1)

A) Keeps bias fixed but reduces variance
B) Increases variance and reduces bias
C) Reduces variance but increases bias
D) Tends to keep variance the same and bias increases
E) Reduces bias but keeps variance fixed
F) n - increase
λ - Increase
d - reduce
c - reduce
α - no impact

2)

$$\frac{\partial}{\partial B_j}\left(\lambda \sum_{j=1}^{d} B_j^2\right) = 2\lambda B_j$$

$$\frac{\partial}{\partial B_j}\left(\lambda \sum_{j=1}^{d} |B_j|\right) = \begin{cases} \lambda & \text{if } B_j > 0 \\ -\lambda & \text{if } B_j < 0 \end{cases}$$

b) The gradient of the 1st term is used by the model to determine optimality of the mse surface. During gradient descent, the hyperparameter α is often gradually reduce until the model's gradient converges to zero.

When considering both terms the optimal point of the descent is where the two surfaces meet and are minimized. During gradient descent with a sufficiently large λ, for the non-important features, the B_j is shrunk to minimize the overall loss but since the L_1 surface is linear when B_j is shrunk more it is pushed abruptly to zero resulting in feature selection.

c) In L_2, sparsity is not encouraged because as non-important weights are shrunk they do shrink much more smoothly due to the smoothness of the L_2 concentric circles and so

It's strength of reducing / increasing depends on B_j value

As B_j gets smaller, the penalty term $2\lambda B_j$, the pull shrinks as well, making the coefficients pull smaller and smaller but not pushing B_j all the way to zero.

3) a)

$$\nabla_{w_0^*}(J(w)) = \nabla_{w_0^*}\left(\frac{1}{n}\| Y - w_0 - w_1 X\|_2^2\right)$$

$$= 2 \cdot \frac{1}{n} \nabla_{w_0^*}(Y - w_0 - w_1 X) \cdot (Y - w_0 - w_1 X)$$

$$= \frac{2}{n} \cdot -1 \cdot (Y - w_0 - w_1 X)$$

$$= -\frac{2}{n} Y + \frac{2}{n} w_0 + \frac{2}{n} w_1 X$$

$$= \frac{2}{n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)(-1) \quad //$$

$$\nabla_{w_1^*}(J(w)) = \nabla_{w_1^*}\left(\frac{1}{n}\|Y - w_0 - w_1 X\|_2^2\right)$$

$$= 2 \cdot \frac{1}{n} \nabla_{w_1^*}(Y - w_0 - w_1 X) \cdot (Y - w_0 - w_1 X)$$

$$= \frac{2}{n} - X (Y - w_0 - w_1 X)$$

$$= -\frac{2}{n} X^T Y + \frac{2}{n} X w_0 + \frac{2}{n} X^T X w_1$$

$$= \frac{2}{n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)(-x_i) \quad //$$

b) From part (a)

$$\frac{\partial J}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)(-1) \Rightarrow 0 \cdot \frac{1}{2} \cdot (-1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)$$

$$\bar{x} \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)\right) = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)\bar{x} = 0$$

$$\frac{\partial J}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)(-x_i) \Rightarrow 0 \cdot \frac{1}{2} \cdot (-1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i)x_i$$

$$\frac{1}{n}\left(\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)x_i\right)$$

$$\frac{1}{n}\left(\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)x_i\right) - \frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)\bar{x} = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w_0^* - w_1^* x_i)(x_i - \bar{x}) = 0$$

c) No because if the matrix $X^T X$ in the closed form equation $\beta(2) = X^T Y (X^T X)^{-1}$ is not invertible there are infinitely many solutions or no solution d)

$$\nabla_w (J(w)) = \nabla_w \left(\frac{1}{n} (y - Xw)^T (y - Xw) \right)$$

$$= \frac{1}{n} \nabla_w \left(y^T y - 2(Xw)^T y + X^T w Xw \right)$$

$$= \frac{1}{n} \left(-2X^T y + 2X^T X w \right) = \frac{2}{n} \left(-X^T y + X^T X w \right) = 0$$

$$\cancel{\frac{2}{n}} (X^T y) = \cancel{\frac{2}{n}} (X^T X w) \Rightarrow X^T y = X^T X w$$

$$w^* = X^T y (X^T X)^{-1} \quad \text{if } X^T X \text{ is invertible}$$

4)

a) $L(\beta : Z) = \sum_{i=1}^{n} (y_i - w_1 x_1 + w_2 x_2) + \lambda (W_1^2 + w_2^2)$

$= \sum_{i=1}^{n} (y_i - 0 \cdot 1 + 0 \cdot (-1)) + \lambda (0^2 + 0^2)$

$= \sum_{i=1}^{n} (y_i - 0) + \lambda \cdot 0$

$= \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{1}{2}(0 + 1) = \frac{1}{2} \cdot 1 = 0.5$

b)

