# CSI 5340 Homework Exercise 1

Name: Jayshil Patel

Student Number 300312550

## Introduction

The problem at hand involves understanding the relationship between a real-valued random variable X and its corresponding dependent variable Y, which is affected by both a cosine function of X and a zero-mean Gaussian random variable Z.

The objective is to design and analyze polynomial regression models without prior knowledge of the true underlying relationship between X and Y. Specifically, we seek to investigate the impact of model complexity, sample size, and noise variance on the model's fitting capabilities and generalization performance. The experiment is structured to evaluate the Mean Squared Error (MSE) as a metric for both in-sample ($\bar{E}_{in}$) and out-of-sample ($\bar{E}_{out}$) performance, shedding light on the trade-offs and optimizations necessary for constructing effective regression models.

## Initial considerations

Because of Efficient computation and almost similar outcomes, I have used Mini-Batch Gradient descent because that was the fastest when compared to Gradient Descent and Stochastic Gradient descent. The Learning Rate for all the gradient descent was 0.001 and did 2000 iterations for all 3 types of Gradient Descent methods. The batch size for Mini-Batch Stochastic Gradient Descent was 50.

## Effect of Dataset Length on Mean-Squared Error with a variance of 0.2. And different model complexities. **No Regularization**

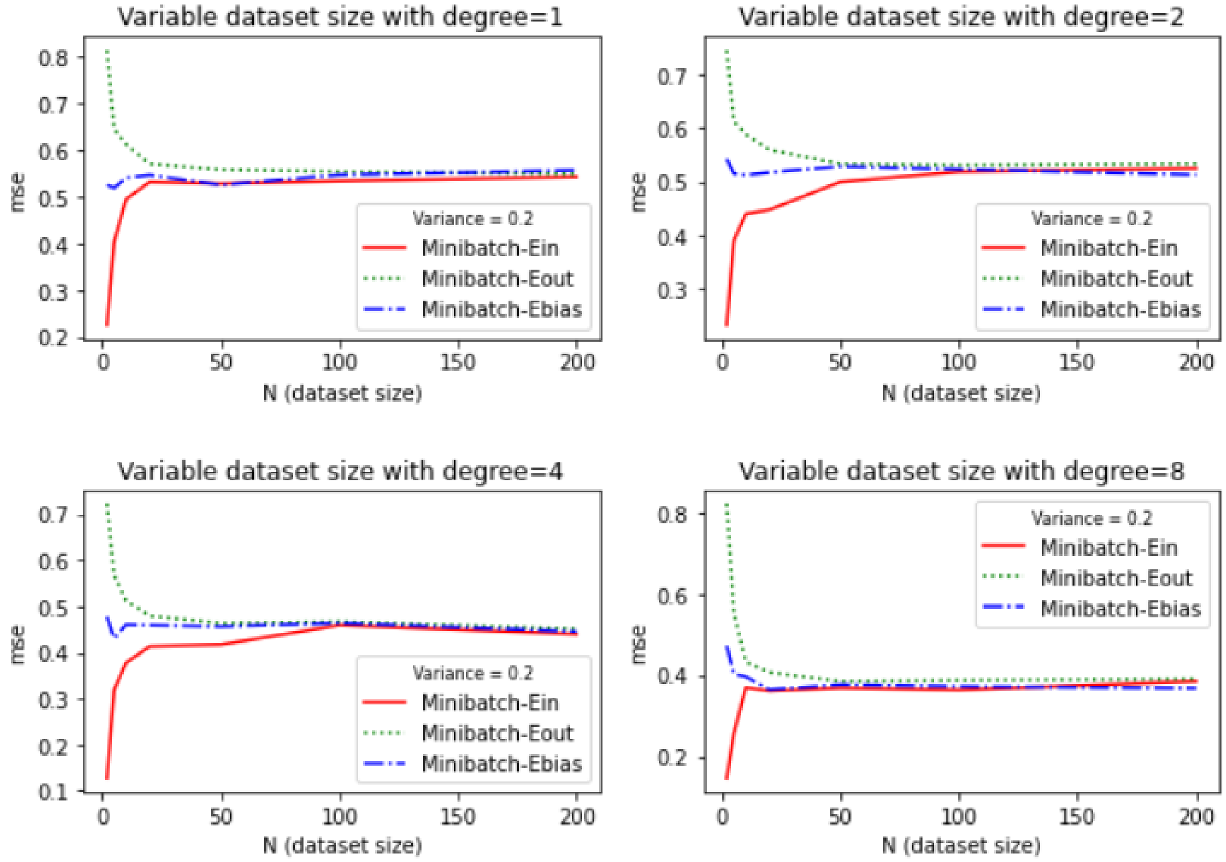


*Figure 1*

Here in Figure 1, we can observe that $E_{in}$(In-sample error or Training error) and $E_{out}$(Out-of-sample error or Testing error) converge at MSE = 0.6 to 0.5 for degrees 1,2 and 4. The convergence level reduces when the degree of complexity of the model increases to degree 8. For models with less complexity (degrees 1, 2and 4) the MSE was not high, but definitely higher than degree 8 because of underfitting, which indicates the model is too generalized to capture the features as well, and that is the reason why lower complexity models converge at higher MSE when compared to higher complexity models.

