Exploring the Effects of Expressiveness and Regularization on Linear Regression Models

COMP551 - Assignment 2

Kaibo Zhang
Email: kaibo.zhang@mail.mcgill.ca

Mingshu Liu
Email: mingshu.liu@mail.mcgill.ca

Alek Bedard
Email: alek.bedard@mail.mcgill.ca

October 22, 2024

Abstract

This report focuses on applying linear regression with non-linear basis functions to synthetic data, exploring model complexity and its impact on performance, and studying the bias-variance tradeoff. Additionally, it investigates the effects of L1 and L2 regularization using cross-validation and the impact of regularization on loss landscapes. Through the experiments, we observed that increasing the number of Gaussian basis functions initially improves model performance but eventually leads to overfitting. Regularization techniques effectively mitigated this by promoting model sparsity and reducing large coefficient values. The analysis illustrates how balancing model complexity and regularization strength results in improved generalization.

1 Linear Regression with Non-Linear Basis Functions

We generated 100 synthetic data points using the following non-linear function with noise:

$$y(x) = \sin(\sqrt{x}) + \cos(x) + \sin(x) + \epsilon$$

The linear regression model was then fitted using Gaussian basis functions, with the number of basis functions varying from 0 to 100. The sum of squared errors (SSE) was computed for both the training and validation sets. Plots of the Gaussian basis functions across the range of generated data are shown in 1.

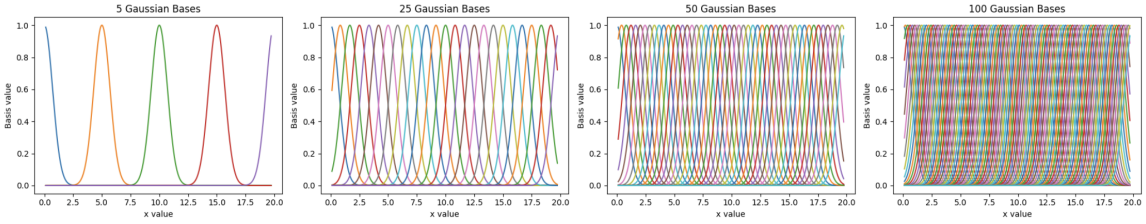

Figure 1: Plots of varying Gaussian basis functions

1.1 Model Fit and Number of Basis Transformation

The graph 2 contains the fit of the models with different numbers of bases, illustrating how the model complexity changes with different choices of D in the range of [0, 100, 10]. When D is small, such as 0 or 10, the model is too simple to capture the underlying patterns in the data, resulting in underfitting. The fit remains smooth but fails to match the true function. As D increases to moderate values, around 20 to 40, the model begins to better capture the true function. The fit aligns more closely with the true data distribution without overfitting the noise, reflecting a well-optimized model. However, as D becomes larger, particularly beyond 50 and up to 100, the model becomes excessively complex. At this stage, the fit starts to follow the noisy data points more closely, resulting in an erratic and fluctuating line. This is a clear sign of overfitting, where the model not only captures the true function but also fits the random noise introduced during the data generation process. Thus, as the number of basis functions increases, the model becomes more prone to fitting the noise rather than the true underlying function.

Figure 2: Linear regression model fit with varying number of Gaussian basis functions

Number of Basis Functions	Training SSE	Validation SSE
0.0	204.28398940705955	36.30039484278573
10.0	93.18330679223516	19.86189602224994
20.0	49.57295040611672	19.401712302912433
30.0	43.66046825993248	31.334773247199468
40.0	35.438874419635916	25.403336667752757
50.0	25.68607687318712	36.748715056629294
60.0	19.07443141250739	1078.5893932255835
70.0	11.622723070992958	16942.24387169697
80.0	3.7817236286219464	17851389.5355622
90.0	3.878609017997214	34106149.7597189
100.0	3.9185183623006914	43085568.977724165

Figure 3: Training and validation SSE for varying number of Gaussian basis functions

1.2 Optimal Number of Gaussian basis

Table 3 provides numerical support for the patterns observed in the graphs. With few basis functions (e.g., 0 or 10), both training and validation SSEs start high, indicating underfitting. Around 20 basis functions, both SSEs drop, showing optimal generalization. Beyond 30 functions, the training SSE keeps falling, but the validation SSE rises, signaling overfitting as the model begins to capture noise. The steep rise in validation SSE after 60 functions further reinforces that the ideal number of basis functions is around 20 for balancing bias and variance. The validation set is crucial in selecting the optimal model because it reflects how well the model can generalize to new, unseen data. While the training SSE alone might suggest that more complex models (with more basis functions) perform better, the validation SSE tells a different story. It shows that after a certain complexity threshold (around 20 basis functions in this case), the model starts overfitting. Thus, the validation SSE helps us pinpoint the model with the right balance of complexity—avoiding both underfitting and overfitting—making the case for selecting a model with around 20 basis functions as optimal for this synthetic dataset.

