Machine Learning

Linear Regression

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

Supervised Learning

- Regression
- Classification

example

Notation:

m: number of training samples

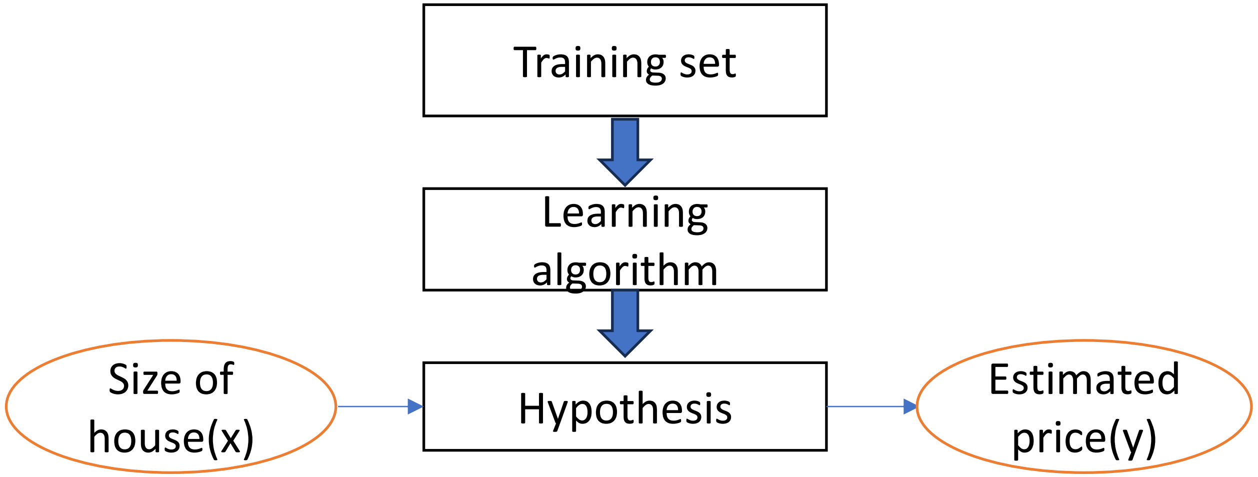
x: input variable

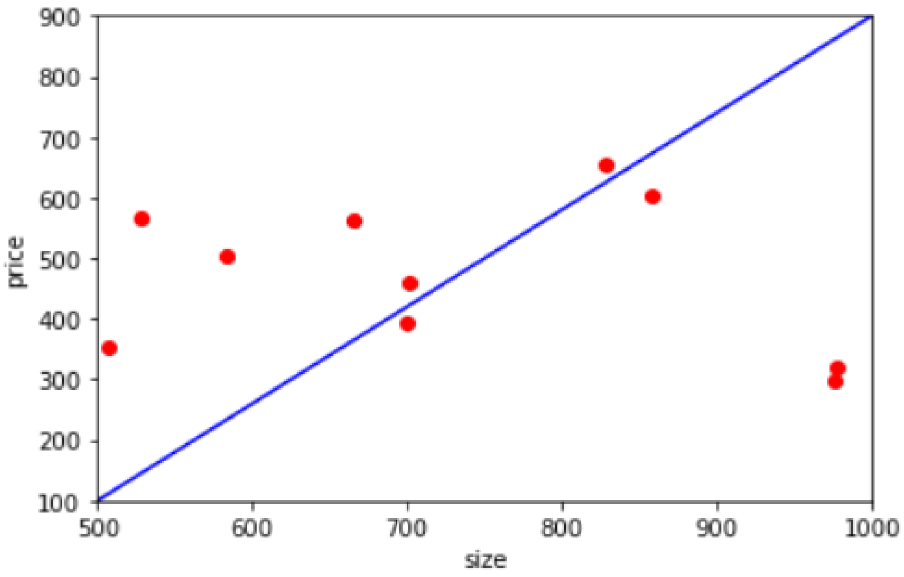
y: output variable
Or
target variable

(x_i, y_i): i th training sample

number	Size (x variable)	Price (y variable)	
1	100	500	(x_1, y_1)
2	750	2000	(x_2, y_2)
3	852	178	(x_3, y_3)
	…	…	
m	3210	870	(x_m, y_m)

