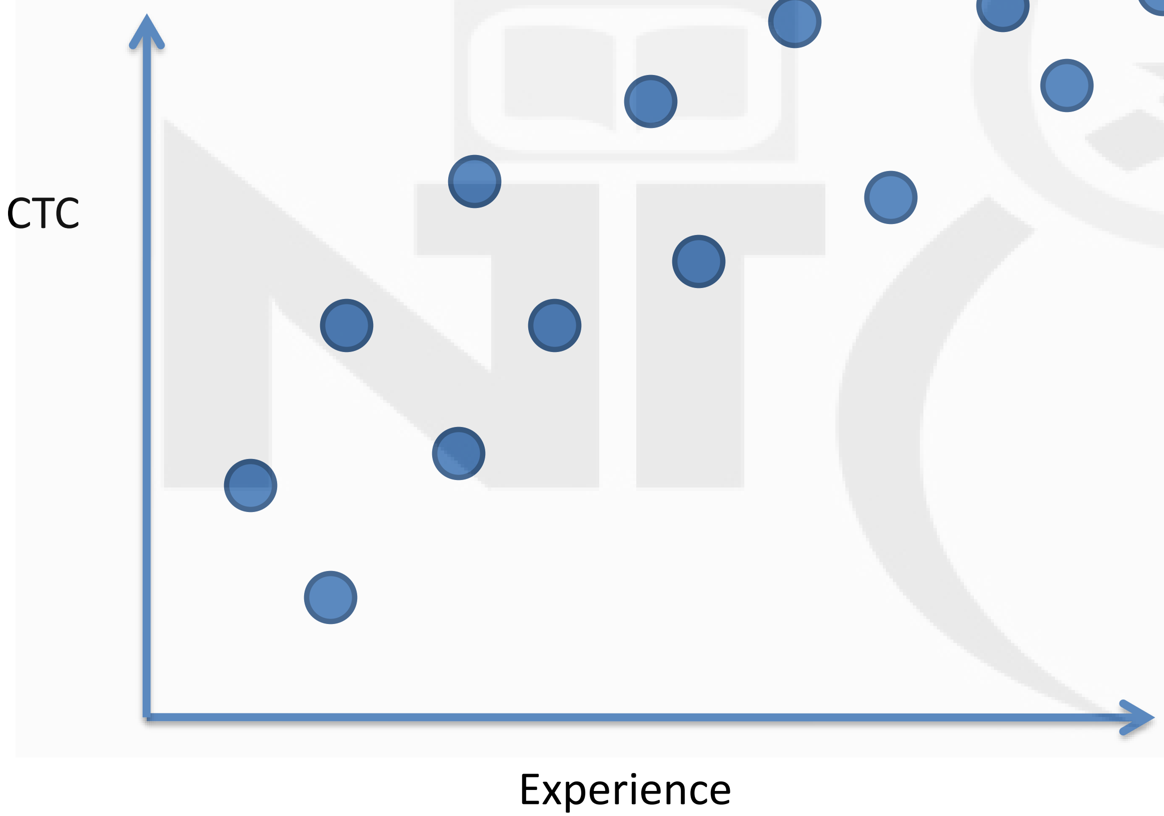
REGULARIZATION

-MUKESH KUMAR

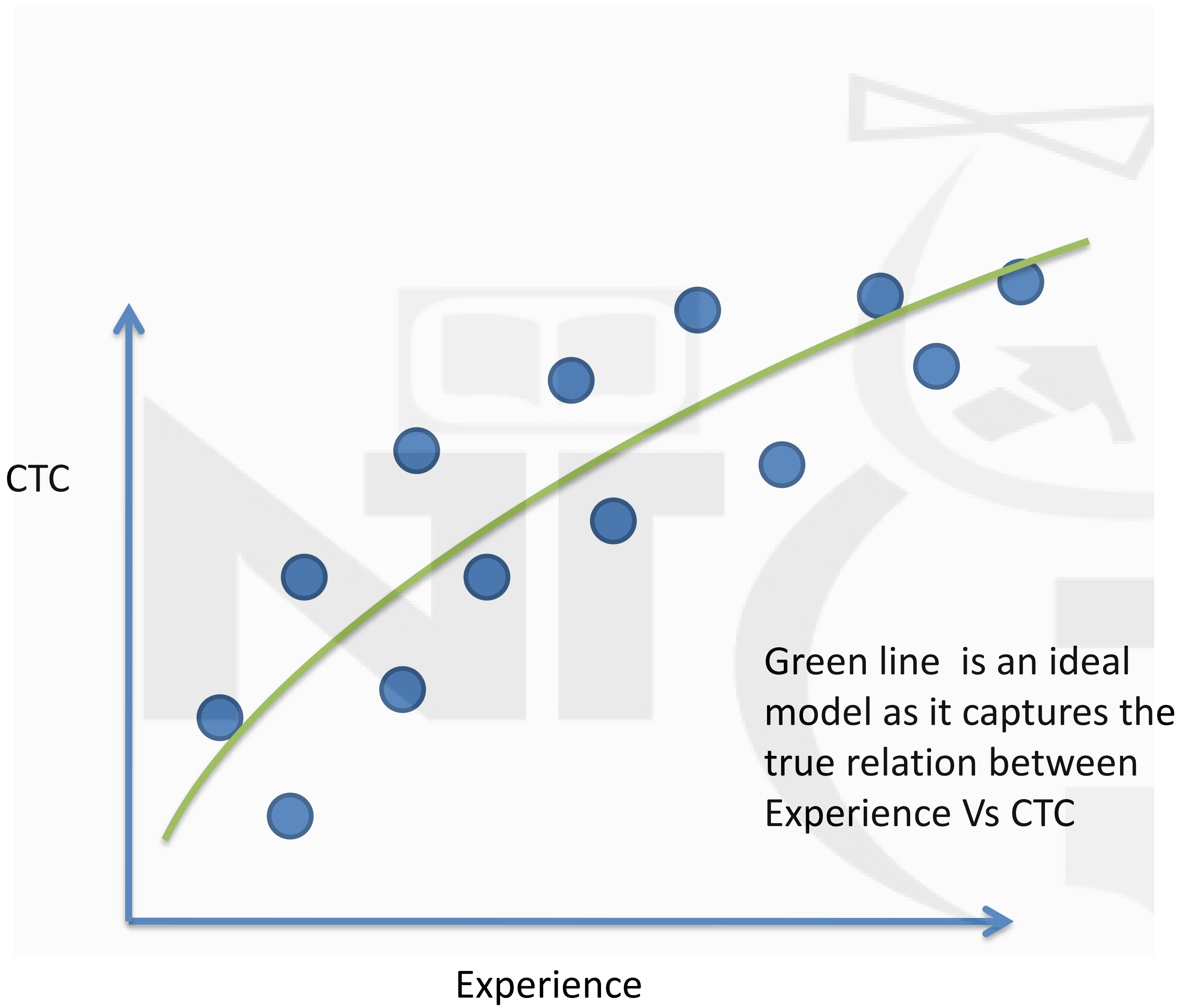
AGENDA

- Bias &Variance
- Overfit & Underfit
- Regularization
- Ridge Regression
- Lasso Regression

