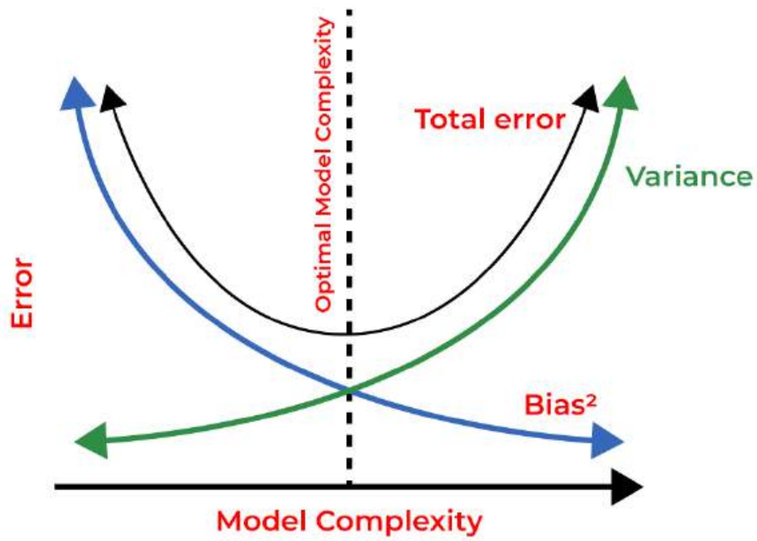
13.02.2024

BITS F464: Machine Learning

REGRESSION MODELS

Chittaranjan Hota, Sr. Professor
Dept. of Computer Sc. and Information Systems
hota@hyderabad.bits-pilani.ac.in

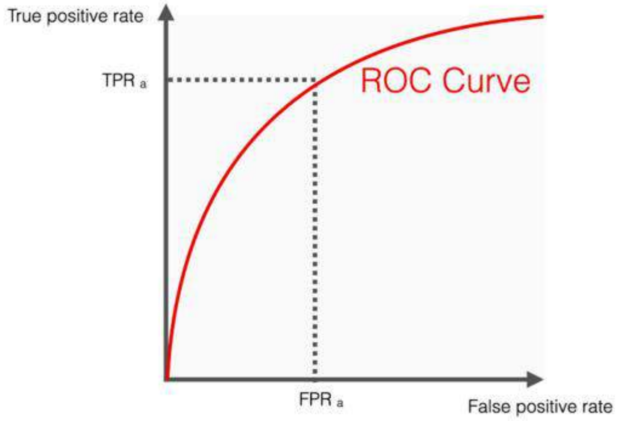
Recap:



Confusion Table:
truth\prediction

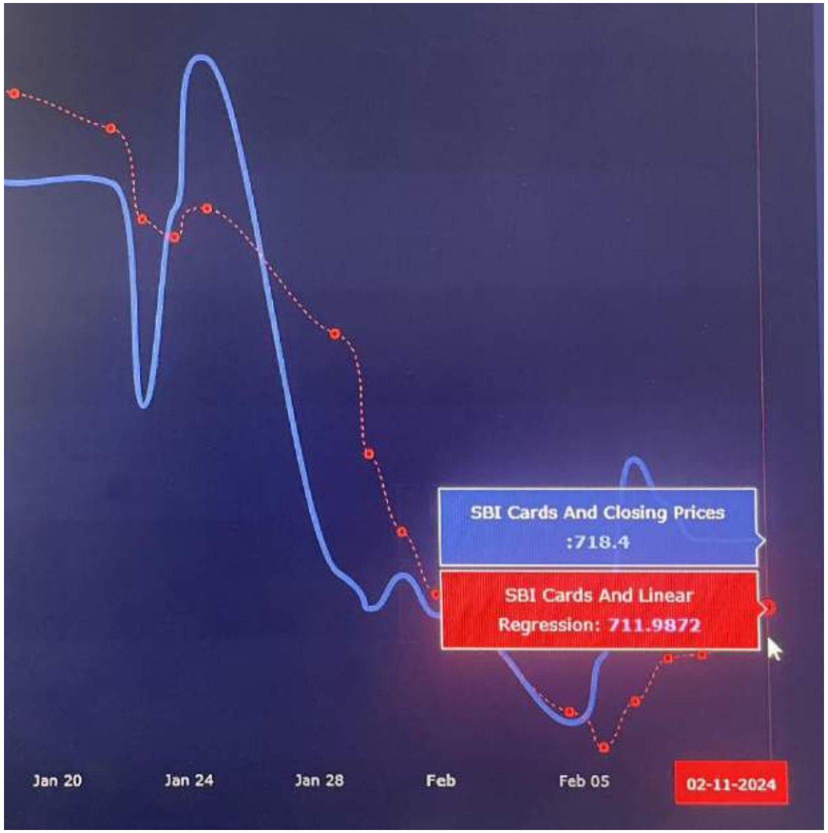
	1	2	3	4	5	6	7	8
1	29	0	0	0	0	2	0	0
2	0	100	0	0	9	0	3	0
3	0	1	65	1	0	0	5	0
4	0	0	4	13	0	0	0	2
5	0	6	0	0	61	3	0	0
6	0	0	0	0	7	23	0	0
7	0	9	6	0	0	0	42	0
8	0	0	0	1	0	0	0	7

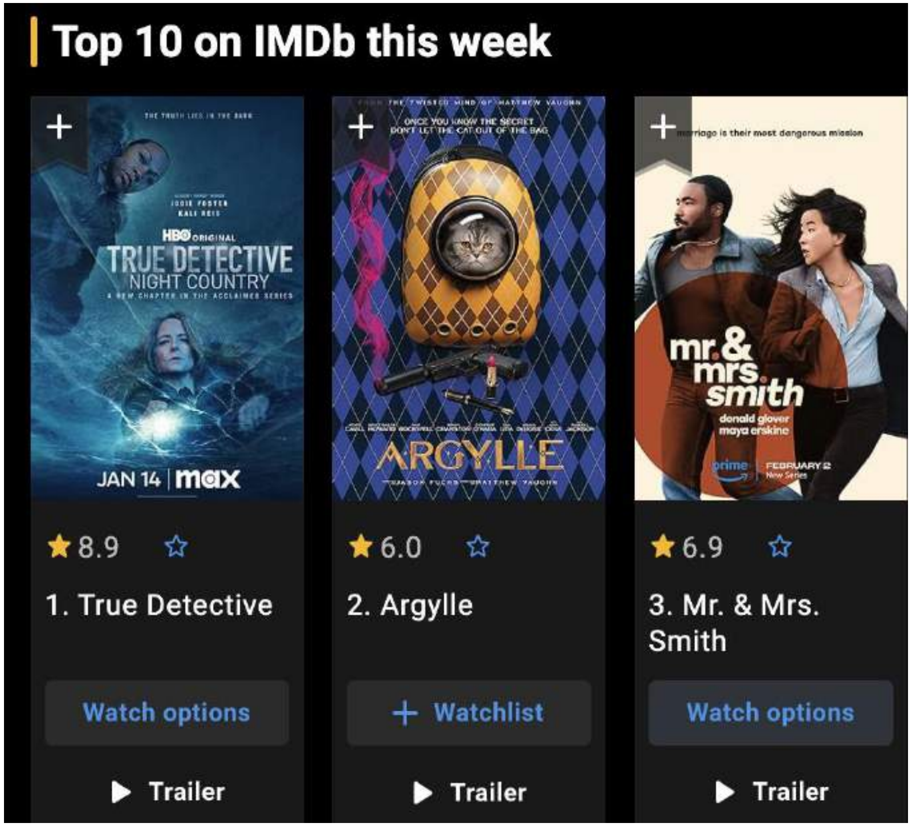
Total: 399



$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

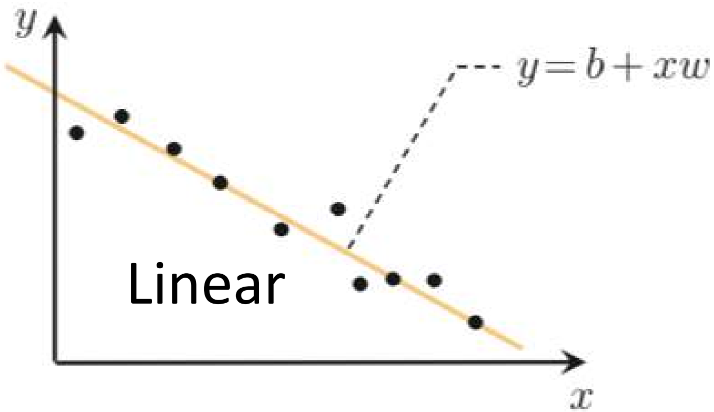
What Type of Problems can you solve?

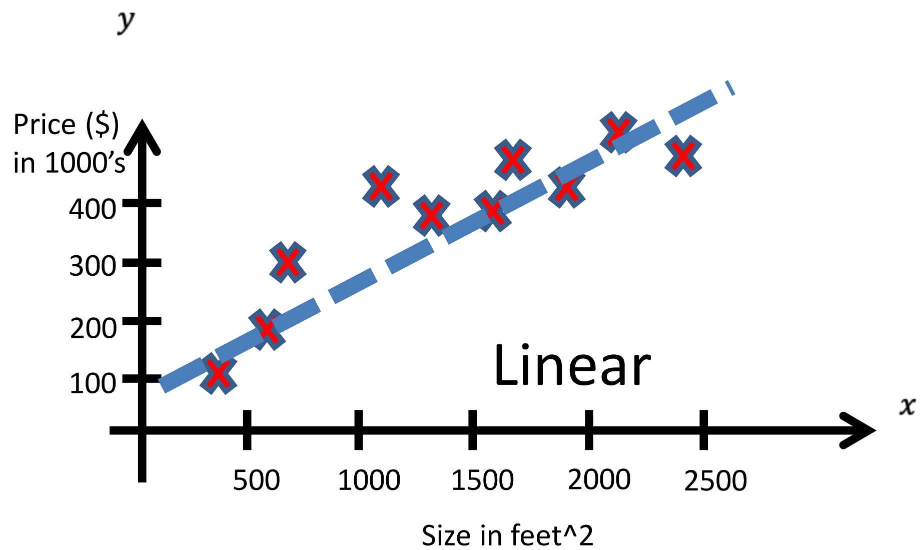


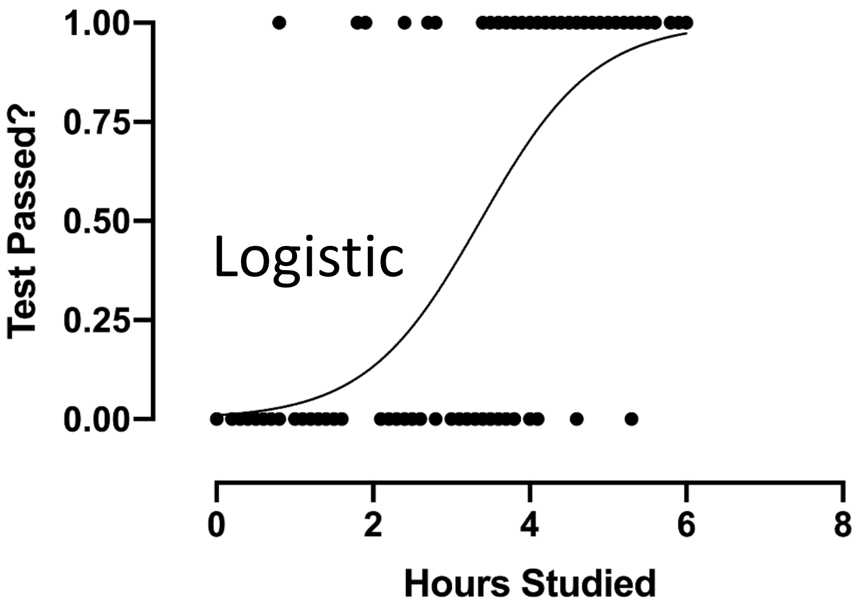


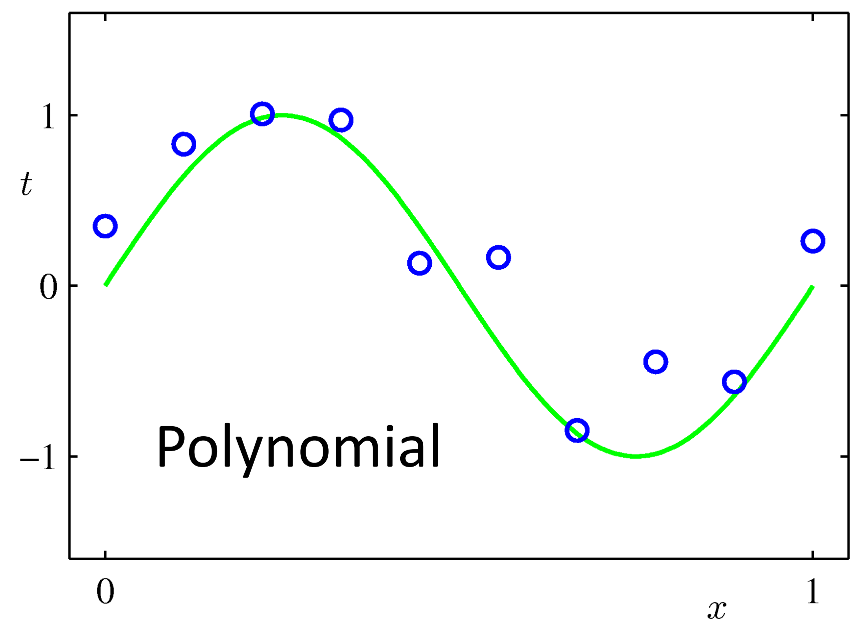
Source: www.macroaxis.com/stocks/

https://www.imdb.com/



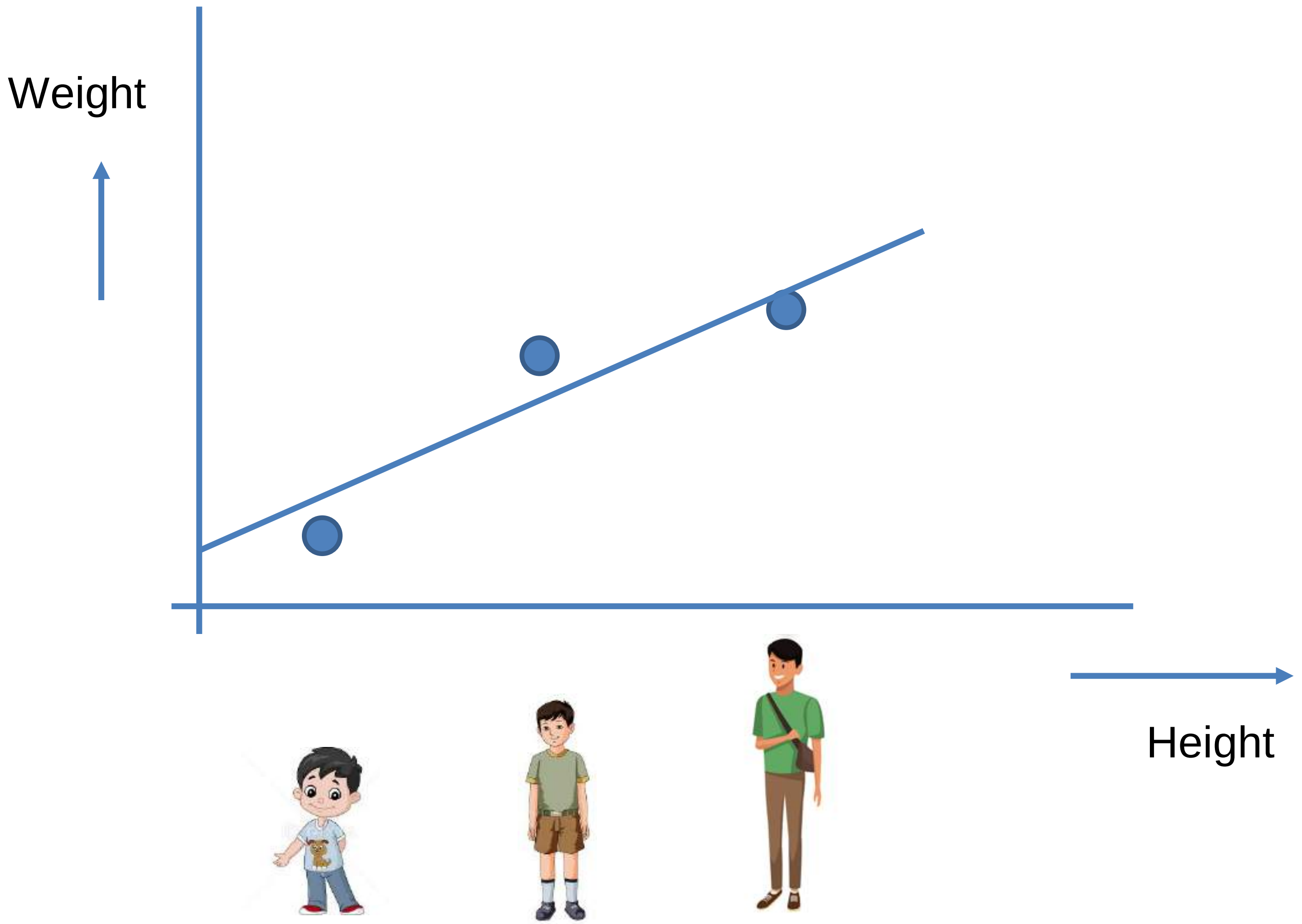






Different types of Regression for different purposes. Ridge, Lasso, Bayesian, …

Regression with Scalar Input(Univariate)

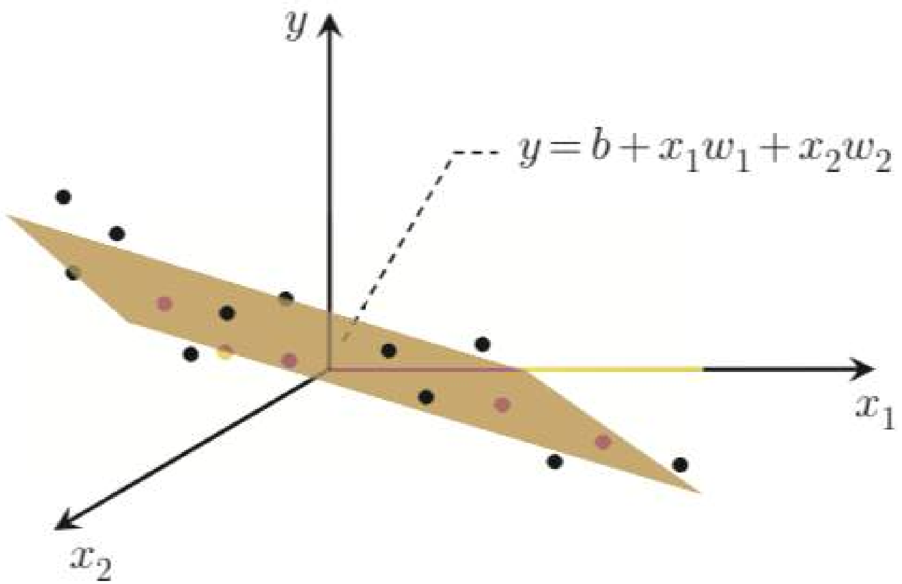


Simple Linear Regression

With Vector inputs (more covariates)

