Regularization and Model Evaluation

Abstract

This assignment explores key concepts in linear regression, including the impact of model complexity, the bias-variance trade-off, and the effects of regularization through L1 (Lasso) and L2 (Ridge) methods. Using synthetic data generated from a non-linear function, we examined how varying the number of Gaussian basis functions influences model performance, transitioning from underfitting to overfitting. Through iterative model fitting and evaluation using SSE on training and validation sets, we identified the optimal complexity that balances these extremes. Additionally, by resampling data multiple times, we visualized the bias-variance trade-off and assessed how model variability affects fitting accuracy. We then implemented regularization techniques and evaluated the optimal regularization strength using 10-fold cross-validation, revealing how different values of λ affect bias, variance, and overall model performance. Finally, we analyzed the effects of L1 and L2 regularization on loss functions. The results obtained highlight the relationship between model complexity, regularization techniques and the inherent bias-variance trade-off.

Task 1

We observe in the graphs below that with very few bases, both the training and validation SSE are high. This indicates underfitting, where the model is too simple to capture the underlying patterns in the data. Around D = 20, the training SSE and the validation SSE are minimized, suggesting that this is the optimal number of bases for capturing the data pattern without overfitting. Beyond this number of bases, although the training SSE continues to decrease, the validation SSE starts to increase significantly. This indicates overfitting, where the model fits the training data too well to generalize new data. The model becomes too complex and starts capturing noise in the training data rather than the true underlying patter.

The validation set is crucial for identifying overfitting. While the training SSE alone might suggest that increasing complexity always improves performance, as it keeps decreasing, the validation SSE reveals when additional complexity starts to harm the model's generalization ability to unseen data. By monitoring the validation SSE, we can select the model with the best generalization ability, which in our case is D = 20. In short, the transition from underfitting to overfitting is effectively managed by evaluating both training and validation errors. The validation set helps determine the optimal model complexity that balances fitting accuracy with generalization capability.

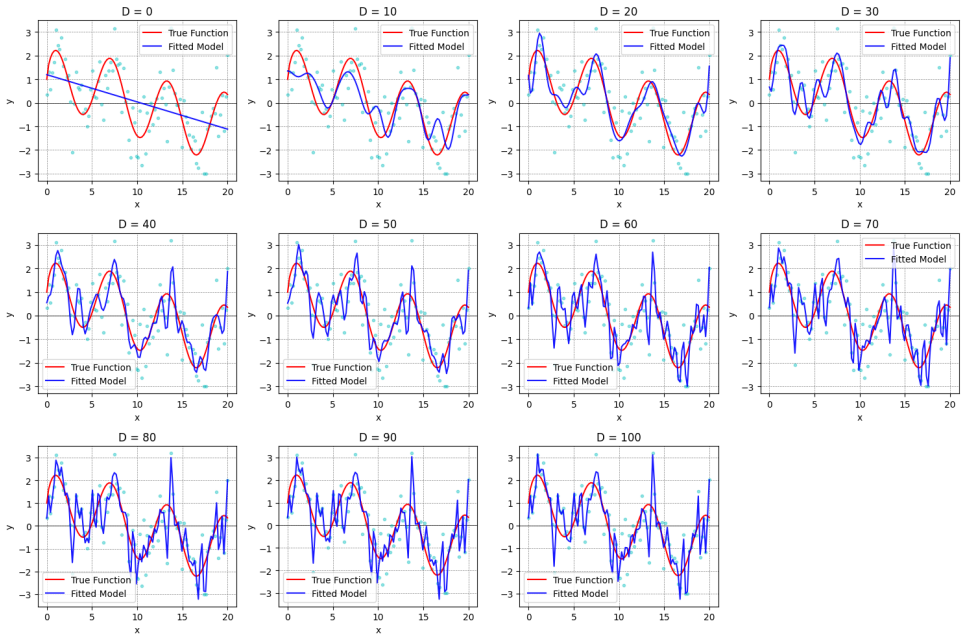


Figure 1: Fitted Models with Different Number of Bases

D	Training SSE	Validation SSE
0	195.101211	5.918599e+01
10	89.113432	5.613664e+01
20	63.782760	5.319490e+01
30	55.146587	5.554818e+01
40	46.006862	5.903792e+01
50	28.085429	9.455887e+02
60	20.215959	4.292276e+02
70	6.919290	1.340185e+06
80	3.882299	2.821291e+07
90	3.547272	8.086495e+07
100	3.305981	2.429992e+08

Table 1: SSE vs Number of Gaussian Bases

Task 2

Repeating the first task 10 times, we observe a clear transition from high bias to high variance as the number of Gaussian bases (D) increases. At D = 0, the model shows high bias as it is underfitting the data. The simpler and flatter model fails to capture the complexity of the true function. On the other end, when D increases beyond 20, the model exhibits high variance, overfitting the data. It starts to capture noise, leading to erratic predictions, especially at the edges of the input range. We observe that the optimal number of bases is around 20, where the model achieves a good balance, by closely approximating the true function without adding excessive fluctuations.

