Investigating Model Complexity, Bias-Variance Trade-off, and Regularization Techniques in Linear Regression

COMP551 A2 Report

Nicholas Milin (261 106 314) – Diana Covaci (261 086 280) – Viktor Allais (261 148 866)

1 Abstract

This project investigates the performance of linear regression under varying model complexities, examines the bias-variance trade-off, L1 and L2 regularization, regularization with cross-validation, and the effects of L1 and L2 regularization on loss using two synthetic data sets.

(1) We first analyzed how the number of Gaussian non-linear bases influences the performance on training and validation sets, identifying an optimal configuration with 11 bases that best generalizes to unseen data. (2) We explored the bias-variance trade-off, demonstrating that moderate model complexity achieves the best generalization performance by balancing high bias from underfitting with high variance from overfitting. (3) We then investigated the combined effects of L1 and L2 regularization through the use of Elastic Net, finding optimal model performance when $\lambda_1 = 0$ and $\lambda_2 = 0.01$. (4) Finally, we analyzed the effects of L1 and L2 regularization, observing that both methods improve gradient descent convergence, while L1 regularization further encourages sparsity by driving weights to zero. Overall, our experiments demonstrate the importance of careful model selection for achieving an optimal bias-variance balance and highlight the value of thorough regularization testing with cross-validation to achieve strong predictive generalization.

2 Results

2.1 Linear Regression with Non-Linear Basis Functions

We first generated 100 uniformly spaced data points from the non-linear function

$$y(x) = (\log x + 1)\cos(x) + \sin(2x) + \epsilon$$

over the range [0,10]. The noise (ϵ) was sampled randomly and independently from a unit Gaussian distribution. We then fit the synthetic data using linear regression models with 0 to 45 Gaussian basis functions. Figure 1 and Figure 2 illustrate the generated data and the Gaussian bases functions plot ($D = 45$). Figure 3 visualizes the model fits across a varying number of basis functions.

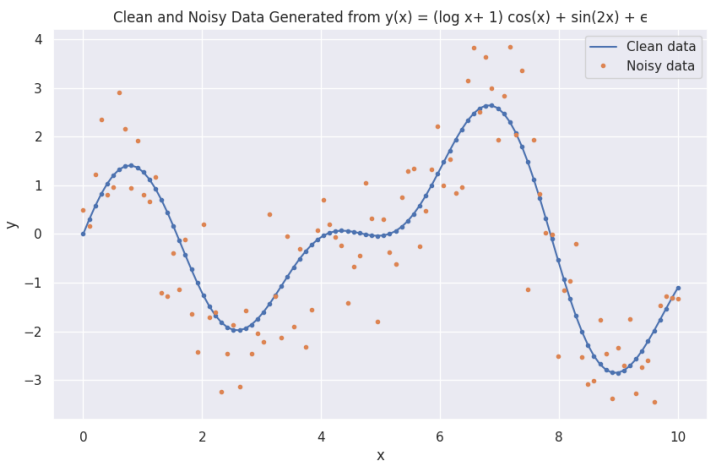

Fig 1. Clean and Generated Noisy Data

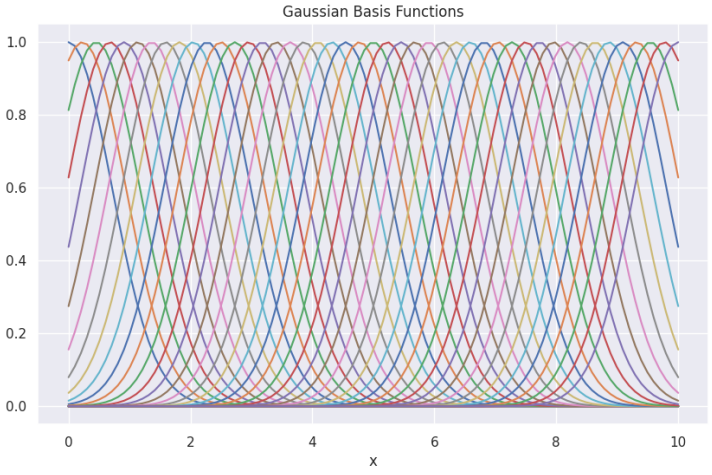

Fig 2. Gaussian Basis Functions (D=45)

