


Figura 208: curva di fase nel piano  $(x, z)$

Anche in questo caso è possibile osservare la limitatezza della curva e l'evoluzione concentrata sempre nello stesso volume dello spazio delle fasi, coerentemente con quanto mostrato. Infine, nel piano  $(y, z)$  si è ottenuto quanto segue.

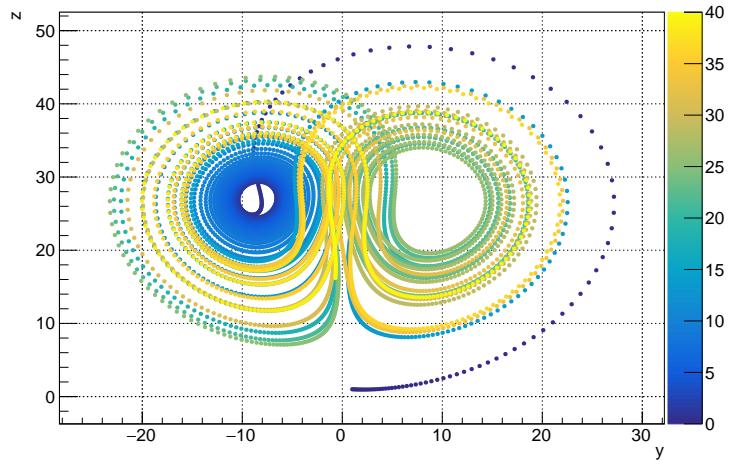


Figura 209: curva di fase nel piano  $(y, z)$

L'evoluzione presenta, anche in questo caso, caratteristiche qualitative analoghe a quelle osservate in precedenza. Evidentemente, le curve ottenute presentano alcune caratteristiche singolari: anzitutto, una particolare simmetria rispetto ad un asse. In secondo luogo, le soluzioni tendono a disporsi sempre intorno a due lobi centrati in due punti dello spazio delle fasi. Non è difficile verificare, anche solo osservando le scale dei grafici, che tali punti corrispondono proprio ai punti di equilibrio  $P_{\pm}$  studiati in precedenza. Al fine di avere una visione globale dell'andamento della soluzione, si è quindi deciso di plottare l'intera curva di fase nello spazio delle fasi  $(x, y, z)$ , ottenendo il seguente grafico.

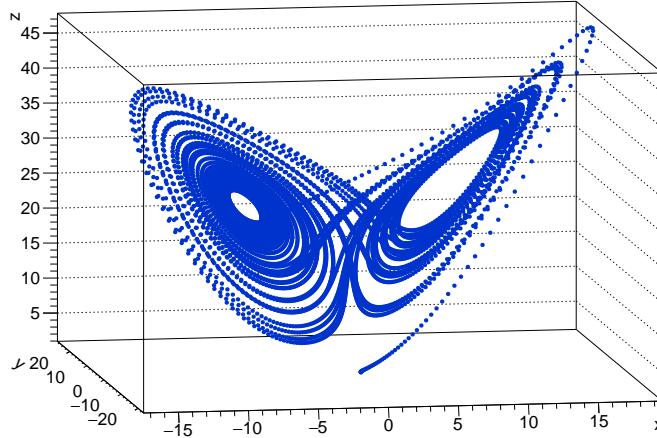


Figura 210: curva di fase nello spazio delle fasi  $(x, y, z)$

Come si può notare, le caratteristiche globali appaiono consistenti con quelle osservate nelle proiezioni bidimensionali della curva. Un fatto rilevante che può essere verificato per mezzo di considerazioni analitiche riguarda la tendenza del campo vettoriale a ruotare intorno ai punti  $P_{\pm}$ . Si noti, infatti, che

$$\begin{aligned}\nabla \times F &:= \det \begin{pmatrix} \hat{i} & \hat{j} & \hat{k} \\ \partial_x & \partial_y & \partial_z \\ \dot{x} & \dot{y} & \dot{z} \end{pmatrix} = \\ &= (\partial_y \dot{z} - \partial_z \dot{y}) \hat{i} + (\partial_z \dot{x} - \partial_x \dot{z}) \hat{j} + (\partial_x \dot{y} - \partial_y \dot{x}) \hat{k} = \\ &= (2x, -y, \rho - \sigma - z)\end{aligned}$$

Sostituendo i valori dei parametri fissati avremo allora

$$\nabla \times F = (2x, -y, 18 - z)$$

Valutando il rotore nel punto  $P_+$  avremo

$$(\nabla \times F)_{P_+} = (12\sqrt{2}, -6\sqrt{2}, -9)$$

Valutando il rotore nel punto  $P_-$  avremo

$$(\nabla \times F)_{P_-} = (-12\sqrt{2}, 6\sqrt{2}, -9)$$

Abbiamo allora ottenuto i due vettori  $(\nabla \times F)_{P_{\pm}}$  applicati ai punti  $P_{\pm}$  che rappresentano il rotore del campo associato al sistema valutato nei punti di equilibrio. Il rotore ha un'interpretazione dinamica interessante: la sua direzione coincide con l'asse di rotazione, il suo modulo dà indicazioni circa la "velocità" di rotazione del flusso e, infine, il verso è coerente con il verso in cui avviene la rotazione secondo la regola della mano destra. Ma allora, le quantità vettoriali determinate possono essere usate come verifica del fatto che il campo tenda a ruotare intorno a  $P_{\pm}$ . Si sono quindi rappresentati i due vettori determinati insieme alla soluzione nello spazio delle fasi, ottenendo quanto segue.

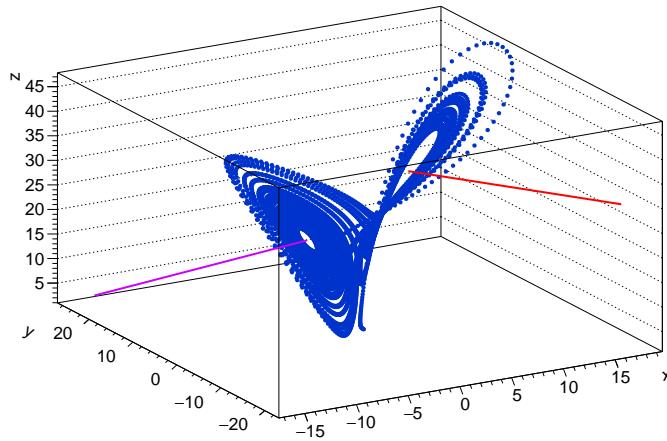


Figura 211: curva di fase con rotori  $(\nabla \times F)_{P\pm}$

Si noti che non sono state rappresentate le punte delle frecce poiché, nel caso dei vettori applicati in esame, il verso è già univocamente determinato osservando che l'inizio del vettore corrisponde al centro dei due lobi, ossia ai punti di equilibrio  $P_\pm$ . Di seguito è riportato lo stesso grafico da un'angolazione diversa.

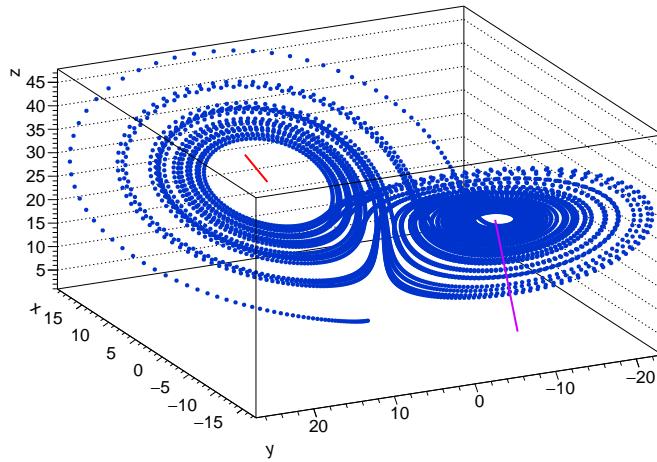


Figura 212: curva di fase con rotori  $(\nabla \times F)_{P\pm}$  con diversa angolazione

Come è possibile notare dalla lunghezza di  $(\nabla \times F)_{P\pm}$ , il campo tende a ruotare in modo considerevole intorno ai due punti in esame, coerentemente con quanto si osserva dalle simulazioni numeriche. Inoltre, ricordando la regola della mano destra, il verso di rotazione risulta consistente con i risultati ottenuti nelle figure 207, 208 e 209, nelle quali l'informazione sui tempi rende conto in modo visivo del verso di rotazione. Infine, anche la direzione dei rotori risulta coerente, in quanto appare qualitativamente ortogonale al piano in cui tende ad evolvere la soluzione al sistema in un intorno dei punti di equilibrio.

Da quanto si ha avuto modo di osservare fino a questo punto, le proprietà delle soluzioni all'attrattore di Lorenz sono ora più chiare. In particolare, visti i risultati ottenuti, ci aspettiamo che il sistema in esame presenti qualche insieme *attrattore*, ossia una regione finita dello spazio delle fasi verso cui converge la soluzione dopo un certo tempo finito. Tuttavia, non è chiaro se tale insieme possa esistere qualunque sia il set di dati iniziali considerato. Ricordando che  $F$  è il campo vettoriale associato al sistema dinamico si consideri, quindi, una generica superficie chiusa  $S(t)$  nello spazio delle fasi avente volume  $V(t)$ . Si supponga allora di fissare un set di dati iniziali in tale volume. Vogliamo capire come evolve nel tempo il volume in cui è contenuta la soluzione. Si può dimostrare che vale l'uguaglianza

$$\dot{V}(t) = \int_{S(t)} F \cdot dS = \int_{V(t)} \nabla \cdot F dV$$

per teorema della divergenza. Complessivamente, questo risultato è meglio noto in letteratura come teorema di Liouville. Sostituendo il valore della divergenza ricavato in precedenza si avrà

$$\dot{V}(t) = -(\sigma + \beta + 1) \int_{V(t)} dV = -(\sigma + \beta + 1)V(t)$$

Abbiamo allora ottenuto l'equazione differenziale ordinaria al primo ordine

$$\dot{V}(t) = -(\sigma + \beta + 1)V(t)$$

che descrive la variazione del volume in cui sono presenti le traiettorie nel tempo. Posto  $V_0 := V(0)$  il volume all'istante iniziale, il problema di Cauchy risultante è risolubile analiticamente e la soluzione ha la forma

$$V(t) = V_0 \exp(-(\sigma + \beta + 1)t)$$

Tuttavia, possiamo notare agilmente che

$$\lim_{t \rightarrow +\infty} V(t) = 0$$

Abbiamo allora mostrato che, qualunque sia l'insieme di dati iniziali considerato, la traiettoria descritta dalla soluzione del sistema tenderà a convergere ad un insieme attrattore con un tempo esponenziale. Questo fatto poteva già essere intuito in quanto la divergenza del campo è negativa in ogni punto dello spazio delle fasi, ma in tal modo siamo riusciti a mostrare più in dettaglio che la soluzione tende convergere ad un insieme attrattore con una rapidità che sappiamo ora quantificare. Al fine di verificare quanto detto, si sono quindi svolte diverse simulazioni con il metodo di Runge-Kutta 4 fino al tempo  $\bar{t} = 40$  e con un numero di passi  $N = 7000$ . In particolare, si sono generate 3 diverse soluzioni corrispondenti a diversi dati iniziali, anche molto distanti dall'origine:

$$I_1 := (83, 32, 8) \quad \text{e} \quad I_2 := (0, 21, -4) \quad \text{e} \quad I_3 := (-54, -2, -41)$$

dove  $I$  è una terna che rappresenta il dato iniziale. Di seguito sono riportati i risultati ottenuti direttamente nello spazio delle fasi  $(x, y, z)$ .

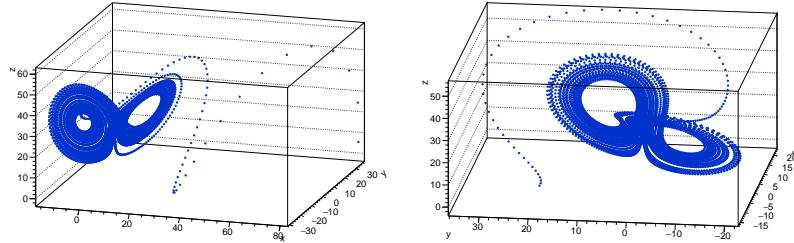


Figura 213: dati iniziali  $I_1$

Figura 214: dati iniziali  $I_2$

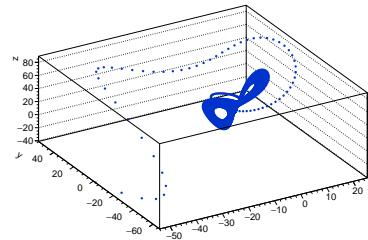


Figura 215: dati iniziali  $I_3$

Come è possibile notare, le curve di fase forniscono una verifica numerica dei risultati ottenuti: le soluzioni tendono molto rapidamente a convergere ad un insieme attrattore in prossimità dell'origine degli assi, qualunque sia il set di dati iniziali considerato. La rapidità con cui avviene il fenomeno è quantificabile anche osservando che i punti che costituiscono la soluzione risultano molto meno addensati per i primi istanti temporali, coerentemente con una rapidità esponenziale. Una volta raggiunto l'insieme attrattore, le soluzioni tendono poi a disporsi sempre secondo la stessa geometria, ruotando intorno ai punti di equilibrio  $P_{\pm}$ . Esiste un modo per determinare l'insieme attrattore analiticamente, ma ometteremo per il momento questa parte di studio.

### Caos deterministico

Abbiamo quindi giustificato il termine attrattore, che viene spesso accostato al sistema di Lorenz. Più precisamente, il sistema dinamico in esame appartiene alla classe degli *attrattori strani*, che consiste in tutti quegli attrattori caratterizzati da una forte sensibilità ai dati iniziali. Il sistema di Lorenz si configura quindi come un esempio di sistema dinamico caotico dove il caos, in questo caso, è dato dalla grande sensibilità ai dati di Cauchy. Di fatto, l'interesse verso questo sistema è dato quasi esclusivamente da questa proprietà. Gli attrattori strani sono anche caratterizzati da una struttura frattale e da una complessa dinamica non periodica. Intuitivamente, un sistema dinamico ha struttura frattale quando ammette un insieme attrattore e quando ingrandendo a piacere in questo insieme è sempre possibile trovare una porzione della curva soluzione del sistema. In altre parole, per  $t \rightarrow \infty$ , la soluzione deve essere densa nell'insieme attrattore, nel significato analitico del termine. Diamo questa precisazione in

quanto si avrà modo di incontrare di nuovo il termine frattale, ma con un significato non del tutto sovrapponibile al significato che assume nel caso in esame. Ai nostri fini, ci concentreremo solo sulla forte sensibilità alle condizioni iniziali: la proprietà caratterizzante del sistema dinamico in esame. Si consideri, dunque, una curva  $\gamma : \mathbb{R} \rightarrow \mathbb{R}^3$  soluzione del sistema con dati iniziali  $A$ . Sia poi  $\gamma_\varepsilon : \mathbb{R} \rightarrow \mathbb{R}^3$  una soluzione ottenuta con dati iniziali  $A_\varepsilon = A + \varepsilon$ , dove  $\varepsilon$  è un vettore con entrate date da numeri piccoli a piacere, ossia un vettore di norma piccola. Stiamo quindi considerando due soluzioni tali che una sia data, rispetto all'altra, dallo stesso dato iniziale perturbato di una quantità infinitesima. Un sistema sensibile ai dati iniziali può essere definito informalmente come un sistema tale che  $\gamma_\varepsilon$  si discosti significativamente e caoticamente da  $\gamma$  già a partire da tempi piccoli. Si è quindi deciso di verificare numericamente questa proprietà, costruendo due soluzioni con il metodo di Runge-Kutta 4 e fissando il numero di passi a  $N = 7000$ . In particolare, si sono considerati i vettori

$$A = (14, 2, 21) \quad \text{e} \quad \varepsilon = (0.1, 0.1, 0.1)$$

Di seguito sono riportati gli andamenti di  $\gamma$  e  $\gamma_\varepsilon$  a confronto per diversi valori di tempo finale  $\bar{t}$ .

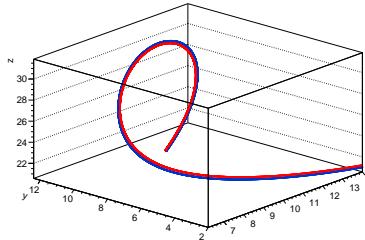


Figura 216:  $\gamma$  e  $\gamma_\varepsilon$  per  $\bar{t} = 0.5$

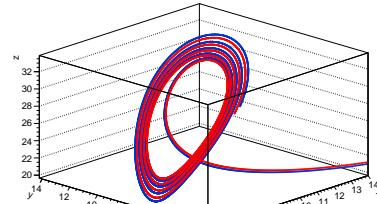


Figura 217:  $\gamma$  e  $\gamma_\varepsilon$  per  $\bar{t} = 3$

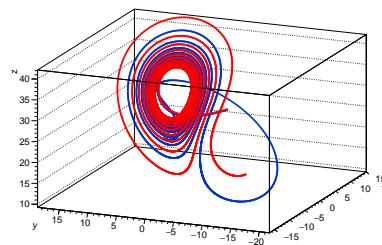


Figura 218:  $\gamma$  e  $\gamma_\varepsilon$  per  $\bar{t} = 7$

Come è possibile notare, per un breve istante di tempo iniziale, le due soluzioni di fatto coincidono, per poi iniziare a differire significativamente non appena vengono attratte dall'insieme attrattore. Rapidamente le due soluzioni tendono ad assumere andamenti non sovrapponibili, per poi divergere caoticamente per  $\bar{t} = 7$ . Risultati del tutto sovrapponibili possono essere ottenuti variando il dato iniziale  $A$  a piacere. Ritornando alla simulazione svolta, ad un tempo  $\bar{t} = 18$  la situazione è la seguente.

