Machine Learning -
Theory & Practice

Module 3: Linear Regression

Lecture 1: *Introduction*

Lecture 1 Outline

► Topic 1: Basic concepts

► Topic 2: Model Estimation

Introduction

- **Linear Regression (LR) is a supervised machine learning algorithm for regression tasks.**

- **There are one or more independent variables called *Regressors, Explanatory Variables or Attributes*:**

  $$X = \{x_1, x_2 \ldots\}$$

- **The response variable *y,* is a continuous-valued random variable**

- **The distribution of *y* is dependent on the Regressors**

Hypothesis

- The hypothesis, that is the assumed relationship between response *y and* one or more regressors $X =$ $\{x_1,...,x_m\}$, m$\geq$1, is described by the following:

Consider $i = 1..n$ instances of data:

$$y_i = f(X_i) = w_0 + \sum_{j=1..m} w_j x_{i,j}^k + \epsilon_i$$

- **Note that in Linear Regression, the relationship is linear in weights, $\{w_0,w_1,...,w_m\}$. w_0 is called bias, and the remaining weights are called *Regression Coefficients* - w_j** quantifies the strength of the relationship between y and regressor x_j.

Hypothesis Space & Search Strategy

- The Hypothesis space is an infinite set of all possible representations of the hypothesis, given training data D={X,y}
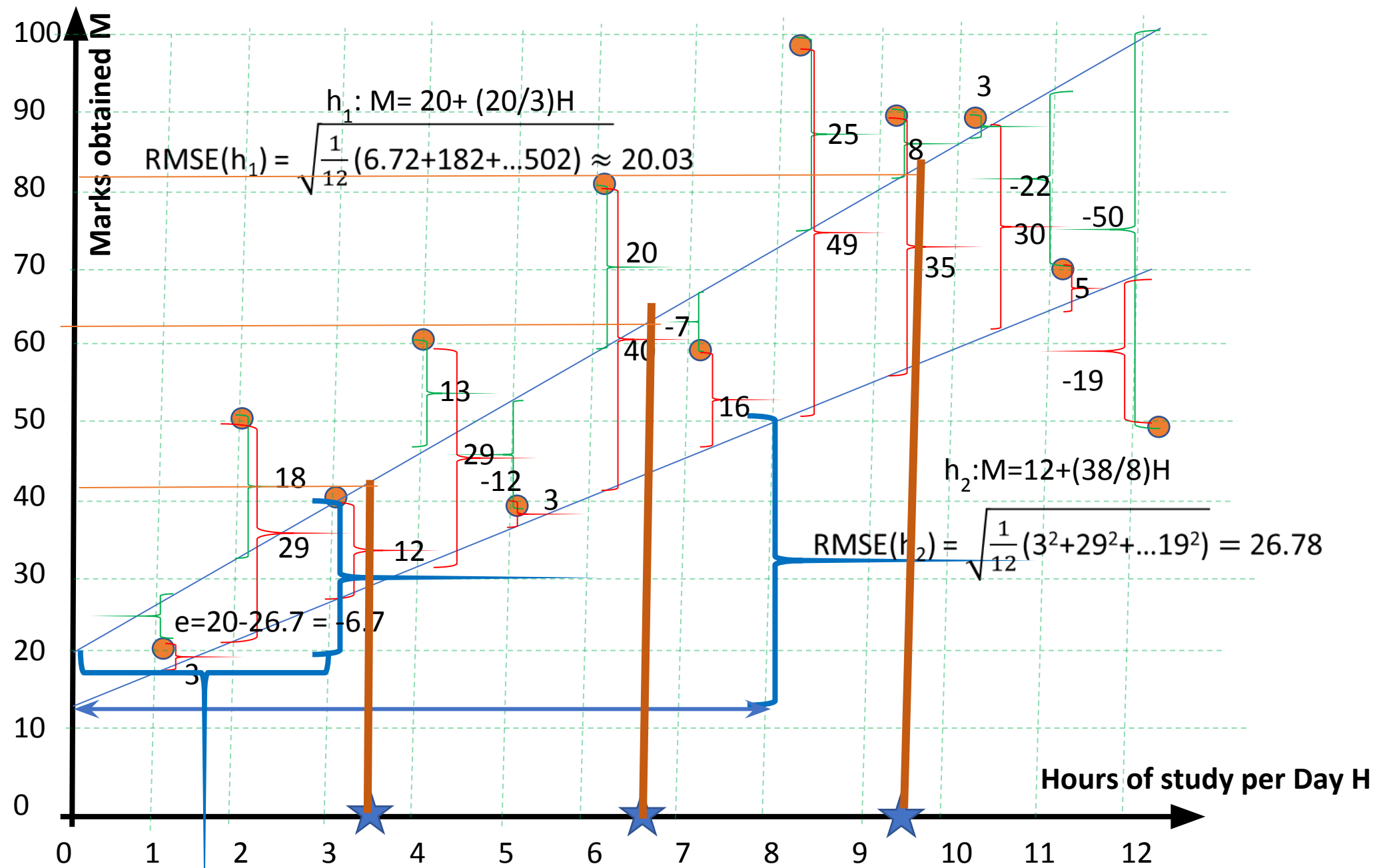
$$y_i = f(X_i) = w_0 + \sum_{j=1..m} w_j x_{i,j}^k + \epsilon_i$$