2 Bias-Variance Tradeoff with Multiple Fits

The analysis is based on running each model 10 times with 10 different datasets generated from the same true distribution. The plots highlight the bias-variance trade-off as the number of basis functions (D) increases. For low D values (0 to 10), the models show high bias and underfitting, failing to capture the complexity of the data. This is evident from the high training and validation errors and the smooth but inaccurate average fit lines (red). Bias dominates here, as the models consistently miss key data features. As D increases to moderate values (10 to 20), bias decreases, and the models better capture the true distribution. Validation errors drop, and the fits align closely with the data as well as their average fit, showing an optimal balance between bias and variance. Both the fit plots and validation SSE reflect this improved generalization. However, beyond D=40, overfitting occurs as model complexity increases, leading to high variance. Training error continues to decrease, but validation error rises sharply, indicating the model is fitting noise rather than the true pattern. The fluctuating fit lines (green) around the average fit and the rise in validation SSE highlight this shift from low bias to high variance, resulting in poor generalization.

3 Regularization with Cross-Validation

We implemented L1 and L2 regularization to control model complexity and applied 10-fold cross-validation to evaluate different regularization strengths (λ). The mean squared error (MSE) for both training and validation sets was plotted against λ, allowing for the selection of the optimal λ that minimizes the validation error.

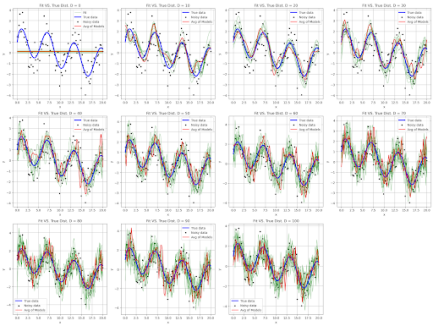

Figure 4: Linear Regression models fit with varying number of gaussian bases

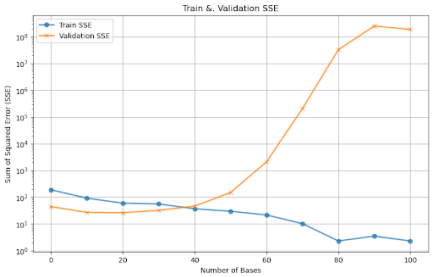

Figure 5: Training and validation SSE for varying number of Gaussian basis functions

3.1 Optimal lambda Selection

The validation error plots for both L1 and L2 regularization show that the optimal λ values occur where the validation MSE is minimized—around $\lambda \approx 0.2682$ for L1 and $\lambda \approx 1.9306$ for L2. At these points, the model achieves a balance between bias and variance, generalizing well without overfitting. With small λ values (near zero), regularization is weak, leading to overfitting and high variance. The key difference between L1 and L2 lies in how they control complexity. L1 tends to force coefficients to zero with greater regularization powers, leading to sparsity, while L2 shrinks all coefficients more gradually. At high λ values, both regularizations exhibit high bias, but L1's impact is more abrupt, and when reaching a certain threshold, the curve actually becomes too sparse which pushes the model to become a flat line at zero. L2's effect is much smoother and gradual, reflecting their distinct approaches to controlling model complexity.

