Gradient Descent

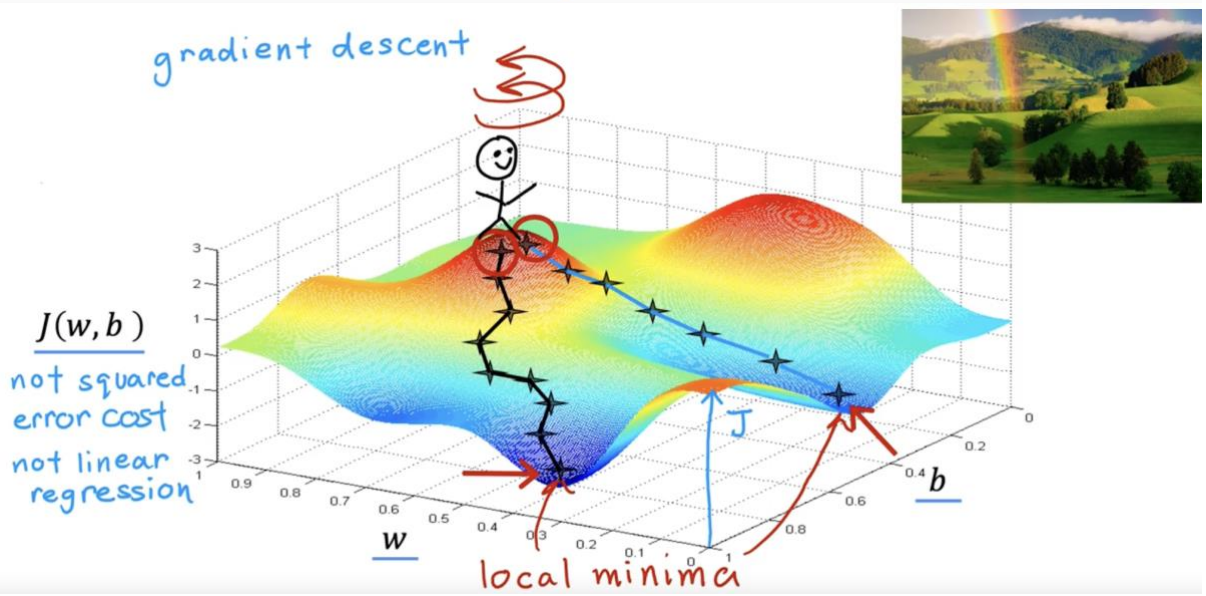
Have some function $J(w,b)$ for linear regression or any function

Want $\min_{w,b} J(w,b)$ $\min_{w_1,\ldots,w_n,b} J(w_1,w_2,\ldots,w_n,b)$

Outline:

- Start with some w,b (set $w=0$, $b=0$)
- Keep changing w,b to reduce $J(w,b)$
- Until we settle at or near a minimum

J not always



For each and every step, you will look around at 360 degree and decides to go in which direction.

