Ridge and LASSO Regression

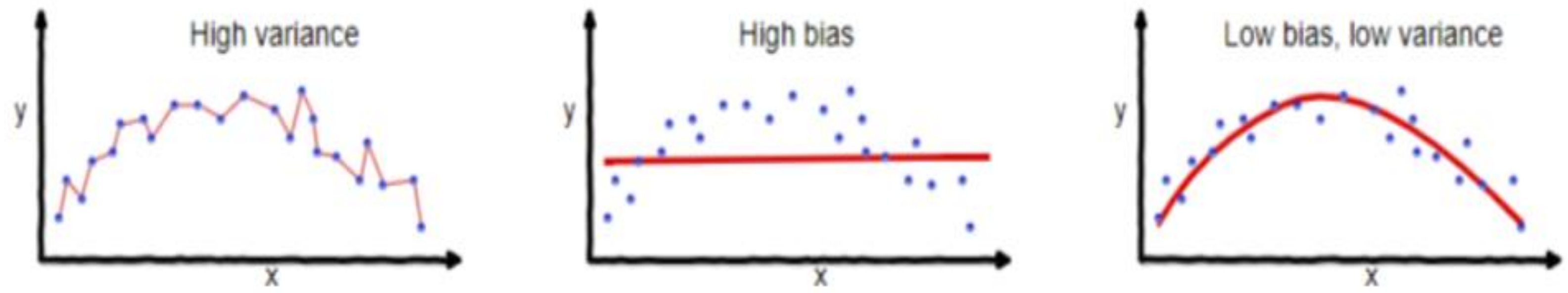
By Dr. Bidisha Bhabani

Bias

- **Bias:** Biases are the underlying assumptions that are made by data to simplify the target function.
- Bias does help us generalize the data better and make the model less sensitive to single data points.
- It also decreases the training time because of the decrease in complexity of target function High bias suggest that there is more assumption taken on target function.
- This leads to the underfitting of the model sometimes.
- Examples of High bias Algorithms include Linear Regression, Logistic Regression etc.

Variance

- **Variance:** In machine learning, Variance is a type of error that occurs due to a model's sensitivity to small fluctuations in the dataset.
- The high variance would cause an algorithm to model the outliers/noise in the training set.
- This is most commonly referred to as overfitting.
- In this situation, the model basically learns every data point and does not offer good prediction when it tested on a novel dataset.
- Examples of High variance Algorithms include Decision Tree, KNN etc.



overfitting   underfitting   Good balance

Regularization

- Let us consider that we have a very accurate model, this model has a low error in predictions and it’s not from the target (which is represented by bull’s eye).
- This model has low bias and variance.
- Now, if the predictions are scattered here and there then that is the symbol of high variance, also if the predictions are far from the target then that is the symbol of high bias.
- Sometimes we need to choose between low variance and low bias.
- There is an approach that prefers some bias over high variance, this approach is called **Regularization**.
- It works well for most of the classification/regression problems.

