CS/ECE/ME532 Period 18 Activity

Estimated time: 15 mins for P1, 20 mins for P2, 15 mins for P3, 20 mins for P4

1. A breast cancer gene database has approximately 8000 genes from 100 subjects. The label y_i is the disease state of the ith subject (+1 if no cancer, -1 if breast cancer). Suppose we build a linear classifier that combines the 8000 genes, say $\boldsymbol{g}_i, i = 1, 2, \ldots, 100$ to predict whether a subject has cancer $\hat{y}_i = \text{sign}\{\boldsymbol{g}_i^T \boldsymbol{w}\}$. Note that here $\boldsymbol{g}_i$ and $\boldsymbol{w}$ are 8000-by-1 vectors. You recall from the previous period that the least-squares problem for finding classifier weights has no unique solution.

Your hypothesis is that a relatively small number of the 8000 genes are predictive of the cancer state. Identify a regularization strategy consistent with this hypothesis and justify your choice.

SOLUTION: The classification problem can be written as

$$\min_{\boldsymbol{w}} \left|\left| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{100} \end{bmatrix} - \begin{bmatrix} \boldsymbol{g}_1^T \\ \boldsymbol{g}_2^T \\ \vdots \\ \boldsymbol{g}_{100}^T \end{bmatrix} \boldsymbol{w} \right|\right|_2^2$$

Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{g}_1^T \\ \boldsymbol{g}_2^T \\ \vdots \\ \boldsymbol{g}_{100}^T \end{bmatrix}$ is a 100 by 8000 matrix.

Since there are 100 equations and 8000 unknowns, there is no unique solution. $\boldsymbol{A}$ is at most rank 100. We need regularization. Ridge regression will produce a dense solution with many nonzero terms in $\boldsymbol{w}$. The LASSO or least squares with an ℓ_1 regularizer will produce a sparser solution and is more consistent with the hypothesis.

2. Consider the least-squares problem $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$ where $\boldsymbol{y} = 4$ and $\boldsymbol{X} = \begin{bmatrix} 2 & 1 \end{bmatrix}$.

a) Does this problem have a unique solution? Why or why not?

b) Sketch the contours of the cost function $f(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2$ in the $w_1 - w_2$ plane.

c) Now consider the LASSO $\min_{\boldsymbol{w}} ||\boldsymbol{w}||_1$ subject to $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 < 1$. Find the solution using the following steps

i. Repeat your sketch from part **b)**.

ii. Add a sketch of $||\boldsymbol{w}||_1 = c$

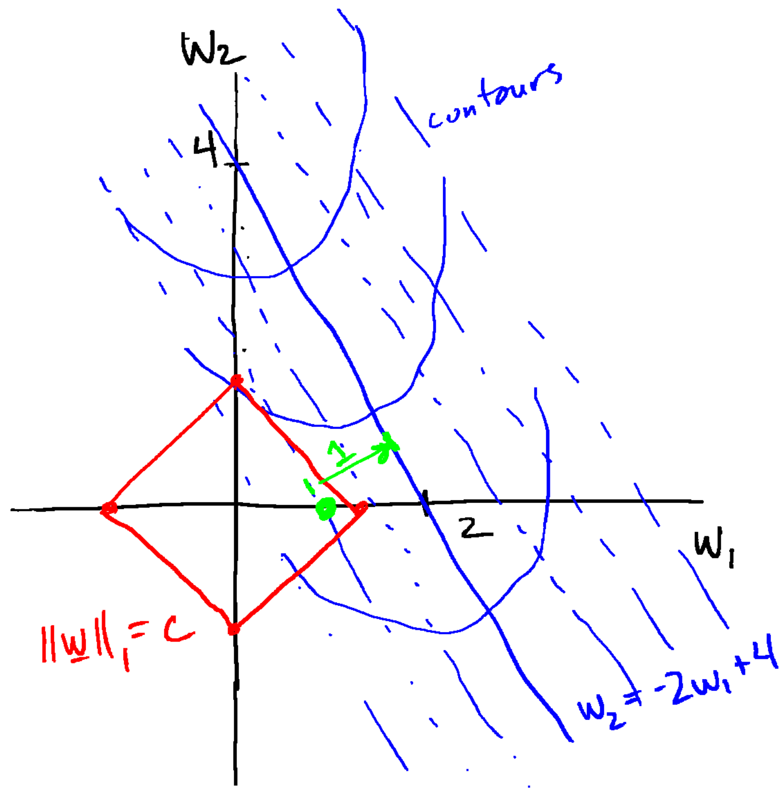
iii. Find the $\boldsymbol{w}$ that satisfies $||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 = 1$ with the minimum possible value of $||\boldsymbol{w}||_1$.

d) Use your insight from the previous part to sketch the set of solutions to the problem $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1$ for $0 < \lambda < \infty$.