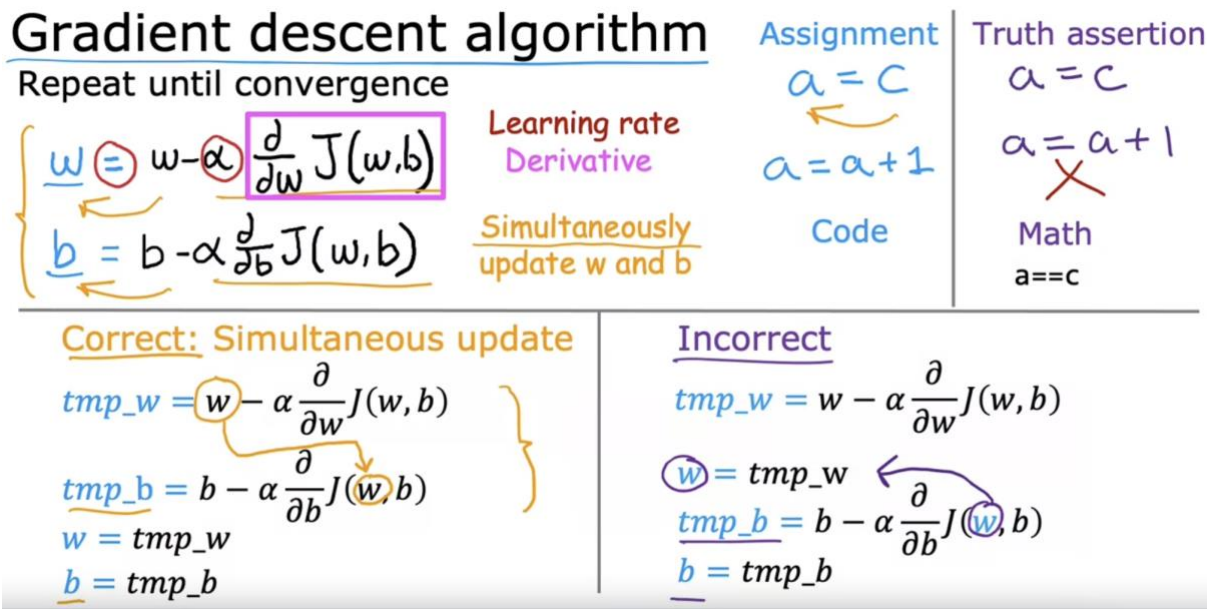
Implementing gradient descent

$\frac{d}{dw} J(w,b)$ → After looking around for 360 degree, Which direction to go ??

α → How many steps you would take in that particular direction ??

$\alpha \frac{d}{dw} J(w,b)$ Learning rate, Derivative

→ Steps take in a particular direction



temp_b takes the old w value

temp_b takes the updated w value

Question

Gradient descent is an algorithm for finding values of parameters w and b that minimize the cost function J. What does this update statement do? (Assume α is small.)

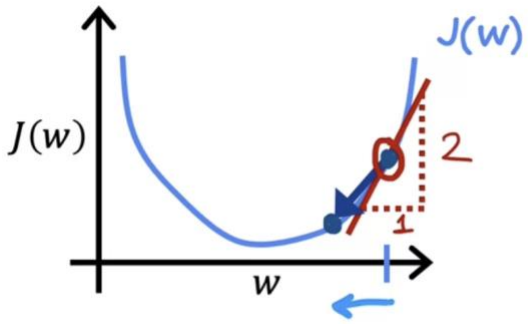
$$w = w - \alpha \frac{\partial J(w,b)}{\partial w}$$

- (●) Updates parameter w by a small amount
- ( ) Checks whether w is equal to $w - \alpha \frac{\partial J(w,b)}{\partial w}$

✓ **Correct**
This updates the parameter by a small amount, in order to reduce the cost J.

Gradient Descent Intuition

Now starting value of w is too high. Will get a positive slope. So w is reducing.

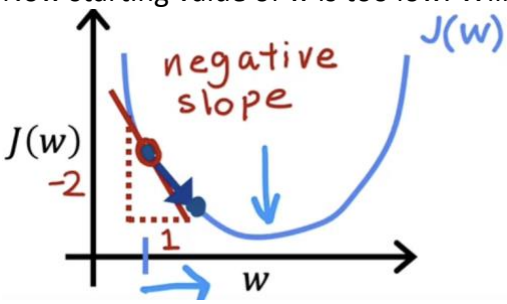


$$w = w - \alpha \frac{d}{dw} J(w)$$

$$> 0$$

$$w = w - \alpha \cdot (positive\ number)$$

Now starting value of w is too low. Will get a negative slope. So w is increasing.



