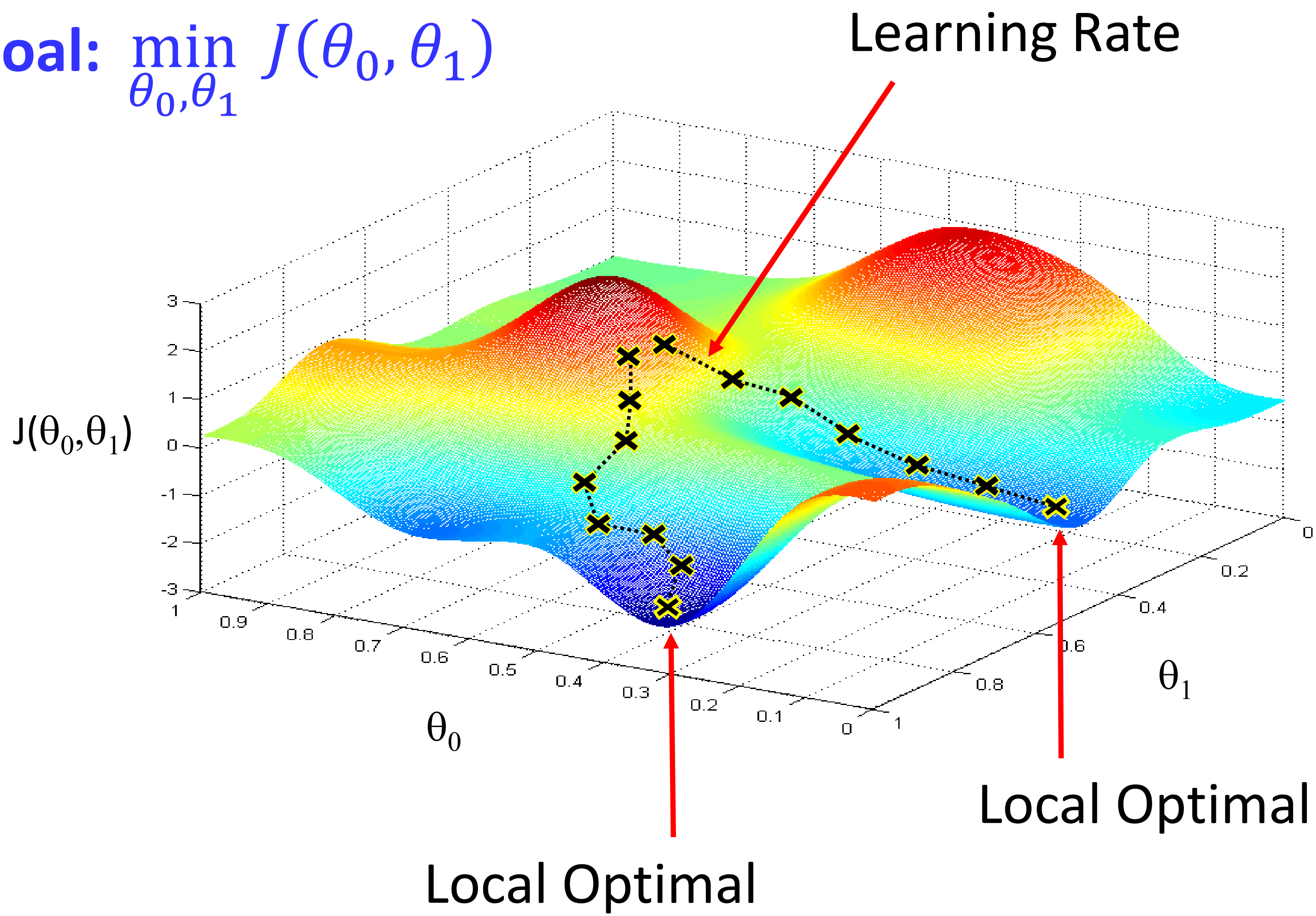
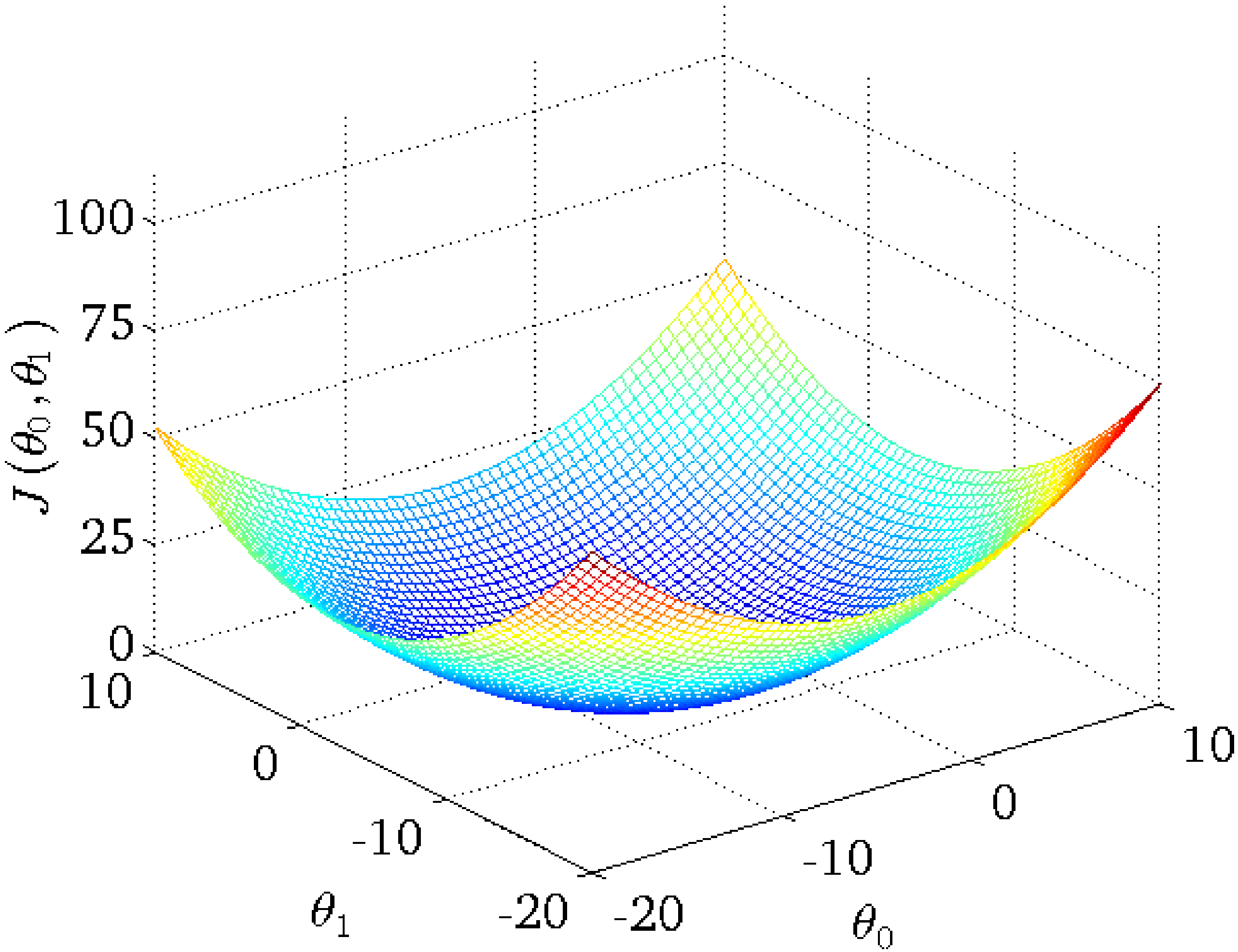
Gradient Descent

