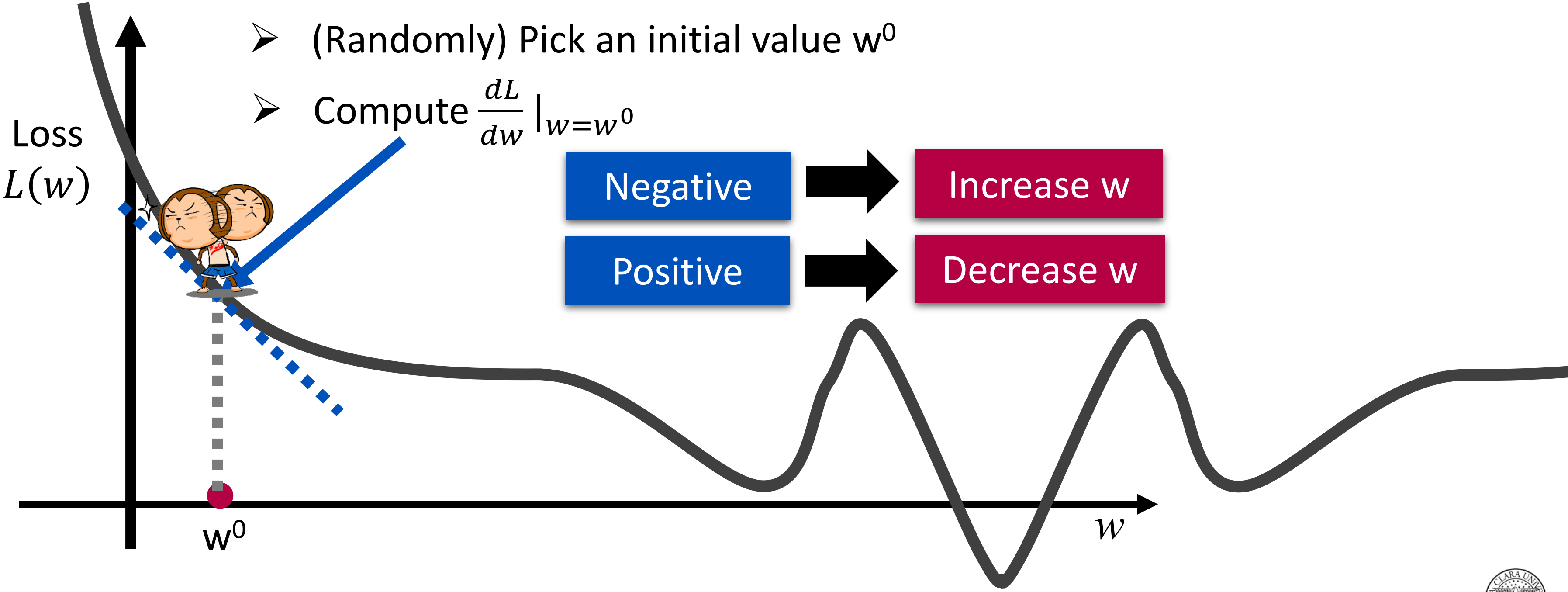
Gradient Descent

Gradient Descent: Part I

$$w^* = arg \min_w L(w)$$

Consider loss function $L(w)$ with one parameter w:

