Analysis of Least Squares Regularization Methods

Steven Y. Liu[1], Luke Stuckenbruck[2], and Alex Ojemann[3]

1 Department of Aerospace, University of Colorado Boulder

2 Department of Applied Mathematics, University of Colorado Boulder

3 Department of Computer Science, University of Colorado Boulder

In this paper, we investigate the behaviors of different types of regularization of least squares regression. Regularization, in a general sense, serves to limit the size of a model to limit overfitting of the training data. For a least squares regression model, the loss function is the residual sum of squares. Regularizing this model involves the addition of a penalty term to the loss function. Some forms of regularization of least squares regression models also allow for custom weights for each feature that determine how strong the regularization is for each of them rather than applying the same regularization strength for all values. In this paper we investigate the use of Ridge, Tikhonov, and LASSO regularization on least squares regression models and compare the effects of the regularization on the model parameters.

1. Introduction

The Ordinary Least Squares (OLS) regression is a technique commonly used for data fitting of over determined linear systems in numerical analysis. An ordinary least squares regression minimizes the residual sum of squares of the expression $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are features and labels respectively we wish to fit $\mathbf{x}$ to. However, it is possible for a model derived from the ordinary least squares regression to have undesirable characteristics due to ill conditioned problems or noisy data. To resolve this, we add a weight expression to the least squares regression formulation that restricts the solution. Regularization of least squares is a technique in which a penalty term is added to the loss function of ordinary least squares regression. This penalty term generally consists of a hyperparameter γ that scales a norm of the solution of the model adjusted by a weight function. The weight expression is determined by the type of regularization being used. We will be exploring 3 methods of regularizing least squares regressions in this project: Tikhonov Regression (TR), Ridge Regression (RR), and LASSO Regression.

In Ridge Regression, the norm is the 2-norm of the solution. Ridge Regression aims to solve a problem with the least squares approach. Specifically, systems with a lot of noise may cause regular least squares to produce large regression coefficients as it attempts to fit the uneven data. This will produce a solution that is ill-fit to the data set.

According to the Eberly College of Science at Penn State University, Hoerl and Kennard (1970) theorised that a constant could be added to the diagonal entries of the matrix $\mathbf{A^T A}$ in the least squares method in order to add a stabilizing effect [4]. This constant is chosen to be one that minimizes the sum of the least squares.

The applications of the Ridge Regression method and data fitting in general range across a wide variety of fields. Virtually any topic or institution that handles data can benefit from line fitting methods. One example of a recent use for ridge regression is in the field of genetic studies. Arashi et al. describe how scientists seek to model the production of riboflavin associated with certain genes in bacteria with a linear system [1]. However, the number of genes is much larger than the number of samples taken, necessitating the use of linear regression models such as ridge regression. The article concludes that the Ridge Regression was effective in modeling the data, more so than the previously used methods.

Tikhonov Regression is a generalized form of the Ridge Regression. The norm is also the 2-norm of the coefficients, multiplied by a matrix of weight values instead of a single value, allowing for differing levels of regularization for each feature. Generally, the Tikhonov matrix is chosen to be the first or second derivative matrix operator on a vector function. This matrix enforces a level of smoothness along the solution vector, preventing the solution terms from being too far apart. This prevents large changes between terms and reduces oscillatory behavior in the solution vector.

In this project, we will explore the derivation of each of these estimators and demonstrate their effects on the least squares solution with numerical examples.

2. Derivations

2.1. **Tikhonov's Regularization.** A common method of regularization for the least squares problem is the *Tikhonov Regularization (TR)*. This method adds a weight matrix $\mathbf{D}$ to the least squares formulation in order to constrain the solution by some metric. Often the weight matrix is a first or second derivative matrix operator. If we take the weight matrix to be the identity matrix scaled by a factor, the regularization method is known as a *Ridge Regression*, which is a special case of TR, and expanded on in Section 2.2.

