

# Machine Learning

by Stanford University

Week1

Jose Vicente Yago Martínez

April 5, 2019

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	What is machine learning? . . . . .	3
1.2	Supervised learning . . . . .	3
1.3	Unsupervised learning . . . . .	4
<b>2</b>	<b>Model and cost function</b>	<b>4</b>
2.1	Model representation . . . . .	4
2.2	Cost function . . . . .	5
2.2.1	Intuition I . . . . .	5
2.2.2	Intuition II . . . . .	7
<b>3</b>	<b>Parameter learning</b>	<b>10</b>
3.1	Gradient descent intuition . . . . .	10
3.2	Gradient descent for linear regression . . . . .	13

# 1 Introduction

## 1.1 What is machine learning?

Two definitions of Machine Learning are offered. Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.

Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ."

Example: playing checkers.

$E$  = the experience of playing many games of checkers

$T$  = the task of playing checkers.

$P$  = the probability that the program will win the next game.

In general, any machine learning problem can be assigned to one of two broad classifications:

Supervised learning and Unsupervised learning.

## 1.2 Supervised learning

In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.

Supervised learning problems are categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

### Example 1:

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete categories.

### Example 2:

(a) Regression - Given a picture of a person, we have to predict their age on the basis of the given picture

(b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

## 1.3 Unsupervised learning

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

We can derive this structure by clustering the data based on relationships among the variables in the data.

With unsupervised learning there is no feedback based on the prediction results.

### Example:

Clustering: Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.

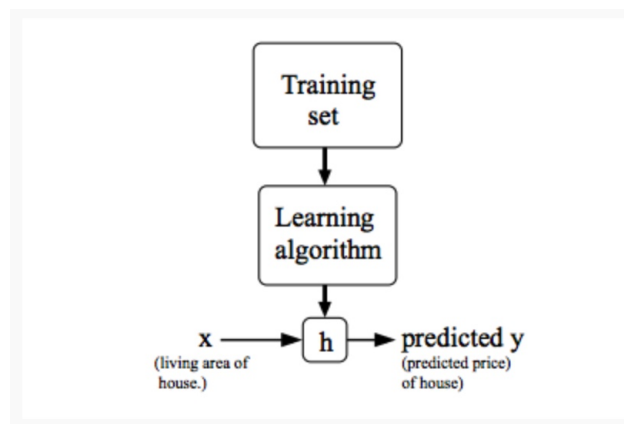
Non-clustering: The "Cocktail Party Algorithm", allows you to find structure in a chaotic environment.

## 2 Model and cost function

### 2.1 Model representation

To establish notation for future use, we'll use  $x^{(i)}$  to denote the "input" variables (living area in this example), also called input features, and  $y^{(i)}$  to denote the "output" or target variable that we are trying to predict (price). A pair  $(x^{(i)}, y^{(i)})$  is called a training example, and the dataset that we'll be using to learn—a list of  $m$  training examples  $(x^{(i)}, y^{(i)}); i = 1, \dots, m$ —is called a training set. Note that the superscript "(i)" in the notation is simply an index into the training set, and has nothing to do with exponentiation. We will also use  $X$  to denote the space of input values, and  $Y$  to denote the space of output values. In this example,  $X = Y = \mathbb{R}$ .

To describe the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function  $h : X \rightarrow Y$  so that  $h(x)$  is a "good" predictor for the corresponding value of  $y$ . For historical reasons, this function  $h$  is called a hypothesis. Seen pictorially, the process is therefore like this:



When the target variable that we're trying to predict is continuous, such as in our housing example, we call the learning problem a regression problem. When  $y$  can take on only a

small number of discrete values (such as if, given the living area, we wanted to predict if a dwelling is a house or an apartment, say), we call it a classification problem.

