2.2

$$y = w^T \phi(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

$$\hat{w}(w) = \underset{w}{\operatorname{argmin}} \left[\|y - Xw\|^2 + \lambda \|w\|_1 \right]$$

Prior on $w \sim$ Laplace $(0, \theta)$

$$\rightarrow \text{Laplace}(w_i \mid 0, b) = \frac{1}{2b} e^{-\frac{|w_i - 0|}{b}}$$

Now

$$P(\theta \mid D) \propto P(D \mid \theta) P(\theta)$$

$$\log(P(\theta \mid D)) \propto \log(P(D \mid \theta)) + \log(P(\theta))$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\log(P(D \mid \theta)) + \log(P(\theta)) \right]$$

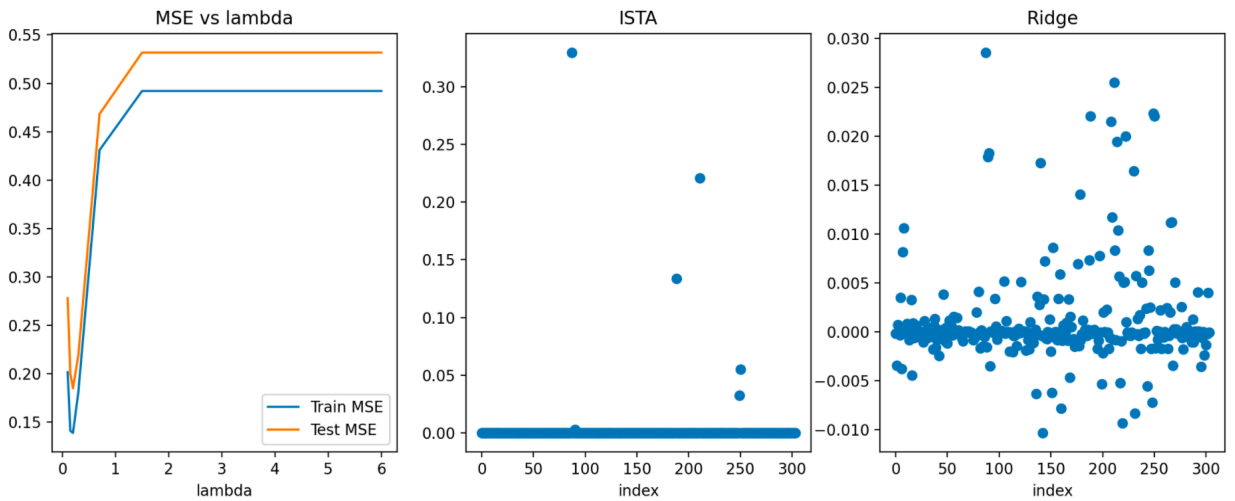
$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmax}} \left[\log \left(\prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - w^T x_i)^2}{2\sigma^2}} \right) + \log \left(\prod_{i=1}^{M} \frac{1}{2b} e^{-\frac{|w_i|}{b}} \right) \right]$$

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmax}} \left[-\sum_{i=1}^{N} \frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\|w\|_1}{b} \right]$$

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmin}} \left[\sum (y_i - w^T x_i)^2 + \lambda \|w\|_1 \right]$$

$$\lambda = \frac{2\sigma^2}{b}$$

$\hat{w}_{MAP}$ is same as $\hat{w}(w)$ - hence proved


MSE vs lambda
Train MSE
Test MSE
lambda
ISTA
index
Ridge
index

1.2 (b) PLOT Explanation

When λ is in range 0.1 to 0.2
Ideally if we let the model converge we should observe increasing train error but since maxiter is capped at 10,000 we observe higher error then we should but test error is dire as increase in lambda reduces overfitting and therefore we observe decrease in test mse initially
Further increasing value of lambda leads to increase in error as sparsity in weights is increased i.e features with non zero weights reduce. Error later on becomes constant as model is not learning.

1.2 (c) For weightvector of ISTA the number of components with a non zero value is very less compared to that of Ridge. This is expected and justified as loss function of LASSO is such that it only selects very few features.

