ECE C147/C247, Winter 2023 HW #2

UCLA ECE

1) Noisy Linear Regression

$D = \{(x^{(1)}, y^{(1)}), \ldots, y^{(N)}, y^{(N)}$

size of $D = N \times (d+1)$, +1 for labels

$x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$ where i is the ith house.

$\Rightarrow$ small weights

Recall: $$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)})^T\theta)^2 = \frac{1}{N}(Y - X\theta)^T(Y - X\theta)$$

Add noise:

$$\tilde{\mathcal{L}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - (x^{(i)} + \delta^{(i)})^T\theta)^2$$

$\delta^{(i)}$, i.i.d and $\sim N(0, \sigma^2 I) \in \mathbb{R}^d$

a) $$E[\tilde{\mathcal{L}}(\theta)] = E\left[\frac{1}{n}\sum_{i=1}^{n}(\underbrace{y^{(i)} - (x^{(i)})^T\theta}_{a \in \mathbb{R}} - \delta^{(i)T}\theta)^2\right]$$

$$E[\tilde{\mathcal{L}}(\theta)] = E\left[\frac{1}{N}\sum_{i=1}^{n}(a^2 - 2a\delta^{(i)T}\theta + \theta^T\delta^{(i)}\delta^{(i)T}\theta)\right.$$

$$E[\tilde{\mathcal{L}}(\theta)] = \frac{1}{N}\sum_{i=1}^{n}E\left[(y^{(i)} - (x^{(i)})^T\theta)^2 - E 2(y^{(i)} - (x^{(i)})^T\theta)\delta^{(i)T}\theta + \theta^T\delta^{(i)}\delta^{(i)T}\theta)\right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - (x^{(i)})^T \theta)^2 - 2(y^{(i)} - (x^{(i)})^T \theta) \underbrace{\underbrace{E[\delta^{(i)}]}_{\vec{0}}{}^T \theta}_{0} + \theta^T \underbrace{E[\delta^{(i)} \delta^{(i)T}]}_{\sigma^2 I_{d \times d}} \theta$$

$$= \ell(\theta) - 0 + \frac{1}{N} \sum_{i=1}^{N} \sigma^2 \theta^T \theta$$

$$= \ell(\theta) + \sigma^2 \theta^T \theta$$

$$\boxed{= \ell(\theta) + \sigma^2 \|\theta\|_2^2} \rightarrow \text{Ridge regression} \rightarrow \text{smaller weights}$$

b) Additional term will have the same effect of Ridge regression with $\lambda = \sigma^2 > 0$. Ridge regression is mainly used to shrink the coefficients and avoid overfitting.

c) $\sigma = 0$, is regular linear regression with $\tilde{\ell}(\theta) = \ell(\theta)$. Therefore, there will be no regularization.

d) $\sigma = \infty$, all coefficients should be zero to minimize the cost function