Final Report

Introduction:

Project Title:

Geometric Partition Entropy for Identifying Optimal Training Sets for Classification Tasks

Motivation/background/business problem:

In machine learning classification tasks (could be expanded to regression as well), obtaining sufficient labeled data to train models is often crucial for models to achieve accurate prediction results. However, due to various constraints such as time limitations or hardware restrictions in the real world, acquiring enough labeled data can be challenging. In such cases, it becomes imperative for us to be able to still maintain a decent prediction accuracy of the machine learning model despite limitations on sample size we could use to train them. In order to achieve this goal, we often need to **optimize the distribution of clustered data within a high-dimensional feature space** given the quantity of data is limited. Even though entropy-based measures offer a natural approach for handling such distributions, existing estimators have encountered difficulties, particularly in dealing with the curse of dimensionality, sparse samples, and the presence of outliers etc.

This project aims to address these challenges by introducing a data-efficient learning approach, specifically a novel form of entropy quantification called Boltzmann-Shannon Interaction Entropy used for datasets in high-dimensional feature spaces. The project will explore the relationship between entropy measures of training sets and the performance of models trained on those sets. **The ultimate goal is to provide insights and guidance on optimal sub-sampling or data collection strategies for creating minimally effective training sets.** The study will focus on well-known datasets, namely CIFAR-10 and CIFAR-100, which are extensively studied and widely available.

The computation of the entropy metric itself is straightforward, but the project's **main challenge lies in feature extraction and other preprocessing tasks** essential for robust classification tasks. In other words, **we should be focusing on the data collection, cleaning, and preprocessing phases** (instead of the ML model training phase) and **apply further mathematical analysis on it**.

This capstone project offers a comprehensive introduction to image classification tasks using highly regarded datasets, complemented by more practical benefits (using limited

amount of training data to train machine learning models efficiently and effectively). Additionally, it provides a unique perspective for further analysis in the field of machine learning and information theory.

Problem statement:

Identify smaller subsets from extensive training datasets (such as CIFAR-10) that yield machine learning classifiers' performance nearly equivalent to that achieved when trained on the complete original dataset.

Existing work or literature review:

1. Geometric Partition Entropy: Coarse-Graining a Continuous State Space by Christopher Tyler Diggans and Abd AlRahman R. AlMomani. This literature mentions that entropy serves as a measure of uncertainty and ignorance in predicting continuous phenomena. However, the traditional approaches, based on thermodynamics and Shannon's theory, are fundamentally discrete. This means they are not ideally suited for continuous phenomena. Therefore, the literature proposes a new approach named Geometric Partition Entropy to quantify uncertainty in predicting continuous phenomena using entropy, which is more suitable, consistent, and informative compared to traditional methods, especially in cases of complex distributions or limited sampling.

2. Boltzmann–Shannon interaction entropy: A normalized measure for continuous variables with an application as a subsample quality metric by Christopher Tyler Diggans and Abd AlRahman R. AlMomani. It introduces the concept of Boltzmann–Shannon interaction entropy (BSIE) as a parameter-free measure for quantifying uncertainty in continuous probability distributions. It highlights its advantages over traditional entropy estimators (in handling sparse samples and extreme outliers) and demonstrates its utility in evaluating subsampling quality in regression tasks. Furthermore, unlike the Geometric Partition Entropy approach, this method is normalized and unbiased, so the entropy values won't be impacted by the amount of variance due to the distinction in sample size.

3. Reviewing Singular Value Decomposition (SVD) (especially the parts about what information the decomposed eigenvectors and singular values provide) as well as the concept of Principal Component Analysis (PCA) would also be beneficial to attain more valuable insight with respect to the overall approach and workflow of the project.

Overall approach:

We currently are using the CIFAR10 dataset and a sample CNN from Kaggle to do the followings:

1. Train the sample CNN with training set in CIFAR10 and make sure the classification accuracy on the test set is good enough (around 80 percent). We save this trained model which serves as our embedding function to extract feature space of CIFAR10 images.

2. Map all the images in the CIFAR10 training set into the pretrained model to attain the intermediate feature space (just before the final classification layer of the CNN model) and therefore have a matrix with the dimension of number of features times number of training images.