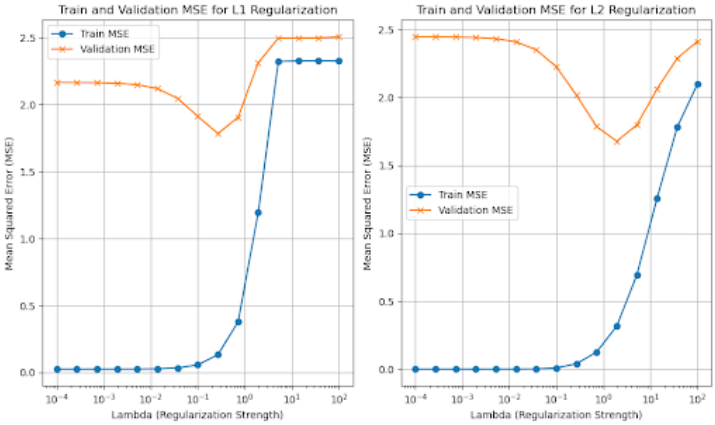

Figure 6: Training and validation MSE for L1 and L2 regularization

3.2 Effect of lambda on Bias and Variance

In the first-row plots, the bias-variance decomposition for both L1 and L2 regularization highlights how the choice of λ directly impacts the trade-off between bias and variance. For very small λ values (close to 10^{-4}), the models exhibit low bias but high variance. This is typical when regularization is weak, allowing the model to overfit the training data by capturing noise, thus resulting in high variance. As λ increases, the variance steadily decreases, while bias increases, particularly around λ=1and beyond. This reflects the regularization effect, where stronger regularization reduces the model's flexibility, increasing bias but lowering variance, as the model becomes too simple to overfit the data. For both L1 and L2, after a certain point (above $\lambda = 10$), the bias dominates as regularization is too strong, and variance remains low. At these high λ values, the error increases primarily due to underfitting, as evidenced by the increase in the bias + variance curves.

3.3 Cross-Validation Enhancements and Curve Shifts

The cross-validation (CV, k=10) plots in the second row reinforce the insights derived from the train-test split results. Cross-validation provides a more robust estimate of model performance by averaging over multiple data

splits, leading to a more generalized view of bias and variance trends. In these plots, the curves shift downward compared to the train-test split plots, indicating that cross-validation reduces both the overall bias and variance, resulting in a lower total error. This downward shift occurs because CV better estimates the model's ability to generalize by mitigating the chance that the chosen train-test split was not representative of the full dataset. The benefit of this approach is that it minimizes the risk of overfitting or underfitting tied to a particular train-test split, offering more reliable parameter selection and performance evaluation across different λ values.

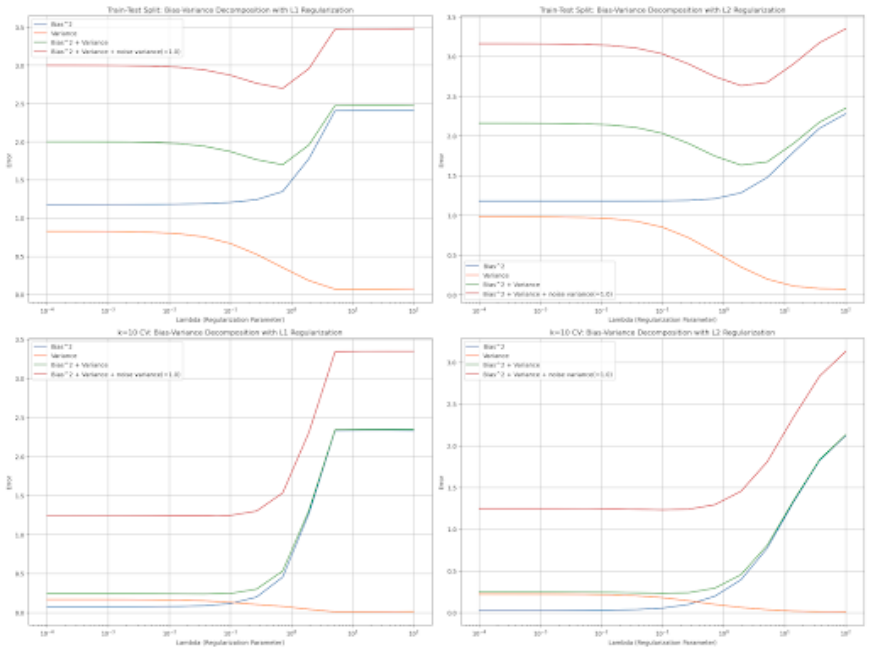

Figure 7: Bias-Variance trade-off decomposition graph for L1 and L2 regularizations