SOLUTION:

a) No unique solution. There is one equation and two unknowns.

b) The squared error is zero along the line $2w_1 + w_2 = 4$ or $w_2 = -2w_1 + 4$. This has vertical intercept $w_2 = 4$ and slope -2. The contours of $f(\boldsymbol{w})$ are lines parallel to the line of zero squared error, with the height (value) of $f(\boldsymbol{w})$ given by the squared distance from the zero squared error line. Consequently, $f(\boldsymbol{w})$ describes a U-shaped surface, that is, a valley with low point along the line $w_2 = -2w_1 + 4$.



Solution lies on w_1 axis, since the corner of $\|w\|_1 = c$ is closest to $w_2 = -2w_1 + 4$ on the positive w_1 axis. The solution is one unit of distance from $w_2 = -2w_1 + 4$

c) sketch.png

d) The solution will lie on the w_1 axis, since that corner of $||\boldsymbol{w}_1||_1 = c$ is closest to the zero squared error line. Note that when $\lambda \approx 0$ the solution is $\boldsymbol{w} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, while for λ very large the solution approaches $\boldsymbol{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ along the w_1 axis.

3. The script provided has a function that will compute a specified number of iterations of the proximal gradient descent algorithm for solving the ℓ_1-regularized least-squares problem

$$\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda ||\boldsymbol{w}||_1$$

The script will get you started displaying the path taken by the weights in the proximal gradient descent iteration superimposed on a contour plot of the squared error surface for the cost function defined in problem **2.** part **b)** starting from $\boldsymbol{w}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The script assumes $\lambda = 4$ and $\tau = 1/4$.

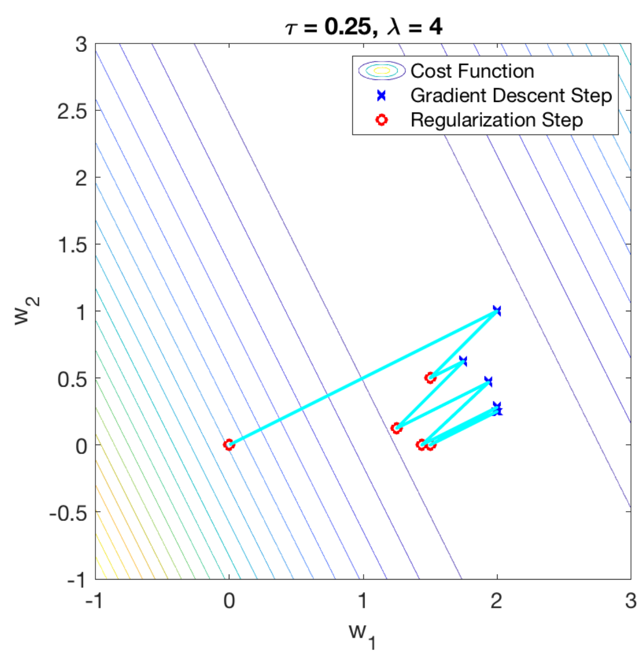
Include the plots you generate below with your submission.

a) How many iterations does it take for the algorithm to converge to the solution? What is the converged value for $\boldsymbol{w}$?

b) Change to $\lambda = 2$. How many iterations does it take for the algorithm to converge to the solution? What is the converged value for $\boldsymbol{w}$?

