

Report Template

FYS-STK3155 - Project X

Theodor Waalberg
University of Oslo
(Dated: September 23, 2025)

In this project, we investigate the application of supervised machine learning techniques to regression problems, with a particular focus on understanding the bias/variance trade-off and the effects of regularization. Using synthetic datasets, we implement and compare Ordinary Least Squares (OLS) and Ridge regression models across varying levels of model complexity and regularization strength. Our results demonstrate how model selection and hyperparameter tuning influence prediction accuracy, and we visualize these effects through informative figures. The findings provide practical insights into balancing bias and variance for optimal model performance.

I. INTRODUCTION

Machine learning has become an essential tool for analyzing and modeling complex data across science and industry. In regression analysis, the goal is to predict a continuous outcome variable from one or more input features, while avoiding both underfitting and overfitting. Achieving robust generalization requires careful control of model complexity and a clear understanding of the bias-variance trade-off. In this project, we explore these concepts by implementing and evaluating Ordinary Least Squares (OLS) [1] and Ridge regression. We systematically study how polynomial model degree and regularization strength (ridge penalty) affect prediction error measured by mean squared error (MSE). To make these effects tangible, we present visualizations of the bias-variance trade-off and a heatmap of MSE across degrees and regularization parameters. Together, these results illustrate how model selection and hyperparameter tuning can improve predictive performance.

The rest of the report is organized as follows. Section II describes the methods and implementation details used in our experiments. Section III presents and discusses the results, including the figures that illustrate the trade-offs inherent to model complexity and regularization. Finally, Section IV summarizes the main findings and outlines potential directions for future work.

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 used ChatGPT to assist with drafting text and troubleshooting minor coding issues. All generated suggestions were reviewed, verified, and edited by us to ensure correctness and relevance to the project.

III. RESULTS AND DISCUSSION

- Present your results
- Give a critical discussion of your work and place it in the correct context.
- Relate your work to other calculations/studies
- An eventual reader should be able to reproduce your calculations if she/he wants to do so. All input variables should be properly explained.
- Make sure that figures and tables contain enough information in their captions, axis labels etc. so that an eventual reader can gain a good impression of your work by studying figures and tables only.
- Figure 1 [2] illustrates the bias-variance tradeoff, demonstrating how model complexity affects the balance between bias and variance in predictive modeling.

- Figure 2 shows the MSE of a Ridge regression model for various polynomial degrees and lambda values, highlighting how regularization strength impacts model performance across different complexity levels.

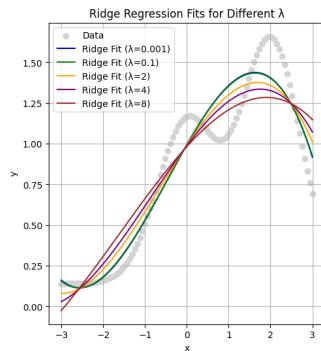


Figure 1: Variance tradeoff for different normalizing values of lambda.

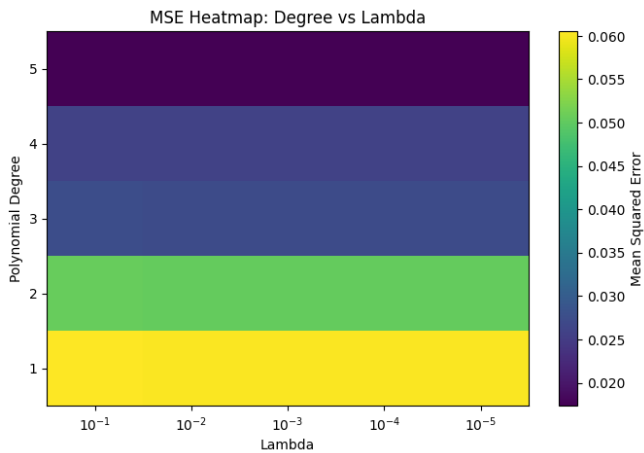


Figure 2: Heatmap showing the MSE of a Ridge regression model for various polynomial degrees and lambda values.

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

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics (Springer, New York, 2009), URL <https://link.springer.com/book/10.1007/978-0-387-84858-7>.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Journal of Machine Learning Research* **12**, 2825 (2011).