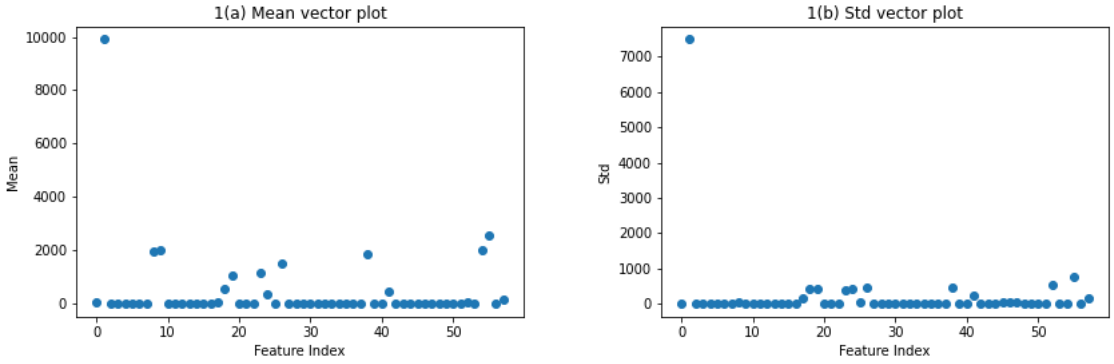
EECS 545 Homework 3 Solution (F21)

1. Ridge Regression

(a) Grading rubrics

- 1.5 pts for each plot that is approximately correct, e.g., look for similar patterns instead of exact numbers.
- 0 pt if no effort or completely wrong.

Mean, standard deviation plots:



(b) Grading rubrics

- 1.5 pts for correct offset (w_0) within ± 10 and 1.5 pts for correct weights ($w_1, \ldots, w_5$) within ± 0.5.
- 0 pt if no effort or completely wrong.

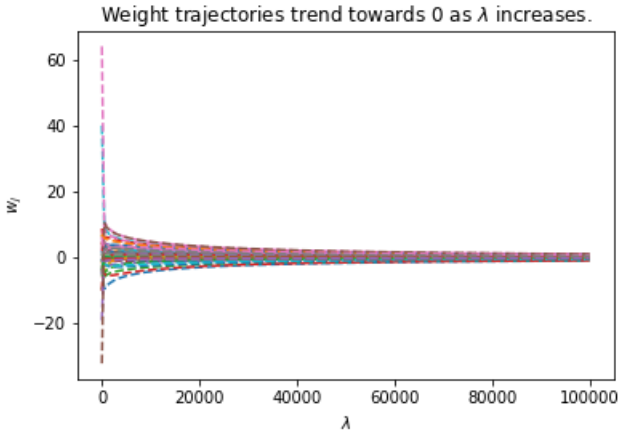
$w_0 = 181.7$
$w_1 = 3.04$, $w_2 = 4.38$, $w_3 = 2.28$, $w_4 = 0.13$, $w_5 = -0.45$

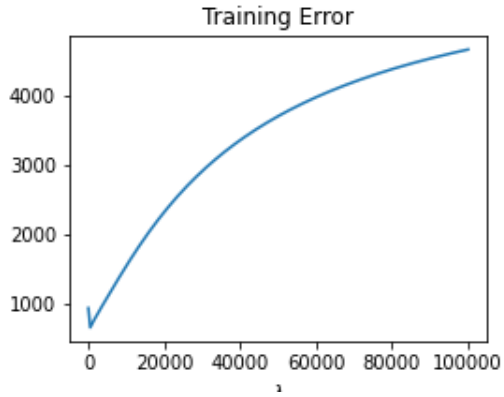
(c) Grading rubrics

- 1.5 pts for correct pattern in the weights trajectories, i.e., larger at the beginning and converge towards 0.
- 1.5 pts for correct pattern in the training MSE plot, i.e., increases as λ increase.
- 0 pt if no effort or completely wrong.

