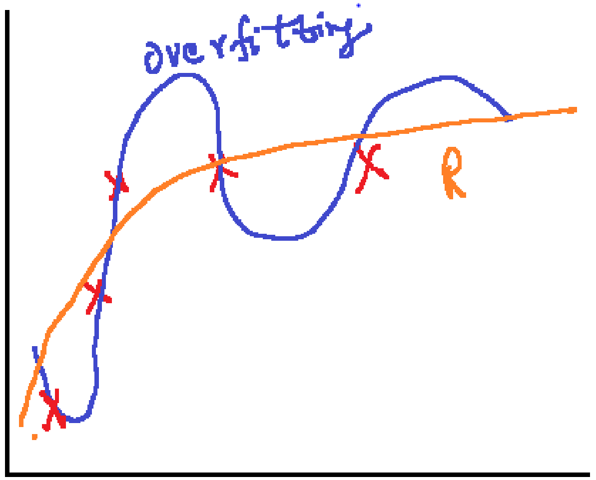
Regularization

Agenda

- Overfitting, Underfitting
- Bias, Variance
- Regularization
- L1 & L2 Loss Function
- Lasso – Ridge Regression
- Heavy Coding - Lasso – Ridge Regression

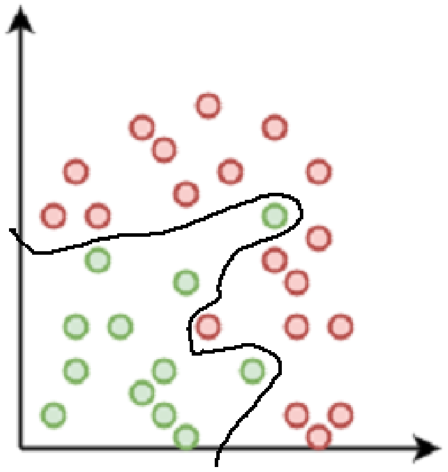
Overfitting

- Complex Decision Boundry
- Good fit of training data
- Poor fit of test Data



Fixing Overfitting

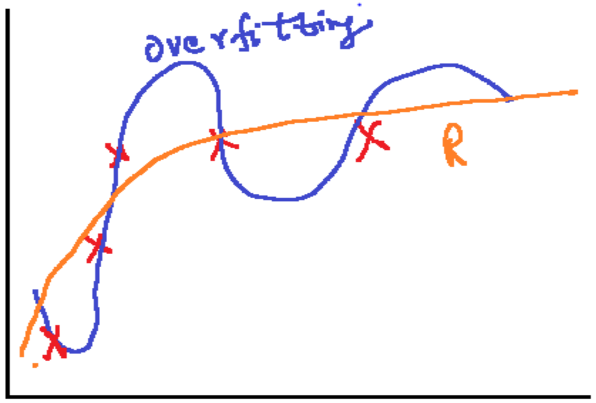
- Regularization Hyperparameter
- Cross Validation
- Bias – Variance trade off



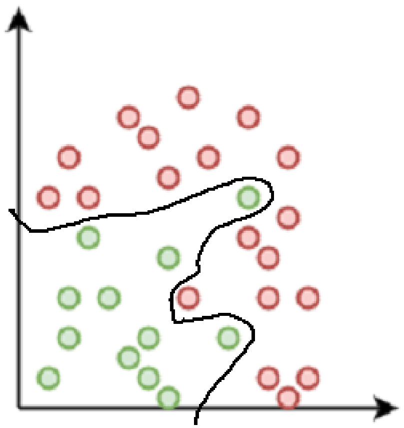
Underfitting

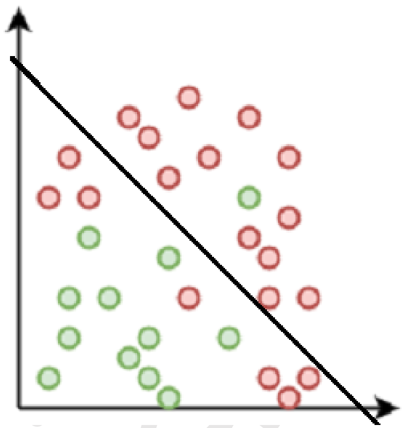
Underfitting Happens when:

