6. (a) In your own words, describe how to use the update equation in the gradient descent algorithm.

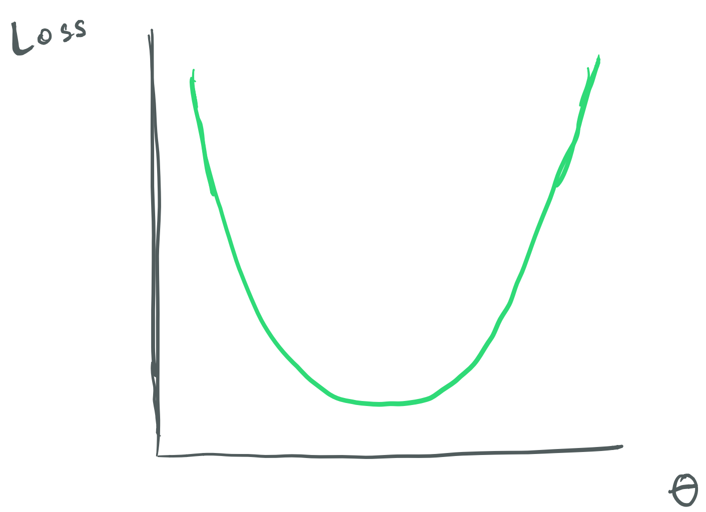
(b) Say that x and y are your model parameters and f as defined in question 1 is your loss function. Describe in your own words what happens "visually" as the gradient descent algorithm runs.

Recall the general update equation for gradient descent:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(t)}}$$

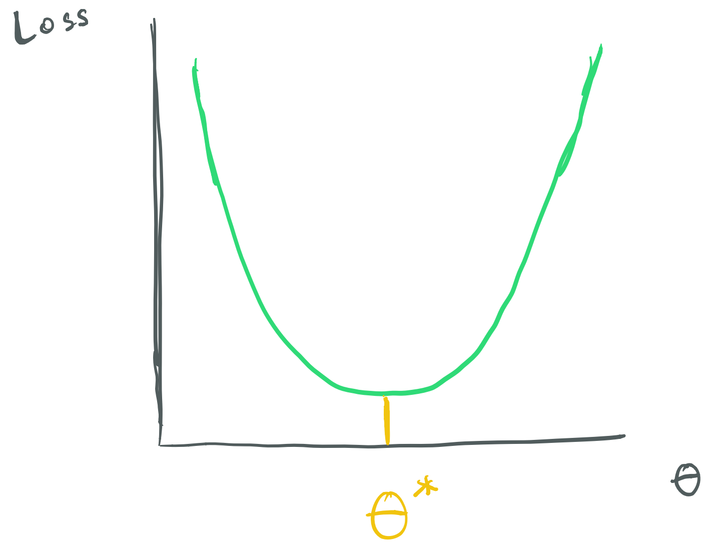
$\frac{\partial L}{\partial \theta}$ is a function of θ, so evaluate it at $\theta = \theta^{(t)}$

Let's look at an example of gradient descent in 2-D to see what gradient descent is actually doing. Suppose we use the average squared loss, so our loss function looks something like:



Remember, the goal of gradient descent is to find the θ that minimizes the loss function. I'll call that optimal value θ^*.

