Analyzing Different Regression and Resampling Methods

Stan Daniels, Francesco Minisini, Teresa Ghirlandi, and Carolina Ceccacci
University of Oslo
Data Analysis and Machine Learning (FYS-STK3155/FYS4155)
(Dated: October 6, 2025)

Regression in machine learning is a fundamental technique for predicting outcomes based on input features. It finds relationships between variables so that predictions on unseen data can be made. A major challenge arises as model complexity increases: low-degree models may underfit, while high-degree polynomial regression can become unstable and overfit the data. In this project, we study the Runge function, a well-known function that highlights the difficulties of high-degree polynomial interpolation. We apply Ordinary Least Squares, Ridge, and Lasso regression, complemented by gradient descent and its variants, including momentum, Adagrad, RMSprop, and ADAM. Resampling techniques such as bootstrap and cross-validation are used to evaluate model generalization and analyze the bias-variance trade-off. The results show that OLS fits become highly unstable for high polynomial degrees, while Ridge and Lasso regularization significantly improve stability and predictive accuracy. Gradient descent methods reproduce the analytical results, though their performance depends strongly on learning-rate strategies. Overall, the study highlights the importance of regularization and resampling for controlling overfitting and improving the reliability of regression models.

I. INTRODUCTION

The aim of this project is to study various regression methods, such as Ordinary Least Squares, Ridge Regression, and Lasso Regression. It focuses on fitting polynomials to a specific one-dimensional function, the Runge function:

$$\frac{1}{1+25x^2}$$

The Runge function shows the difficulties of high-degree polynomial interpolation and this makes it an ideal test case to compare the performances of the different methods. First, an OLS regression analysis is performed, exploring the dependence on the number of data points and the degree of polynomial. The analysis is then extended to Ridge and Lasso regressions, which add a regularization parameter λ. Gradient descent methods are implemented. The analysis starts with the standard gradient descent method, but then, to improve efficiency and convergence, several variants of the gradient descent have been developed, such as momentum, stochastic gradient descent and adaptive methods, including Adagrad, RMSprop, ADAM. The performance of OLS, Ridge and Lasso is then compared with the gradient descent-based optimization methods. In order to evaluate model performance and investigate bias-variance trade-off, resampling techniques such as bootstrap and cross-validation are applied, highlighting how different choices of model complexity and regularization affect the trade-off between bias and variance. These techniques provide insight into the stability of the models and the reliability of their predictions. Overall, this project aims to illustrate the strengths and the limitations of each method.
The structure of this project is as follows:

- Section II "Methods", describes the regression techniques and optimization algorithms, as well as the resampling methods.
- Section III "Results and Discussion", presents the numerical results, compares the performance of the different methods and discusses their implications in terms of bias-variance trade-off
- section IV "Conclusion", summarizes the main results and the insights gained from the methods studied.

II. METHODS

Let $\mathbf{y} \in \mathbb{R}^n$ denote the vector of target values and $\mathbf{X} \in \mathbb{R}^{n\times p}$ the design matrix containing p predictors for n observations. The following linear model is assumed:

$$\mathbf{y} = \tilde{\mathbf{y}} + \boldsymbol{\epsilon}, \quad \text{with} \quad \tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta},$$

where $\tilde{\mathbf{y}}$ represents the predictions of the model, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the vector of unknown coefficients to be estimated and $\boldsymbol{\epsilon}$ is a vector of errors, typically assumed to be independent and identically distributed with zero mean and variance σ^2.

The goal of regression is to find an estimate of the optimal parameter $\boldsymbol{\theta}$ that best explains the observed data according to a chosen criterion.

A detailed description of the methods follows.

A. Ordinary Least Squares

Ordinary Least Squares (OLS) is the classical method for linear regression and it estimates $\boldsymbol{\theta}$ by minimizing the mean squared error. This is the cost function that is going to be optimize:

$$C(\boldsymbol{\theta}) = \frac{1}{n}\left\{(\boldsymbol{y}-\boldsymbol{X\theta})^T(\boldsymbol{y}-\boldsymbol{X\theta})\right\}$$

It means that it's required that the derivative with respect to $\boldsymbol{\theta}$ be set equal to zero:

$$\frac{\partial C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta_j}} = 0 = \mathbf{X}^\top(\mathbf{y}-\mathbf{X\theta})$$

$$\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X\theta}$$

For a full-rank design matrix, this has the closed-form solution:

$$\hat{\boldsymbol{\theta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

OLS provides a simple solution, suitable when the features are few and not highly correlated.

1. *Implementation*

The inputs of the function are a feature matrix X and a vector of targets y. The output is the vector of the parameters $\boldsymbol{\theta}$:

B. Ridge regression

A regularization parameter λ can be introduced by defining a new cost function to be optimized, that is:

$$C(\boldsymbol{\theta}) = \frac{1}{n}||\boldsymbol{y}-\boldsymbol{X\theta}||_2^2 + \lambda||\boldsymbol{\theta}||_2^2$$

where the second term represents an L^2 penalty on the size of the coefficients. This leads to the Ridge regression minimization problem where it is required that $|\boldsymbol{\theta}||_2^2 \leq t$, where t is a finite positive number. One of the main motivations behind Ridge is its ability to resolve the problem of non-invertibility of $\mathbf{X}^\top\mathbf{X}$, which often arises when features are highly correlated. Ridge regression resolves this problem by adding the parameter $\boldsymbol{\lambda}$ to the diagonal of $\mathbf{X}^\top\mathbf{X}$ before inverting it. Taking the derivatives with respect to $\boldsymbol{\theta}$ the optimal parameters are obtained:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{Ridge}} = (\mathbf{X}^\top\mathbf{X} + \boldsymbol{\lambda}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

with $\mathbf{I}$ being a $p \times p$ identity matrix.