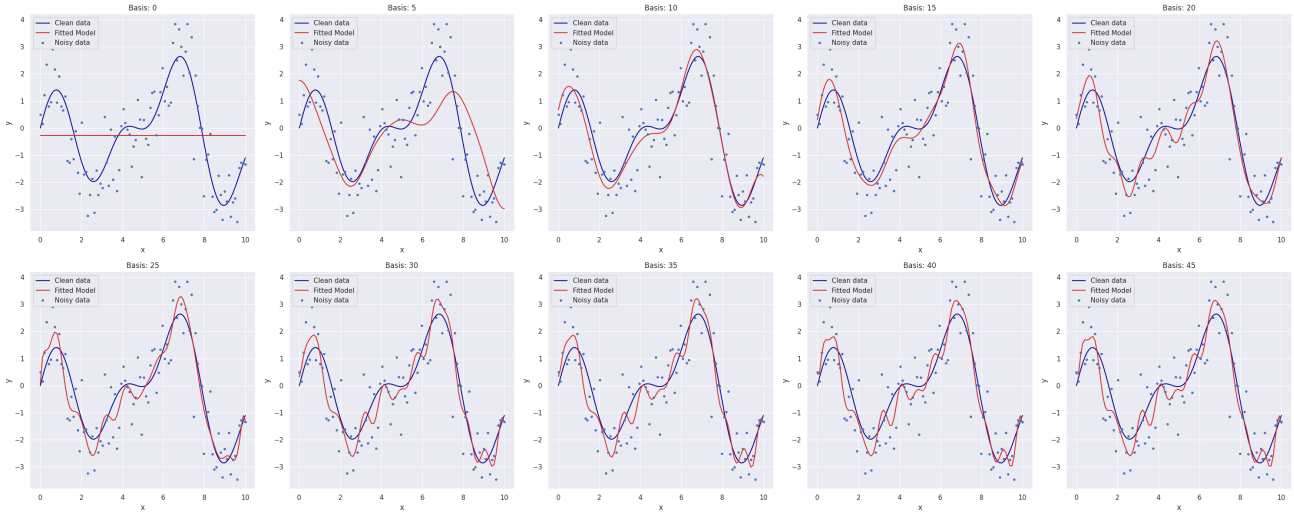

Fig 3. Fitted Models using Varying Number of Gaussian Bases

Table 1 summarises the closed-form linear regression training and validation sum of squared errors (SSE) on an 80/20 train/validation data split. Initially, at a low number of bases, the model exhibits high training (263.76) and validation (53.60) SSEs due to high bias and low complexity, causing it to underfit. As we increase the number of non-linear bases, the model is able to better capture the non-linear relationship, reducing both training and validation SSEs. However, past a certain point ($D \approx 20$), training SSE continues to drop while validation SSE begins to increase, indicating overfitting to the noise in the training data and increased variance. The validation set identifies the model configuration that achieves the optimal bias-variance trade-off and generalizes best to unseen data from the same distribution. From Table 2, the minimum validation SSE (12.77) occurs when the number of non-linear basis functions is 11. This complexity level is selected as it represents the point where the model maintains the strongest generalization performance.

Num. Bases	Train SSE	Validation SSE
0	263.76	53.60
5	159.14	36.84
10	62.08	14.16
15	57.31	17.08
20	51.86	20.93
25	49.35	39.12
30	48.45	28.58
35	48.45	28.03
40	48.45	27.85
45	48.46	27.77

Table 1: Train and Validation SSE Over Varying Number of Gaussian Bases D=(0,5,...,45)

Num. Bases	Train SSE	Validation SSE
5	159.14	36.84
6	131.62	38.48
7	79.25	23.37
8	68.46	13.73
9	67.83	12.90
10	62.08	14.16
11	61.96	12.77
12	61.94	14.00
13	60.60	15.16
14	58.31	15.88
15	57.31	17.08

Table 2: Train and Validation SSE Over Varying Number of Gaussian Bases M=(5,6,...,15)