BIAS & VARIANCE

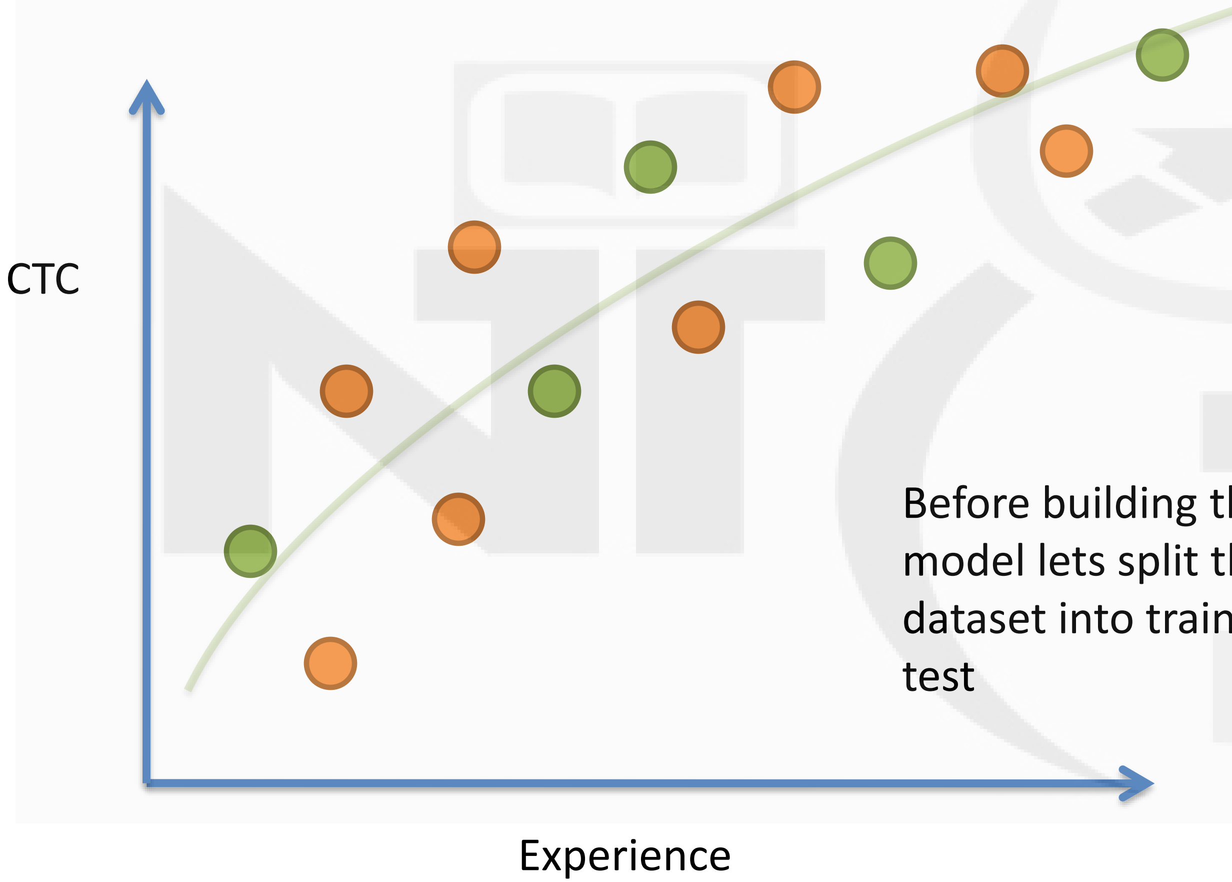

CTC
Experience

- To find the relationship between **CTC and Experience** lets build two models


CTC
Experience
Green line is an ideal model as it captures the true relation between Experience Vs CTC