example



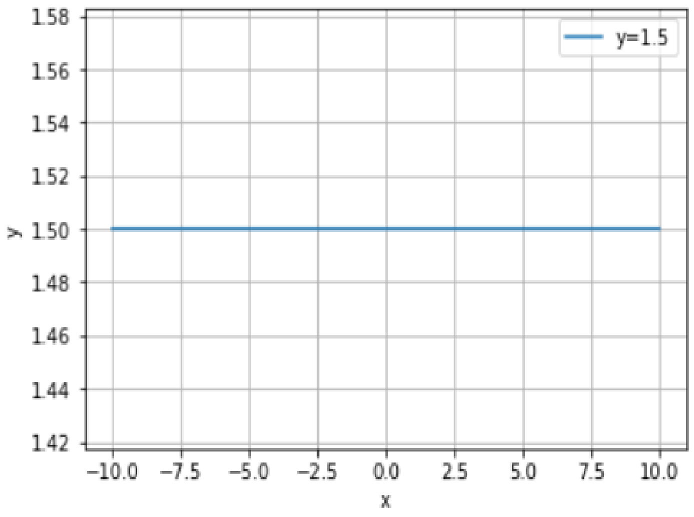


$h(x) = \theta_0 + \theta_1 x$

parameters= $\{ \theta_0 , \theta_1 \}$

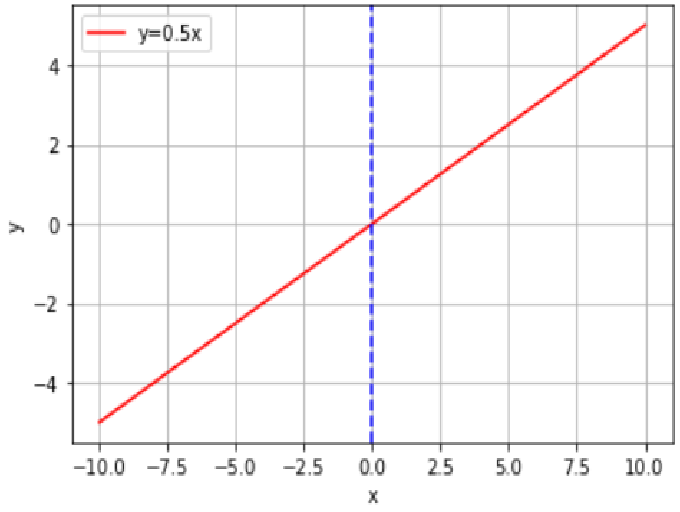
$$h(x) = \theta_0 + \theta_1 x$$

parameters



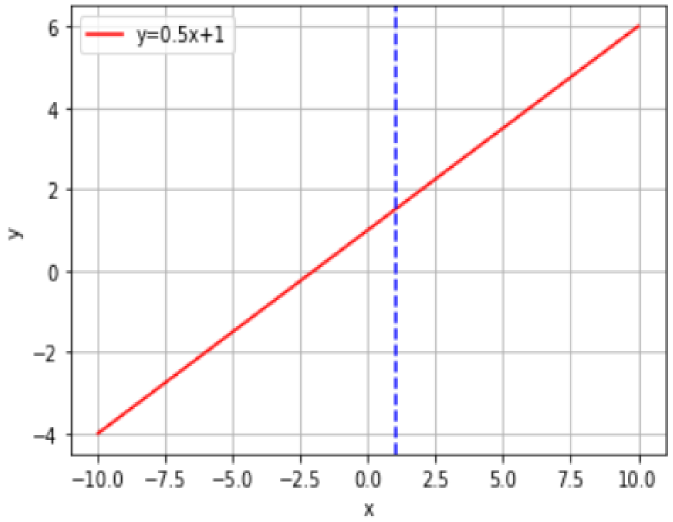
$h(x) = 1.5$

$\theta_0 = 1.5$
$\theta_1 = 0$



$h(x) = 0.5x$

$\theta_0 = 0$
$\theta_1 = 0.5$



$h(x)=0.5x+1$

$\theta_0 = 1$
$\theta_1 = 0.5$

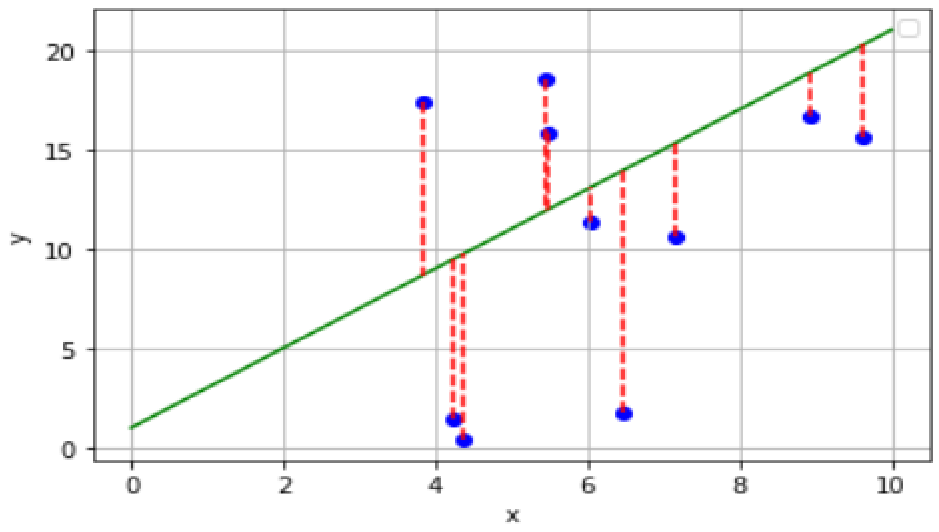
Cost Function

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

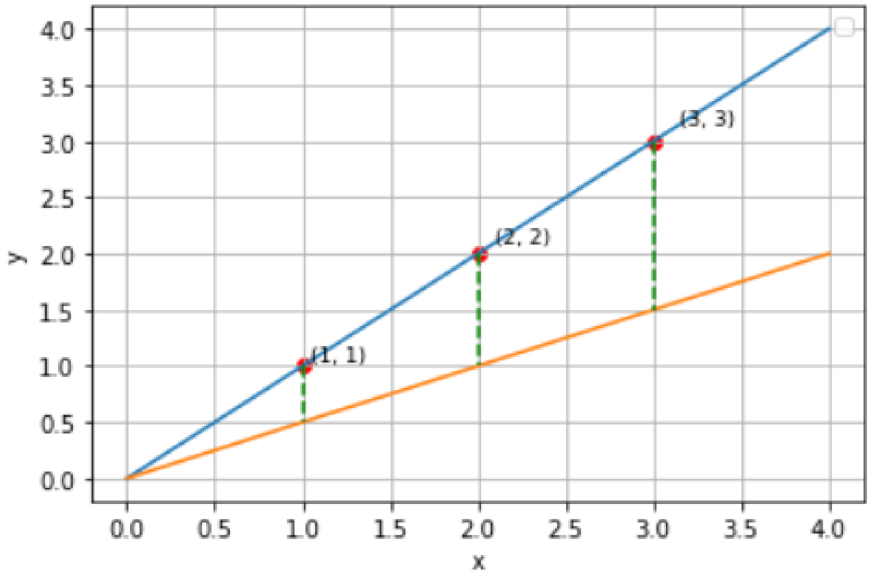
Mean square error(MSE)

Minimize $\boldsymbol{J(\theta_0 , \theta_1)}$

$\boldsymbol{\theta_0 , \theta_1}$



example



$$J(\theta_0{=}0\ ,\ \theta_1{=}0.5\ ) = \frac{1}{2m}\sum_{i=1}^{m}(0.5x_i - y_i)^2$$