- (Randomly) Pick an initial value w^0
- Compute $\frac{dL}{dw}|_{w=w^0}$



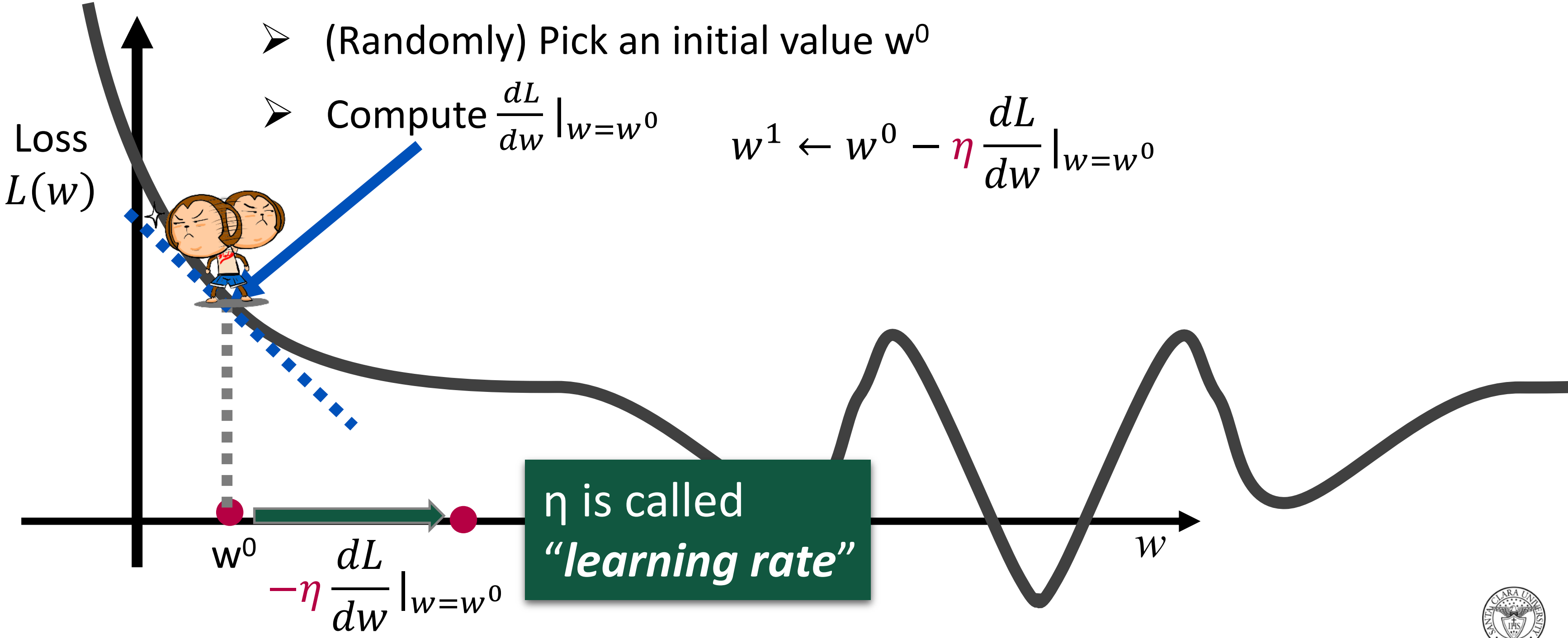
Gradient Descent: Part II

$$w^* = arg\min_{w} L(w)$$

Consider loss function $L(w)$ with one parameter w:

- (Randomly) Pick an initial value w^0
- Compute $\frac{dL}{dw}|_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta \frac{dL}{dw}|_{w=w^0}$$



η is called *"learning rate"*

Gradient Descent: Part III

$$w^* = arg\min_w L(w)$$

Consider loss function $L(w)$ with one parameter w:

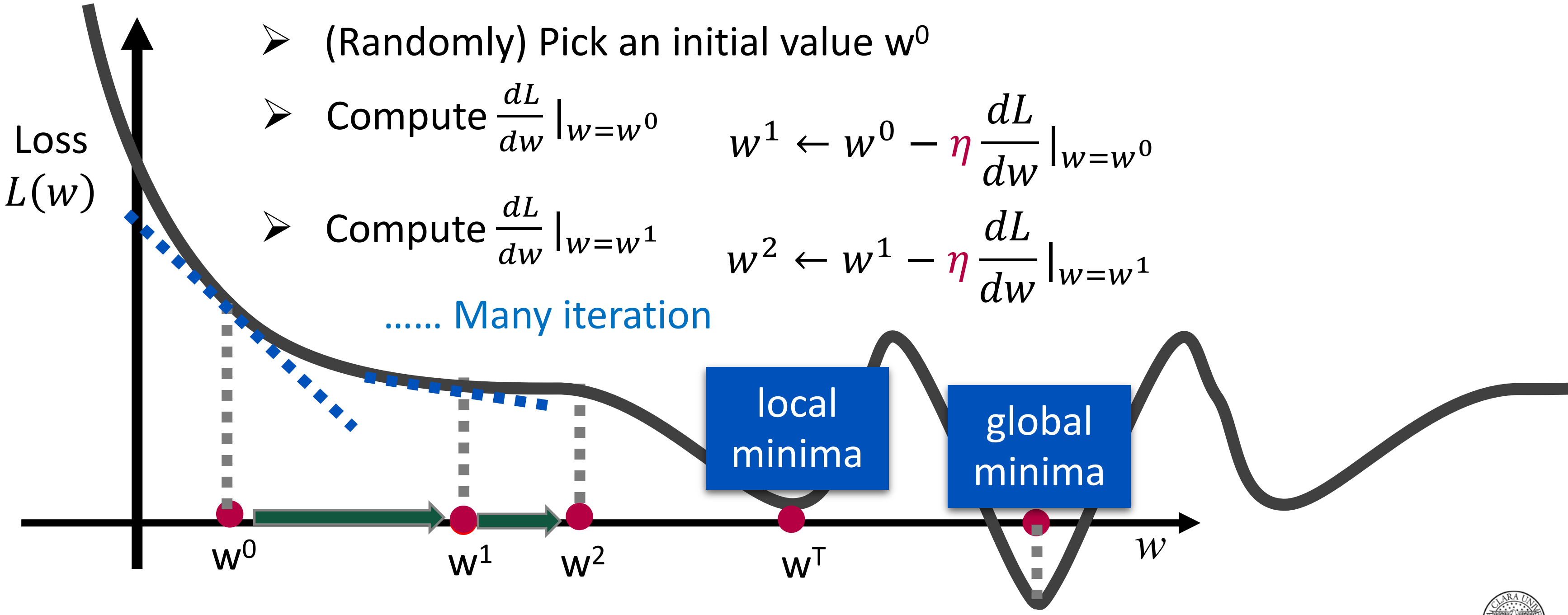
- (Randomly) Pick an initial value w^0
- Compute $\frac{dL}{dw}|_{w=w^0}$

$$w^1 \leftarrow w^0 - \eta\frac{dL}{dw}|_{w=w^0}$$

- Compute $\frac{dL}{dw}|_{w=w^1}$

$$w^2 \leftarrow w^1 - \eta\frac{dL}{dw}|_{w=w^1}$$

…… Many iteration



Gradient Descent: Part IV

$$\begin{bmatrix} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \end{bmatrix}_{\text{gradient}}$$

How about two parameters?

$$w^*, b^* = arg \min_{w,b} L(w, b)$$

- (Randomly) Pick an initial value w^0, b^0
- Compute $\frac{\partial L}{\partial w}|_{w=w^0,b=b^0}, \frac{\partial L}{\partial b}|_{w=w^0,b=b^0}$

$$w^1 \leftarrow w^0 - \eta \frac{\partial L}{\partial w}|_{w=w^0,b=b^0} \qquad b^1 \leftarrow b^0 - \eta \frac{\partial L}{\partial b}|_{w=w^0,b=b^0}$$