2.2 Bias-Variance Tradeoff with Multiple Fits

To illustrate the bias-variance tradeoff, the fitting process detailed in Section 2.1 was repeated ten times for each number of Gaussian basis functions, with new data sampled in each repetition. This approach allowed us to examine how model complexity and data variation jointly affect the bias and variance of the fitted models. The resulting plots of multiple fits (Figure 4) show that with very few basis functions, the fits are smooth and fail to capture the underlying trend of the true function, indicating high bias and underfitting. As the number of basis functions increases, the fits begin to follow the true curve more closely, and the average of the fitted models (in red) aligns more accurately with the ground truth (in blue), demonstrating reduced bias. However, when too many basis functions are used, the individual fits become highly variable, with large fluctuations between repetitions, reflecting increased variance and the onset of overfitting.

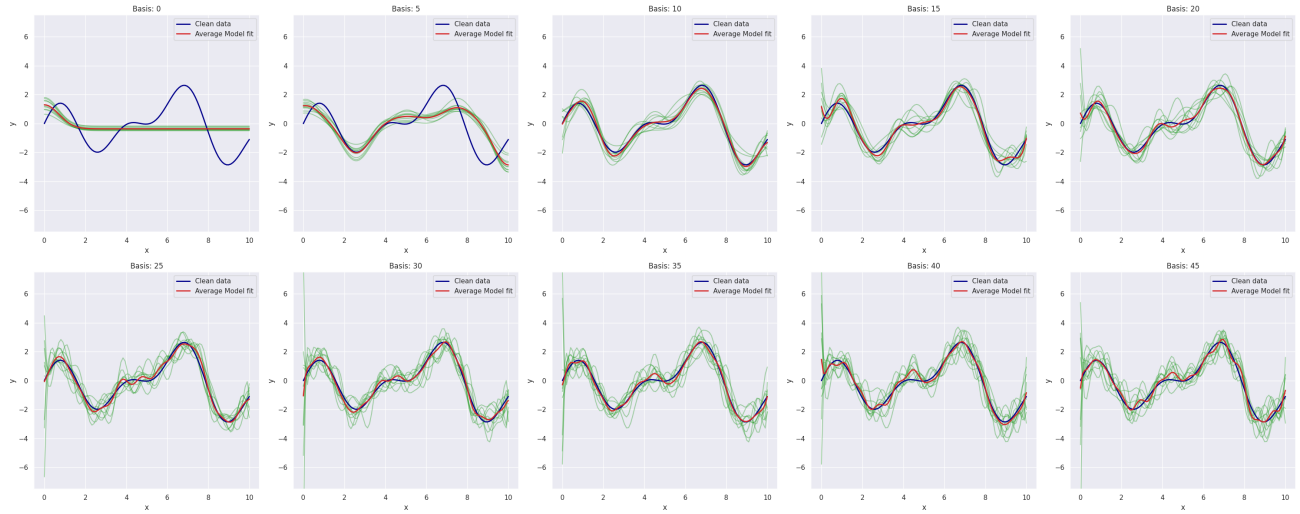

Fig 4. Bias and Variance Across Multiple Fits on Varying Number of Gaussian Bases

The graph in Figure 5 shows the average training and validation errors, supporting earlier observations. This visual further highlights the tradeoff and illustrates the existence of a middle ground between bias and variance. Both errors decrease initially as the model gains flexibility and learns the underlying data patterns. However, beyond 20-25 basis functions, the validation error rises sharply while the training error continues to decrease. This turning point represents the optimal balance in model fitting, where the model is complex enough to capture the true function but not so flexible that it overfits noise in the data. The results demonstrate the expected bias-variance tradeoff: simpler models suffer from high bias, overly complex models suffer from high variance, and an intermediate level of model complexity achieves the best generalization performance, effectively balancing bias and variance.

