LECTURE 4

SPRING 2021
APPLIED MACHINE LEARNING
CIHANG XIE

SLIDE CREDIT:
NARGES NOROUZI

TODAY

- Linear Regression
- Least Squares Method
- Gradient Descent Algorithm

THE FIRST ORDER LINEAR MODEL

Image credit: https://nextjournal.com/intelrefinery/simple-linear-regression

REGRESSION HYPOTHESIS

- Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d = \sum_{j=0}^{d} \theta_j x_j$$

Our data has d-dimension

e.g., {house size, house location, …, year built}

LEAST SQUARES LINEAR REGRESSION

- Cost Function

Summation over the whole dataset

$$Cost(\theta) = \underbrace{\frac{1}{2\times n}}_{\text{Averaging}} \sum_{i=0}^{n} \underbrace{\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2}_{\text{The squared error on a single data point}}$$

- Fit by solving

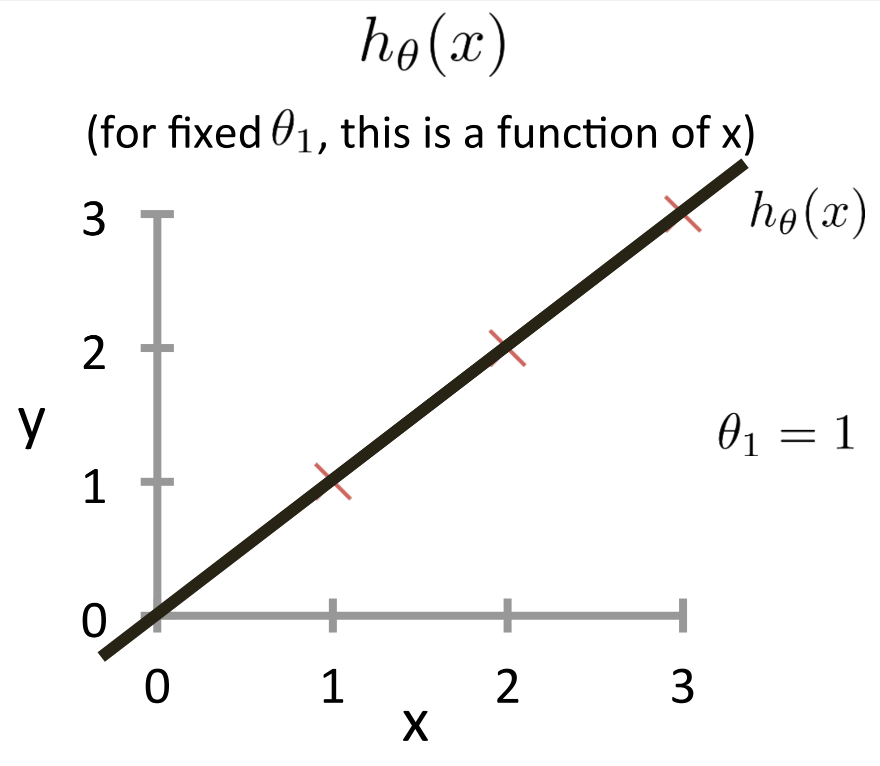
$$\min_{\theta} Cost(\theta)$$

INTUITION BEHIND COST FUNCTION

$$\boldsymbol{h_\theta(x)} = \theta_0 + \theta_1 X$$

$\boldsymbol{h_\theta(x)}$

FOR A FIXED θ_1, THIS IS A FUNCTION OF X

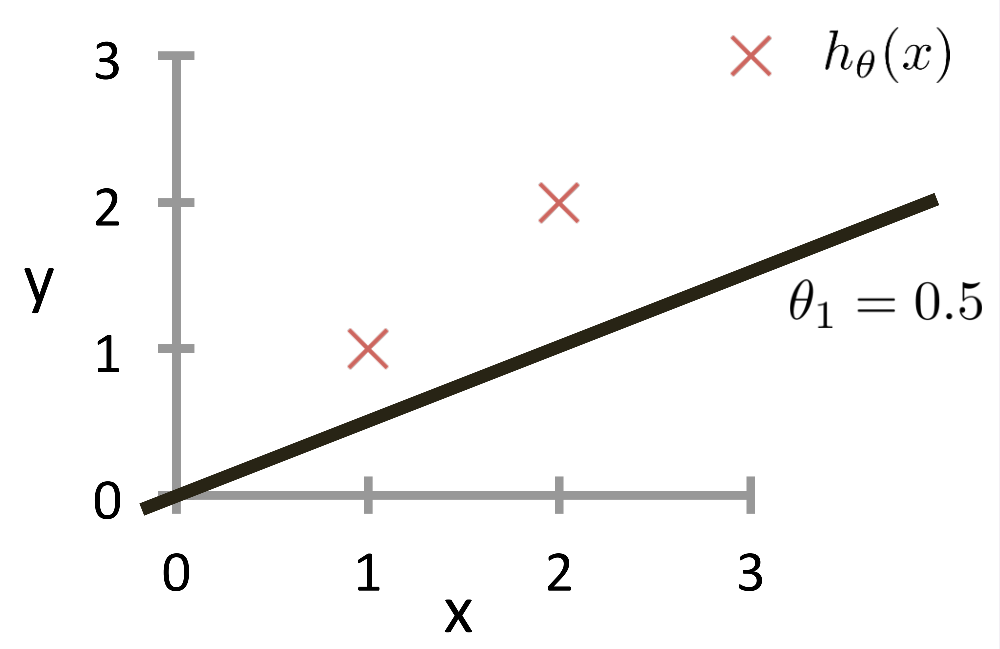


INTUITION BEHIND COST FUNCTION

$$\boldsymbol{h_\theta(x)} = \theta_0 + \theta_1 X$$

$\boldsymbol{h_\theta(x)}$

FOR A FIXED θ_1, THIS IS A FUNCTION OF X



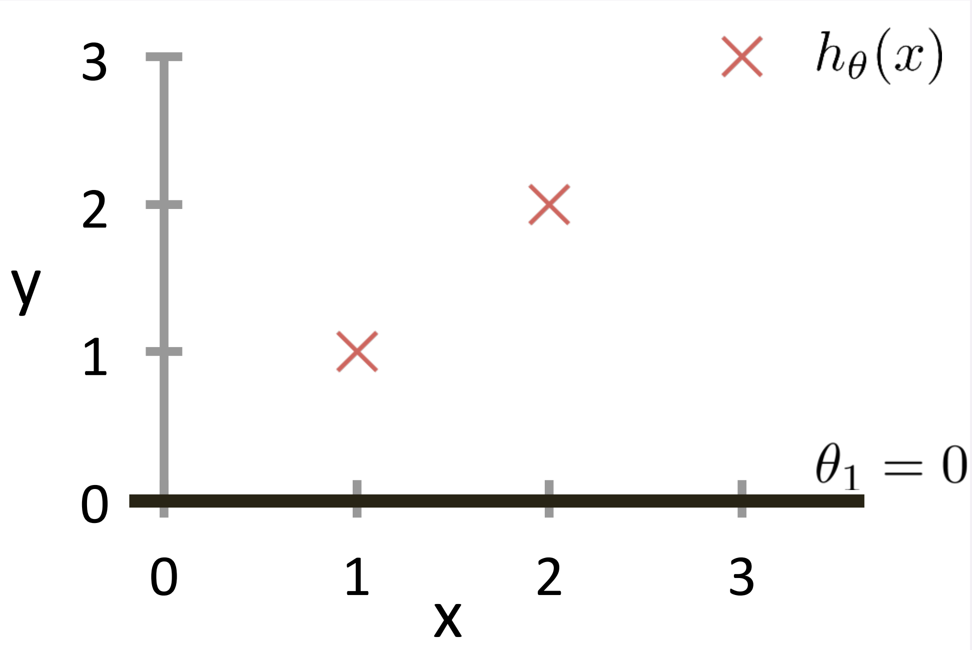
$$Cost(\theta) = Cost([\theta_0, \theta_1]) = Cost([0, 0.5]) = \frac{1}{2\times 3}\left((0.5\ - 1)^2 + (1 - 2)^2 + (1.5\ - 3)^2\right) = 0.58$$

