Gradient Descent Algorithm

Dr. Muhammad Wasim

Intuition behind Gradient Descent Algorithm

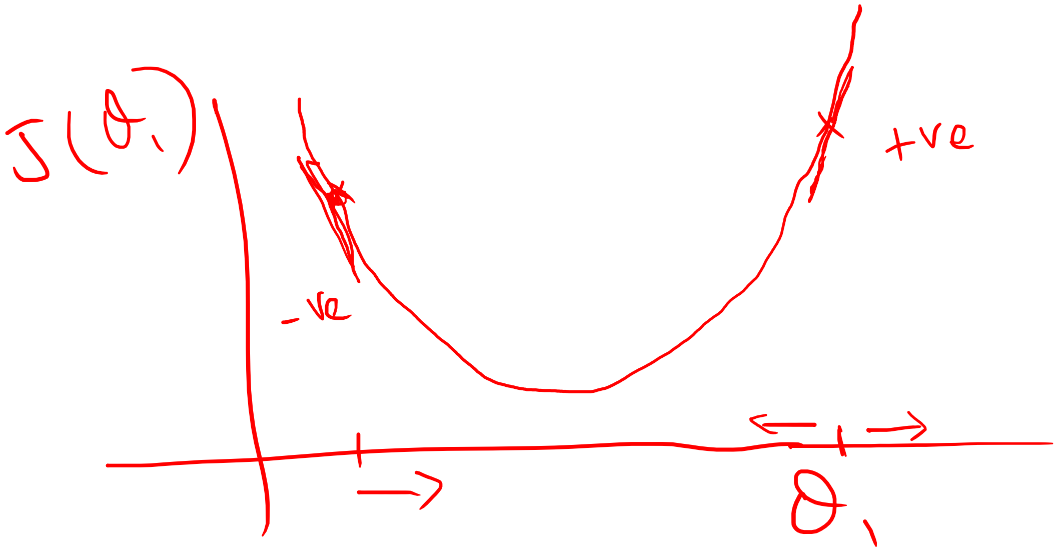
$$\min_{\theta_1} \boxed{J(\theta_1)}$$

$$\theta_1 \in \mathbb{R}$$

−α × −ve = +ve

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

−No ? +ve

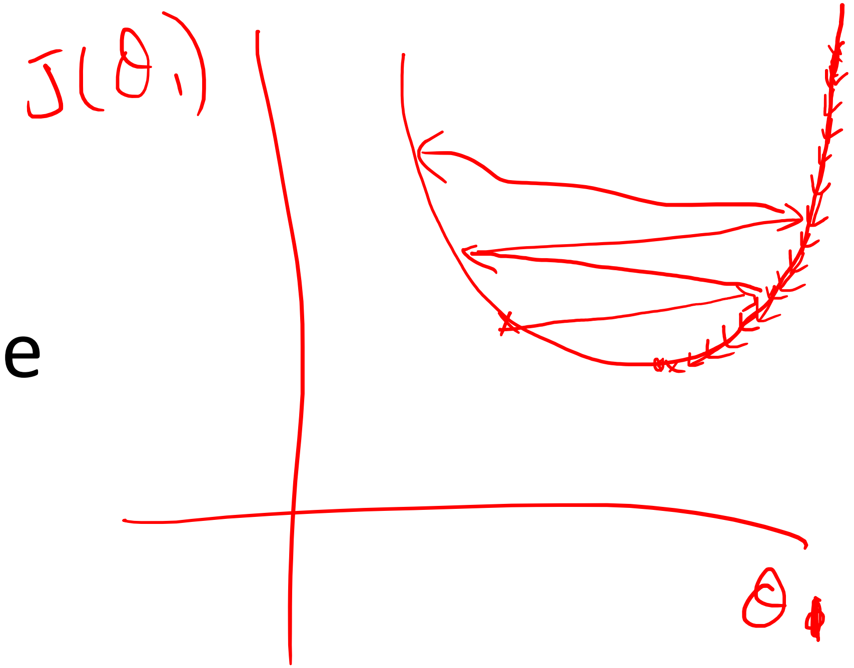


The role of Alpha – The Learning Rate

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

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



Computing the Derivative

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

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

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

Gradient Descent for Linear Regression with one Feature

Repeat until convergence {

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

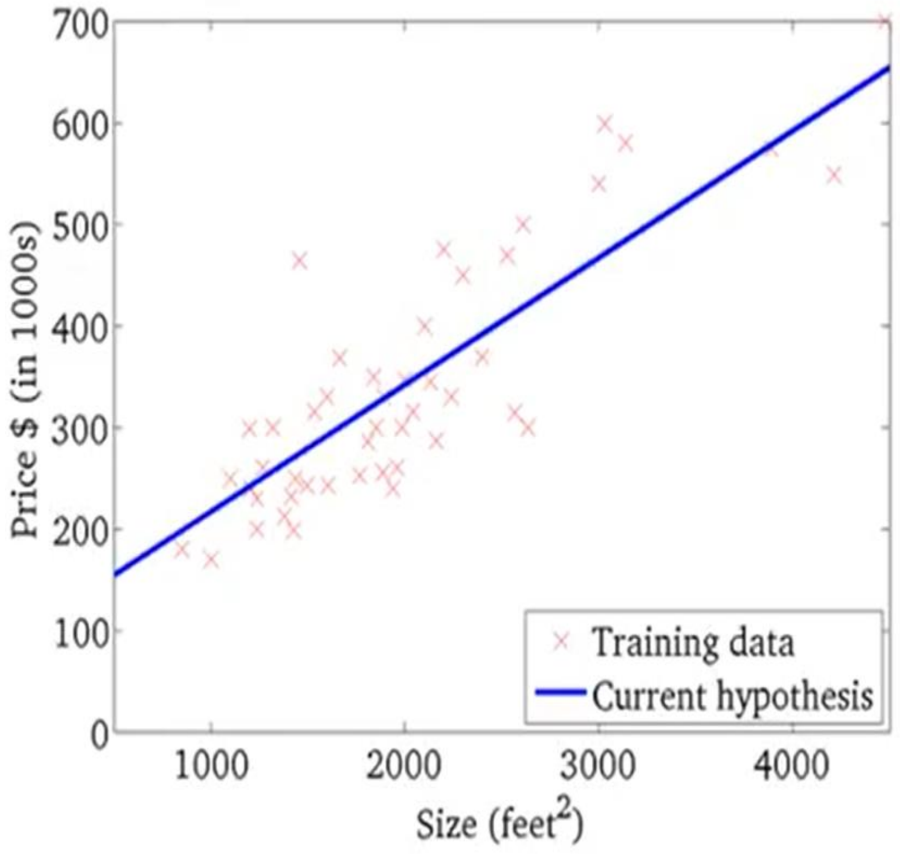
$$\theta_1 = \theta_1 - \alpha \boxed{\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) . x^{(i)}}$$

}

How gradient descent for regression works

$h_\theta(x)$

(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)