

# Linear Regression

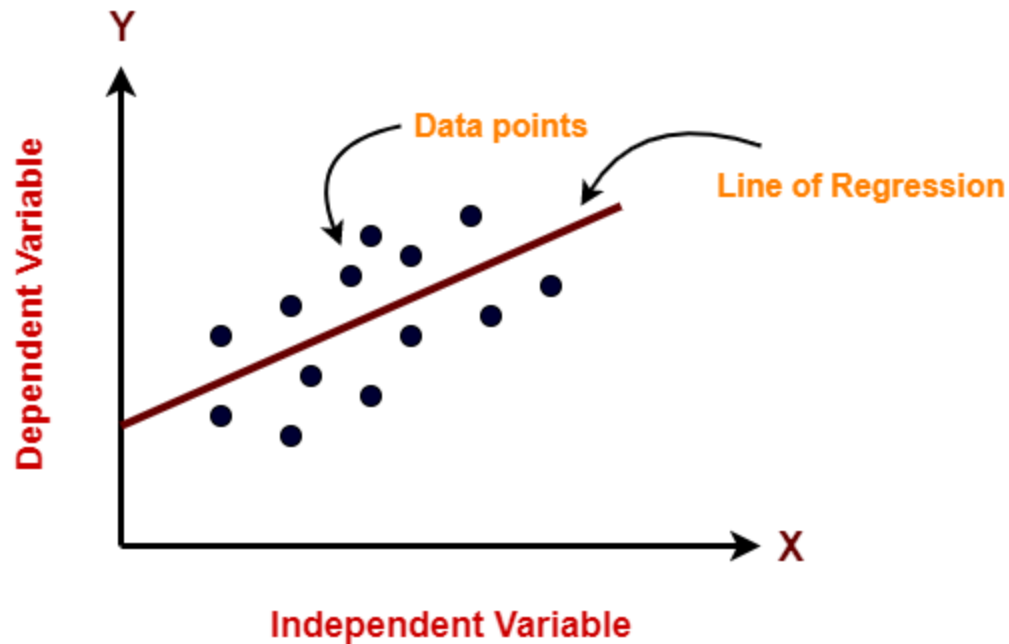
# Contents

- \* Linear Regression
- \* Types of Linear Regression
- \* Simple Linear Regression
- \* Multiple Linear Regression

# Linear Regression: Definition

- \* Regression analysis is used to Predict the value of a dependent variable based on the value of at least one independent variable
- \* Explain the impact of changes in an independent variable on the dependent variable
- \* Linear Regression is a supervised machine learning algorithm.
- \* It tries to find out the best linear relationship that describes the data you have.
- \* It assumes that there exists a linear relationship between a dependent variable and independent variable(s).
- \* The value of the dependent variable of a linear regression model is a continuous value i.e. real numbers.
- \* Dependent variable/ Output Variable :
  - \* the variable we wish to predict or explain
- \* Independent variable/ Input Variable :
  - \* the variable used to explain the dependent variable

# Simple example of Linear Regression



# Types of Linear Regression:

- \* Simple Linear regression
- \* Multiple Linear Regression

# Simple Linear Regression

- \* In simple linear regression, the **dependent variable depends only on a single independent variable.**
- \* For simple linear regression, the form of the model is-
- \*  $Y = \beta_0 + \beta_1 X$
- \* Y is a dependent variable.
- \* X is an independent variable.
- \*  $\beta_0$  and  $\beta_1$  are the regression coefficients.
- \*  $\beta_0$  is the intercept or the bias that fixes the offset to a line.
- \*  $\beta_1$  is the slope or weight that specifies the factor by which X has an impact on Y.
- \*

# Possible values of $\beta_1$

- \* Case-01:  $\beta_1 < 0$

- \*

- \* It indicates that variable X has negative impact on Y.
- \* If X increases, Y will decrease and vice-versa.

- \* Case-02:  $\beta_1 = 0$

- \*

- \* It indicates that variable X has no impact on Y.
- \* If X changes, there will be no change in Y.

# Possible values of $\beta_1$

- \* Case-03:  $\beta_1 > 0$

- \*

- \* It indicates that variable X has positive impact on Y.

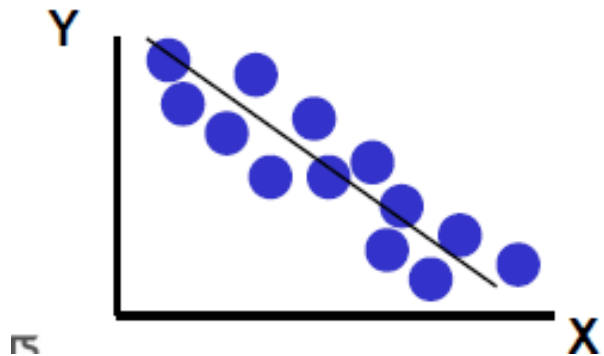
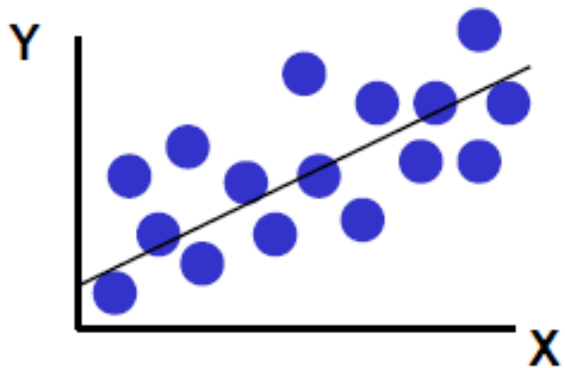
- \* If X increases, Y will increase and vice-versa.

- \*

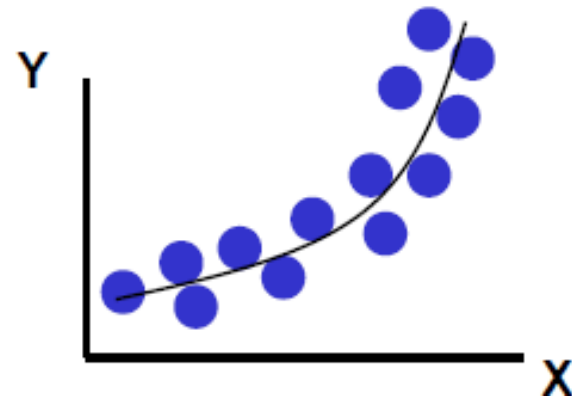
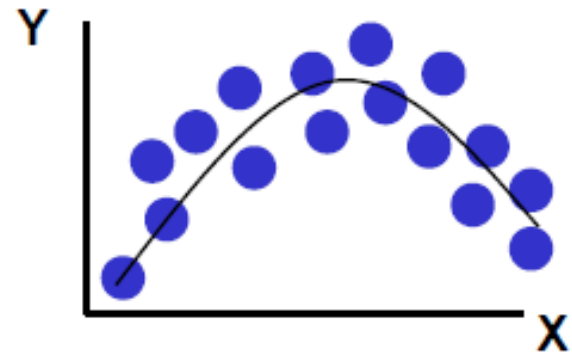


# Types of Relationships

Linear relationships

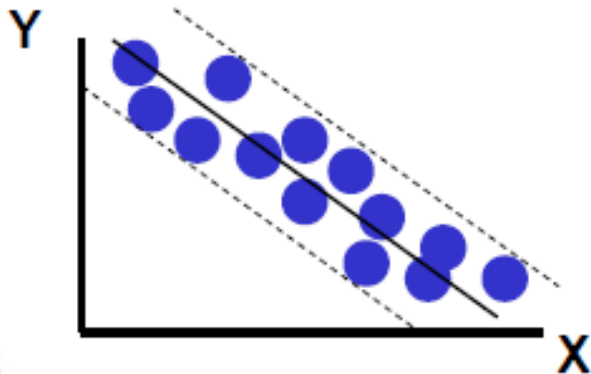
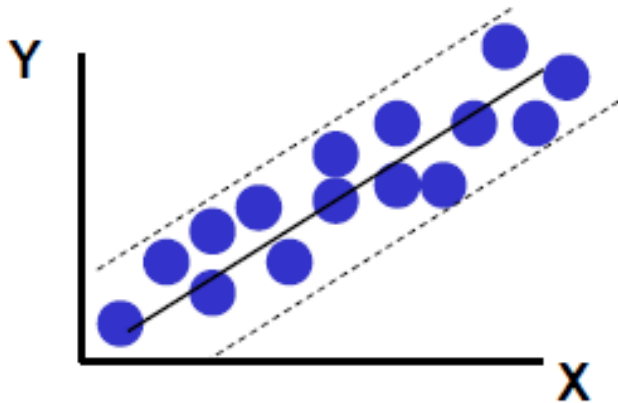


Curvilinear relationships

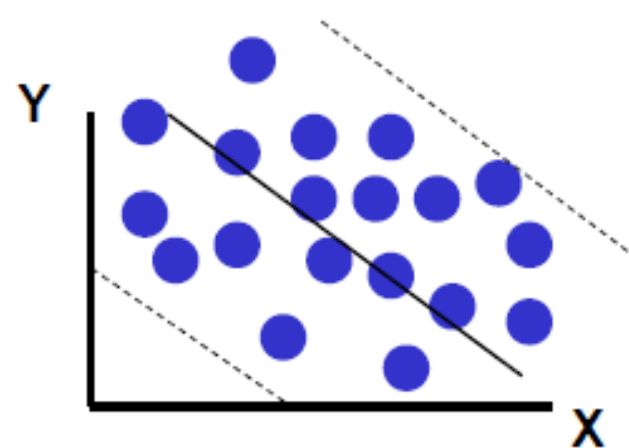
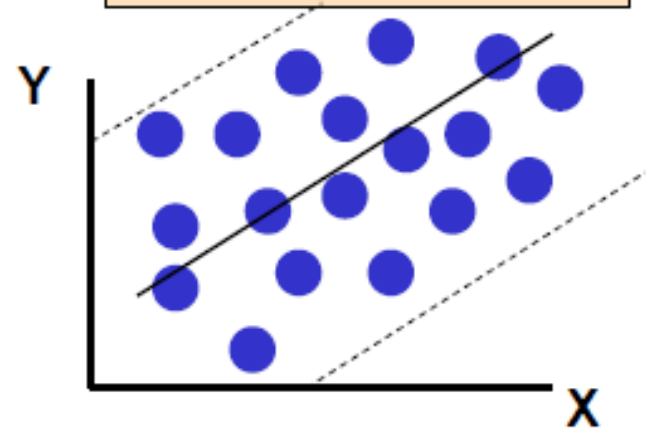


# Types of Relationships

Strong relationships

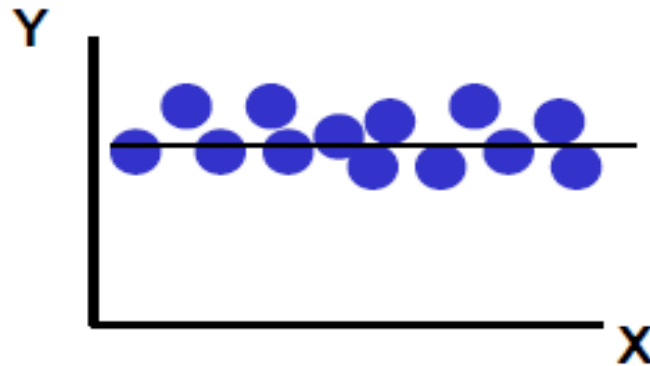
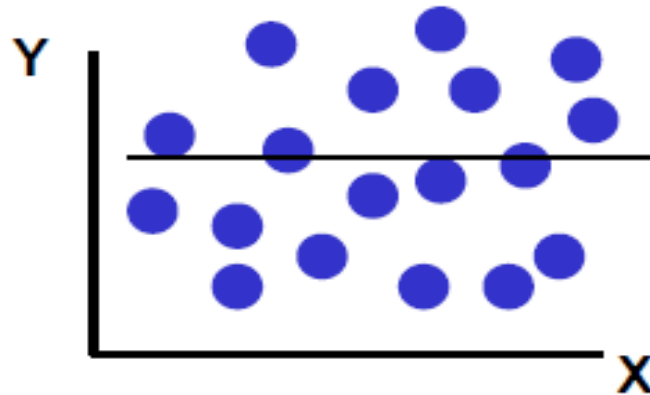