1. *Implementation*

C. Lasso regression

Here the regularization term is based on the L_1 norm of the parameters. The cost function is defined as

$$C(\boldsymbol{\theta}) = \frac{1}{n}||\boldsymbol{y}-\boldsymbol{X\theta}||_2^2 + \lambda\,\|\boldsymbol{\theta}\|_1,$$

where $\|\boldsymbol{\theta}\|_1 = \sum_j |\theta_j|$. This formulation leads to the Lasso minimization problem where it is required that $\|\boldsymbol{\theta}\|_1 \leq t$, with t being a finite positive number.

Unlike Ridge regression, taking the derivatives with respect to $\boldsymbol{\theta}$ does not lead to an analytical solution. The equation can however be solved by using standard convex optimization algorithms.

The key feature of Lasso lies in its ability to shrink some estimated coefficients $\hat{\theta}_j$ exactly to zero. When this happens, the corresponding predictor is completely removed from the model. In contrast, Ridge regression never eliminates variables: it only shrinks the coefficients $\hat{\theta}_j$ towards zero but keeps all predictors in the model. Typically, Lasso Regression is preferred when the goal is to simplify the model and improve interpretability, especially when there are a lot of features. On the other hand, Ridge regression is better for handling multicollinearity among features.

1. *Implementation*

D. Gradient descent and its variants

Although OLS and Ridge regression have analytical solutions, such solutions are not always available in general, so a numerical approach is often needed to optimize the same cost function.

Consider the cost function $C(\boldsymbol{\theta})$ that has to be minimized with respect to the parameters $\boldsymbol{\theta}$. A second-order Taylor expansion around a point $\boldsymbol{\theta}_n$ is performed:

$$C(\boldsymbol{\theta}) \approx C(\boldsymbol{\theta}_n) + (\boldsymbol{\theta}-\boldsymbol{\theta}_n)^\top\nabla_{\boldsymbol{\theta}}C(\boldsymbol{\theta}_n) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_n)^\top\mathbf{H}(\boldsymbol{\theta}_n)(\boldsymbol{\theta}-\boldsymbol{\theta}_n),$$

where $\mathbf{H}(\boldsymbol{\theta}_n)$ is the Hessian matrix of second derivatives at $\boldsymbol{\theta}_n$.

Neglecting the second-order term (or assuming it is costly to compute), a first-order approximation gives the update rule in the direction of the steepest descent:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \eta\,\nabla_{\boldsymbol{\theta}}C(\boldsymbol{\theta}_n),$$

where $\eta > 0$ is a learning rate controlling the step size.

This iterative procedure moves the parameters towards the minimum of $C(\boldsymbol{\theta})$. For convex functions such as the mean squared error in linear regression, convergence is guaranteed if η is chosen appropriately.

A limitation of this method is the fixed learning rate η:

- if η is too large, the updates can overshoot the minimum, causing oscillations or divergence
- if η is too small, convergence is very slow

Moreover, for a function with steep directions and flat directions, a single global η may be inappropriate: steep coordinates require a smaller step size to avoid oscillation

and flat coordinates could use a larger step to speed up progress.

In order to mitigate this problem, gradient descent with momentum is introduced: it refers to a method that smoothens the optimization trajectory by adding a term that helps the optimizer remember the past gradients.

Mathematically, let $\boldsymbol{v}_n$ denote the velocity (or accumulated gradient) at iteration n. The update rules for gradient descent with momentum are:

$$\boldsymbol{v}_{n+1} = \gamma\, \boldsymbol{v}_n - \eta\, \nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}_n),$$

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \boldsymbol{v}_{n+1},$$

where $\gamma \in [0, 1)$ is the momentum coefficient, which controls how much of the past gradients are remembered in the current update. A value close to 1 means the optimizer will have more inertia while a value closer to 0 means less reliance on past gradients. This mechanism enables the algorithm to suppress oscillations along steep directions while simultaneously accelerating progress across flatter regions of the cost surface. The result is a convergence process that is both faster and more stable than standard gradient descent.
This leads to more advanced optimization methods which use an adaptive learning rate for each parameter that depends on the history of gradients.
First of all, Adagrad:

The iteration start from an initial guess $\boldsymbol{\theta}^{(0)}$, the parameters are updated iteratively according to the rule:

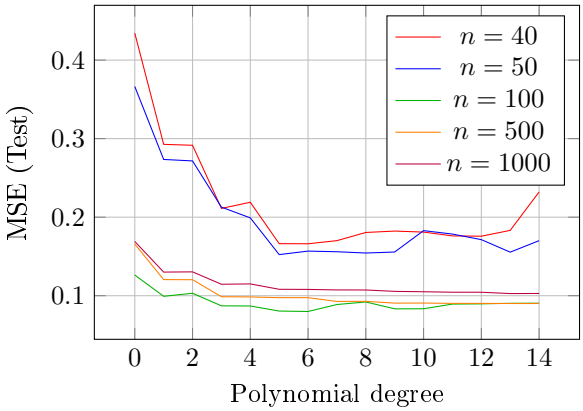
$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} - \eta\, \nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}^{(n)}),$$

where $\eta > 0$ is the learning rate and $\nabla_{\boldsymbol{\theta}} C(\boldsymbol{\theta}^{(t)})$ is the gradient of the cost function with respect to $\boldsymbol{\theta}$ at iteration t.

III. RESULTS

Insert figures, tables, and discussions.

Test MSE vs Polynomial degree



IV. DISCUSSION AND CONCLUSION

Summarize findings, bias-variance trade-off, comparisons.