CS 4104
APPLIED MACHINE LEARNING

Dr. Hashim Yasin

National University of Computer and Emerging Sciences,

Faisalabad, Pakistan.

REGRESSION

Classification vs Regression

Classification problem

Features | Label

#	Height (inches)	Weight (kgs)	B.P. Sys	B.P. Dia	Heart disease
1	62	70	120	80	No
2	72	90	110	70	No
3	74	80	130	70	No
4	65	120	150	90	Yes
5	67	100	140	85	Yes
6	64	110	130	90	No
7	69	150	170	100	Yes
8	66	125	145	90	?
9	74	67	110	60	?

Feature vector (4-dimensional)

Label vector

Training Data

Test Data

Classification vs Regression

Regression problem

#	Height (inches)	Weight (kgs)	B.P. Sys	B.P. Dia	Cholesterol Level
1	62	70	120	80	150
2	72	90	110	70	160
3	74	80	130	70	130
4	65	120	150	90	200
5	67	100	140	85	190
6	64	110	130	90	130
7	69	150	170	100	250
8	66	125	145	90	?
9	74	67	110	60	?

LINEAR REGRESSION

Linear Regression with One Variable

Training set of housing prices

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

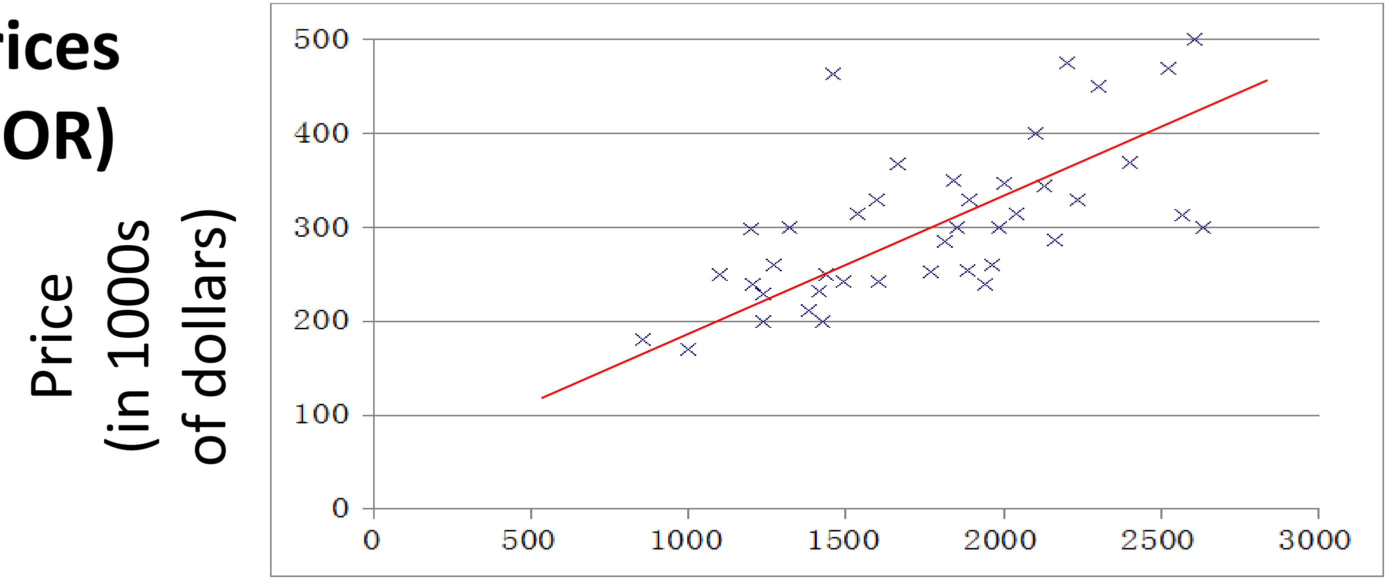
Notation:

m = Number of training examples
x's = "input" variable / features
y's = "output" variable / "target" variable

One Training example (x, y)
i^{th} training example $(x^{(i)}, y^{(i)})$

Linear Regression with One Variable

Housing Prices
(Portland, OR)



Size ($feet^2$)

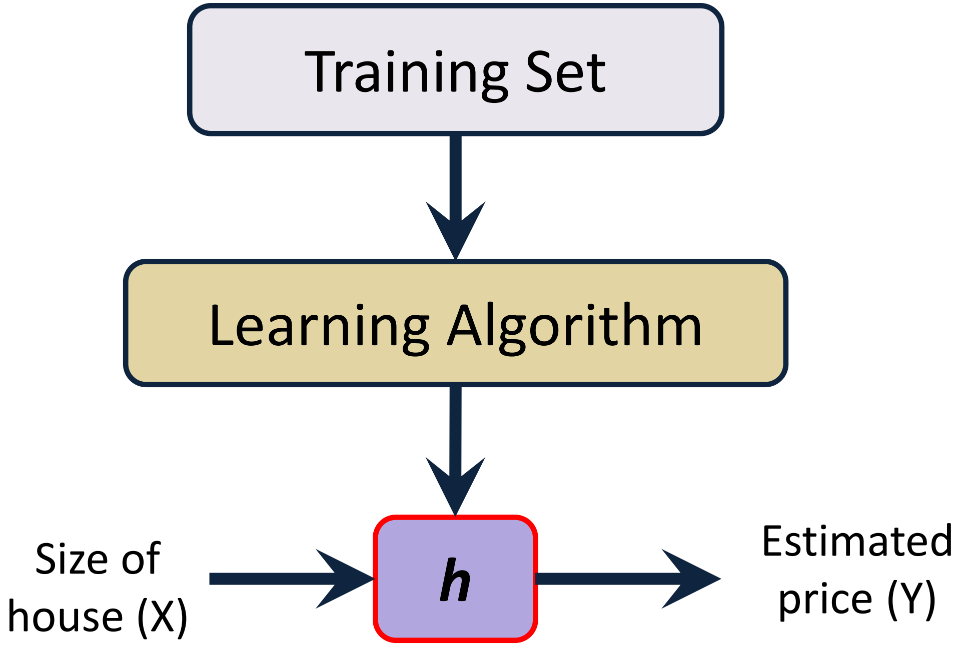
Supervised Learning

Given the "right answer" for each example in the data.