Weak relationships

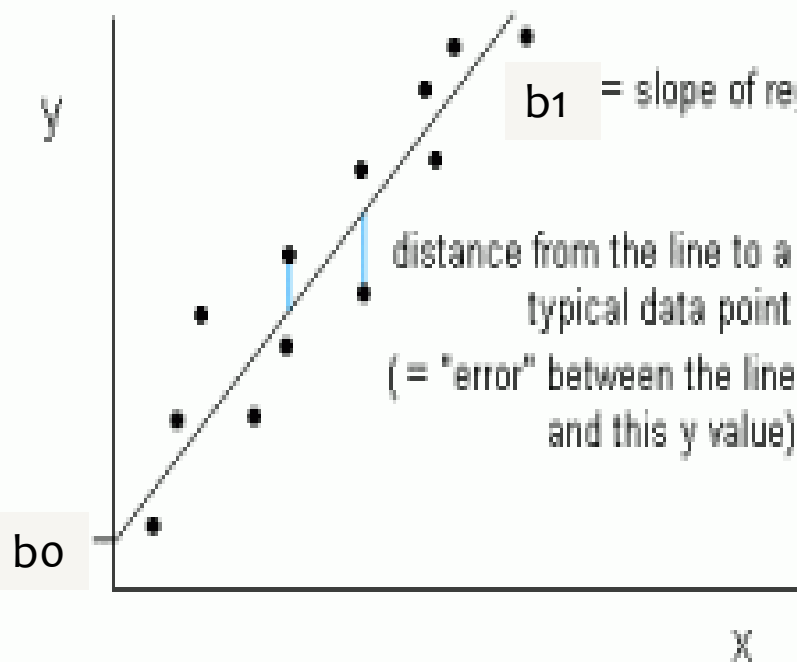


# Types of Relationships

No relationship



# Simple Linear Regression Equation



Estimated  
(or predicted)  
Y value for  
observation  $i$

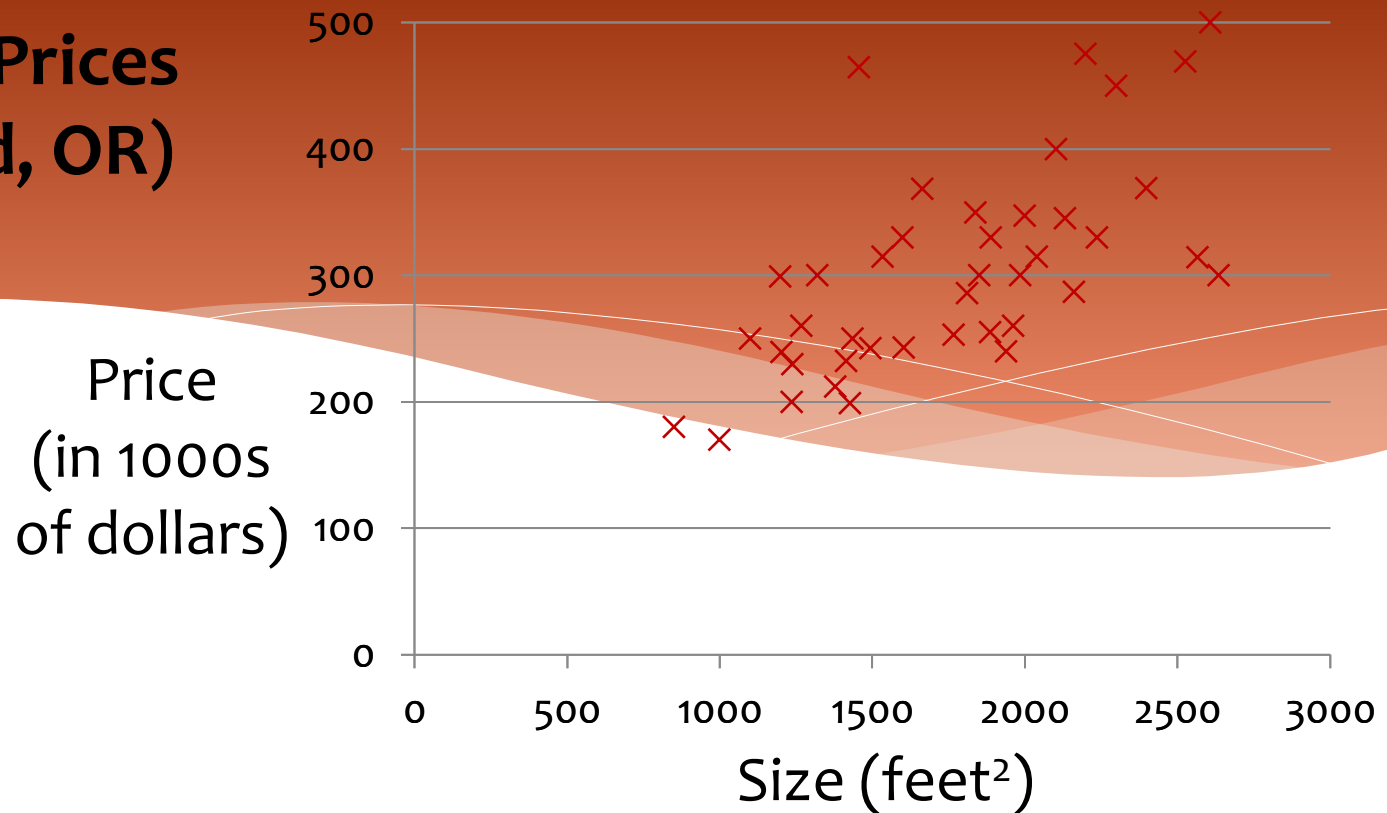
Estimate of  
the regression  
intercept

Estimate of the  
regression slope

Value of X for  
observation  $i$

$$\hat{Y}_i = b_0 + b_1 X_i$$

# Housing Prices (Portland, OR)



## Supervised Learning

Given the “right answer” for each example in the data.

## Regression Problem

Predict real-valued output

**Training set of  
housing prices  
(Portland, OR)**

**Size in feet<sup>2</sup>  
(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

Training Set

Size in feet<sup>2</sup>  
(x)

Price (\$) in  
1000's (y)

2104

460

1416

232

1534

315

852

178

...

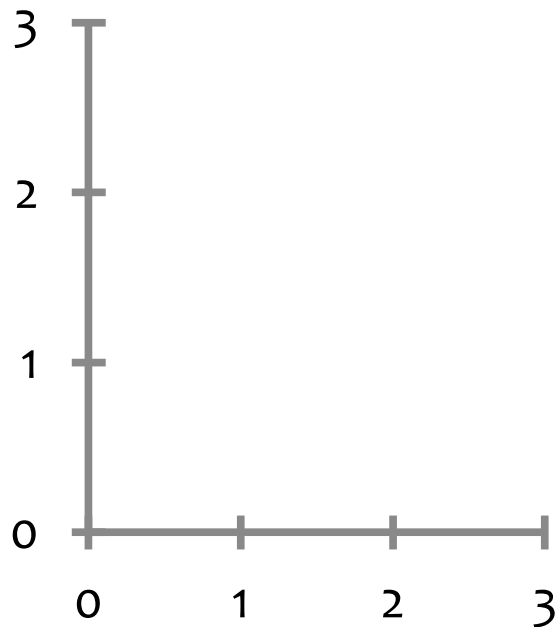
...

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

$\theta_i$ 's: Parameters

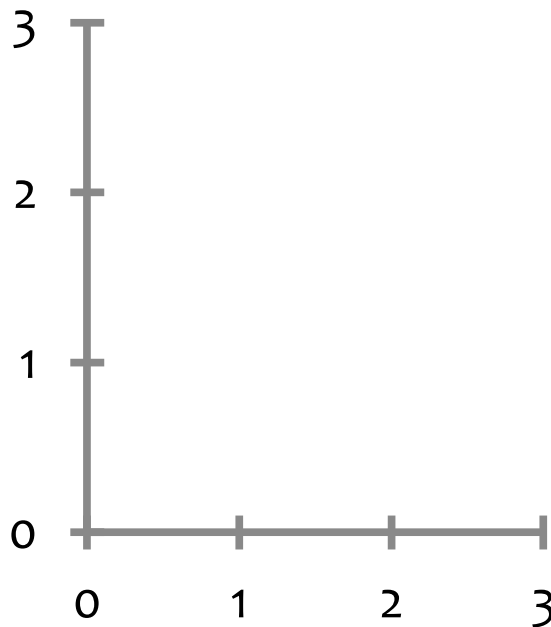
How to choose  $\theta_i$ 's ?

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



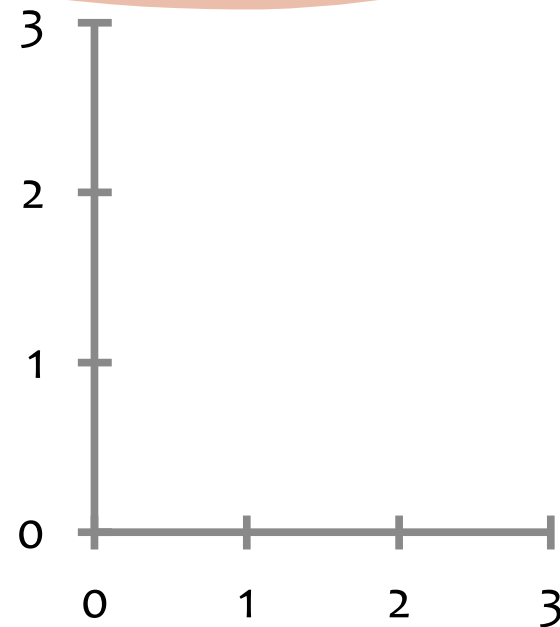
$$\theta_0 = 1.5$$

$$\theta_1 = 0$$



$$\theta_0 = 0$$

$$\theta_1 = 0.5$$

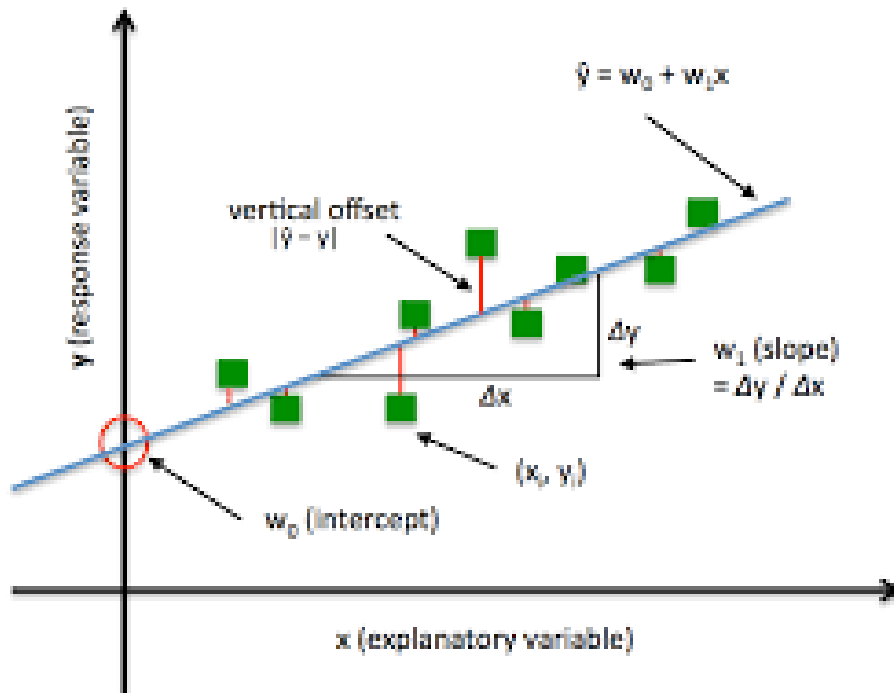


$$\theta_0 = 1$$

$$\theta_1 = 0.5$$



# Error between Actual and Predicted value/ Cost function

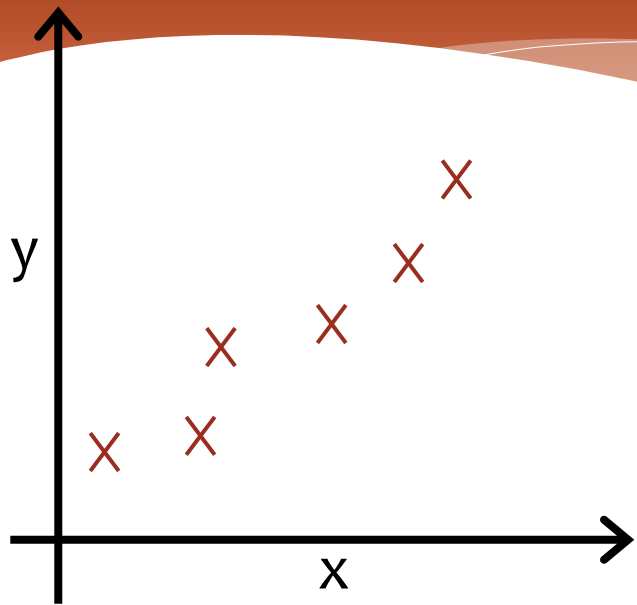


$$J = \frac{1}{2m} \sum_{i=1}^m (\hat{y} - y)^2$$

$\hat{y}$  is predicted value

$Y$  is observed or actual value

Goal is to minimize error or cost function between predicted and actual value



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

Hypothesis:

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

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

Parameters:

$$\theta_0, \theta_1$$

$$\theta_1$$

Cost Function:

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

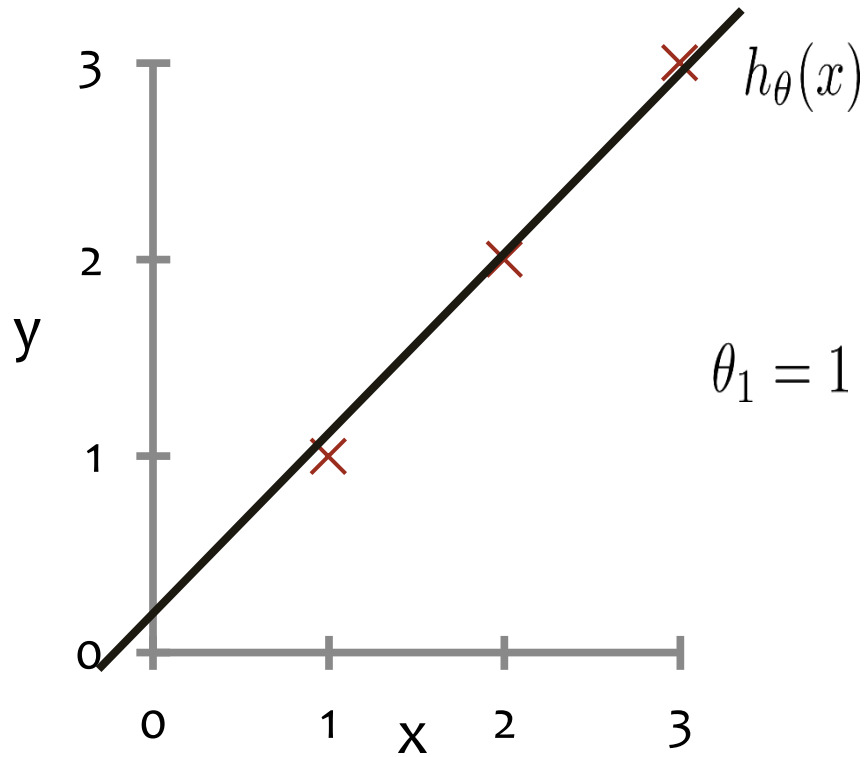
$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

