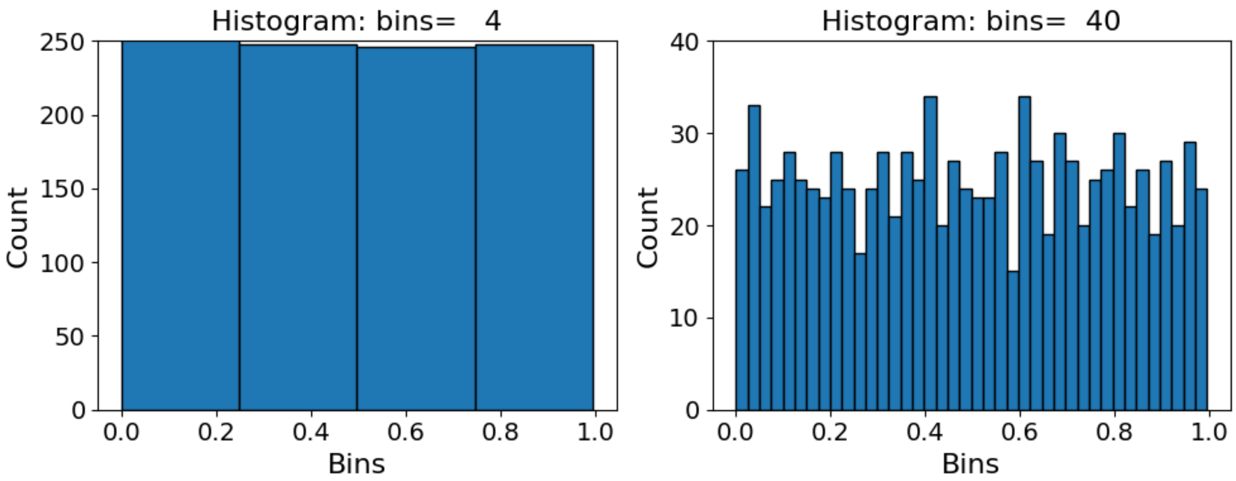
CO544: Machine Learning and Data Mining
Machine Learning Lab One

Ranage R.D.P.R. - E/19/310

1. Random Numbers and Uni-variate Densities

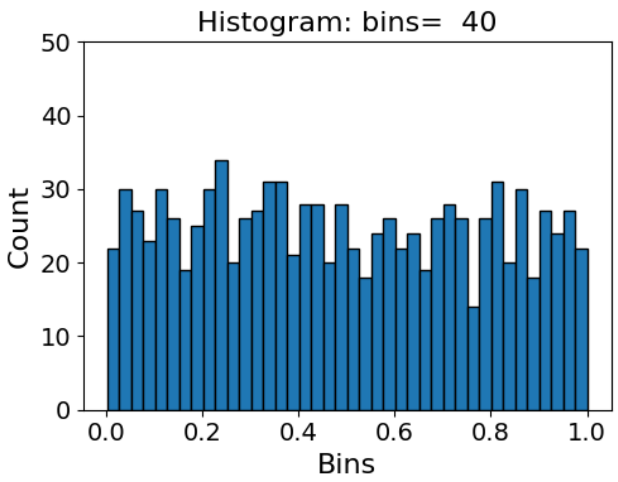
The code produces a visual representation with two histograms placed next to each other. Each histogram illustrates the spread of 1000 random numbers obtained from a uniform distribution. One histogram divides the data into 4 bins while the other uses 40 bins.

