Report

CS F320 (Foundation of Data Science)

Assignment 1

Our approach:

- The data preprocessing involved loading the dataset, normalizing the feature variable by using the formula $X' = (X - \mu) / \sigma$, shuffling the data by index, and splitting it into training (80%) and testing (20%).
- We then built the regression model for degrees 1 to 9 with 500 epochs.
- Then we implemented L2 regularization and built 9 new models, and found optimal lambda using Bias-Variance Decomposition, minimizing the total error (sum of bias squared and variance).
- Finally, we visualized the results.

1. Without regularization:
    - We implemented batch gradient descent for 500 epochs to arrive at the minimal weights/coefficients and bias term w0.
    - We plotted the mean squared error against degree of polynomial and observed that MSE was minimum for polynomial regression of degree 5.

2. With L2 regularization (ridge regression):

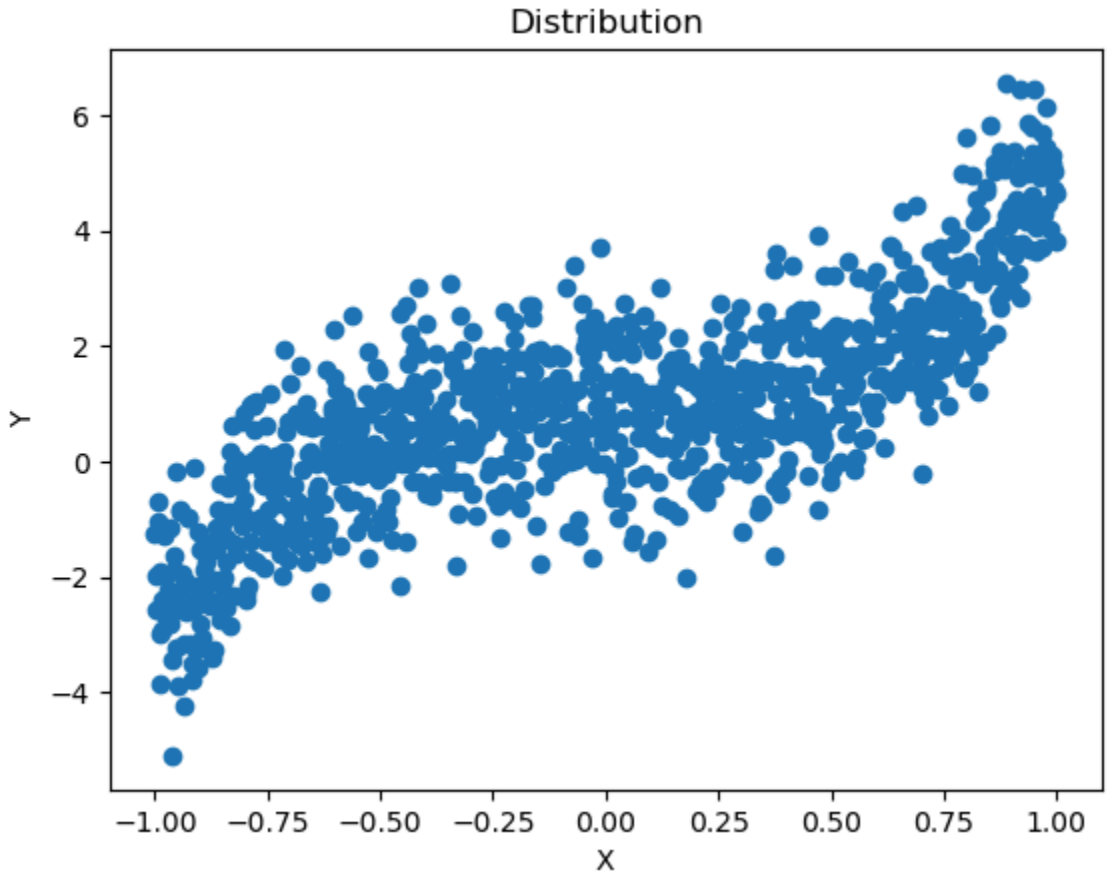
- We added the regularization term (lambda * sum of squares of weights) to the gradient descent algorithm.
- For each degree from 1 to 9, we found the square of bias and variance for each lambda value using Numpy's logspace function (from 10^-8 to 1) to find the model with least error.
- We observed least MSE for degree 2 as lambda approaches 1.