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

minimize  $J(\theta_1)$   
 $\theta_1$

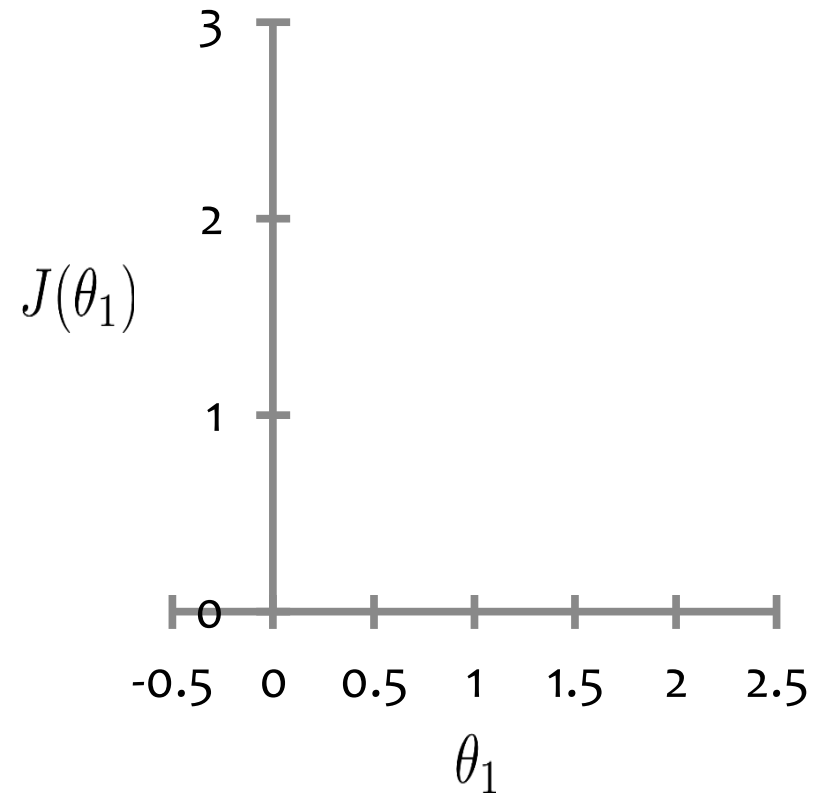
$$h_{\theta}(x)$$

(for fixed  $\theta_1$ , this is a function of  $x$ )



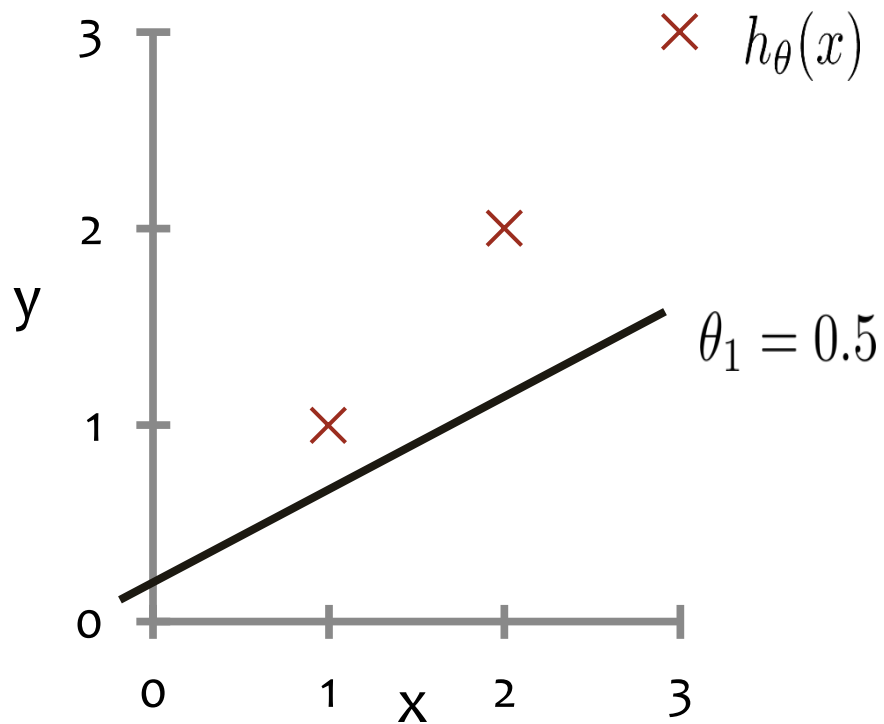
$$J(\theta_1)$$

(function of the parameter  $\theta_1$ )



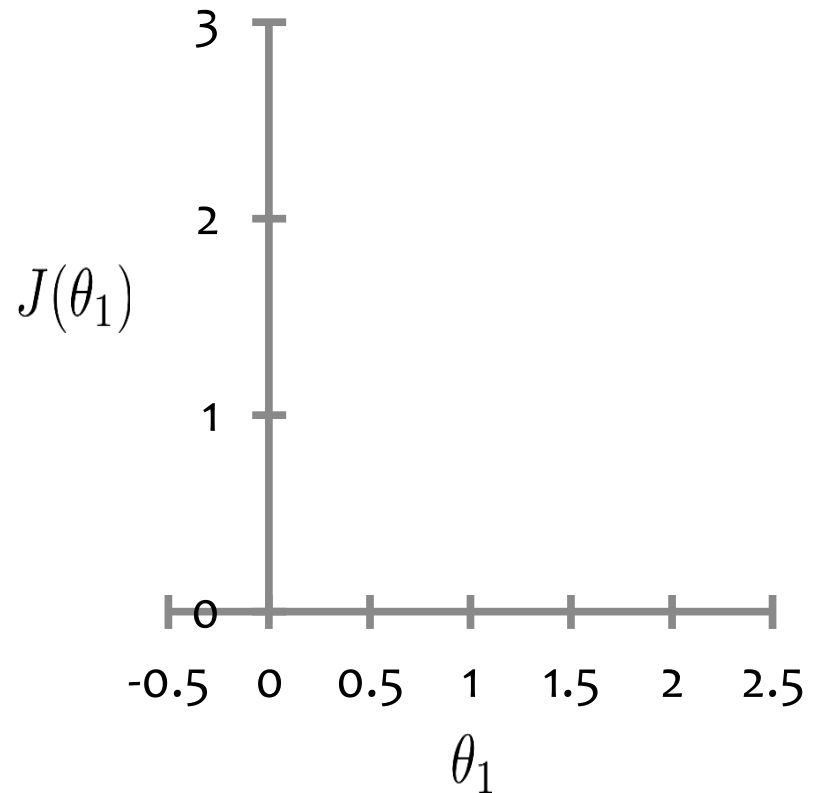
$$h_{\theta}(x)$$

(for fixed  $\theta_1$ , this is a function of  $x$ )



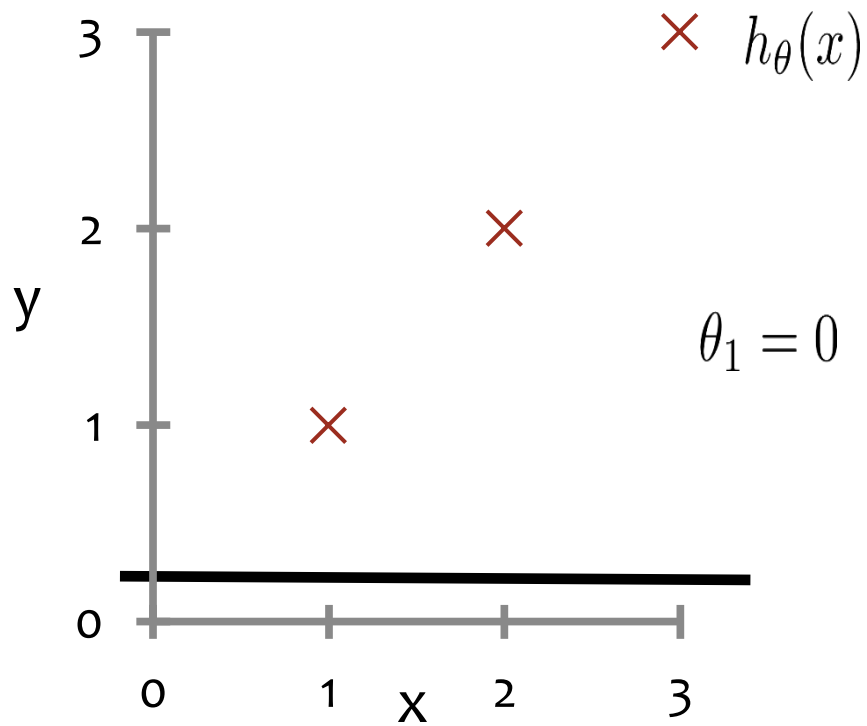
$$J(\theta_1)$$

(function of the parameter  $\theta_1$ )



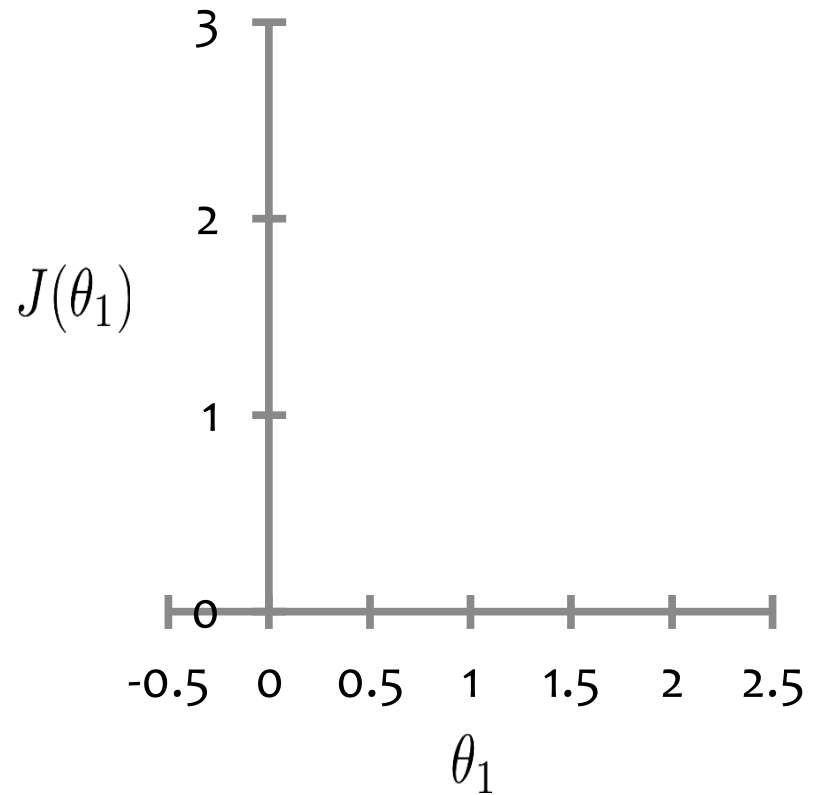
$$h_{\theta}(x)$$

(for fixed  $\theta_1$ , this is a function of  $x$ )



$$J(\theta_1)$$

(function of the parameter  $\theta_1$ )



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

Parameters:  $\theta_0, \theta_1$

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

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

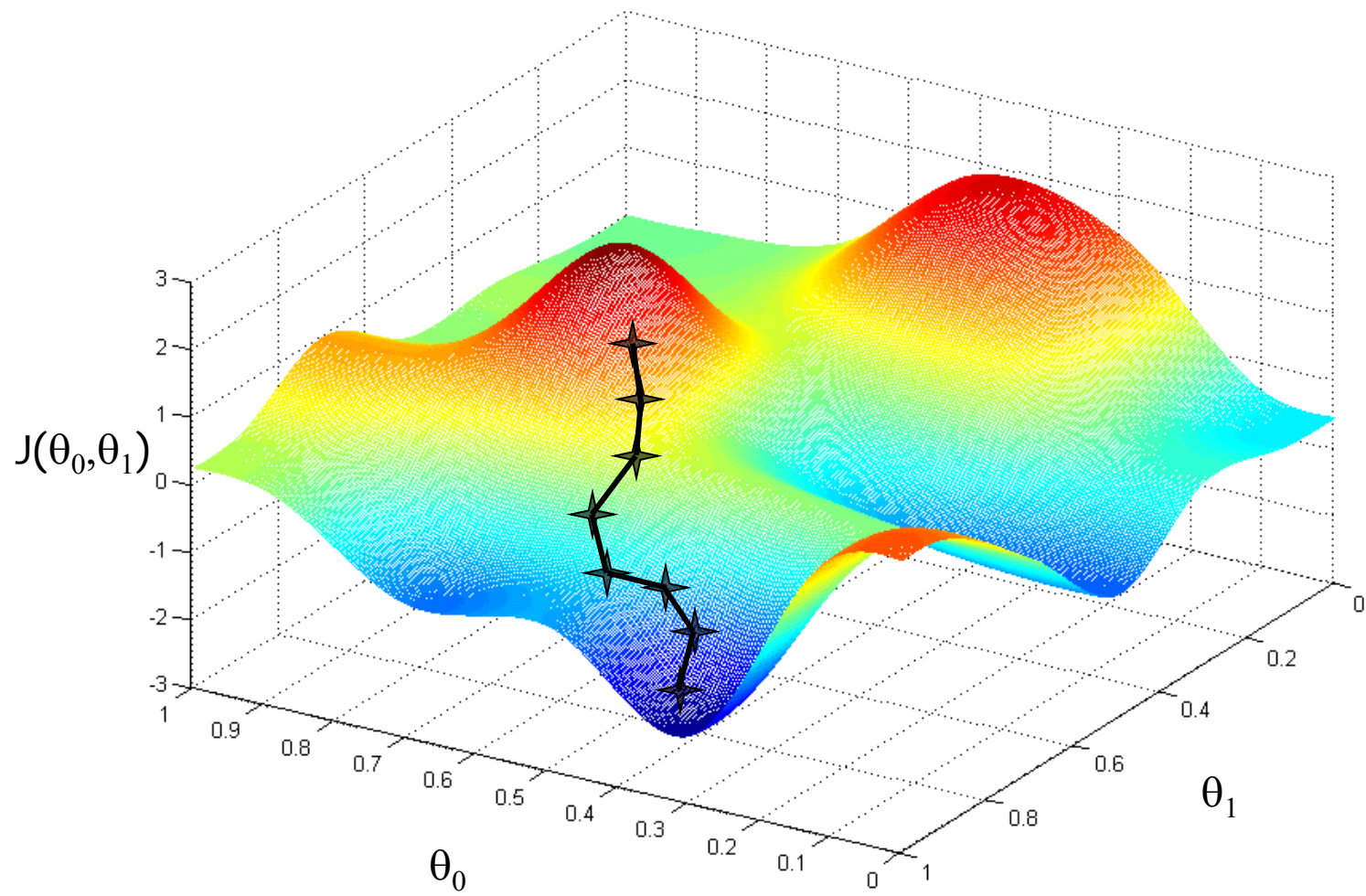
Have some function  $J(\theta_0, \theta_1)$

