Lecture 2

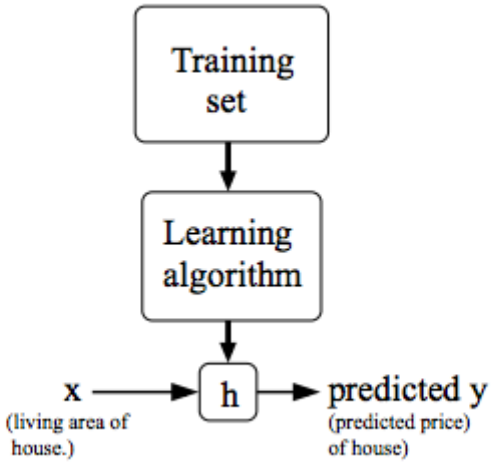
Notation

- $x^{(i)}$: input (features) variables
- $y^{(i)}$: output (target) variable, i.e., what we are trying to predict
- a pair $(x^{(i)}, y^{(i)})$ is called a training example
- a list of m training examples $(x^{(i)}, y^{(i)}); i = 1, \dots, m$ is called a training set, i.e., the dataset we'll be using to learn
- X: space of input values
- Y: space of output values

Supervised Learning (formally)

- **goal:** given a training set, learn a function $h : X \to Y$ so that $h(x)$ is a "good predictor" of y.

> For historical reasons, the function h is called a hypothesis.



- **regression problem:** when the target variable is continuous.
- **classification problem:** when y can take on only a small number of discrete values.

Cost function

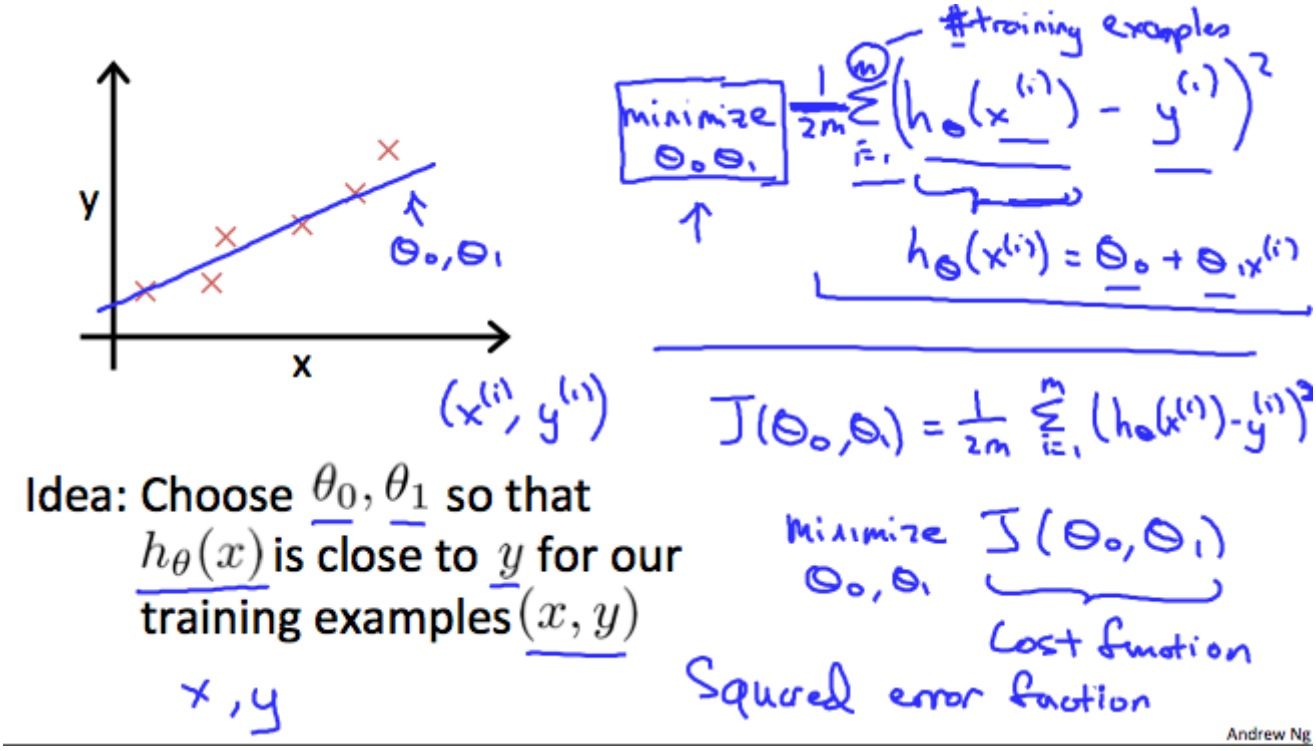
Measures de accuracy of the hypothesis function. In the case of Linear Regression, the cost function can be defined as

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 \tag{1}$$

It is $\frac{1}{2}\bar{x}$, where $\bar{x}$ is the mean of the squares of $h_\theta(x_i) - y_i$.