Regression Problem

Predict *real-valued* output

Regression



Question : How to describe *h*?

$$h: X \rightarrow Y$$

Regression Example

Training set of housing prices

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

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

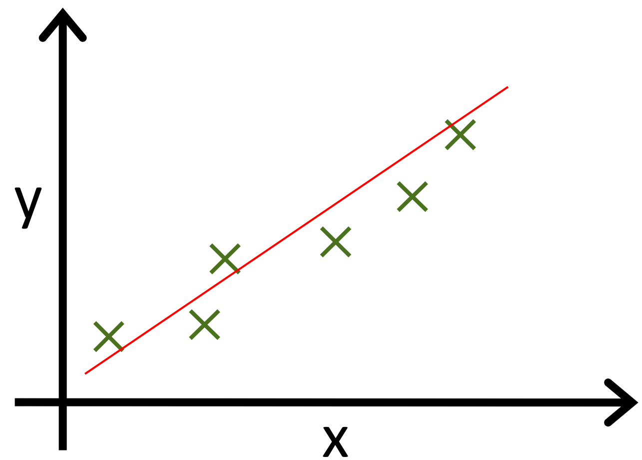
θ_i's: Parameters

How to choose θ_i's ?

Regression

- How to choose these parameters , θ (regression coefficient)?
- The standard approach is the **least square method,** through which parameters are minimized
- The machine learning program optimizes the parameters, θ, such that the approximation error is minimized.

Regression



Idea: Choose θ_0, θ_1 so that
$h_\theta(x)$ is close to y for our
training examples (x, y)

Cost Function

Simplified:

Hypothesis:

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

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

Parameters:

$$\theta_0, \theta_1$$

$$\theta_1$$

Cost Function:

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

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

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

$$\underset{\theta_1}{\text{minimize}}\ J(\theta_1)$$

REGRESSION EXAMPLE

Cost Function … Example

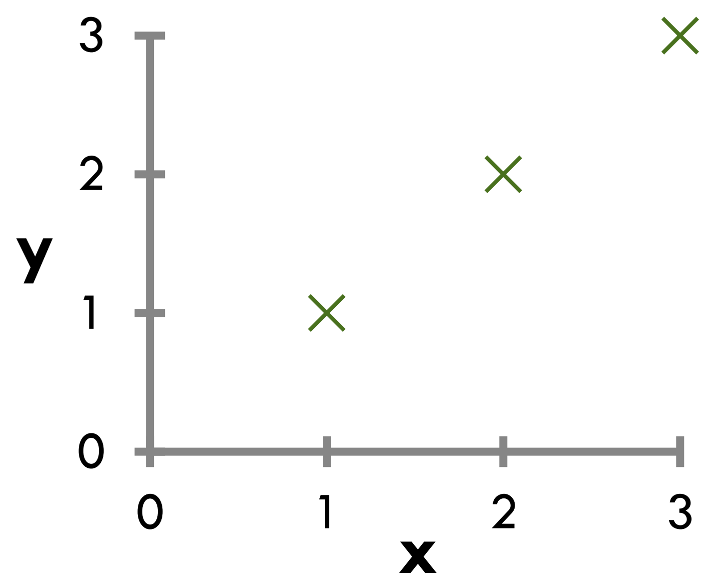
Consider the given cost function, hypothesis and the datapoints. Find out the cost function values, when the parameter (θ_1) values are: 0, 0.5 and 1,

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

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

x	y
1	1
2	2
3	3

$$\theta_1 = 0,$$
$$\theta_1 = 0.5$$
$$\theta_1 = 1$$

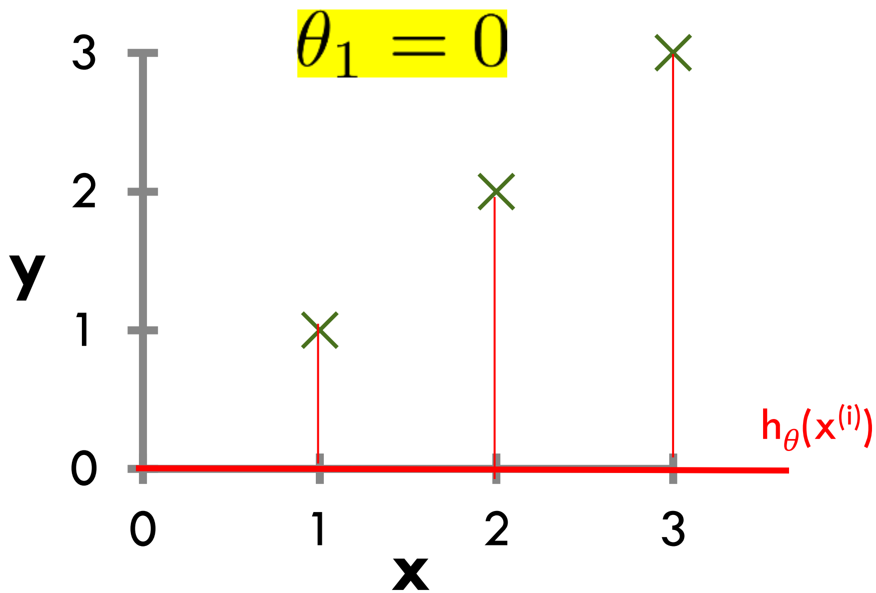


$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost Function

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x)



