Gradient Descent

Gradient

$$f(x,y) = x^2y + \sin(y)$$

$$\frac{\partial f}{\partial x} = 2xy$$

$$\frac{\partial f}{\partial y} = x^2 + cos(y)$$

– Gradient puts these two partial derivatives together in a vector as follows:

$$\nabla f(x,y) = \nabla\, x^2y + \sin(y) = \begin{bmatrix} 2xy \\ x^2 + \cos(y) \end{bmatrix}$$

Cost Function Linear Regression

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

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

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

Derivative of cost function for one training example (x, y)

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{n} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}$$

Gradient Descent for Linear Regression

- Initialize $\boldsymbol{\theta}$
- Repeat until convergence

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

simultaneous update for j = 0 ... d

- Initialize $\boldsymbol{\theta}$
- Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right) x_j^{(i)}$$

simultaneous update for j = 0 ... d

Gradient Descent for Linear Regression

- To achieve simultaneous update
  - At the start of each GD iteration, compute $h_\theta\left(x^{(i)}\right)$
  - Use this stored value in the update step loop

- Assume convergence when $\|\boldsymbol{\theta}_{new} - \boldsymbol{\theta}_{old}\|_2 < \epsilon$

L_2 norm: $\|\boldsymbol{v}\|_2 = \sqrt{\sum_i v_i^2} = \sqrt{v_1^2 + v_2^2 + \ldots + v_{|v|}^2}$

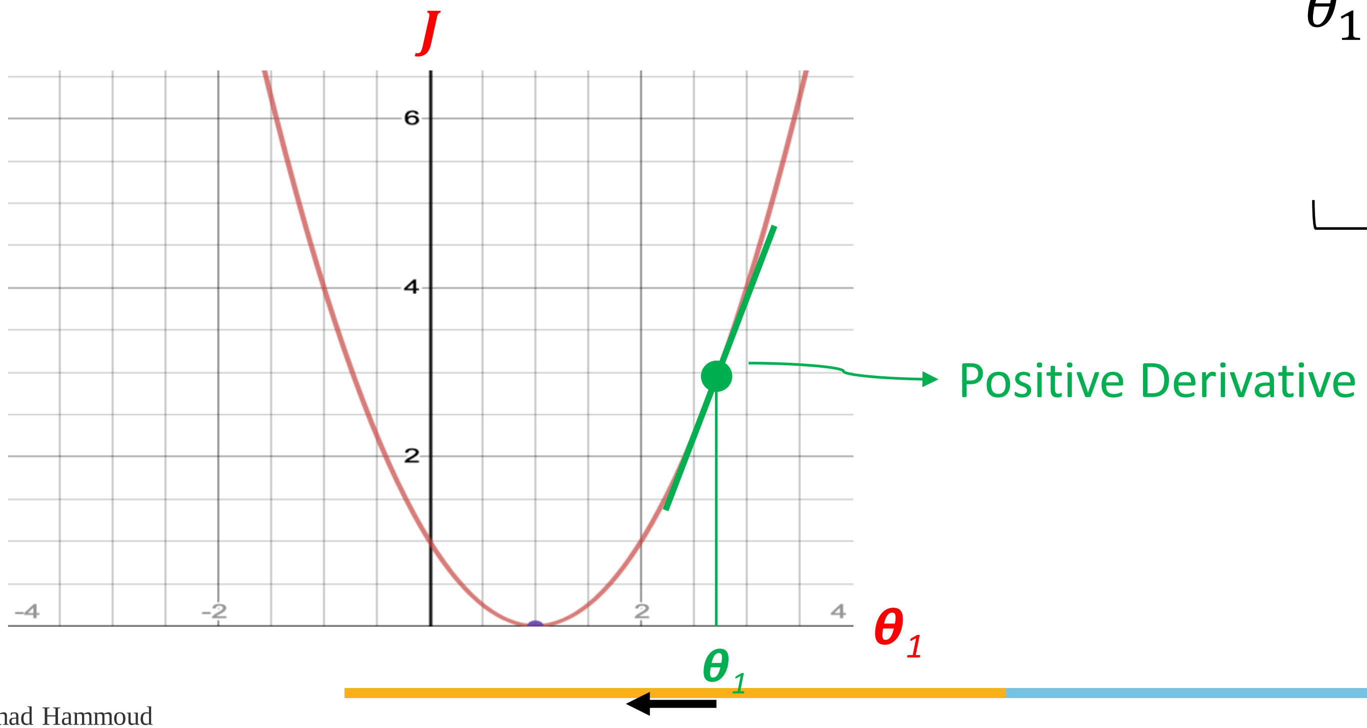
Gradient Descent Intution

- optimization objective is to minimize $\underset{\theta_0,\theta_1}{} J(\theta_1)$

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

$$\underbrace{= \theta_1 - \alpha\, (Positive\ Number)}$$

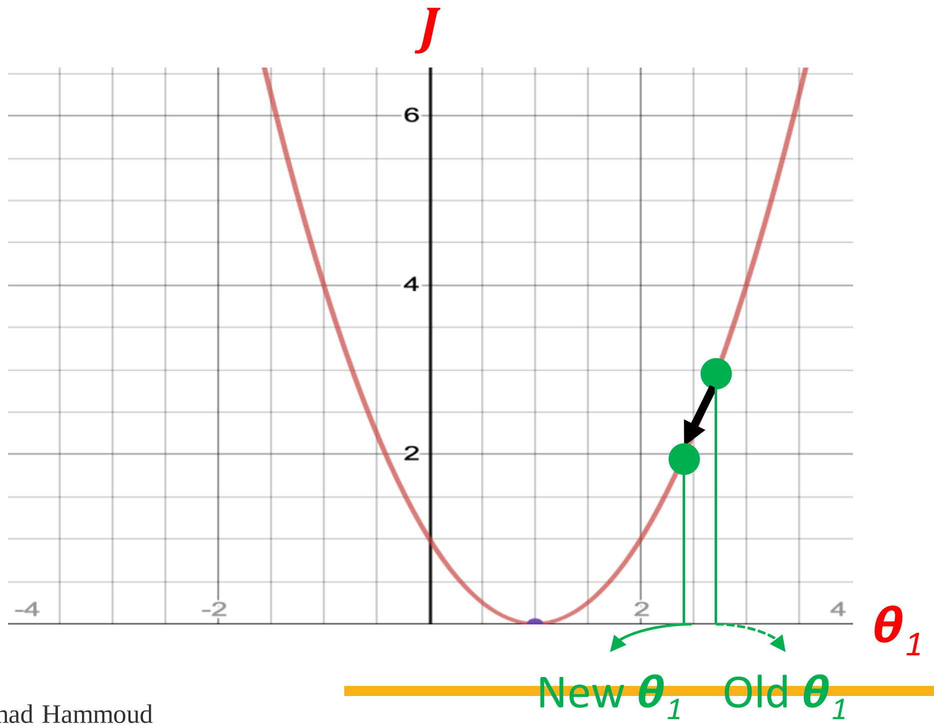
Decrease θ_1 by a certain value



Credit: Prof. Mohammad Hammoud

The Impact of Partial Derviative

- optimization objective is to minimize$_{\theta_0,\theta_1}$ $\boldsymbol{J(\theta_1)}$

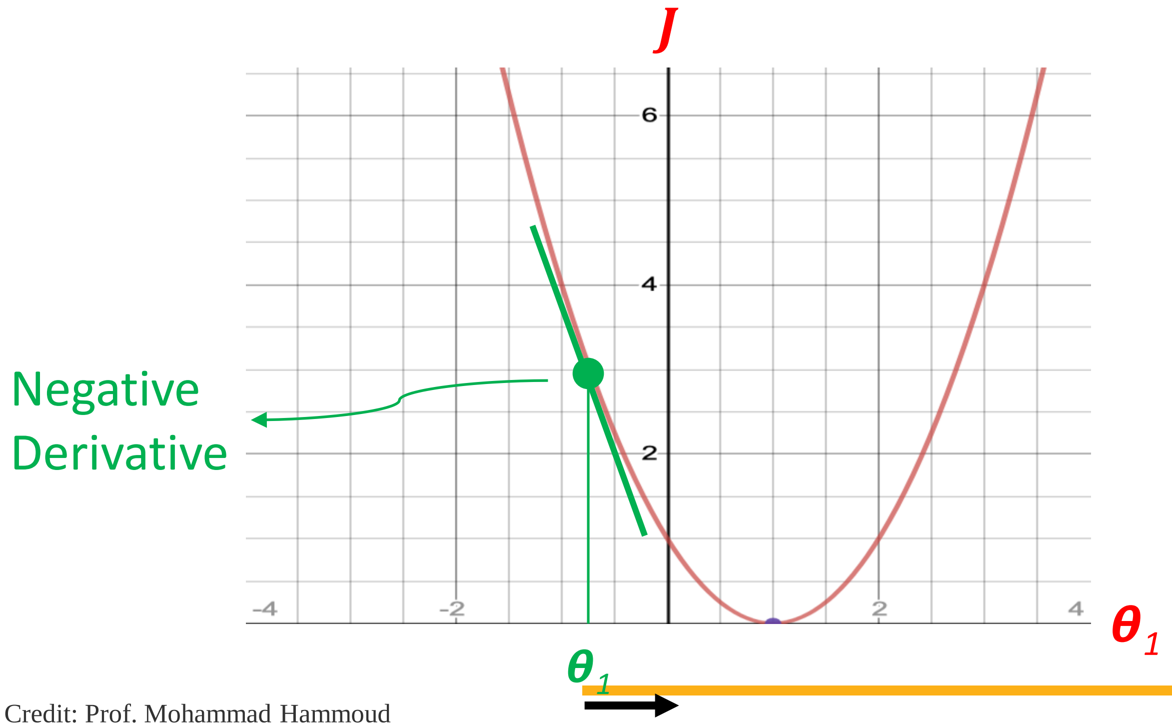


$$\theta_1 = \theta_1 - \alpha \frac{\boldsymbol{d}\, \boldsymbol{J(\theta_1)}}{\boldsymbol{d}\, \theta_j}$$

$$\underbrace{= \theta_1 - \alpha\, (Positive\ Number)}_{\text{Decrease } \theta_1 \text{ by a certain value}}$$

The Impact of Partial Derviative

- optimization objective is to $\min_{\theta_0,\theta_1} \boldsymbol{J(\theta_1)}$