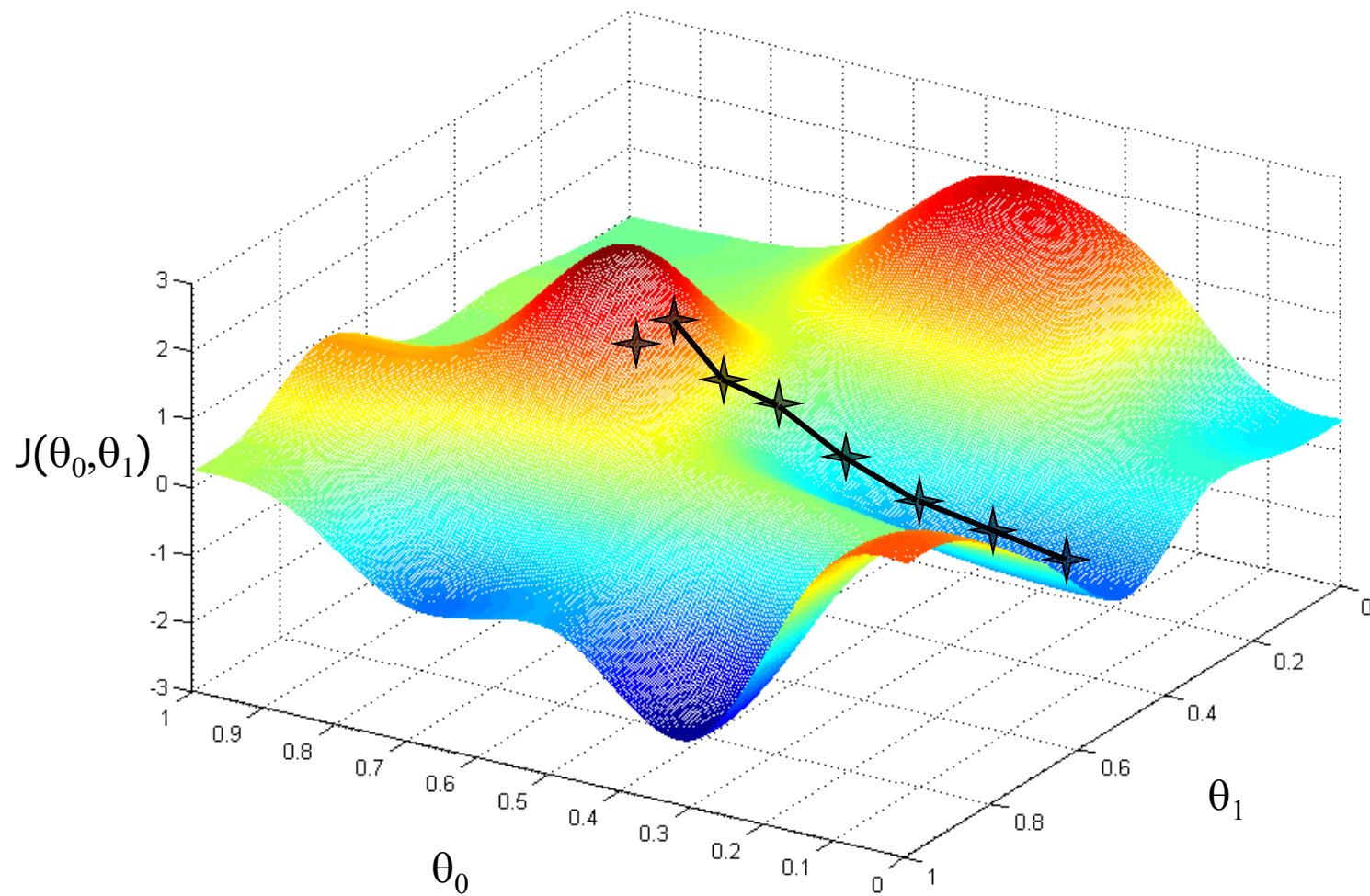
Want  $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

## Outline:

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





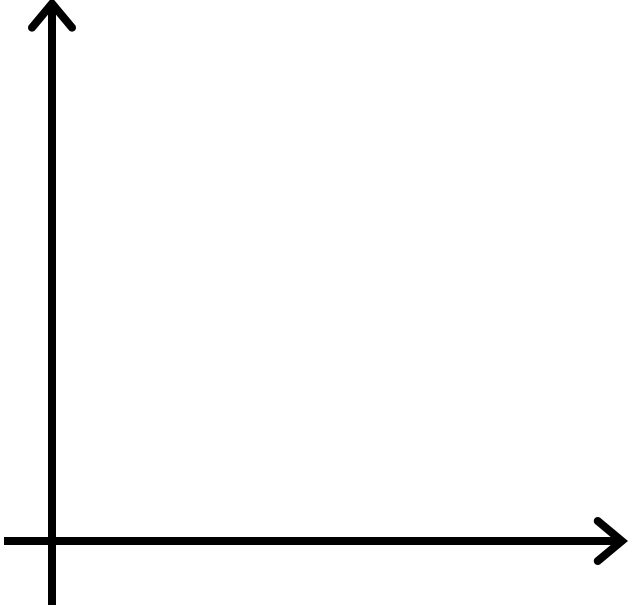
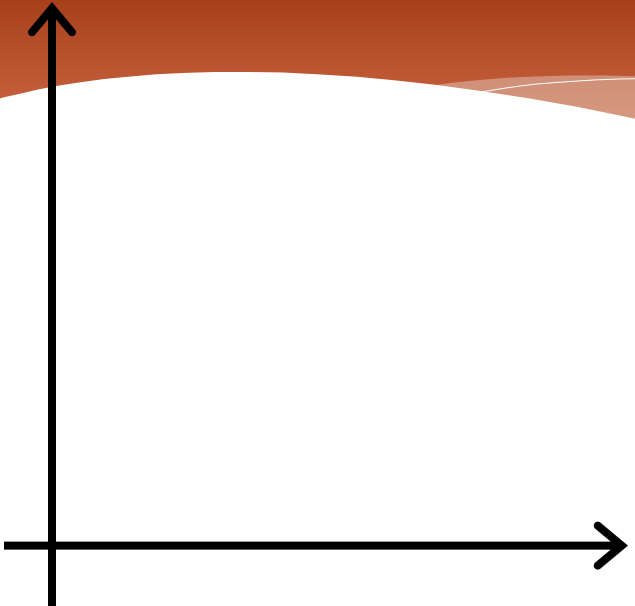
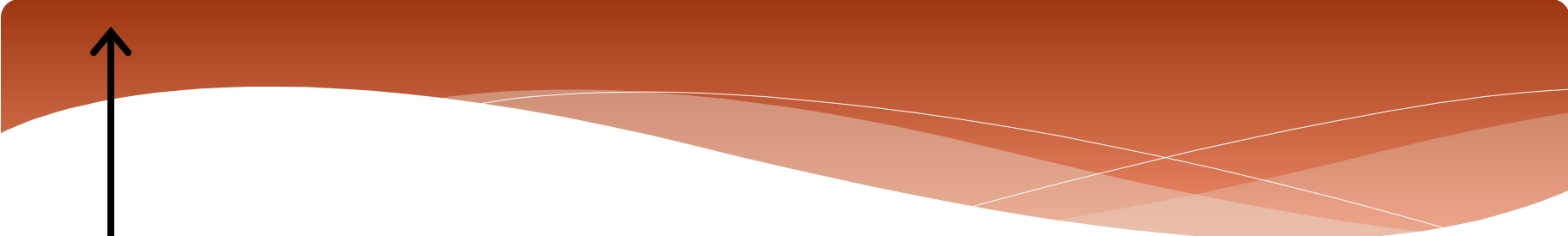


# Gradient Descent Algorithm

- \* **Gradient descent** is **used** to minimize a cost function  $J(\theta)$  or  $J(w)$  parameterized by a model parameters  $\theta$  or  $w$ . The **gradient** (or derivative) tells us the incline or slope of the cost function. Hence, to minimize the cost function, we move in the direction opposite to the **gradient**.

# Gradient Descent Algorithm

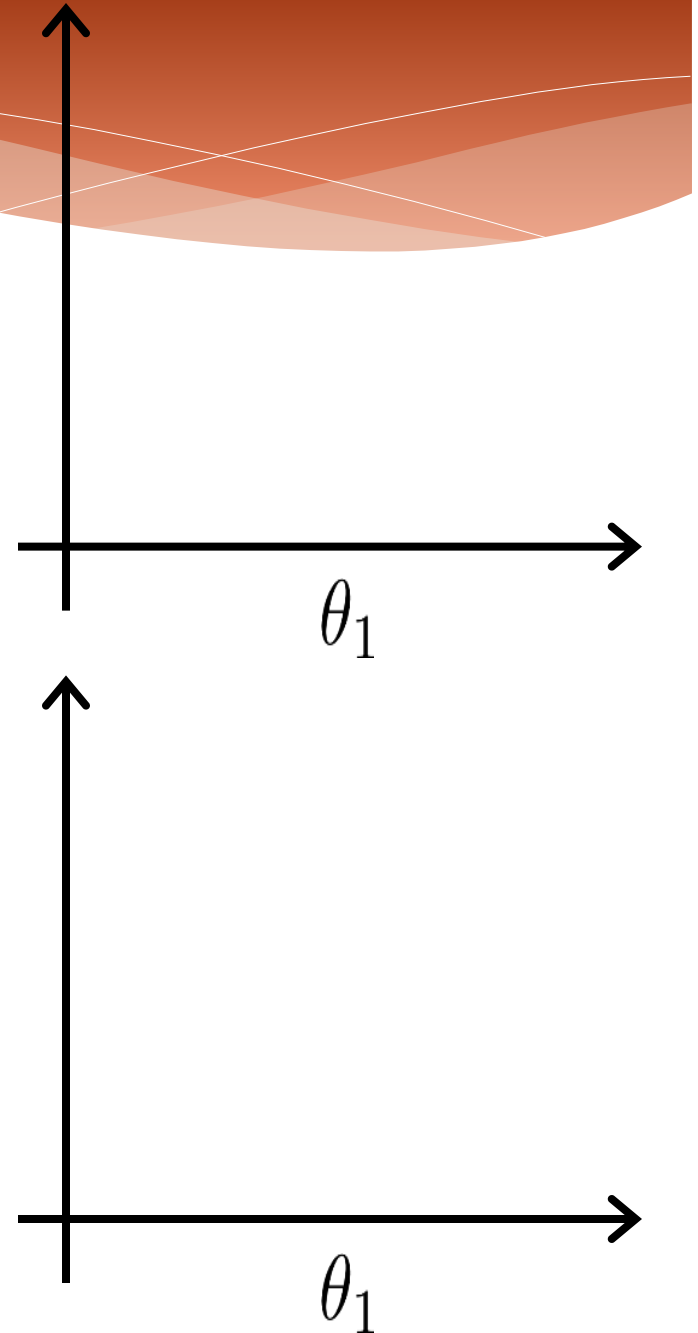
repeat until convergence {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$       (for  $j = 0$  and  $j = 1$ )  
}



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

If  $\alpha$  is too small, gradient descent can be slow.

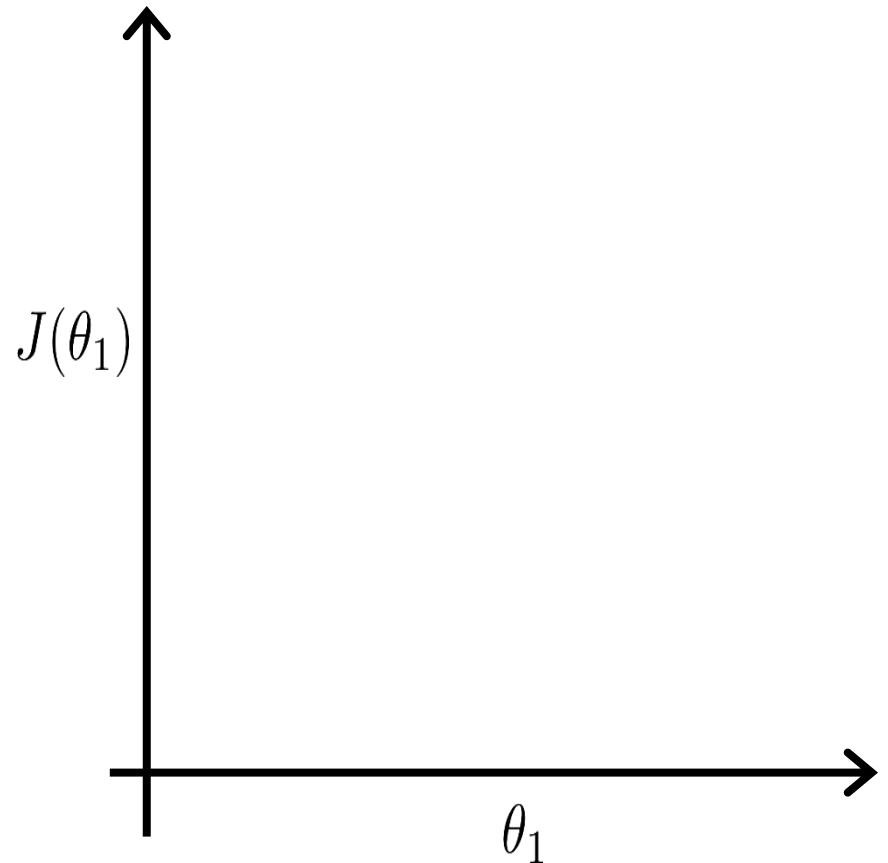
If  $\alpha$  is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.



Gradient descent can converge to a local minimum, even with the learning rate  $\alpha$  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  $\alpha$  over time.



