# Report 2: Regularization and Model Evaluation

Student Daria Goptsii (261275056), Giane Mayumi Galhard (51261276747), Yixuan Qin (261010963)
Course COMP 551

## Abstract

This project implements linear regression using synthetic data, focusing on non-linear basis functions, the bias-variance trade-off, and the effects of L1 (Lasso) and L2 (Ridge) regularization. The experiments show how increasing model complexity through non-linear bases increases expressiveness but can lead to overfitting, while excessive regularization results in underfitting. To balance this trade-off, cross-validation can be applied to select the optimal regularization strength that minimizes validation error. The effects of L1 regularization promote sparsity by setting some coefficients to zero, whereas L2 penalizes large weights more smoothly. Overall, the tasks identify underfitting and overfitting, and uses that to select hyperparameters to achieve better generalization performance.

## Tasks

### Task 1

In this experiment, we generate synthetic data and fit a linear regression model with non-linear basis functions to analyze how varying model complexity influences performance.

To see how noise affects model behavior, we repeated the experiment with three noise levels: $\sigma^2 = 0.1$, 1.0, and 3.0, keeping the model complexity fixed at $D = 20$. When the noise was small ($\sigma^2 = 0.1$), the model followed the true function very closely. With moderate noise ($\sigma^2 = 1.0$), the fitted curve became smoother and slightly less precise but generalized better. At high noise ($\sigma^2 = 3.0$), the data became very scattered, and the model could no longer capture the fine details of the true pattern, although it avoided extreme overfitting. Overall, more noise made the model less accurate but also less likely to overfit, showing how both noise and model complexity influence generalization (Figure 1).

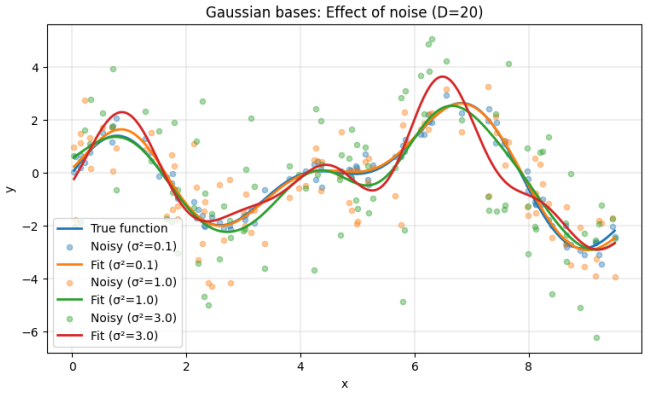


Figure 1: Effect of Noise Variance on Model Fit

Although the slides suggest picking the simplest model within one standard deviation of the model with the lowest validation error, in our analysis, we use the one standard error (1-SE) rule instead because it provides a more accurate way to measure uncertainty. As noted by Chen and Yang (2021)[1], "*When a regression procedure produces the regression estimator converging relatively fast to the true regression function, the standard error estimation formula in the 1-SE rule is justified asymptotically.*"

One standard deviation shows how much the errors vary across folds, which makes the range too wide and can lead to choosing a model that is too simple (e.g., $D = 0$, Figures 6-7). The standard error, on the other hand, measures how certain we are about the average validation error, giving a smaller and more reliable range for identifying the optimal model complexity.

As the number of Gaussian basis functions $D$ increases, training SSE decreases monotonically because the model becomes more flexible. For small $D$, the model underfits — it is too simple to capture the nonlinear structure of the data. Validation SSE reaches its minimum at $D = 5$ ($\text{SSE}_{\text{val}} = 25.62$) and remains nearly constant for larger $D$. Using