Furthermore, we observe on the log-scale plot of MSE vs D that the training MSE consistently decreases as D increases, indicating better fit to training data. As for the validation MSE, it initially slightly decreases before reaching a minimum around D = 20 and sharply increasing for higher D values. The divergence between training and validation MSE for large D values is a clear sign of overfitting.

These plots effectively demonstrate the importance of model selection to balance bias and variance. The optimal model provides the best trade-off between fitting the training data and generalizing to new data, highlighting the importance of using a validation set to guide model complexity selection.

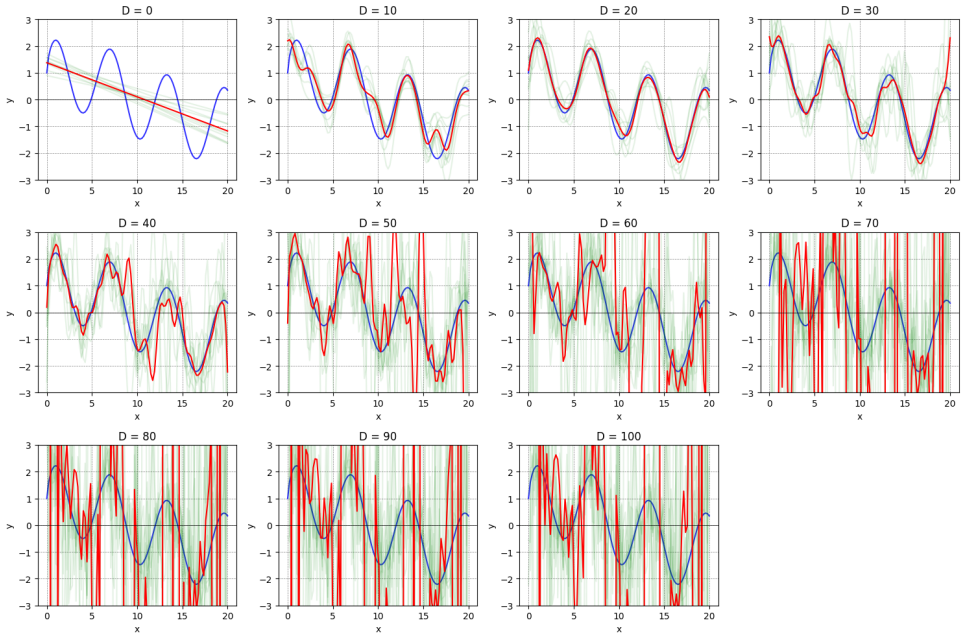


Figure 2: Multiple Fitted Models with Different Number of Bases

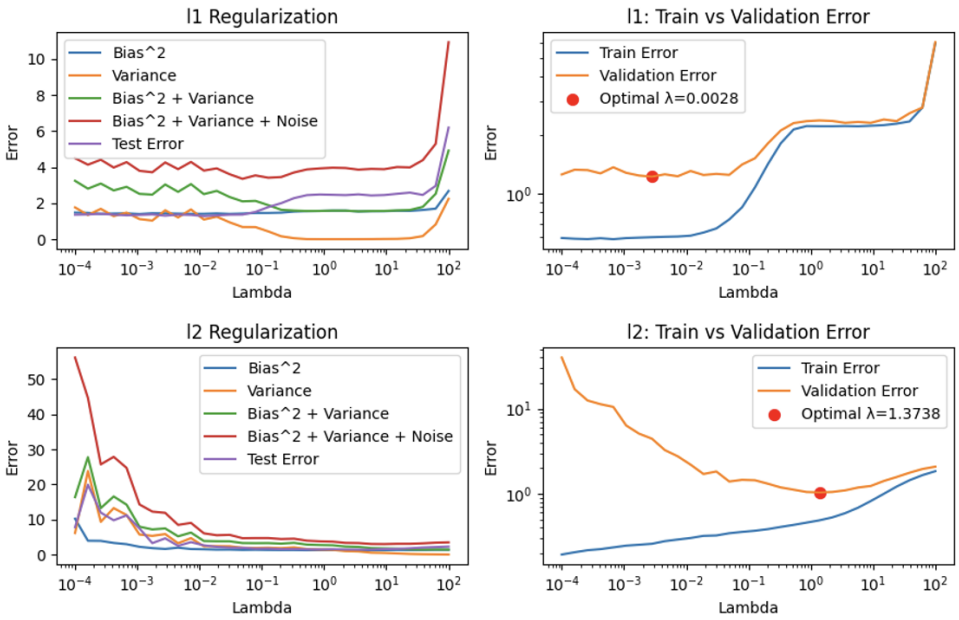
Task 3

In this task we implemented L1 and L1 regularization to our model and used 10-fold cross-validation to a series of λ values to find the optimal regularization strength that balances the trade-off between low bias and low variance. In our case, the validation error is minimized for L1 around λ=0.0028 and for L2, around λ=1.3738.

