Q(3)

$$X = \begin{bmatrix} \text{---} x^{(1)} \text{---} \\ \text{---} x^{(2)} \text{---} \\ \text{---} x^{(3)} \text{---} \\ \vdots \\ \text{---} x^{(m)} \text{---} \end{bmatrix}_{m \times (n+1)}, \quad Y = \begin{bmatrix} Y^{(1)} \\ \vdots \\ Y^{(m)} \end{bmatrix}$$

$(n+1)$ indicates we have an extra feature containing only '1's.

~~this removes the~~

This ensures that $\hat{Y} = X\theta$ works (instead of $Y = W^T X + b$)

Parameters $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$

In BGD

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}, \text{ where } J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(x^{(i)}\theta - Y^{(i)}\right)^2$$

Date
Page

Repeat

$$\Rightarrow \Theta_j \leftarrow \Theta_j - \alpha \sum_{i=1}^{m} (X^{(i)}\Theta - Y^{(i)}) X_j^{(i)}, \; j = 1, 2, \ldots, n+1$$

until $J(\Theta)$ stops ~~improving~~ reducing.

SGD

Repeat

$$\Theta_j \leftarrow \Theta_j - \alpha (X^{(i)}\Theta - Y^{(i)}), \; j = 1, 2, \ldots, n+1$$

until $\mathcal{L}(\Theta) = (X^{(i)}\Theta - Y^{(i)})^2$ stops reducing

Repeat this for each training example i.e $i = 1, 2, \ldots, m$.

For complex (large) datasets, SGD is better as it provides ~~better~~ faster convergence.

~~Math~~ ~~A mathematical proof is~~

~~Intuit~~

Intuition

In BGD, we have to calculate $\sum_{i=1} X^{(i)}\Theta - Y^{(i)}$ by iterating over all training examples before updating Θ_j.

Whereas in SGD, an ~~update~~ optimization is made for each training example. ~~This makes sure It is highly li~~

It is highly likely that optimizing Θ for each training example is "close" to the dirn when we optimize entire cost fn.