INTUITION BEHIND COST FUNCTION

$$\boldsymbol{h_\theta(x)} = \theta_0 + \theta_1 X$$

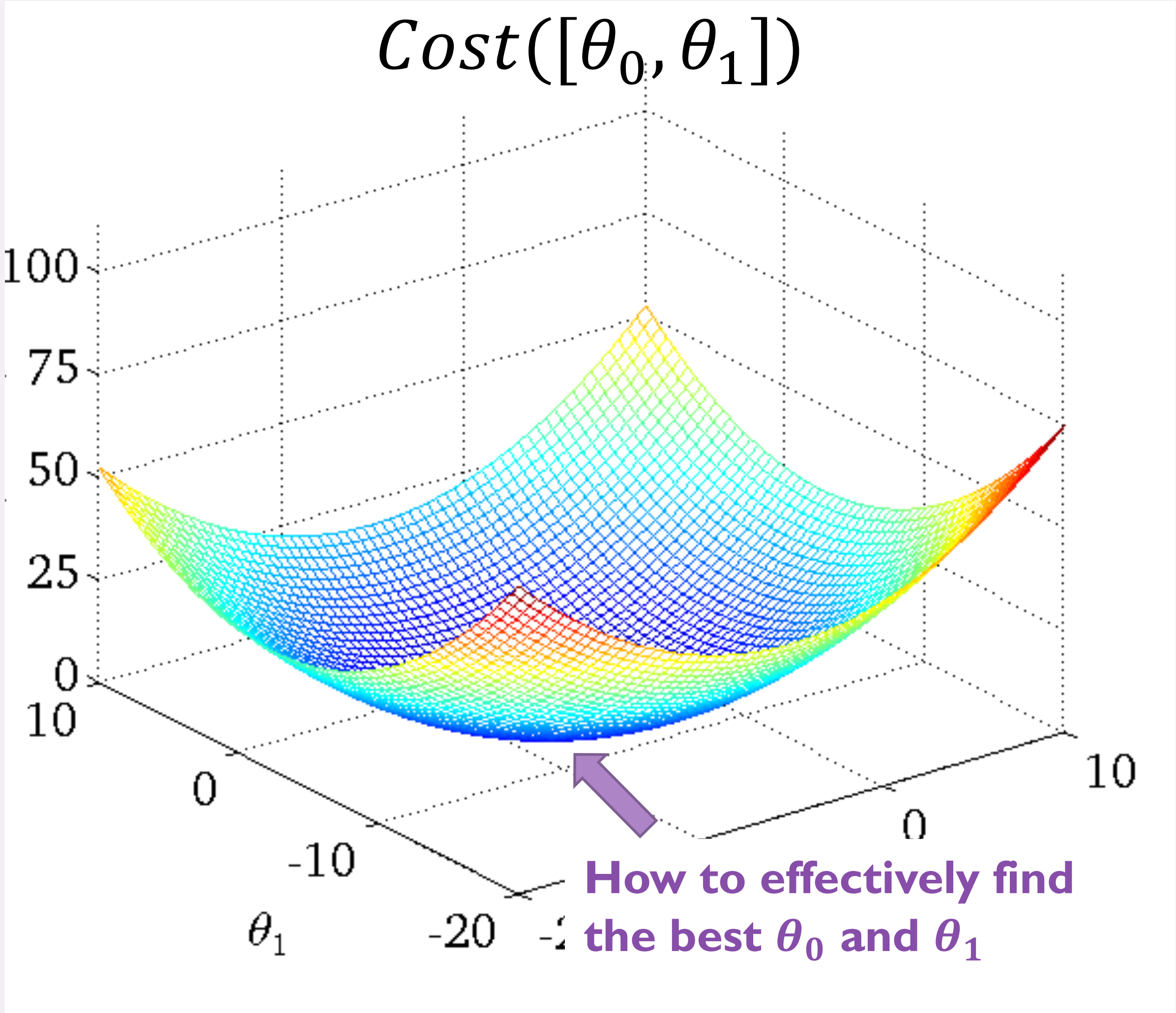
$\boldsymbol{h_\theta(x)}$

FOR A FIXED θ_1, THIS IS A FUNCTION OF X

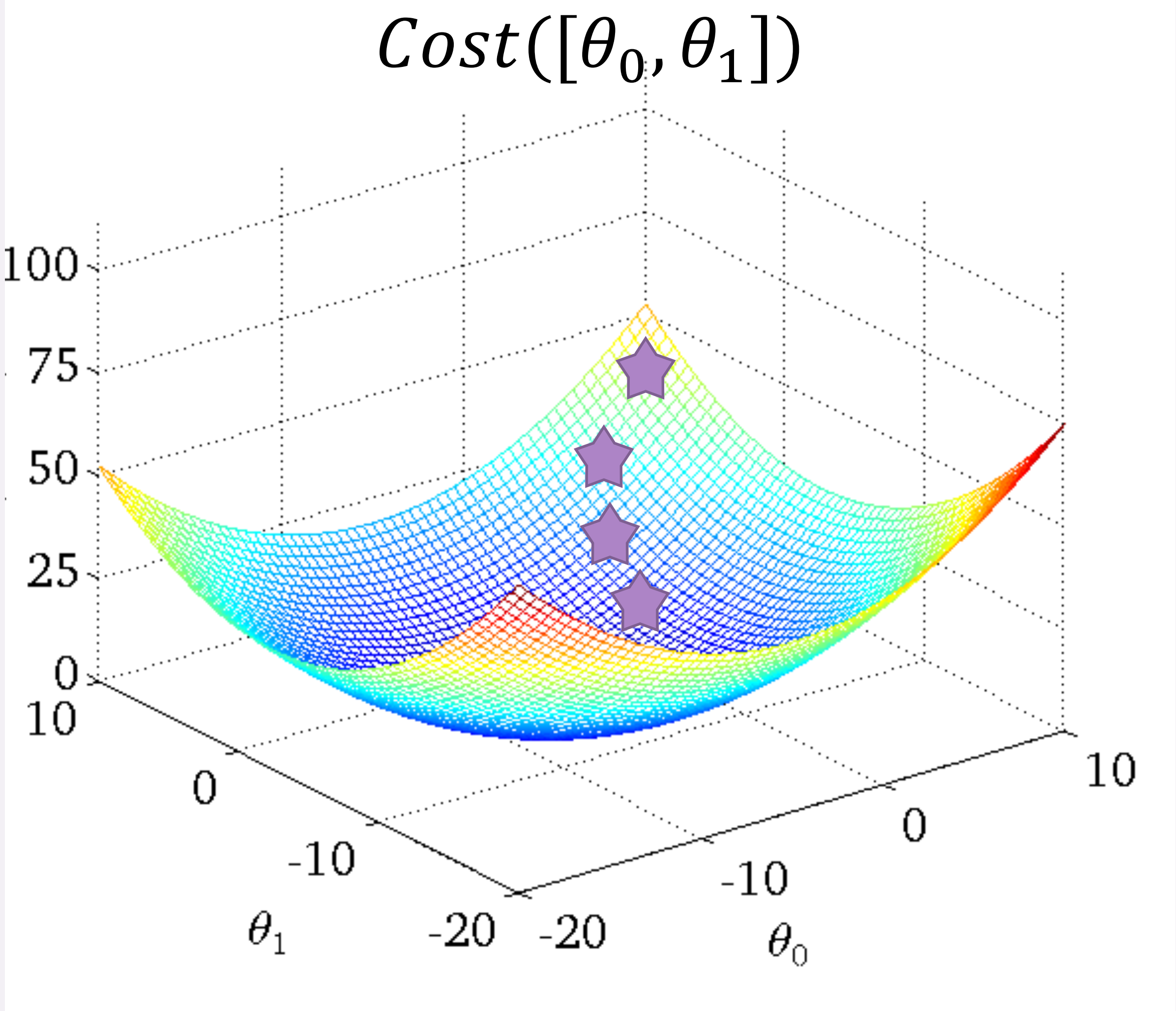


$$Cost(\theta) = Cost([\theta_0, \theta_1]) = Cost([0,0]) = \frac{1}{2\times 3}\left((0-1)^2 + (0-2)^2 + (0\ -3)^2\right) = 2.33$$

