CS/ECE/ME 532
Homework 6: Iterative Algorithms for Regularized LS

1. **Landweber convergence**. Consider the Landweber iteration for solving a standard least-squares problem with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, and $\boldsymbol{A}$ has full column rank. Recall that this iteration begins with some initial $\boldsymbol{x}_0$ and then:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \tau \boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) \qquad \text{for } k = 0, 1, \ldots \tag{1}$$

   a) We expect the algorithm to converge to $\boldsymbol{x}_\star = (\boldsymbol{A}^\mathsf{T}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{b}$. Define the *error* as $\boldsymbol{e}_k := \boldsymbol{x}_k - \boldsymbol{x}_\star$. Show how to rewrite (1) in the form $\boldsymbol{e}_{k+1} = \boldsymbol{P}\boldsymbol{e}_k$. What is the matrix $\boldsymbol{P}$?

   b) Define the *residual* $\boldsymbol{r}_k := \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}$. Show how to rewrite (1) in the form $\boldsymbol{r}_{k+1} = \boldsymbol{Q}\boldsymbol{r}_k$. What is the matrix $\boldsymbol{Q}$?

   c) Let $\{\sigma_i\}$ be the singular values of $\boldsymbol{A}$. Prove that when $0 < \tau < \frac{2}{\sigma_1^2}$, we have $\lim_{k\to\infty} \boldsymbol{e}_k = \boldsymbol{0}$. **Hint:** substitute the SVD of $\boldsymbol{A}$ into your expression for $\boldsymbol{P}$.

   d) Prove that if $\boldsymbol{A}$ is rank-deficient and $\boldsymbol{x}_0 = \boldsymbol{0}$, then the Landweber iteration converges to the minimum norm solution. **Hint:** redo part **a)** using $\boldsymbol{x}_\star = \boldsymbol{A}^\dagger\boldsymbol{b}$ and see how this affects part **c)**.

2. **Data Fitting vs. Sparsity Tradeoff**. For this problem, we'll use the dataset `BreastCancer.mat` (see the journal paper posted on Moodle for more details about the dataset). The goal is to solve the Lasso problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

   Here $\boldsymbol{\beta}$ is the weight vector applied to the expression levels of 8141 genes and $\boldsymbol{y}$ is the vector of labels ($+1$ and -1) for each of the 295 patients. In this problem we will vary λ to explore the tradeoff between data-fitting and sparsity, and we will compare the Lasso to Ridge Regression.

   a) Write an implementation of the ISTA (iterative soft-thresholding) as covered in class (and in the posted notes on the Moodle) that solves the above Lasso problem.

   b) For each λ, use your code to find the optimal weights using only the first 100 patients (first 100 rows). Plot a trade-off curve with the residual $\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2$ on the vertical-axis and $\|\boldsymbol{\beta}\|_1$ on the horizontal-axis and. The vertical-axis should contain the residual from the training data. Explain what you see. **Note:** you'll have to play around with how you choose λ to see the whole curve!

   c) With your solutions from part **b)**, produce a new trade-off curve. This time, plot the error rate on the vertical-axis versus the sparsity on the horizontal-axis. Here, the error rate is the number of correct predictions divided by the total number of predictions. The sparsity is the number of nonzero entries in β. For this purpose, we'll say an entry β_i is nonzero if $|\beta_i| > 10^{-6}$. Again, the vertical-axis should contain the error rate from the training data.

d) Repeat parts **b)** and **c)**. This time for the vertical-axis use the residual (respectively the error rate) from the validation data. For validation, use the remaining rows of the data matrix (the ones not used for training). Again, explain what you see in both trade-off plots.

e) Compare the performance of the Lasso and Ridge Regression in this application. Do this by the following steps. Randomly split the set of 295 patients into 10 subsets of size 29-30. Use these subsets for training, tuning, and testing. Use the data in 8 subsets to find a solution $\widehat{\boldsymbol{\beta}}$ to the Lasso optimization above and to the ridge regression

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2 \ .$$

Repeat this for a range of λ values, each yielding a solution $\widehat{\boldsymbol{\beta}}_\lambda$. Then compute the prediction error using each $\widehat{\boldsymbol{\beta}}_\lambda$ on one of the remaining two subsets. Select the $\widehat{\boldsymbol{\beta}}_\lambda$ that has the smallest prediction error (one for the Lasso and one for ridge regression). Finally, compute the test error on the final subset of data by comparing both the squared error and the count of how labels were incorrectly predicted, i.e., $\text{sign}(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_\lambda)\neq y_i$. To get a better sense of the performances, you can repeat this entire process by taking different subsets of 8, 1, and 1 for training, tuning, and testing, and compute the average squared errors and average number of misclassifications.