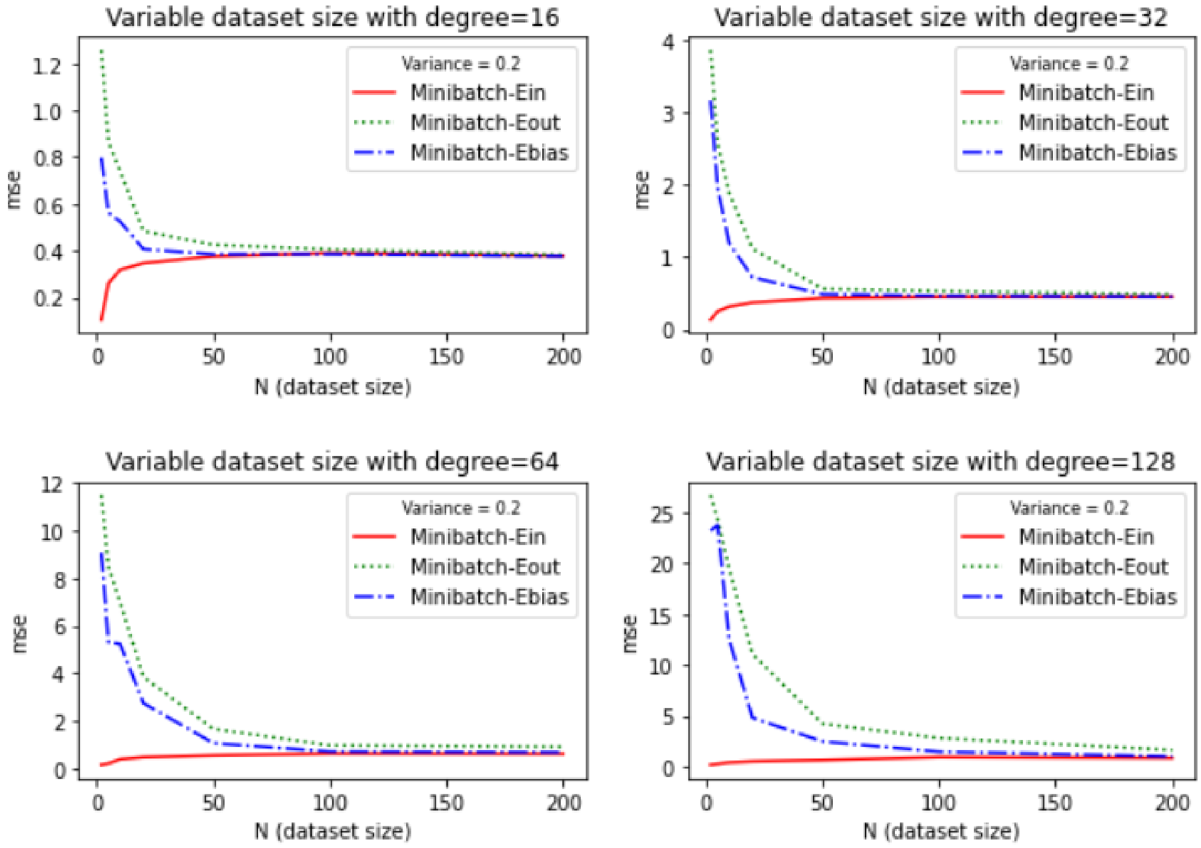


*Figure 2*

As seen in the following figure (Figure 2) the Ein and Eout converge and remain the same at ≈ 0.4 for degree 16. And this level increases to almost 1 for degrees 32 and 64. For model complexity 128, it is very high (around 5). This indicates overfitting of the model which tends to capture the noise as well. When MSE is compared to dataset size, we can observe that for smaller dataset the $E_{in}$ and $E_{out}$ is very high and they both reduces when the dataset size increases. However, datasets larger than 100 are unnecessary because no major changes can be observed, and takes a lot of computation time. For higher complexity models the convergence of $E_{in}$ and $E_{out}$ takes more datasets because of overfitting and converges

Effect of Dataset length on Mean-Squared Error with variance of 0.05. And different model complexities. **No Regularization**

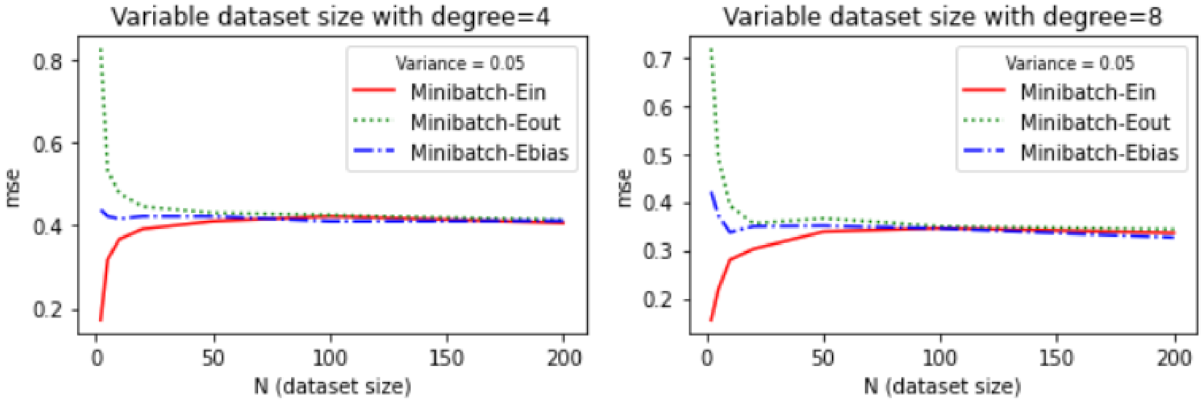


*Figure 3*

From Figure 3 we can observe that the convergence of $E_{in}$ and $E_{out}$ takes less dataset because of low noise or variance in the dataset and convergence is smoother when compared to Variance = 0.2. An interesting trend in Figure 4 is observed for model complexity = 64 and variance 0.05 the graph of $E_{bias}$ looks smoother when compared to $E_{bias}$ for model complexity = 64 and variance 0.2 because of less noise. One more observation is the decrease in $E_{bias}$ for model complexity = 8 (variance = 0.2 vs variance = 0.05) we can observe that for lower variance or noise in the dataset, the decrease in $E_{bias}$ is quicker(takes less number of data) when compared to higher noise graph.