$$= \frac{1}{2*3}\ [(0.5-1)^2 + (1-2)^2 + (1.5-3)^2]$$

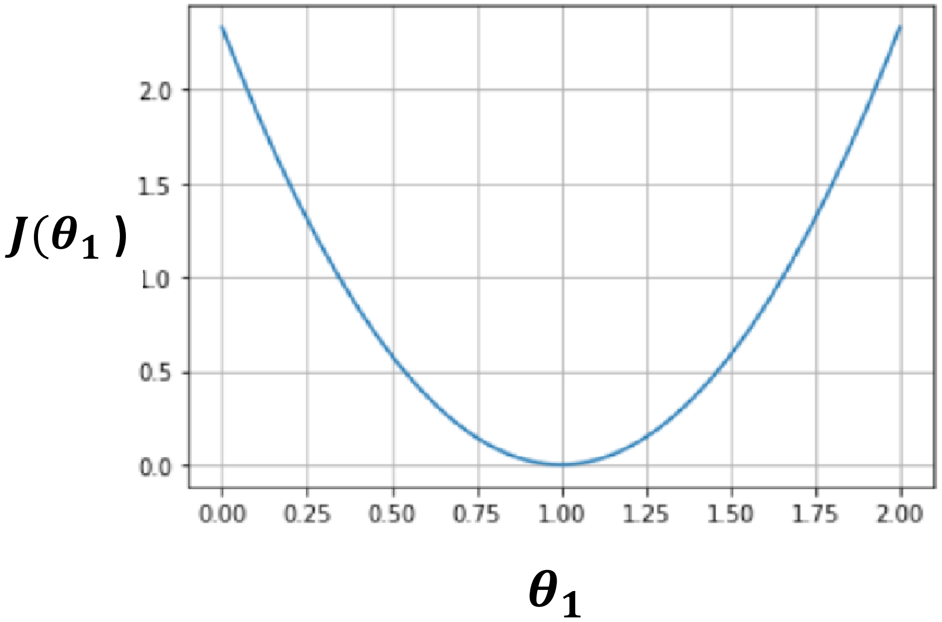
$$= \frac{1}{6}\,(3.5) = 0.58$$

$$J(\theta_0{=}0\ ,\ \theta_1{=}1\ ) = \frac{1}{2m}\sum_{i=1}^{m}(x_i - y_i)^2$$

$$= \frac{1}{2*3}\ [(1-1)^2 + (2-2)^2 + (3-3)^2]$$

$$= \frac{1}{6}\,(0) = 0$$

example



θ_1	$J(\theta_1)$
0	14/6
0.5	0.58
1	0
1.5	0.58
2	14/6

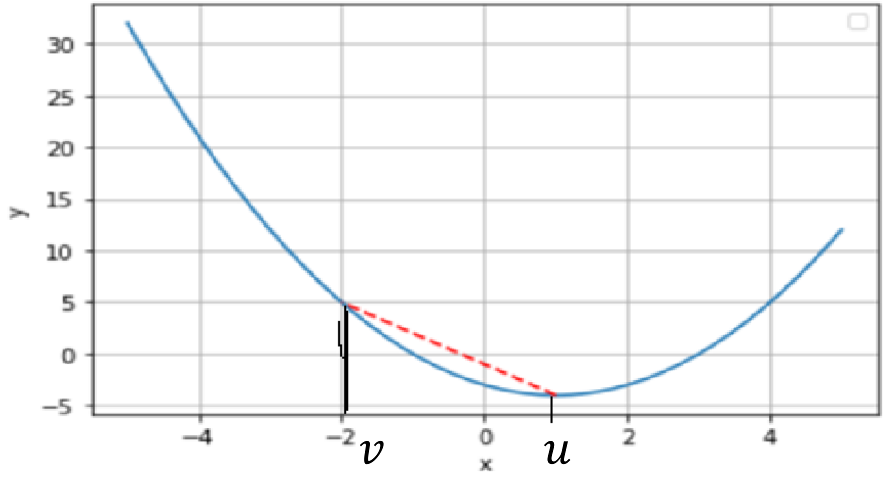
- Plotting the cost for each value of θ_1
- The minimum point: θ_1=1
- Using Grid Search to find best values of parameters

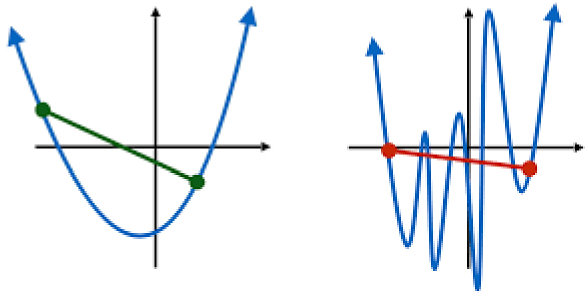
Cost Function

- $J(\theta_0 , \theta_1) = \frac{1}{m} \sum_{i=1}^{m} |h(x_i) - y_i|$ Mean absolute error(MAE)

- Better for outliers compared with MSE

Convexity



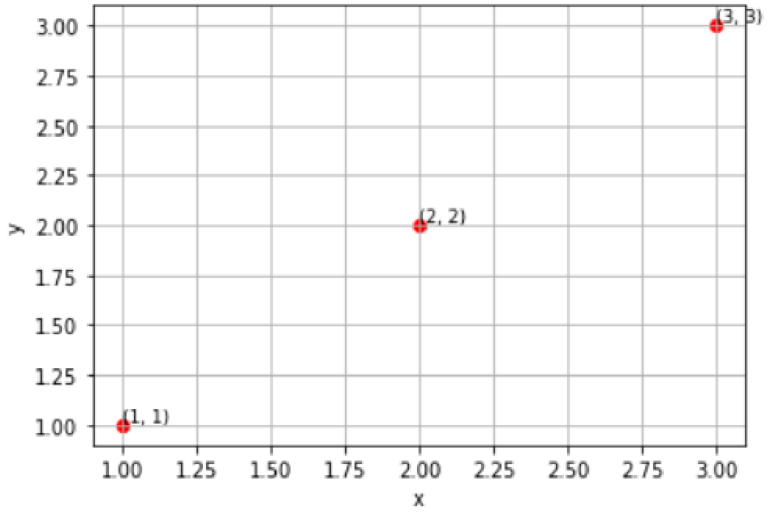


Function h(u) with u$\in X$ is convex if for any $u, v \in X$ and for any $0 \leq \lambda \leq 1$ we have:

$$\mathbf{h(\lambda u + (1-\lambda)v) \leq \lambda\, h(u) + (1-\lambda)\, h(v)}$$

برای توابع محدب هر بهینه محلی یک بهینه سراسری است.

example



$if\ \theta_1 = -1$:

MAE= $\frac{1}{3}$ [|1 - (-1)| + |2 - (-2)| +|3 - (-3)|] = 4

$if\ \theta_1 = 0$:

MAE= $\frac{1}{3}$ [|1 - 0| + |2 - 0| +|3 - 0|] = 2

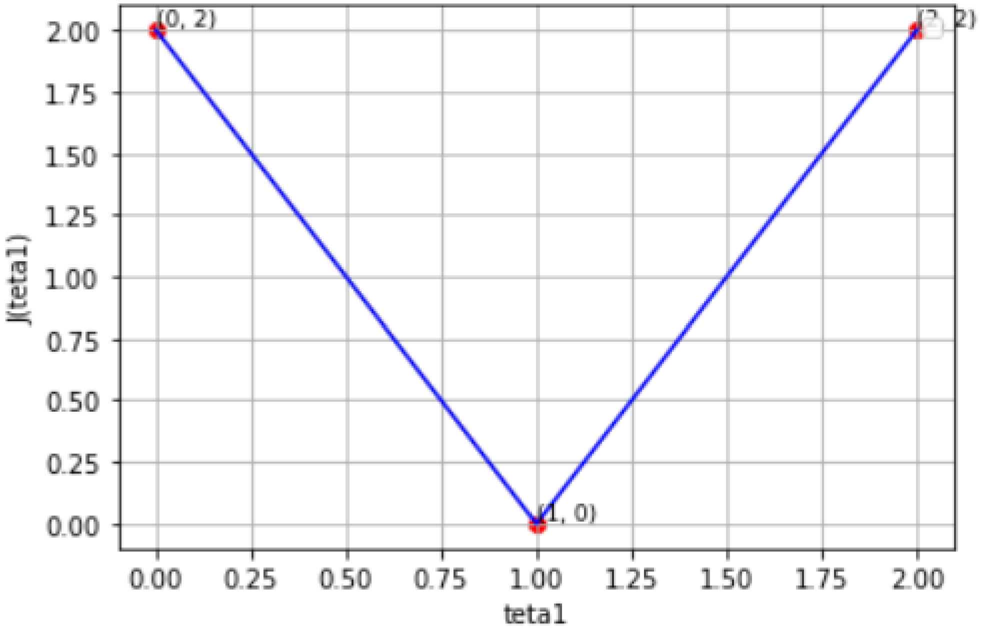
$if\ \theta_1 = 1$:

MAE= $\frac{1}{3}$ [|1 - 1| + |2 - 2| +|3 - 3|] = 0

$if\ \theta_1 = 2$:

MAE= $\frac{1}{3}$ [|1 - 2| + |2 - 4| +|3 - 6|] = 2

example



MAE is convex

$\boldsymbol{\theta_1}$	$\boldsymbol{J(\theta_1)}$
-1	4
-0.5	3
0	2
0.5	1
1	0
1.5	1
2	2
2.5	3
3	4