- To search through the hypothesis space, *LR adjusts w_0 and the weight of each regressor,* to achieve minimum prediction error. This is an optimization process.

- The search process finds the best hypothesis *h(X) = y,* that matches the mapping *c(X) = y,* in the training data.

Tuning the Weight Parameters

Aim is to minimize prediction error. Weights must be tuned by considering:

- Which regressor(s) have a strong impact on the response, and are therefore significant?

- Which regressor(s) barely have any influence and can be eliminated?

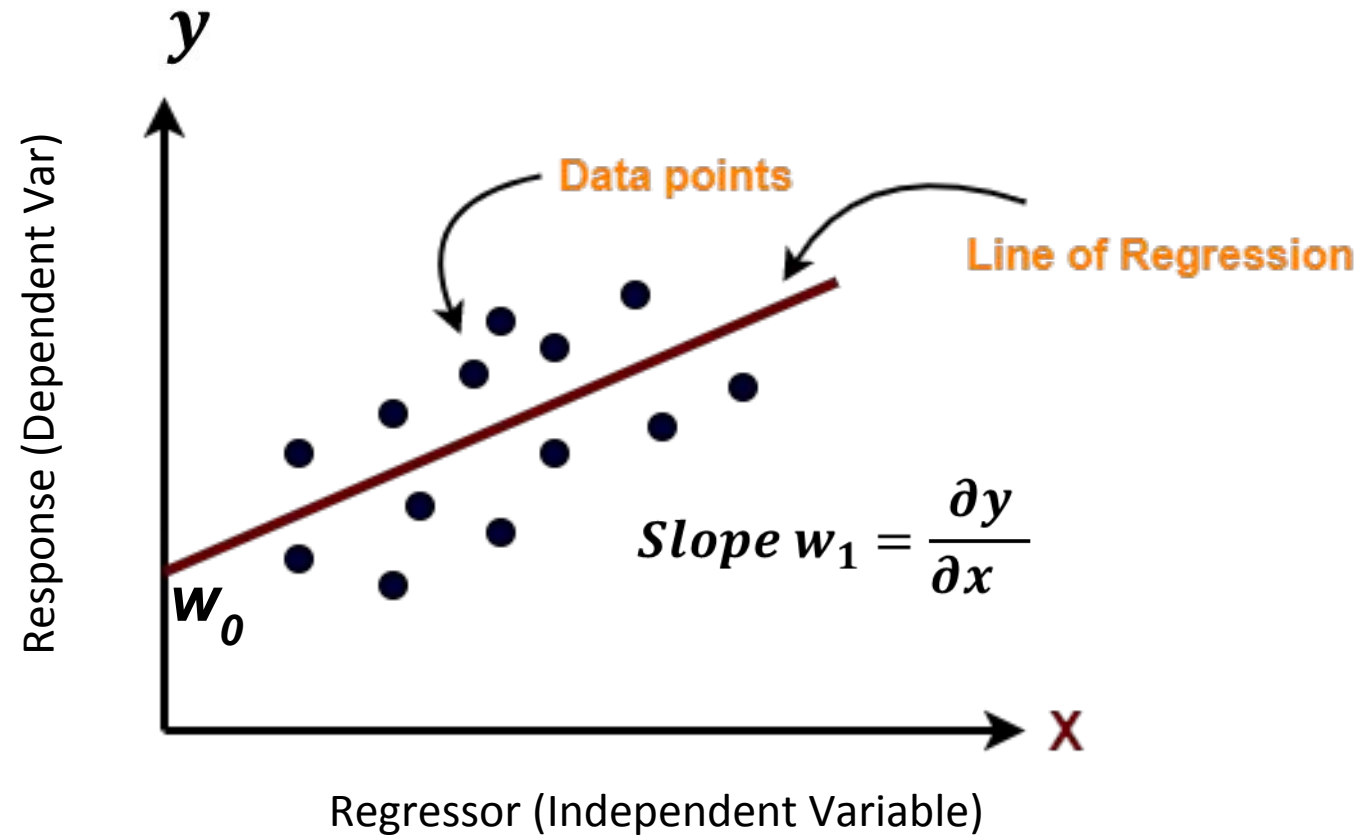- Which regressor(s) hold redundant information already captured by another attribute?