3. Apply Singular Value Decomposition (SVD) to this matrix:

    i) Will have one matrix U (containing orthonormal vectors) in the feature space.

    ii) Will have a matrix of singular values (the middle diagonal matrix). These values are the variances associated with the data in the direction of the singular vector. Singular values are arranged in descending order along the diagonal.

    iii) Will have one matrix V^T (containing orthonormal vectors) in the data space. (images)

    iv) We will focus mainly on the diagonal middle matrix (singular values) analysis and try to explain how it can help us to identify an optimal training subset. However, the matrices U and V should also be used later to supplement our research purpose especially *when the information provided by the diagonal matrix is not sufficient to help us identify an optimal training subsample*.

4. Randomly subsample an equal number of images from each category in the training data and combine them together to form one subsample. Or this subsample can be randomly drawn as well from the whole training set without the need to maintain the same number of images from each class. **This is especially true when we want to draw a biased subsample intentionally** (that is, we oversample from one or two particular classes while under sample from the rest of the classes) Repeat this process multiple times.

5. Apply SVD to these subsamples of the CIFAR-10 dataset. So, we will have singular values with respect to the subsamples as well as the original singular values with respect to the whole dataset.

6. Figure out whether the singular vectors of these subsamples are pointing in the same direction as the singular vector of the CIFAR-10 dataset (can use dot product etc.) Note: later in the research, this step is not actually being used nor it is necessary to do so, but it provides me with some innovative ideas when it comes to exploring the approaches to strengthen the research goal when the current methods may not be enough to produce a convincing result.

7. If they share similar directions as the whole dataset one, then figure out whether they are similar to the whole dataset in geometric distribution by inputting the singular values of both subsamples and the whole dataset into Geometric Partition Entropy/Boltzmann Shannon interaction Entropy and compare their output values.

8. If they are similar to the output entropy value of the singular values of the whole training set, then we can expect these subsamples to have similarly good training performance on machine learning models as the entire training dataset of CIFAR-10 as they share a similar data structure as the whole training set.

9. Other more innovative approaches are in the future work section.

Methods:

Data (if applicable):

CIFAR-10 dataset. The CIFAR-10 dataset is a widely used benchmark dataset in the domain of machine learning and computer vision. It is frequently employed for tasks such as image classification and object recognition. In this project, we will use it to train and further evaluate the performance of sample Convolutional Neural Networks from Kaggle in the image classification task. We will also extract the internal feature space of these images to do deeper analysis.

Overview:

- CIFAR-10 stands for "Canadian Institute For Advanced Research" and "10" signifies that it contains 10 classes.
- The dataset consists of 60,000 32x32 color images for 10 classes, with 6,000 images per class.
- These classes include common objects and animals: airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks.

- The dataset is divided into training and testing sets, with 50,000 images in the training set and 10,000 images in the test set.
- It is a balanced dataset, meaning each class contains an equal number of images.
- The CIFAR-10 dataset is freely available for academic and research purposes, and it can be accessed through various platforms and repositories, including the official CIFAR website or the TensorFlow and PyTorch libraries.

Note: we could also use CIFAR-100 dataset for this project if time permits.
Update: after discussion with Dr.Diggans and his colleagues in machine learning, it turns out that the Cifar-10 dataset may not be ideal for us do the task in this project. We might want to consider that datasets as well such as NMIST etc.

Processing:

We shuffle the original 50,000 training images and divide them into two parts: training set with 45,000 images and validation set of 5,000 images. The remaining 10,000 images belong to the test set (to evaluate and compare the classification accuracy of CNN models we will train with the whole training set and with subsamples).
But because we compare the entropy values of subsamples with the entropy of the whole training set (meaning containing all the 50000 images), we might want to instead use the whole 50000 images to train the embedding function.

Analytic Methods:

- After extracting the feature spaces of both subsamples and whole training set of CIFAR-10, we will then attain their respective singular vectors via Singular Value Decomposition (SVD) operation and followed by Boltzmann Shannon Interaction Entropy to get their corresponding output values. (The appropriate inputs to the BSIE function might subject to change because it depends on the singular values after applying SVD on the feature space)
- We will then plot these entropy values of subsamples as well as the entropy value of the CIFAR-10 training set. Through comparison, we can discover which subsample entropy values are closest to the entropy value of the entire training set.
- Regarding these closest entropy values, we will retrieve their associated subsamples and use these subsamples to train new Convolutional Neural Networks respectively.
- Compare the performance (image classification accuracy) between these newly trained networks with the original network. We can also compare them to the performance of newly trained networks from the subsamples whose entropy values might not be so close to the entropy value of the original training set.