LARGEST ECONOMIES IN THE WOR		
Rank	Country	GDP (in USD Bil
1.	United States of America	26,954
2.	China	17,786
3.	Germany	4,430
4.	Japan	4,231
5.	India	3,730
6.	United Kingdom (UK)	3,332
7.	France	3,052
8.	Italy	2,190
9.	Brazil	2,132
10.	Canada	2,122

https://currentaffairs.adda247.com/



- Unemployment rate, education level, population count, land area, income level, investment rate, life expectancy, … (Multiple Linear Regression: Multi-variate)

Another Example of Multi-variate Regression

Sales = b + w_1 weather + w_2 money + w_3 day

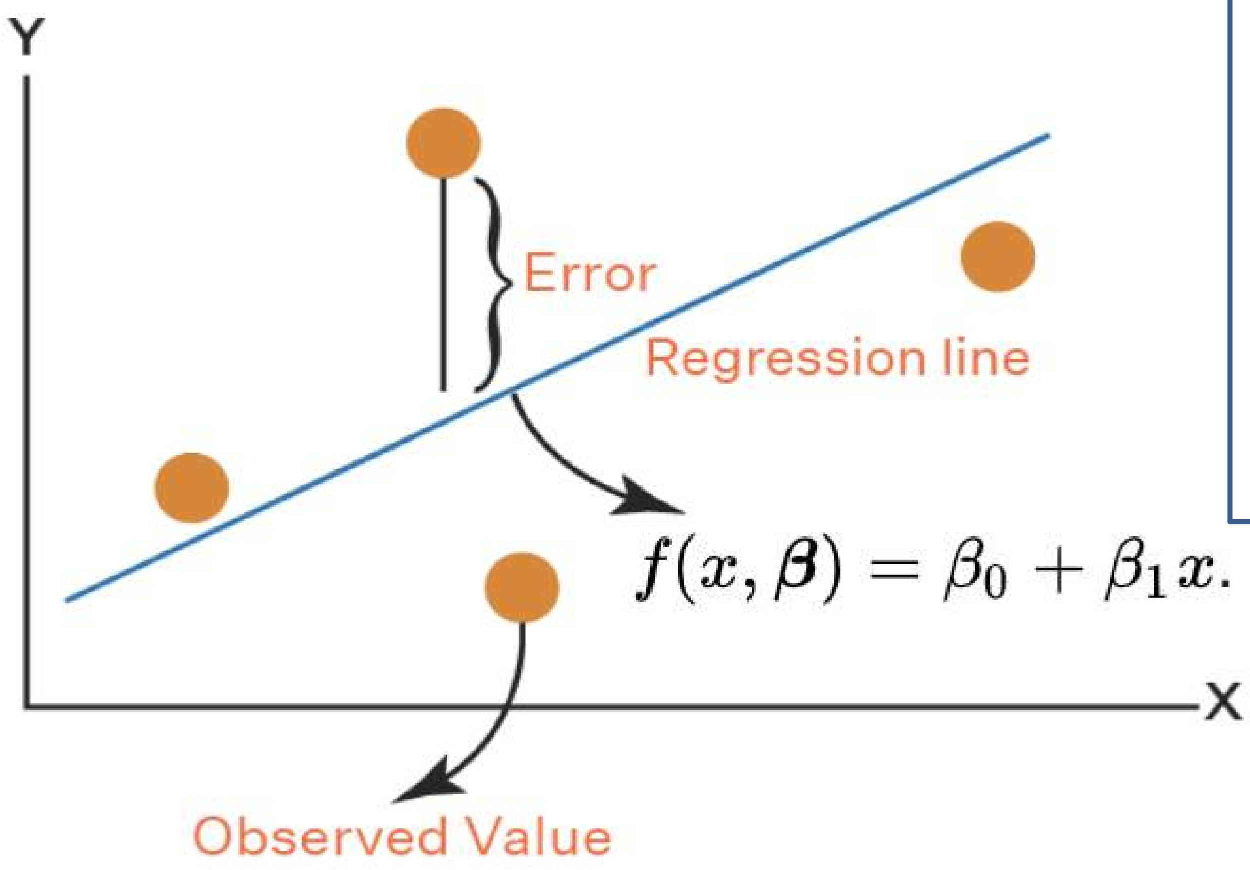
Regression:

Process of finding out relationship between a dependent variable (outcome/ response/ label) and one or more independent variables (predictors/ covariates/ explanatory variables/ features)

Independent variables (X): weather (rainy, sunny, cloudy), amount in hand, day type (working, holiday), Dependent variable: Y (Sales)

How the dependent variable (Y) will react to each variable X taken independently?

Best Fitting a Line: Least Squares Method



If $\beta_1 > 0$

How are X and Y related?

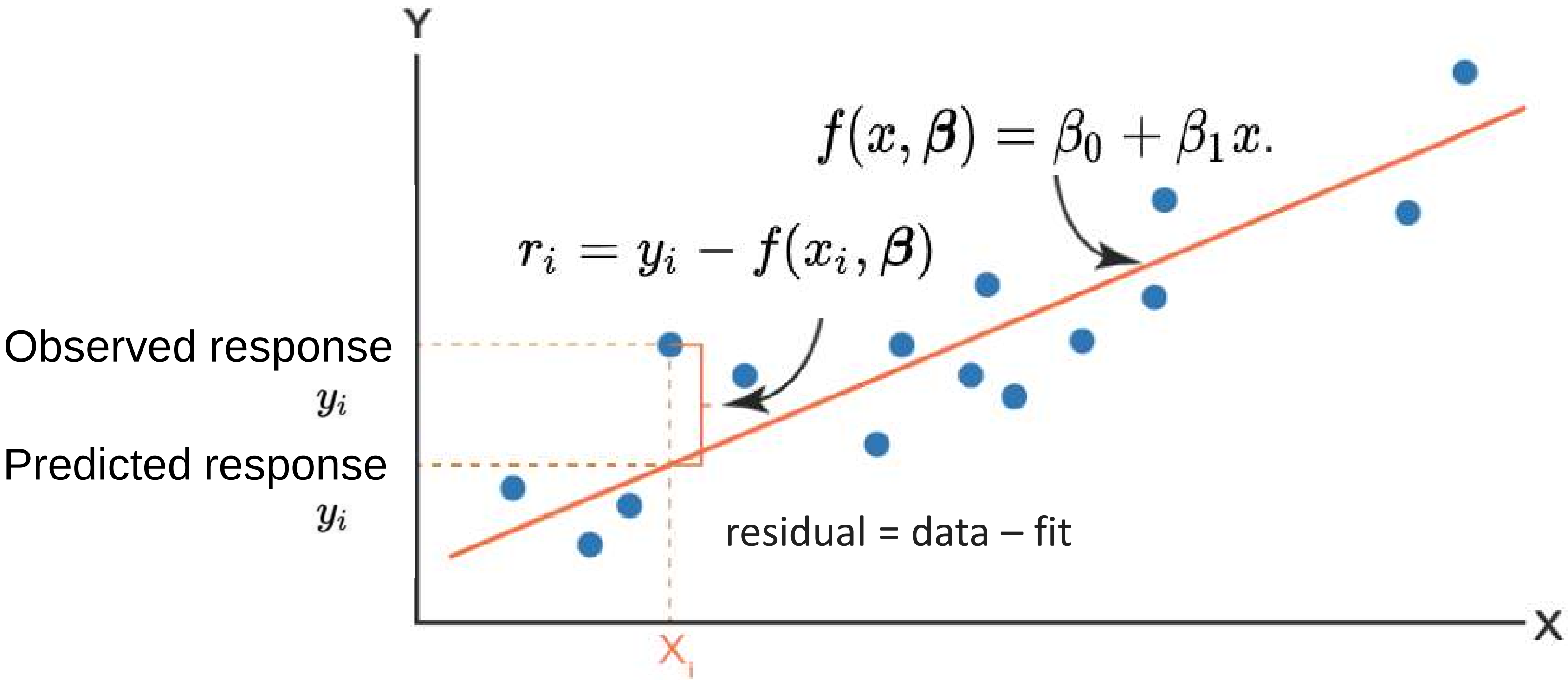
If $\beta_1 < 0$?

If $\beta_1 == 0$?

The target function: $f(x, \boldsymbol{\beta})$, where m adjustable parameters are held in vector $\boldsymbol{\beta}$.

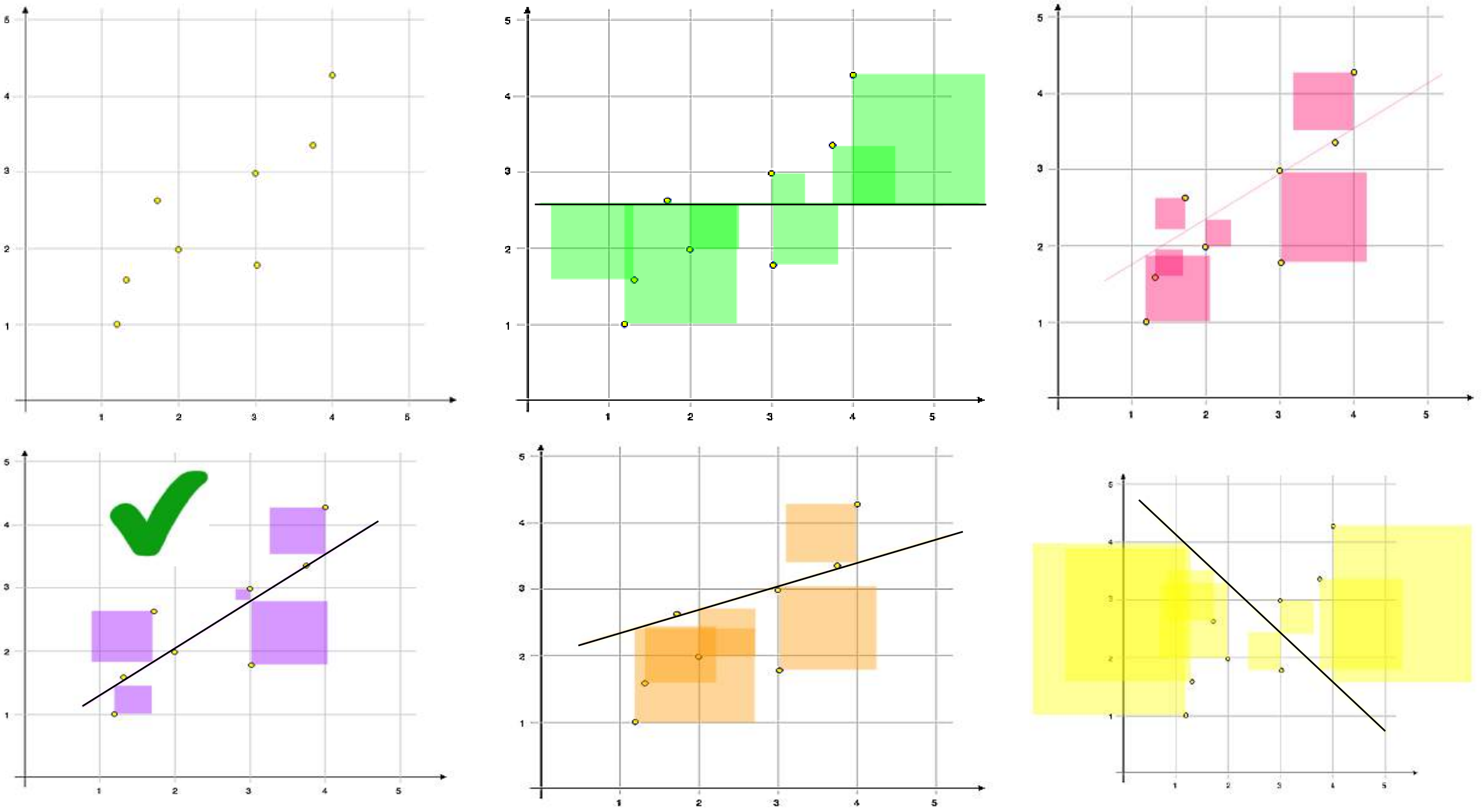
Simple Linear Regression

Best Fitting a Line: Least Squares Method



Find out the optimal parameter values by minimizing the sum of squared residuals $S = \sum_{i=1}^{n} r_i^2$

Can you choose the best-fit line?

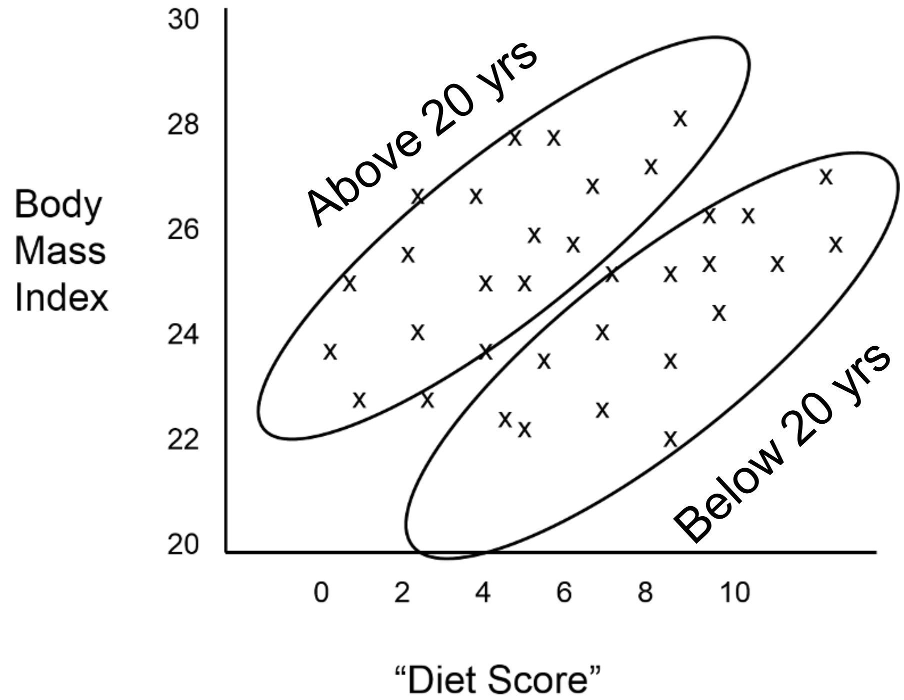


Hypothetically: Say, weight = 2 + 1.5 height

Multiple Linear Regression Analysis

Diet Score	Age>20	BMI
4	1	27
7	1	29
6	0	23
2	0	20
3	1	21
...	...	...

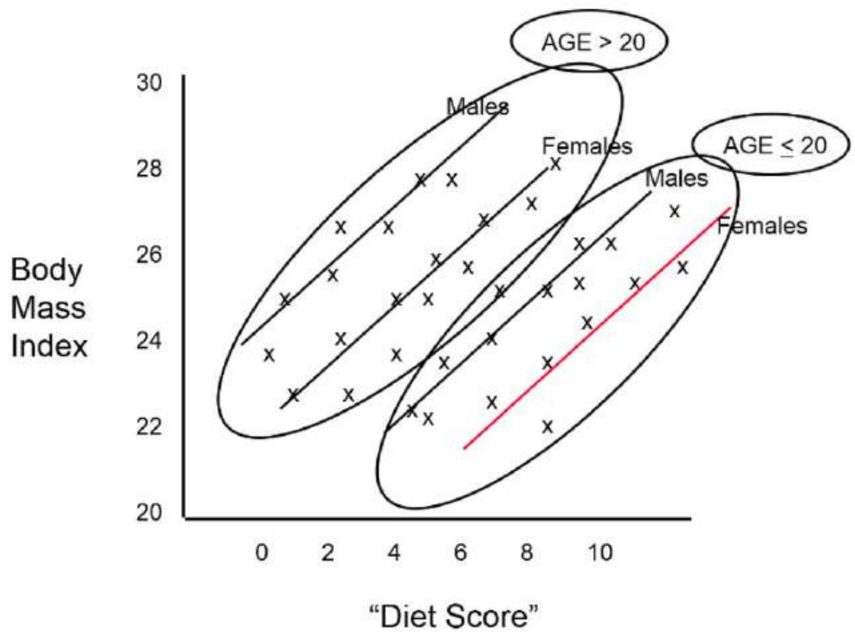
(hardly any association between the two)



(People are clustered based on age)

Img. Source: https://sphweb.bumc.bu.edu/

Continued…



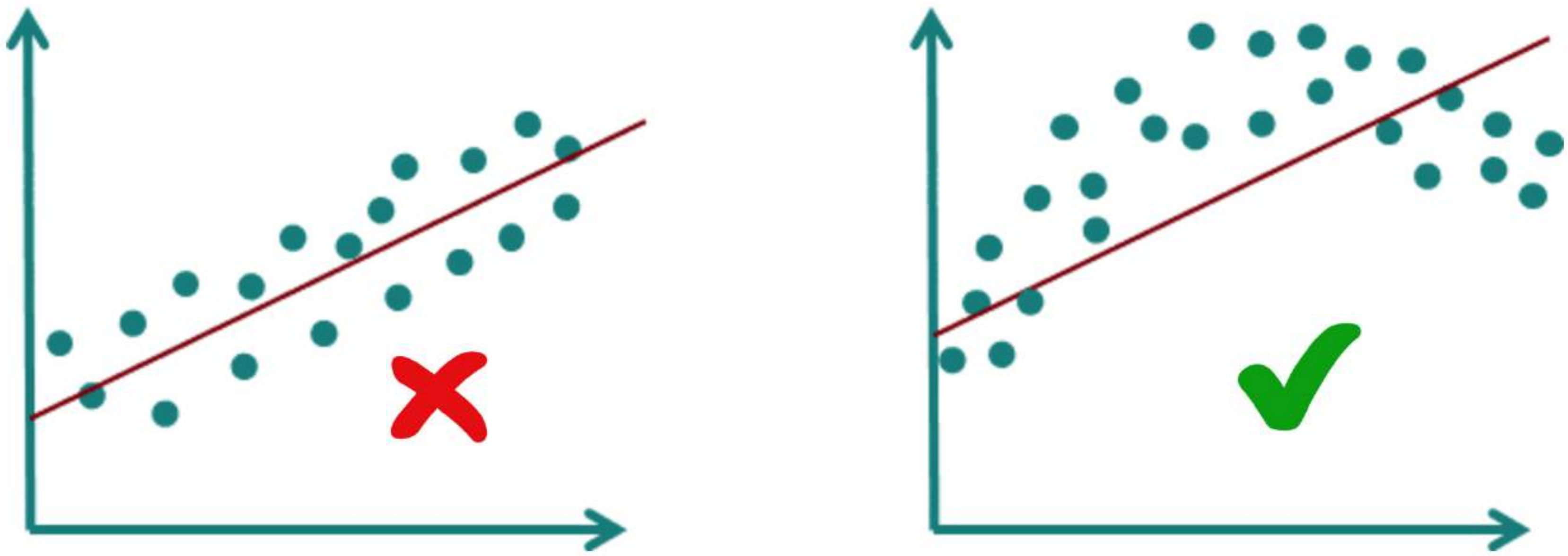
Diet Score	Male	Age>20	BMI
4	0	1	27
7	1	1	29
6	1	0	23
2	0	0	20
3	0	1	21
...	...	...	...