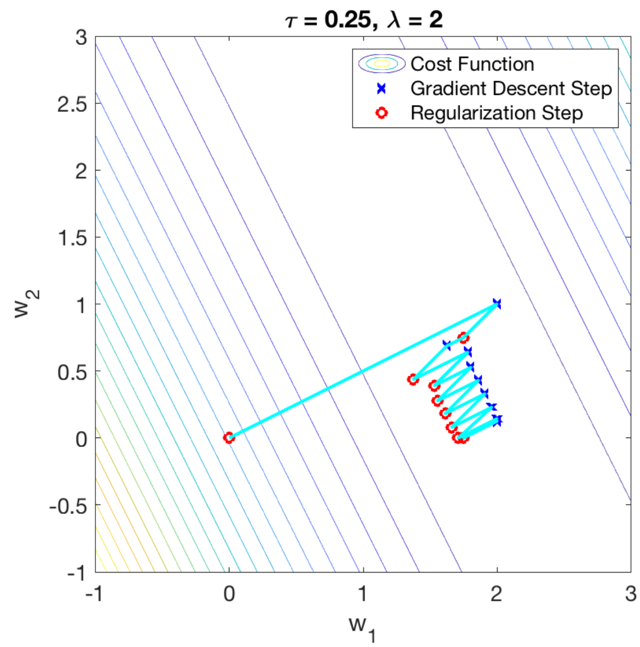
c) Explain what happens to the weights in the regularization step.

SOLUTION:



a)

Converges to $w_1 = 1.5, w_2 = 0$ in four iterations.



b)

Converges to $w_1 = 1.75, w_2 = 0$ in eight iterations.

c) The soft thresholding in the regularization step shrinks both w_1 and w_2 if they are larger than $\lambda\tau/2 = 1/2$ in the first case and $1/4$ in the second. If one of the coordinate values is less than $\lambda\tau/2$, then that coordinate is set to zero. This is why after a couple steps $w_2 = 0$.

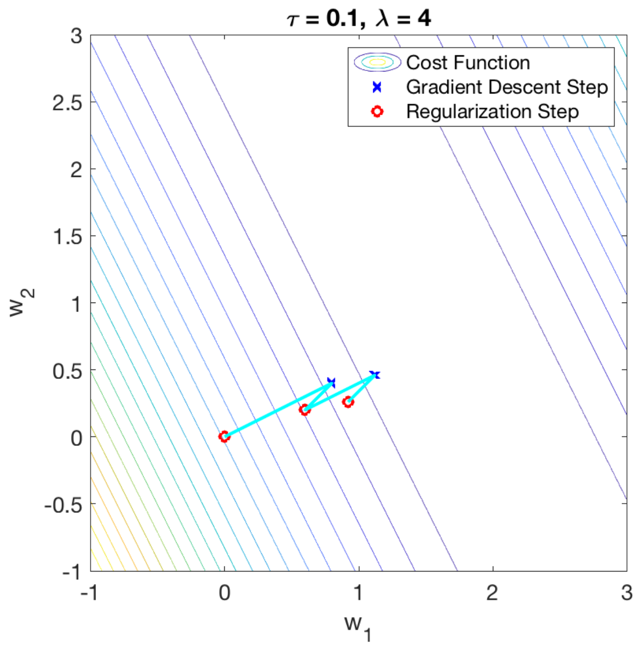
4. Use the proximal gradient algorithm to solve $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + 4||\boldsymbol{w}||_1$ for the parameters defined in problem **2.**

a) What is the maximum value for the step size in the negative gradient direction, τ?

b) Suppose $\tau = 0.1$ and you start at $\boldsymbol{w}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Calculate the first two complete iterations of the proximal gradient algorithm and depict $\boldsymbol{w}^{(0)}, \boldsymbol{z}^{(1)}, \boldsymbol{w}^{(1)}, \boldsymbol{z}^{(2)}$ and $\boldsymbol{w}^{(2)}$ on a sketch of the cost function identical to the one you created in problem **2.b)**.

SOLUTION:

a) The proximal gradient approach was derived assuming $\tau < 1/||\boldsymbol{X}||_{op}^2$. in this case $||\boldsymbol{X}||_{op} = \sqrt{5}$, so $\tau < 0.2$. Note that the step size used in Problem 3 is larger than the step size for which stability is guaranteed, yet the algorithm converges. The proximal gradient algorithm may converge for larger values, but such convergence is only guaranteed if $\tau < 1/||\boldsymbol{X}||_{op}^2$.



b)