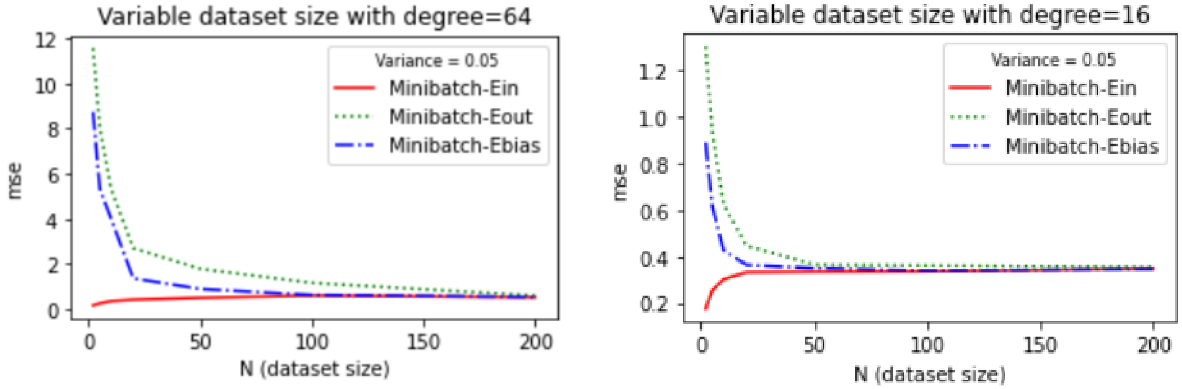


*Figure 4*

Effect of Dataset Length on Mean-Squared Error. And different model