## 2.2 Cost function

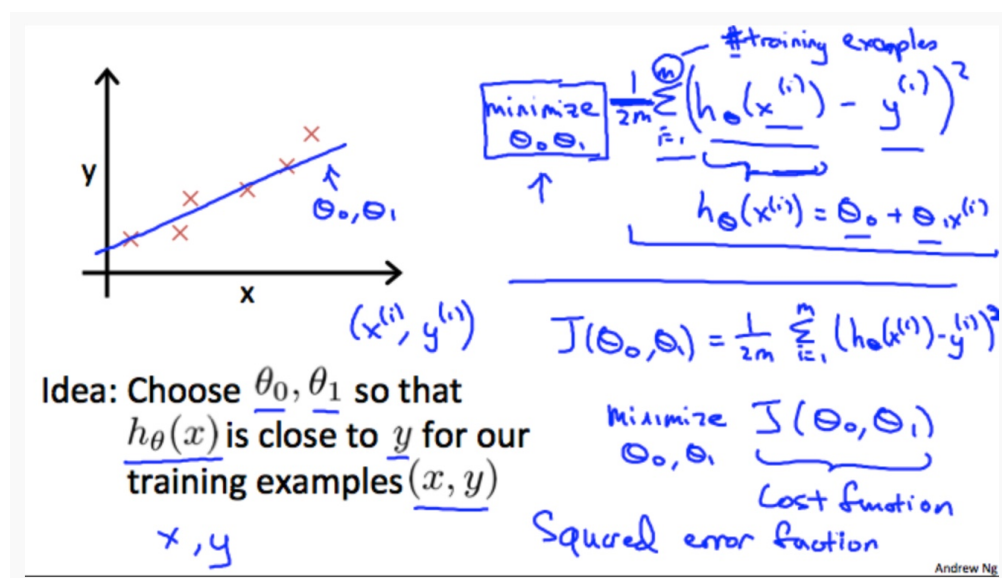
### 2.2.1 Intuition I

We can measure the accuracy of our hypothesis function by using a **cost function**. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from  $x$ 's and the actual output  $y$ 's.

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

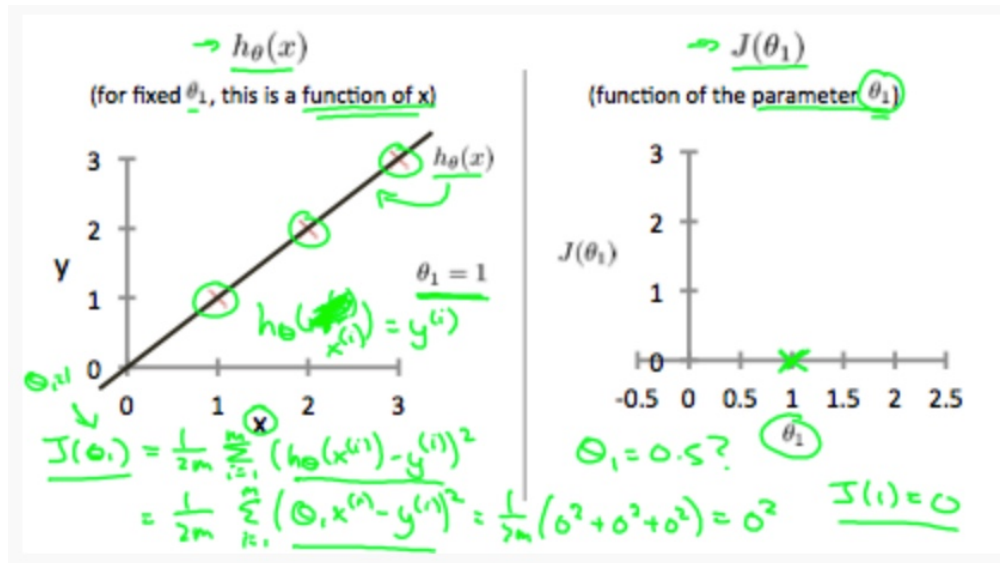
To break it apart, it is  $\frac{1}{2} \bar{x}$  where  $\bar{x}$  is the mean of the squares of  $h_{\theta}(x_i) - y_i$ , or the difference between the predicted value and the actual value.

This function is otherwise called the "Squared error function", or "Mean squared error". The mean is halved ( $\frac{1}{2}$ ) 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. The following image summarizes what the cost function does:

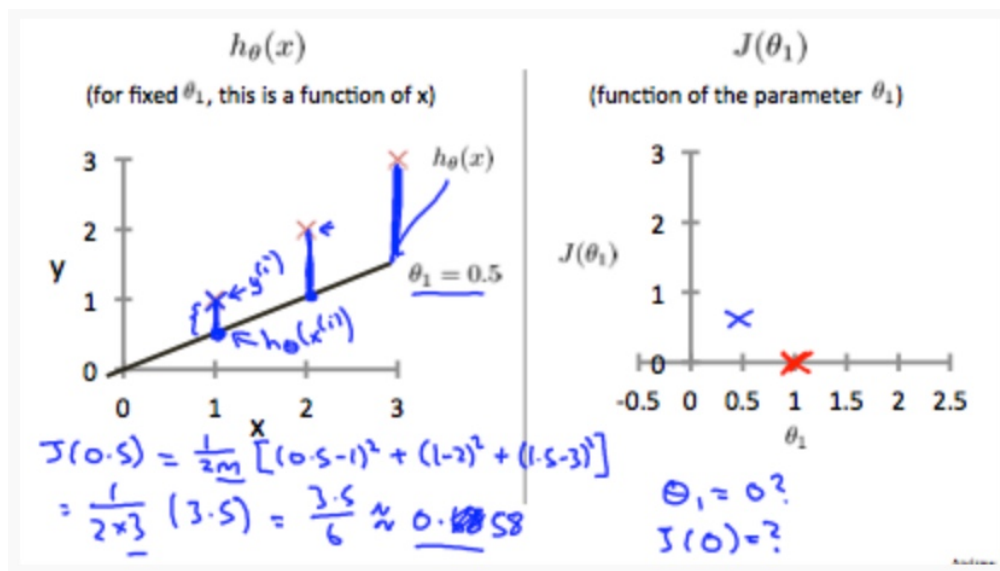


If we try to think of it in visual terms, our training data set is scattered on the  $x$ - $y$  plane. We are trying to make a straight line (defined by  $h_{\theta}(x)$ ) which passes through these scattered data points.

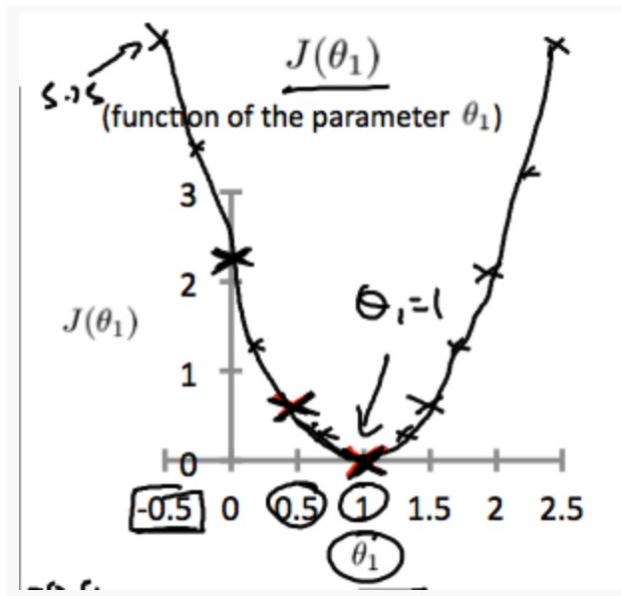
Our objective is to get the best possible line. The best possible line will be such so that the average squared vertical distances of the scattered points from the line will be the least. Ideally, the line should pass through all the points of our training data set. In such a case, the value of  $J(\theta_0, \theta_1)$  will be 0. The following example shows the ideal situation where we have a cost function of 0.



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.



This increases our cost function to 0.58. Plotting several other points yields to the following graph:



Thus as a goal, we should try to minimize the cost function. In this case,  $\theta_1 = 1$  is our global minimum.

