#### Geometric Partition Entropy for identifying Optimal Training Set for Classification Tasks

## Overall Project Plan

Identify subsamples of large training datasets (these subsamples are assumed to share a similar structure and distribution as the whole dataset) so that neural network classifiers perform almost as well as they would if trained on the full original training data.

## Logistics

Email for communicating progress and asking questions during the project. (Maybe need a more convenient way like Slack)

GitHub for sharing the code.

Meetings: every Thursday 11AM EST → on Zoom

## Weekly Progress Log

|  |  |  |
| --- | --- | --- |
| Time | To-Do | Done |
| Feb 08, 2024 | Xiao Wen:   1. Understand the overall structure of the AFRL project. 2. Decided the image dataset and CNN architecture | Xiao Wen:   1. Done 2. Done (Cifar-10 dataset and one sample CNN from Kaggle) |
| Feb 15, 2024 | Xiao Wen:   1. Train the sample CNN with 45000 training images, 5000 validation images, and test on 10000 testing images (make sure to achieve a decent classification accuracy) 2. Modified the sample PyTorch code and extracted the intermediate feature space before the final classification layer for the dataset and subsets. 3. Clean the code and upload to GitHub | Xiao Wen:   1. Done 2. Done (need to verify it is done in the correct way) 3. Done |
| Feb 22, 2024 | No meeting this week  Xiao Wen   1. Random sample 5000 images multiple times as subsets 2. Extract their intermediate feature spaces. 3. Apply Boltzmann-Shannon Interaction on singular values after applying SVD to the subsets. |  |