Let us recall the cost function with a simplified hypothesis.

$$H(x) = Wx$$

$$\therefore \text{cost}(W) = \frac{1}{m}\sum_{i=1}^{m}(Wx_i - y_i)^2$$

Note that now the cost function can be expressed as a function of W.

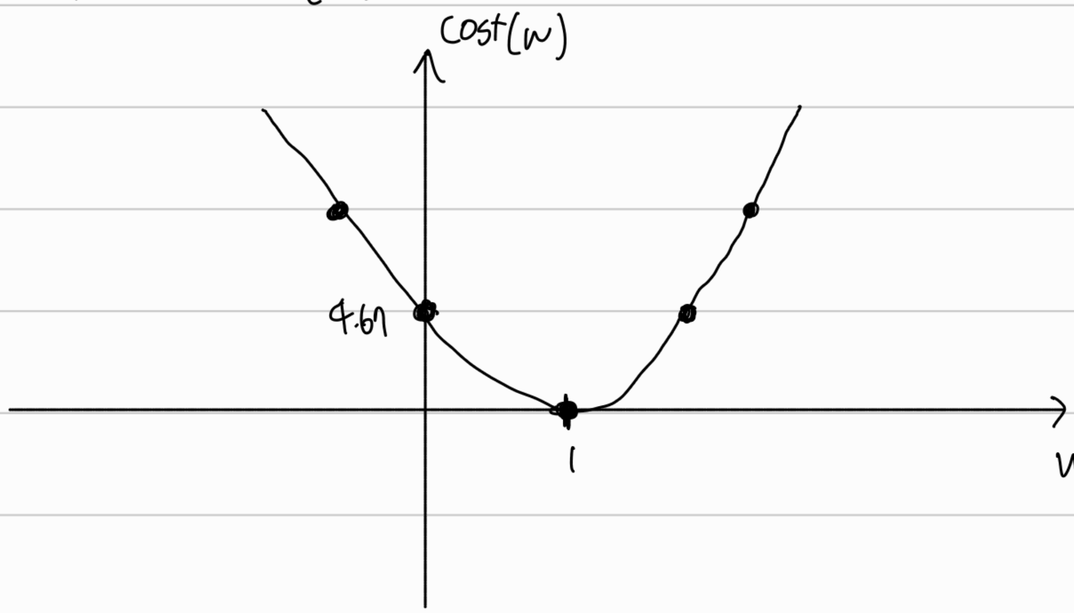
Given the following data and W value, what is cost (w)?

x	y
1	1
2	2
3	3

- $W=1$, $\text{Cost}(w) = \frac{1}{3}((1\cdot 1-1)^2 + (1\cdot 2-2)^2 + (1\cdot 3-3)^2) = 0$
- $W=0$, $\text{cost}(w) = \frac{1}{3}((0\cdot 1-1)^2 + (0\cdot 2-2)^2 + (0\cdot 3-3)^2) = 4.67$
- $W=2$, $\text{cost}(w) = \frac{1}{3}((2\cdot 1-1)^2 + (2\cdot 2-2)^2 + (2\cdot 3-3)^2) = 4.67$

The pattern resembles a parabola

What does cost (W) look like?



Question arises: How do we find the values of W, b to minimize cost(w, b)? what about a more general function like Cost(w_1, w_2, ...)?

Gradient Descent Algorithm

slope come down

- Minimize cost is goal
- Gradient Descent is used in many minimization problems
- For given cost(w,b), it finds w and b that minimizes cost(w,b)
- Can be applied to more general functions : cost$(w_1, w_2, \ldots)$

How does Gradient Descent work?

↳ How would you find the lowest point?

Suppose you're at the top of a mountain. How do you get down?

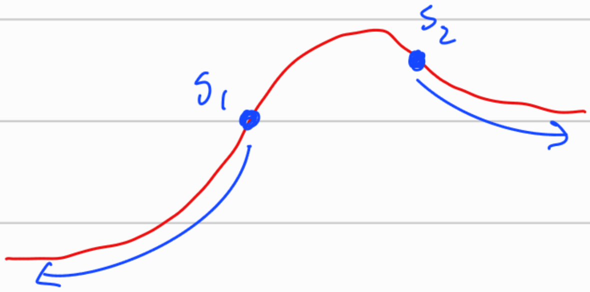
A: follow the track with a downwards slope until you reach the ground

✯ slope = derivative, ground = local minimum (derivative is 0)

- start w/ initial guess
  - Start at 0,0 (or any other value)
  - Keep changing w and b a little bit to reduce cost(w,b)
- Each time you change the parameters, choose the gradient which reduces cost(w,b) the most
- Repeat until converging to local minimum

✯ where you start can change which minimum you end up on

Ex.



