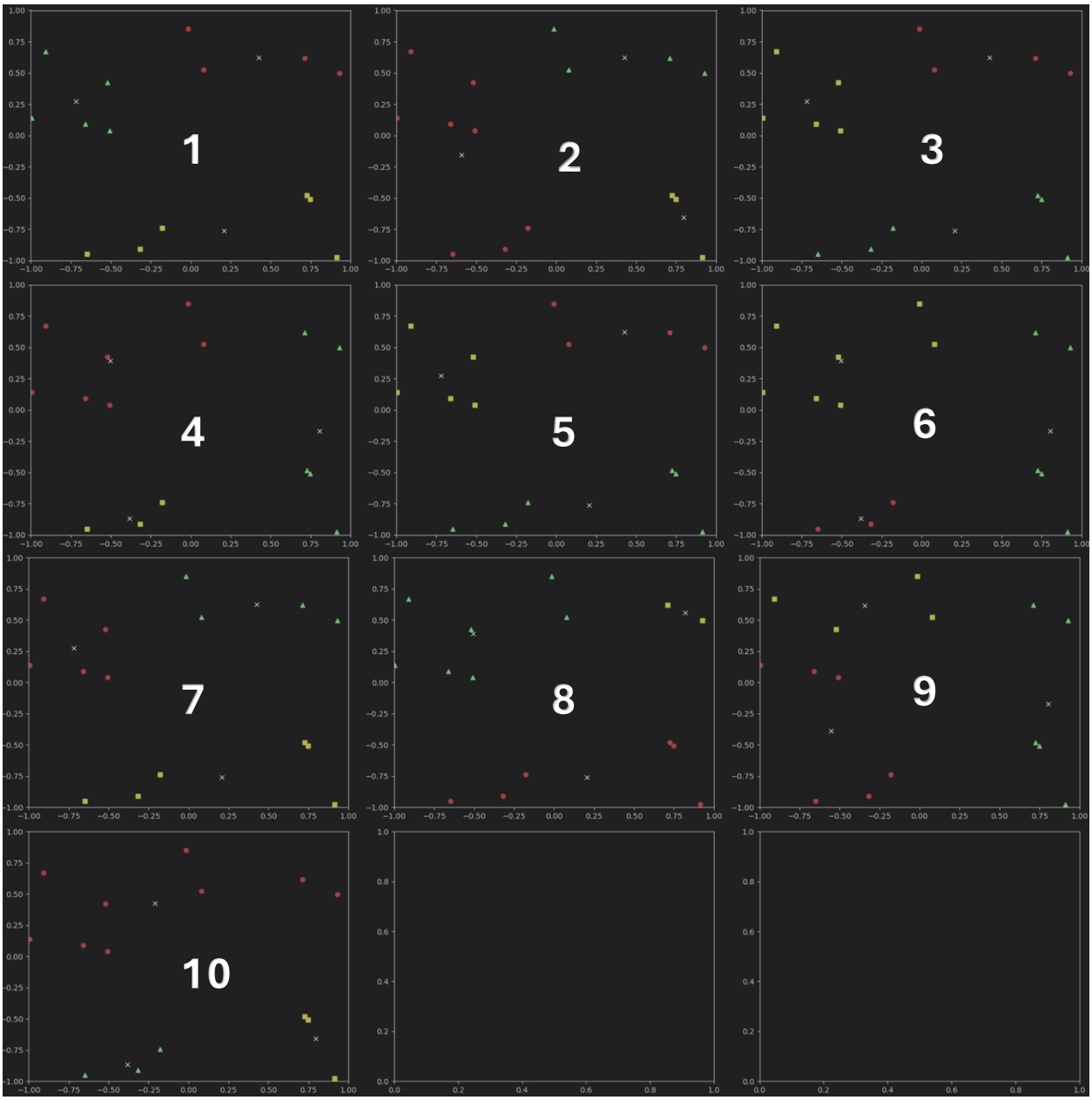
Lucas Sipos

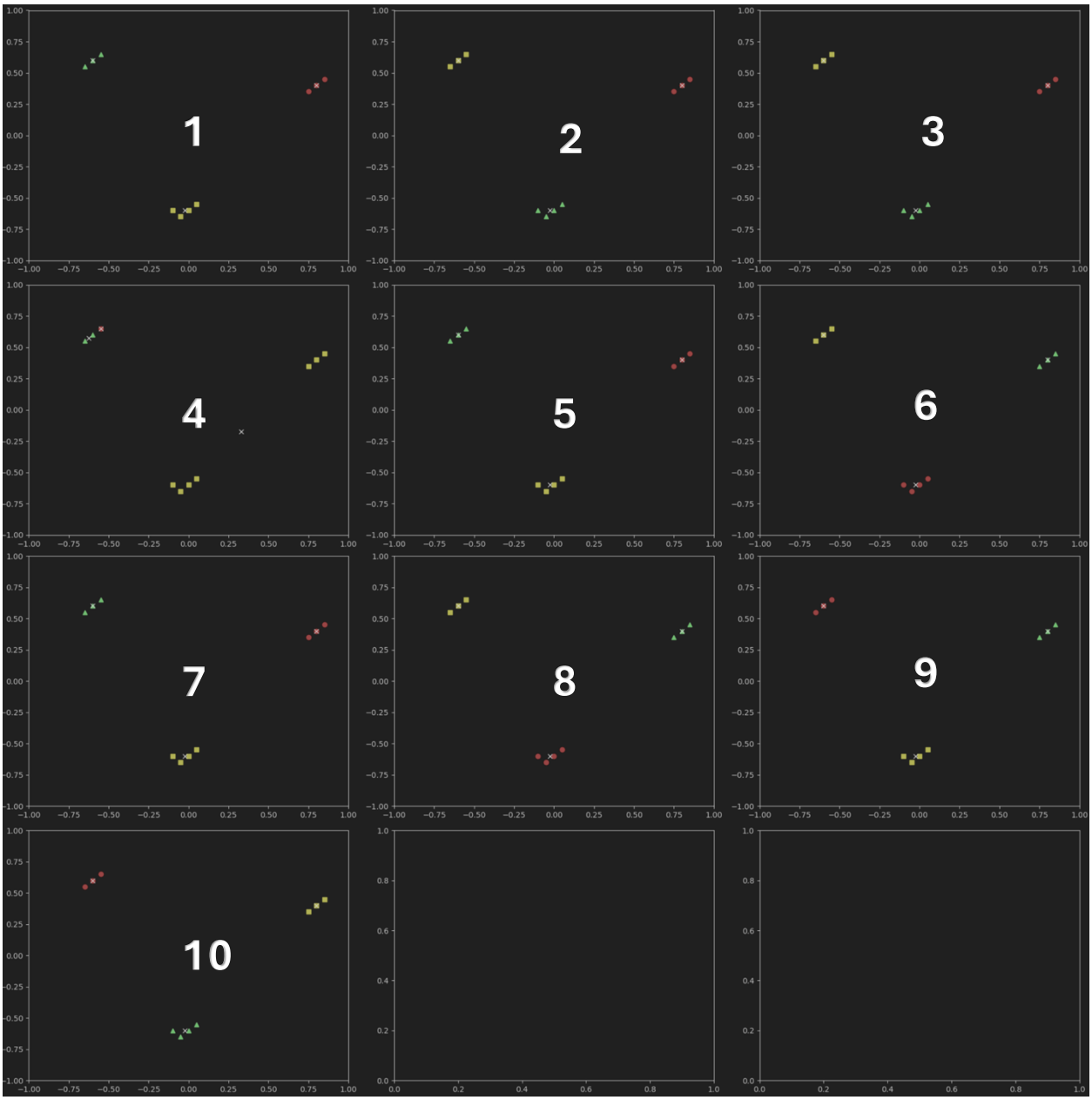
Problem 1:



As we can see from the plotted graphs, 1,3,5,7 have the same, and the most common pattern of clusters; 4 and 6 have the same clusters, but a little different from the previous ones. The rest are different from each other, but something noticeable is that 8 is the closest to the most common pattern of clusters, the difference is that the centroid from the green part is slightly too much to the right as well as the yellow one, if they both were a bit to the left it would've been the same pattern of clusters as the most common one. The weirdest one seems to be the 10th one because the red centroid is dominating the whole positive part on the y-axis, and that shouldn't be the case because the points are too far

apart to have a link/meaning between them. The 2nd one is not doing so shabby either because it takes almost all the negative values on the x-axis, which is not good either, given the fact that the points are too far away from each other, just like the case in the 10th image.

Problem 2:



It was almost a perfect first try, but it had to be that one case that destroyed the streak. In order to achieve this output, I thought about the following things:

1. The points should be at a significant distance from each other