$h_1: M = 20 + (20/3)H$

$RMSE(h_1) = \sqrt{\dfrac{1}{12}(6.72+182+\ldots 502)} \approx 20.03$

$h_2: M = 12 + (38/8)H$

$RMSE(h_2) = \sqrt{\dfrac{1}{12}(3^2 + 29^2 + \ldots 19^2)} = 26.78$

$e = 20 - 26.7 = -6.7$

25
8
-22
-50
49
35
30
20
-7
40
16
5
13
-19
29
-12
18
3
29
12
3

Marks obtained M

Hours of study per Day H

Types of Linear Regression

❑ *Simple LR*: One regressor with degree 1

❑ *Multiple LR:* Multiple regressors all of degree 1

❑ *Polynomial simple/multiple Linear Regression:*
Regressors can have degree > 1 (x^2, x^3, x_1x_2…..)

❑ *Multivariate simple/multiple Linear Regression:* More than
one response variables that are coordinated

Simple LR $y = w_0 + w_1 x_1 + \epsilon$

y

Response (Dependent Var)

Data points

Line of Regression

$Slope\ w_1 = \dfrac{\partial y}{\partial x}$

w_0

X

Regressor (Independent Variable)

Simple LR example:

Example: Car rental y = base rental w_0 + rate w_1*miles covered x + ϵ

- w_0 may be initial conditions or aggregate effect of all factors
- w_1 is the strength by which X has an impact on y.

Multiple LR

- $$y = w_0 + w_1 x_1 + w_2 x_2 \ldots + \cdots w_n x_n + \epsilon$$

- Now, there is a **Plane of Regression**

- w_j is the slope or weight by which x_i has an impact on y
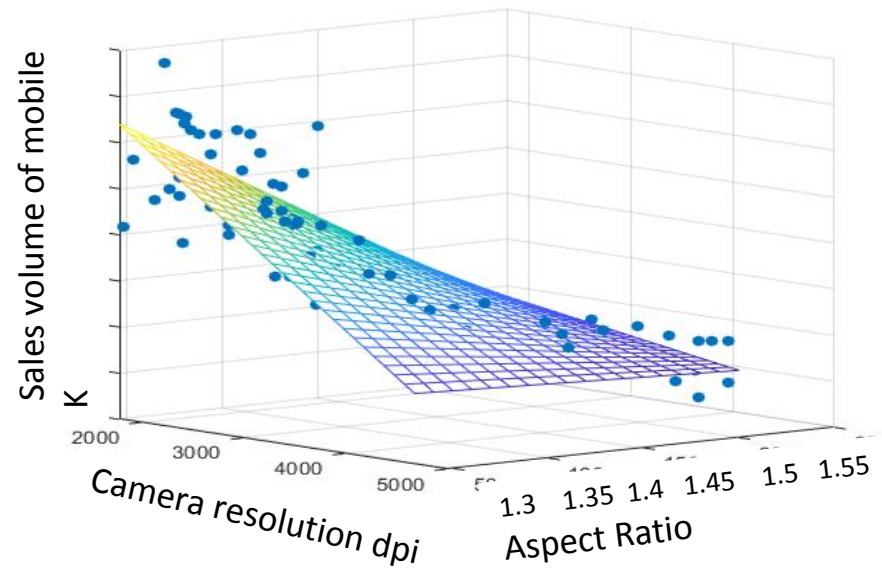
Polynomial LR

$$y = w_0 + w_1 x_1 + w_3 x_1{}^2 + w_4 x_1 x_2 + \varepsilon$$

Box-Office-Collections = w_0 + w_2*Num-of-viewers + w_3*Actor-popularity_Rating^2

