CS251 Assignment-1
Problem 5

Josyula Venkata Aditya[1] and Kartik Sreekumar Nair[2]

[1]210050075
[2]210050083

August 22, 2022

Problem 5

Generate a dataset comprising a set of real numbers drawn from the uniform distribution on $[0, 1]$. Consider various dataset sizes

$$N = 5, 10, 20, 40, 60, 80, 100, 500, 10^3, 10^4$$

For each N, repeat the following experiment $M := 100$ times:

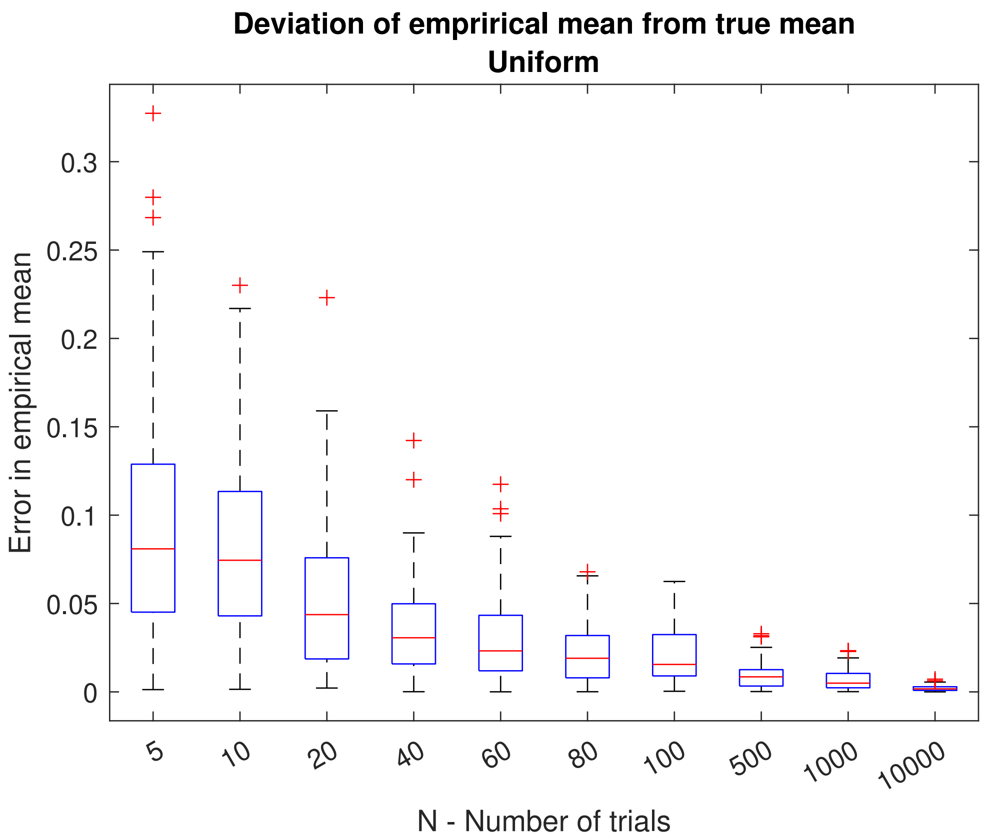
1. generate the data
2. compute the average $\hat{\mu}$
3. measure the error between the computed average $\hat{\mu}$ and the true mean μ as $|\hat{\mu} - \mu_{true}|$.

   Refer to `problem5.mlx` for the generated data, and implementation.

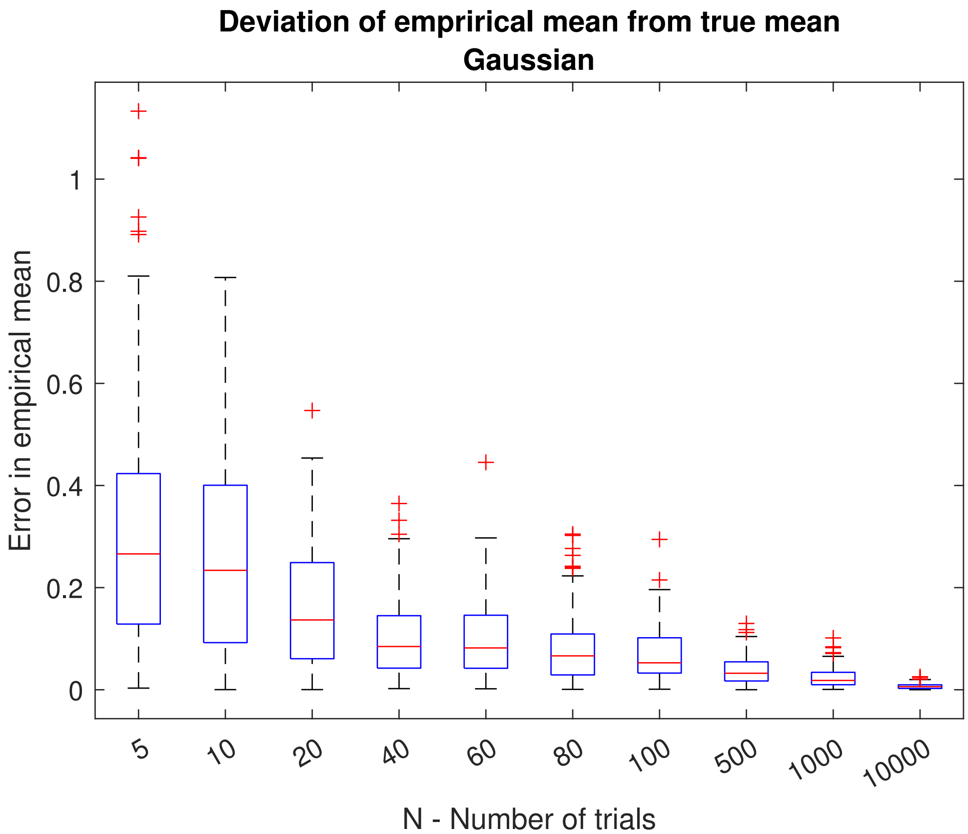
Approach

We generated the M experiments for each value of N as a single matrix, with each column corresponding to a particular N (the boxplot also takes input in this way). Then looped over each value of N, for which we computed the error M times , stored as a 1D array. And concatenated the arrays to get the desired matrix. We used the same procedure for both parts, with just the random generator functions differing.

- For the uniform distribution, plot a single graph that shows the distribution of errors (across M repeats) for all values of N using a box-and-whisker plot. You may use the boxplot$(\cdot)$ function in Matlab.



- Repeat the above question by replacing the uniform distribution by a Gaussian distribution with $\mu := 0$ and $\sigma^2 := 1$.



- Interpret what you see in the graphs. What happens to the distribution of error as N increases ? This graph is a direct reflection of the law of large numbers. i.e. the mean converges to the theoretical mean as $N \to \infty$

  When the value of N is small, the data that is generated is more 'random' than the data generated when N is large. So for small N, the spread in the deviations from the true mean are relatively high. As N increases to large enough values(say 1000 and 10000), the mean values obtained over the $M(100)$ iterations are almost equal to the true mean value. In fact, there is barely any data set that is conspicuously farther from 0 when $N = 10000$; as is evident from the graph.

  In conclusion, the distribution of error decreases as N increases.