Lecture 1.5 and 1.6

- Regression
  - Linear Regression
  - Intuition
  - Loss Function
- Regularization
  - Lasso Regularization
  - Ridge Regularization

Regression

- **Regression** is a statistical method for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors)
- In **linear regression**, the relationship is **modeled as a straight line**
- **Intuition** in regression is the idea of using a linear approach to predict a continuous value for a data point by generalizing over the data

Linear Regression

- Linear regression assumes a linear relationship between the dependent variable y and the independent variable(s) x
- The equation of a simple linear regression model is

$$y = mx + c$$

$m\ (w_1)$: slope of the line (weight of the predictor)

$c\ (or\ w_0\ )$: y-intercept

x: input feature(s)

y: predicted output (target)

- For multiple predictors (features), the equation becomes

$$y = w_0\ + w_1\, x_1\ + w_2\, x_2\ + \cdots + w_n\, x_n$$

Loss Function

- The performance of the linear regression model is evaluated using the cost function, commonly the Mean Squared Error (MSE)

$$E(w_0, w_1, \dots, w_n) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

where,

E: Loss function

N: Number of data points

$\hat{y}_i$:Predicted value

y_i: Actual value

Example 1

SL No.	x (data)	y_{actual}
1	1	2
2	2	3

$$y = w_1 x + w_0$$

Initial Setup

$$w_0 = 0$$
$$w_1 = 0$$

$Learning\ rate\ (\eta) = 0.1$

$Loss\ function{:}\ E(w_0, w_1, \dots, w_n) = \frac{1}{2N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2$

Iteration 1

$$w_0 = 0, w_1 = 0, E(w_0, w_1, \dots, w_n) = \frac{1}{2N}\sum_{i=1}^{N}(w_1 x_i + w_0 - y_i)^2$$

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)$$

$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)x_i$$

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i) = \frac{1}{2}(-2 + -3) = -2.5$$

$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)x_i = \frac{1}{2}(-2 \times 1 + -3 \times 2) = -4$$

x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$\boldsymbol{w_1 = w_1 - \eta \nabla E(w_1)}$	$\boldsymbol{w_0 = w_0 - \eta \nabla E(w_0)}$
1	2	0 x 1 + 0 = 0	0 – 2 = -2	**0 – 0.1x-4= 0.4**	**0 – 0.1x-2.5= 0.25**
2	3	0 x 2 + 0 = 0	0 – 3 = -3		

Iteration 2

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i) = \frac{1}{2}(-1.35 - 1.95) = -1.65$$

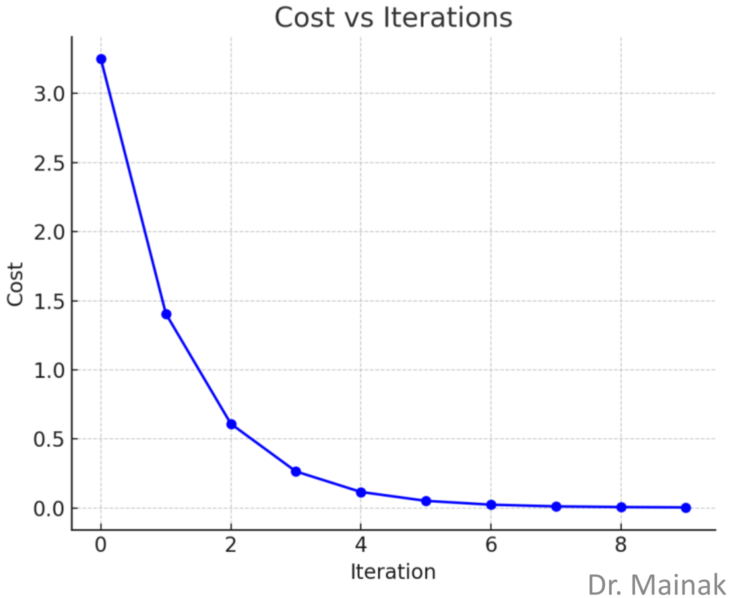
$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)x_i = \frac{1}{2}(-1.35 \times 1 - 1.95 \times 2) = -2.625$$