- y Being Linear function of Regression Coefficients, one can solve the polynomial regression problem as a linear problem w.r.t {W}

- Each product term such as $x_1^2, x_1 x_2$ etc can be considered a separate regressor

Practical Applications

I. **Predicting response for new $\vec{X}$**

   ➢ **Predict price of new property**

   ➢ **Predict sales volume of new product**

II. **Explain variation in the response variable** y **due to variation in the explanatory variables** $\vec{X} =$ **$\{x_1, x_2\ldots\ldots:\}$**

   ➢ **Explain Stock price variation**

Topic 2: *Model Estimation*

Ordinary Least Squares

- The learning model is found by tuning the regression coefficients: $W=\{w_1,w_2...\}$.

- The Objective is to *Minimize{Loss Function J(W)}*

- Error $e_i = y_i - \hat{y}_i$. Loss Function is ½ of MSE over all *n* training samples. Hence OF is:

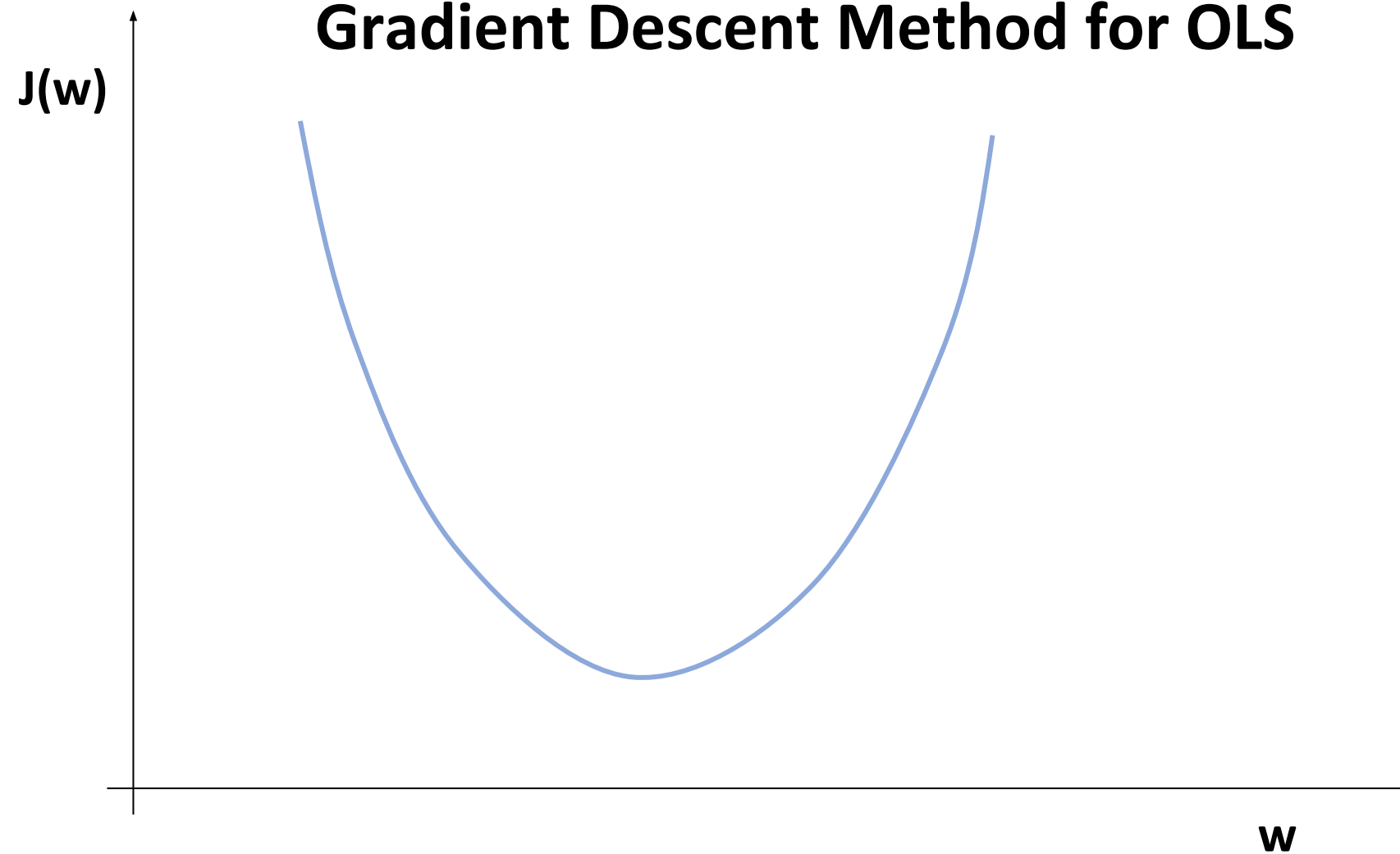$$Minimize\{J(W)\}=Minimize\left\{\left(\frac{1}{2}\times\frac{\sum_{i=1,n}e_i^2}{n}\right)\right\}$$