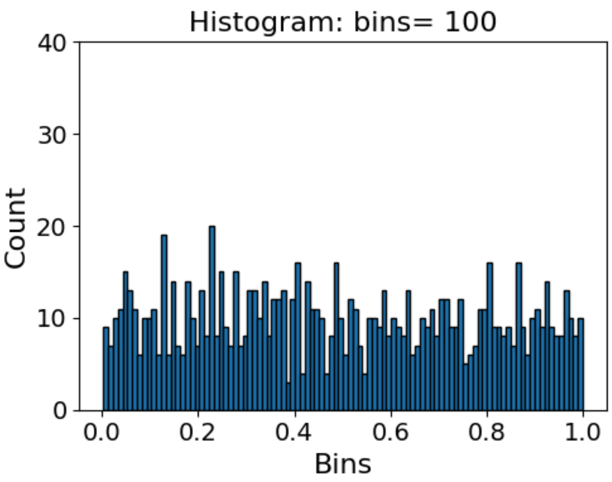


i. Though the data is from a uniform distribution, the histogram does not appear flat. Why?
   - Due to the limited data and the choice of bin sizes, even in cases where the underlying distribution is uniform, the histograms may display slight irregularities or bumps. These deviations from a perfectly flat distribution can occur due to insufficient sample size or inappropriate bin selection.

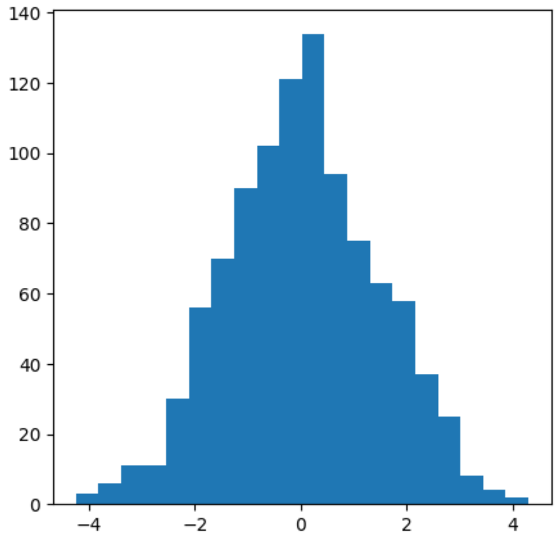
ii. Every time you run it, the histogram looks slightly different? Why?
   - The histogram looks different each time because the data changes with every run of the code. This variability is expected when working with random data and highlights the importance of setting a random seed for consistent results.

iii. How do the above observations change (if so how) if you had started with more data?
   - Using more data points leads to more reliable results and a better understanding of the distribution. With a larger dataset, the histogram looks smoother and more consistent, reducing the influence of random fluctuations caused by binning. Shown below is an image of the histograms after increasing number of bins.

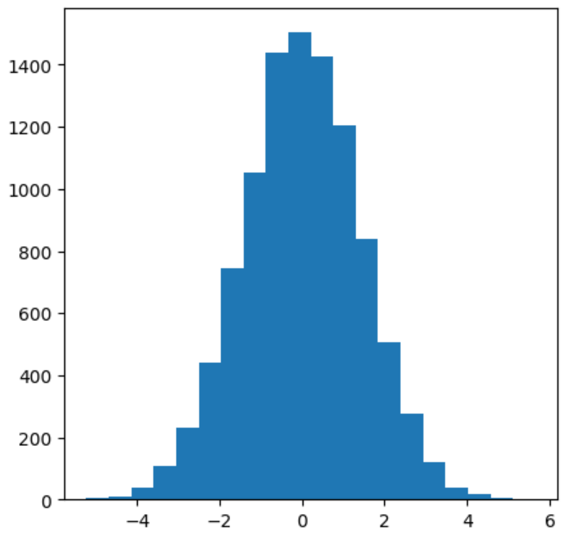




After adding and subtracting some uniform random numbers:



N = 1000



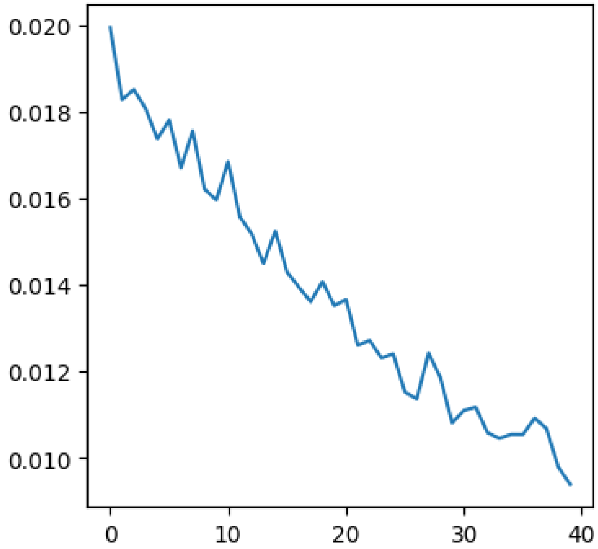
N = 10000

i. What do you observe? How does the resulting histogram change when you change the number of uniform random numbers you add and subtract? Is there a theory that explains your observation?

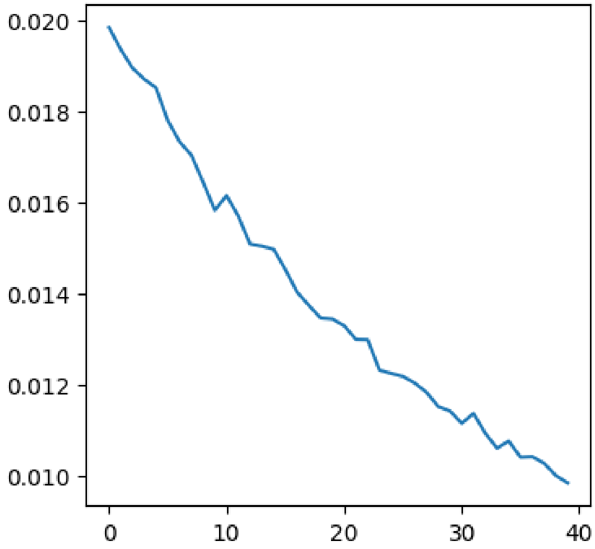
- Observations:
  The histogram we see looks like it's shaped like a bell, centered around 0. When we add and subtract lots of random numbers that are evenly spread out, the histogram starts to look more like a bell-shaped curve. This happens because of something called the Central Limit Theorem (CLT). It says that when we add up or average a bunch of random numbers, even if they're from different distributions, the result tends to look like a bell curve.

- Explanation:
  When we add and subtract many evenly spread-out random numbers, the total tends to look like a bell curve because of the CLT. Even though each individual addition or subtraction adds some randomness, when we do a lot of them (like in a loop), they add up in a way that makes the total look like a bell curve. So, even if the original numbers were evenly spread out, the final result ends up looking like a bell curve when we do lots of adding and subtracting.

2. Uncertainty in Estimation
When we calculate the variance from various sets of samples, we'll likely get slightly different results each time. However, if we had more data, we would anticipate this variation to be minimal.