BMI = 18 + 1.5 (diet score) + 1.6 (male) + 4.2 (age > 20)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

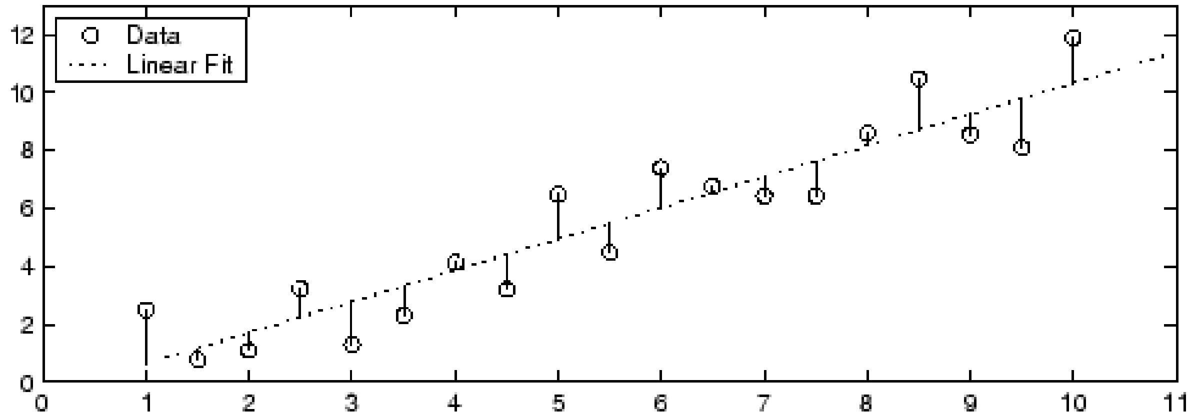
Img. Source: https://sphweb.bumc.bu.edu/

Non-linear relationships

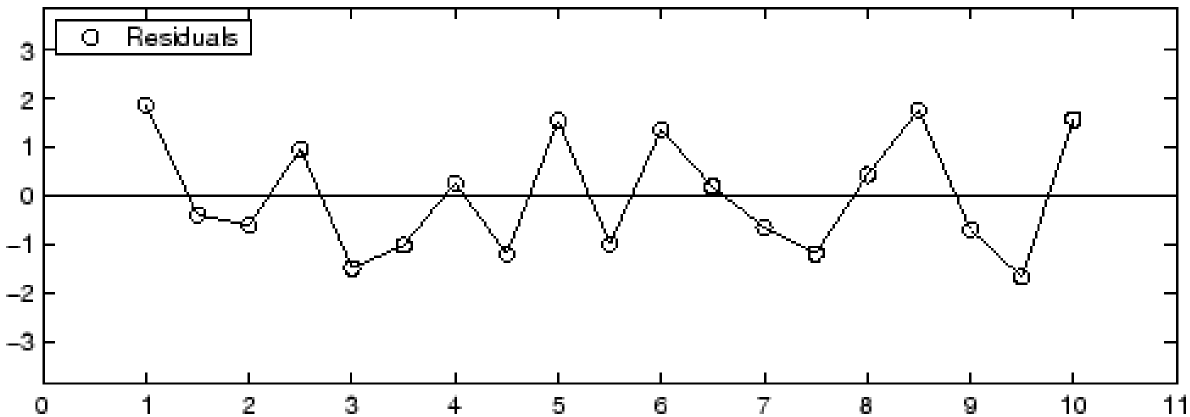


Examples: House price based on Floor area, Electricity consumption based on no. of household members and appliances being used.

Analyzing Residuals

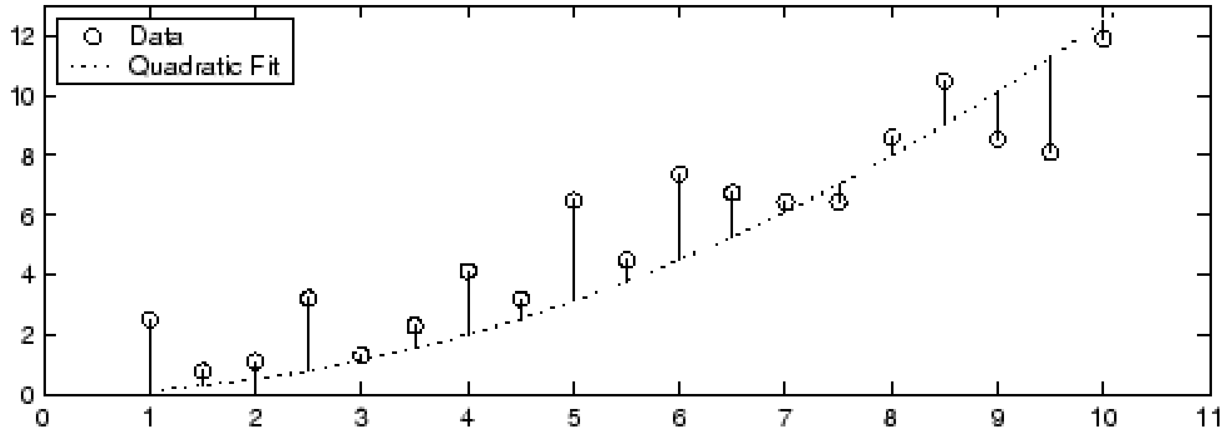


Model describes data well or poor?

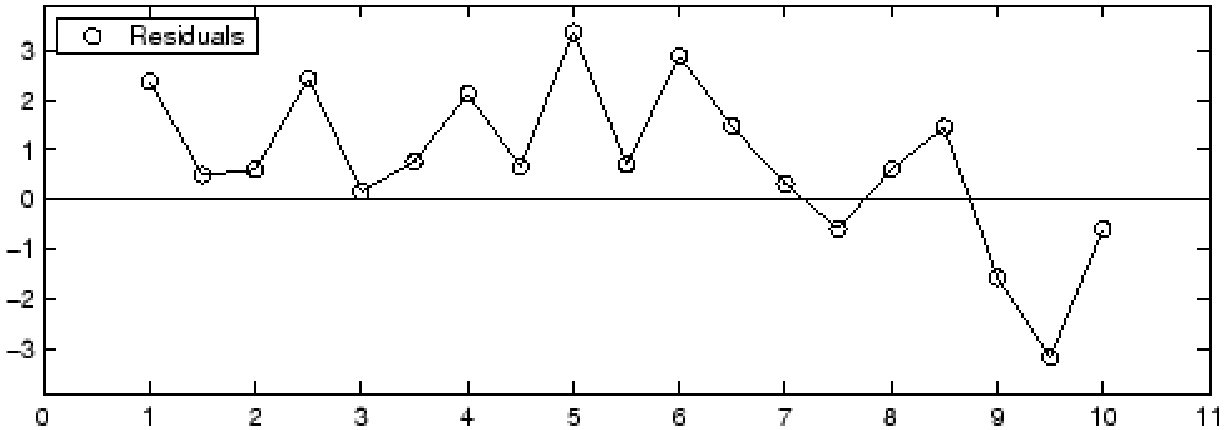


Randomly scattered around zero

Continued...



Model includes a Second-degree polynomial (quadratic term)



Systematically positive for much of the data.

Good or bad fit?

Non-linear relations using Linear models?

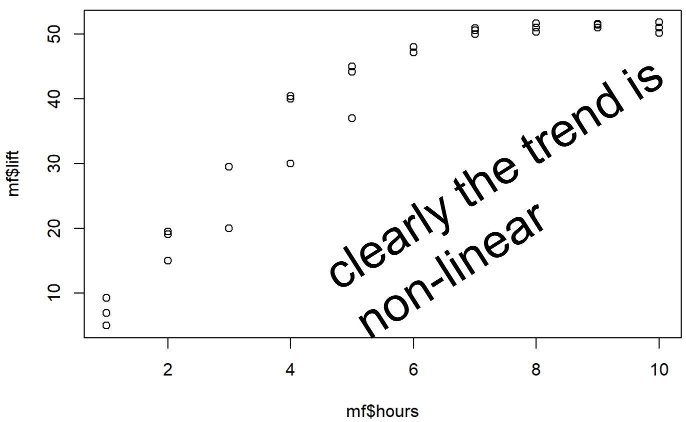
- Feature Engineering: Engineer new features by transforming the existing ones to capture non-linear relationships, e.g, you can include polynomial features (e.g., quadratic, cubic).
- Using Basis Functions: Instead of using the original features, you can use basis functions, which are transformations of the original features, e.g polynomial basis functions, Gaussian radial basis functions, or sigmoidal basis functions.
- Regularization: Ridge regression (L2 regularization) or Lasso regression (L1 regularization) to penalize large coefficients.
- Non-linear Regression Models: If the relationship is highly non-linear, use decision trees, random forests, support vector machines, or neural networks.

We will see some of these…

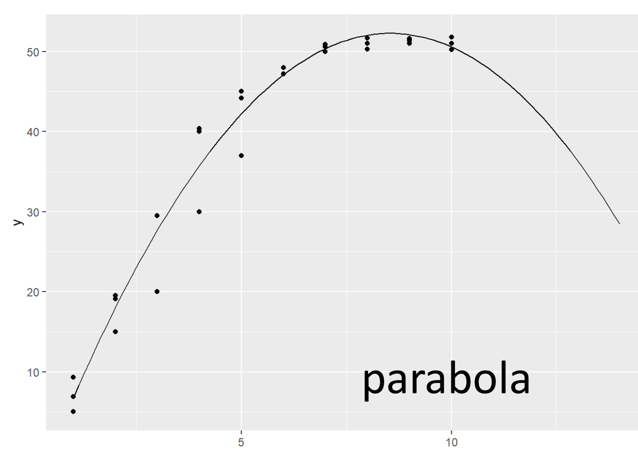
An Example

##	name	lift	hours
## 1	Person 01	5.0	1
## 2	Person 02	15.0	2
## 3	Person 03	20.0	3
## 4	Person 04	30.0	4
## 5	Person 05	37.0	5
## 6	Person 06	48.0	6
## 7	Person 07	50.0	7
## 8	Person 08	51.0	8
## 9	Person 09	51.0	9
## 10	Person 10	51.0	10
## 11	Person 11	6.9	1
## 12	Person 12	19.5	2
## 13	Person 13	29.5	3
## 14	Person 14	40.4	4
## 15	Person 15	45.0	5
## 16	Person 16	48.0	6
## 17	Person 17	50.9	7
## 18	Person 18	50.3	8
## 19	Person 19	51.4	9
## 20	Person 20	51.8	10
## 21	Person 21	9.3	1
## 22	Person 22	19.1	2
## 23	Person 23	29.5	3
## 24	Person 24	40.0	4
## 25	Person 25	44.2	5
## 26	Person 26	47.2	6
## 27	Person 27	50.6	7
## 28	Person 28	51.7	8
## 29	Person 29	51.6	9
## 30	Person 30	50.2	10

- *lift* is the dependent variable, and the independent variable is the '*hours*', i.e the time spent in weight lifting.



- We add a quadratic term as an independent variable in the model. y = x^2

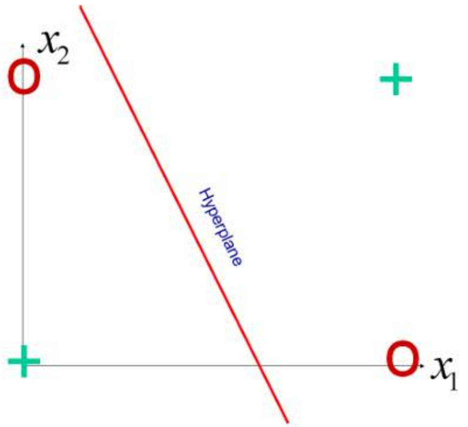


$$\hat{lift} = -6.13 + 13.67 * hours - 0.8 * hours^2$$

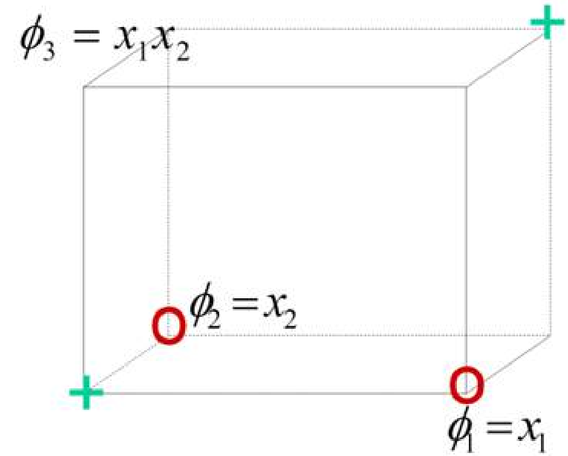
##	name	lift	hours	hoursSq
## 1	Person 01	5.0	1	1
## 2	Person 02	15.0	2	4
## 3	Person 03	20.0	3	9
## 4	Person 04	30.0	4	16
## 5	Person 05	37.0	5	25
## 6	Person 06	48.0	6	36
## 7	Person 07	50.0	7	49
## 8	Person 08	51.0	8	64
## 9	Person 09	51.0	9	81
## 10	Person 10	51.0	10	100
## 11	Person 11	6.9	1	1
## 12	Person 12	19.5	2	4
## 13	Person 13	29.5	3	9
## 14	Person 14	40.4	4	16
## 15	Person 15	45.0	5	25
## 16	Person 16	48.0	6	36
## 17	Person 17	50.9	7	49
## 18	Person 18	50.3	8	64
## 19	Person 19	51.4	9	81
## 20	Person 20	51.8	10	100
## 21	Person 21	9.3	1	1
## 22	Person 22	19.1	2	4
## 23	Person 23	29.5	3	9
## 24	Person 24	40.0	4	16
## 25	Person 25	44.2	5	25
## 26	Person 26	47.2	6	36
## 27	Person 27	50.6	7	49
## 28	Person 28	51.7	8	64
## 29	Person 29	51.6	9	81
## 30	Person 30	50.2	10	100

Source: https://bookdown.org/

Basis Functions: Why are they needed?



Linear or non-linear?



Let us add a basis function x_1x_2 into the input (this term couples two terms non-linearly)

With the third input $z = x_1x_2$ the XOR becomes linearly separable.

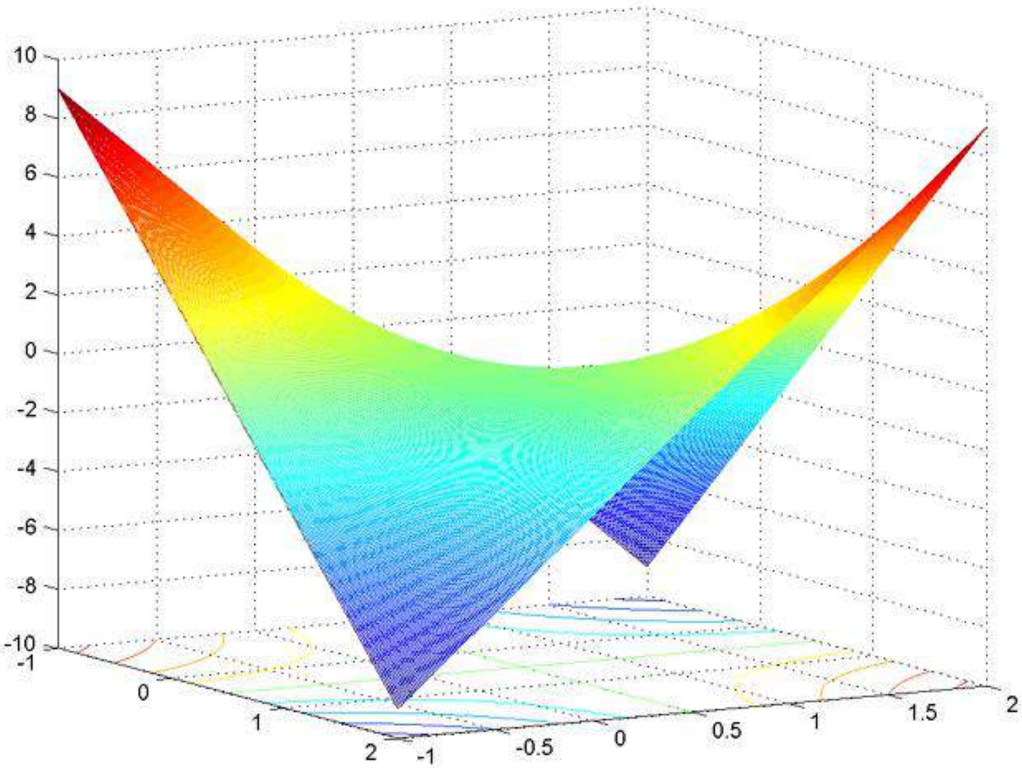
$$f(\mathrm{x}) = 1 - 2x_1 - 2x_2 + 4x_1x_2 = \phi_1(x) - 2\phi_2(x) - 2\phi_3(x) + 4\phi_4(x)$$

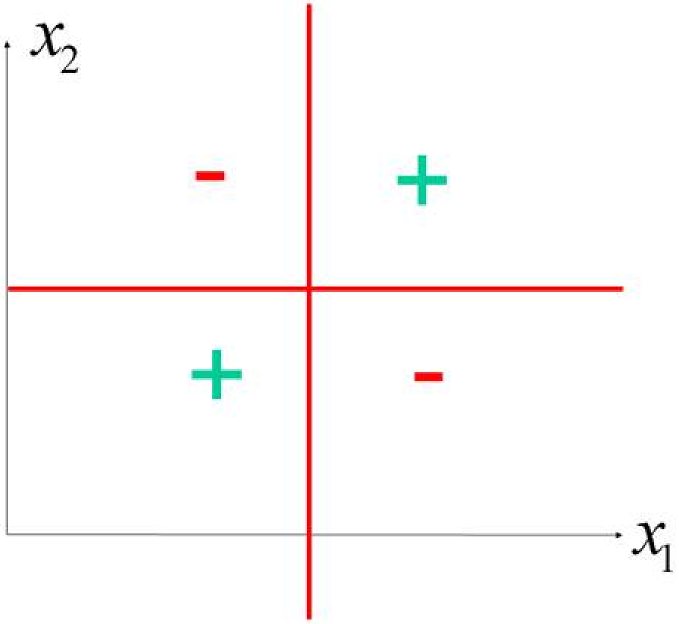
with $\phi_1(x) = 1, \phi_2(x) = x_1, \phi_3(x) = x_2, \phi_4(x) = x_1x_2$

Acknowledgement: Volker Tresp's presentation

Continued…

$$f(\mathbf{x}) = 1 - 2x_1 - 2x_2 + 4x_1x_2$$





Acknowledgement: Volker Tresp's presentation

What are Basis Functions?