Cost Function

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

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

Minimize $\boldsymbol{J(\theta_0 , \theta_1)}$

$\boldsymbol{\theta_0 , \theta_1}$

If $J(\theta_1) = (\theta_1 -2)^2$

$\frac{dJ(\theta_1)}{d\theta_1} = 0$ $\Rightarrow$ $\frac{dJ(\theta_1)}{d\theta_1} = 2\ (\theta_1 -2) = 0$ $\Rightarrow$ $\theta_1 = 2$

Gradient Descent

Minimize $\boldsymbol{J(\theta_0 , \theta_1)}$

$$\boldsymbol{\theta_0 , \theta_1}$$

Minimize $\boldsymbol{J(\theta_0 , \theta_1, \dots , \theta_n)}$

$$\boldsymbol{\theta_0 , \theta_1, \dots , \theta_n}$$

Repeat until convergence: {

For j=0,…,n

$$\boldsymbol{\theta_j} = \boldsymbol{\theta_j} - \alpha \frac{d\boldsymbol{J(\theta_0 , \theta_1, \dots , \theta_n)}}{d\boldsymbol{\theta_j}}$$

}

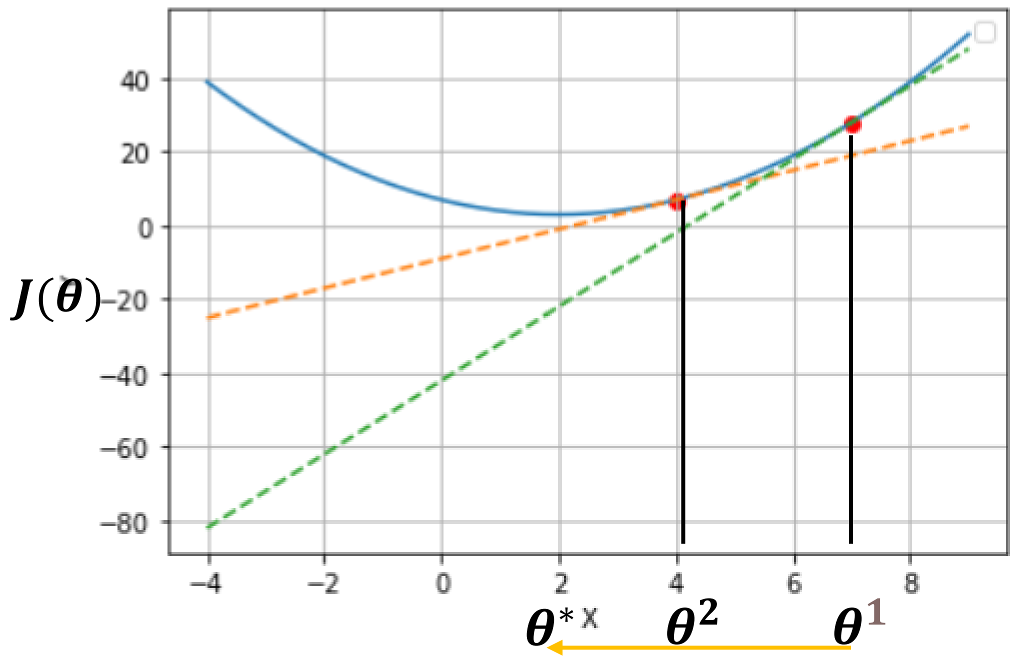
α **is learning rate**

Updating all $\boldsymbol{\theta_j}$ $\boldsymbol{Simultaneous}$**ly**

Convergence condition:

$$\|\theta^{t+1} - \theta^{t}\|_2 \leq \varepsilon$$

Gradient Descent



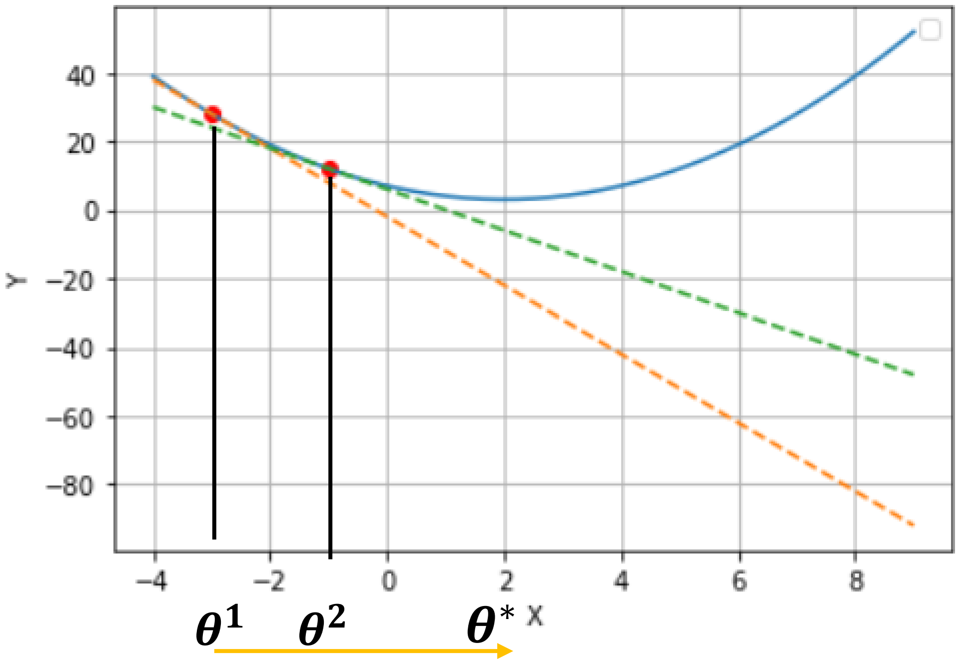
خطوط مماس نشان داده شده دارای شیب یا مشتق مثبت هستند.
درنتیجه:

$$\frac{dJ(\theta^1)}{d\theta^1} > 0 \,,\ \alpha > 0 \Rightarrow \alpha \frac{dJ(\theta^1)}{d\theta^1} > 0$$

$$\Rightarrow \theta^2 = \theta^1 - \alpha d\theta^1$$

θ کوچکتر میشود و به سمت چپ حرکت میکنیم.

