Policies

- You are free to collaborate on all of the problems, subject to the collaboration policy stated in the syllabus.

- Please submit your report as a single .pdf file to Gradescope under "Homework 4" or "Homework 4 Corrections". **In the report, include any images generated by your code along with your answers to the questions.** For instructions specifically pertaining to the Gradescope submission process, see https://www.gradescope.com/get_started#student-submission.

- Please submit your code as a .zip archive to Gradescope under "Homework 4 Code" or "Homework 4 Code Corrections". The .zip file should contain your code files. Submit your code either as Jupyter notebook .ipynb files or .py files.

1 Autoencoder vs. PCA [80 points]

This problem revisits the hands-on notebook https://github.com/jmduarte/phys139_239/blob/main/notebooks/07_Autoencoder.ipynb. For this problem, we will use 50k jet images total (`max_-jets = 50000`) as was done in class. We *highly recommend* using the GPU-enabled DataHub with the latest `jmduarte/phys139_239:latest` image for this problem.

Problem A [15 points]: As discussed in class, the reconstruction loss, i.e. the mean-squared error (MSE), can be interpreted as an *anomaly score*, which roughly tells you how anomalous a jet is. We can use this anomaly score as a discriminator between QCD jets (background) and top quark jets (signal). Plot the ROC curve for the MSE loss of the autoencoder on the merged testing sample of QCD and top quark jets. Assume a label of 1 for top quark jets and a label of 0 for QCD jets. You can use the existing functions from earlier notebooks or `sklearn.metrics.roc_curve` and `sklearn.metrics.auc` directly. Report the AUC.

Problem B [25 points]: Principal component analysis (PCA) is another way of doing dimensionality reduction. Perform (fit) a PCA on only the QCD training images using `sklearn.decomposition.PCA` with 2 components. Note you will have to reshape the image tensors so that they are 2D (as required by PCA) instead of 4D (as required by the autoencoder), e.g. `qcd_test_images.reshape(-1, im_size * im_size))`.

Plot the distribution of the PCA reconstruction losses for top quark jets and QCD jets, separately, using the testing data. To evaluate the reconstruction loss, you will have to first apply the PCA transformation (`transform`) to the QCD and top quark jets, then apply the inverse transformation (`inverse_-transform`) and compare the output with the original input. You can follow the loss calculation as was done earlier in the notebook for the autoencoder. The key is to make sure the tensor shapes are correct: at the end you want a single number for each data point representing the loss value for that image.

Hint: Review https://web.archive.org/web/20230329163611/https://rittikghosh.com/autoencoder.html, which is also relevant for the following parts.

Problem C [15 points]: Plot the ROC curve for the PCA similar to how you did in Problem A. Report the AUC. Is it greater or smaller than the AUC achieved by the autoencoder?

Problem D [25 points]: Create a scatter plot for the QCD and top quark test images in the 2D latent space for both the autoencoder and the PCA. Color the data points depending on whether they are QCD or top quarks (i.e. one color for QCD and a different color for top quarks).

For the autoencoder, this will require using the `encoder` model's predict function with the appropriate input(s).

What do you observe about the difference between QCD and top quark jets in these two spaces generally?

How is the 2D space of the autoencoder different from the 2D space of PCA?