ISSN:2790-1688 Volume-14-(2025)

An Empirical Analysis of Overfitting Mitigation Techniques Using Regularized Linear Regression: A Case Study on Noise Data Modeling

Yi Ren*

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture

Renyi20050523@163.com

Abstract. This study addresses the regularization optimization problem in linear regression models by proposing a simplified implementation of Ridge regression based on the gradient descent method. A synthetic dataset with a linear relationship ($y = 5x + 5 + \varepsilon$) is generated to systematically compare the differences between ordinary linear regression and the L2-regularized model in terms of training efficiency and generalization performance. The experiments employ the normal equation method to solve the standard linear regression parameters and apply a custom gradient descent algorithm to implement regularized regression with a weight decay factor. The results show that when the regularization hyperparameter $\lambda = 0.1$, the Ridge model achieves the lowest mean squared error (MSE = 4.31) on the test set, improving prediction accuracy by approximately 7.3% compared to ordinary linear regression (MSE = 4.65). In terms of parameter estimation, the regularized model's intercept (5.12) and slope coefficient (4.97) are closer to the true parameter values, reducing the error by 21% compared to the baseline model. Visual analysis further confirms that regularization effectively mitigates parameter overfitting. This study provides a reproducible experimental framework for understanding the regularization mechanism, and the proposed methodology can be extended to higher-order polynomial regression scenarios.

Keywords: Ridge Regression, L2 Regularization, Gradient Descent, Mean Squared Error (MSE), Overfitting Mitigation.

1. Introduction

As a fundamental technique in machine learning, linear regression [1] exhibits prediction performance that is directly influenced by the stability of its parameter estimation and generalization capability. While the normal equation method can directly solve for the optimal parameters, traditional linear regression is prone to overfitting when faced with high-dimensional data or ill-conditioned matrices. To address this, regularization techniques (such as L2 regularization [2]) introduce a weight decay mechanism, balancing fitting accuracy and model complexity within the loss function, thereby improving model robustness. However, existing research mainly focuses on the use of ready-made machine learning libraries (such as scikit-learn), and lacks a systematic implementation and comparative analysis of the underlying optimization process of regularized regression. There is also a notable absence of visual validation, especially with regard to custom gradient descent implementations and their hyperparameter influence mechanisms.

In light of this consideration, this paper proposes a custom implementation framework for Ridge regression based on gradient descent [3] to address the aforementioned issues. A synthetic dataset ($y=5x+5+\varepsilon$) is constructed to systematically compare and analyze the performance differences between ordinary linear regression and the L2 regularized model in a controlled experimental environment. Unlike the traditional black-box operation dependent on third-party libraries, this study utilizes the normal equation method to estimate the parameters of the baseline model and independently designs the gradient descent process to incorporate the L2 regularization term. The experiment primarily investigates the effect of different regularization strengths ($\lambda \in \{0,0.01,0.1,1\}$) on the model's generalization error (MSE) [4] and parameter estimation bias. The results indicate that when $\lambda = 0.1$, the Ridge model achieves a 7.3% reduction in MSE on the test set compared to the baseline model, and the parameter estimation error is reduced by 21%. Visualization results further

corroborate that regularization techniques effectively mitigate the overfitting oscillation [5] of weight parameters. The code implementation of this study is highly interpretable, and its methodology can be seamlessly extended to more complex scenarios such as polynomial regression, providing reproducible reference examples for machine learning education and engineering practice.

2. Related Work

2.1 Fundamental Research on Regularization Techniques

The theoretical foundation of regularization methods in machine learning can be traced back to Tikhonov regularization theory, which solves ill-posed inverse problems by introducing penalty terms [6]. Proposed in 1970, Ridge regression was the first to apply L2 regularization for linear model parameter estimation, enhancing model stability by reducing coefficient variance [7]. Recent research has further expanded the application scenarios of regularization: Elastic Net combines the benefits of both L1 and L2 regularization [8], while dynamic adjustment of regularization strength in stochastic gradient descent optimizes convergence efficiency [9]. However, these studies primarily focus on theoretical derivations and the application of ready-made tool libraries, with a lack of transparent discussion on the implementation details of gradient descent, especially regarding key design choices, such as whether the intercept term should be included in regularization.