Simplest model of Linear Regression: $y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$

Key Property: Linear function of parameters. Also, it is a linear function of its input variables → Imposes serious limitations on the model.

Basis functions come to rescue (called derived features in machine learning) are building blocks for creating more complex functions

For example, individual powers of x: the basis functions 1, x, x^2, x^3… can be combined together to form a polynomial function.

Basis functions $\phi(x)$ extend this class of models by considering linear combinations of handpicked fixed nonlinear functions of the input variables.

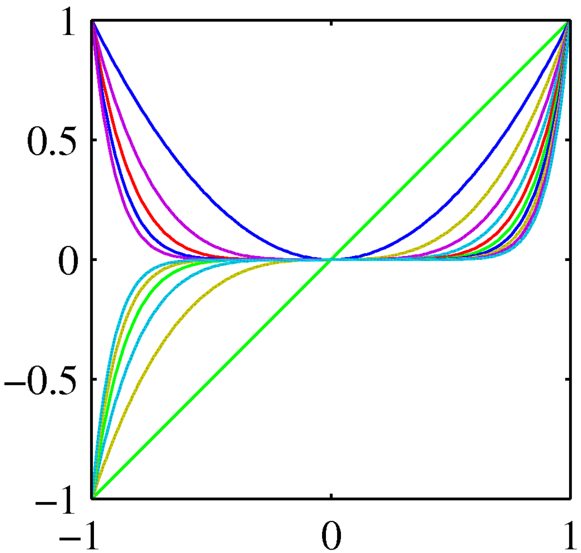
Non linearity in the data while keeping linearity in parameters.

(vector form) $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$ or $y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$

Where, $\phi(\mathbf{x}) = [\phi_0(x_1), \phi_1(x_2), \ldots, \phi_{M-1}(x_n)]^T$ and $\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}}$

Basis functions for Non-linearity

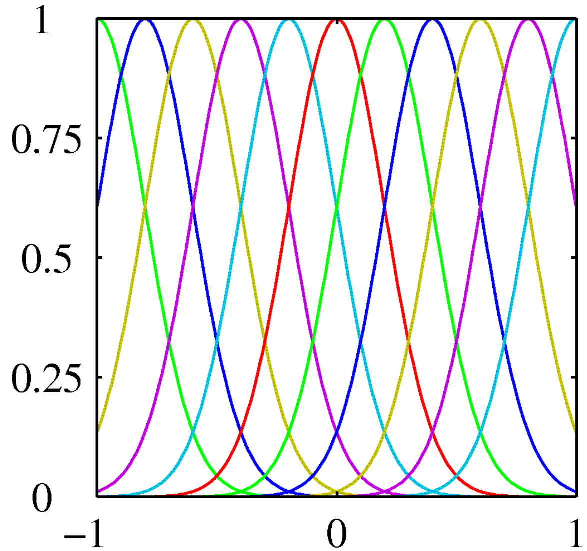
$$\phi_j(x) = x^j$$



(Polynomial basis function)

Global: a small change in x affects all basis functions

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$



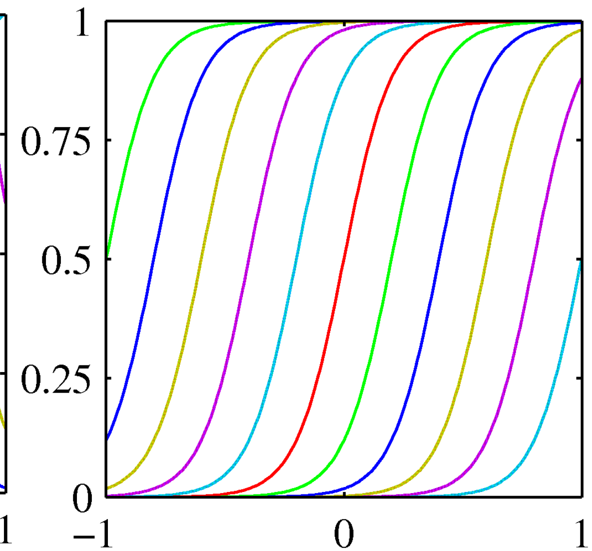
(Gaussian basis function)

Local: a small change in x only affects nearby basis functions.

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

Where,

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

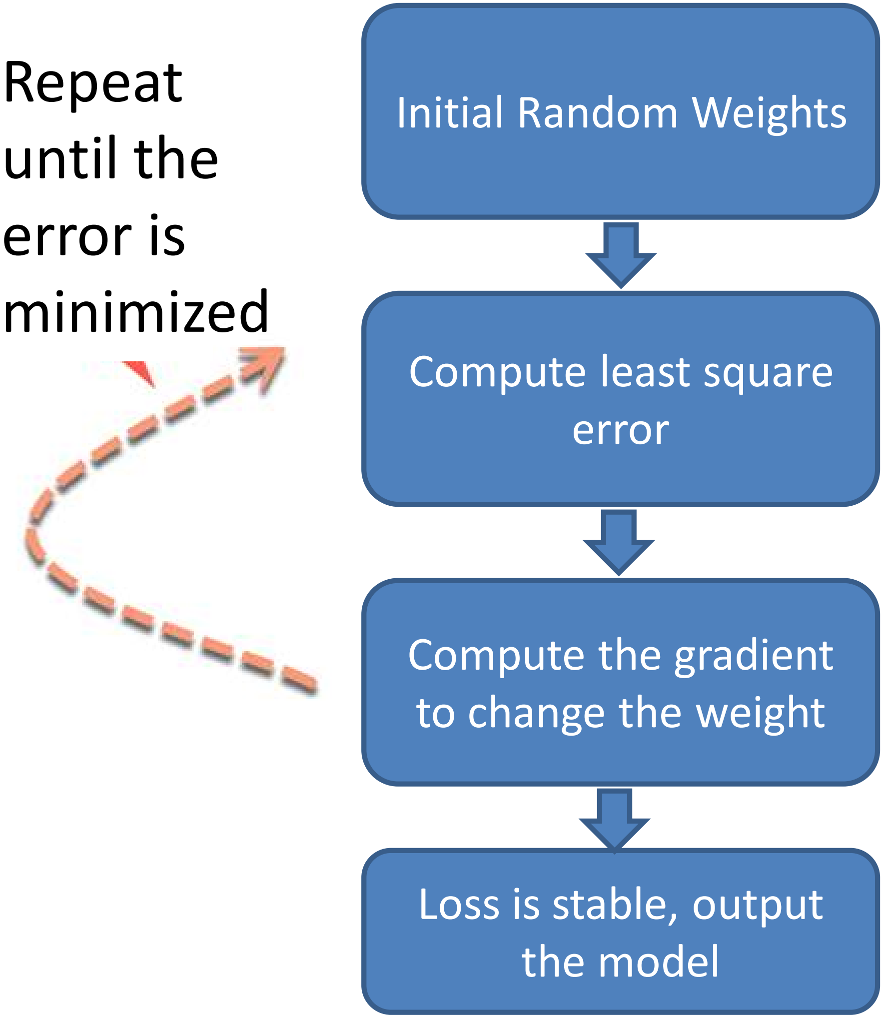


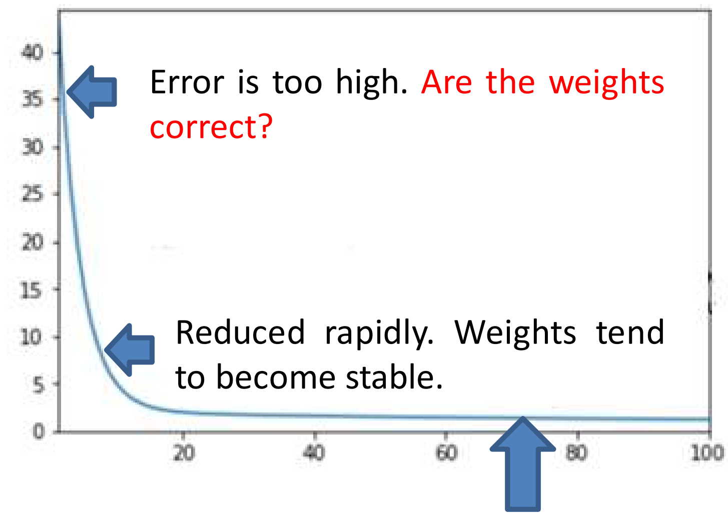
(Sigmoidal basis function)

Local: a small change in x only affects nearby basis functions.

The Learning Algorithm

Repeat until the error is minimized





No more change of the loss/ cost function. Model found best weights.

An Example of house price prediction

Size in sq. feet (x)	Price in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

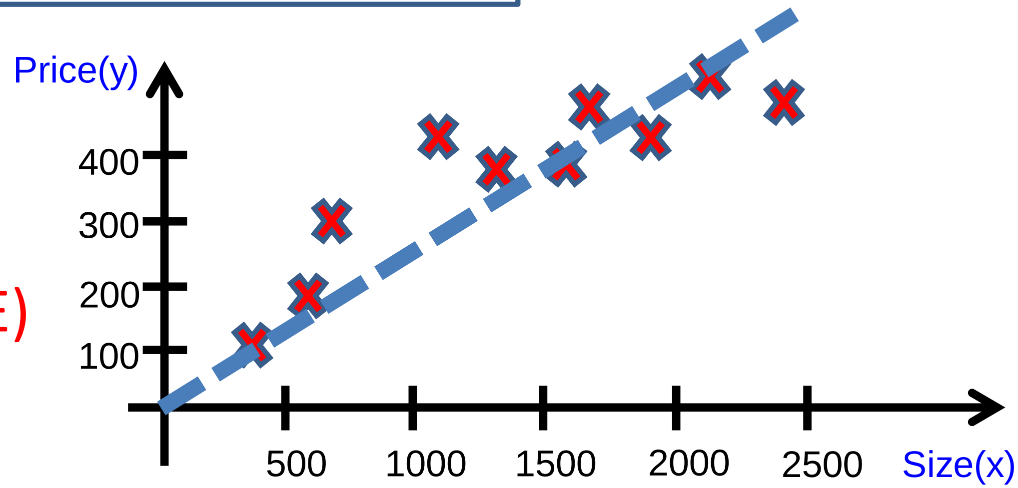
Training Set

$$y = h_\theta(x) = \theta_0 + \theta_1 x$$

What is the value of θ_0 ?

Minimize Cost/ Loss: (MSE)

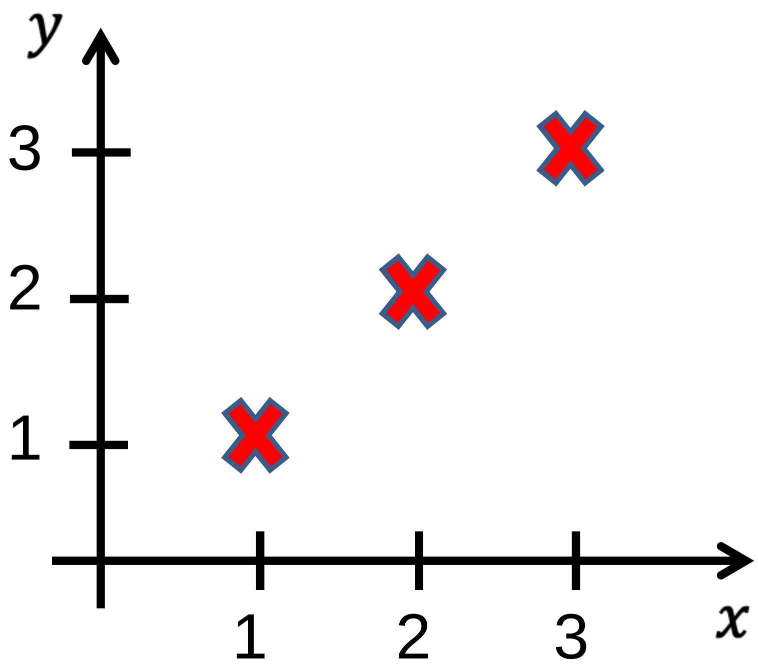
$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$



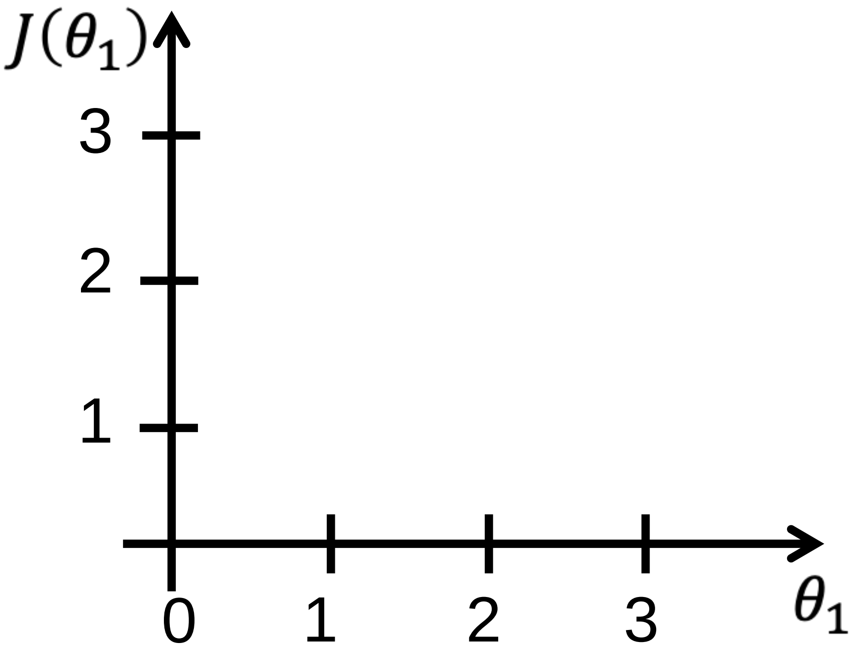
The division by 2 is for convenience and doesn't fundamentally change the result; it simplifies the derivative computation when optimizing models.

Minimizing the Cost Function

$h_\theta(x)$, function of x



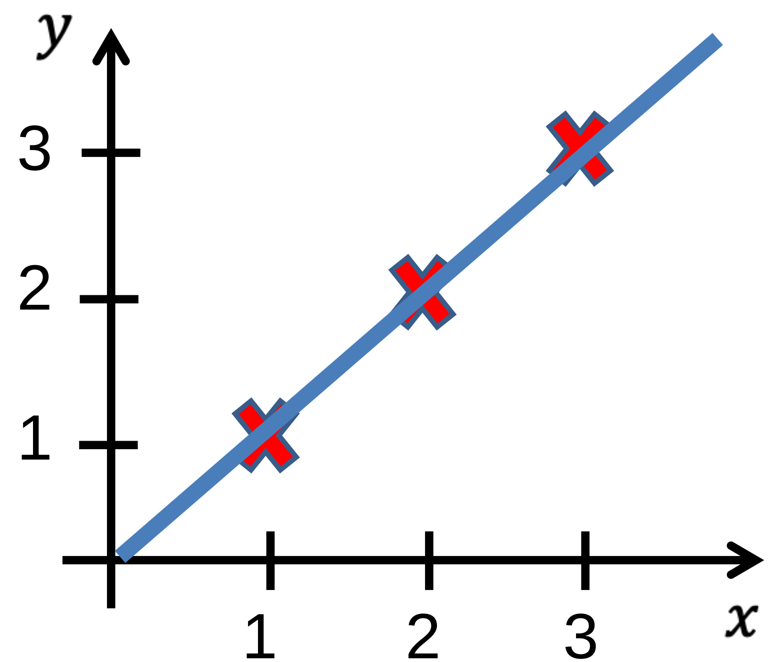
$J(\theta_1)$, function of θ_1



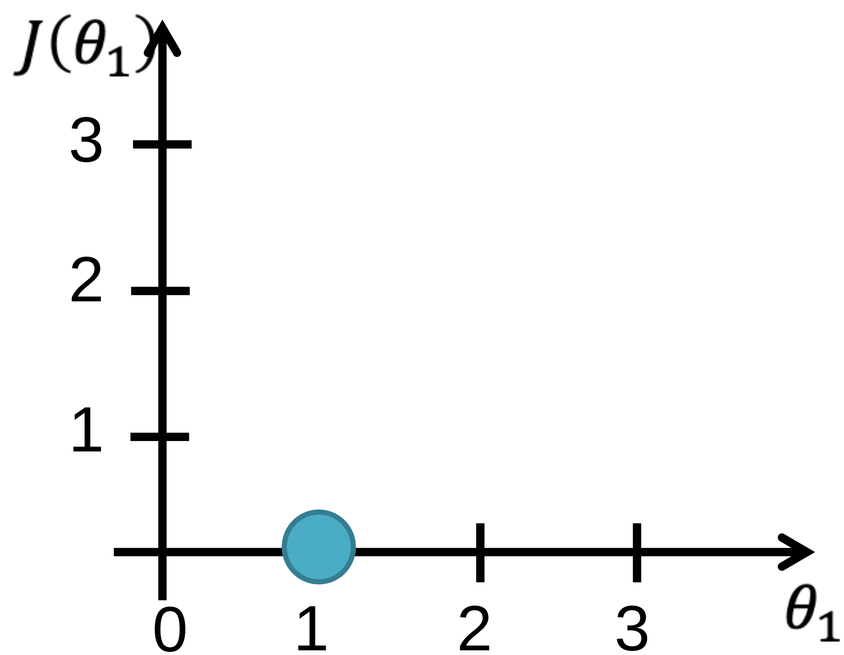
Acknowledgement: Andrew Ng, Stanford

Continued...