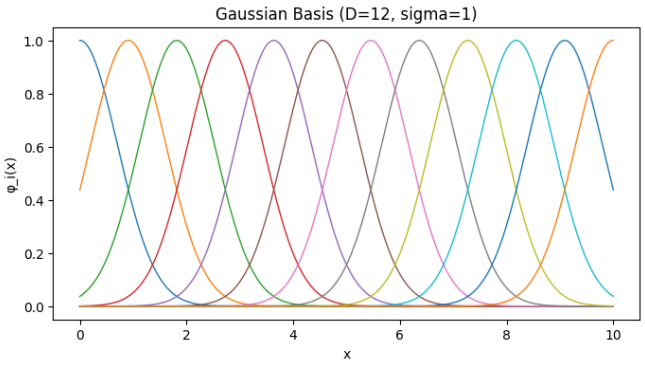


Figure 2: Gaussian basis functions

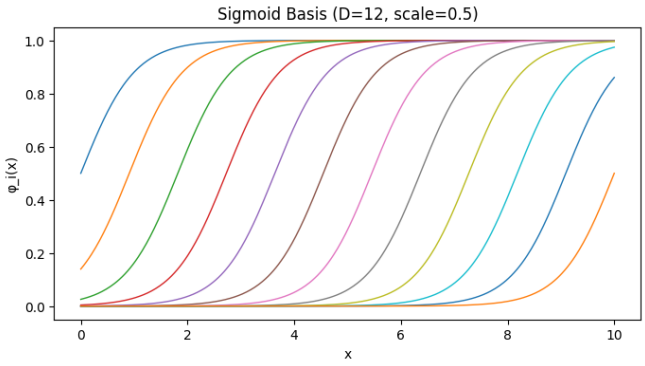


Figure 3: Sigmoid basis functions

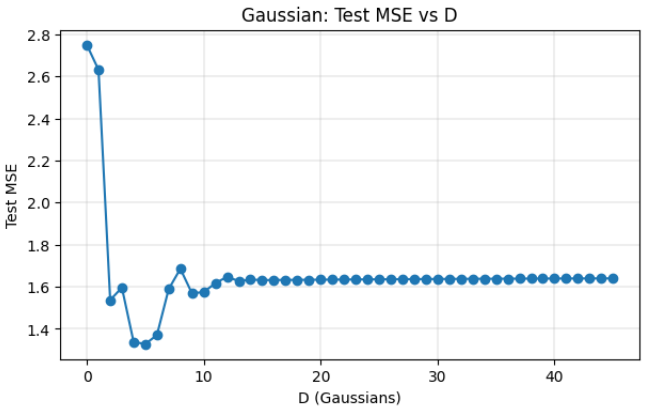


Figure 4: Test MSE vs Number of Gaussian Basis

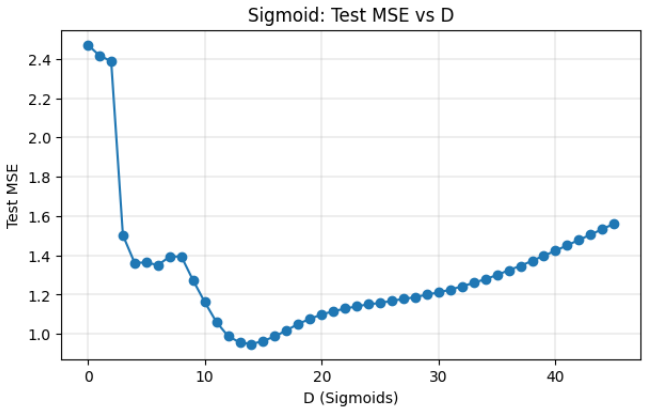


Figure 5: Test MSE vs Number of Sigmoid Basis

| D | SSE_train | SSE_val | ±1 SE | ±1 SD |
|---|---|---|---|---|
| 0 | 309.68 | 34.94 | 2.54 | 8.04 |
| 5 | 190.80 | 25.62 | 7.07 | 22.35 |
| 10 | 89.20 | 26.78 | 8.25 | 26.10 |
| 15 | 86.68 | 26.55 | 8.35 | 26.39 |
| 20 | 86.48 | 26.59 | 8.37 | 26.48 |
| 25 | 86.37 | 26.61 | 8.38 | 26.49 |
| 30 | 86.27 | 26.62 | 8.38 | 26.50 |
| 35 | 86.20 | 26.63 | 8.38 | 26.50 |
| 40 | 86.13 | 26.63 | 8.38 | 26.50 |
| 45 | 86.07 | 26.63 | 8.38 | 26.51 |

Best D (min mean SSE_val): 5 | SSE_val = 25.62 (±22.35 SD, ±7.07 SE)
1-SE rule -> choose D = 5 (threshold = 32.69)
1-SD rule -> choose D = 0 (threshold = 47.97)