The minimization problem for TR can be written:

$$\arg\min_x ||\mathbf{Ax} - \mathbf{b}||_2^2 + \lambda||\mathbf{Dx}||_2^2 \tag{1}$$

Given that $||\mathbf{x}||_2^2 = \mathbf{x^T x}$, the Tikhonov Regularization in matrix form is derived [3]

$$\arg\min_x ||\mathbf{Ax} - \mathbf{b}||_2^2 + \lambda||\mathbf{Dx}||_2^2$$

$$\arg\min_x (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) + \lambda(\mathbf{Dx})^T(\mathbf{Dx})$$

$$\arg\min_x (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) + \lambda\mathbf{x^T D^T Dx}$$

$$\arg\min_x \mathbf{x^T A^T Ax} - \mathbf{x^T A^T b} - \mathbf{b^T Ax} + \mathbf{b^T b} + \lambda(\mathbf{x^T D^T Dx})$$

$$\arg\min_x \mathbf{x^T}(\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})\mathbf{x} + \mathbf{b^T b} - 2\mathbf{b^T Ax}$$

Note $\mathbf{b^T Ax} = \mathbf{x^T A^T b} = (\mathbf{b^T Ax})^\mathbf{T}$ since all of these expressions result in the same scalar value. The expression being minimized is essentially a positive quadratic equation, so it will be minimized when its gradient is 0 with respect to $\mathbf{x}$.

$$2(\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})\mathbf{x} - 2\mathbf{b^T A} = 0$$

$$2(\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})\mathbf{x} = 2\mathbf{b^T A}$$

$$\mathbf{x} = (\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})^{-\mathbf{1}}\mathbf{b^T A}$$

$$\mathbf{x} = (\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})^{-\mathbf{1}}\mathbf{A^T b}$$

Where the matrix estimator $E_{Tikhonov}$ used to find $\mathbf{x}$ is

$$E_{Tikhonov} = (\mathbf{A^T A} + \lambda^\mathbf{2}\mathbf{D^T D})^{-\mathbf{1}}\mathbf{A^T} \tag{2}$$

$E_{Tikhonov}$ is the equivalent to the normal equation matrix $(\mathbf{A^T A})^{-1}\mathbf{A^T}$ when solving OLS. If $\mathbf{D}$ is the identity matrix scaled by some factor γ, this regularization is known as the standard form of TR, or the Ridge Regression. The standard form TR limits the magnitude of the solution vector $\mathbf{x}$ depending on the magnitude of γ. Another common form of the Tikhonov matrix is a derivative operator such as

$$\mathbf{D}_1 = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{3}$$

$$\mathbf{D}_2 = \begin{bmatrix} \frac{1}{2} & -1 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & \frac{1}{2} & -1 & \frac{1}{2} & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix} \tag{4}$$

Which approximate the first and second order derivatives respectively. Increasing the scaling λ of the weight function results in the solution $\mathbf{x}$ approaching a horizontal line with the first derivative operator, and a slanted line with the second derivative operator. I.e. when plotting the elements of the solution with respect to the index, the resultant shape will approach a horizontal or slanted line as λ increases for the first and second derivative operators [6].

2.2. **Ridge Regression.** To determine the equation of the line that best fits a set of data, the expression $||\mathbf{Ax} - \mathbf{b}||_2^2 + \gamma||\mathbf{x}||_2^2$ must be minimized for $\mathbf{x}$, where $\mathbf{x}$ represents the solution of the least squares equation. Because $||\mathbf{x}||_2^2 = \mathbf{x^T x}$ we can rewrite this equation as $(\mathbf{Ax} - \mathbf{b})^\mathbf{T}(\mathbf{Ax} - \mathbf{b}) + \gamma \mathbf{x^T x}$. Using the fact that $(\mathbf{A^T x^T b})^\mathbf{T} = \mathbf{b^T Ax}$, this expression simplifies to:

$$\mathbf{A^T x^T Ax} - 2\mathbf{b^T Ax} + \mathbf{b^T b} + \gamma \mathbf{x^T x}$$

To minimize this expression for $\mathbf{x}$, we must take the derivative with respect to $\mathbf{x}$, bearing in mind that $\boldsymbol{A^T A}$ is symmetric:

$$2\mathbf{A^T Ax} - 2\mathbf{b^T A} + 2\gamma \mathbf{x}$$

Setting the equation equal to zero and factoring out $\mathbf{x}$ and 2:

$$(2\mathbf{A^T A} + 2\gamma \mathbf{I})\mathbf{x} - 2\mathbf{b^T A} = 0$$

$$(\mathbf{A^T A} + \gamma \mathbf{I})\mathbf{x} = \mathbf{b^T A}$$

Solving for $\mathbf{x}$ we find:

$$\mathbf{x} = (\mathbf{A^T A} + \gamma \mathbf{I})^{-1} \mathbf{b^T A}$$

The solution to this equation is $\boldsymbol{x}^*$, which represents the coefficients that minimize the original ridge regression expression.