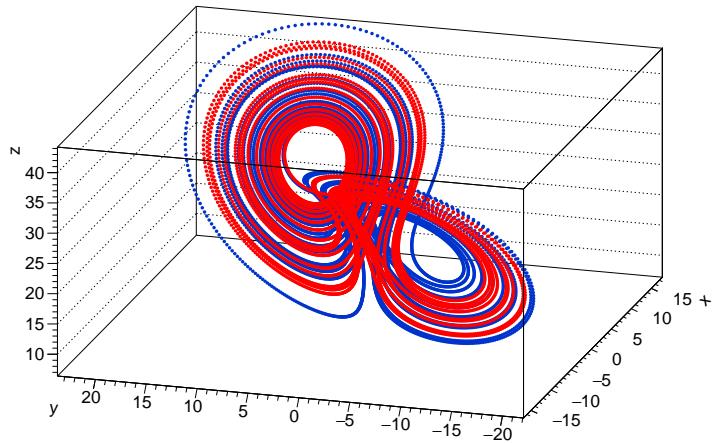


Figura 219:  $\gamma$  e  $\gamma_\varepsilon$  per  $\bar{t} = 18$

Risulta quindi possibile concludere che la non linearità del sistema di Lorenz renda la soluzione caotica, minando alla base la possibilità di compiere previsioni accurate sulle traiettorie nel lungo periodo. L'esistenza di attrattori strani, o più in generale di sistemi sensibili ai dati iniziali, rende necessaria una riflessione circa i fenomeni aventi natura deterministica. Siamo infatti portati a pensare che ogni fenomeno la cui soluzione sia descritta da una legge deterministica come un'equazione differenziale sia perfettamente predicibile, posti i dati iniziali. La celebre convinzione di Laplace, in un mondo dove le condizioni iniziali sono conoscibili con precisione illimitata, risulta perfettamente sensata. Tuttavia, nell'universo imperfetto in cui viviamo, dove ogni misura (al calcolatore o in laboratorio) è affetta da un errore, la caoticità di sistemi dinamici come quello di Lorenz rende, di fatto, impossibile fare previsioni ad ogni tempo. Per la natura stessa del problema, ciò che potremo fare sarà operare previsioni accurate nel breve termine, utilizzando metodi numerici e dati iniziali quanto più precisi possibile. Il fenomeno descritto è individuato, talvolta, con l'osimoro *caos deterministico*: espressione che adesso assume un significato chiaro e preciso.

## Esercizio 18

Si considerino  $N$  masse  $m_i$  univocamente individuate nello spazio, in cui è posto un sistema di riferimento cartesiano, dai vettori posizione  $\vec{r}_i = \vec{r}_i(t)$ . Si vuole studiare numericamente la soluzione al sistema di equazioni differenziali

$$\frac{d^2\vec{r}_i}{dt^2} = - \sum_{\substack{j=1 \\ j \neq i}}^3 m_j \frac{\vec{r}_i - \vec{r}_j}{\|\vec{r}_i - \vec{r}_j\|^3} \quad \text{con } i = 1, 2, 3 \quad (79)$$

utilizzando il metodo di Runge-Kutta 4. Si vuole quindi studiare la soluzione al *problema dei tre corpi* con costante di attrazione gravitazionale normalizzata  $G = 1$ . In particolare, studieremo la soluzione per due diversi set di dati iniziali e valori delle masse. Per ognuno dei due casi siamo interessati ad affrontare il problema sia da un punto di vista di analisi del metodo numerico, sia da un punto di vista strettamente fisico.

A meno di alcuni casi caratterizzati da particolari simmetrie, per generici dati iniziali e valori delle masse non trascurabili, il problema dei tre corpi non ammette soluzione analitica esprimibile per mezzo di una combinazione di funzioni elementari: risulta allora necessario procedere per via numerica. Una versione semplificata del problema è nota in letteratura come *problema ristretto dei tre corpi* e consiste nel considerare una delle tre masse come trascurabile rispetto alle due masse principali. Sotto questa ipotesi, l'analisi del problema da un punto di vista meccanico risulta più accessibile. Anzitutto, si noti che quello in esame consiste in un sistema di tre equazioni differenziali vettoriali al secondo ordine. Più esplicitamente, la (79) si scrive quindi come

$$\begin{cases} \ddot{\vec{r}}_1 = -m_2 \frac{\vec{r}_1 - \vec{r}_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{\vec{r}_1 - \vec{r}_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \ddot{\vec{r}}_2 = -m_3 \frac{\vec{r}_2 - \vec{r}_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{\vec{r}_2 - \vec{r}_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \ddot{\vec{r}}_3 = -m_1 \frac{\vec{r}_3 - \vec{r}_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{\vec{r}_3 - \vec{r}_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \end{cases}$$

Posto  $\vec{r}_i = (x_i, y_i, z_i)$  il vettore posizione della  $i$ -esima particella, esplicitando la dinamica di ogni singola componente otterremo il sistema

$$\begin{cases} \ddot{x}_1 = -m_2 \frac{x_1 - x_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{x_1 - x_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \ddot{y}_1 = -m_2 \frac{y_1 - y_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{y_1 - y_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \ddot{z}_1 = -m_2 \frac{z_1 - z_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{z_1 - z_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \ddot{x}_2 = -m_3 \frac{x_2 - x_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{x_2 - x_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \ddot{y}_2 = -m_3 \frac{y_2 - y_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{y_2 - y_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \ddot{z}_2 = -m_3 \frac{z_2 - z_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{z_2 - z_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \ddot{x}_3 = -m_1 \frac{x_3 - x_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{x_3 - x_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \\ \ddot{y}_3 = -m_1 \frac{y_3 - y_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{y_3 - y_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \\ \ddot{z}_3 = -m_1 \frac{z_3 - z_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{z_3 - z_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \end{cases}$$

La forma in cui siamo riusciti a riscrivere il sistema è ora più simile ai casi affrontati fino a qui. In particolare, abbiamo un sistema di 9 equazioni differenziali scalari al secondo ordine. L'ultimo passaggio consisterà nella trasformazione di

ogni singola ODE del sistema in un sistema dinamico in  $\mathbb{R}^2$ . Per proposizione (0.16) avremo allora il sistema dinamico in  $\mathbb{R}^{18}$

$$\left\{ \begin{array}{l} \dot{x}_1 = v_{x_1} \\ \dot{v}_{x_1} = -m_2 \frac{x_1 - x_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{x_1 - x_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \dot{y}_1 = v_{y_1} \\ \dot{v}_{y_1} = -m_2 \frac{y_1 - y_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{y_1 - y_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \dot{z}_1 = v_{z_1} \\ \dot{v}_{z_1} = -m_2 \frac{z_1 - z_2}{\|\vec{r}_1 - \vec{r}_2\|^3} - m_3 \frac{z_1 - z_3}{\|\vec{r}_1 - \vec{r}_3\|^3} \\ \dot{x}_2 = v_{x_2} \\ \dot{v}_{x_2} = -m_3 \frac{x_2 - x_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{x_2 - x_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \dot{y}_2 = v_{y_2} \\ \dot{v}_{y_2} = -m_3 \frac{y_2 - y_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{y_2 - y_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \dot{z}_2 = v_{z_2} \\ \dot{v}_{z_2} = -m_3 \frac{z_2 - z_3}{\|\vec{r}_2 - \vec{r}_3\|^3} - m_1 \frac{z_2 - z_1}{\|\vec{r}_2 - \vec{r}_1\|^3} \\ \dot{x}_3 = v_{x_3} \\ \dot{v}_{x_3} = -m_1 \frac{x_3 - x_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{x_3 - x_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \\ \dot{y}_3 = v_{y_3} \\ \dot{v}_{y_3} = -m_1 \frac{y_3 - y_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{y_3 - y_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \\ \dot{z}_3 = v_{z_3} \\ \dot{v}_{z_3} = -m_1 \frac{z_3 - z_1}{\|\vec{r}_3 - \vec{r}_1\|^3} - m_2 \frac{z_3 - z_2}{\|\vec{r}_3 - \vec{r}_2\|^3} \end{array} \right.$$

che rappresenta una riscrittura equivalente dell'equazione originaria. Una volta fornito un set di condizioni iniziali

$$\vec{r}_i^0, \vec{v}_i^0 \in \mathbb{R}^3 \quad \text{con} \quad i = 1, 2, 3$$

e specializzati i valori parametrici delle masse, il sistema determinato renderà possibile l'applicazione diretta della (73) con  $m = 18$ . Nonostante l'elevata dimensionalità del problema in esame, siamo quindi ora in grado di simulare numericamente l'evoluzione della soluzione al problema dei tre corpi utilizzando i soliti algoritmi risolutivi.

### Primo set di dati iniziali

Vogliamo simulare l'evoluzione della soluzione al problema dei tre corpi con il set di dati iniziali

$$\left\{ \begin{array}{l} \vec{r}_1^0 = (1, 0, 0) \\ \vec{r}_2^0 = (-1, 0, 0) \\ \vec{r}_3^0 = (0, 0, 0) \end{array} \right. \quad \text{e} \quad \left\{ \begin{array}{l} \vec{v}_1^0 = (0, 0.4, 0) \\ \vec{v}_2^0 = (0, -0.8, 0.7) \\ \vec{v}_3^0 = (0, -0.8, -0.7) \end{array} \right.$$

e con valori delle masse dei punti materiali fissati a

$$m_1 = 1.6 \quad \text{e} \quad m_2 = m_3 = 0.4$$

utilizzando il più preciso metodo di Runge-Kutta 4.

Anzitutto, similmente a quanto fatto fino a questo punto, si è deciso di operare un'analisi in precisione del metodo numerico in esame. Si sono allora generate coppie di soluzioni ad un passo raddoppiato l'una rispetto all'altra fino al tempo  $\bar{t} = 3$  nel range di  $N$  tale che

$$1500 \leq N < 3000 \quad \text{con} \quad N_{i+1} = N_i + 50$$

Fissato uno dei tre corpi, si sono poi calcolate le distanze (in metrica euclidea) tra la soluzione  $\vec{r}_N$  costruita per un numero di passi  $N$  e la soluzione  $\vec{r}_{2N}$  costruita per un numero di passi raddoppiato  $2N$  come

$$\Delta(\bar{t}) = \|\vec{r}_N(\bar{t}) - \vec{r}_{2N}(\bar{t})\|$$

per ogni  $N$  nel range selezionato. In questo caso, per ragioni di semplicità, si è deciso di svolgere l'analisi solo per il corpo  $m_1$ . Con passaggi del tutto analoghi è possibile ottenere risultati praticamente identici anche per  $m_2$  e  $m_3$ . Chiaramente, visti i risultati riportati nella prefazione teorica, ci si aspetta che l'errore scali in funzione del passo come la (72). Si sono quindi calcolati i logaritmi delle dispersioni e dei passi corrispondenti, per poi interpolare i dati con una retta della forma  $y = mx + q$  come segue.

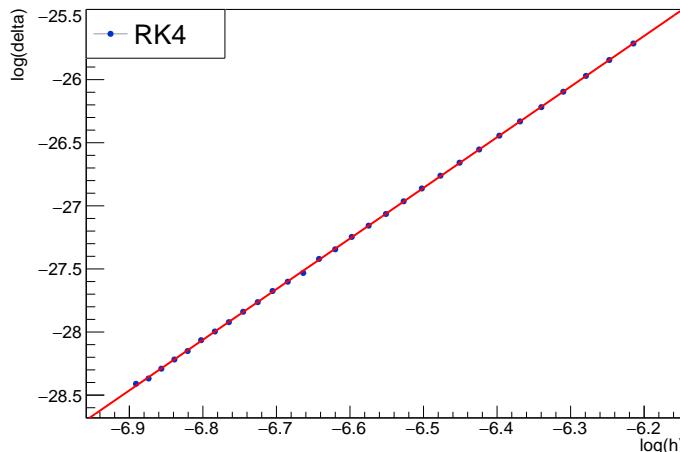


Figura 220: andamento  $\log \Delta(\bar{t})$  con RK4 per  $m_1$ : fit

Si è ottenuta la stima di parametri che segue.

$$q = -0.776 \quad \text{e} \quad m = 4.01 \approx 4$$

Il valore del coefficiente angolare stimato permette di concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 4 risulta essere al quarto ordine in  $h$ . Si noti che, vista la complessità del sistema di equazioni differenziali implementato, la verifica del corretto andamento dell'errore in RK4 può essere anche utilizzata al contrario, ossia come verifica della corretta implementazione del sistema. Evidentemente, se si procede per questa strada è necessario assumere che il metodo numerico sia stato implementato correttamente: fatto ormai verificato diverse volte negli esercizi precedenti.

Come secondo step, siamo ora interessati alla visualizzazione delle soluzioni al sistema in esame nello spazio tridimensionale. Si sono quindi generate le soluzioni con il metodo di Runge-Kutta 4 per diversi valori temporali. Per semplicità, si è scelto un valore del numero di passi fissato a  $N = 10000$ , che ha permesso di lavorare con un passo  $h$  sufficientemente piccolo per tutti i tempi finali selezionati. Di seguito è riportata l'evoluzione delle traiettorie dei tre corpi per alcuni valori temporali.

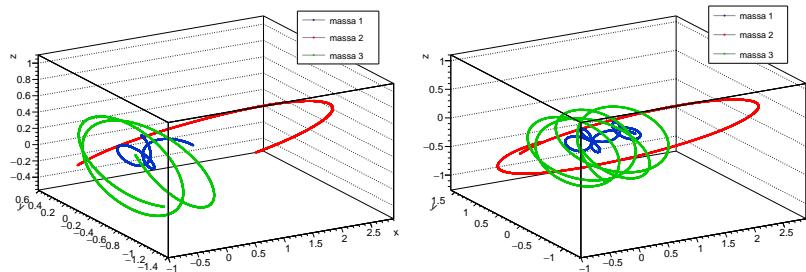


Figura 221: traiettorie per  $\bar{t} = 8$

Figura 222: traiettorie per  $\bar{t} = 15$

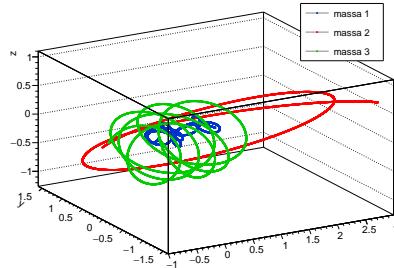


Figura 223: traiettorie per  $\bar{t} = 20$

Come è possibile notare, le traiettorie dei tre punti materiali soggetti alla sola interazione gravitazionale reciproca evolvono in modo caotico nello spazio tridimensionale. Il caso in esame, allora, fa pensare di essere di fronte ad un nuovo caso di sistema dinamico caotico. Questo fatto suscita particolare interesse in quanto l'evoluzione del moto nel caso di due soli corpi appare, invece, ben più regolare e prevedibile, qualunque sia il valore dei dati iniziali. Inoltre, si noti che, fino al tempo  $\bar{t} = 20$ , le traiettorie di  $m_1$  e  $m_3$  sembrano rimanere confinate in una regione limitata dello spazio fisico tridimensionale. A differenza del caso dell'esercizio precedente, tuttavia, la possibilità di giustificare o garantire un certo tipo di comportamento qualitativo per mezzo di considerazioni analitiche è minata alla base dalla complessità del sistema dinamico che governa il moto. Come già accennato in precedenza, un'analisi più sistematica può essere effettuata sotto ipotesi di una massa trascurabile rispetto alle altre due. Sfortunatamente, nel caso in esame, le tre masse presentano tutte un ordine di grandezza comparabile. Procediamo allora con considerazioni qualitative a par-

tire dai risultati delle simulazioni. Di seguito è riportato un plot più dettagliato e da una diversa angolazione per  $\bar{t} = 20$ .

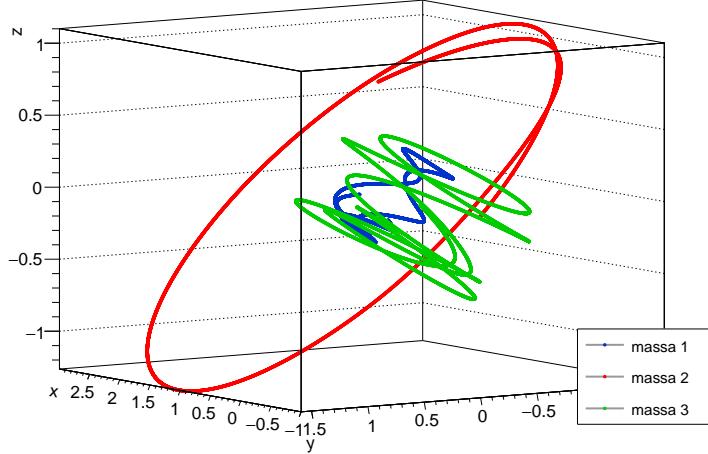


Figura 224: traiettorie per  $\bar{t} = 20$ , diversa angolazione

Come è possibile apprezzare da questa visuale, l'unica massa che tende a seguire una traiettoria più regolare è la massa  $m_2$ . Inoltre, a differenza delle altre due masse,  $m_2$  sembra non restare confinata in una regione limitata dello spazio fisico. Al fine di verificarlo numericamente, si è simulato il moto per un tempo più grande, di  $\bar{t} = 80$ , ottenendo i risultati che seguono.

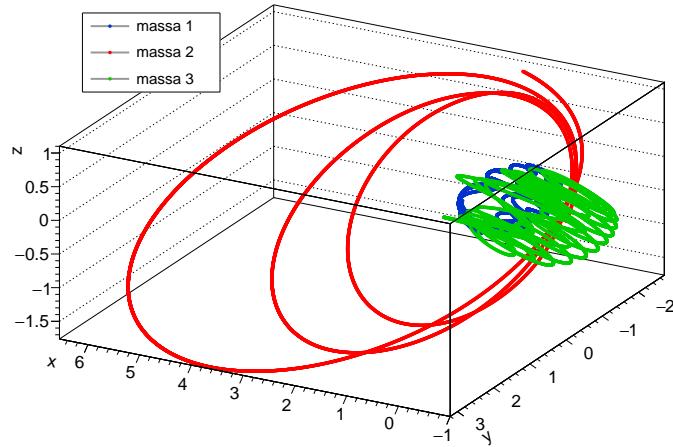


Figura 225: traiettorie per  $\bar{t} = 80$

I risultati ottenuti rendono conto qualitativamente di quanto si è ipotizzato: anche per tempi grandi,  $m_1$  e  $m_3$  rimangono confinate in una regione limitata dello spazio, mentre  $m_2$  segue una traiettoria più regolare, caratterizzata da una pseudo-periodicità.