Weight trajectories, train MSE Weight trajectories should get close to zero by $\lambda = 10^5$. The plot may have a skewed scale if the offset w_0 trajectory is included; if included, w_0 should be a horizontal line.

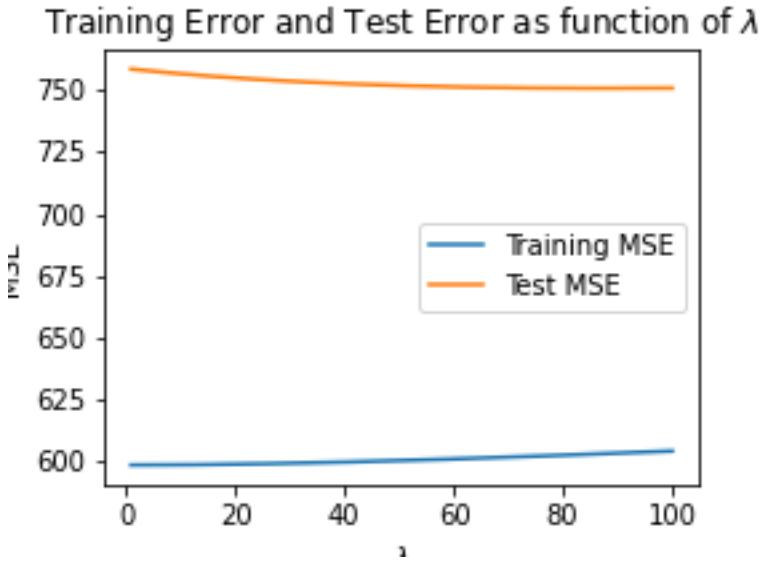
Training MSE should start around 900-1000 at $\lambda = 0$, reach a local minimum around $\lambda = 500, 1000$, then increase to over 4000 by $\lambda = 10^5$.





(d) Grading rubrics

- 1.5 pts for correct pattern described in the solution.
- 1.5 pts for correct optimal λ within ± 100.
- 0 pt if no effort or completely wrong.



Optimal training $\lambda = 1.0$. Optimal testing $\lambda = 88.56$

The plot should show training MSE monotonically increasing from $\lambda = 1$, while test MSE initially decreases to a local minima (at $\lambda = 88.56$) before increasing.

2. Optimal soft-margin hyperplane

(a) Grading rubrics

- 1.5 pts for correctly using the conditions, e.g., $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) \leq 0$ and $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) \geq 1 - \xi_i^*$
- 1.5 pts for using $\xi^* \geq 1$ to show its greater than or equal to the misclassification error.
- 0 pt if no effort or completely wrong.

If $\mathbf{x}_i$ is misclassified, then

$$y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) \leq 0$$

Combined with the constraint:

$$y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) \geq 1 - \xi_i^*$$

We have: $\xi_i^* \geq 1$, then we have:

$$\frac{1}{n}\sum_i \boldsymbol{\xi}^* \geq \frac{\text{number of } \mathbf{x}_i \text{ misclassified}}{n} = \text{training error}$$

(b) Grading rubrics

- 3 pts for recognizing the correct proportional constant, i.e., $\|\boldsymbol{w}^*\|$ or $1/\|\boldsymbol{w}^*\|$. Partial pts can be given for, e.g.,
  - 1 pt for recognizing $\xi_i = 1 - y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*)$
  - 1 pt for correct distance between the data point and the margin
  - 1 pt for recognizing the relationship $1 - y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) = |y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) - 1|$
- 0 pt if no effort or completely wrong.

When $\xi_i > 0$, we have $\xi_i = 1 - y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*)$

Solution 1

As given in the problem statement the hyperplane represent margin is given by $\{\boldsymbol{x} : (\boldsymbol{w}^*)^T\boldsymbol{x} + b^* = y_i\}$, multiply both side by y_i and we can rewrite the margin in the standard form: $\{\boldsymbol{x} : (y_i\boldsymbol{w}^*)^T\boldsymbol{x} + (y_i b^*) - 1 = 0\}$.