### 2.2.2 Intuition II

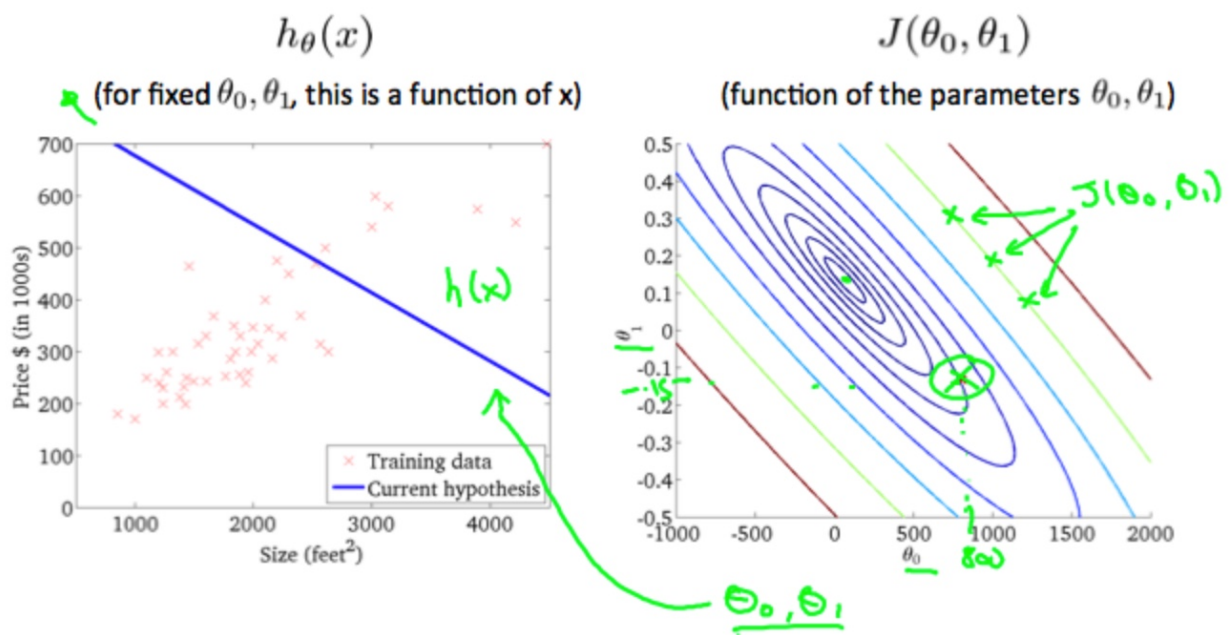
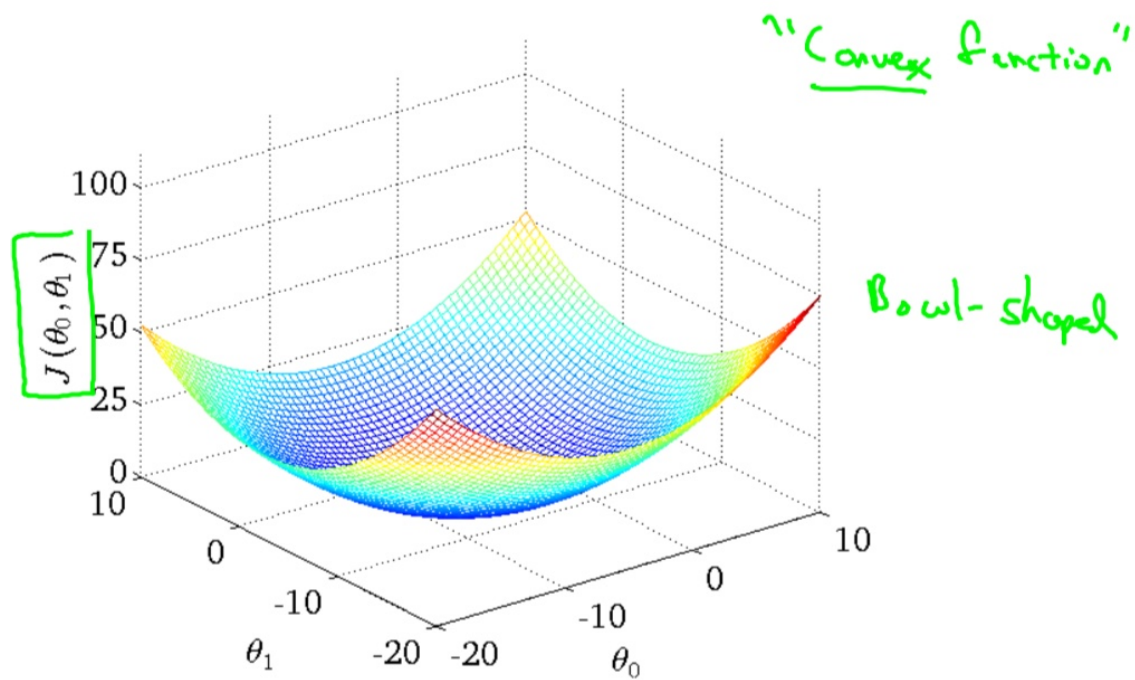
**Hypothesis:**  $h_{\theta}(x) = \theta_0 + \theta_1 x$