- Compute $\frac{\partial L}{\partial w}|_{w=w^1,b=b^1}, \frac{\partial L}{\partial b}|_{w=w^1,b=b^1}$

$$w^2 \leftarrow w^1 - \eta \frac{\partial L}{\partial w}|_{w=w^1,b=b^1} \qquad b^2 \leftarrow b^1 - \eta \frac{\partial L}{\partial b}|_{w=w^1,b=b^1}$$

Gradient Descent: Part V

Formulation of $\partial L / \partial w$ and $\partial L / \partial b$

$$L(w, b) = \sum_{n=1}^{m} \left(\hat{y}^n - \left(b + \underline{w \cdot x_{cp}^n} \right) \right)^2$$

$$\frac{\partial L}{\partial w} =? \sum_{n=1}^{m} 2 \left(\hat{y}^n - \left(b + w \cdot x_{cp}^n \right) \right)$$

$$\frac{\partial L}{\partial b} =?$$

Gradient Descent: Part VI

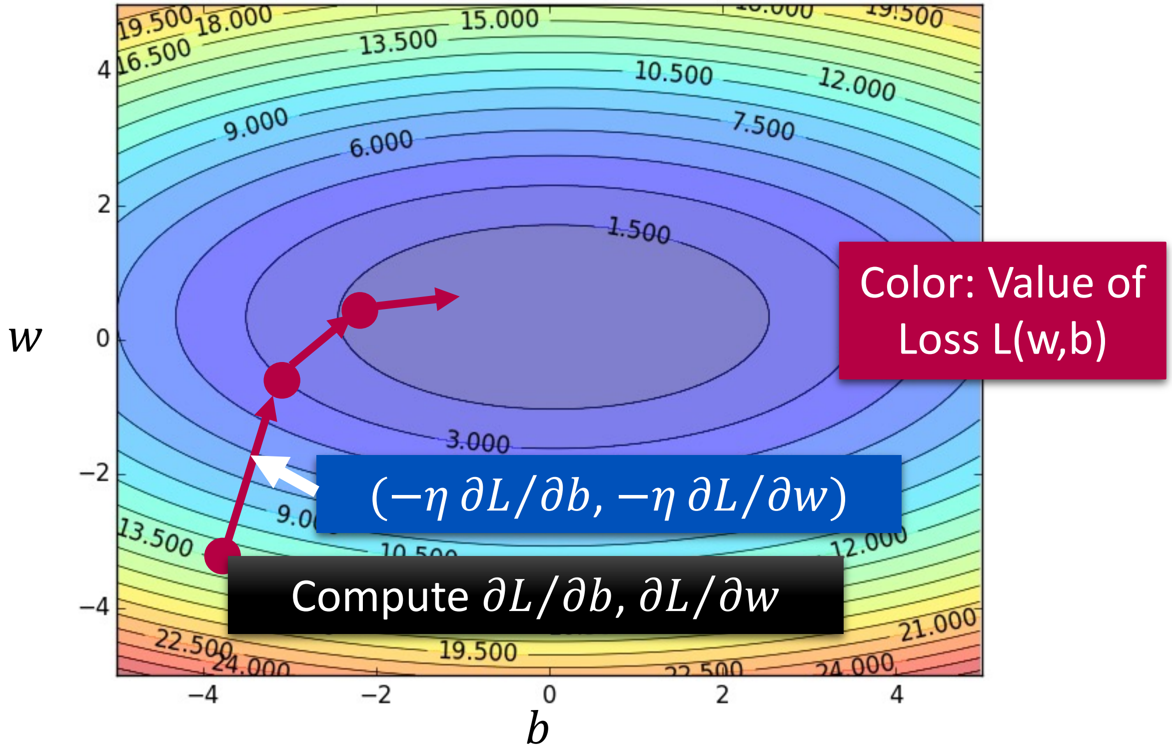
Formulation of $\partial L / \partial w$ and $\partial L / \partial b$

$$L(w, b) = \sum_{n=1}^{m} \left(\hat{y}^n - \left(\underline{b} + w \cdot x_{cp}^n\right)\right)^2$$

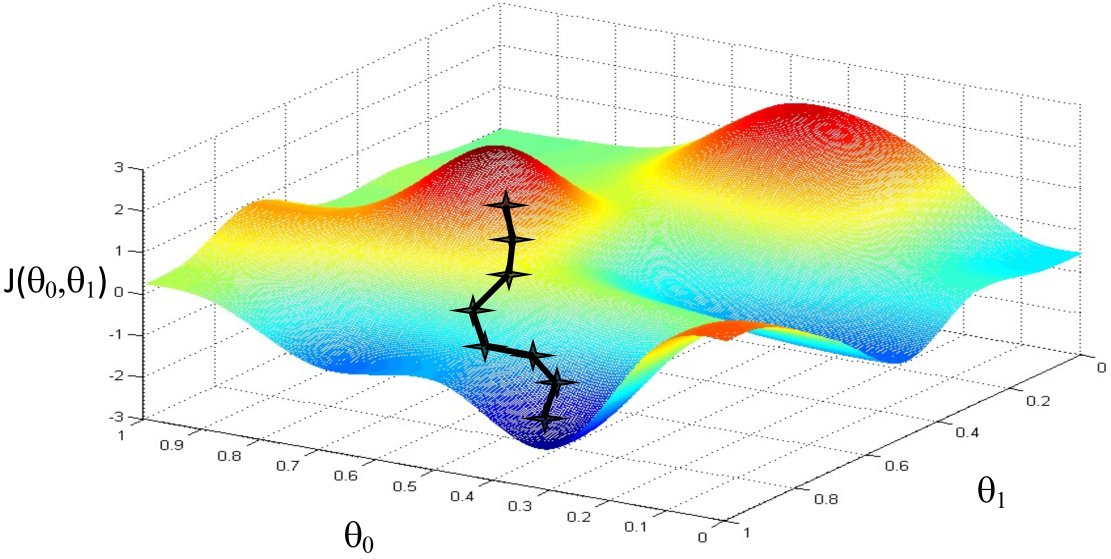
$$\frac{\partial L}{\partial w} =? \sum_{n=1}^{m} 2\left(\hat{y}^n - \left(b + w \cdot x_{cp}^n\right)\right)\left(-x_{cp}^n\right)$$