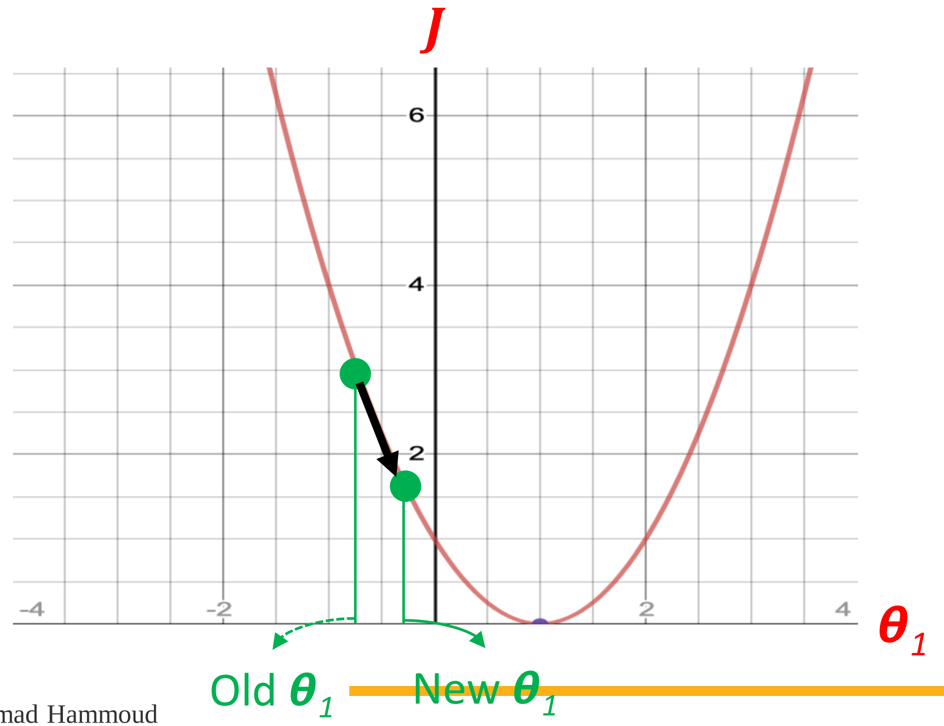
$$\theta_1 = \theta_1 - \alpha \frac{\boldsymbol{d}\, \boldsymbol{J(\theta_1)}}{\boldsymbol{d}\, \theta_j}$$

$$\underbrace{= \theta_1 - \alpha\, (Negative\ Number)}_{\text{Increase } \theta_1 \text{ by a certain value}}$$

Credit: Prof. Mohammad Hammoud

The Impact of Partial Derviative

- optimization objective is to $\underset{\theta_0,\theta_1}{\text{minimize}}\, \boldsymbol{J(\theta_1)}$



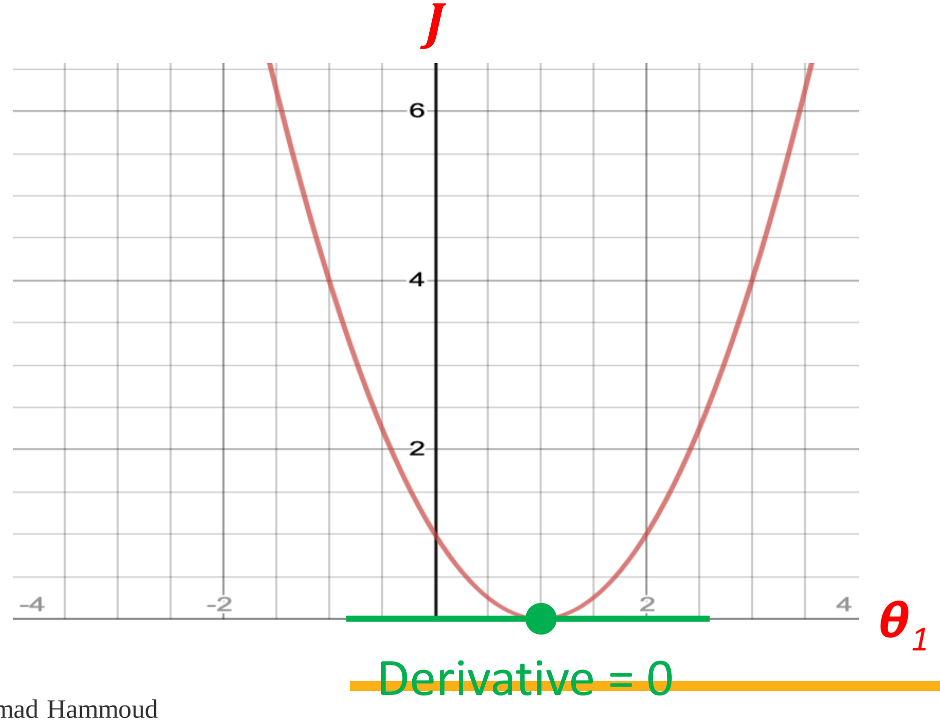
$$\theta_1 = \theta_1 - \alpha \frac{\boldsymbol{d}\, \boldsymbol{J(\theta_1)}}{\boldsymbol{d}\, \theta_j}$$

$$\underbrace{= \theta_1 - \alpha\, (Negative\ Number)}_{\text{Increase } \theta_1 \text{ by a certain value}}$$

Credit: Prof. Mohammad Hammoud

The Impact of Partial Derviative

- optimization objective is to $\underset{\theta_0,\theta_1}{\text{minimize}}\ \boldsymbol{J(\theta_1)}$



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

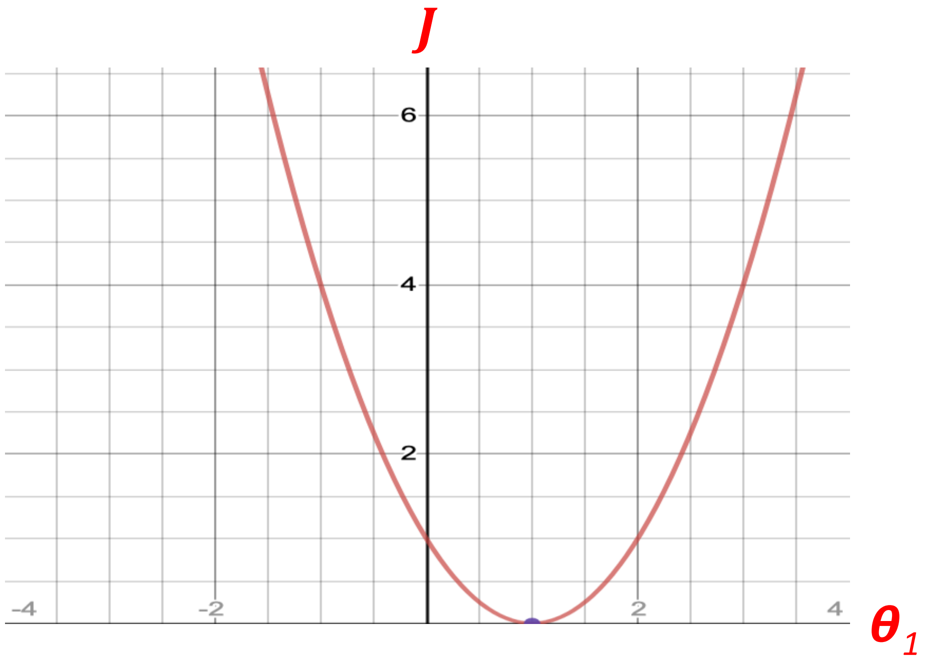
$$\underbrace{= \theta_1 - \alpha\ (Zero)}$$

θ_1 remains the same, hence, gradient descent *converges*

Credit: Prof. Mohammad Hammoud

The Impact of Learning Rate

- optimization objective is to minimize$_{\theta_0,\theta_1}$ $\boldsymbol{J(\theta_1)}$



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

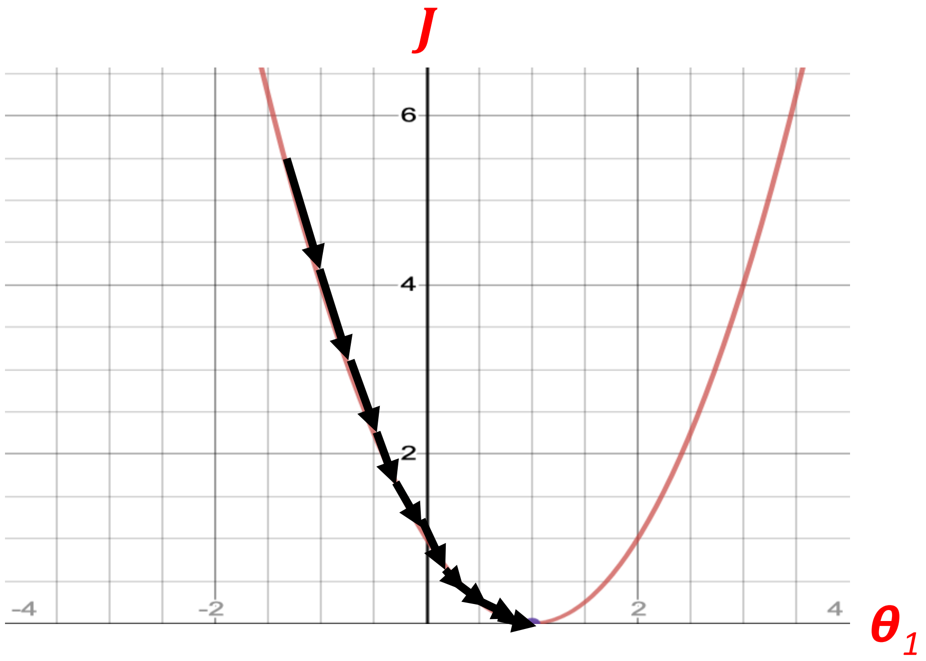
Learing Rate

What happens if α is too small?