$h_\theta(x)$, function of x



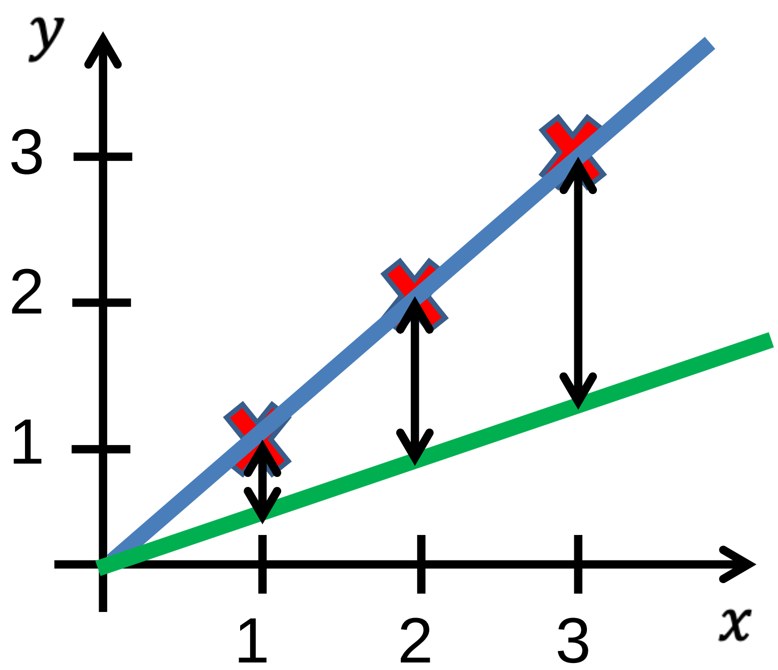
$J(\theta_1)$, function of θ_1



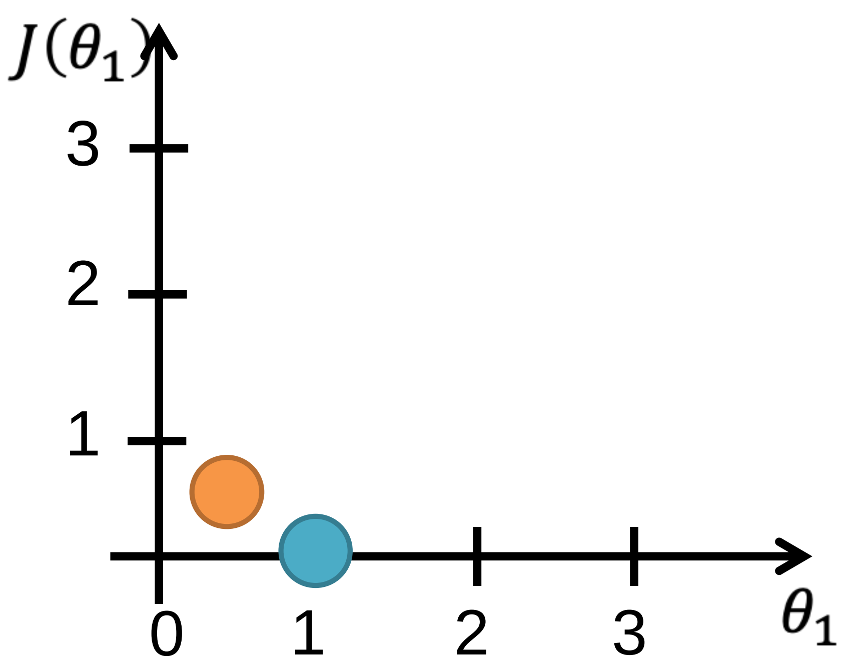
Acknowledgement: Andrew Ng, Stanford

Continued…

$h_\theta(x)$, function of x



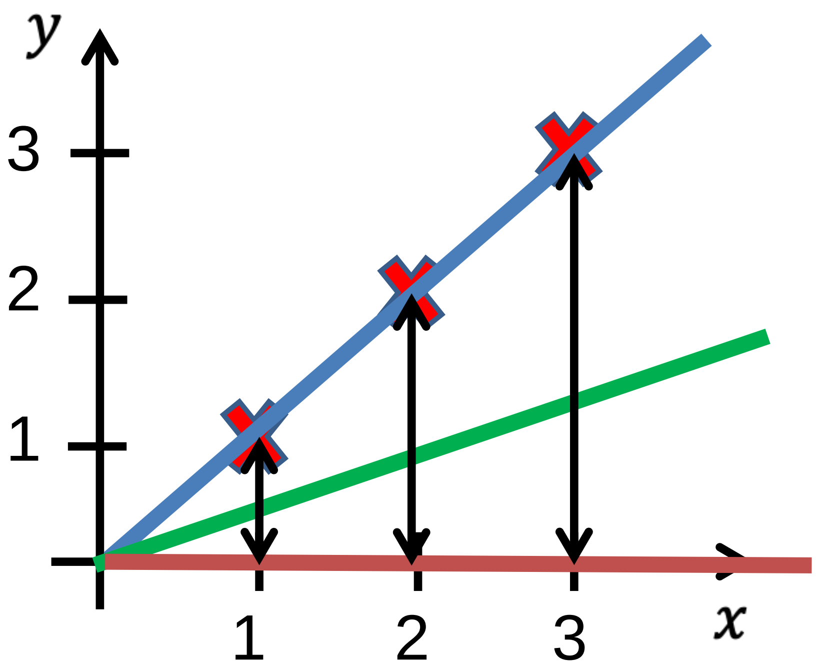
$J(\theta_1)$, function of θ_1



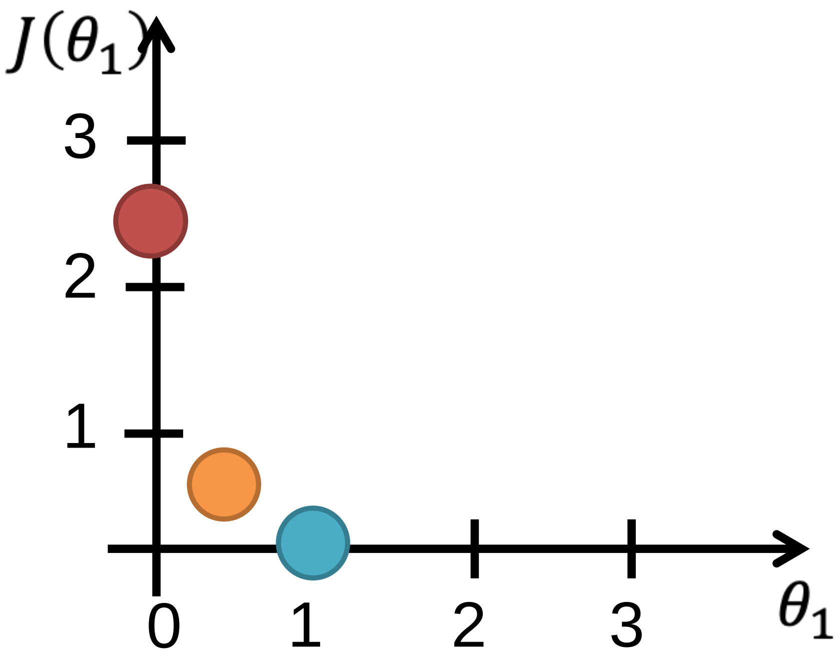
Acknowledgement: Andrew Ng, Stanford

Continued…

$h_\theta(x)$, function of x



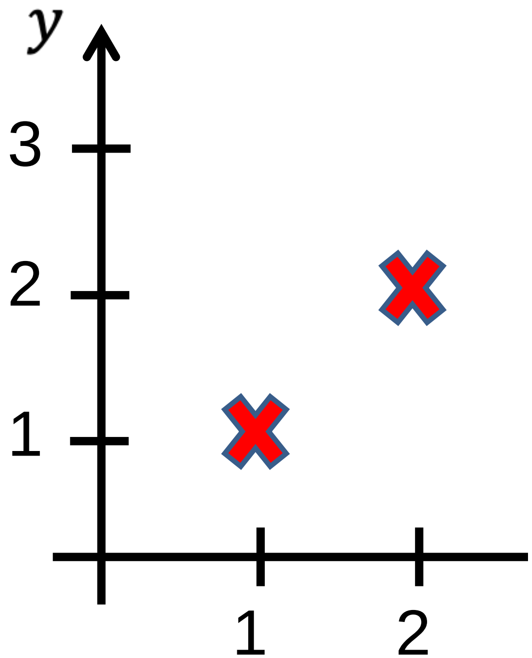
$J(\theta_1)$, function of θ_1

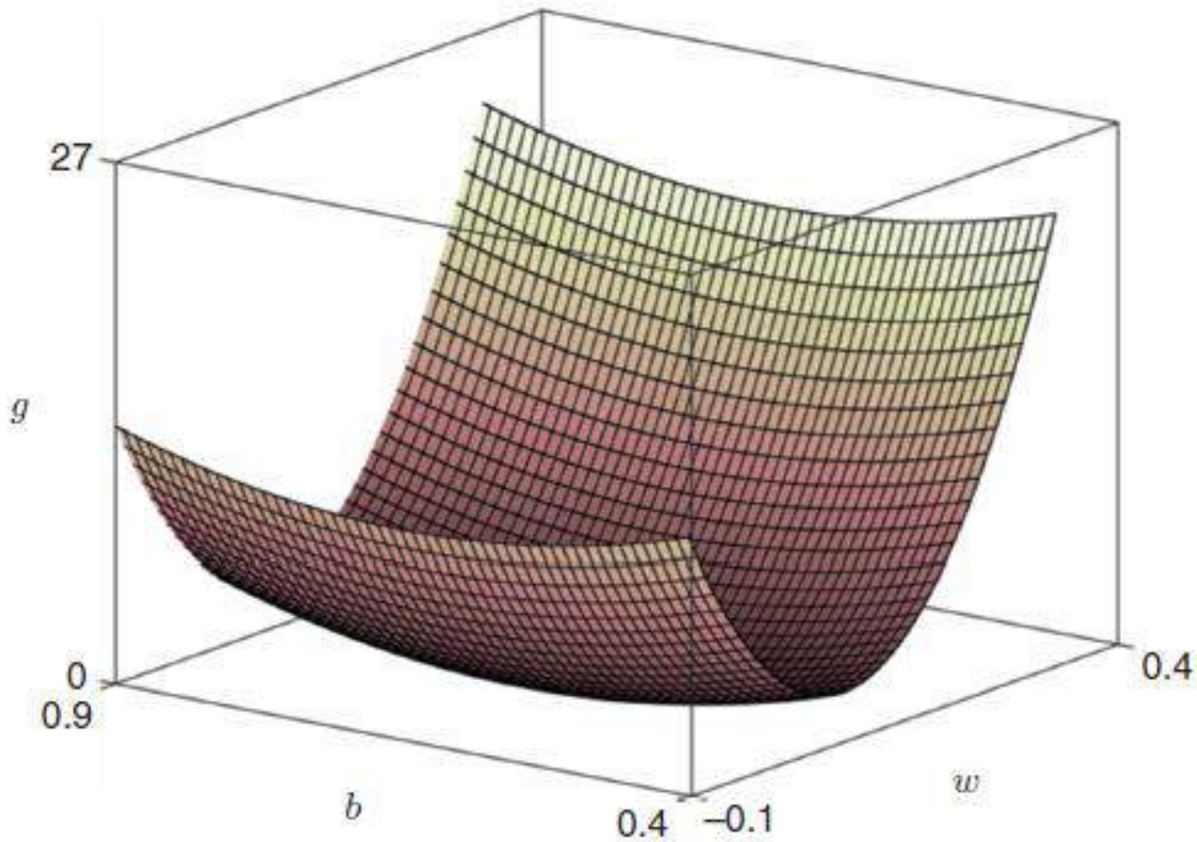


Acknowledgement: Andrew Ng, Stanford

Continued...

$h_\theta(x)$, functi





MSE cost function for linear regression is always Convex.

Gradient Descent: Minimizing the MSE

- Optimization algorithm used to minimize the MSE function by iteratively adjusting parameters in the direction of the negative gradient, aiming to find the optimal set of parameters.

Img. Source: https://www.analyticsvidhya.com/

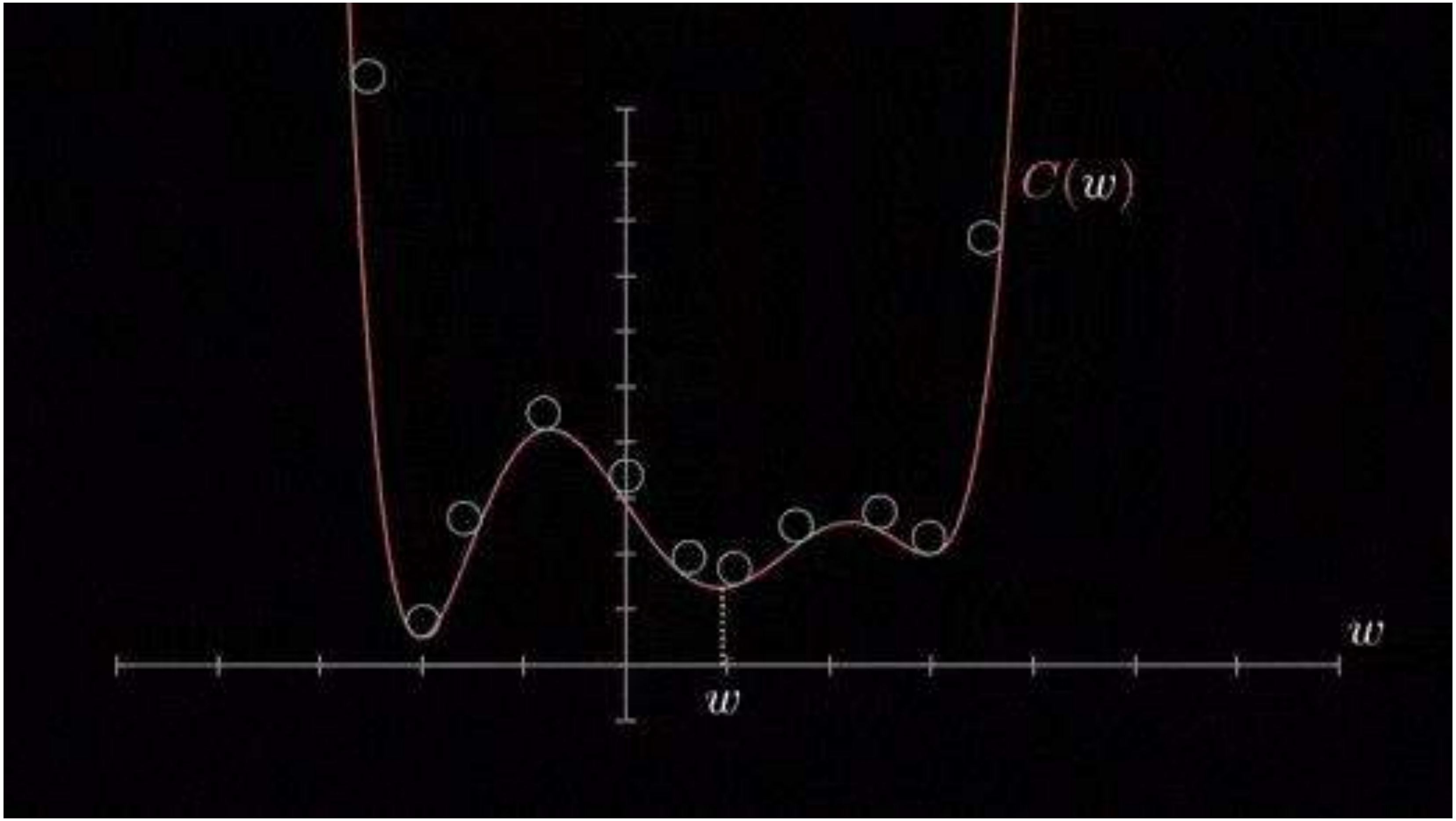
If we represent the gradient of the loss function as ∇L, and the parameters we are optimizing as θ:

Then the update rule for gradient descent is:

$$\theta\_new = \theta\_old - \alpha * \nabla L$$

Move in the opposite direction of the gradient.

Many local minima in gradient descent



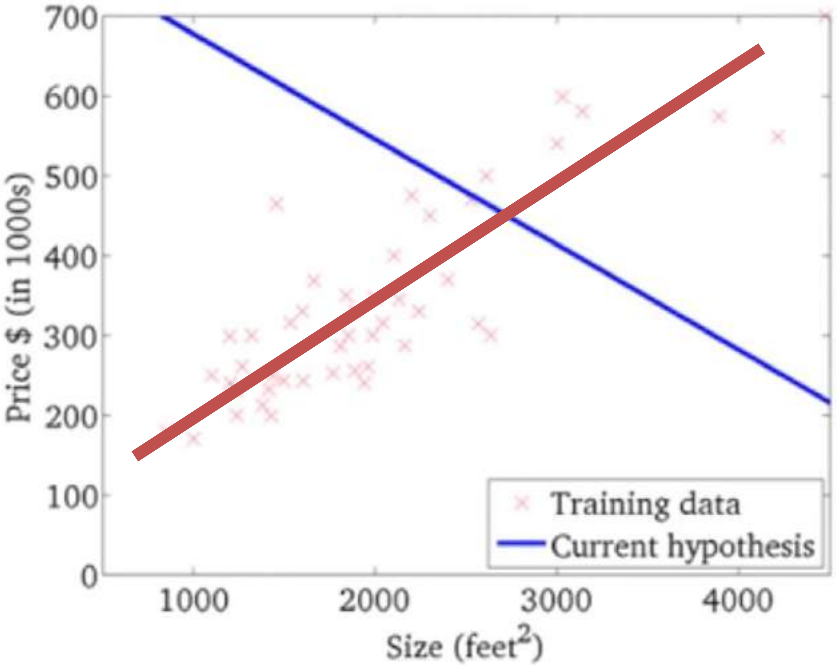
MSE cost function is Convex. Will you get many local minima? No, only one global minima.

Reason: If you pick any two points on the curve, the line joining them will never cross the curve.

