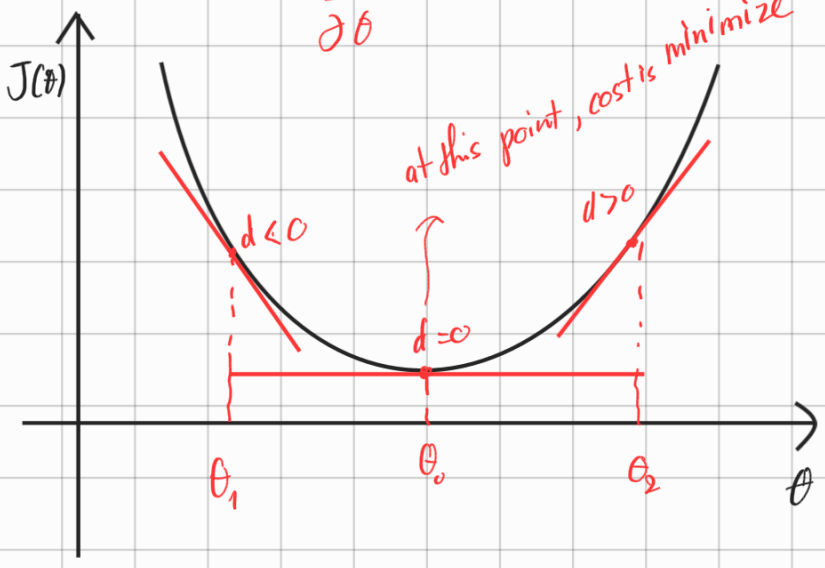
Gradient Descent

- The idea is If you change a little bit model parameter θ → your cost function will change

- And How you change θ to make your cost reduce.

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

let $d = \frac{\partial J}{\partial \theta}$



- $d < 0$: If θ belong to this part then we want it to increase to θ_0

- $d > 0$: If θ belong to this part then we want to It to decreare to θ_0

- Then we can illustrate this as: $\theta = \theta - \alpha . d$
  + α : learning rate
  + d : change and sign of θ

- First, we initialize value for θ
- Second, we choose appropriate α
  + If too large → θ can not converge
  + If too low → θ converge very slow

- Third, Calculate gradient descent (d)

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2$$

temporary forget this constant

$$\rightarrow \frac{\partial J}{\partial \theta_j} = \frac{2}{m} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)} \quad (j = 0..n)$$

For example with $j = 0$:

$$\frac{\partial J}{\partial \theta_0} = (\theta^T x^{(1)} - y^{(1)}) x_0^{(1)} + (\theta^T x^{(2)} - y^{(2)}) x_0^{(2)} + \cdots + (\theta^T x^{(m)} - y^{(m)}) x_0^{(m)}$$

Rewritten in matrix form:

$$\text{gradient} = \underset{n \times 1}{\begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_n} \end{bmatrix}} = \underset{n \times 1}{\begin{bmatrix} \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_0^{(i)} \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)}) x_n^{(i)} \end{bmatrix}} \quad (I)$$

we have

$$X = \underset{m \times n}{\begin{bmatrix} x_0^1 & x_1^1 & \cdots & x_n^1 \\ x_0^2 & x_1^2 & \cdots & x_n^2 \\ \vdots & & & \\ x_0^m & x_1^m & \cdots & x_n^m \end{bmatrix}}$$

$$X\theta - \hat{y} = \begin{bmatrix} \theta_0 x_0^{(1)} + \theta_1 x_1^{(1)} + \cdots + \theta_n x_n^{(1)} - y^{(1)} \\ \theta_0 x_0^{(2)} + \theta_1 x_1^{(2)} + \cdots + \theta_n x_n^{(2)} - y^{(2)} \end{bmatrix}$$

$$\begin{bmatrix} \\ \\ \\ \theta_0 x_0^{(m)} + \theta_1 x_1^{(m)} + \cdots + \theta_n x_n^{(m)} - y^{(m)} \end{bmatrix}$$

- then (I) can be written as:

$$\text{gradient} = X^T_{n\times m} \left(X_{m\times n}\, \theta_{n\times 1} - \hat{y}_{m\times 1} \right)$$

- Remember put $\frac{2}{m}$ again:

$$\text{gradient} = \frac{2}{m} X^T (X\theta - \hat{y})$$

- Finally, we have the formular:

$$\theta = \theta - \alpha \cdot \text{gradient}$$

(α: choosing)