x	y
1	1
2	2
3	3

$$h_1 = 0 \times 1 = 0$$

$$h_2 = 0 \times 2 = 0$$

$$h_3 = 0 \times 3 = 0$$

$J(\theta_1)$ (function of the parameter θ_1)

$$J(\theta_1) = J(0) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

$$= \frac{1}{2m}\sum_{i=1}^{m}(\theta_1 x^i - y^i)^2$$

$$= \frac{1}{2m}[(0-1)^2 + (0-2)^2 + (0-3)^2]$$

$$= \frac{1}{2 \times 3}[1 + 4 + 9]$$

$$= \frac{1}{6} \times 14 = 2.3$$

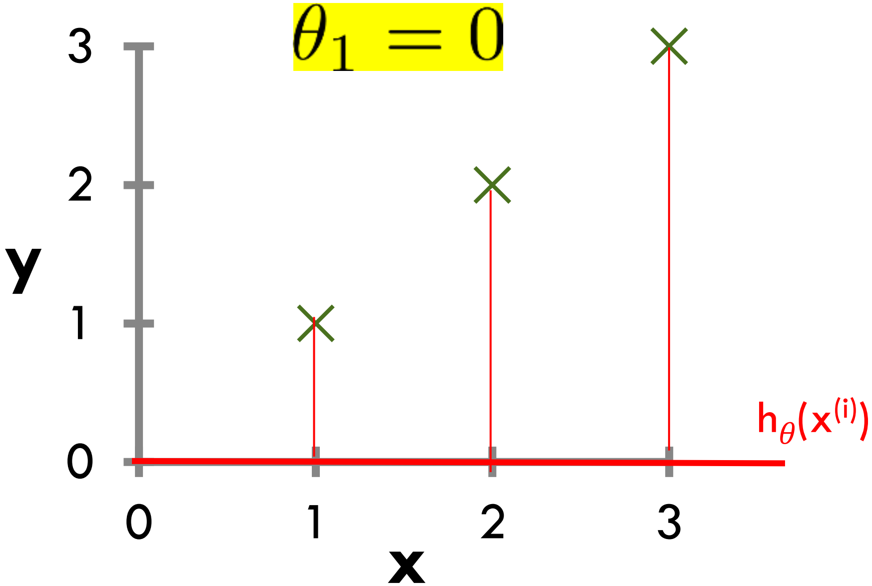
Cost Function

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

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x)

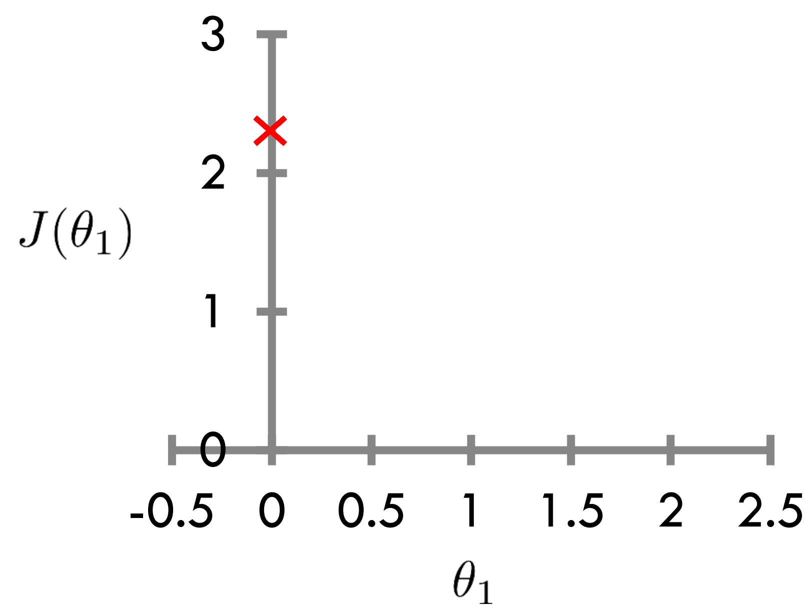
$\theta_1 = 0$



$$J(0) = \frac{1}{2\times 3} \sum_{i=1}^{3} [1^2 + 2^2 + 3^2]$$

$$= \frac{1}{6} \times 14 = 2.3$$

$J(\theta_1)$ (function of the parameter θ_1)

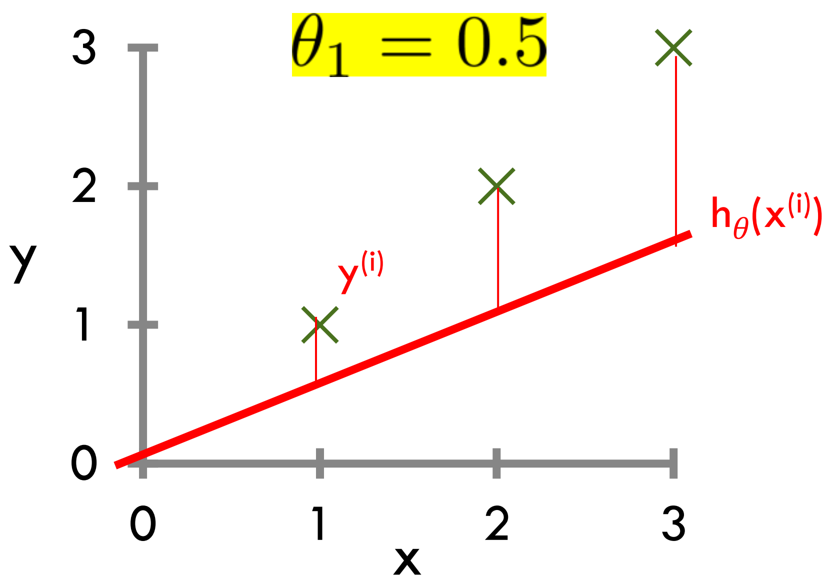


Cost Function

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

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x)