Gradient Descent



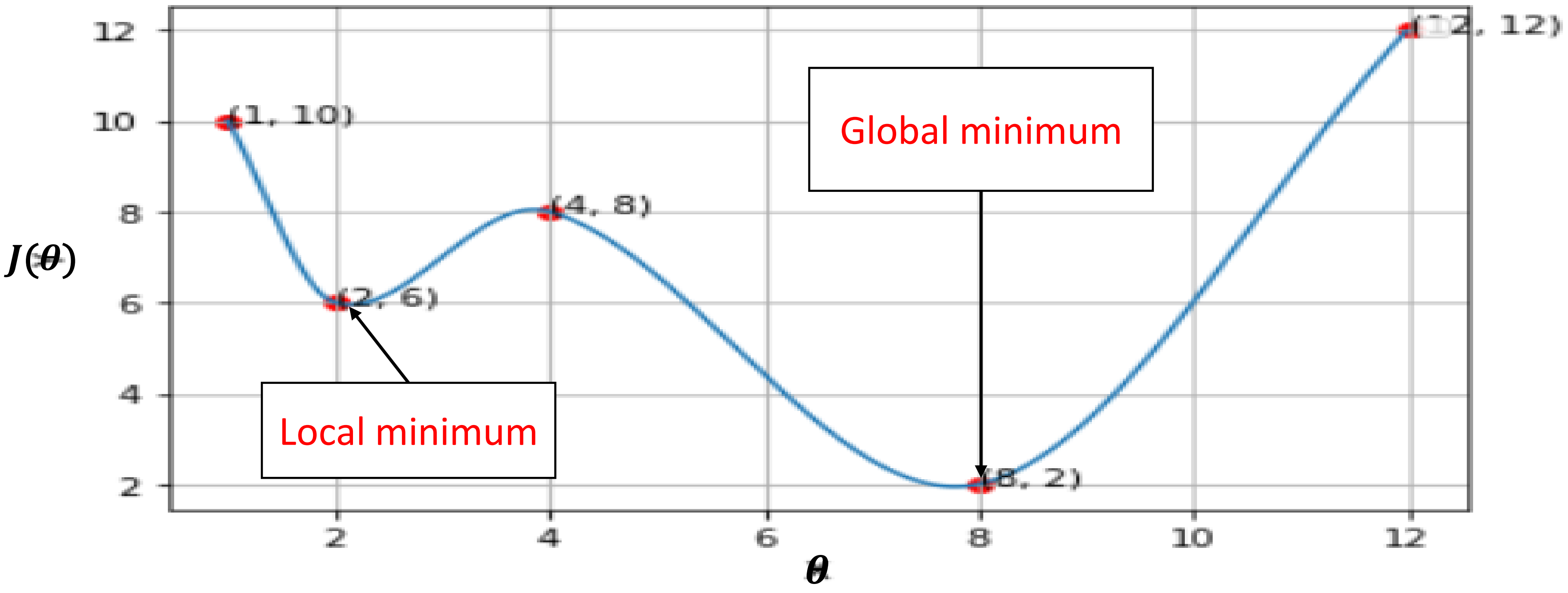
خطوط مماس نشان داده شده دارای شیب یا مشتق مثبت هستند.
درنتیجه:

$$\frac{dJ(\theta^1)}{d\theta^1} < 0 \,,\ \alpha > 0 \Rightarrow \alpha \frac{dJ(\theta^1)}{d\theta^1} < 0$$

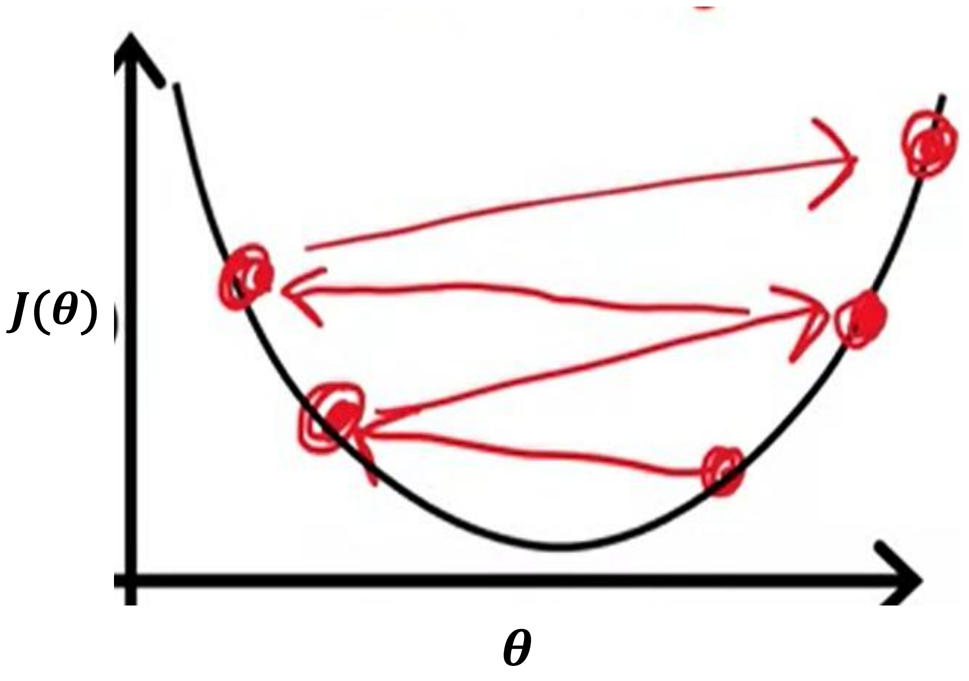
$$\Rightarrow \theta^2 = \theta^1 - \alpha d\theta^1$$

θ بزرگتر میشود و به سمت راست حرکت میکنیم.

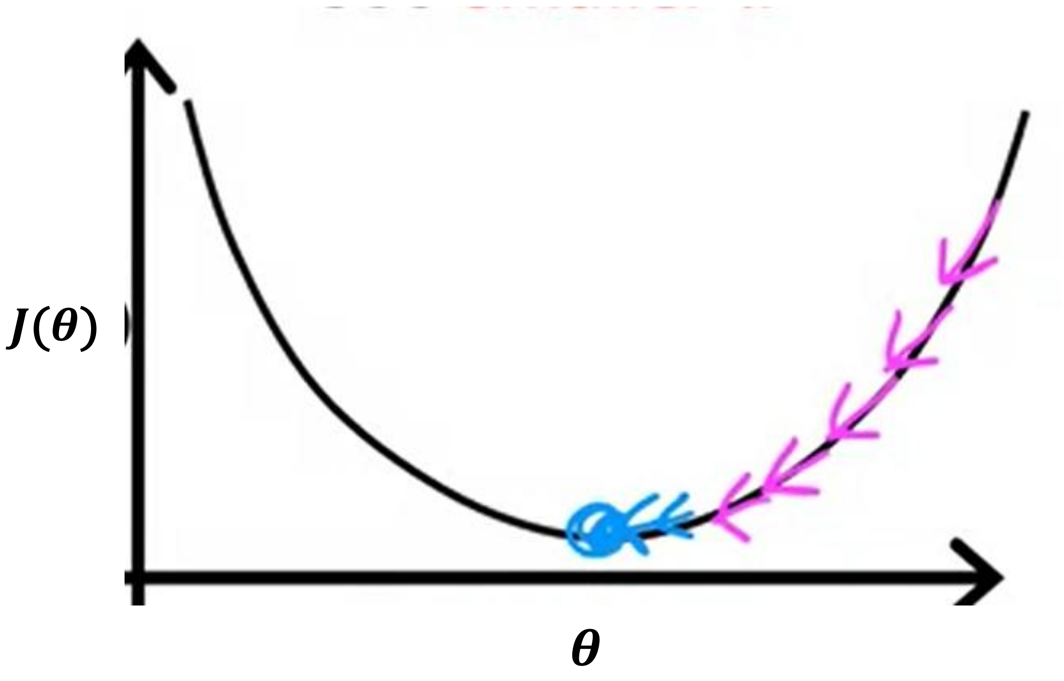
Gradient Descent Weakness



Choosing Learning Rate



α is too large



α is small

Gradient Descent

Correct form

temp0 = $\boldsymbol{\theta_0} \; - \; \alpha \dfrac{dJ(\boldsymbol{\theta_0} \,, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_0}}$

temp1 = $\boldsymbol{\theta_1} \; - \; \alpha \dfrac{dJ(\boldsymbol{\theta_0} \,, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_1}}$