$$\frac{\partial L}{\partial b} =? \sum_{n=1}^{m} 2\left(\hat{y}^n - \left(b + w \cdot x_{cp}^n\right)\right)$$

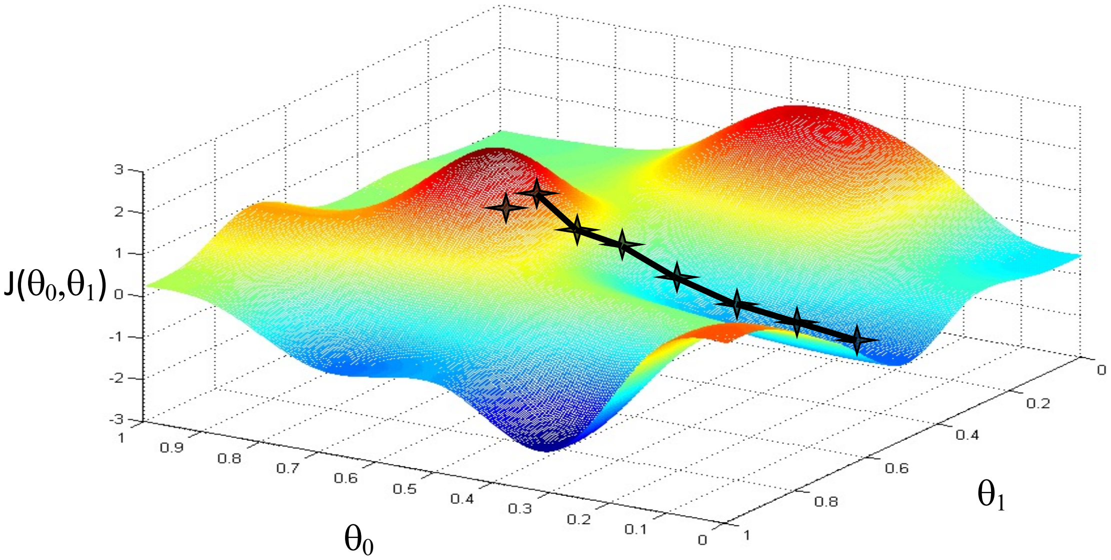
Gradient Descent: Part VII



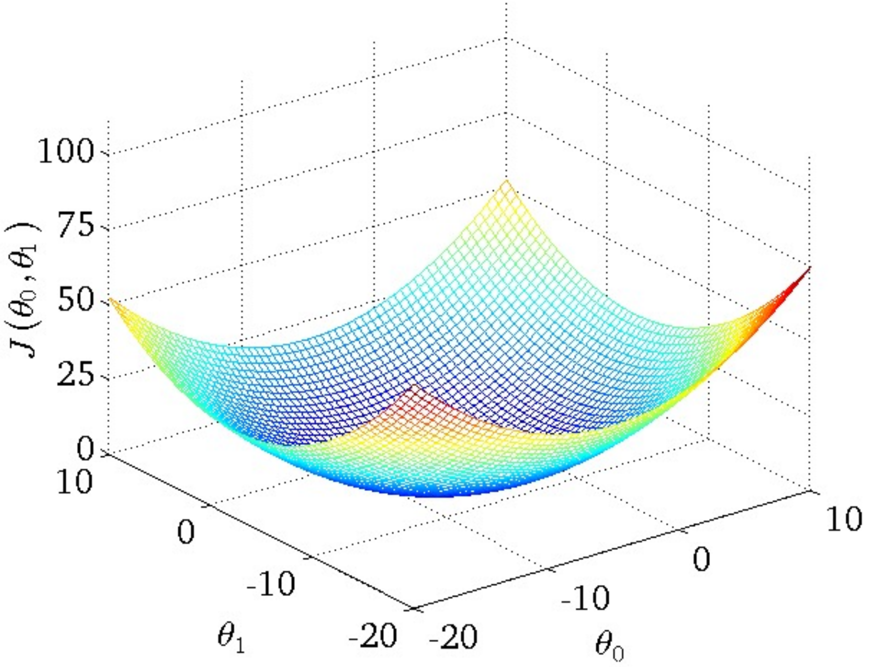
Gradient Descent: Part VIII



Gradient Descent: Part IX



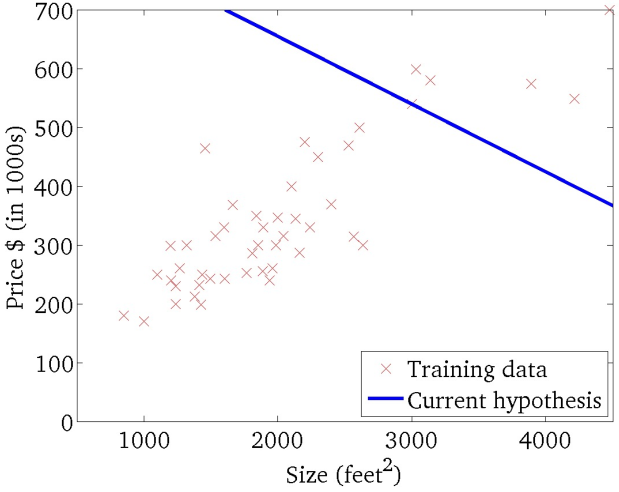
Gradient Descent for Linear Regression: Part I



