CS/ECE/ME532 Activity 18

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.

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 $0 < \lambda < \infty$.

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.

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?

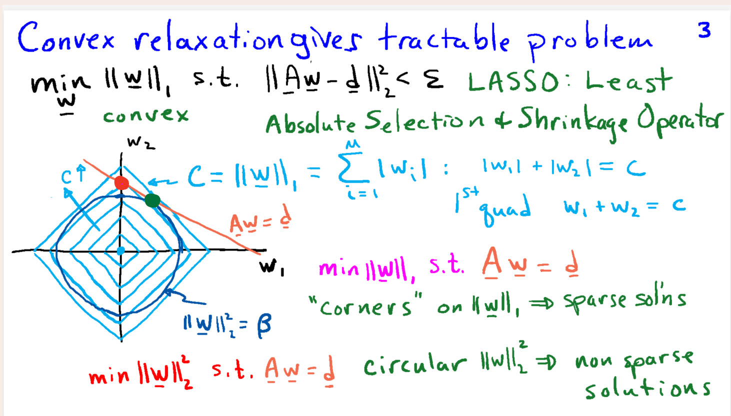
b) Suppose $\tau = 0.1$ and you start at $\boldsymbol{w}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Calculate the first two complete iterations of the proximal gradient algorithm and depict $\boldsymbol{w}^{(0)}, \boldsymbol{z}^{(1)}, \boldsymbol{w}^{(1)}, \boldsymbol{z}^{(2)}$ and $\boldsymbol{w}^{(2)}$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

532 Activity 18 DEVIN BRESSER

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.

A good regularization strategy to eliminate features is l1-regularization aka LASSO regularization. This is because of the geometry of the unit norm "ball" for the l1-norm, which is more of a 45-degree rotated square. The lowest l1-norms lie at the "corners" of the unit diamond (along axes), and thus, l1-norm minimization forces some dimensions (features) to zero.



- smallest l_1 norm is along corners, where some dimensions are 0

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 $||\boldsymbol{w}||_1$.

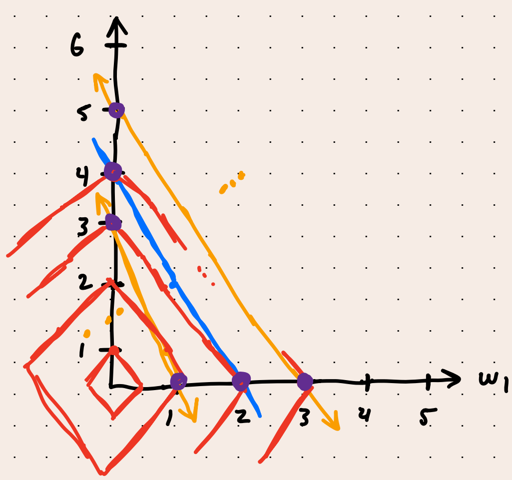
d) Use your insight from the previous part to sketch the set of solutions to the problem $\min_{\boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1$ for $0 < \lambda < \infty$.

a.) $\text{rank}(\underline{X}) = 1 < \dim(\underline{w}) = 2$. No unique solution.

$\underset{1\times 2}{\underline{X}}\ \underset{2\times 1}{\underline{w}} = \underset{1\times 1}{\underline{y}}$.

b.) $f(\underline{w}) = ||y - \underline{X}\underline{w}||_2^2 = ||4 - [2\ 1]\underline{w}||_2^2 = (4 - 2w_1 - w_2)^2$

$(4 - 2w_1 - w_2)^2 = c \rightarrow c = 0, 1, 2, 3 \ldots$

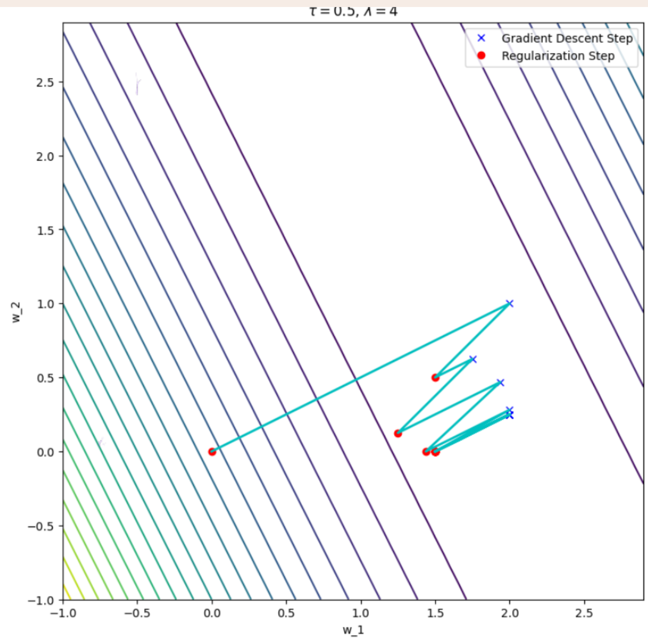


c.) Added ℓ_1 norm rectangles. (red)

As we can see the minimum ℓ_1-norm is achieved along the w_1 & w_2 axes

d.) These are the purple dots • they will be spaced apart further & further with a squared relationship.