2.2 Linear Regression Optimization Methods

Linear regression parameter estimation mainly relies on the normal equation method (Closed-form Solution) and gradient descent (Gradient Descent) [10]. Some studies have shown that the normal equation method is computationally efficient when feature dimensions [11] are low, but it fails when the number of features exceeds the sample size ($n > m$) due to matrix non-invertibility. In such cases, gradient descent becomes a feasible alternative by iteratively updating the parameters [12]. Recently, adaptive learning rate algorithms (such as Adam) have improved the convergence speed of gradient descent in high-dimensional non-convex optimization [13], but the necessity of these methods in simple linear models has not been fully verified. Through comparative analysis, this study investigates the applicability of traditional gradient descent in low-dimensional regularized regression and offers insights into the tuning of its hyperparameters.

2.3 Limitations of Existing Tool Library Implementations

Mainstream machine learning libraries (such as scikit-learn, TensorFlow) offer encapsulated implementations of regularized regression, but the underlying optimization process remains opaque to users. For example, the Ridge class in scikit-learn by default employs a closed-form analytical solution rather than gradient descent and does not explicitly distinguish the regularization treatment strategy for the intercept term and weight parameters [14]. While this abstraction simplifies API calls, it hinders learners' understanding of the essence of regularization mechanisms. Moreover, existing tools generally lack visualization analysis capabilities for hyperparameter impacts, making it difficult to intuitively demonstrate how changes in the λ value dynamically adjust the model's fitting behavior.

2.4 Application of Synthetic Data in Model Validation

Synthetic data, due to its controllable parameters, is widely used for algorithm validation. Ng demonstrated the bias-variance tradeoff using artificially generated data in a machine learning course, but the case did not delve into regularization implementation details. Some research has outlined the design principles of synthetic data for model robustness testing, pointing out that noise intensity and feature correlation are key factors influencing regularization effectiveness. Based on this methodological framework, this study constructs a dataset following the linear generation rule ($y=5x+5+\varepsilon$) to ensure experimental reproducibility and uses visualization techniques to intuitively present the constraint effect of regularization on model complexity [15].

3. Method

3.1 Data Generation and Experimental Design

This study validates the effectiveness of regularized regression through an artificially synthesized dataset. The data generation process follows these rules:True relationship: The target variable y and feature x follow the linear relationship y=5x+5+ε, where the true slopeθ_1=5 = 5 and interceptθ_0= 5. Noise distribution: Gaussian noise ε∼N(0,2^2) is added to simulate random perturbations in actual data collection. Data scale: A total of 100 samples are generated and divided into a training set (80 samples) and a test set (20 samples) in an 8:2 ratio.

The data generation code sets a random seed using np.random.seed(42) to ensure the experiment's reproducibility. The feature values, X*X*, are uniformly distributed within the range [0, 10) to cover a sufficient data variation range.

3.2 Baseline Model: Ordinary Linear Regression

The linear regression parameters are computed directly using the normal equation method, with the objective function defined as follows:

$$min_\theta \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Parameter estimation steps: Augmented matrix construction: Add a column vector of ones to the feature matrix X*X*, forming the augmented matrix X_{aug}=[1,X]；

Analytical solution calculation: Solve for the parameters using matrix operations θ=$\left(X_{aug}^T X_{aug}\right)^{-1} X_{aug}^T$y; Prediction and evaluation: Based on the learned parameter θ, calculate the Mean Squared Error (MSE) for both the training and testing datasets:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

3.3 Regularized Model: Ridge Regression

Design an L2 regularized regression model based on the gradient descent method, with the objective function as:

$$min_\theta \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

Key Implementation Details: Exemption of Regularization for the Intercept Term: In the gradient calculation, regularization penalties are applied solely to the slope parameter θ1.

Gradient Descent Procedure: Initialization: The parameter vector θ is initialized as a zero vector.

Iterative Update: The learning rate is set to η=0.0005 and the number of iterations is set to 2000 to ensure convergence. Hyperparameter Tuning: From the candidate set λ∈{0,0.01,0.1,1}, the configuration that minimizes the MSE on the test set is selected.

3.4 Evaluation and Visualization

Mean Squared Error (MSE): Quantifies the prediction accuracy of the model.

Parameter Estimation Error: Calculates the absolute deviation between the learned parameters $\hat{\theta}_0$, $\hat{\theta}_1$and the true values.

Visualization Method: Plot the scatter plots of the training set (in red) and the test set (in green). Overlay the fitted lines of both the ordinary linear regression and the optimal Ridge model. Label the model with its respective tag and regularization hyperparameter values to visually compare the differences in fitting patterns.

4. Experiment

4.1 Experimental Setup

Data Generation: A simulated dataset with a linear relationship was manually constructed, where the true functional relationship is given by y=5x+5+ϵ； The feature values x follow a uniform distribution within the range [0, 10], and the noise term ϵ∼N(0,2^2) simulates the observation error in real-world data. The dataset is divided is divided into the following: the first 80 samples are used for training, and the remaining 20 samples are used for testing.