- We can derive various methods to test whether a subsample whose entropy value is closer to that of the training set will result in a CNN model with higher prediction accuracy. For example, we can select one sample from each class of the training set, then compute the entropy value for each of them. We compare them with the entropy values of their parent classes one by one correspondingly. After we combine these samples from each class to form a subset to train a CNN model, we can evaluate this model's performance on each of the classes to see if a representative sample (similar entropy value to its parent class) from one particular class will result to an associated higher prediction accuracy of the model on that class.

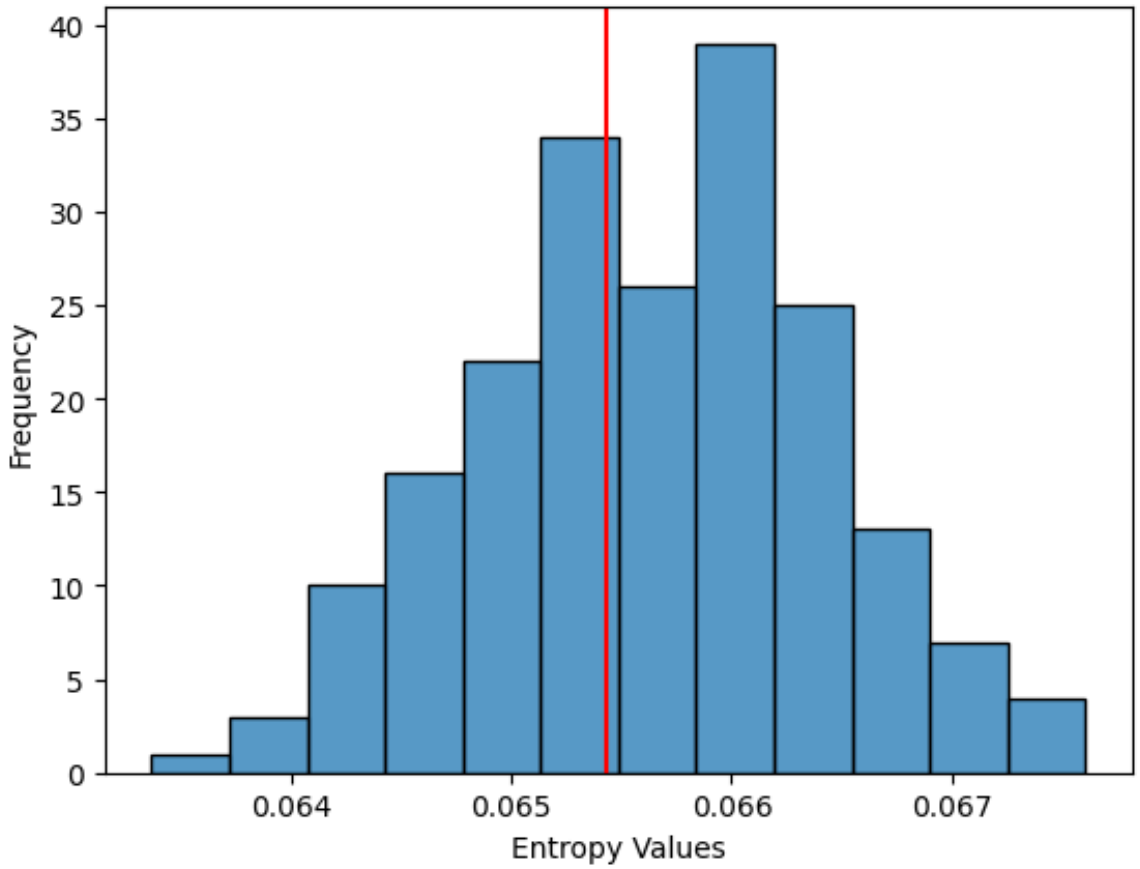
Evaluation:

We are trying to figure out if subsamples (given they are good representation of the original dataset in terms of data geometric distribution) could also be used to train Convolutional Neural Network and achieve decent performance. Hence, the evaluation is to observe the classification accuracy of these newly trained models on the CIFAR-10 training and testing sets and compare them with the performance of the original Convolutional Neural Network trained with the whole training set.