Re-arranging with linear algebra methods we have the final form:

$$\mathbf{x}^* = (\mathbf{A^T A} + \gamma \mathbf{I})^{-1} \mathbf{A^T b}$$

3. Numerical Experiments

3.1. **Ridge Regression.** The first test of Ridge Regression involves fitting the line $y = 3x + 2$ with Gaussian noise added. A range of γs from 0 to 5000 were tested and the one that best optimized the data was 12.75. The results of this fit compared to least squares and the actual function line are shown in Figure 1:

The next test involves fitting the line $y = x^2$ with Gaussian noise added. The γ that best optimized the data was 40.75. The results of this fit compared to least squares and the actual function line are shown in Figure 2:

As can be seen in the plot, ridge regression is more accurate than least squares when modeling a straight line. For an exponential function the results are a little less clear. However, a careful analysis of the R^2 factor for each reveals that the ridge regression is slightly more accurate than the least squares method.

3.2. **Tikhonov's Regularization.** A common Tikhonov matrix $\mathbf{D}$ used for Tikhonov regularization is the derivative operator. Using this matrix enforces smoothness along a vector function $\mathbf{x}$, where "smoothness" in this case refers to the behavior of the elements of the vector solution $\mathbf{x}$. For the first order derivative operator defined in Equation 5, the greater factor λ, the closer the resultant elements of $\mathbf{x}$ will be. This reduces large oscillations and large changes between the elements of $\mathbf{x}$.

$$\mathbf{D} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{5}$$

Now we can construct an arbitrary order approximation of a given function with the Tikhonov Regression and test its accuracy. We sample the functions in Section 3.1 with added Gaussian noise, and take the regularized least squares regression with the Tikhonov matrix specified.