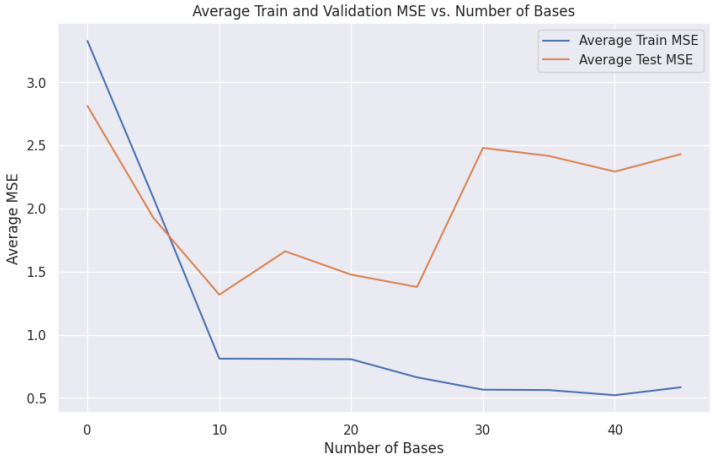

Fig 5. Average MSE for Train and Validation Sets Across Varying Number of Gaussian Bases

2.3 Regularization with Cross-Validation

Using the same data from Section 2.1, we next investigated Elastic Net regularization with k-fold cross-validation and 11 non-linear bases. We used gradient descent with a learning rate of 0.5 and 10,000 epochs to implement the regularization. The two hyperparameters, λ_1 and λ_2, controlling the strengths of L1 and L2 regularization terms, were tuned through a 10-fold cross-validation grid search across 10 different values in the range [0, 1].

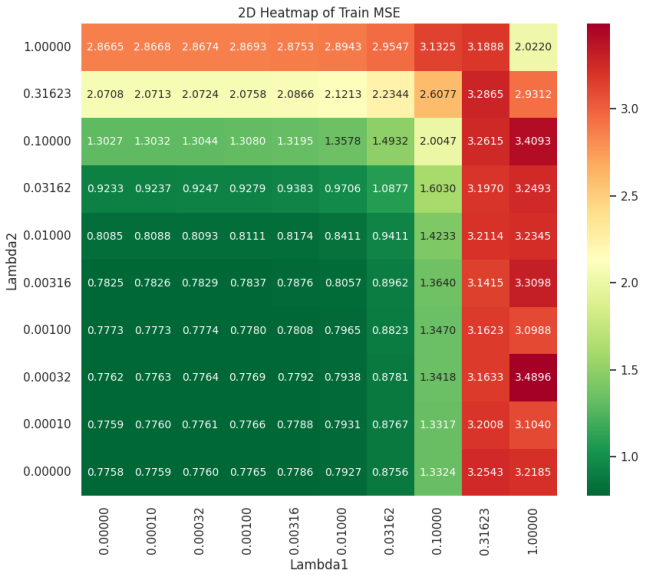

Fig 6. 2D Heatmap of Train MSE

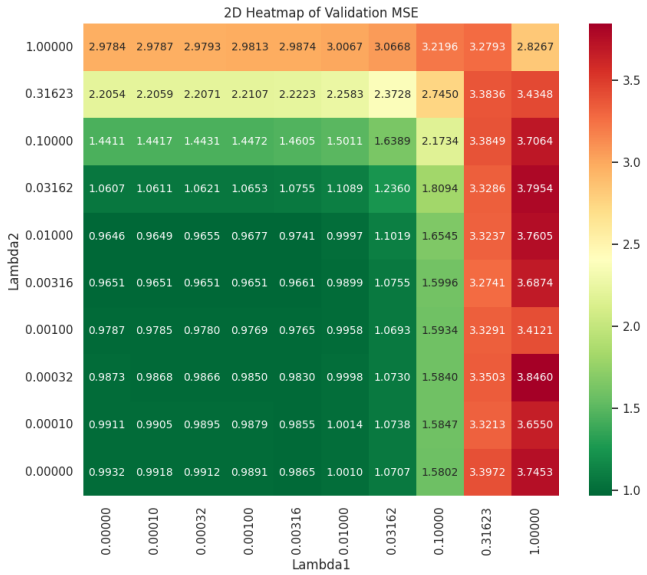

Fig 7. 2D Heatmap of Validation MSE

Figures 6 and 7 show the training and validation MSE heatmaps. The minimum training MSE (0.7758) occurs when $\lambda_1 = 0$ and $\lambda_2 = 0$. However, the minimum validation MSE (0.9646) occurs at $\lambda_1 = 0$ and $\lambda_2 = 0.01$, indicating that mild L2 regularization improves generalization. When considered independently, optimal L1 regularization achieved a minimum MSE of 0.9864 ($\lambda_1 = 0.0032$), while L2 regularization achieved a minimum MSE of 0.9645 ($\lambda_2 = 0.0032$).