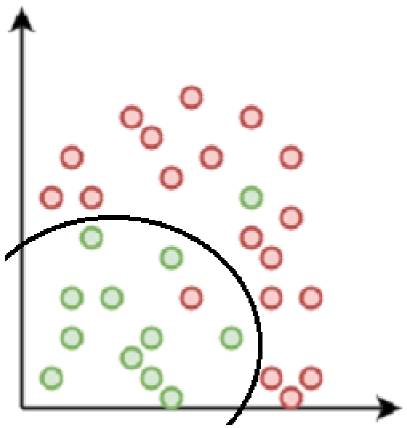
- Less amount of data to train
- Try to build linear model on non-linear data



Bias and Variance







- Overfitting
- High Variance

- Underfitting
- High Bias

- Good Model

Bias & Variance

- **Bias**: Difference between average prediction and the correct value which we are trying to predict
- **Variance**: Its the change in the amount of estimate of the target function on changing the training dataset.

- **High Bias** means model pays very little attention to training data and oversimplifies the model (Underfitting)
- **High Variance**: large change in estimate of the target function on changing training dataset (Overfitting)

- **Low Bias, High Variance**: Decision Trees, Simple Vector Machine and k-Nearest Neighbors
- **High Bias, Low Variance**: Linear Regression, Linear Discriminant Analysis and Logistic Regression

Regularization

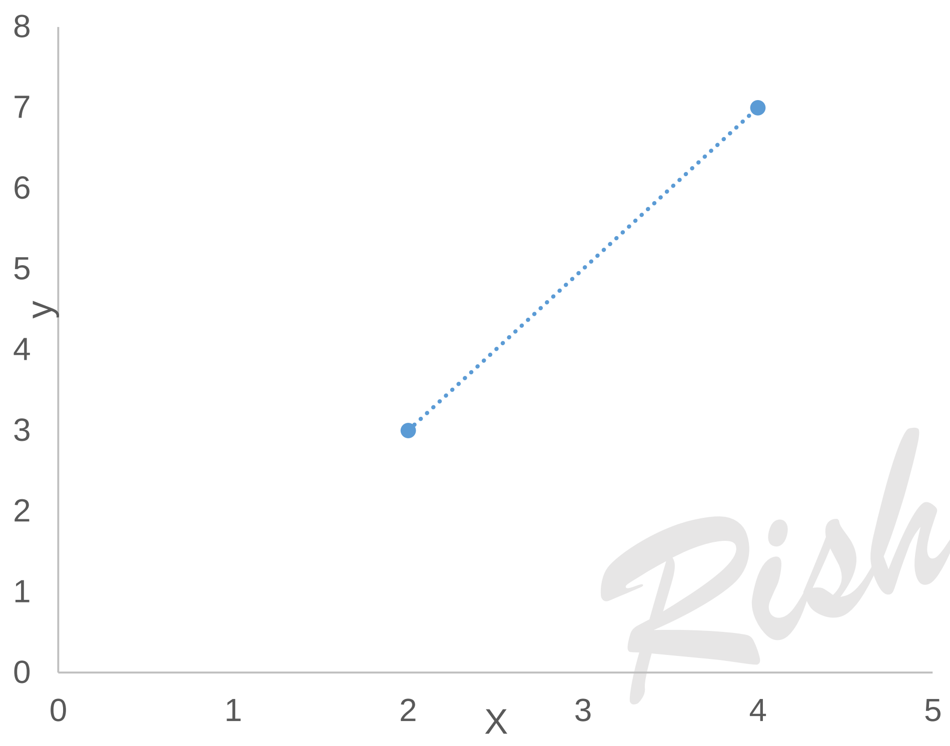
- Suppose we have a equation: $y = \theta_o + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
- Chances are there that above equation will overfit the training data.
- Intuition:
  If we penalize and make θ_3, θ_4 very small than effectively above equation will behave like $y = \theta_o + \theta_1 x + \theta_2 x^2$ without removing higher polynomial terms.

Cost Function: $J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{n} \theta_j^2]$