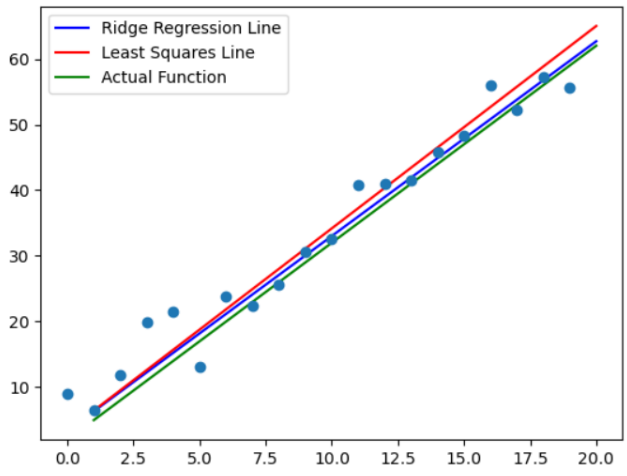

FIGURE 1. Ridge and OLS 1st order fit to a linear function sampled with noise

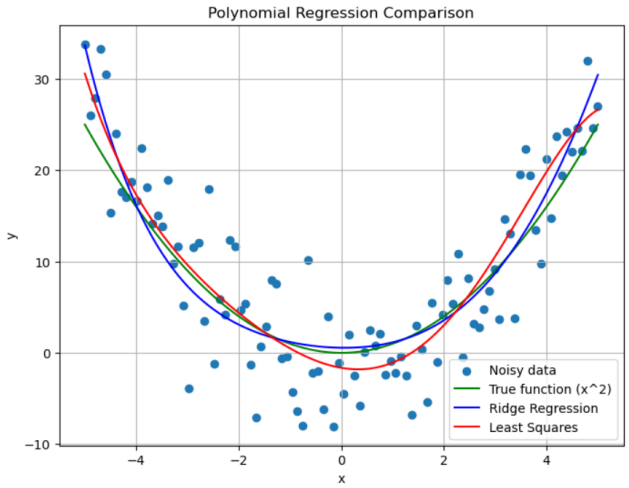

FIGURE 2. 5th order fit of OLS and RR to x^2 with noise

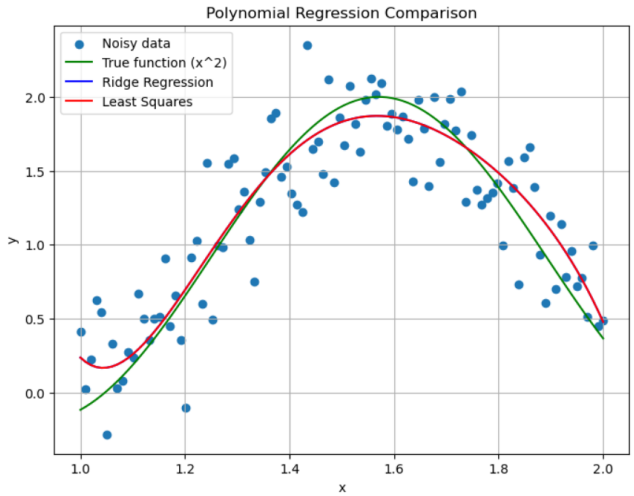

FIGURE 3. 5th order fit of OLS and RR to a sinusoidal function sampled with noise

Similar to our tests with Ridge regression, we searched through a range of λ values when evaluating the Tikhonov Regularization in order to find a value that fits best with the sampled data. To do this, we split the sampled data points into a test and training set, where we trained the model using the training set, then evaluated its performance against the test set. We then selected λ that minimized the error with the test set. This way, we create a regression that is likely to fit the best with any sampled data, and not just the set it was trained on.

Figure 4 shows an example of how Tikhonov Regularization performs compared to OLS when making a linear regression. The function $y = 2x + 3$ was sampled 20 times at regular intervals from $x = 0$ to $x = 20$ with added noise. The noise was sampled as a normal distribution with standard deviation $\sigma = 4$. In this example,

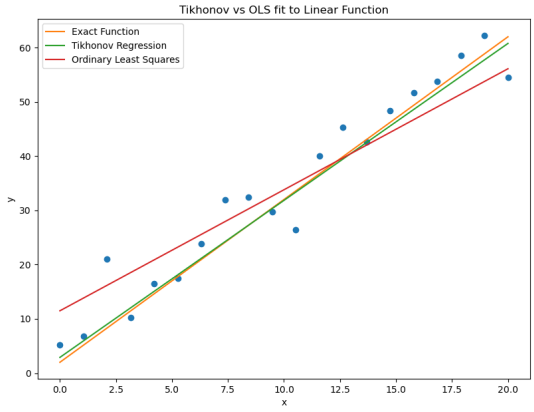

FIGURE 4. 1st Order fit to $y = 3x + 2$ with OLS and Tikhonov

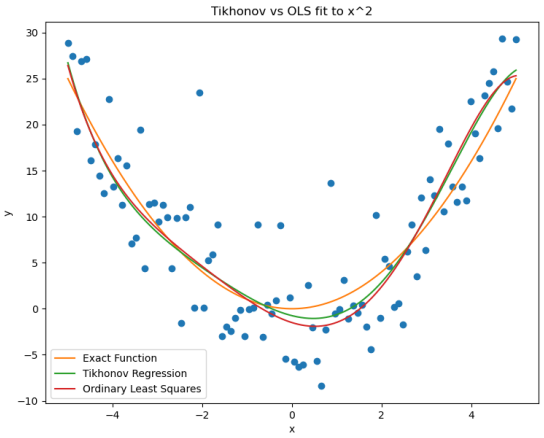

FIGURE 5. 5th order fit of x^2 with OLS and Tikhonov

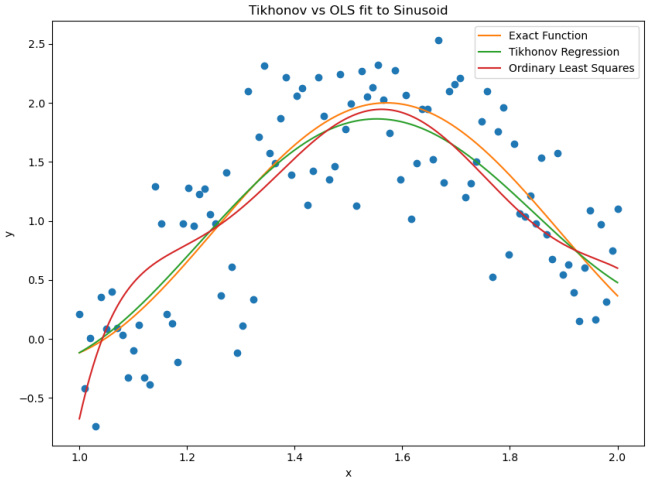

FIGURE 6. 5th order sine fit with OLS and Tikhonov

the ideal λ value was around 26, however other tests returned λ values ranging between 0 and 600. We can see in this case that TR found a solution that is much closer to the exact function, while OLS overfitted to the sampled data and resulted in a much different slope.

