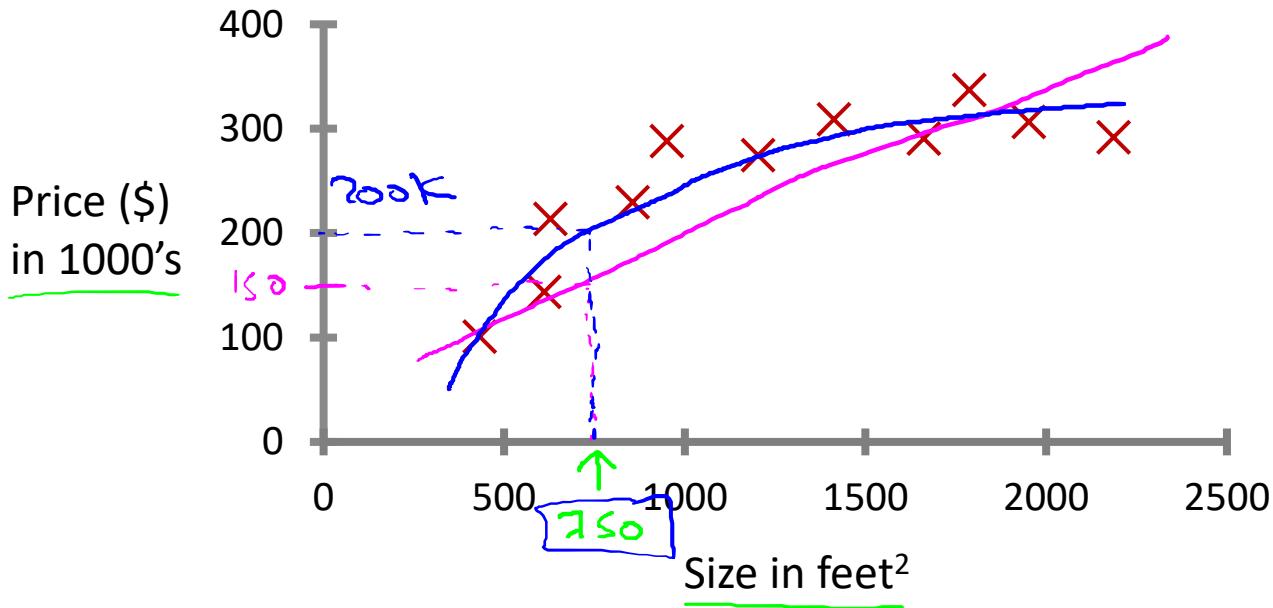


Chapter 3: Supervised Machine Learning

Artificial Intelligence and Machine Learning

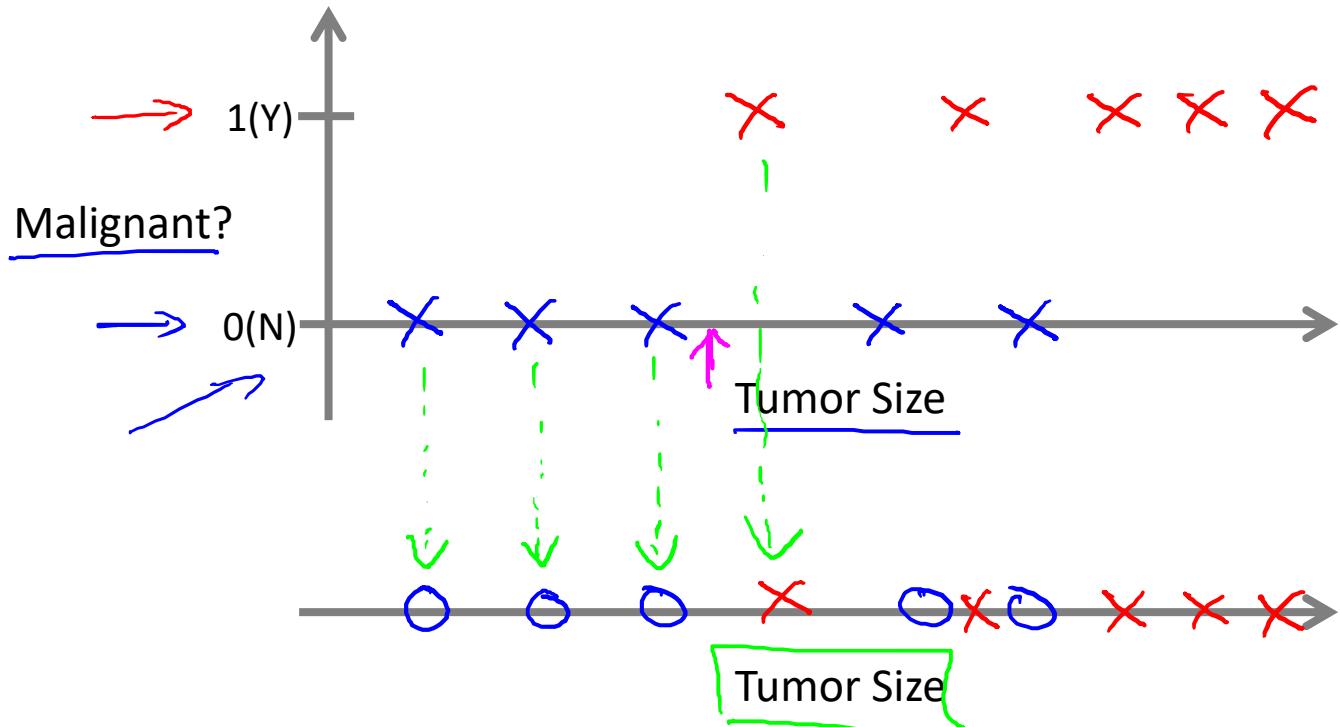
Housing price prediction.



Supervised Learning
'right answers' given

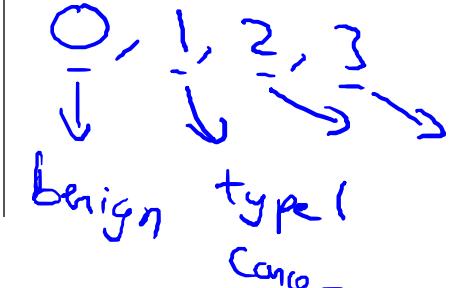
Regression: Predict continuous valued output (price)

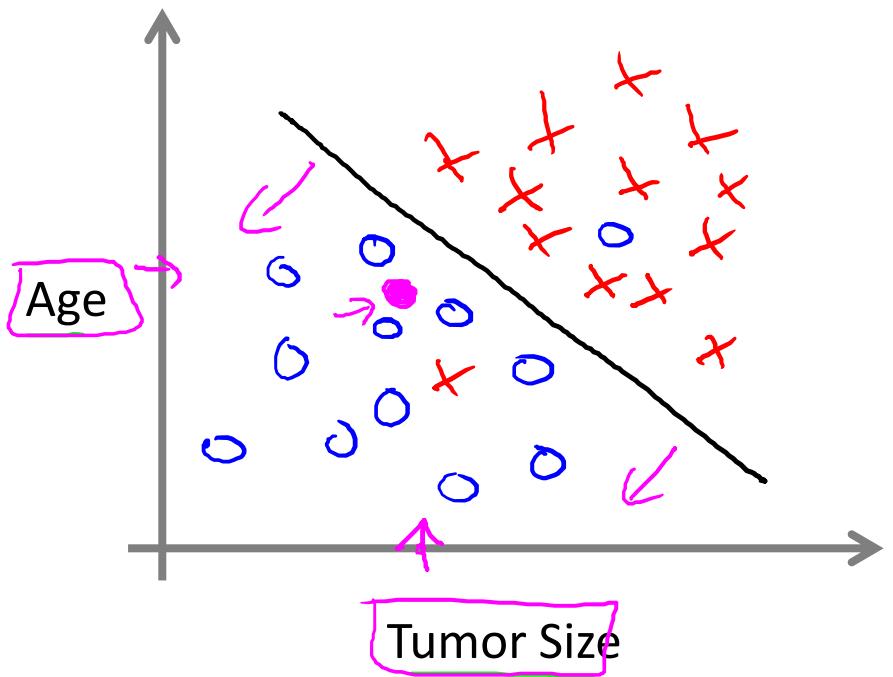
Breast cancer (malignant, benign)



Classification

Discrete valued output (0 or 1)





- Clump Thickness
 - Uniformity of Cell Size
 - Uniformity of Cell Shape
- ...

You're running a company, and you want to develop learning algorithms to address each of two problems.

1000's

→ Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

→ Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

→ 0 - not hacked
→ 1 - hacked

Should you treat these as classification or as regression problems?

Treat both as classification problems.

Treat problem 1 as a classification problem, problem 2 as a regression problem.

→ Treat problem 1 as a regression problem, problem 2 as a classification problem.

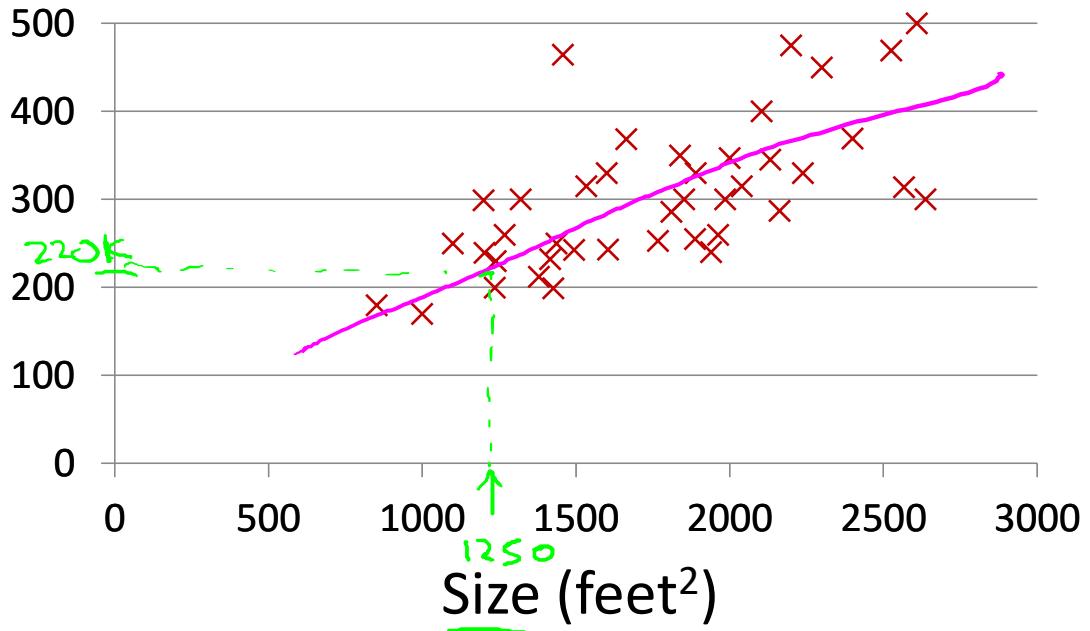
Treat both as regression problems.

Linear regression with one variable

Model representation

Housing Prices (Portland, OR)

Price
(in 1000s
of dollars)



Supervised Learning

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

Regression Problem

Predict real-valued output

Classification: Discrete-valued output

Training set of housing prices (Portland, OR)

Size in feet ² (x)	Price (\$) in 1000's (y)
→ 2104	460
→ 1416	232
→ 1534	315
852	178
...	...
\vdots	\vdots

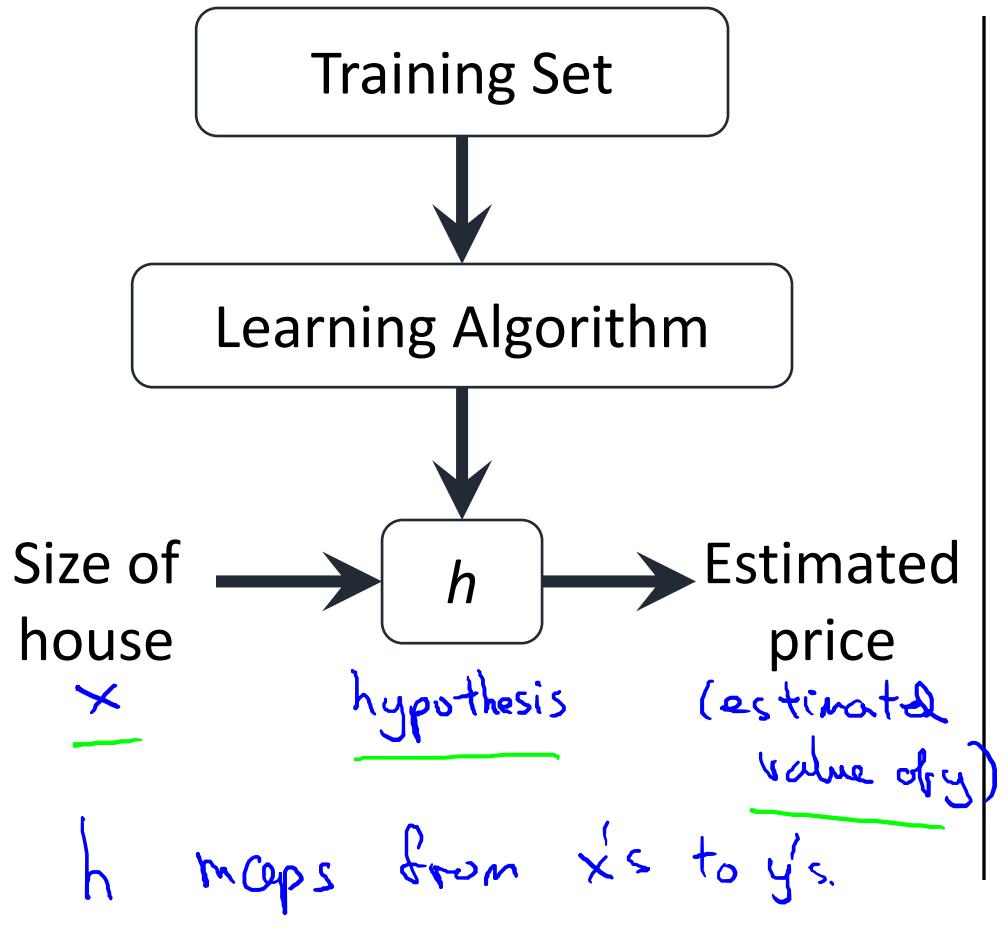
Notation:

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

(x, y) - one training example

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

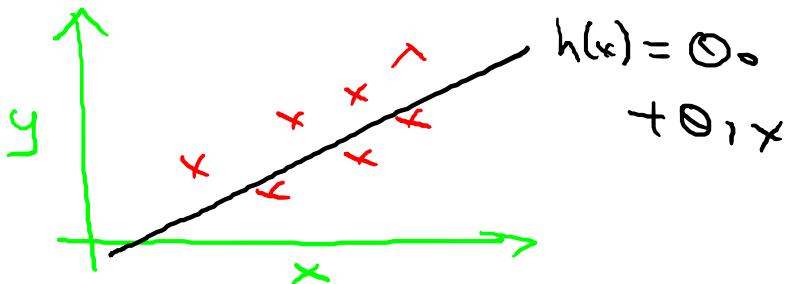
$$\begin{cases} x^{(1)} = 2104 \\ x^{(2)} = 1416 \\ \vdots \\ y^{(1)} = 460 \end{cases}$$



How do we represent h ?

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

Shorthand: $h(x)$



Linear regression with one variable.
 Univariate linear regression.

one variable

Linear regression with one variable

Cost function

Training Set

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

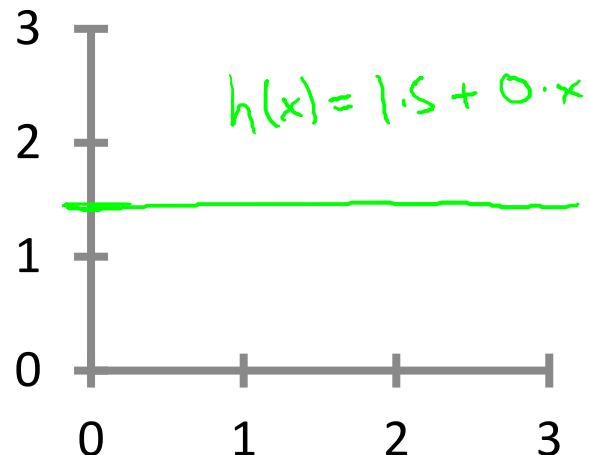
$m = 47$

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

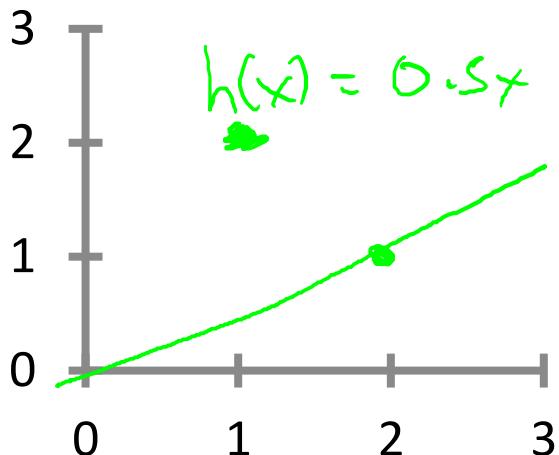
θ_i 's: Parameters

How to choose θ_i 's ?