INTUITION BEHIND COST FUNCTION



BASIC SEARCH PROCEDURE

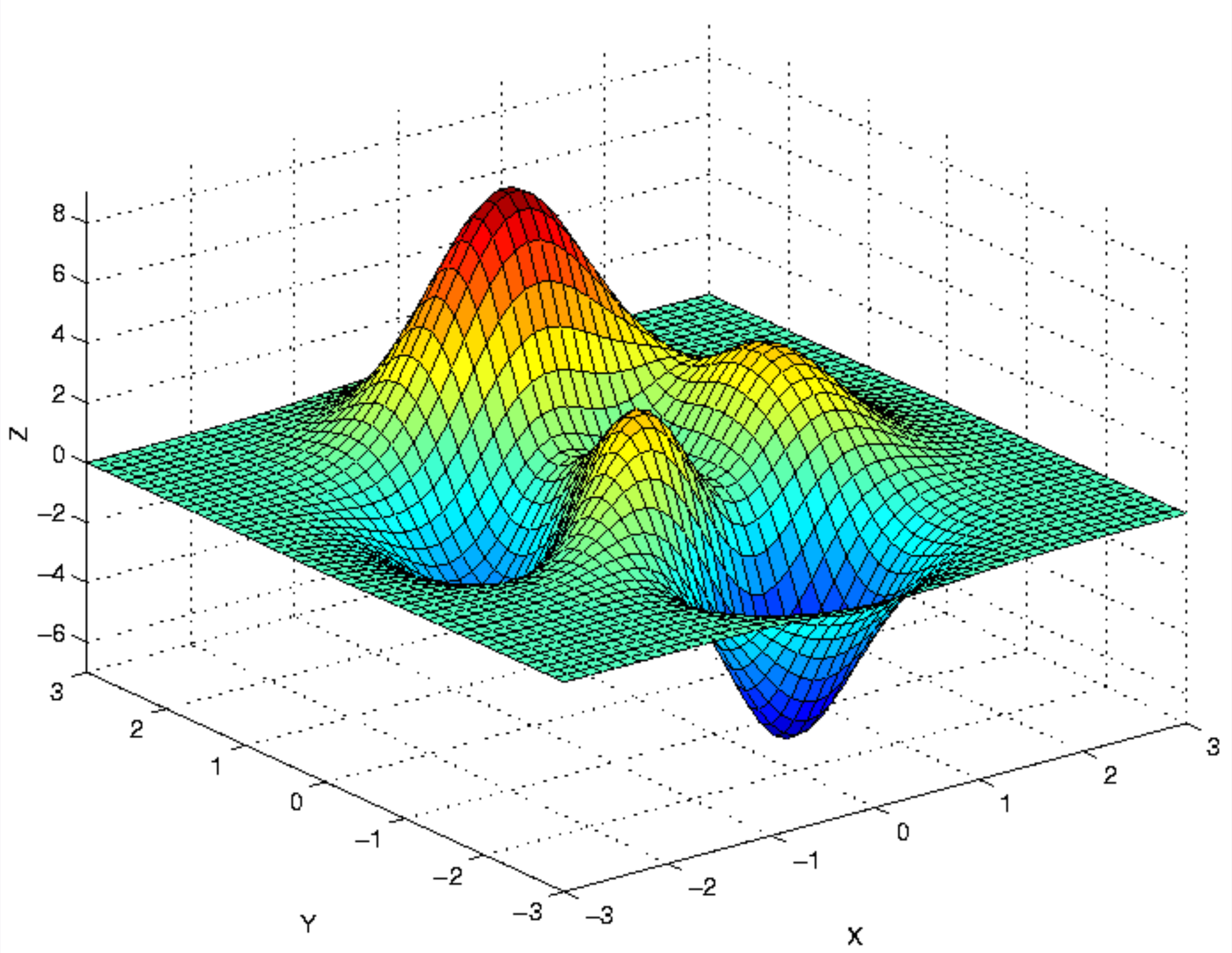


- Choose initial value for θ
- Choose a new value for θ to reduce $Cost\ (\theta_0, \theta_1)$

  -- repeat until we reach a minimum

Since the least squares objective function for linear regression is convex, we don't need to worry about local minima

A TEASER

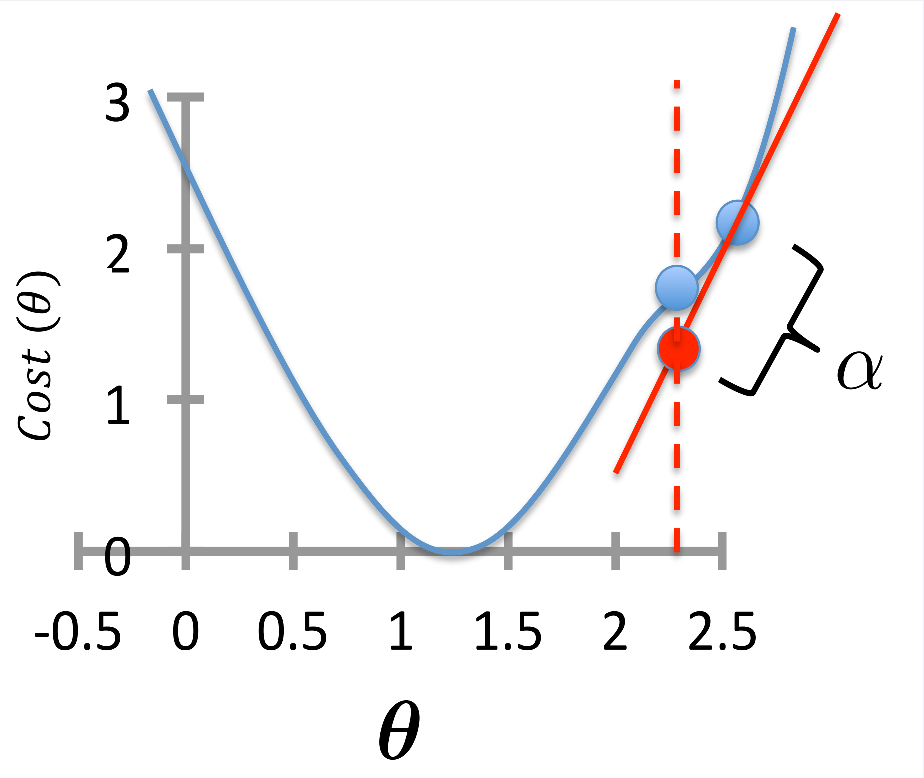


GRADIENT DESCENT

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial Cost(\theta)}{\partial \theta_j}$$

(simultaneous update for $\theta_0, \theta_1, \dots, \theta_d$)

Learning rate



GRADIENT DESCENT

- Initialize θ
- Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial Cost(\theta)}{\partial \theta_j} \qquad \text{(simultaneous update for } \theta_0, \theta_1, \dots, \theta_d\text{)}$$