Recall from the lecture the distance between $\mathbf{x}_i$ and the margin then is given by:

$$\frac{|(y_i\boldsymbol{w}^*)^T\boldsymbol{x}_i + (y_i b^*) - 1|}{||\mathbf{y_i}\mathbf{w}^*||} = \frac{|y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) - 1|}{||\mathbf{w}^*||} = \frac{1 - y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*)}{||\mathbf{w}^*||} = \frac{\xi_i}{||\mathbf{w}^*||}$$

Solution 2

Case 1: when $0 \leq y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) < 1$, x_i lies within the margin

Then $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) = |y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*)| = |\mathbf{w}^{*T}\mathbf{x}_i + b^*|$

Thus $\xi_i = 1 - |\mathbf{w}^{*T}\mathbf{x}_i + b^*| = ||\mathbf{w}^*||(\frac{1-|\mathbf{w}^{*T}\mathbf{x}_i+b^*|}{||\mathbf{w}^*||})$, where $\frac{1-|\mathbf{w}^{*T}\mathbf{x}_i+b^*|}{||\mathbf{w}^*||})$ is the distance between x_i and the margin.

Case 2: when $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) < 0$, x_i lies outside the margin, and it's misclassified.

Then $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) = -|y_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*)| = -|\mathbf{w}^{*T}\mathbf{x}_i + b^*|$

Thus $\xi_i = 1 + |\mathbf{w}^{*T}\mathbf{x}_i + b^*| = ||\mathbf{w}^*||(\frac{1+|\mathbf{w}^{*T}\mathbf{x}_i+b^*|}{||\mathbf{w}^*||})$, where $\frac{1+|\mathbf{w}^{*T}\mathbf{x}_i+b^*|}{||\mathbf{w}^*||})$ is the distance between x_i and the margin.

In summary when $\xi_i > 0$ is propositional to the distance of x_i to the margin, and the propotion constant is $||\mathbf{w}^*||$.

3. Subgradient methods for the optimal soft margin hyperplane

(a) Grading rubrics

- 1 pt for correct gradient/subgradient in each scenario, e.g., $y_i(\mathbf{w}^T\mathbf{x}_i+b) < 1, > 1$, or $= 1$. Dock 0.5 pt for each minor error, e.g, missing/incorrect sign, missing constant term, etc.
- 0 pt if no effort or completely wrong.

The equation

$$J_i(\mathbf{w}, b) = \frac{1}{n}(L(y_i, \mathbf{w}^T\mathbf{x}_i + b) + \frac{\lambda}{2}\|\mathbf{w}\|^2)$$

satisfies

$$\sum_{i=1}^{n} J_i(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \mathbf{w}^T\mathbf{x}_i + b) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Since the non-differentiability of the hinge loss $L(y_i, \mathbf{w}^T\mathbf{x}_i+b) = \max(0, 1-y_i(\mathbf{w}^T\mathbf{x}_i+b))$ occurs when $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$, we will consider the following three regions individually:

$$\text{(I) } y_i(\mathbf{w}^T\mathbf{x}_i + b) < 1, \quad \text{(II) } y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1, \quad \text{(III) } y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1.$$

In region I, we have $J_i(\mathbf{w}, b) = \frac{1}{n}\left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) + \frac{\lambda}{2}\|\mathbf{w}\|^2\right)$. Since this is differentiable, the subgradient in this region is given by its gradient:

$$\mathbf{u}_i^{(\text{I})} = \nabla_{\boldsymbol{\theta}} J_i(\mathbf{w}, b) = \begin{bmatrix} \frac{\partial}{\partial b} J_i(\mathbf{w}, b) \\ \nabla_{\mathbf{w}} J_i(\mathbf{w}, b) \end{bmatrix} = \frac{1}{n}\begin{bmatrix} -y_i \\ -y_i\mathbf{x}_i + \lambda\mathbf{w} \end{bmatrix}. \tag{1}$$

