Gradient Descent for Multiple Variables

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$

Our formal hypothesis in multivariable linear regression where we've adopted the convention that $x_0 = 1$. The parameters of this model are θ_0 through θ_n, but instead of thinking of this as n separate parameters, which is valid, we are going to think of the parameters as θ where theta is a n+1-dimensional vector.

Parameters: $\theta_0, \theta_1, \ldots, \theta_n = \theta$ (n + 1 dimensional vector)

$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Our cost function is J of θ_0 through θ_n which is given by this usual sum of square of error term. But again instead of thinking of J as a function of these n + 1 numbers, we're going to more commonly use J as just a function of the parameter vector θ.

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

Video Question: When there are n features, we define the cost function as

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

For linear regression, which of the following are also equivalent and correct definitions of $J(\theta)$?

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

> $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} ((\sum_{j=0}^{n} \theta_j x_j^{(i)}) - y^{(i)})^2$ (Inner sum starts at 0)

- $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} ((\sum_{j=1}^{n} \theta_j x_j^{(i)}) - y^{(i)})^2$ (Inner sum starts at 1)
- $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} ((\sum_{j=0}^{n} \theta_j x_j^{(i)}) - (\sum_{j=0}^{n} y_j^{(i)}))^2$

The gradient descent equation itself is generally the same form; we just have to repeat it for our 'n' features:

$$
\begin{aligned}
&\text{repeat until convergence: } \{ \\
&\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)} \\
&\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)} \\
&\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)} \\
&\cdots \\
&\}
\end{aligned}
$$

In other words:

$$
\begin{aligned}
&\text{repeat until convergence: } \{ \\
&\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)} \qquad \text{for j := 0...n} \\
&\}
\end{aligned}
$$

The following image compares gradient descent with one variable to gradient descent with multiple variables: