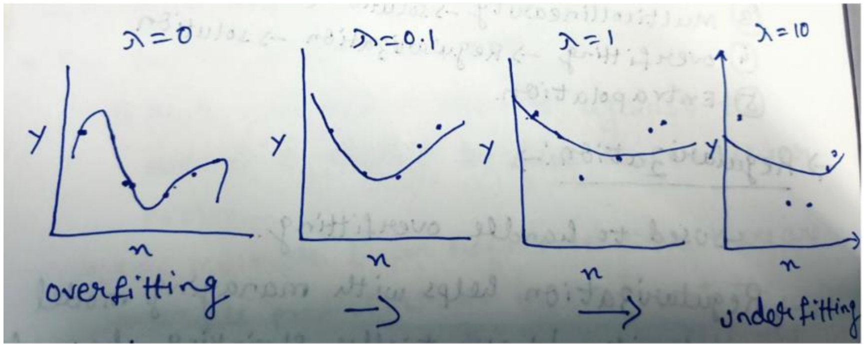
# Subjective Questions

1. **What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?**

**Ans:**

- The optimal value of alpha is –
  - For Ridge regression - **4.0**
  - For Lasso regression - **0.0001**
- If I double the value of alpha –



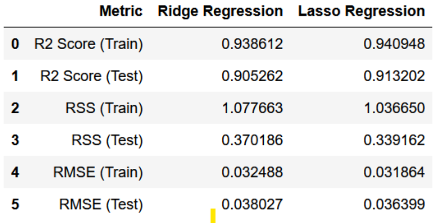
As the alpha value increases, the model will become more simpler which lead to underfitting.
In our case, if alpha value is doubled –

  - For Ridge regression, if alpha value is changed from 4 to 8 –
    - R2 score of train data is reduced from 0.9386 to 0.9313.
    - R2 score of test data is reduced from 0.9052 to 0.9037
  - For Lasso regression, if alpha value is changed from 0.0001 to 0.0002 –
    - R2 score of train data is reduced from 0.9409 to 0.9300.
    - R2 score of test data is changed from 0.9132 to 0.9135
  - I have observed the magnitude of coefficients also decreased.
- Important predictor variables after alpha is doubled –
  - For Ridge regression –
    - When alpha is 4 -> OverallQual (0.096803)
    - When alpha is 8 -> OverallQual (0.080534)
  - For Lasso Regression –
    - When alpha is 0.0001 -> GrLivArea(0.252663)
    - When alpha is 0.0002 -> GrLivArea(0.241984)
  - Even though in both the cases the important predictor variable did not change **but the coefficient value in both the cases is penalized(reduced)**

2. **You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?**

**Ans:**

- I have got the following test results after evaluating the model.



| | Metric | Ridge Regression | Lasso Regression |
|---|---|---|---|
| 0 | R2 Score (Train) | 0.938612 | 0.940948 |
| 1 | R2 Score (Test) | 0.905262 | 0.913202 |
| 2 | RSS (Train) | 1.077663 | 1.036650 |
| 3 | RSS (Test) | 0.370186 | 0.339162 |
| 4 | RMSE (Train) | 0.032488 | 0.031864 |
| 5 | RMSE (Test) | 0.038027 | 0.036399 |

-

- We can clearly observe that both Ridge and Lasso regression models provide same accuracy.
- However, Ridge model uses almost all the features(predictors) available where as **Lasso does the Feature elimination** and uses only 131 features for prediction.
- Thus we can choose Lasso regression model as our final model.

3. **After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?**

**Ans:**

- Initial 5 most important predictor variables for Lasso regression model are –

| | | |
|---|---|---|
| 1. | GrLivArea | 0.252663 |
| 2. | Condition2_PosN | -0.190465 |
| 3. | OverallQual | 0.156167 |
| 4. | TotalBsmtSF | 0.104795 |
| 5. | OverallCond | 0.099088 |

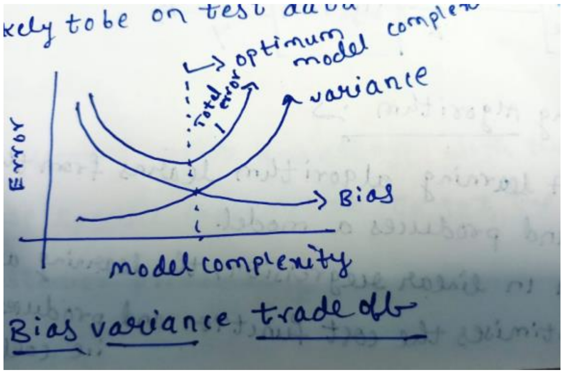
- After, I removed these predictors and build the Lasso regression model again, I got the same optimal value of alpha. Now the new important predictor variables are –

| | | |
|---|---|---|
| 1. | 1stFlrSF | 0.207689 |
| 2. | 2ndFlrSF | 0.101392 |
| 3. | Functional_Maj2 | -0.086102 |
| 4. | MSZoning_RL | 0.083527 |
| 5. | MSZoning_FV | 0.081519 |

4. **How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?**

**Ans:**

- If the model memorizes the entire training data, if you change the dataset slightly then this model will also need to change drastically. This model is therefore less robust and is sensitive to changes in training data and is called high variance. This situation is called overfitting.
- Bias quantifies how accurate the mode is likely to be on test data.
- A model should not be either overfit or underfit. It should have less bias and less variance. And hence it should be optimal complexity.



- 
- **Regularization** is the process of simplifying models to achieve the correct balance between keeping the model simple and not to naïve.
- **Hyperparameters** are the ones we pass on the learning algorithm to control the complexity of the final mode.