Comp551 - Assignment 2

Members:
Tran Tuan Khai Tuan Phan (260916543)
Adam Dufour(261193949)
Gabriel Caballero (261108565)

October 22, 2025

Abstract

Regularization is a key concept in machine learning to obtain the best models possible and to refrain from over-training. In this work, lasso regularization (L1) and ridge regularization (L2) have been studied on two different models. The first model of interest in this work is a non-linear base implementation of a linear regression model. The second model, is a more basic linear regression model, this allowed the study of bias and variance. From these studies, the complexity of the models were visualized and it allowed the following conclusions : bias gives high training and validation error as variance leads to super complex models that suffers over-fitting. It was also possible to perform a hyperparameter optimization study that allowed to follow the gradient descent into the optimization of λ for our model at different regularization strength and mode (L1 or L2). It was observed that 15 Gaussian bases gave the best performance for the first model while avoiding over-fitting. Regarding the second model, it was found that stronger regularization lead to a worse model and that L1 and L2 performed similarly. Finally, the best λ for L1 and L2 regularization were evaluated to be $\lambda_{L1} = 0.0010$ and $\lambda_{L2} = 0.0028$.

Results

Task 1 - Linear Regression with non-linear basis functions

Generating the ground truth distribution according to the following equation :

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

Where ϵ is the noise added to the distribution that is sampled from a normal distribution of $\mu = 0$ and $\sigma = 1$. The points were generated from $0 + 1 \times 10^{-9} \leq x \leq 10$ since when $x = 0$ the model suffered from the following concept :

$$\lim_{x \to 0^+} \log(x) = -\infty$$

The correction of 1×10^{-9} is then applied to avoid this issue. The data can be visualized in fig. 1. Applying non-linear Gaussian bases to the noisy data from fig. 1 using an increasing number of bases gives the results observed in fig. 2. The expression of the Gaussian bases can be shown as such :

$$\phi(x, \mu, \sigma) = \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) \tag{2}$$

The plot of the sum of squared error (SSE) for all different basis functions averaged over 10 runs produced fig. 4. The process illustrates that the number of basis functions (D) leads directly to the complexity of a machine learning model. As complexity increases, it is expected that the error on the training data tends to go to zero as the error on the validation set jumps up significantly. This clearly shows the concept of over-fitting, as the model's variance or complexity increases, the model no longer learns, but only remembers the data. In addition, the spikes observed as D increases can be explained by the fact that the model is fitting outlier values. All in all, since our ground truth is a simple function, it is expected that a smaller D would give the best model, in this case $D = 15$ gives the best results.

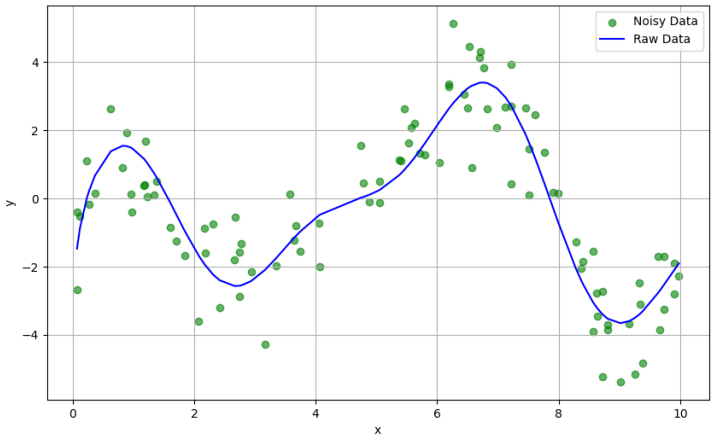

Figure 1: Generated data distribution from equation eq. (2)