2.1

1 vs 1

→ Splits a multiclass classification into one binary classification problem per each paired class

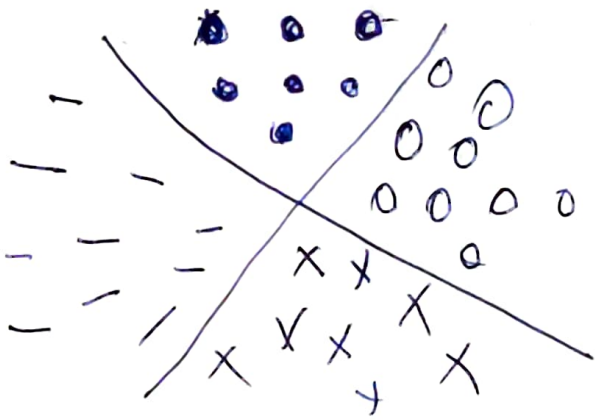
→ It can classify even when classes are mutually linearly seperable but not 1 vs rest linearly seperable

example →

1 vs Rest

→ Splits a multiclass classification into one binary classification problem per class

→ If not linearly seperable as 1 vs rest, Then it can't classify correctly



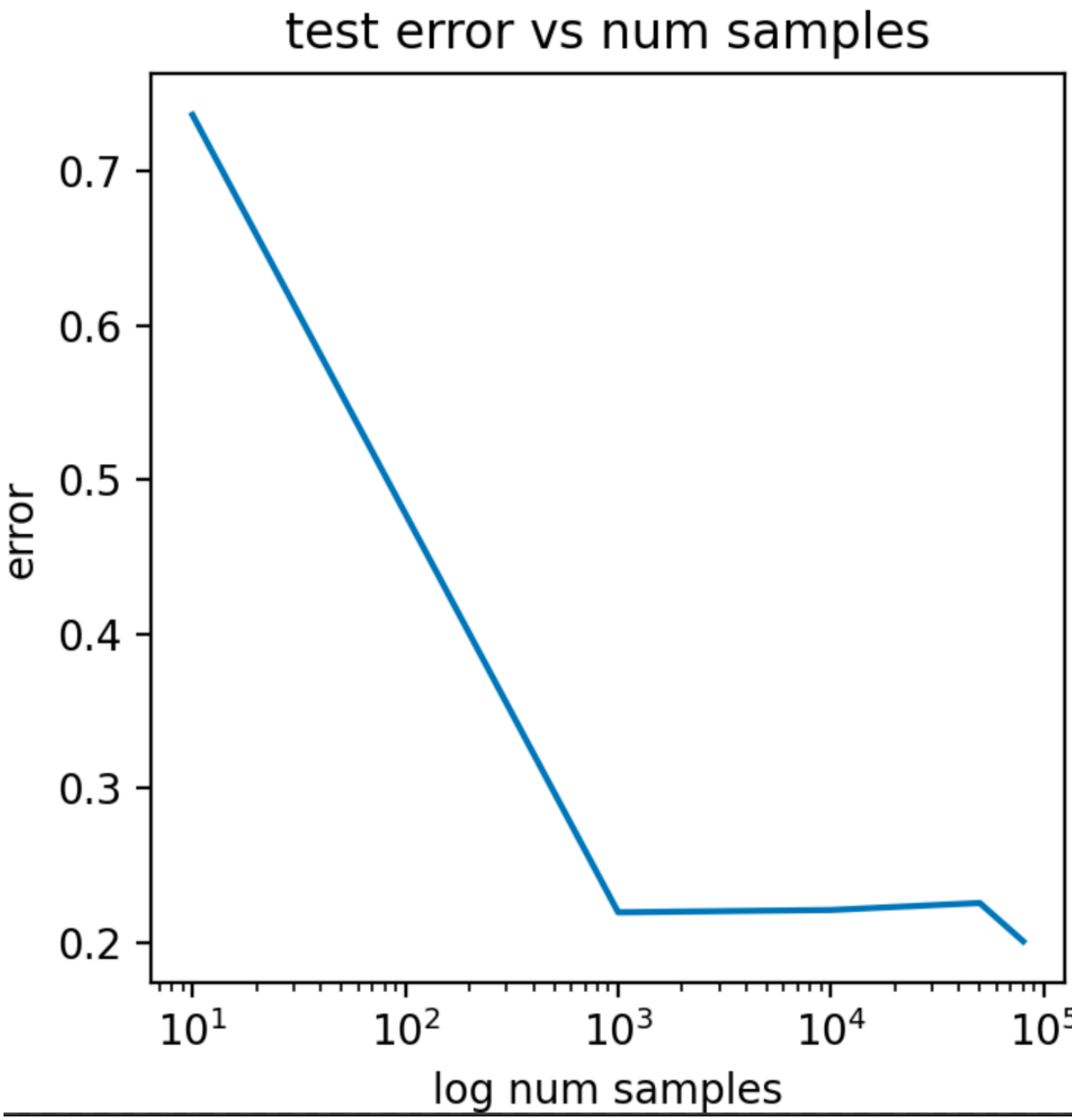
3.1 (a) increase in lambda reduces overfitting which has direct relation to decrease in variance and increase in bias. To be clear overfitting was reduced as in lasso increase of lambda increases sparsity in weight vector i.e components with zero weight increase hence reduced overfitting