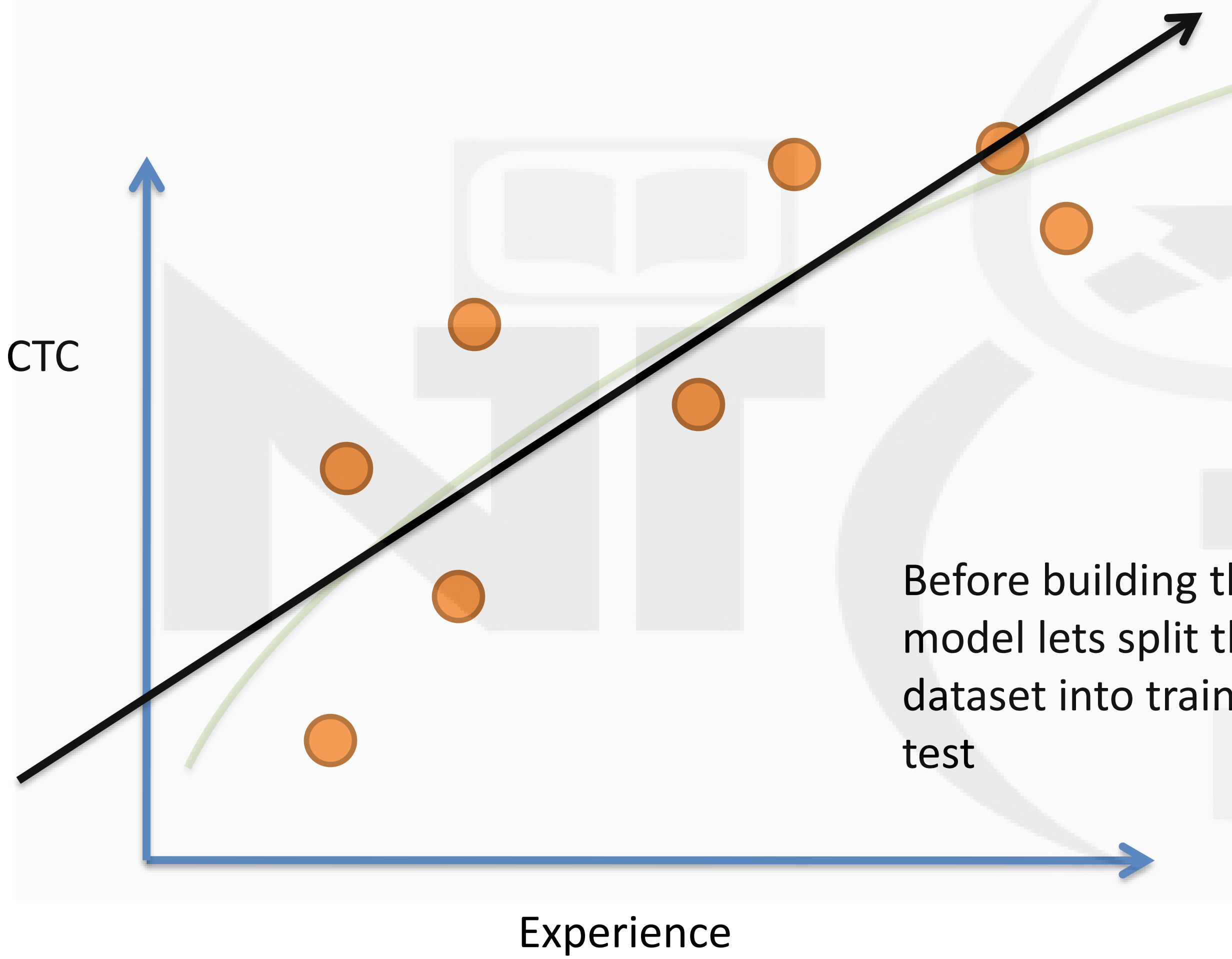
Orange Dots – Training data

Green Dots – Testing data

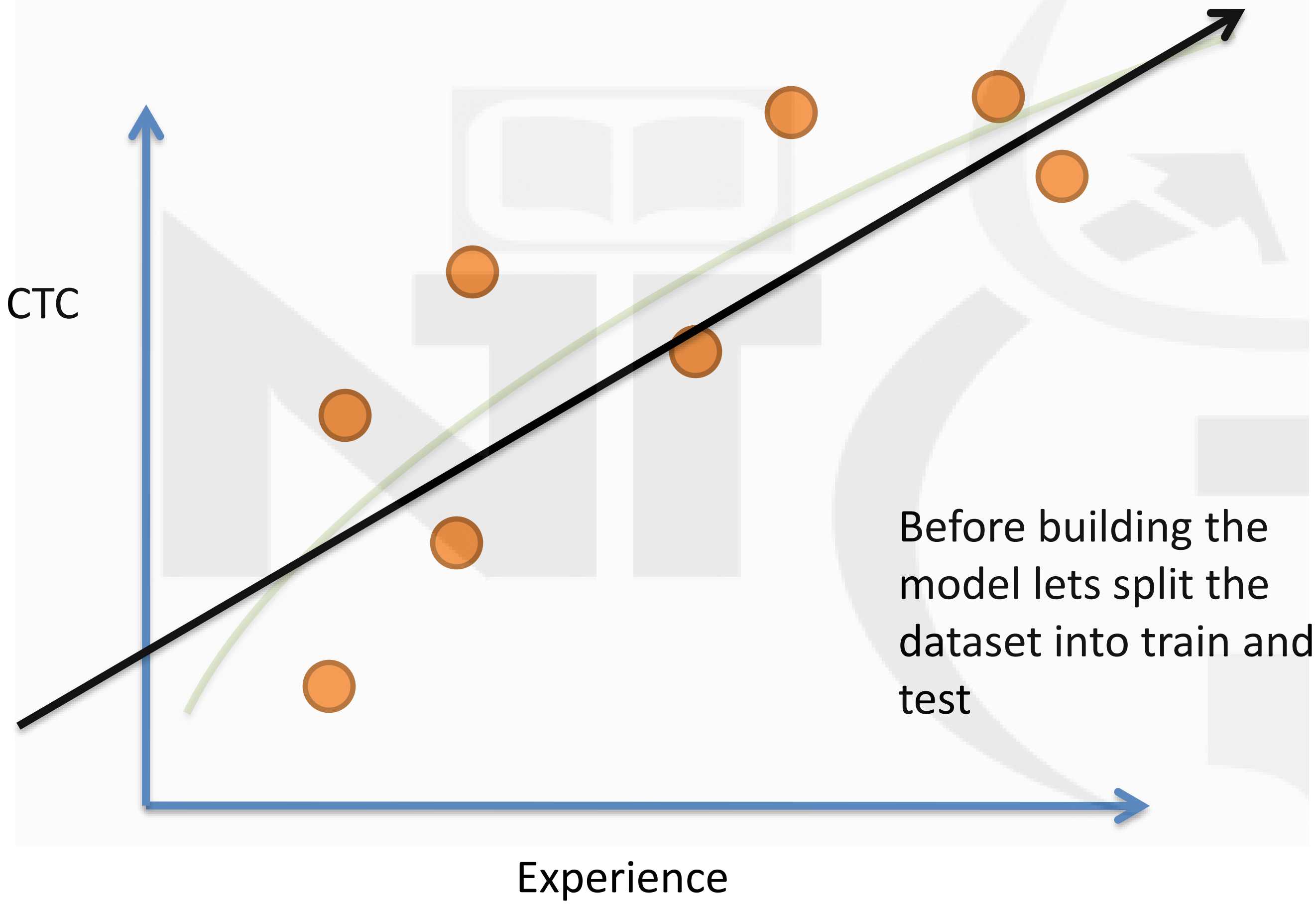


Before building the model lets split the dataset into train and test

Black line represents LR model



Before building the model lets split the dataset into train and test



Before building the model lets split the dataset into train and test

BIAS

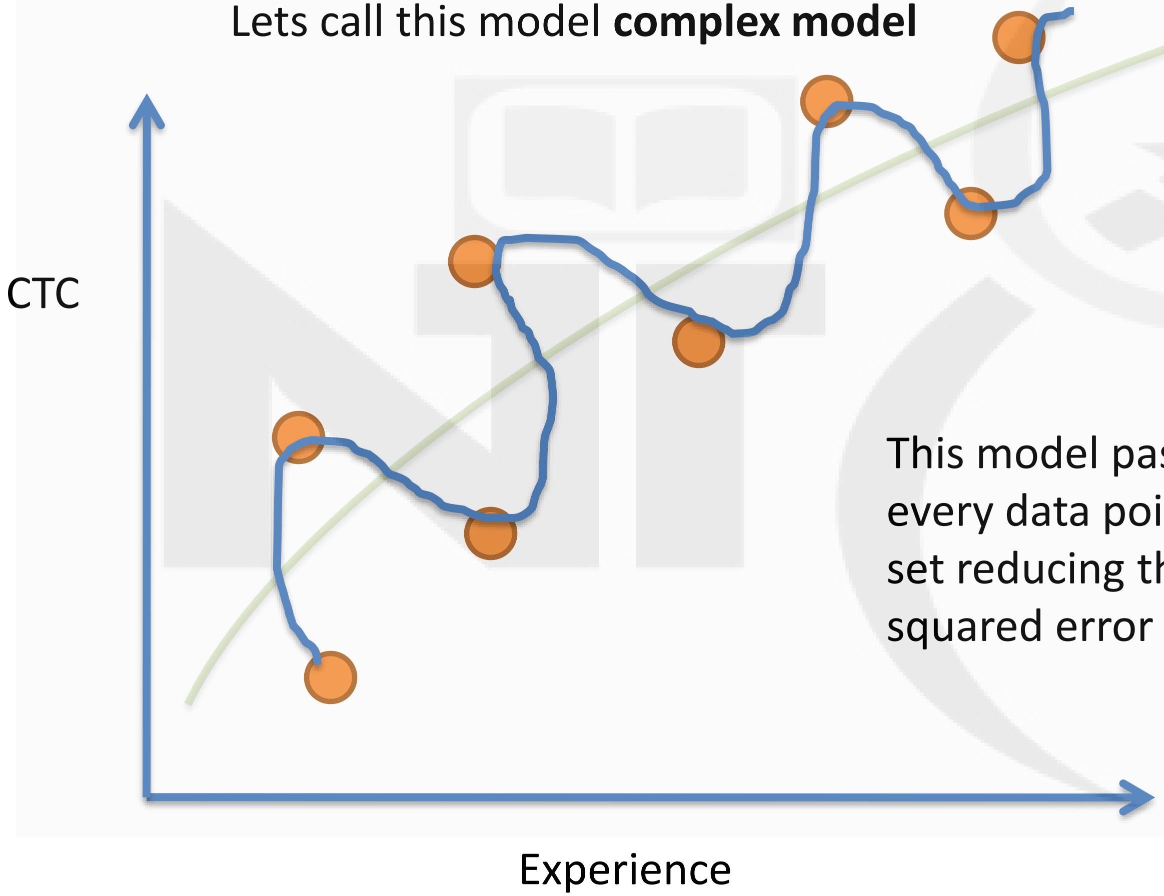
- LR model can never capture the true relationship between the CTC & Experience.

- When the model is too simple and fails to capture the underlying patterns in the data, Model is said to have **HIGH BIAS**

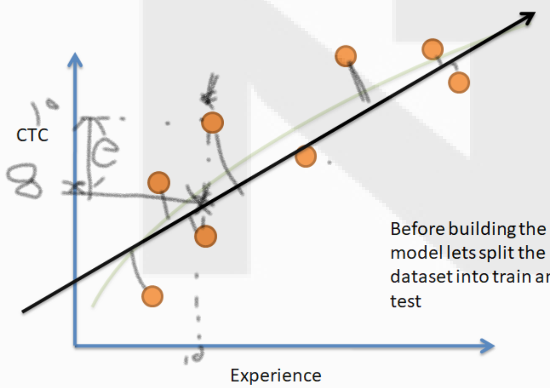
- Inability of a ML algorithm to capture the true pattern in the data is **BIAS**

Lets look at another model

Lets call this model **complex model**

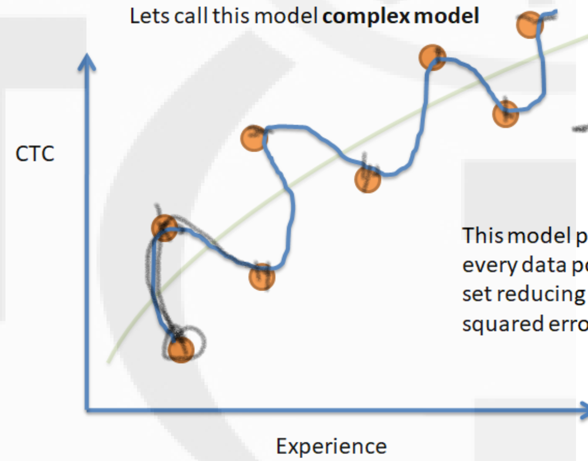


This model passes through every data point in training set reducing the sum of squared error to zero



Before building the model lets split the dataset into train and test

Lets call this model **complex model**



TE=0

This model passes through every data point in training set reducing the sum of squared error to zero