Note: because we only draw a subsample from its training set to train a CNN model, it means the **remaining images not in this subsample can also be used new data the model hasn't seen so far to evaluate the model performance**, which even is a better case in a sense that it will strengthen our hypothesis even more if the result is promising as we expect.

Results:

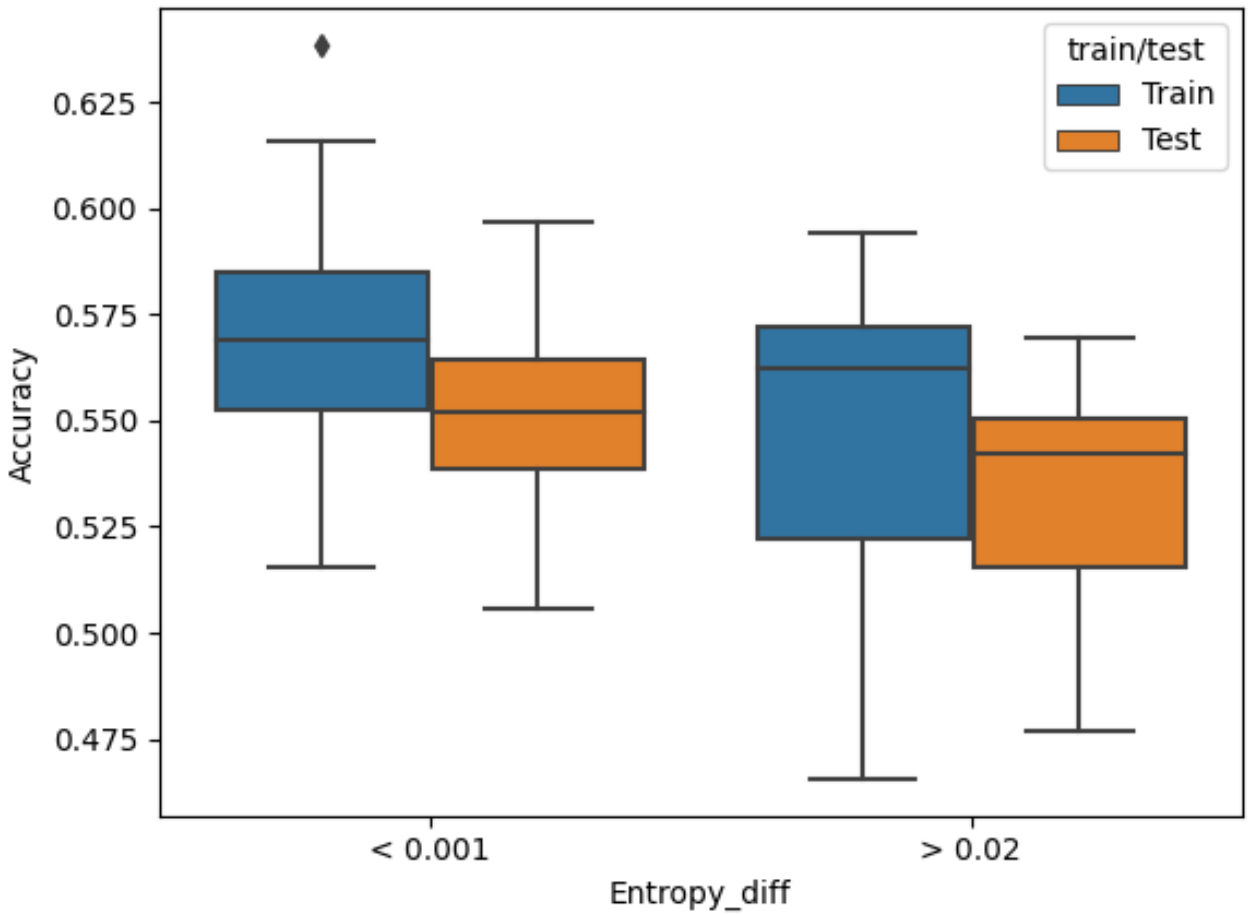
Histogram of BSIE values of 200 subsamples and the BSIE value of original training dataset



As we can see, the entropy value of the whole training dataset (red line) is almost like a median value with respect to entropy values for other subsamples. The next step would be to use each subsample to train the Convolutional Neural Network respectively, test their classification accuracy on the test set, and compare the results with the performance of Convolutional Neural Network trained with the original entire training set.