complexities. **With Regularization**.

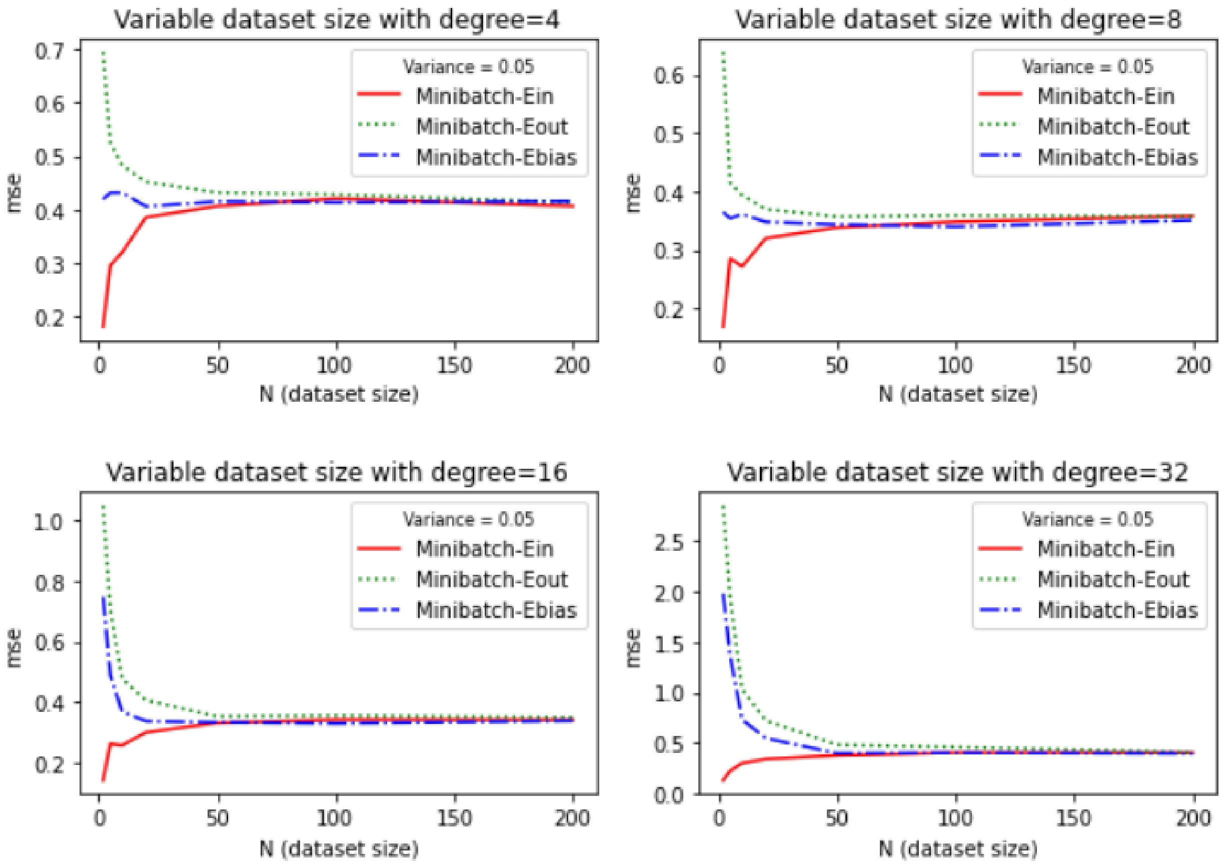
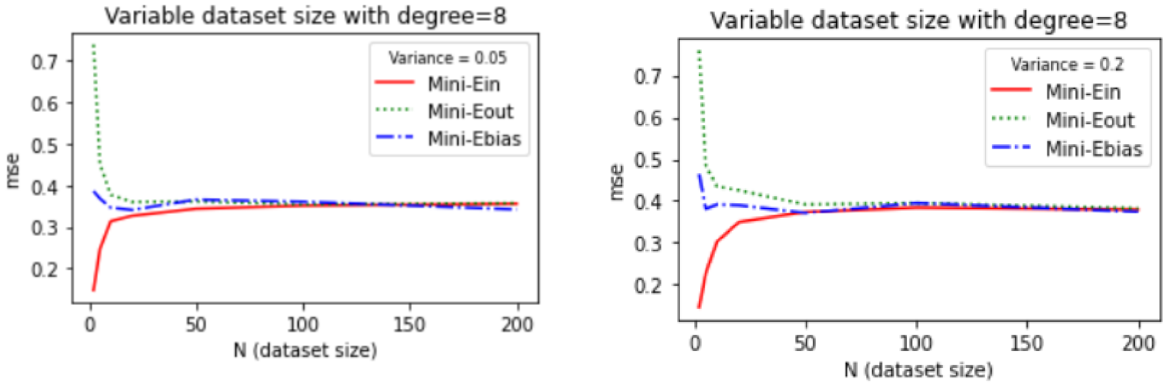


