Data 598 (Winter 2023): Homework 2

1 Performance vs. Width

In this exercise, we will experiment with the number of hidden units in a multilayer perceptron (MLP) with a single hidden layer. The number of hidden units is also referred to the **width** of the hidden layer.
Here are the details:

- The setup is identical to the demo/lab and you may reuse that code. Take the FashionMNIST dataset and randomly subsample 12% of its training set to work with. As a test set, we will use the full test set of FashionMNIST.
- Define a MLP with one hidden layer of width $h = 16$. Find the divergent learning rate η^* for this model and use a fixed learning rate of $\eta^*/2$, as we discussed in class.
- Train the model for 120 passes over the data.
- Repeat this procedure for widths $h = 8, 32, 128, 512$ with the same learning rate $\eta^*/2$ as above (i.e., you do not need to find the divergent learning rate of each model for this exercise).

The deliverables for this exercise are:

1. Make 4 plots, one each for the train loss, train accuracy, test loss and test accuracy over the course of training (i.e., the metric on the y-axis and number of effective passes on the x-axis). Plot all 4 lines, one for each value of h on the same plot.
2. When the training accuracy is 100%, the model is said to **interpolate** the training data. What is the smallest width at which we observe perfect interpolation of the training data?
3. As we vary the width of the network, at which training epoch do we observe perfect interpolation of the data? That is, make a plot with h on the x-axis and number of passes over the data required for interpolation on the y axis.

Some notes for broader context:

- For the full dataset, we can make the same observation, but the smallest width of the network which can perfectly interpolate the training data is much larger.
- While networks of a range of widths can perfectly interpolate the training data, the test errors they obtain can be different. This is one of the observations that makes tuning a neural network architecture more of an art than a science.
- A general rule of thumb is that the neural network should be large enough to interpolate the training data.

2 (Bonus) Divergent Learning Rate, Accuracy and Width

Consider the same setting as in Exercise 1 above.

Part 2.1 Find the divergent learning rate η_h^* for width $h \in [4, 8, 16, 32, 128, 512, 2048]$. Make a plot for the divergent learning rate versus the hidden width.

Part 2.2 For a given width h, run SGD for 1 pass over the data with learning rate $\eta \in [\eta_h^*, \eta_h^*/2, \eta_h^*/4, \eta_h^*/8]$, where η_h^* is from Part 2.1. Measure the test accuracy for each of these learning rates. Let A_h denote the best accuracy obtained here. Repeat this procedure for each of the widths considered in Part 2.1.

Part 2.3 Make a plot of the **best** test accuracy A_h at the end of one pass over the data as the width h is varied.