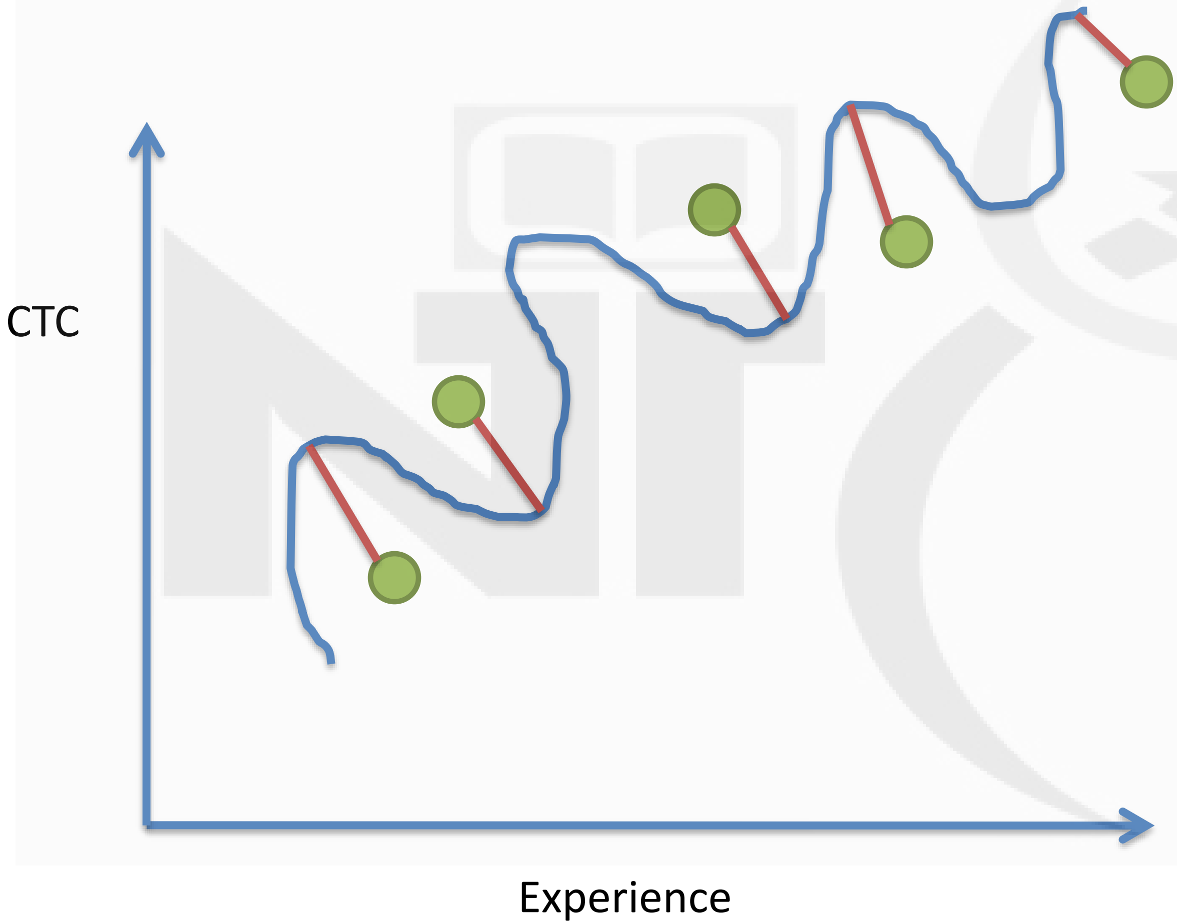
- Because Complex model can capture the pattern in the data perfectly it will have a very low bias.

- So the SSE on LR is very high hence it has a HIGH BIAS

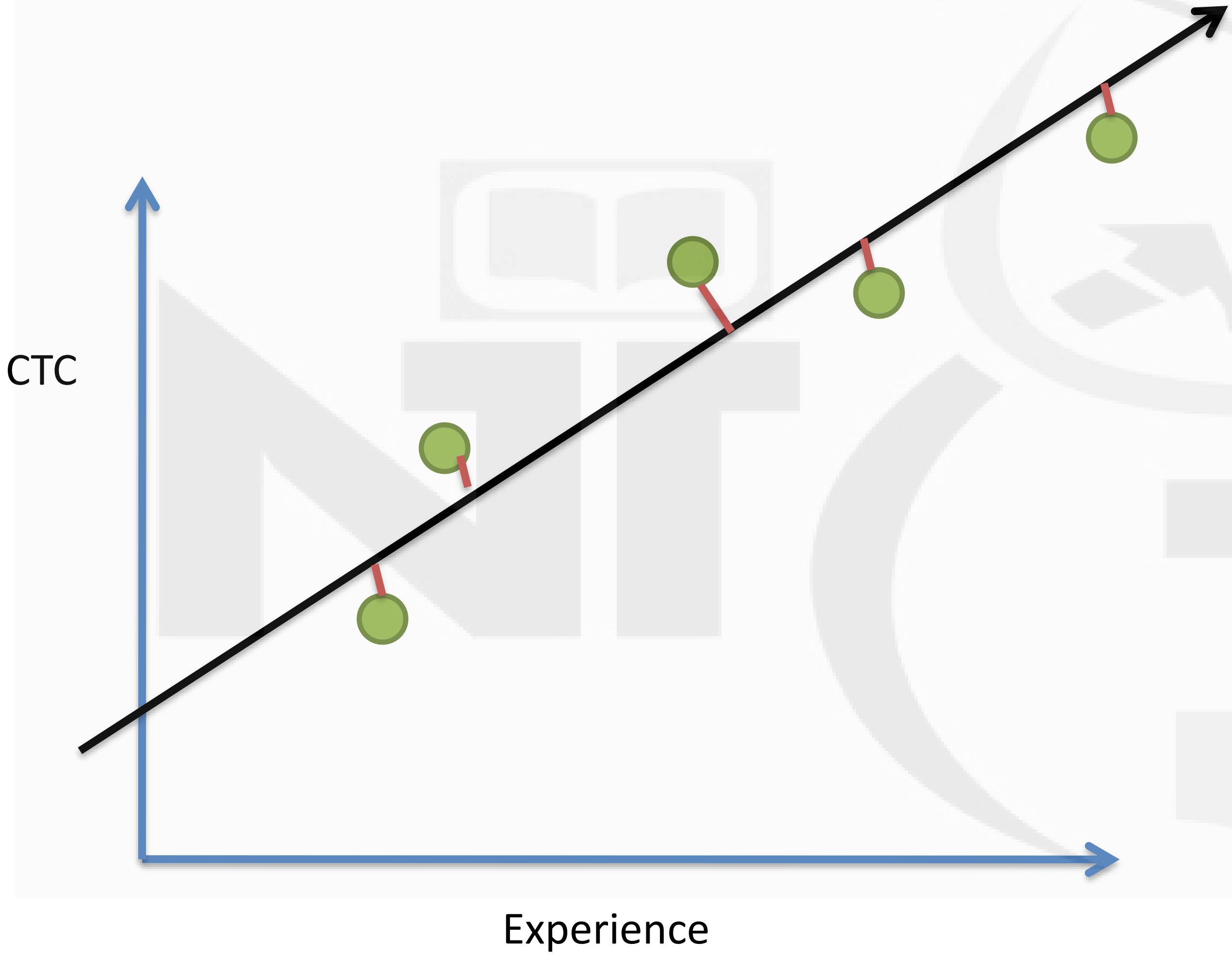
- Where as model2 has a very low SSE so it has a very LOW BIAS

LETS FIT BOTH THE MODELS ON TESTING SET

Model2 doesn't fit the test data that well



LR model fits fairly good on the test data



Variance

- Variance is a measure of the model's ability to generalize from the training data to unseen data.

- If a model has low variance, it means that it performs consistently across different datasets (including the testing dataset), suggesting that it has learned more general patterns.

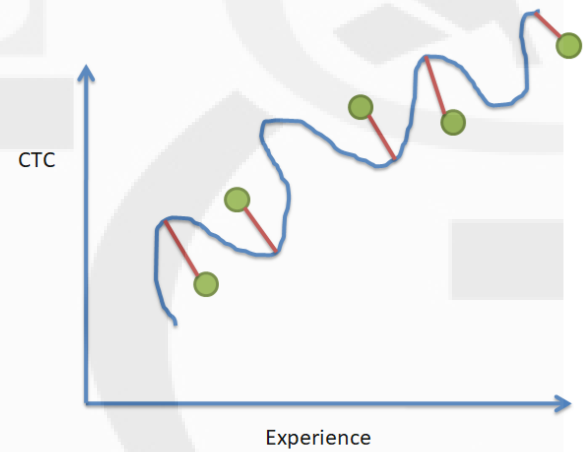
- When we use a trained model on a new dataset every time its gives a different accuracy which is very less compared to its training accuracy , this type of model is said to have a high variance

SUMMARY

Model with High Variance low bias

- Its hard to predict how it will perform on new unseen data

- Sometime it might perform well and sometimes it might do worse



Model with high Bias and Low Variance

- It never going to give you great predictions
- It will give good prediction consistently over any future dataset

- High variance(Low bias): Model performs exceptionally well on training set but fails on testset, or future datasets

- Low variance & high bais:
- - model doesn’t perform that well on training data
- However , it gives you a fairly consistent prediction over future dataset

- **High Bias = Underfit = Low variance= 2 Simple model**
- Low variance = more or less same accuracy on all future datasets
- **Low Bias = Overfit = High Variance = 2 Complex Model**
- High Variance = poor & unpredictable accuracy on future data sets

BIAS-VARIANCE TRADEOFF

- **High Bias** typically leads to **underfitting**, where the model is too simple and fails to capture the underlying patterns in the data. This results in poor performance on both training and test data.
- **Low Bias** is generally desirable, but if the model has too low a bias, it may become too complex, leading to **overfitting**, where the model performs well on training data but poorly on unseen data.

REGULARIZATION

What is Regularization?

- Regularization is a technique used in machine learning to prevent overfitting.
- Regularization helps to improve the model's generalization to unseen data.