In fact, we can expect that if a certain subsample possesses an entropy value close to the whole dataset one, then the new model trained with this subsample could have a similarly good classification performance like the original model even though such classification performance is not entirely guaranteed.

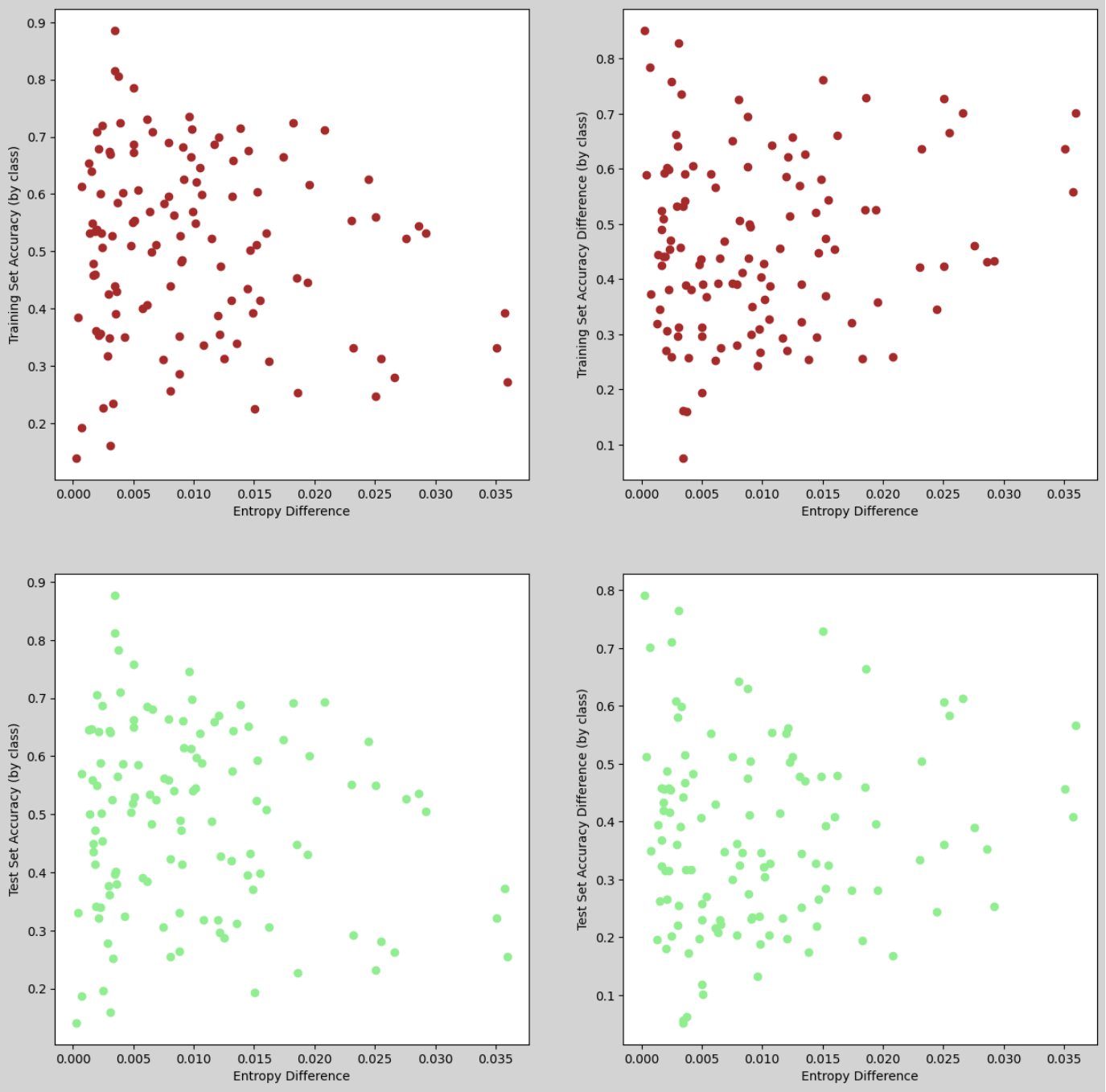
We can use a boxplot to display the result we attained so far:



That is, when we have subsamples whose entropy values are closer to the entropy value of the whole training set, the ML models trained with these subsamples often outperform the ML models trained with subsamples whose entropy values are more distinct from that of the whole training set.

Note: this trend is more apparent when it comes to the training set accuracy compared to the test set accuracy.

Another way to show the trend is to select an equal amount of data points (subsample) from each class. If one subsample from a particular class shows an entropy value similar to that of its parent class, then after we combine these subsamples to train a ML model, we want to figure out if that model will have higher prediction accuracy for other data points (it hasn't seen) in that class.



The scatterplots above somehow demonstrated our hypothesis subtly even if the correlation between the entropy difference (between subsample and its parent class) and the train/test accuracy is still a bit weak.

In the end, we can try different approaches to construct our embedding functions (which is just the machine learning model we train and used for extracting feature space) or change

the way to draw subsamples etc. and these approaches will likely yield distinctive model performance.

App or Tool (if applicable):

The machine learning architecture we chose for the AFRL project is one sample Convolutional Neural Network implemented with PyTorch from Kaggle. This model is deployed to conduct image classification tasks upon training. The robustness of this architecture can be guaranteed in a sense that it follows the conventional structure (which proves to be good) to build Convolutional Neural Network (Convolution layer, Activation function, Pooling layer, Fully Connected layer etc.) and it consistently maintains a classification accuracy above 75 percent on the unseen testing images. The same architecture is used whenever we are building the embedding function or train a CNN model with subsamples.

Discussion:

Summary:

This project aims to employ different data-efficient learning approaches to select representative subsamples to train machine learning models (Convolutional Neural Network in this case). The resulting models maintain a similarly good performance in prediction accuracy as if the models are trained with the original entire training dataset in Cifar-10. Simultaneously, because we only use a subsample to train the ML model, we also optimize the training time during this process.

Conclusion/Take home messages:

- There exist multiple ways to figure out if a subset of data shares a similar structure as the original dataset. Using entropy method is simply one of them. However, at the same time, what is so special about the BSIE approach is that this method will generate values to allow you to compare the similarities in data structures between datasets **regardless of the size of the datasets** (in other methods, this metric will often heavily impact the similarity comparison).
- If a subset of data is a good structural representation of the original dataset, the machine learning model trained on it could also display decent performance as if trained on the original dataset.

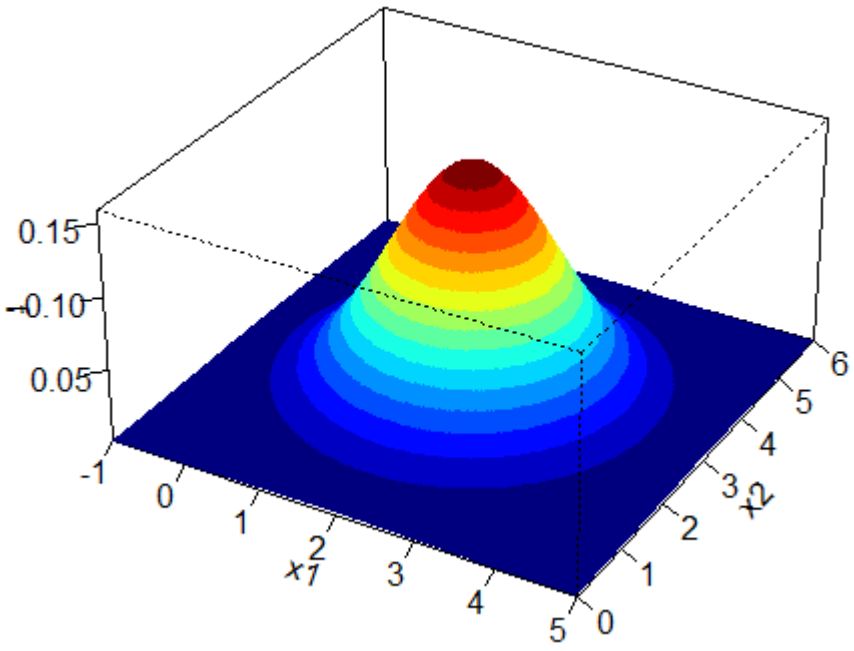
Future work:

- Use pretrained models from PyTorch torchvision class to extract feature spaces of images.