In region II, since $L(y_i, \mathbf{w}^T\mathbf{x}_i + b) = 0$, we simply have $J_i(\mathbf{w}, b) = \frac{\lambda}{2n}\|\mathbf{w}\|^2$. Once again this is differentiable, so the subgradient in this region is also given by its gradient:

$$\mathbf{u}_i^{(\text{II})} = \nabla_{\boldsymbol{\theta}} J_i(\mathbf{w}, b) = \begin{bmatrix} \frac{\partial}{\partial b} J_i(\mathbf{w}, b) \\ \nabla_{\mathbf{w}} J_i(\mathbf{w}, b) \end{bmatrix} = \frac{1}{n}\begin{bmatrix} 0 \\ \lambda\mathbf{w} \end{bmatrix}. \tag{2}$$

Finally, region III involves a non-differentiable point $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$. Here we can take either (1), (2), or any convex combination of the two $\tau\mathbf{u}_i^{(\text{I})} + (1-\tau)\mathbf{u}_i^{(\text{II})}$, $\tau \in [0, 1]$, which can be written:

$$\mathbf{u}_i^{(\text{III})} = \frac{1}{n}\begin{bmatrix} -\tau y_i \\ -\tau y_i\mathbf{x}_i + \lambda\mathbf{w} \end{bmatrix} \text{ for any } \tau \in [0, 1]. \tag{3}$$

We select $\tau = 0$ for our code in part **f.**, which gives us $\mathbf{u}_i^{(\text{III})} = \frac{1}{n}[0, \lambda\mathbf{w}^T]^T$.
To summarize, a subgradient $\mathbf{u}_i$ for J_i with respect to $\boldsymbol{\theta}$ can be written as:

$$\mathbf{u}_i = \begin{cases} \frac{1}{n}[-y_i,\ -y_i\mathbf{x}_i^T + \lambda\mathbf{w}^T]^T & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + b) < 1 \\ \frac{1}{n}[-\tau y_i,\ -\tau y_i\mathbf{x}_i^T + \lambda\mathbf{w}^T]^T, \tau \in [0, 1] & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1 \\ \frac{1}{n}[0,\ \lambda\mathbf{w}^T]^T & \text{if } y_i(\mathbf{w}^T\mathbf{x}_i + b) > 1. \end{cases}$$

To receive full credit, students may select any subgradient from above.

(b) Grading rubrics

- 1.5 pts for reporting correct hyperplane, e.g., w, b within ± 0.5, and margin within ± 0.01, and data plot with the separating hyperplane.
- 1.5 pts for correct pattern (no need to check exact values) in the objective vs. iteration plot and correct objective value within ± 0.1 in the end.
- 0 pt if no effort or completely wrong.

Figure 1 shows the results for the subgradient method. The estimated hyperplane parameters are $w = [-17.8163 - 9.1171]^T$ and $b = 12.0680$, margin $\rho = 0.04996$, and the final objective function value is 0.4498.