Gradient Descent for Linear Regression: Part II

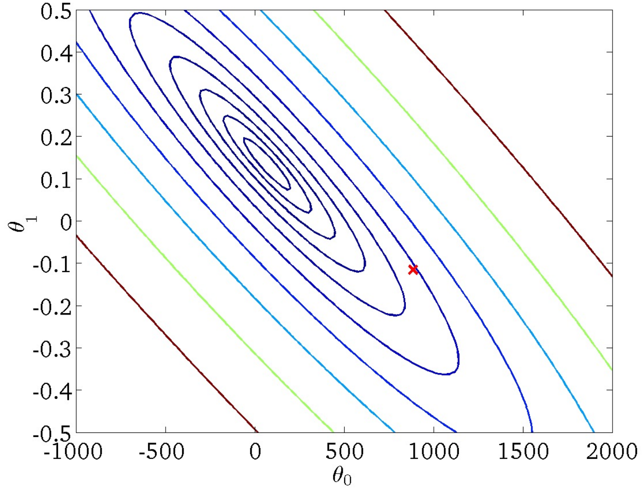
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

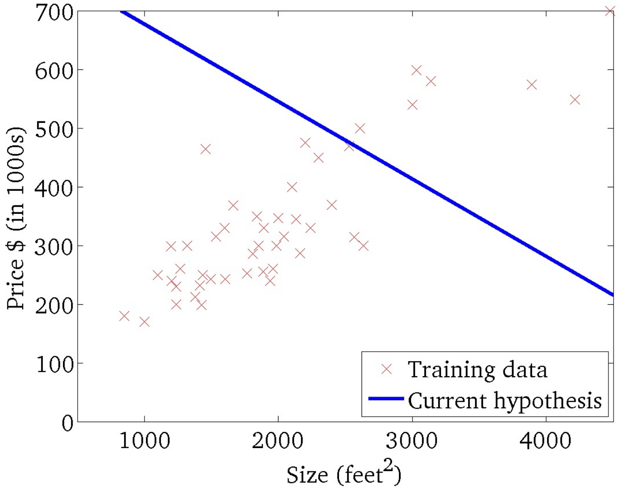
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part III

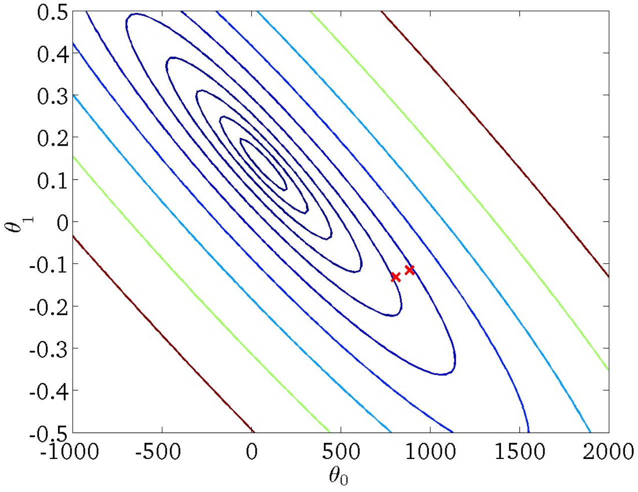
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

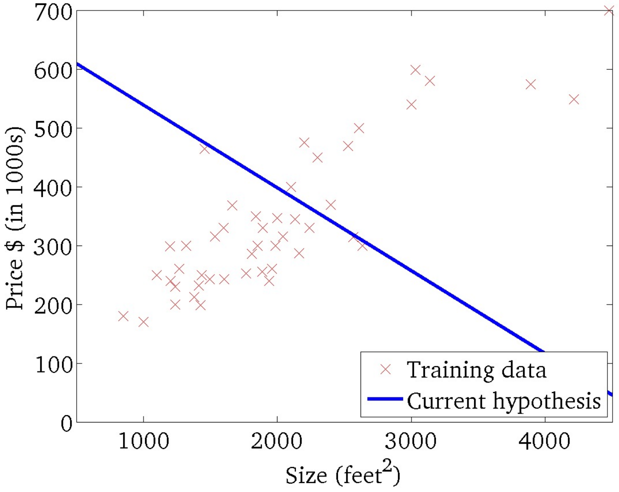
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part IV