Figure 6: Cross-Validation Results and Model Selection Rules (Gaussian Basis)

| D | SSE_train | SSE_val | ±1 SE | ±1 SD |
|---|---|---|---|---|
| 0 | 359.27 | 41.06 | 3.76 | 11.89 |
| 5 | 148.03 | 36.23 | 7.76 | 24.53 |
| 10 | 85.87 | 32.39 | 5.87 | 18.56 |
| 15 | 78.15 | 34.48 | 7.31 | 23.12 |
| 20 | 76.71 | 36.72 | 8.53 | 26.99 |
| 25 | 75.00 | 38.52 | 9.34 | 29.53 |
| 30 | 72.89 | 39.92 | 9.84 | 31.13 |
| 35 | 70.69 | 40.94 | 10.19 | 32.22 |
| 40 | 68.32 | 41.73 | 10.43 | 32.99 |
| 45 | 65.75 | 42.48 | 10.62 | 33.59 |

Best D (min mean SSE_val): 10 | SSE_val = 32.39 (±18.56 SD, ±5.87 SE)
1-SE rule -> choose D = 5 (threshold = 38.25)
1-SD rule -> choose D = 0 (threshold = 50.94)

Figure 7: Cross-Validation Results and Model Selection Rules (Sigmoid Basis)