Gradient Descent

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



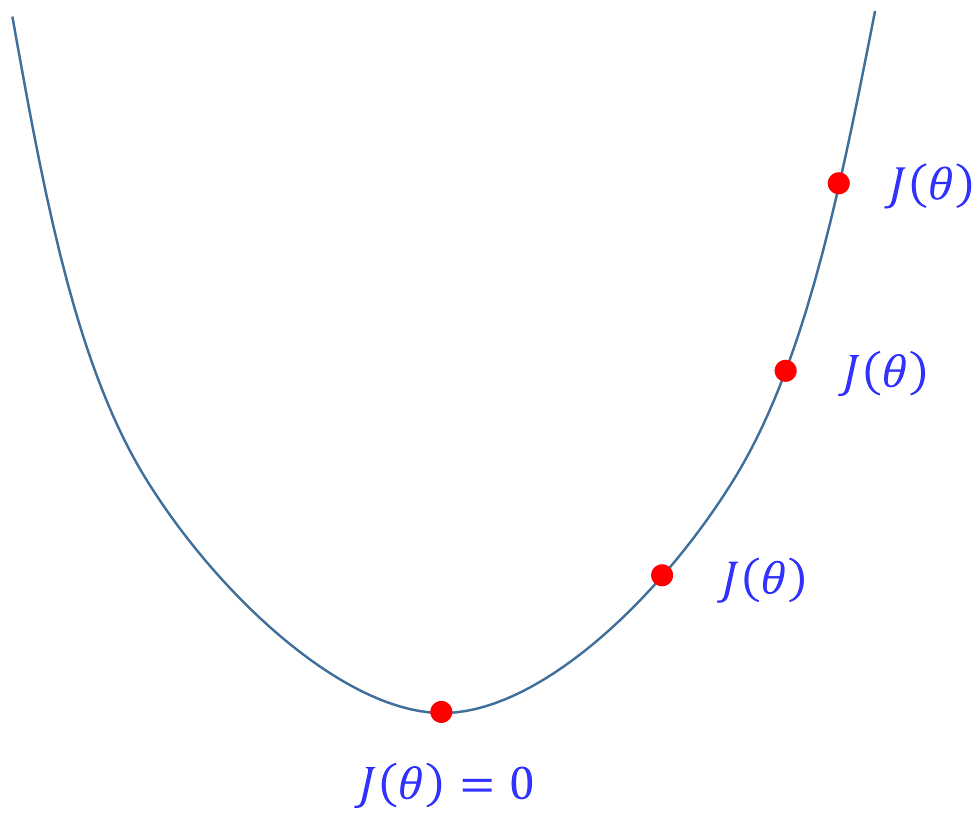
Cost Function With 2 Parameters



Cost Function With 1 Parameter

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

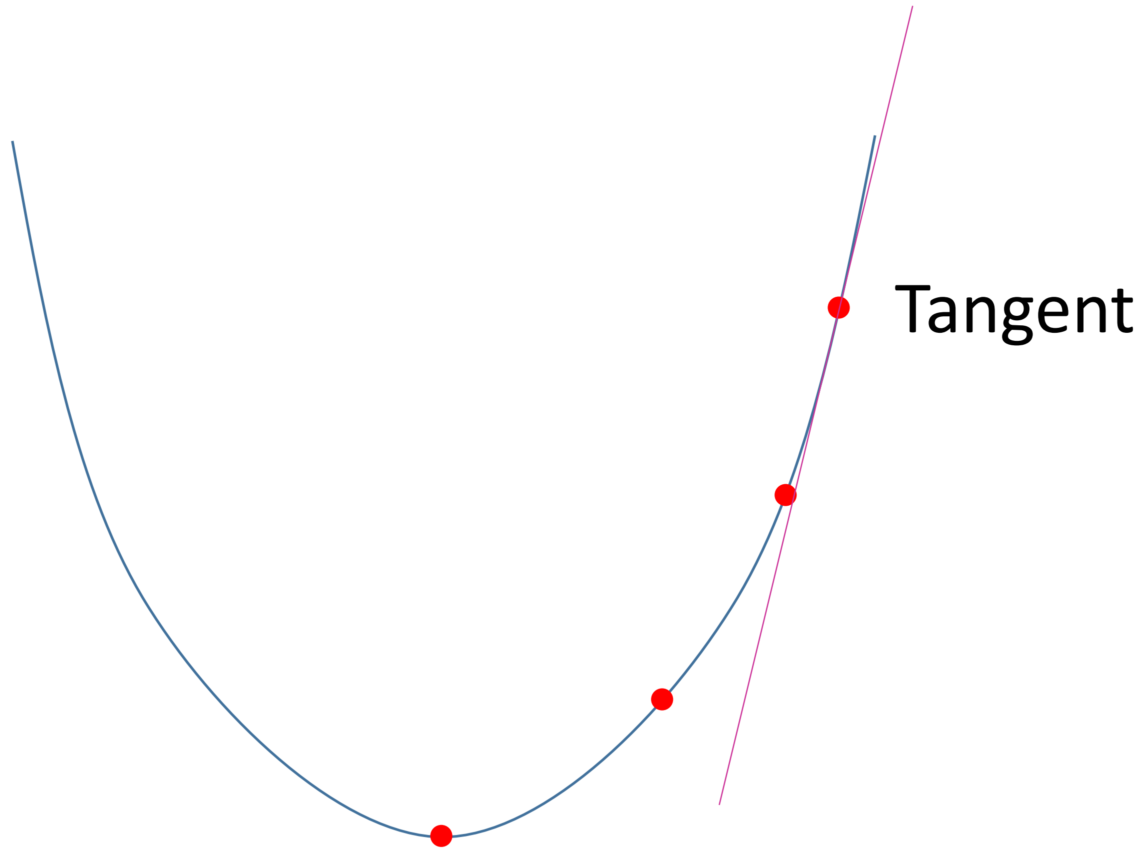
Will be using only one Parameter (θ) to simplify the explanation.



Gradient Descent

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

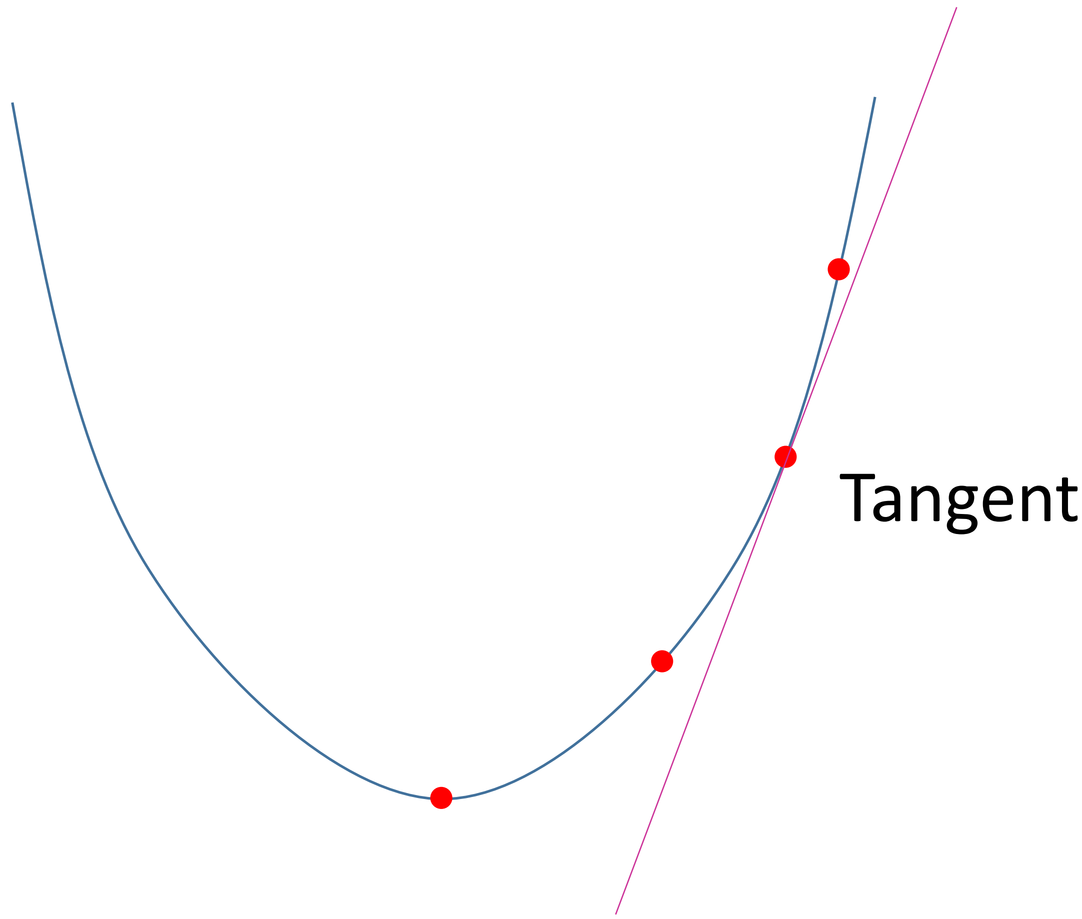
Will be using only one Parameter (θ) to simplify the explanation.



Gradient Descent

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

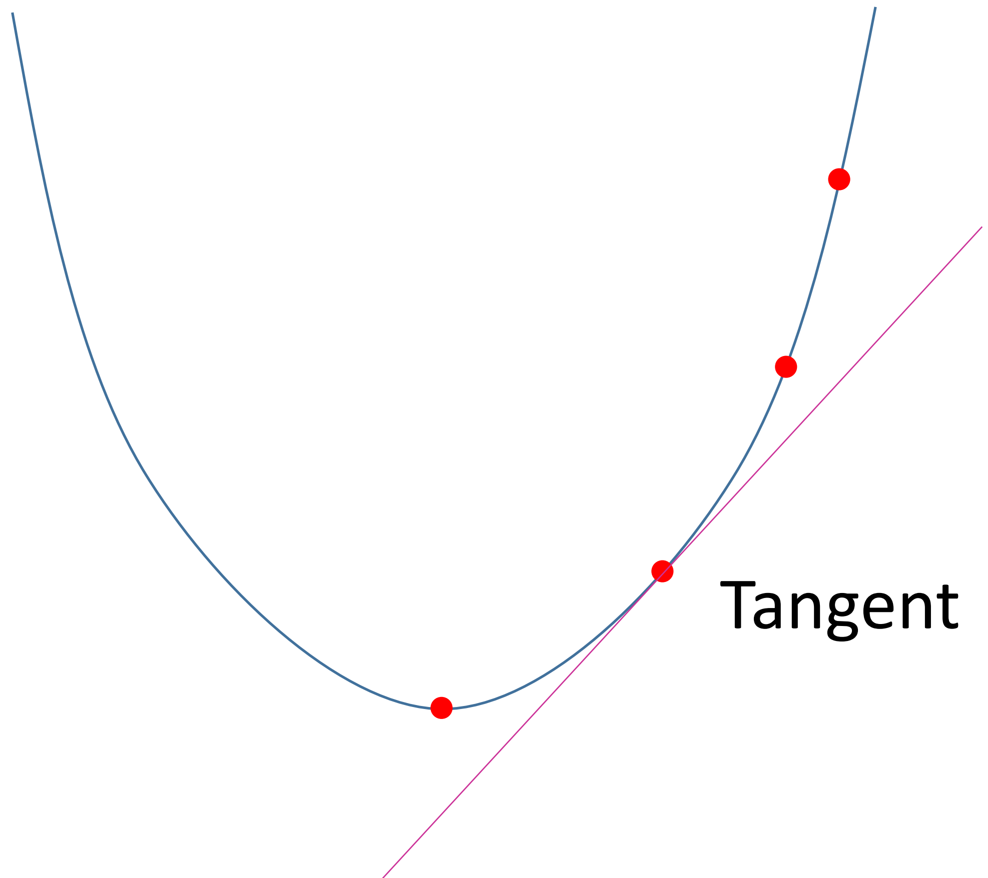
Will be using only one Parameter (θ) to simplify the explanation.



Gradient Descent

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

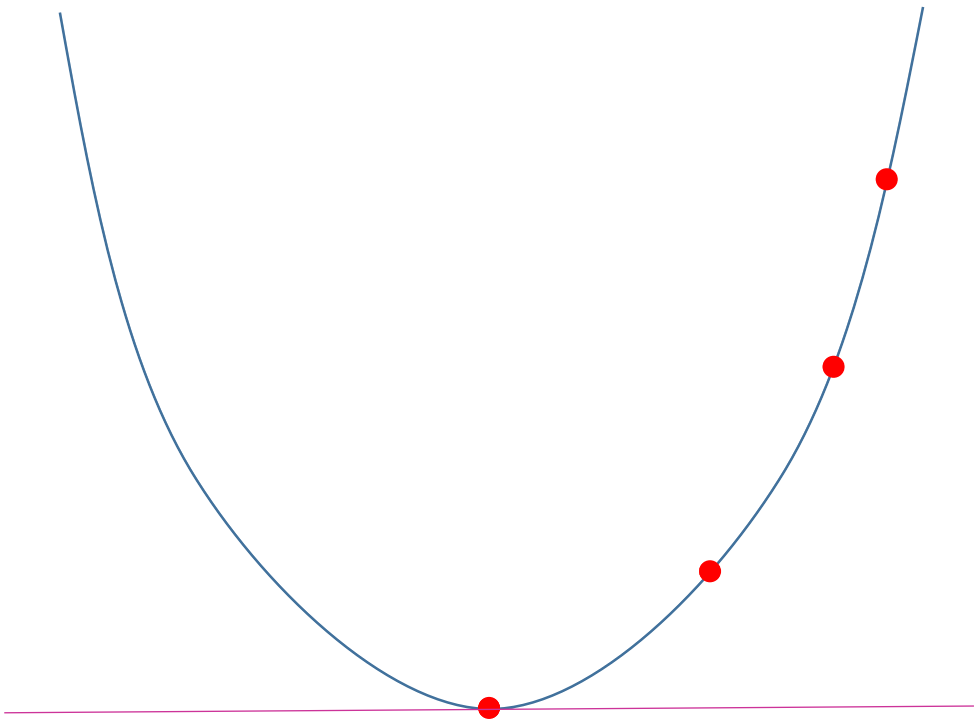
Will be using only one Parameter (θ) to simplify the explanation.