x	y
1	1
2	2
3	3

$$h_1 = 0.5 \times 1 = 0.5$$

$$h_2 = 0.5 \times 2 = 1$$

$$h_3 = 0.5 \times 3 = 1.5$$

$J(\theta_1)$ (function of the parameter θ_1)

$$J(\theta_1) = J(0.5) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

$$= \frac{1}{2m}\sum_{i=1}^{m}(\theta_1 x^i - y^i)^2$$

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

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

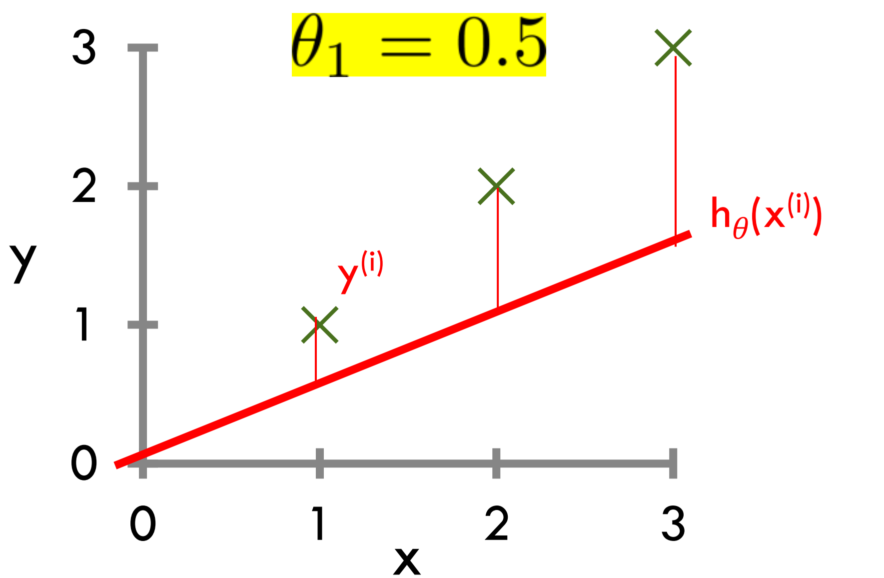
$$= \frac{1}{6}\times(3.5) = 0.58$$

Cost Function

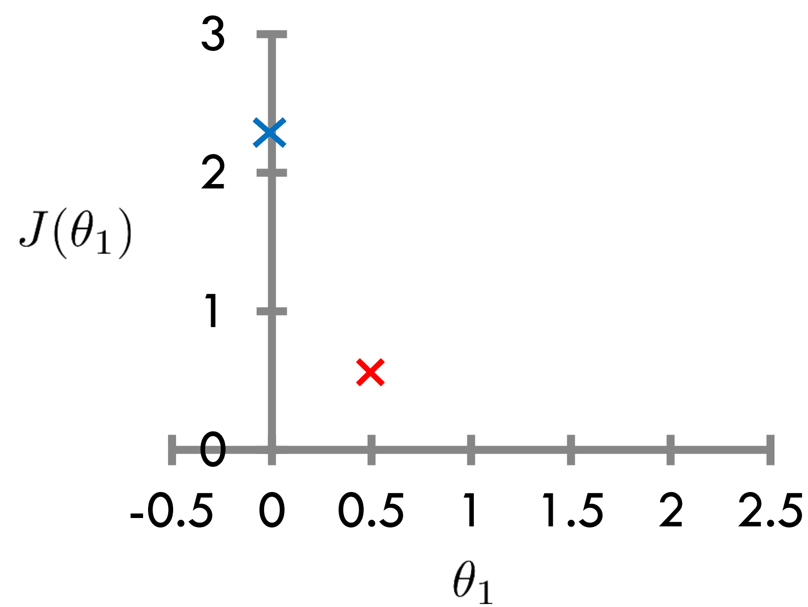
$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x)



$J(\theta_1)$ (function of the parameter θ_1)



$$J(0.5) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right]$$

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

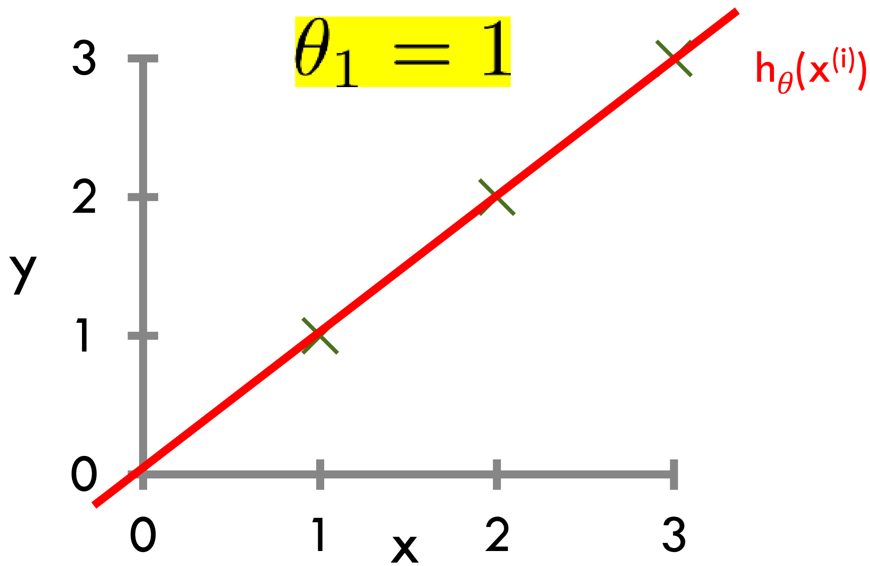
$J(0.5) = 0.58$

Cost Function

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

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x)



x	y
1	1
2	2
3	3

$h_1 = 1 \times 1 = 1$

$h_2 = 1 \times 2 = 2$

$h_3 = 1 \times 3 = 3$

$J(\theta_1)$ (function of the parameter θ_1)