Credit: Prof. Mohammad Hammoud

The Impact of Learning Rate

- optimization objective is to $\underset{\theta_0,\theta_1}{\text{minimize}}\ \boldsymbol{J(\theta_1)}$



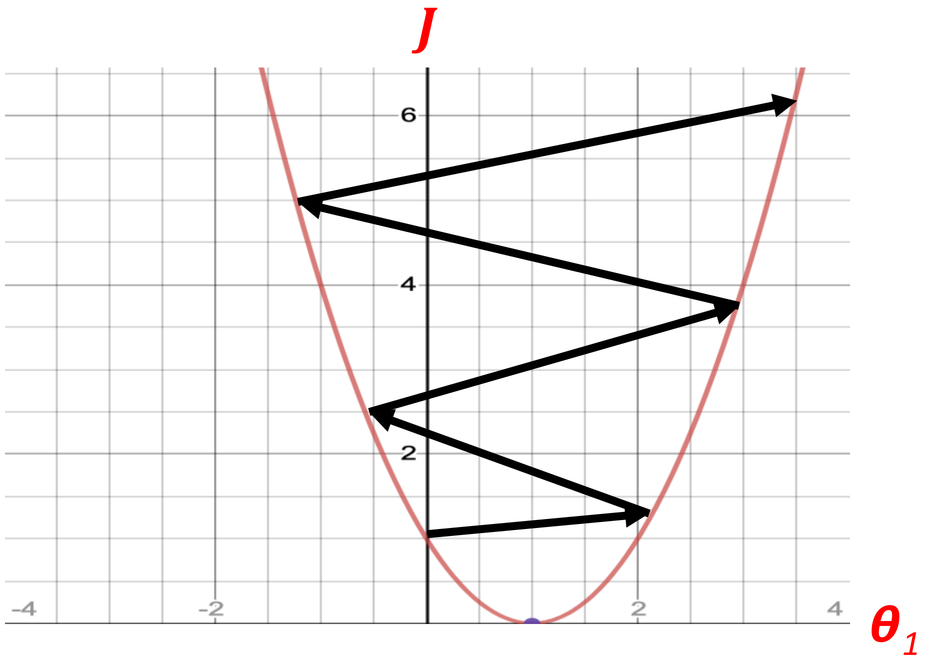
$$\theta_1 = \theta_1 - \alpha \frac{\boldsymbol{d\,J(\theta_1)}}{\boldsymbol{d}\,\theta_j}$$

$$= \theta_1 - (Too\ Small\ Number)\ \frac{\boldsymbol{d\,J(\theta_1)}}{\boldsymbol{d}\,\theta_j}$$

θ_1 changes only a tiny bit on each step, hence, gradient descent *will render slow (will take more time to converge)*

The Impact of Learning Rate

- optimization objective is to minimize$_{\theta_0,\theta_1}$ $\boldsymbol{J(\theta_1)}$



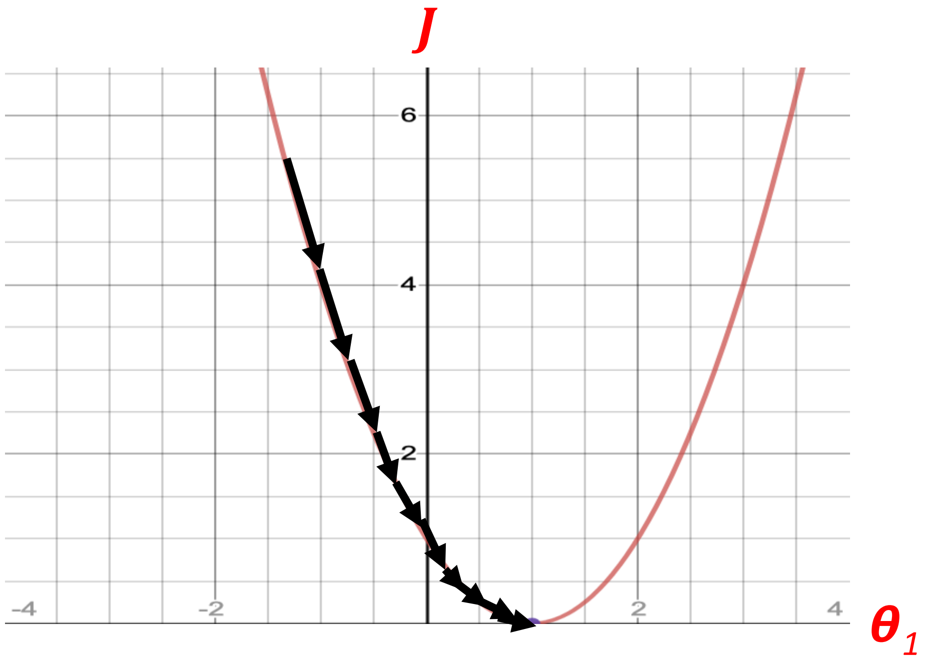
$$\theta_1 = \theta_1 - \alpha \frac{\boldsymbol{d}\, \boldsymbol{J(\theta_1)}}{\boldsymbol{d}\, \theta_j}$$

$$\underbrace{= \theta_1 - (Too\ Large\ Number) \frac{\boldsymbol{d}\, \boldsymbol{J(\theta_1)}}{\boldsymbol{d}\, \theta_j}}$$

θ_1 changes a lot (and probably faster) on each step, hence, gradient descent *will potentially overshoot the minimum and, accordingly, fail to converge (or even diverge)*

The Impact of Learning Rate

- optimization objective is to minimize $\underset{\theta_0,\theta_1}{\text{minimize}}\ J(\theta_1)$



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

We can also ***fix α*** because as we approach the (global) minimum, gradient descent will automatically start taking smaller steps (i.e., θ_1 will start changing at a slower pace because the derivative will become less steep)

Linear Regression Example

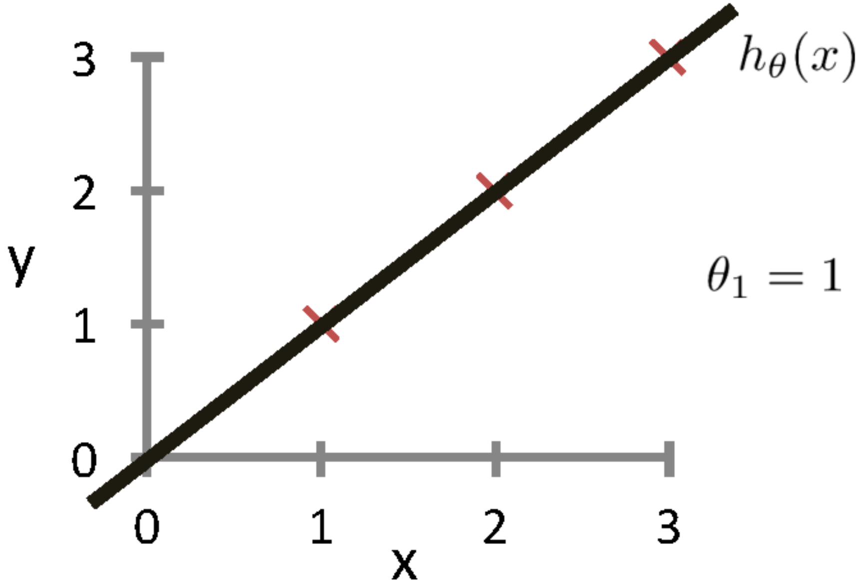
Intuition Behind Cost Function

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

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

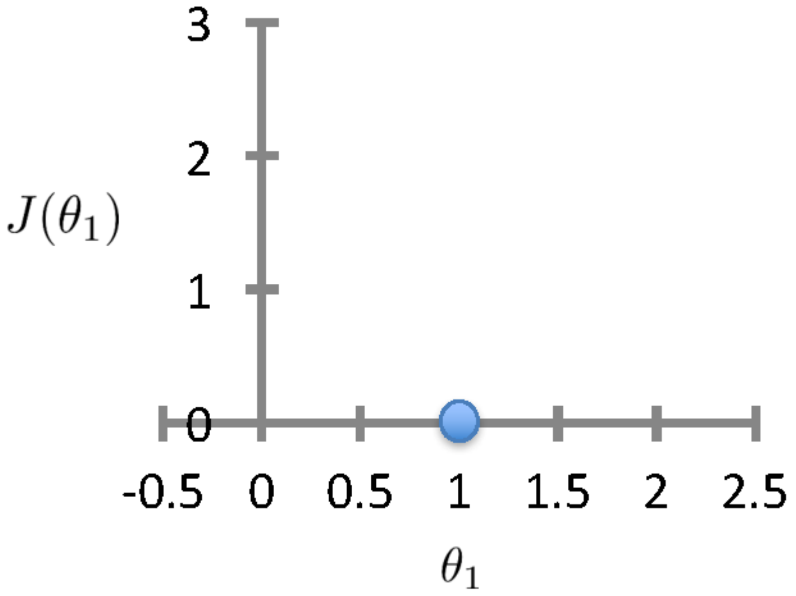
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Based on example
by Andrew Ng

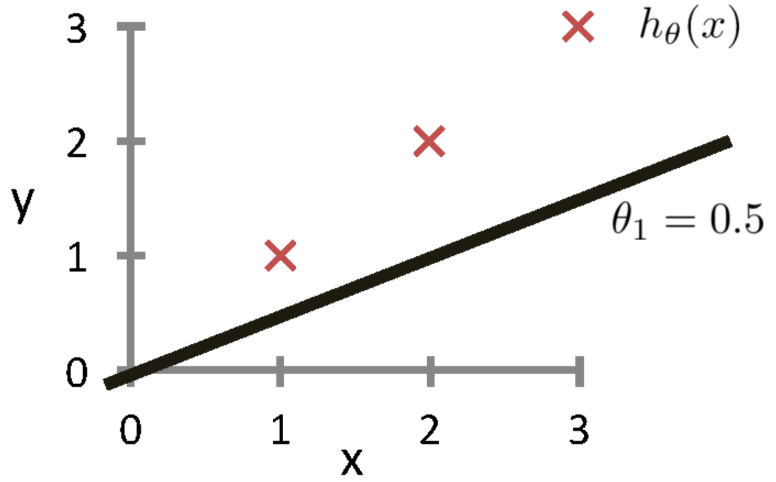
Intuition Behind Cost Function

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

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

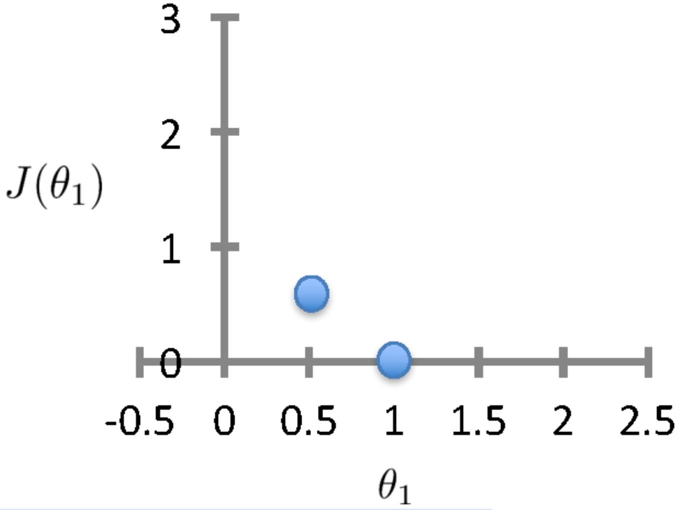
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Based on example by Andrew Ng