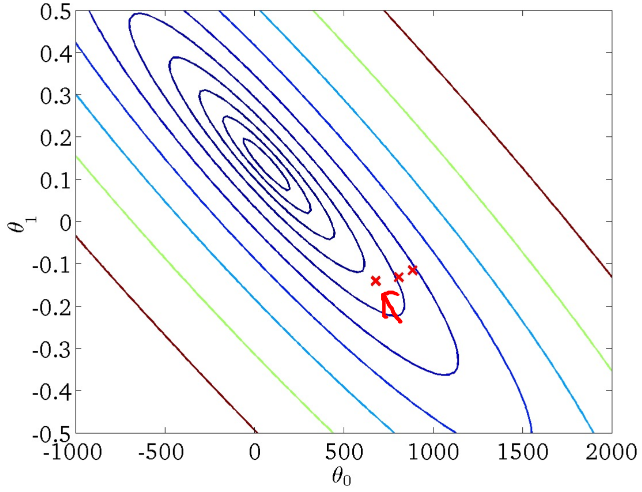
$$h_\theta(x)$$

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



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

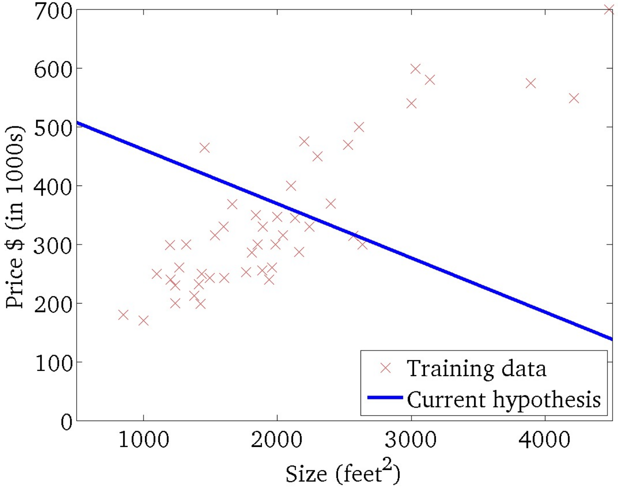
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part V

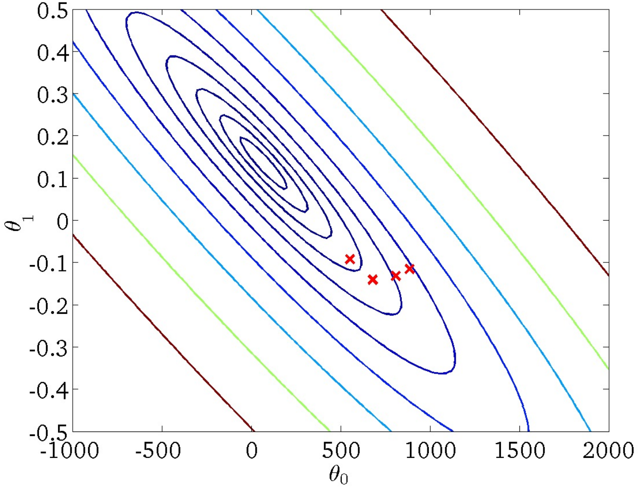
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

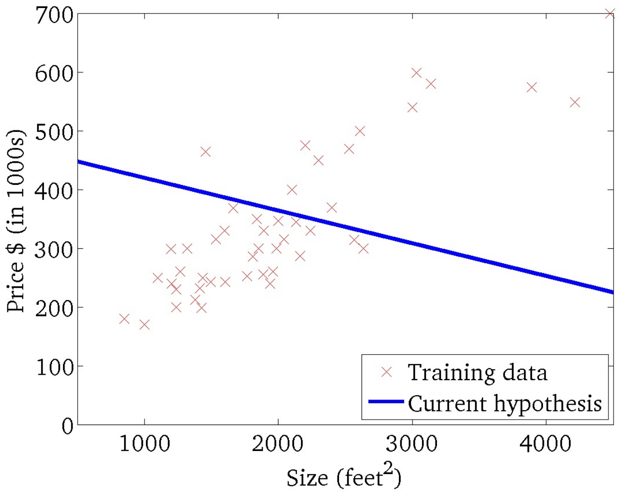
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part VI

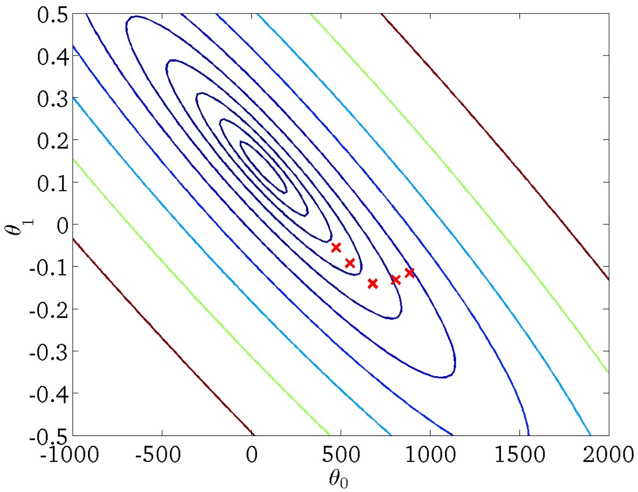
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

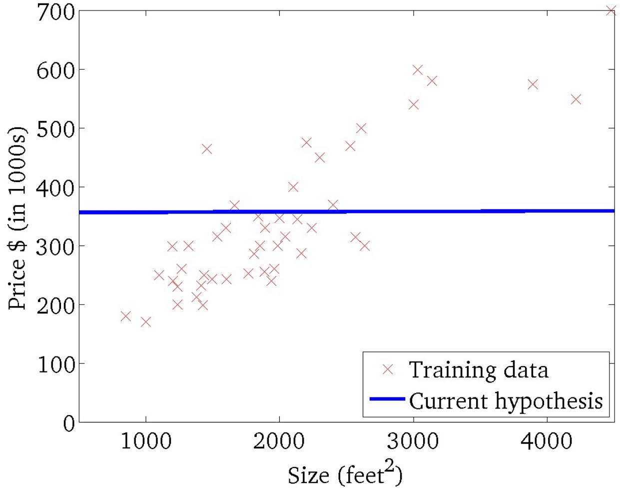
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part VII