Nonostante non sia di fatto possibile disporre dell'analisi per giustificare i comportamenti descritti, il sistema in esame ha il pregio di essere soggetto alle sole interazioni gravitazionali. Sappiamo che, date due masse  $m_i$  e  $m_j$ , posta la normalizzazione  $G = 1$  e posto  $\hat{r}$  il versore radiale, la forza gravitazionale generata da  $m_i$  che agisce su  $m_j$  si scrive come

$$\vec{F}_{i,j} = -\frac{m_i m_j}{\|\vec{r}_i - \vec{r}_j\|^2} \hat{r}$$

Scrivendo più compattamente la distanza (in norma euclidea) tra le due masse come  $r_{i,j} := \|\vec{r}_i - \vec{r}_j\|$  avremo che

$$\vec{F}_{i,j} = -\frac{m_i m_j}{r_{i,j}^2} \hat{r}$$

In questo modo, è ora più visibile il fatto che  $F_{i,j}$  dipenda dalla sola posizione reciproca tra  $m_i$  e  $m_j$ . Questo ci garantisce, di fatto, che l'interazione gravitazionale ammette potenziale. Fissato lo zero del potenziale all'infinito, ossia posto  $U_{i,j}(\infty) = 0$ , per definizione di energia potenziale avremo

$$U_{i,j}(r_{i,j}) = - \int_{\infty}^{r_{i,j}} -\frac{m_i m_j}{r^2} dr = -\frac{m_i m_j}{r_{i,j}}$$

Ma allora, il sistema in esame è un sistema conservativo, da cui segue che vale il teorema di conservazione dell'energia meccanica. A differenza dei casi studiati in precedenza, caratterizzati da un singolo punto materiale in uno spazio monodimensionale (inteso come singolo grado di liberà), in questo caso dovremo combinare le energie potenziali e cinetiche dei tre corpi che compongono il sistema. L'energia cinetica totale del sistema si scriverà come

$$T_{tot} = T_1 + T_2 + T_3 = \frac{1}{2} m_1 \|\vec{v}_1\|^2 + \frac{1}{2} m_2 \|\vec{v}_2\|^2 + \frac{1}{2} m_3 \|\vec{v}_3\|^2$$

L'energia potenziale, invece, andrà contata una sola volta per ogni coppia di interazioni. Avremo quindi

$$U_{tot} = U_{1,2} + U_{2,3} + U_{1,3} = - \left( \frac{m_1 m_2}{r_{1,2}} + \frac{m_2 m_3}{r_{2,3}} + \frac{m_1 m_3}{r_{1,3}} \right)$$

Per la conservazione dell'energia meccanica ci aspettiamo, allora, che l'energia totale del sistema gravitazionale a tre corpi data da

$$E := T_{tot} + U_{tot}$$

si conservi nel tempo lungo le soluzioni del sistema dinamico. Al fine di verificare questo fatto, si sono generate le soluzioni fino al tempo  $\bar{t} = 15$  con RK4, fissando  $N = 10000$  come nei precedenti casi, assicurandoci di lavorare con un valore del passo  $h$  sufficientemente piccolo. Si è quindi valutata la funzione energia

$$E = E(\vec{r}_j, \vec{v}_j) \quad \forall j = 1, \dots, N$$

nelle soluzioni generate, per poi interpolare l'energia in funzione del tempo con la mappa affine  $E = mt + q$ . Di seguito sono riportati i risultati ottenuti.

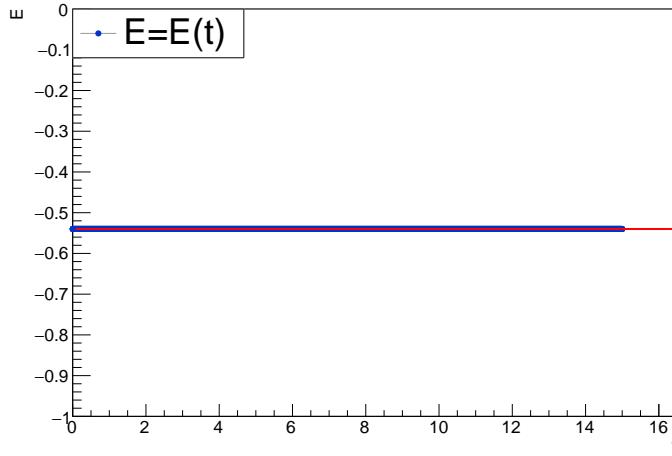


Figura 226:  $E(t)$  con Runge-Kutta 4 fino a  $\bar{t} = 15$ : fit

Si è ottenuta la stima di parametri che segue.

$$q = -0.54 \quad \text{e} \quad m = -7.9 \cdot 10^{-14} \approx 0$$

Tenendo conto dell'ordine della precisione doppia utilizzata, possiamo concludere che i valori stimati dei parametri rappresentino una verifica quantitativa del teorema di conservazione dell'energia meccanica. Si è notato, tuttavia, che piazzando i risultati nella loro scala naturale il risultato è il seguente.

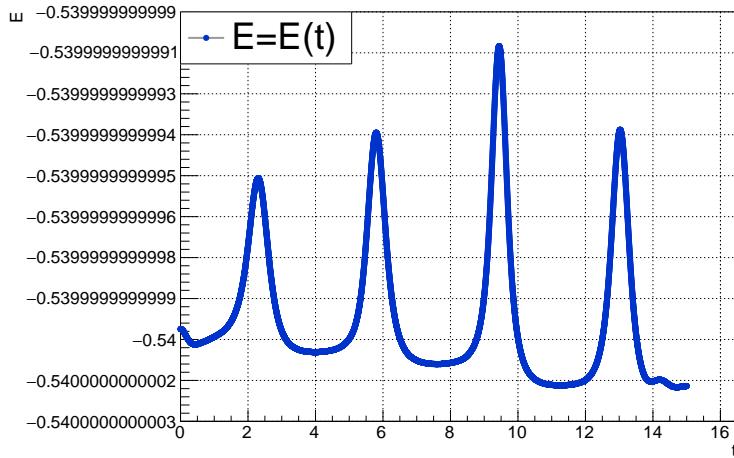


Figura 227:  $E(t)$  con Runge-Kutta 4 fino a  $\bar{t} = 15$ : ingrandimento

Come è possibile notare, l'andamento dell'energia valutata nella soluzione costruita numericamente appare oscillante e quasi periodico. Non è difficile notare, però, che l'ordine di grandezza dell'ampiezza delle oscillazioni, visibile dalla scala verticale del grafico, risulta  $\sim 10^{-13}$ , ossia confrontabile con la doppia precisione utilizzata per il calcolo. Segue che le fluttuazioni osservate non hanno

alcuna rilevanza o corrispondenza fisica: come si ha già avuto modo di notare in precedenza, la loro presenza è solo un sintomo del fatto che RK4 commette un errore nella costruzione delle soluzioni, congiuntamente alla presenza di intervalli in cui le soluzioni stesse variano bruscamente. D'altra parte, in effetti, è proprio la convergenza del fit di figura 226 che ci assicura la poca rilevanza fisica dell'effetto che si osserva in figura 227.

### Secondo set di dati iniziali

Vogliamo simulare l'evoluzione della soluzione al problema dei tre corpi con il set di dati iniziali

$$\begin{cases} \vec{r}_1^0 = (1, 0, 0) \\ \vec{r}_2^0 = (-1, 0, 0) \\ \vec{r}_3^0 = (0, 0, 0) \end{cases} \quad \text{e} \quad \begin{cases} \vec{v}_1^0 = (0, 0.15, -0.15) \\ \vec{v}_2^0 = (0, -0.15, 0.15) \\ \vec{v}_3^0 = (0, 0, 0) \end{cases}$$

e con valori delle masse dei punti materiali fissati a

$$m_1 = m_2 = m_3 = 0.3$$

utilizzando il più preciso metodo di Runge-Kutta 4.

Anche in questo caso, si è deciso di operare un'analisi in precisione del metodo numerico in esame. Si sono quindi eseguiti passaggi analoghi a quelli svolti per il primo set di dati, calcolando i logaritmi delle distanze tra la soluzione  $\vec{r}_N$  costruita per un numero  $N$  di passi e la soluzione  $\vec{r}_{2N}$  costruita per un numero di passi raddoppiato  $2N$ . L'operazione è stata eseguita per la sola massa  $m_1$  e fino al tempo  $\bar{t} = 2$ , nel range di  $N$  tale che

$$1500 \leq N < 3000 \quad \text{con} \quad N_{i+1} = N_i + 50$$

interpolando i risultati con la retta  $y = mx + q$ , ottenendo i seguenti risultati.

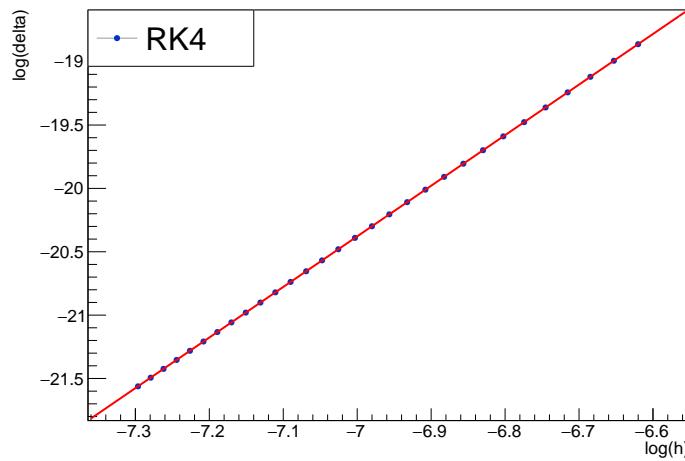


Figura 228: andamento  $\log \Delta(\bar{t})$  con RK4 per  $m_1$ : fit

Si è ottenuta la stima di parametri che segue.

$$q = 7.57 \quad \text{e} \quad m = 3.99 \approx 4$$

Il valore del coefficiente angolare stimato permette di concludere che, coerentemente con quanto ci si aspetta, il metodo di Runge-Kutta 4 risulta essere al quarto ordine in  $h$ .

Anche in questo caso, come secondo step, siamo interessati alla visualizzazione delle soluzioni al sistema in esame nello spazio tridimensionale. Si sono quindi generate le soluzioni con il metodo di Runge-Kutta 4 per diversi valori temporali. Per semplicità, si è scelto un valore del numero di passi fissato a  $N = 10000$ , che ha consentito di lavorare con un passo  $h$  sufficientemente piccolo per tutti i tempi finali selezionati. Di seguito è riportata l'evoluzione delle traiettorie dei tre corpi per alcuni valori temporali.

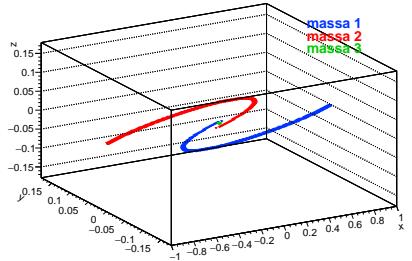


Figura 229: traiettorie per  $\bar{t} = 2$

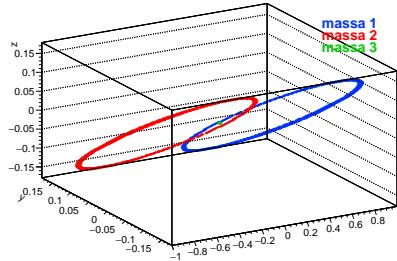


Figura 230: traiettorie per  $\bar{t} = 5$

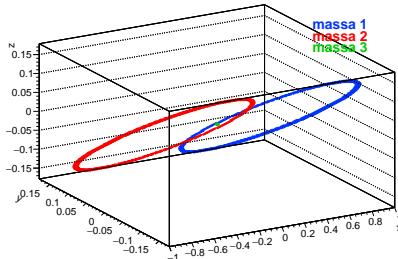


Figura 231: traiettorie per  $\bar{t} = 15$

Come è possibile notare, a differenza del caso precedente, per questo set di dati iniziali la traiettoria di ogni massa mostra un andamento molto più regolare e prevedibile. Inoltre, è facile verificare che il moto avviene, per ogni tempo, sempre nello stesso piano dello spazio tridimensionale. Questo fatto suggerisce che, per questo set di dati iniziali, sia possibile ridefinire il sistema di riferimento cartesiano ruotandolo rigidamente di un certo angolo  $\theta$  rispetto a quello iniziale, al fine di ottenere una riformulazione equivalente del problema con un numero ridotto di gradi di libertà. In altre parole, il problema fisico in esame vive in uno spazio bidimensionale. Di seguito sono riportate le traiettorie per  $\bar{t} = 20$ .

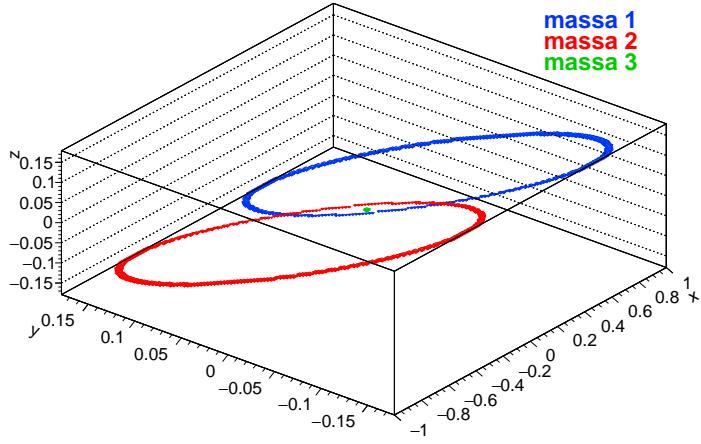


Figura 232: traiettorie per  $\bar{t} = 20$

Come si nota dal grafico, le masse  $m_1$  e  $m_2$  assumono un andamento a forma di ellisse nello spazio fisico tridimensionale. Il loro moto risulta, quindi, limitato e periodico per ogni tempo. Dalle simulazioni numeriche, invece, la massa  $m_3$  non si muove, restando in quiete nella posizione  $\vec{r} = \vec{r}_3^0$  coincidente ad uno dei dati iniziali. Al fine di apprezzare e verificare più nel dettaglio le caratteristiche di regolarità discusse, di seguito è riportata l'evoluzione per  $\bar{t} = 50$ .

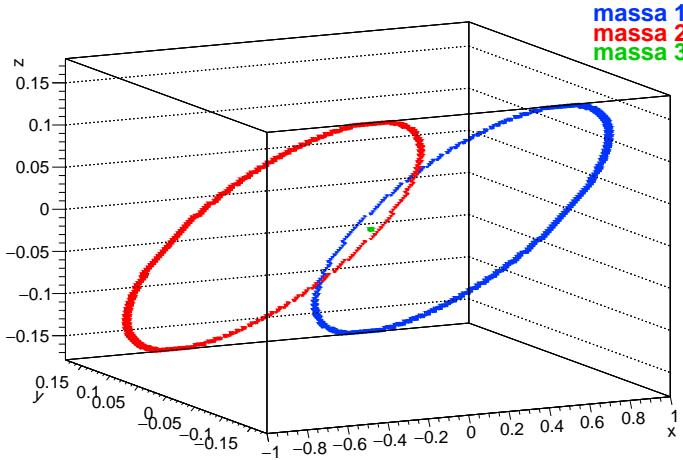


Figura 233: traiettorie per  $\bar{t} = 50$

Anche da questa angolazione e per un tempo maggiore, risultano confermate tutte le caratteristiche qualitative del moto delle tre masse. Non è difficile verificare che il moto è il medesimo per ogni tempo selezionato. Ovviamente, al fine di svolgere questa verifica in modo consistente, risulta necessario diminuire il passo di integrazione all'aumentare del tempo finale, tenendo conto dell'analisi in precisione effettuata in precedenza.

Siamo quindi interessati a studiare più nel dettaglio il comportamento della massa  $m_3$ . Evidentemente, dalle simulazioni numeriche, la massa in esame risulta in equilibrio statico per ogni tempo. Vogliamo, dunque, trovare una caratterizzazione dell'equilibrio che sia semplice da verificare computazionalmente, al fine di chiarire meglio le condizioni sotto cui vale il fenomeno osservato. A seguito di diverse simulazioni si è notato un fatto rilevante, che introduciamo con la seguente proposizione.

**Proposizione 0.18** (condizione necessaria di equilibrio). *Si considerino tre masse  $m_1$ ,  $m_2$  e  $m_3$  soggette alla sola interazione gravitazionale. Detto  $U_{1,3}$  il potenziale dato dall'interazione tra  $m_1$  e  $m_3$ , e  $U_{2,3}$  il potenziale dato dall'interazione tra  $m_2$  e  $m_3$ , si consideri la quantità*

$$U_3 := U_{1,3} - U_{2,3}$$

*Se  $m_3$  è in equilibrio statico e  $m_1 = m_2$ , allora*

$$\frac{dU_3}{dt} = 0 \quad \forall t \geq 0$$

*lungo le soluzioni del sistema.*

*Dimostrazione.* Anzitutto, notiamo che per potenziali gravitazionali vale

$$U_{1,3} = U_{1,3}(r_1(t)) \quad \text{e} \quad U_{2,3} = U_{2,3}(r_2(t))$$

dove  $r_1$  è la distanza tra  $m_1$  e  $m_3$ , e  $r_2$  è la distanza tra  $m_2$  e  $m_3$ . Ma allora, calcolando più esplicitamente la derivata temporale di  $U_3$  avremo che

$$\begin{aligned} \frac{dU_3}{dt} &= \frac{dU_{1,3}}{dt} - \frac{dU_{2,3}}{dt} = \\ &= \frac{dU_{1,3}}{dr_1} \frac{dr_1}{dt} - \frac{dU_{2,3}}{dr_2} \frac{dr_2}{dt} \end{aligned} \quad (80)$$

per linearità dell'operatore derivata e regola della catena. Per la prima ipotesi, la massa  $m_3$  è in equilibrio: siccome le forze sono dirette lungo la stessa direzione per ogni tempo possiamo scrivere

$$F_{1,3} - F_{2,3} = 0$$

per il primo principio della dinamica. Siccome entrambe le forze gravitazionali ammettono potenziale potremo scrivere

$$-\frac{dU_{1,3}}{dr_1} + \frac{dU_{2,3}}{dr_2} = 0 \quad \iff \quad \frac{dU_{1,3}}{dr_1} - \frac{dU_{2,3}}{dr_2} = 0$$

Inoltre, notiamo che, esplicitando la relazione di equilibrio sulle forze avremo

$$-\frac{m_1 m_3}{r_1^2} = -\frac{m_2 m_3}{r_2^2}$$

Ma per la seconda ipotesi introdotta, si ha che  $m_1 = m_2$ , da cui segue che

$$r_1(t) = r_2(t) \quad \implies \quad \frac{dr_1}{dt} = \frac{dr_2}{dt} = \frac{dr}{dt}$$

Ma allora, la relazione (80) si potrà scrivere come

$$\dot{U}_3 = \frac{dr}{dt} \left( \frac{dU_{1,3}}{dr_1} - \frac{dU_{2,3}}{dr_2} \right) = 0 \quad \forall t \geq 0$$

per la condizione di equilibrio sui potenziali ricavata in precedenza.  $\square$

In altre parole, siamo riusciti a mostrare che, sotto le due ipotesi esplicitate, la quantità  $U_3$  è una costante del moto per il sistema. Ma allora, siccome l'ipotesi sulle masse, per il set di condizioni iniziali in esame, è sempre verificata, disponiamo adesso di una condizione necessaria di equilibrio per la massa  $m_3$ . Il fatto rilevante consiste nel notare che tale condizione è di banale verifica computazionale una volta costruite le soluzioni. Come prima cosa, allora, si è verificata tale condizione costruendo le soluzioni fino a  $\bar{t} = 15$  con un numero di passi pari a  $N = 10000$ . Si è quindi valutato per punti

$$U_3(t_i) = U_{1,3}(t_i) - U_{2,3}(t_i)$$

per poi interpolare i dati ottenuti in funzione del tempo con la retta  $U_3 = mt + q$ . Di seguito sono riportati i risultati ottenuti.