# Gradient Descent For Linear Regression

Gradient descent algorithm

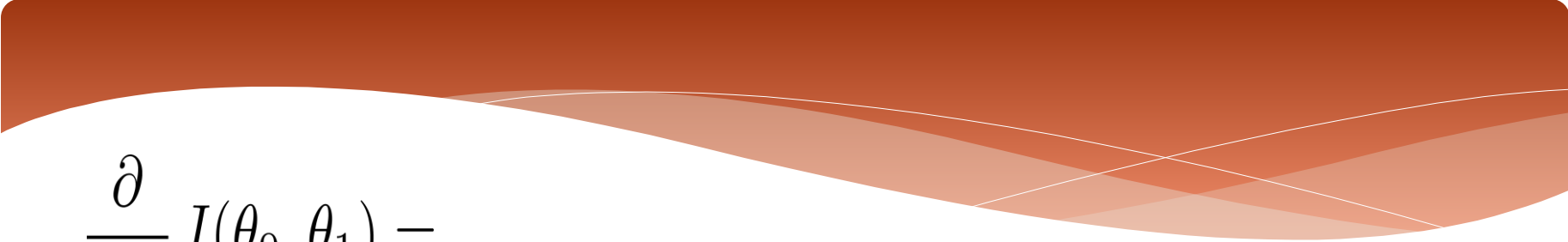
repeat until convergence {  
     $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$   
    (for  $j = 1$  and  $j = 0$ )  
}

Linear Regression Model

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

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




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

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

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

# Gradient descent algorithm

repeat until convergence {

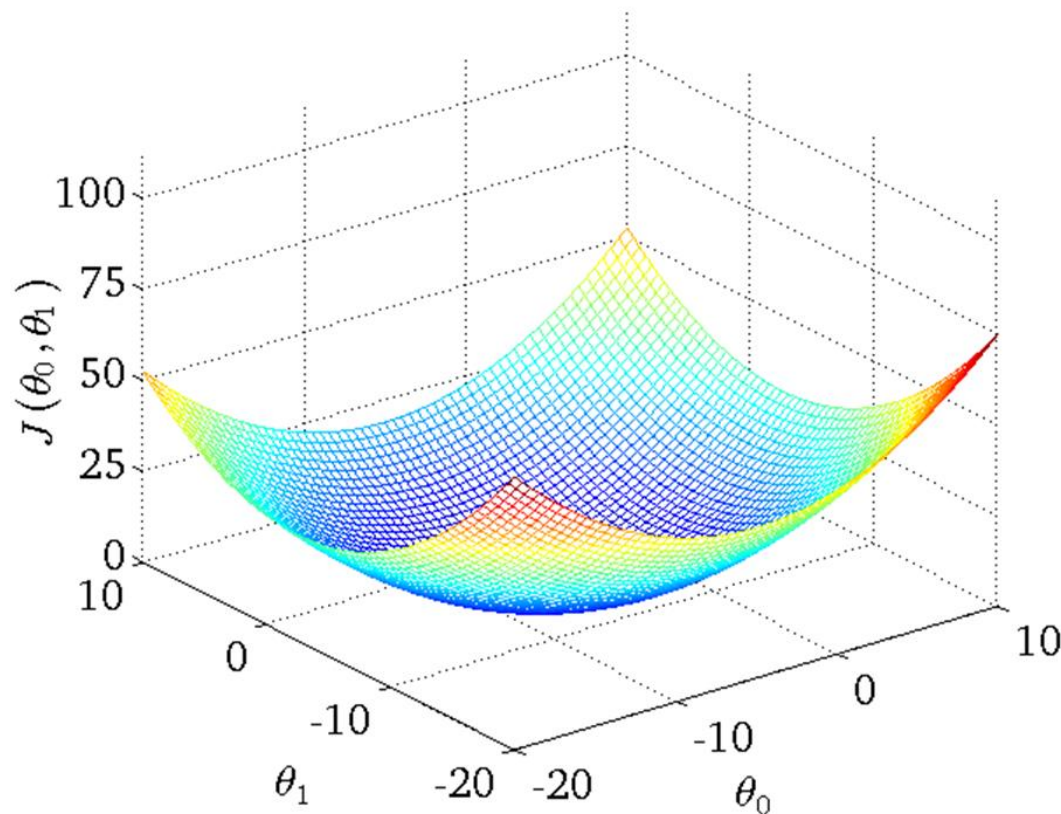
$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})$$

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

update  
and  
 $\theta_0$  and  $\theta_1$   
simultaneously

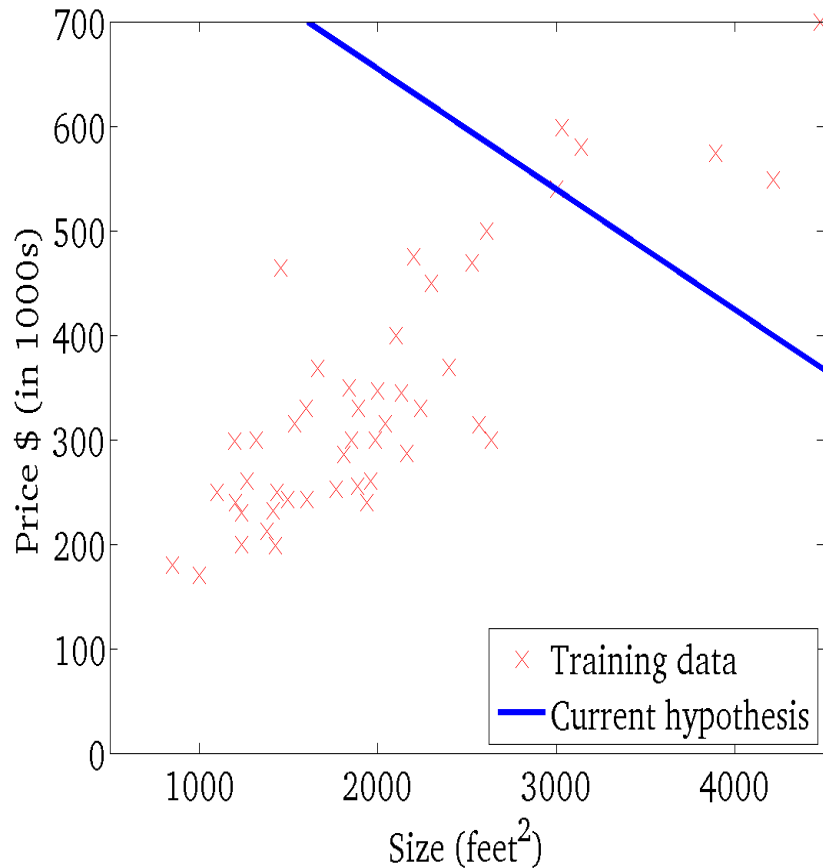
}

# Convex Function



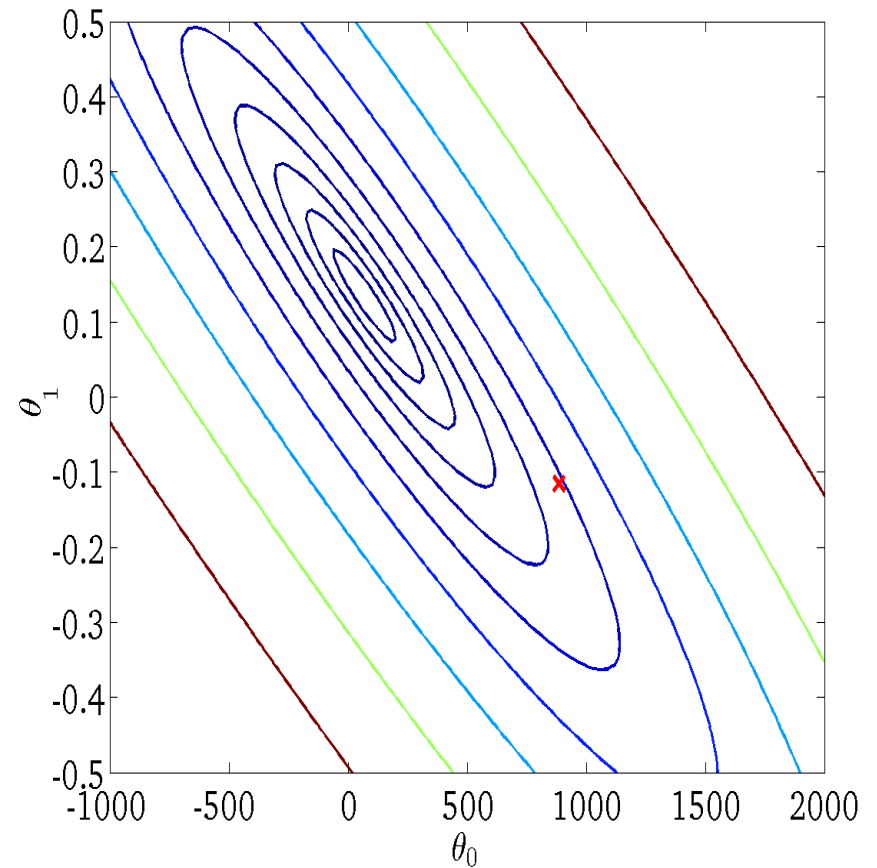
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



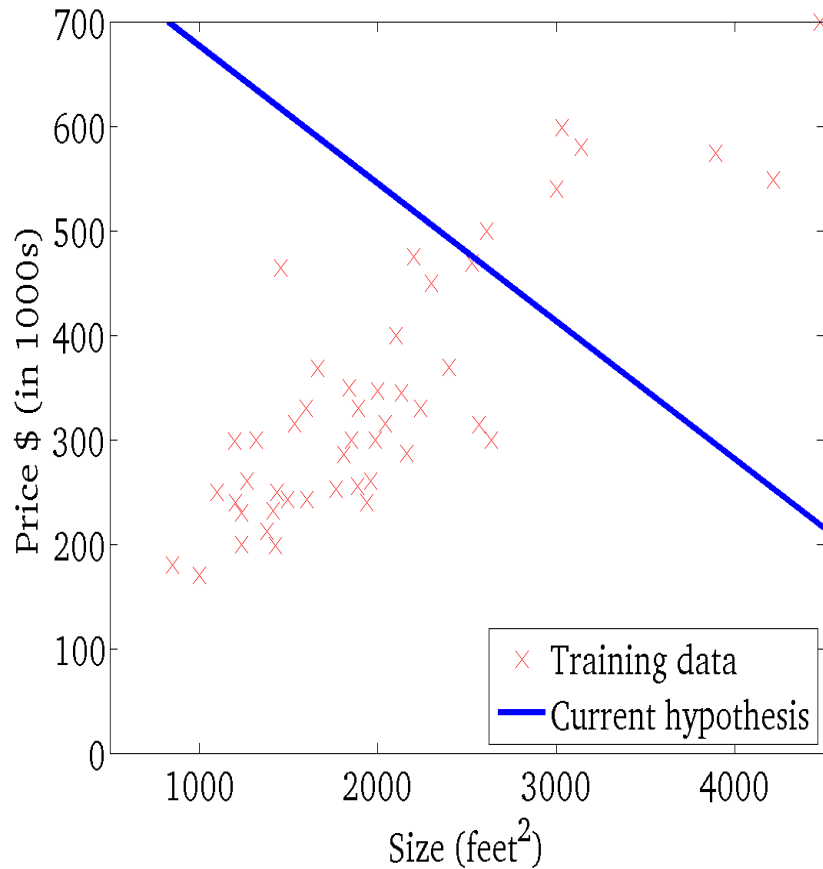
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



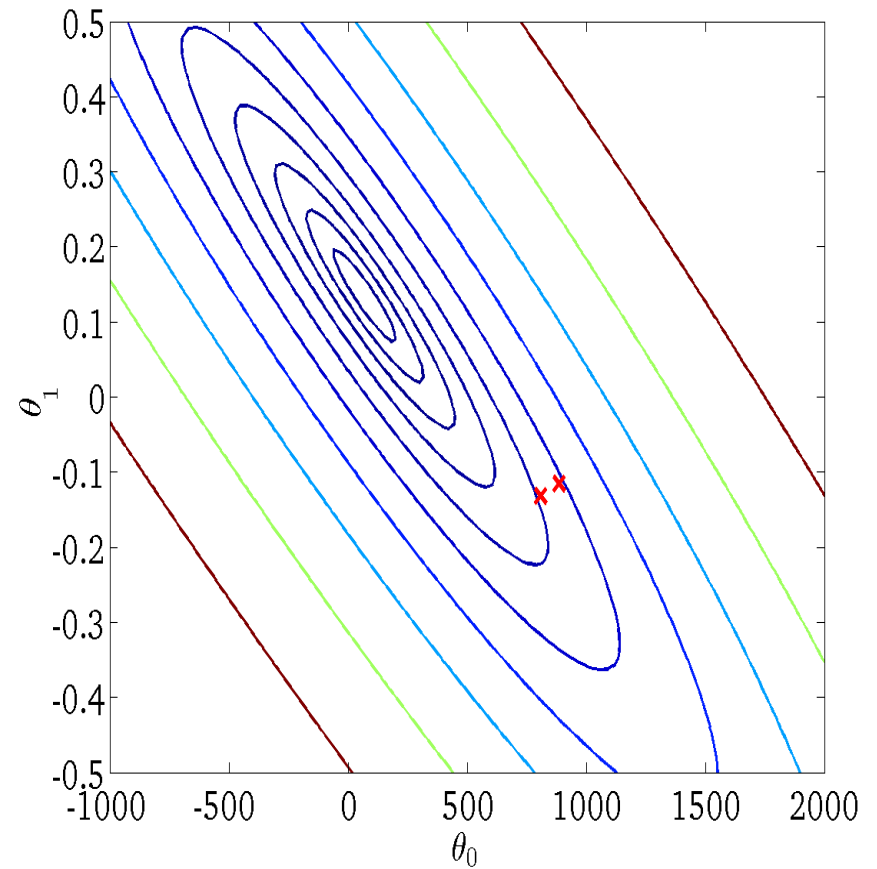
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



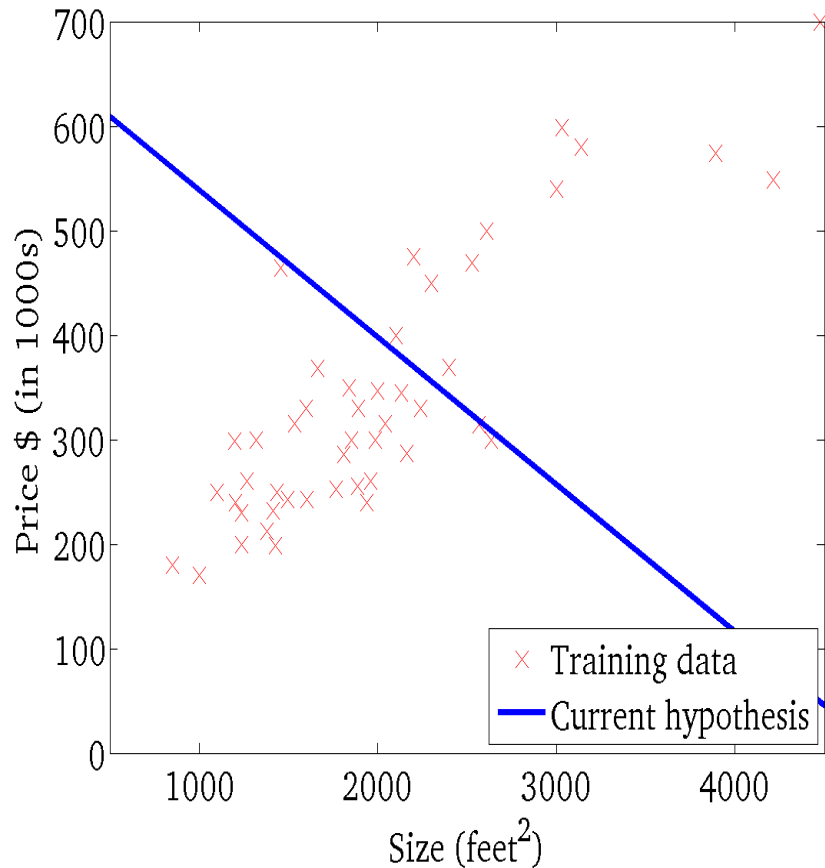
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