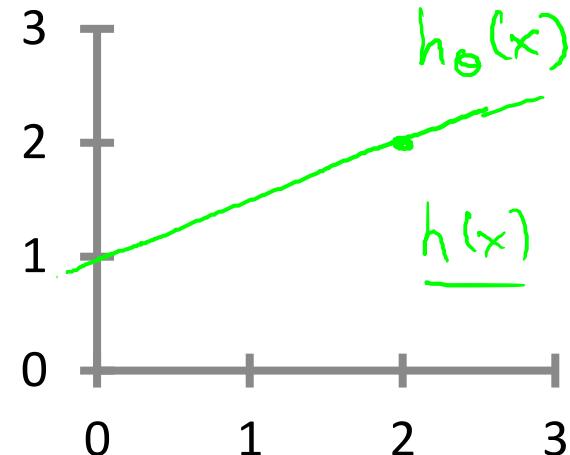
$$\underline{h_{\theta}(x) = \theta_0 + \theta_1 x}$$



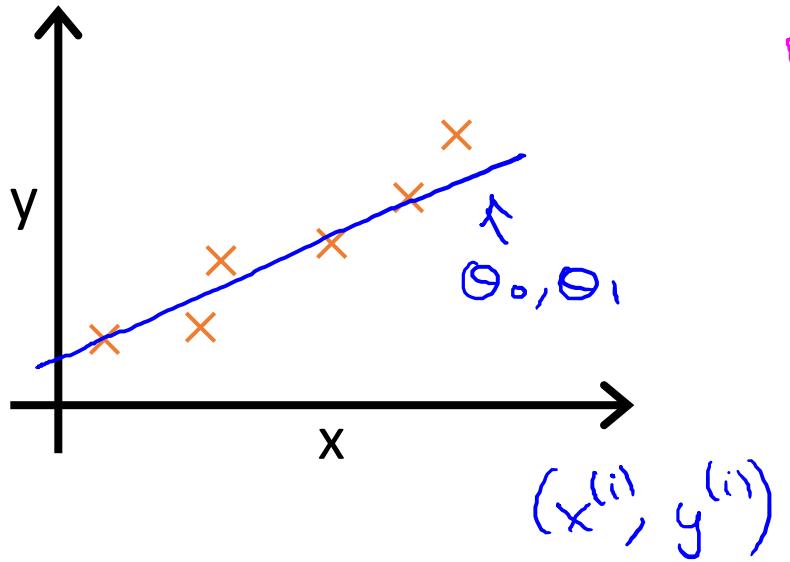
$$\rightarrow \theta_0 = 1.5$$
$$\rightarrow \theta_1 = 0$$



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



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



Idea: Choose $\underline{\theta_0}, \underline{\theta_1}$ so that
 $\underline{h_\theta(x)}$ is close to y for our
training examples $\underline{(x, y)}$

x, y

minimize $\underline{\theta_0, \theta_1}$

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

↑

$h_\theta(\underline{x^{(i)}}) = \underline{\theta_0 + \theta_1 x^{(i)}}$

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

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

Cost function

Squared error function

Linear regression with one variable

Cost function intuition I

Hypothesis:

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

Parameters:

$$\underline{\theta_0, \theta_1}$$



Cost Function:

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

Goal: minimize $J(\theta_0, \theta_1)$

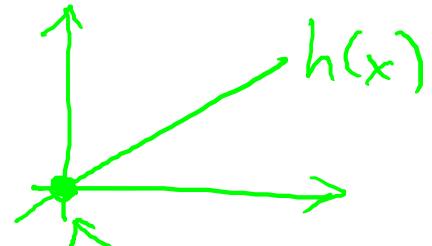
$$\nearrow \theta_0, \theta_1$$

Simplified

$$h_{\theta}(x) = \underline{\theta_1 x}$$

$$\underline{\theta_0} = 0$$

$$\underline{\theta_1}$$



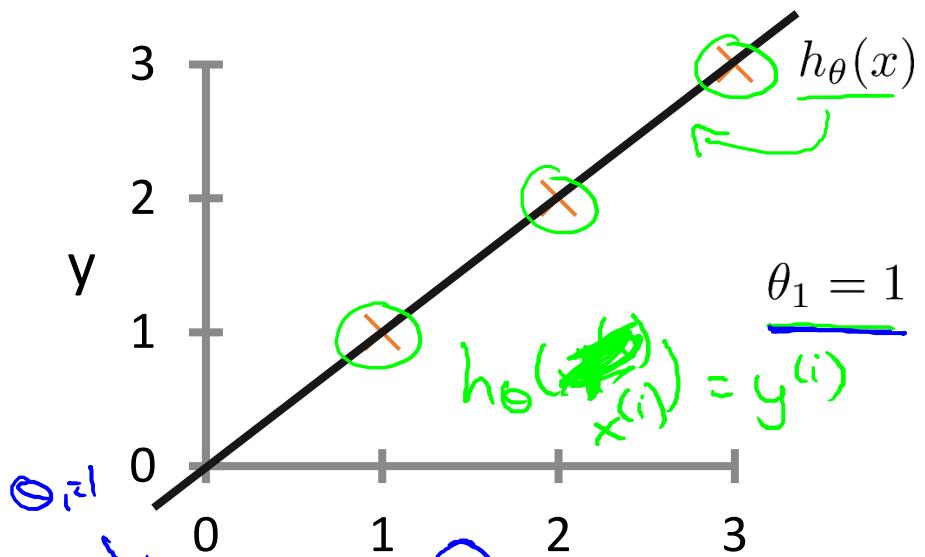
$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^m (\underline{h_{\theta}(x^{(i)}) - y^{(i)}})^2$$

minimize $\underline{\theta_1}$ $J(\theta_1)$

$$\underline{\theta_0, x^{(i)}}$$

→ $h_\theta(x)$

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

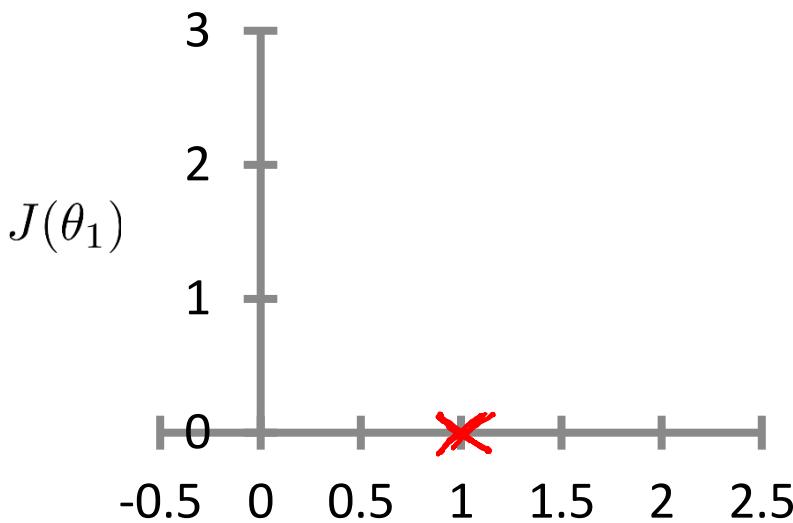


$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{2m} \sum_{i=1}^m (\theta_0 x^{(i)} - y^{(i)})^2 = \frac{1}{2m} (0^2 + 0^2 + 0^2) = 0^2$$

→ $J(\theta_1)$

(function of the parameter θ_1)

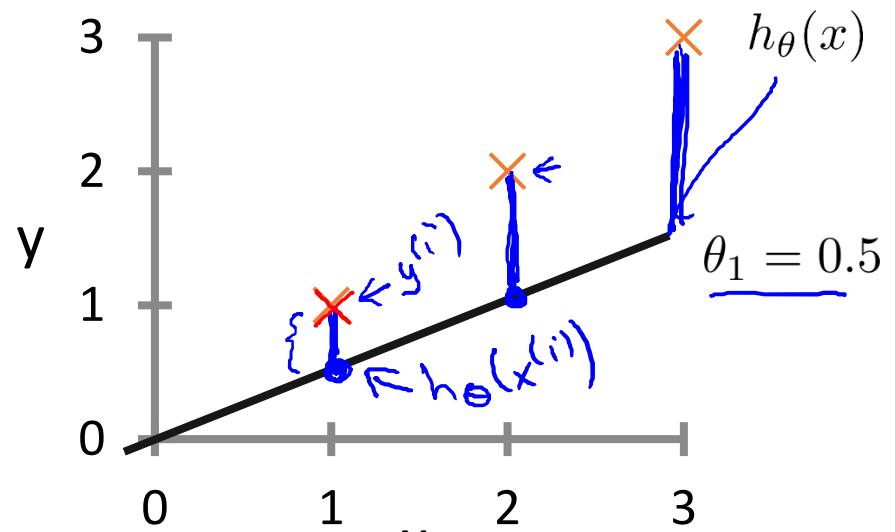


$$\theta_1 = 0.5?$$

$$J(1) = 0$$

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

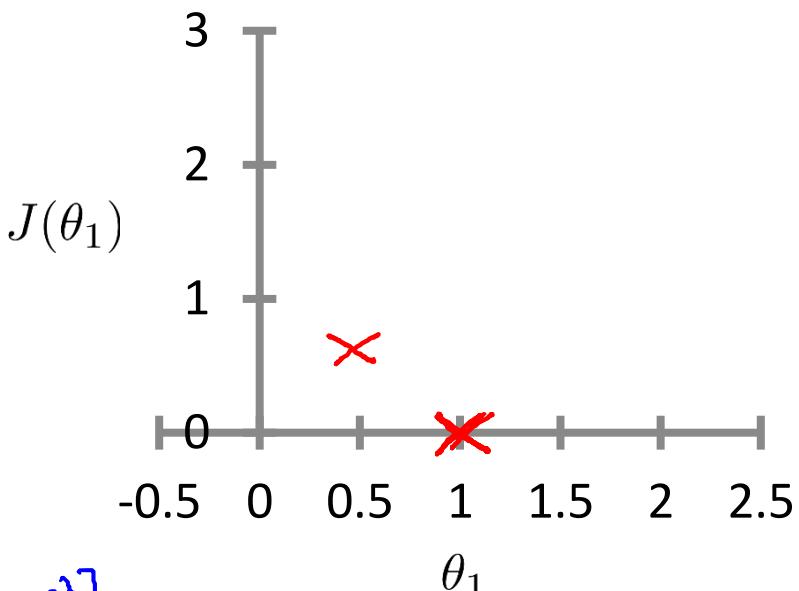
(for fixed θ_1 , this is a function of x)



$$\begin{aligned}
 J(0.5) &= \frac{1}{2m} [(0.5-1)^2 + (1-2)^2 + (1.5-3)^2] \\
 &= \frac{1}{2 \times 3} (3.5) = \frac{3.5}{6} \approx \underline{0.58}
 \end{aligned}$$

$$J(\theta_1)$$

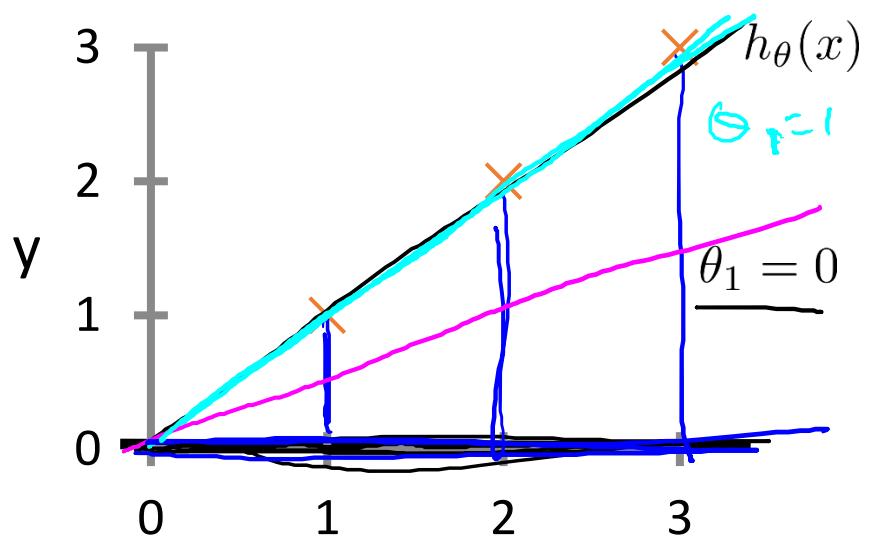
(function of the parameter θ_1)



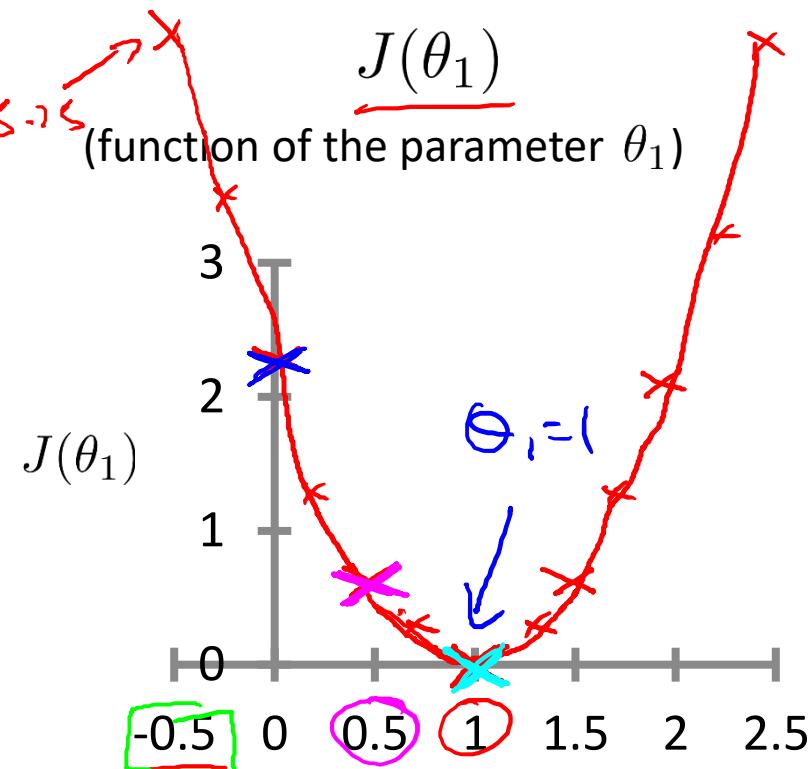
$$\begin{aligned}
 \theta_1 &=? \\
 J(0) &=?
 \end{aligned}$$

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

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



$$\begin{aligned} J(0) &= \frac{1}{2m} (1^2 + 2^2 + 3^2) \\ &= \frac{1}{6} \cdot 14 \approx 2.3 \end{aligned}$$



Linear regression with one variable

Cost function intuition II

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

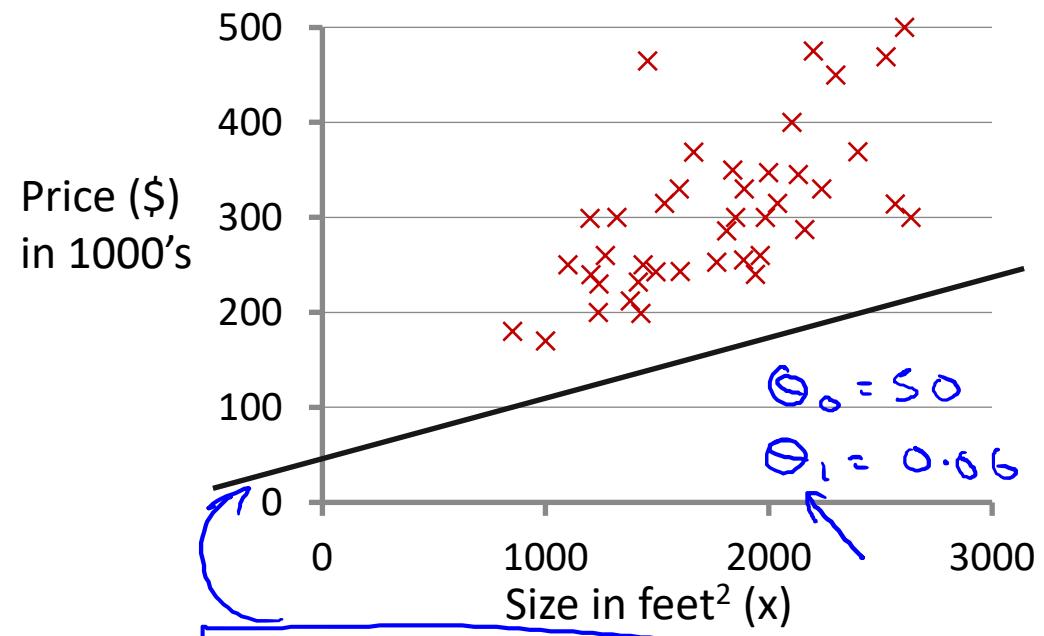
Parameters: θ_0, θ_1

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

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

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

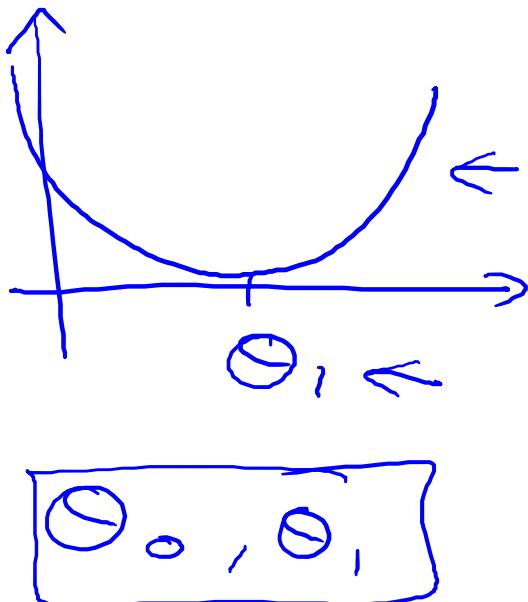
(for fixed θ_0, θ_1 , this is a function of x)