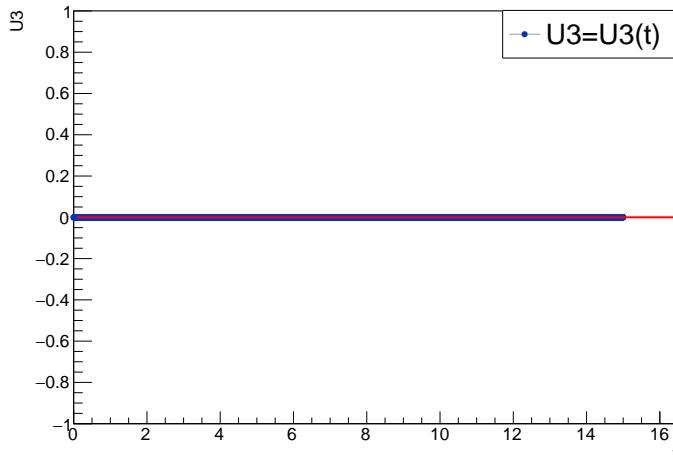


Figura 234:  $U_3(t)$  con Runge-Kutta 4 fino a  $\bar{t} = 15$

Si è ottenuta la stima di parametri che segue.

$$q = 0 \quad \text{e} \quad m = 0$$

Come è possibile notare, la funzione  $U_3$  risulta identicamente nulla nel tempo, da cui segue che la sua derivata temporale sarà anch'essa nulla. Dalla verifica numerica della proposizione 0.18, evidentemente, non segue necessariamente che  $m_3$  sia in equilibrio: per essere certi di questo fatto bisognerebbe mostrare l'implicazione inversa della proposizione. Tuttavia, possiamo concludere che sia più che ragionevole supporre l'equilibrio statico di  $m_3$ . A questo punto, siamo interessati a studiare la natura dell'equilibrio della terza massa. A tale scopo, si sono generate le soluzioni al problema dei tre copri con posizione iniziale di  $m_3$  perturbata di una quantità  $\varepsilon$  piccola in modulo, ossia con dati iniziali

$$\begin{cases} \vec{r}_1^0 = (1, 0, 0) \\ \vec{r}_2^0 = (-1, 0, 0) \\ \vec{r}_3^\varepsilon = (\varepsilon, \varepsilon, \varepsilon) \end{cases} \quad \text{e} \quad \begin{cases} \vec{v}_1^0 = (0, 0.15, -0.15) \\ \vec{v}_2^0 = (0, -0.15, 0.15) \\ \vec{v}_3^0 = (0, 0, 0) \end{cases}$$

e medesimi valori delle masse. In particolare, si è fissato  $\varepsilon = 10^{-6}$ , generando soluzioni al problema dei tre corpi fino a  $\bar{t} = 3$  con  $N = 10000$ , come segue.

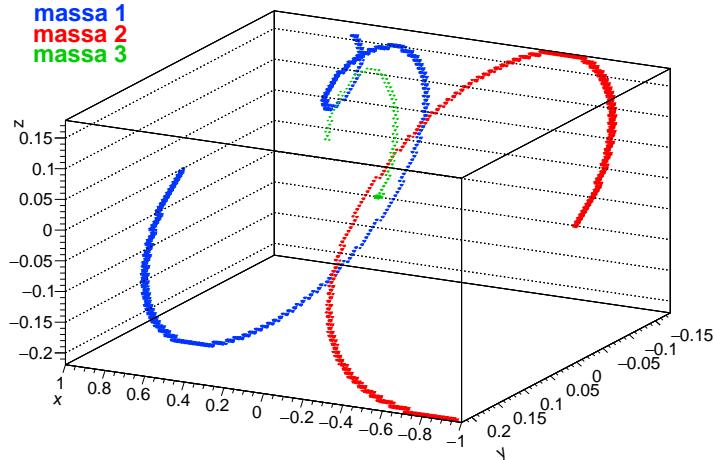


Figura 235: traiettorie con dati iniziali perturbati per  $\bar{t} = 3$

Come è possibile notare, per una variazione infinitesima della posizione iniziale di  $m_3$ , si sono ottenute traiettorie rapidamente divergenti da un intorno del punto di equilibrio stabile nello spazio tridimensionale. D'altra parte, è immediato verificare che l'andamento di  $U_3(t)$  nelle soluzioni perturbate è il seguente.

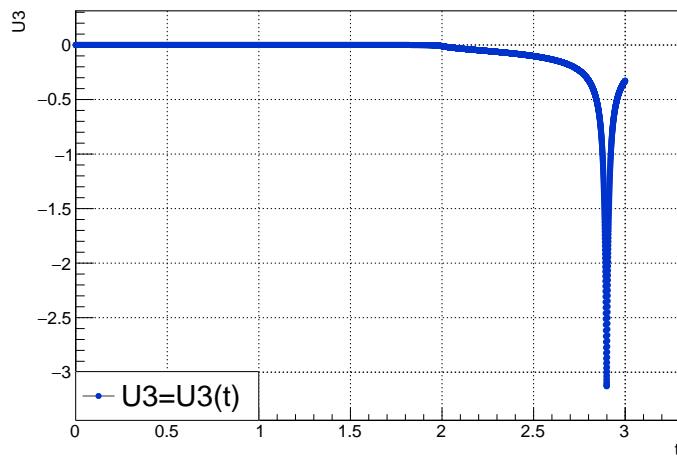


Figura 236:  $U_3(t)$  con dati iniziali perturbati fino a  $\bar{t} = 3$

Evidentemente, in questo caso, l'andamento di  $U_3$  non risulta più costante per ogni tempo positivo, ma assume una forma irregolare a partire da  $t \approx 2$ , da

cui segue che la derivata sarà, in generale, diversa dal valore nullo a partire da quel valore temporale. Per la contronominale della proposizione 0.18 avremo che, per i nuovi dati iniziali perturbati,  $m_3$  non è più in equilibrio statico, coerentemente con quanto si osserva dalla figura 235. Inoltre, vista la rapida divergenza dall'equilibrio, risulta possibile classificare l'equilibrio di  $m_3$  come instabile. Si noti, tuttavia, che l'andamento evidenziato dal grafico 236 risulta, quantomeno, insolito. Visto l'andamento costante per i primi valori temporali, il grafico porta a pensare che, per tutti i tempi più piccoli del tempo critico, la massa  $m_3$  continui a restare in equilibrio. In effetti, svolgendo una simulazione delle traiettorie per  $\bar{t} = 1.8$  il risultato è il seguente.

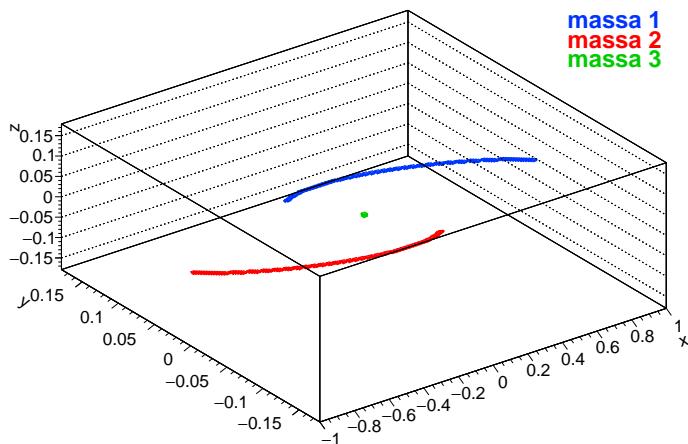


Figura 237: traiettorie con dati iniziali perturbati per  $\bar{t} = 1.8$

Coerentemente con quanto ipotizzato, la massa  $m_3$  rimane in equilibrio statico per tutti i valori più piccoli del tempo critico anche per i dati iniziali perturbati. Quanto si osserva è più che ragionevole se si pensa al fatto che, per i primi istanti temporali, le masse  $m_1$  e  $m_2$  risultano ancora distanti dalla massa  $m_3$ . L'effetto dell'interazione, allora, risulterà trascurabile fino a quando la distanza reciproca non sarà tale da smuovere  $m_3$  dalla posizione di equilibrio in modo significativo. Vista la natura instabile dell'equilibrio, dall'istante critico in poi il sistema perderà irreversibilmente la propria simmetria, evolvendo in modo caotico ed imprevedibile.

Nota la natura del punto di equilibrio siamo ora interessati, come nel caso precedente, a verificare il teorema di conservazione dell'energia meccanica. Si sono quindi generate le soluzioni fino al tempo  $\bar{t} = 15$  con RK4, fissando  $N = 10000$  come nei precedenti casi, al fine di assicurarsi un valore del passo  $h$  sufficientemente piccolo. Si è quindi valutata la funzione energia

$$E = E(\vec{r}_j, \vec{v}_j) \quad \forall j = 1, \dots, N$$

nelle soluzioni generate, per poi interpolare l'energia in funzione del tempo con la mappa affine  $E = mt + q$ . Di seguito sono riportati i risultati ottenuti.

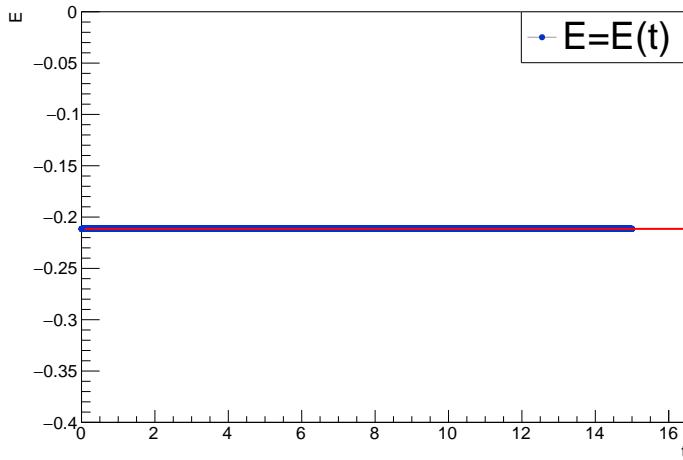


Figura 238:  $E(t)$  con Runge-Kutta 4 fino a  $\bar{t} = 15$ : fit

Si è ottenuta la stima di parametri che segue.

$$q = -0.211 \quad \text{e} \quad m = -3.07 \cdot 10^{-8} \approx 0$$

Possiamo quindi concludere che i valori stimati dei parametri rappresentino una verifica quantitativa del teorema di conservazione dell'energia meccanica. In questo caso, tuttavia, a differenza del set di dati precedente, la stima di  $m$  non raggiunge la precisione doppia utilizzata per il calcolo. Al fine di indagare meglio il fenomeno, si sono quindi plottati gli stessi dati nella loro scala naturale, ottenendo quanto segue.

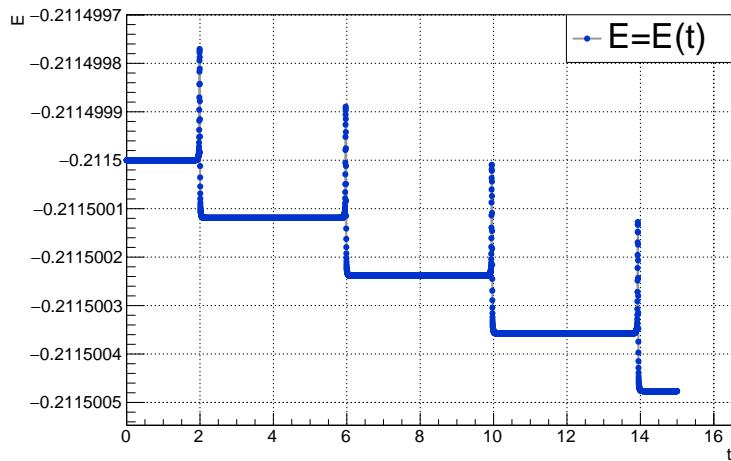


Figura 239:  $E(t)$  con Runge-Kutta 4 fino a  $\bar{t} = 15$ : ingrandimento

Come si nota, anche in questo caso, i risultati in una scala più dettagliata mostrano un andamento non rettilineo ma periodico. In particolare, l'energia sembra assumere un andamento decrescente a gradini per ogni tempo. In questo

caso, non è possibile liquidare il fenomeno con la stessa facilità del caso precedente, in quanto le variazioni di energia non sono confrontabili con la precisione doppia utilizzata per il calcolo. Notiamo, quindi, che la funzione energia assume dei picchi per tutti i valori di tempo  $t$  tali che

$$t = 4n + 2 \quad \forall n \in \mathbb{N}$$

Osservando la posizione finale delle masse  $m_1$  e  $m_2$  per i valori temporali critici appena individuati si è notato che, in corrispondenza di tali valori, le masse in moto ellittico si trovano sempre a distanza minima dalla massa in quiete  $m_3$ . Si è quindi deciso di verificare computazionalmente questa congettura, calcolando

$$r_{1,3}(t_i) = \|\vec{r}_1(t_i) - \vec{r}_3(t_i)\| \quad \forall i = 1, \dots, N$$

con  $N = 10000$ . Per uno degli steps centrali della dimostrazione della 0.18 sappiamo, infatti, che per il nostro set di dati iniziali vale

$$r_1(t) = r_2(t) \implies \|\vec{r}_1 - \vec{r}_3\| = \|\vec{r}_2 - \vec{r}_3\|$$

per ogni tempo, poiché  $m_3$  è in quiete per ogni tempo. Ma allora, lo studio della quantità calcolata sarà completamente informativa sia per lo studio della distanza tra  $m_1$  e  $m_3$ , sia per lo studio della distanza tra  $m_2$  e  $m_3$ . Si sono quindi plottate le distanze  $r_{1,3}$  al variare del tempo, ottenendo l'andamento mostrato nel grafico che segue.

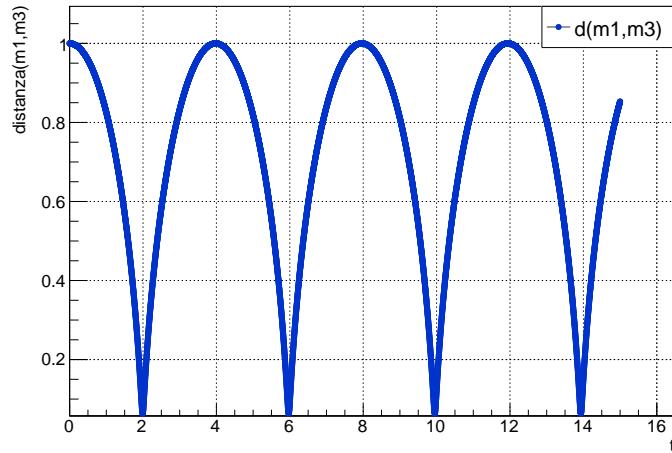


Figura 240:  $d(m_1, m_3) = d(m_2, m_3)$  fino a  $\bar{t} = 15$

Come ipotizzato, la distanza tra  $m_1$  e  $m_3$  è minima per tutti i valori dei tempi critici individuati. Dal confronto tra il grafico 239 e il grafico 240 possiamo allora concludere che l'andamento inatteso dell'energia si verifichi in corrispondenza di tutti e i soli minimi delle distanze tra le masse in moto e quella in quiete. Evidentemente, in corrispondenza dei minimi di distanza, l'interazione gravitazionale sarà più forte. Possiamo allora ipotizzare che per tali valori le soluzioni del sistema dinamico siano soggette a variazioni più brusche: per quanto già sappiamo, questo fatto, combinato all'errore di un metodo approssimato come

RK4, causa una ricostruzione meno precisa delle soluzioni numeriche. Pertanto, l'andamento inatteso non trascurabile dell'energia meccanica nel tempo può essere spiegato in questo modo.

In definitiva, abbiamo avuto modo di osservare che il sistema dinamico che descrive l'evoluzione delle soluzioni al problema dei tre corpi risulta molto più complesso da studiare rispetto al solo caso di una coppia di masse. In particolare, il sistema si presenta, di fatto, nella forma di un sistema caotico. Questa affermazione si può verificare in modo più sistematico facendo uno studio simile a quello dell'esercizio precedente, anche se già l'evoluzione delle traiettorie per dati perturbati mostrata in figura 235 permette di verificare qualitativamente tale fenomeno. Inoltre, confrontando due set diversi di dati iniziali, si ha avuto modo di verificare che non sempre il sistema evolve in modo poco prevedibile: la regolarità delle traiettorie è strettamente legata ai dati iniziali, coerentemente con la definizione di caoticità di un sistema. Esistono valori dei parametri, infatti, tali che la simmetria del sistema porti alla generazione di costanti del moto, come quella individuata in proposizione 0.18, che ha permesso uno studio più approfondito delle condizioni di equilibrio di una delle tre masse in gioco. Per casi come questi, nei quali l'analisi del sistema meccanico è tutto meno che banale, i metodi numerici risultano di capitale importanza per affrontare il problema.

## Zeri di funzione

Siamo interessati allo studio di diversi metodi numerici in grado di determinare gli zeri di una funzione.

**Definizione 0.19.** Si consideri la funzione complessa  $f : \Omega \subseteq \mathbb{C} \rightarrow \mathbb{C}$ . Diremo che  $z \in \mathbb{C}$  è uno *zero* o una *radice* di  $f$  se vale

$$f(z) = 0$$

Si consideri, dunque, il caso particolare di una funzione reale di variabile reale della forma  $f : \Omega \subseteq \mathbb{R} \rightarrow \mathbb{R}$ . In tal caso, la definizione 0.19 si riduce a quella usuale con l'interpretazione geometrica nota, posta la sostituzione  $z = x \in \mathbb{R}$ .

**Teorema 0.20** (di Bolzano/degli zeri). *Si consideri una funzione reale di variabile reale  $f : [a, b] \rightarrow \mathbb{R}$  tale che  $f$  sia continua in  $[a, b]$  e  $f(a)f(b) < 0$ , ossia tale che che assuma valori opposti agli estremi. Allora*

$$\exists x_0 \in [a, b] \quad \text{tale che} \quad f(x_0) = 0$$

Si noti che il teorema 0.20 non garantisce l'unicità dello zero, ma soltanto l'esistenza. Per l'unicità è necessaria l'aggiunta dell'ipotesi che  $f$  sia una funzione strettamente monotona nell'intervallo di definizione  $[a, b]$ . Si noti, inoltre, che queste sono tutte condizioni sufficienti ma non necessarie per l'esistenza degli zeri: basti pensare ad una funzione definita a tratti o che presenti punti isolati appartenenti all'asse delle ascisse.