This function is called the **"Squared Error Function"**, or **"Mean Squared Error"**.

Note: the mean is halved $\left(\frac{1}{2}\right)$ as a convenience for the computation of the gradient descent, as the derivative term of the square function will cancel out the $\frac{1}{2}$ term.

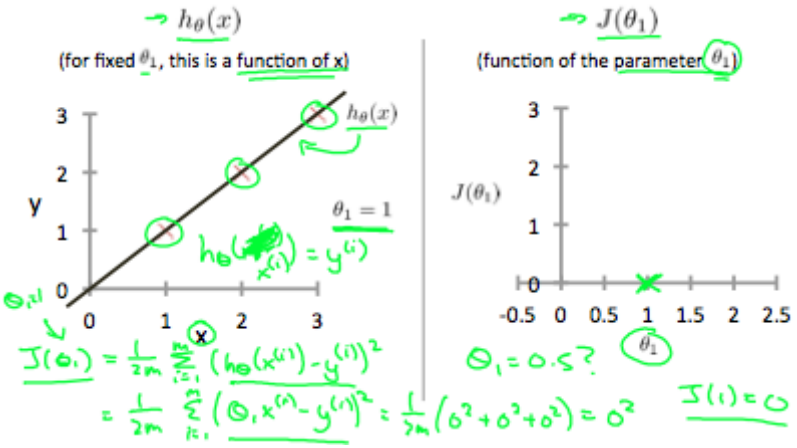


Cost Funcion - Intuition I

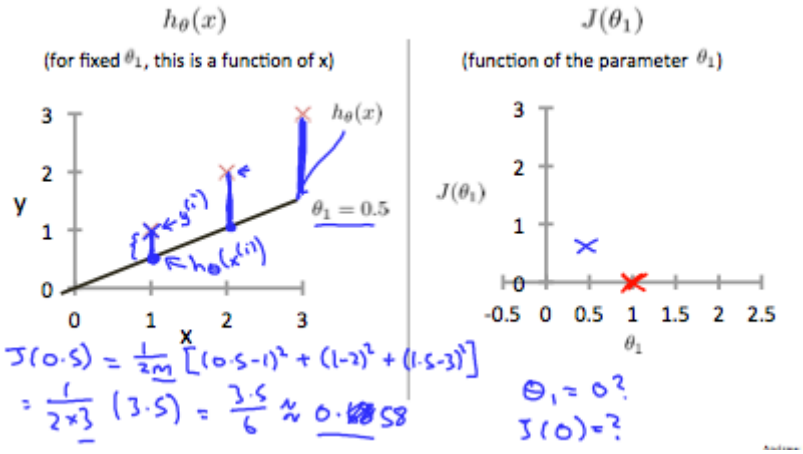
Note: in this example, suppose $\theta_0 = 0$, i.e., $h_\theta(x_i) = \theta_1 x_i$

Objective: get the best possible line which passes through the scattered points.

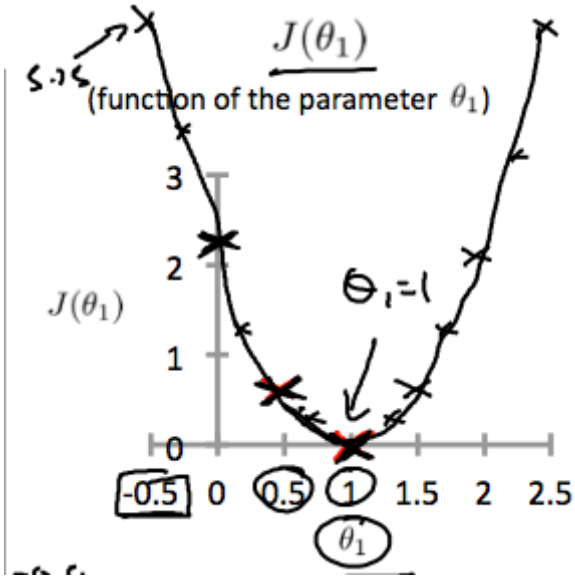
The best possible line will be such so that the average squared vertical distances of the scattered points from the line will be the least. Ideally, the line should pass through all the points of our training data set. In such a case, the value of $J(\theta_0, \theta_1) = 0$. The next image illustrates such situation.



When $\theta_1 = 1$, we get a slope of 1 which goes through every single data point in our model. Conversely, when $\theta_1 = 0.5$, we see the vertical distance from our fit to the data points increase.