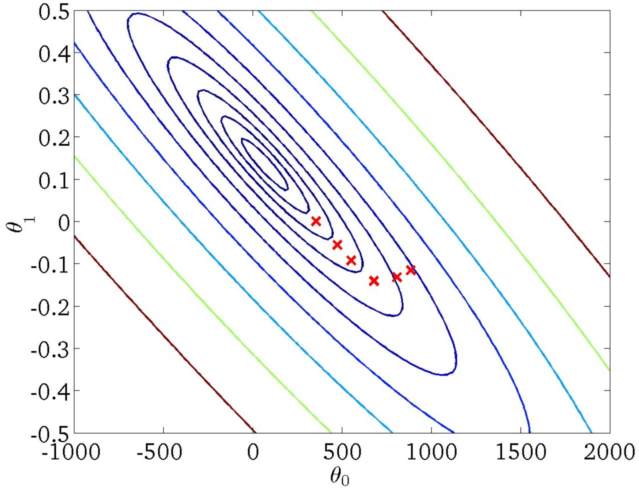
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

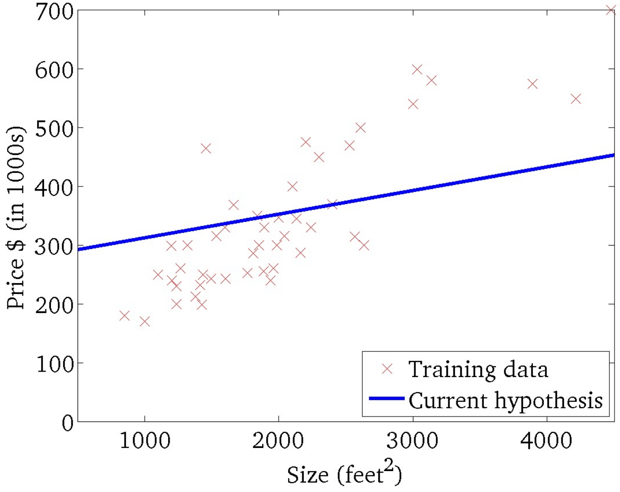
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part VIII

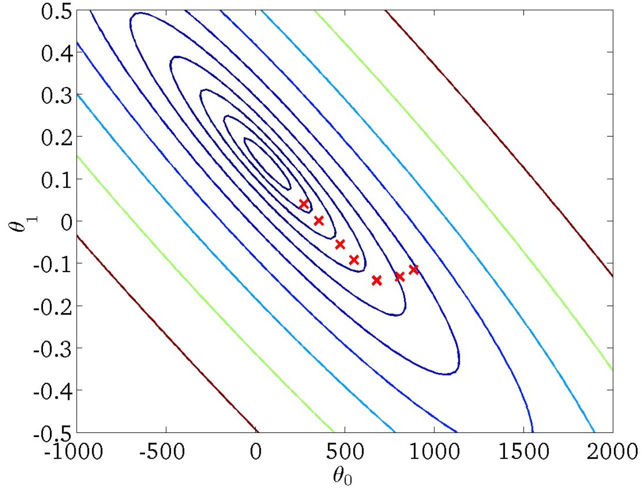
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

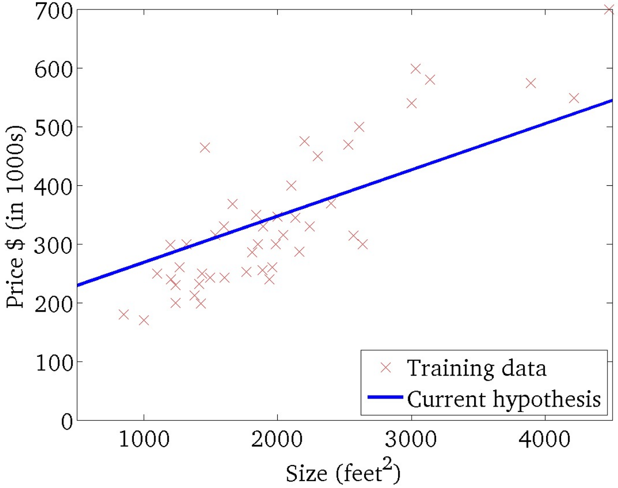
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part IX

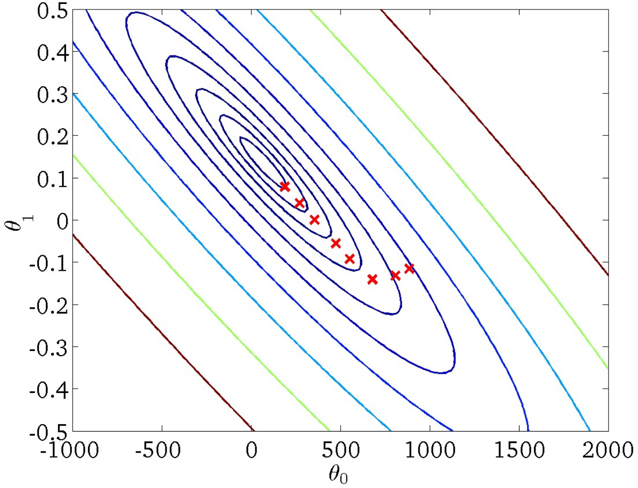
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