- For linear regression:

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

With $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \dots + \theta_d x_d = \ \sum_{j=0}^{d} \theta_j x_j$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \boxed{\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2}$$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \quad \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \quad \boxed{\frac{\partial}{\partial \theta_j} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2}$$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \qquad \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \qquad \frac{\partial}{\partial \theta_j} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \qquad \boxed{2\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \frac{\partial}{\partial \theta_j} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)}$$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \quad \frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \quad \frac{\partial}{\partial \theta_j}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \quad 2\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$= \boxed{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{d}\theta_k x_k^{(i)} - y^{(i)}\right)} \quad \boxed{\frac{\partial}{\partial \theta_j}\left(\sum_{k=0}^{d}\theta_k x_k^{(i)} - y^{(i)}\right)}$$

GRADIENT DESCENT

$$\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \frac{\partial}{\partial \theta_j}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

$$= \frac{1}{2n} \sum_{i=1}^{n} 2\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)}\right) \frac{\partial}{\partial \theta_j}\left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)}\right)$$

$$= \boxed{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)}\right) x_j^{(i)}}$$

GRADIENT DESCENT

Only one data point: n=1. & This data only has one dimension: d=1

$$\frac{\partial Cost(\theta)}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}\frac{1}{2}(\theta_0 + \theta_1 x \; - y)^2$$

$$= \frac{\partial}{\partial \theta_0}\frac{1}{2}[{\theta_0}^2 + 2\theta_0(\theta_1 x - y) + (\theta_1 x - y)^2]$$

$$= \theta_0 + (\theta_1 x - y)$$

Recall $\frac{\partial Cost(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(\sum_{k=0}^{d}\theta_k x_k^{(i)} - y^{(i)})\, x_j^{(i)}$

With n=1, j=0 $\Rightarrow (\sum_{k=0}^{d}\theta_k x_k \; - y \quad)\, x_0$

With $d = 1 \Rightarrow (\theta_0 x_0 \; + \theta_1 x_1 \quad - y)x_0$

Recall $x_0 \; = 1 \Rightarrow (\theta_0 + \theta_1 x_1 \; - y)$

CONTOUR PLOT

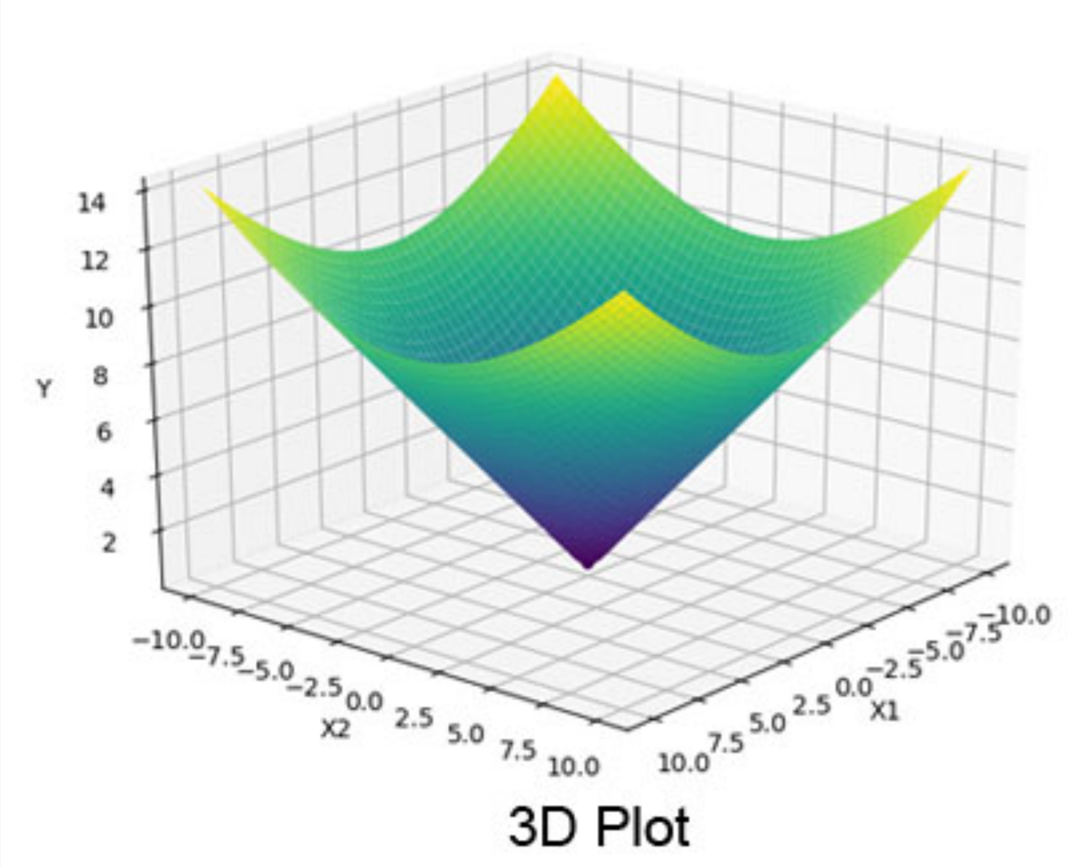
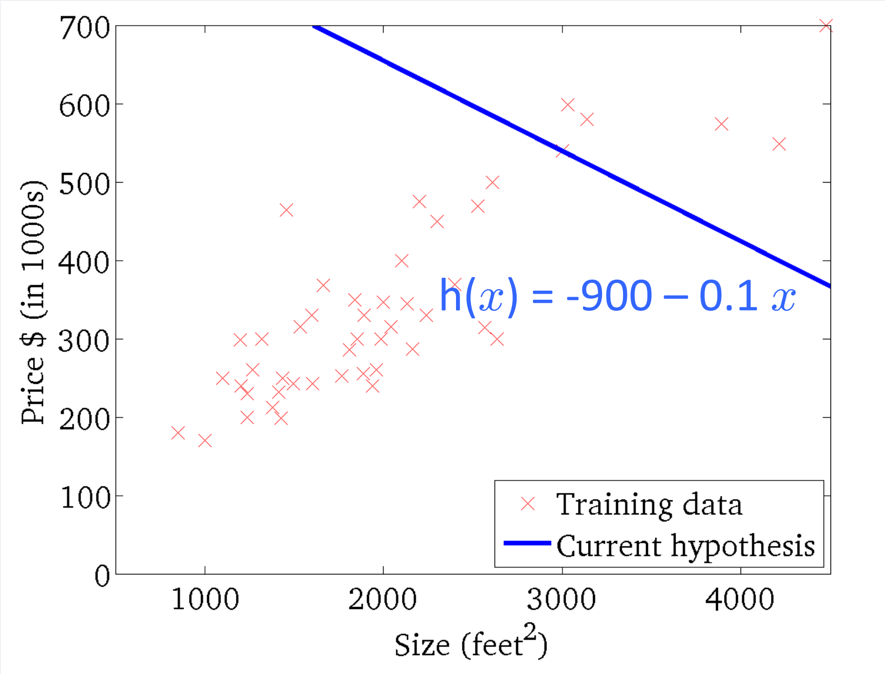