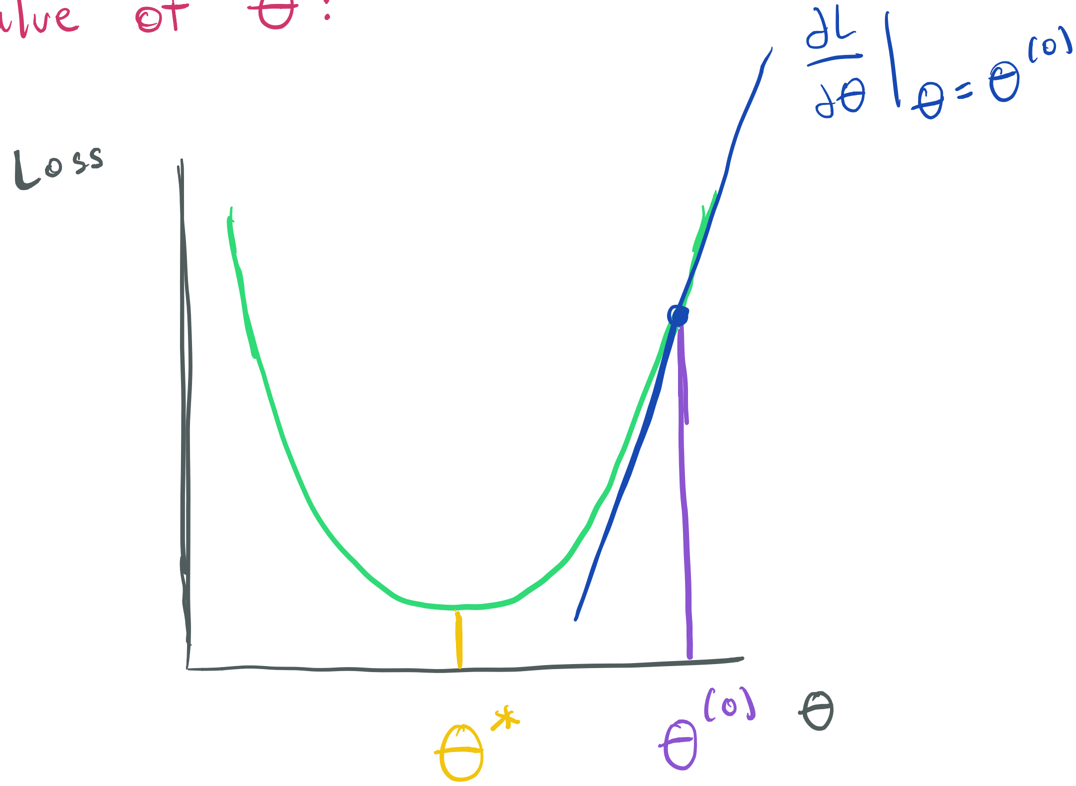
On the graph, θ^* looks like:



The way gradient descent works is that starts with a random guess for θ^*, which is the initial value $\theta^{(0)}$. On the graph, $\theta^{(0)}$ could be anywhere since it is a random guess.

Obviously, $\theta^{(0)}$ is not θ^*, as we can see from the graph. But how does gradient descent know that $\theta^{(0)} \neq \theta^*$? That's where $\frac{\partial L}{\partial \theta}$ comes in.

In our 2D example, $\frac{\partial L}{\partial \theta}$ represents the slope of the tangent line to the loss function at our current value of θ:



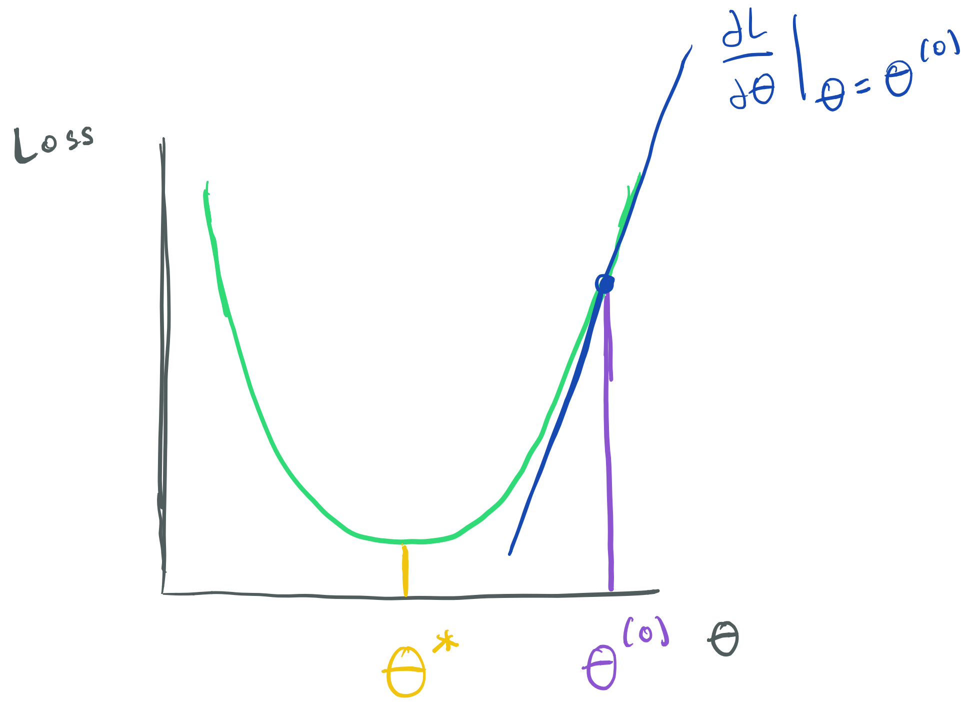
Since $\frac{\partial L}{\partial \theta}$ at $\theta = \theta^{(0)}$ is large and positive, gradient descent knows that $\theta^{(0)} \neq \theta^*$. This is because $\frac{\partial L}{\partial \theta} = 0$ at $\theta = \theta^*$, since θ^* is the minimum for L.

