Part 1:

Question 1: According to your results (i.e., elbow_k), are there 3 species of iris represented in the iris data set?

Answer:

Yes. Based on the reconstruction-error (inertia) curve obtained from running K-Means for k = 1 through 20, the “elbow” occurs at k = 3. This indicates that the data naturally clusters into three distinct groups, which aligns with the three known species in the Iris dataset: *Setosa*, *Versicolor*, and *Virginica*.
Indeed, the confusion matrix for k = 3 shows that:

- **Cluster 1** corresponds mainly to *Setosa* samples.
- **Cluster 2** corresponds mostly to *Versicolor*.
- **Cluster 3** corresponds primarily to *Virginica*.

Part 2:

Question 2a: According to your AIC results (i.e., aic_elbow_k), are there 3

Answer:

Yes. The AIC curve reached its minimum (elbow) at **k = 3**, indicating that three Gaussian components provide the best balance between model fit and complexity. According to the GMM confusion matrix for k=3, one component corresponds almost perfectly to *Iris setosa* (49/49 correct). Another corresponds primarily to *Iris versicolor* (49 correct, 1 misclassified). The third mostly captures *Iris virginica* (36 correct, 14 overlapping with versicolor).

Question 2b: According to your BIC results (i.e., bic_elbow_k), are there 3

Answer:

Yes. The BIC curve also indicated an elbow at k = 3, producing the same clustering structure and accuracy as AIC. The identical confusion matrix shows that both model selection criteria converge on the same interpretation.

Part 3:

Question 3a: Use the quantization error vs grid sizes graph to identify the 'elbow' for grid size using the quantization error in the same way you found the elbow in k-means (reconstruction error) and GMM (AIC/BIC), what grid size would you select based on this elbow?

Answer: From the quantization error curve, the most significant reduction occurs between the 3×3 and 7×7 grids, after which the curve begins to flatten. Beyond 7×7, increasing the grid size leads to only marginal improvements in quantization error. Therefore, using the same “elbow” reasoning applied in K-Means and GMM, **the optimal grid size is 7×7**, as it achieves a strong reduction in error before further increases provide diminishing improvements.

Question 3b: How does grid size affect SOM performance?

Answer: Increasing the grid size **monotonically decreases QE** (finer quantization/fewer representation errors). Indeed, looking at the QE vs Gird Size as well as the different U-matrices:

- Very small grids (e.g., 3×3) **underfit** with a more coarse map, high QE, blurred boundaries on the U-Matrix.
- Very large grids (e.g., 25×25) can become **over-granular/sparse** for Iris (150 samples vs 625 neurons which is approximately 0.24 samples/neuron). The U-Matrix looks noisy/fragmented and many neurons are rarely BMUs which is an indicator of a less stable topology and longer training.
- Medium grids (7×7 to 15×15) give a **good balance**: clear cluster regions on the U-Matrix with substantially lower QE than smaller grids.

Question 3c: Which grid size (between 7×7 and 25×25) is a “perfect fit” for Iris, and why?

Answer: The 7×7 grid is the best fit for the Iris dataset under our setup because it offers the strongest quantization-error vs. granularity trade-off. It delivers a clear, well-balanced U-Matrix with distinct cluster regions (QE 0.0721) while avoiding the over-granularity and sparsity seen at 25×25 (QE is 0.0164 but there are many lightly used neurons).