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

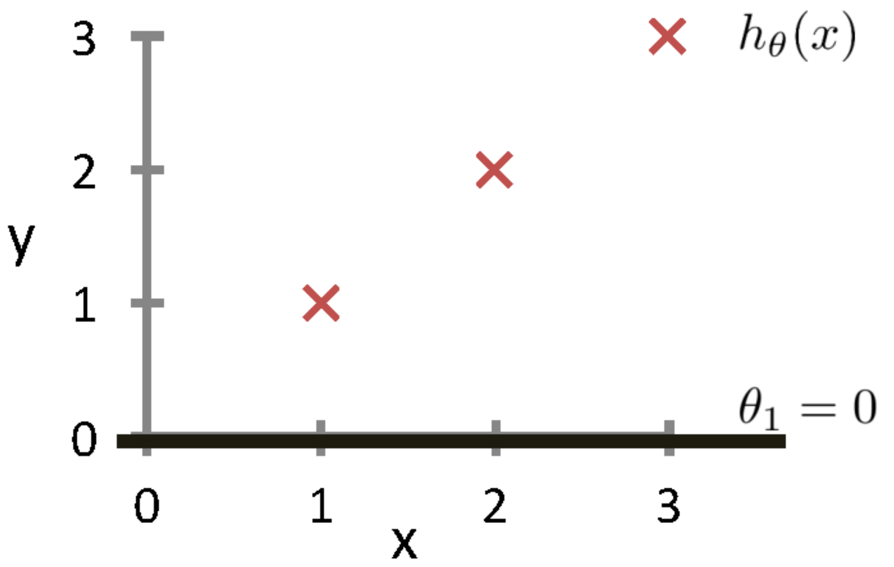
Intuition Behind Cost Function

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

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

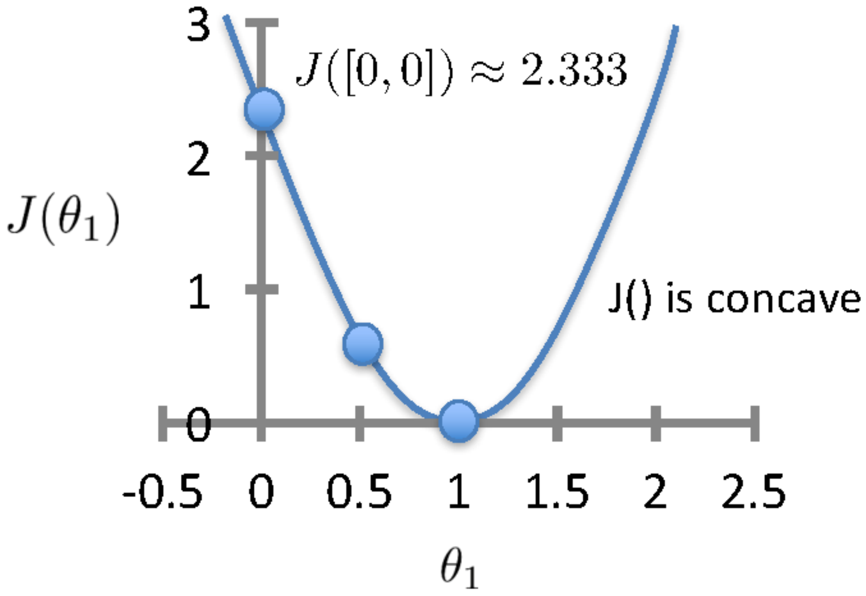
$h_\theta(x)$

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



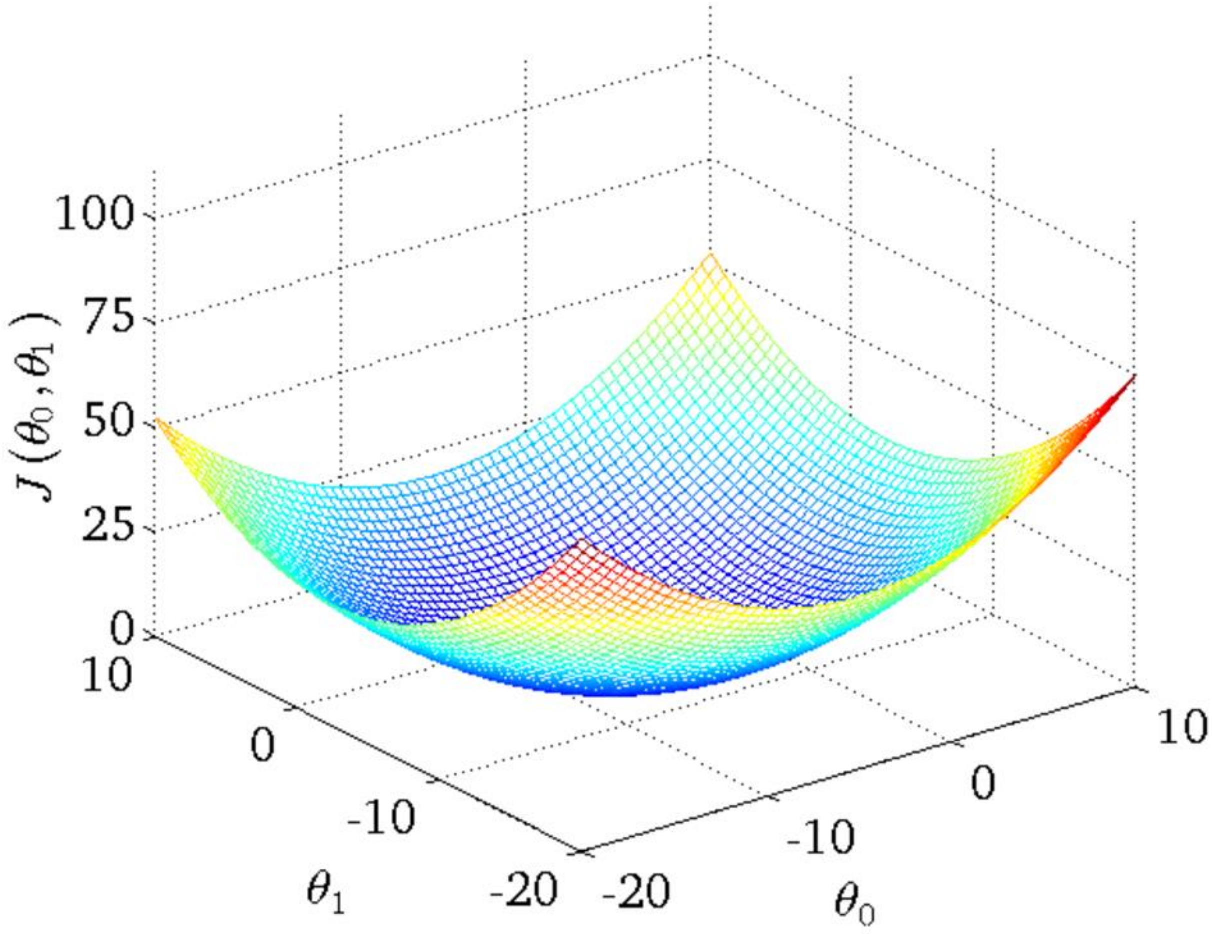
$J(\theta_1)$

(function of the parameter θ_1)



Based on example
by Andrew Ng

Intuition Behind Cost Function

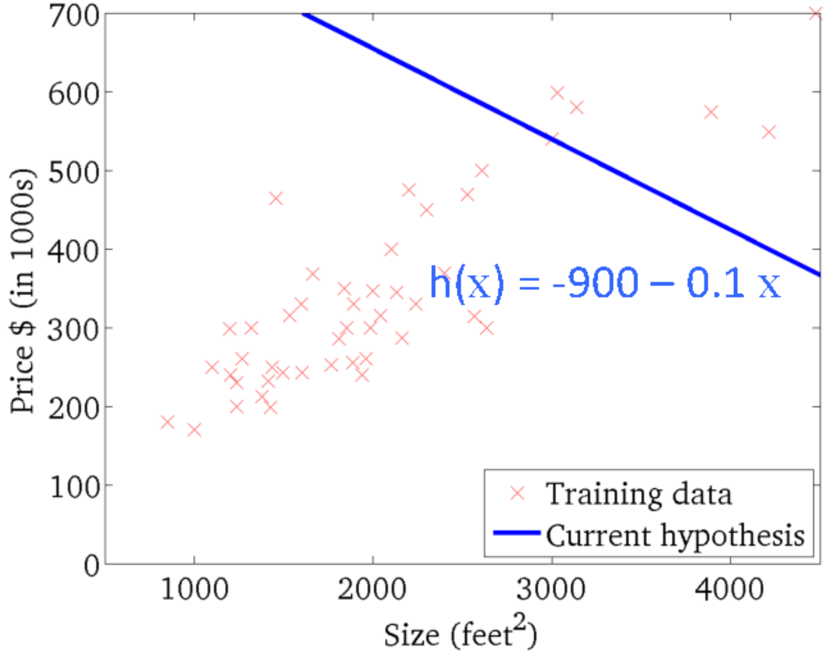


Slide by Andrew Ng

Gradient Descent

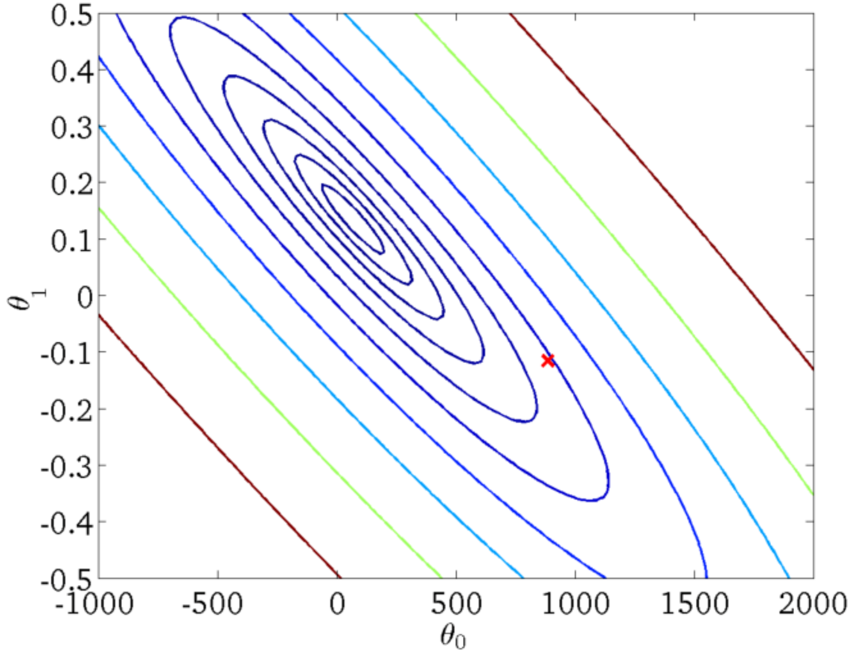
$$h_\theta(x)$$

(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)