Figure 8 presents the bias-variance decomposition over a 10-fold cross-validation. The plots reveal that increasing λ_1 generally raises the test MSE, as stronger L1 regularization induces sparsity and underfitting. In contrast, a small λ_2 slightly increases bias while reducing variance, resulting in improved overall generalization.

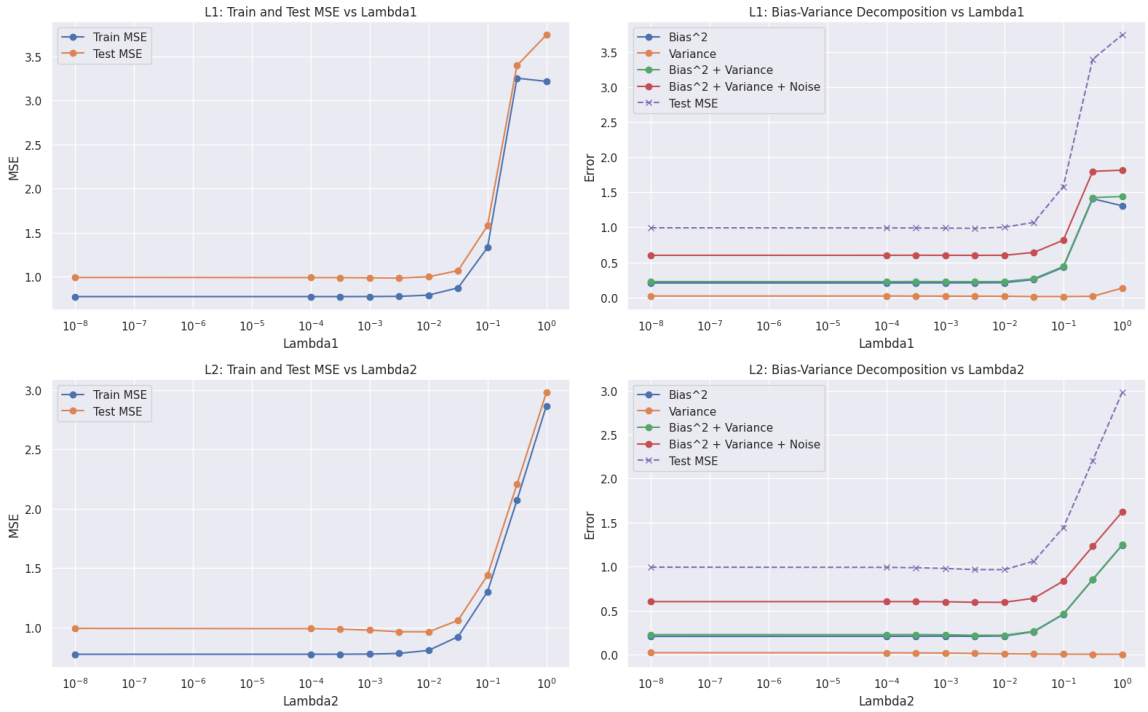

Fig 8. Bias-Varaince Decomposititon for L1 and L2 regularization

2.4 Effect of L1 and L2 Regularization on loss

To further demonstrate the effects of L1 and L2 regularization on loss, we generated synthetic data following a linear relationship with Gaussian noise, defined as

$$y = -3x + 8 + 2\epsilon$$

where x is uniformly distributed between 0 and 10, and ϵ is Gaussian noise with a mean of 0 and a variance of 1. Figure 9 illustrates the generated synthetic data.