$\boldsymbol{\theta_0}$ = temp0
$\boldsymbol{\theta_1}$ = temp1

✓

Incorrect form

$\boldsymbol{\theta_0} = \boldsymbol{\theta_0} \; - \; \alpha \dfrac{dJ(\boldsymbol{\theta_0} \,, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_0}}$

$\boldsymbol{\theta_1} = \boldsymbol{\theta_1} \; - \; \alpha \dfrac{dJ(\boldsymbol{\theta_0} \,, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_1}}$

✗

Linear regression model

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

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

$$\frac{dJ(\boldsymbol{\theta_0}, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_0}} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$$

$$\frac{dJ(\boldsymbol{\theta_0}, \boldsymbol{\theta_1})}{d\boldsymbol{\theta_1}} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)\, x_i$$

Linear regression model

Repeat until convergence: {

$$\boldsymbol{\theta_0} = \boldsymbol{\theta_0} - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$$

$$\boldsymbol{\theta_1} = \boldsymbol{\theta_1} - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)\, x_i$$

}

بروز رسانی همزمان

$$\theta^t = \begin{bmatrix} \boldsymbol{\theta_0} \\ \boldsymbol{\theta_1} \end{bmatrix}, \quad \theta^{t+1} = \begin{bmatrix} \boldsymbol{\theta_0} \\ \boldsymbol{\theta_1} \end{bmatrix}, \quad d\theta = \begin{bmatrix} d\boldsymbol{\theta_0} \\ d\boldsymbol{\theta_1} \end{bmatrix}$$

Convergence condition:

- $\|\theta^{t+1} - \theta^t\|_2 = \sqrt[2]{(\theta_0^{t+1} - \theta_0^t)^2 + (\theta_1^{t+1} - \theta_1^t)^2} < \varepsilon$
- $\|d\theta\|_2 < \varepsilon$

Batch Gradient Descent

$$\frac{d\boldsymbol{J}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}(h_{\theta}(x_i) - y_i)$$

Batch Gradient Descent

$\theta_0 \leftarrow random$, $\theta_1 \leftarrow random$

Repeat until convergence: {

J $\leftarrow$ 0 , $d\theta_1 \leftarrow$ 0 , $d\theta_0 \leftarrow$ 0

For i = 1 to m:

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

j += $(h_\theta(x_i) - y_i)^2$

$d\theta_1$ += $2\ (h_\theta(x_i) - y_i)\ x_i$

$d\theta_0$ += $2\ (h_\theta(x_i) - y_i)$

J /= 2m

$d\theta_1$ /= 2m

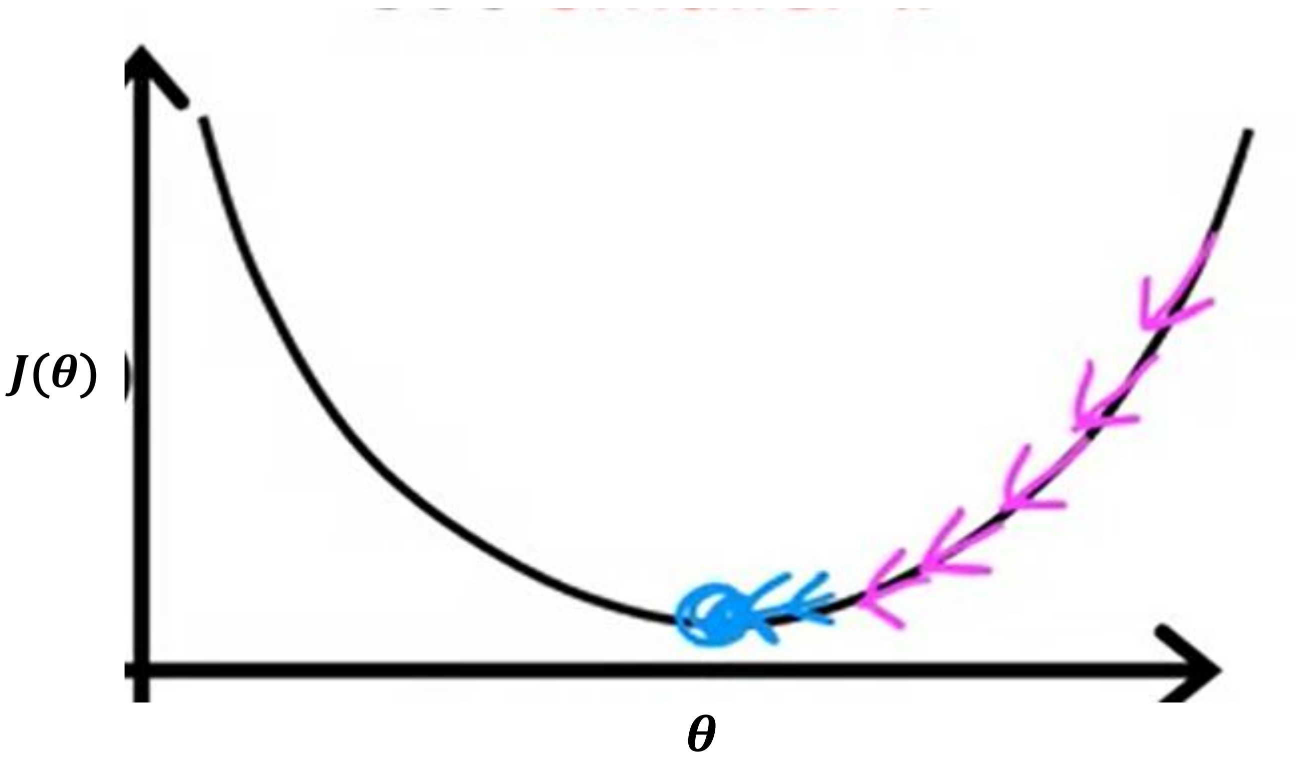
$d\theta_0$ /= 2m

$\theta_1 = \theta_1 - \alpha d\theta_1$

$\theta_0 = \theta_0 - \alpha d\theta_0$

}

Gradient Descent



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

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

number	size	#bedrooms	# floors	Price(y)
1	100	2	1	10000
2	150	3	2	175000
…	…	…	…	…
m	…	…	…	…

n: #features = 3

m: #training data

x_i: i th data in training set

x_j^i: j th feature of i th data in training set

$h_\theta(x^i) = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + ... + \theta_n x_n^i$

$y = [y^1 , y^2 , ... , y^m]^T \in R^{m*1}$

$X = [\boldsymbol{x}^1 , \boldsymbol{x}^2 , ... , \boldsymbol{x}^m]^T \in R^{m*(n+1)}$

$$\vec{x} = \begin{bmatrix} \boldsymbol{x_0} \\ \boldsymbol{x_1} \\ \boldsymbol{x_2} \\ . \\ . \\ . \\ \boldsymbol{x_n} \end{bmatrix} \in R^{n+1} \quad , \quad \overrightarrow{\theta} = \begin{bmatrix} \boldsymbol{\theta_0} \\ \boldsymbol{\theta_1} \\ \boldsymbol{\theta_2} \\ . \\ . \\ . \\ \boldsymbol{\theta_n} \end{bmatrix} \in R^{n+1} \Rightarrow h_\theta(x) = x^T\theta = \theta^T x$$