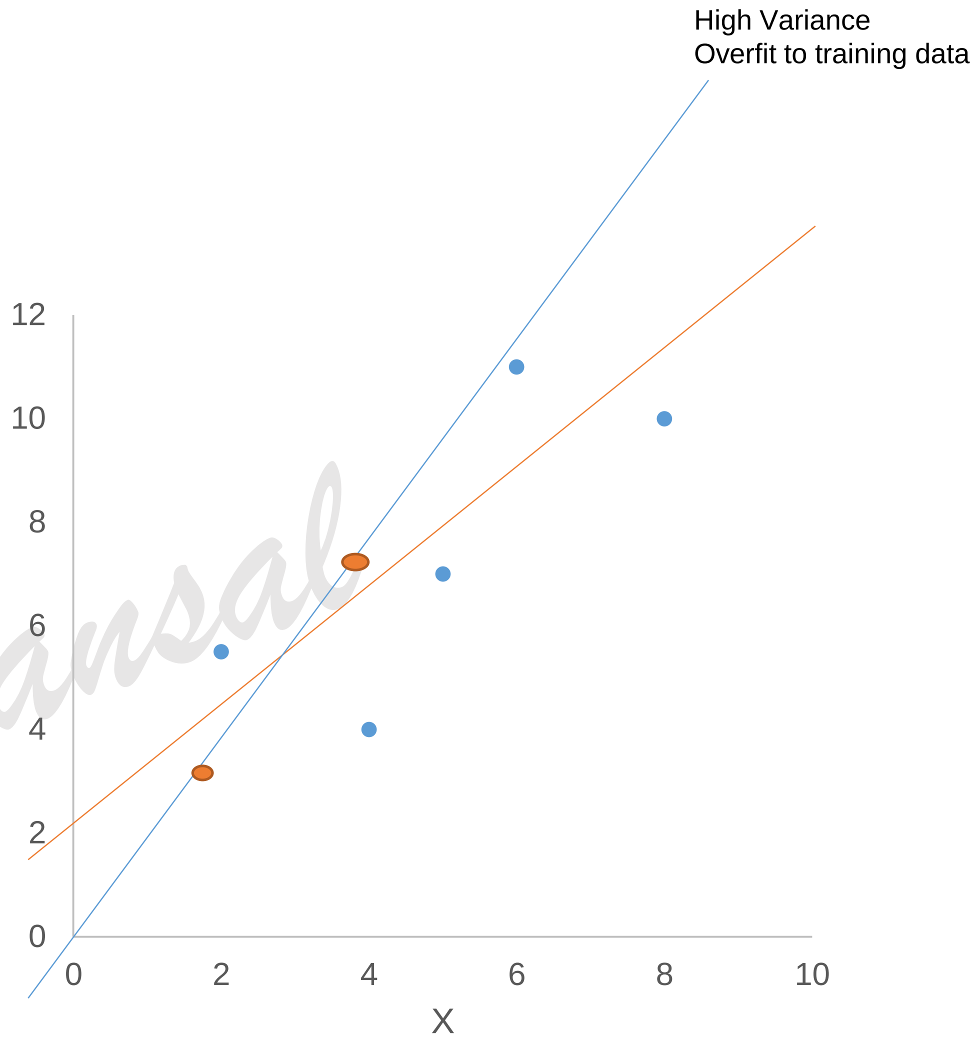
λ is the regularization parameter. It makes θ reduce after every iteration.

Regularization

- Keep all the features but reduce the magnitude/values of parameter θ.
- Works well when we have a lot of features, each of which contributes a bit to predicting y.
- Penalize Complex models
- Reduces variance error but increases bias
- E.g: Lasso, Ridge (Modeling Techniques)
- important when you have a dataset with 100,000+ features.



Training
RSS = 0



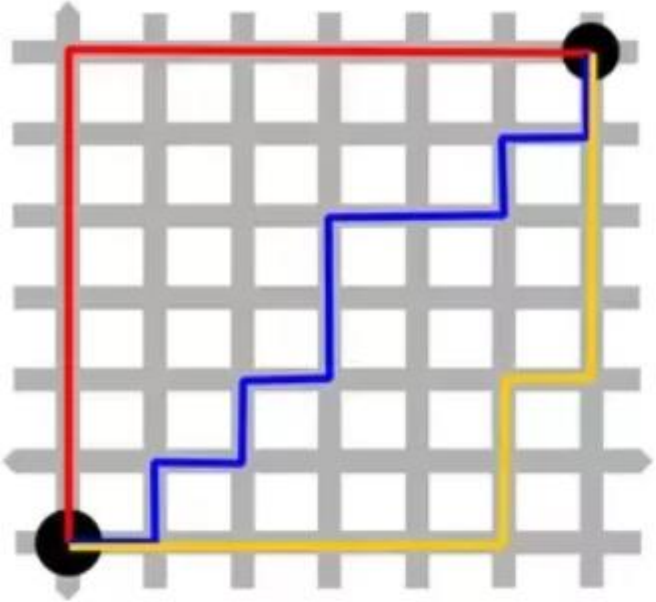
Testing
RSS - Large

L1 Norm Loss Function

- Least Absolute Deviations (LAD)
- Minimize error which is sum of all the absolute differences between the Actual value and