Cost Function of a linear regression model

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h\theta(x)^i - y^i)^2$$

$h\theta(x)^i$ → Predicted Points

y^i → Actual Points

$\sum_{i=1}^{m} (h\theta(x)^i - y^i)^2$ → Mean Squared Error

Ridge Regression

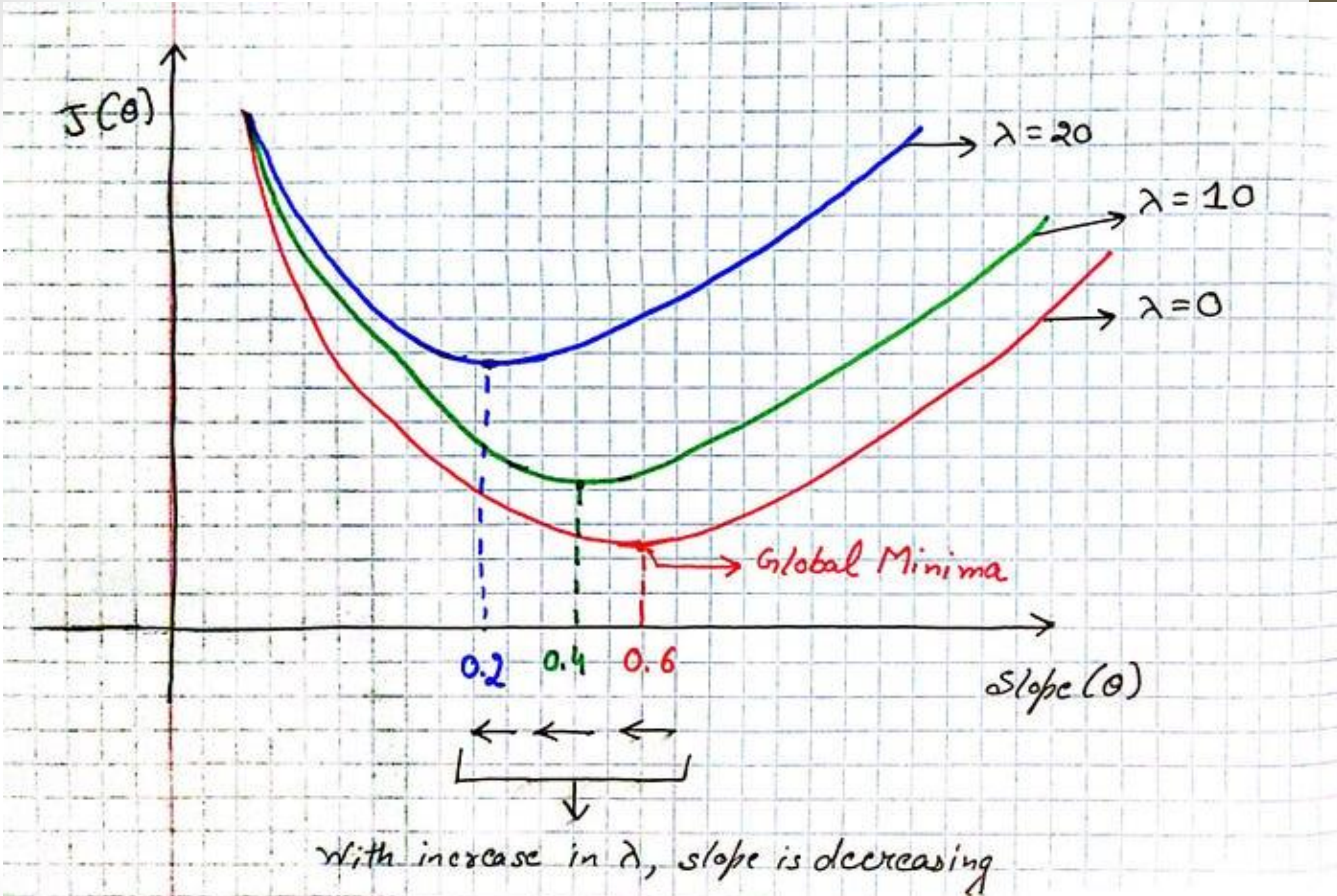
- To combat the issue of overfitting in linear regression models, ridge regression is a regularization approach.
- The size of the coefficients is reduced and overfitting is prevented by adding a penalty term to the cost function of linear regression.
- The penalty term regulates the magnitude of the coefficients in the model and is proportional to the sum of squared coefficients. The coefficients shrink toward zero when the penalty term's value is raised, lowering the model's variance.

$$Cost\ Function = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x)^i - y^i\right)^2 + \lambda\,(slope)^2$$

$$\lambda = Hyperparameter$$

Ridge Regression

- where y is the actual value, h(x) denotes the predicted value
- If λ = 0 then the cost function of Ridge Regression and the cost function of linear regression is the same.
- If even adding the hyperparameter to the original cost function the training accuracy won't improve then it will keep on changing the value of λ and try to find the best λ parameter.
- As the cost function is changing after adding a hyperparameter, the model will also change the best-fit line.
- **Relationship between λ and slope**
- The slope is inversely proportional to the λ. This means as the hyperparameter λ increases slope θ decreases and vice-versa.


J(θ)
λ=20
λ=10
λ=0
Global Minima
0.2
0.4
0.6
slope(θ)
With increase in λ, slope is decreasing

Limitation of Ridge Regression

- **Limitation of Ridge Regression:** Ridge regression decreases the complexity of a model but does not reduce the number of variables since it never leads to a coefficient be zero rather only minimizes it. Hence, this model is not good for feature reduction.