As we increase the degree, the error remains the same across lambda values which is due to overfitting.
Regularization penalizes large coefficients, thus reducing overfitting.

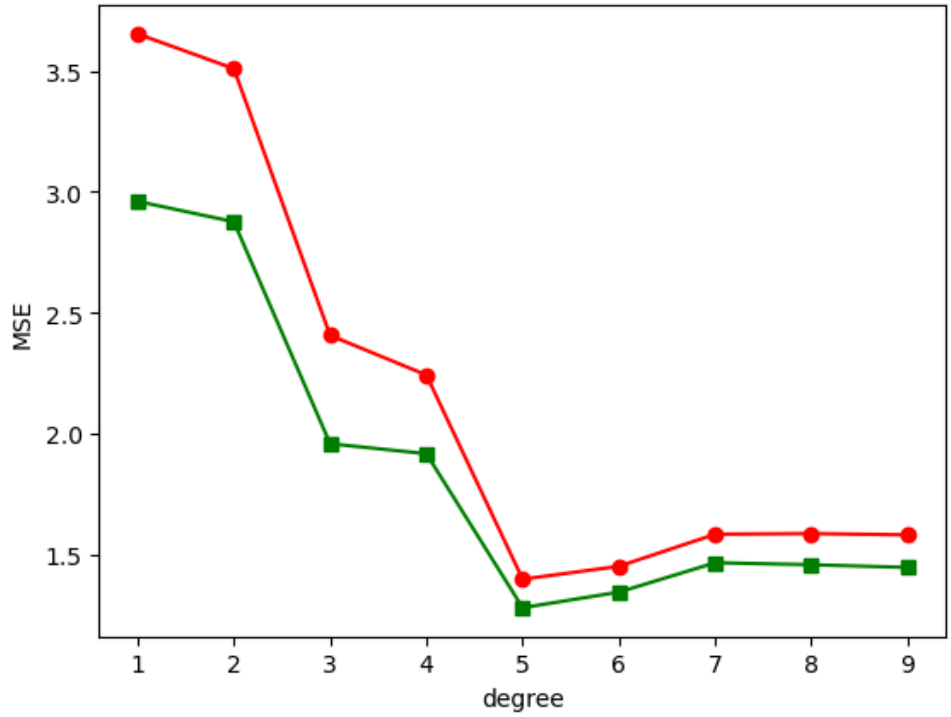
In case of underfitting, bias is high (close to the total error) whereas for overfitting (higher degrees of polynomial), bias is low and variance is high, as we can observe from the plots.

Small values of lambda leads complex models, which causes overfitting. Larger values on the other hand give very simplified models, risking underfitting.

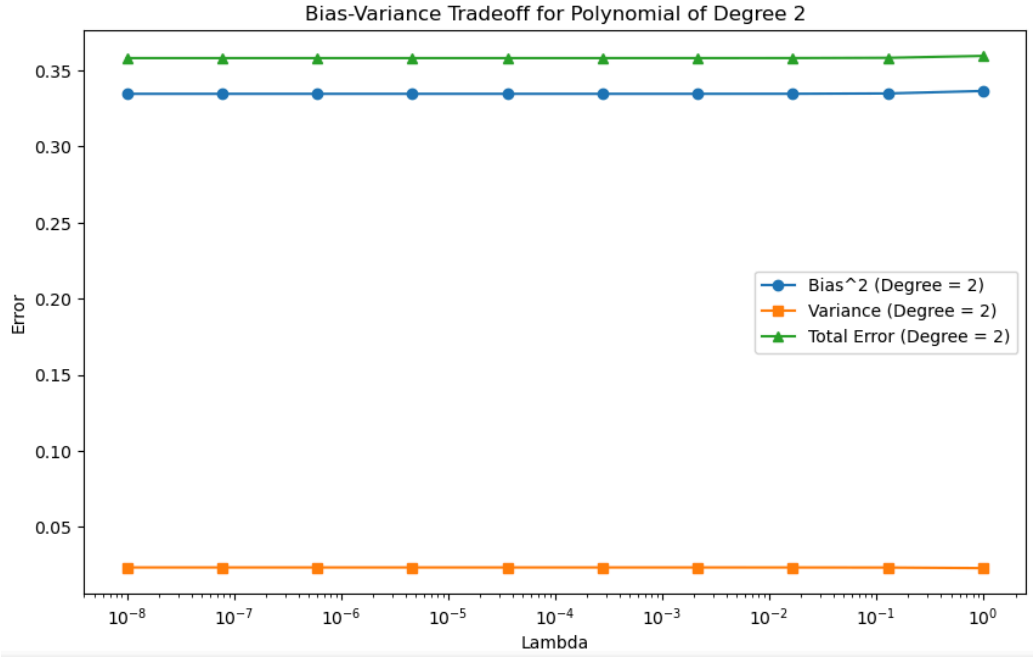
Plot 1:



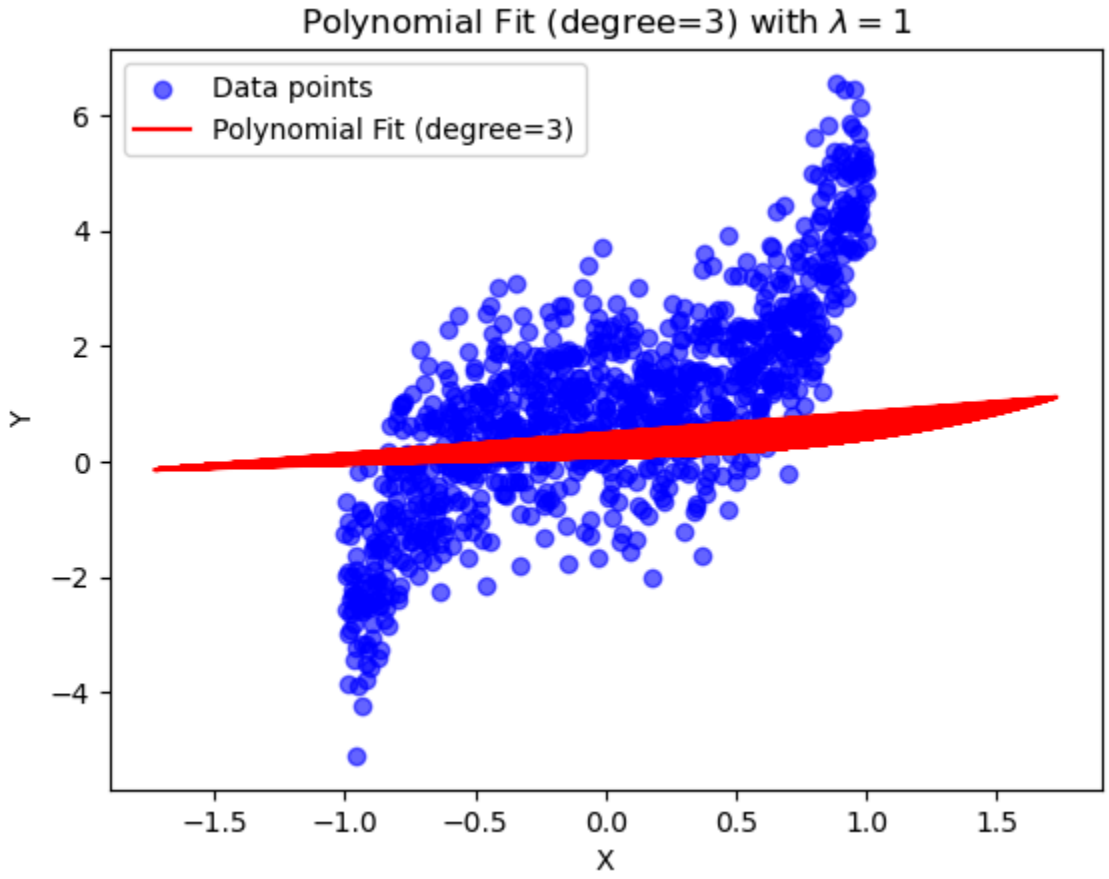
Plot 2: Without regularization



Plot 3: With L2 regularization



Plot 4: Best fit model



YASH KANTAMNENI - 2022A7PS0120H
SAVITRA ROY - 2022A4PS1490H
MOKSH JAIN - 2022A7PS0186H