Lecture 6 (Gradient Descent)

1 Finding $argmin J(\theta)$ - Gradient Descent

The idea of *gradient descent* is used, since it might not always be possible to find **zeroes** of the gradient of $J(\theta)$. The algorithm is something like:

1. θ is initialised to a random value, call it θ^0
2. Now, θ is updated as

$$\theta^{(t+1)} = \theta^t - \eta \cdot \left.\frac{\partial f(\theta)}{\partial \theta}\right|_{\theta^t}$$

In $n+1$ dimensional space, it looks like:

$$\theta^{(t+1)} = \theta^t - \eta \cdot \nabla_\theta f(\theta^t)|_{\theta^t}$$

(θ^k is a $n+1$ dimensional vector, ∇_θ is computed by taking partial derivate for each term individually)