5. Coordinate Descent Lasso Regression. Here you will derive and implement the lasso "shooting method" discussed in class. The lasso objective function is:

$$F_{lasso}(\mathbf{w}) = \sum_{j=1}^{n}(y_j - \hat{y}_j(\mathbf{w}))^2 + \lambda \sum_{i=1}^{p} |w_i|$$

where $\lambda > 0$ and $\hat{y}_j(\mathbf{w}) = w_0 + \sum_{i=1}^{p} w_i x_{ij} = w_0 + \mathbf{x}^T \mathbf{w}_1$. As usual x_{ij} is the ij element of the $p \times n$ feature matrix $\mathbb{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$.

(a). As $|w_i|$ is not differentiable, the optimality condition for w_i to minimize the ith coordinate of $F_{lasso}(\mathbf{w})$ is that $\mathbf{0} \in \partial_{w_i} F_{lasso}(\mathbf{w})$, where $\partial_{w_i} F_{lasso}(\mathbf{w})$ is the subdifferential with respect to the i-th coordinate w_i, $i = 1, \ldots, d$. We stated in class that the subdifferential is

$$\partial_{w_i} F_{lasso}(\mathbf{w}) = a_i w_i - c_i + \lambda \partial_{w_i} |w_i|, \qquad i \geq 1.$$

We also stated that the solution w_i to the subdifferential equation $\mathbf{0} \in \partial_{w_i} F_{lasso}(\mathbf{w})$ is the soft thresholded w_i for $i \geq 1$

$$w_i = \text{soft}\left(\frac{c_i}{a_i}; \frac{\lambda}{a_i}\right) = \begin{cases} \frac{c_i - \lambda}{a_i}, & c_i > \lambda \\ 0, & c_i \in [-\lambda, \lambda] \\ \frac{c_i + \lambda}{a_i}, & c_i < -\lambda \end{cases}$$

where $\text{soft}\,(a; \delta) = \text{sign}(a) \max\{0, |a| - \delta\}$. By checking all three cases of c_i in the above expression, verify that if w_i is so specified it satisfies the subdifferential optimality condition

$$\mathbf{0} \in \partial_{w_i} F_{lasso}(\mathbf{w}).$$

Hint: use the definition of the subdifferential of $|w|$ (derived in class)

$$\partial_w |w| = \begin{cases} 1, & w > 0 \\ [-1, 1], & w = 0 \\ -1, & w < 0 \end{cases}$$

(b). Using the results of part (a) and the fact (recall part (a) of Problem 4) that for sphered data the optimum offset weight w_0 that minimizes F_{lasso} is $w_0^* = \frac{1}{n} \sum_{j=1}^{n} y_j$, implement the CD Lasso regression in python. Run your CD Lasso regression code on the **sphered** data you created in Problem 1 for 2900 iterations with $\lambda = 100$ and with $\mathbf{w}$ initialized to be a vector of all 1's.

As in Problem 4, plot the weight trajectories over iteration, plot the learning curve (MSE on the training data). Also report final test MSE.

Comment on the differences between the results of the ridge regression solution and the CD lasso solution. Are there any values of the lasso weight vector that have been thresholded to zero after 2900 iterations?