*Figure 5*

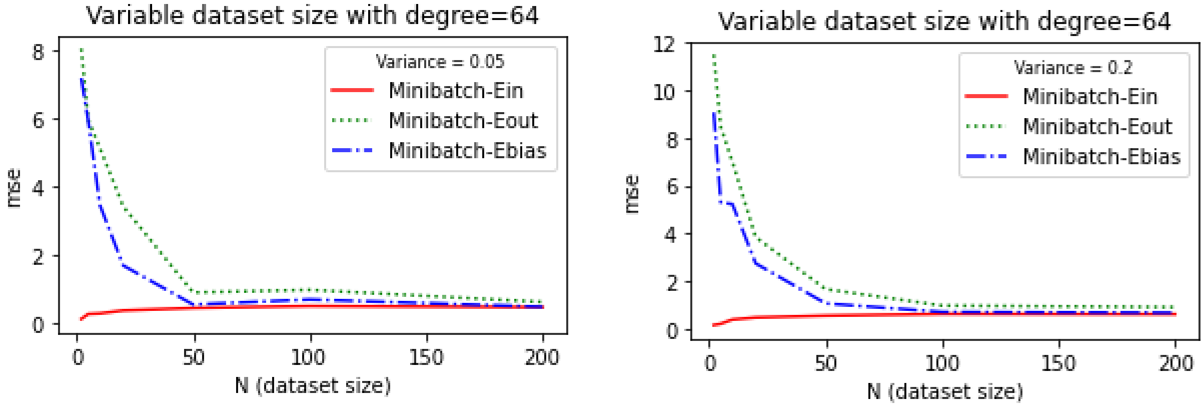
Here from Figure 5, we can observe that the graphs are smoother when compared to the graphs with no regularization of the same model complexity and variance. An interesting observation here is the trend in the MSE vs Dataset size graph for variance = 0.05 and model complexity = 8 the $E_{bias}$, $E_{in}$, $E_{out}$ merges to a lower value of MSE compared to the one with the same parameters and no regularization as shown in Figure 1. For higher complexity models, regularization works better. In Figure 5, we can observe that the model complexity for degree = 32 for the Errors converges to 0.5 whereas the graph for error converges above MSE = 0.5 in Figure 1 for the non-regularized model and the convergence curve is smoother.