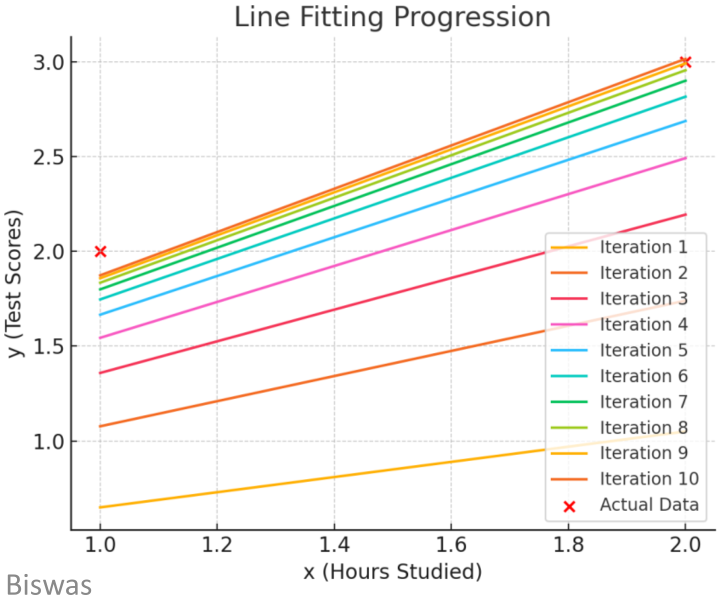
x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$w_1 = w_1 - \eta \nabla E(w_1)$	$w_0 = w_0 - \eta \nabla E(w_0)$
1	2	0.4 x 1 + 0.25 = 0.65	0.65 – 2 = -1.35	**0.4 – 0.1x-2.625= 0.6625**	**0 .25– 0.1x-1.65= 0.415**
2	3	0.4 x 2 + 0.25 = 1.05	1.05 – 3 = -1.95		