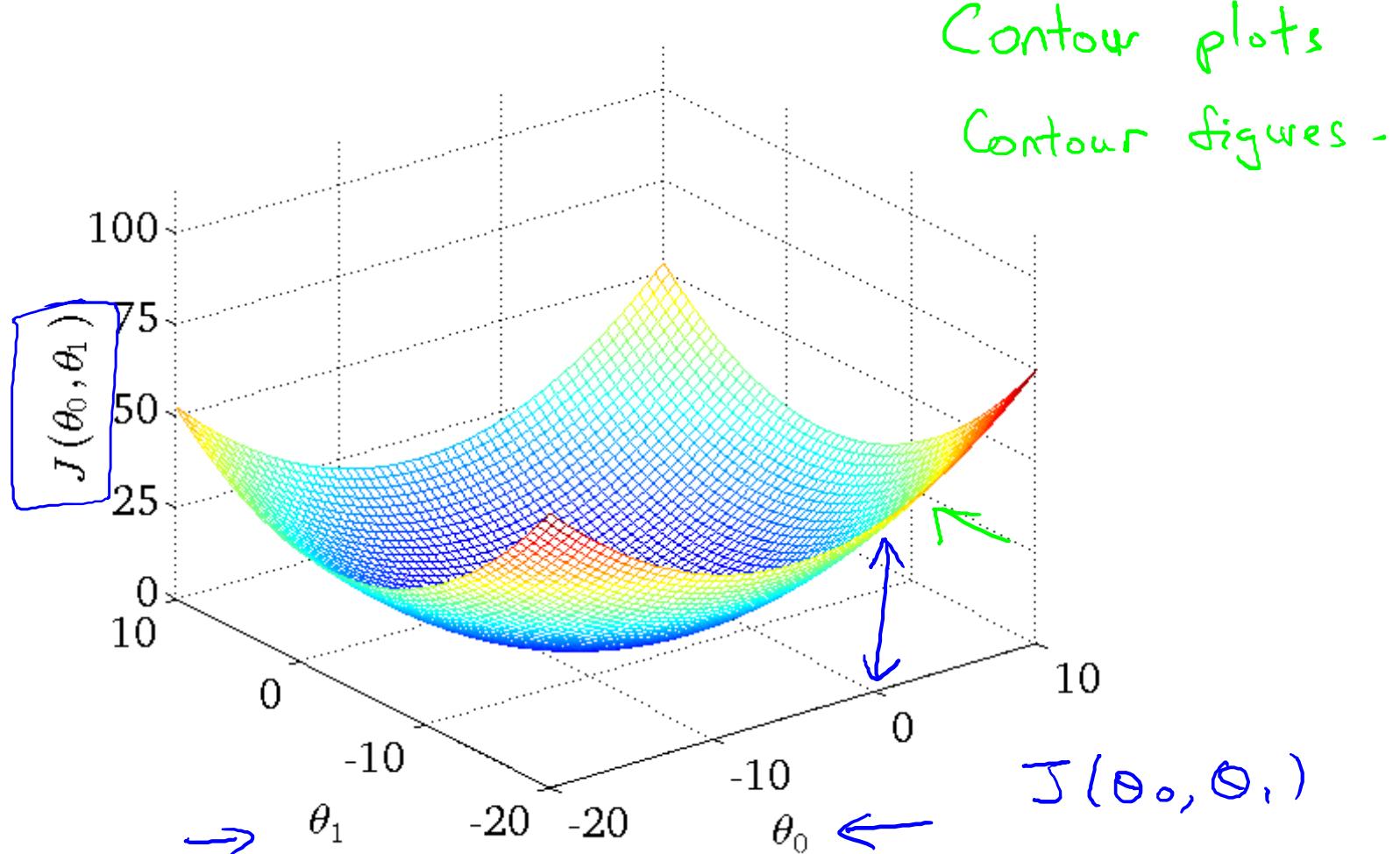


$$h_{\theta}(x) = 50 + 0.06x$$

$$J(\theta_0, \theta_1)$$

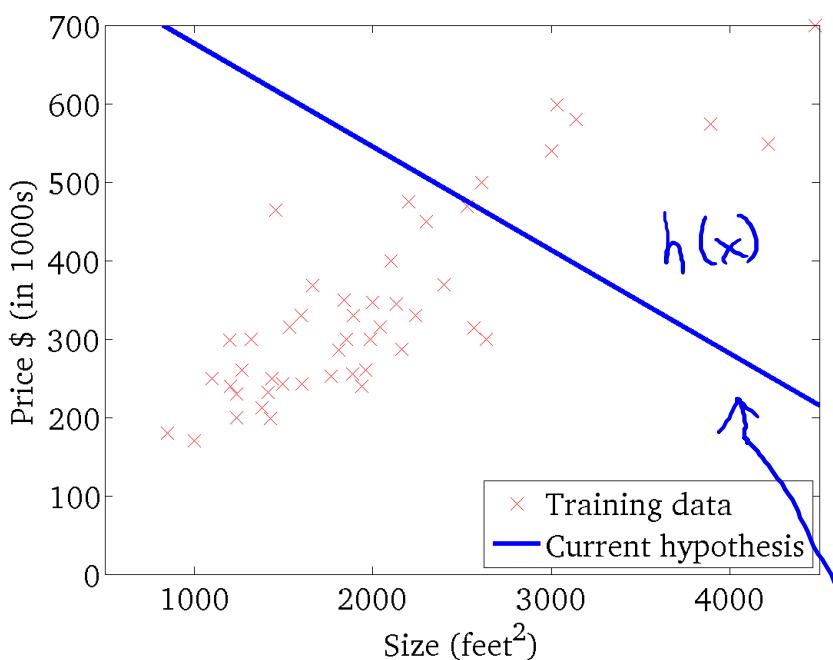
(function of the parameters θ_0, θ_1)





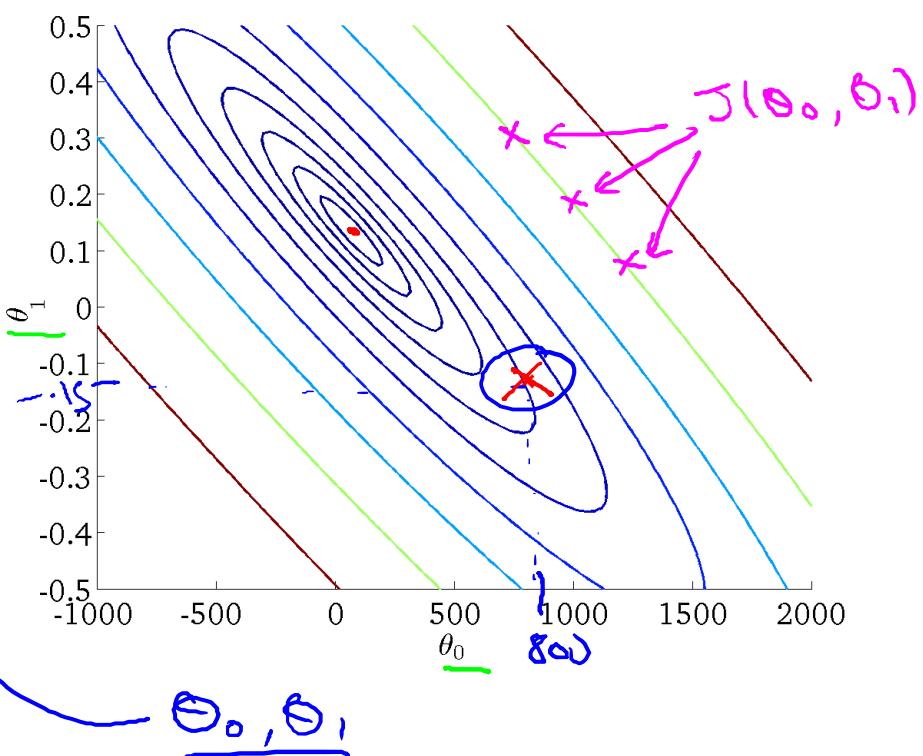
$$h_{\theta}(x)$$

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



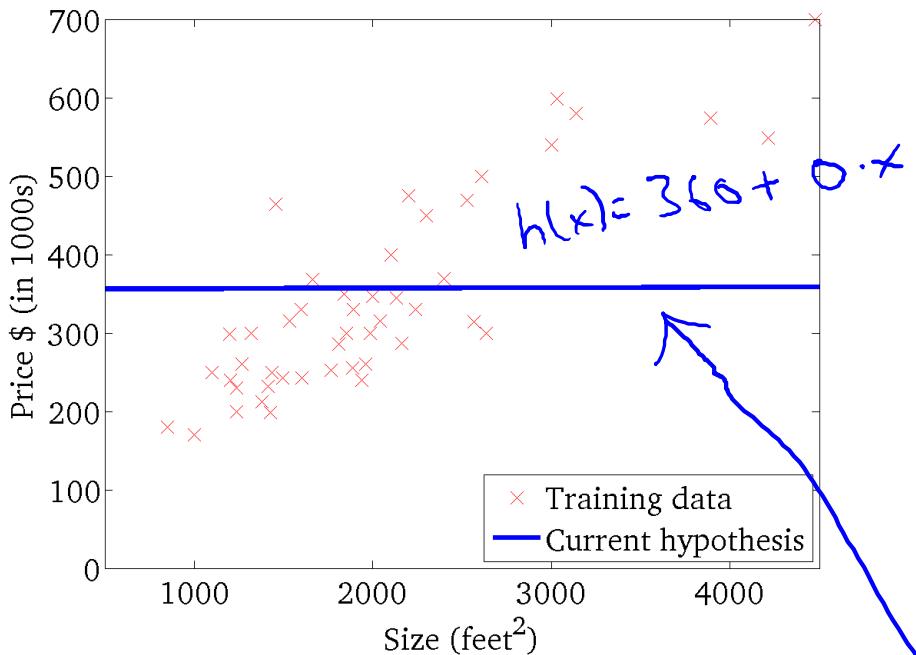
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



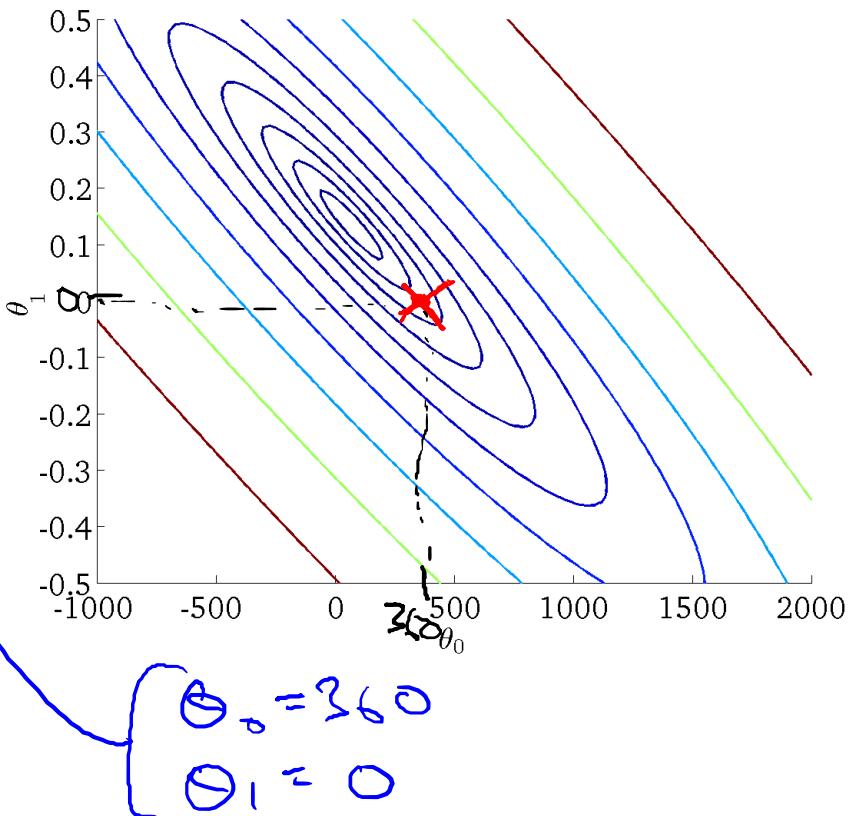
$$h_{\theta}(x)$$

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



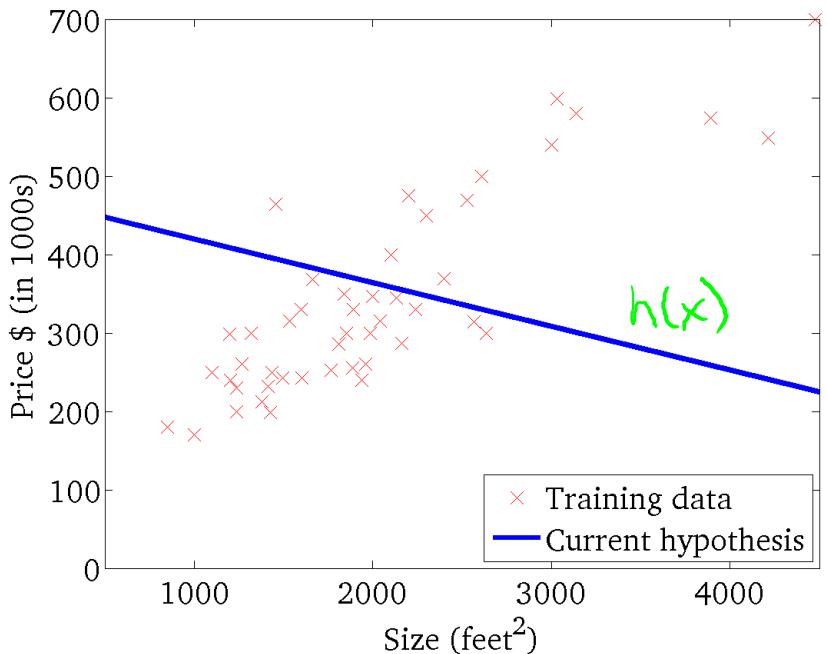
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



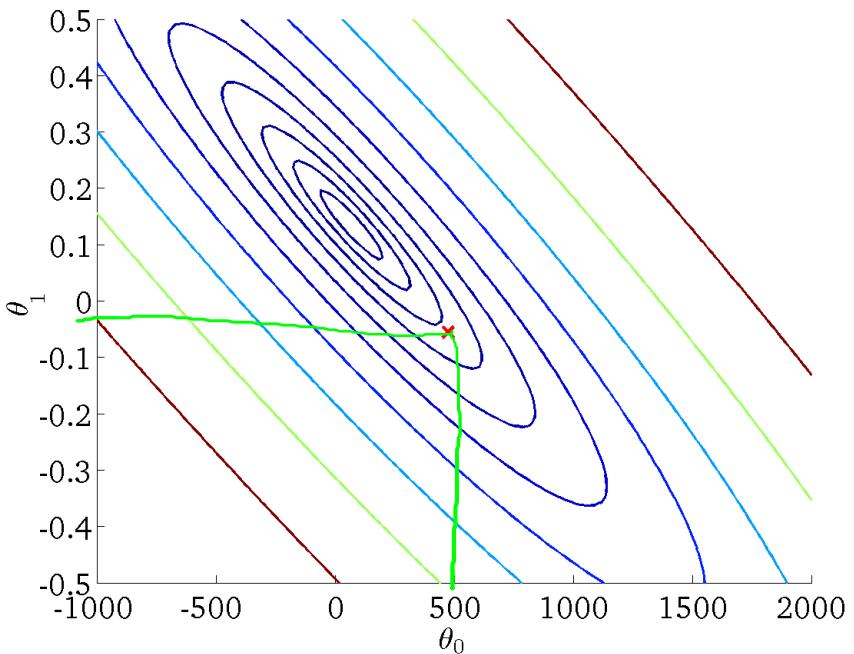
$$h_{\theta}(x)$$

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



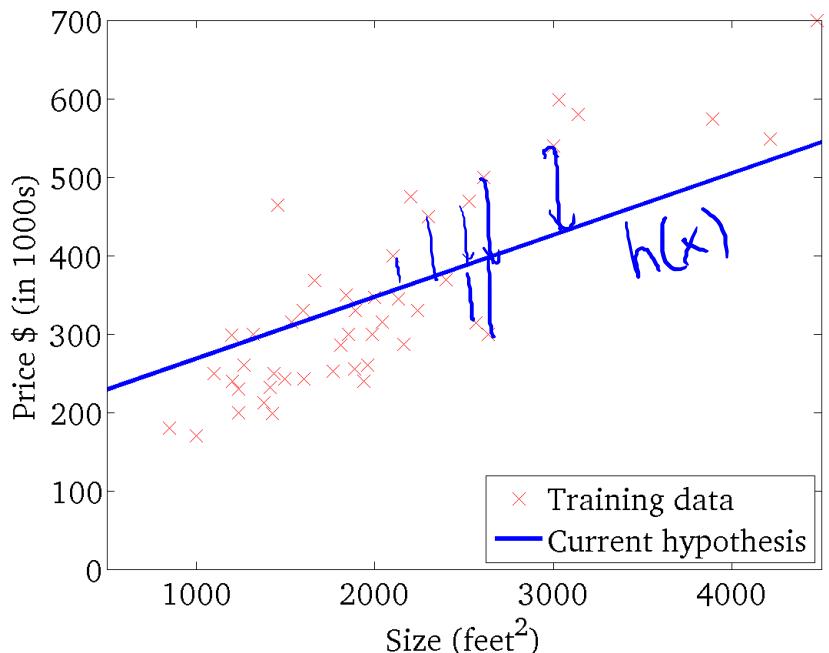
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