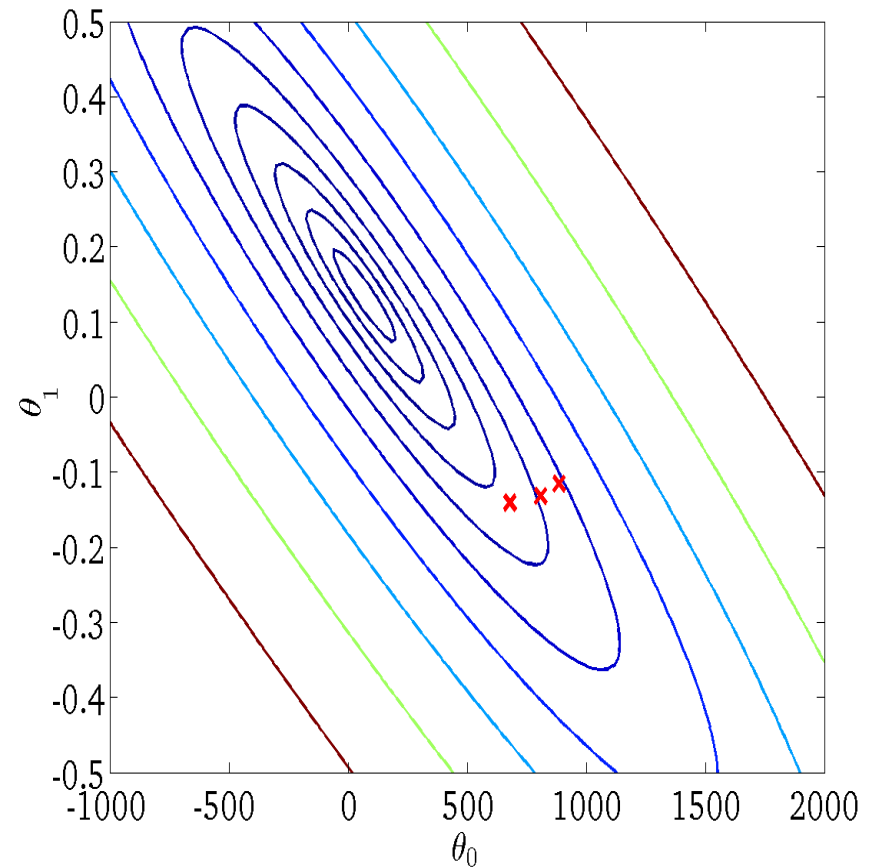
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



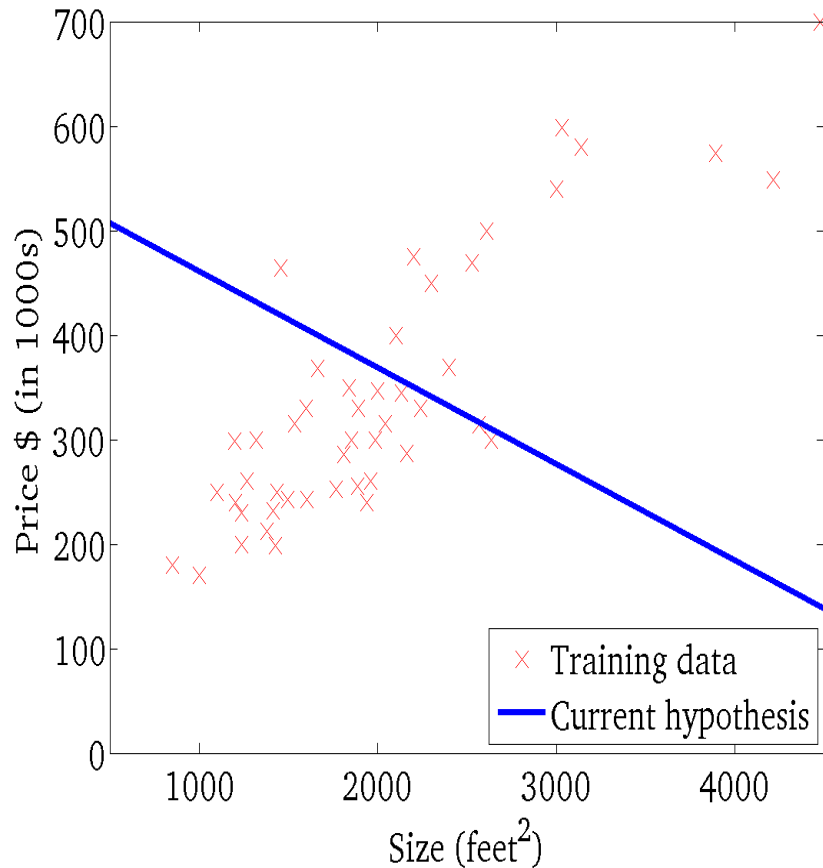
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



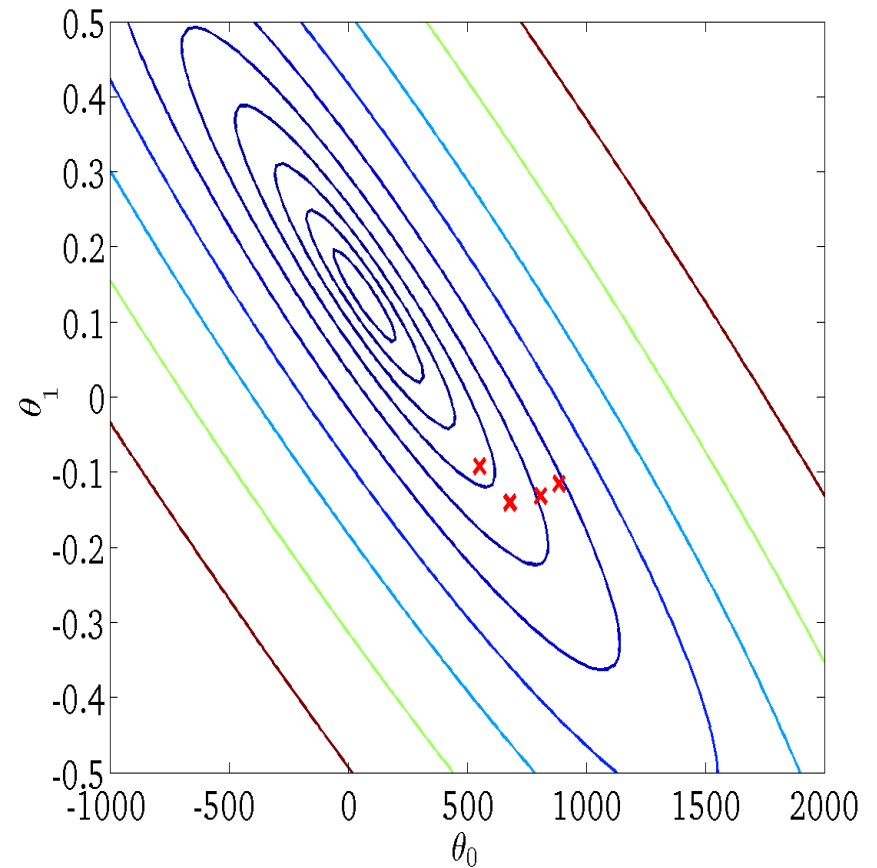
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



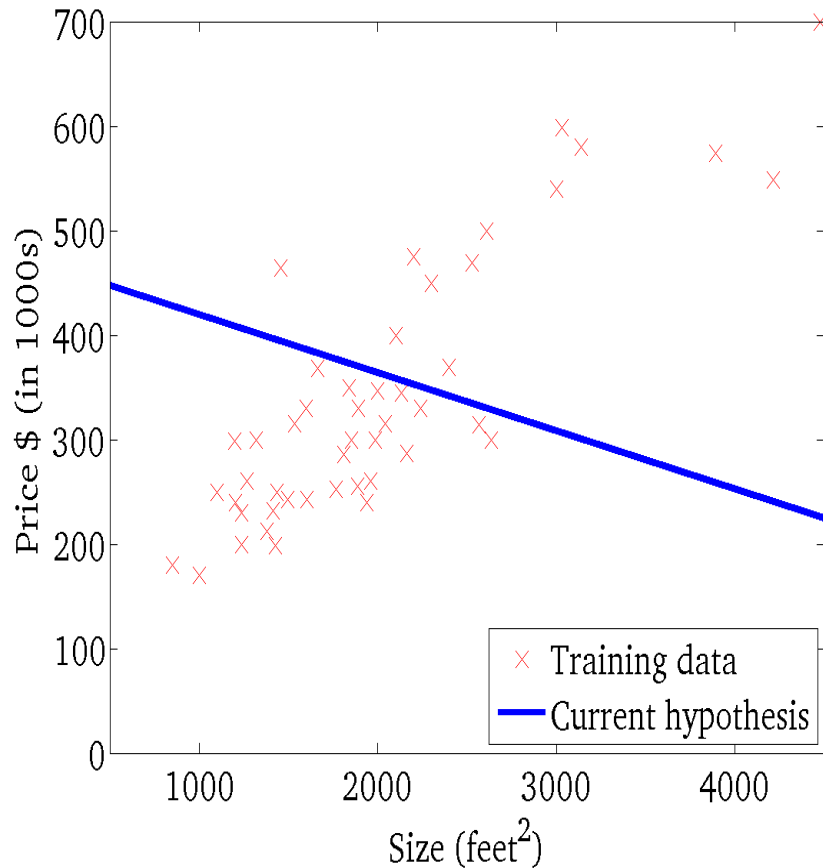
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



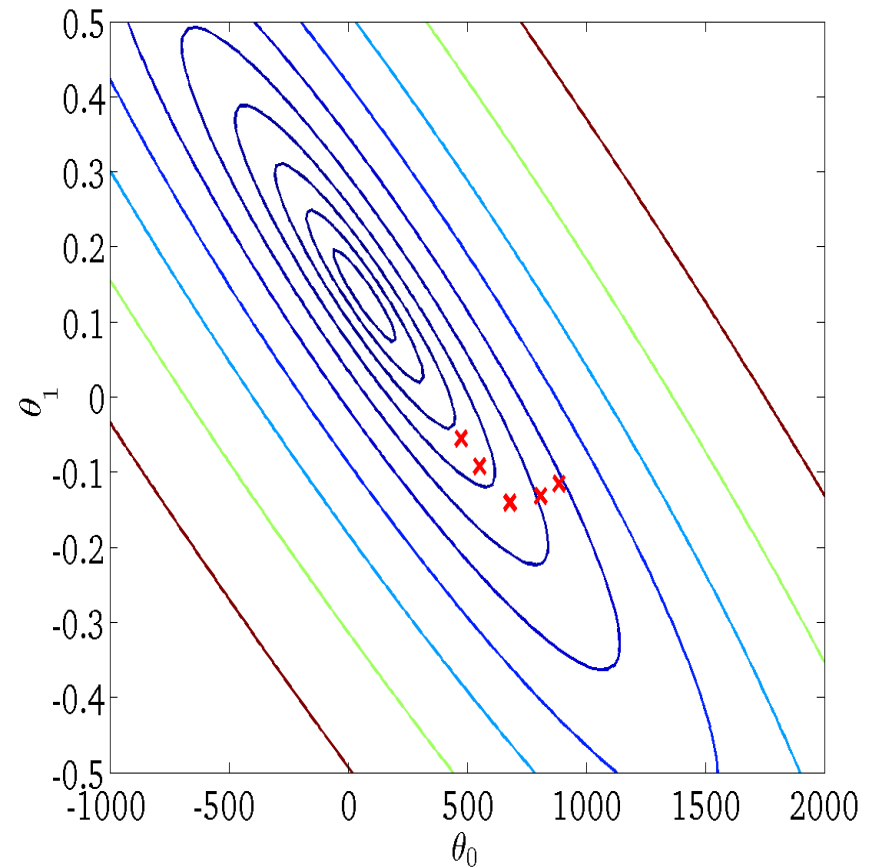
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



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

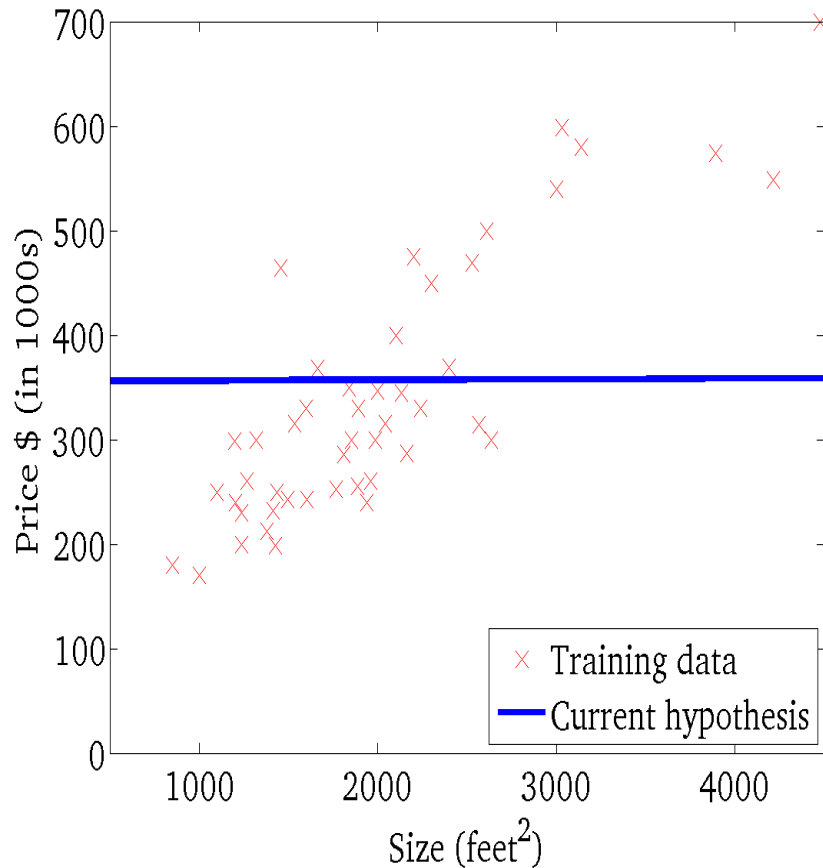
(function of the parameters  $\theta_0, \theta_1$ )