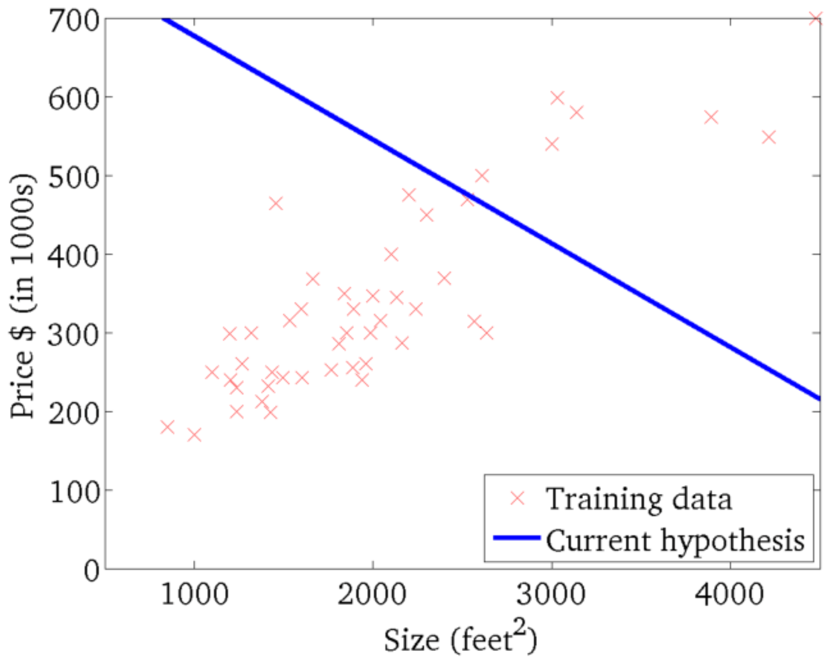


Slide by Andrew Ng

Gradient Descent

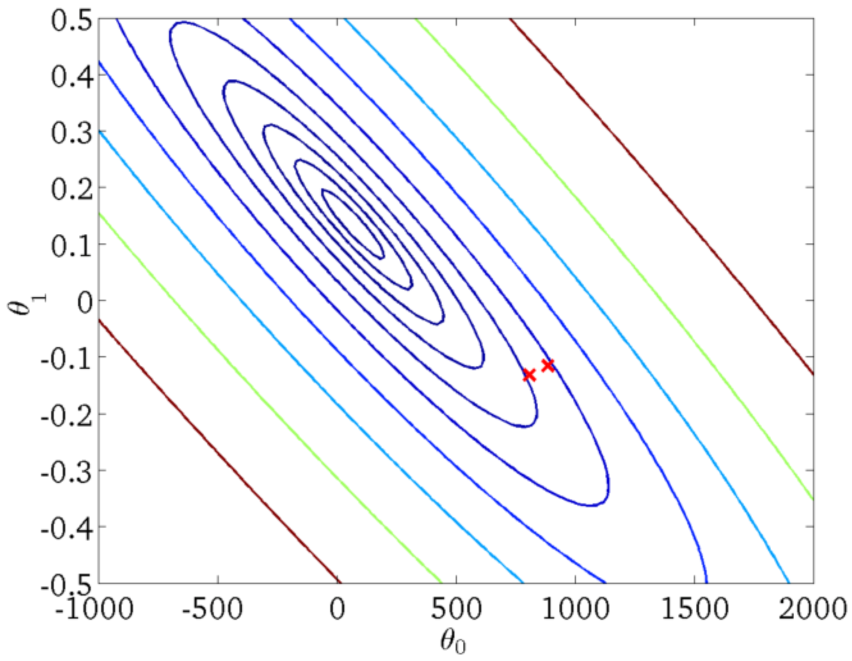
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

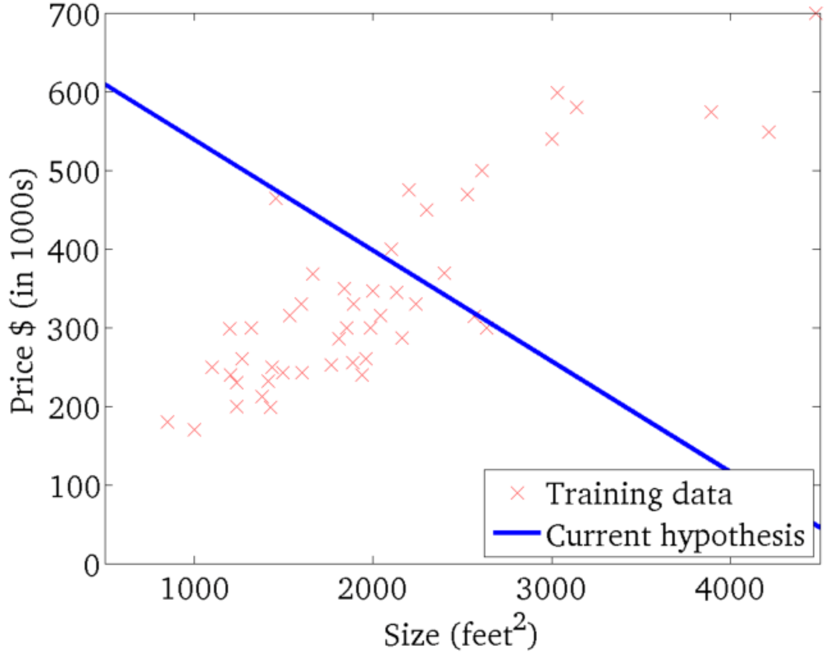
(function of the parameters θ_0, θ_1)



Gradient Descent

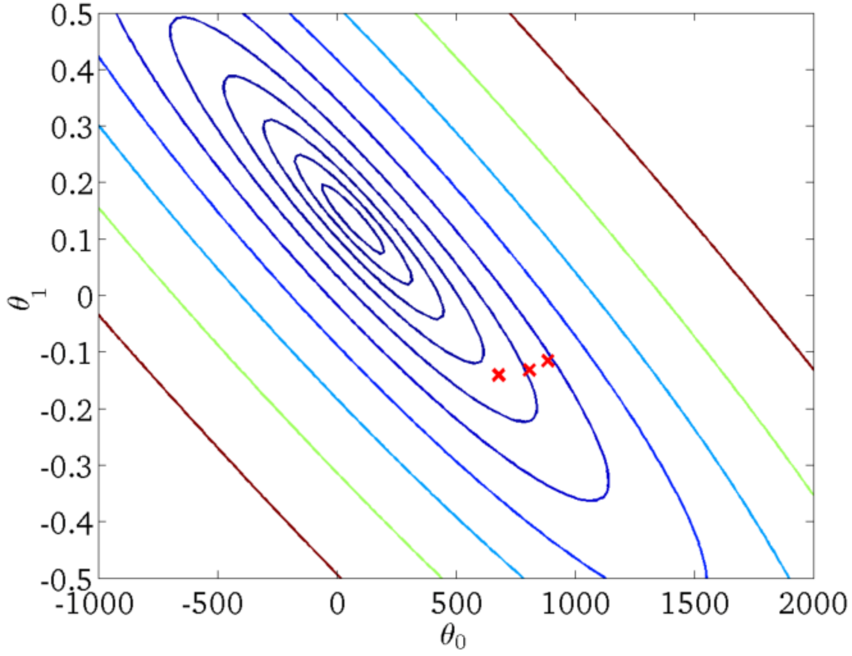
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

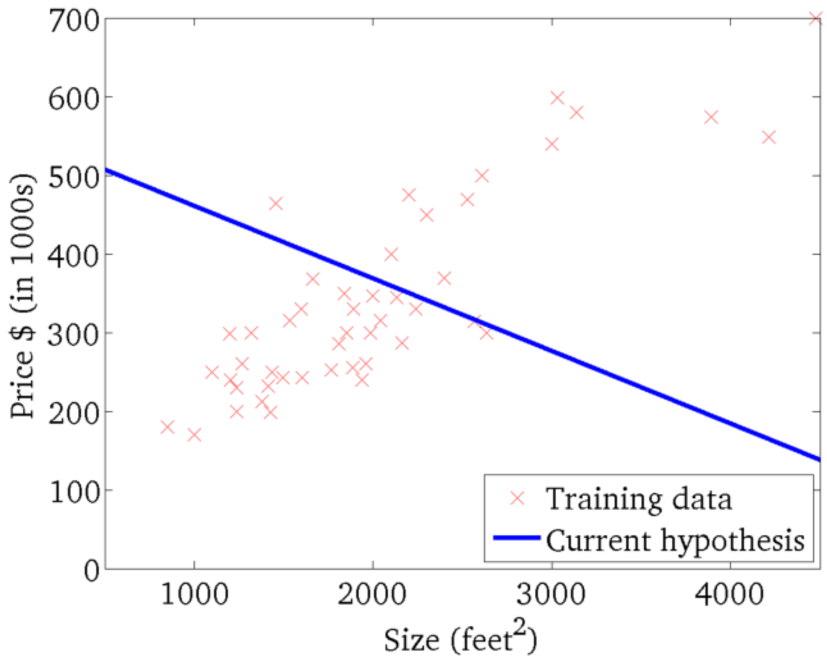
(function of the parameters θ_0, θ_1)