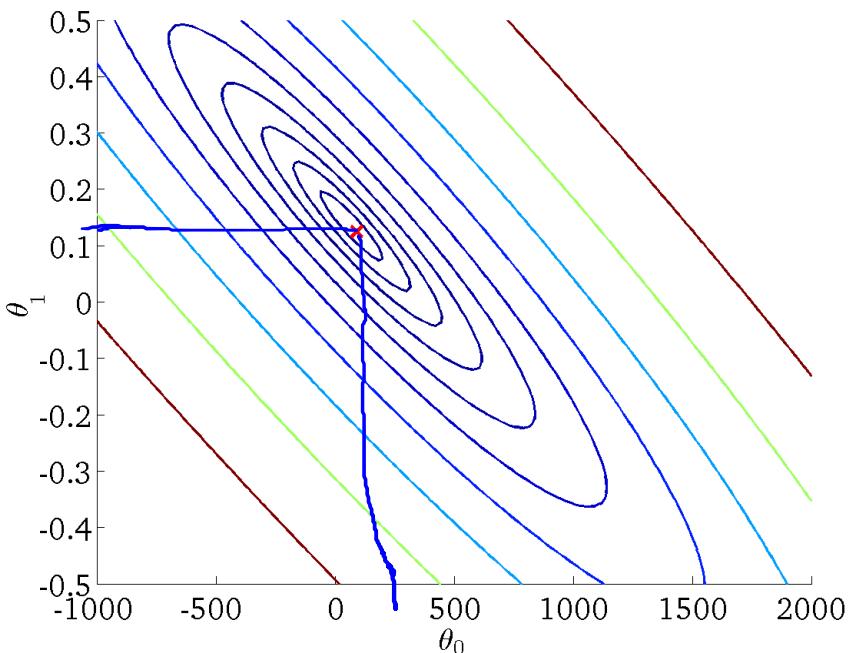
$$h_{\theta}(x)$$

(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)



Linear regression with one variable

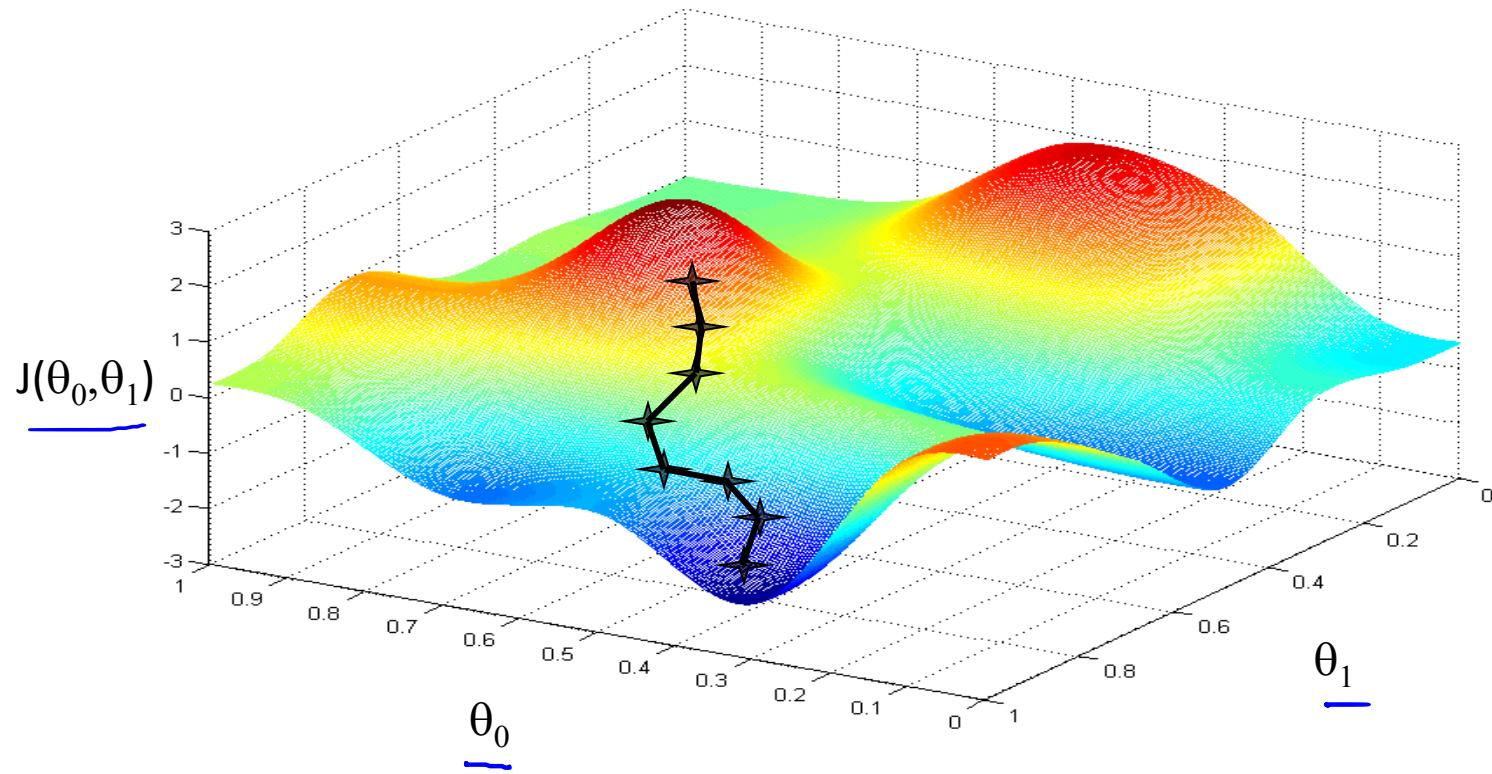
Gradient descent

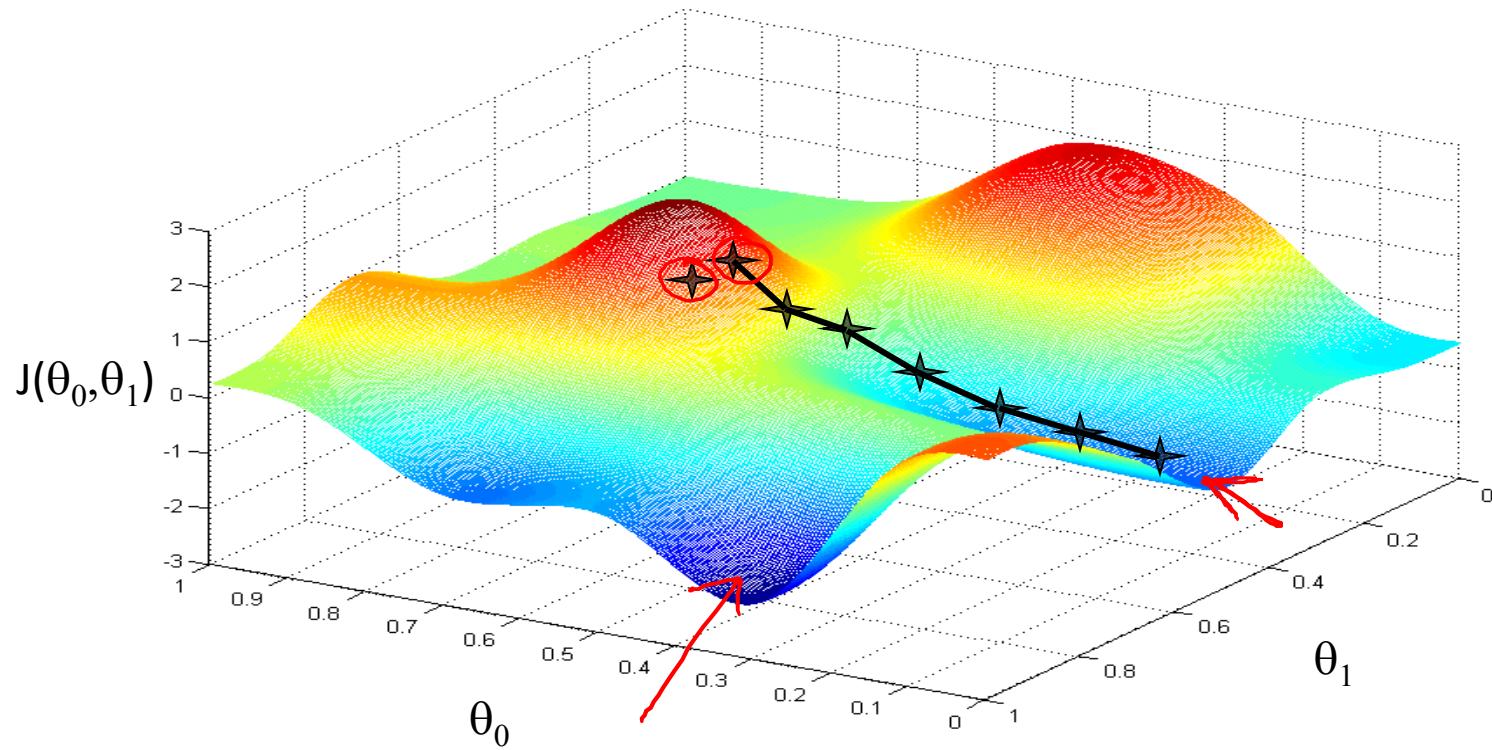
Have some function $\underline{J(\theta_0, \theta_1)}$ $J(\theta_0, \theta_1, \theta_2, \dots, \theta_n)$

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

Outline:

- Start with some $\underline{\theta_0, \theta_1}$ (say $\theta_0 = 0, \theta_1 = 0$)
- Keep changing $\underline{\theta_0, \theta_1}$ to reduce $\underline{J(\theta_0, \theta_1)}$
until we hopefully end up at a minimum





Gradient descent algorithm

repeat until convergence {

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

learning rate

θ_0, θ_1

Assignment

$$a := b$$

$$a := a + 1$$

Truth assertion

$$a = b$$

$$a = a + 1 \times$$

(for $j = 0$ and $j = 1$)

Simultaneously update
 θ_0 and θ_1

Correct: Simultaneous update

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



Incorrect:

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



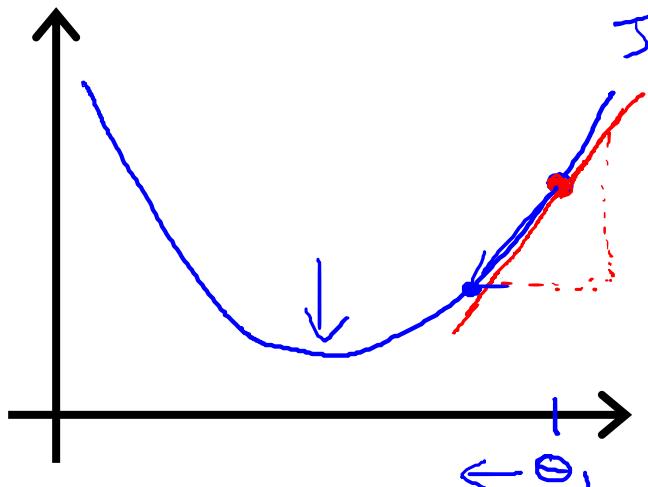
Linear regression with one variable

Gradient descent intuition

Gradient descent algorithm

repeat until convergence {
 → $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ (simultaneously update
 } $j = 0$ and $j = 1$)
 ↓ ↑
 learning rate derivative
}

$$\min_{\theta_1} J(\theta_1) \quad \theta_1 \in \mathbb{R}.$$

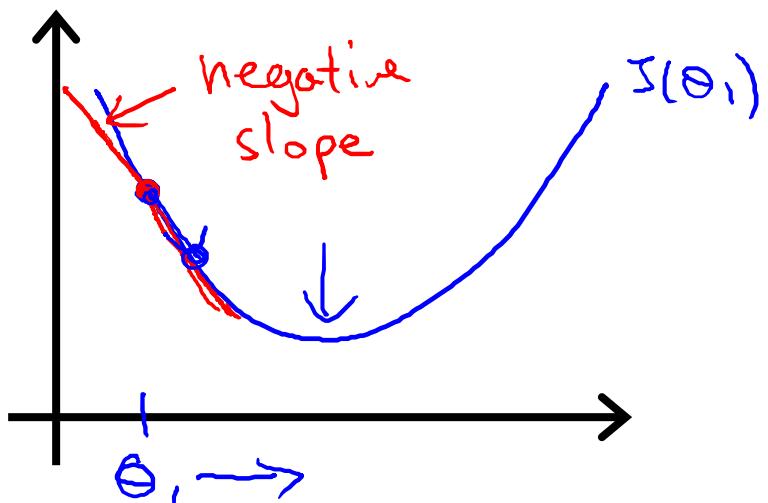


$J(\theta_0)$ ($\theta_0 \in \mathbb{R}$)

$$\theta_1 := \theta_0 - \frac{\alpha}{\frac{d}{d\theta_1} J(\theta_1)} \geq 0$$

Red circle highlights the term $\frac{d}{d\theta_1}$.

$\theta_1 := \theta_1 - \frac{\alpha}{\text{(positive number)}}$



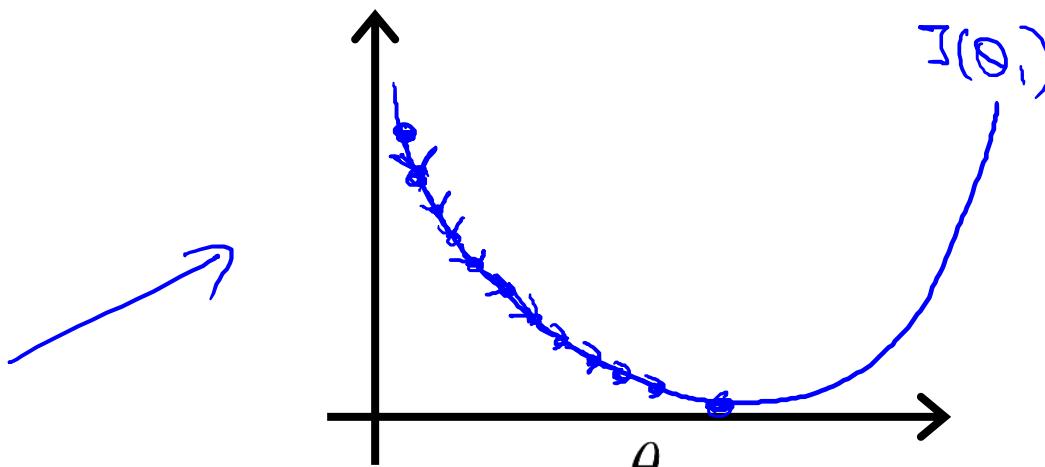
$J(\theta_0)$

$$\frac{\frac{d}{d\theta_1} J(\theta_1)}{\leq 0}$$

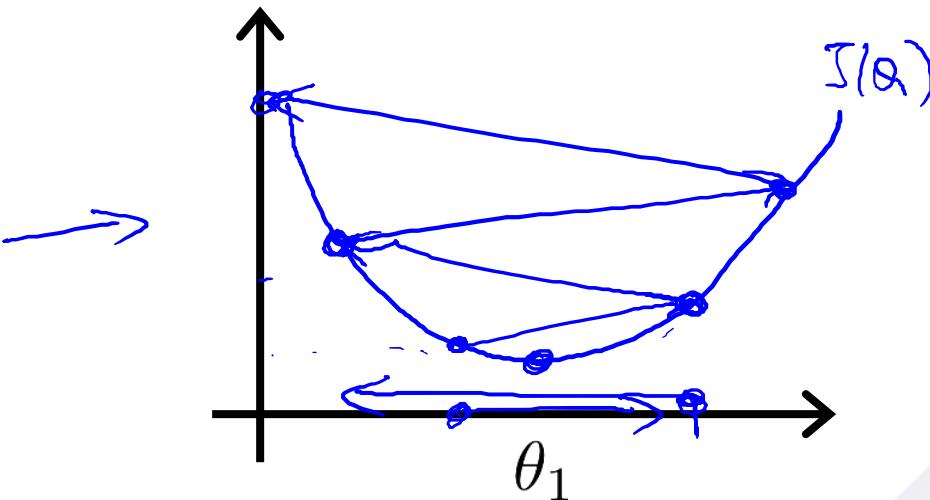
$\theta_1 := \theta_1 - \frac{\alpha}{\text{(negative number)}}$

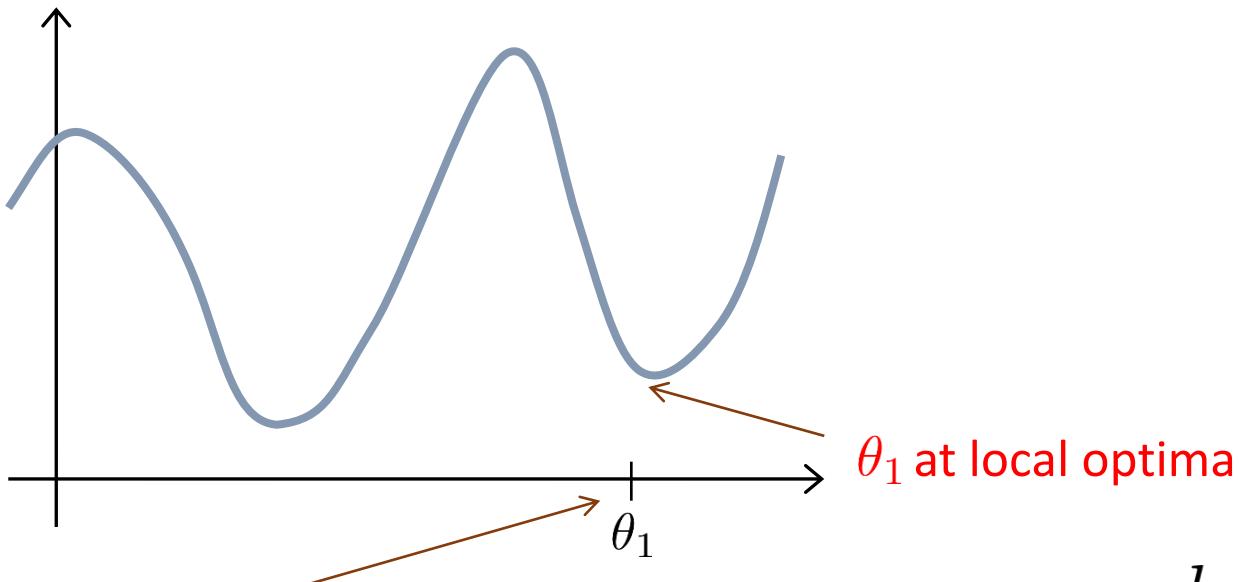
$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

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, or even diverge.