Iteration 10

- $w_0 = 0.7322, w_1 \ = 1.1411,$

$$y = 1.1411x + 0.7322$$

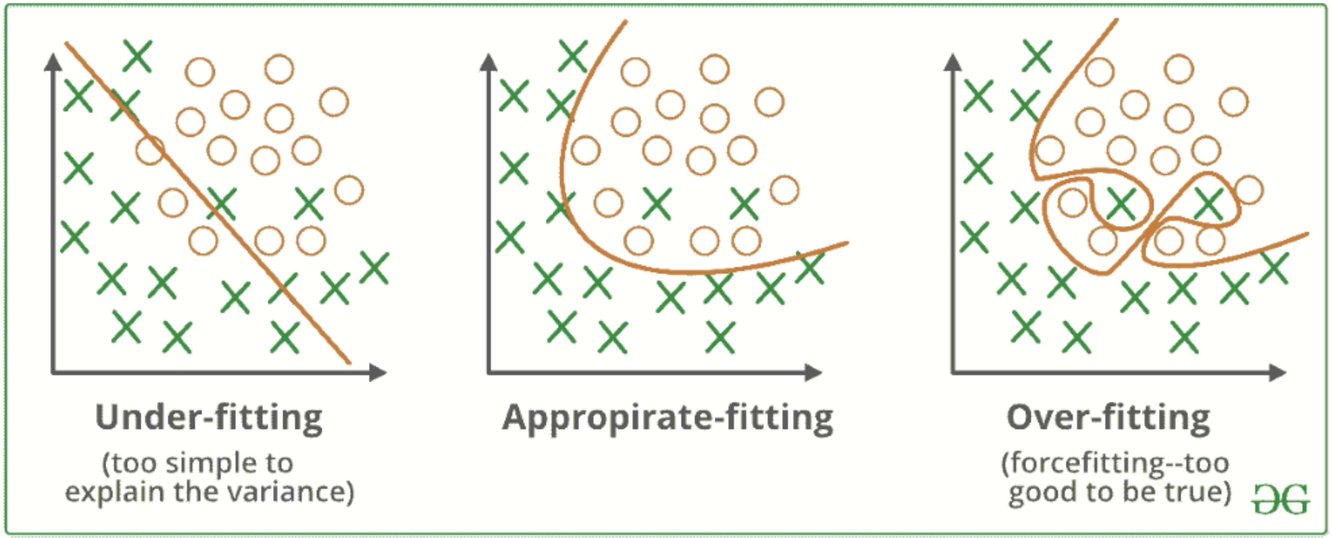


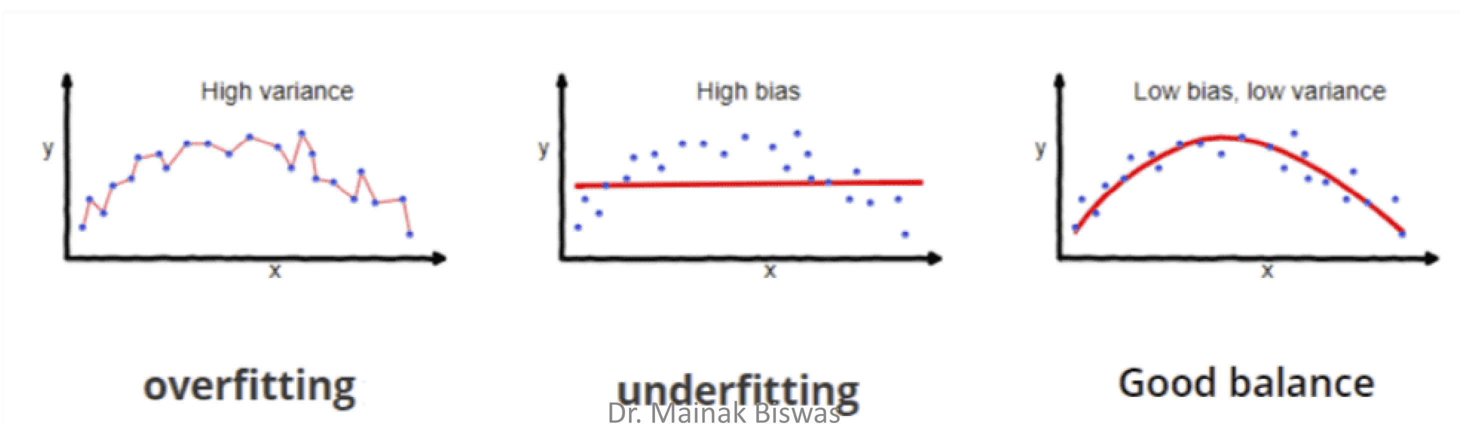


Regularization

- **Regularization** is a technique used in machine learning and statistics to **prevent overfitting** of models by adding a penalty term to the model's loss function
- Regularization provides this increased generalizability at the sake of increased training error
- Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting. The commonly used regularization techniques are :
  - Lasso Regularization – L1 Regularization
  - Ridge Regularization – L2 Regularization

Overfitting





Role of Regularization

- **Complexity Control**: Regularization helps control model complexity by preventing overfitting to training data, resulting in better generalization to new data.
- **Preventing Overfitting**: One way to prevent overfitting is to use regularization, which penalizes large coefficients and constrains their magnitudes, thereby preventing a model from becoming overly complex and memorizing the training data instead of learning its underlying patterns.
- **Balancing Bias and Variance**: Regularization can help balance the trade-off between model bias (underfitting) and model variance (overfitting) in machine learning, which leads to improved performance.
- **Feature Selection**: Some regularization methods, such as L1 regularization (Lasso), promote sparse solutions that drive some feature coefficients to zero. This automatically selects important features while excluding less important ones.
- **Handling Multicollinearity**: When features are highly correlated (multicollinearity), regularization can stabilize the model by reducing coefficient sensitivity to small data changes.
- **Generalization**: Regularized models learn underlying patterns of data for better generalization to new data, instead of memorizing specific examples.