$$\frac{d}{dw} J(w) < 0$$

$$w = w - \alpha \cdot (negative\ number)$$

Question

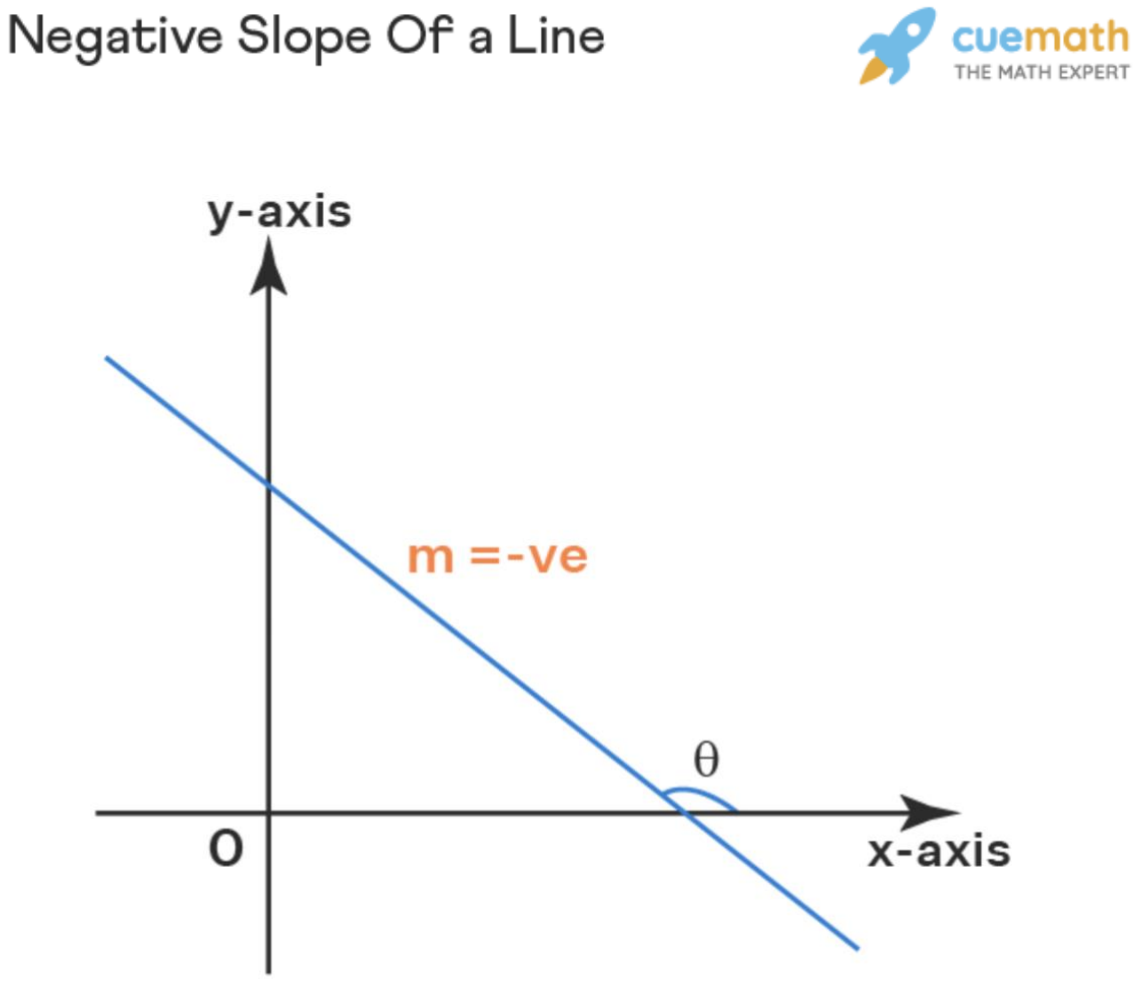
Assume the learning rate α is a small positive number. When $\frac{\partial J(w,b)}{\partial w}$ is a positive number (greater than zero) -- as in the example in the upper part of the slide shown above -- what happens to w after one update step?

- w increases
- It is not possible to tell if w will increase or decrease.
- w stays the same
- w decreases.

✓ **Correct**
The learning rate α is always a positive number, so if you take W minus a positive number, you end up with a new value for W that is smaller

Graph of Negative Slope

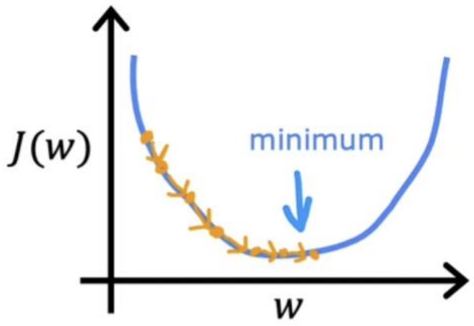
The concept of the negative slope gives an inverse relationship between two quantities. The two quantities are represented graphically across the x-axis and the y-axis, and the line is plotted to represent the relationship between these two variables. As the value of the quantity represented along with the x-axis increases, the value of the other quantity represented along with the y-axis decreases. The inverse relationship of the increase of the x value, and the decrease of the y value is represented by the negative slope of the line.