Gradient Descent

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

Will be using only one Parameter (θ) to simplify the explanation.



Tangent

Gradient Descent

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

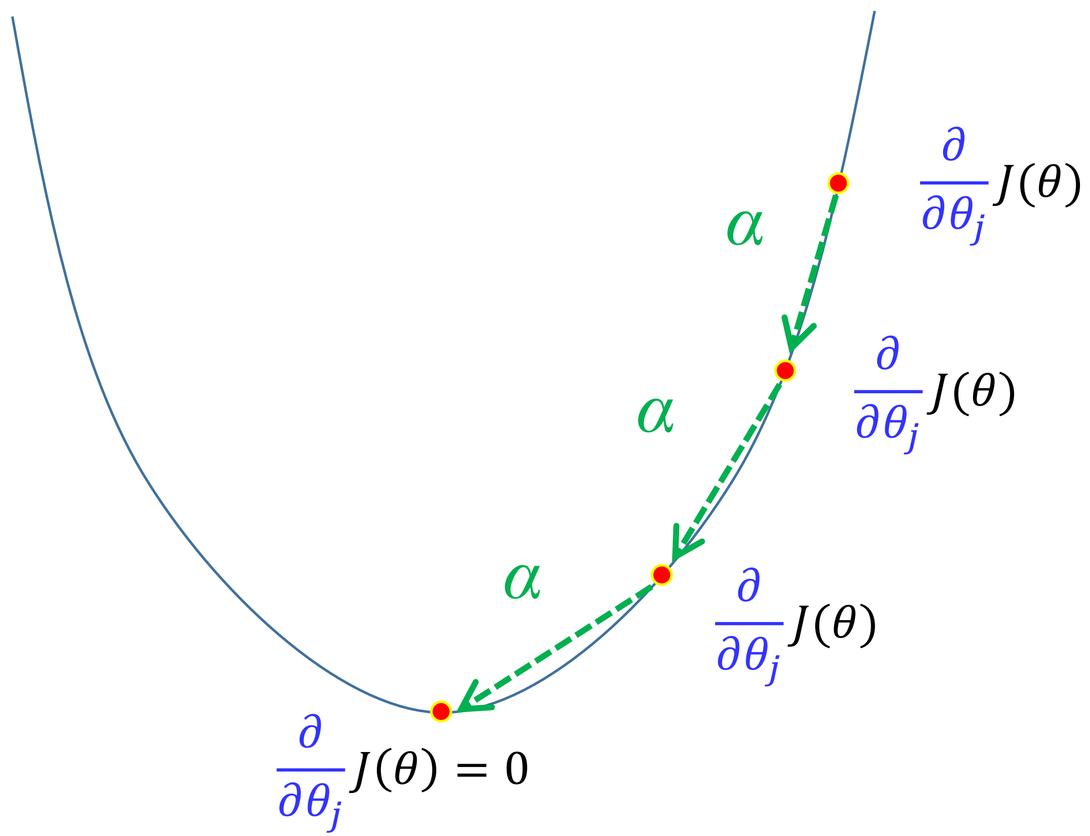
$$Tangent \rightarrow \frac{\partial}{\partial \theta_j} J(\theta)$$

Derivative

Partial (∂) / Total (d)

Gradient Descent

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



Gradient Descent Algorithm

Resuming the explanation with 2 Parameters

$j = 0,\ j = 1$

repeat until convergence {

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

}

Gradient Descent Algorithm

$j = 0,\ j = 1$

repeat until convergence {

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

}

Update **MUST** be SIMULTANEOUS

Gradient Descent Algorithm Implementation

Incorrect

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

Correct

$$
\begin{aligned}
&\textit{repeat until convergence } \{ \\
&\qquad t0 := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) \\
&\qquad t1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) \\
&\qquad \theta_0 := t0 \\
&\qquad \theta_1 := t1 \\
&\}
\end{aligned}
$$

Gradient Descent Algorithm

$j = 0,\ j = 1$

repeat until convergence {

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

}

Gradient Descent Algorithm

$$j = 0,\ j = 1$$

$$\textit{repeat until convergence } \{$$

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

$$\}$$

Update **MUST** be
SIMULTANEOUS

Gradient Descent Algorithm

$$j = 0,\ j = 1$$

$$\textit{repeat until convergence } \{$$

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

$$\}$$

Update MUST be SIMULTANEOUS

Learning Rate

Gradient Descent Algorithm

$$j = 0,\ j = 1$$

$$repeat\ until\ convergence\ \{$$

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

$$\}$$

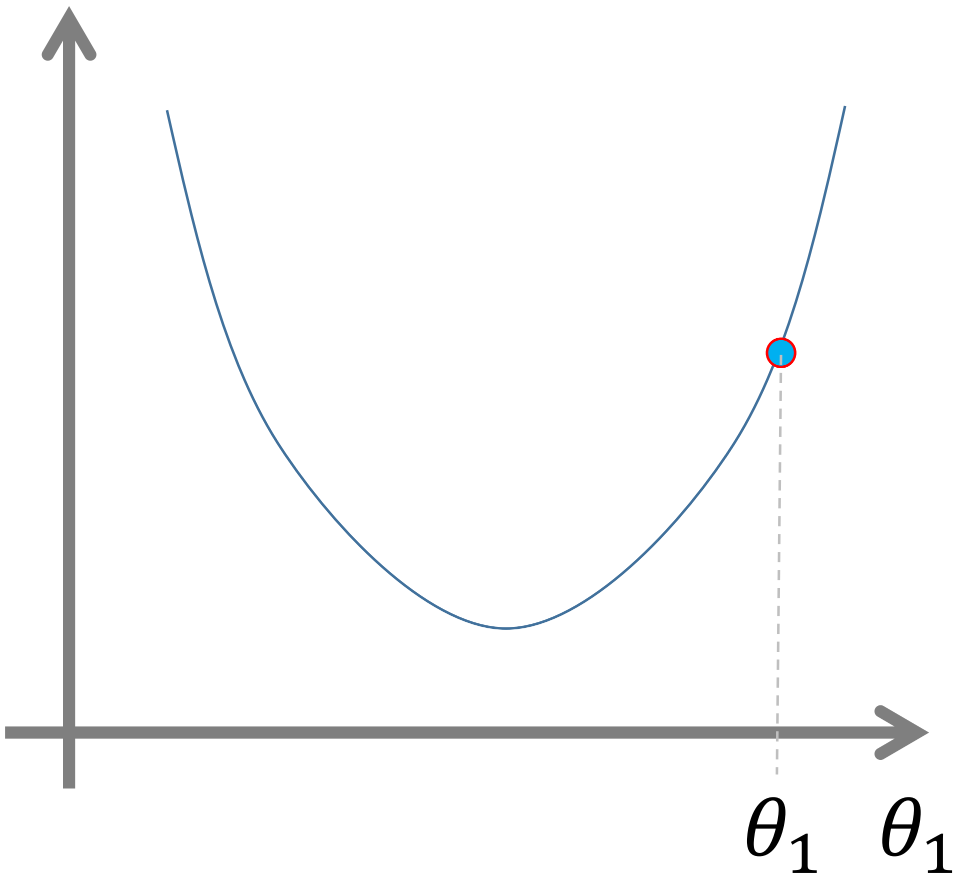
Update MUST be SIMULTANEOUS

Learning Rate

Tangent Value

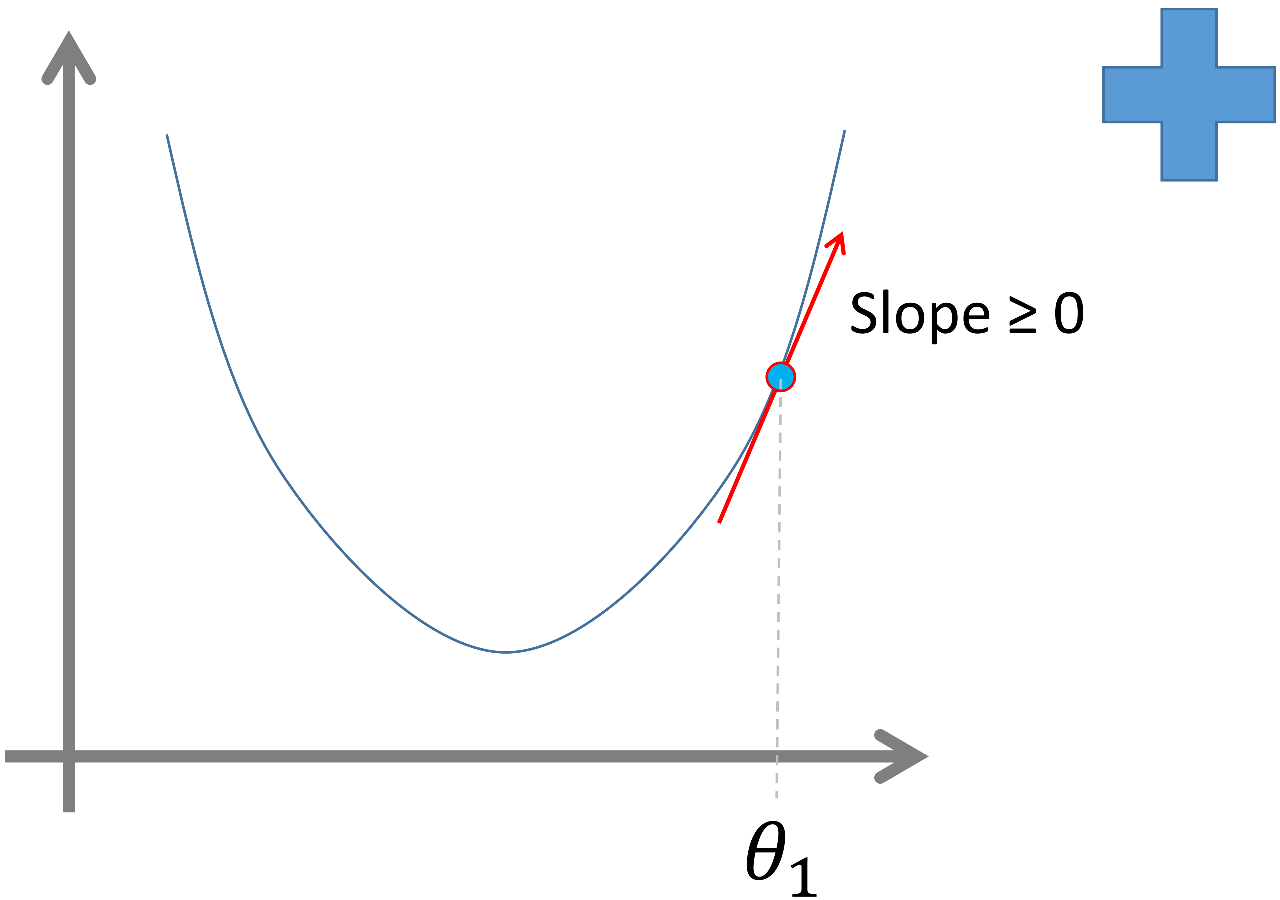
Gradient Descent

$$\color{magenta}{Let\ \theta_0 := 0;}\ \theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$$