**Parameters:**  $\theta_0, \theta_1$

**Cost Function:**  $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

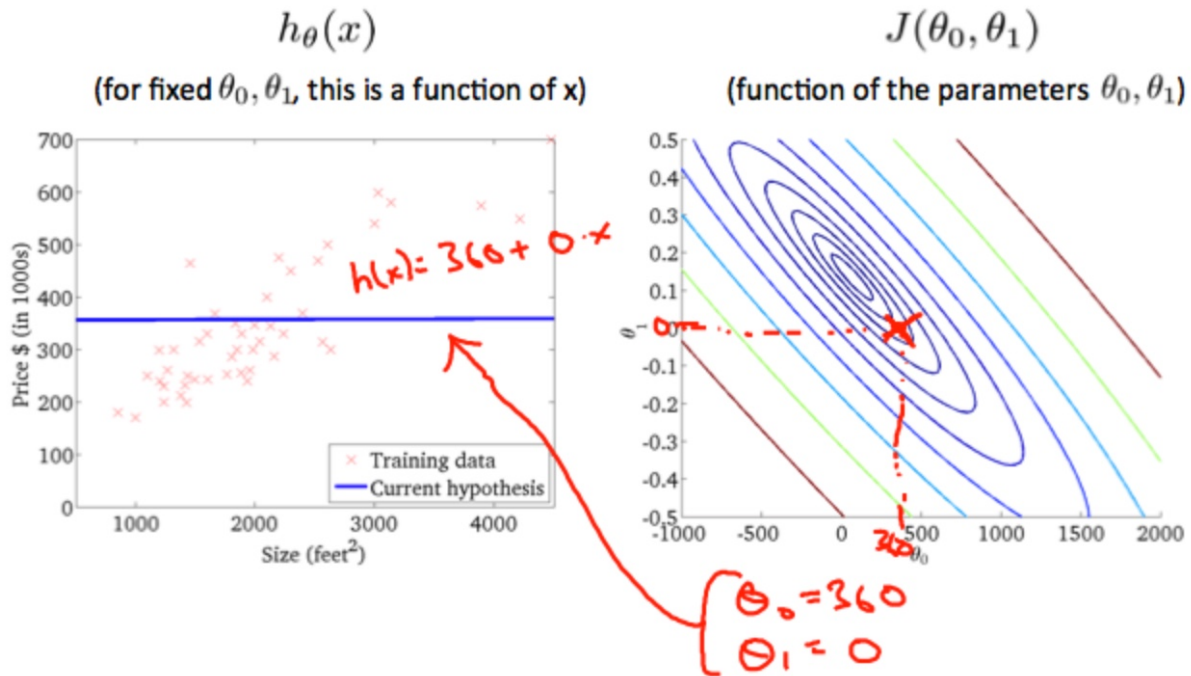
**Goal:**  $\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$

A 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.

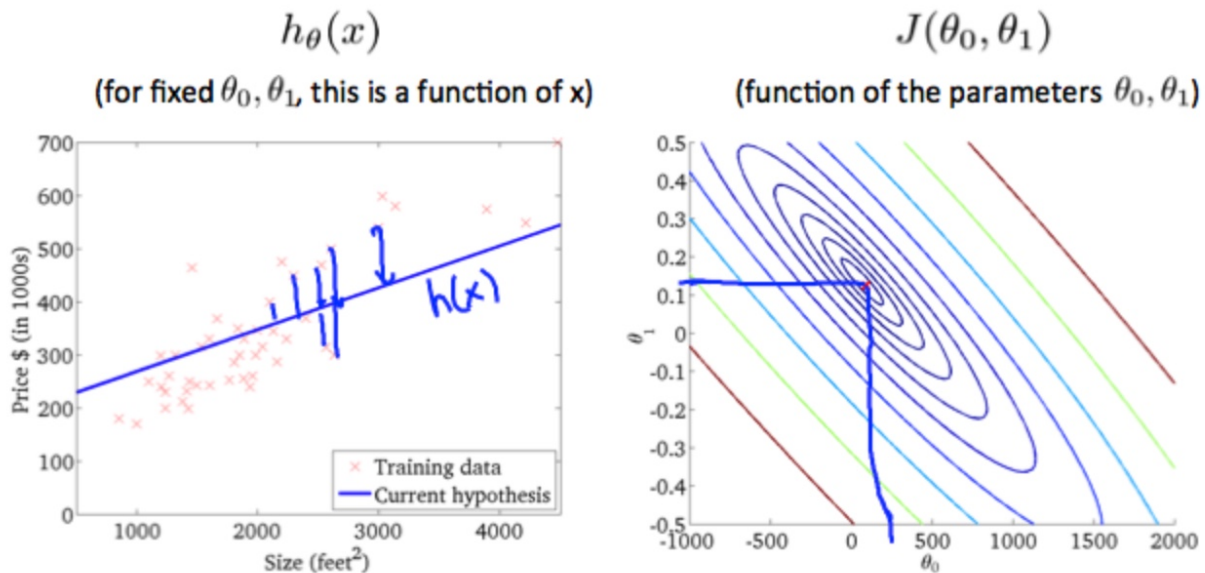


Taking any color and going along the 'circle', one would expect to get the same value of the cost function. For example, the three green points found on the green line above have the same value for  $J(\theta_0, \theta_1)$  and as a result, they are found along the same line. The circled  $x$  displays 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  $\theta_1$  and  $\theta_0$  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'.