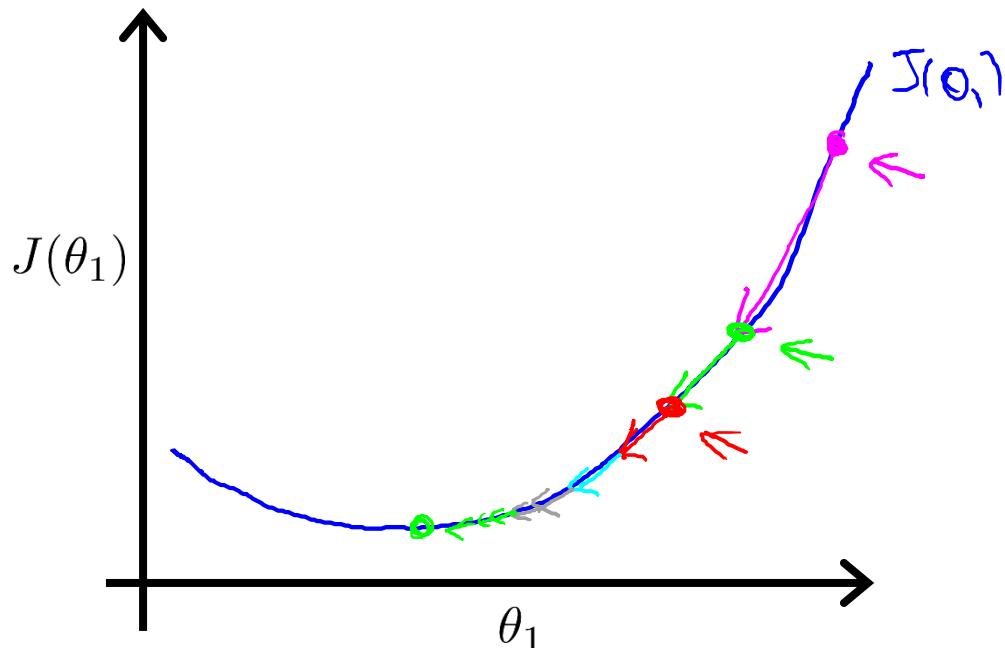


$$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$$

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

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

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



Linear regression with one variable

Gradient descent for
linear regression

Gradient descent algorithm

```
repeat until convergence {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$   
    (for  $j = 1$  and  $j = 0$ )  
}
```

Linear Regression Model

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

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

$$\begin{aligned}
 \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) &= \frac{2}{2m} \frac{1}{m} \sum_{i=1}^m \underline{(h_{\theta}(x^{(i)}) - y^{(i)})^2} \\
 &= \frac{2}{2\theta_j} \frac{1}{2m} \sum_{i=1}^m \underline{(h_{\theta_0 + \theta_1 x^{(i)}} - y^{(i)})^2}
 \end{aligned}$$

$$j = 0 : \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

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

Gradient descent algorithm

repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \right]$$

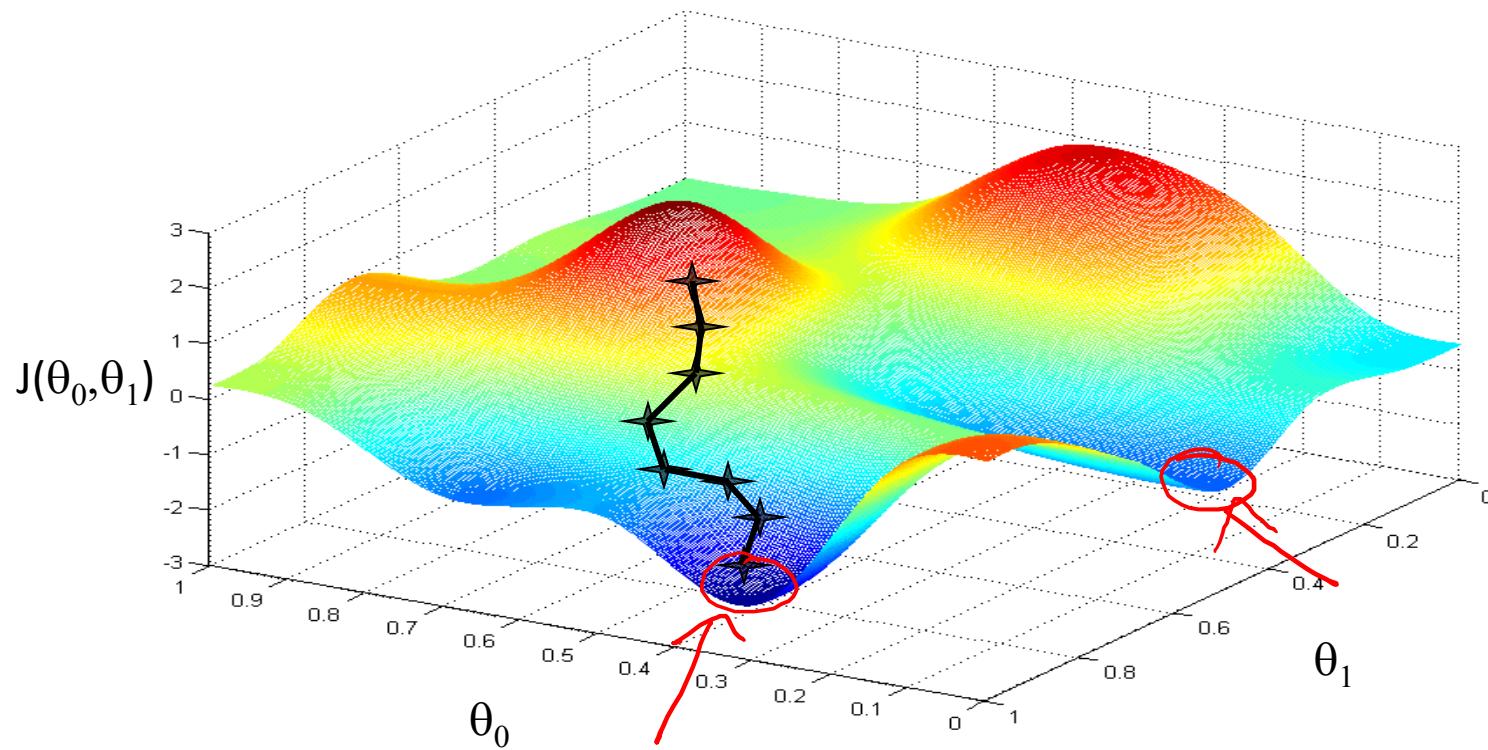
$$\theta_1 := \theta_1 - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)} \right]$$

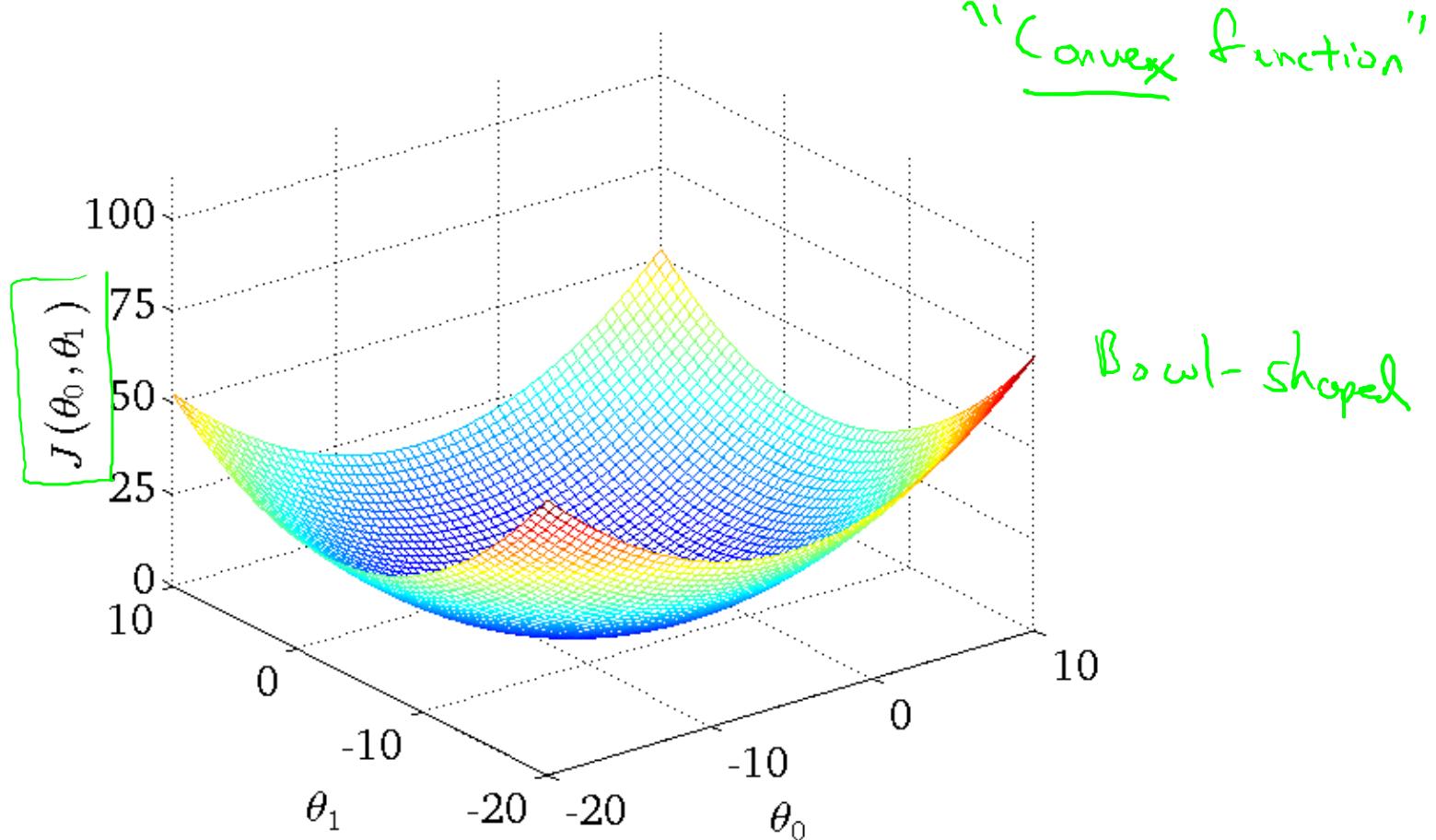
}

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

update
 θ_0 and θ_1
 simultaneously

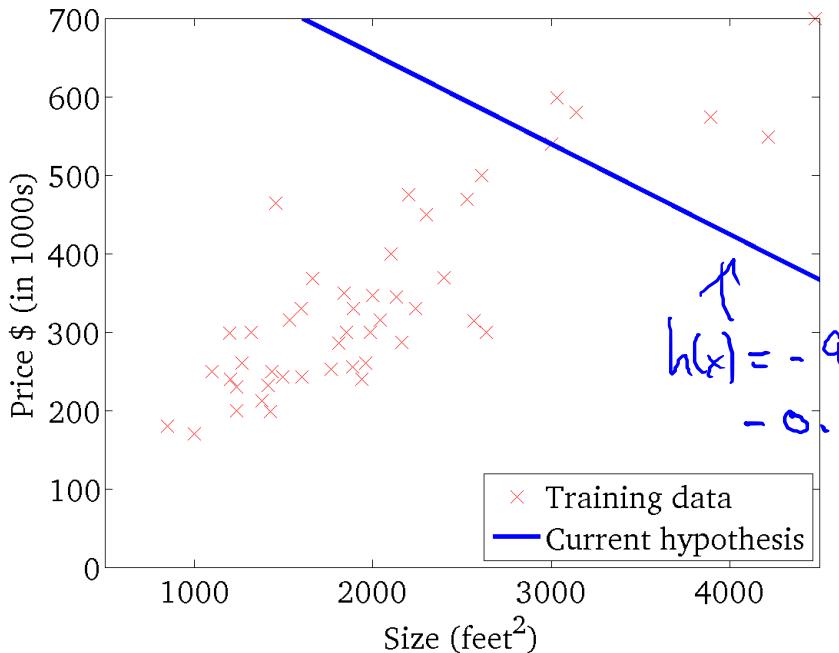
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$





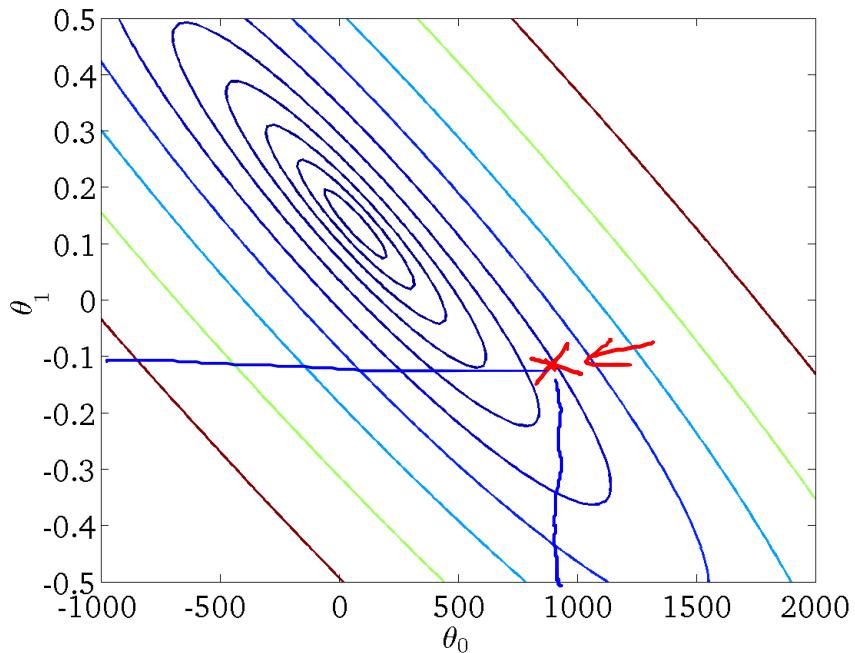
$$\underline{h_{\theta}(x)}$$

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



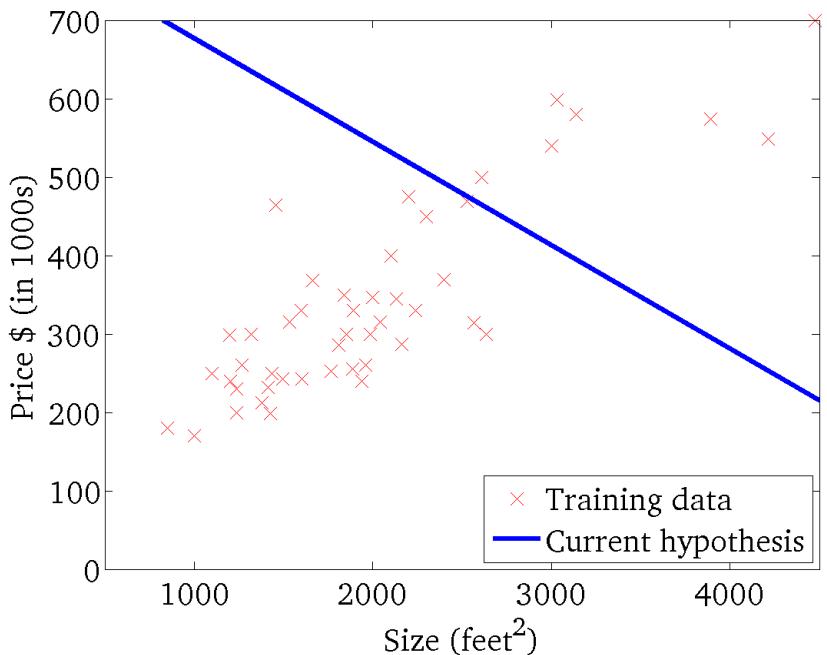
$$\underline{J(\theta_0, \theta_1)}$$

(function of the parameters θ_0, θ_1)