Figure 5 compares a 5th order polynomial TR and OLS fit to $y = x^2$, with 100 regular samples along $x \in [-5, 5]$ and sampling standard devation $\sigma = 5$. The ideal λ value found in this case was about 22. However, similar to the linear regressions, ideal values differed significantly with no clear pattern based on the sampled data. This range was similar to the range found when testing linear regressions, not exceeding a factor of 1000. It is less obvious in this example, but when comparing R^2 of the regressions to the exact function, TR performed

slightly better, with $R^2 = 46.37$ compared to $R^2 = 54.81$ with OLS. R^2 was calculated by taking 1000 samples along the exact values for the regression and true functions.

Finally, we tested how TR would behave with a different function basis. For this test, we sampled $y = \sin(x) + \sin(5x)$ for 100 evenly spaced x values between 1 and 2. The noise added had a standard distribution $\sigma = 0.5$. We used the set of basis functions $\sin(nx)$ when calculating the set of regression coefficients, since we knew the exact function is a sum of sine functions. Figure 6 shows the plots of OLS and TR compared to the exact function. Here it is obvious that TR produces an approximation much closer to the exact function we sampled over the given interval. Additionally, the ideal λ value from this example was 0.90, and λ values in other tests only ranged between 0 and 10. It seems the magnitude of the ideal λ value is related to the size of the sampled interval.

From our tests, it is difficult to give a good heuristic method for selecting λ when performing TR. General trends observed show that higher order approximations may require larger λ, as well as an inverse correlation between sample size and λ's magnitude.

4. Intro to Independent Extension

Our independent extension will be to explore LASSO regression. LASSO regression is very similar to ridge regression in that the loss function is the residual sum of squares plus a regularization parameter, γ, times a norm of the coefficients, but the L1 norm of the coefficients is used instead of the L2 norm. The nature of this constraint tends to produce sparser models, or models with more coefficients with values of exactly 0, than other forms of regularization of least squares [7]. In our independent extension we will describe some of the mathematical formulation behind LASSO regression, explain how solutions to LASSO regression are found, and demonstrate how LASSO generates sparser models than ordinary least squares, Ridge Regression, and Tikhonov Regression.

5. Mathematical Formulation for Independent Extension

Similar to Ridge regression, we can take the derivative with respect to x of the loss function, which in LASSO regression is the residual sum of squares plus the L1 norm of the coefficients times the hyperparameter γ.

$$Loss = \sum_{i=1}^{k}((\sum_{j=1}^{n}(A_{ij} \cdot x_j + x_0) - b_i)^2) + \gamma \sum_{j=1}^{n}(x_j) \tag{6}$$

$$= \mathbf{b^T b} - 2\mathbf{x^T A^T b} + \mathbf{x^T A^T A x} + \gamma \| |\mathbf{x^T}| \|_1 \tag{7}$$

We then take the derivative with respect to x as follows.

$$\frac{\partial Loss}{\partial x} = 2\mathbf{A^T Ax} + 2\mathbf{A^T b} + \gamma sgn(\mathbf{I}_0\mathbf{x}) \tag{8}$$

Where $sgn(\mathbf{I}_0\mathbf{x})$ is the sign function with respect to each of the components of $\mathbf{x}$. There does not exist an analytical expression for the solution for LASSO regression like there exists for other forms of regularized least squares regression in the case where there are multiple predictors because the presence of the sign function means that when one tries to solve it for one component, the solution for this component is dependent on all other components[2]. Solutions for LASSO regression are instead found using numerical algorithms[5].