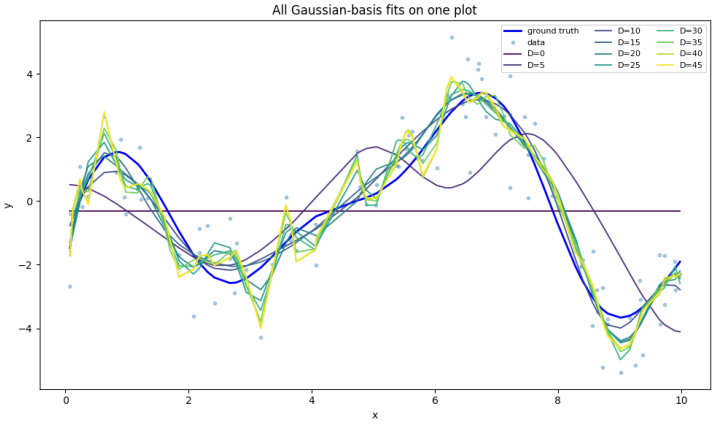

Figure 2: Generated data distribution from equation eq. (2)

Task 2 - Bias-Variance tradeoff with Multiple fits

The following figures show different models obtained for a varying number of basis functions across 10 iterations, fig. 3 combined with fig. 4, demonstrating the behavior of loss and variance for multiple basis functions. Once again, this section clearly shows the impact of variance and bias on the results of the model. It is possible to observe that when the number of basis functions is very small, the model is simple and does a poor job of estimating the ground truth. The model, however, is very repetitive across all iterations, this illustrates a high bias of the model. As D increases, the model does a better job of fitting the ground truth, but as it increases even more, the iterations become more and more variable between iterations and stretches further from the ground truth. This illustrates the variance, as the iterations are more different across all runs.

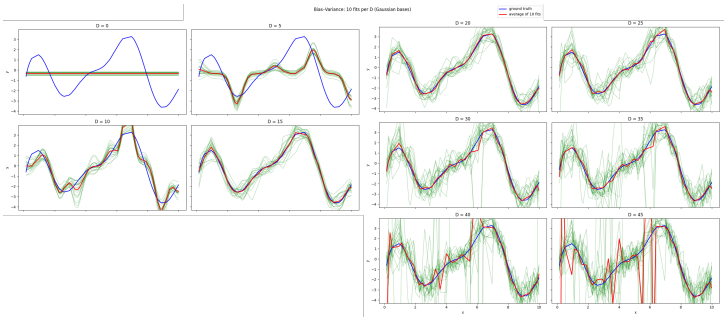

Figure 3: Model results across 10 repetitions

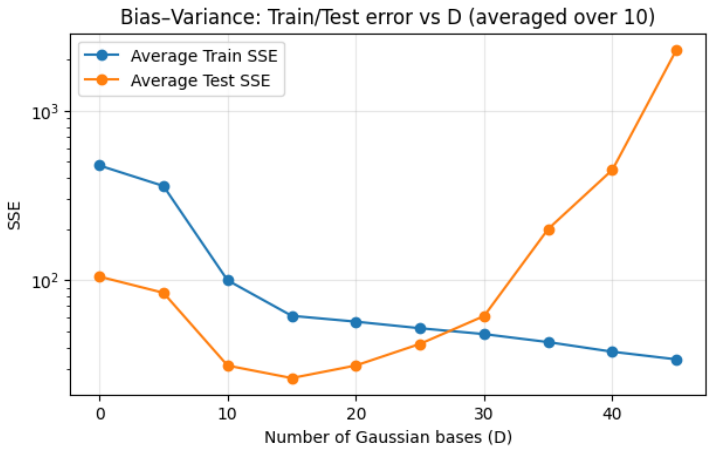

Figure 4: Loss for the model across the different basis set averaged over 10 iterations