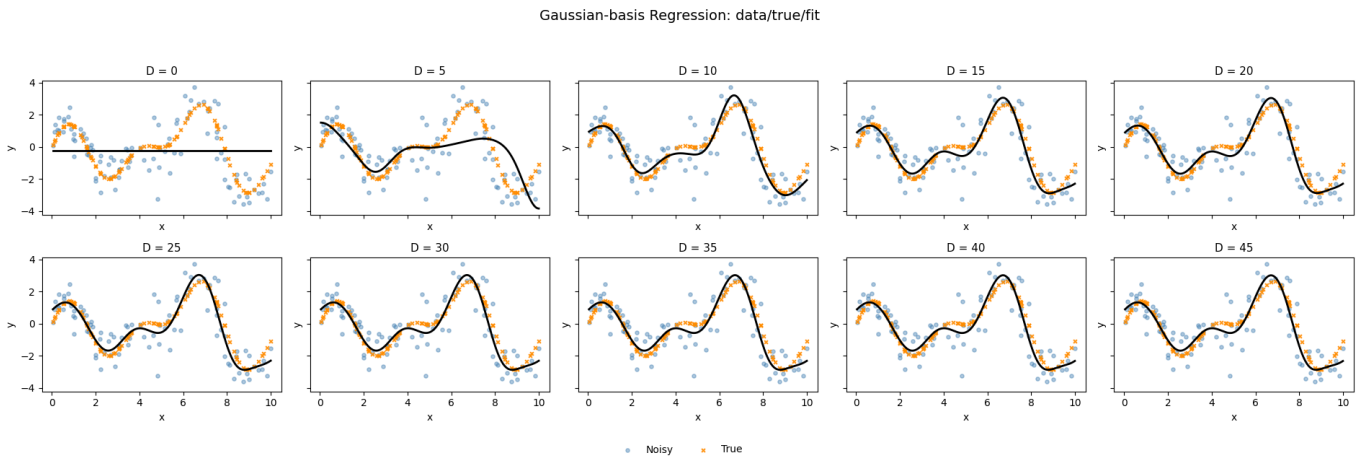


Figure 8: Gaussian Basis Fits for Various D Values

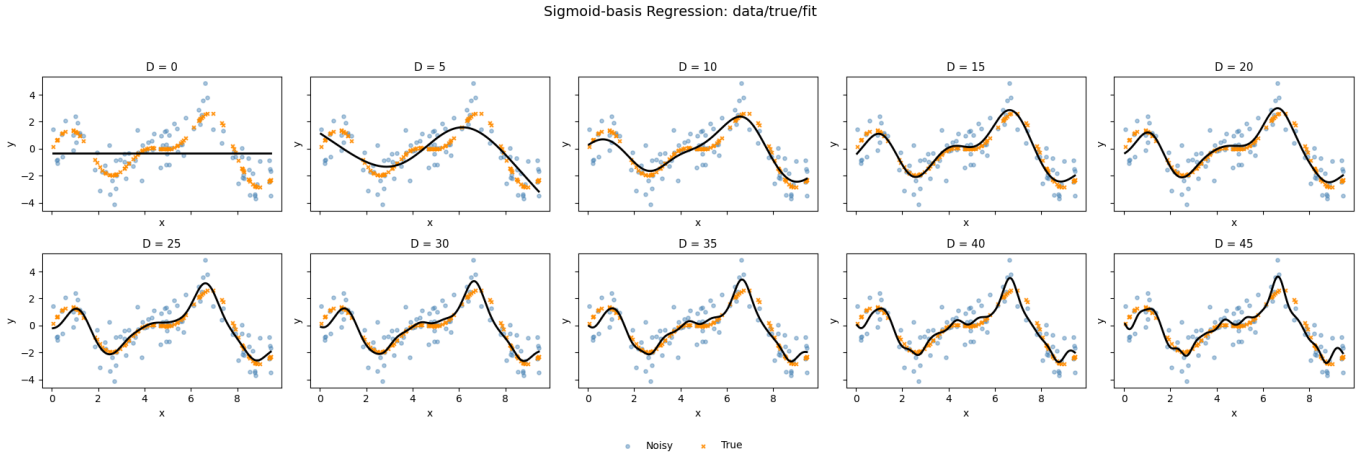


Figure 9: Sigmoid Basis Fits for Various D Values

the 1-SE rule (threshold = 32.69), the optimal model is $D = 5$, which provides the best balance between underfitting and overfitting (Figure 4, 6).

Similarly to Gaussian bases, as the number of sigmoid basis functions $D$ increases, training SSE decreases because the model becomes more flexible and fits the training data more closely. For small $D$, the model underfits and cannot capture the nonlinear structure of the data. Validation SSE reaches its minimum at $D = 10$ ($\text{SSE}_{\text{val}} = 32.39$) and increases slightly afterward, showing the start of overfitting. Using the 1-SE rule (threshold = 38.25), the optimal model is $D = 5$, which provides a good balance between bias and variance (Figure 5, 7).

Although the single-fit plots (Figures 8-9) might visually suggest that $D = 10$ provides the best fit to the data, cross-validation results indicate otherwise: for the Gaussian basis $D = 5$ reaches the minimum validation error, and for the sigmoid basis $D = 5$ lies within the 1-SE band of the minimum at $D = 10$, so the 1-SE rule selects the simpler $D = 5$ in both cases.