*(a)REGULARIZED* *(b)NON-REGULARIZED*

*Figure 6*

Comparing Figure 6(a) and Figure 6(b) we can observe a clear difference in noise and regularization on the model. Figure 6(a) is a regularized model with a variance of 0.05 and Figure 6(b) is of model with no regularization and a variance of 0.2. The convergence of Ein and Eout is smoother for 6(a) when compared to 6(b).

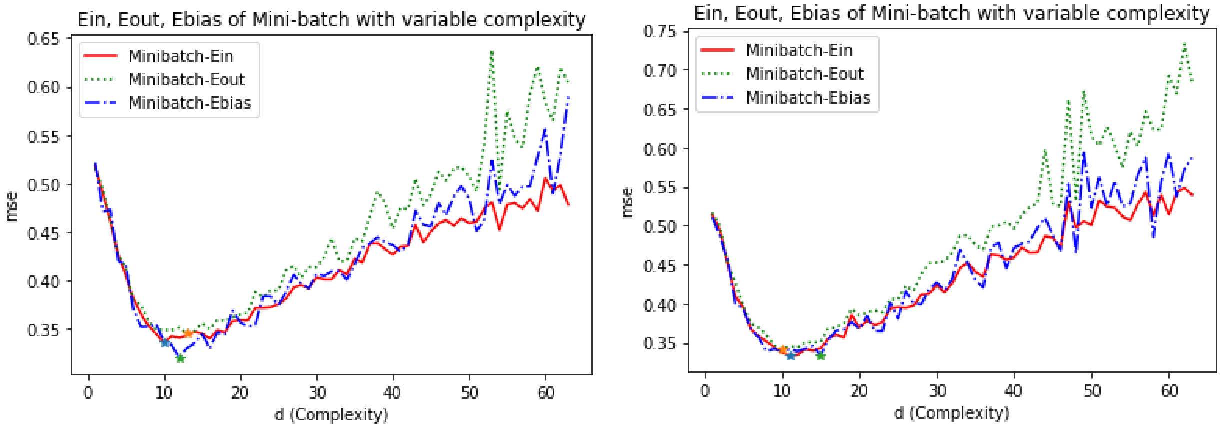


*(a)REGULARIZED* *(b)NON-REGULARIZED*

*Figure 7*

The Same observation can be made on higher complexity models of degree 64. Here the $E_{bias}$ curve for the regularized model is smoother when compared to the non-regularized model for the different Variance in the dataset.

# Effect of Model Complexity on MSE

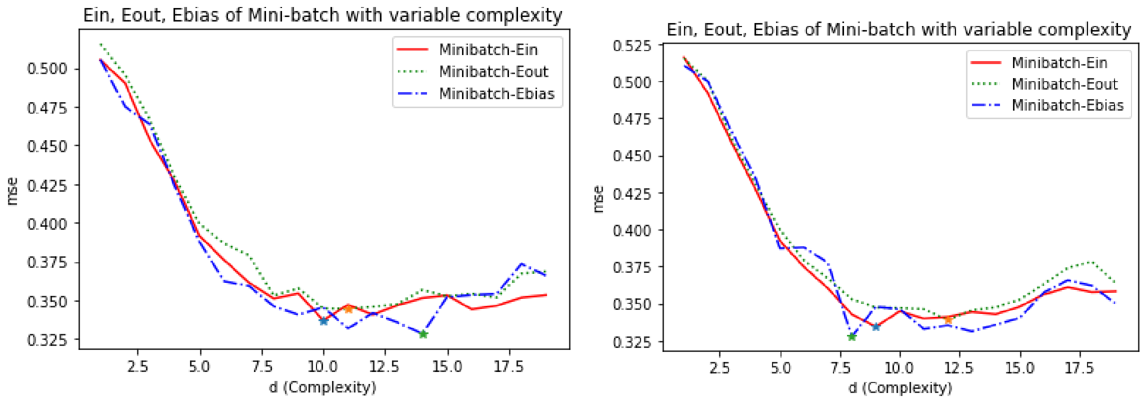


*(a)REGULARIZED*

*(b)NON-REGULARIZED*

*Figure 8*

The plot illustrates the Mean Squared Error (MSE) as a function of model complexity (d) for a regularized model. As the complexity of the model increases, represented by an increasing number of features or higher degrees of polynomials, the MSE initially decreases. This reduction in MSE indicates better fitting to the training data. However, beyond a certain point, the MSE starts to increase as the model becomes overly complex and begins to overfit the training data. Regularization helps prevent excessive model complexity by imposing a penalty on the size of the model coefficients.