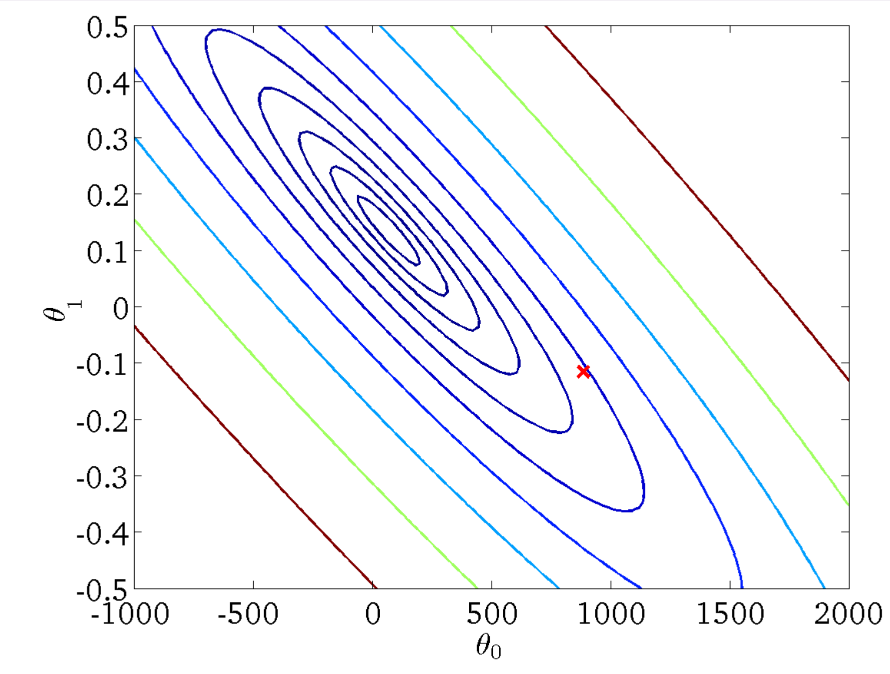
Image credit: http://www.adeveloperdiary.com/data-science/how-to-visualize-gradient-descent-using-contour-plot-in-python/

EXAMPLE

$$h_\theta(x)$$

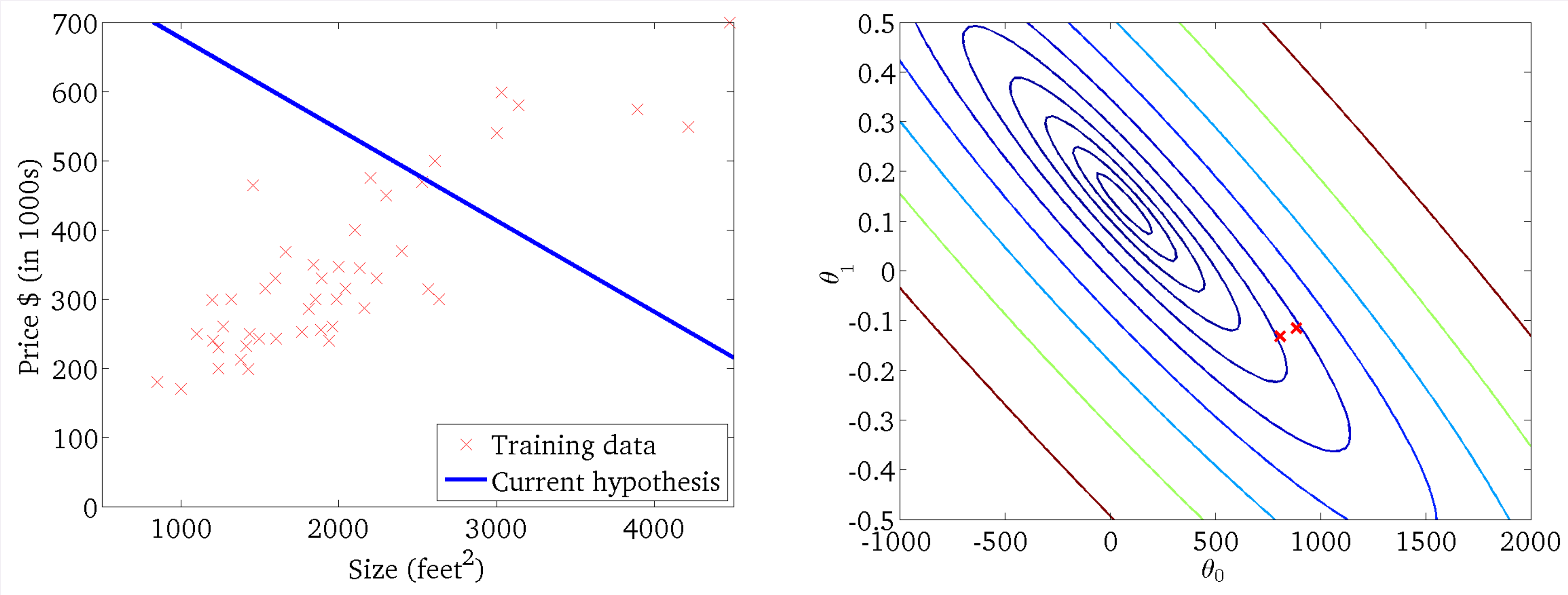


$$Cost([\theta_0, \theta_1])$$



EXAMPLE

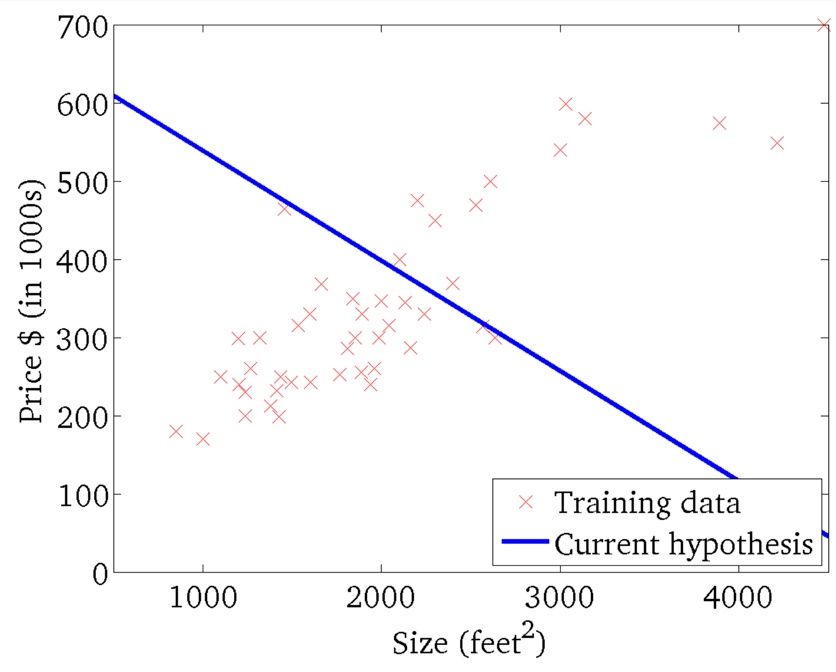
$h_\theta(x)$ $\quad$ $Cost([\theta_0, \theta_1])$