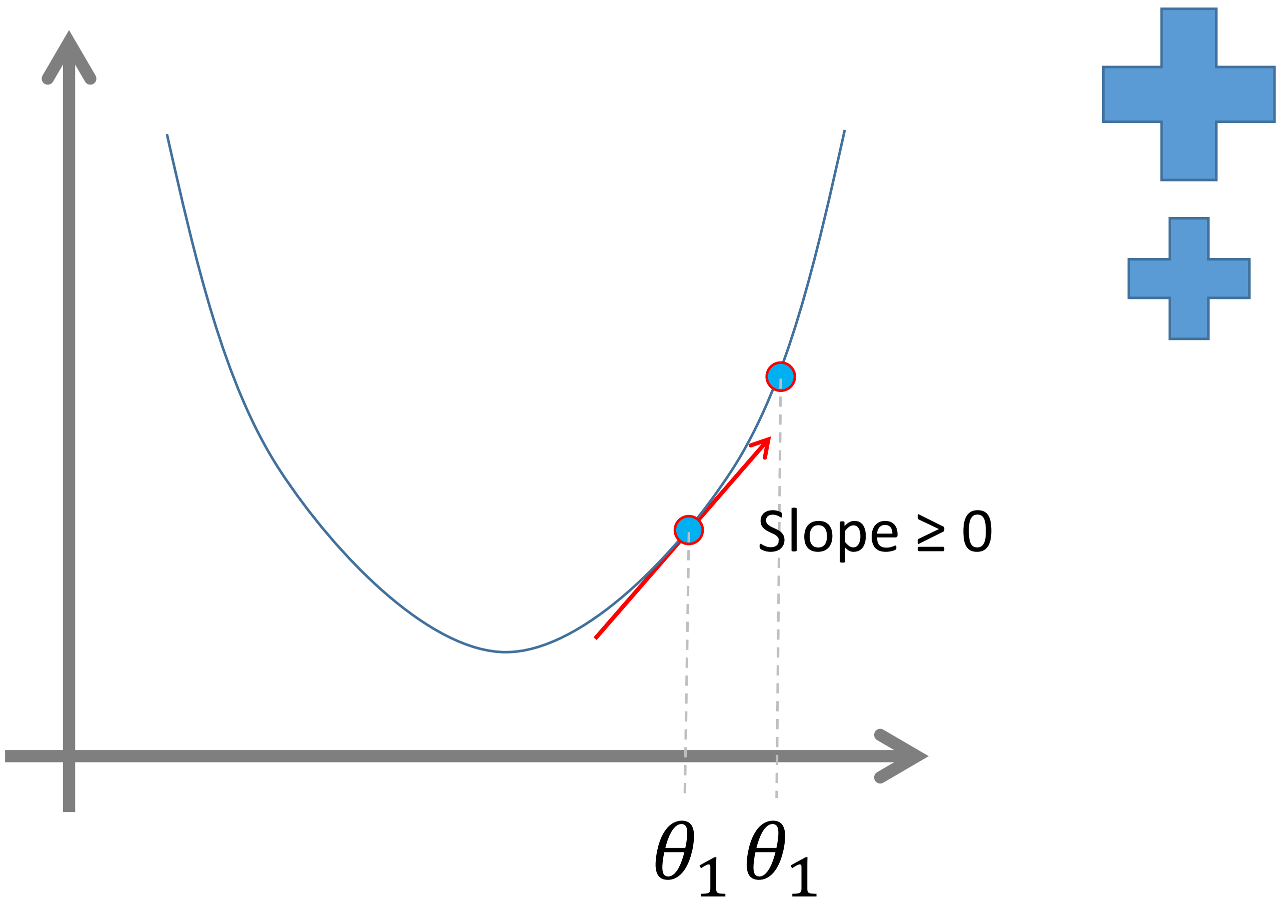
Gradient Descent

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



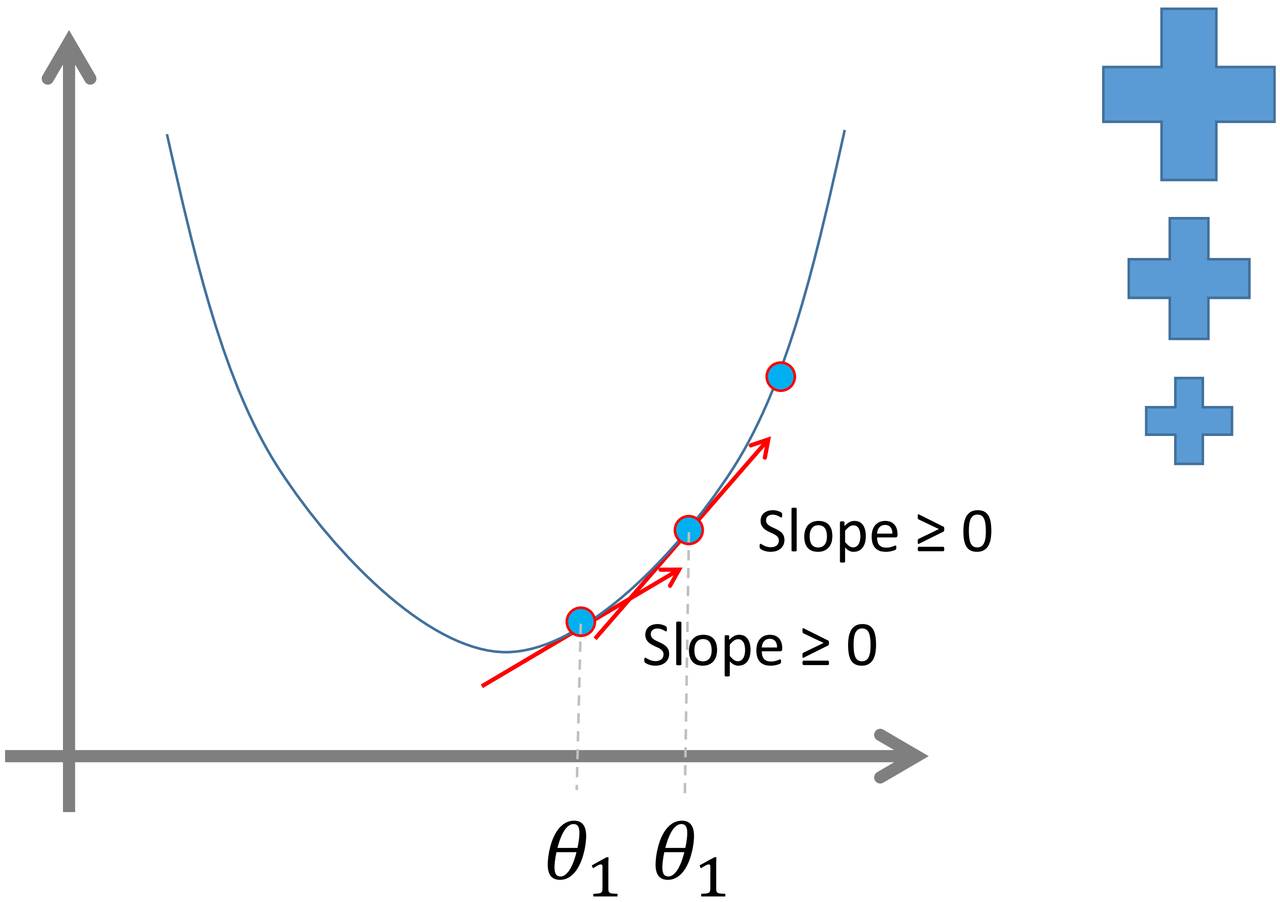
Gradient Descent

$$\color{magenta}{Let\ \theta_0 := 0;}\ \theta_1 := \theta_1 - \alpha \color{red}{\frac{d}{d\theta_1} J(\theta_1)}$$