Regularization Techniques:

- L1 Regularization (Lasso)
- L2 Regularization (Ridge)

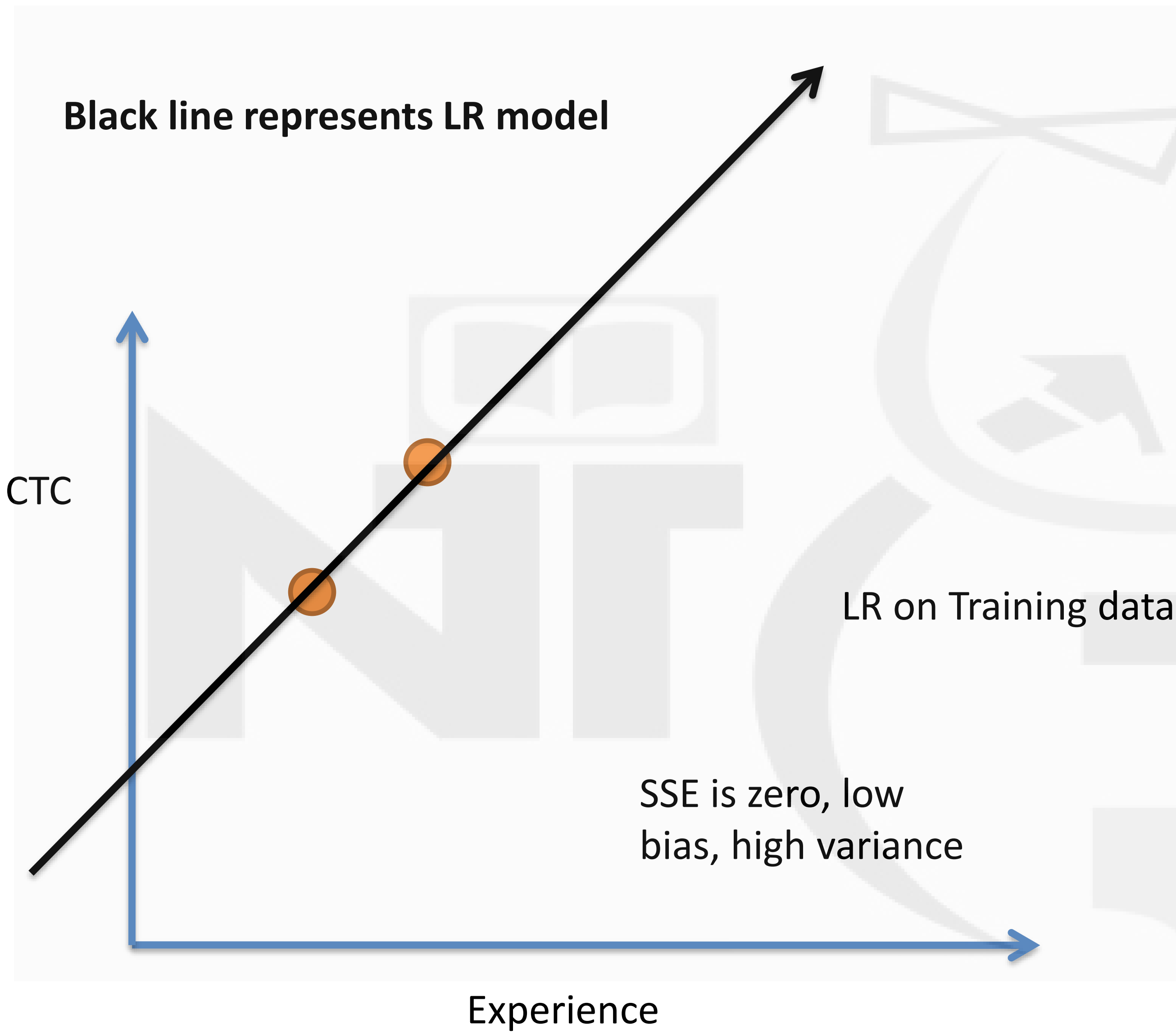
- **L1 Regularization (Lasso)**: This technique adds a penalty equal to the absolute value of the magnitude of coefficients. It can lead to sparse models where some feature weights are exactly zero, effectively performing feature selection.

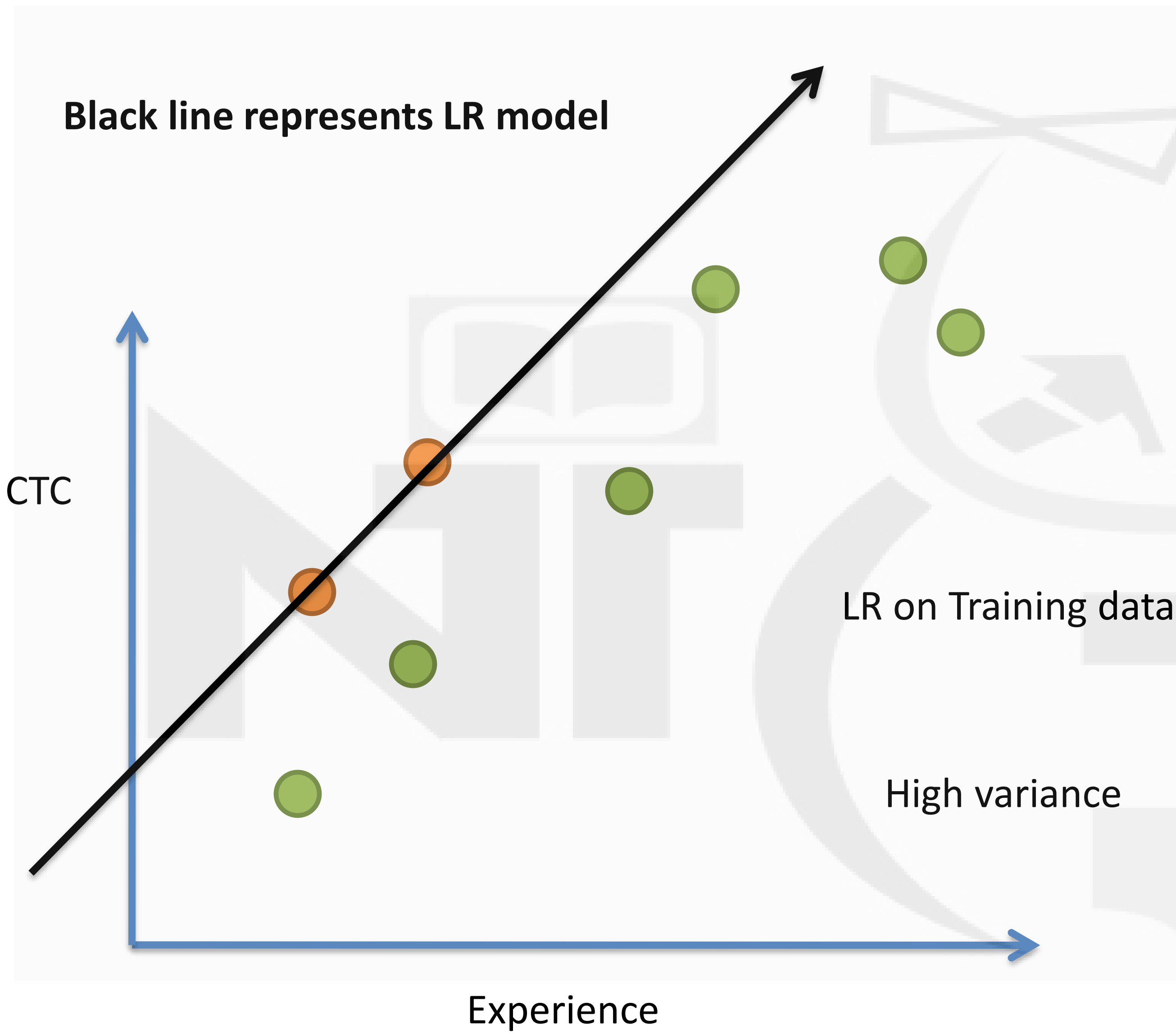
$$\text{Cost function} = \text{Loss function} + \lambda \sum_i |w_i|$$