The predicted value

$$L1\, Loss\, Function = \sum_{i=1}^{n} |Y_{actual} - Y_{predict}|$$



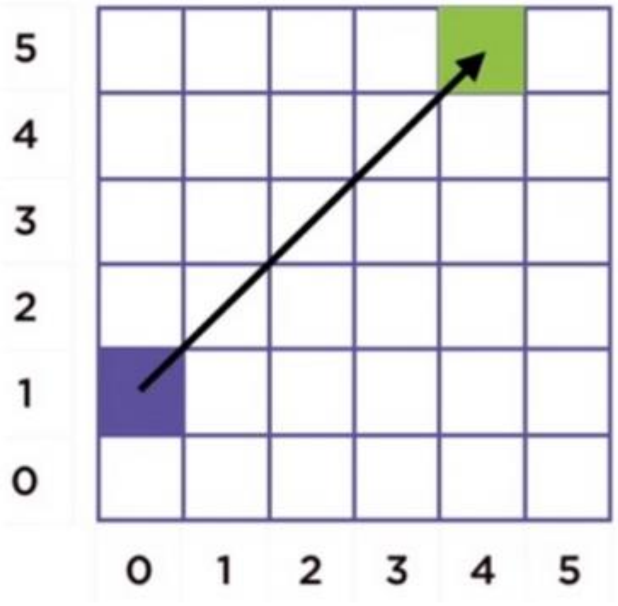
L2 Norm Loss Function

- Least squares error (LSE)
- Minimize error which is sum of all the squared differences between the Actual value and

The predicted value

$$L2LossFunction = \sum_{i=1}^{n}(Y_{actual} - Y_{predict})^2$$

- Penalize outliers heavily



Lasso Regression

- Cost Function: RSS + α*(sum of absolute values of coefficients)

$$\alpha(|\theta_0| + |\theta_1|)$$

- Add penalty for large coefficients
- Penalty function – L1 norm of regression coefficients
- Regularization hyperparameter (alpha) – how much severe penalty will be
- Larger values of α should result in fewer coefficients as the cost function needs to be minimized

Ridge Regression

Cost Function: RSS + α*(sum of squares of coefficients)

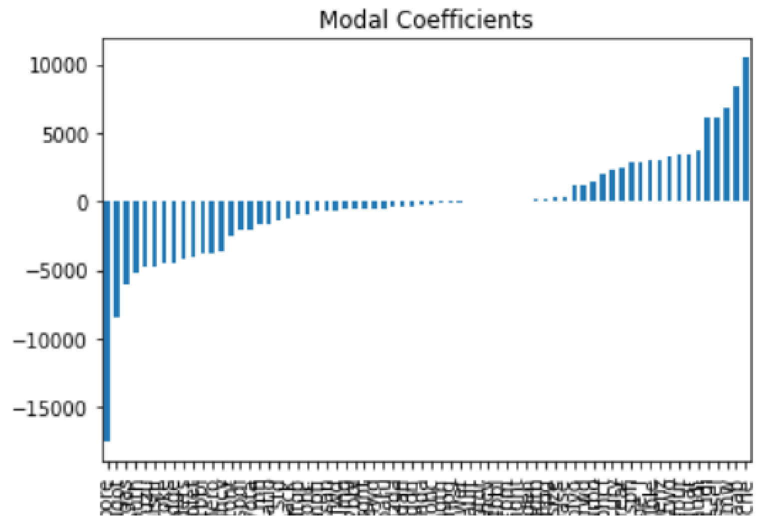
$$\alpha(|\theta_0|^2 + |\theta_1|^2)$$

- Larger values of α should result in smaller coefficients as the cost function needs to be minimized