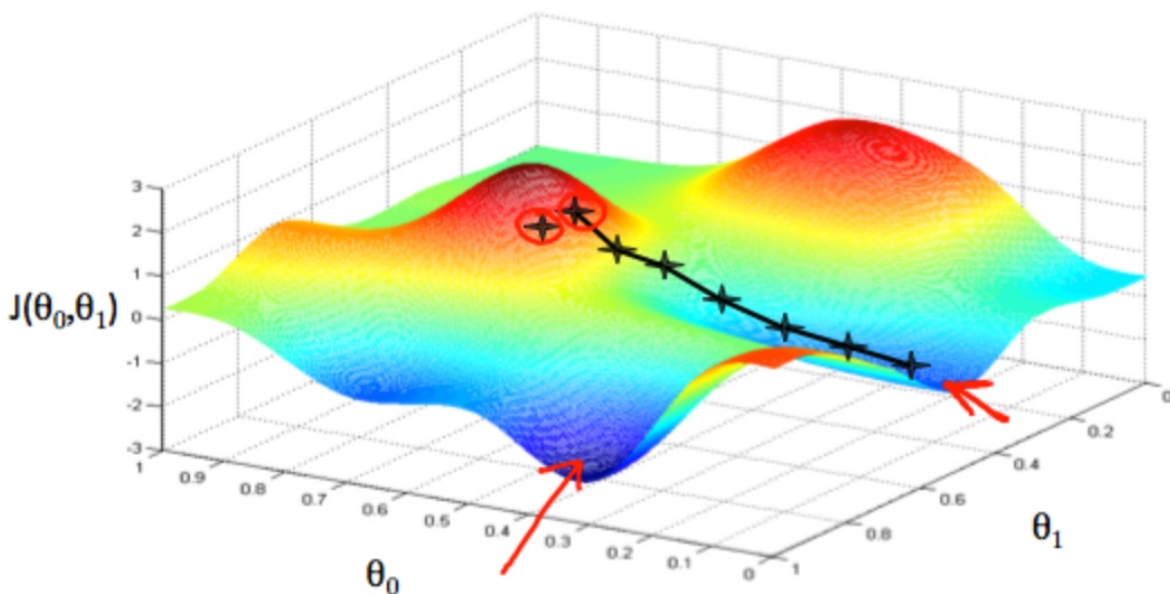
## 3 Parameter learning

### 3.1 Gradient descent intuition

So we have our hypothesis function and we have a way of measuring how well it fits into the data. Now we need to estimate the parameters in the hypothesis function. That's where gradient descent comes in.

Imagine that we graph our hypothesis function based on its fields  $\theta_0$  and  $\theta_1$  (actually we are graphing the cost function as a function of the parameter estimates). We are not graphing  $x$  and  $y$  itself, but the parameter range of our hypothesis function and the cost resulting from selecting a particular set of parameters.

We put  $\theta_0$  on the  $x$  axis and  $\theta_1$  on the  $y$  axis, with the cost function on the vertical  $z$  axis. The points on our graph will be the result of the cost function using our hypothesis with those specific theta parameters. The graph below depicts such a setup.



We will know that we have succeeded when our cost function is at the very bottom of the pits in our graph, i.e. when its value is the minimum. The red arrows show the minimum points in the graph.

The way we do this is by taking the derivative (the tangential line to a function) of our cost function. The slope of the tangent is the derivative at that point and it will give us a direction to move towards. We make steps down the cost function in the direction with the steepest descent. The size of each step is determined by the parameter  $\alpha$ , which is called the learning rate.

For example, the distance between each 'star' in the graph above represents a step determined by our parameter  $\alpha$ . A smaller  $\alpha$  would result in a smaller step and a larger  $\alpha$  results in a larger step. The direction in which the step is taken is determined by the partial derivative of  $J(\theta_0, \theta_1)$ . Depending on where one starts on the graph, one could end up at different points. The image above shows us two different starting points that end up in two different places.

The gradient descent algorithm is:

repeat until convergence:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

where

$j=0,1$  represents the feature index number.

At each iteration  $j$ , one should simultaneously update the parameters  $\theta_1, \theta_2, \dots, \theta_n$ . Updating a specific parameter prior to calculating another one on the  $j^{(th)}$  iteration would yield to a wrong implementation.