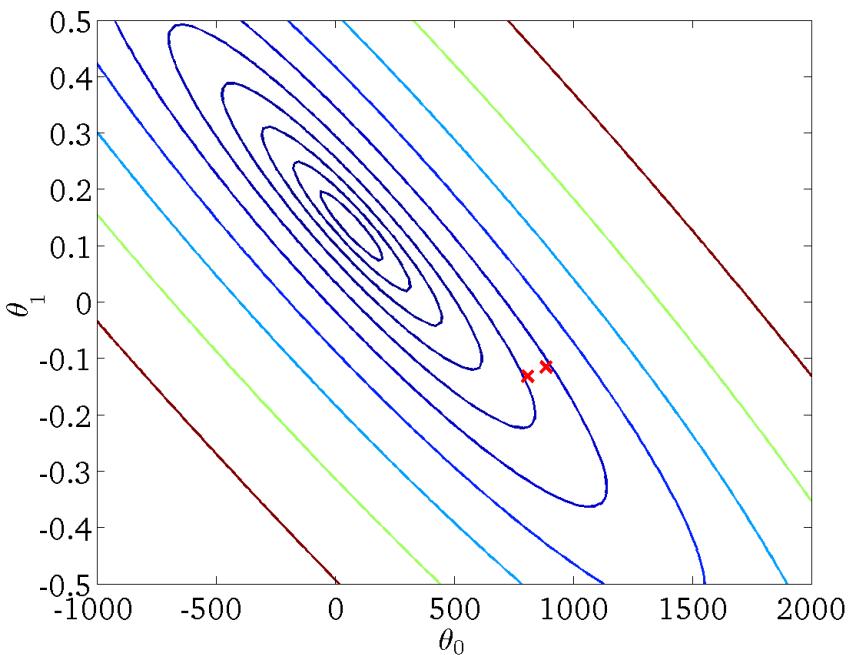
$$h_{\theta}(x)$$

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



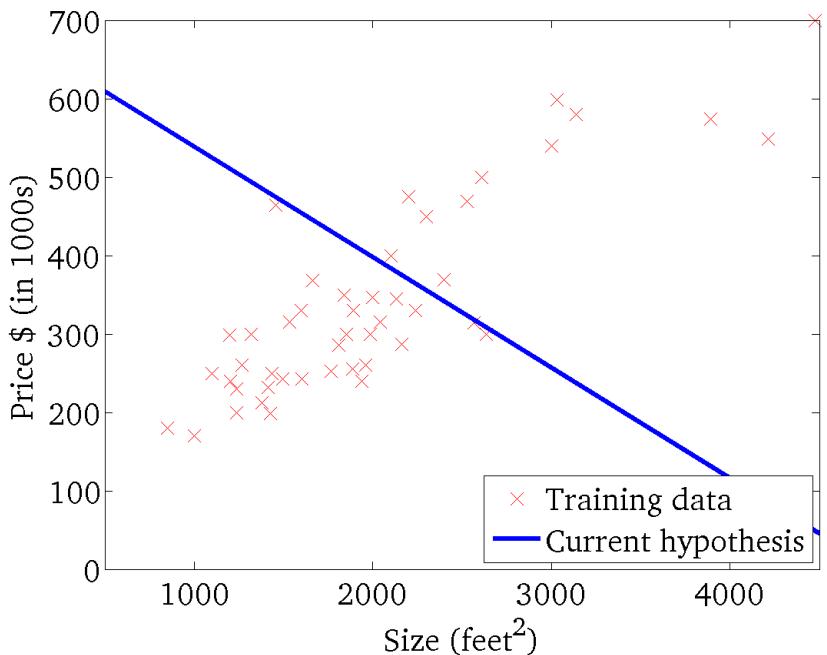
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



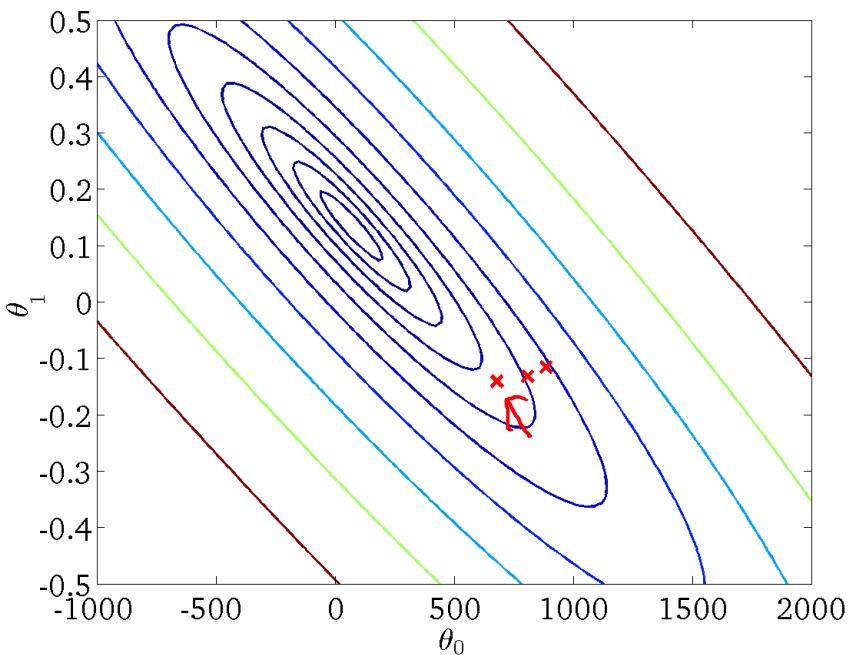
$$h_{\theta}(x)$$

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



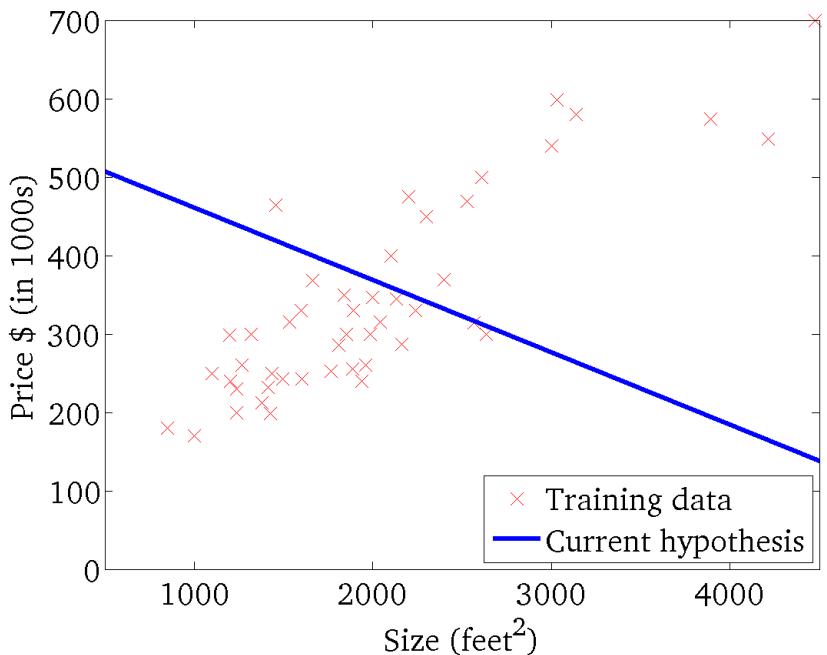
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



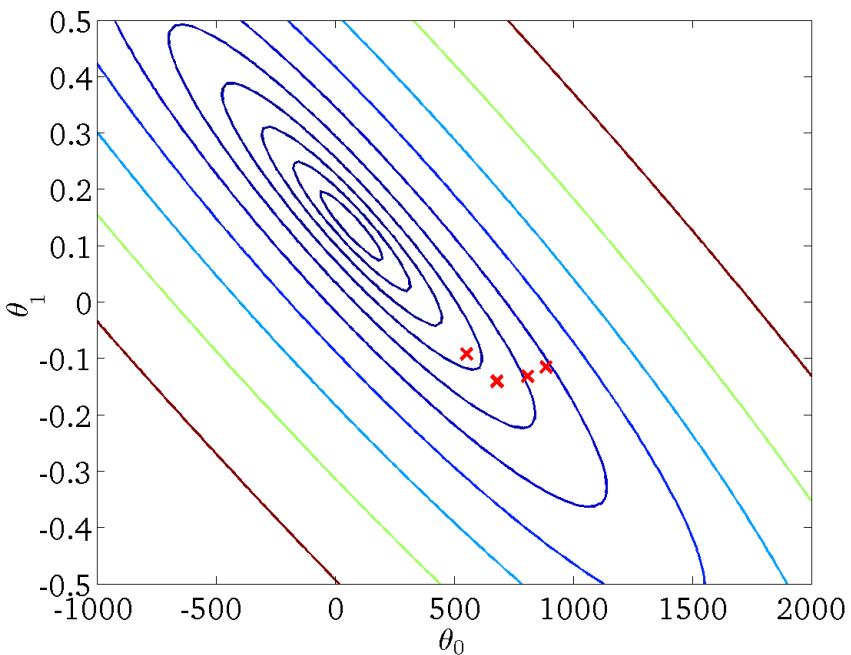
$$h_{\theta}(x)$$

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



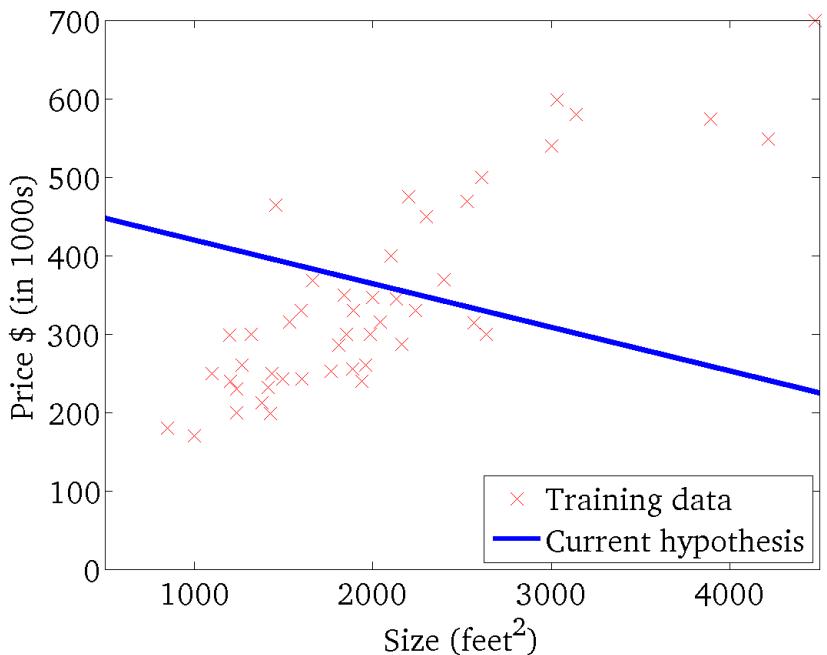
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



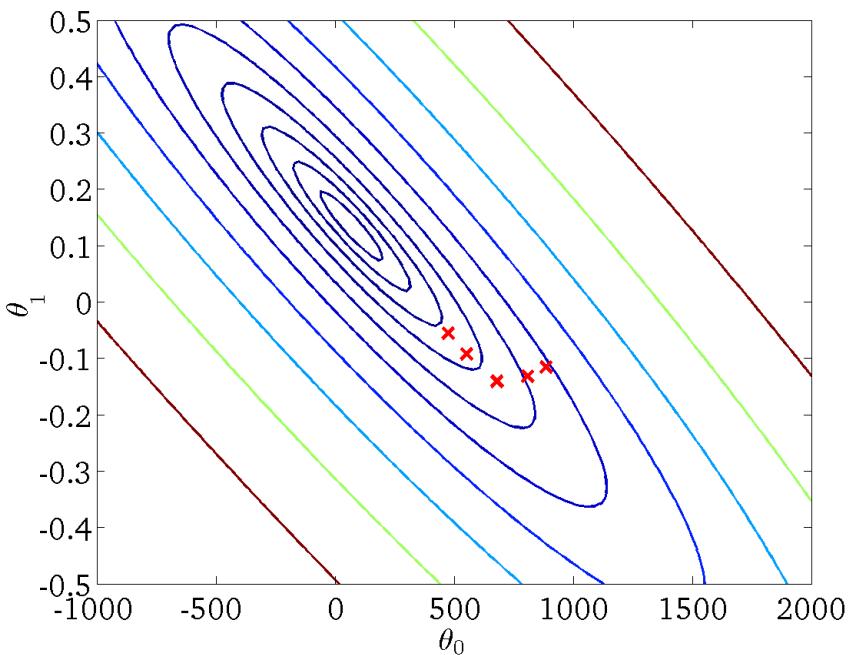
$$h_{\theta}(x)$$

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



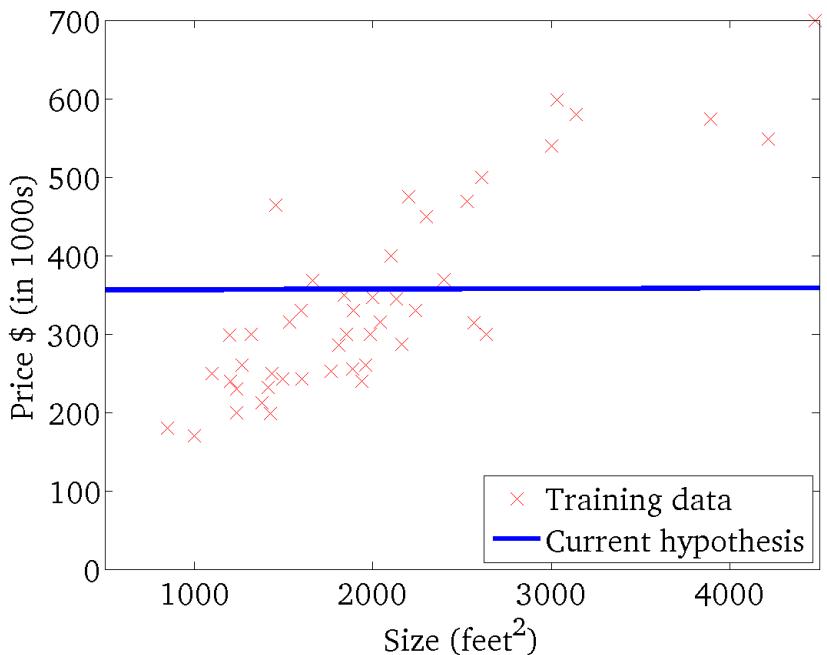
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



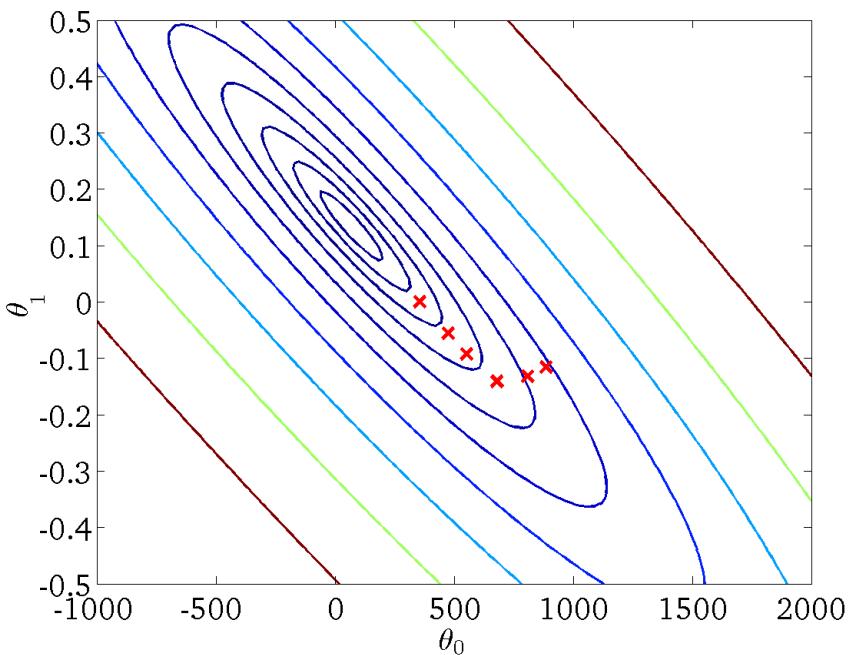
$$h_{\theta}(x)$$

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



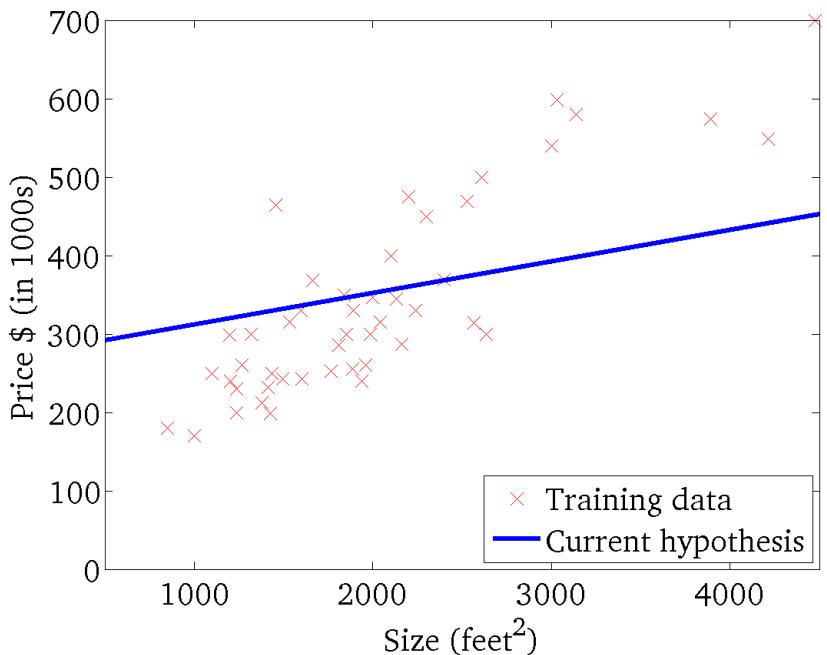
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