A negative slope graph is a line with a negative slope, which falls as it moves from left to right. The line with a negative slope makes an obtuse angle with the positive x-axis, in the anti-clockwise direction.

Learning Rate

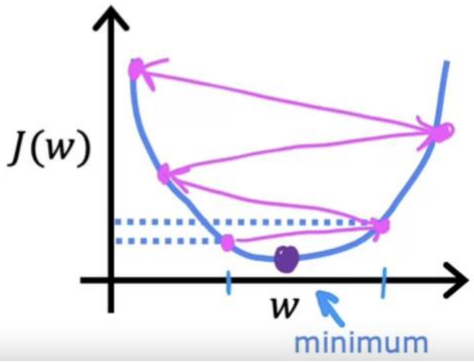
$$w = w - \alpha \frac{d}{dw} J(w)$$



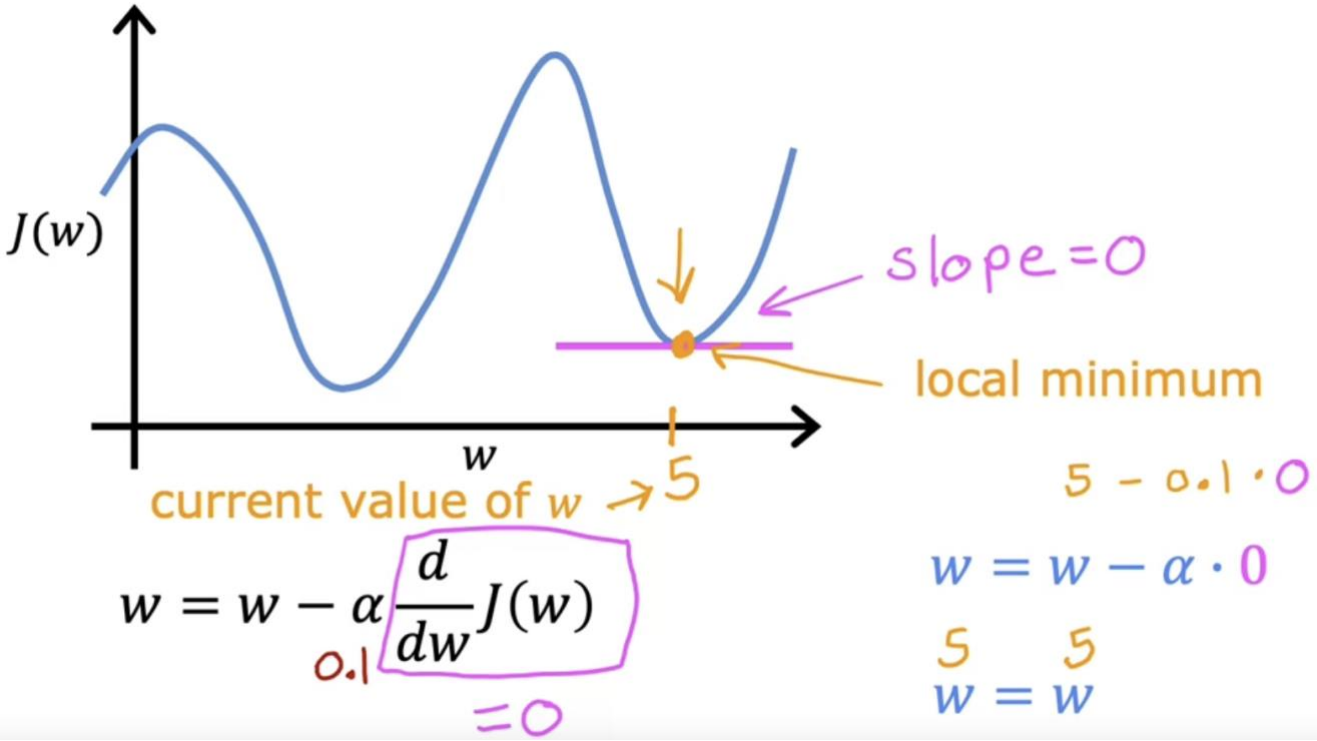
If α is too small...
Gradient descent may be slow.