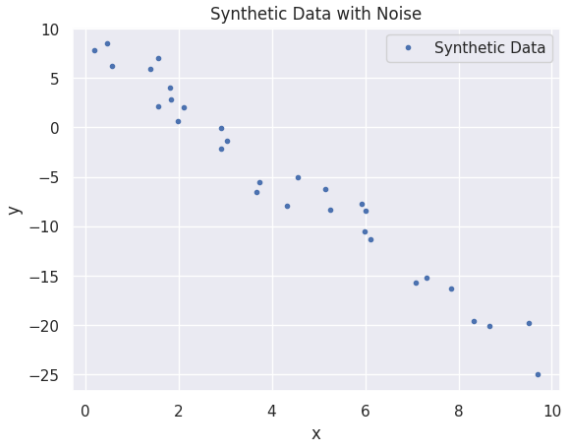

Fig 9. Synthetically Generated Noisy Data

We then implemented L1 and L2 regularization using gradient descent, with a learning rate of 0.01 and a maximum of 50,000 iterations. For both L1 and L2 regularization, we set the regularization parameter to $\lambda = 0.5$, representing moderate regularization. Figures 10 and 11 show the model predictions obtained using L1 and L2 regularization, respectively.

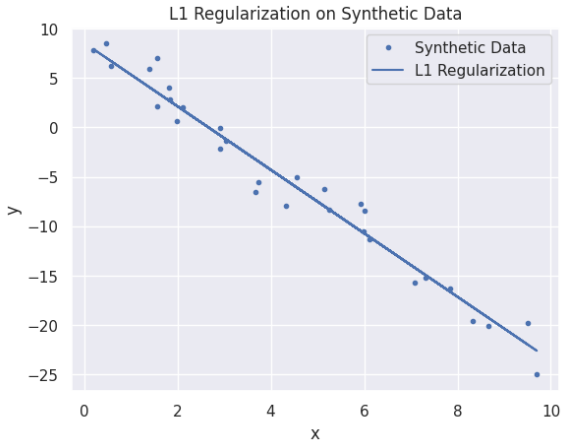

Fig 10. Synthetic Data Fitted by Linear Regression with L1 Regularization

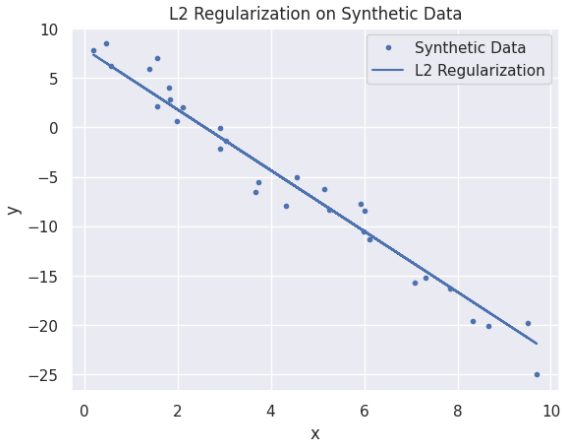

Fig 11. Synthetic Data Fitted by Linear Regression with L2 Regularization

From the plots above, moderate L1 and L2 regularization perform very well, accurately capturing the linear relationship. This is expected because we are working with a simple linear regression model, where regularization has less impact than it would on an overfitted model. Next, we plot the loss contours of L1 and L2 regularization at regularization strengths $\lambda = [0, 5, 10, 30]$. We note that in order to illustrate sparsity, Figure 12 penalizes both weights, while Figure 13 does not penalize the bias (in accordance with the assignment).

