# Regression models
## FYS-STK3155 - Project 1

Herman Nissen-Sollie
*University of Oslo*

(Dated: September 26, 2025)

Regression models are used to predict based on correlation between different variables, and this can be used to make valuable predictions, especially in datasets that are large or have more complex relations which could be hard to pick up or quantify. This paper investigates the performance of three widely used regression models: Ordinary Least Squares (OLS), Ridge regression, and LASSO regression. Synthetic datasets were generated using the Runge function, with varying levels of noise, to test the models under controlled conditions. We analyze the bias–variance tradeoff, evaluate the role of hyperparameters, and examine the effectiveness of optimization techniques such as gradient descent and resampling (bootstrap and cross-validation). Quantitative comparisons are carried out to highlight the strengths and limitations of each approach. The results, while ongoing, aim to provide insight into the practical behavior of these regression methods beyond their theoretical properties. All codes used to produce the results in this paper will be available at: `https://github.com/pushing-py/Project_1`

## I. INTRODUCTION

Machine Learning (ML) is a field concerned with developing algorithms that can automatically learn patterns from data [1]. Unlike traditional programming, where explicit rules are coded, ML algorithms improve their performance by identifying structures in datasets. patterns that are often too complex for humans to detect directly. This ability is particularly valuable given the vast amounts of data that can now be processed efficiently by modern computational methods.

Regression analysis is one of the fundamental approaches in ML, with widespread applications in fields such as medicine and economics. The primary objective of regression models, as well as many other ML algorithms, is to learn from existing data and predict outcomes for unseen cases. More specifically, regression aims to capture the relationship between a set of independent variables and a dependent variable, allowing new predictions to be made when new inputs are provided. However, this approach comes with challenges, including how to handle noise in the data and how to manage the computational cost of the algorithms when applied to large datasets. Choosing and applying a model is therefore not a "one-size-fits-all" task, but rather a matter of understanding the nature of the data and balancing different compromises to obtain satisfactory results.'

In this report, we focus on three regression models: Ordinary Least Squares (OLS), Ridge regression, and LASSO regression. These models are applied to data generated from the Runge function, a classical test case in numerical analysis due to its sensitivity to polynomial approximation. We compare the performance of the different regression methods, analyze how hyperparameters affect their predictions, and explore optimization techniques such as gradient descent and resampling (bootstrap and cross-validation). We also investigate how different levels of noise influence the models' behavior and performance.

The goal of this study is to develop a deeper understanding of regression models, connect theoretical insights to practical results, and critically assess the methods and choices made throughout the analysis.

## II. METHOD

### A. AI declaration

AI has been used to polish the language in my writing, where i have gone trough some drafts of certain sections of the text, given them as prompts and implemented some of the LLM's response. AI has also been used to help with Latex, scince I'm not very profient in this. I have also used AI to help with coding, since it is good at detecting the errors and sometimes make my code more readable and easier to use.

## III. RESULTS AND DISCUSSION

Figure 1 shows that the MSE on the training data decreases monotonically as polynomial degree increases, since more complex models can always fit the training points better. For the test data, the MSE first decreases (reduced bias), but eventually starts to increas, which is a clear indication of overfitting: the model begins to capture noise rather than the underlying signal. This illustrates the bias–variance tradeoff: higher complexity reduces bias but increases variance, leading to worse generalizability when the model becomes too flexible. The splitting of the dataset was done wit the use of scikit-learn[2].

On the ridge reegrassion model, we could see on our data the lambda values, and the change to them, had little effect. This is shown clearlyu in Figure 2

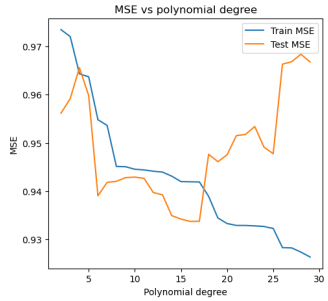
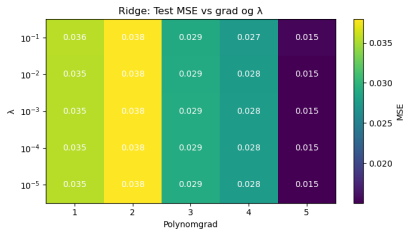
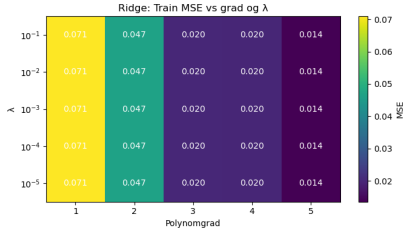



Figure 1: The bias variance tradeoff is her exemplified on the data generated from the runge function, where the OLS model is being used. We can tell that Train MSE keeps going down, whilst the Test MSE starts ioncreasing again at the polynomial degree of 17. We cant tell that this value have the best balance between bias and variance.




(a) Trainingdata




(b) Testdata

Figure 2: Train and test MSE for Ridge regression as a function of polynomial degree (x-axis) and regularization parameter lambda (y-axis). The color scale and cell values show the mean squared error on the test set. The figure illustrates that increasing the polynomial degree reduces the error, while variations in lambda have little effect in this case.

[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%2F978-0-387-84858-7`.

[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).