If α is too large...

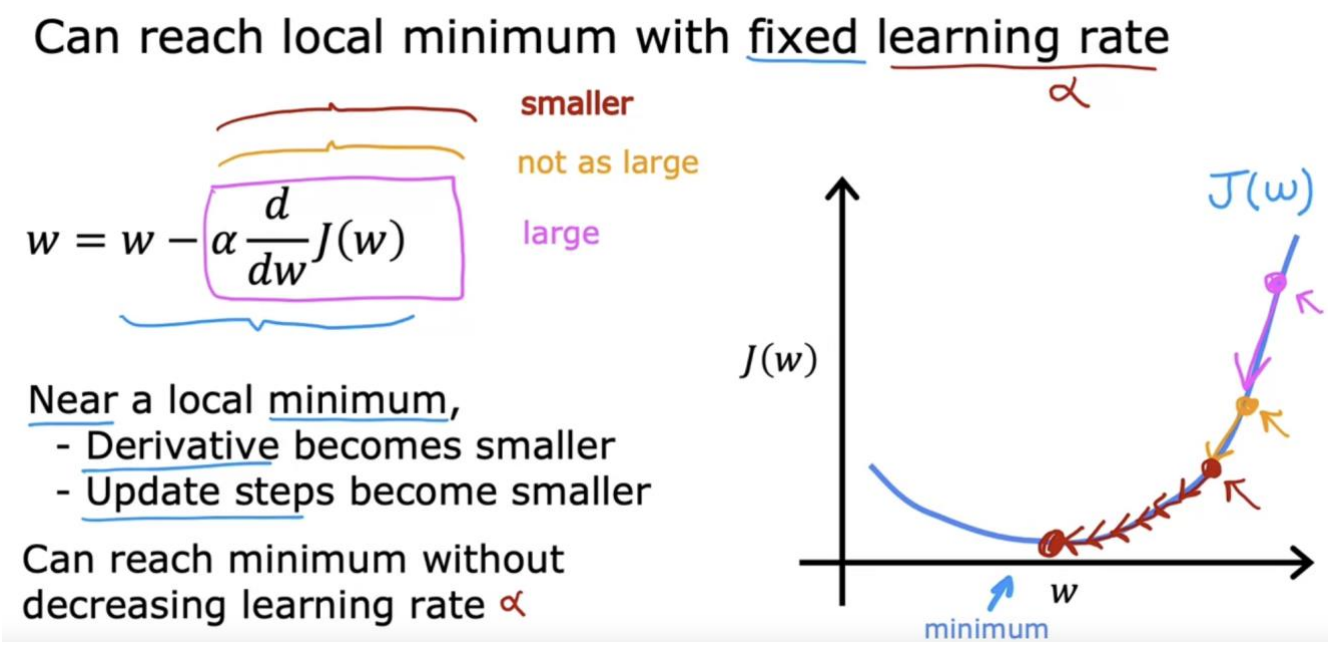
Gradient descent may:
- Overshoot, never reach minimum
- Fail to converge, diverge



When the w is already at the local minimum ??



So gradient descent makes w remains unchanged



pink → slope is large, So derivative is also large, So large step.
orange → slope is small, So derivative is also small, So small step.
Red → slope is too small, So derivative is also too small, So too small step.

So as we approach the minimum. The derivative gets closer and closer to zero.
So as we run gradient descent, eventually we're taking very small steps until you finally reach a local minimum.