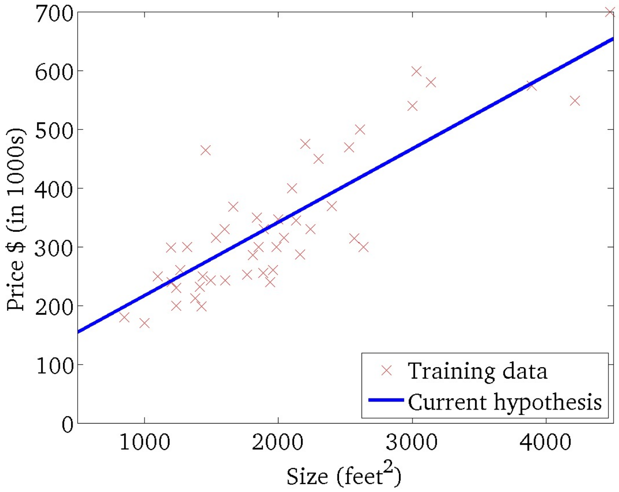
(function of the parameters θ_0, θ_1)



Gradient Descent for Linear Regression: Part X

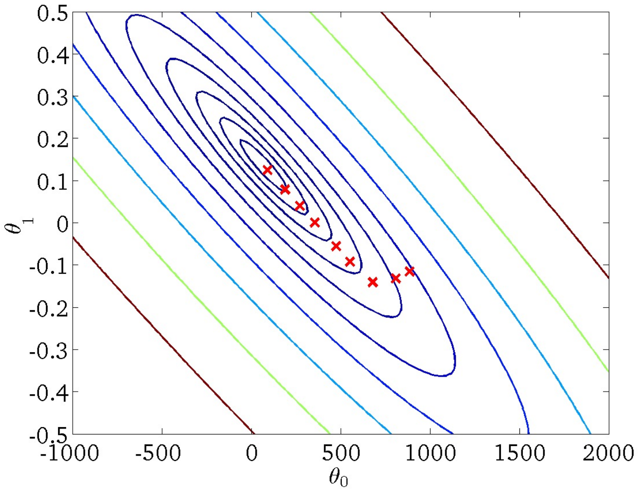
$h_\theta(x)$

(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)



Review: Gradient Descent

Optimization problem:

$$\theta^* = \arg\min_{\theta} L(\theta)$$ L: loss function θ: parameters

Suppose that θ has two variables $\{\theta_1, \theta_2\}$

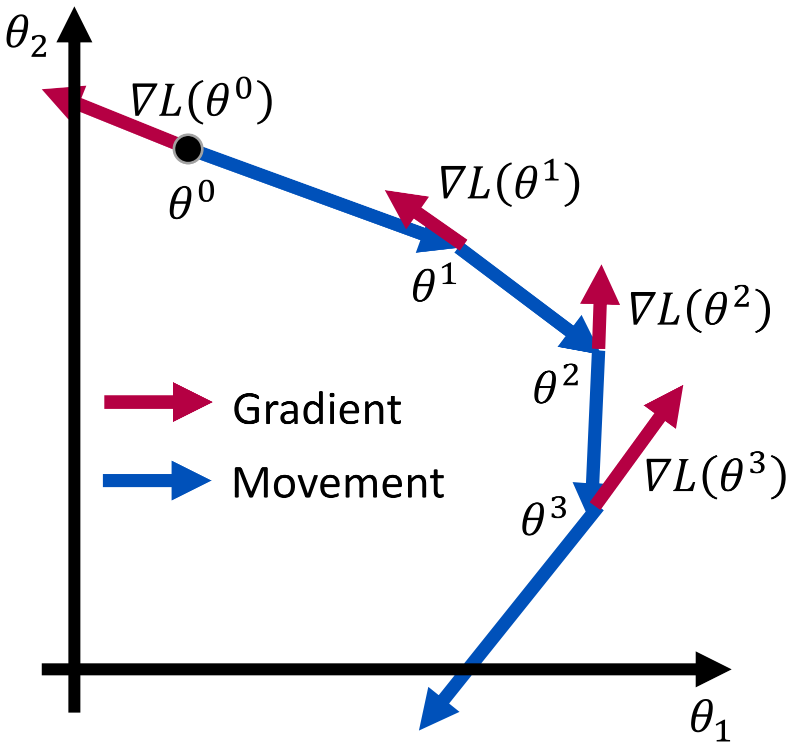
Randomly start at $\theta^0 = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix}$

$$\nabla L(\theta) = \begin{bmatrix} \partial L(\theta_1)/\partial\theta_1 \\ \partial L(\theta_2)/\partial\theta_2 \end{bmatrix}$$

$$\begin{bmatrix} \theta_1^1 \\ \theta_2^1 \end{bmatrix} = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix} - \eta \begin{bmatrix} \partial L(\theta_1^0)/\partial\theta_1 \\ \partial L(\theta_2^0)/\partial\theta_2 \end{bmatrix} \Rightarrow \theta^1 = \theta^0 - \eta\nabla L(\theta^0)$$

$$\begin{bmatrix} \theta_1^2 \\ \theta_2^2 \end{bmatrix} = \begin{bmatrix} \theta_1^1 \\ \theta_2^1 \end{bmatrix} - \eta \begin{bmatrix} \partial L(\theta_1^1)/\partial\theta_1 \\ \partial L(\theta_2^1)/\partial\theta_2 \end{bmatrix} \Rightarrow \theta^2 = \theta^1 - \eta\nabla L(\theta^1)$$

Review: Gradient Descent, Continued



1. Start at position θ^0
2. Compute gradient at θ^0
3. Move to $\theta^1 = \theta^0 - \eta\nabla L(\theta^0)$
4. Compute gradient at θ^1
5. Move to $\theta^2 = \theta^1 - \eta\nabla L(\theta^1)$

⋮