Gradient Descent

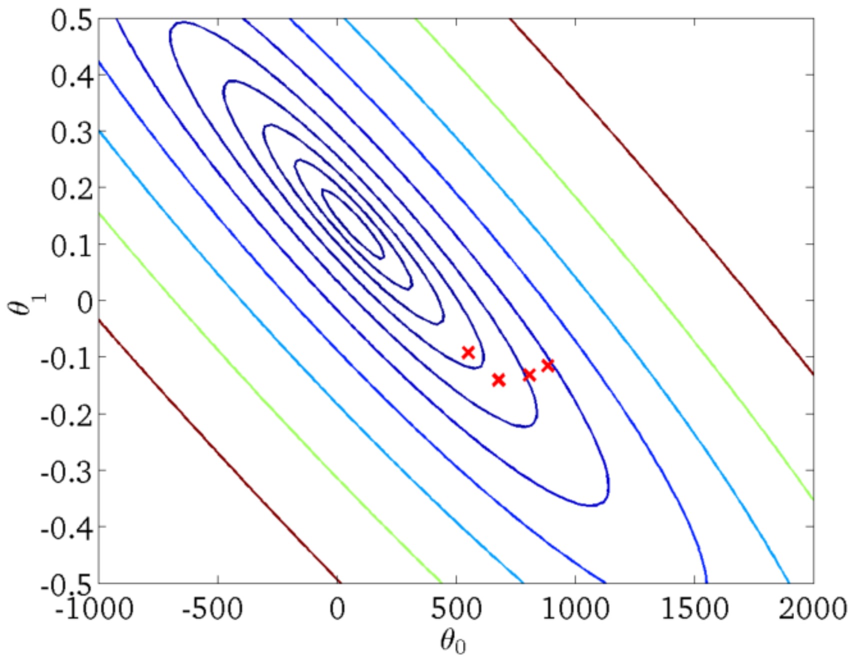
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

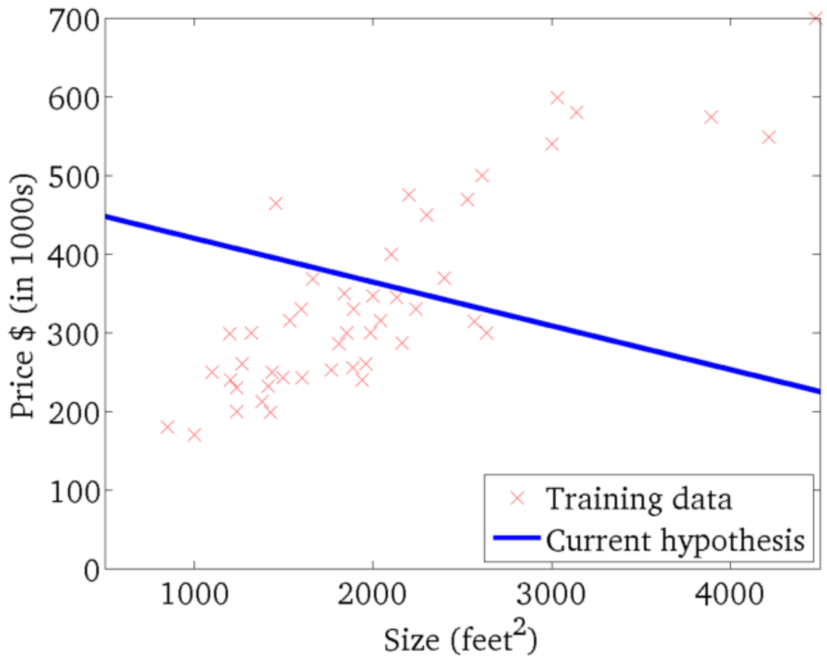
(function of the parameters θ_0, θ_1)



Gradient Descent

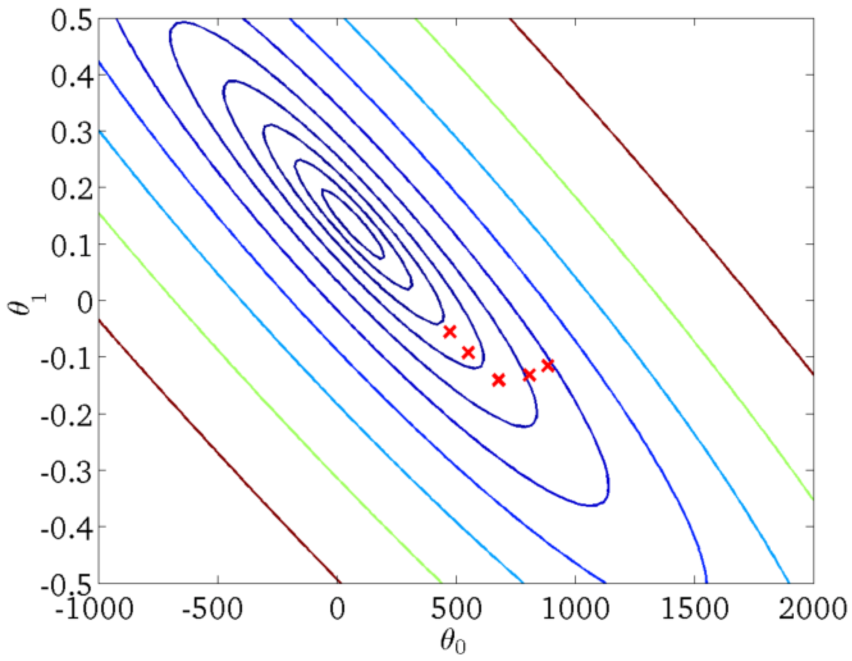
$h_\theta(x)$

(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)

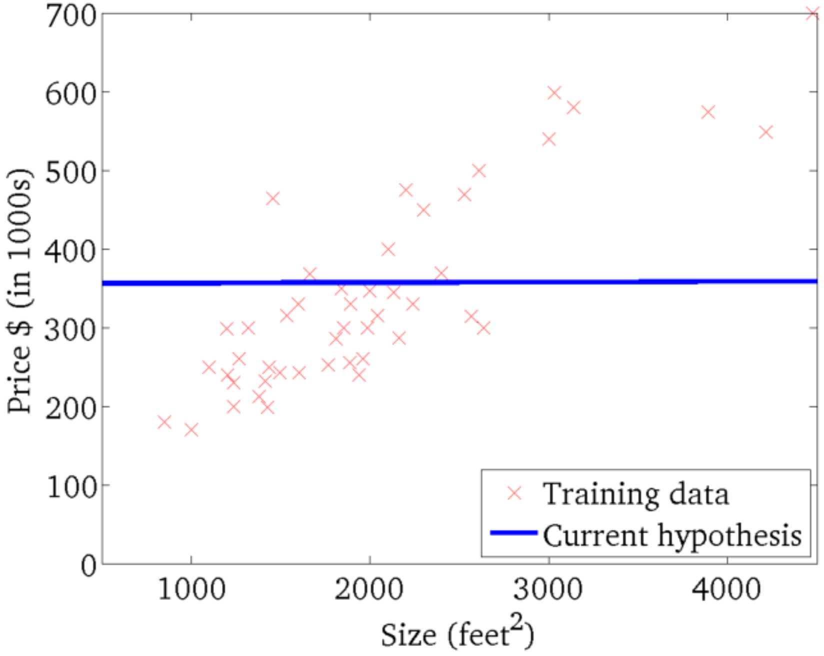


Slide by Andrew Ng

Gradient Descent

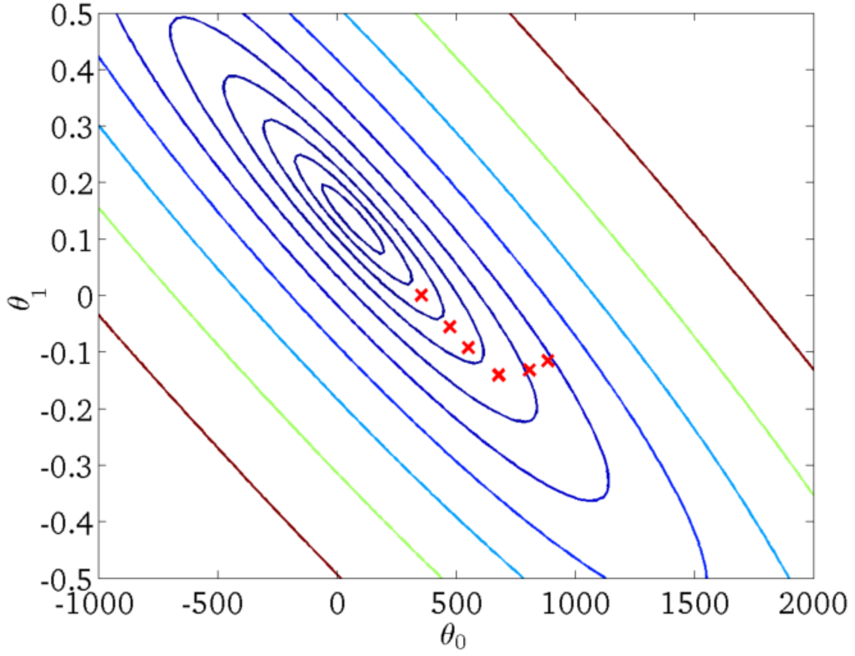
$h_\theta(x)$

(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)