### Correct: Simultaneous update

→  $\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$   
→  $\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$   
→  $\theta_0 := \text{temp0}$   
→  $\theta_1 := \text{temp1}$

### Incorrect:

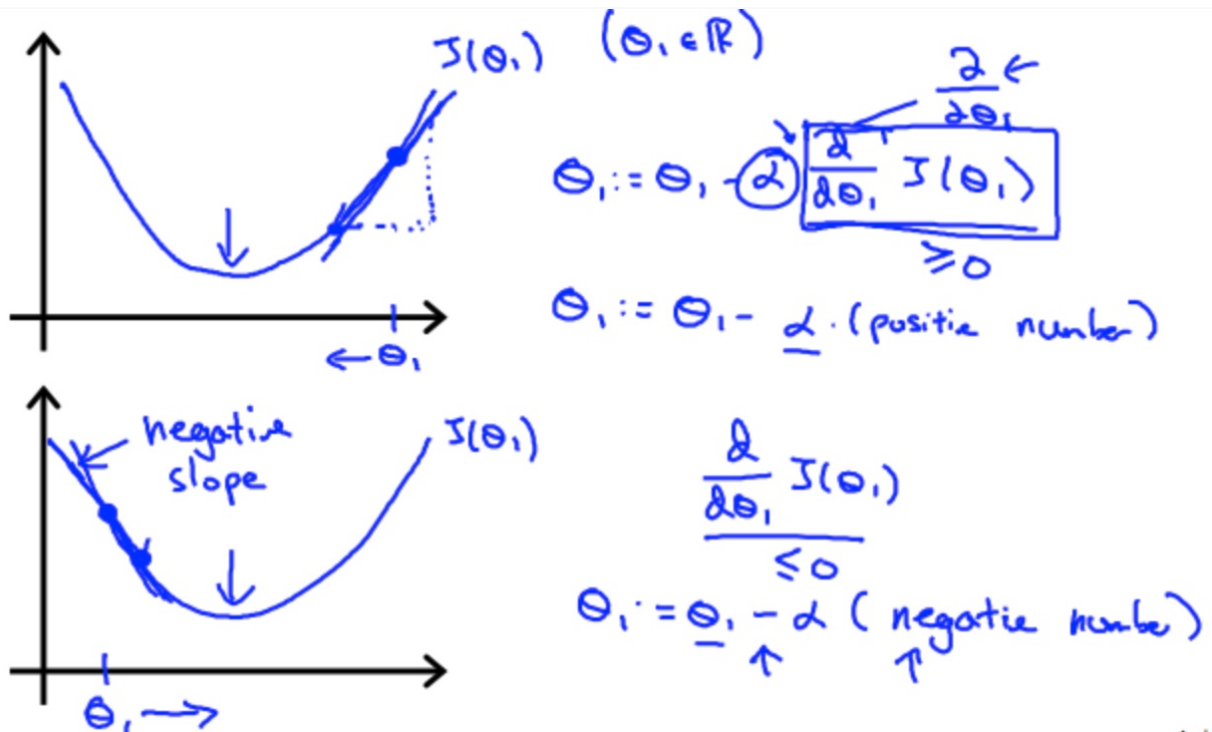
→  $\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$   
→  $\theta_0 := \text{temp0}$   
→  $\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$   
→  $\theta_1 := \text{temp1}$

In this video we explored the scenario where we used one parameter  $\theta_1$  and plotted its cost function to implement a gradient descent. Our formula for a single parameter was :

Repeat until convergence:

$$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$$

Regardless of the slope's sign for  $\frac{d}{d\theta_1} J(\theta_1)$ ,  $\theta_1$  eventually converges to its minimum value. The following graph shows that when the slope is negative, the value of  $\theta_1$  increases and when it is positive, the value of  $\theta_1$  decreases.

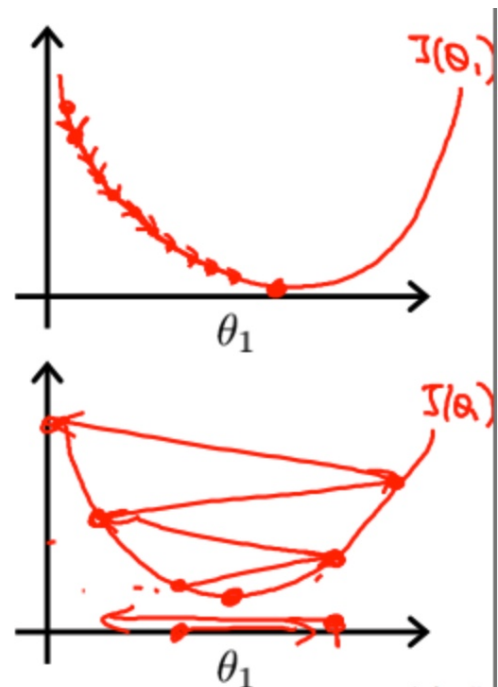


On a side note, we should adjust our parameter  $\alpha$  to ensure that the gradient descent algorithm converges in a reasonable time. Failure to converge or too much time to obtain the minimum value imply that our step size is wrong.