$\nabla_{w_1}(L) = \frac{1}{N} \nabla_{w_1} \left(\sum_{i=1}^{N} (y_i - w x_i)^2 + \lambda \|w\|^2 \right.$

$= \frac{1}{N} \sum_{i=1}^{N} 2 \cdot \nabla_{w_1} (y_i - w_1 x_{i,1} + w_2 x_{i,2}) + \lambda \nabla_{w_1} (w_1^2 + w_2^2)$

$= \frac{1}{N} \sum_{i=1}^{N} 2 \cdot -x_{1,1} (y_i - (w_1 x_{1,1} + w_2 x_{1,2})) + \lambda \nabla_{w_1} (w_1^2 + w_2^2)$

$= \frac{1}{N} \sum_{i=1}^{N} -2x_{i,1} (y_i - w x_i) + 2\lambda w_1$

$\nabla_{w_2}(L) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{w_2} (y_i - (w_1 x_{1,1} + w_2 x_{1,2}))^2 + 2\lambda w_2$

$\nabla_{w_2}(L) = \frac{1}{N} \sum_{i=1}^{N} -2x_{1,2} (y_i - w x_i) + 2\lambda w_2$

$$\nabla_{w_1}(L) = \frac{1}{N}\sum_{i=1}^{N} -2x_{i,1}(y_i - wx_i) + 2\lambda w_1$$

$$\nabla_{w_2}(L) = \frac{1}{N}\sum_{i=1}^{N} -2x_{i,2}(y_i - wx_i) + 2\lambda w_2$$

Step 1 : Plugged values

$$\nabla_{w_1}(L) = \frac{1}{2}\left(-2(1)\ (0-0)\right)$$

$$+ \frac{1}{2}\left(-2(-1)\ (1-0)\right)$$

$$1 \cdot (2) \cdot \frac{1}{2} = 1$$

$$\nabla_{w_1}(L) = 0 + 1 = 1 + 0 = 1$$

$$\nabla_{w_2}(L) = \frac{1}{2}(-2)(-1)\ (0-0) + 2(1)(0)$$

$$+ \frac{1}{2}(-2)(-1)(1-0) + 2(1)(0)$$

$$\nabla_{w_2}(L) = 0 + 1 = 1 + 0 = 1$$

$$W_{new} = W_{old} - \alpha \nabla_w(L)$$

$$\alpha = 1$$

$$W_{new} = [0, 0] - 1\ [1, 1]$$

$$= [0, 0] - [1, 1] = [-1, -1]$$

$w \leftarrow [-1, -1] \quad x_1 \leftarrow (1, -1) \quad y_1 \leftarrow 0 \quad x_2 \leftarrow (-1, -1) \quad y_2 \leftarrow 1$

$$\nabla_{w_1}(L) = \frac{1}{N} \sum_{i=1}^{N} -2x_{i,1}(y_i - w x_i) + 2\lambda w_1$$

$$\nabla_{w_2}(L) = \frac{1}{N} \sum_{i=1}^{N} -2x_{i,2}(y_i - w x_i) + 2\lambda w_2$$

Step 2

$$\nabla_{w_1}(L) = \frac{1}{2}(-2)(1)[0 - (-1 \cdot 1 + -1 \cdot -1)]$$
$$+ \frac{1}{2}(-2)(-1)[1 - (-1 \cdot -1 + -1 \cdot -1)]$$

$$\nabla_{w_1}(L) = 0 + -1 + -2 = -3$$

$$\nabla_{w_1}(L) = -3$$

$$\nabla_{w_2}(L) = \frac{1}{2}(-2)(-1)[0 - (-1 \cdot 1 + -1 \cdot -1)] + 2(1)(-1)$$
$$+ \frac{1}{2}(-2)(-1)[1 - (-1 \cdot -1 + -1 \cdot -1)] + 2(1)(-1)$$

$$\nabla_{w_2}(L) = (0 + -1) + -2 = -3$$

$$\nabla_{w_2}(L) = -3$$

$$w_{new} = [-1, -1] - 1[-3, -3] = [2, 2]$$

$$L(\beta : Z) = \frac{1}{N}\sum_{i=1}^{n} (y_i - w_1 x_1 + w_2 x_2)^2 + \lambda (w_1^2 + w_2^2)$$

$$= \frac{1}{2}(0 - (2 \cdot 1 + 2 \cdot (-1)))^2 + \frac{1}{2}(1 - (2 \cdot -1 + 2 \cdot (-1))^2$$

$$(-2 + -2)$$

$$= \frac{1}{2}(0 - 0)^2 + \frac{1}{2}(1+4)^2$$

$$= \frac{1}{2}(0)^2 + \frac{1}{2}(5)^2 = 0 + 12.5 = 12.5$$

Reg term $= 1(2^2 + 2^2) = 8$

$$12.5 + 8 = \underline{\underline{20.5}}$$

(c)