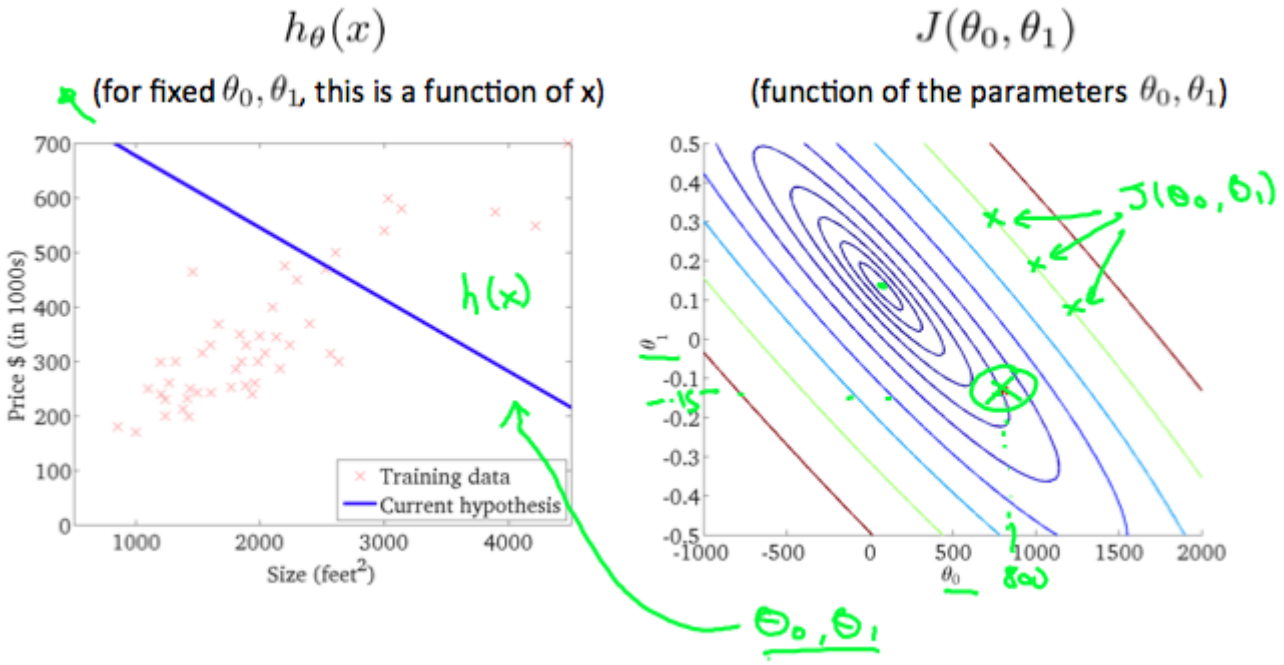
So, in this example, if we keep plotting several points, our cost function should look like this:



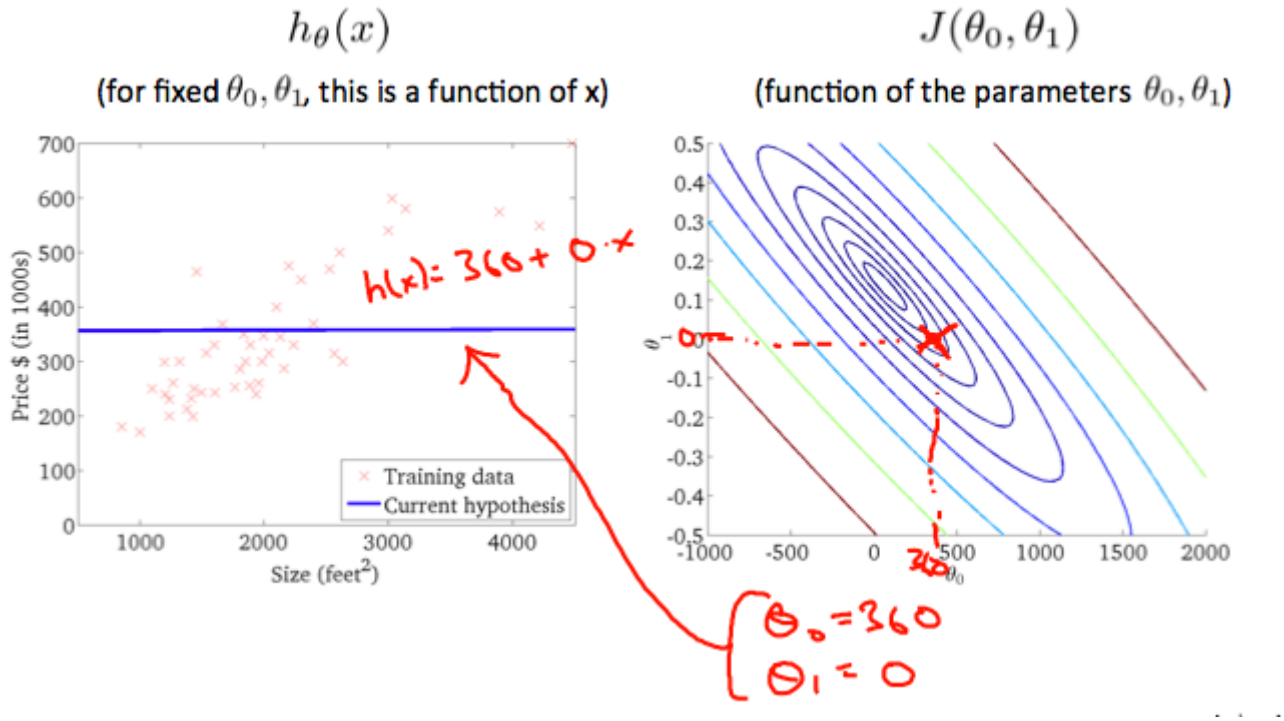
Then, our goal is to minimize the cost function. As we can see, $\theta_1 = 1$ is our global minimum.