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If  $\alpha$  is too small, gradient descent can be slow.

If  $\alpha$  is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.





How does gradient descent converge with a fixed step size  $\alpha$ ?

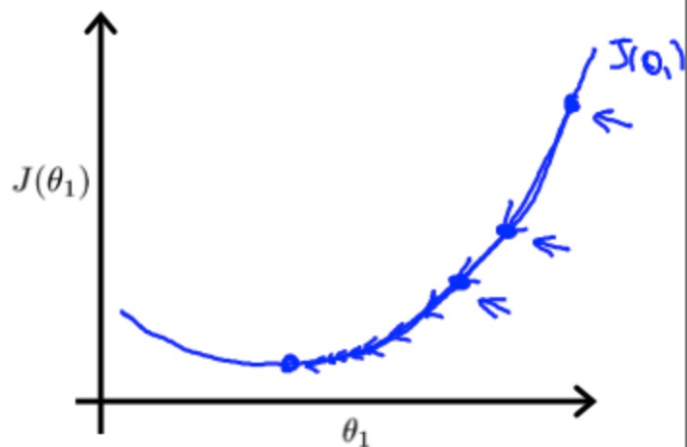
The intuition behind the convergence is that  $\frac{d}{d\theta_1} J(\theta_1)$  approaches 0 as we approach the bottom of our convex function. At the minimum, the derivative will always be 0 and thus we get:

$$\theta_1 := \theta_1 - \alpha * 0$$

**Gradient descent can converge to a local minimum, even with the learning rate  $\alpha$  fixed.**

$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$

**As we approach a local minimum, gradient descent will automatically take smaller steps. So, no need to decrease  $\alpha$  over time.**



### 3.2 Gradient descent for linear regression

When specifically applied to the case of linear regression, a new form of the gradient descent equation can be derived. We can substitute our actual cost function and our actual hypothesis function and modify the equation to :

repeat until convergence: {

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^m ((h_{\theta}(x_i) - y_i)x_i)$$

}

where  $m$  is the size of the training set,  $\theta_0$  a constant that will be changing simultaneously with  $\theta_1$  and  $x_i, y_i$  are values of the given training set (data).

Note that we have separated out the two cases for  $\theta_j$  into separate equations for  $\theta_0$  and  $\theta_1$ ; and that for  $\theta_1$  we are multiplying  $x_i$  at the end due to the derivative. The following is a derivation of  $\frac{\partial}{\partial \theta_j} J(\theta)$  for a single example :

where  $m$  is the size of the training set,  $\theta_0$  a constant that will be changing simultaneously with  $\theta_1$  and  $x_i, y_i$  are values of the given training set (data).

Note that we have separated out the two cases for  $\theta_j$  into separate equations for  $\theta_0$  and  $\theta_1$ ; and that for  $\theta_1$  we are multiplying  $x_i$  at the end due to the derivative. The following is a derivation of  $\frac{\partial}{\partial \theta_j} J(\theta)$  for a single example :

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_{\theta}(x) - y)^2 \\ &= 2 \cdot \frac{1}{2} (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_{\theta}(x) - y) \\ &= (h_{\theta}(x) - y) \cdot \frac{\partial}{\partial \theta_j} \left( \sum_{i=0}^n \theta_i x_i - y \right) \\ &= (h_{\theta}(x) - y) x_j \end{aligned}$$

The point of all this is that if we start with a guess for our hypothesis and then repeatedly apply these gradient descent equations, our hypothesis will become more and more accurate.

So, this is simply gradient descent on the original cost function  $J$ . This method looks at every example in the entire training set on every step, and is called **batch gradient descent**. Note that, while gradient descent can be susceptible to local minima in general, the optimization problem we have posed here for linear regression has only one global, and no other local, optima; thus gradient descent always converges (assuming the learning rate  $\alpha$  is not too large) to the global minimum. Indeed,  $J$  is a convex quadratic function. Here is an example of gradient descent as it is run to minimize a quadratic function.

