

# ST 453 Advanced computing for statistical reasoning

## Homework problem set 8

October 28, 2024

**No R packages are permitted for use in this assignment.**

1. Fixing some  $\beta \in \mathbb{R}^p$  with  $p = 1000$  and  $\mathbf{n} = \mathbf{300}$ , generate a synthetic data set from a logistic regression model. Next, use a low-rank SVD approximation to the design matrix (similar to what we have done previously for linear regression) to fit the logistic regression model with a **gradient descent** algorithm using the full data set (i.e., **not stochastic nor batch gradient descent**). After you fit the low-rank logistic regression model, transform the coefficient estimates from the low-rank model to coefficient estimates in the full model (note that these coefficients are not identifiable), and (write an R function to) plot the receiver operating characteristic (ROC) curve ([https://en.wikipedia.org/wiki/Receiver\\_operating\\_characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)) on both the training data and on an out-of-sample test set of size 300.
2. Repeat problem 1, but with a training sample size of  $\mathbf{n} = \mathbf{2000}$  and using **stochastic gradient descent**. Decide on an optimal choice of the subsample size,  $r \in \{1, \dots, n\}$ , to evaluate the gradient at each iteration.
3. Repeat problem 1, but with a training sample size of  $\mathbf{n} = \mathbf{2000}$  and using **mini batch gradient descent**. Decide on an optimal choice of the batch size,  $r \in \{1, \dots, n\}$ , to evaluate the gradient at each iteration.