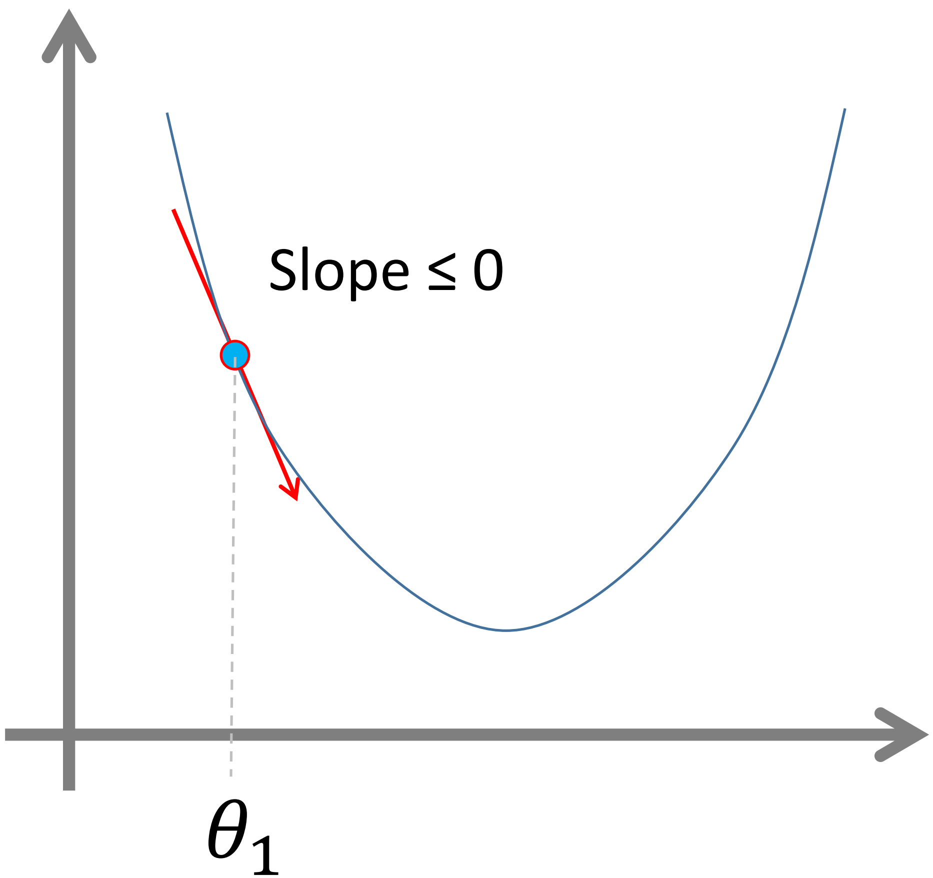
Gradient Descent

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



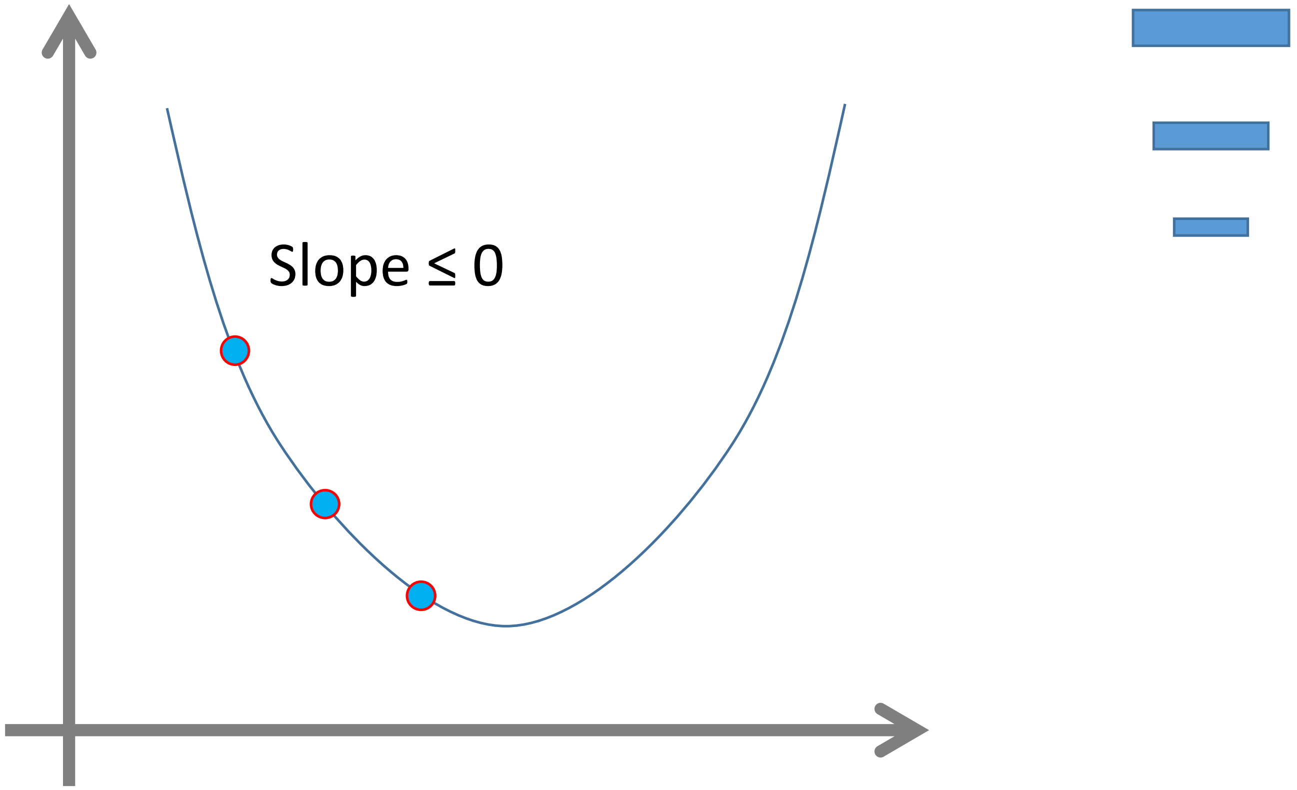
Gradient Descent

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



Gradient Descent

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



Exercise

- As we approach a local minimum, will Gradient Descent automatically take smaller steps?

- Do we need to adjust the Learning Rate (α) manually over time?

Solution

- As we approach a local minimum, will Gradient Descent automatically take smaller steps?

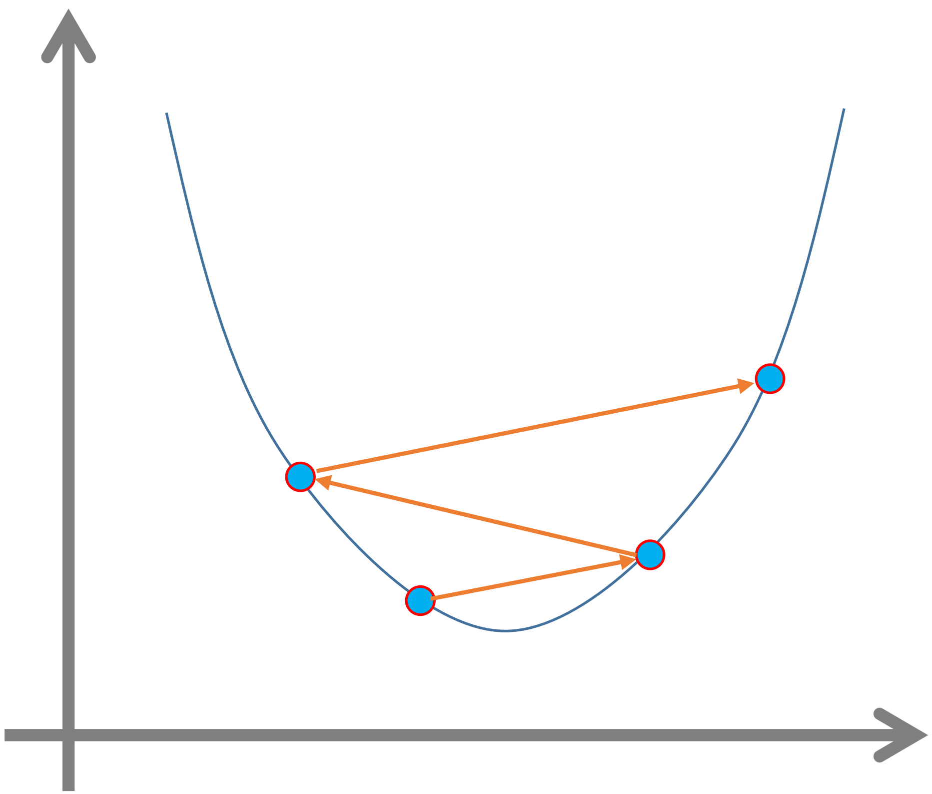
    Yes

- Do we need to adjust the Learning Rate (α) manually over time?

    No

Gradient Descent

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



If α is too high?

Gradient Descent can overshoot the minimum; it may fail to converge, or even diverge.

Linear Regression Model

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

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

Gradient Descent Algorithm –> $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

$$j = 0,\ j = 1$$

$$\textit{repeat until convergence } \{$$

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

$$\}$$

Derivative

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

Derivative

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

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

Derivative

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

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

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

Derivative

$$\frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2 \; ; \; j \in \{0, 1\}$$