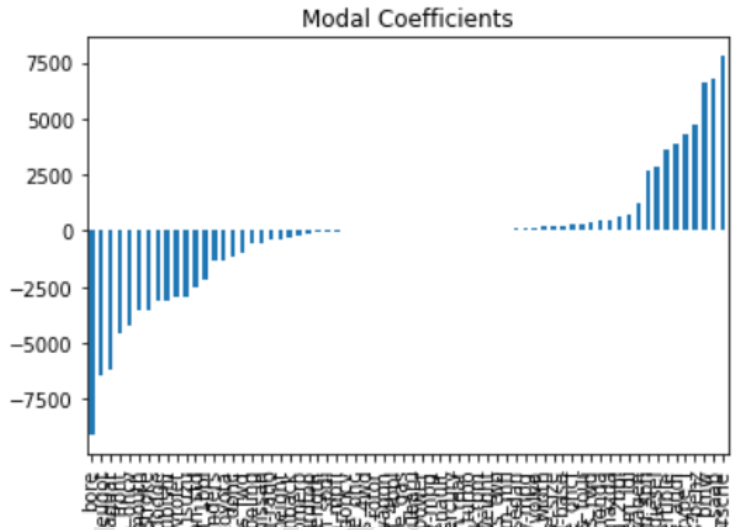
- Ridge Regression penalizes large coefficients even more than Lasso as coefficients are squared in cost function