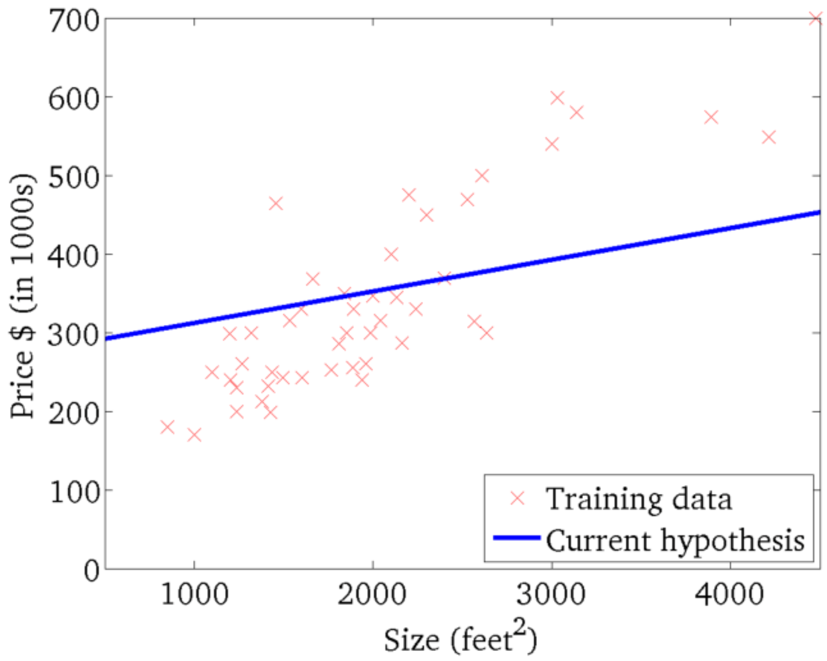


Slide by Andrew Ng

Gradient Descent

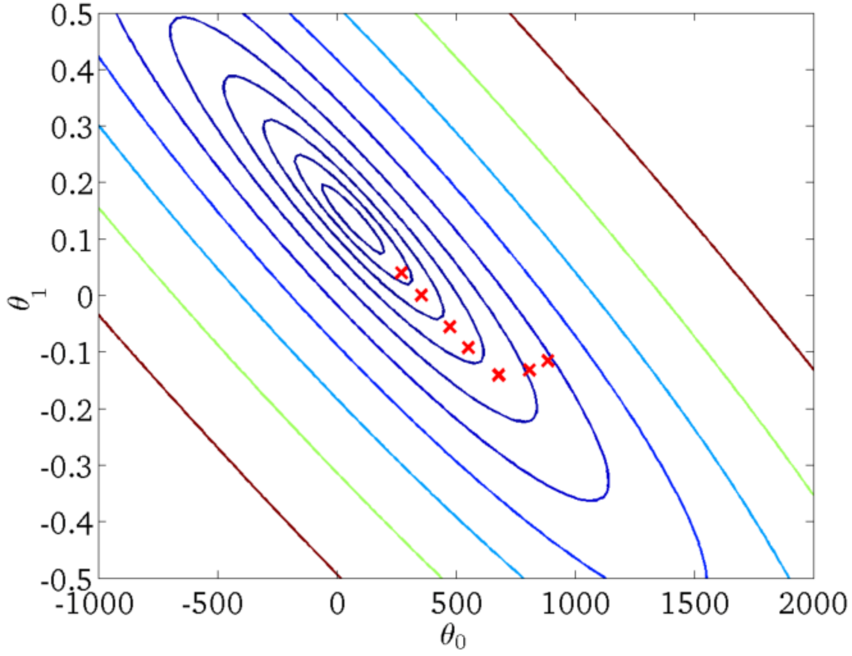
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

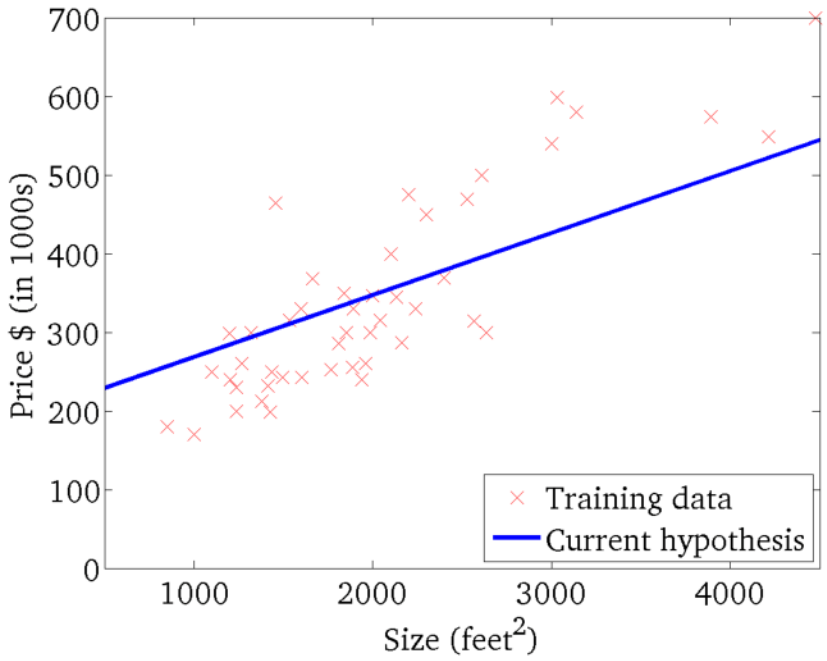
(function of the parameters θ_0, θ_1)



Gradient Descent

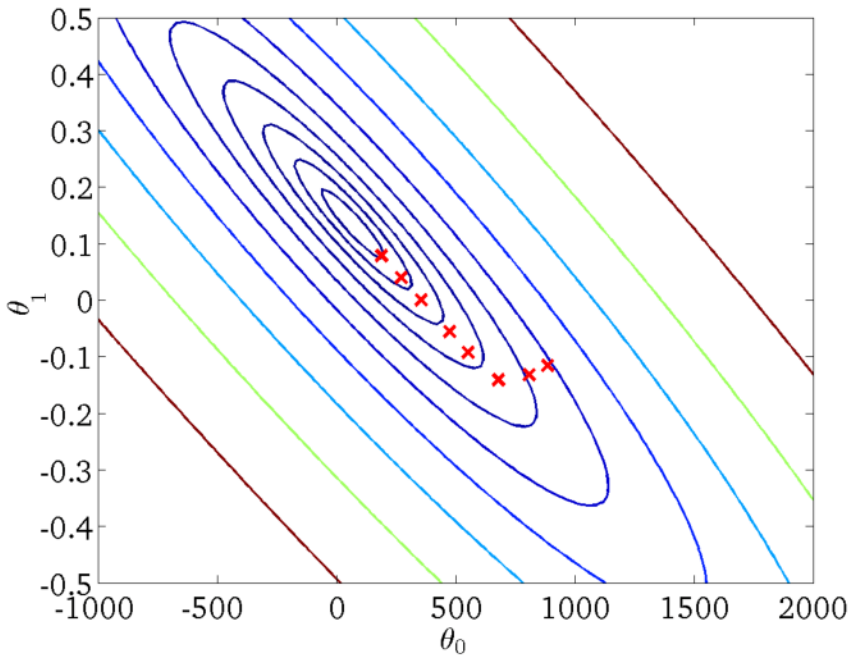
$h_\theta(x)$

(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)

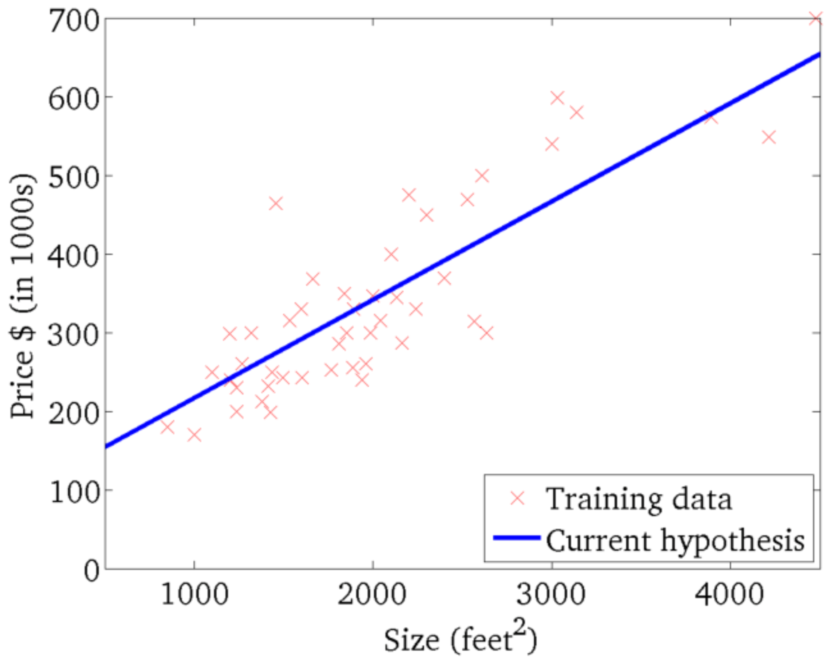


Slide by Andrew Ng

Gradient Descent

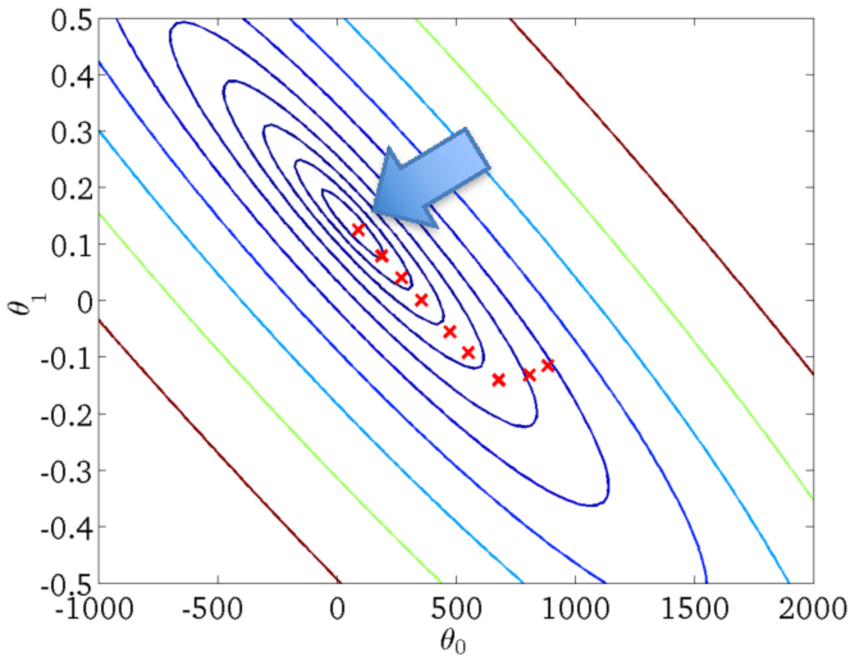
$h_\theta(x)$

(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)



Basic Search Procedure

- Choose initial value for θ
- Until we reach a minimum:
  - Choose a new value for $\boldsymbol{\theta}$ to reduce $J(\boldsymbol{\theta})$