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - wx_i)^2$$

$$\nabla L(w) = \frac{1}{N} \nabla_w (\| y - w^* X \|_2^2) + \lambda \| w^* \|_2^2$$

$$= \frac{2}{N} \nabla w (y - w^* X) \cdot (y - w^* X) + \lambda 2 w^*$$

$$= \frac{2}{N} \cdot -X \cdot (y - w^* X) + \lambda 2 w^*$$

$$= -\frac{2}{N} X^T y + \frac{2}{N} X^T X w^* + \lambda 2 w^*$$

Setting to zero

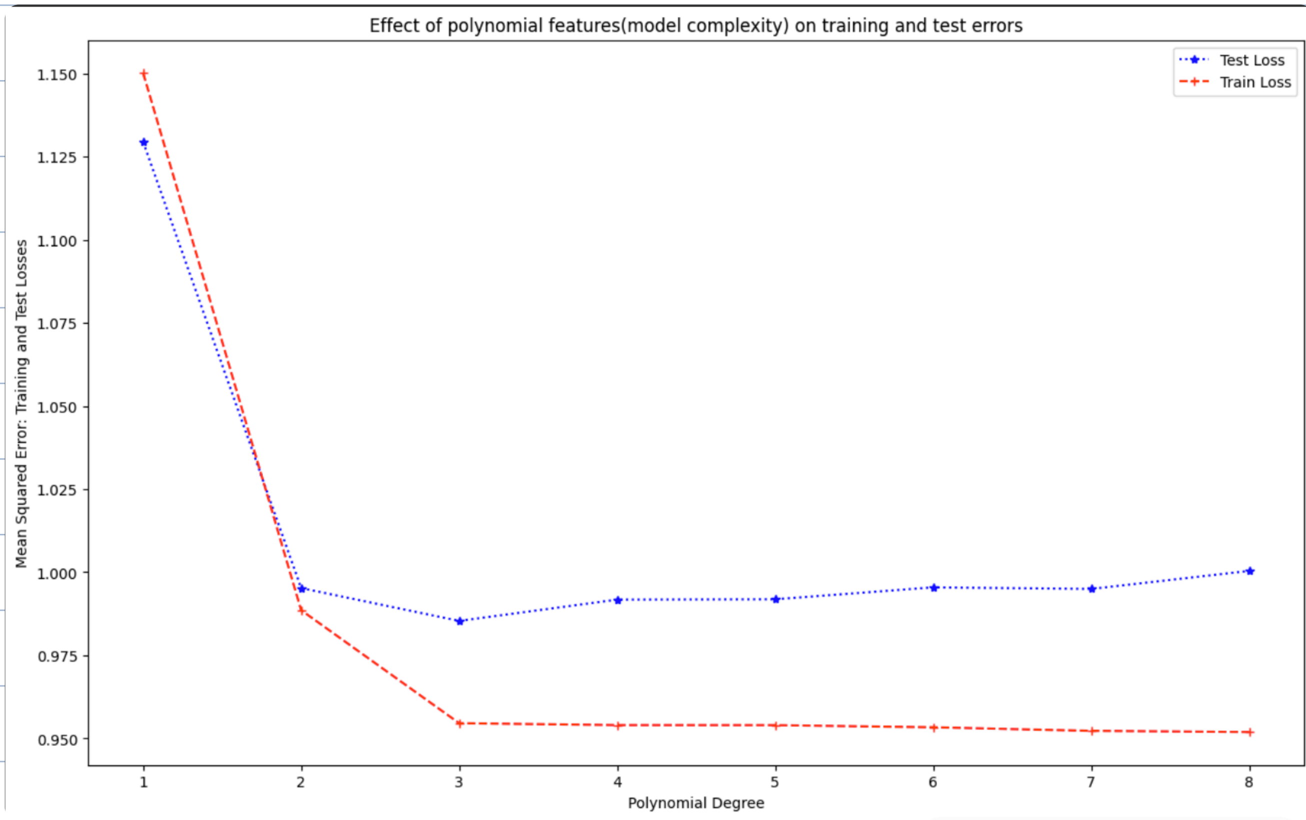
$$-\frac{2}{N} X^T y + \frac{2}{N} X^T X w^* + \lambda 2 \cdot w^* = 0$$

$$\frac{2}{N} X^T y = \frac{2}{N} X^T X w^* + 2\lambda w^*$$

$$\frac{2}{N} (X^T y) = \frac{2}{N} (X^T X + \underline{\lambda I N}) w^*$$

$$w^* = (X^T Y)(X^T X + N\lambda I)^{-1}$$

Coding Question : Section 1.3



Trend and reasoning

- At lower polynomial degrees (1–3), the model is too simple (underfitting) increasing the polynomial degree allows the model to capture more patterns in the data which reduces both training and testing errors. The model is too simple resulting in high error due to bias.
- At degrees (3-4) the training loss keeps decreasing so the model is getting better at fitting the training data. Bias & variance balance. The training loss stops improving and starts to slightly increase

