

→ Here we will see how different learning rate effects the cost function graph.

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

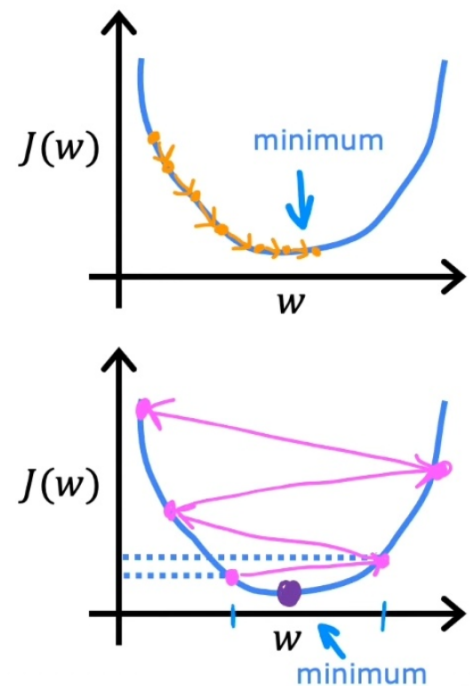
If α is too small...

Gradient descent may be slow.

If α is too large...

Gradient descent may:

- Overshoot, never reach minimum
- Fail to converge, diverge



→ If we choose α as a small value, it will take tiny steps toward local minima

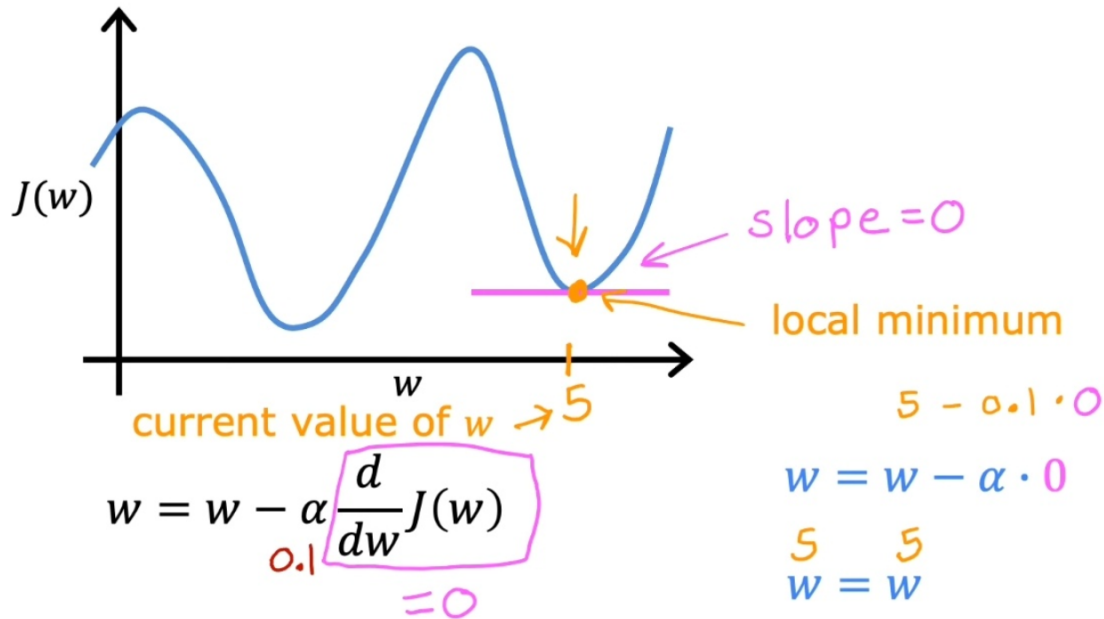
→ so it will be slow.

→ If α is too high, it may work opposite.

→ Cause it will take huge steps and it may increase the $J(w)$. since the

derivatives term is changing.

→ The value of w and b will increase.



→ The main drawback of this formula is, it always sticks on local minima.

→ Cause, at the bottom of the curve the slope is 0. So, it becomes $w = w - 0$

→ the value of w won't change.

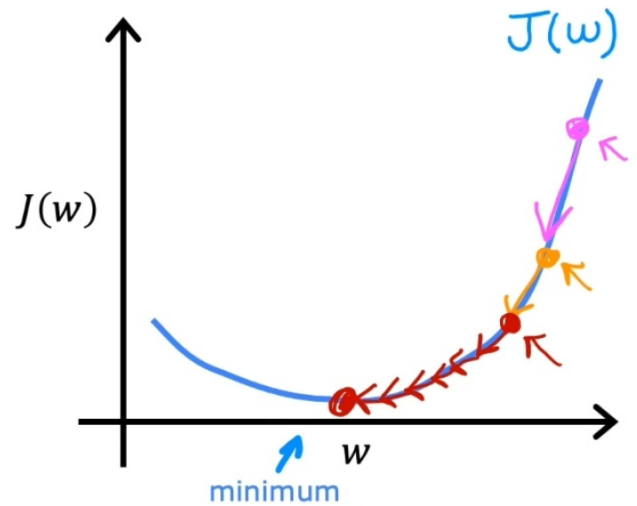
Can reach local minimum with fixed learning rate

$$w = w - \underbrace{\alpha}_{\text{smaller}} \underbrace{\frac{d}{dw} J(w)}_{\text{not as large}} \underbrace{\phantom{\alpha \frac{d}{dw} J(w)}}_{\text{large}}$$

Near a local minimum,

- Derivative becomes smaller
- Update steps become smaller

Can reach minimum without decreasing learning rate α



→ We do not need to change the learning rate. It will work with fixed α

→ At first step it will take large step. after that it will take small steps as reaching toward local minima.

→ cause the slope value is decreasing as we reaching toward local minima.