Visualizing Gradient Descent

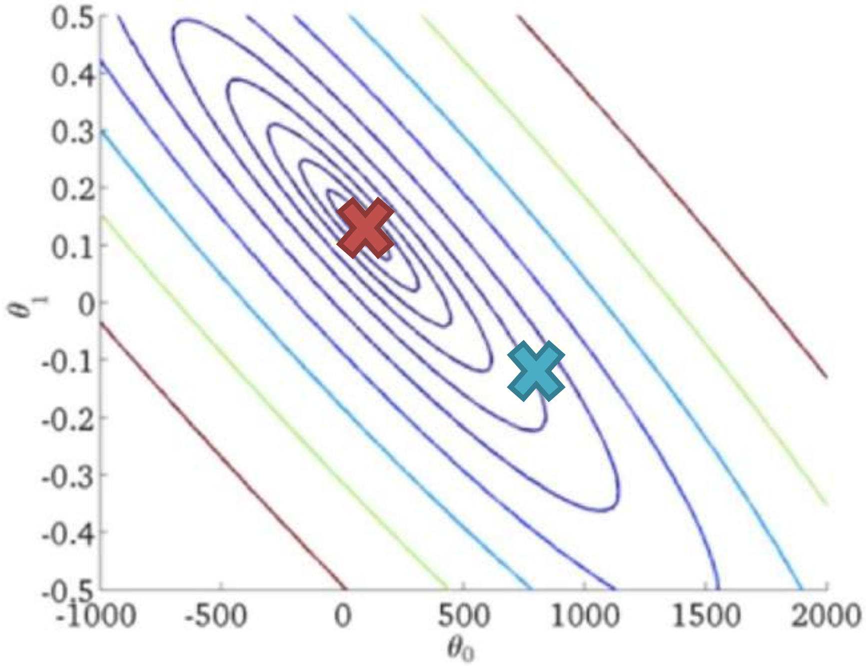
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

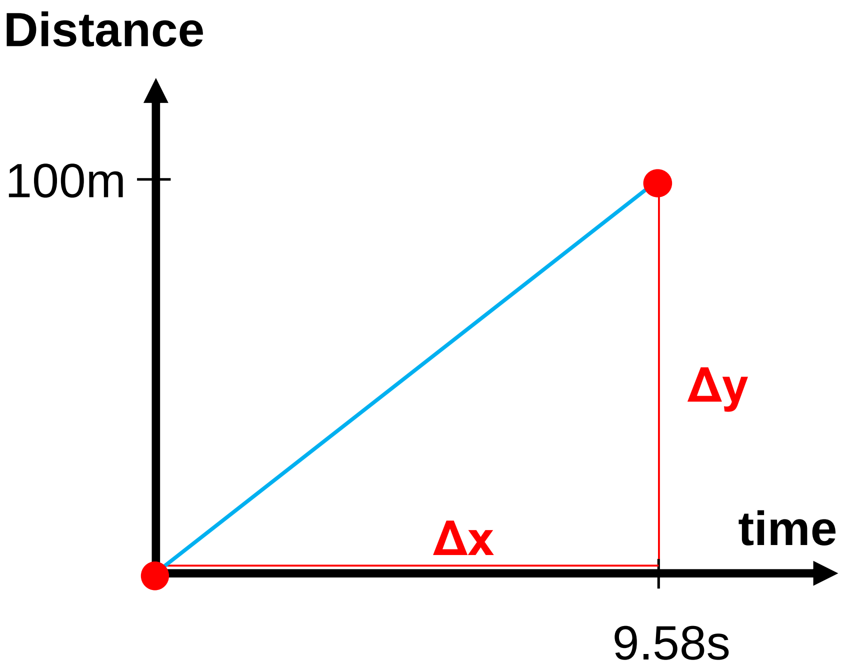
(function of the parameters θ_0, θ_1)



(visualized by using Contours)

Acknowledgement: Andrew Ng, Stanford

A bit of Math: Derivative of a Function?



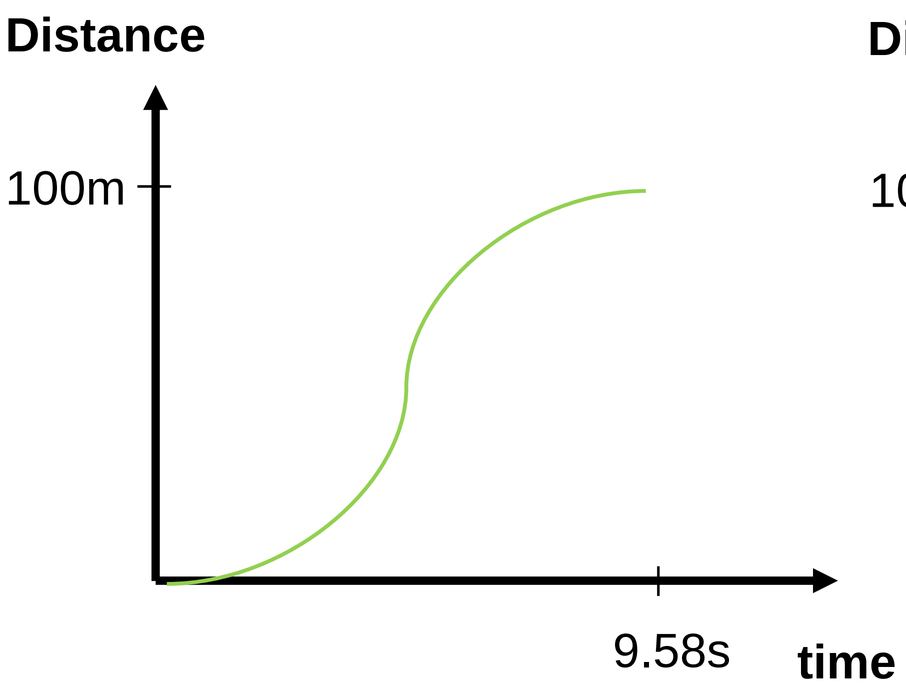
World's fastest man on the earth?

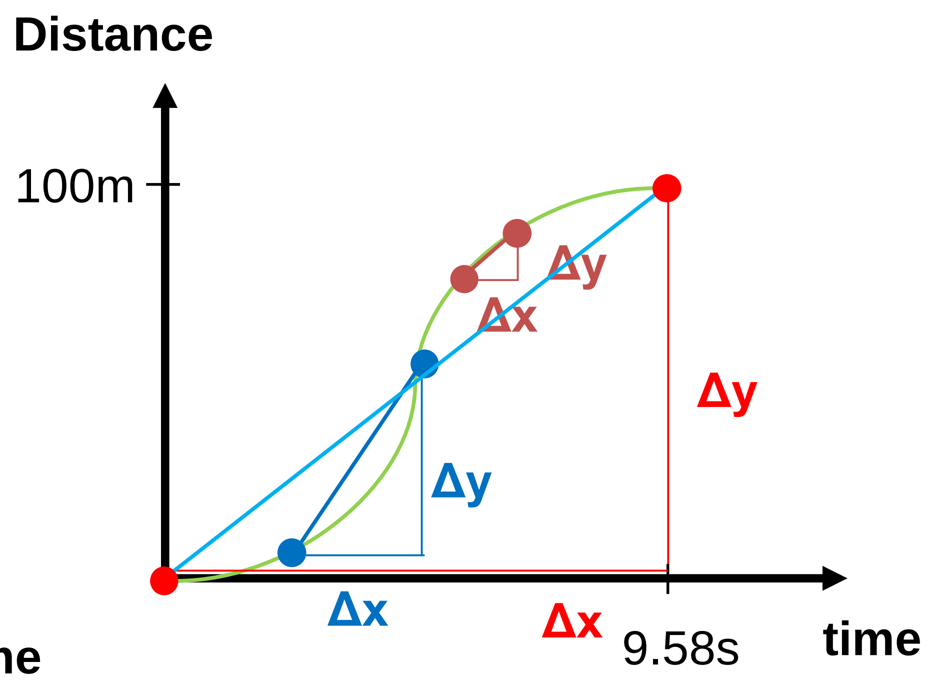
What is his Average Speed?

= Change in Distance/Change in Time
= $\Delta y/\Delta x$
= 100/9.58
= 10.43m/s

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Instantaneous Speed Vs Average Speed



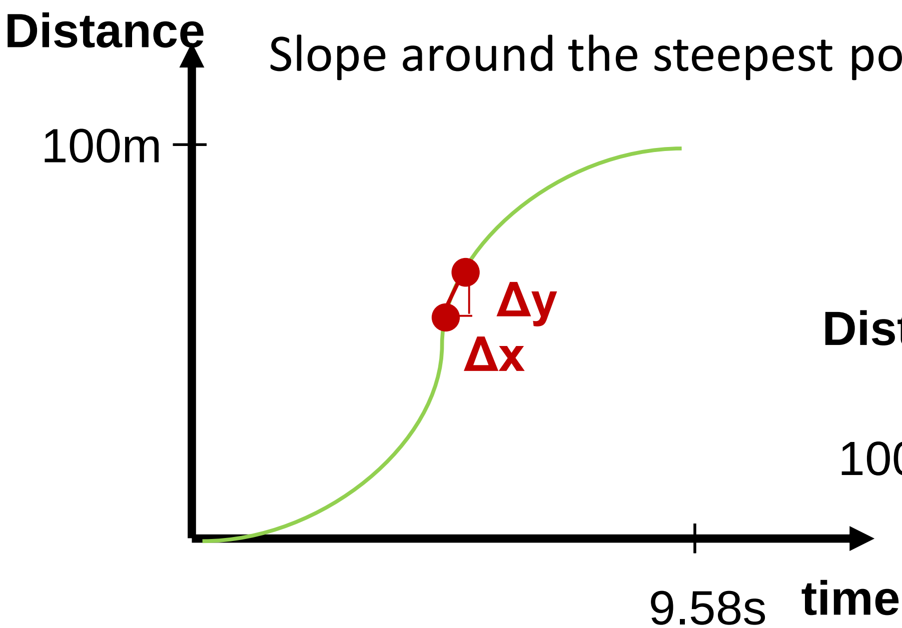


Will the Δy/Δx or Δy/Δx be different than the average slope, i.e., Δy/Δx? ✔

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

What would be really the Instantaneous speed?

Slope around the steepest point.



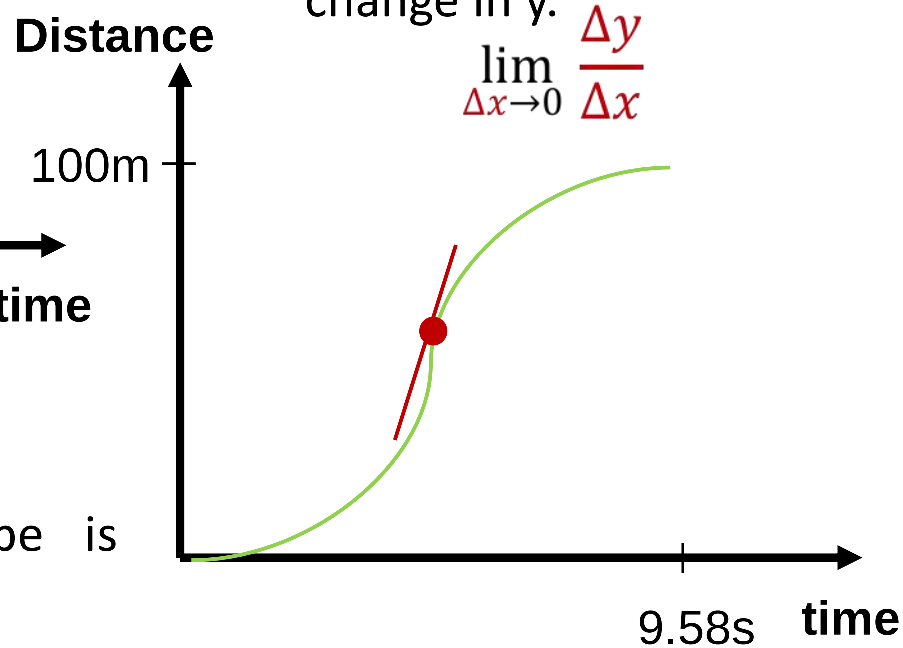
Fastest Instantaneous speed?

An Approximation: As the slope is changing constantly.

Better approximation:

Measure the slope with a smaller and smaller change in x that yields a smaller and smaller change in y.

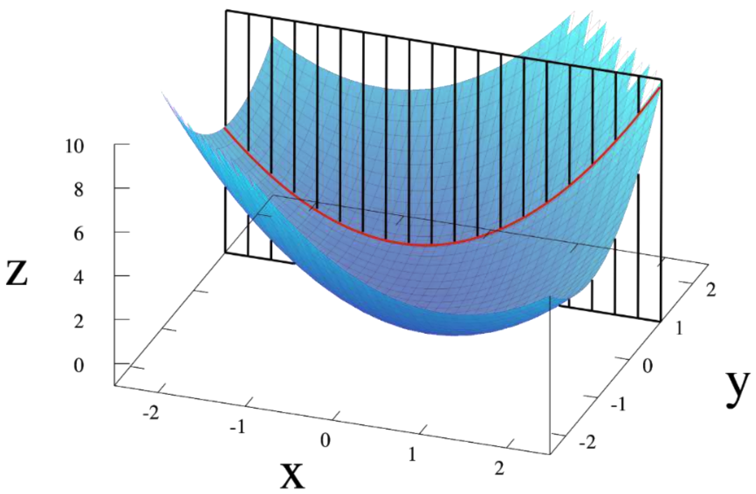
$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$



Instantaneous Slope is called **Derivative**: $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ = dy/dx

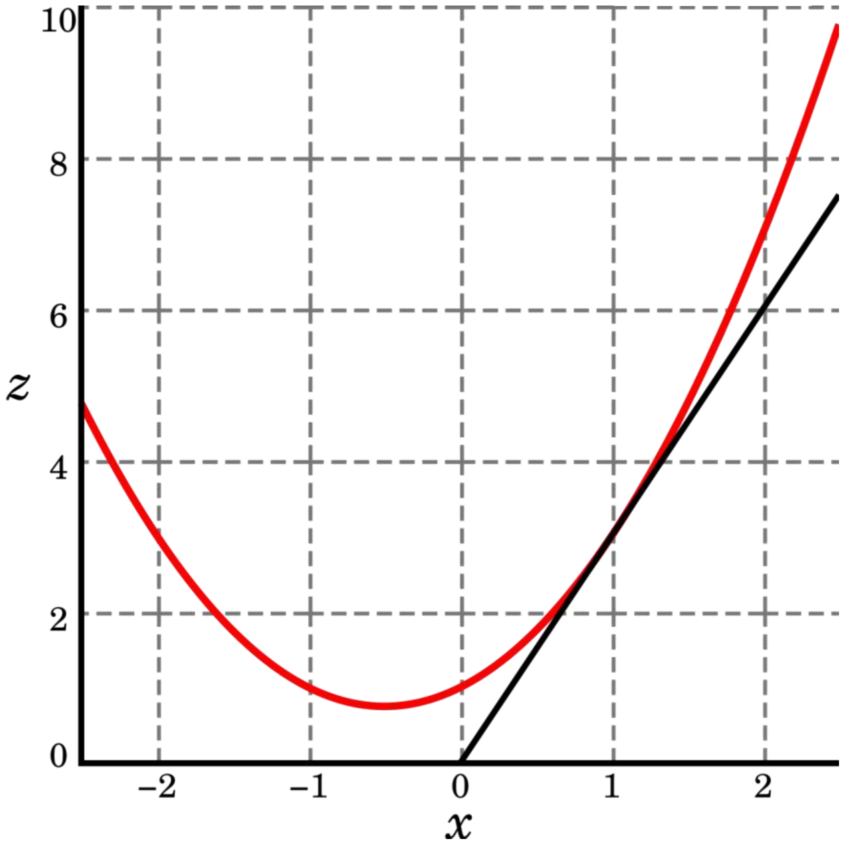
What is Partial Derivative?