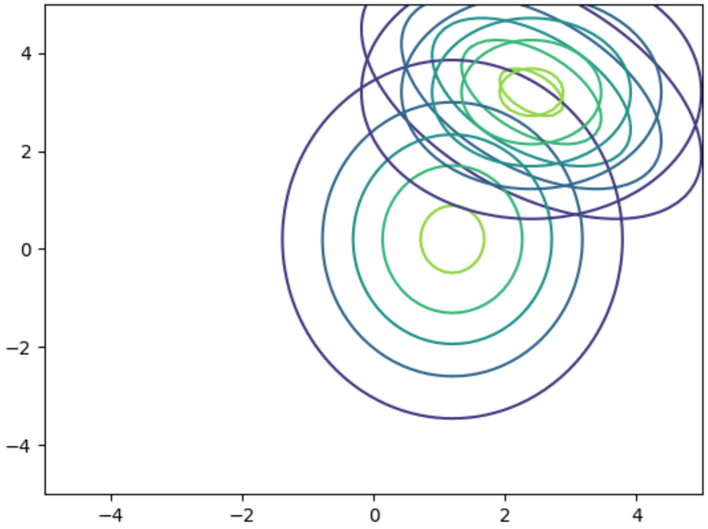
Trials = 2000



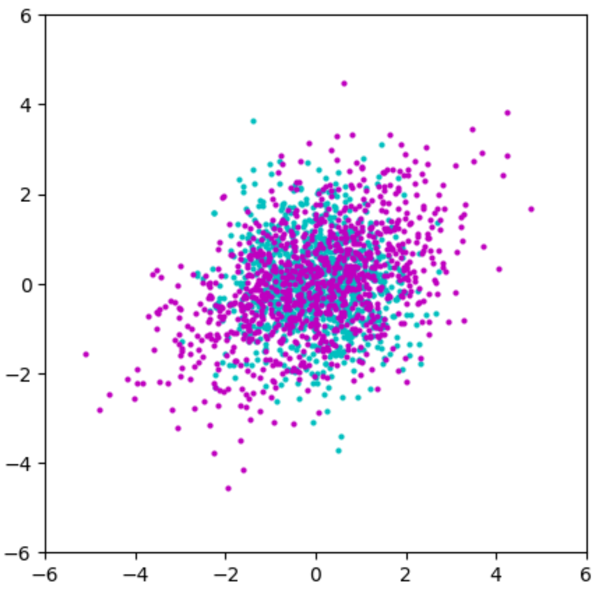
Trials = 20000

3. Bi-variate Gaussian Distribution

In here, the gauss2D function finds the value of a 2D Gaussian probability density function at a given point x, with a mean m and covariance matrix C. The twoGaussianPlot function creates a grid in the x-y plane and calculates the 2D Gaussian value at each point on this grid using gauss2D. Then, it gives back X, Y, and Z coordinates as arrays, defining the plot of the 2D Gaussian.

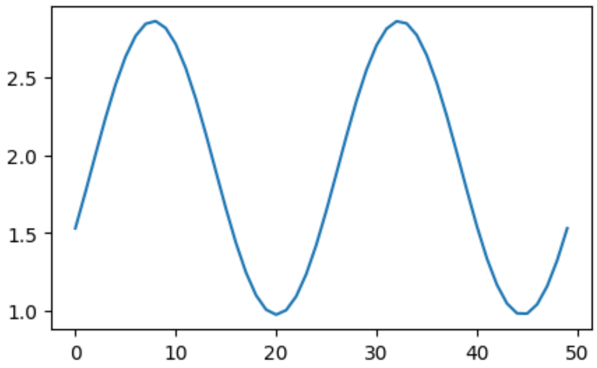


4. Sampling from a multi-variate Gaussian



The cyan plot displays data obtained from a standard normal distribution, where each sample is independent and evenly spread across the plot without any specific pattern or correlation. On the other hand, the purple plot shows data with a diagonal distribution, indicating correlation between the samples. This correlation arises from using the Cholesky decomposition of the covariance matrix C, which introduces correlation between the dimensions of the samples specified by C.

5. Distribution of Projections



i. What are the maxima and minima of the resulting plot?

Maximum variance: 2.858012294369908
Minimum variance: 0.9767323398899298

ii. Compute the eigenvalues an eigenvectors of the covariance matrix C

Eigenvalues: [3.   1.]
Eigenvectors: [[0.70710678   -0.70710678]
[0.70710678   0.70710678]]

iii. Can you see a relationship between the eigenvalues and eigenvectors and the maxima and minima of the way the projected variance changes?

The values are nearly the same because projections indicate where each random variable changes the most or least. Eigenvalues and eigenvectors tell us where variance changes the most. Basically, the eigenvalues of the covariance matrix show the highest and lowest variance in projections, while the corresponding eigenvectors point out the directions of these changes.

iv. The shape of the graph might have looked sinusoidal for this two-dimensional problem. Can you analytically confirm if this might be true?

Since the projected variance is affected by the cosine of theta, and the cosine function fluctuates sinusoidally with theta, we anticipate seeing a sinusoidal pattern in the plot of projected variance.