Task 3 - Regularization with Cross-Validation

fig. 5 shows that the implementation of the regularization of L1 and L2 allows the optimization of the hyperparameter λ. The experiment allowed the bias and variance to be plotted in fig. 6 allowing the analysis of the behavior and the selection of the best value for the hyperparameter λ. The value for each regularization strategy was chosen as the minimum point of the curve shows in fig. 5, this value clearly gives the smallest error for the model independent of complexity. In general, increasing λ shows that a higher regularization strength leads to a worse model. This makes sense as the regularization strength increases, the model is strongly penalized for higher weights, and prevents the model from truly learning and leads to poor predictions. This can be explained by the tradeoff between bias and variance, as a strong λ leads to a simple model, a model with high bias. A small or zero λ can cause over-fitting of the model as there are few or no constraints on the model's learning process. Ultimately, λ allows to obtain a better model to a point, where the regularization is too restrictive for high λ and leads to a worse model.

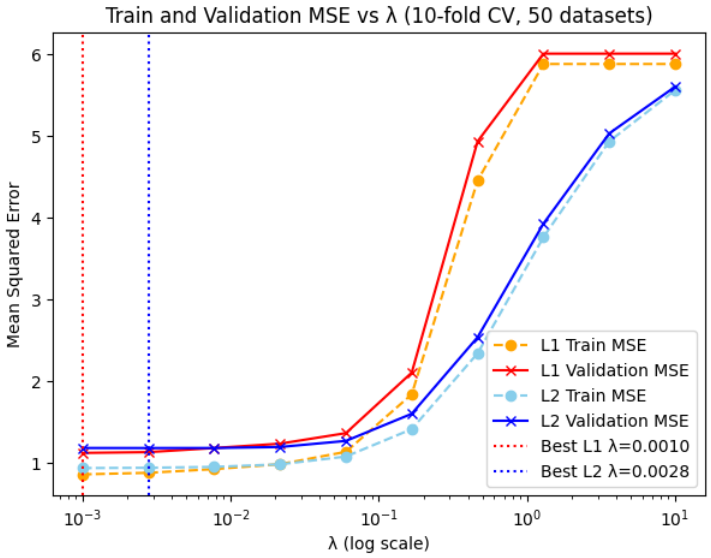

Figure 5: Hyperparameter optimization for each regularization approach

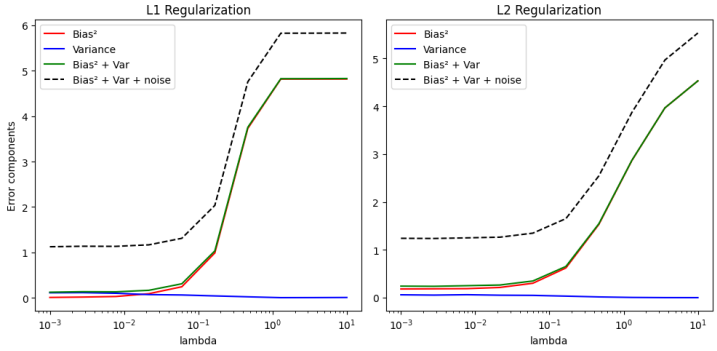

Figure 6: Bias and variance plotted for each regularization types

Task 4 - Effect of L1 and L2 Regularization on Loss

In this section, a new dataset was generated using the equation :

$$y(x) = -3x + 8 + \epsilon \tag{3}$$

Where ϵ is sampled as it is in eq. (2) The optimization path is shown in fig. 7. It is possible to observe that the stronger λ is, the shorter the optimization path is as the penalty is stronger. The smaller λ gave the best results as the weights were closer to the ground truth (no noise) and the loss was smaller as shown in fig. 8.