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

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

List of Derivative Rules

https://www.math.ucdavis.edu/~kouba/Math17BHWDIRECTORY/Derivatives.pdf

Power Rule: $\mathrm{f}(x) = x^n$ then $\mathrm{f}'(x) = nx^{n-1}$

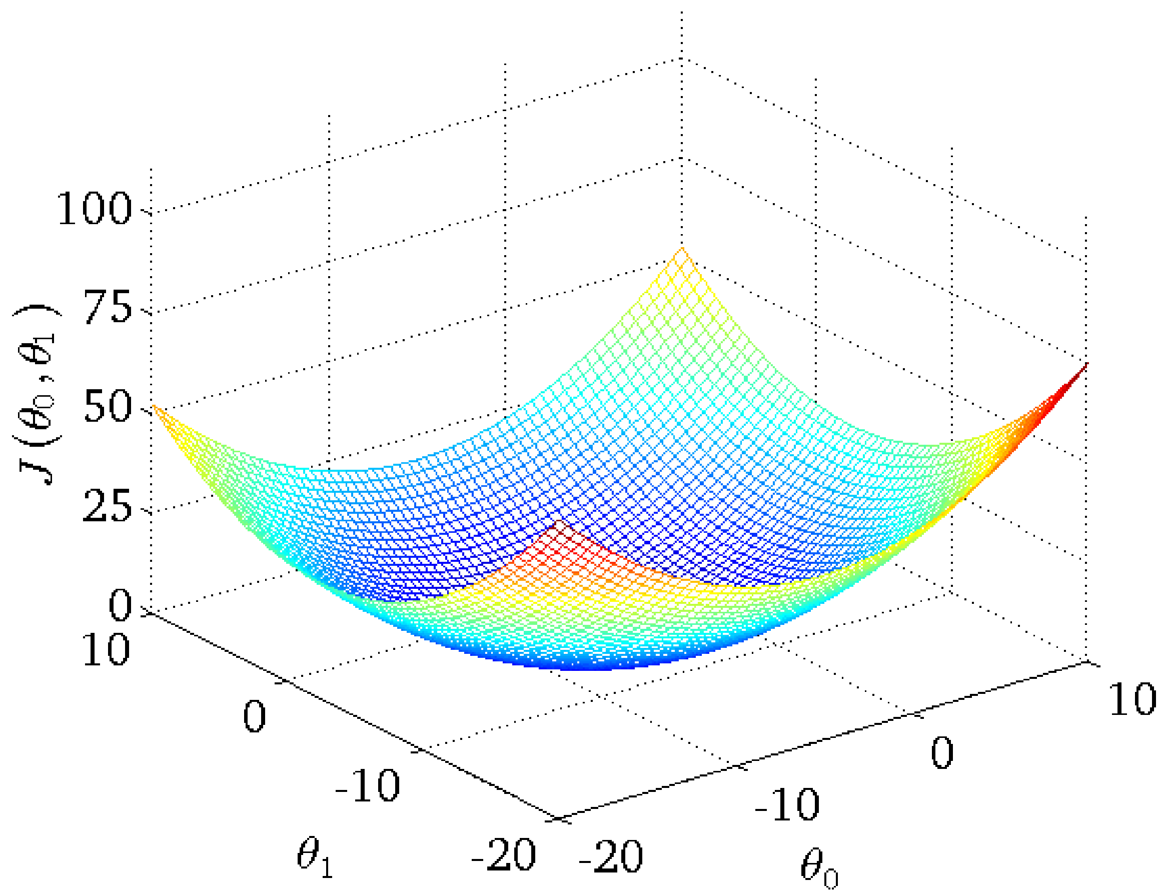
Batch Gradient Descent

Batch = use **ALL** Training Examples

in each step of Gradient Descent

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

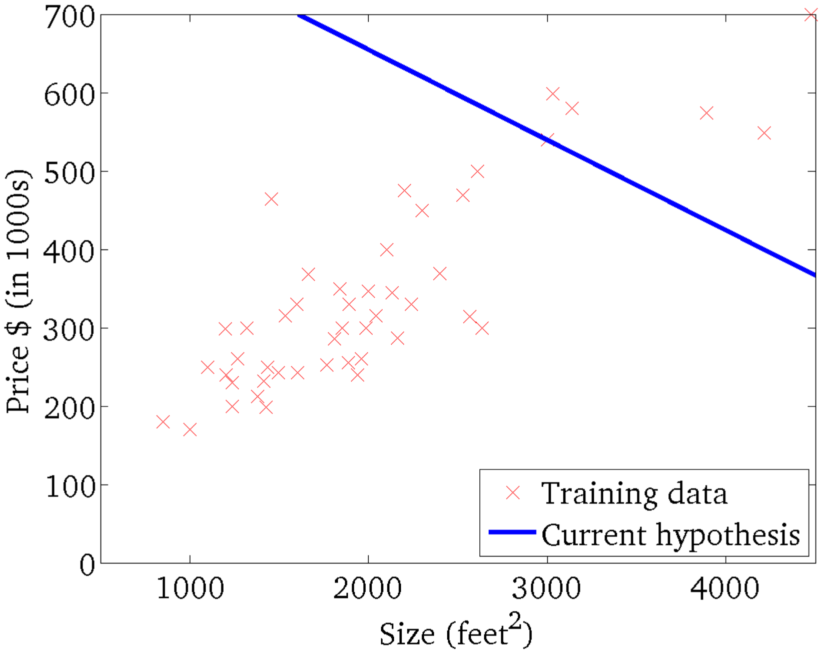
Batch Gradient Descent



Batch Gradient Descent

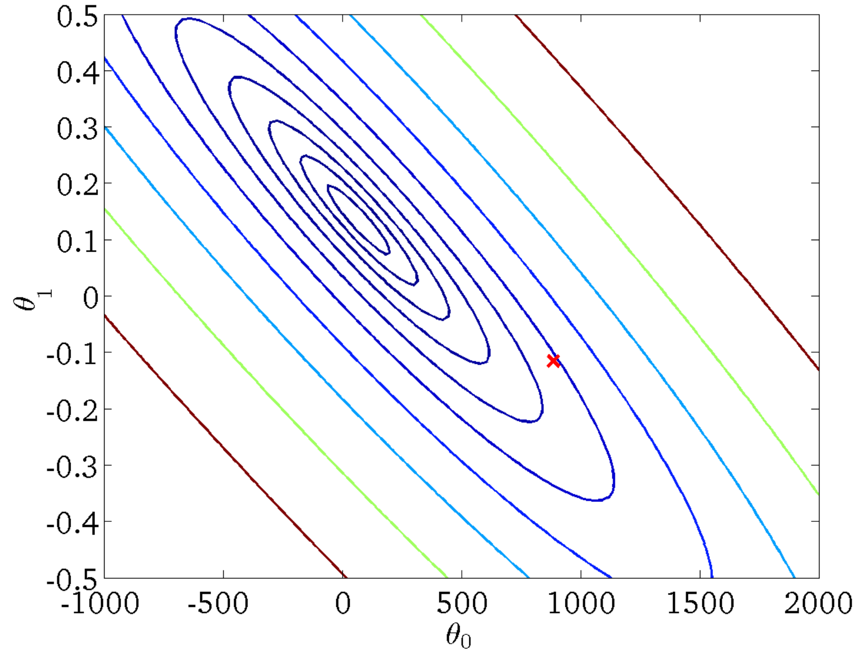
$h_\theta(x)$

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



[redacted]

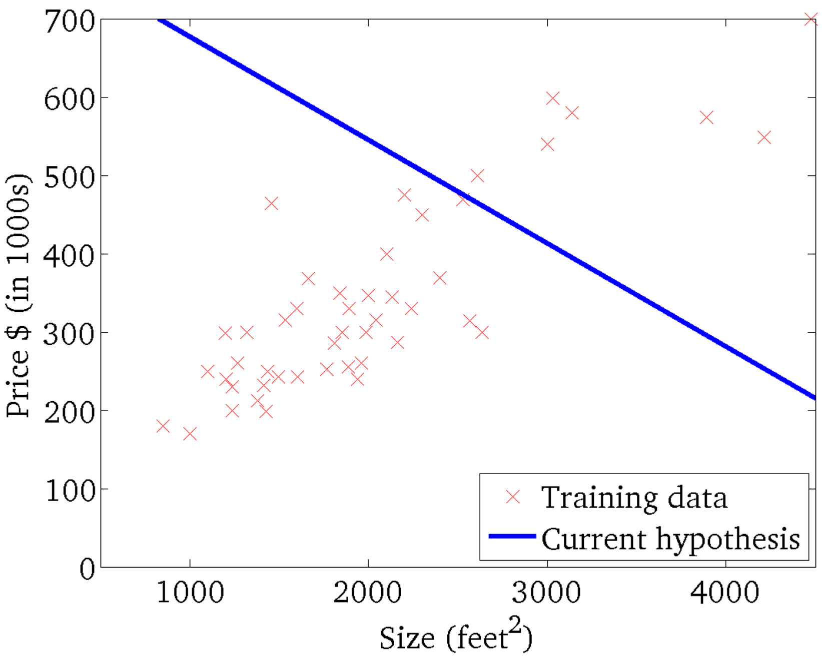
(function of the parameters θ_0, θ_1)



Batch Gradient Descent

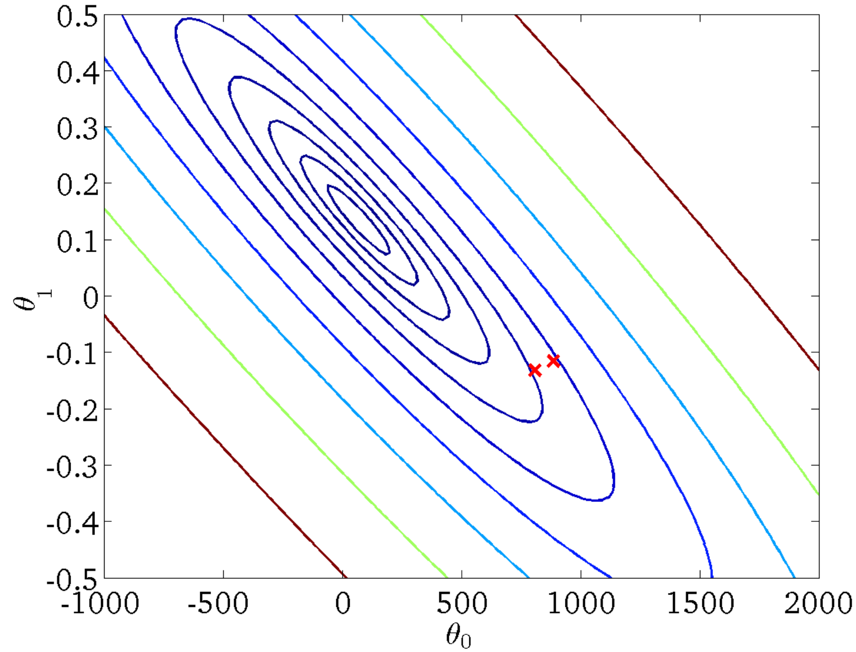
$h_\theta(x)$

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



[redacted]

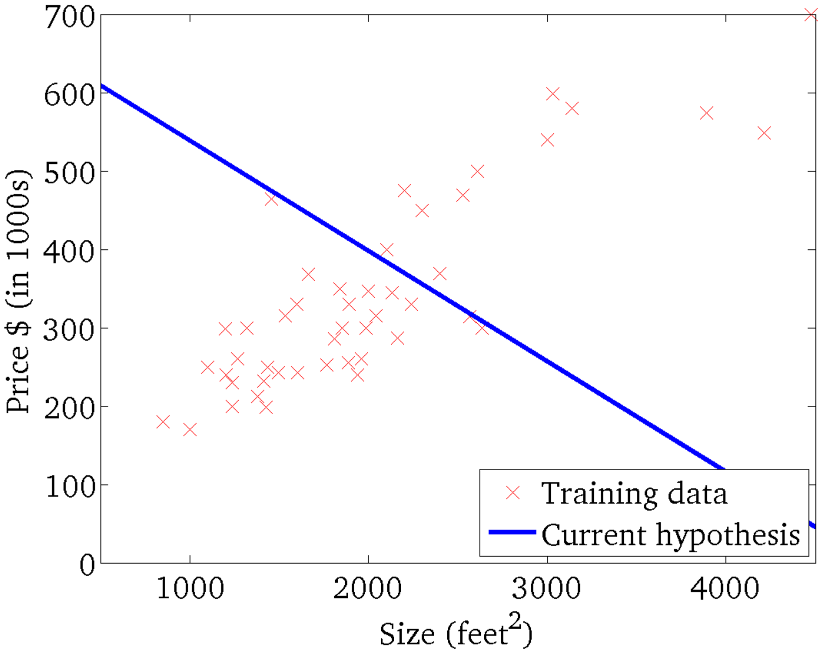
(function of the parameters θ_0, θ_1)



Batch Gradient Descent

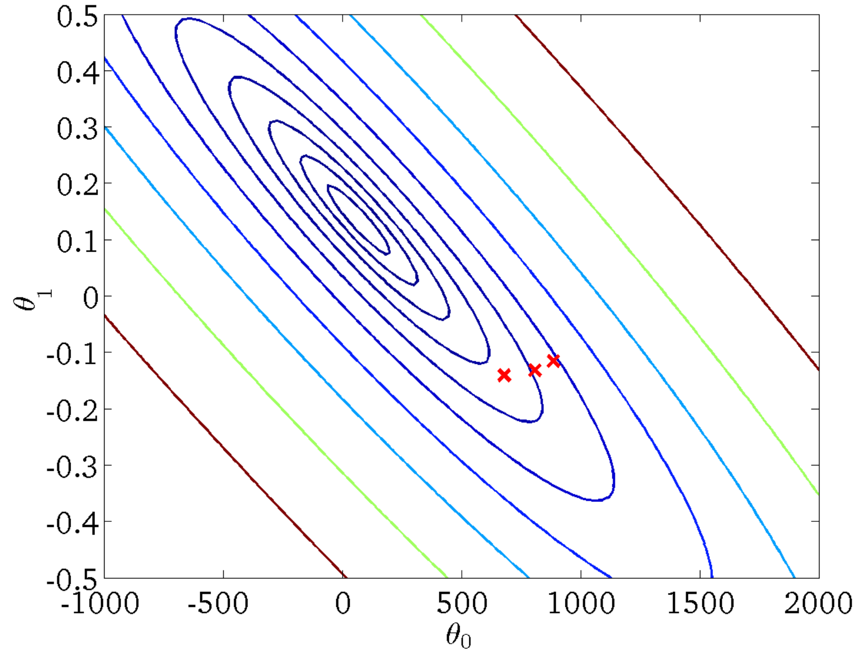
$h_\theta(x)$

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



[redacted]