2. Points from different clusters shouldn't be aligned on the same x/y axis

Problem 3:

Paper citation: Z. Chen and S. Xia, "K-means Clustering Algorithm with Improved Initial Center," *2009 Second International Workshop on Knowledge Discovery and Data Mining*, Moscow, Russia, 2009, pp. 790-792, doi: 10.1109/WKDD.2009.210.
keywords: {Clustering algorithms;Partitioning algorithms;Data mining;Iterative algorithms;Computer science;Electronic mail;Switches;Clustering methods;Convergence;Algorithm design and analysis;data clustering;k-means;initial center},

Link: https://ieeexplore.ieee.org/abstract/document/4772054?casa_token=ne1-hxkSmzwAAAAA:_kAz7LmL8tFM0E-M3C8Y0ovldHpAQYyYdQgGSXO8SZ1nAreIKanPdGc4fK_IXfnagEekbJFhWg

The Zhang Chen and Xia Shixiong improvement to K-means clustering focuses on enhancing the initialisation process of cluster centres.

What I have understood:

- Instead of pre-specifying the K number of clusters, the algorithm starts with each data point as its own cluster
- It iteratively merges clusters until it achieves an optimal number of clusters, determined by the data distribution
- The centres are chosen based on a neighbourhood density
- The initial clusters are assigned on a radius defined around each centre calculated using the maximum distance within clusters
- In the end, for optimal clustering, the algorithm iteratively merges clusters based on similar criteria.