## Task 2

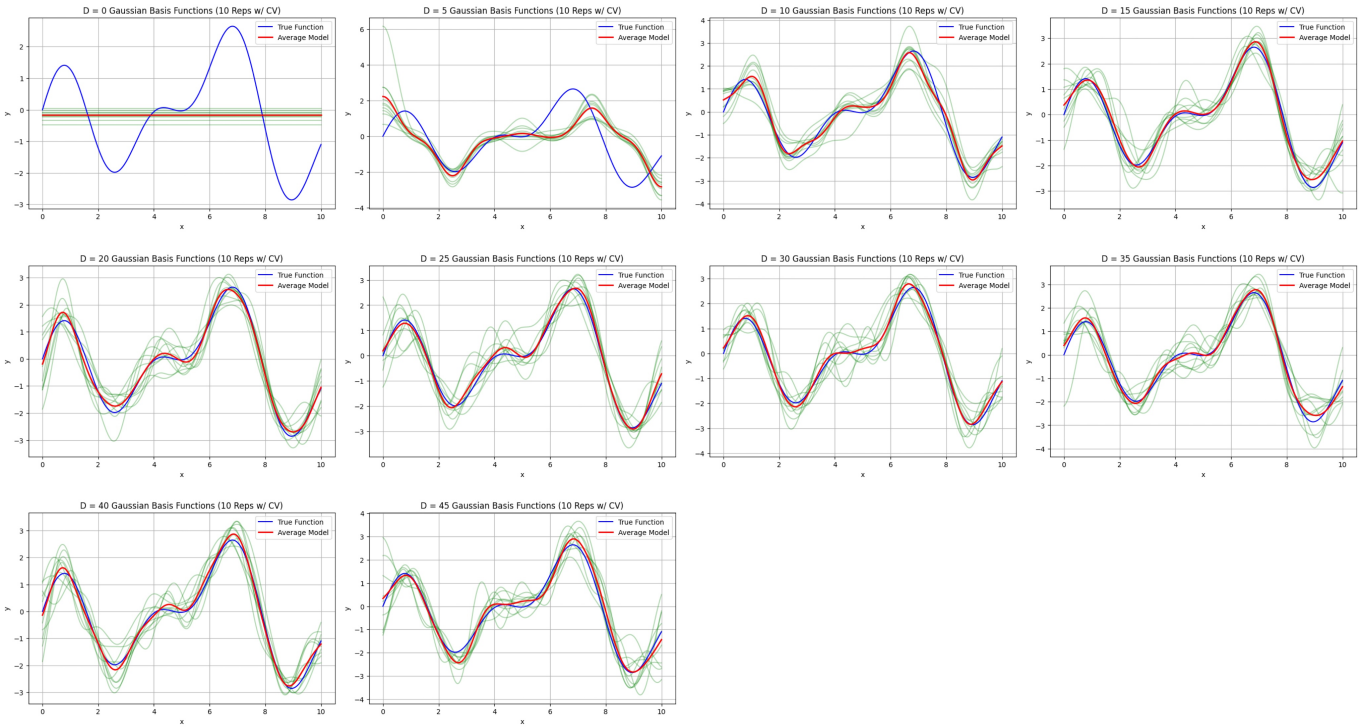


Figure 10: Visualizing model bias and variance

With fewer Gaussian basis functions, the model suffers from underfitting (exhibiting high bias and low variance), as the averaged fit deviates from the true function, but individual fits remain close to each other. Increasing the number of basis functions reduces bias yet increases variance (notably from 5 to 25), after which adding bases brings little change since the model already captures the complexity of the true function.

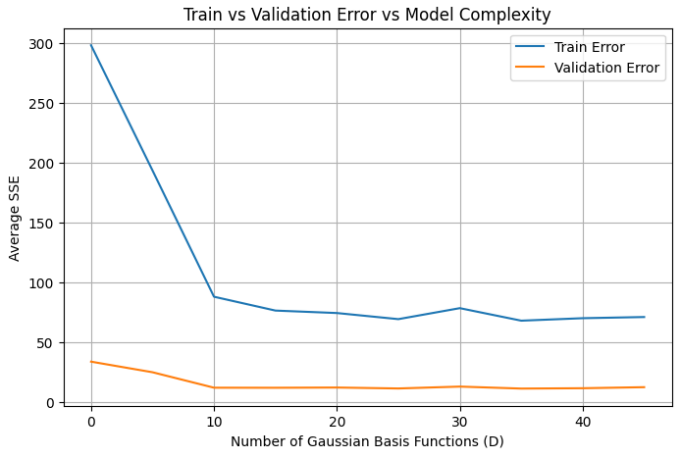


Figure 11: Average training and test error

The training cost decreases significantly as the number of Gaussian bases increases from 0 to 30, and the model is recovering from underfitting. After that, the model soon reached its optimal complexity.

## Task 3

In this experiment, the goal was to implement L1 (Lasso) and L2 (Ridge) regularization to a linear model with a Gaussian basis number 45 using 10-fold cross-validation to determine the optimal regularization strength, sampling 50 datasets from the ground truth distribution, and then training a model on each dataset. The range of $\lambda$ is 10 logarithmically spaced values between $10^{-3}$ to $10^{1}$.

As observed in Figures 9 and 10, when the value of $\lambda$ is small (from $10^{-3}$ to around $10^{-2}$), the training error is low while the validation error increases. This behavior indicates that the model is too expressive and overfits the training dataset. On the other hand, as $\lambda$ becomes too large (from $10^{0}$ to $10^{1}$), both training and validation errors increase, as the model becomes too simple to represent the data, leading to underfitting.