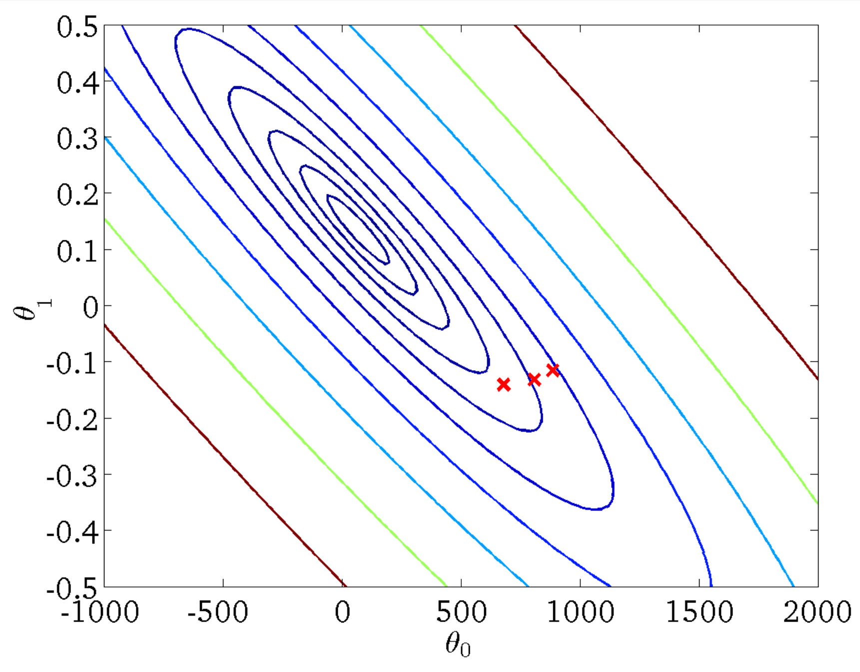


EXAMPLE

$h_\theta(x)$

$Cost([\theta_0, \theta_1])$

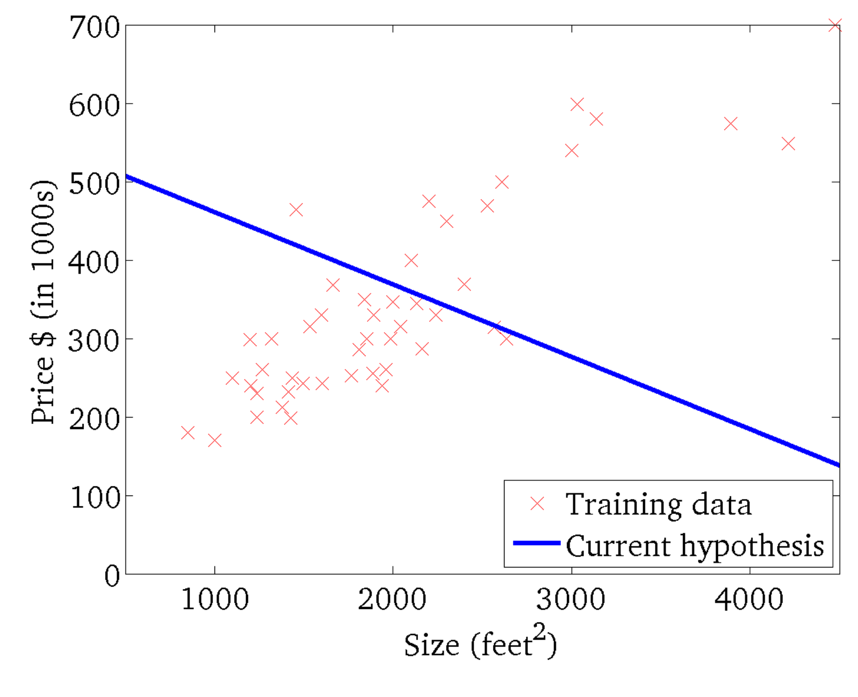


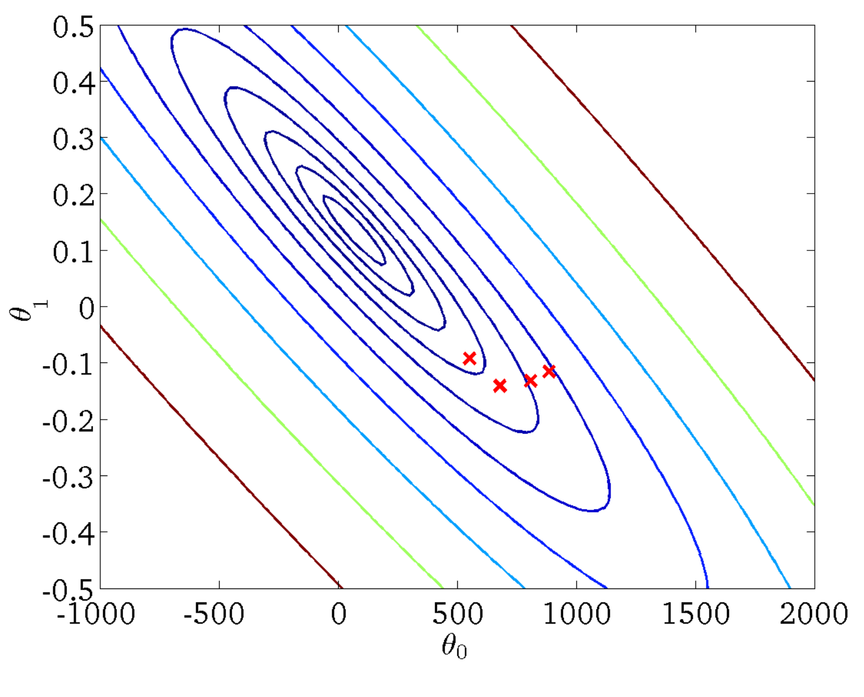


EXAMPLE

$h_\theta(x)$

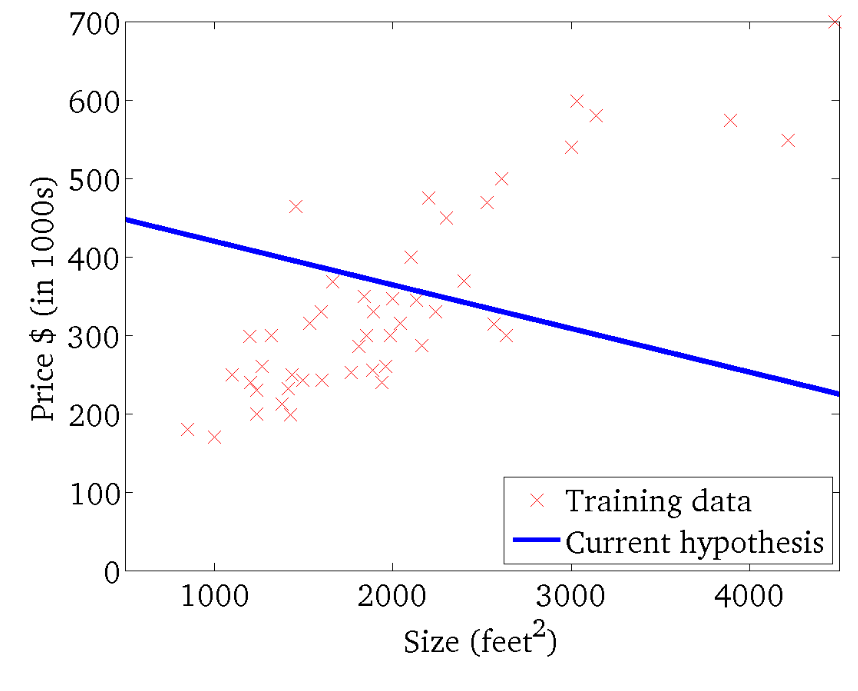
$Cost([\theta_0, \theta_1])$



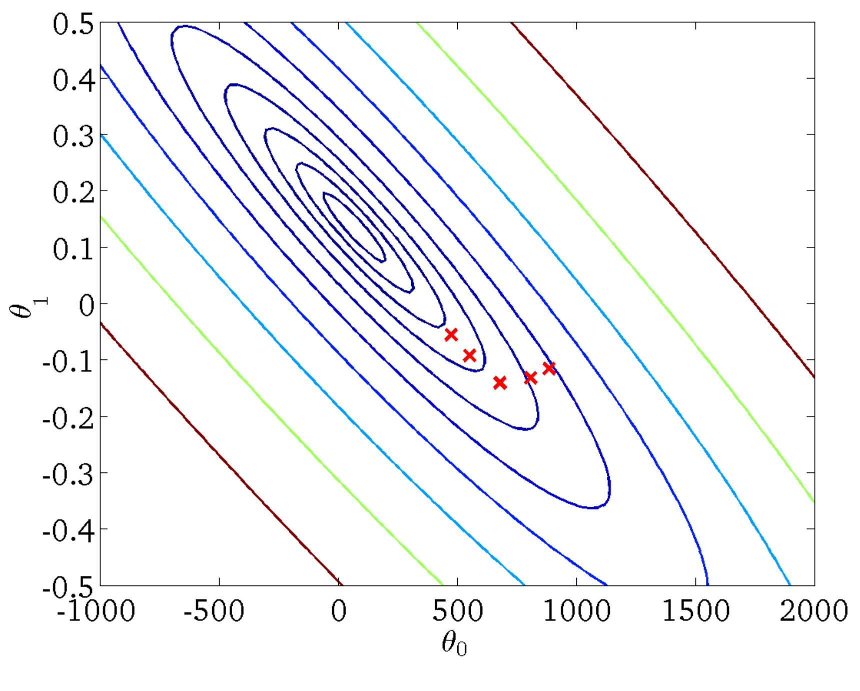