Lasso – Ridge Comparison

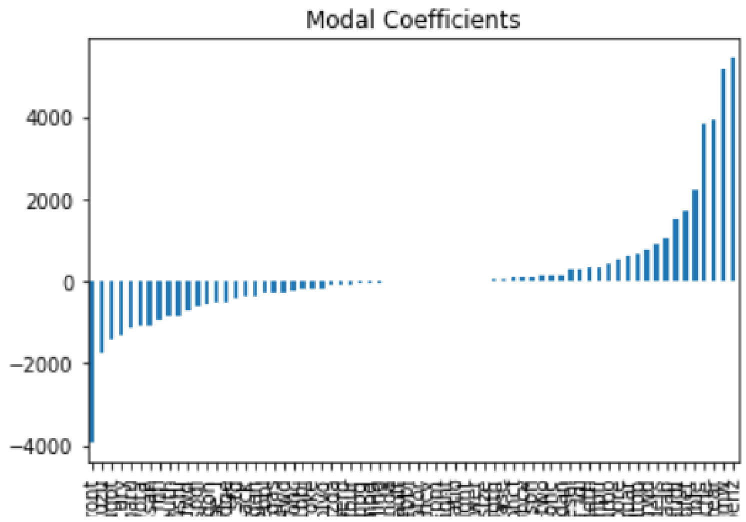
Without Lasso



Lasso, alpha = 0.5

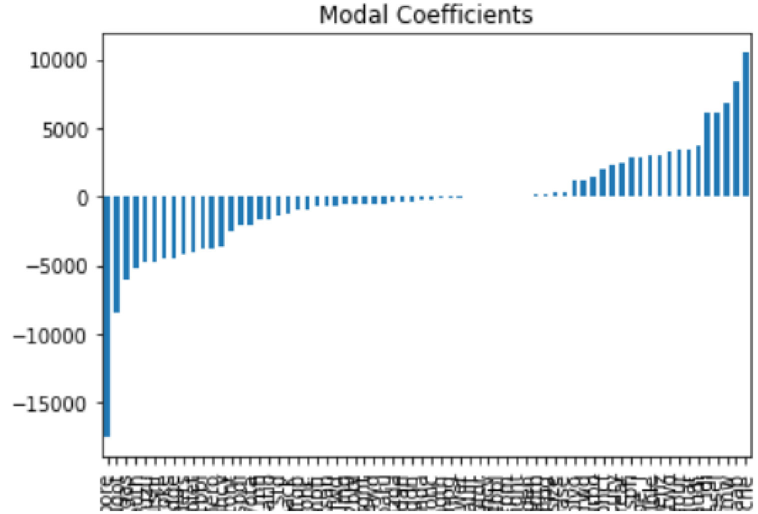


Ridge, alpha = 0.5

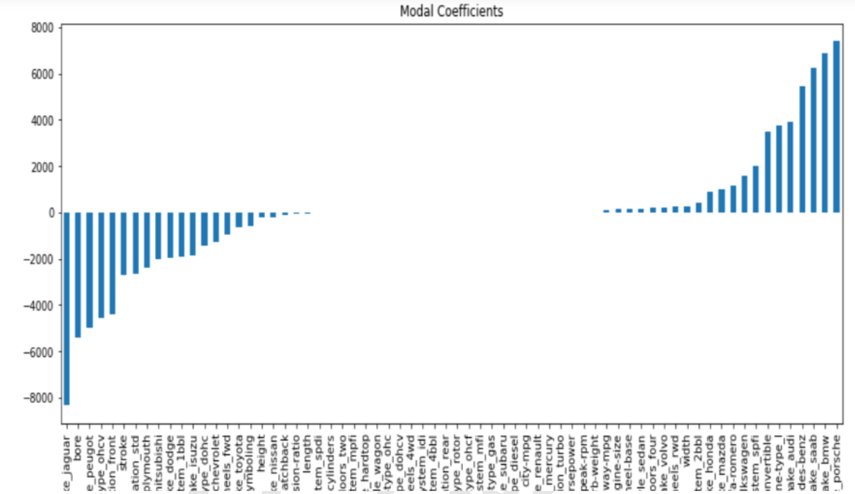


Lasso – Ridge Comparison

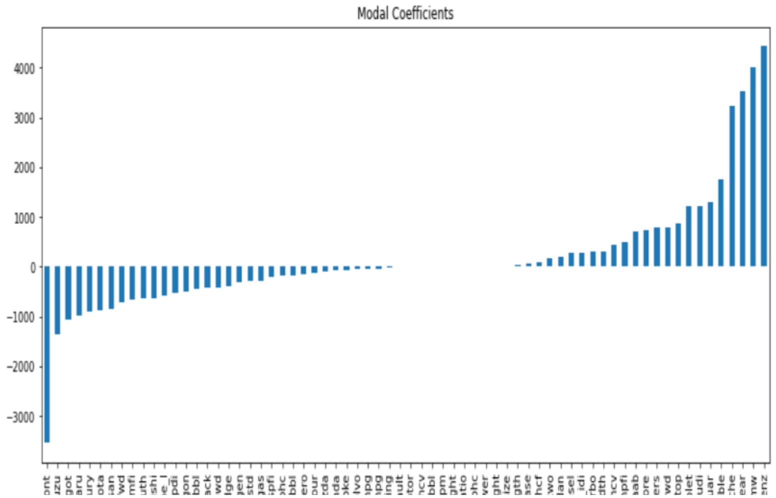
Without Lasso



Lasso, alpha = 1



Ridge, alpha = 1



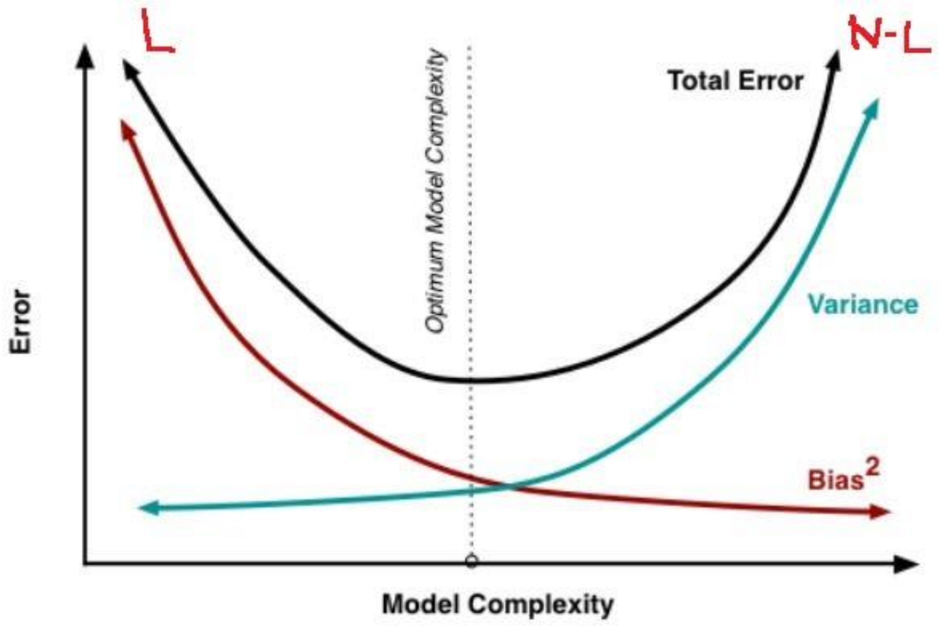
As alpha increases:
Lasso: coefficients are reducing to absolute zero
Ridge: coefficients are approaching zero, it penalizes large coefficients more compare to Lasso

Lasso is having a property called feature selection. Where, it selects only some of the features while reduces the coefficients of others to zero.
Ridge is not having any such property.

Bias Variance Trade off

- Goal of Supervised Algorithm is have low bias and low variance
- Linear Models - High Bias and Low Variance
- Non Linear Models - Low Bias and High Variance
- Increasing the bias -> decrease the variance
- Increasing the variance -> decrease the bias

- Total Error = bias^2 + variance + Irreducible error



- Parameterization used to balance bias and variance