- When constructing the embedding function:
  1) We can also try to use a subsample (rather than the whole training dataset) to do so even if this subsample may not be a perfect representation of the whole training dataset.
  2) We can even try to use the raw pixels of the Cifar-10 images. In other words, we don't use the embedding CNN function to extract the feature space (therefore we could avoid the cost to train it), and we simply flatten pixels of each image into a column vector and combine them to form a matrix. Then we can apply SVD and BSIE to this matrix as usual. (However, this approach hasn't produced any promising results so far)

- Still use the whole training dataset to construct the embedding function, but this time extract the feature space (and do SVD and BSIE) with respect to each class in the training set. After we attain a good interpretation on the data geometric distribution regarding each class, we can find the most representative subset from each class and combine them together to form a combined subsample. We can also select the most unrepresentative ones from each class deliberately and combine them to form a bad subsample to set up a contrast in prediction accuracy between the most representative subsample and the most unrepresentative subsample. We can slightly modify the approach as well. For example, we can try to select only one representative subset from one class and combine it with the subsets (that are not very representative) from other classes. Then we can train the CNN model on this combined subsample and observe if the model performs well only in classifying images from that particular class whose subset is representative.

- Currently, when we compare the entropy value of one subsample against the entropy value of another, we still only concentrate on the singular values of the middle diagonal matrix after applying SVD to the subsample feature space. In other words, the left U and right V^T matrices are left unused. These two matrices contain the orthonormal eigenvectors which tell us about really essential information about data distribution (in a new eigenspace, not in the original coordinates), so they should not be neglected. In other words, besides using entropy value to distinguish whether a subset is representative or not, the eigenvector similarity between the subset and the overall training set should also be an important metric for us to consider. In the end, this approach will generate a new metric and it could be better to use it along with the entropy metric so that we can have a 3 dimensional plot that whenever we have a point close to the origin (in x-y plane), we can regard this point (representing a subsample) is close to the overall training set in terms of the entropy metric and the eigenvector metric. It looks something like the following:



As you might guess already, the third axis (height) represents the prediction accuracy of each of the models we train with subsamples.

Ethical considerations:

None and all the data we used is publicly available currently. We might consider other approaches to test our hypothesis and use other datasets (besides Cifar-10), but the ethical issue still seems trivial.

Other considerations:

The goal of the project is to identify an optimal subset (or multiple of them) so that we can use less data to train ML models in an efficient way without substantially decreasing its prediction accuracy. We can do this by extracting the feature space and analyzing geometric structure and data distribution of subsamples, but to extract such information from subsamples, we have to train the model first on the whole dataset anyway to attain the embedding function.

Concession:

1) The correlation between the entropy value and accuracy seems to be weak, so we might want to add other metrics. (Plus, it is hard to attain a subsample whose entropy value is very different from the entropy value of its parent dataset.
2) The improvement in prediction accuracy by using this method to identify an optimal training subset is still limited.

This project is a solo project, and the only group member is me (Xiao Wen, UNI: xw2943). Currently, I have done all the work up appointed by my mentor Dr.Diggans. However, the preliminary result I attained so far is not promising and convincing, so I need to further reexamine my code and come up with new approaches.

Reference:

- Christopher Tyler Diggans, Abd AlRahman R. AlMomani. 2022. Geometric Partition Entropy: Coarse-Graining a Continuous State Space
- Christopher Tyler Diggans, Abd AlRahman R. AlMomani. 2023. Boltzmann–Shannon interaction entropy: A normalized measure for continuous variables with an application as subsample quality metric
- 3D Normal Distribution plot https://datasciencegenie.com/wp-content/uploads/2020/04/DifferentMeans3D.png