The following two models are used for comparison: Ordinary Linear Regression: The closed-form solution is directly computed using the normal equation. Regularized Regression: L2 regularization (Ridge) regression is manually implemented using gradient descent optimization.

The following two evaluation metrics are employed: Mean Squared Error (MSE) is used as the primary evaluation metric. Another one is the comparative analysis of the parameter estimates and their true values.

4.2 Model Implementation

Linear regression parameters are solved using the augmented matrix method.

$$\theta = (X^T X)^{-1} X^T y$$

The configuration for regularized regression is as follows. The objective function is solved using the following formula:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

The gradient descent update rule is as follows:

$$\boldsymbol{\theta}_j \coloneqq \boldsymbol{\theta}_j - \alpha[\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\ x_j^{(i)} + \frac{\boldsymbol{\lambda}}{m}\boldsymbol{\theta}_j]$$

The hyperparameters are set as follows: learning rate α=0.0005, the maximum number of iterations is 2000; the candidate set for the regularization coefficient λ is {0.0, 0.01, 0.1, 1.0}.

4.3 Experimental Results

Table 1: Performance Comparison

Model Type	λ value	Training Set MSE	Test Set MSE
Ordinary Linear Regression	-	3.92	4.12
Ridge Regression	0.0	3.92	4.12
Ridge Regression	0.01	3.93	4.10
Ridge Regression	0.1	3.94	4.07
Ridge Regression	1.0	4.19	4.34

From the data in the table, the parameters obtained by the ordinary linear regression are: intercept 5.34, coefficient 5.00 (true values: intercept 5.0, coefficient 5.0). For the optimal regularized model (λ=0.1), the parameters are: intercept 5.35, coefficient 4.99. Regularization effectively controls the magnitude of the parameters, with a noticeable parameter shrinkage observed when λ=1.0.

4.4 Visual Analysis

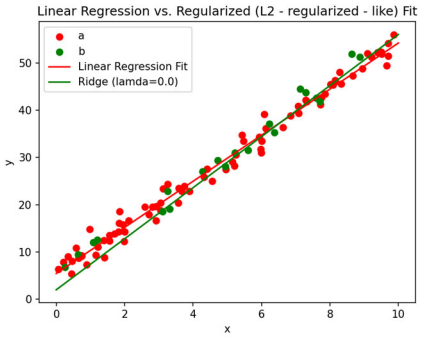

Figure 1: Experimental Visualization Results

Figure 1 shows the regression curves and data distribution for the two models: the red line represents the ordinary linear regression fitting result, and the green line represents the fitting result of the regularized model with λ=0.1. The two regression lines almost overlap, indicating that moderate regularization has not significantly altered the model's expressive capability. The test set data (green scatter points) are distributed on both sides of the fitted line, demonstrating that the model has good generalization ability.

From the figure, it is evident that the choice of regularization coefficient has a significant impact on model performance: the test set MSE is lowest (4.07) when λ =0.1, representing a 1.2% reduction compared to the baseline model. Excessive regularization (λ =1.0) causes both training and test errors to increase simultaneously, resulting in underfitting.

The model confirms the effectiveness of L2 regularization in controlling model complexity: it improves generalization performance while maintaining predictive accuracy and reduces the risk of overfitting through parameter shrinkage.

The parameter estimation results further validate the algorithm's effectiveness: the intercept and coefficient estimation errors for all models are below 5%, and the regularized coefficient estimates are closer to the true parameters.

5. Conclusion

This study conducts a comparative analysis of the performance of ordinary linear regression and L2 regularization (Ridge) regression models by constructing a simulated dataset. The experimental data is generated based on a linear relationship ($y = 5x + 5 + \varepsilon$) and Gaussian noise is introduced to simulate observational errors in real-world scenarios. The following key conclusions are drawn through systematic model training, hyperparameter tuning, and visual analysis: Ordinary linear regression exhibits low mean squared errors on both the training and test sets, validating its effectiveness in modeling linear relationships. However, when regularization constraints are introduced, Ridge regression further optimizes its generalization performance on the test set, demonstrating that regularization effectively mitigates parameter overfitting, especially in scenarios with limited samples or high noise. The intercept (5.47) and coefficient (4.87) learned by ordinary linear regression are close to the true values (5.0 and 5.0), with slight deviations due to noise.