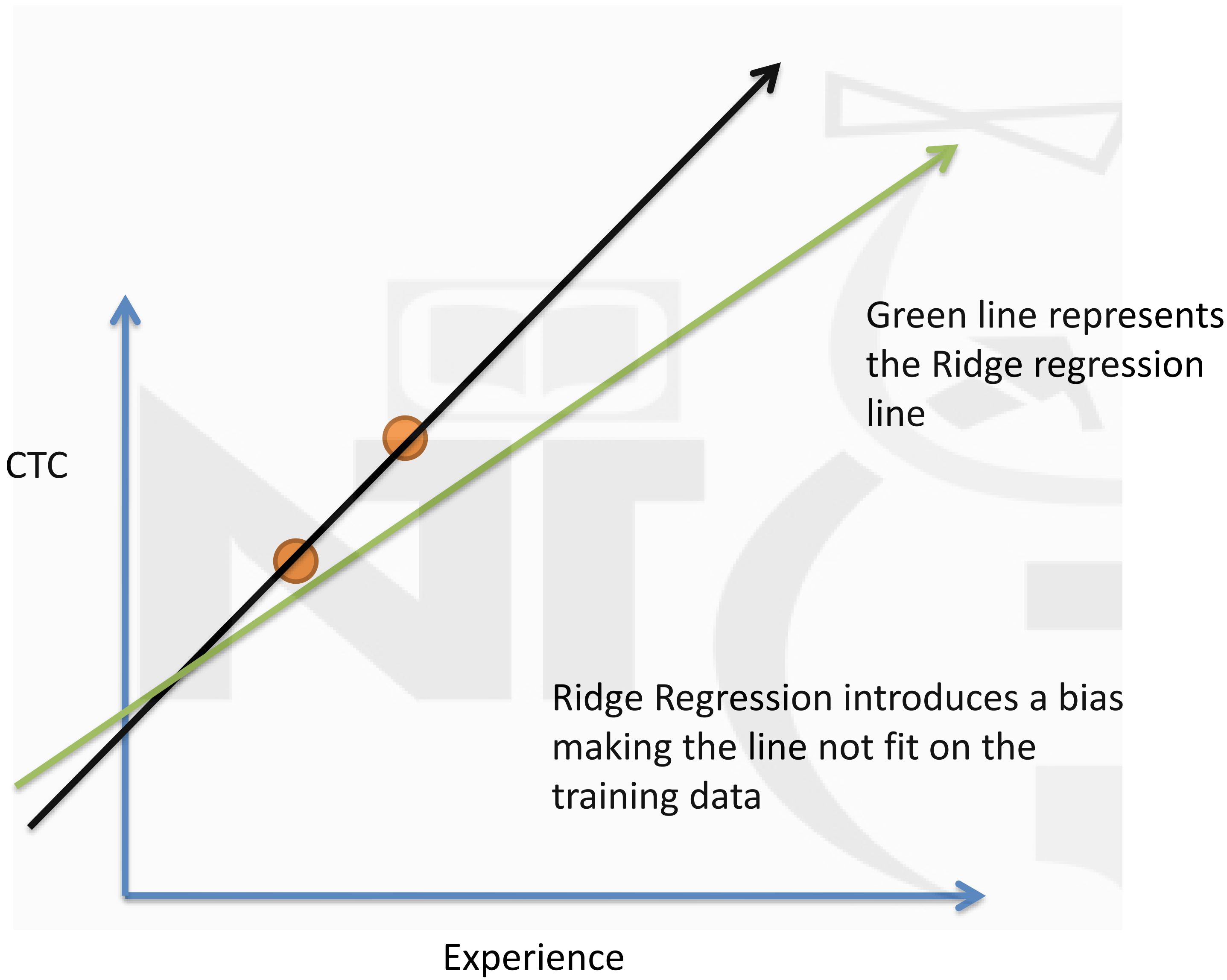
- **L2 Regularization (Ridge)**: This technique adds a penalty equal to the square of the magnitude of coefficients. It tends to shrink the coefficients but does not set them to zero, leading to a more smooth model.

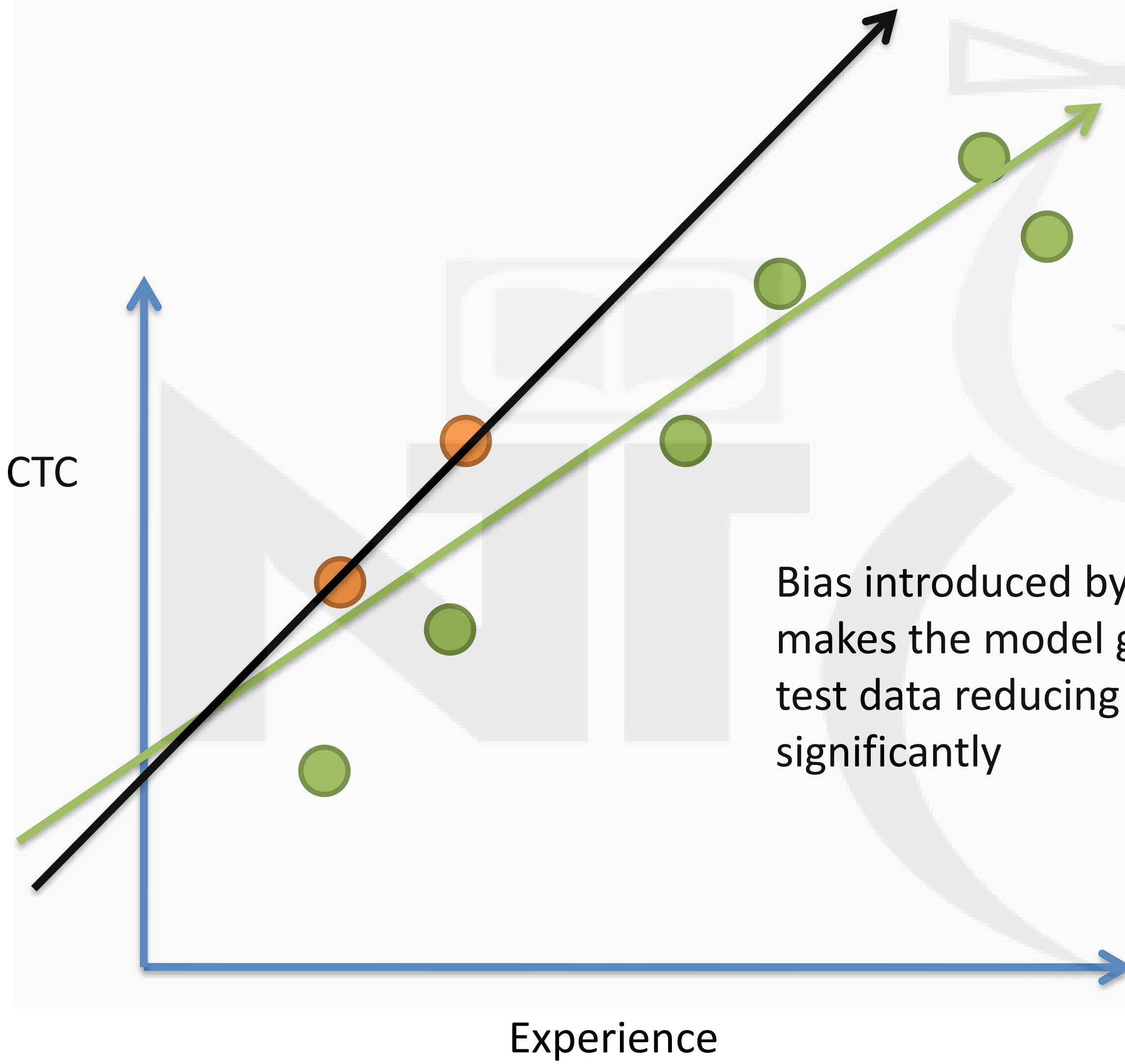
$$\text{Cost function} = \text{Loss function} + \lambda \sum_i w_i^2$$

RIDGE REGRESSION


Black line represents LR model
CTC
LR on Training data
SSE is zero, low bias, high variance
Experience


Black line represents LR model
CTC
LR on Training data
High variance
Experience


CTC
Green line represents the Ridge regression line
Ridge Regression introduces a bias making the line not fit on the training data
Experience



Bias introduced by the Ridge makes the model generalize on test data reducing the variance significantly