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

$= \frac{1}{2m}\sum_{i=1}^{m}(\theta_1 x^i - y^i)^2$

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

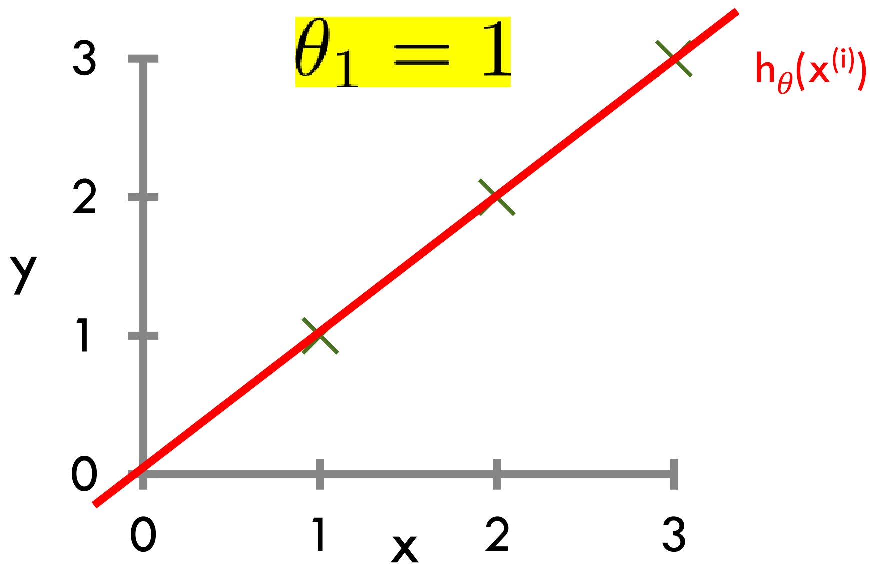
$= \frac{1}{2m}(0^2 + 0^2 + 0^2) = 0$

Cost Function

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

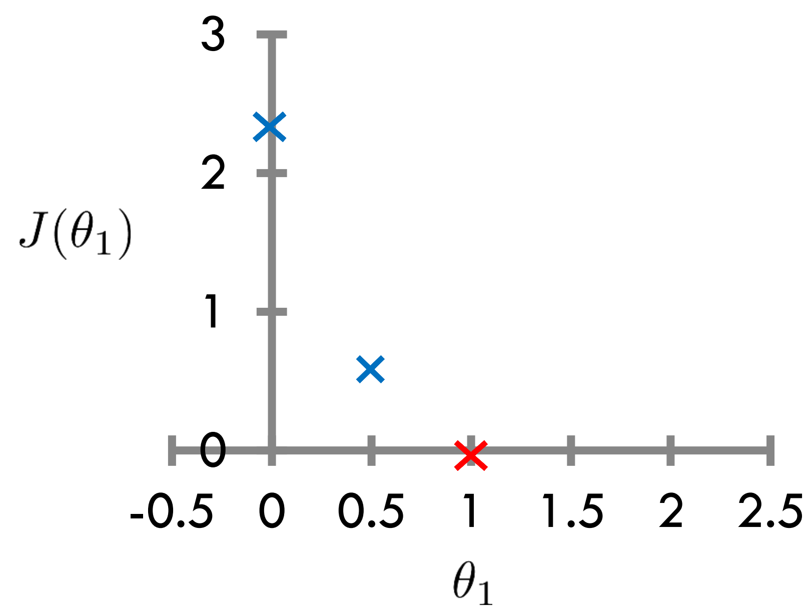
$$h_\theta(x) = \theta_1 x$$

$h_\theta(x)$ (for fixed θ_1, this is a function of x)



$$J(\theta_1) = J(1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$
$$= \frac{1}{2m}\sum_{i=1}^{m}(\theta_1 x^i - y^i)^2$$
$$= \frac{1}{2m}(0^2 + 0^2 + 0^2) = 0$$

$J(\theta_1)$ (function of the parameter θ_1)



$J(1) = 0$

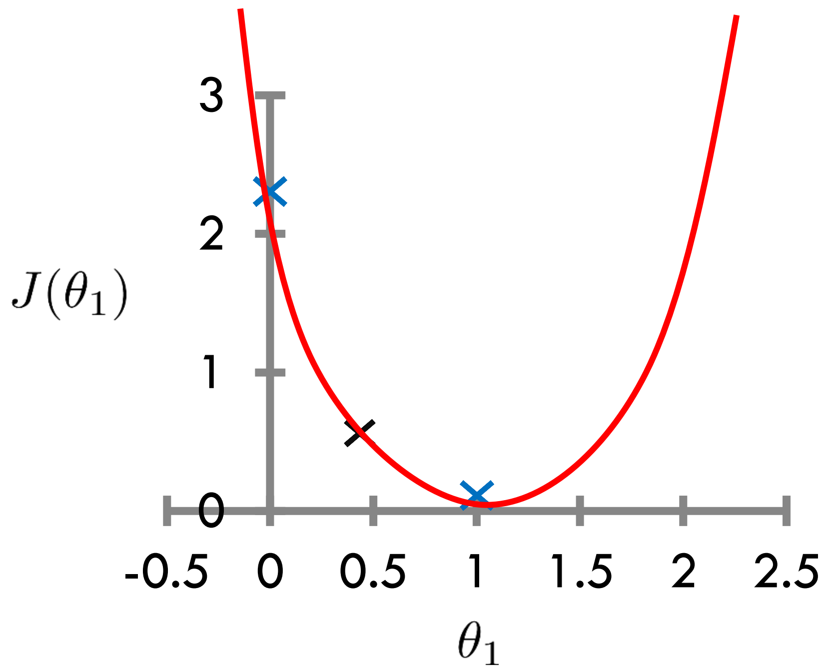
$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost Function

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

$h_\theta(x)$ (for fixed θ_1, this is a function of x).