*(a)REGULARIZED*

*(b)NON-REGULARIZED*

*Figure 9*

The comparison highlights that the regularized model achieves the minimum MSE at a later point, demonstrating the effectiveness of regularization in delaying overfitting and finding an optimal trade-off between model complexity and performance. On the other hand, the non-regularized model reaches the minimum MSE earlier but is more susceptible to overfitting as complexity increases.

# Effect of Noise on MSE

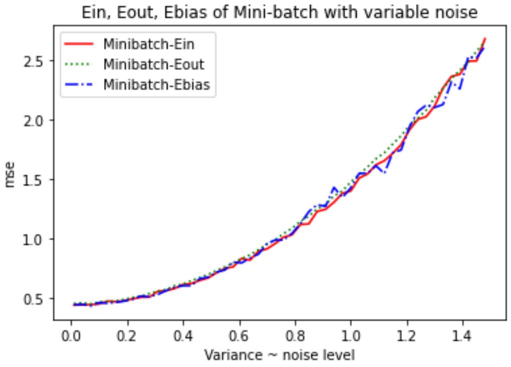
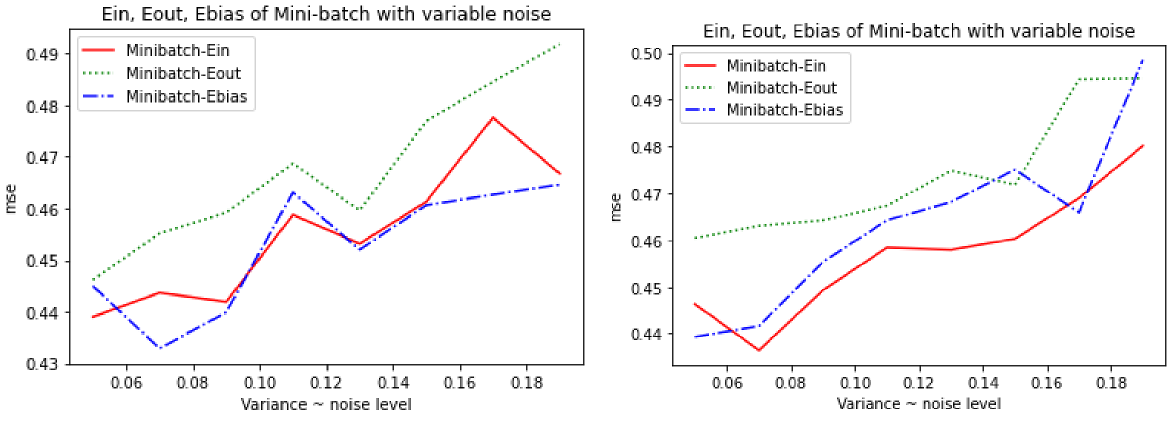


*Figure 10*

The plot showcases the relationship between the Mean Squared Error (MSE) and increasing variance for a given model. As variance increases, the MSE consistently rises for all three error metrics: in-sample error ($E_{in}$), out-of-sample error ($E_{out}$), and bias error ($E_{bias}$). This indicates that higher variance in the model leads to increased errors across both the training set ($E_{in}$) and unseen data ($E_{out}$), as well as an increase in the bias error, representing the difference between the true model and the expected model.



*(a)NON-REGULARIZED* *(b)REGULARIZED*

*Figure 11*

Regularized Model: The plot for the regularized linear regression model showcases smoother curves as the variance increases. This smoothness is attributed to the effect of regularization, which helps to prevent extreme fluctuations in the model's performance, even with varying levels of noise.

Non-Regularized Model: In contrast, the plot for the non-regularized linear regression model exhibits more pronounced and fluctuating curves as variance increases. Without regularization, the model is more sensitive to noise, resulting in larger fluctuations in MSE as the noise levels rise.