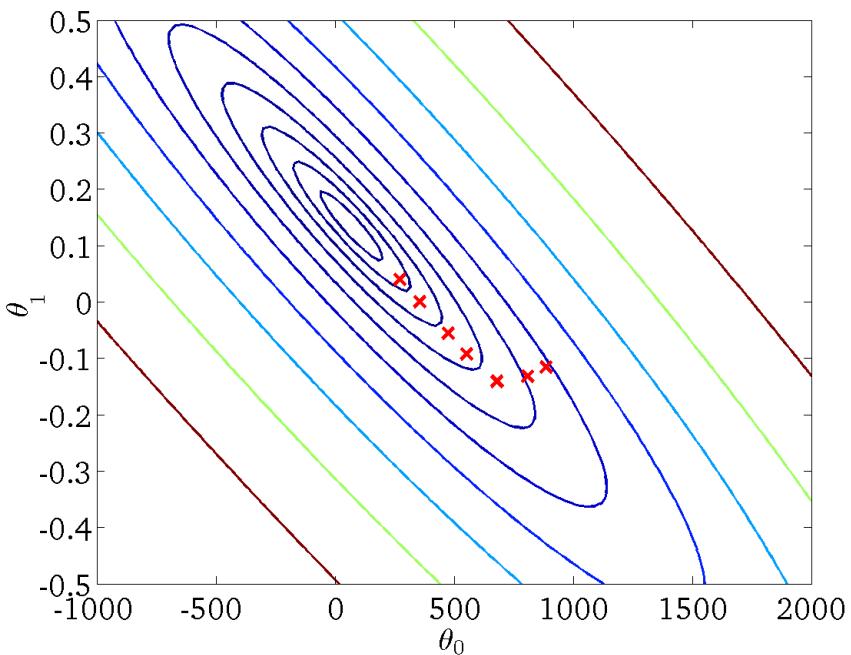
$$h_{\theta}(x)$$

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



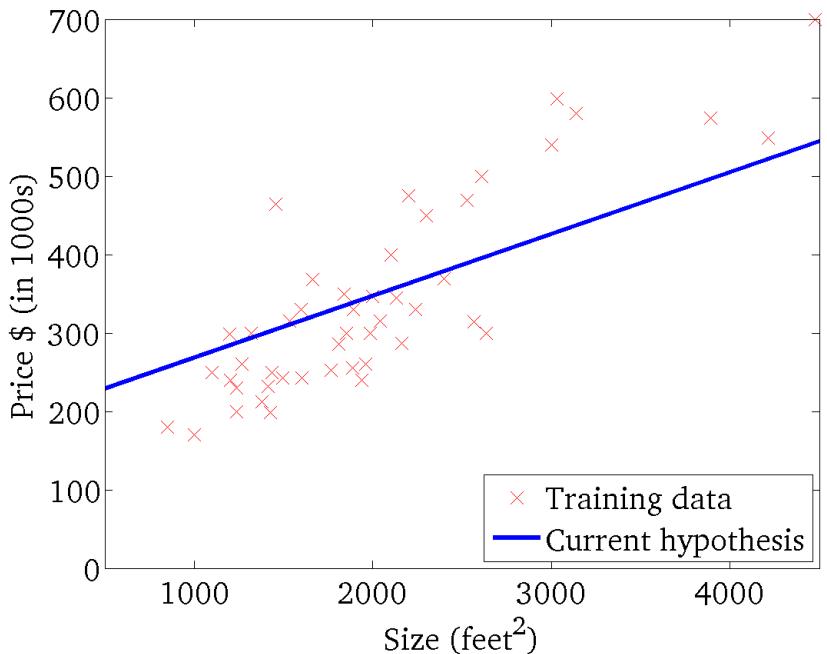
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



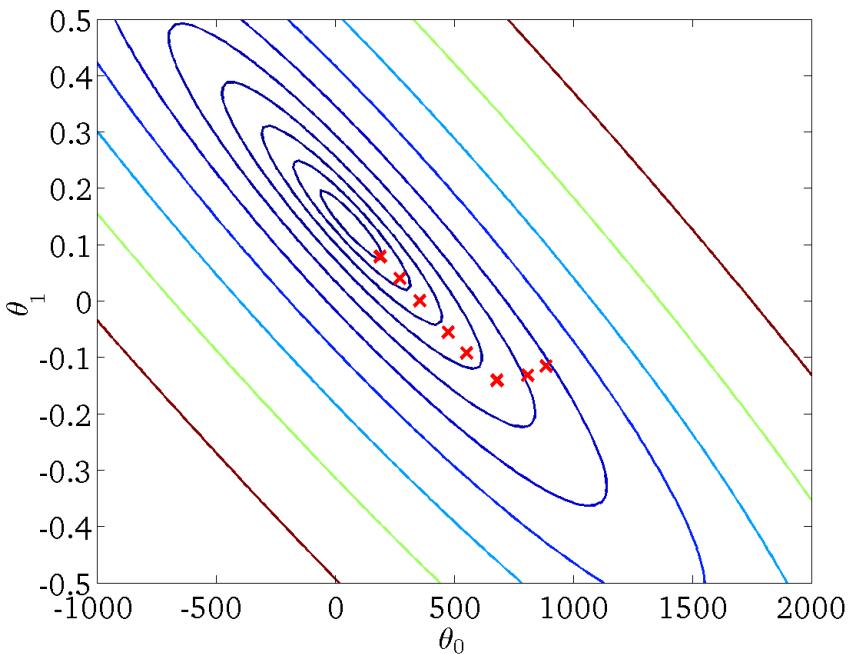
$$h_{\theta}(x)$$

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



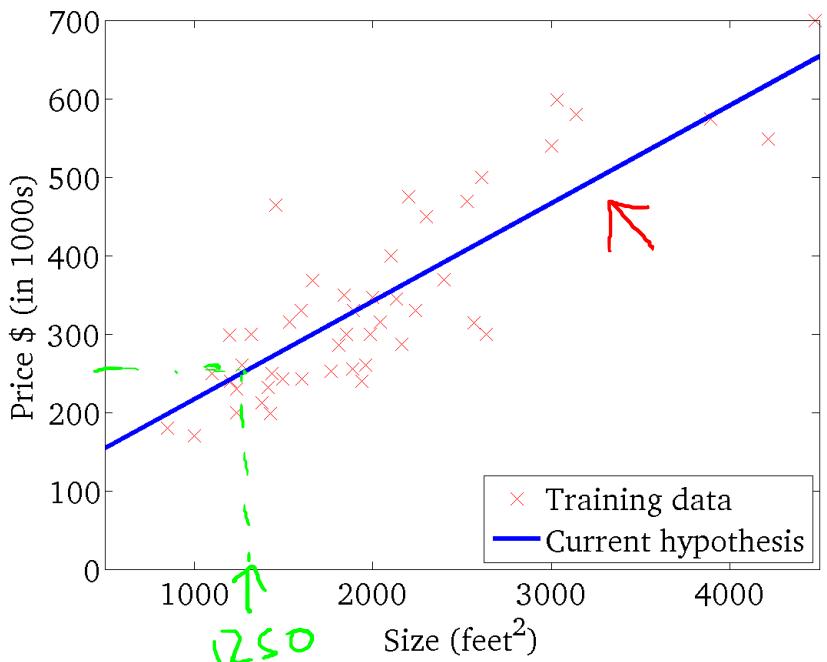
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



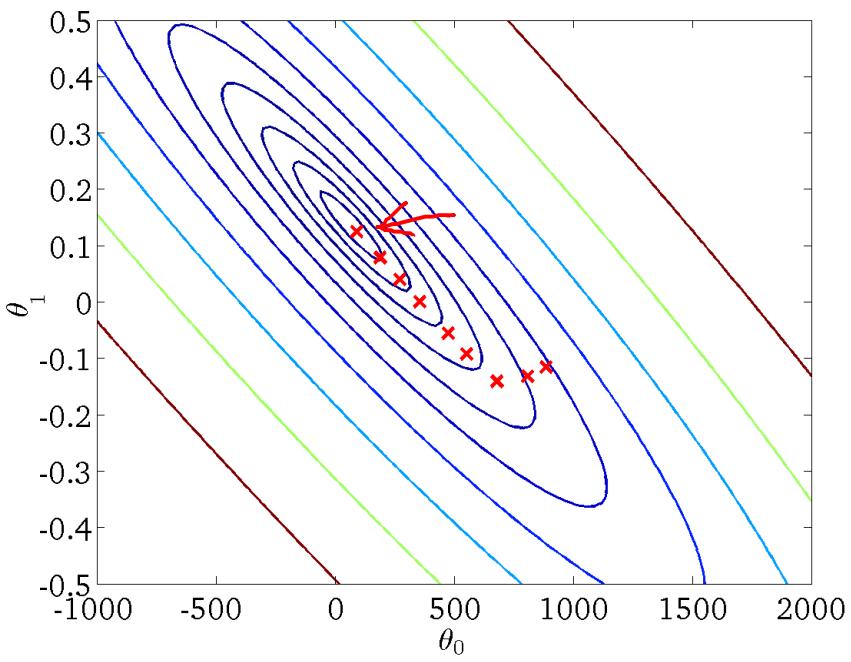
$$h_{\theta}(x)$$

(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)



Linear Regression with multiple variables

Multiple features

Multiple features (variables).

Size (feet ²)	Price (\$1000)
\xrightarrow{x}	$y \xleftarrow{ }$
2104	460
1416	232
1534	315
852	178
...	...

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

Multiple features (variables).

\rightarrow Size (feet ²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price (\$1000) y
2104	5	1	45	460
\rightarrow 1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...

Notation:

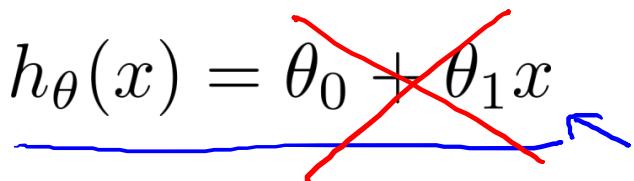
- $\rightarrow n = \text{number of features}$ $n=4$
- $\rightarrow x^{(i)} = \text{input (features) of } i^{\text{th}} \text{ training example.}$
- $\rightarrow x_j^{(i)} = \text{value of feature } j \text{ in } i^{\text{th}} \text{ training example.}$

$\underline{x}^{(2)} = \begin{bmatrix} 1416 \\ 3 \\ 2 \\ 40 \end{bmatrix}$

$x_3^{(2)} = 2$

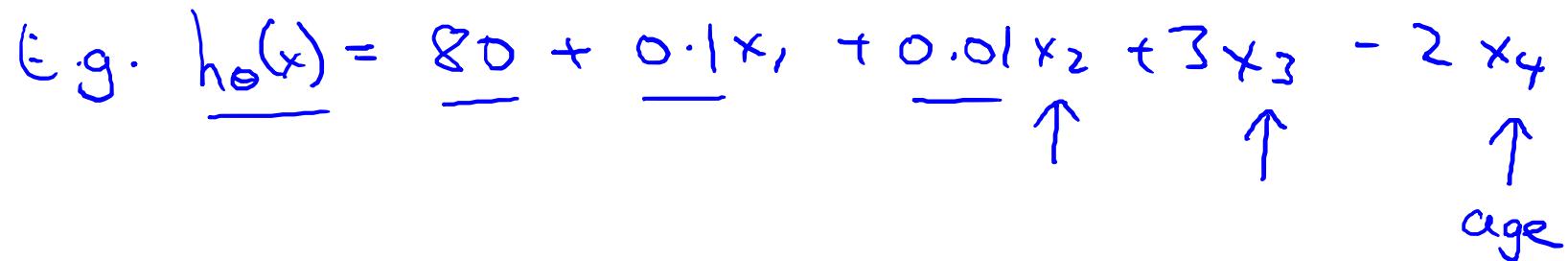
Hypothesis:

Previously:

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


$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

E.g. $\underline{h_{\theta}(x)} = \underline{80} + \underline{0.1x_1} + \underline{0.01x_2} + \underline{3x_3} - \underline{2x_4}$



$$\rightarrow h_{\theta}(x) = \underline{\theta_0} + \underline{\theta_1}x_1 + \underline{\theta_2}x_2 + \cdots + \underline{\theta_n}x_n$$

For convenience of notation, define $x_0 = 1$. ($x_0^{(i)} = 1$)

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$\begin{aligned} h_{\Theta}(x) &= \underline{\Theta_0x_0 + \Theta_1x_1 + \cdots + \Theta_nx_n} \\ &= \boxed{\Theta^T x} \end{aligned}$$

$$\Theta^T = \begin{bmatrix} \Theta_0 & \Theta_1 & \cdots & \Theta_n \end{bmatrix}$$

(n+1) × 1
 matrix
 $\Theta^T x$

Multivariate linear regression. ←

Linear Regression with multiple variables

Gradient descent for
multiple variables

Hypothesis: $h_{\theta}(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

Parameters: $\underline{\theta_0, \theta_1, \dots, \theta_n}$ Θ $n+1$ -dimensional vector

Cost function:

$$\underline{J(\theta_0, \theta_1, \dots, \theta_n)} = \underline{J(\Theta)} = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Gradient descent:

Repeat {

$$\rightarrow \theta_j := \theta_j - \alpha \left[\frac{\partial}{\partial \theta_j} J(\theta_0, \dots, \theta_n) \right] \underline{J(\Theta)}$$

} ↑ (simultaneously update for every $j = 0, \dots, n$)

Gradient Descent

Previously ($n=1$):

Repeat {



$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})$$

$\underbrace{\frac{\partial}{\partial \theta_0} J(\theta)}$



$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)}$$

$\cancel{x}_1^{(i)}$

(simultaneously update θ_0, θ_1)

}

New algorithm ($n \geq 1$):

Repeat {

$$\frac{\partial}{\partial \theta_j} J(\theta)$$



$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for $j = 0, \dots, n$)

}



$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$



$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)}$$



$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)}) x_2^{(i)}$$

...

$$x_0^{(i)} = 1$$

Linear Regression with multiple variables

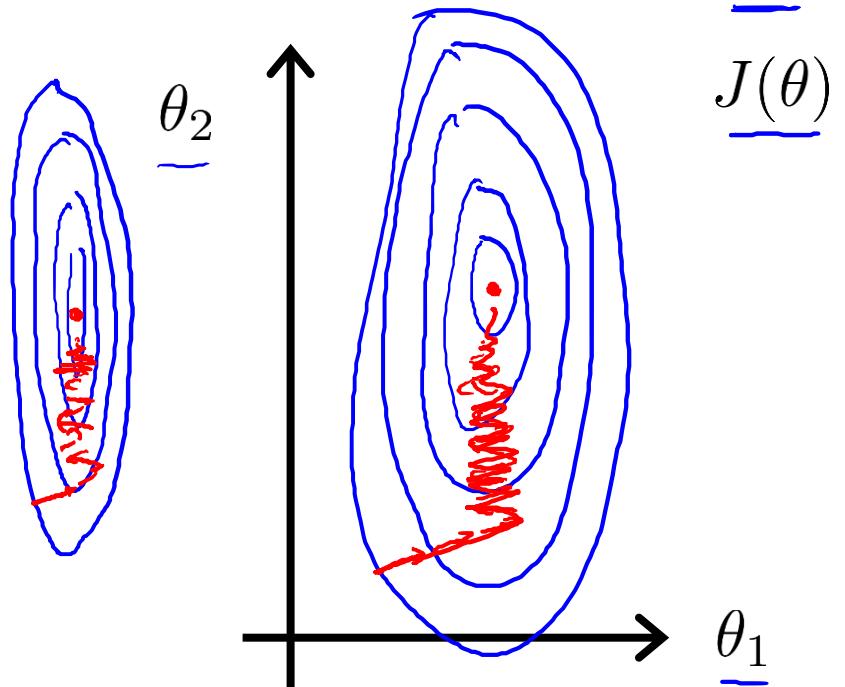
Gradient descent in
practice I: Feature Scaling

Feature Scaling

Idea: Make sure features are on a similar scale.