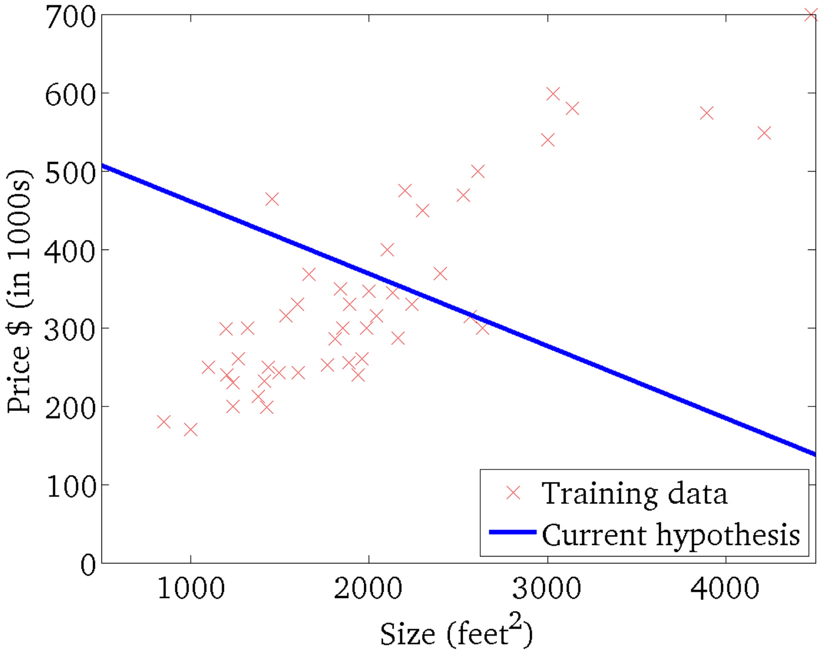
(function of the parameters θ_0, θ_1)



Batch Gradient Descent

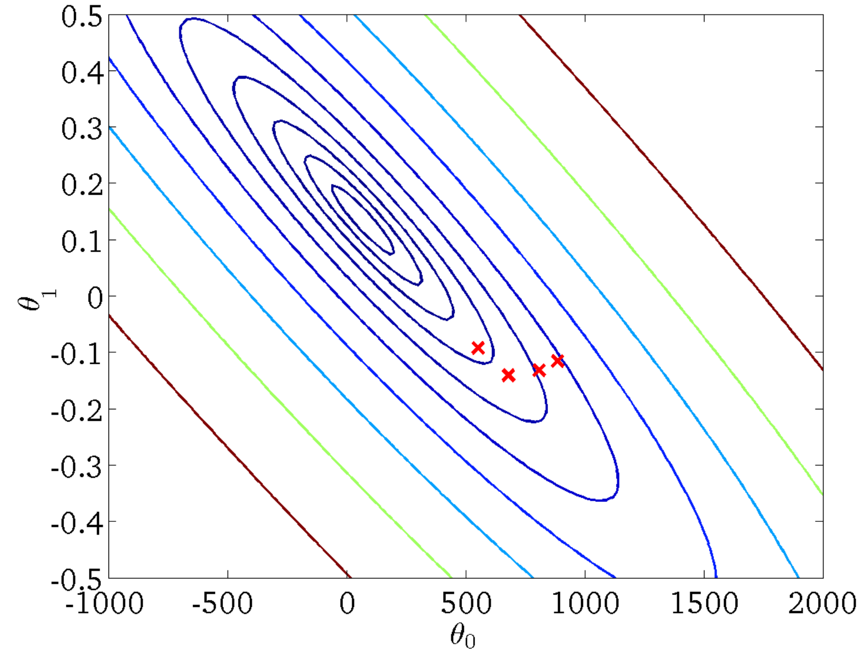
$h_\theta(x)$

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



[redacted]

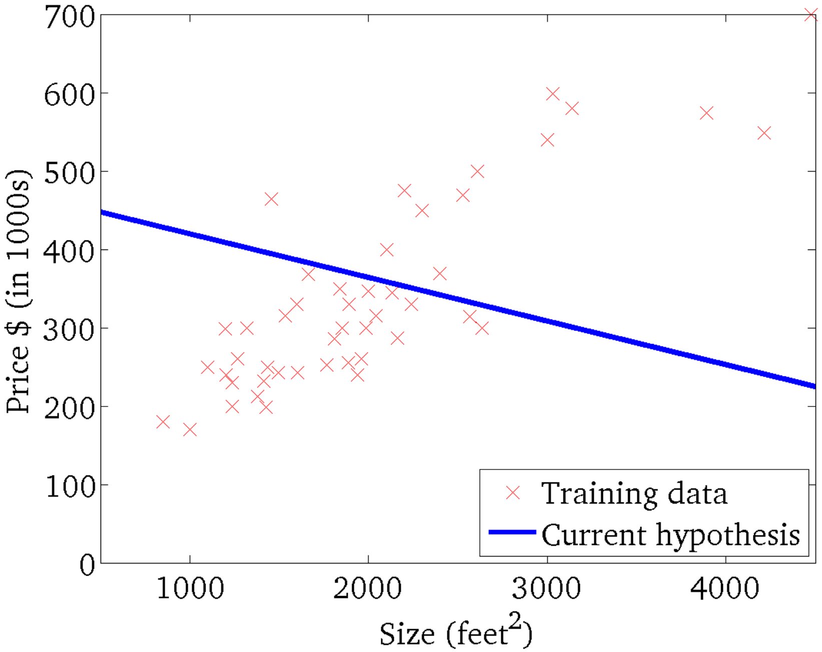
(function of the parameters θ_0, θ_1)



Batch Gradient Descent

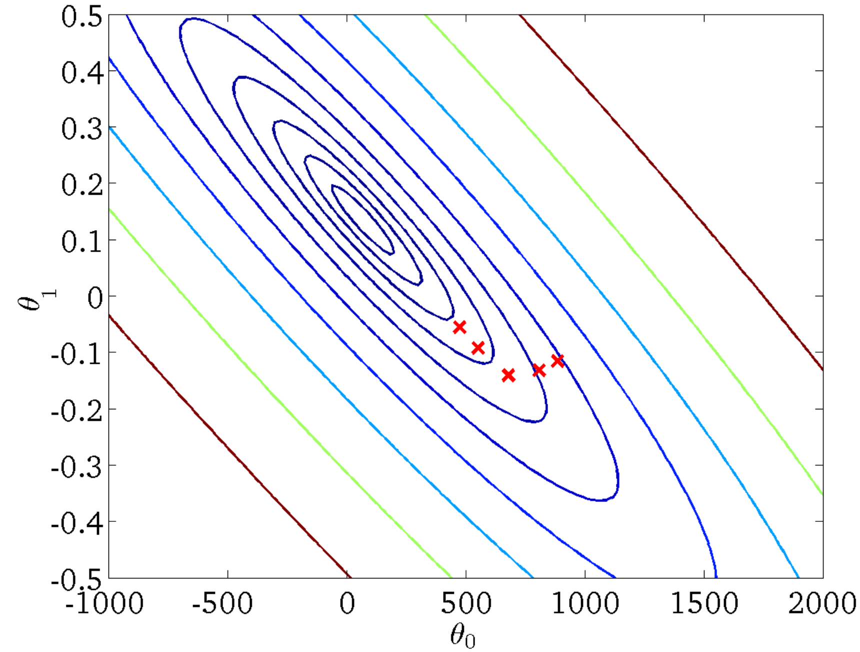
$$h_\theta(x)$$

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



[redacted]

(function of the parameters θ_0, θ_1)



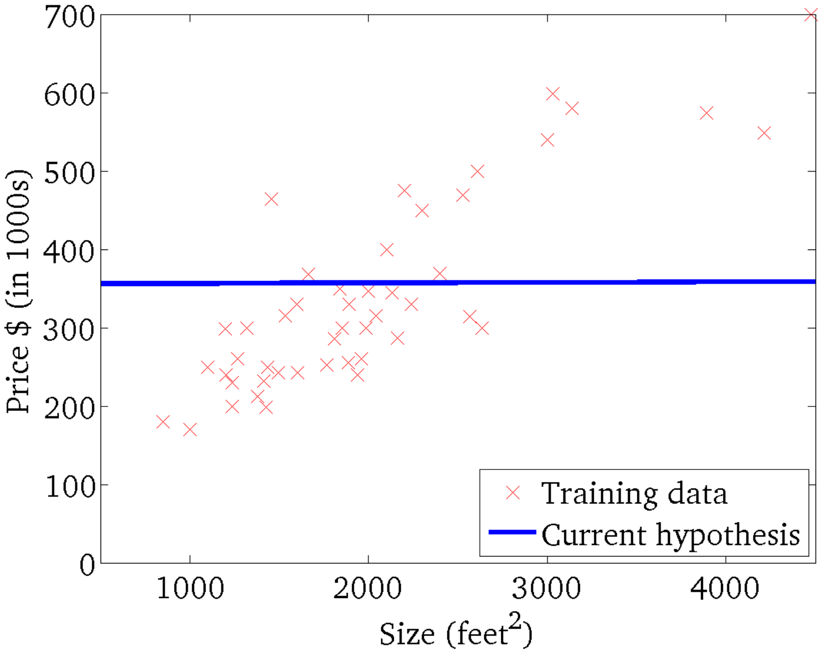
Batch Gradient Descent

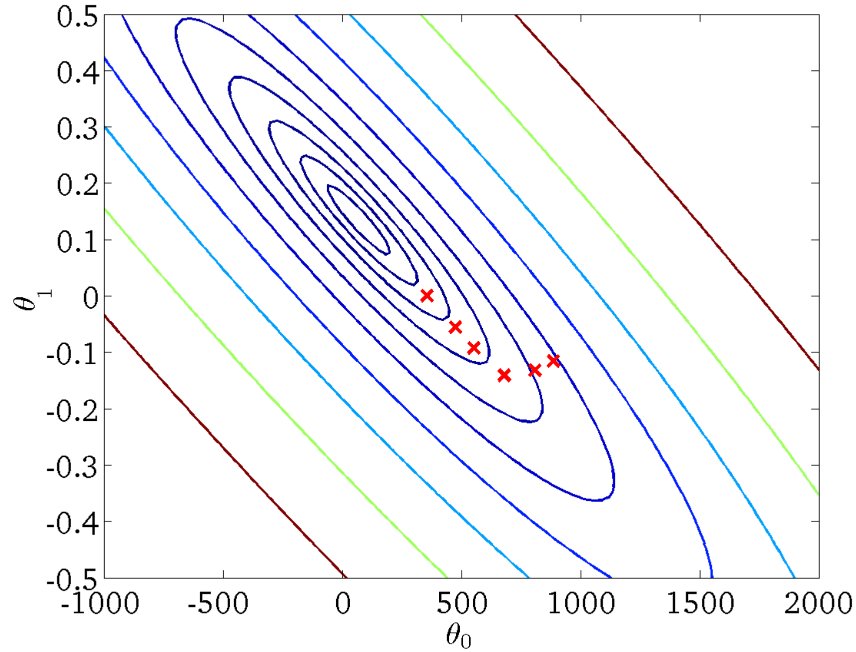
$h_\theta(x)$

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

[redacted]