$J(\theta_1)$ (function of the parameter θ_1)



What's next?

Have some function $J(\theta_0, \theta_1)$

Want $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Outline:

- Start with some random values θ_0, θ_1
- **Keep changing** θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

  until we hopefully end up at a minimum

GRADIENT DESCENT

Partial Derivatives … Preliminaries

Say for instance, we have a function:

$$f(x, y) = x^4 + y^7$$

partial derivative of the function w.r.t 'x' will be :

$$\frac{\partial f}{\partial x} = 4x^3 + 0$$

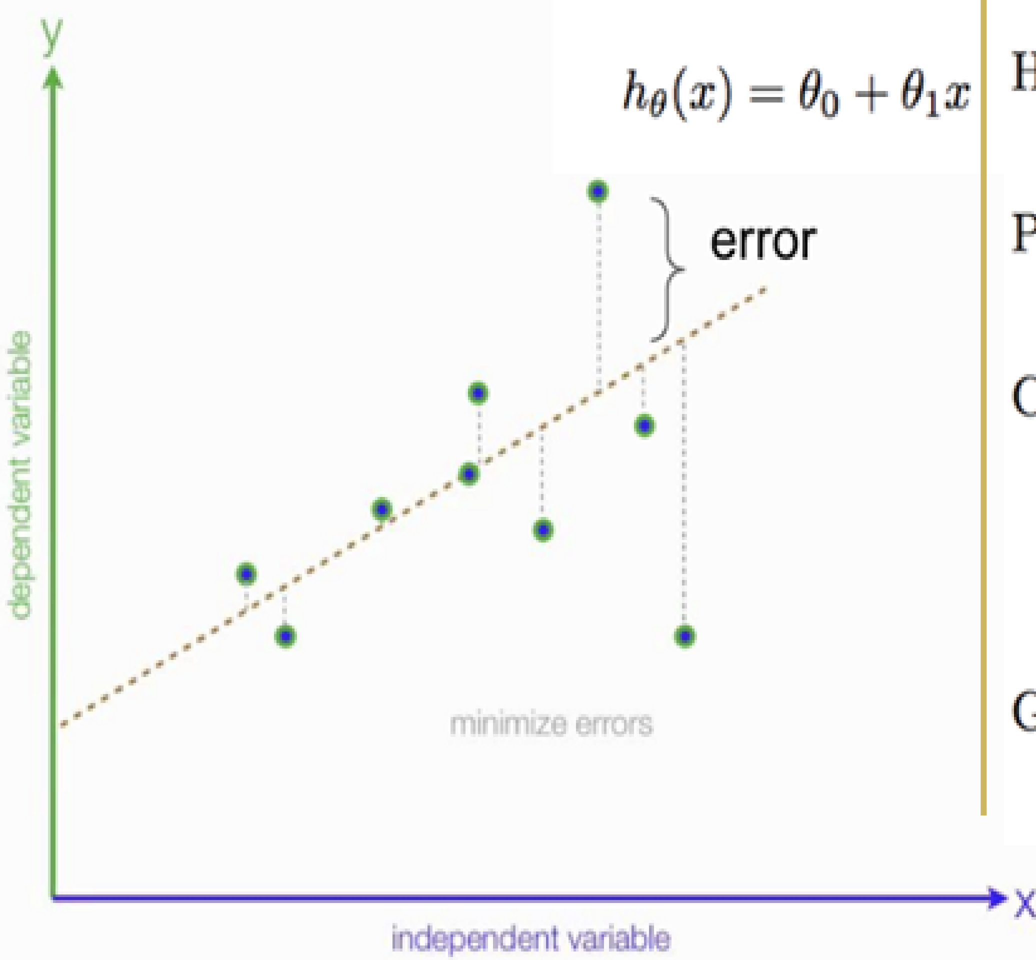
treating 'y' as a constant

And partial derivative of the function w.r.t 'y' will be :

$$\frac{\partial f}{\partial y} = 0 + 7y^6$$

treating 'x' as a constant

Gradient Descent



Hypothesis:

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

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

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

Goal:

$$\underset{\theta_0, \theta_1}{minimize} J(\theta_0, \theta_1)$$

Gradient Descent

Have some function $J(\theta_0, \theta_1)$

Want $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Outline:

- Start with some θ_0, θ_1
- Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

  until we hopefully end up at a minimum

Gradient Descent

Gradient descent algorithm

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad \text{(simultaneously update } j = 0 \text{ and } j = 1) \\
&\}
\end{aligned}
$$

Notice : α is the learning rate.

Gradient descent algorithm

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad (\text{for } j = 0 \text{ and } j = 1) \\
&\}
\end{aligned}
$$

Correct: Simultaneous update

$$
\begin{aligned}
\text{temp0} &:= \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \\
\text{temp1} &:= \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) \\
\theta_0 &:= \text{temp0} \\
\theta_1 &:= \text{temp1}
\end{aligned}
$$

Incorrect:

$$
\begin{aligned}
\text{temp0} &:= \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \\
\theta_0 &:= \text{temp0} \\
\text{temp1} &:= \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) \\
\theta_1 &:= \text{temp1}
\end{aligned}
$$

Gradient Descent

Gradient descent algorithm

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\qquad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad \text{(simultaneously update } j = 0 \text{ and } j = 1) \\
&\}
\end{aligned}
$$

Notice : α is the learning rate.

?