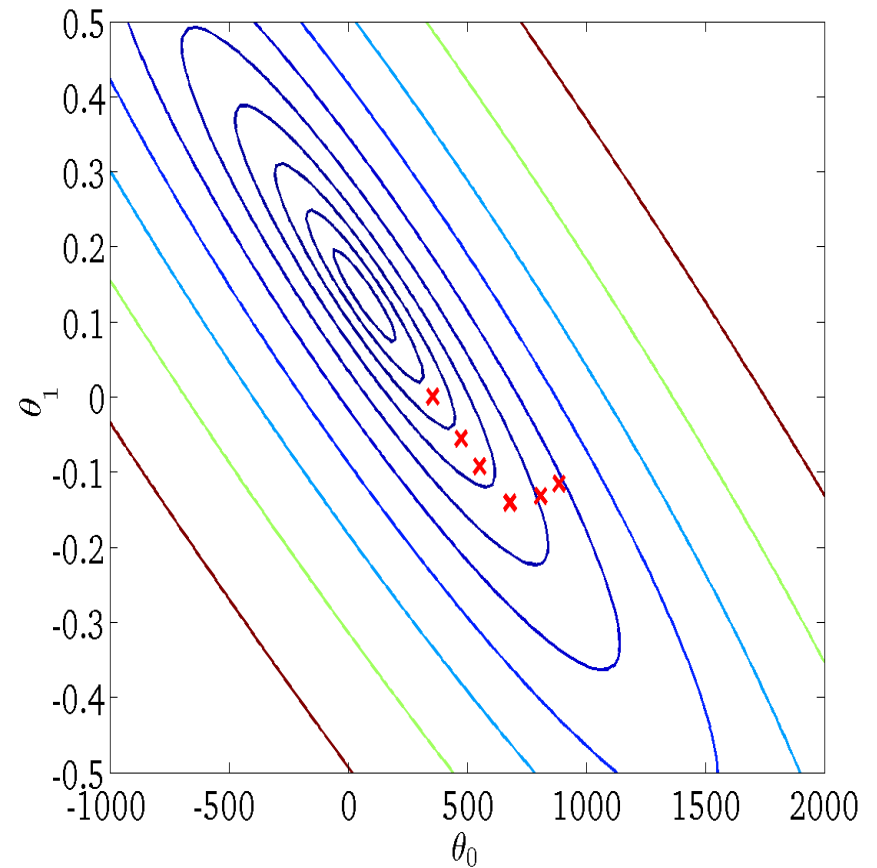
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



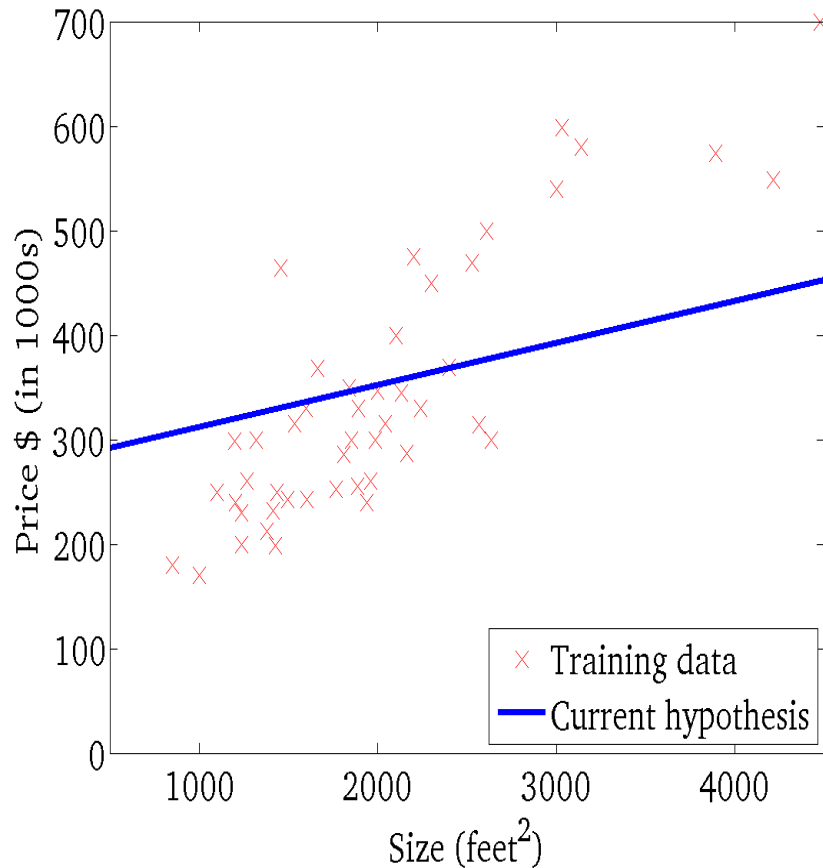
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



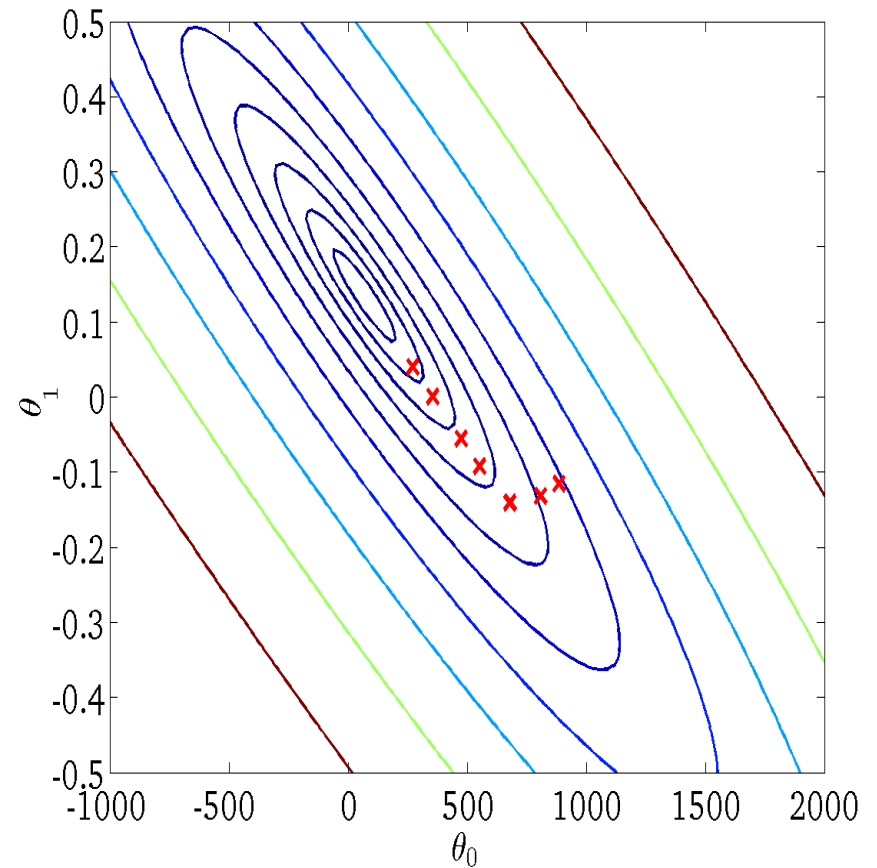
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



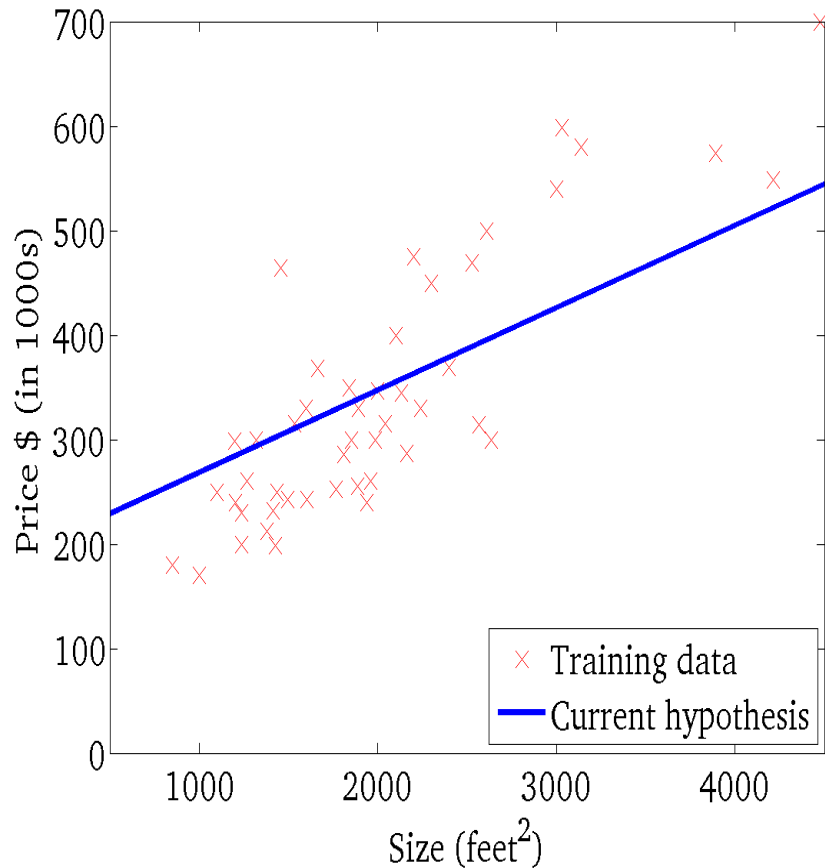
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



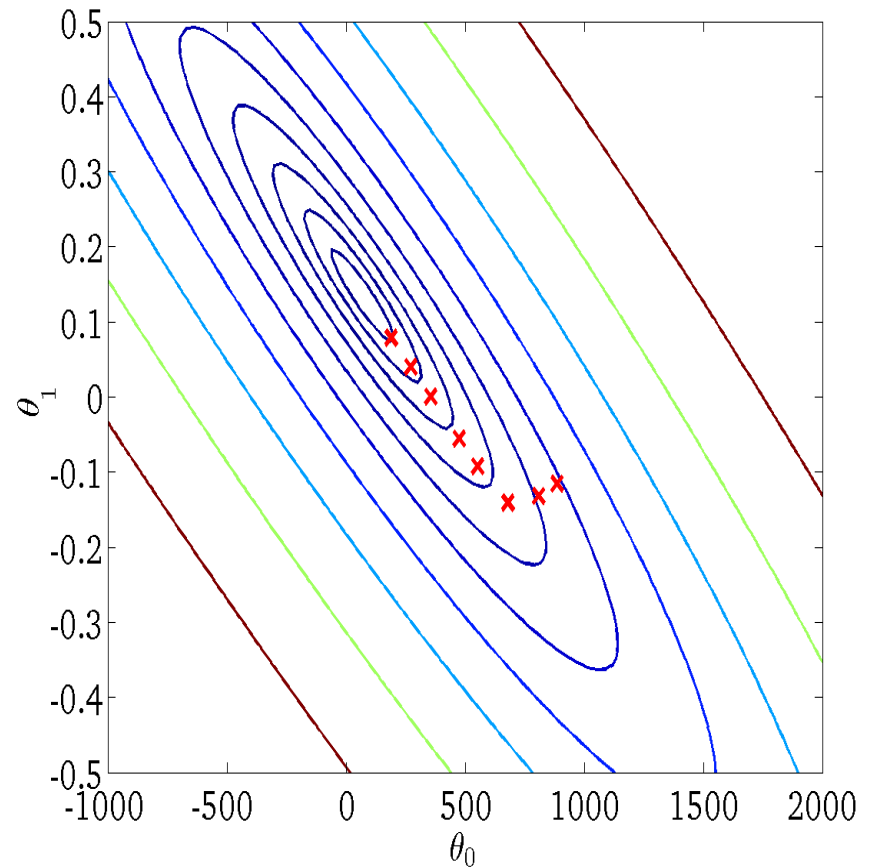
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



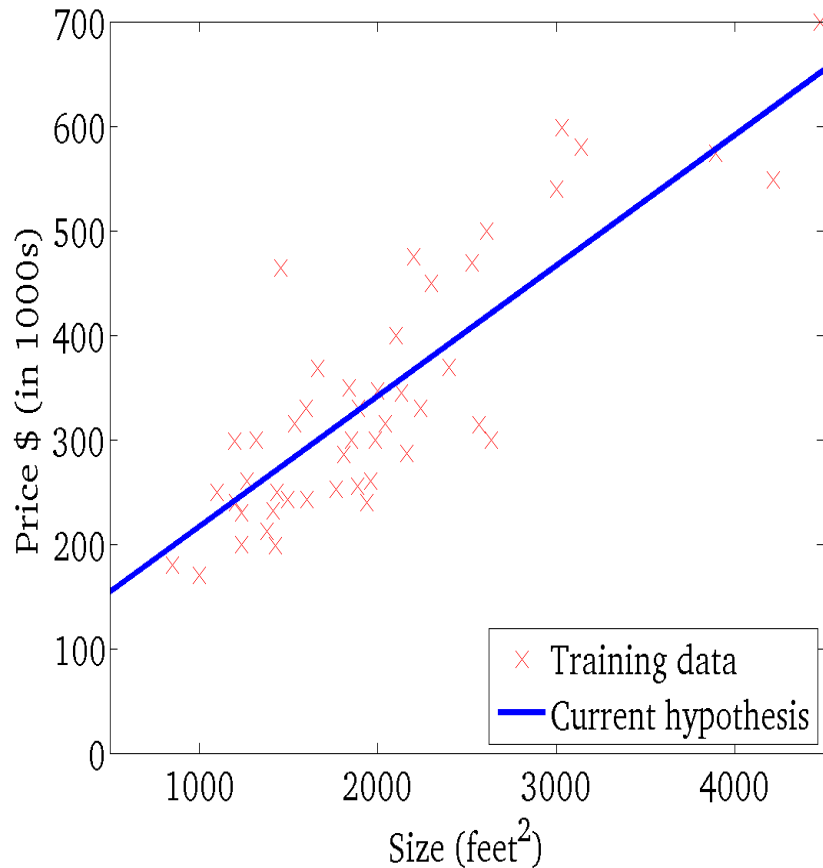
$$J(\theta_0, \theta_1)$$

(function of the parameters  $\theta_0, \theta_1$ )