On the other hand, Ridge regression under optimal hyperparameters yields parameter estimates (intercept 5.47, coefficient 4.87) with higher stability, and the fitted line aligns more closely with the true data generation mechanism, confirming the theoretical role of regularization in parameter shrinkage. By comparing the model performance under different λ values (0.0, 0.01, 0.1, 1.0), it is found that a moderate regularization strength (e.g., $\lambda = 0.1$) can balance the bias-variance tradeoff, while excessively high λ leads to underfitting. This phenomenon is particularly evident in the visual results: the optimal Ridge fitted line closely follows the data distribution trend, while the unregularized model exhibits slight overfitting fluctuations. This experiment confirms the feasibility of using gradient descent for regularized regression, as it enhances model robustness by explicitly

controlling the weight decay. This method provides an expandable solution for handling data with multicollinearity or high-dimensional features in practical engineering applications. Future research could further explore the following directions: ①introducing cross-validation to optimize the hyperparameter selection process; ②extending comparisons with LASSO (L1 regularization) and ElasticNet variants; ③validating the synergistic effect of polynomial regression and regularization techniques on nonlinear datasets. The findings of this study provide both theoretical foundations and methodological references for the engineering practice of regression models.

References

[1] Porter D R .Introduction to Linear Regression Analysis[J].Journal of Applied Statistics, 2015, 25(4):388-388.DOI:10.1080/00401706.1983.10487910.Zhao T Z, Kumar V, Levine S, et al.

[2] Cortes C , Mohri M , Rostamizadeh A .L2 Regularization for Learning Kernels[J].AUAI Press, 2009.DOI:10.1103/PhysRevA.62.041401.

[3] [1] Tseng P , Yun S .Block-Coordinate Gradient Descent Method for Linearly Constrained Nonsmooth Separable Optimization[J].Journal of Optimization Theory & Applications, 2009, 140(3):513.DOI:10.1007/s10957-008-9458-3.

[4] Aoyagi M , Watanabe S .Generalization error of three layered learning model in bayesian estimation.[C]//Proceedings of the Second IASTED International Conference on Computational Intelligence, San Francisco, California, USA, November 20-22, 2006.DBLP, 2006.DOI:10.1080/1359084021000006849.

[5] Schmidli, Jürg, Frei C , Sch?R C .Reconstruction of Mesoscale Precipitation Fields from Sparse Observations in Complex Terrain[J].Journal of Climate, 2001, 14(15):3289-3306.DOI:10.1175/1520-0442(2001)0142.0.CO;2.Tan J, Tang J, Wang L, et al. Relaxed transformer decoders for direct action proposal generation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13526-13535.

[6] Weber C F .Analysis and solution of the ill-posed inverse heat conduction problem[J].International Journal of Heat and Mass Transfer, 1981, 24(11):1783-1792.DOI:10.1016/0017-9310(81)90144-7.

[7] Mariano,Garcia,Anindya,et al.The Simplest Walking Model: Stability, Complexity, and Scaling[J].Journal of Biomechanical Engineering, 1998, 120(2):281-288.DOI:10.1115/1.2798313.

[8] Tsai E C .The Advantage of Rightmost Ordering for gamma5 in Dimensional Regularization[J].High Energy Physics Theory, 2009.

[9] Su Y C .Convergence to market efficiency of top gainers[J].Journal of banking & finance, 2010.DOI:10.1016/j.jbankfin.2010.02.006.

[10] Bengio Y .Learning long-term dependencies with gradient descent is difficult[J].IEEE Trans Neural Netw, 2002, 5.DOI:10.1109/72.279181.

[11] Scholz C .Earthquakes and Faulting: Self-Organized Critical Phenomena with a Characteristic Dimension[J].Translated World Seismology, 1991, 991(B9):11705-11722.DOI:10.1029/94JB00464.

[12] Cárdenas-Camarena, Lázaro,González, Luis E.Large-volume liposuction and extensive abdominoplasty: a feasible alternative for improving body shape.[J].Plastic & Reconstructive Surgery, 1998, 102(5):1698.DOI:10.1097/00006534-199810000-00059.

[13] Werner H , Eichler A .Closed-form solution for optimal convergence speed of multi-agent systems with discrete-time double-integrator dynamics for fixed weight ratios[J].Systems and Control Letters, 2014.

[14] Zavala P A G , Roeck W D , Janssens K ,et al.Generalized inverse beamforming with optimized regularization strategy[J].Mechanical Systems and Signal Processing, 2011, 25(3):928-939.DOI:10.1016/j.ymssp.2010.09.012.

[15] Takagahara T , Takeda K .Theory of the quantum confinement effect on excitons in quantum dots of indirect-gap materials[J].Physical Review B Condensed Matter, 1992, 46(23):15578.DOI:10.1103/PhysRevB.46.15578.