EXAMPLE

$h_\theta(x)$

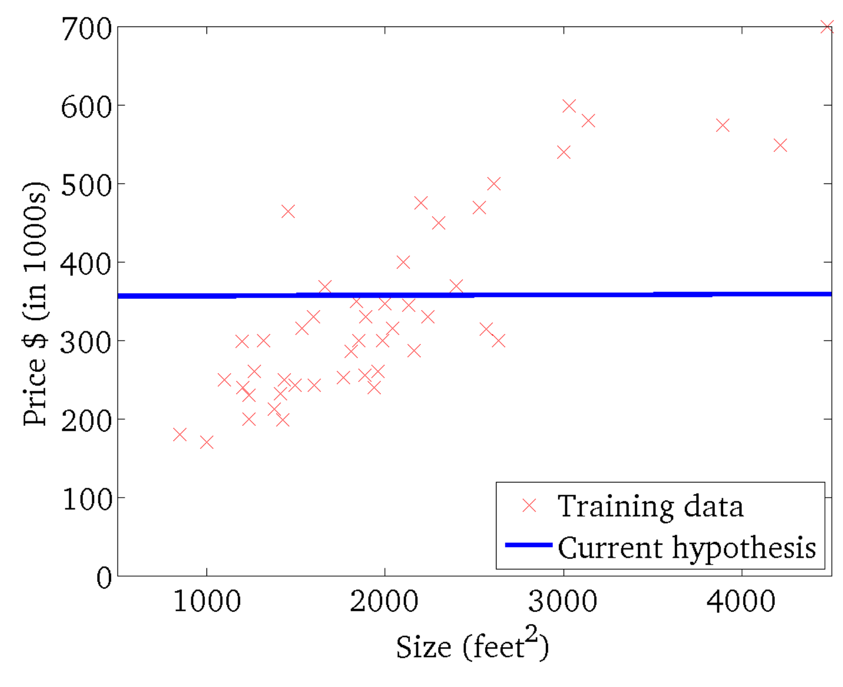


$Cost([\theta_0, \theta_1])$

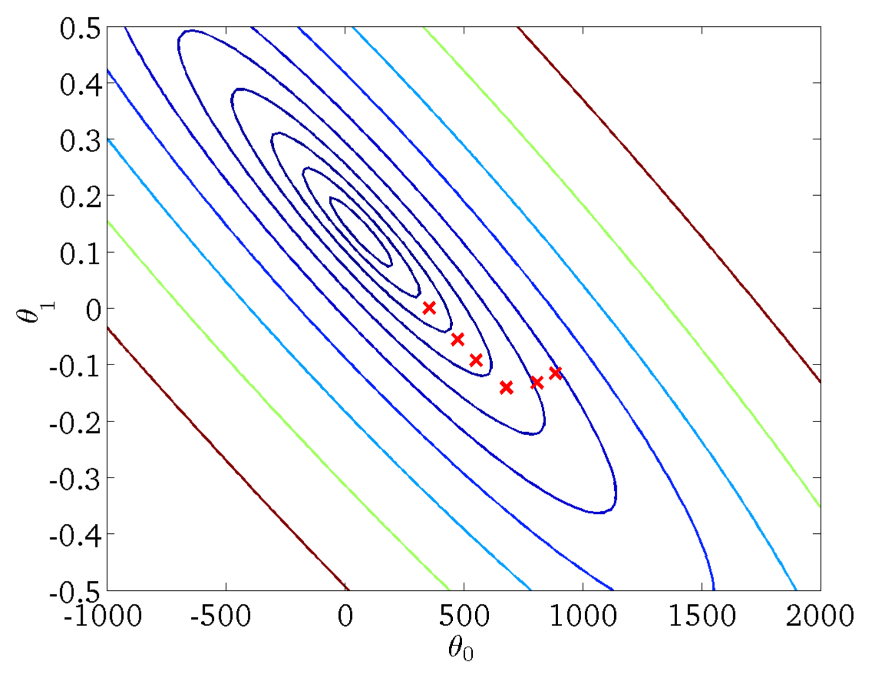


EXAMPLE

$h_\theta(x)$

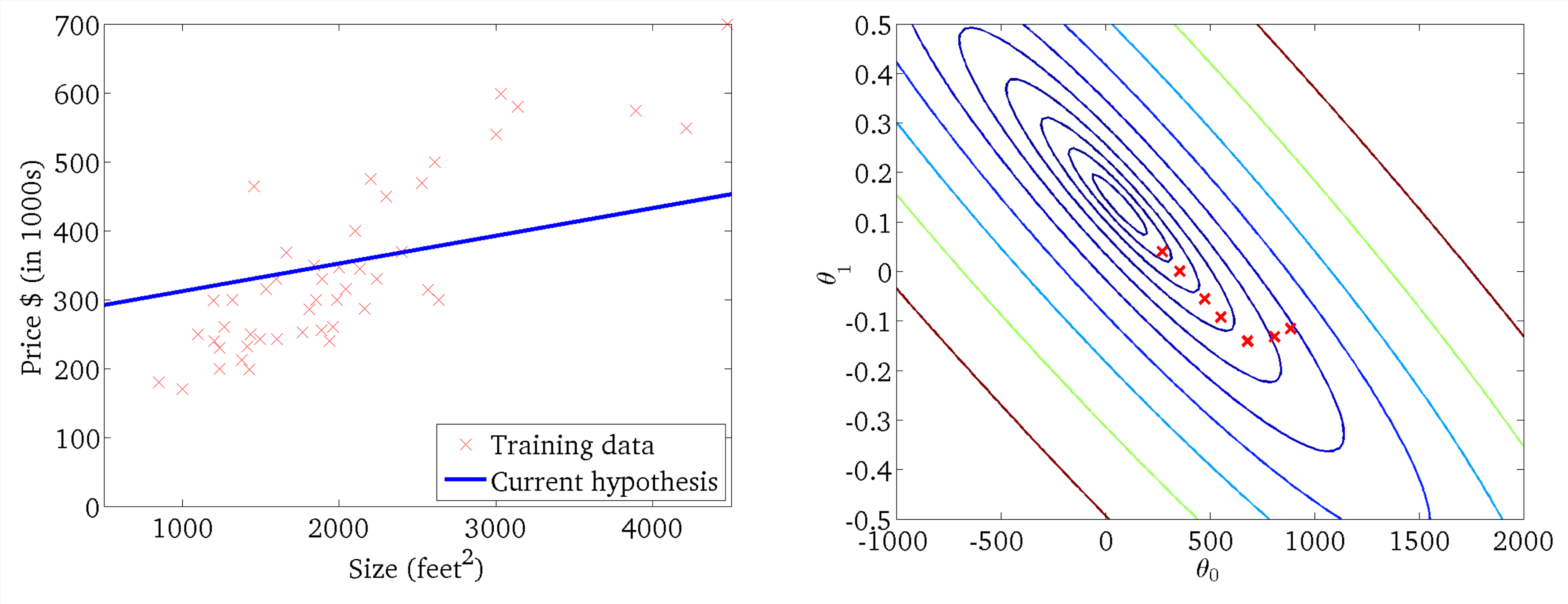


$Cost([\theta_0, \theta_1])$



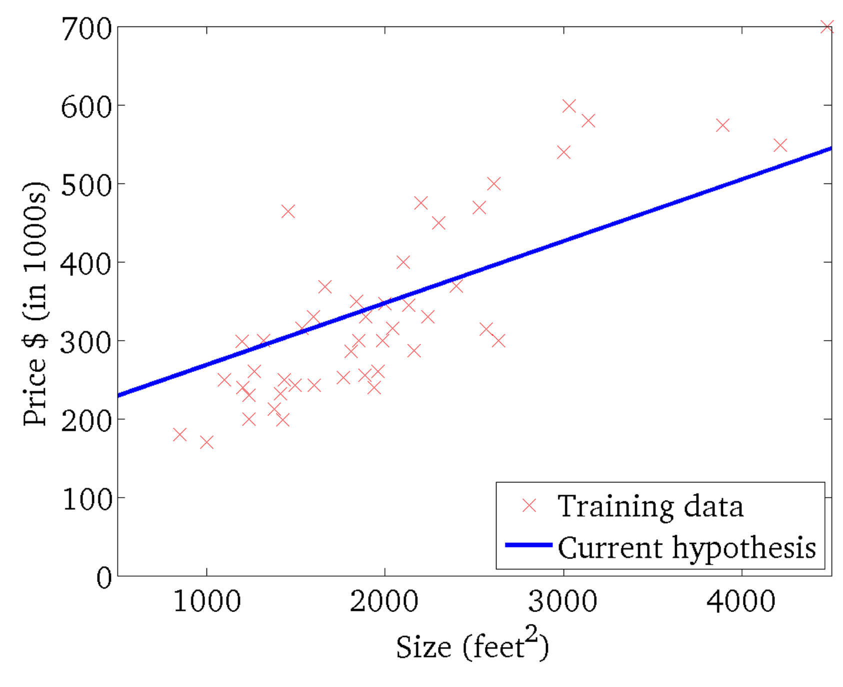
EXAMPLE

$h_\theta(x)$ $Cost([\theta_0, \theta_1])$

