Relazione: LS Quicksort

Lorenzo Livio Vaccarecci (matr. 5462843)

A.A. 2023/2024

1 Codice

```python
import random
from tqdm import tqdm
import matplotlib.pyplot as plt
import math
import numpy as np

X = 0

def confronta(s, pos, Sequence):
    global X
    sequenceMin = []
    sequenceMax = []
    for i in range(len(Sequence)):
        if (i != pos):
            if Sequence[i] < s:
                sequenceMin.append(Sequence[i])
                X += 1
            else:
                sequenceMax.append(Sequence[i])
                X += 1
    return [sequenceMin, sequenceMax]

def LVQuickSort(Sequence):
    if (len(Sequence) <= 1): return Sequence
    pos = random.randint(0, len(Sequence)-1)
    s = Sequence[pos]
    confronto = confronta(s, pos, Sequence)
    sequenceMin = LVQuickSort(confronto[0])
    sequenceMax = LVQuickSort(confronto[1])
    return np.concatenate([sequenceMin, [s], sequenceMax])

def valore_medio(R, Xr):
    sommatoria = sum(Xr)
    return (1/R) * sommatoria

def varianza(R, Xr, u):
    sommatoria = 0
    for i in Xr:
        sommatoria += ((i-u)**2)
    return (1/(R-1)) * sommatoria
```

```python

def markov(mu, k):
    return mu/(k*mu)

def chebyshev(mu, k, varianza):
    return varianza / (((k-1)**2)*(mu**2))

def conta_frequenze(Xr, n, val_medio):
    k = 0
    for x in Xr:
        if x >= n*val_medio:
            k += 1
    return k

n = 10**4
R = 10**5
Xr = []
array = np.random.randint(0, n, size=n)

for i in tqdm(range(R)):
    X = 0
    LVQuickSort(array)
    Xr.append(X)

val_medio = valore_medio(R, Xr)
varianza = varianza(R, Xr, val_medio)
dev_standard = math.sqrt(varianza)

print("Valore medio: ", val_medio)
print("Varianza: ", dev_standard)
print("Deviazione standard: ", math.sqrt(dev_standard))

plt.hist(Xr, edgecolor="black", bins=50)
plt.xlabel("Numero di confronti")
plt.ylabel("Frequenza")
plt.show()

v1 = 2
v2 = 3

print(markov(val_medio, v1))
print(markov(val_medio, v2))

print(chebyshev(val_medio, v1, varianza))
print(chebyshev(val_medio, v2, varianza))

print("Frequenza empirica di X per il doppio: ", conta_frequenze(Xr, v1,
    val_medio)/R)
print("Frequenza empirica di X per il triplo: ", conta_frequenze(Xr, v2,
    val_medio)/R)
```

2 Obiettivo

Determinare il numero medio di confronti effettuati dall'algoritmo Quicksort Las Vegas per ordinare un array di n elementi e calcolare la deviazione standard di tale valore. Inoltre, verificare la validità delle disuguaglianze di Chebyshev e Markov per il doppio e il triplo del valore medio confrontandole con la frequenza di valori superiori o uguali al doppio o al triplo del valore medio.

3 Formule utilizzate

3.1 Formule principali

3.1.1 Valore medio e deviazione standard

$$\hat{\mu} = \frac{1}{R}\sum_{r=1}^{R} X_r \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{R-1}\sum_{r=1}^{R}(X_r - \hat{\mu})^2 \tag{2}$$

Dove X_r è il numero di confronti effettuati al passo r-esimo e R è il numero di esecuzioni dell'algoritmo.

3.1.2 Disuguaglianze

Chebyshev:

$$Pr\left\{X \geq v\mu\right\} \leq \frac{\hat{\sigma}^2}{(v-1)^2\hat{\mu}^2} \tag{3}$$

Markov:

$$Pr\left\{X \geq v\mu\right\} \leq \frac{\hat{\mu}}{v\hat{\mu}} \tag{4}$$

3.1.3 Formule ausiliarie

$$\mathbb{E}[X] = \sum_{i=1}^{n-1} 2\ln(n-i+1) = 2\sum_{i=1}^{n-1}\ln(n-i+1) \tag{5}$$

