Exercise 1 (Lesson 5, 10 points)**.** Read either https://arxiv.org/abs/1411.6652 or https://arxiv.org/abs/2002.04805 carefully and answer the six questions on page 3 of the syllabus to the extent possible given what we have learned so far in the course.

1. **What is your dataset and why is it interesting?**

   They use vision datasets MNIST, SVHN and CIFAR10 with small sample sizes. Computer vision is becoming increasingly important for everything from facial recognition to self-driving cars, so being able to generalize better (the goal of this paper) in vision classification tasks could result in more reliable models in a variety of common situations. The fact that the traininig is restricted to using small sample sizes makes the problem more interesting, as machine learning methods typically require mass amounts of data to generalize well.

2. **What type of insight do you hope to gain using TDA methods?**

   TDA methods can quantify how close the representations of data are in some latent space. The closer training samples are in this space, the better our model should generalize. To formalize this, we can construct a graph in our latent space using the training samples. If this space has a metric, we can measure how close together the nodes are by finding unit balls of fixed radius β (centered at each vertex) that cover the whole graph.

3. **What other methods have people used, or might be natural to use, to gain similar insight on this type of data?**

   Various implicit regularization strategies, such as the usual norm penalties on model weights, have been used. Explicit regularization strategies have been used to directly change the representation of the training samples. Other methods include clustering the representations and matching the representation distributions to Gaussians.

4. **How did you actually use TDA methods? Please describe both conceptual and implementation details.**

   They say that a graph is β-connected if the open balls of radius β centered at each vertex form a single connected component. This relies on the latent space $\mathcal{Z}$ being a metric space, but this isn't a problem since we can treat $\mathcal{Z}$ as a subset of $\mathbb{R}^n$.

   Conceptually, the smaller β is, the more tightly clustered a β-connected graph is. Thus if the latent representations of training samples have a low β, the probability density induced on $\mathcal{Z}$ is more concentrated. This will have a beneficial effect on generalization.

   To encourage β-connectivity during training, they tack on the following penalty to the loss:

   $$\mathcal{L}(B) = \sum_{i=1}^{n} \sum_{d \in \mathcal{D}_B} |d - \beta|,$$

   where B is some minibatch and $\mathcal{D}_B$ is the set of death times for the 0-dimensional Vietoris-Rips complex built from B. Since this is differentiable, it can be optimized with standard backpropagation techniques.

5. **What were the findings?**

   The experiments showed that the β of the training samples was highly correlated with the β of the testing samples, meaning the mass concentration effects carried over into testing. Cross validation was used for various values of β in order to choose a specific value for the final model.

   With this approximately optimal β, the paper's method outperformed 5 other state-of-the-art regularizers. Interestingly, though, it was noted that penalizing *any* deviation of the death times from β was actually more beneficial than just penalizing death times greather than β.

6. **What else would you try, given more time?**

   It seems natural to extend the notion of β-connectedness to higher dimensions and find β_n such that H_n becomes trivial when the unit balls at each vertex are of radius β_n. This would introduce more hyperparameters into the model, though, as it's not obvious how to weight the 1-dimensional penalties vs the 0-dimensional ones, etc.

   Additionally, since the paper noted that not allowing the death times to vary much from β actually hurt performance, it could be interesting to see how the model generalizes based on how far the death times are from β. It's not immediately obvious to me if there's an "optimal" distribution of death times, but there are standard distributions that could be tried, and their results could help build intuition about that idea.