(function of the parameters θ_0, θ_1)





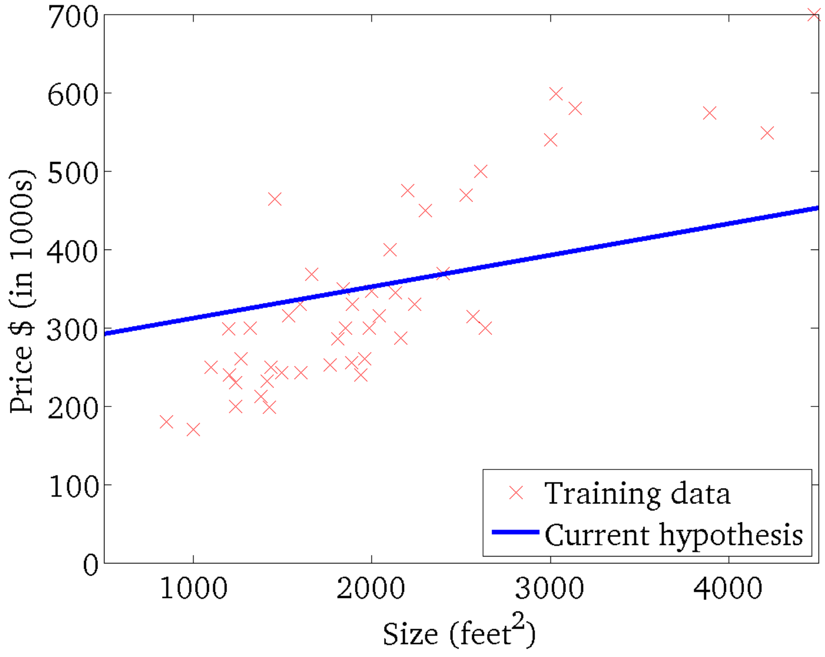
Batch Gradient Descent

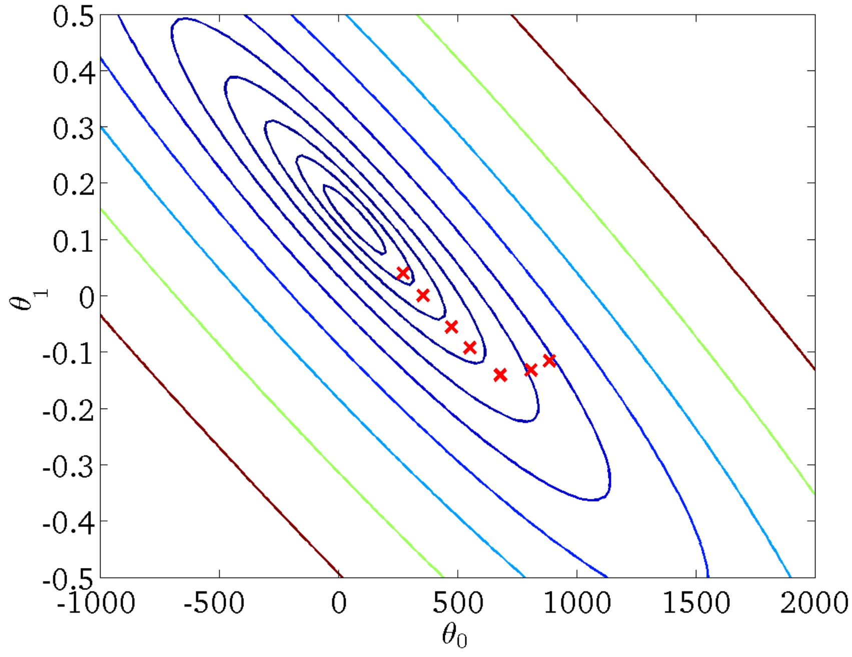
$h_\theta(x)$

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

[redacted]

(function of the parameters θ_0, θ_1)

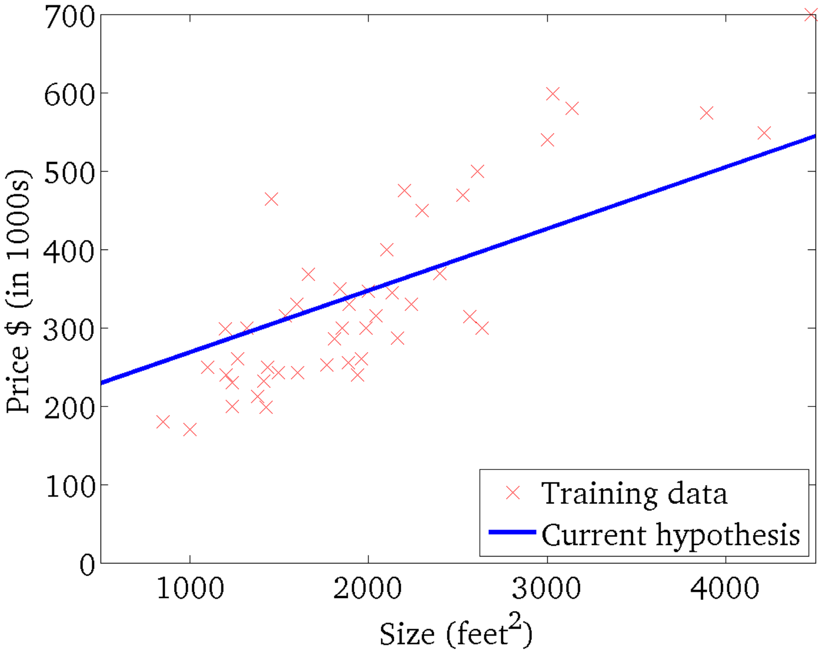




Batch Gradient Descent

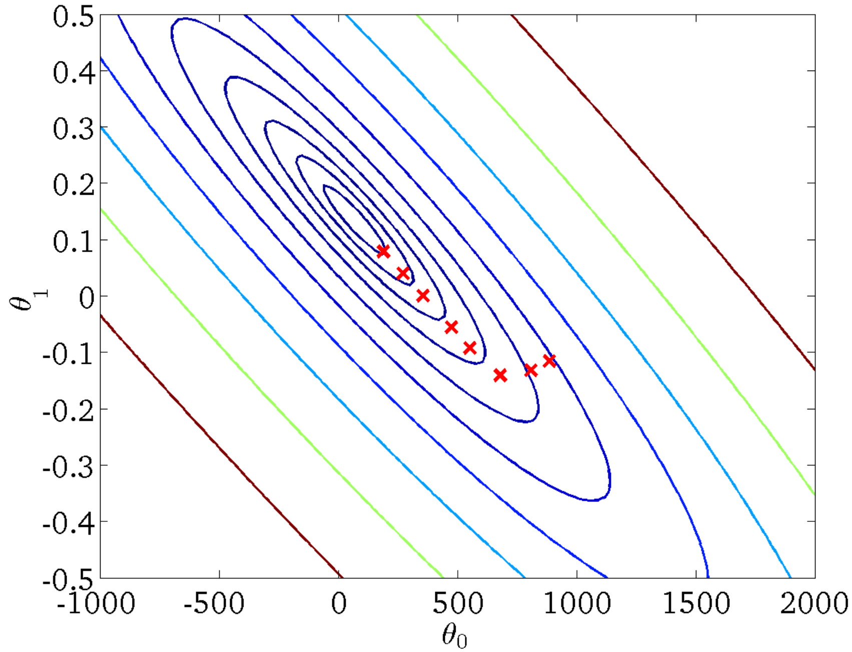
$h_\theta(x)$

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



[redacted]

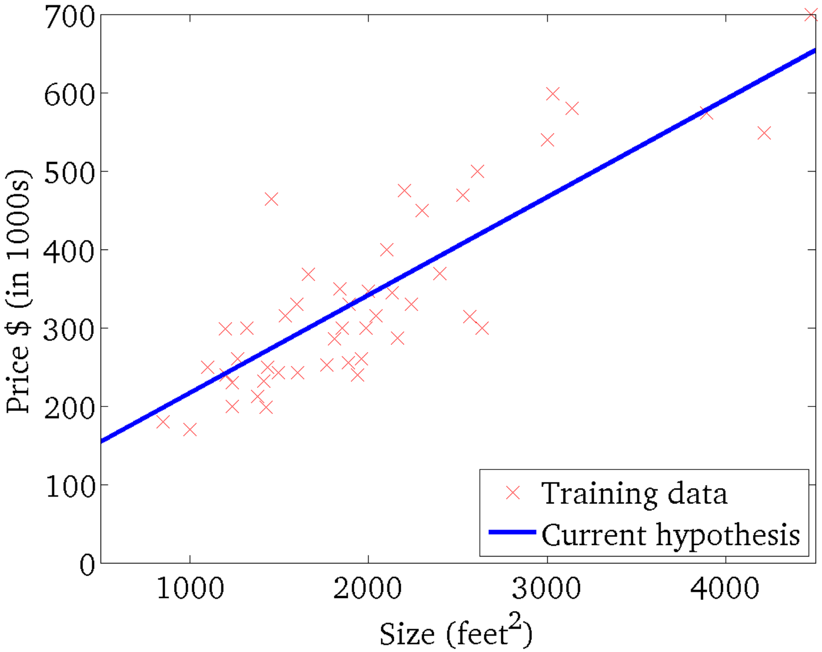
(function of the parameters θ_0, θ_1)



Batch Gradient Descent

$h_\theta(x)$

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



[redacted]

(function of the parameters θ_0, θ_1)