Cost Function - Intuition II

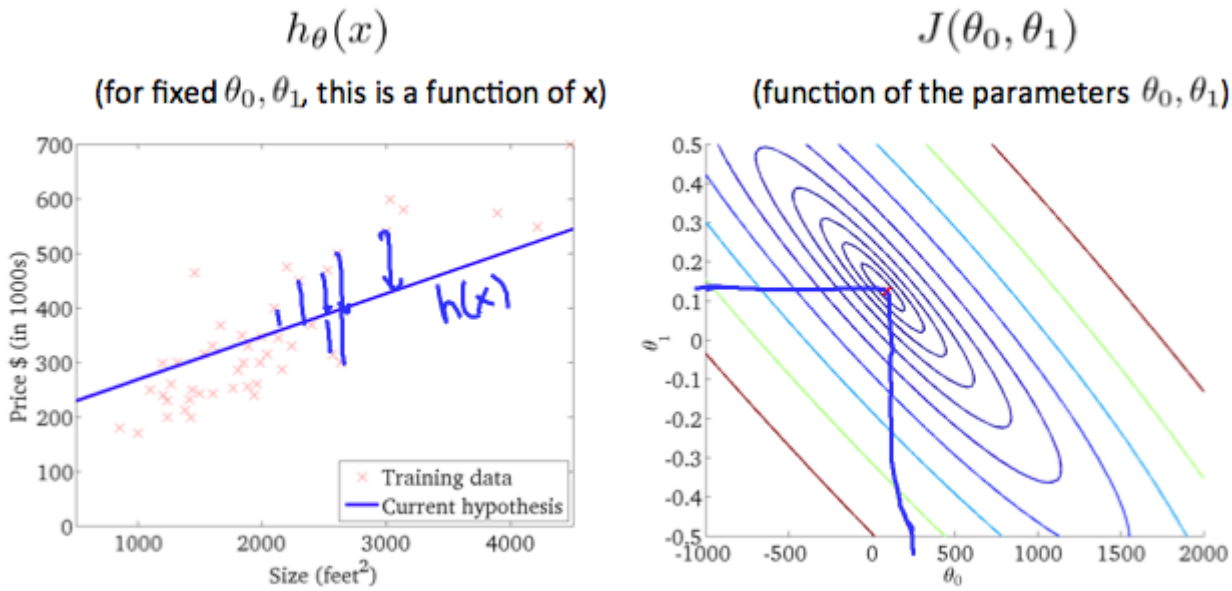
Contour plot is a graph that contains many contour lines. A contour line of a two variable function has a constant value at all points of the same line. An example of such a graph is the one to the right below.



The circled x displays the value of the cost function for the graph on the left when $\theta_0 = 800$ and $\theta_1 = -0.15$. Taking another $h(x)$ and plotting its contour plot, one gets the following graphs:



When $\theta_0 = 360$ and $\theta_1 = 0$, the value of $J(\theta_0, \theta_1)$ in the contour plot gets closer to the center thus reducing the cost function error. Now giving our hypothesis function a slightly positive slope results in a better fit of the data.



The graph above minimizes the cost function as much as possible and consequently, the result of θ_1 and θ_0 tend to be around 0.12 and 250, respectively. Plotting those values on our graph to the right seems to put our point in the center of the inner most 'circle'.

References

[1] Machine Learning - Stanford University (https://www.coursera.org/learn/machine-learning)