DSC 140A - Discussion 05

Problem 1.

Recall that the regularized least squares risk is

$$\tilde{R}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{\phi}(\vec{x}^{(i)}) - y_i)^2 + \lambda\|\vec{w}\|^2$$

Show that

$$\tilde{R}(\vec{w}) = \frac{1}{n}\left(\vec{w}^T\Phi^T\Phi\vec{w} - 2\vec{w}^T\Phi^T\vec{y} + \vec{y}^T\vec{y}\right) + \lambda\vec{w}^T\vec{w},$$

where Φ is the matrix whose ith row is $\vec{\phi}(\vec{x}^{(i)}$, and where $\vec{y} = (y_1, \ldots, y_n)^T$.

Problem 2.

In class, we discussed how L1 regularization encourages sparse solutions and can be seen as a method for feature selection. In this problem, we will explore why L1 regularization promotes sparsity from the perspective of gradient descent.

a) First, write down the partial derivatives of the L1 and L2 regularization terms with respect to a specific weight w_j (you may ignore the case where $w_j = 0$, as the gradient might be undefined there).

- The L1 regularization term is given by:

$$R_1(\vec{w}) = \lambda\sum_{j=1}^{d}|w_j|$$

- The L2 regularization term is given by:

$$R_2(\vec{w}) = \lambda\sum_{j=1}^{d}w_j^2$$

b) Based on these derivatives, which regularizer is more effective at pushing w_j to zero?
Hint: Consider the behavior of the gradients when w_j is already small. For simplicity, assume that the partial derivative $\frac{\partial}{\partial w_j}$ of the Mean Squared Error (MSE) term is zero.