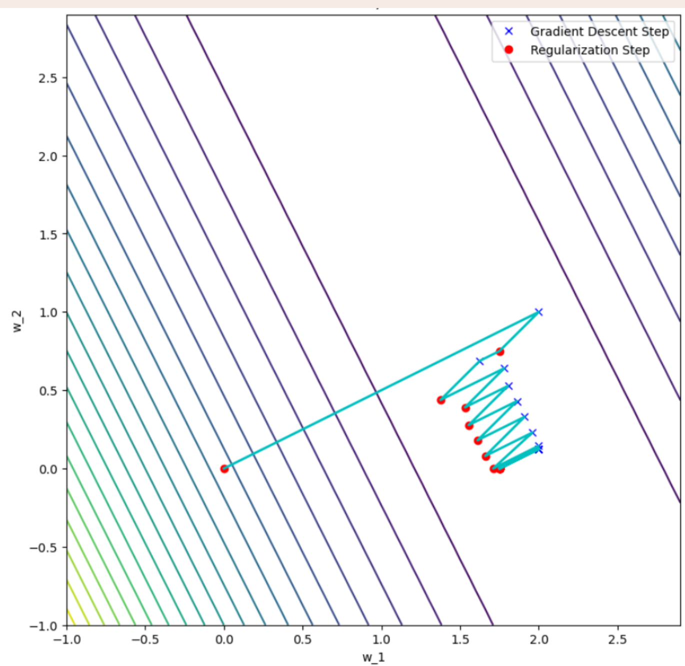
P3




```
[14]: W[:,-1]
```

```
[14]: array([1.5, 0. ])
```

```
[6]: # 3a comment: The algorithm takes 8 steps (counting both G-D steps and reg steps) to converge to a
     # point where it bounces back between a G-D and reg step to infinity.
     # w_conv = [1.5, 0]
```




```
16]: W[:,-1]
```

```
16]: array([1.75, 0.  ])
```

```
[ ]: # Problem 3b comment:
     # Now the algorithm takes 16 steps (total, defined the same way) before oscillating forever.
     # w_conv = [1.75 0]
```

```
17]: # Problem 3c comment:
     # In the regularization step, the weights bounce back in the opposite direction.
     # They're pulled away from the path directly to the minimum solution.
```

1.) Initialize $\quad \underline{w}_0 = \underline{0}, \quad 0 < \tau < 1/\|\underline{A}\|_{op}^2$

2.) LS-GD $\quad \underline{z}^{(k)} = \underline{w}^{(k)} - \tau \underline{A}^T(\underline{A}\,\underline{w}^{(k)} - \underline{d})$ (grad. descent from current estimate of $\underline{w}$, $\underline{w}^{(k)}$ defines $z^{(k)}$)

3.) Regularize $\quad \underline{w}^{(k+1)} = \underset{\underline{w}}{\text{argmin}} \|\underline{z}^{(k)} - \underline{w}\|_2^2 + \lambda\tau r(\underline{w})$ (Solve the proximal (close) regularized problem. Find a $\underline{w}$ (that has been regularized) that is close to $\underline{z}$.)

4.) Check if converged

$$\text{if } \underbrace{\|\underline{w}^{(k+1)} - \underline{w}^{(k)}\|_2^2}_{\text{If iteration does not result in significant change to } \underline{w}} < \varepsilon, \text{ STOP}$$

else, to 2.)

$$X = [2 \ 1], \quad y = 4$$

4.)

a.) Calculator gives $\delta_1 = 2.236$

$$\rightarrow \tau < 1/\|A\|_{op}^2 = 1/(2.236)^2 = 0.2$$

b.) $\tau = 0.1$, $w^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\underline{z}^{(0)} = \underline{w}^{(0)} - 2\tau \underline{X}^T(\underline{X}\,\underline{w}^{(0)} - y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}(-4) = \begin{bmatrix} 1.6 \\ 0.8 \end{bmatrix}$$

$$\underline{w}^{(1)} = \underset{\underline{w}}{\text{argmin}} \|z^{(0)} - \underline{w}\|_2^2 + 0.4\|\underline{w}\|_1$$

$$\rightarrow \begin{cases} w_1^{(1)} = \left(|z_1^{(0)}| - \frac{0.4}{2}\right)_+ \text{sign}(z_1^{(0)}) = (1.6 - 0.8)_+ \cdot 1 = 0.8 \\ w_2^{(1)} = \left(|z_2^{(0)}| - \frac{0.4}{2}\right)_+ \text{sign}(z_2^{(0)}) = (0.8 - 0.8)_+ \cdot 1 = 0 \end{cases} \rightarrow \underline{w}_1 = \begin{bmatrix} 0.8 \\ 0 \end{bmatrix}$$