$\boldsymbol{x_0} = 1$

θ_0 ***is bias***

Cost function

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

$$e^i = (x^i)^T \ \theta \ - y^i \Longrightarrow \quad e = X\theta - y \quad \Longrightarrow \quad J(\theta) = \frac{1}{2m} e^T e$$

$$e \ , \ X\theta \ , y \in R^m$$

Gradient Descent

Repeat until convergence: {

For j=0,…,n

$$\boldsymbol{\theta_j} = \boldsymbol{\theta_j} - \alpha \frac{d\boldsymbol{J}(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_n})}{d\boldsymbol{\theta_j}}$$

}

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

$$\frac{d\boldsymbol{J}(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_n})}{d\boldsymbol{\theta_j}} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)\, x_j^i$$

(j=0,…,n , x_0^i = 1)

$$\frac{d\boldsymbol{J}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{1}{m} X^T \mathrm{e}$$

(n+1)*1

m*1

$$\frac{d\boldsymbol{J}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)\, x^i$$

حجم محاسبات ضرب ماتریس

$A \in R^{a*b}$, $B \in R^{b*c}$ $\Rightarrow$ $\underset{O(abc)}{AB} \in R^{a*c}$ $(2b-1)\,ac$ *flops*

Calculating : $e = X\theta - y$

a=m
b=n+1
c=1

$m(2n+1)$ (ضرب و جمع)
m (تفریق)

2m(n+1) (مجموع) $\Rightarrow$ O(mn)

$$\frac{dJ(\theta)}{d\theta} : (2m-1)(n+1) + (n+1) = 2m(n+1)$$ $\Rightarrow$ O(mn) در مجموع

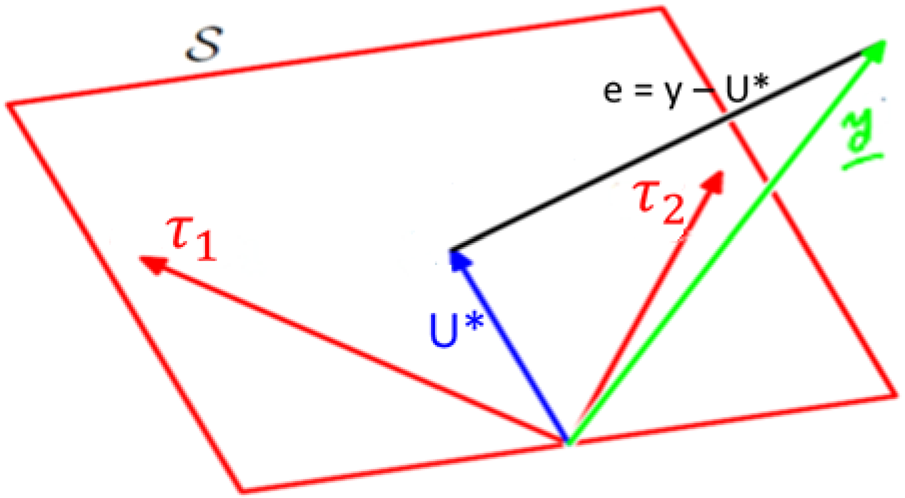
تقسیم بر m

$\frac{1}{m} X^T e$

a=n+1
b=m
c=1

مفهوم هندسی

$$\min_{\theta} \|y - X\theta\|_2 = \min_{\theta} \|e\|_2$$



Span of X:

فضایی که توسط ستون های X پوشش داده می شود. هر بردار در این فضا به صورت U = Xθ نشان داده می شود. و به آن span(X) می گویند.

U بهینه که به صورت U* نشان داده می شود برداری است که e = y – U* بر span(X) عمود باشد. یا به عبات دیگر U*ای باید انتخاب شود که برابر با نگاشت y در span(X) باشد.

$$X = \begin{bmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \\ x_3^1 & x_3^2 \end{bmatrix} \epsilon \mathcal{R}^{m\times n}$$