Esistono almeno due metodi iterativi che possiamo individuare per la ricerca di zeri nel caso di funzioni di variabile reale.

### Metodo di bisezione

Si supponga una funzione reale  $f$  tale che valgano le ipotesi del teorema di Bolzano. Si supponga, inoltre, che lo zero all'interno dell'intervallo  $[a, b]$  sia unico, aggiungendo, ad esempio, l'ipotesi di stretta monotonia. Il metodo di bisezione consiste nel calcolare il punto medio di  $[a, b]$  come

$$x_{\text{med}} := \frac{a + b}{2}$$

Se  $f(x_{\text{med}}) = 0$  allora, per definizione,  $x_{\text{med}}$  è lo zero di  $f$  cercato, altrimenti si riapplica la medesima procedura reiterando il metodo in uno dei due intervalli  $[a, x_{\text{med}}]$  oppure  $[x_{\text{med}}, b]$ , scegliendo quello nel quale  $f$  assume valori discordi agli estremi. Se lo zero in  $[a, b]$  è unico, infatti, l'algoritmo restringe iterativamente l'intervallo in cui siamo certi essere presente lo zero per teorema di Bolzano. L'algoritmo di bisezione si interrompe quando

$$|a_n - b_n| < \varepsilon$$

ossia quando la misura dell'intervallo all' $n$ -esima iterazione è più piccola di un valore di precisione  $\varepsilon$  scelto, e restituisce come stima dello zero il punto medio

$$\tilde{x}_0 = \frac{a_n + b_n}{2}$$

Dalla descrizione dell'algoritmo è evidente che il metodo di bisezione abbia struttura ricorsiva. La sua implementazione può quindi essere compattata nella scrittura di una funzione che operi per ricorsione. Algoritmi, come quello della bisezione, che necessitano di specificare due valori iniziali che racchiudono ciò che si vuole cercare, sono detti *metodi bracketing*. Il metodo di bisezione è il metodo algoritmico più semplice da utilizzare per la ricerca degli zeri di funzione. Tipicamente, tuttavia, risulta essere particolarmente inefficiente, in quanto necessita di un numero di iterazioni non trascurabile o comunque superiore ad altri metodi per stimare uno zero a precisioni elevate.

### Metodo di Newton-Raphson

Si supponga una funzione reale  $f$  tale che ammetta un unico zero all'interno dell'intervallo di definizione  $[a, b]$ . Sia  $f$  derivabile all'interno dell'intervallo e sia  $f'$  la sua funzione derivata tale che  $f'(x) \neq 0$  per ogni  $x \in [a, b]$ . Il metodo di Newton-Raphson consiste nell'approssimare la funzione  $f$  nella sua retta tangente in un punto di partenza  $x_0 \in [a, b]$ , per poi calcolare lo zero  $x_1$  di quest'ultima. Viene poi calcolata di nuovo la tangente ad  $f$  nel punto  $x_1$ , calcolandone il nuovo zero. Al generico passo  $n$  si ha quindi la formula ricorsiva

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (81)$$

che all'aumentare di  $n$  restituisce un valore  $x_n$  che converge allo zero di  $f$ . Il metodo procede in modo iterativo fino a quando

$$|x_{n+1} - x_n| < \varepsilon$$

ossia fino a quando lo scarto tra una stima dello zero e la stima al passo successivo all' $n$ -esima iterazione è più piccolo di un valore di precisione  $\varepsilon$  scelto, e restituisce come stima dello zero il punto  $x_n$ . Si noti che il metodo di Newton-Raphson richiede ipotesi più stringenti rispetto al metodo di bisezione per garantire l'implementazione. Oltre all'ipotesi di esistenza di un solo zero in  $[a, b]$  si richiede che  $f$  sia derivabile in tutto l'intervallo di definizione e che sia nota la forma analitica della derivata prima. Il fatto che sia necessario un quantitativo maggiore di informazione su  $f$  rende intuitivo il fatto che, in generale, il metodo di Newton-Raphson abbia una velocità di convergenza maggiore rispetto al metodo di bisezione a parità di precisione  $\varepsilon$  scelta. Equivalentemente, Newton-Raphson necessita di un numero minore di iterazioni per convergere. D'altra parte, come si avrà modo di verificare, il metodo di Newton-Raphson non è un metodo globalmente convergente, ossia la convergenza (o la sua velocità) e il corretto funzionamento dipendono dal valore del dato iniziale  $x_0$  con il quale viene svolta la prima iterazione. Il metodo di bisezione, invece, si dimostra essere globalmente convergente. Si noti, infine, che il funzionamento del metodo richiede l'esistenza di un solo zero nell'intervallo. Questo fatto è ottenibile richiedendo, ad esempio, che sia verificato il teorema 0.20 con l'aggiunta della stretta monotonia. Tuttavia, il teorema dà soltanto delle condizioni sufficienti, ossia esistono funzioni che ammettono un solo zero in  $[a, b]$  tali da essere discontinue nell'intervallo o tali da assumere valori concordi agli estremi. La conoscenza, anche solo qualitativa, della forma della funzione  $f$  permette di applicare il metodo di Newton-Raphson anche nei casi in cui non valgono le

ipotesi del teorema 0.20. Nel caso del metodo di bisezione, invece, le ipotesi del teorema di Bolzano risultano necessarie in quanto la struttura stessa del metodo (il criterio di scelta di quale sotto-intervallo considerare per riapplicare la procedura iterativamente) si fonda sul fatto che tale teorema sia verificato.

Visti i vantaggi e gli svantaggi, la scelta su quale metodo applicare dipenderà dal caso specifico analizzato, tenendo conto delle esigenze del calcolo in un dato contesto.

## Esercizio 19

Si vogliono stimare numericamente gli zeri di due funzioni reali di variabile reale: una funzione quadratica e il polinomio di Legendre di ordine 10, utilizzando il metodo di bisezione e il metodo di Newton-Raphson. Si vuole, infine, operare un'analisi in efficienza dei due algoritmi proposti.

### Estensione a zeri multipli

Anzitutto, si noti che, in entrambi i casi, si vogliono determinare gli zeri di una certa funzione che, in generale, possono essere multipli visti i gradi dei polinomi in esame. Il metodo di bisezione e quello di Newton-Raphson sono stati definiti sotto ipotesi di esistenza di un unico zero all'interno dell'intervallo in cui si vuole applicare l'algoritmo. Risulta quindi necessario generalizzare i due metodi discussi al caso di zeri multipli. Sia  $f : \Omega \subseteq \mathbb{R} \rightarrow \mathbb{R}$  continua e sia  $[a, b] \subseteq \Omega$  l'intervallo all'interno del quale si vogliono determinare gli zeri multipli di  $f$ . Si supponga che  $f$  non ammetta zeri con molteplicità maggiore di 1 in tutto  $[a, b]$ . Si consideri poi una partizione dell'intervallo sufficientemente fitta, ossia  $N$  abbastanza grande tale che

$$[a, b] = \bigcup_{i=1}^N [c_i, c_{i+1}] \quad \text{con} \quad c_1 = a \quad \text{e} \quad c_{N+1} = b$$

Siccome  $f$  è continua in  $\Omega$ , allora lo sarà anche in ogni sotto-intervallo  $[c_i, c_{i+1}]$  in cui è stato diviso  $[a, b]$ . Ma allora avremo che, se è verificata la condizione

$$f(c_i)f(c_{i+1}) < 0 \tag{82}$$

esisterà almeno uno zero all'interno dell'intervallo  $[c_i, c_{i+1}]$  per teorema di Bolzano. Se la partizione di  $[a, b]$  è sufficientemente fitta si ha che  $[c_i, c_{i+1}]$  è sufficientemente piccolo, ossia al suo interno  $f$  è strettamente monotona per continuità. In tal modo è garantita l'unicità dello zero all'interno del sotto-intervallo che verifica la condizione di valori discordi agli estremi. Sempre per continuità, se  $f$  è strettamente monotona in  $[c_i, c_{i+1}]$  e ammette uno zero in questo sotto-intervallo, allora si avrà  $f(c_i)f(c_{i+1}) < 0$ . Vista la doppia implicazione basterà, a questo punto, applicare i due metodi all'interno di tutti e i soli sotto-intervalli  $[c_i, c_{i+1}]$  che verificano la condizione (82).

Si noti che la generalizzazione derivante dalla partizione effettuata presenta un possibile problema a livello operativo: può accadere, infatti, che esistano uno o più sotto-intervalli tali che  $f(c_i) = 0$  oppure  $f(c_{i+1}) = 0$ , ossia che uno degli estremi di qualche sotto-intervallo cada in corrispondenza di uno zero di  $f$ . In tal modo si avrebbe  $f(c_i)f(c_{i+1}) = 0$ , e lo zero non sarebbe visto dal metodo, in quanto la condizione (82) non risulterebbe verificata. Evidentemente, la probabilità che questo accada aumenta all'aumentare del numero di sotto-intervalli con cui si decide di ricoprire  $[a, b]$ . Questo rappresenta un problema in quanto la funzionalità del metodo di partizione dell'intervallo è strettamente legata, come si è mostrato, al numero elevato di sotto-intervalli  $[c_i, c_{i+1}]$  che vengono generati. Esistono almeno due modi possibili per risolvere il problema discusso: studiamo dunque nel dettaglio le due strade.

### 1) Controllo agli estremi

La via più facile consiste nell'effettuare, a seguito della partizione di  $[a, b]$ , un controllo su tutti gli estremi dei sotto-intervalli generati. In particolare, se

$$f(c_i) = 0$$

il metodo implementato restituirà direttamente  $c_i$  come zero di  $f$ . La condizione scritta, per essere resa operativa, al calcolatore dovrà essere implementata come

$$-\varepsilon < f(c_i) < \varepsilon$$

dove  $\varepsilon$  è un numero molto piccolo, dell'ordine della precisione con cui si sta operano. La necessità di questo fatto può essere verificata banalmente facendo stampare al calcolatore, ad esempio, il valore  $\sin(2\pi)$ . Si avrà modo di osservare che il numero restituito risulta dell'ordine della precisione utilizzata, ma diverso da zero. Questo accade per via dell'aritmetica finita del calcolatore: nell'esempio proposto il calcolatore commetterà sia un errore di arrotondamento, approssimando il numero trascendente  $\pi$  ad una successione decimale finita, sia un errore di troncamento, approssimando il seno al suo sviluppo in Taylor fino ad un certo ordine.

### 2) Partizione pseudo-casuale

Una seconda strada possibile per ridurre la probabilità di incorrere in questo errore consiste nell'effettuare una partizione di  $[a, b]$  sfruttando le sequenze pseudo-casuali. In particolare, si genera la sequenza ordinata

$$\{x_i\}_{i=1,\dots,N} \quad \text{con} \quad \text{pdf}(x) = U(a, b) \quad \text{e} \quad x_i < x_{i+1}$$

ossia una sequenza di  $N$  numeri distribuiti uniformemente all'interno dell'intervallo  $[a, b]$ . Viene poi effettuata la partizione

$$[a, b] = [a, x_1] \cup [x_1, x_2] \cup \dots \cup [x_N, b]$$

per poi ripetere il medesimo procedimento di ricerca precedente verificando, per ogni sotto-intervallo, la condizione (82). Questo algoritmo viene ripetuto per un numero  $M$  di volte, producendo l'insieme di vettori degli zeri

$$\{\vec{x}_i\}_{i=1,\dots,M}$$

I vettori  $\vec{x}_i$  avranno, in generale, dimensione diversa a seconda del numero di volte in cui si è perso uno zero a causa della caduta di qualche estremo di  $[c_i, c_{i+1}]$  su uno zero di  $f$ . Il vettore con maggiore probabilità di contenere tutti gli zeri di  $f$  in  $[a, b]$  sarà, dunque, il vettore che ha per componenti tutti gli elementi di  $\vec{x}_i$  che sono comparsi almeno una volta. La partizione non deterministica dell'intervallo può essere una soluzione in quanto, ad ogni chiamata della funzione generatrice, viene generata una sequenza diversa, ma equamente distribuita su tutto  $[a, b]$ . Chiaramente, può sempre accadere che un estremo di qualche sotto-intervallo cada in corrispondenza di uno zero di  $f$ , producendo un vettore dei risultati di dimensione inferiore a quella reale. Per tale ragione si esegue la procedura un numero  $M$  di volte, per poi unire tutti gli zeri ottenuti

in un unico vettore globale dei risultati. Non è difficile convincersi del fatto che questa seconda strada risulti particolarmente dispendiosa da un punto di vista computazionale, soprattutto in presenza di un numero molto elevato di zeri multipli.

Si ricordi il fatto che, entrambe le generalizzazioni, funzionano a meno dei casi in cui  $f$  ammetta anche radici con molteplicità maggiore di 1 nel campo di applicazione degli algoritmi. Per l'analisi di queste situazioni il codice necessita di essere raffinato ulteriormente con altre considerazioni. Per radici aventi molteplicità pari, ad esempio, il metodo di bisezione non è applicabile in quanto non sarà verificata l'ipotesi di valori discordi agli estremi. In tal caso sarà necessario isolare un intorno della radice e procedere con l'applicazione del metodo di Newton-Raphson. In ogni caso è sempre utile avere un'idea qualitativa dell'andamento della funzione in esame, al fine di applicare i metodi opportuni ad intervalli limitati nei quali è garantita la convergenza degli algoritmi. Si tralascia uno studio accurato dei casi di radici con molteplicità multipla in quanto non saranno oggetti degli esercizi qui proposti.

Si noti, inoltre, che per conseguenza immediata del teorema di Rolle, una mappa  $f$  di classe (almeno)  $C^1$  che ammette zeri multipli dovrà ammettere, in qualche punto del suo dominio, punti stazionari, ossia punti a derivata nulla. La probabilità di cadere in una divisione per zero nell'applicazione del metodo di Newton dipenderà dalla forma analitica di  $f$ , ma soprattutto dal dato iniziale da cui si parte per l'applicazione dell'algoritmo. Da questi due elementi dipenderà, infatti, il punto in cui la tangente interseca l'asse delle ascisse, che potrà eventualmente essere un punto stazionario. Anche in questo caso si può pensare di risolvere il problema eseguendo un controllo sulla derivata prima ad ogni nuovo punto calcolato con l'algoritmo di Newton: se il controllo fallisce, individuando un punto a derivata nulla, l'algoritmo partirà da capo con un nuovo dato iniziale che differirà dal precedente di una quantità  $\varepsilon$  piccola. Il problema descritto e la sua possibile risoluzione fanno pensare che la scelta della partizione pseudocasuale come metodo di ricerca di zeri multipli possa essere, in alcuni casi, la strategia migliore se si parla dell'algoritmo di Newton: ad ogni nuova partizione effettuata, infatti, il dato iniziale sarà leggermente diverso dal precedente, aumentando la probabilità che tutti gli zeri cercati siano presenti nell'unione degli  $M$  vettori dei risultati. Ad ogni modo, nei risultati che seguono non si è mai verificata una situazione che potesse far pensare al fallimento dell'algoritmo: per tale ragione, si è preferito non raffinare il codice con la procedura descritta. Risulta comunque importante sottolineare il problema in quanto questo fatto dà una forte ragione di credere a quanto si è accennato nell'introduzione: il metodo di Newton-Raphson non è un metodo globalmente convergente, ma dipende fortemente dal dato iniziale selezionato. La sensibilità del metodo al dato iniziale e le relative conseguenze saranno oggetto di studio dell'esercizio successivo.

Nelle analisi che seguono si è deciso di implementare la prima strada per la risoluzione del possibile problema di perdita di alcuni zeri, facendo quindi una singola partizione deterministica e svolgendo un controllo su ogni estremo, in quanto risulta più immediata e meno costosa computazionalmente.

## Funzione quadratica

Si vogliono determinare gli zeri della parabola

$$f(x) = 2x^2 - 3x + 1$$

Anzitutto, si noti che  $\Delta > 0$ , da cui segue che la funzione ammette due radici reali distinte. Applicando la formula risolutiva per equazioni di secondo grado è possibile determinare analiticamente gli zeri

$$x_1 = \frac{1}{2} \quad \text{e} \quad x_2 = 1$$

Si sono quindi applicati i metodi di bisezione e di Newton-Raphson generalizzati all'interno del compatto  $[-10, 10]$  in cui siamo certi essere presenti tutti gli zeri. In particolare, si è eseguita una partizione dell'intervallo con  $N = 100$ , ossia si sono generati 100 sotto-intervalli disgiunti, impostando una precisione di  $\varepsilon = 0.0000001$  in entrambi i casi. La tabella che segue mostra i risultati numerici ottenuti denotati con una tilde.

	$\tilde{x}_1$	$\tilde{x}_2$
Bisezione	0.5	1
Newton-Raphson	0.5	1

Evidentemente, i risultati ottenuti risultano consistenti con quelli attesi, dando prova del corretto funzionamento della generalizzazione a zeri multipli discussa. Si sono quindi plottati i punti  $(x_i, f(x_i))$  ottenuti dal metodo di bisezione sovrapponendo la funzione quadratica al fine di verificare, anche visivamente, il fatto che  $\tilde{x}_1$  e  $\tilde{x}_2$  fossero zeri per  $f$ . La scelta di quale dei due metodi utilizzare per il plot è del tutto arbitraria e del tutto equivalente: gli zeri ottenuti sono stati stimati a meno di una precisione  $\varepsilon$  identica in entrambi i casi.