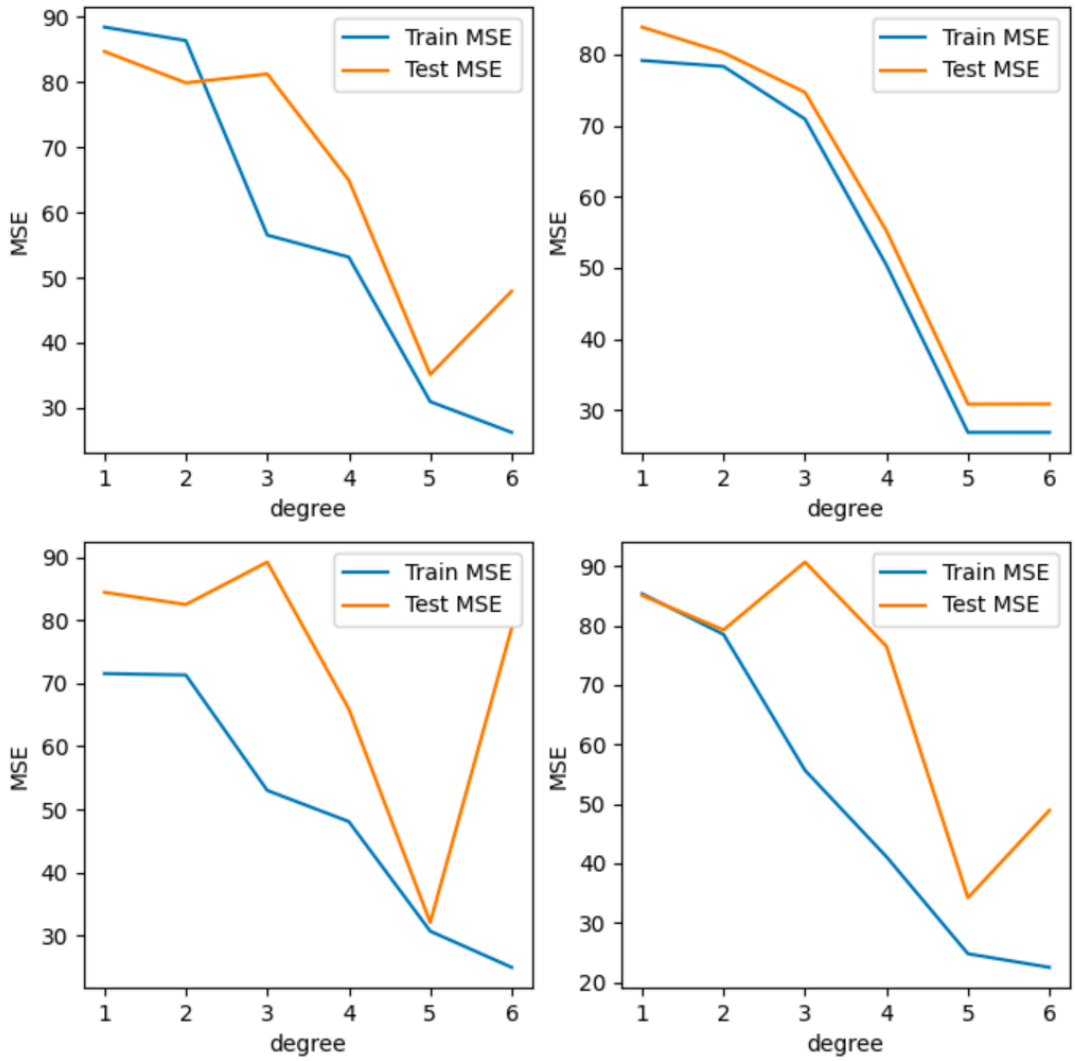
(b) The variance decreases and bias remains the same as the classification planes are more defined

~~(c) Such a modification to data will lead to overfitting faster~~

~~Thereof increase in variance and decrease in bias~~

(c) both bias and variance remain the same as the newly added features are redundant and not have any effect on output


test error vs num samples
error
0.7
0.6
0.5
0.4
0.3
0.2
10^1
10^2
10^3
10^4
10^5
log num samples


Train MSE
Test MSE
MSE
degree
Train MSE
Test MSE
MSE
degree
Train MSE
Test MSE
MSE
degree
Train MSE
Test MSE
MSE
degree

3.2

The test error will decrease as observed in graph as variance is reduced and bias remains the same.

Theoretically at degree = 6 we should obtain the least train error as it has the least bias of all

As we can see empirically in the plots that test MSEs seems to minimise at degree = 5 The bias decreases with increase in degree but variance increases as well so degree 5 is optimal for test mse. degree 6 is overfitting the data