τ_1 τ_2

m: تعداد داده

n: تعداد ویژگی

Feature Scaling

$$x_1^i \ , x_2^i \ , \ \dots \ , x_n^i$$

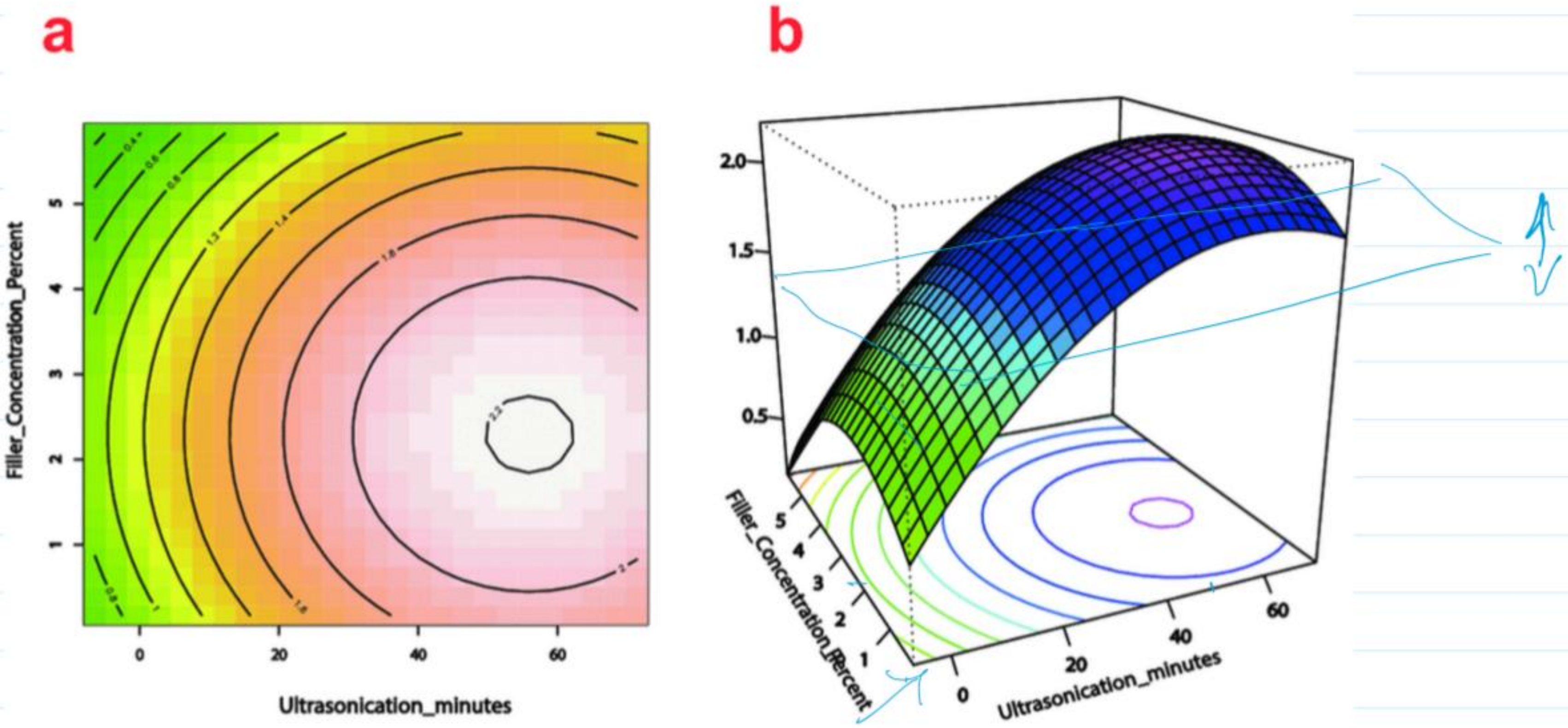
$$-1 \leq x_j \ \leq 1$$

$0 < \ x_1 \ < 1000$

$0 < \ x_2 \ < 5$

$x_1 : \dfrac{size}{1000}$

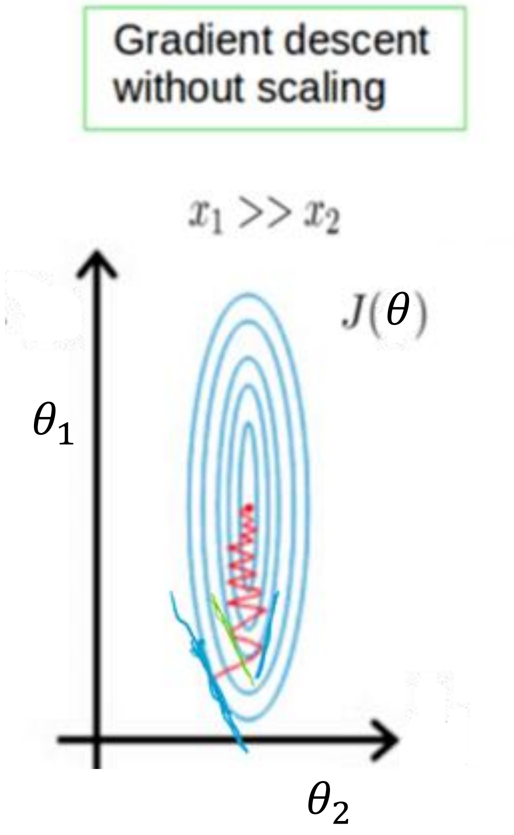
$x_2 : \dfrac{\#bedrooms}{5}$

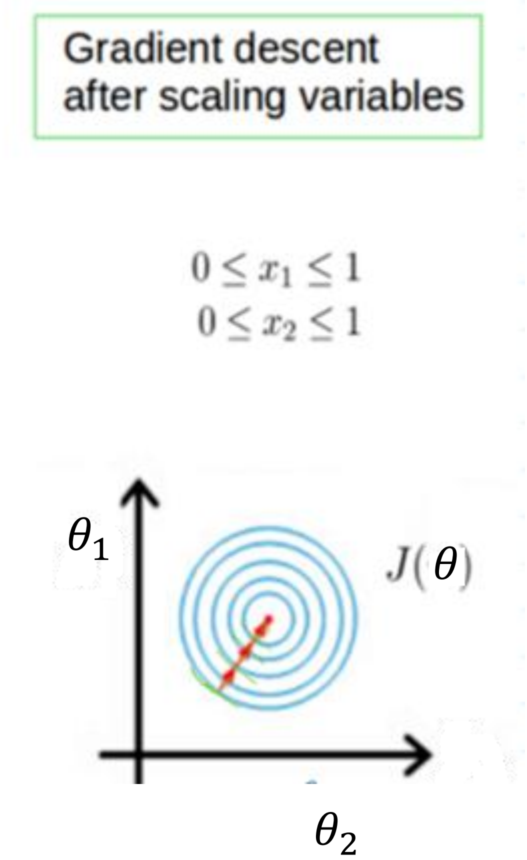

a
b
Filler_Concentration_Percent
Ultrasonication_minutes
0
20
40
60
1
2
3
4
5
2.0
1.5
1.0
0.5

Contour Plot

$$\frac{\theta_1}{b^2} + \frac{\theta_2}{a^2} = 1$$

2a: قطر بزرگ
2b: قطر کوچک





$$\frac{\theta_1}{a^2} + \frac{\theta_2}{a^2} = 1$$

Feature Scaling

Scaled features:

- $0 \leq x_1 \leq 3$ ✓
- $-3 \leq x_1 \leq 3$ ✓
- $-2 \leq x_2 \leq 0.5$ ✓
- $-\frac{1}{3} \leq x_2 \leq \frac{1}{3}$ ✓

Need scaling:

$-100 \leq x_3 \leq 100$ ✗

$-0.001 \leq x_4 \leq 0.001$ ✗

Feature Scaling

$$x_1^* = \frac{x_1 - \mu_1}{standard\_deviation}$$

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^i$$

$$bedroom^* = \frac{bedroom - 2.5}{5}$$

$$size^* = \frac{size - 300}{2000}$$

Creating New Features

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

قیمت خانه

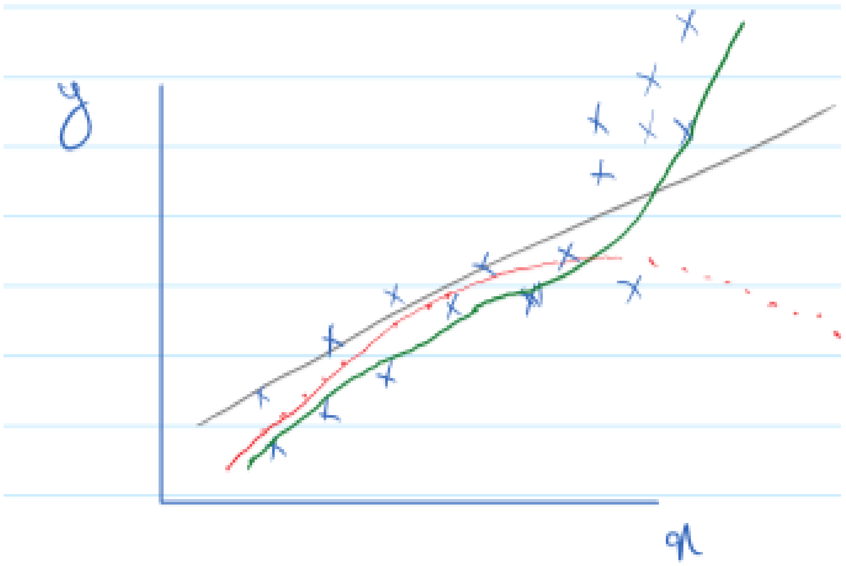
طول خانه

عرض خانه

(مساحت خانه) $x^* = x_1 * x_2$

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

Creating New Features



درجه 2:

$$\theta_0 + \theta_1 x + \theta_2 x^2$$

درجه 3:

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

Need scaling:

x: 0,…,1000

x^2: 0,…, 10^6

x^3: 0,…, 10^9

We can use:

x , x^2 , x^3 , $\sqrt{x}$

$$\theta_0 + \theta_1 x + \theta_2 \sqrt{x}$$