This take-home portion of the midterm consists of 2 exercises. Recall that there is no collaboration allowed. Copying solutions from elsewhere (e.g., online or in books) is also forbidden. However, you are allowed to post questions in the discussion forum on Canvas.

The submission deadline is **Sunday May 17th, 11:59pm**. No extensions are allowed and no late submissions will be accepted, regardless of the excuse. Please upload to Canvas: (1) An html or pdf file with all of your output and comments; and (2) The code file(s), which may have the extension .ipynb or .py.

1 Exercise 1

In this exercise, you will implement in **Python** a first version of your own ℓ_2^2-regularized binary classification with exponential loss. You will write your own codes for all functions: accelerated gradient algorithm, ℓ_2^2-regularized binary classification with exponential loss, hyper-parameter search on a validation set.The ℓ_2^2-regularized binary classification with exponential loss is a supervised binary classification method, similar to ℓ_2^2-regularized binary logistic regression.

$$\min_{\beta \in \mathbb{R}^d} F(\beta) := \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i x_i^T \beta\right) + \lambda \|\beta\|_2^2 \,. \tag{1}$$

- Compute the gradient $\nabla F(\beta)$ of F.
- Consider the Spam dataset from *The Elements of Statistical Learning*. Standardize the data, if you have not done so already. Be sure to use the training and test splits from the website. You can find the link to the train/test split here: `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`
- Write a function *myclassifier* that implements the accelerated gradient algorithm to train the ℓ_2^2-regularized binary classification with exponential loss. The function takes as input the initial step-size for the backtracking rule, the ε for the stopping criterion based on the norm of the gradient of the objective.
- Train your ℓ_2^2-regularized binary classification with exponential loss and $\varepsilon = 0.005$ on the Spam dataset for $\lambda = 1$. Report your misclassification error for this value of λ.
- Write a function *hyperparamsearch* that implements grid search to find the best value of λ in terms of misclassification error on a logarithmic grid of values on a validation set.
- Find the optimal value of λ in terms of misclassification error using grid search on a validation set, with a 60%/40% split for training/validation. Report your misclassification error for the best value of λ found. Report the specificity and the sensitivity for the best value of λ found.

2 Statistical machine learning inquiries

In this exercise, you are facing simple scenarios and asked to tackle them based on your knowledge of statistical machine learning and data science in general. Please answer the questions for each case in a few lines. Please feel free to write equations or draw figures if it helps.

Contagion. Sharon and Billie, two Stanford coronavirus researchers, heard you were taking DATA 558 at UW. They also heard you had learned about performance metrics. They recruited you to help them get the statistics correct in their revised report. This time, they report that among 3324 pre-COVID-19 samples tested using an antibody test, 16 tested positive. Estimate the sensitivity and specificity of the antibody test based on this information.

Flying saucers. Steven and Kathleen are passionate about flying saucers. They have designed a machine-learning based system that allows them to detect whenever a UFO passes by in the sky. It just so happens that you are good friends with an extraterrestrial being who happens to know all flight times and locations of flying saucers. How would you evaluate Steve and Kathleen's flying saucer detector? What performance metric could you use to assess whether their flying saucer detector is a fake or not?