Gradient Descent

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

Partial Derivative w.r.t. θ_0

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

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

$$\frac{\partial\, J(\theta_0, \theta_1)}{\partial \theta_0} = \frac{1}{2m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_0} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial\, J(\theta_0, \theta_1)}{\partial \theta_0} = 2\,\frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \frac{\partial}{\partial \theta_0} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$

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

Gradient Descent

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

Partial Derivative w.r.t. θ_1

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

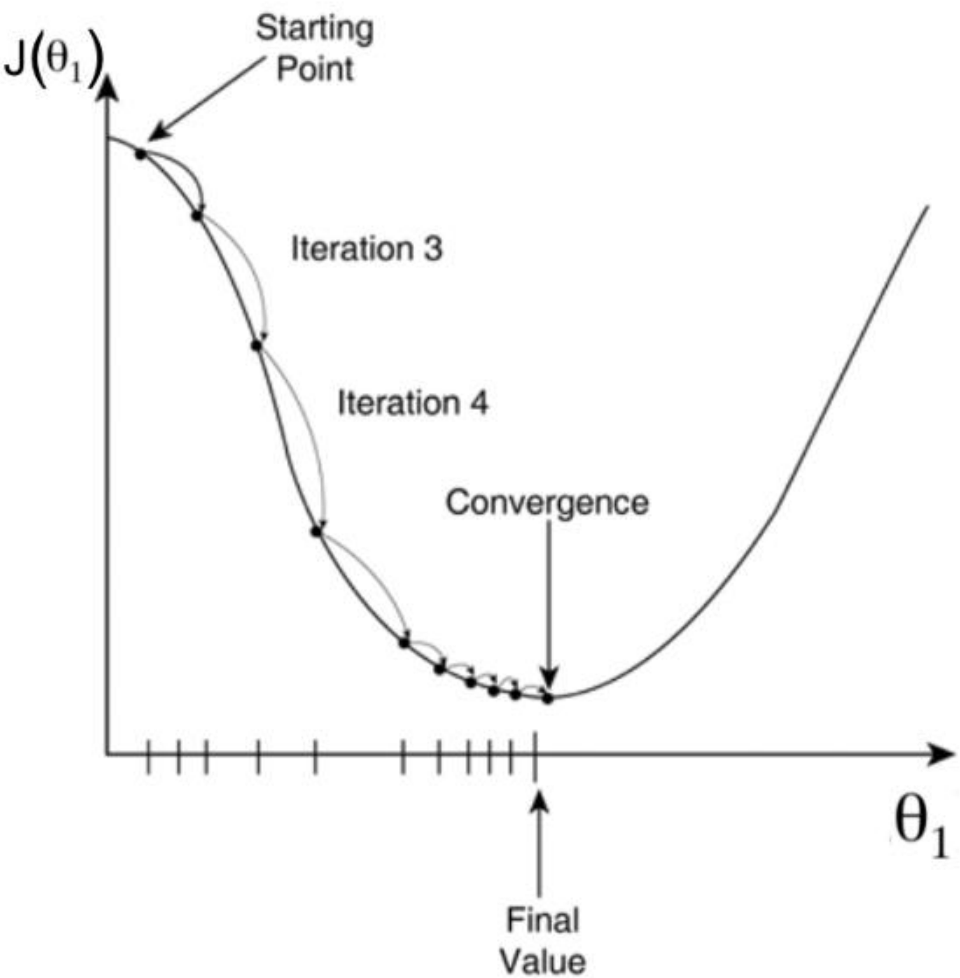
$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{\partial}{\partial \theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{1}{2m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_1} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = 2 \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \frac{\partial}{\partial \theta_1} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$

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

Gradient Descent



Cost Function – "One Half Mean Squared Error":

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

Objective:

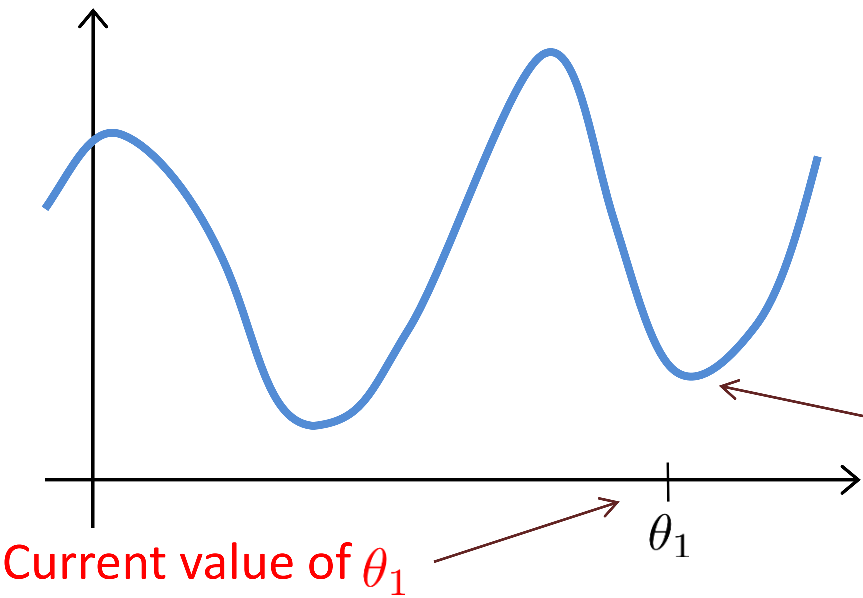
$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Derivatives:

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

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \cdot x^{(i)}$$

Gradient Descent … Learning Rate



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

Unchange

θ_1 at local optima

Gradient descent can converge to a local minimum, even with the learning rate α fixed.