6. Numerical Results for the Independent Extension

6.1. **Single Linear Feature.** In Section 3.1, Ridge regression was applied to generate a model with a single feature. Here, we will replicate this with LASSO regression and compare the results. The line $y = 3x + 2$ was sampled 20 times over the interval [0,20] with Gaussian noise added. These 20 samples were split into 10 training and 10 testing samples and a Ridge, LASSO, and least squares regression model were fit on the training samples. The γ values for the ridge and LASSO models were set at 2.48 and 0.68 respectively based on what values resulted in the highest R^2 value on the test set of the values we tested which were from 0 to 50 in increments of 0.01. Table 1 shows the model parameters, in this case the intercept and the coefficient of x, and R^2 value on the test set for each model.

Table 1. Model Parameters and R^2 for Single Feature

Model	Ordinary Least Squares Regression	Ridge Regression	LASSO Regression
Intercept	6.64	7.23	7.23
x	2.77	2.74	2.74
R^2 on test set	0.955	0.955	0.955

The R^2 values on the test set are the same down to three decimal places for all three models. Ridge and LASSO regression are not particularly helpful in this case because there is only one feature to estimate and it is directly correlated with the response variable.

6.2. **Polynomial Features.** When there are multiple features, particularly when some have little to no true correlation with the output and appear significant in an ordinary least squares regression model due to noise, regularization becomes much more practical. In sections 3.1 and 3.2, a Ridge regression model and a Tikhonov regression model respectively were used to estimate the coefficients of $y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$ when the response data was 100 samples of the function $y = x^2$ with added Gaussian noise. Here, we again fit a LASSO regression model to the same data using the same features. These 100 samples were split into 50 training and 50 testing samples and a Ridge, LASSO, Tikhonov, and least squares regression model were

fit on the training samples. The γ values for the ridge, LASSO, and Tikhonov models were set at 37.1483, 1.0040, and 18.0721 respectively based on what values resulted in the highest R^2 value on the test set of the values we tested which were from 0 to 500. Table 2 shows the parameters, R^2 value, and the norm of the error relative to the true function on the test set for each model.

TABLE 2. Model Parameters and Evaluation Metrics for Polynomial Features

Model	Tikhonov Regression	Ordinary Least Squares Regression	Ridge Regression	LASSO Regression
Intercept	-0.0199	-0.4606	0.0192	0.1986
x	0.2382	0.4034	0.1045	-0.0000
x^2	1.0311	1.1350	0.9669	0.9006
x^3	-0.1013	-0.1279	-0.0786	-0.0520
x^4	-0.0040	-0.0079	-0.0010	0.0017
x^5	0.0045	0.0053	0.0037	0.0027
R^2 on test set	0.8100	0.8088	0.8104	0.8138
Norm of error from true function	22.8005	22.8742	22.7778	22.6632

Figure 7 shows the training and testing data points, the ordinary least squares regression model, the ridge regression model, and the LASSO regression model, the Tikhonov Regression model.

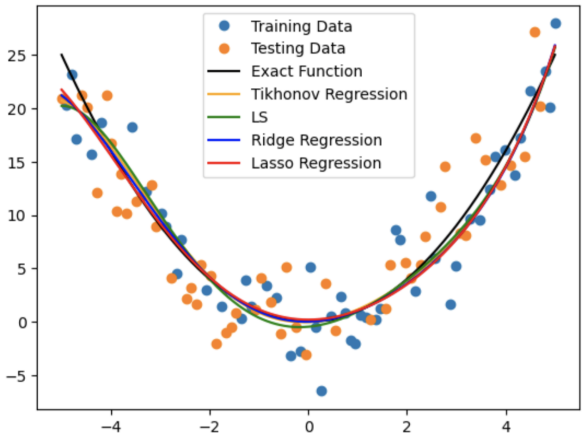

FIGURE 7. Model Comparison for $y = x^2$ with Polynomial Features

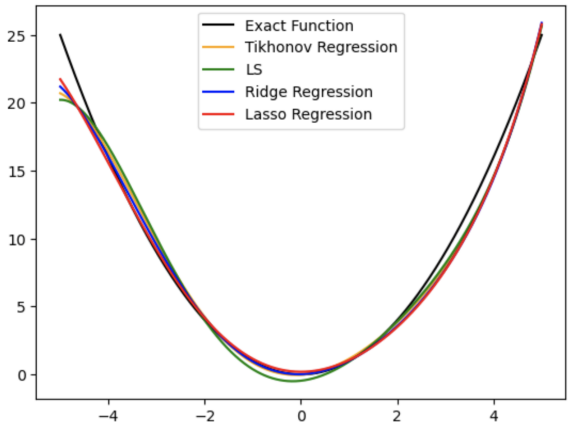

FIGURE 8. Model Comparison for $y = x^2$ with Polynomial Features

Figure 8 shows the same model comparison as Figure 7 but without the data to more clearly illustrate the effect of the regularization on the models. We can see that the OLS model has more significant deviations from the true function than the regularized models.