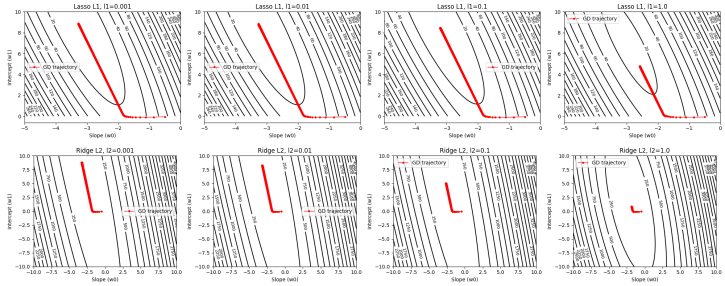

Figure 7: Model path for L1 regularization

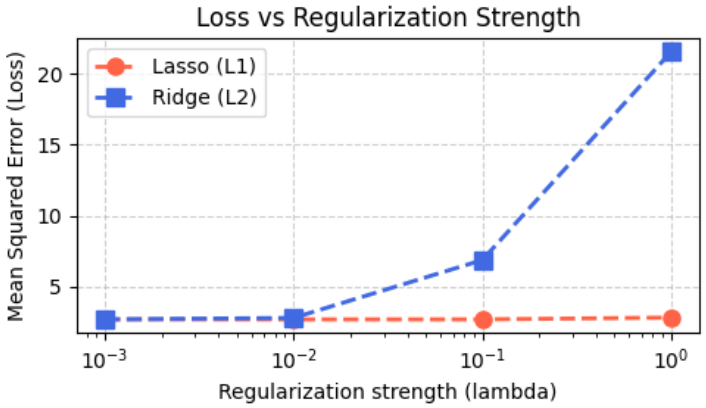

Figure 8: Loss reported across different regularization strength

Discussions and Conclusions

Finally, this experience has allowed the group to understand and comprehend better how bias and variance impacts a model. As it is clearly not a simple problem, it is crucial to evaluate this metric as it allows for a better model. As Task 1 and 2 shows, this tradeoff can be obtained by evaluating the loss and by looking at the models' fit to the ground truth for different levels of complexity. However, in this experiment, the ground truth distribution was available and allowed to get a clear and complete picture of the models' performance. As it is not always available in other problems, this indicates the difficulty of

evaluating and obtaining an absolute answer to this problem. Task 3 provided insights on how the different regularization techniques allowed for the optimization of the models' hyperparameters λ, once again illustrating the bias variance tradeoff. Task 4 presented visuals of the paths the models took to minimize λ and obtain the best model for different regularization strengths.

Additional work

Additionally, the model weights were studied and compared for each type of regularization. As shown in fig. 9, it is observed that the weights in L1 regularization approach a value of zero for many more weights than in L2. This is explainable in the way L1 differs from L2. In fact, when looking at the penalty shape for each type of regularization. In L1, the shape is much more abrupt as it is square, L2's penalty is smoother and forms a circles. Then, the penalty applied to bigger weights is then stronger in L1 than in L2. However, this difference can be relevant when discriminating L1 and L2. If a simpler a more scarce model is needed for computational resources reasons, L1 is the way to go. If the user wants every weights to be considered in the process, L2 can assure this is the case.

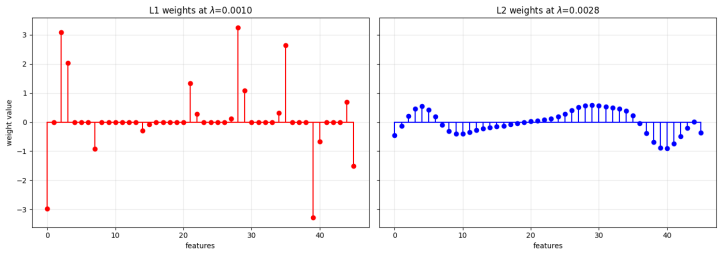

Figure 9: Model weights reported in accordance to regularization types

Statement of Contribution

Tran Tuan Khai Tuan Phan: Bias and variance plots; study on size of basis functions; effect of σ and μ on fit; code for T1 and T2.

Gabriel Caballero: Coding in T1, T2, T3, and T4, additional work section, README and requirements files.

Adam Dufour: Report writing; task 3 and 4 coding sections; plots for loss and model path in T3 and T4.