Report
FYS-STK4155 - Project 1

Malene Fjeldsrud Karlsen and Birk Tininus Skogsrud
University of Oslo
(Dated: September 26, 2025)

This report focuses on the fundamentals of Machine Learning through mathematical and statistical aspects by producing a dataset and creating algorithms for various regression methods. The regression methods used in the report are Ordinary Least Squares, Ridge Regression, and Lasso Regression. Furthermore, we will introduce methods of Gradient Descent and momentum to enhance the learning, and resampling methods to reproduce data sets efficiently and profitably. The goal is to get a better understanding of the fundamentals of Machine Learning and compare and analyze how these algorithms and methods work. To solve this, the report will contain a numerical solution of the varying methods and a results and discussion section to discuss and highlight what we can learn from these algorithms. The main results and implications will be discussed later, as I am sadly not done with the tasks yet :(

I. INTRODUCTION

Using Runges function, we will study how we can use Machine Learning theory to best fit the data of this one-dimensional polynomial. To generate a model for the polynomial, this report will focus on the fundamentals of Machine Learning. Using a supervised learning method, we will in this project create algorithms for various regression methods; the Ordinary Least Squares, Ridge Regression, and Lasso regression, allowing us to compare the outcomes and decide on which one of the regressions suits our data the best. To achieve this we must identify possible sources of errors in our model, such as overfitting the model with too complex data, or underfitting our data with too little data. Overfitting will not allow our model to capture the underlying patterns of our dataset, while underfitting our data does not provide enough information for our model to learn the real behavior of the provided data points. To find these sources of errors and navigate through these challenges, we perform statistical analysis of the mean squared error and the R2-score to evaluate the fit and study how well our model performs. The Mean Squared Error and R2-score, also known as a cost function and metrics, allows us to evaluate the errors in our model by comparing our predicted data with our data. Finding the data points (parameters) that suits our data is achieved through the regression algorithms, while gradient descent and optimization methods such as AdaGrad, RMSProp, and ADAM is used to find the parameters that minimizes the difference between our predicted values and our real data. Finding the minimum of our cost functions helps in creating the best model for the data. This report aims to provide a better understanding of how to increase efficiency in our models and how to navigate through the challenges of overfitting, underfitting, high variance and high bias through the bias-variance tradeoff. In the end, the report will include methods on how to resample our datasets using the Bootstraps method and cross-validation to ensure our model understands and adapts to the variability of the dataset while producing new datasets in an profitable manner. Ensuring this process ensures the model can be used in the future for new datasets. In the Methods Section (II) of this paper we will provide a mathematical explanation of the methods used, as well as the numerical solutions. All results will be discussed and reviewed to compare the methods used and their efficiency as well as their ability to fit a model to our given data in Section III of the report. The challenges faced in the numerical solutions will be presented if they affect our model output and critically reviewed. Furthermore, we will conclude on our most important findings in section IV and what to focus on in future projects of similar nature.

II. METHODS

A. Method 1/X

- Describe the methods and algorithms, including the motivation for using them and their applicability to the problem
- Derive central equations when appropriate, the text is the most important part, not the equations.

B. Implementation

- Explain how you implemented the methods and also say something about the structure of your algorithm and present very central parts of your code, not more than 10 lines
- You should plug in some calculations to demonstrate your code, such as selected runs used to validate and verify your results. A reader needs to understand that your code reproduces selected benchmarks and reproduces previous results, either numerical and/or well-known closed form expressions.

C. Use of AI tools

We have used chat for help to understand what you meant with upload an LLM. It is now uploaded to the GitHub Repository.

- Describe how AI tools like ChatGPT were used in the production of the code and report.

III. RESULTS AND DISCUSSION

This is us referring to the figures. Figure one is found at 1 and figure two is found at 2.

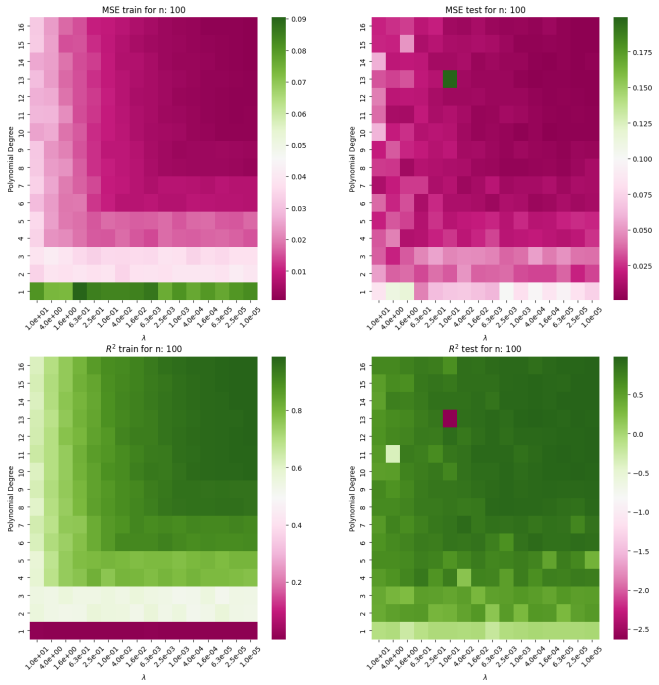

Figure 1: Heatmap of MSE and R2 score of a Ridge Regression model. The MSE and R2 results are found for different λ on the x-axis and for an increasing polynomial degree on the y-axis. The colorbar explains the MSE values for the two first plots and the R2 scores for the last two. (For the resulting project we will index the subplots with figure names such as a, b, c and d. We dont have time now though. We will also label the colorbar.)

IV. CONCLUSION

- State your main findings and interpretations
- Try to discuss the pros and cons of the methods and possible improvements
- State limitations of the study
- Try as far as possible to present perspectives for future work

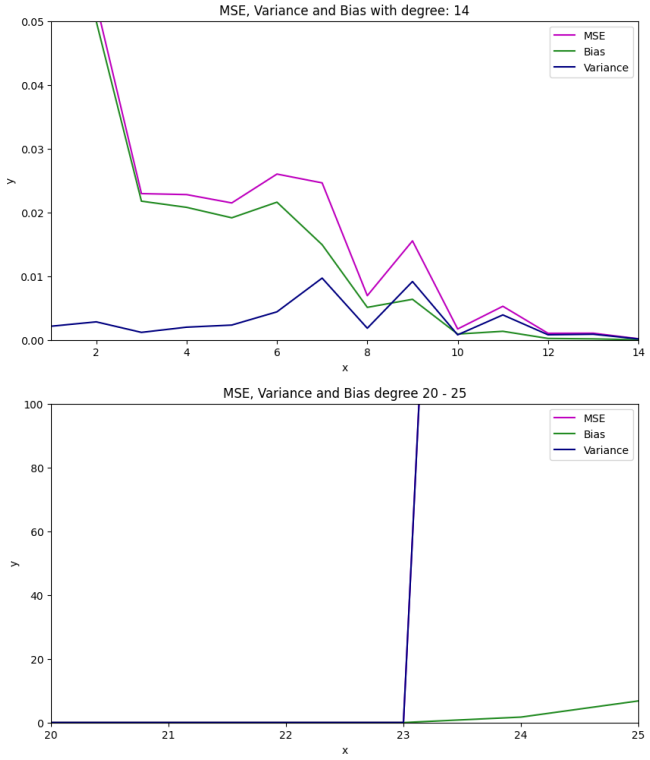

Figure 2: The variance bias tradeoff. One can see that the variance goes higher after the polynomial degree 23, but both the variance and bias lowers in value until then.