In this case, all three of the regularized regression models outperform ordinary least squares on the test set and in terms of their proximity to the true function. This effect becomes even more dramatic if the number of points are reduced because the ordinary least squares model overfits the training data even to an even higher degree. In addition, the LASSO model has a 0 coefficient for x while the ridge regression has generally smaller coefficients than OLS for each feature due to the penalty term but still nonzero. This aligns with what was expected that LASSO regression results in sparser models than Ridge regression.

6.3. **Sinusoidal Features.** We also implemented a comparison of Ridge, LASSO, and Tikhonov regression using the example function $y = sin(x) + sin(5x)$ with added Gaussian noise, similar to the experiments in sections 3.1 and 3.2. The coefficients estimated by each of these models are $a * sin(5x) + b * sin(4x) + c * sin(3x) + d * sin(2x) + e * sin(x) + f$, similar to those estimated in the polynomial example but incrementing the periodicity of the sine function from 0 to 5 instead of the degree. The 100 samples of the function with added noise were split into 50 training and 50 testing samples and a Ridge, LASSO, Tikhonov, and least squares regression model were fit on the training samples. The γ values for the ridge, LASSO, and Tikhonov models were set at 1.7836, 0.0200, and 3.6673 respectively based on what values resulted in the highest R^2 value on the test set of the values we tested which were from 0 to 10. Table 3 shows the parameters, R^2 value, and the norm of the error relative to the true function on the test set for each model.

TABLE 3. Model Parameters and Evaluation Metrics for Sinusoidal Features

Model	Tikhonov Regression	Ordinary Least Squares Regression	Ridge Regression	LASSO Regression
Intercept	0.1252	-47.4141	0.6070	0.9084
$sin(x)$	1.0077	57.6306	0.0695	0.0000
$sin(2x)$	-0.4790	-1.9713	-0.3115	-0.0435
$sin(3x)$	0.2717	10.9047	-0.4805	-0.0000
$sin(4x)$	-0.2820	-1.2977	-0.1745	0.0000
$sin(5x)$	1.0202	2.6010	0.7257	1.0119
R^2 on test set	0.5986	0.5451	0.6064	0.6134
Norm of error from true function	3.8686	4.1187	3.8310	3.7967

Figure 9 shows the training and testing data points, the ordinary least squares regression model, the ridge regression model, and the LASSO regression model, the Tikhonov Regression model.

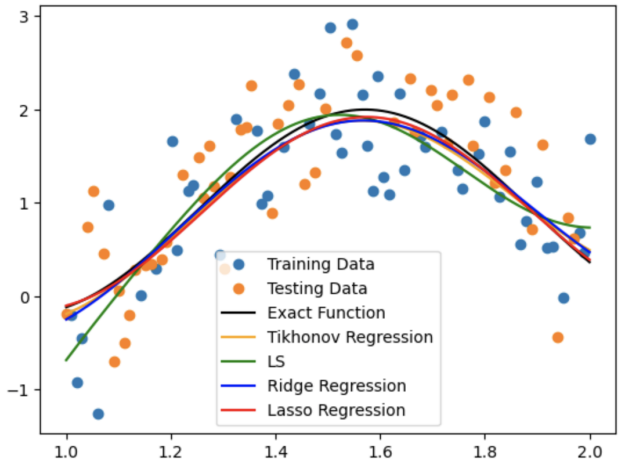

FIGURE 9. Model Comparison for $y = \sin(x) + \sin(5x)$ with Sinusiodal Features

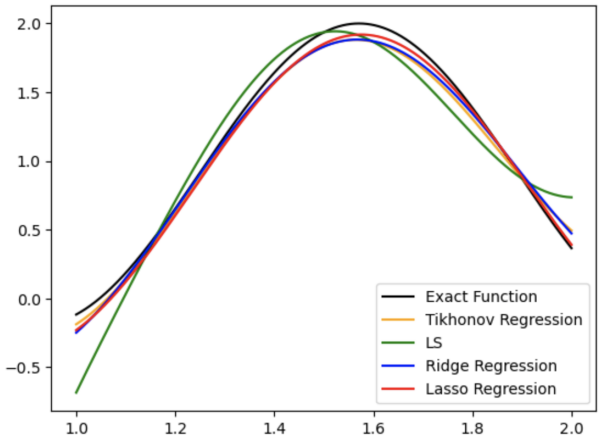

FIGURE 10. Model Comparison for $y = \sin(x) + \sin(5x)$ with Sinusiodal Features