Gradient Descent for Linear Regression

(Optional)

$$\frac{\partial}{\partial w}J(w,b) = \frac{\partial}{\partial w}\frac{1}{2m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)^2 = \frac{\partial}{\partial w}\frac{1}{2m}\sum_{i=1}^{m}\left(wx^{(i)}+b-y^{(i)}\right)^2$$

$$= \frac{1}{2m}\sum_{i=1}^{m}\left(wx^{(i)}+b-y^{(i)}\right)2x^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)x^{(i)}$$

$$\frac{\partial}{\partial b}J(w,b) = \frac{\partial}{\partial b}\frac{1}{2m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)^2 = \frac{\partial}{\partial b}\frac{1}{2m}\sum_{i=1}^{m}\left(wx^{(i)}+b-y^{(i)}\right)^2$$

$$= \frac{1}{2m}\sum_{i=1}^{m}\left(wx^{(i)}+b-y^{(i)}\right)2 = \frac{1}{m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)$$

no $x^{(i)}$

Gradient descent algorithm

$\frac{\partial}{\partial w}J(w,b)$

repeat until convergence {

$$w = w - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)x^{(i)}$$

$$b = b - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(f_{w,b}(x^{(i)})-y^{(i)}\right)$$

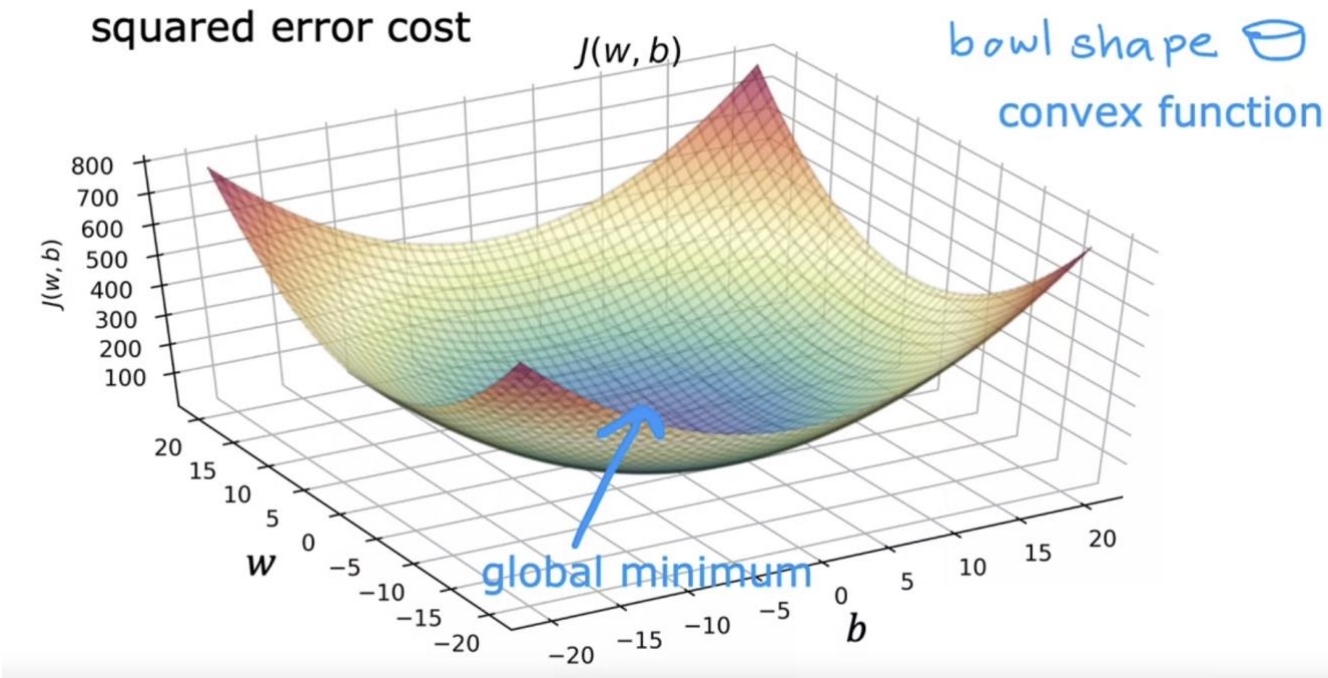