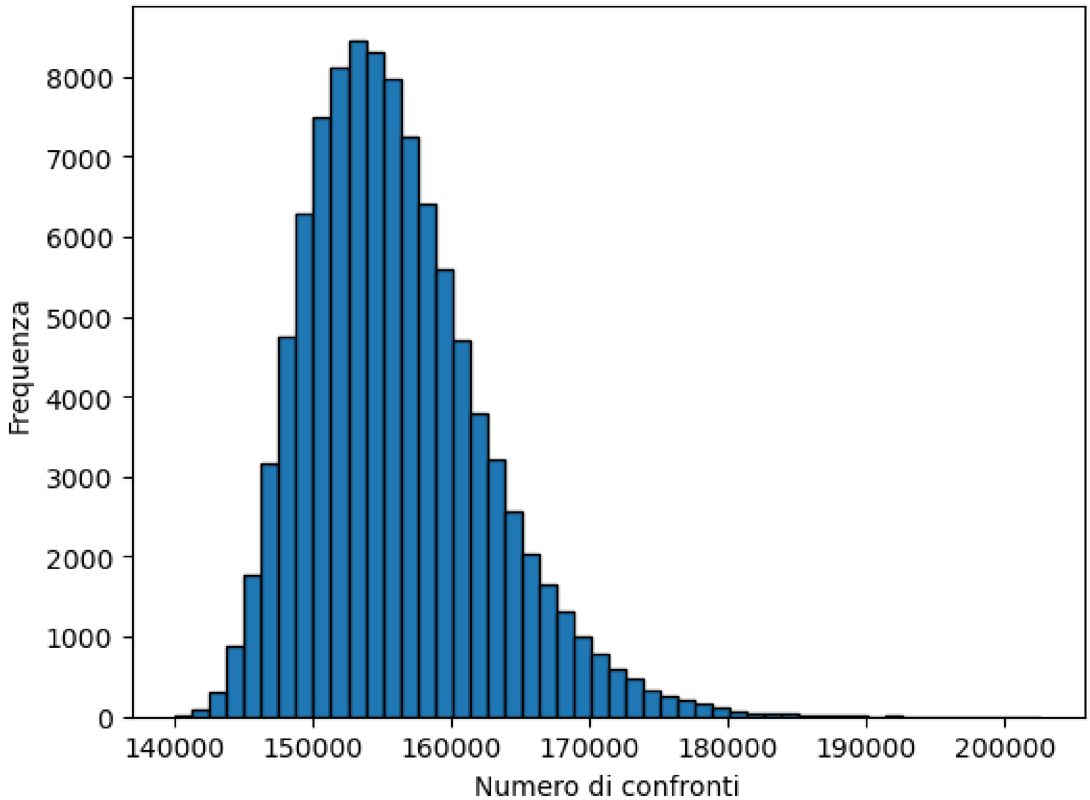
$$\mathbb{E}[X] = 2n\ln(n) \tag{6}$$

4 Risultati e Conclusioni

Dopo aver implementato l'algoritmo Quicksort Las Vegas, si è utilizzata una sequenza di numeri casuali di cardinalità 10^4 eseguendo l'algoritmo per 10^5 volte e si sono ottenuti i seguenti valori con la formula (1) e la radice della formula (2):

- Valore medio ($\hat{\mu}$) $\simeq$ 156533.23
- Deviazione standard ($\hat{\sigma}$) $\simeq$ 6498.49

Usando il numero di confronti X effettuati si può generare il seguente istogramma:



Il grafico mostra che la distribuzione dei valori è ampia (questo conferma la deviazione standard elevata), il picco si trova in un valore vicino a 150000 confermando il valore medio ottenuto.
Attenzione: questa parte qua è da togliere, le formule 5 e 6 sono per calcolare il costo computazionale O
Un'ulteriore conferma che il valore medio è quello atteso è data dalla formula (5):

$$2 \sum_{i=1}^{10^4-1} \ln(10^4 - i + 1) \simeq 164216.47$$

E in modo simile dalla formula (6):

$$2 \cdot 10^4 \ln(10^4) \simeq 184206.81$$

Si può notare che i valori medi ottenuti dalle formule teoriche e quello calcolato dall'implementazione dell'algoritmo sono molto simili. Questa discrepanza può essere attribuita a diversi fattori, tra cui la natura approssimativa delle formule teoriche e dalla scelta del pivot. Tuttavia, il valore medio ottenuto dal programma è ragionevole e conferma l'efficienza dell'algoritmo.
Calcolata la disuguaglianza di Chebyshev usando la formula (3) con $v = 2 \quad v = 3$ rispettivamente:

$$\frac{6498.49^2}{(2-1)^2 \cdot 156533.23^2} \simeq 0.0017 \simeq 0.17\%$$

$$\frac{6498.49^2}{(3-1)^2 \cdot 156533.23^2} \simeq 0.0004 \simeq 0.04\%$$

Calcolata anche la disuguaglianza di Markov per le stesse v usando la formula (4) si ottengono facilmente le due probabilità:

$$\frac{156533.23}{2 \cdot 156533.23} = 0.5 = 50\%$$

$$\frac{156533.23}{3 \cdot 156533.23} \simeq 0.33 \simeq 33\%$$

Secondo la disuguaglianza di Markov, la probabilità che il numero di confronti sia superiore a $2\hat{\mu}$ può arrivare fino al 50%, mentre la probabilità che sia superiore a $3\hat{\mu}$ fino al 33%. Questo significa che la disuguaglianza di Markov è molto meno cauta rispetto a quella di Chebyshev che, invece, fornisce un limite superiore alla probabilità che il numero di confronti sia superiore a $2\hat{\mu}$ del 0.17% e del 0.04% per $3\hat{\mu}$. Le frequenze del numero di confronti superiore al doppio o al triplo di $\hat{\mu}$ sono state pari a 0 ed effettivamente $0 \leq 0.17\%, 0 \leq 0.04\%$ (per Chebyshev) e $0 \leq 50\%, 0 \leq 33\%$ (per Markov).

5 Conclusioni

In conclusione, i risultati ottenuti dall'implementazione dell'algoritmo Quicksort Las Vegas confermano la sua efficienza per la risoluzione di problemi di ordinamenti di array molto grandi. Il valore medio e la deviazione standard calcolati indicano che la distribuzione dei valori è ampia ma concentrata attorno al valore medio come confermato dall'istogramma.
Le disuguaglianze di Chebyshev e Markov confermano che la probabilità che il numero di confronti sia superiore al doppio e al triplo del valore medio è molto bassa ed è stata confermata dalla frequenza dei valori superiori al doppio e al triplo del valore medio.