4 Effect of L1 and L2 regularization on Loss

In this section, we explore how L1 and L2 regularization techniques affect the gradient descent process with respect to the loss function for linear regression. 50 datapoints were sampled from the following distribution:

$$y = -4x + 10 + 2\epsilon$$

Five loss contour plots were generated for L1 and L2 regularization schemes with varying regularization strengths. As shown in figure 8, L1 tends to drive the weights toward 0, while impacting negatively the smoothness of the contours, making it more difficult to optimize. On the other hand, figure 9 shows how larger regularization

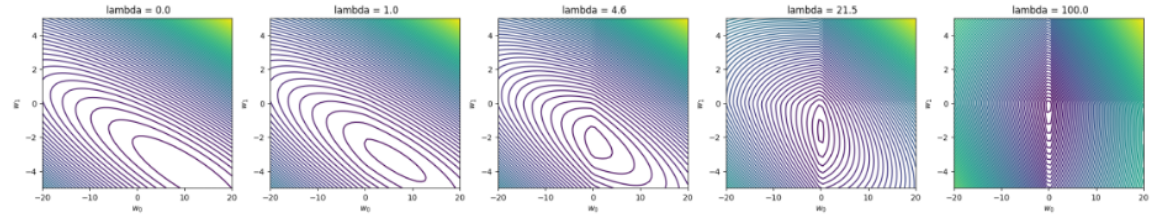

Figure 8: Loss contours for L1 regularization

strengths in L2 penalize larger weights, while keeping the contours relatively smooth. In L1, the contours are more diamond-shaped, reflecting the nature of the regularization technique, which tends to drive some parameters to zero (inducing sparsity). As λ increases, the gradient descent path moves vertically, with w_1 being more significantly shrunk towards zero. For high λ (eg. $\lambda = 100$), L1 regularization forces the parameters to lie mostly along the axes, a characteristic of L1's sparsity-promoting behavior (figure 10). We can observe that L2 also constrains the weights uniformly as lambda increases. As λ increases (from left to right), the contour lines become tighter, indicating stronger regularization. For $\lambda = 0$, there is no regularization, and the gradient descent path moves

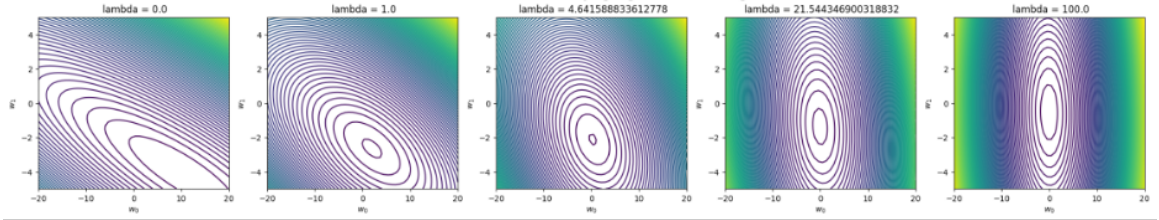

Figure 9: Loss contours for L2 regularization

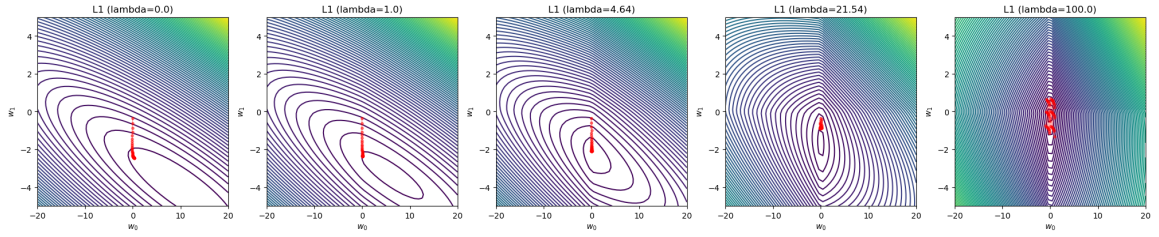

Figure 10: Gradient descent path for L1 regularization

smoothly towards the center. As λ increases (e.g., $\lambda = 21.54$), the contours become more circular, and the path converges more directly toward the center (figure 11). Contrarily to L1, we can observe that w_1 is not pushed to 0, even with an exceptionally large regularization strength ($\lambda = 100$). The descent paths shown figures 10 and