}

Update w and b simultaneously

$\frac{\partial}{\partial b}J(w,b)$

$f_{w,b}(x^{(i)}) = wx^{(i)}+b$

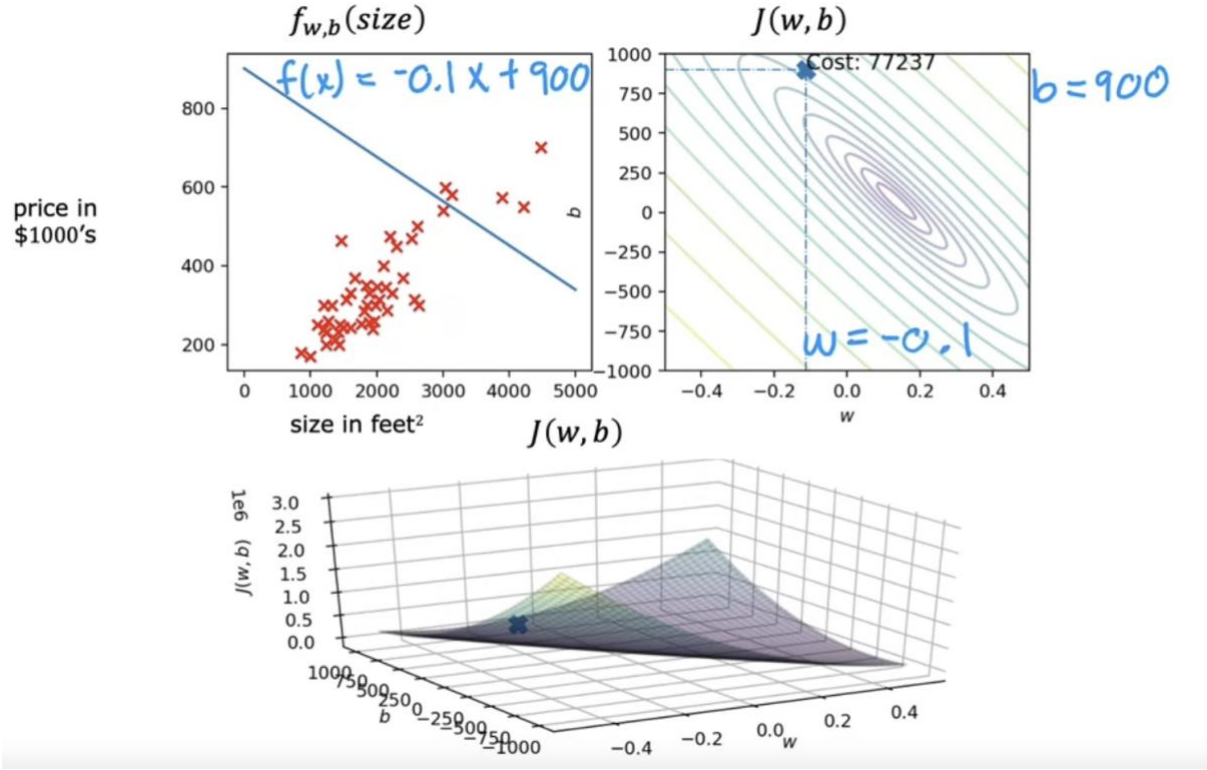
When using square errored cost function



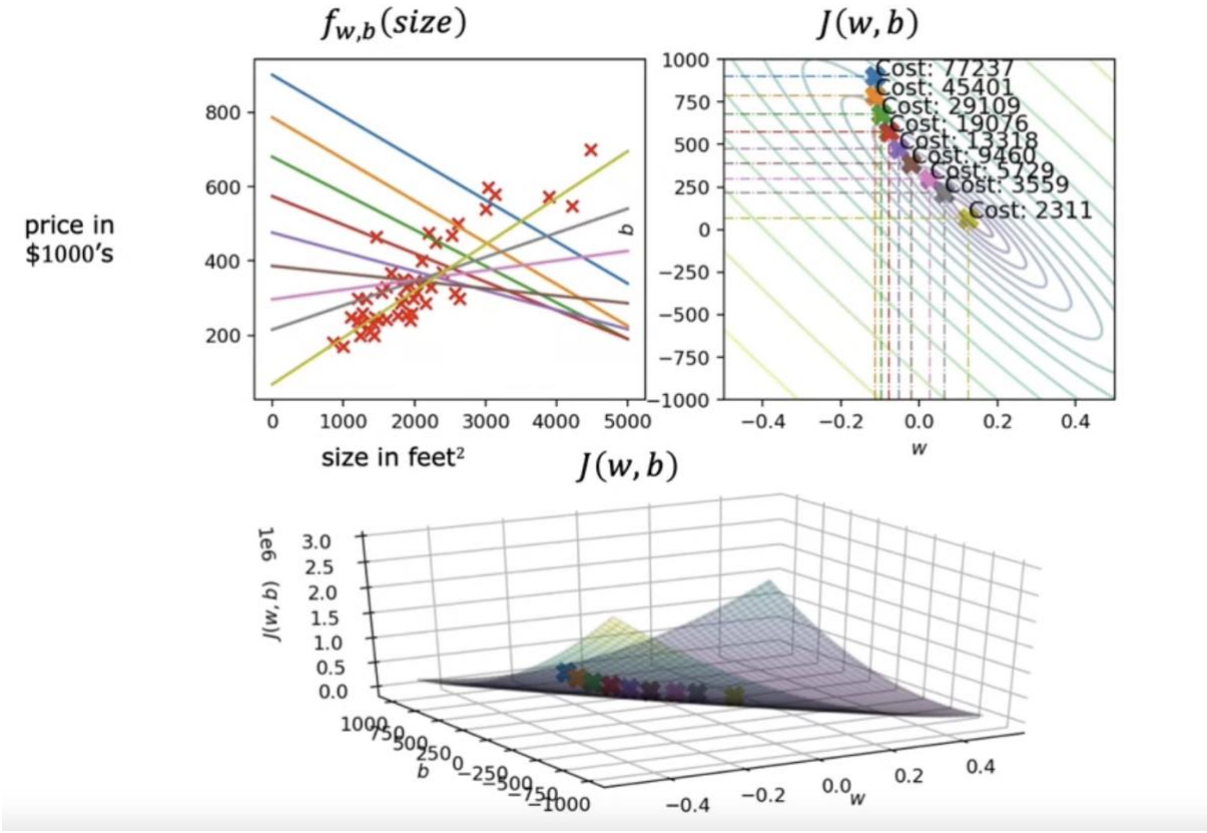
Convex function is a bowl shaped function which have only one global minimum and no local minimum it-self.

When you implement gradient descent on a convex function, one nice property is that so long as you're learning rate is chosen appropriately, it will always converge to the global minimum.

Running Gradient Descent



After running gradient descent,



"Batch" gradient descent

"Batch": Each step of gradient descent uses all the training examples.

other gradient descent: subsets

	x size in feet²	y price in $1000's
(1)	2104	400
(2)	1416	232
(3)	1534	315
(4)	852	178
...	...	...
(47)	3210	870

$m = 47$

$$\sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)})^2$$

Refer → 5_Gradient_Descent_Solution.ipynb

Quiz

1. 1 / 1 point

Gradient descent is an algorithm for finding values of parameters w and b that minimize the cost function J.

repeat until convergence {

$$w = w - \alpha \frac{\partial}{\partial w} J(w,b)$$

$$b = b - \alpha \frac{\partial}{\partial b} J(w,b)$$

When $\frac{\partial J(w,b)}{\partial w}$ is a negative number (less than zero), what happens to w after one update step?

- ○ w stays the same
- ◉ w increases.
- ○ It is not possible to tell if w will increase or decrease.
- ○ w decreases

✓ Correct
The learning rate is always a positive number, so if you take W minus a negative number, you end up with a new value for W that is larger (more positive).

2. 1 / 1 point

For linear regression, what is the update step for parameter b?

- ○ $b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)}) x^{(i)}$
- ◉ $b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)})$

✓ Correct
The update step is $b = b - \alpha \frac{\partial J(w,b)}{\partial w}$ where $\frac{\partial J(w,b)}{\partial b}$ can be computed with this expression: $\sum_{i=1}^{m} (f_{w,b}(x^{(i)}) - y^{(i)})$

References

Gradient descent | Coursera
https://www.cuemath.com/geometry/negative-slope/