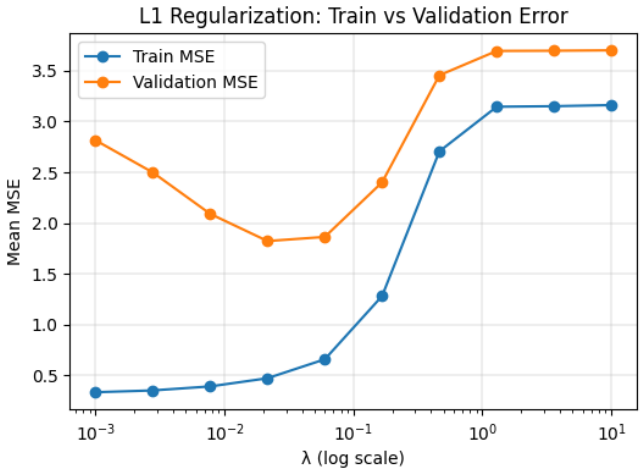


Figure 12: Train and validation error vs. $\lambda$ (L1)

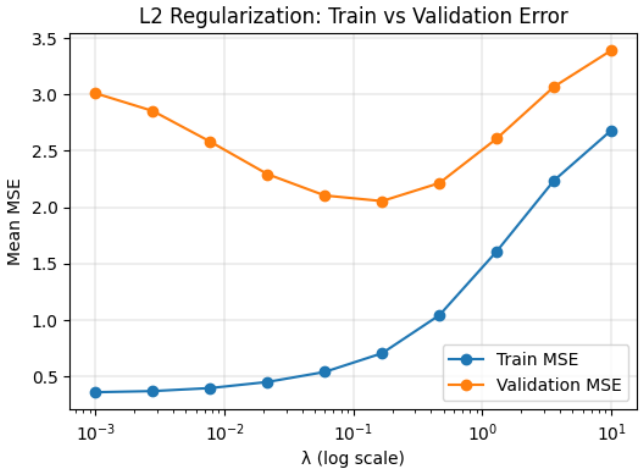


Figure 13: Train and validation error vs. $\lambda$ (L2)

According to Figures 11 and 12, for small $\lambda$ values (from $10^{-3}$ to around $10^{-2}$), the bias is low. At the same time, the variance is high, indicating that the model is too expressive and slightly overfits the training data. As $\lambda$ increases, the variance decreases and the bias increases, leading to a region where their sum and total error get to a minimum. For L1 (Lasso) regularization, which enforces stronger sparsity, the optimal $\lambda$ is smaller ($\approx 10^{-2}$ - $10^{-1}$). While, for L2 (Ridge) regularization, the optimal is when $\lambda \approx 10^{-1}$ - $10^{0}$. Beyond these points, the bias gets bigger and the model underfits, as stronger regularization may simplify the model too much and reduce its ability to capture data variability in a general way.

Finally, to choose a specific optimal regularization strength value, it is possible to apply a "rule of thumb" that selects the simplest model within one standard deviation (for this experiment) of the model with the lowest validation error. The simplest model means the one with lower variance; therefore, the optimal value is the largest $\lambda$ within one standard deviation of the minimum validation error.

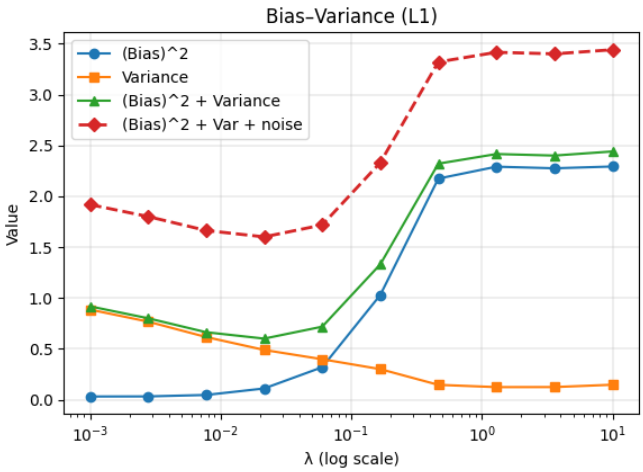


Figure 14: Bias-Variance trade-off (L1)