- This is Ordinary Least Squares (OLS) approach

Parabolic relationship between Loss Function and weights

$$J(w) = \left(\frac{\sum_{i=1,n} e_i^2}{2*n} \right)$$

$$= \left(\frac{\sum_{i=1,n}(y_i - \widehat{y}_i)^2}{2*n} \right)$$

$$= \left(\frac{\sum_{i=1,n}(y_i - (w_0 + w_1 x_1 + \cdots.))^2}{2*n} \right)$$

Gradient Descent Method for OLS

Learning the Bias w_0

- **GD or Delta Learning Rule:** $\Delta w_0 \propto - \dfrac{\partial J(w)}{\partial w_0}$

- **Gradient of Loss function J(W) w.r.t. w_0 is :**

$$\frac{\partial J(w)}{\partial w_0} = \frac{\partial}{\partial w_0} \left(\frac{\sum_{i=1,n}\left(y_i - (w_0 + w_1 x_{1,i} + \cdots)\right)^2}{2 * n} \right)$$

Therefore,

$$\Delta w_0 = -\eta \frac{1}{n} \sum_i \left(y_i - \left(w_0 + w_1 x_{1,i} + \cdots\right)\right)$$

$$\Delta w_0 = -\eta \frac{1}{n} \sum_i e_i$$

Delta Learning Rule for w_1

- GD or Delta Learning Rule: $\quad \Delta w_1 \propto -\dfrac{\partial J(w)}{\partial w_1}$

- Gradient of J(W) w.r.t. w_1 is :