## Error comparison with variance for different types of Gradient Descent

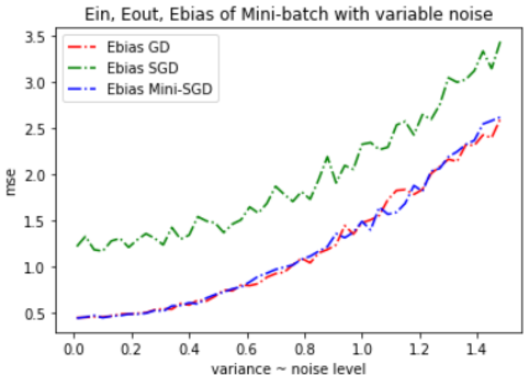


*Figure 12*

**Gradient Descent (GD):** Utilizes the complete dataset for precise gradient estimation, resulting in lower bias error, even with increased noise.

**Stochastic Gradient Descent (SGD):** Operates on one example at a time, leading to more frequent but noisy updates, causing a higher bias error, especially with heightened noise.

**Minibatch Gradient Descent:** Strikes a balance by using a small batch of examples for updates, achieving a bias error lower than SGD but slightly higher than GD, especially in noisier scenarios.

In summary, Gradient Descent (GD) excels in maintaining a lower bias error (Ebias) by virtue of its precise estimation of the gradient derived from the complete dataset. Conversely, Stochastic Gradient Descent (SGD) is more susceptible to noise due to its stochastic updates, potentially resulting in a higher bias error. Minibatch Gradient Descent bridges these extremes by utilizing a batch of examples for updates, striking a balance that yields a bias error lower than SGD but marginally higher than GD, especially when the variance is heightened. These nuances underscore the critical role of the chosen optimization algorithm and its sensitivity to noise levels in the dataset.

# Generalization Error of Regularized and Non-regularized Models

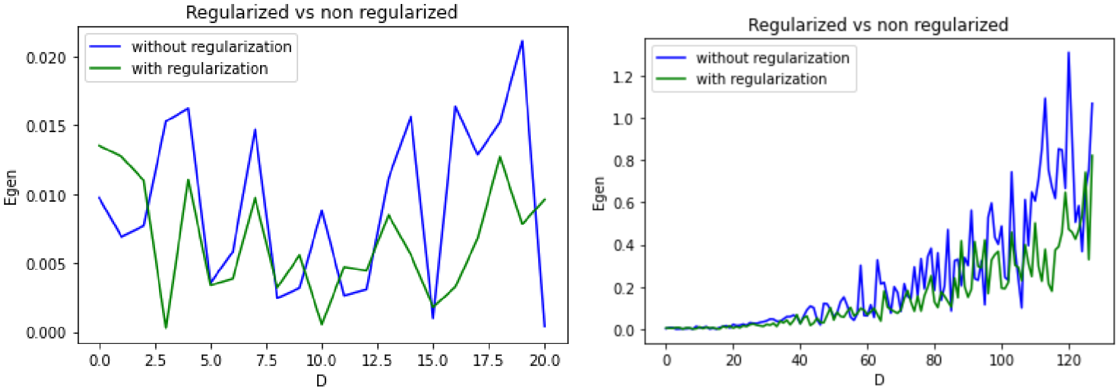


*Figure 12*

The plot illustrates the relationship between generalization error and model complexity for both a regularized and a non-regularized model.

**Regularized Model:**

- The generalization error gradually decreases with increasing model complexity, reaching its lowest points at certain levels of complexity.
- The regularization mechanism effectively controls overfitting, allowing the model to achieve lower generalization error as it becomes more complex.
- The lowest generalization error is attained with a moderate level of model complexity, showcasing the regularization's role in finding a balance between complexity and generalization.

**Non-Regularized Model:**

- In contrast, the generalization error for the non-regularized model exhibits a different trend.
- As the model complexity increases, the generalization error initially decreases, showcasing better fitting to the training data.
- However, beyond a certain point, the generalization error begins to rise as the model becomes overly complex and starts to overfit the training data.