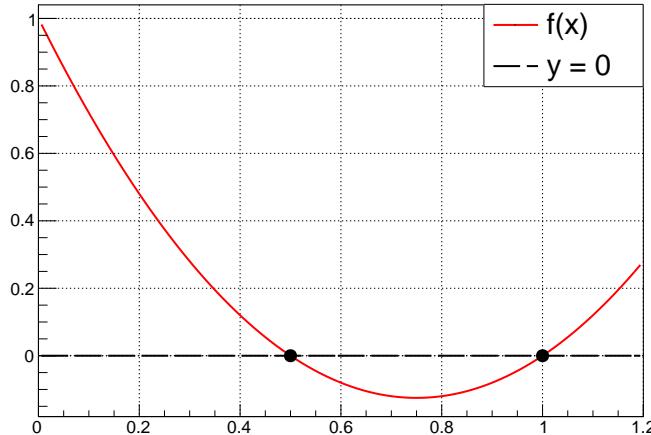


Figura 241: zeri  $(x_i, f(x_i))$  con bisezione e funzione  $f$

Come è possibile notare anche dal plot, entrambi i metodi hanno prodotto due valori consistenti rispetto a quelli attesi calcolati analiticamente. Al fine di

effettuare un confronto in efficienza tra i due metodi si è deciso di calcolare e stampare il numero di iterazioni necessarie alla convergenza in entrambi i casi. In particolare, si sono considerati gli intervalli  $[-3, 0.9]$  e  $[0.6, 1.4]$ , il primo contenente il primo zero  $x_1$  e il secondo contenente  $x_2$ . Si sono quindi applicati i due metodi all'interno di questi sotto-intervalli imponendo una precisione  $\varepsilon = 0.0000001$  dello zero cercato, come segue.

	$\tilde{x}_1 \in [-3, 0.9]$	$\tilde{x}_2 \in [0.6, 1.4]$
Bisezione	26	23
Newton-Raphson	8	5

Come è possibile notare, il numero di iterazioni necessario al metodo Newton-Raphson per convergere risulta, in entrambi i casi, sensibilmente più piccolo rispetto al numero di iterazioni necessario al metodo di bisezione. I risultati ottenuti sono indipendenti dal valore di precisione  $\varepsilon$  selezionato ad intervallo fissato. Si è deciso di rendere lo studio sul numero di iterazioni più sistematico mostrando questo fatto, considerando l'intervallo  $[-3, 0.9]$ . In particolare, si è calcolato il numero di iterazioni  $\eta$  con entrambi i metodi al variare di  $\varepsilon$  nel range

$$0.000001 \leq \varepsilon < 0.1 \quad \text{con} \quad \varepsilon_{i+1} = \varepsilon_i + 0.0001$$

Si sono quindi plottati i punti  $(\varepsilon_i, \eta_i)$  al fine di effettuare un confronto visivo. Il grafico che segue mostra i risultati ottenuti.

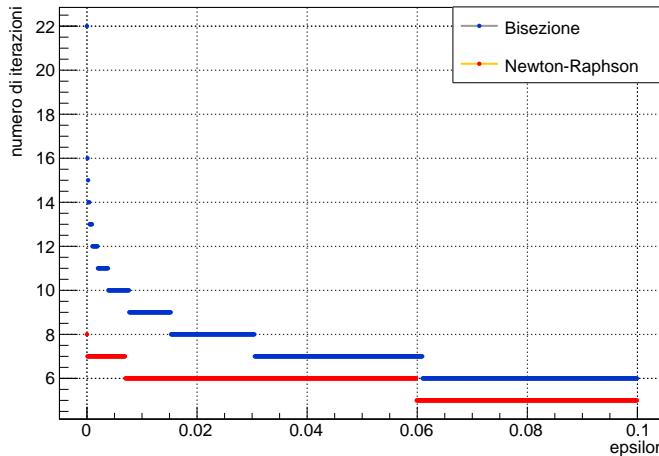


Figura 242: confronto efficienza dei due metodi  $\eta(\varepsilon)$

Come è possibile notare, al variare della precisione il metodo di Newton-Raphson risulta sempre più efficiente rispetto al metodo di bisezione. In particolare, Newton-Raphson risulta essere particolarmente vantaggioso quando si vogliono determinare zeri a precisioni elevate, mentre tende ad essere confrontabile con la bisezione a precisioni molto ridotte. Quanto ottenuto risulta consistente con quanto discusso inizialmente: il metodo di Newton-Raphson ha un'efficienza maggiore in quanto richiede una quantità di informazione maggiore sulla funzione  $f$  rispetto alla bisezione, ossia la conoscenza della derivata prima. Si avrà modo di verificare che, a fronte di questo vantaggio, il metodo di Newton-Raphson risulta particolarmente sensibile al dato iniziale.

## Polinomio di Legendre

Si vogliono determinare gli zeri del polinomio di Legendre di ordine 10

$$P_{10}^L(x) = (46189x^{10} - 109395x^8 + 90090x^6 - 30030x^4 + 3465x^2 - 63)/256$$

Anzitutto, si noti che l'equazione  $P_{10}^L(x) = 0$  non ammette soluzioni analitiche, in quanto non esiste una formula chiusa risolutiva per equazioni algebriche di grado 10. Casi come questi, per i quali il supporto analitico viene meno, rendono conto dell'utilità dei metodi numerici studiati. Notiamo, dunque, che vale il seguente teorema, la cui rilevanza risiede soprattutto nella costruzione del metodo di quadratura gaussiana.

**Teorema 0.21** (zeri dei polinomi ortogonali). *Sia  $P_n(x)$  un polinomio ortogonale di grado  $n$  in  $(a, b)$  rispetto alla sua funzione peso. Allora  $P_n$  ammette esattamente  $n$  radici reali, con molteplicità 1, tutte contenute in  $(a, b)$ .*

Il teorema degli zeri dei polinomi ortogonali dà quindi indicazioni sul numero di zeri che i metodi che verranno applicati dovranno produrre, fornendo anche un importante indizio circa l'intervallo in cui è sensato applicare gli algoritmi. Inoltre, il teorema garantisce la condizione primaria di applicabilità dei metodi, ossia l'assenza di radici con molteplicità superiore a 1. Sappiamo che gli zeri dei polinomi di Legendre sono contenuti nell'intervallo  $[-1, 1]$ . Si sono quindi applicati i metodi di bisezione e di Newton-Raphson generalizzati all'interno del compatto  $[-1.5, 1.5]$  in cui siamo certi essere presenti tutti gli zeri. In particolare, si è eseguita una partizione dell'intervallo con  $N = 100$  sufficientemente fitta, impostando una precisione di  $\varepsilon = 0.0000001$  in entrambi i casi, ottenendo i risultati sintetizzati nella seguente tabella.

	Bisezione	Newton-Raphson
$\tilde{x}_1$	-0.973907	-0.973907
$\tilde{x}_2$	-0.865063	-0.865063
$\tilde{x}_3$	-0.67941	-0.67941
$\tilde{x}_4$	-0.433395	-0.433395
$\tilde{x}_5$	-0.148874	-0.148874
$\tilde{x}_6$	0.148874	0.148874
$\tilde{x}_7$	0.433395	0.433395
$\tilde{x}_8$	0.67941	0.67941
$\tilde{x}_9$	0.865063	0.865063
$\tilde{x}_{10}$	0.973907	0.973907

Coerentemente con quanto ci si aspetta, i due metodi hanno prodotto valori identici, in quanto la precisione è stata fissata uguale in entrambi i casi. Inoltre, gli zeri trovati sono simmetrici rispetto all'asse delle ordinate, coerentemente con il fatto che  $P_{10}^L(x)$  risulta essere una funzione pari del suo argomento. Gli zeri trovati, infine, sono esattamente 10 tutti contenuti nell'intervallo  $[-1, 1]$  associato ai polinomi di Legendre, coerentemente con il teorema 0.21. Si sono quindi plottati i punti  $(x_i, f(x_i))$  ottenuti dal metodo di bisezione sovrapponendo il polinomio di Legendre al fine di verificare, anche visivamente, il fatto che quelli trovati fossero zeri per  $P_{10}^L$ .

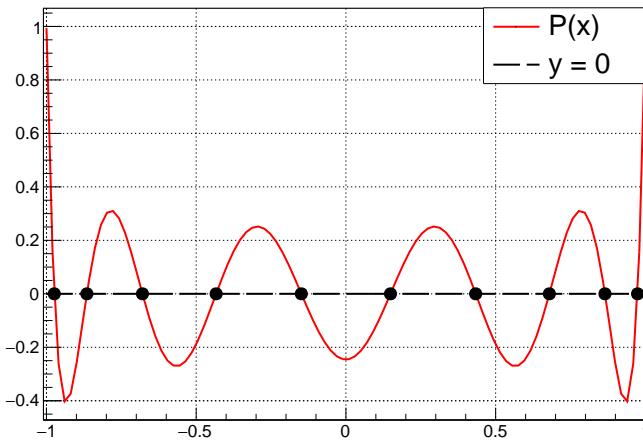


Figura 243: zeri  $(x_i, P_{10}^L(x_i))$  con bisezione e funzione  $P_{10}^L$

Dal grafico ottenuto si ha una un'evidente verifica qualitativa della correttezza dei risultati prodotti. Al fine di svolgere un confronto in efficienza si è deciso di calcolare e stampare, anche in questo caso, il numero di iterazioni necessario ad entrambi i metodi. Si è quindi considerato l'intervallo  $[0.4, 0.6]$  contenente la radice  $x_7$ , per poi applicare i metodi imponendo  $\varepsilon = 0.0000001$ . Si sono ottenute 21 iterazioni per il metodo di bisezione e 4 per il metodo di Newton. Anche in questo caso si è verificato che i risultati ottenuti fossero indipendenti dal valore di  $\varepsilon$  ad intervallo fissato. Si è quindi considerato  $[0.4, 0.6]$ , calcolando il numero di iterazioni  $\eta$  con entrambi i metodi al variare di  $\varepsilon$  nel range

$$0.000001 \leq \varepsilon < 0.1 \quad \text{con} \quad \varepsilon_{i+1} = \varepsilon_i + 0.0001$$

Si sono quindi plottati i punti  $(\varepsilon_i, \eta_i)$ , ottenendo quanto segue.

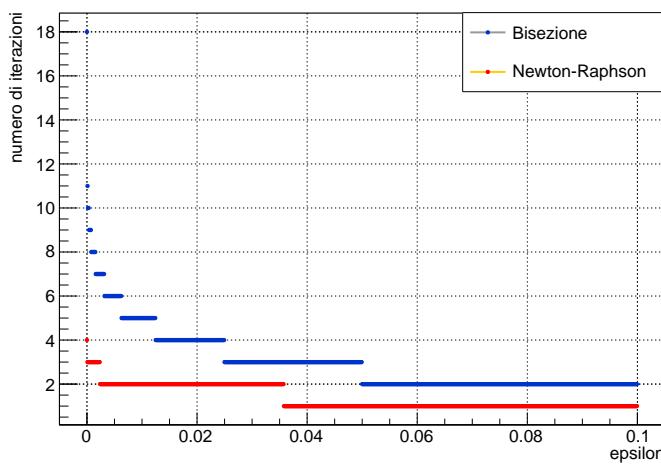


Figura 244: confronto efficienza dei due metodi  $\eta(\varepsilon)$

I risultati ottenuti sono quindi del tutto analoghi ai precedenti: il metodo di Newton-Raphson converge più velocemente a parità di precisione, soprattutto per valori di  $\varepsilon$  molto piccoli.

In definitiva, possiamo affermare che il metodo di Newton-Raphson sia più efficiente del metodo di bisezione a parità di condizioni, in quanto la sua stessa definizione richiede una conoscenza più approfondita della funzione  $f$  di cui vogliamo determinare gli zeri. Di contro, il metodo di Newton non converge per ogni valore del dato iniziale. Questa spiacevole proprietà ci obbliga a prestare particolare attenzione ogni volta che si procede all'utilizzo di tale metodo. La verifica e le conseguenze di questa forte sensibilità ai dati iniziali sarà oggetto principe dell'esercizio che segue.

## Esercizio 20

Si vogliono stimare numericamente gli zeri della funzione complessa

$$f(z) = z^3 - 1$$

utilizzando il metodo di Newton-Raphson, studiandone la convergenza in una certa regione del piano complesso.

**Teorema 0.22** (radici n-esime di un numero complesso). *Sia  $z = \rho e^{i\theta} \in \mathbb{C}$  tale che  $z \neq 0$ . Sia  $n \in \mathbb{N}$  tale che  $n \geq 2$ . Allora esistono esattamente  $n$  radici n-esime di  $z$  date da*

$$w_k = \sqrt[n]{\rho} \exp \left[ i \left( \frac{\theta}{n} + \frac{2\pi}{n} k \right) \right] \quad \forall k = 0, \dots, n-1$$

Inoltre,  $w_k$  si dispongono ai vertici di un poligono regolare centrato nell'origine del piano di Gauss.

Gli zeri della mappa complessa in esame sono dati da  $z \in \mathbb{C}$  tali che

$$z^3 = \rho e^{i\theta} = 1 \quad \Rightarrow \quad \theta = 0 \quad \text{e} \quad \rho = 1$$

L'applicazione della formula risolutiva permette, dunque, di determinare esplicitamente gli zeri di  $f$  per via analitica come

$$z_0 = 1 \quad \text{e} \quad z_1 = e^{i\frac{2}{3}\pi} = -\frac{1}{2} + \frac{\sqrt{3}}{2}i \quad \text{e} \quad z_2 = e^{-i\frac{2}{3}\pi} = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$$

Risulta immediato verificare che le radici trovate si dispongono ai vertici di un poligono regolare centrato nell'origine del piano complesso. Si noti, inoltre, che il numero di soluzioni ottenute è consistente con la seconda forma del teorema fondamentale dell'algebra.

Il metodo di Newton-Raphson può essere generalizzato in campo complesso in modo molto naturale. Posto un dato iniziale  $\bar{z} \in \mathbb{C}$ , la formula iterativa (81) si estende al piano complesso come

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)} \tag{83}$$

Scrivendo i numeri complessi in forma algebrica

$$z := x + iy \quad \text{e} \quad \frac{f(z)}{f'(z)} := \alpha(x, y) + i\beta(x, y)$$

la (83) può essere riscritta come

$$\begin{aligned} x_{n+1} + iy_{n+1} &= x_n + iy_n - \alpha_n(x, y) - i\beta_n(x, y) = \\ &= x_n - \alpha_n(x, y) + i[y_n - \beta_n(x, y)] \end{aligned}$$

Da cui segue che, l'equazione ricorsiva di Newton-Raphson può essere scritta come il sistema di equazioni

$$\begin{cases} x_{n+1} = x_n - \alpha_n(x, y) \\ y_{n+1} = y_n - \beta_n(x, y) \end{cases} \tag{84}$$

che fornisce separatamente il passo  $(n + 1)$ -esimo in funzione del passo  $n$ -esimo di parte reale e immaginaria di  $z$ , posto il dato iniziale  $\bar{z} = (\bar{x}, \bar{y})$ .

I numeri complessi possono essere definiti in modo costruttivo come coppie ordinate di numeri reali (definizione di Hamilton). Da questo fatto si può mostrare che esiste un naturale isomorfismo tra il campo complesso  $\mathbb{C}$  e lo spazio vettoriale  $\mathbb{R}^2$ , una volta definito il prodotto complesso come

$$(a, b) \cdot (c, d) = (ac - bd, ad + bc)$$

Per tale ragione, un numero complesso può essere legittimamente trattato, al calcolatore o su carta, come una coppia ordinata di numeri reali. La struttura dell'oggetto numero complesso si presta, dunque, ad essere definita come struttura indipendente, dotata delle proprie operazioni e dei propri metodi. Si è quindi deciso di implementare una classe in grado di operare con i numeri complessi come oggetti indipendenti, composti da una parte reale e da una parte immaginaria. Si sono poi implementate le operazioni di somma e prodotto che forniscono a  $\mathbb{C}$  struttura di campo. Si sono, infine, definiti alcuni metodi utili come il calcolo del modulo di un numero complesso o il calcolo della fase. L'utilizzo di una struttura dati in grado di trattare i numeri complessi come oggetti indipendenti rende possibile, infatti, l'implementazione diretta della formula ricorsiva (83) senza passare per il sistema equivalente (84), permettendo di mantenere un formalismo più pulito durante le operazioni di calcolo.

Nota la formula ricorsiva, il metodo di Newton-Raphson per funzioni complesse procede in modo del tutto analogo al caso reale a meno del fatto che, la condizione di arresto dell'algoritmo

$$|z_{n+1} - z_n| < \varepsilon$$

data una precisione  $\varepsilon$  selezionata, sarà ora da interpretarsi in senso complesso, ricordando che il modulo di  $z \in \mathbb{C}$  è definito come

$$|z| := \sqrt{\operatorname{Re}(z)^2 + \operatorname{Im}(z)^2}$$

e identifica la distanza di  $z$  dall'origine del piano di Gauss. Si noti che la definizione di modulo di un numero complesso è del tutto equivalente alla definizione di norma di un vettore in  $\mathbb{R}^2$ , con il medesimo significato geometrico. Tale fatto conferisce un'ulteriore verifica dell'isomorfismo di cui sopra.

Calcolando la derivata prima di  $f$  rispetto a  $z$ , nel caso in esame si dovrà implementare la relazione ricorsiva

$$z_{n+1} = z_n - \frac{z_n^3 - 1}{3z_n^2}$$

che convergerà ad uno dei tre zeri di  $f$  a seconda del valore del dato iniziale  $\bar{z}$ . Siamo quindi interessati prima ad una stima statica degli zeri al fine di avere una verifica numerica della corretta implementazione, per poi studiare nel dettaglio la forte dipendenza dell'algoritmo dal dato iniziale.