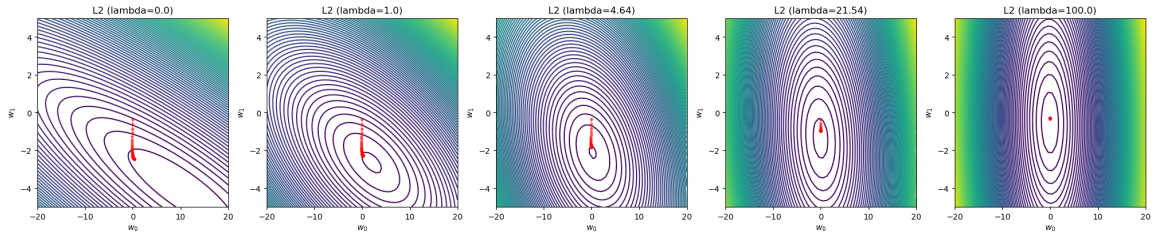

Figure 11: Gradient descent path for L2 regularization

11 were drawn from a vanilla linear regression model. For comparison, we implemented a gradient descent with ADAM optimization. Contrarily to vanilla gradient descent, ADAM yields smoother and more efficient paths for both regularization techniques. We can observe constant convergence toward the middle of the contour lines in appendix figures 12 and 13.

5 Conclusion

The assignment examined the bias-variance tradeoff and the impact of regularization in linear regressions. The results highlight that tuning model complexity is crucial to balancing bias and variance—simple models underfit, while overly complex models overfit to noise. Experiments with L1 and L2 regularization show that large constraints increase bias by making the model too sparse. Bias-variance decomposition confirmed that both regularizations shift from low bias/high variance to high bias/low variance as λ increases. L1 promotes sparsity by forcing coefficients to zero, while L2 shrinks coefficients more gradually. Gradient descent paths also confirmed L1's sparsity and L2's smoother shrinkage, with ADAM optimization providing more efficient convergence. Future investigations could explore the effects of combining both regularization types to fine-tune this balance further as well as other basis transformation techniques.

6 Statement of contributions

Mingshu took the lead in writing the report and contributed with some parts of coding, mainly with dataset exploration. Alek took the lead in formatting the report in Latex and coding some other parts of the experiments and graph developments. Kaibo took the lead in model implementation and coding with a focus on experiments.

7 Appendix

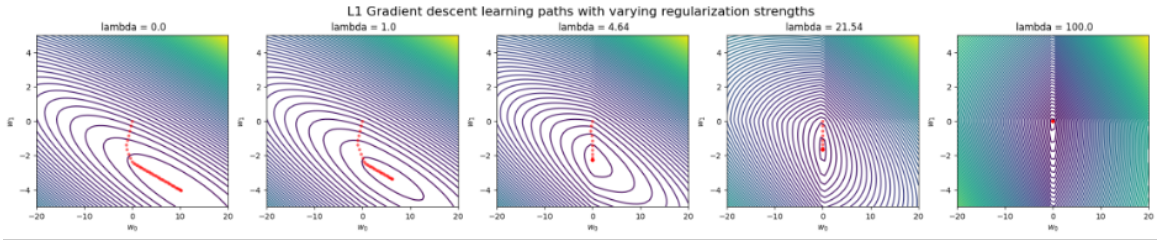

Figure 12: Gradient descent path for L1 regularization using ADAM optimization

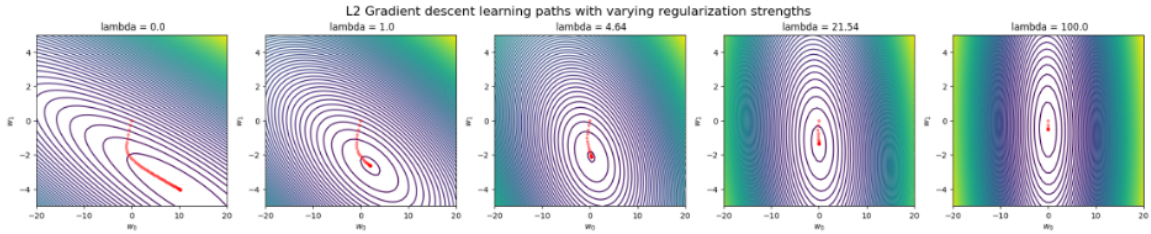

Figure 13: Gradient descent path for L2 regularization using ADAM optimization