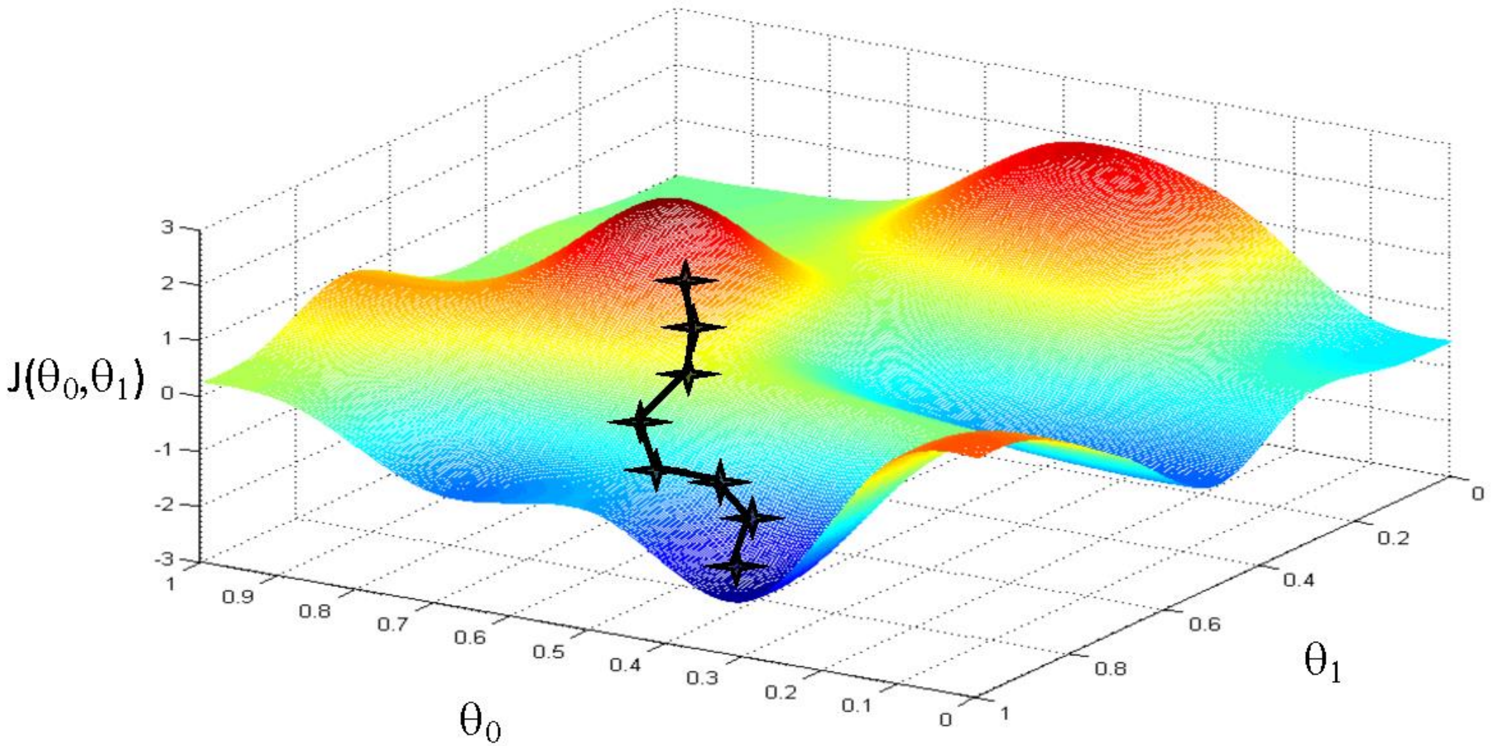


Figure by Andrew Ng

Basic Search Procedure

- Choose initial value for θ
- Until we reach a minimum:
  - Choose a new value for θ to reduce $J(\theta)$

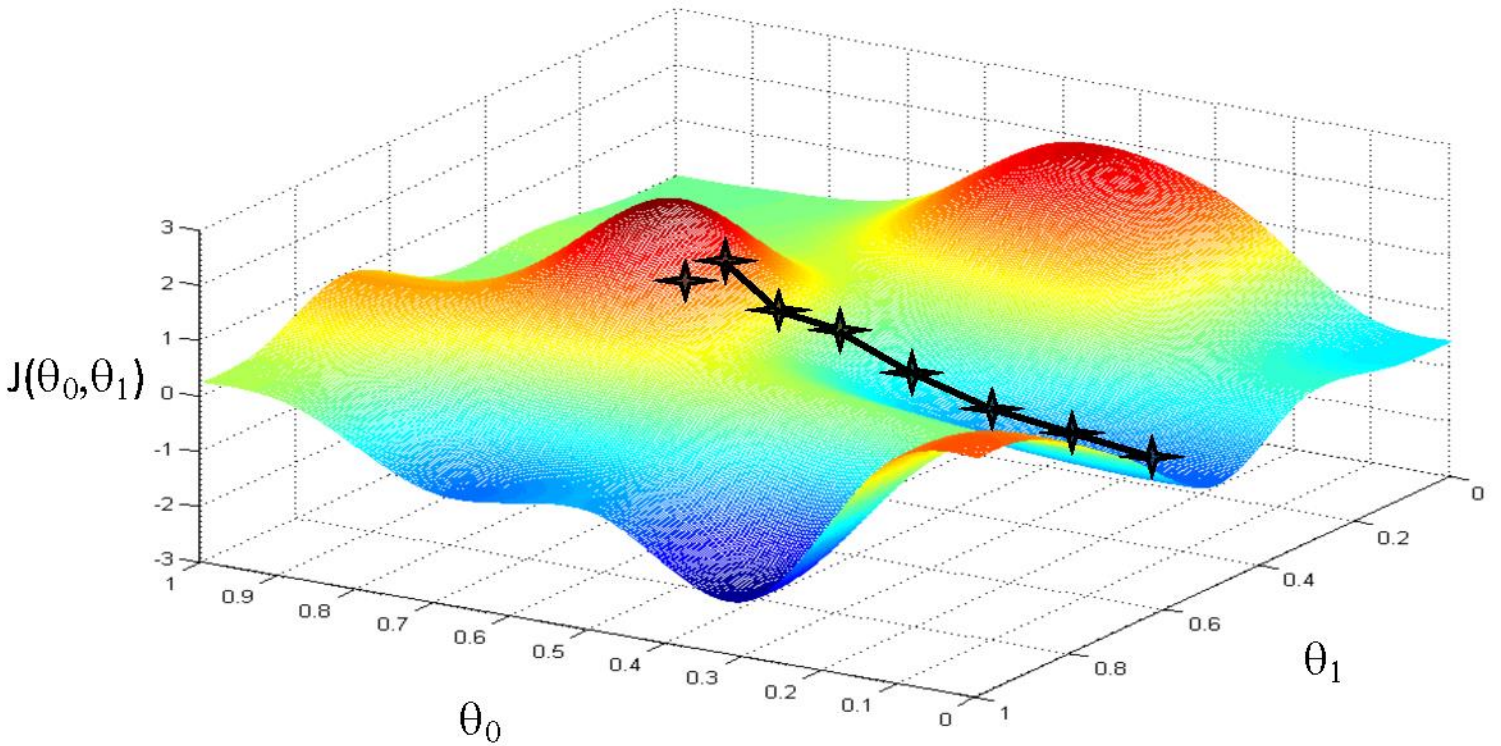
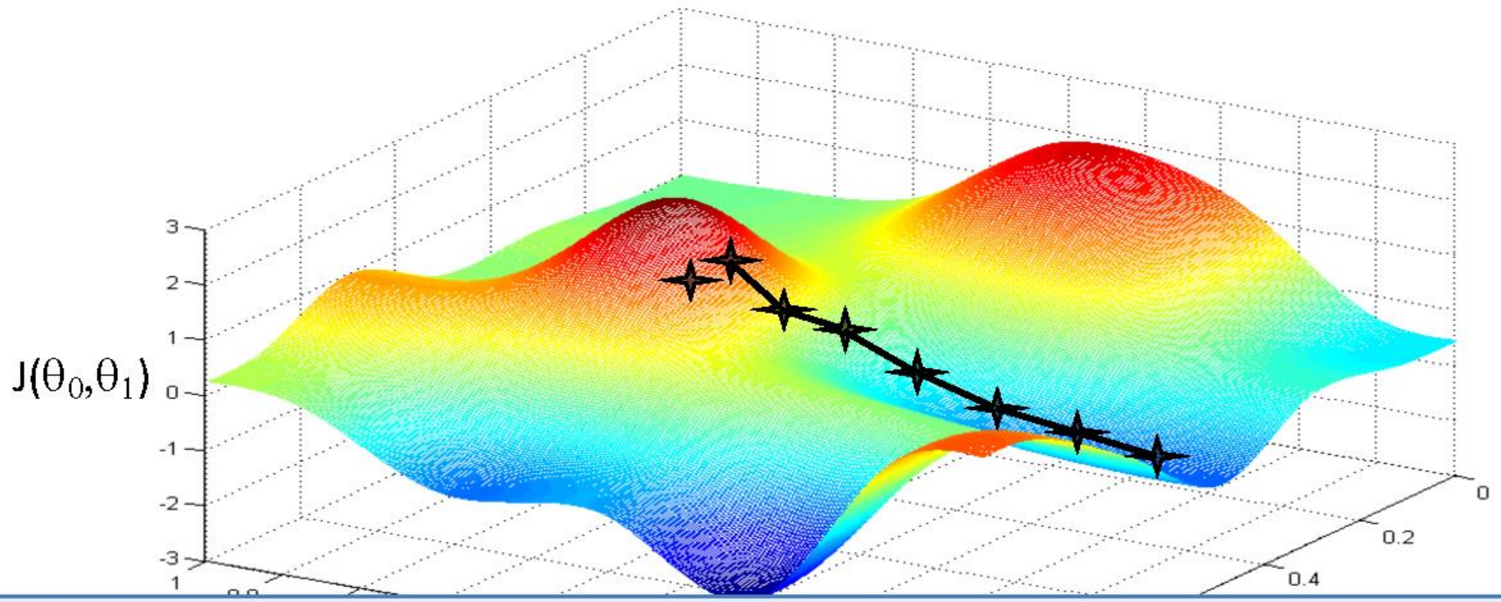


Figure by Andrew Ng

Basic Search Procedure

- Choose initial value for θ
- Until we reach a minimum:
  - Choose a new value for θ to reduce $J(\theta)$



Since the least squares objective function is convex (concave), we don't need to worry about local minima

Figure by Andrew Ng