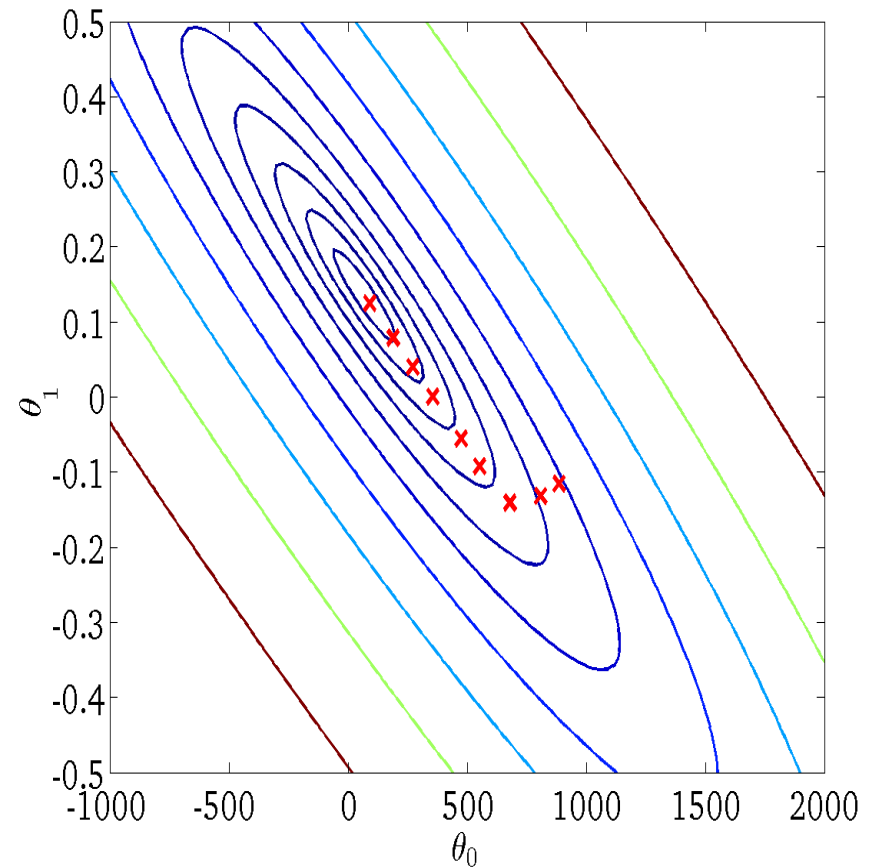
$$h_{\theta}(x)$$

(for fixed  $\theta_0, \theta_1$ , this is a function of  $x$ )



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

(function of the parameters  $\theta_0, \theta_1$ )



# Multiple Linear Regression

- \* In multiple linear regression, the dependent variable depends on more than one independent variables.
- \*
- \* For multiple linear regression, the form of the model is-
- \*  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n$
- \* Y is a dependent variable.
- \*  $X_1, X_2, \dots, X_n$  are independent variables.
- \*  $\beta_0, \beta_1, \dots, \beta_n$  are the regression coefficients.
- \*  $\beta_j$  ( $1 \leq j \leq n$ ) is the slope or weight that specifies the factor by which  $X_j$  has an impact on Y.

## Multiple features (variables).

Size (feet <sup>2</sup> )	Number of bedroom s	Number of floors	Age of home (years)	Price (\$1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Notation:

$n$  = number of features

$x^{(i)}$  = input (features) of  $i^{th}$  training example.

$x_j^{(i)}$  = value of feature  $j$  in  $i^{th}$  training example.

## Hypothesis:

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

For Multiple Linear Regression

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

# Gradient Descent For Multiple Regression

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:  $\theta_0, \theta_1, \dots, \theta_n$

Cost function:

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

Gradient descent:

Repeat {

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

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



# Gradient Descent

Previously (n=1):

Repeat {

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

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

(simultaneously update  $\theta_0, \theta_1$ )

}

New algorithm ( $n \geq 1$ ):

Repeat {

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

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

}

---

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

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

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

...

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

For convenience of notation, define  $x_0 = 1$  .

Multivariate linear regression.

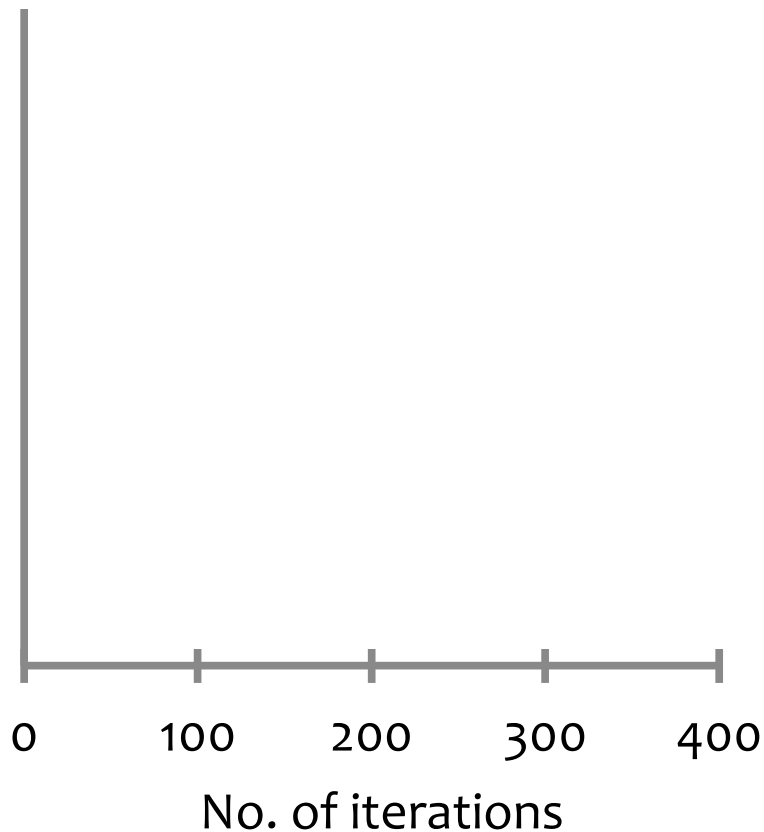
# Gradient descent

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

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

## Making sure gradient descent is working correctly.

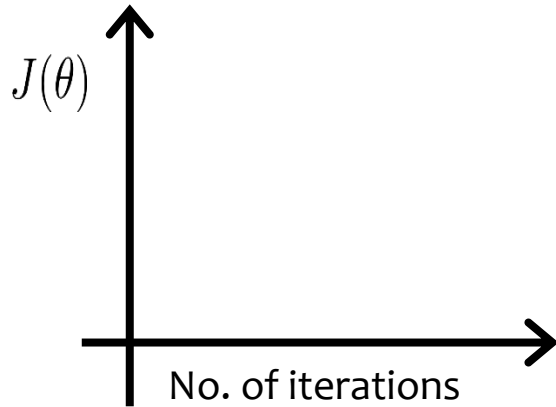
$$\min_{\theta} J(\theta)$$



Example automatic  
convergence test:

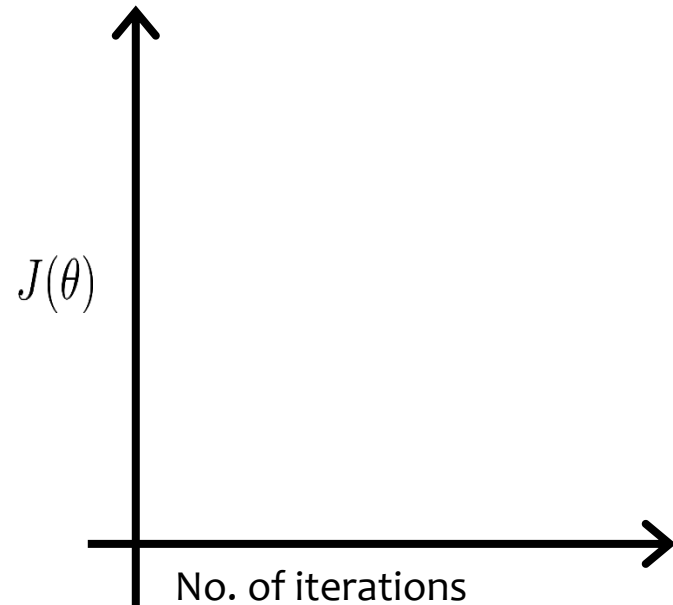
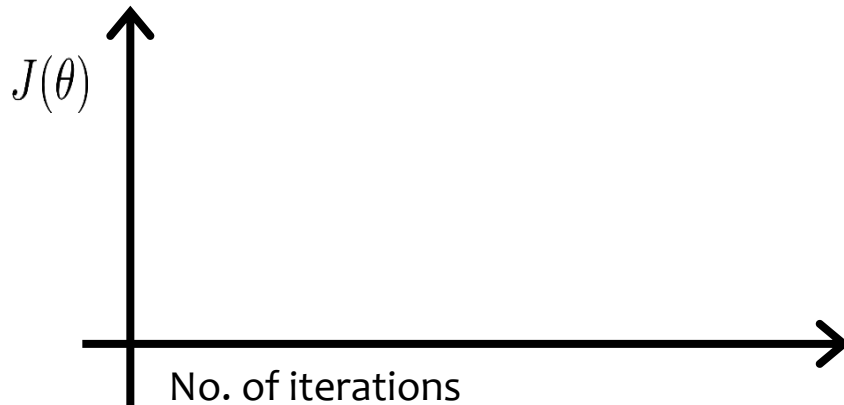
Declare convergence if  $J(\theta)$  decreases by less than  $10^{-3}$  in one iteration.

## Making sure gradient descent is working correctly.



Gradient descent not working.

Use smaller  $\alpha$ .



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

## Summary:

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

To choose  $\alpha$ , try

$\dots, 0.001, \quad , 0.01, \quad , 0.1, \quad , 1, \dots$

# R<sup>2</sup> value for Regression

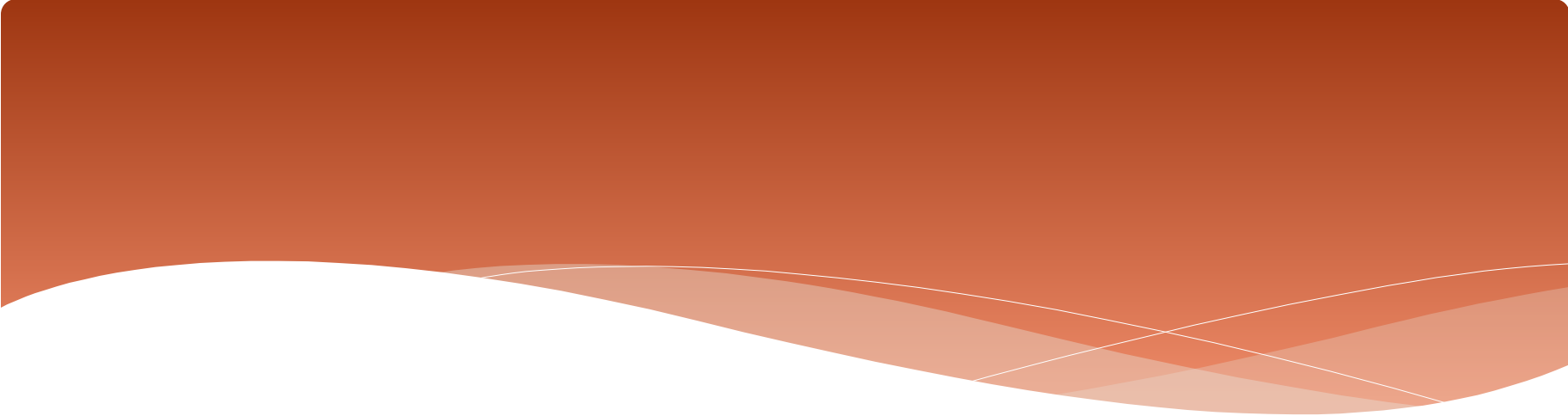
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

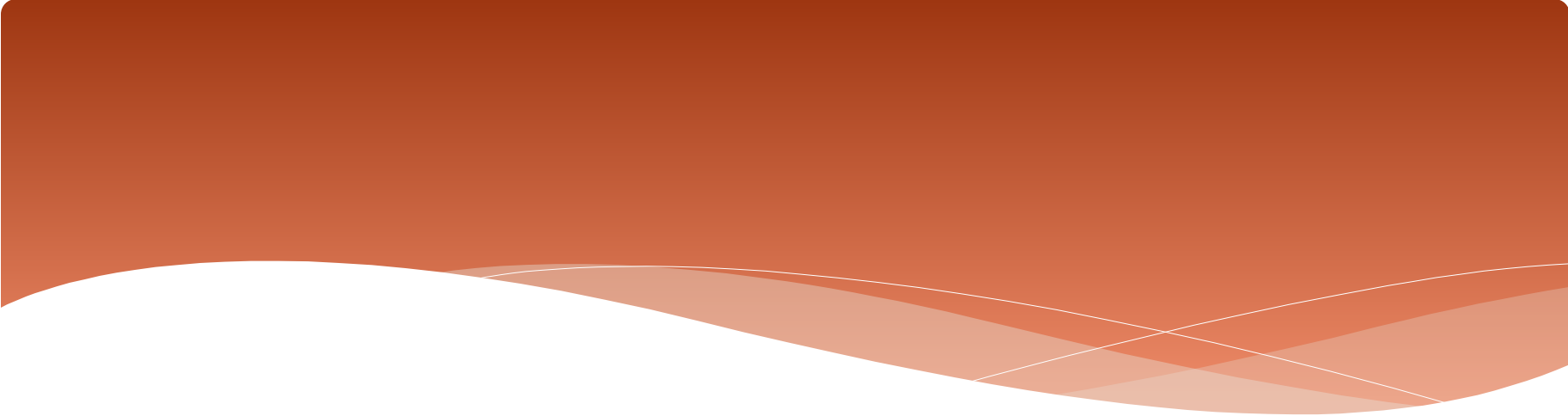
- \* As we keep on adding new features, R square value increases.

## \* R-squared and the Goodness-of-Fit

- \* R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.
- \* R-squared is the percentage of the dependent variable variation that a linear model explains.
- \* R-squared is always between 0 and 100%:



- 
- \* If independent variables are correlated, then  $R^2$  value is high
  - \* Hence  $(1 - R^2)$  in Numerator of adjusted  $R^2$  becomes smaller
  - \* Hence adjusted  $R^2$  decreases

- 
- \*  $R^2$  value increases when features are added, irrespective of significance of features
  - \* Adjusted  $R^2$  value increases after adding features, only if features are significant i.e. affects dependent variable