Lasso Regression

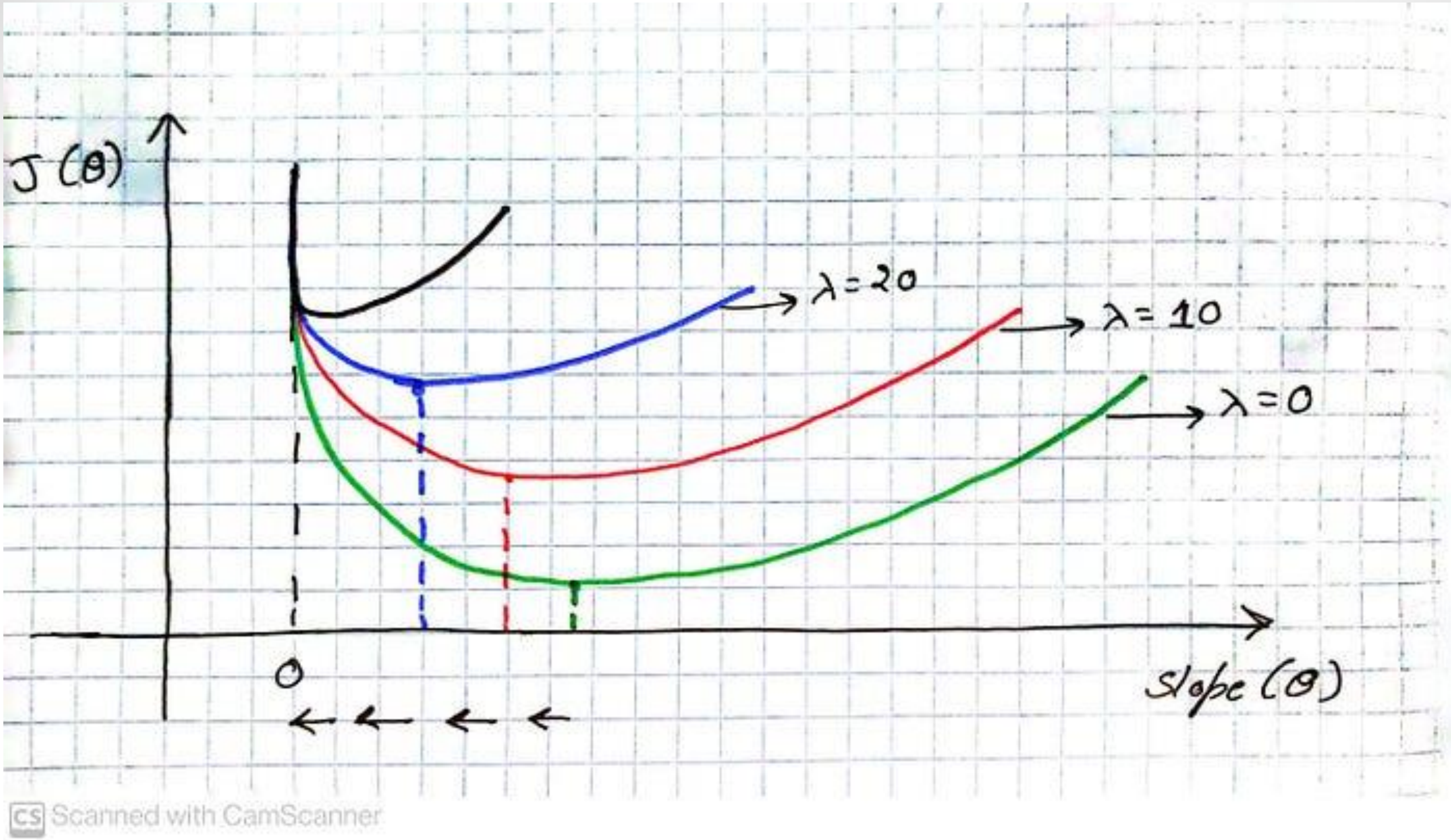
- Lasso regression stands for Least Absolute Shrinkage and Selection Operator.
- It adds penalty term to the cost function. This term is the absolute sum of the coefficients.
- As the value of coefficients increases from *0* this term penalizes, cause model, to decrease the value of coefficients in order to reduce loss.
- The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.
- *The main aim of Lasso Regression is to reduce the features and hence can be used for Feature Selection.*

Lasso Regression

$$Cost\ Function = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta(x)^i - y^i\right)^2 + \lambda \sum_{i=1}^{n}|slope|$$

$$\lambda = Hyperparameter$$

- It is used to reduce the features because in this case slope can become zero.



- Here, with the increase in the value of λ, the slope θ is decreasing.
- But when we will increase the value of λ again and again, at one point the slope will become 0 (hence feature will be removed), and even after that if we will increase the value of λ, the slope will be stuck at 0.

Limitation of Lasso Regression:

- Lasso sometimes struggles with some types of data. If the number of predictors *(p)* is greater than the number of observations *(n)*, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant (or may be used in the test set).
- If there are two or more highly collinear variables then LASSO regression select one of them randomly which is not good for the interpretation of data

Difference b/w Ridge and Lasso Regression

- The main difference between Ridge and Lasso regression is the way they shrink the coefficients. Ridge regression can reduce all the coefficients by a small amount but Lasso can reduce some features more than others and hence can completely eliminate those features.
- Both Ridge and Lasso regression are quite popular algorithms. Ridge regression is used to avoid Overfitting while Lasso Regression can be used for feature selection. Both are regularization techniques. Both methods are useful in their own way and we can choose one over the other according to our needs.