Taking a closer look at how different values of λ affect the model's performance in terms of bias and variance, we observe that for L1, both the bias and variance decrease as λ increases before spiking for large λ values, as the model becomes too simple and fails to capture important patterns (underfitting). For L2, the bias and variance drop consistently with increasing , as L2 regularization reduces model complexity by shrinking weights. This behavior is expected since the model becomes less complex.

Now looking at how different values of λ affect the model's training and validation errors, we observe for L1 that as increases, the training and validations error remain rather constants before increasing sharply. This is because L1 regularization shrinks many weights to zero, simplifying the model, again, leading to underfitting. For L2, the training error increases gradually with , but not as sharply as in L1 regularization. This is because L2 regularization shrinks weights without forcing them to zero, maintaining some model flexibility. The validation initially decreases before reaching a minimum at the optimal λ value and increasing again. For large λs, the over-regularization of the model compromises the model's ability to generalize.

In conclusion, choosing the optimal value is crucial for balancing the bias-variance trade-off. Too low a value leads to high variance (overfitting), while too high a value increases bias (underfitting). The plots suggest that a value around $10^{-}2$ for L1 and around 10^{0} for L2 provides the best performance for this particular model and dataset.



Task 4

When applying L1 and L2 regularization, we observe a few differences in how the models behave under varying regularization strengths. In the plots below, we can see that L1 regularization originally shrinks the bias term than L2 regularization but brings it to 0 faster than L2. That is because for L1 regularization, the key result is that it promotes sparsity by driving some coefficients to exactly 0. As a result, L1 regularization is useful when we want to perform feature selection or reduce the number of non-zero coefficients in the model. In contrast, L2 regularization penalizes large weights but does not promote sparsity as strongly as L1. Instead of setting weights to 0, L2 gradually shrinks all coefficients, resulting in smaller weights across the board, making it suitable when all features are potentially important but should be weighted appropriately to avoid overfitting. For both L1 and L2 regularization, the optimization paths for gradient descent show how regularization impacts the convergences of the model. As λ increases, the paths become steeper and more constrained, reflecting a stronger penalty on the weights. The graphs show that both regularization techniques make the weights converge to similar values.

We suspect that our results could not showcase the the full utility of L1 and L2 regularization because regularization is only useful when we need to simplify an overly complex model. However, in our case, since the model is

already linear, applying regularization doesn't make much sense. We thus observe, in the graph below, an increase in MSE when using stronger regularization strengths as we are simplifying a linear regression model fitted to linear data.

