Week 1.3 Parameter Learning

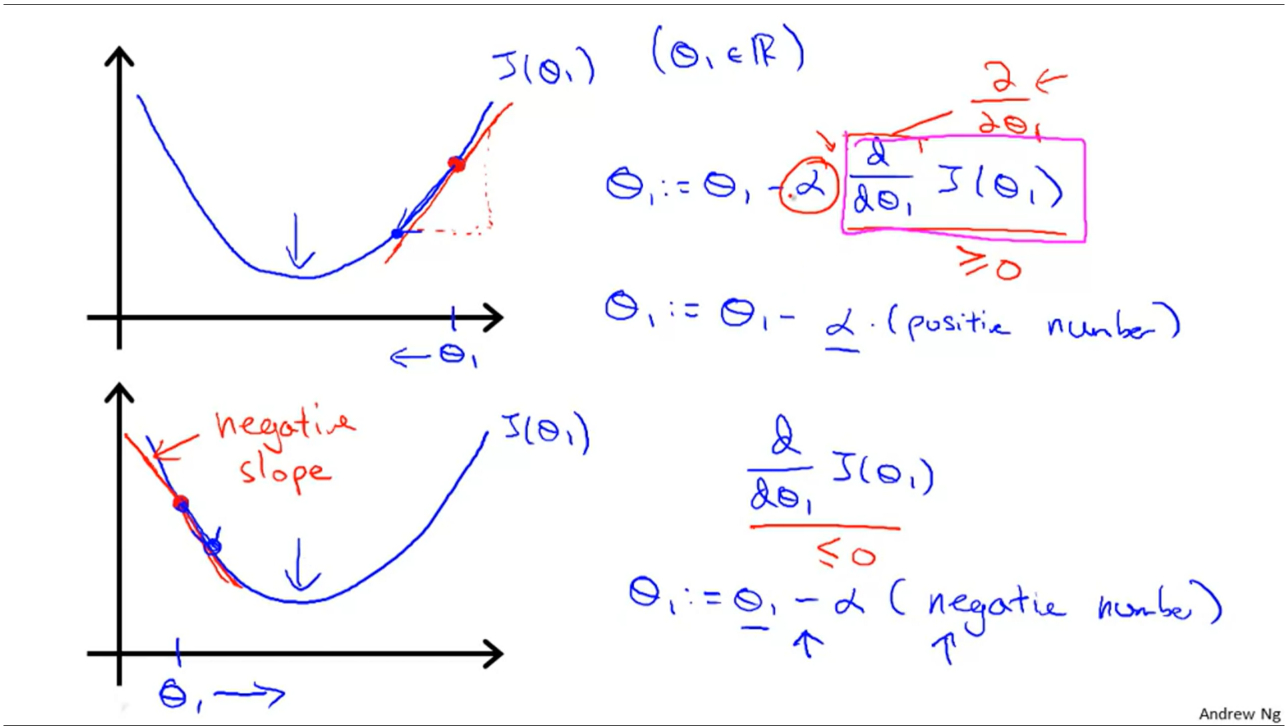
Gradient Descent

1. Start with some θ_0, θ_1
2. Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we hopefully end up at a minimum

- repeat until convergence {

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

  } for ($j = 0$ and $j = 1$)
- α: learning rate
- Simultaneously: update θ_0 and θ_1
- Intuition:



Intuition

- too small

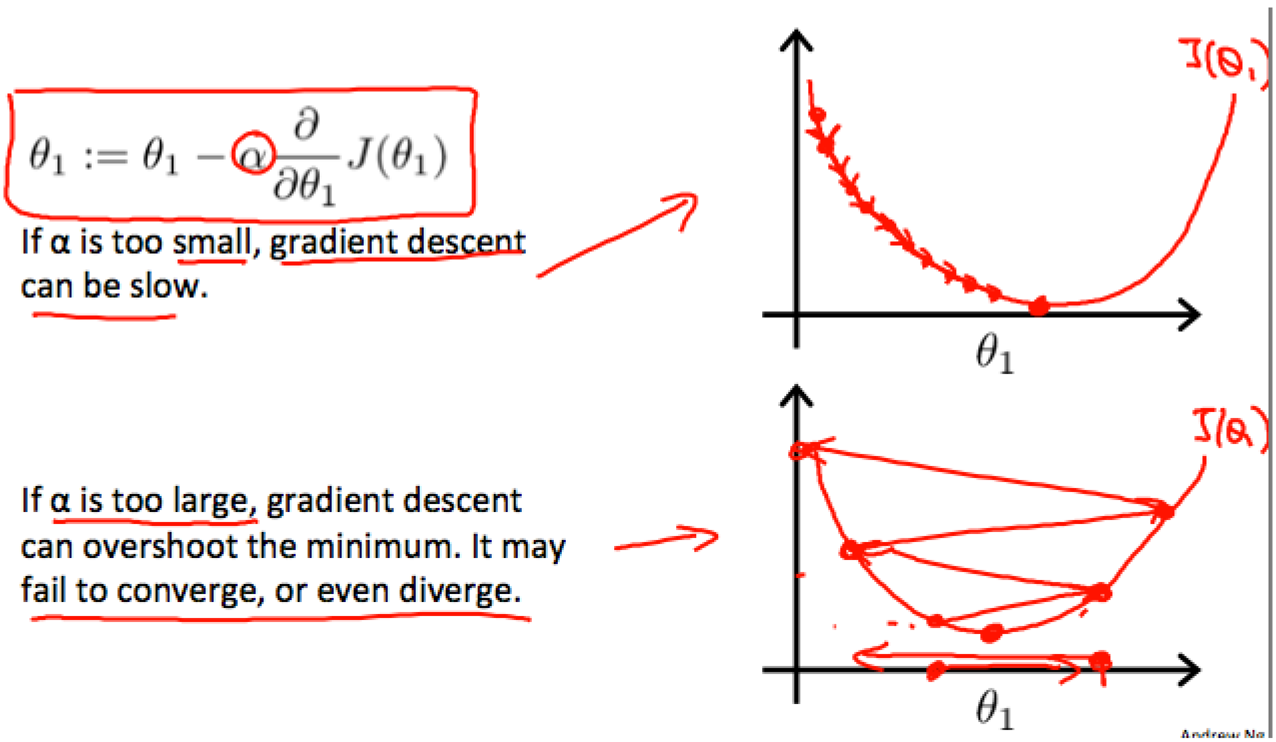
> if α is too small, gradient descent can be slow

- too large:

> if α is too large, gradient descent can overshoot minimum. It may fail to converge, or even diverge.

- Gradient descent can converge to a local minimum, even with the learning rate α fixed

> As we approach a local minimum, gradient descent will automatically take smaller steps. So, no need to decrease α over time.



Gradient Descent for Linear Regression

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

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

- "convex function": bowl-shaped, only one global optimum
- "Batch" Gradient Descent: each step of gradient descent uses all the training examples