$$\frac{\partial J(w)}{\partial w_1} = \frac{\partial}{\partial w_1}\left(\frac{\sum_{i=1,n}(y_i - (w_0 + w_1 x_{1,i} + \cdots))^2}{2 * n}\right)$$
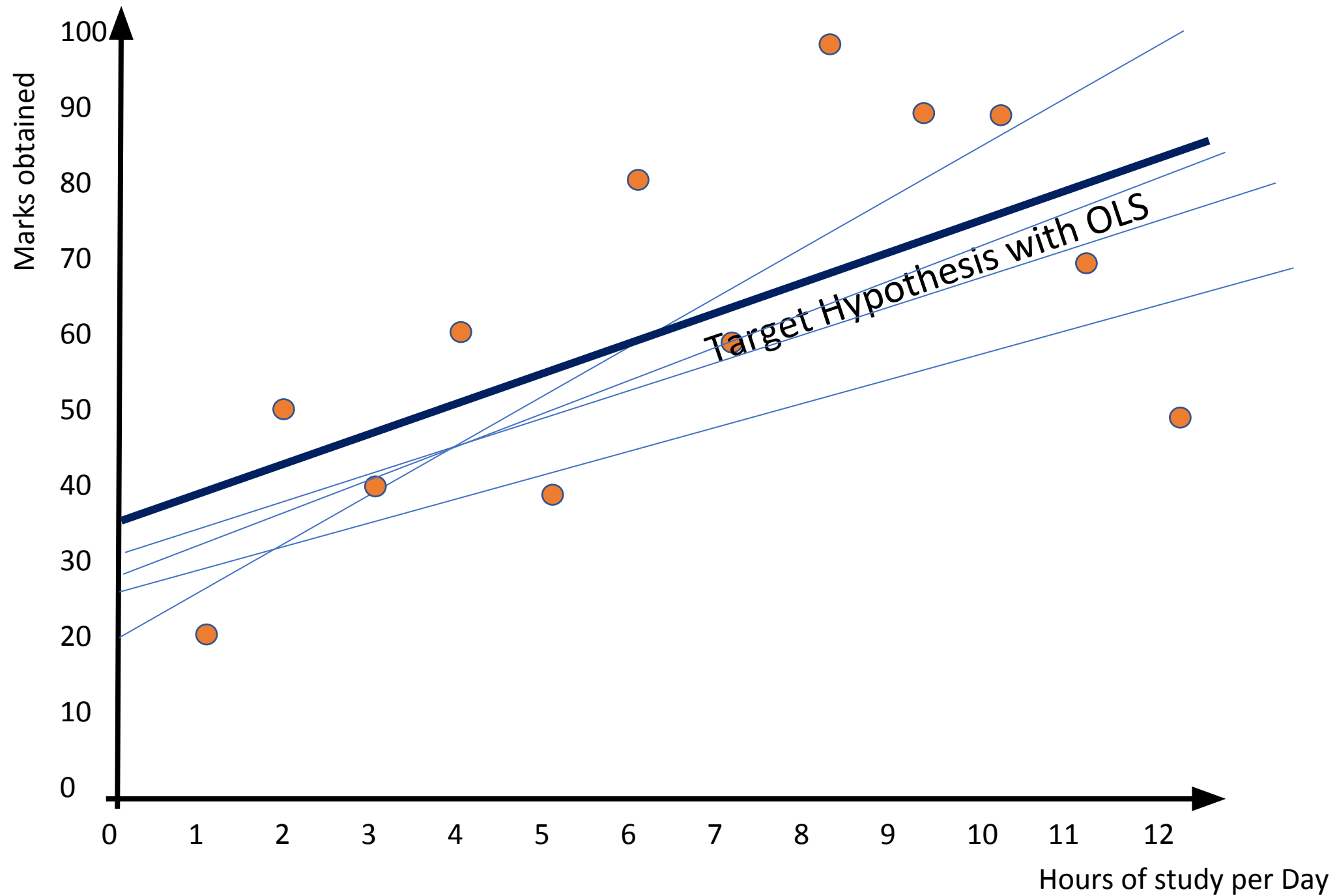
Therefore,

$$\Delta w_1 = -\eta \frac{1}{n}\sum_i (y_i - \hat{y}_i)\, x_{1,i}$$

$$\Delta w_1 = -\eta \frac{1}{n}\sum_i e_i\, x_{1,i}$$

Weight Readjustments

$$w_x(t+1) \leftarrow w_x(t)$$

$$w_0(t+1) = w_0(t) - \eta \frac{1}{n} \sum_i (y_i - \hat{y}_i)$$

$$w_i(t+1) = w_i(t) - \eta \frac{1}{n} \sum_i (y_i - \hat{y}_i) \, x_i$$

Target Hypothesis with OLS

Types of Gradient Descent

- **GD – use all n training samples to readjust weights:**

$$\Delta w_1 = -\eta \frac{1}{n} \sum_{i=1,n} e_i x_i \qquad \Delta w_0 = -\eta \frac{1}{n} \sum_{i=1,n} e_i$$