Starting at s_1 and s_2 different minimums

Formal definitions:

For simplicity in the derivation of cost, we slightly alter cost(w).

$$Cost(w) = \frac{1}{2m} \sum_{i=1}^{m} (wx_i - y_i)^2$$

Note that since our goal is to minimize cost, dividing cost by 2 and then finding the minimum is also allowed

Now we reassign W after each iteration of gradient descent

$$W := W - \alpha \frac{\partial}{\partial w} cost(w)$$

★ α is what is called "learning rate." For now, think of it as a constant. (Bonus: think about how it affects changes in W)

Note that we take the derivative of cost w/ respect to W b/c we want to find min of cost(w)

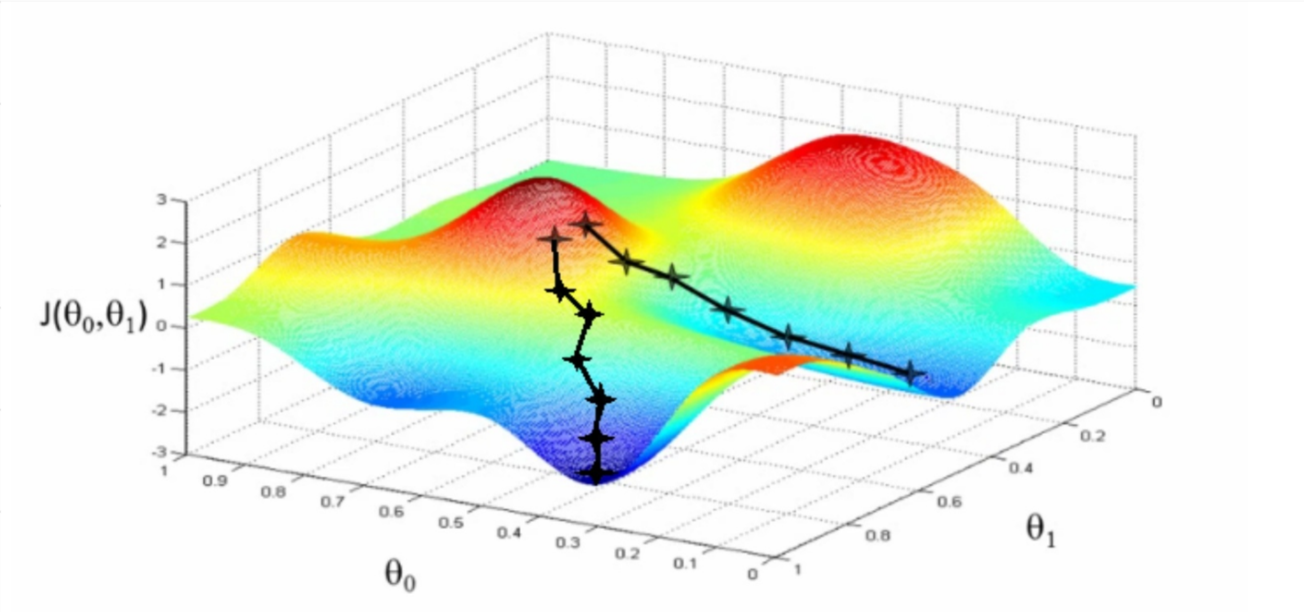
$$\therefore W := W - \alpha \frac{\partial}{\partial w} \frac{1}{2m} \sum_{i=1}^{m} (Wx_i - y_i)^2$$

$$W := W - \alpha \frac{1}{\cancel{2}m} \sum_{i=1}^{m} \cancel{2} (Wx_i - y_i) x_i$$

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

The Gradient Descent algorithm can therefore be expressed by the above equation

Gradient Descent visualization



www.holehouse.org/mlclass