### Stima degli zeri

Anzitutto, al fine di verificare la corretta implementazione della classe e ricavare alcune informazioni qualitative circa l'andamento di  $f$  si è deciso di approcciare lo studio e l'analisi del problema da un punto di vista grafico. Si noti che le mappe complesse necessitano, in generale, di 4 dimensioni per rendere conto graficamente di tutte le informazioni in esse contenute: due dimensioni per il dominio e due per il codominio. Esistono vari modi per ovviare al problema della loro rappresentazione grafica, come l'utilizzo delle superfici di Riemann. In alternativa è possibile pensare all'immagine  $f(z)$  in forma esponenziale: in tal modo risulta possibile costruire un grafico tridimensionale dove due dimensioni rappresentano il dominio  $\mathbb{C}$ , e la terza rappresenta il modulo di  $f$  calcolato in ogni punto. L'informazione sulla fase viene poi aggiunta assegnando un colore differente alla superficie creatasi, univocamente associato ad un dato valore angolare. Dato un numero complesso  $z \in \mathbb{C}$ , vale la caratterizzazione

$$z = 0 \iff |z| = 0$$

Data una funzione complessa di variabile complessa  $f : \Omega \subseteq \mathbb{C} \rightarrow \mathbb{C}$  vale allora la doppia implicazione

$$f(z) = 0 \iff |f(z)| = 0 \quad (85)$$

Da questo fatto segue che è possibile avere una rappresentazione grafica della funzione  $f$  che mostri visivamente gli zeri a cui siamo interessati. In particolare, il grafico di  $|f|$  può essere costruito applicando l'idea di rappresentazione tridimensionale di cui sopra. Nel nostro caso siamo interessati al solo studio degli zeri, pertanto, l'informazione sulla fase può essere trascurata. Al fine di avere una rappresentazione qualitativa dell'andamento di  $f$  e dei suoi zeri si è quindi definita una griglia regolare di  $N^2$  punti nel quadrato

$$Q := [-2, 2] \times [-2, 2] \subset \mathbb{C}$$

Si sono poi calcolati i moduli

$$|f_j| = |f(z_j)| \quad \forall j = 1, \dots, N^2$$

associando ad ogni  $z_j \in Q$  il corrispondente valore di  $|f_j|$ . Si è selezionato un valore sufficientemente grande di  $N = 500$ , al fine di garantire una visualizzazione dettagliata. Si noti che il procedimento qui descritto può essere considerato come un metodo alternativo di ricerca degli zeri entro una certa precisione grazie alla caratterizzazione (85). Tuttavia, è facile immaginare che il calcolo di  $|f|$  per ogni punto di una certa regione del piano complesso risulti particolarmente dispendioso computazionalmente rispetto, ad esempio, al metodo di Newton-Raphson. Inoltre, questo procedimento presuppone di conoscere con certezza il fatto che almeno uno zero si trovi in una certa regione del piano: fatto per nulla scontato se si pensa a mappe più complicate, per le quali gli zeri non sono ricavabili per via analitica. Se non si ha un'idea almeno qualitativa delle regioni del piano che possono contenere zeri, infatti, il calcolo di  $|f|$  per punti risulta, di fatto, un'idea inapplicabile vista la vastità del piano di Gauss. Ad ogni modo, si sono plottate le terne ordinate  $(\operatorname{Re}(z_j), \operatorname{Im}(z_j), |f_j|)$  nello spazio, ottenendo i seguenti risultati.

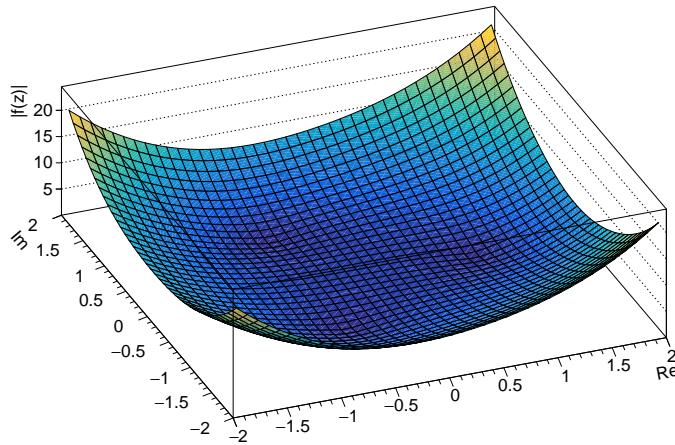


Figura 245: plot  $(\operatorname{Re}(z_j), \operatorname{Im}(z_j), |f_j|)$  dall'alto

Si noti che i colori assegnati alla superficie risultante danno la medesima informazione dell'asse verticale, ossia rappresentano il valore del modulo di  $f$ . A colori più scuri è associato un valore del modulo minore, a colori più chiari un valore del modulo maggiore. Visto da una diversa prospettiva, il grafico assume anche la forma che segue.

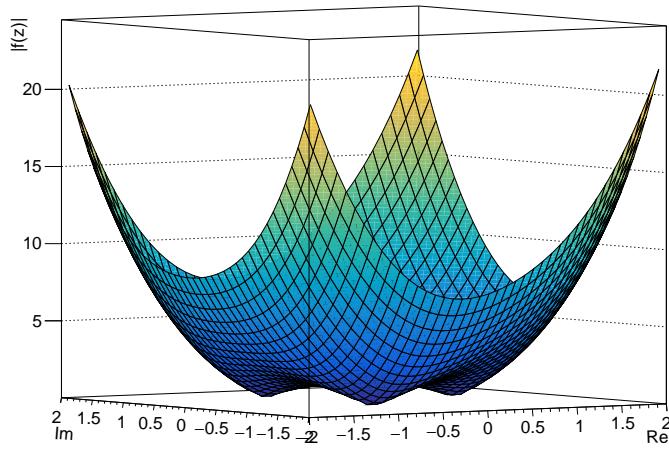


Figura 246: plot  $(\operatorname{Re}(z_j), \operatorname{Im}(z_j), |f_j|)$  dal basso

Come è possibile notare da entrambi i plot, la mappa  $|f|$  presenta visivamente tre punti nei quali si annulla, come ci si aspetta. Per la caratterizzazione (85), infatti, i punti del piano complesso che annullano il modulo di  $f$  coincidono con gli zeri della funzione complessa. La posizione degli zeri di  $f$  visibile dai grafici appare, qualitativamente, consistente con gli zeri determinati per via analitica. La consistenza dei risultati ottenuti permette, inoltre, di dare un certo grado di fiducia alla classe dei numeri complessi implementata.

Si è quindi deciso di svolgere uno studio più quantitativo stimando i tre zeri attesi di  $f$  utilizzando il metodo di Newton-Raphson. Non avendo ancora, a questo livello dello studio, informazioni precise sulle aree di  $\mathbb{C}$  nelle quali il metodo potesse convergere ad un dato zero, si è supposto che la convergenza potesse avvenire, ragionevolmente, almeno per valori di  $\bar{z}$  iniziali appartenenti ad un disco sufficientemente piccolo centrato nello zero cercato. Si sono quindi selezionati i valori iniziali

$$\bar{z}_0 = 0.8 + 0.2i \quad e \quad \bar{z}_1 = -0.8 + 1.2i \quad e \quad \bar{z}_2 = -0.8 - 1.2i$$

sufficientemente vicini, nel piano complesso, ai rispettivi zeri determinati per via analitica. Si è poi impostata arbitrariamente una precisione  $\varepsilon = 10^{-9}$ , sufficientemente piccola da consentire una buona stima. La tabella che segue mostra i risultati ottenuti dall'applicazione dell'algoritmo.

$z_0$	$z_1$	$z_2$
$(1, 10^{-32})$	$(-0.5, 0.866025)$	$(-0.5, -0.866025)$

Tenendo conto della precisione macchina si ha che  $10^{-32} \approx 0$ . Inoltre, si noti che vale  $\sqrt{3}/2 \approx 0.866025$ . Segue che è possibile concludere la compatibilità dei risultati ottenuti con i valori attesi. I risultati ottenuti, oltre a garantire la corretta implementazione, danno anche un indizio riguardo alle aree del piano complesso in cui avviene la convergenza ad un dato zero. Risulta infatti possibile ipotizzare che il metodo di Newton-Raphson converga ad uno zero cercato per valori  $\bar{z}$  iniziali appartenenti (almeno) ad un disco sufficientemente piccolo centrato nello zero stesso. Si sono quindi plottati i punti stimati nel piano complesso, ottenendo il seguente grafico.

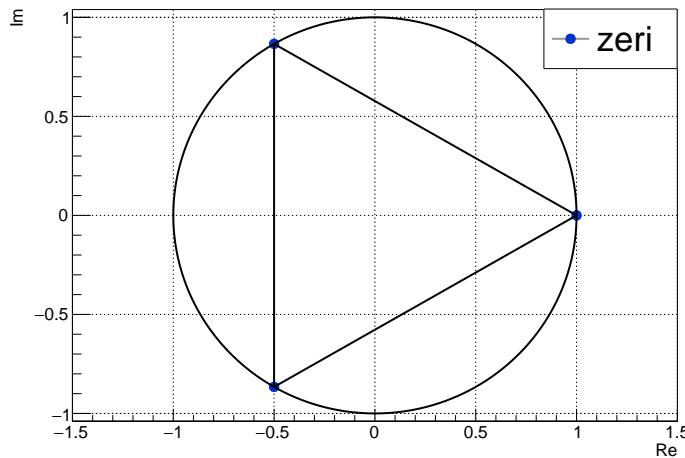


Figura 247: triangolo equilatero centrato in  $(0, 0)$  dato dagli zeri di  $f$

Anche visivamente è possibile apprezzare come gli zeri stimati dal metodo utilizzato si dispongano ai vertici di un triangolo equilatero centrato nell'origine del piano complesso, coerentemente con il teorema 0.22.

## Frattale di Newton

Sappiamo che il metodo di Newton-Raphson, a differenza della bisezione, non è un metodo globalmente convergente. Nei casi in cui  $f'$  è nulla, ad esempio, il metodo non convergerà ad alcuna radice, in quanto l'operazione di divisione per 0 in un campo è priva di significato. I punti in cui Newton-Raphson faticherà a convergere per  $f : \mathbb{R} \rightarrow \mathbb{R}$  corrisponderanno allora a  $x \in \mathbb{R}$  tali che

$$f'(x) \sim 0$$

Nel caso di  $f : \mathbb{C} \rightarrow \mathbb{C}$ , chiaramente, varrà lo stesso fatto, in quanto anche  $\mathbb{C}$  possiede una struttura algebrica di campo. Siamo però interessati a scoprire se un disco sufficientemente piccolo  $B_\varepsilon(z)$  con  $z = (0, 0)$  sia l'unico insieme di punti del piano complesso in cui il metodo fatichi a convergere. Al fine di studiare la velocità di convergenza in funzione del dato iniziale  $z_s \in \mathbb{C}$  si è allora costruita una griglia regolare di  $N^2$  punti nel quadrato

$$Q_2 := [-2, 2] \times [-2, 2] \subset \mathbb{C}$$

Per ognuno degli  $N^2$  punti  $z_s \in Q_2$  si è poi applicato il metodo di Newton-Raphson dato dalla (83). Fissato un valore di precisione  $\varepsilon$ , si è poi fermata l'iterazione secondo la solita condizione di arresto

$$|z_{j+1} - z_j| < \varepsilon$$

per poi associare il numero di iterazioni  $j + 1$  al punto iniziale  $z_s$  considerato. Si è infine plottata la griglia di punti in  $Q_2$  associando, ad ogni punto, un colore diverso in base al numero di iterazioni necessario per la convergenza. In tal modo, regioni del piano caratterizzate dallo stesso colore corrisponderanno a punti in cui il numero di iterazioni è il medesimo. Si è fissato  $N = 1500$  al fine di ottenere visivamente un piano omogeneo di punti, e  $\varepsilon = 10^{-9}$  al fine di rendere particolarmente evidenti gli sbalzi cromatici nelle varie zone del piano. Di seguito è riportato il grafico ottenuto.

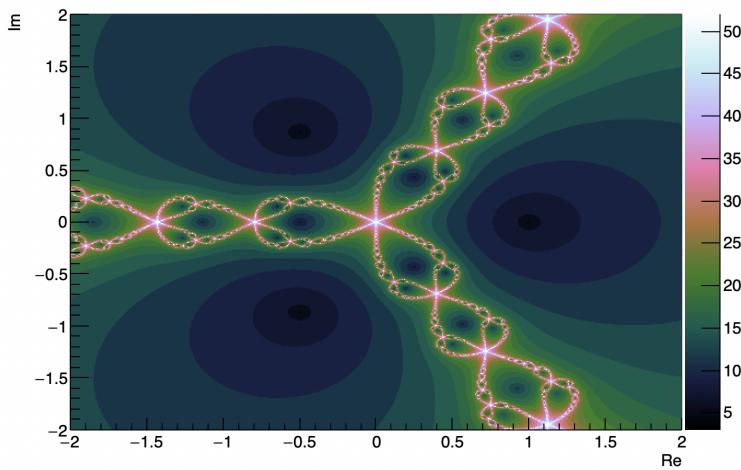


Figura 248: frattale di Newton con numero di iterazioni

Come è possibile notare, il grafico ottenuto mostra una figura particolarmente bizzarra e caotica, simmetrica rispetto all'asse orizzontale. Anzitutto, è possibile notare che il numero di iterazioni necessario alla convergenza ad uno dei tre zeri del polinomio in esame risulta molto elevato in un disco sufficientemente piccolo centrato nello zero, come preventivato. Inoltre, il metodo converge molto rapidamente, con un numero di iterazioni inferiore a 5, in un disco sufficientemente piccolo centrato nelle radici del polinomio in esame. Questo fatto permette di verificare quanto si è ipotizzato nella sezione precedente di studio statico degli zeri. Mano a mano che ci allontana dagli zeri, invece, il metodo necessita di un numero di iterazioni più elevato per convergere, coerentemente con l'intuizione. Tuttavia, il grafico mostra l'esistenza di numerosi altri punti in cui il numero di iterazioni è elevato allo stesso modo del numero di iterazioni intorno all'origine, in particolare superiore a 50, come si nota dalla scala a destra di figura 248. L'insieme di questi punti ad elevato numero di iterazioni costituisce una figura caratterizzata da una particolare geometria, che va sotto il nome di *frattale di Newton*, dal metodo utilizzato per la sua generazione.

**Definizione 0.23.** Chiameremo *frattale* ogni oggetto geometrico dotato di omotetia interna.

L'operazione di omotetia consiste nella dilatazione o nella contrazione di oggetti. Un frattale è allora un oggetto geometrico che si ripete uguale a se stesso su scale diverse, ossia un oggetto invariante sotto ridimensionamenti di scala. La proprietà di omotetia interna è anche detta proprietà di *auto-similarità* o *autosomiglianza*. Al fine di mostrare visivamente la struttura frattale si è allora svolta la medesima operazione di studio del numero di iterazioni nel quadrato

$$Q_{1/2} := [0.5, 1] \times [-1.5, -1] \subset Q_2$$

con gli stessi valori di  $N$  e  $\varepsilon$ , ottenendo il seguente grafico.

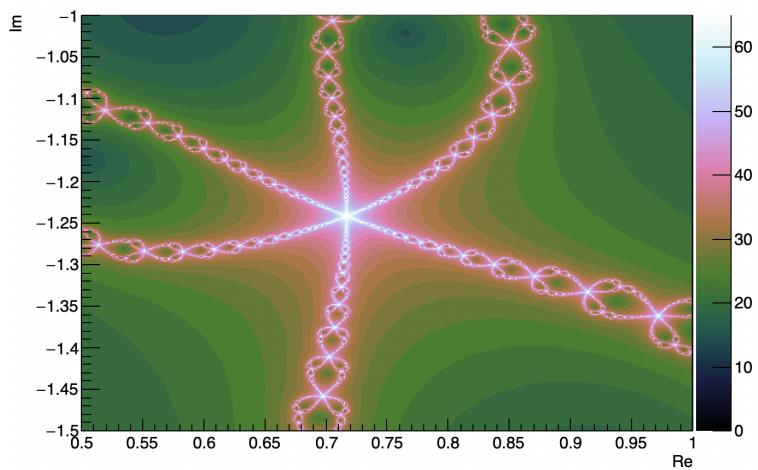


Figura 249: ingrandimento frattale di Newton con numero di iterazioni

Come si nota, la geometria dei punti associati ad un numero elevato di iterazioni rimane invariata rimpicciolendo la scala, ossia rifacendo il medesimo studio

in un sottoinsieme del quadrato  $Q_2$ . Siamo allora interessati a studiare più in profondità le ragioni di questo fenomeno. A priori, infatti, a meno dei punti addensati in un disco sufficientemente piccolo centrato nello zero, non ci si aspetta di osservare un numero elevato di iterazioni in zone del piano complesso così estese. Inoltre, siamo anche interessati alle ragioni per le quali l'applicazione di un algoritmo così elementare produca una tale ricchezza di immagine con una geometria atipica. Al fine di introdurre ad uno studio più dettagliato si è deciso di svolgere un'operazione simile a quella che ha generato il frattale di figura 248. In particolare, si è generata una griglia di  $N = 1500$  punti per direzione coordinata in  $Q_2$ . Per ogni  $z_s \in Q_2$  si è poi applicato il metodo di Newton-Raphson dato dalla (83), come in precedenza. Tuttavia, questa volta, si è associato un colore al punto iniziale  $z_s$  in funzione dello zero a cui il metodo converge entro lo stesso valore di precisione  $\varepsilon$  fissato in precedenza, a prescindere dal numero di iterazioni effettuate. Di seguito si è mostrato quanto ottenuto.

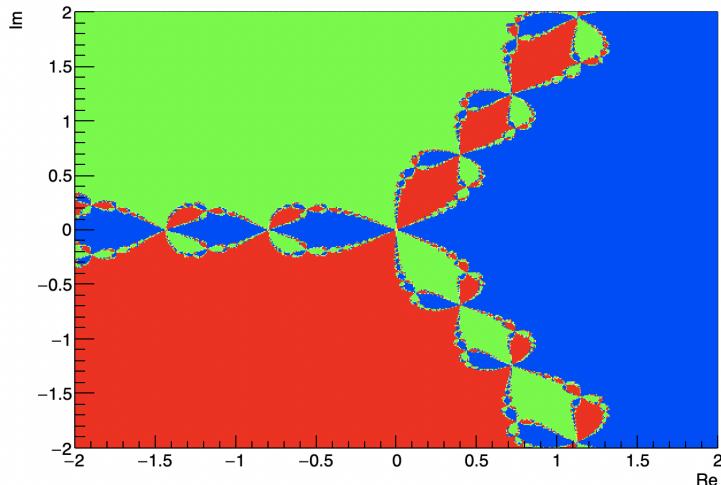


Figura 250: frattale di Newton con zeri di convergenza

Come è possibile notare dalla figura, anche l'assegnazione di un colore in base allo zero a cui il punto iniziale converge genera la medesima struttura frattale nel piano complesso. Si noti che in figura sono presenti 3 colori, in quanto 3 sono gli zeri del polinomio in esame: le aree colorate in blu contengono i dati iniziali che convergono a  $z_0$ , quelle colorate in verde a  $z_1$  e, infine, quelle colorate in rosso contengono i punti che convergono a  $z_2$ . Come ci si aspetta, il metodo converge sempre allo zero  $z_i$  in un disco  $B_\varepsilon(z_i)$  sufficientemente piccolo centrato nello zero cercato, coerentemente con quanto supposto in precedenza. Tuttavia, esistono regioni di convergenza che appaiono disporsi in modo caotico, molto lontane dallo zero a cui i punti al loro interno convergono. Questi fatti risultano, a priori, difficilmente spiegabili con l'intuizione. Al fine di giustificare alcuni dei risultati ottenuti è necessario introdurre alcune nozioni utilizzate in *dinamica olomorfa*: la branca della matematica che si occupa dello studio dell'applicazione ricorsiva di una mappa  $f$  differenziabile in senso complesso. Vedremo che la definizione in modo opportuno di nuovi oggetti permetterà di fare luce su gran parte dei risultati.