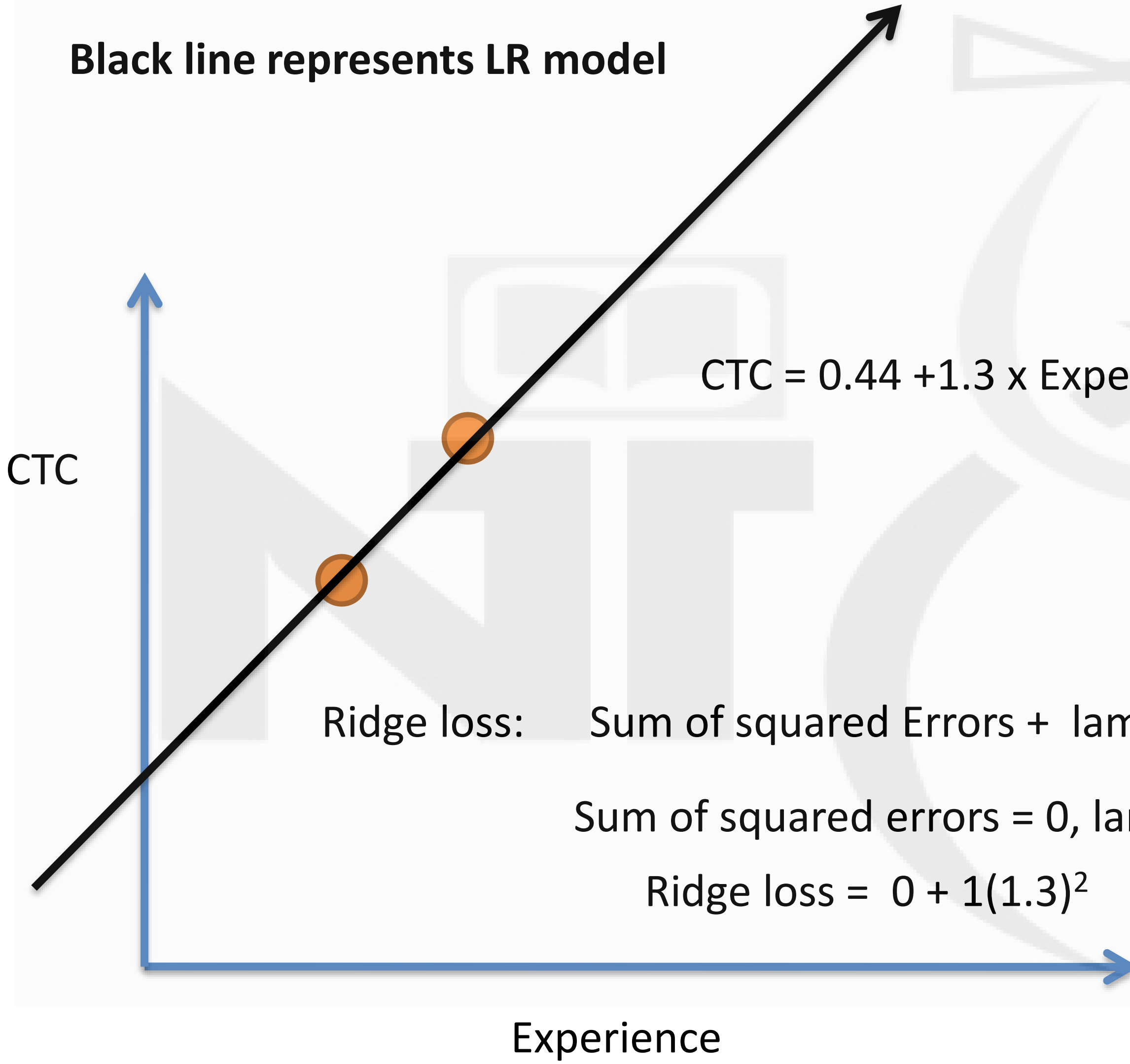
How Ridge Regression works?

- Loss function for LR is:
  - Sum of squrared errors/residuals
- Loss function of Ridge Regression:
  - Sum of squared errors + Lambda*slope^2