Figure 10 shows the same model comparison as Figure 9 but without the data to more clearly illustrate the effect of the regularization on the models. We can see that the OLS model has more significant deviations from the true function than the regularized models, to an even higher degree than it does in the polynomial example.

In this case, all three of the regularized regression models outperform ordinary least squares on the test set and in terms of their proximity to the true function. This effect becomes even more dramatic if the number of points are reduced because the ordinary least squares model overfits the training data to an even higher degree. It is worth noting that the ordinary least squares model also had coefficients that differ much more from the true function than the coefficients of the regularized models, suggesting that the ordinary least squares model would generalize far worse to data outside of the range of x values, in this case $[1, 2]$. In addition, the LASSO model has a 0 coefficient for $\sin(3x)$ and $\sin(4x)$ while the ridge regression has generally smaller coefficients than OLS for each feature due to the penalty term but still nonzero. This aligns with what was expected that LASSO regression results in sparser models than Ridge regression. However, it should also be noted that the LASSO model, despite performing the best of all the models in this experiment both in terms of accuracy on the test set and proximity to the true function, it also had a 0 coefficient on the $\sin(x)$ term which was not zero in the true function. This illustrates that LASSO regression can be prone to drop the coefficients of features that are truly significant to 0 in addition to those that aren't truly significant when the distinction between each feature is small. This is the case in this example because the range of x values is relatively small so the periodicity of the sinusoidal terms is not as evident in the data.

7. Discussion and Conclusion

These results demonstrate some of the valuable aspects of each of these types of regularization of least squares regression. All showed meaningful improvements in solving certain problems where there is significant noise. This can be very useful when dealing with rank deficient problems, or problems where the number of features is greater than the number or rows of data. Performing ordinary least squares does not always result in a unique solution when the system is rank deficient. This is because the solution to an ordinary least squares problem is $\mathbf{x}^* = (\mathbf{A^T A})^{-1}\mathbf{A^T b}$ and $\mathbf{A^T A}$ is at most rank p thus degenerate and non-invertible. Ridge regression, however, has a solution of $\mathbf{x}^* = (\mathbf{A^T A} + \gamma\mathbf{I})^{-1}\mathbf{A^T b}$ as shown in section 2.2 and the addition of $\gamma\mathbf{I}$ to $A^T A$ lowers its condition number and makes its inversion numerically stable [8]. Another application of regularized least squares is to select the most important features in a data set to be included in a more advanced machine learning model so that this model is less likely to overfit the training data. LASSO regression is particularly useful for this because as demonstrated in Section 6.2, it results in sparser models than other forms of regularized least squares. While our work demonstrated the effects of regularized least squares models in relatively simple problems that are easy to visualize, in the future it would be useful to test these findings on real world data sets where the number of features is much larger and the interactions between features are more complex.s

References

[1] Hamzah NA Gasparini M Arashi M, Roozbeh M, *Ridge regression and its applications in genetic studies*, PLoS One **16(4)** (2021), no. e0245376.

[2] Niharika Gauraha, *Introduction to the lasso*, 2018.

[3] Per Christian Hansen Gene H. Golub and Dianne P. O'Leary, *Tikhonov regularization and total least squares*, SIAM Journal on Matrix Analysis and Applications **21** (1999).

[4] Penn State Eberly College of Science, *Applied data mining and statistical learning - 5.1 ridge regression*, 2018, Last accessed 2 December 2023.

[5] Ritwick Roy, *Regularization in machine learning*, 2022.

[6] Forrest Stout and John H. Kalivas, *Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts*, Journal of Chemometrics (2006).

[7] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) **58** (1996), 267–288.

[8] Binxu Wang, *Why does least-squares need regularization?*, 2022.