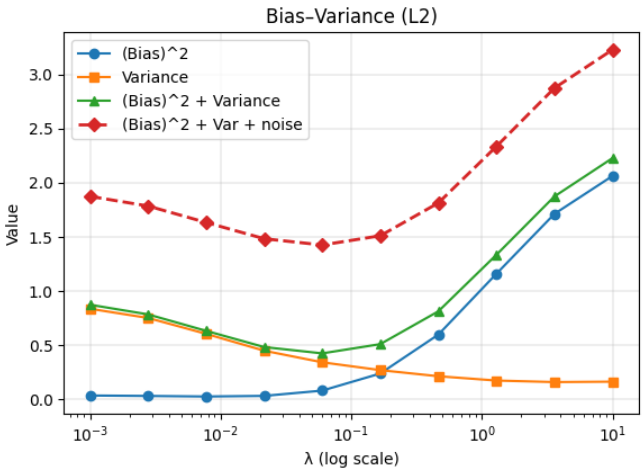


Figure 15: Bias-Variance trade-off (L2)

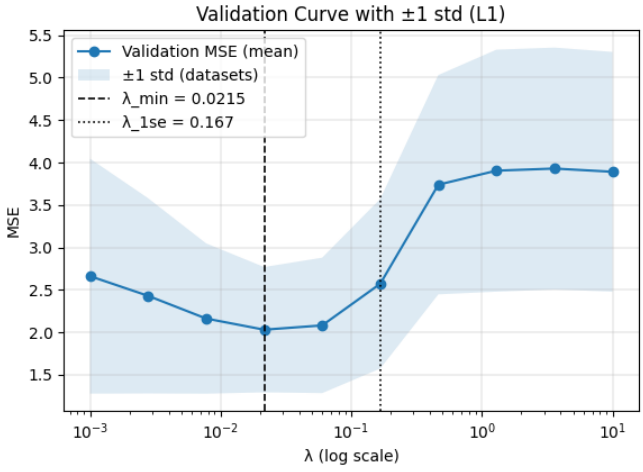


Figure 16: Validation curve and standard deviation (L1)

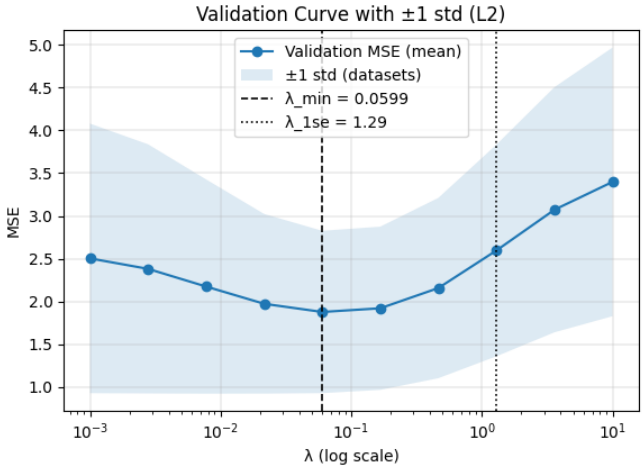


Figure 17: Validation curve and standard deviation (L2)

In Figure 7, the model with the lowest validation error for L1 regularization is the one for which $\lambda$ is 0.0215 for L1. Following the rule of thumb, the optimal $\lambda$ is **0.167**, as it is the simplest model (i.e., higher regularization strength) within one standard deviation of the minimum validation error. For L2, as observed in Figure 8, the lowest $\lambda$ value is 0.0599, and the optimal value is **1.29**.