L1 (LASSO) Regularization

- A regression model which uses the L1 Regularization technique is called LASSO(Least Absolute Shrinkage and Selection Operator) regression
- **Lasso Regression** adds the "**absolute value of magnitude**" of the **coefficient as a penalty** term to the loss function(L)

$$E(w_0, w_1, \dots, w_n) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{i=1}^{m}|w_i|$$

λ: Regularization parameter

$| w_i |$: Absolute value of the i-th weight (coefficient) of the model

m: Number of features

How it affects the gradient?

- In Lasso regularization, the gradient of the cost function with respect to a parameter w includes the term λ·sign(w)
- When computing the gradient of the cost function with respect to ww, the derivative of the absolute value is not defined at w=0, but it is approximated using the sign function:

$$\frac{\delta|w|}{\delta w} = sign(w)$$

When w>0, λ pushes w to decrease

When w<0, λ pushes w to increase

When w=0, the penalty term does not contribute to the gradient

Example 2

SL No.	x (data)	y_{actual}
1	1	2
2	2	3

$$y = w_1 x + w_0$$

Initial Setup

$$w_0 = 0$$
$$w_1 = 0$$

$$Learning\ rate\ (\eta) = 0.1$$

$$Regularization\ Parameter\ (\lambda) = 0.1$$

$$Loss\ function:$$

$$E(w_0, w_1) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2 + \lambda \sum_{i=1}^{m}|w_i|$$

Iteration 1

$w_0 = 0, w_1 = 0,$

$E(w_0, w_1) = \frac{1}{2N}\sum_{i=1}^{N}(w_1 x_i + w_0 - y_i)^2 + \lambda(|w_0| + |w_1|)$

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) + \lambda\, sign\,(w_0)$$

$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i + \lambda\, sign\,(w_1)$$

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) = \frac{1}{2}(-2 + -3) + 0.1 \times 0 = -2.5$$

$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i = \frac{1}{2}(-2 \times 1 + -3 \times 2) + 0.1 \times 0 = -4$$

x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$w_1 = w_1 - \eta\nabla E(w_1)$	$w_0 = w_0 - \eta\nabla E(w_0)$
1	2	0 x 1 + 0 = 0	0 – 2 = -2	**0 – 0.1x-4= 0.4**	**0 – 0.1x-2.5= 0.25**
2	3	0 x 2 + 0 = 0	0 – 3 = -3		

Dr. Mainak Biswas

Iteration 2

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i) = \frac{1}{2}(-1.35 - 1.95) + 0.1 \times 1 = -1.55$$

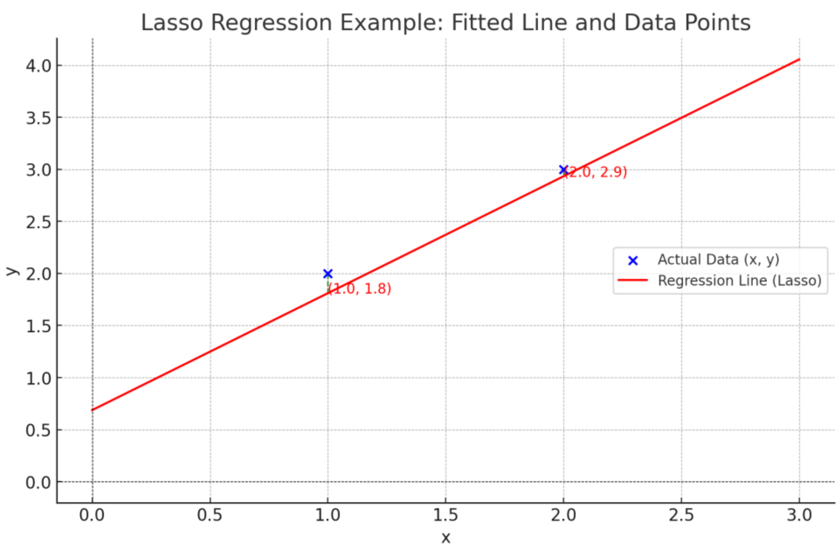
$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i)x_i = \frac{1}{2}(-1.35 \times 1 - 1.95 \times 2) + 0.1 \times 1 = -2.525$$