Anzitutto, la prima nozione essenziale consiste nel fatto che oggetti aventi geometria frattale non possono essere generati da mappe della forma

$$\begin{aligned} g : \mathbb{C} &\rightarrow \mathbb{C} \\ z &\mapsto g(z) \end{aligned}$$

In altre parole, non è possibile costruire un frattale assegnando banalmente un numero complesso ad ogni punto del piano complesso. Ogni oggetto frattale è prodotto da un algoritmo che deve essere applicato ricorsivamente. In particolare, l'oggetto con la proprietà di autosomiglianza si ha per un numero di iterazioni infinito di tale algoritmo, da cui segue immediatamente che, da un punto di vista computazionale, ogni frattale che tentiamo di visualizzare non è che un'approssimazione: fatto sicuramente non sorprendente. Vogliamo allora concentrarci sullo studio di mappe ricorsive della forma

$$z_{n+1} = f(z_n)$$

ossia di algoritmi applicati iterativamente. Sia allora  $P$  un polinomio complesso. Notiamo subito che la mappa

$$z_{n+1} \xrightarrow{f} z_n - \frac{P(z_n)}{P'(z_n)} \quad (86)$$

che definisce il metodo di Newton-Raphson in campo complesso assume una forma di questo tipo. Definiamo quindi due oggetti centrali nella trattazione che seguirà.

**Definizione 0.24.** Sia  $P$  un polinomio complesso e  $w_0$  una sua radice. Sia  $z_0$  il dato iniziale del metodo di Newton. Chiameremo

- *orbita* di  $z_0$  l'insieme

$$\Gamma := \left\{ z_{n+1} \in \mathbb{C} \mid z_{n+1} = z_n - \frac{P(z_n)}{P'(z_n)} \right\} \quad (87)$$

- *bacino di attrazione* di  $w_0$  l'insieme

$$B := \left\{ z_0 \in \mathbb{C} \mid \lim_{n \rightarrow \infty} z_n = w_0 \right\} \quad (88)$$

Si noti che  $P$  ammette sicuramente almeno una radice  $w_0$  per la prima forma del teorema fondamentale dell'algebra. Nella pratica, un'orbita non è altro che la "traiettoria" che segue il dato iniziale  $z_0$  a seguito dell'applicazione ricorsiva della mappa  $f$  di Newton. L'insieme dei punti iniziali tali che questa traiettoria converga al limite alla radice  $w_0$  è il bacino di attrazione. In figura 250, ad esempio, le aree rosse, verdi e blu formano 3 diversi bacini di attrazione per 3 diverse radici di un polinomio di terzo grado. Evidentemente, per ogni punto iniziale, il limite della (88) può esistere o non esistere. Tuttavia, sappiamo che se il limite esiste, sicuramente convergerà ad una delle radici del polinomio  $P$ , per costruzione stessa del metodo di Newton. Per fissare i concetti, di seguito è riportato un esempio grafico di orbita del dato iniziale  $z_s = (0.5, -1.1)$ .

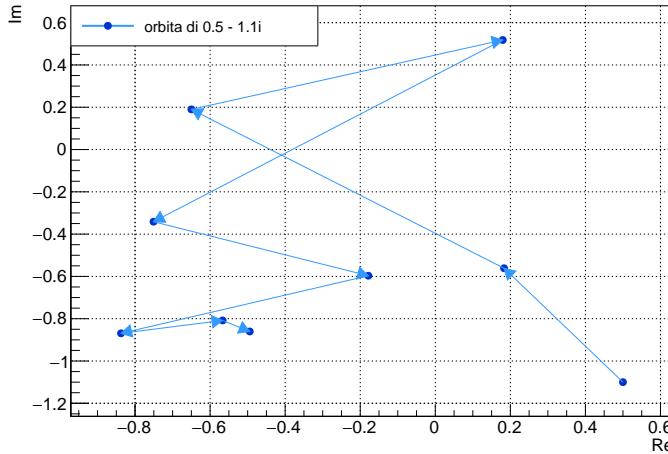


Figura 251: orbita di  $z_s = (0.5, -1.1)$  dopo 9 iterazioni

Si noti che l'orbita riportata converge alla radice  $z_2$  del polinomio oggetto dello studio. Ma allora, dalle definizioni date, segue che il punto  $z_s$  farà parte del bacino di attrazione di  $z_2$ . Introduciamo ora un ultimo ingrediente fondamentale per la trattazione.

**Definizione 0.25.** Sia  $g : \mathbb{C} \rightarrow \mathbb{C}$  una mappa complessa di variabile complessa. Chiameremo *punto fisso* di  $g$  un punto  $z \in \mathbb{C}$  tale che

$$g(z) = z$$

Un punto fisso è quindi un elemento del piano complesso che, sotto l'azione della mappa  $g$ , viene mandato in se stesso. Notiamo subito che vale, dunque, la seguente proposizione.

**Proposizione 0.26.** *Sia  $P$  un polinomio complesso tale che  $P' \neq 0$  e sia  $f$  la mappa ricorsiva data dal metodo di Newton. Allora, i punti fissi di  $f$  sono tutti e soli gli zeri di  $P$ .*

*Dimostrazione.* Il punto  $z \in \mathbb{C}$  è punto fisso per  $f$  se e solo se, per definizione

$$f(z) = z \iff z - \frac{P(z)}{P'(z)} = z \iff P(z) = 0$$

poiché  $P'(z) \neq 0$  per ipotesi. □

Siccome ogni polinomio  $P$  ammette almeno uno zero complesso per teorema fondamentale dell'algebra, possiamo allora concludere che la mappa  $f$  data dal metodo di Newton ammetta almeno un punto fisso per proposizione 0.26. Si noti che il fatto che i punti fissi di  $f$  coincidano con gli zeri del polinomio è un risultato che ci aspettiamo da un algoritmo di ricerca di zeri di funzione. Infatti, se la mappa  $f$  viene applicata ad uno zero di  $P$ , ovviamente questa manderà lo zero in se stesso: non è necessario avvicinarsi allo zero cercato se già stiamo applicando la mappa ad uno zero del polinomio.

**Definizione 0.27.** Sia  $z$  un punto fisso per  $g : \mathbb{C} \rightarrow \mathbb{C}$ . Diremo che  $z$  è

- *attrattivo* se  $\exists \varepsilon > 0$  tale che le orbite  $\Gamma$  di  $w$  abbiano  $z$  come bacino di attrazione  $\forall w \in B_\varepsilon(z)$ .
- *repulsivo* se  $\exists \varepsilon > 0$  tale che le orbite  $\Gamma$  di  $w$  divergano da  $z$  per  $n \rightarrow \infty$  e  $\forall w \in B_\varepsilon(z)$ .

In altre parole, un punto fisso per una mappa complessa è attrattivo se le orbite dei dati iniziali in un disco centrato nel punto fisso tendono ad avvicinarsi al punto stesso, viceversa, si dice repulsivo. Si noti, dunque, che l'operazione di derivazione in senso complesso può essere interpretata come un riscalamento (dilatazione o contrazione) congiunto ad una traslazione dell'insieme di punti in cui viene valutata. Da questo fatto segue la seguente proposizione.

**Proposizione 0.28.** *Sia  $g : \mathbb{C} \rightarrow \mathbb{C}$  una mappa complessa e  $z$  un suo punto fisso. Allora, se*

- $|g'(z)| < 1 \implies z$  è attrattivo
- $|g'(z)| > 1 \implies z$  è repulsivo

Si noti che derivando la mappa  $f$  di Newton rispetto al punto fisso  $z$  si ha

$$f'(z) = \frac{P(z)P''(z)}{P'(z)^2}$$

Siccome ogni punto fisso di  $f$  è uno zero per il polinomio  $P$ , per proposizione 0.26 si avrà  $P(z) = 0$ , e dunque

$$f'(z) = 0 \iff |f'(z)| = 0 < 1$$

Abbiamo allora raggiunto un risultato cruciale: ogni zero del polinomio  $P$  che si vuole stimare utilizzando il metodo di Newton è un punto fisso attrattivo per proposizione 0.28. In particolare, i punti fissi che verificano la condizione  $|f'(z)| = 0$  sono spesso detti *superattrattivi*. Il nome evocativo, chiaramente, si riferisce al fatto che, sotto questa condizione, i punti vicini allo zero convergono con grande rapidità allo zero stesso. Siamo allora riusciti a mostrare la ragione per la quale, in figura 248, in un disco sufficientemente piccolo dei tre zeri del polinomio in esame, il numero di iterazioni necessario alla convergenza risulta molto piccolo, inferiore a 5. Un vantaggio sostanziale del metodo di Newton consiste dunque nel fatto che, se si ha già un'idea qualitativa della posizione dello zero cercato, basterà applicare il metodo con un dato iniziale vicino allo zero stesso per sfruttare la proprietà di superattrattività dei punti fissi. In tutto questo, tuttavia, mancano ancora spiegazioni per una questione centrale: il motivo dell'esistenza di regioni del piano diffuse in cui il metodo fatica a convergere, e la ragione della loro geometria frattale. Abbiamo già accennato al fatto che il limite che compare nella (88) possa non esistere. Ci chiediamo allora se possa configurarsi la situazione in cui un dato iniziale  $z_0$  possa entrare, sotto l'applicazione ricorsiva di  $f$ , in un loop/ciclo a due. In altri termini, ci chiediamo se possa accadere che

$$z_0 \xrightarrow{f} z_1 \xrightarrow{f} z_0 \xrightarrow{f} z_1 \xrightarrow{f} \dots$$

Evidentemente, questo equivale a chiedersi se esista  $z \in \mathbb{C}$  tale che

$$f(f(z)) = z$$

Per induzione possiamo allora chiederci se possa configurarsi la situazione in cui  $z_0$  entri in un loop a  $n$ , che equivale a determinare  $z \in \mathbb{C}$  tali che

$$\underbrace{f(f(\dots f(z) \dots))}_{n \text{ volte}} = z \quad (89)$$

Evidentemente, per proposizione 0.26, sarà sempre vero che gli zeri di  $P$  sono soluzione della (89), in quanto coincidono con i punti fissi di  $f$ . D'altra parte, alla luce di quanto visto, questo fatto è ora intuitivo: se  $w_0$  è uno zero di  $P$ , in quanto punto fisso attrattore, l'applicazione ricorsiva di  $f$  a  $w_0$  manderà ogni volta in  $w_0$  stesso. Tuttavia, notiamo che esistono valori iniziali non banali del piano complesso che verificano tale condizione. In particolare, per funzioni razionali della forma di  $f$ , per teorema fondamentale dell'algebra si ha che

$$\#\{z_0 \in \mathbb{C} \mid (89) \text{ e } P(z_0) \neq 0\} \sim D^n$$

dove  $D := \deg P = 3$ . In altre parole, il numero di dati iniziali non banali la cui orbita entra in un ciclo a  $n$  va esponenzialmente con il numero di steps che compongono il loop. Se ci chiediamo allora quale sia il numero di punti nel piano che verifica la condizione d'ingresso nel loop infinito

$$\underbrace{f(f(\dots f(z) \dots))}_{\infty \text{ volte}} = z \quad (90)$$

il passaggio al limite mostra che dovrà valere

$$\#\{z_0 \in \mathbb{C} \mid (90) \text{ e } P(z_0) \neq 0\} = \infty$$

Questo ci dice che esiste un numero infinito di punti iniziali  $z_0$  nel piano complesso le cui orbite entrano in un loop infinito senza mai convergere al alcuno zero di  $P$ . D'altra parte, questi punti si dispongono a formare un frattale, come spiegheremo a breve, ossia costituiscono il bordo di un insieme bidimensionale che sapremo definire in modo più formale. Siccome la misura di Lebesgue di un qualunque insieme di dimensione inferiore dello spazio in cui vive è nulla, segue che la probabilità di beccare un punto che entri in un loop infinito scegliendo casualmente un punto nel piano complesso sarà anch'essa nulla. Idealmente, se si considera come dato iniziale un punto che costituisce il frattale, il metodo di Newton entrerebbe in un loop infinito. Risulta allora importante domandarsi se qualcuno di questi punti sia un attrattore. Se così fosse, avremmo probabilità non nulla di cadere in un loop infinito, in quanto i dischi nel piano complesso hanno misura non nulla. La risposta, in generale, è affermativa. Per verificarlo, la proposizione 0.28 ci dice che è sufficiente verificare che

$$\left| \frac{d}{dz} f(f(\dots f(z) \dots)) \right| < 1$$

Per nostra fortuna, esiste un risultato che permette di verificare se  $P$  ammetta un ciclo attrattore in modo molto più semplice.

**Teorema 0.29.** *Sia  $P$  un polinomio complesso e  $f$  la mappa di Newton. Siano  $z_i$  le  $n$  radici di  $P$ . Se esiste un ciclo attrattore, allora il centro di massa*

$$\bar{z} := \frac{1}{n} \sum_{i=1}^n z_i$$

*cadrà al suo interno.*

In altre parole, il teorema 0.29 ci assicura che se  $P$  ammette un ciclo attrattore, allora la media aritmetica delle radici del polinomio complesso entrerà nel loop dato dal ciclo. Ai nostri fini, l'aspetto importante del teorema appena enunciato è la sua contronomiale. Il teorema 0.29 equivale allora alla seguente proposizione: se  $\bar{z}$  non cade all'interno di un ciclo attrattore, allora  $P$  non ammette cicli attrattori. Nel caso del polinomio in esame si ha che

$$\bar{z} = 0$$

poiché le radici sono disposte ai vertici di un poligono regolare centrato nell'origine del piano complesso. Tuttavia, in questo caso, la mappa di Newton non può essere valutata in 0, poiché questo causerebbe una divisione per l'elemento neutro rispetto alla somma in un campo, producendo un'espressione priva di significato. Ma allora, per la contronomiale del teorema 0.29 siamo riusciti a mostrare che il polinomio  $P$  oggetto dell'esercizio non ammette cicli attrattori. Per quanto spiegato in precedenza, da questo segue che la probabilità di entrare in un loop infinito scegliendo come punto iniziale un punto casuale del piano risulta nulla. Questo fatto garantisce il corretto funzionamento del metodo di Newton-Raphson da un punto di vista computazionale, e spiega la ragione per la quale, nei risultati ottenuti, non si è mai ottenuto un valore che si avvicinasse al numero massimo di iterazioni  $K = 10000$  impostato manualmente, ossia ad un valore che potesse far pensare ad un numero infinito di iterazioni. Evidentemente, questo risultato non vale in generale, ma è immediato capire che vale in tutti i casi in cui il polinomio oggetto dello studio assume la forma

$$P(z) = z^n + a \quad \text{con} \quad a \in \mathbb{C}$$

ossia in tutti i casi in cui vale il teorema 0.22, che ci assicura la disposizione delle radici ai vertici di un poligono regolare centrato nell'origine. Per genericci polinomi che possiedono anche termini di grado minore del grado massimo, il metodo di Newton-Raphson è molto più sensibile alle condizioni iniziali, in quanto esiste una probabilità non nulla di assegnare un valore che generi un numero infinito di iterazioni. Il frattale di Newton è allora costituito da tutti e soli i punti che verificano la condizione (90). Al fine di dare un'idea del motivo per il quale l'applicazione di Newton-Raphson produca proprio un frattale si consideri la figura 250. Si può dimostrare che una proprietà caratterizzante di un frattale ottenuto come quello in figura è la seguente. Fissato un disco nel piano, i punti al suo interno devono rispettare una delle seguenti condizioni, mutuamente esclusive tra loro:

- convergere ad un solo zero di  $P$
- convergere a tutti gli zeri di  $P$

Visivamente, riferendoci alla figura 250 deve valere che, fissato un disco nel piano, questo deve necessariamente contenere o un colore solo, oppure tutti i colori. Non può accadere che in un disco vi siano punti che convergono ad un numero di zeri più piccolo del numero massimo, ossia non possono configurarsi situazioni intermedie ai due casi esplicitati. Con uno sforzo di astrazione si capisce che, assumendo che questa proprietà debba valere, la geometria della figura creatasi deve avere un infinito numero di "punti angolosi". Infatti, se per assurdo il frattale fosse liscio in qualche punto, esisterebbe un disco sufficientemente piccolo centrato nel punto contenente dati iniziali che convergono a due soli zeri: questo contraddice la proprietà di cui sopra. In altre parole, se la proprietà deve valere, non può formarsi alcuna altra geometria se non una geometria frattale. Il problema è quindi ora ricondotto al mostrare che debba valere questa proprietà. Nella letteratura la proprietà menzionata è un corollario di quello che è chiamato *teorema di Montel*, la cui discussione non è banale e richiede studi più profondi dell'analisi complessa e delle dinamiche olomorfe. Si noti che quanto discusso spiega anche la ragione per la quale geometrie frattali si generano tipicamente solo per polinomi di grado maggiore o uguale a 3. Per polinomi complessi di grado 2, infatti, il confine liscio rispetterebbe del tutto la proprietà menzionata. Ad ogni modo, siamo ora in grado di arrivare ad una definizione più precisa del particolare frattale ottenuto: quello di Newton.

**Definizione 0.30.** Sia  $P$  un polinomio complesso e  $f$  la sua mappa di Newton. Siano  $w_i$  le  $n$  radici del polinomio e siano  $B_i$  i corrispondenti bacini di attrazione. Chiameremo *insieme di Fatou* di  $f$  l'insieme

$$F := \bigcup_{i=1}^n B_i$$

In altre parole, l'insieme di Fatou, in quanto unione di tutti i bacini di attrazione per le radici di  $P$ , rappresenta l'insieme dei punti del piano complesso tali che la loro orbita rimanga stabile e prevedibile.

**Definizione 0.31.** Sia  $P$  un polinomio complesso e  $f$  la sua mappa di Newton. Chiameremo *insieme di Julia* di  $f$  l'insieme

$$J := \mathbb{C} \setminus F$$

L'insieme di Julia di un polinomio, invece, è tutto il resto: l'insieme dei punti la cui orbita evolve in modo caotico entrando in un loop infinito. Di fatto, coincide con i punti che compongono il frattale.

**Definizione 0.32.** Sia  $P$  un polinomio complesso. Chiameremo *frattale di Newton* l'insieme di Julia della mappa meromorfa

$$z \xrightarrow{f} z - \frac{P(z)}{P'(z)}$$

Siccome l'insieme di Julia rappresenta il complementare dell'insieme di Fatou, ossia il bordo di ogni bacino di attrazione, siamo allora riusciti a mostrare quanto si era accennato in precedenza: il frattale di Newton ha misura inferiore rispetto allo spazio in cui vive. Per tale ragione, a livello computazionale, sotto ipotesi di non avere cicli attrattori, il metodo convergerà ad uno degli zeri cercati qualunque sia il valore (sensato) del dato iniziale assegnato al calcolatore.