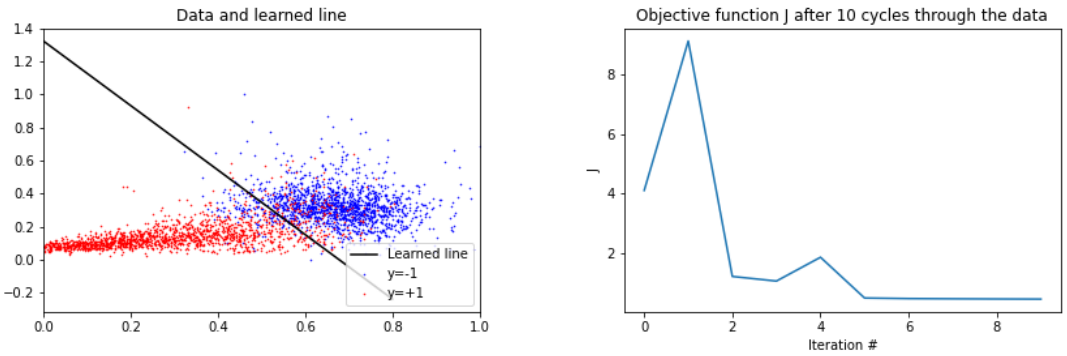


Figure 1: Results from the subgradient method.

(c) Grading rubrics

- 1.5 pts for reporting correct hyperplane, e.g., w, b within ± 0.5, and margin within ± 0.01, and data plot with the separating hyperplane.
- 1.5 pts for correct pattern (no need to check exact values) in the objective vs. iteration plot and correct objective value within ± 0.1 in the end.
- 0 pt if no effort or completely wrong.

Figure 2 shows the results for the stochastic subgradient method. The estimated hyperplane parameters are $w = [-5.8037 \ -4.3894]^T$ and $b = 4.0535$, margin $\rho = 0.1374$ and the final objective function value is 0.2583.

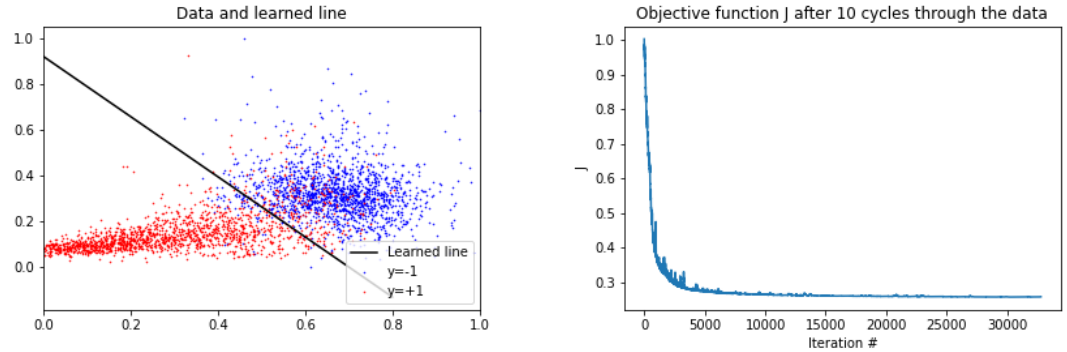


Figure 2: Results from the stochastic subgradient method.

The stochastic subgradient method converges faster than subgradient. It looks like subgradient takes until iteration 5 or 6 to converge, whereas stochastic subgradient converges after 2 or 3 cycles through the data (roughly 6000-9000 iterations).

4. Coordinate Descent Ridge Regression

(a) Grading rubrics

- 3 pts for correct approach, i.e., taking derivative w.r.t. the objective function and solve for $\frac{\partial}{\partial w_0} F_{ridge}(\mathbf{w}) = 0$.
- 0 pt if no effort or completely wrong.

The w_0 that minimizes F_{ridge} can be derived as follows:

$$\begin{aligned}
\frac{\partial}{\partial w_0} F_{ridge}(\mathbf{w}) &= \frac{\partial}{\partial w_0} \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}))^2 \\
&= \frac{\partial}{\partial w_0} \sum_{j=1}^{n} (y_j - w_0 - \mathbf{x}_j^T \mathbf{w}_1)^2 \\
&= -2 \sum_{j=1}^{n} (y_j - w_0 - \mathbf{x}_j^T \mathbf{w}_1) \\
&= -2 \left(\sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \mathbf{x}_j^T \mathbf{w}_1 - n w_0 \right) \\
&= -2 \left(\sum_{j=1}^{n} y_j - n w_0 \right)
\end{aligned} \tag{4}$$

$$\frac{\partial}{\partial w_0} F_{ridge}(\mathbf{w}) = 0 \implies w_0 = \frac{1}{n} \sum_{j=1}^{n} y_j \tag{5}$$