x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$\boldsymbol{w_1 = w_1 - \eta \nabla E(w_1)}$	$\boldsymbol{w_0 = w_0 - \eta \nabla E(w_0)}$
1	2	0.4 x 1 + 0.25 = 0.65	0.65 – 2 = -1.35	**0.4 – 0.1x-2.525= 0.6525**	**0 .25– 0.1x-1.55= 0.405**
2	3	0.4 x 2 + 0.25 = 1.05	1.05 – 3 = -1.95		

Iteration 10

- $w_0 = 0.6874, w_1 = 1.1227$

$$y = 1.1227x + 0.6874$$



L2 (Ridge) Regularization

- **Ridge regression**, also known as L2 regularization, is a technique used to improve the performance and generalization of linear regression models
- It penalizes large coefficients to reduce overfitting while ensuring that all features contribute to the prediction

$$E(w_0, w_1, \dots, w_n) = \frac{1}{2N}\sum_{i=1}^{N}(\widehat{y_i} - y_i)^2 + \lambda \sum_{i=1}^{m} w_i^2$$

λ: Regularization parameter

w_i^2: square of the magnitude of coefficients

m: Number of features

How it affects the gradient?

- If λ=0: Ridge regression reduces to ordinary least squares (OLS) regression with no regularization
- If λ is very large: Coefficients shrink significantly, which may lead to Underfitting (model is too simple)
- Optimal λ: A balance between overfitting and Underfitting is achieved through cross-validation to determine the best λ

Example 3

SL No.	x (data)	y_{actual}
1	1	2
2	2	3

$$y = w_1 x + w_0$$

Initial Setup

$$w_0 = 0$$
$$w_1 = 0$$

$Learning\ rate\ (\eta) = 0.1$

$Regularization\ Parameter\ (\lambda) = 0.1$

$$Loss\ function:$$

$$E(w_0, w_1) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2 + \lambda\sum_{i=1}^{m} w_i^2$$

Iteration 1

$$w_0 = 0, w_1 = 0,$$
$$E(w_0, w_1) = \frac{1}{2N}\sum_{i=1}^{N}(w_1 x_i + w_0 - y_i)^2 + \lambda(w_0^2 + w_1^2)$$

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i) + 2\lambda w_0$$
$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)x_i + 2\lambda w_1$$
$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i) = \frac{1}{2}(-2 + -3) + 2 \times 0.1 \times 0 = -2.5$$
$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)x_i = \frac{1}{2}(-2 \times 1 + -3 \times 2) + 2 \times 0.1 \times 0 = -4$$

x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$w_1 = w_1 - \eta\nabla E(w_1)$	$w_0 = w_0 - \eta\nabla E(w_0)$
1	2	0 x 1 + 0 = 0	0 – 2 = -2	**0 – 0.1x-4= 0.4**	**0 – 0.1x-2.5= 0.25**
2	3	0 x 2 + 0 = 0	0 – 3 = -3		

Iteration 2

$$\frac{\delta E}{\delta w_0} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) = \frac{1}{2}(-1.35 - 1.95) + 2 \times 0.1 \times 0.25 = -1.60$$

$$\frac{\delta E}{\delta w_1} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i = \frac{1}{2}(-1.35 \times 1 - 1.95 \times 2) + 2 \times 0.1 \times 0.4 = -2.545$$

x (data)	y_{actual}	$\hat{y} = w_1 x + w_0$	Residual Error	$w_1 = w_1 - \eta \nabla E(w_1)$	$w_0 = w_0 - \eta \nabla E(w_0)$
1	2	0.4 x 1 + 0.25 = 0.65	0.65 – 2 = -1.35	**0.4 – 0.1x-2.545= 0.6545**	**0 .25– 0.1x-1.60= 0.41**
2	3	0.4 x 2 + 0.25 = 1.05	1.05 – 3 = -1.95		

Iteration 10

w0=3.5089
w1=0.8912