What is the partial derivative of this function at P(1,1)? $\frac{\partial z}{\partial x} = 3$



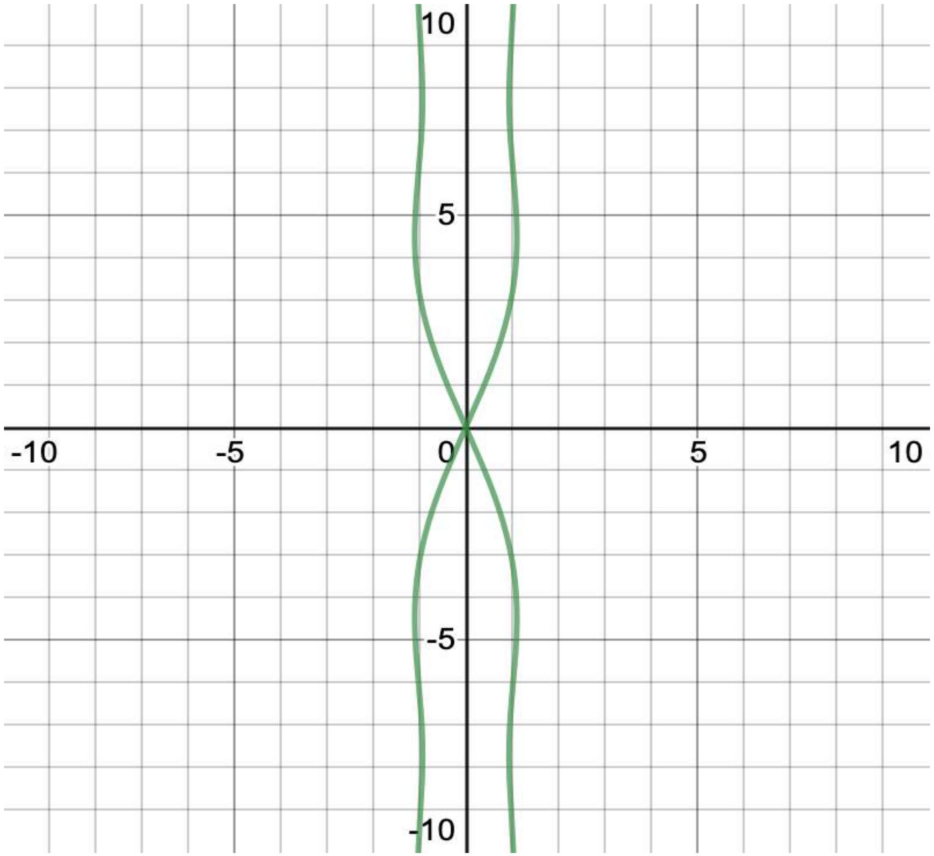
$$z = f(x, y) = x^2 + xy + y^2$$

That is the slope of *f* at the point (*x*, *y*)



(Img. Source: Wiki)

Gradient: All partial derivatives together



Multivariate Function: $f(x, y) = x^2 y + \sin(y)$

$$\frac{\partial f}{\partial x} = 2xy$$

$$\frac{\partial f}{\partial y} = x^2 + cos(y)$$

Gradient

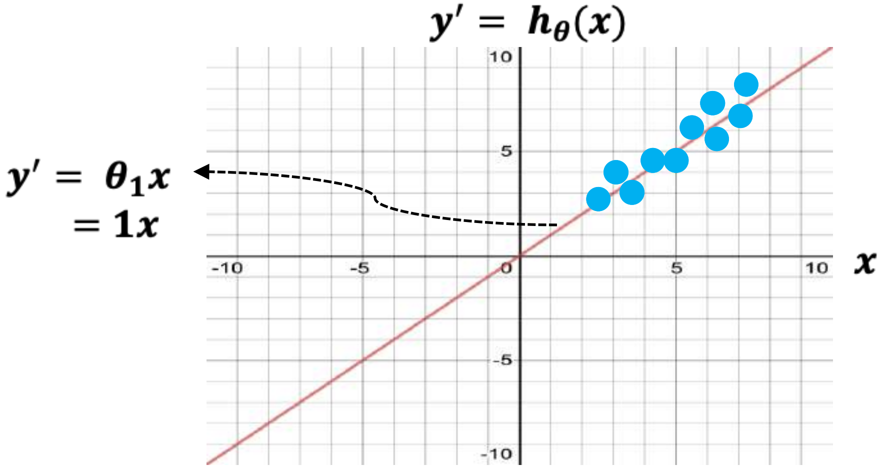
$$\nabla f(x, y) = \nabla\, x^2 y + \sin(y)$$

$$= \begin{bmatrix} 2xy \\ x^2 + \cos(y) \end{bmatrix}$$

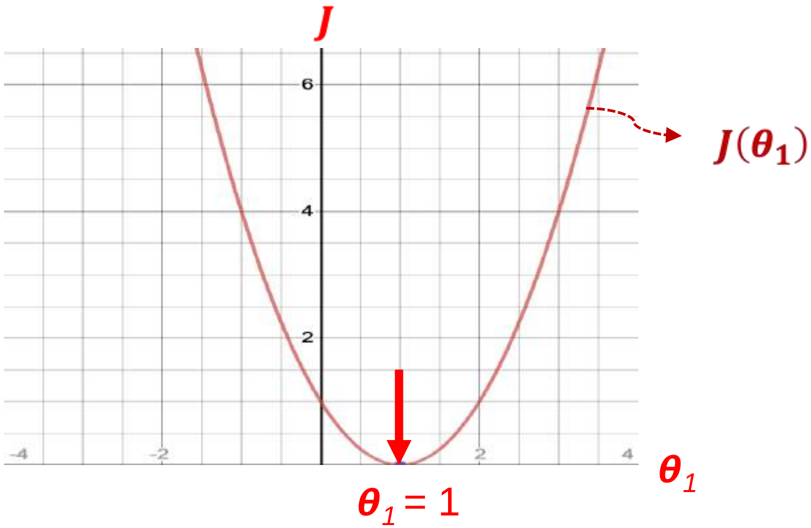
Acknowledgement: Mohammad Hammoud, CMU (Qatar)

The Impact of Partial Derviative

- For simplicity, let us assume our optimization objective is to $\underset{\theta_0,\theta_1}{\text{minimize}}\, J(\theta_1)$, thus, $\theta_0 = 0$



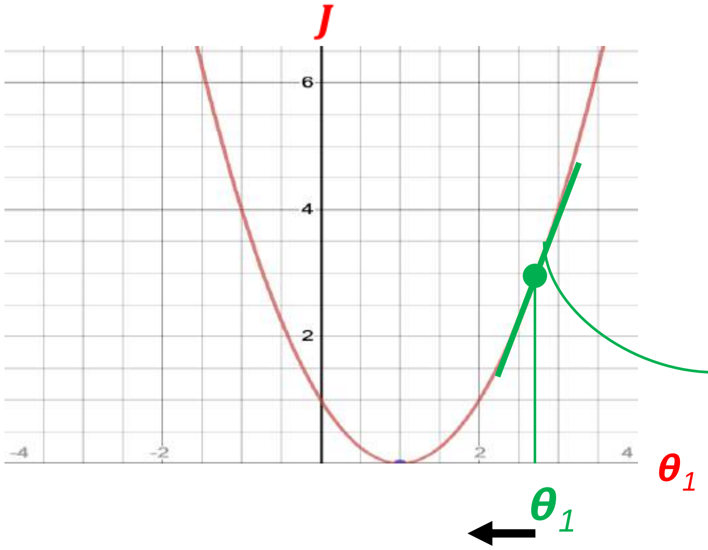
$h_\theta(x)$ is the ***Hypothesis Function***



$J(\theta_1)$ is the ***Cost Function***

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued…



$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

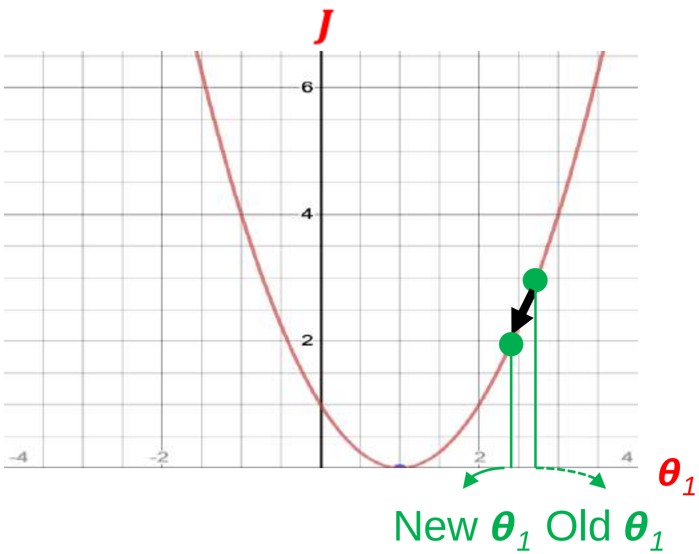
$$= \theta_1 - \alpha\, (Positive\ Number)$$

Decrease θ_1 by a certain value

Positive Derivative

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued...



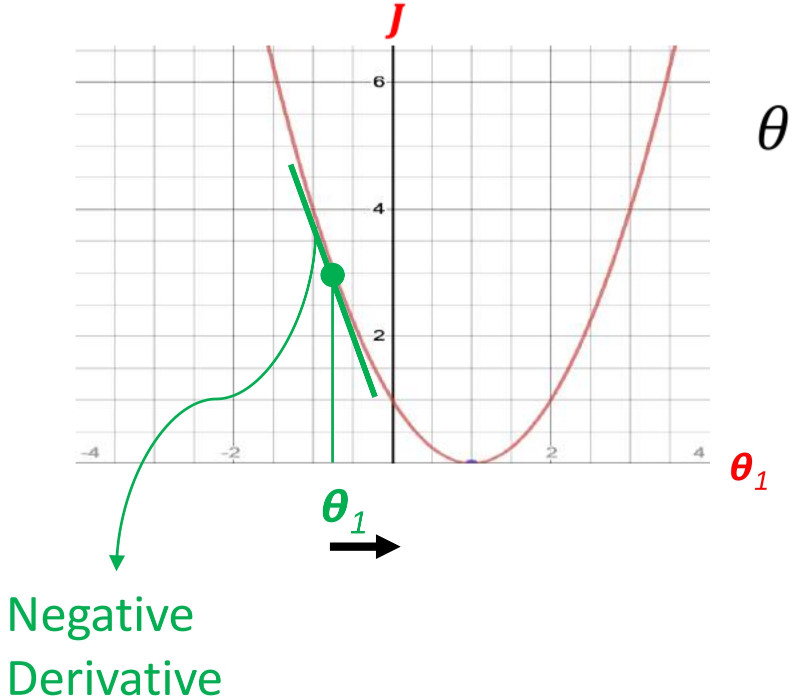
$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

$$= \theta_1 - \alpha\, (Positive\ Number)$$

Decrease θ_1 by a certain value

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued…



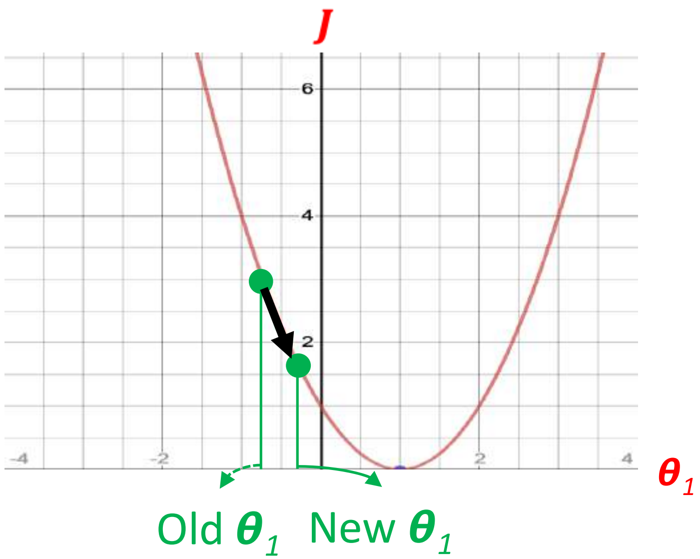
$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

$$= \theta_1 - \alpha\, (Negative\ Number)$$

Increase θ_1 by a certain value

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued…



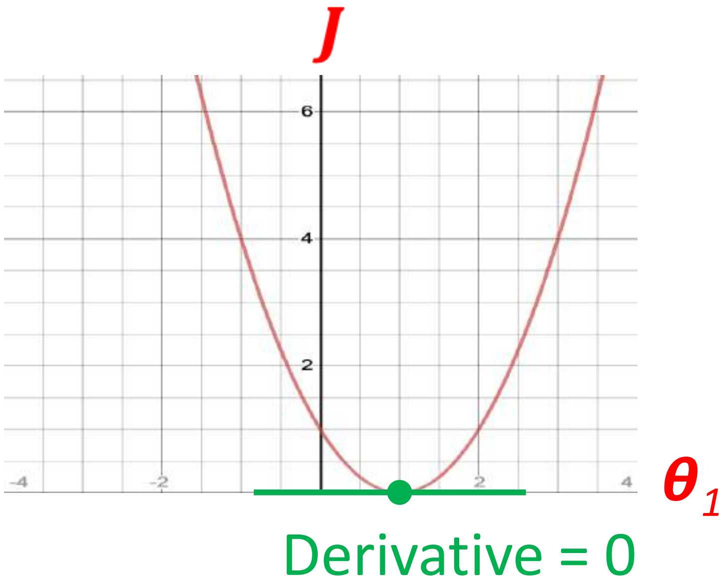
$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

$$= \theta_1 - \alpha\, (Negative\ Number)$$

Increase θ_1 by a certain value

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued…



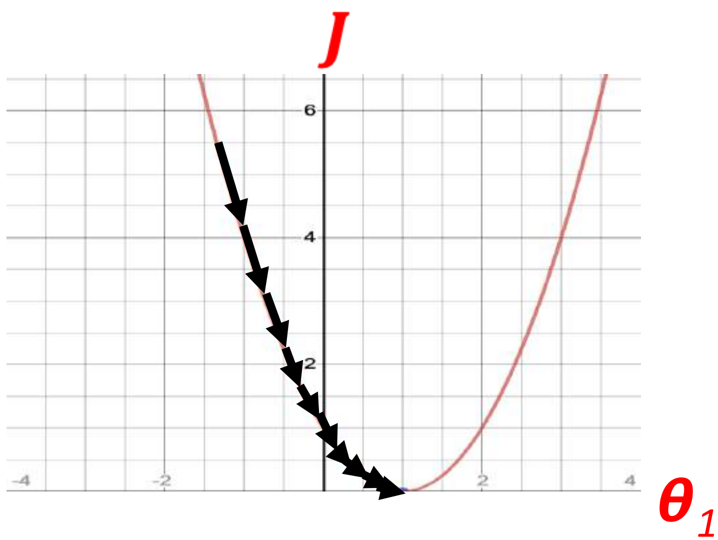
$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

$$= \theta_1 - \alpha\, (\text{Zero})$$

θ_1 remains the same, and hence, gradient descent has converged.

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

The Impact of Learning Rate



Too Small

Learing Rate

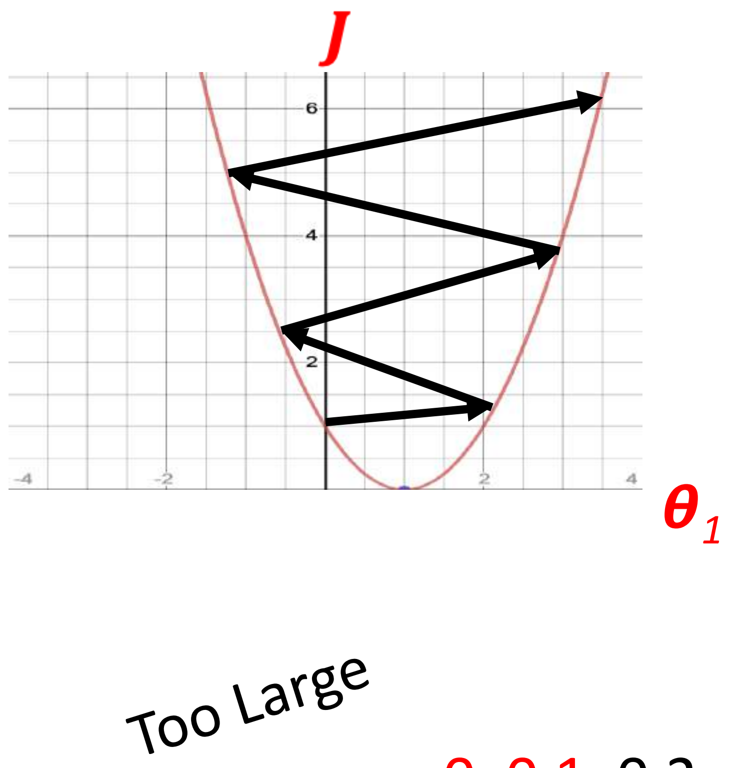
$$\theta_1 = \theta_1 - \alpha \frac{d\,J(\theta_1)}{d\,\theta_j}$$

$$= \theta_1 - (Too\ Small\ Number)\ \frac{d\,J(\theta_1)}{d\,\theta_j}$$

θ_1 changes only a tiny bit on each step, hence, gradient descent will render slow (will take more time to converge)

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Continued...



$$\theta_1 = \theta_1 - \alpha \frac{d\, J(\theta_1)}{d\, \theta_j}$$

$$= \theta_1 - (Too\ Large\ Number) \frac{d\, J(\theta_1)}{d\, \theta_j}$$

θ_1 changes a lot (and probably faster) on each step, hence, gradient descent will potentially overshoot the minimum and, accordingly, fail to converge (or even diverge)

0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, ..., 0.9, 1

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

Gradient Descent for Linear Regression

Linear regression model:

$h_\theta(x) = \theta_0 + \theta_1 x$

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

$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^{m}$

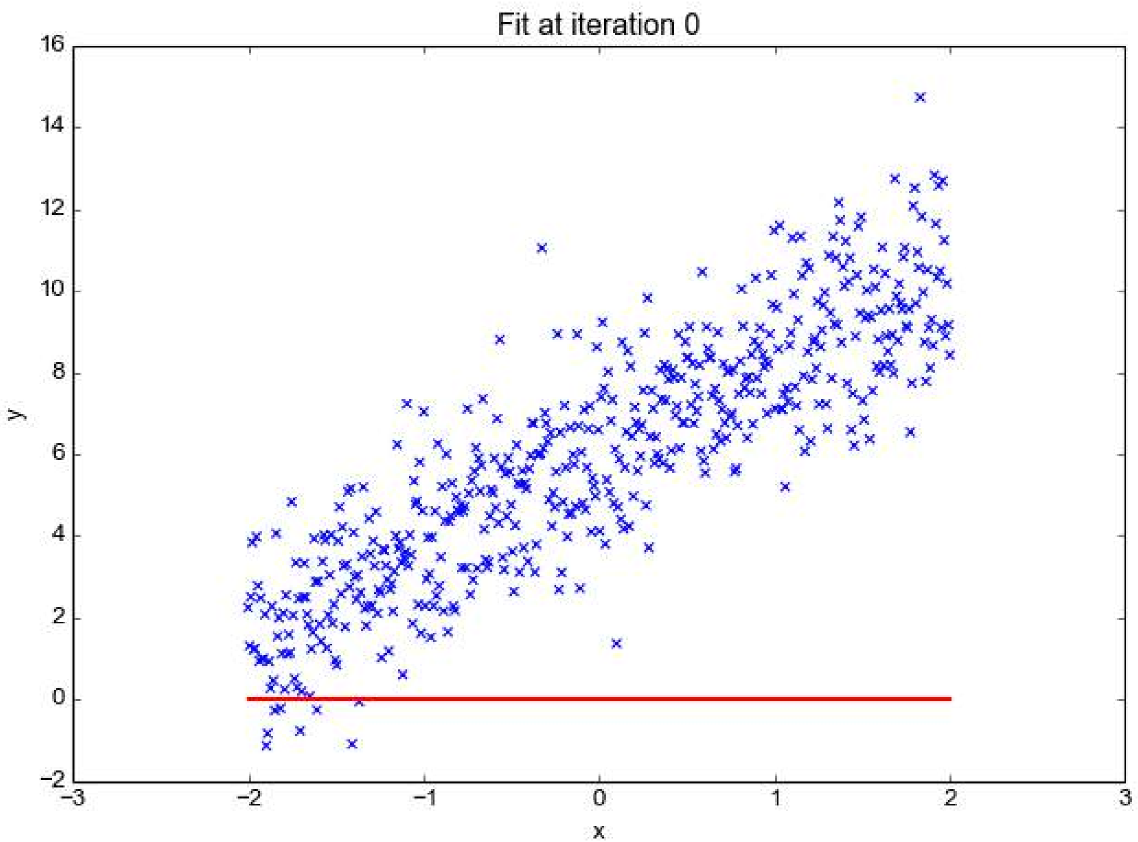
$= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=}^{m}$