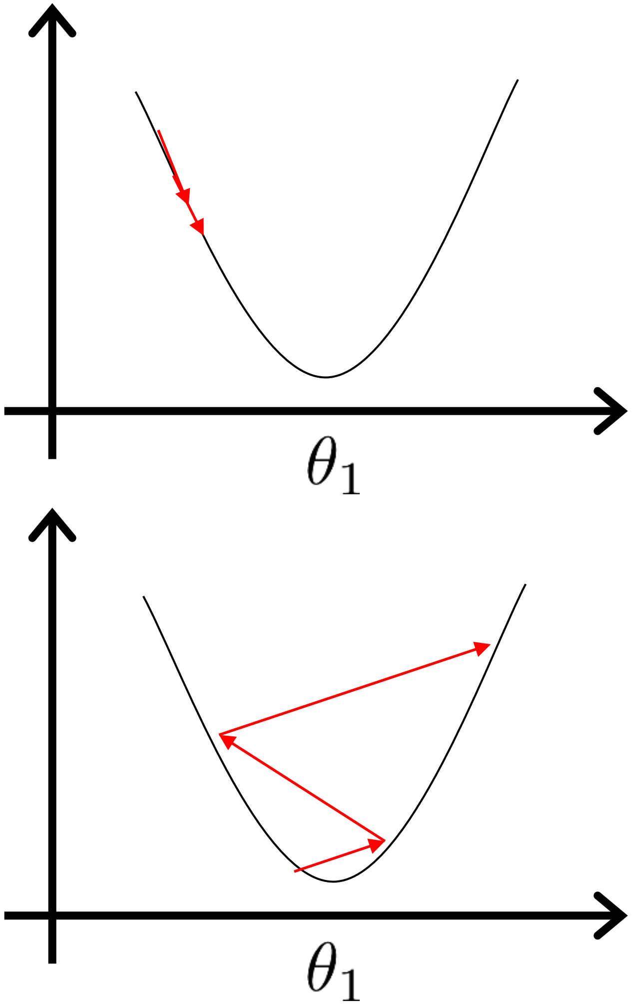
As we approach a local minimum, gradient descent will automatically take smaller steps. So, no need to decrease α over time.

Gradient Descent … Learning Rate

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If α is too small, gradient descent can be slow.

If α is too large, gradient descent can overshoot the minimum. It may fail to converge.



Summary

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

Objective: $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Update rules:

$$\theta_0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

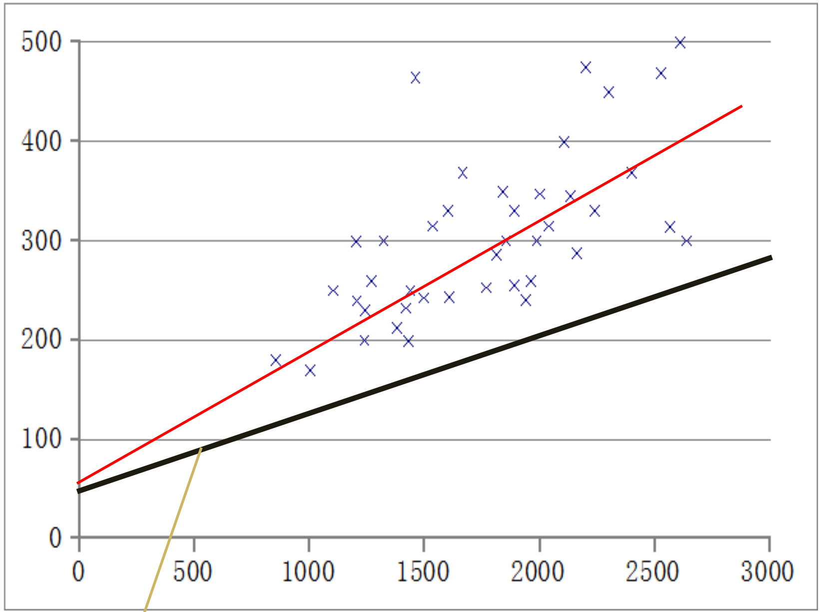
Derivatives:

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

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \cdot x^{(i)}$$

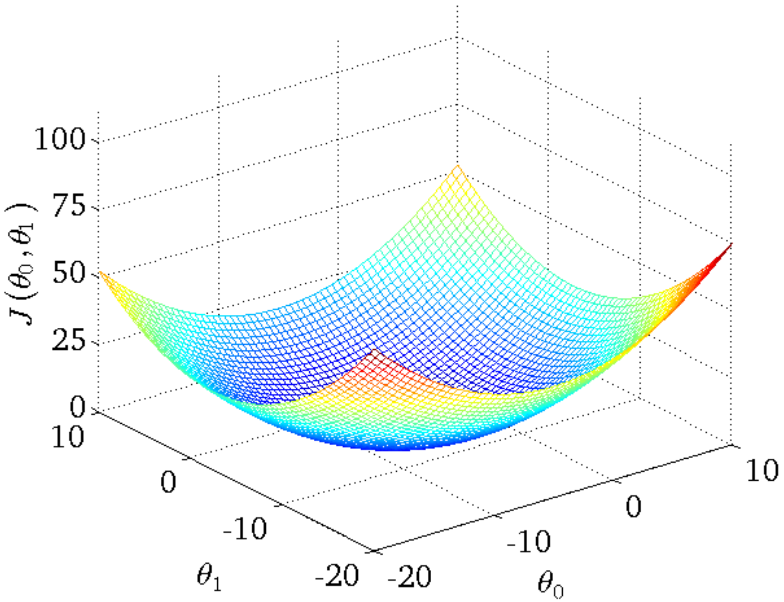
Regression … Example

Price ($) in 1000's



Size in feet² (x)

$$h_\theta(x) = 50 + 0.06x$$



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

$$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$$

Acknowledgement

Tom Mitchel, Russel & Norvig, Andrew Ng, Alpydin & Ch. Eick.