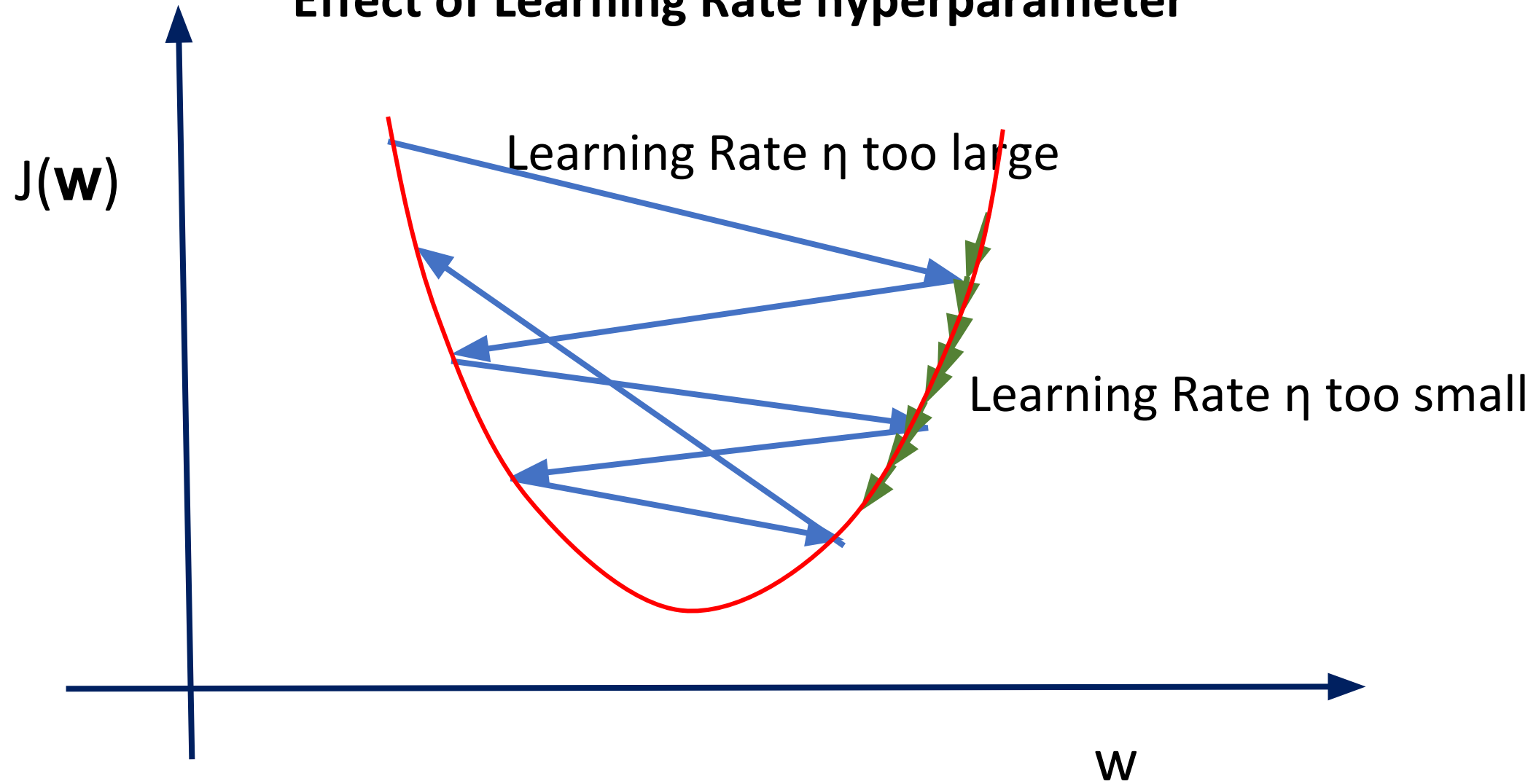$$J(w) = \frac{\sum_{i=1,n}(y_i - \hat{y}_i)^2}{2*n}$$

- **Stochastic GD: use any one sample chosen at random time t_i**

$$\Delta w_1 = -\eta e_i x_i \qquad \Delta w_0 = -\eta e_i$$

- **Mini batch GD: use m<<n samples:**

$$\Delta w_1 = -\eta \frac{1}{m} \sum_{i=1,m} e_i x_i \qquad \Delta w_0 = -\eta \frac{1}{m} \sum_{i=1,m} e_i$$

Effect of Learning Rate hyperparameter

J(**w**)

Learning Rate η too large

Learning Rate η too small

w

Recap

- Linear Regression is a supervised regression method, with continuous response and one or more explanatory variables/ regressors/ attributes

- The hypothesis is given by:

$$y_i = f(X_i) = w_0 + \sum_{j=1..m} w_j x_{i,j}^k + \epsilon_i$$

- Hypothesis space search is carried out by adjusting the regression coefficients w_0, w_1......

- LR may be: simple LR with one regressor, Multiple LR with >1 regressors, Polynomial LR with regressors of higher degree or Multivariate LR with multiple coordinated response variables

Recap

- Ordinary Least Squares (OLS) technique, tunes the weight parameters to find the best mapping between attributes and response, with least MSE loss function.

- Gradient Descent is a method for OLS, based on the idea that weights are adjusted in direct proportion to the gradient of the loss function, and in reverse direction

- Stochastic GD method considers one random data point at a time to calculate and perform weight adjustment, Batch GD takes a few of them, and GD uses all of them.