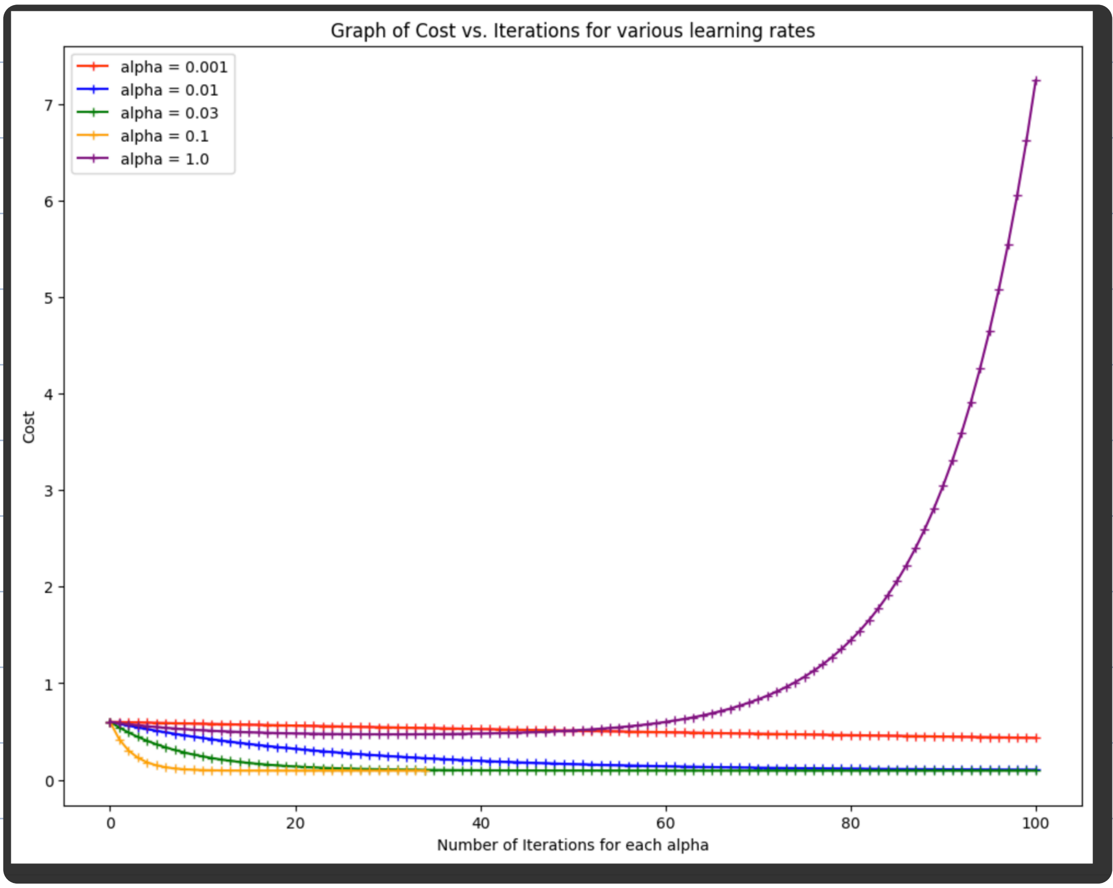
This means the model is starting to memorize data rather than generalizing well to unseen data.

- After degree 4-5, the training loss stays low and stabilizes so the model is probably perfectly fitting the data.

  On the other hand the test loss starts to increase slowly which is a signal that the model is starting to overfit.

  The model becomes too complex capturing noise and small flactuations instead of patterns.

Part 1.4 : Effect of learning rate on Gradient descent.



Comment on Effect:

- The lowest learning rate, <u>0.001</u> is not the best because it simply descends too slowly and gets stopped out by our max-iterations.
- The middle learning rates, 0.03 is much better than the lowest because by the time it get stopped out it achieves a lower cost an descends steadily downwards as well
- The yellow learning rate, 0.1 is the most appropriate choice because the cost reduces much faster and it converges after ≈ 35 iterations.
- The highest learning rate 1.0 is too high causing the cost to increase over time. This is expected because the gradient descent hops are so big that the overall cost <u>increases</u>