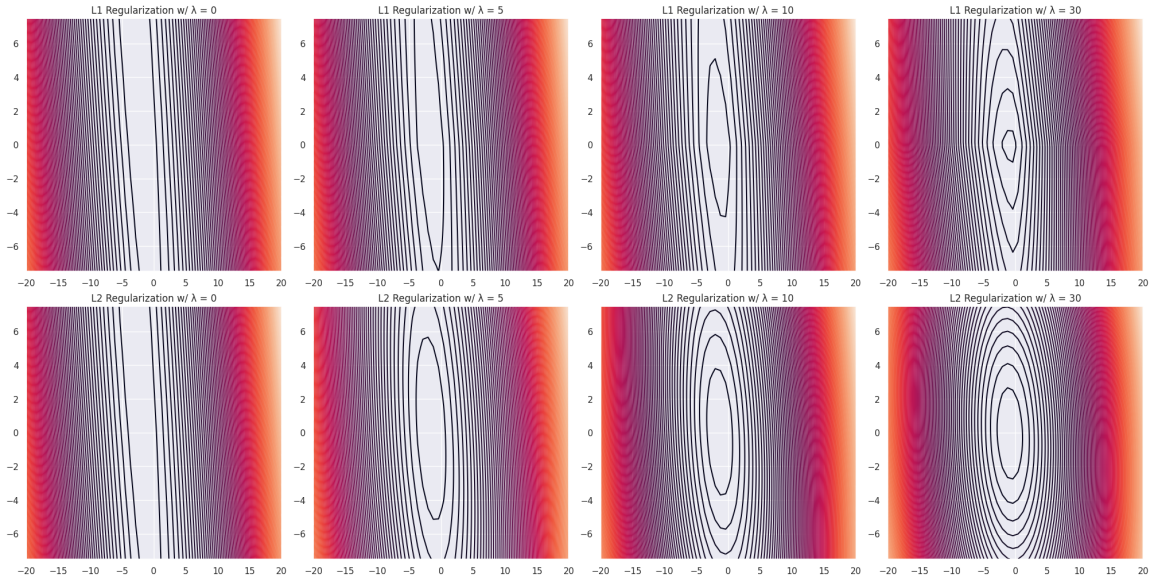

Fig 12. Isocontours of L1 and L2 Regularization Penalizing Both Weights

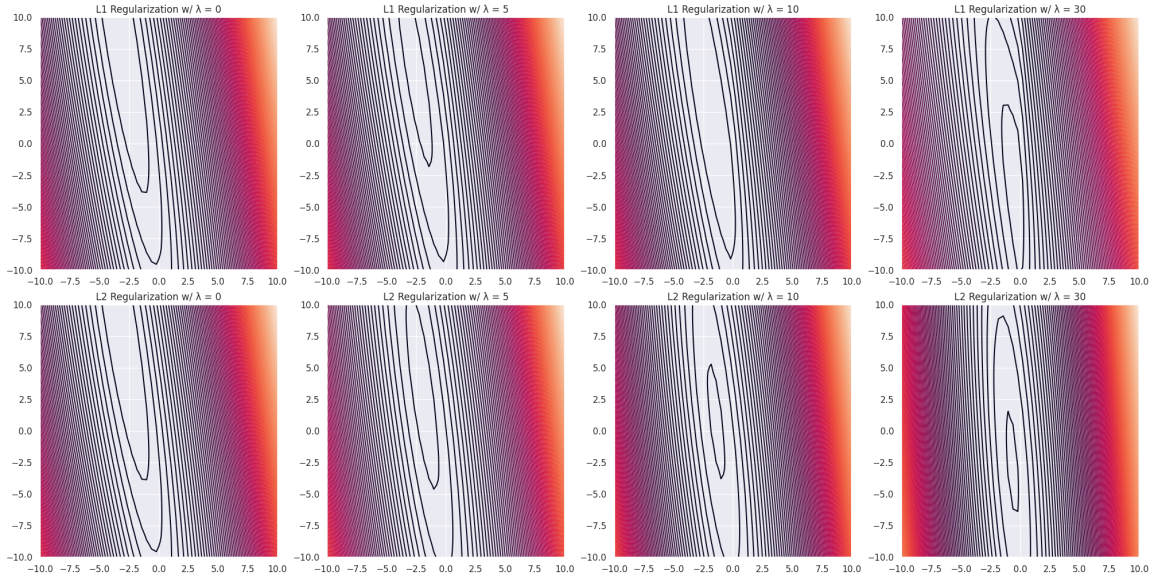

Fig 13. Isocontours of L1 and L2 Regularization Not Penalizing the Bias

As illustrated in the figures above, L1 regularization exhibits its characteristic sharp corners and diamond-shaped contours, while L2 regularization displays circular isocontours. L1 regularization's diamond-shaped contours encourage sparsity; when minimizing the total cost, the optimal solution is more likely to lie on an edge or corner of the diamond, where one of the weights is zero. In contrast, L2 regularization has circular isocontours, which still penalize large weights but are less likely to drive any weights to exactly zero. Next, we plot the trajectory of the gradient descent optimizer on the contour plots in which we don't penalize bias. A learning rate of 0.01 was used, and the results are shown for 225 iterations.

