1 Methods

This project implements a Self-Organising Map (SOM) to cluster flower species from the Iris dataset, demonstrating how this unsupervised learning method organizes high-dimensional data into a lower-dimensional space while preserving relationships between data points. The SOM operates by initializing a grid of (40x40) neurons with random weights. Each input data point is compared to these weights to identify the "winning neuron" which is the most similar to the input, defined by cosine similarity. The algorithm updates the weights of the winning neuron and its neighbors, allowing the map to adapt based on the input data. This approach captures the inherent structure of the data, enabling cluster visualization. As training progresses, closer neurons develop similar weights due to the influence of the Best Matching Unit (BMU). Data points are assigned to neurons based on similarity, resulting in distinct clusters that help identify patterns within the dataset.

2 Results

After training the SOM for 10 epochs on the Iris dataset, two scatter plots were generated:
• Initial Weights Plot: Shows how the data points were distributed based on randomly initialized weights.
• Final Weights Plot: Illustrates the organized clustering of data points after training.

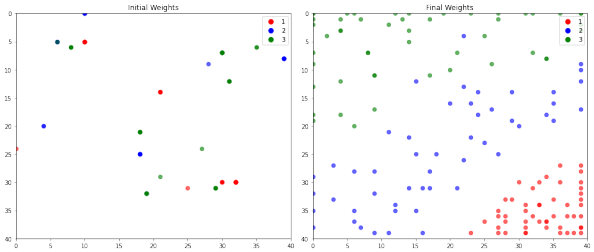

Figure 1: Divergence scores and KL divergence upper bounds for varying M.

The final plot indicates that the SOM effectively grouped similar species based on their features, highlighting its capability to discern patterns within the data. Another important observation is that in both plots, the number of points is smaller than the number of samples. This occurs because points are assigned to the closest neuron in a discrete map, leading to overlaps where different points can share the same neuron. In the initial case, these overlaps are larger due to the random initialization of weights, which does not accurately represent the data distribution. In contrast, the trained weights result in fewer overlaps, as the SOM learns to adjust the weights based on the input data, effectively capturing the similarity between points and within clusters.