E.g. $x_1 = \text{size } (0\text{-}2000 \text{ feet}^2)$ ←

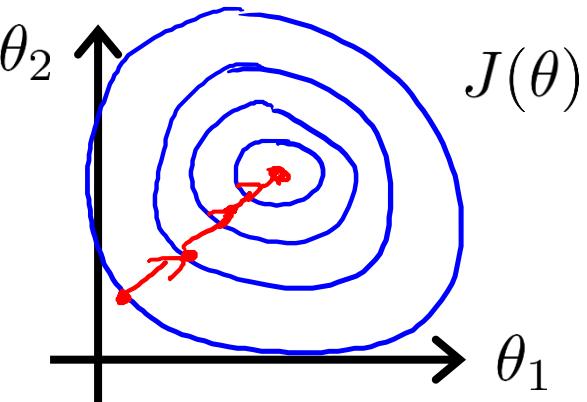
$x_2 = \text{number of bedrooms } (1\text{-}5)$ ←



$$\rightarrow x_1 = \frac{\text{size (feet}^2)}{2000} \quad \swarrow$$

$$\rightarrow x_2 = \frac{\text{number of bedrooms}}{5} \quad \swarrow$$

$$0 \leq x_1 \leq 1 \quad 0 \leq x_2 \leq 1$$



Feature Scaling

Get every feature into approximately a $-1 \leq x_i \leq 1$ range.

$$x_0 = 1$$

$$-1 \leq x_i \leq 1$$

$$6 \leq x_1 \leq 3 \quad \checkmark$$

$$-2 \leq x_2 \leq 0.5 \quad \checkmark$$

$$-100 \leq x_3 \leq 100 \quad \times$$

$$-0.0001 \leq x_4 \leq 0.0001 \quad \times$$

$$-3 \text{ to } 3 \quad \checkmark$$

$$-\frac{1}{3} \text{ to } \frac{1}{3} \quad \checkmark$$

Mean normalization

Replace x_i with $\underline{x_i - \mu_i}$ to make features have approximately zero mean
 (Do not apply to $x_0 = 1$).

E.g. $\rightarrow x_1 = \frac{\text{size} - 1000}{2000}$

Average size = 100

$$x_2 = \frac{\#bedrooms - 2}{5 - 4}$$

$1 \leq$ bedrooms

$\rightarrow [-0.5 \leq x_1 \leq 0.5, -0.5 \leq x_2 \leq 0.5]$

$$x_1 \leftarrow \frac{x_1 - \mu_1}{\sigma_1}$$

avg value
 of x_1
 in training
 set

range $(\max - \min)$
 (or standard deviation)

$x_2 \leftarrow \frac{x_2 - \mu_2}{\sigma_2}$

Linear Regression with multiple variables

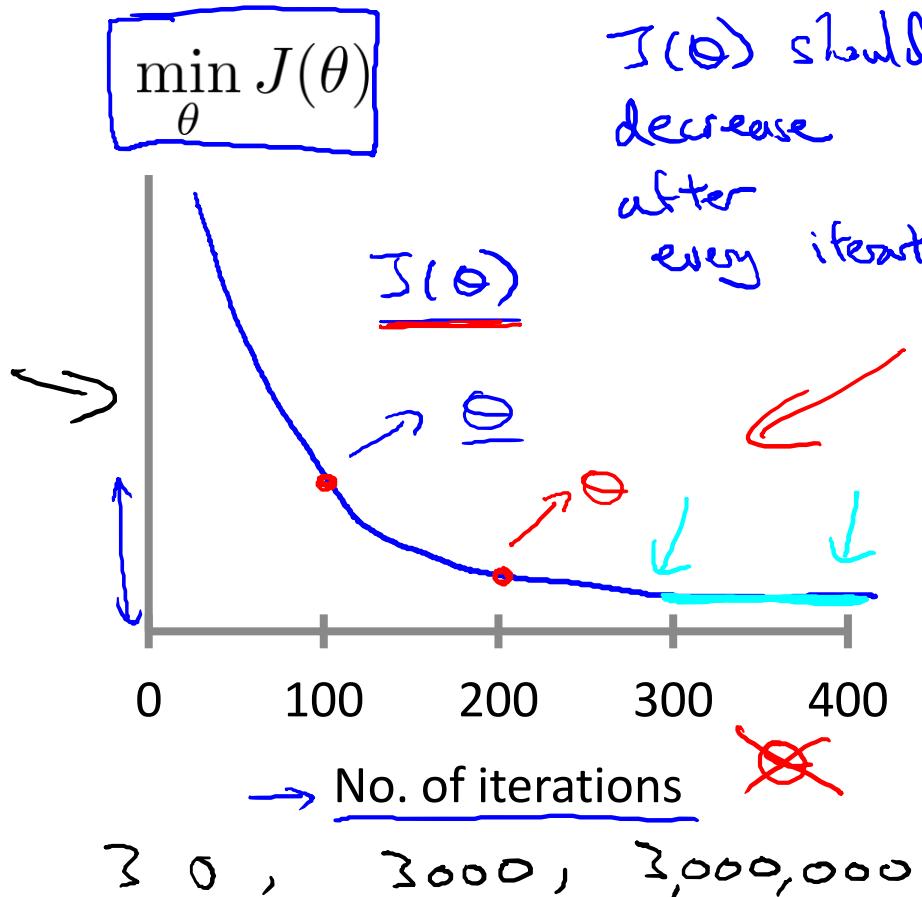
Gradient descent in
practice II: Learning rate

Gradient descent

$$\Rightarrow \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- “Debugging”: How to make sure gradient descent is working correctly.
- How to choose learning rate α .

Making sure gradient descent is working correctly.



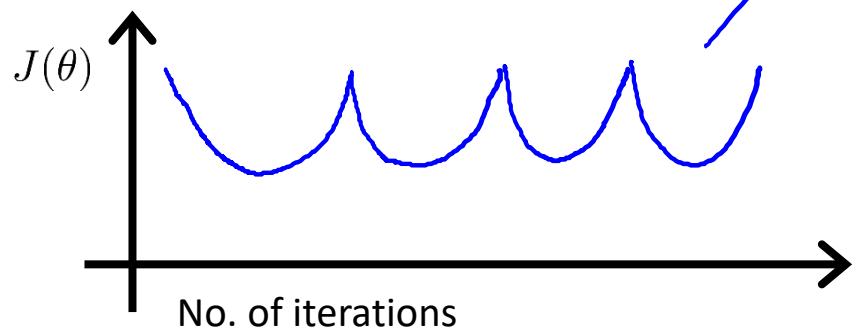
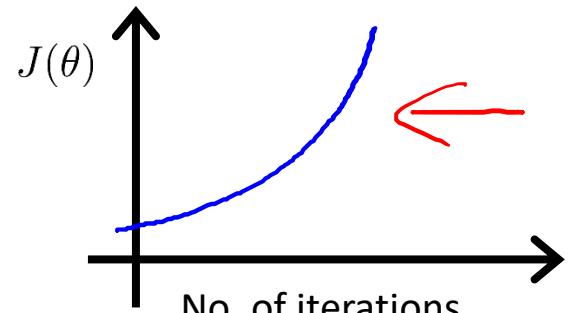
$J(\theta)$ should decrease after every iteration.

→ Example automatic convergence test:

→ Declare convergence if $J(\theta)$ decreases by less than 10^{-3}

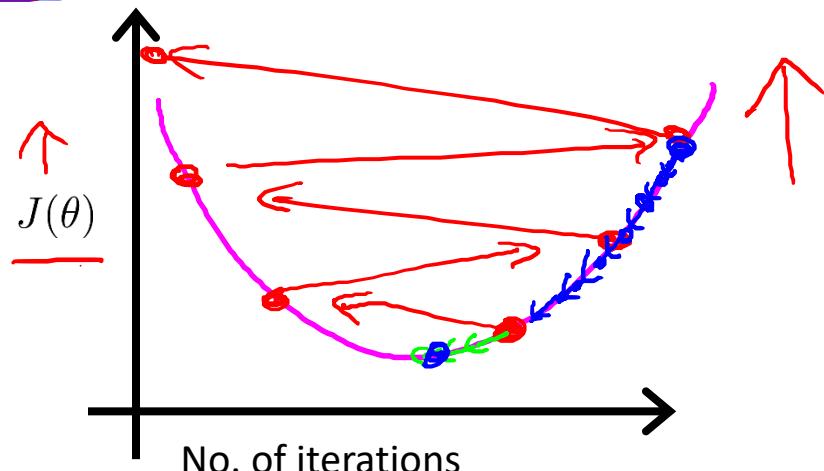
$$10^{-3}$$

Making sure gradient descent is working correctly.



Gradient descent not working.

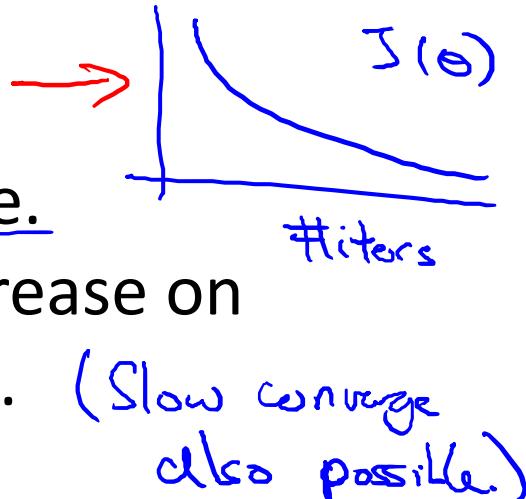
Use smaller α .



- For sufficiently small α , $J(\theta)$ should decrease on every iteration. ↗
- But if α is too small, gradient descent can be slow to converge.

Summary:

- If α is too small: slow convergence.
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge. (Slow converge also possible.)



To choose α , try

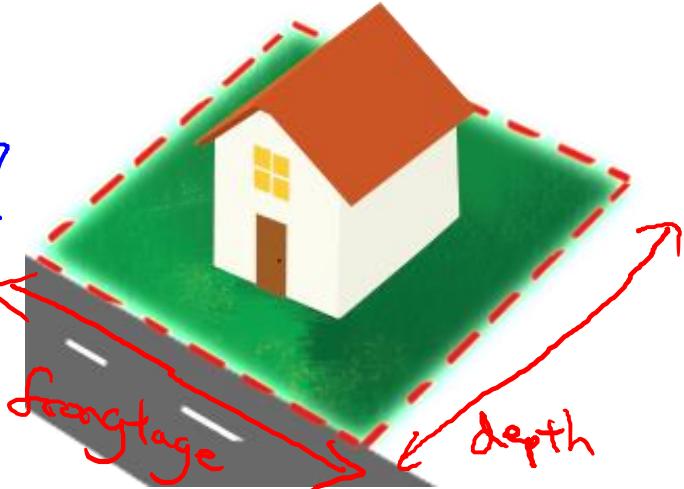
$$\dots, \underbrace{0.001}_{\uparrow}, \underbrace{0.003}_{\approx 2x}, \underbrace{0.01}_{\approx 2x}, \underbrace{0.03}_{3x}, \underbrace{0.1}_{\approx 3x}, \underbrace{0.3}_{3x}, \underbrace{1}_{\approx 3x}, \dots$$

Linear Regression with multiple variables

Features and
polynomial regression

Housing prices prediction

$$h_{\theta}(x) = \theta_0 + \theta_1 \times \underbrace{\text{frontage}}_{x_1} + \theta_2 \times \underbrace{\text{depth}}_{x_2}$$



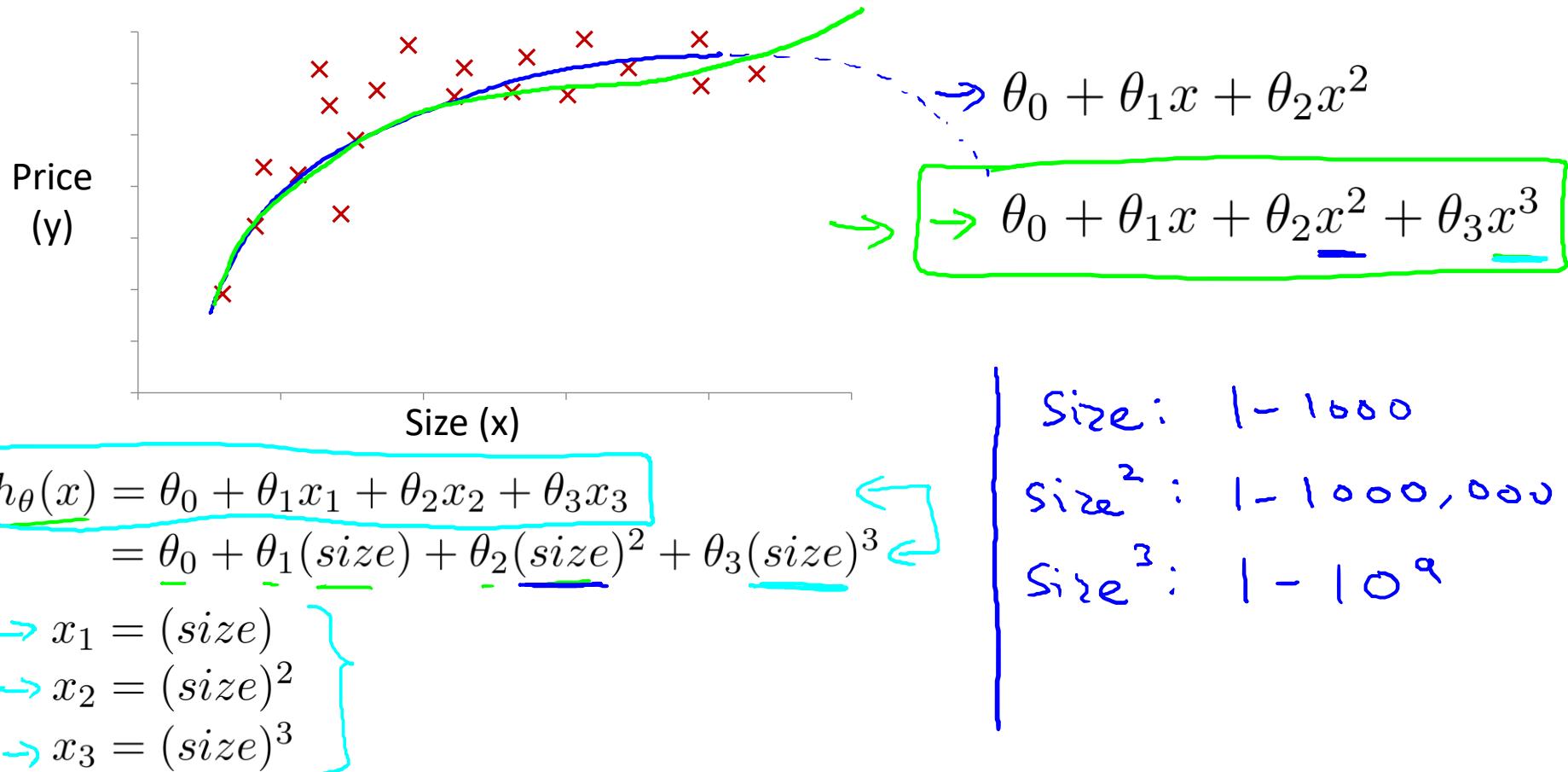
Area

$$x = \underline{\text{frontage} \times \text{depth}}$$

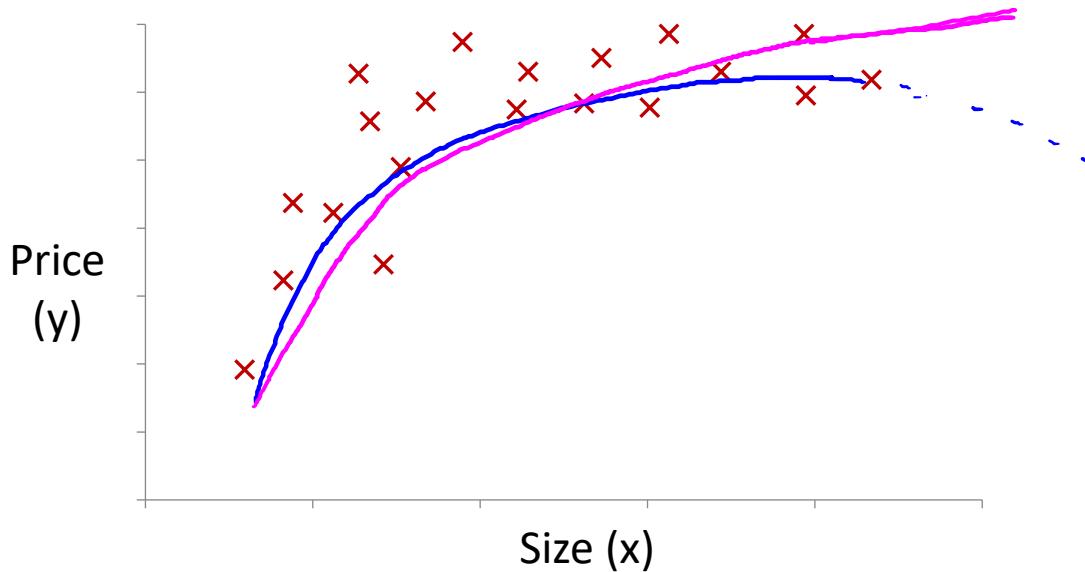
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

\curvearrowleft land area

Polynomial regression



Choice of features



$$\rightarrow h_{\theta}(x) = \theta_0 + \theta_1(\text{size}) + \theta_2(\text{size})^2$$

$$\rightarrow h_{\theta}(x) = \theta_0 + \theta_1(\text{size}) + \theta_2 \sqrt{(\text{size})}$$