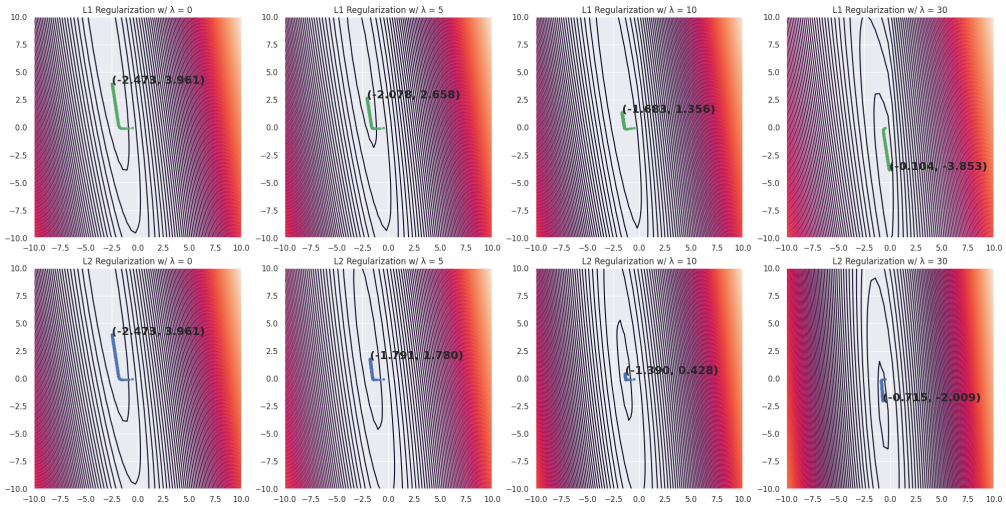

Fig 14. Path of Gradient Descent on the Isocontours of L1 and L2 Regularization

As shown in Figure 14, L1 and L2 regularization both aid the convergence of gradient descent. From the contour plots of L1 and L2 regularization, we observe that increasing the regularization strength drives the optimal solution towards smaller values, while the bias fluctuates since it is not penalized. Furthermore, we note that L1 regularization pushes the penalized weight to almost 0 (-0.104) with $\lambda = 30$, while L2 regularization keeps the weight small (-0.715) but non-zero. Thus, we've shown that L1 regularization encourage sparsity.

3 Discussion and Conclusion

(1) We first showed that increasing the number of non-linear basis functions improves model accuracy until a certain point, after which the model begins to overfit to noise. For the function tested, the optimal complexity of 11 bases achieved a minimal validation error. (2) We next highlighted the bias-variance tradeoff, demonstrating that simple models underfit with high bias, complex models overfit with high variance, and an intermediate level of complexity achieves the best generalization performance. (3) We then illustrated that L2 regularization with a small coefficient ($\lambda_2 = 0.01$) improved validation performance, reducing variance without over-constraining the model. L1 regularization proved less beneficial due to the high correlation between the data. Elastic Net regression allowed for the combination of L1 and L2; however, the best model relied solely on L2 regularization in this case. (4) Finally, our analysis highlighted that L1 and L2 regularization both helped with the convergence of gradient descent, and contrasted the differences between the two regularization methods, specifically that L1 regularization induces sparsity in models.

Some potential directions for future work include:

- Altering the basis functions by using sigmoid or polynomial-based functions. Optimizing the basis functions' placement and scale could yield better fits and a more efficient model.
- Investigating the effects of varying the noise distribution would assess how well the current model selection and regularization generalize to different noise distributions.
- Comparing closed-form L2 regression solutions with gradient-based approaches to examine computational cost and efficiency.

Overall, our results demonstrate that achieving strong predictive performance in linear regression requires thorough optimization of model complexity and regularization to effectively balance bias and variance.

4 Statement of Contributions

Diana Covaci contributed to the implementation and analysis of Task 2. Nicholas Milin contributed to the implementation and analysis of Tasks 1 and 3. Viktor Allais contributed to the implementation of Task 4. Diana Covaci, Nicholas Milin, and Viktor Allais wrote the report. Diana Covaci converted the write-up to LaTeX.