Moreover, $\frac{\partial}{\partial w_0^2} F_{ridge}(\mathbf{w}) > 0$ which implies that $F_{ridge}(\mathbf{w})$ is strictly convex and hence the unique critical point of w_0 is the mean of $\mathbf{y}$

(b) Grading rubrics

- 3 pts for correct approach, i.e., taking derivative of the objective function w.r.t. w_i and re-arrange the terms. Dock 0.5 pt for each minor error, e.g., missing/incorrect sign, missing/incorrect constant terms, wrong summation, etc.
- 0 pt if no effort or completely wrong.

For $i > 0$ the partial derivative of F_{ridge} with respect to w_i can be derived as follows:

$$\begin{aligned}
\frac{\partial}{\partial w_i} \left(\sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}))^2 + \lambda \sum_{k=1}^{p} |w_k|^2 \right) &= -2 \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w})) \frac{\partial}{\partial w_i} (\hat{y}_j(\mathbf{w})) + 2\lambda w_i \\
&= -2 \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}_{-i}) - x_{ij} w_i) x_{ij} + 2\lambda w_i \\
&= -2 \sum_{j=1}^{n} (y_j - \hat{y}_j(\mathbf{w}_{-i})) x_{ij} + 2 w_i \sum_{j=1}^{n} x_{ij}^2 + 2\lambda w_i \\
&= -c_i + a_i w_i + 2\lambda w_i
\end{aligned} \tag{6}$$

where

$$c_i = 2\sum_{j=1}^{n}(y_j - \hat{y}_j(\mathbf{w}_{-i}))x_{ij}, \quad a_i = 2\sum_{j=1}^{n} x_{ij}^2, \quad i \geq 0 \tag{7}$$

with $\hat{y}_j(\mathbf{w}_{-i})$ the predictor with coefficient w_i set to zero.

(c) Grading rubrics

- 3 pts for correct approach, i.e., use the results from the previous part plus the component-wise derivative of the regularization term, i.e., $2\lambda w_i$.
- 0 pt if no effort or completely wrong.

The optimal condition for w_i that minimize F_{ridge} can be found by solving the following equation:

$$\begin{aligned} \frac{\partial}{\partial w_i} F_{ridge}(\mathbf{w}) &= -c_i + a_i w_i + 2\lambda w_i = 0 \\ w_i &= \frac{c_i}{a_i + 2\lambda} \end{aligned} \tag{8}$$

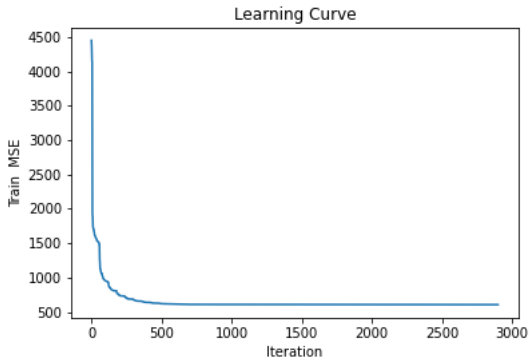
Moreover, $\frac{\partial}{\partial w_i^2} F_{ridge}(\mathbf{w}) > 0$ which makes this point the unique minimizer.

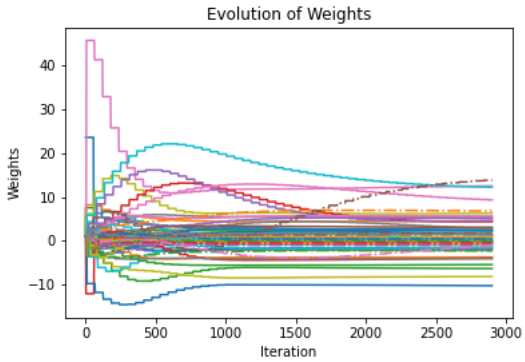
(d) Grading rubrics

- 1.5 pts each for approximately correct pattern (no need to check exact values) in the learning curve plot or the weights plot.
- 0 pt if no effort or completely wrong.