- Lambda is hyperparameter

$$\text{Cost function} = \text{Loss function} + \lambda \sum_i w_i^2$$

Black line represents LR model

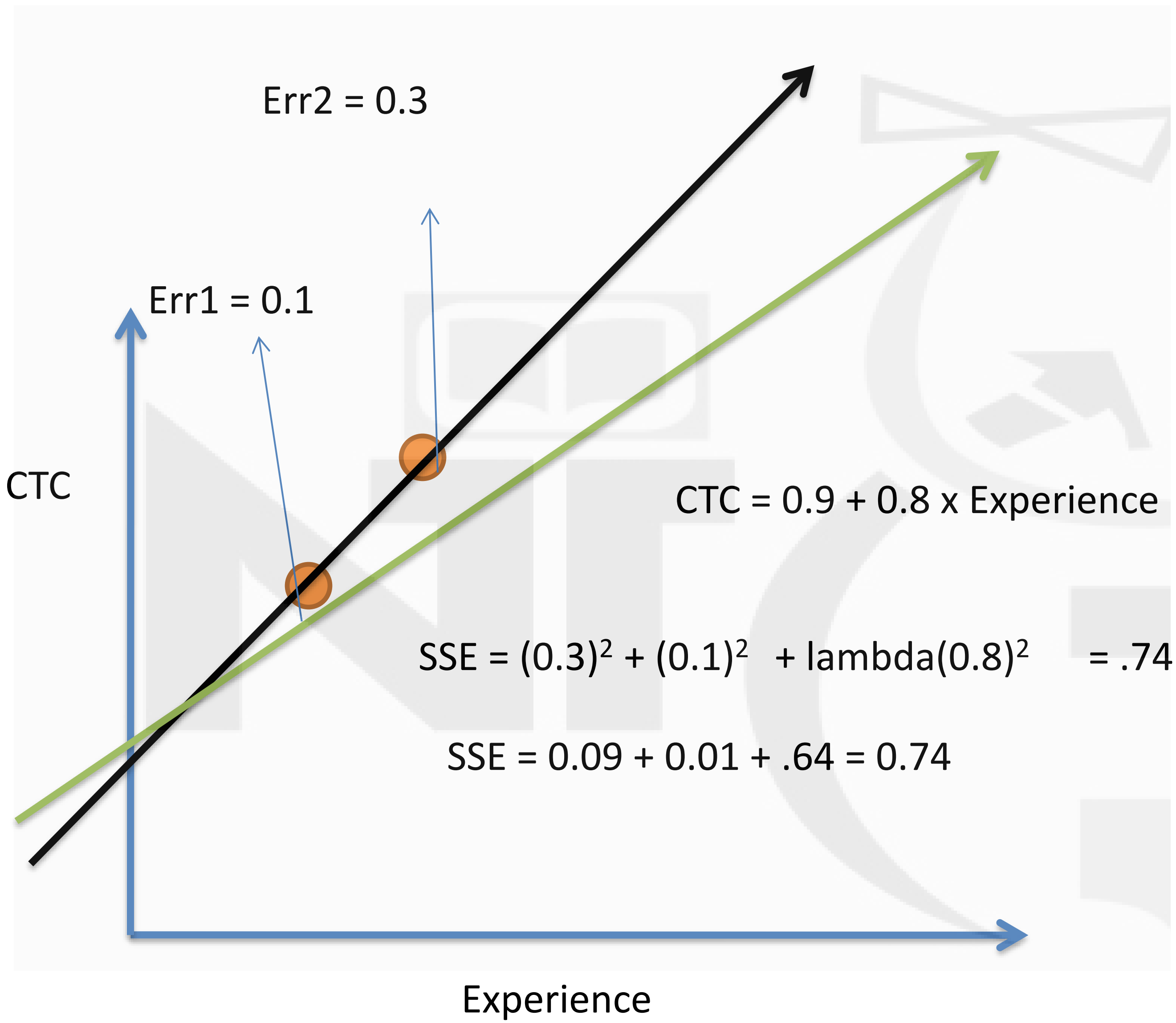


CTC = 0.44 +1.3 x Experience

Ridge loss: Sum of squared Errors + lambda * $(1.3)^2$

Sum of squared errors = 0, lambda =1

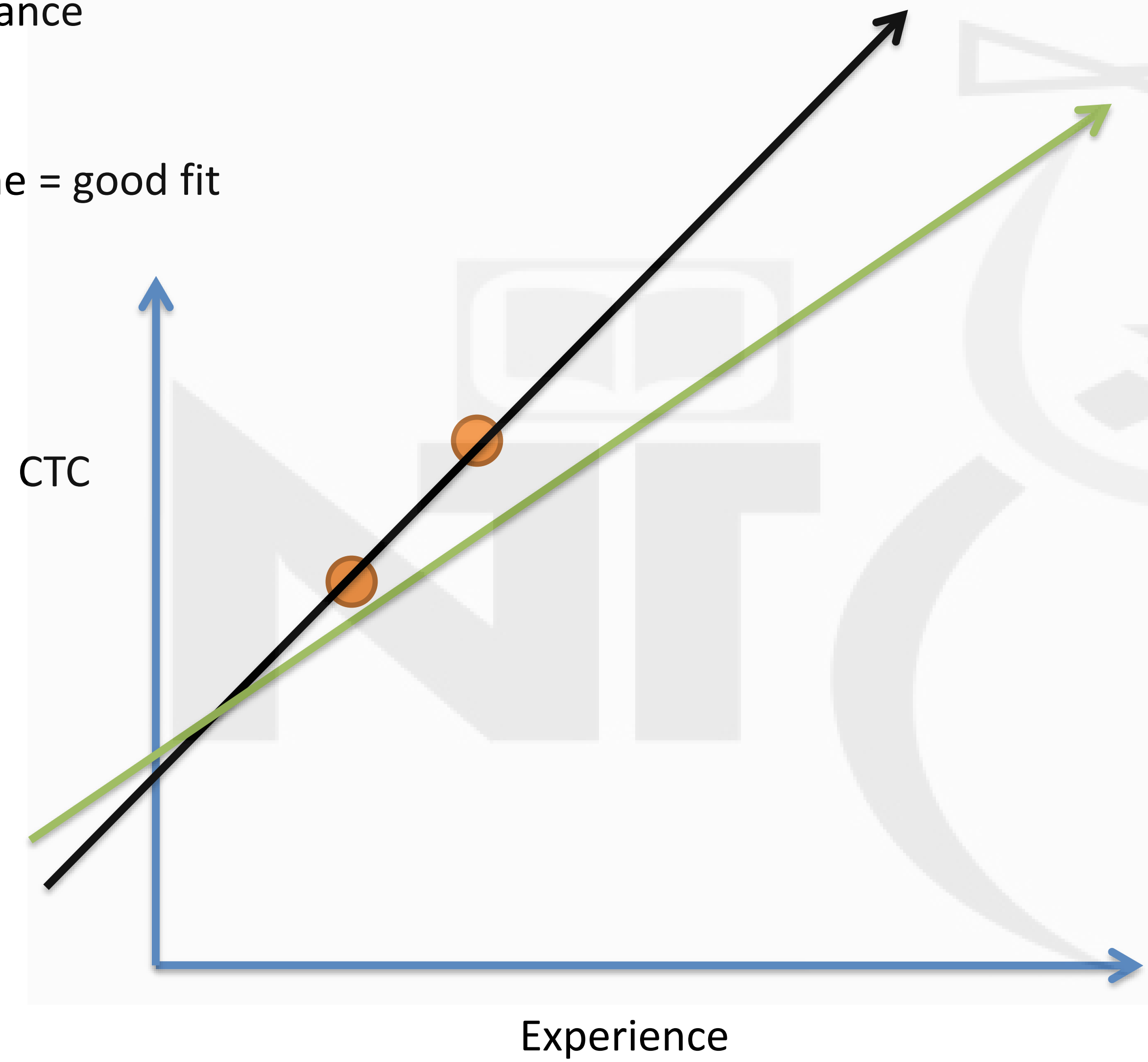
Ridge loss = $0 + 1(1.3)^2$ = 1.69


Err2 = 0.3
Err1 = 0.1
CTC
CTC = 0.9 + 0.8 x Experience
SSE = $(0.3)^2 + (0.1)^2$ + lambda$(0.8)^2$ = .74
SSE = 0.09 + 0.01 + .64 = 0.74
Experience


SSE = 1.69
SSE = 0.74
CTC
Ridge Regression chooses the line
with min SSE
Experience

Black line is overfit , = low bias =
high variance

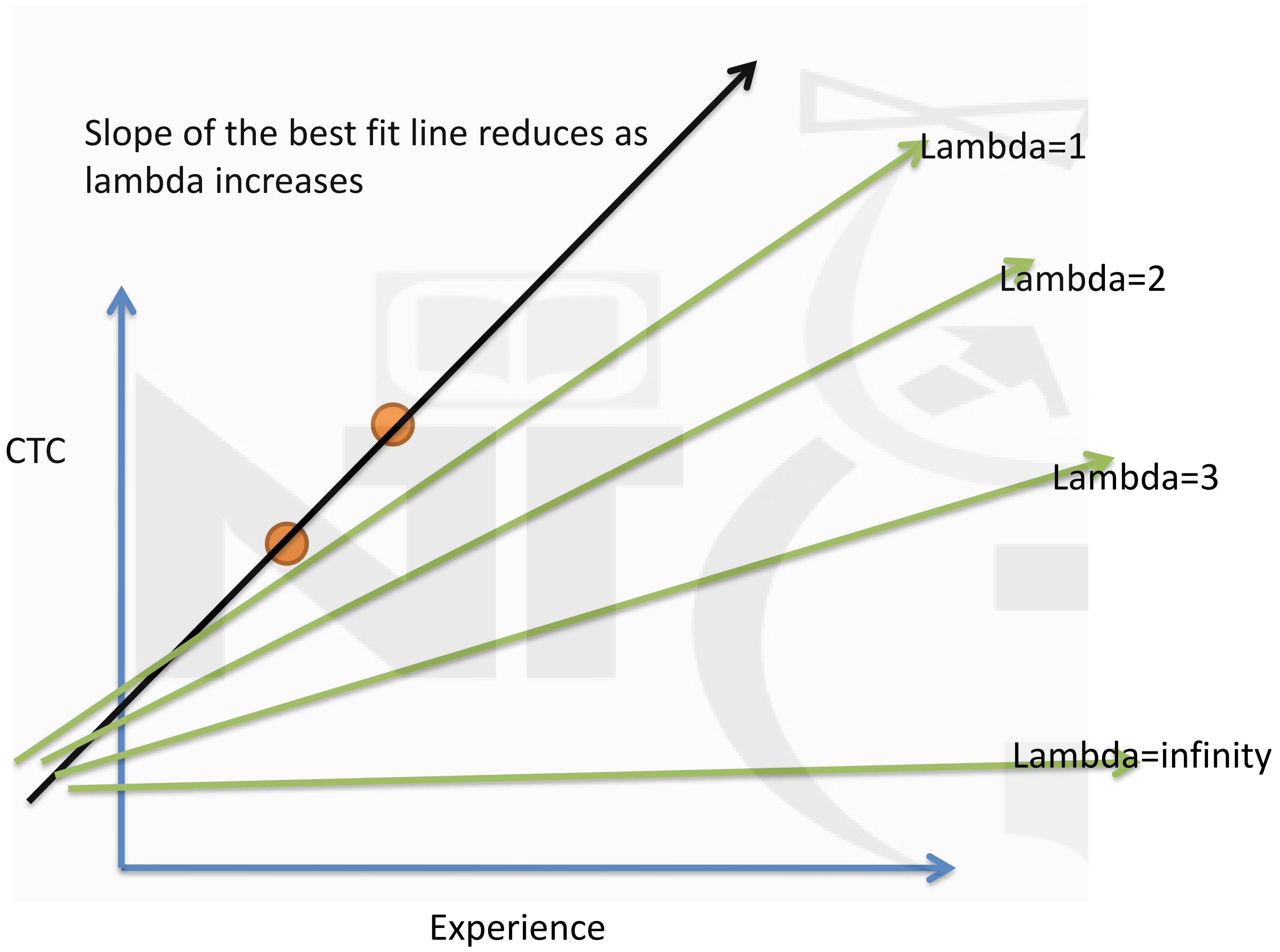
Green line = good fit



Lambda in Ridge Regression

- Lambda can take values from 0 to infinity

- As lambda increases the slope of the line reduces


Slope of the best fit line reduces as lambda increases
Lambda=1
Lambda=2
Lambda=3
Lambda=infinity
CTC
Experience

L2 Regularization

- Ridge regression is called L2 regularization because it involves adding a penalty to the loss function that is proportional to the square of the magnitude of the coefficients.

$$\text{Cost function} = \text{Loss function} + \lambda \sum_i w_i^2$$

LASSO REGRESSION

- **L1 Regularization (Lasso)**: This technique adds a penalty equal to the absolute value of the magnitude of coefficients. It can lead to sparse models where some feature weights are exactly zero, effectively performing feature selection.

$$\text{Cost function} = \text{Loss function} + \lambda \sum_i |w_i|$$

- Similar to Ridge regression however it take the absolute value of slope(coeff) in the penalty compared to squared in Ridge.

Comparison and Choice

- **Ridge Regression** is more appropriate when you have many features and you believe all or most of them have some relevance. It helps manage multicollinearity and avoids overfitting without discarding any features.

- **Lasso Regression** is useful when you suspect that only a few features are important. It not only helps with overfitting but also performs automatic feature selection, which can lead to a simpler and more interpretable model.

Assignment

- Try Ridge and Lasso on USA Housing dataset and compare the models