Repeat until convergence

$j = 0: \quad \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m}$

$j = 1: \quad \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m}$

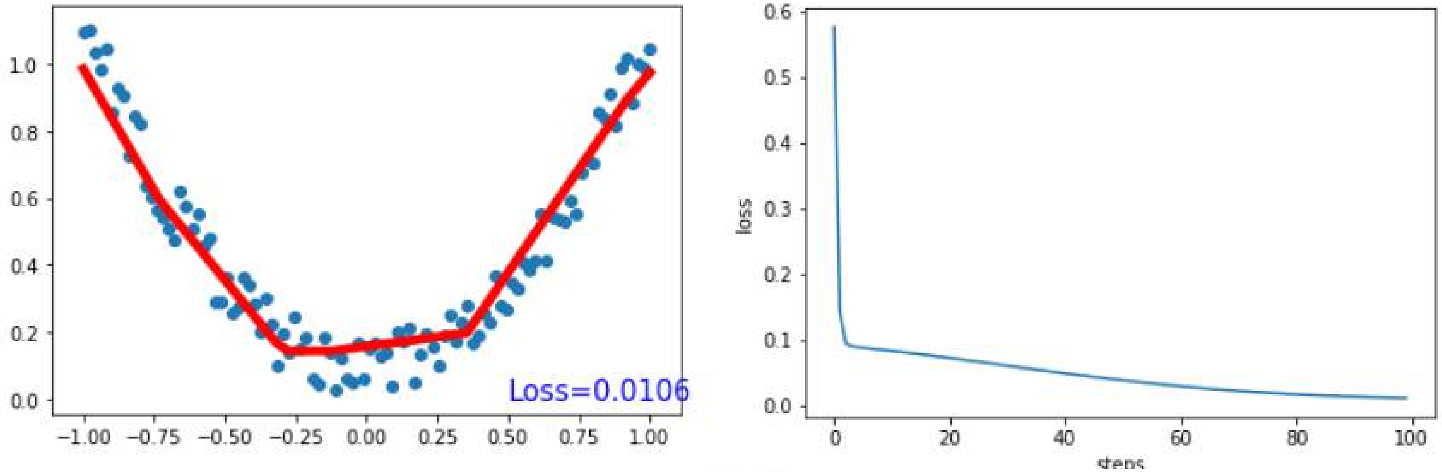
}



Update θ_0 and θ_1 simultaneously

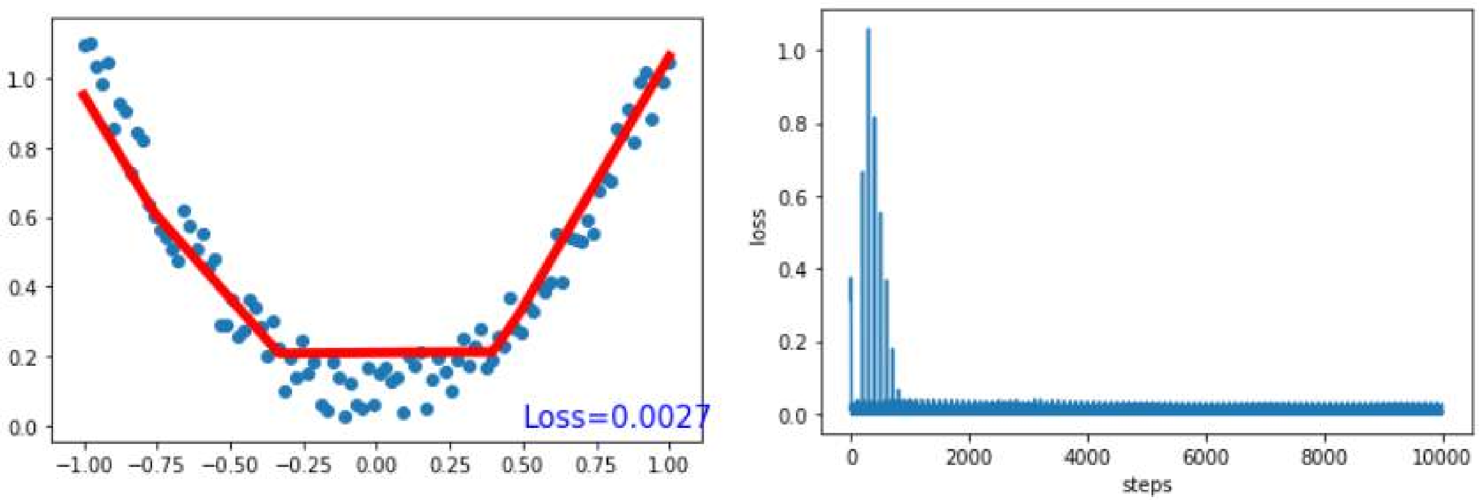
Batch Vs Stochastic Gradient Descent

GD



Very smooth convergence, however using all the data for one update.

SGD

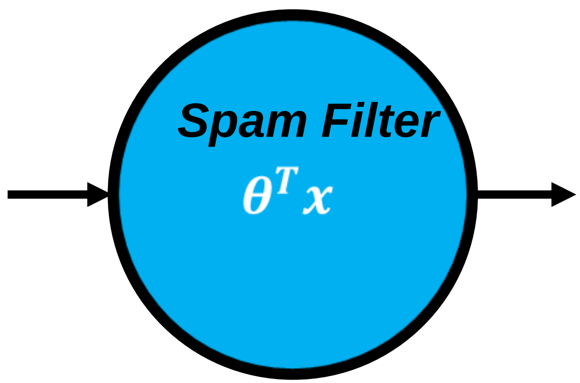


Very noisy convergence, because using only one data point for one update.

Regression vs. Classification

- What are the possible outputs of the linear regression function $\boldsymbol{h_\theta(x) = \theta^T x}$?
  - Real-valued outputs

$\boldsymbol{x = [1, 0, 0, 1, 0]}$



$\boldsymbol{h_\theta(x) = 3.2 + 6.7 = 9.9}$

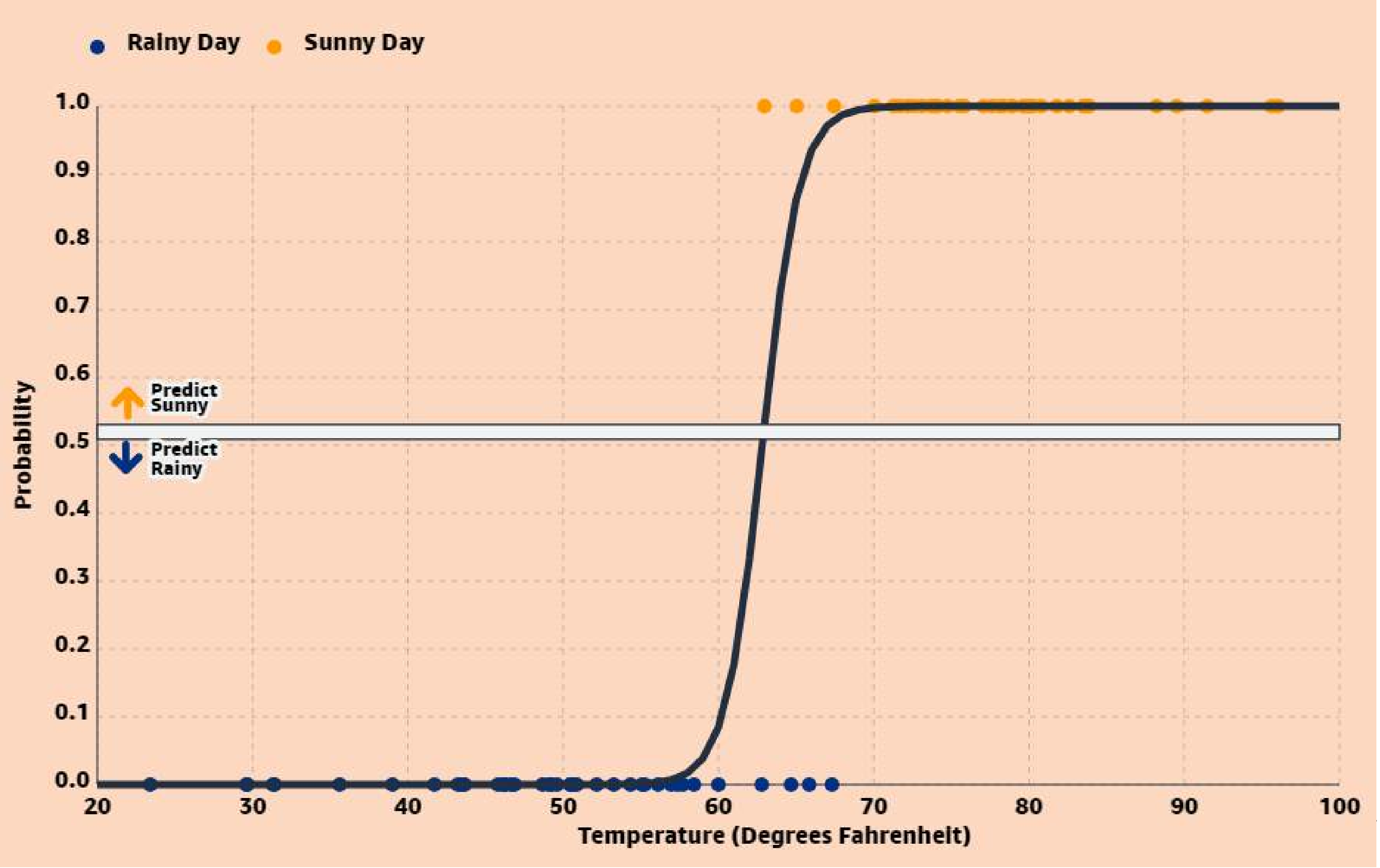
$\in \mathbb{R}$, which makes it a ***regression problem***

A spam or not a spam?

We need a *discrete-valued* output (e.g., 0 or 1)

Acknowledgement: Mohammad Hammoud, CMU (Qatar)

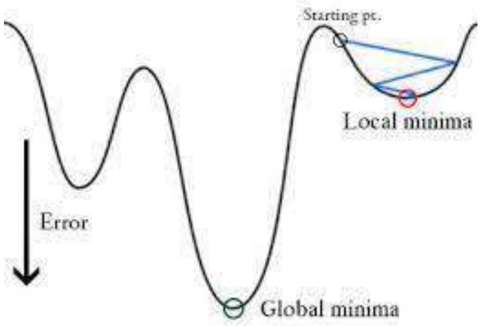
Logistic regression



Source: https://mlu-explain.github.io/logistic-regression/

Loss function for logistic regression

- If you use MSE for Logistic regression, what problems it might create?

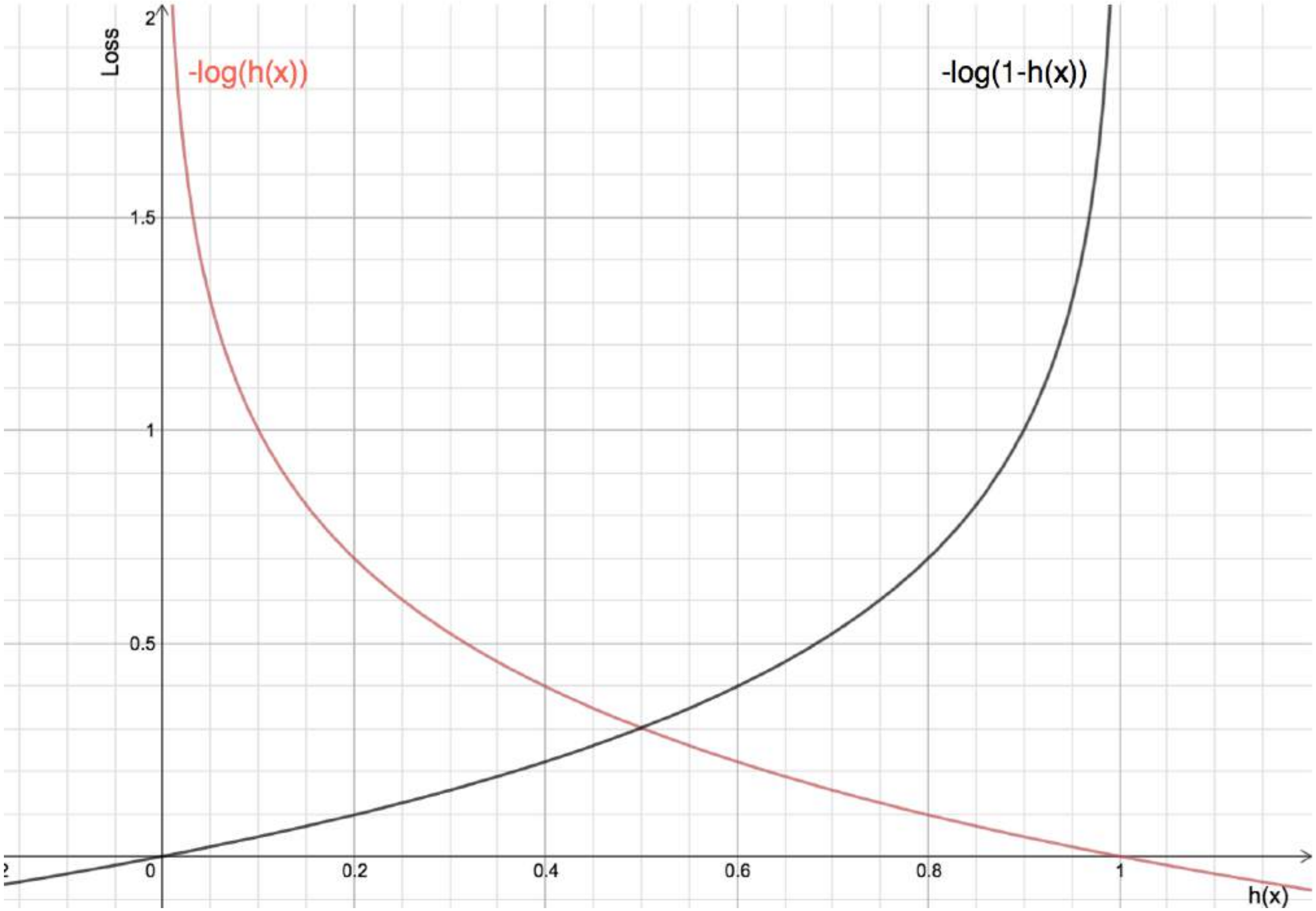


- A suitable loss function in logistic regression is called the Log-Loss, or binary cross-entropy. This function is:

$$\text{Log-Loss} = \sum_{i=0}^{n} -(y_i * \log(p_i) + (1 - y_i) * \log(1 - p_i))$$

- It penalizes deviations, offering a continuous metric for optimization during model training.

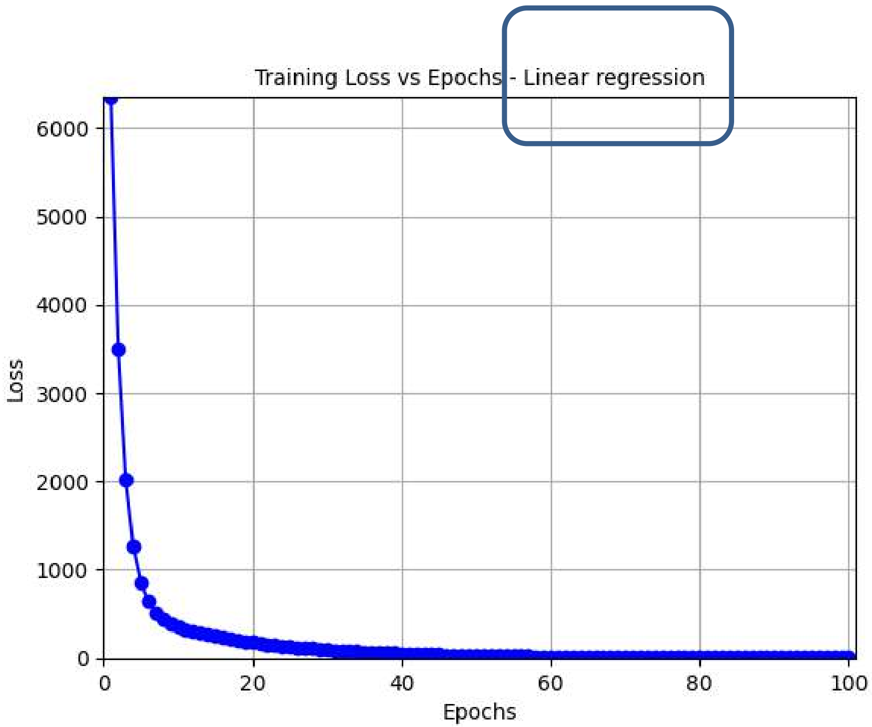
Continued...

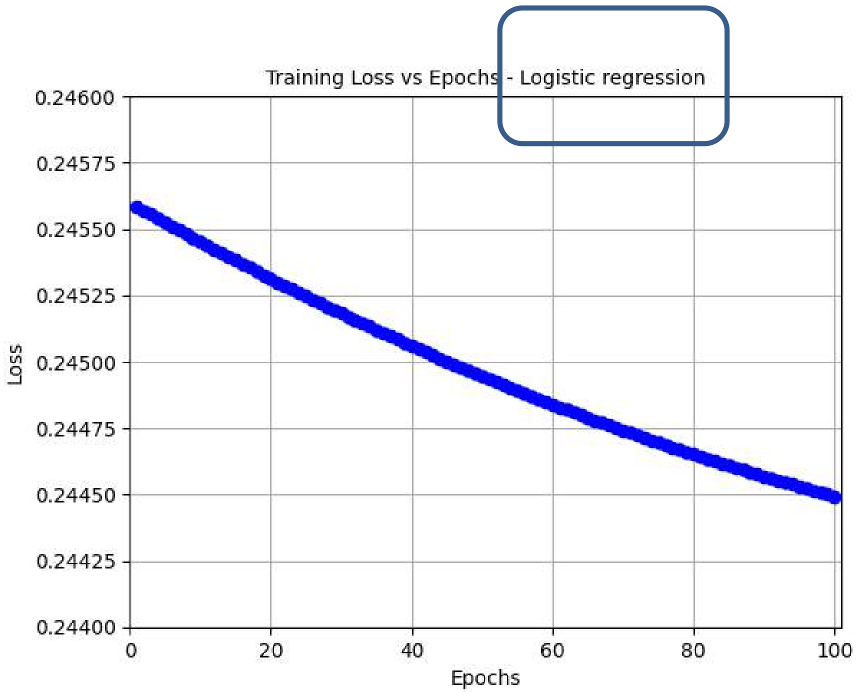


Red:
class 1

Black:
class 0

Assignment 3 (Due date: 1st March 2024)





Thank you!