Train MSE/learning curve should converge quickly. The final value of the objective is 604.7197, but this need not be turned in.

Some weights tend to small values, but they don't all converge to 0 as they did in the first question (much smaller range of λ).





5. Coordinate Descent Lasso Regression

(a) Grading rubrics

- 1 pt for each of the three cases. Give full credits if correctly plug in the soft-thresholded w_i to the gradient function and check that it either equals 0 or contains 0 for the case $c_i \in [-\lambda, \lambda]$. Deduct 0.5 pt for each minor error, e.g., plugged in wrong w_i or wrong $\partial_w |w|$, etc.
- 0 pt if no effort or completely wrong.

For $i \geq 1$ the the partial derivative of F_{lasso} with respect to w_i is as follows:

$$\partial_{w_i} F_{lasso}(\mathbf{w}) = a_i w_i - c_i + \lambda \partial_{w_i} |w_i| \tag{9}$$

where

$$c_i = 2\sum_{j=1}^{n}(y_j - \hat{y}_j(\mathbf{w}_{-i}))x_{ij}, \quad a_i = 2\sum_{j=1}^{n} x_{ij}^2, \quad i \geq 0 \tag{10}$$

with $\hat{y}_j(\mathbf{w}_{-i})$ as the predictor with coefficient w_i set to zero.
And we have the definition of the subdifferential for w_i:

$$\partial_{w_i} |w_i| = \begin{cases} 1, & w > 0 \\ [-1, 1], & w = 0 \\ -1, & w < 0 \end{cases} \tag{11}$$

Case 1: $c_i > \lambda$
Here we have $w_i = \frac{c_i - \lambda}{a_i}$. Since $c_i > \lambda$, we have $\partial_{w_i} |w_i| = 1$ (note that a_i is always positive). Then plugging into 9:

$$\partial_{w_i} F_{lasso}(\mathbf{w}) = a_i \frac{c_i - \lambda}{a_i} - c_i + \lambda = 0 \tag{12}$$

Case 2: $c_i \in [-\lambda, \lambda]$
Then $w_i = 0$ and $\partial_{w_i} |w_i| = [-1, 1]$. Then equation 9:

$$\partial_{w_i} F_{lasso}(\mathbf{w}) = -c_i + \lambda \partial_{w_i} |w_i| \tag{13}$$

Consider a value $z \in \partial_{w_i} |w_i|$, then we can always choose $c_i = \lambda z \in [-\lambda, \lambda]$ to satisfy the optimality condition. Thus $0 \in \partial_{w_i} F_{lasso}$ for $c_i \in [-\lambda, \lambda]$.
Case 3: $c_i < \lambda$ So $w_i = \frac{c_i + \lambda}{a_i}$ and $\partial_{w_i} |w_i| = -1$ (since $w_i < 0$ given the condition on c_i). Plugging into 9:

$$\partial_{w_i} F_{lasso}(\mathbf{w}) = a_i \frac{c_i + \lambda}{a_i} - c_i - \lambda = 0 \tag{14}$$

(b) Grading rubrics

- 2 pts for correct patterns (no need to check exact values) in the two plots.
- 0.5 pt for correct final MSE within ± 10.
- 0.5 pt for correctly identifying the three zero weights.
- 0 pt if no effort or completely wrong.

Final test MSE = 755.0815

Four weights go to zero: 'Paved.Drive', 'Enclosed.Porch', 'Pool.QC'

Ridge and LASSO ave very similar final test MSE. But LASSO also zerod out four features. In other words, for a very marginally worse test MSE on LASSO, we can have a relatively more sparse feature set.