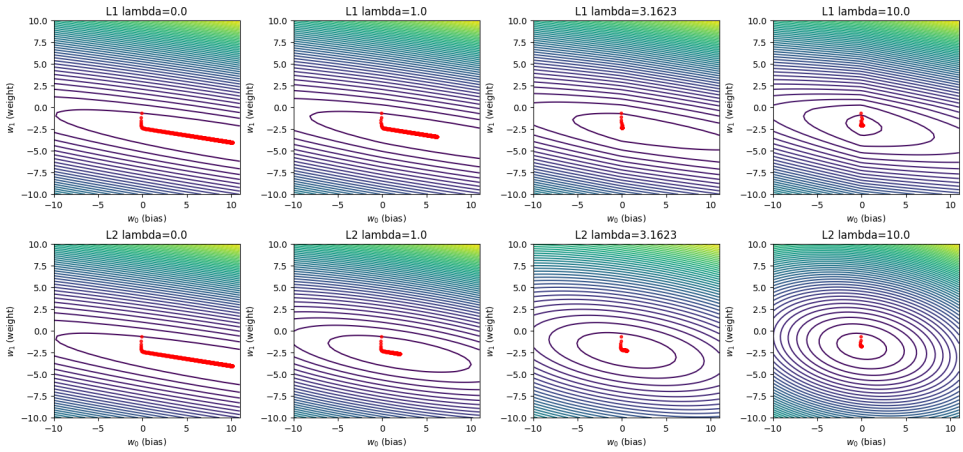
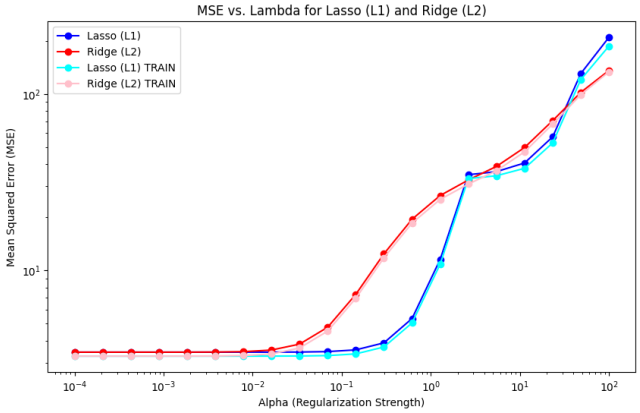
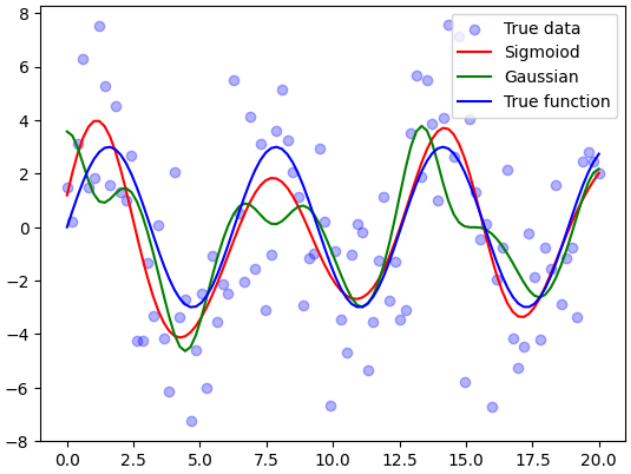


Figure 3: Contour Plots and Optimization Paths



Originality/Creativity

We played around in task 1 with the Sigmoid bases to see what the differences are between those and the Gaussian bases, and whether there are any advantages to using one over the other. We observed that the usage of Sigmoid function over Gaussian can have its advantage in some specific situations. Since the Sigmoid function has constant S-shaped curves, it makes it ideal for situations where there's a need to capture the smooth transition of a data, a gradual shift in of behaviors, for example. On the other hand, the Gaussian function is more bell shape, making it better for capturing region-specific inputs around the centers, optimal to demonstrate peaks or local patterns at specific points. Additionally, since the Sigmoid function only lies between 0 and 1, it offers a by default strong counter against sample with extreme values and outliers, whereas the Gaussian function, although smooth, has more challenges when going into higher dimensions. The Gaussian therefore requires careful hyperparameter selection. Thus, in situation where the data obtained has a high noise level and oscillating, Sigmoid bases can be more appealing than Gaussian bases, as it can capture the oscillating pattern of the dataset while dealing more effectively with outliers and extreme values. In the graph below, we observe that the model fitted with Sigmoid bases has a better fit and MSE = 0.84716 whereas with Gaussian bases, MSE=1.66545



Discussion and Conclusion

This assignment provided valuable insights into the dynamics of linear regression, particularly concerning model complexity, regularization, and the bias-variance trade-off. Throughout the analysis, we observed that as the number of Gaussian basis functions increased, the model's ability to fit the training data improved. However, this trend also led to a greater risk of overfitting, where the model performed poorly on the validation set due to excessive complexity. We identified the optimal number of bases to be around 20, where model complexity and generalization capabilities were balanced. Through repeated model fitting, it was revealed, as anticipated, that models with fewer basis functions exhibited high bias, resulting in systematic errors, while those with an excessive number of bases displayed high variance, leading to fluctuating predictions. The repeated experiments emphasized that achieving a model that minimizes both bias and variance is crucial for robust predictive capabilities. The application of L1 and L2 regularization techniques further illustrated how regularization can effectively manage overfitting. By adjusting the regularization strength, we found that bigger λ values shrunk the weights more and that L2 regularization shrinks the weights more uniformly while L1 regularization encourages sparsity. The cross-validation results reinforced the necessity of the tuning regularization parameter λ to find the optimal balance that enhances model performance, which in our case was around $10^{-}2$ for L1 regularization and around $10^{-}0$ for L2 regularization.

In conclusion, this assignment underscored the interplay between model complexity, regularization, and predictive performance. Understanding these relationships is vital for developing effective regression models in practice. As we mentioned in task 4, the model from which the synthetic data was generated from was linear and future work would involve working with multi-variable functions which would more effectively highlight difference between the 2 regularization techniques, that is, L1 encourages sparsity while L2 penalizes large weights while keeping more of weights.

Statement of Contributions

William: task 1 and task 2
Guillaume: task 3 and task 4
Jessica: assignment write-up