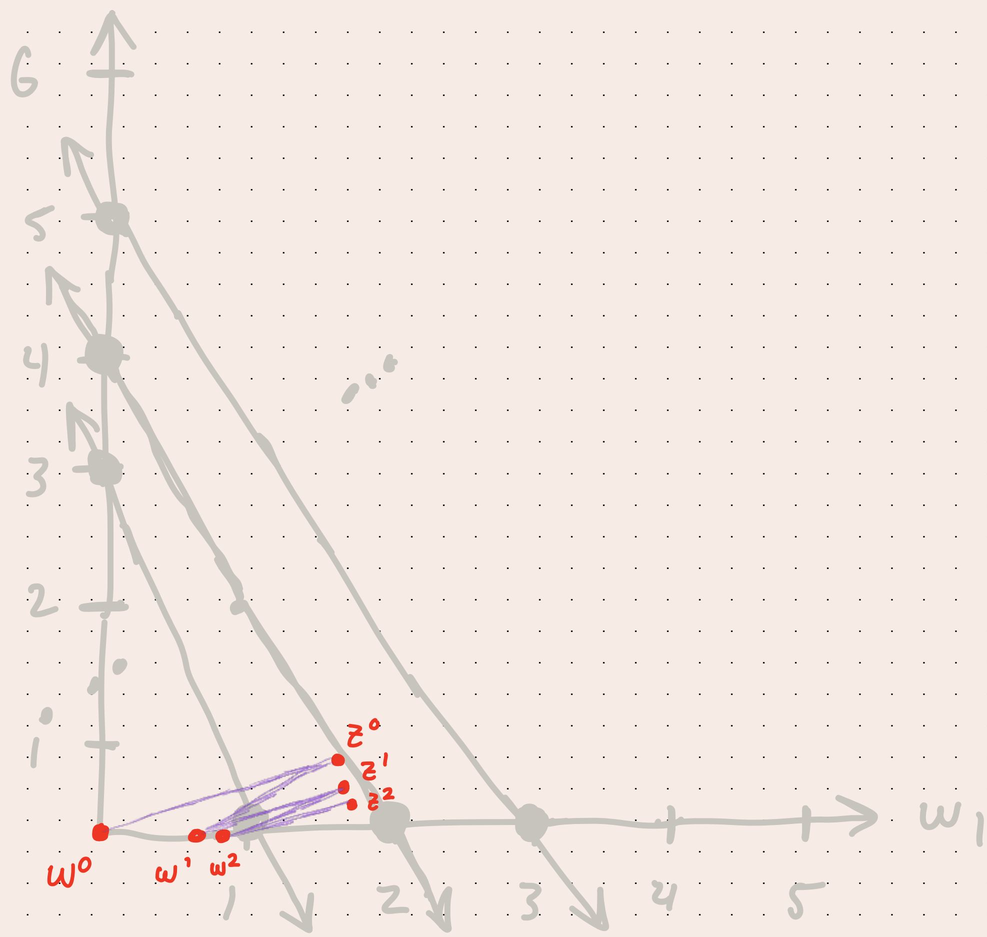
$$\rightarrow \underline{z}^{(1)} = \underline{w}^{(1)} - 2\tau \underline{X}^T(\underline{X}\,\underline{w}^{(1)} - y) = \begin{bmatrix} 0.8 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}\left([2 \ 1]\begin{bmatrix} -0.8 \\ 0 \end{bmatrix} - 4\right)$$

$$= \begin{bmatrix} 0.8 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}(-2.4) = \begin{bmatrix} 1.76 \\ 0.48 \end{bmatrix}$$

$$\rightarrow \begin{cases} w_1^{(2)} = \left(|z_1^{(1)}| - \frac{0.4}{2}\right)_+ \text{sign}(z_1^{(1)}) = (1.76 - 0.8)_+ \cdot 1 = 0.96 \\ w_2^{(2)} = \left(|z_2^{(1)}| - \frac{0.4}{2}\right)_+ \text{sign}(z_2^{(1)}) = (0.48 - 0.8)_+ \cdot 1 = 0 \end{cases} \rightarrow \underline{w}^{(2)} = \begin{bmatrix} 0.96 \\ 0 \end{bmatrix}$$

$$\rightarrow \underline{z}^{(2)} = \underline{w}^{(2)} - 2\tau \underline{X}^T(\underline{X}\,\underline{w}^{(2)} - y) = \begin{bmatrix} 0.96 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}\left([2 \ 1]\begin{bmatrix} 0.96 \\ 0 \end{bmatrix} - 4\right)$$

$$= \begin{bmatrix} 0.96 \\ 0 \end{bmatrix} - \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix}(-2.08) = \begin{bmatrix} 1.792 \\ 0.416 \end{bmatrix}$$



G
5
4
3
2
1
w1
1
2
3
4
5
z0
z1
z2
w0
w1
w2