## Task 4

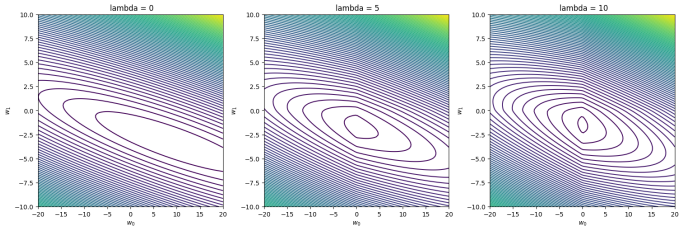


Figure 18: L1 regularization contour plot

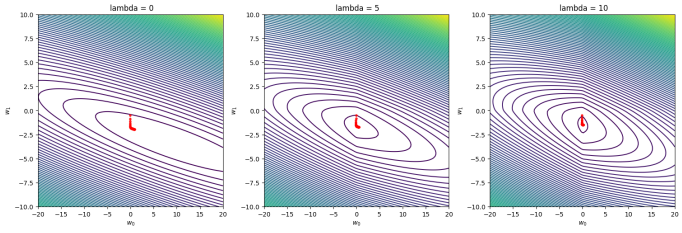


Figure 19: L1 regularization gradient trajectory

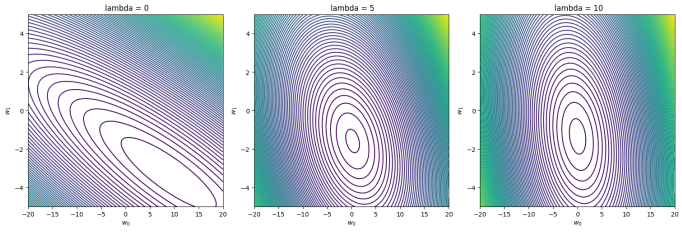


Figure 20: L2 regularization contour plot

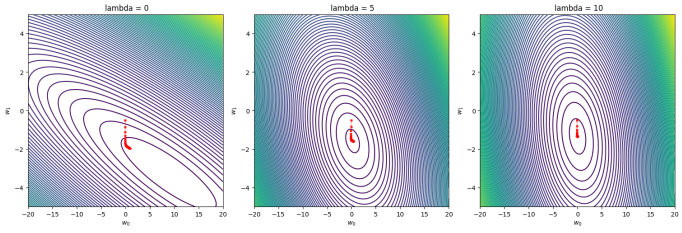


Figure 21: L2 regularization grandient trajectory

L1 regularization encourages sparsity by pushing the optimal weight for $w_0$ towards 0. As $\lambda$ increases, the feasible region becomes more constrained around $w_0 = 0$, and the model's final weights are pulled closer to these sparse optimal values. In contrast, the L2 regularization plots show smoother, more circular contours, and while larger $\lambda$ values pull the weights closer to the origin, they remain nonzero. This demonstrates that L2 penalizes large weights without enforcing sparsity. As $\lambda$ increases from 0 to 10, the loss landscape becomes steeper and more centralized, leading to shorter optimization paths and smaller final weight magnitudes.

# Originality and creativity

To go beyond the minimum requirements, we implemented the k-fold cross-validation logic in the Bias–Variance Decomposition plot for Task 3, which can be verified in the source code. Additionally, in Task 1, the instructions required the

use of Gaussian basis functions, but we also experimented with a Sigmoid basis to compare the model behavior under different nonlinear transformations, as presented in Task 1. Even though the fitted curve is different from the Gaussian basis, the behavior of the number of basis vs. overfitting or underfitting is the same. We also tested how different levels of noise affect model performance and generalization. In task 1 we also compared the 1-SD and 1-SE rules for model selection, with the latter giving a more reliable model selection.

## Discussion and Conclusion

The experiments collectively demonstrated how model complexity, noise, and regularization influence the behavior and generalization of linear regression models. Increasing the number of basis functions improved the model's ability to capture nonlinear patterns but also made it more likely to overfit. The results highlighted the balance between bias and variance: more flexibility lowered bias but increased variance, while regularization helped find the right middle ground. L1 regularization encouraged sparsity by forcing some weights to zero, while L2 regularization produced smoother weight shrinkage without eliminating parameters. Visualizing the loss surfaces confirmed that stronger regularization simplified the model by pulling weights closer to zero. Overall, the results show how choosing the right model complexity and regularization strength with cross-validation leads to more reliable and generalizable results. Future work could focus on applying these approaches to real-world datasets, incorporating nonlinear or Bayesian regularization, and evaluating how model performance varies across different data distributions and in the presence of outliers.

## Statement of Contributions

Daria Goptsii was responsible for Task 1, which inlcuded implementing and analyzing Gaussian and Sigmoid basis functions, as well as studying how different levels of noise affect model performance. Giane Mayumi was responsible for implementing task 3, including the Bias-Variance Decomposition with k-fold. Yixuan was responsible for task 2 and 4, looked at the effect of adding regularization level on weight and costs. Overall, all team members helped each other during implementation, reviewed and discussed the results together, in addition to contributing to the report.

## References

[1] Y. Chen and Y. Yang, *The One Standard Error Rule for Model Selection: Does It Work?*, *Stats*, vol. 4, no. 4, pp. 868–892, 2021. Available at: https://doi.org/10.3390/stats4040051