STAT 102B: **Final Project**

George Michailidis

Due electronically through **BruinLearn/Gradescope**
on June 12 at 5 pm

NO EXTENSIONS WILL BE GRANTED

Problem 1:

Consider the three data sets (`regression_data_node1.csv, regression_data_node2.csv, regression_data_node3.csv`) comprising a response variable y and 600 predictors $X_1, \cdots, X_{600}$.

Part (a): (10 points)

Write code to implement the coordinate descent algorithm for lasso regression.

Discuss which stopping criterion you decided to implement.

For each dataset, use the first 80% of the entries as the training set and the remaining 20% as the validation set. Use the validation data to tune the regularization parameter $\lambda > 0$ for the Lasso regression model.

1. Report the values of the regularization parameter λ that yielded the best models for each of the three data sets, based on validation loss.
2. For each data set, report the **indices of the non-zero regression coefficients** in the final selected model.
3. Identify and report the **indices of the regression coefficients that are non-zero across all three models** (i.e., the intersection of non-zero coefficients).
4. Report the test loss of the final model selected for each data set, using the test data set.

Part (b): 10 points

Since the three data sets differ in sample size, the regression coefficients estimated from each will naturally vary.

To address this, the owners of the three data sets agree to **collaborate in estimating a shared regression coefficient vector**, while strictly avoiding any direct sharing of their respective data sets.

They adopt the following algorithm:

- Each data owner splits their data into 80% for training purposes and 20% for validation purposes (tuning the λ parameter).
- Initializing with the zero vector, each owner runs 5 iterations of coordinate descent for lasso regression using a local regularization parameter λ_k.
- After the local updates, each owner sends their current estimate of the regression vector to a **trusted aggregator**.
- The aggregator computes a weighted average of the received regression vectors, with weights proportional to the respective sample sizes, and broadcasts the resulting global estimate back to all owners.
- This process repeats until the aggregator detects that the global regression vector has changed by less than a predefined tolerance (10^{-6}) between two successive rounds.

Each data owner independently selects their tuning parameter λ_k, for $k = 1, 2, 3$, by minimizing the validation loss on their respective validation sets.

Once the aggregator determines that the algorithm has converged, it evaluates the final aggregated model using the available test data set.

1. Report the values of the regularization parameters λ_k selected by each data owner based on validation loss.
2. Report the **indices of the non-zero regression coefficients** in the final aggregated model.
3. Compute the **confusion matrix** between the regression coefficient reported by the aggregator (treated as the "ground truth") and each of the best three individual models computed in Part (a).
4. Report the test loss of the final selected model, evaluated on the test data set.

How do you conclusions change, if the aggregation occurs **every 10 iterations?**

Organization of the Report

For both Part (a) and (b), your report should be organized as follows:

- Answer the questions asked directly and summarize your findings supported with informative plots and tables.
- Your code should be given in a separate file. The code should carefully documented, where you comment what each piece of the code tries to accomplish (e.g., you have a function that implements soft-thresholding, you have another function that implements coordinate descent, etc.)