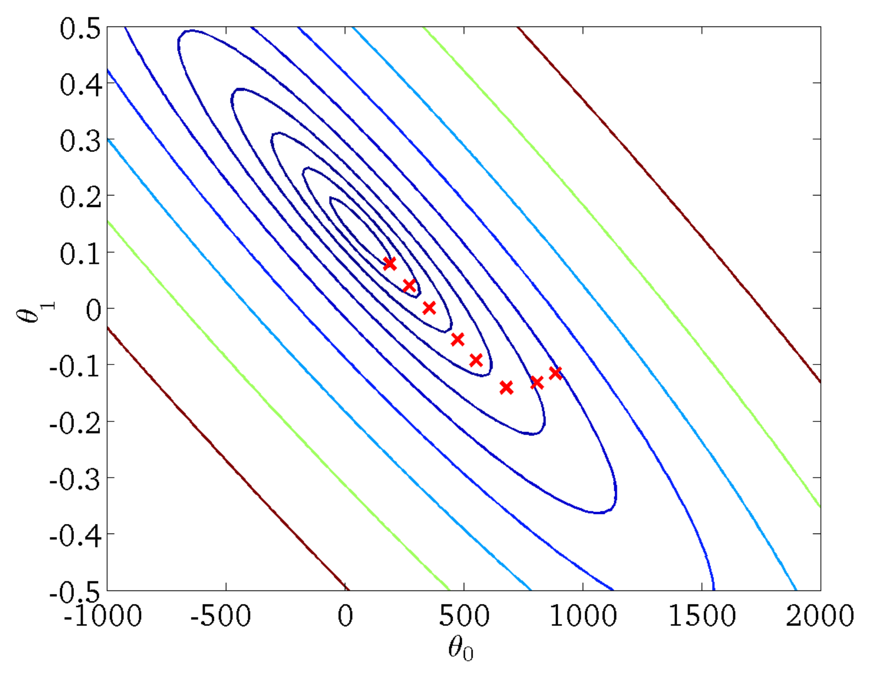


EXAMPLE

$h_\theta(x)$

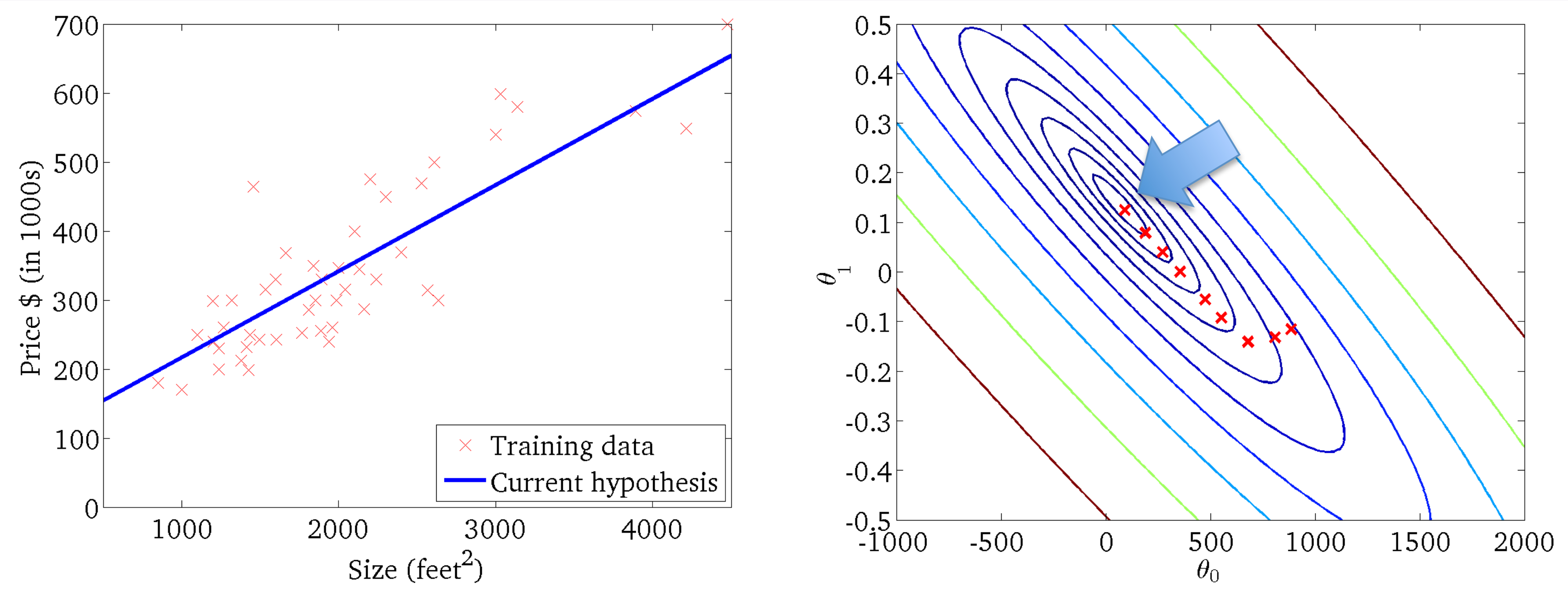


$Cost([\theta_0, \theta_1])$



EXAMPLE

$h_\theta(x)$ $\qquad\qquad$ $Cost([\theta_0, \theta_1])$



EXERCISE

HTTPS://DRIVE.GOOGLE.COM/FILE/D/1EUSN4VPISA31P0_XVLNNL8J89W6WAZNG/VIEW?USP=SHARING

QUESTIONS?