So gradient descent knows that $\theta^{(0)} \neq \theta^*$, so it needs to guess another value for θ, which will be $\theta^{(1)}$. But what should $\theta^{(1)}$ be?

Let's use the information that $\frac{\partial L}{\partial \theta}$ is large and positive at $\theta = \theta^{(0)}$. Since $\frac{\partial L}{\partial \theta}$ is positive, we know that L is increasing at $\theta = \theta^{(0)}$. This means that $\theta^* < \theta^{(0)}$, because functions increase after attaining their minimum value (we can say this based on the definition of what minimum is). Thus, we know that we need to "push back" our estimate for θ^* by choosing $\theta^{(1)}$ to be less than $\theta^{(0)}$.

this is where the - sign comes from in the update equation

But how much less than $\theta^{(0)}$ should $\theta^{(1)}$ be? Let's use the other piece of information we have about $\frac{\partial L}{\partial \theta}$ at $\theta = \theta^{(0)}$, that $\frac{\partial L}{\partial \theta}$ is large (in addition to positive)



Since $\frac{\partial L}{\partial \theta}$ at $\theta = \theta^{(0)}$ is large, we can say that $\theta^{(0)}$ is pretty far from θ^*. The reason for this relies on the interpretation of $\frac{\partial L}{\partial \theta}$ as the slope of the tangent line. If the slope of the tangent line is large and positive, then the function is increasing quickly. You can see this from $\frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(0)}}$ on the graph. Since we are trying to find the value of θ that minimizes the function, we want to move away from values of θ where the function is increasing quickly.

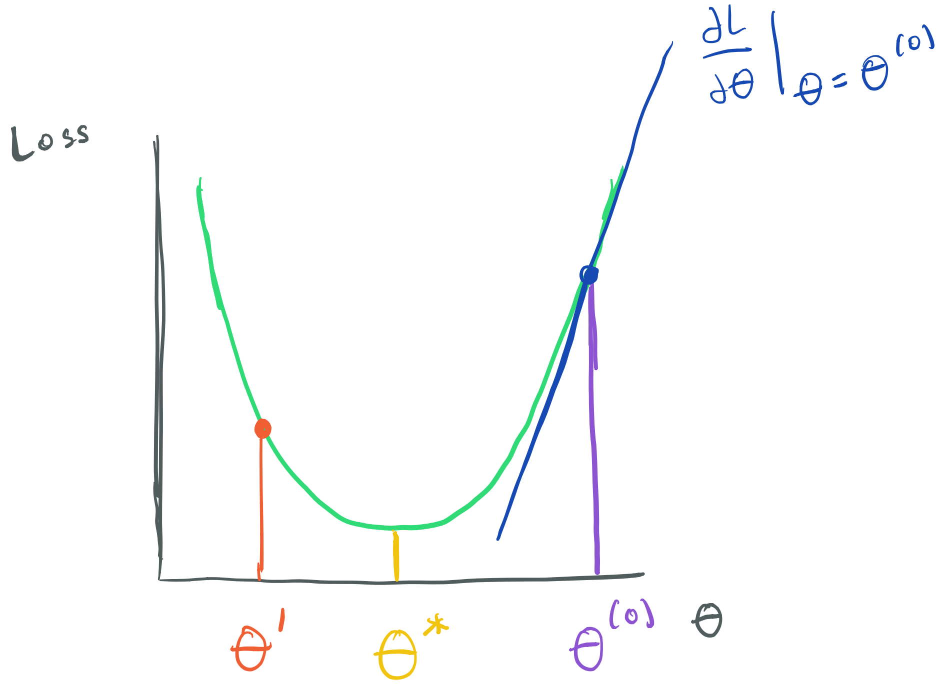
Specifically, we know that when the slope of the tangent line is large, we need to change our θ by a lot. In general, we know that the larger our $\frac{\partial L}{\partial \theta}$, the more we need to change θ. The way the gradient descent update expresses this is by changing θ by an amount proportional to $\frac{\partial L}{\partial \theta}$:

$$\theta^{(t+1)} \leftarrow \underbrace{\theta^{(t)} - \alpha \left.\frac{\partial L}{\partial \theta}\right|_{\theta=\theta^{(t)}}}_{\text{"push back" } \theta \text{ by an amount proportional to } \left.\frac{\partial L}{\partial \theta}\right|_{\theta=\theta^{(t)}}}$$

Let's stop and do a quick summary of what we've said so far. The reason the update rule has a minus sign is because when $\frac{\partial L}{\partial \theta}$ is positive, we've overshot θ^* and have to "push back" our estimate to be less than the previous estimate.

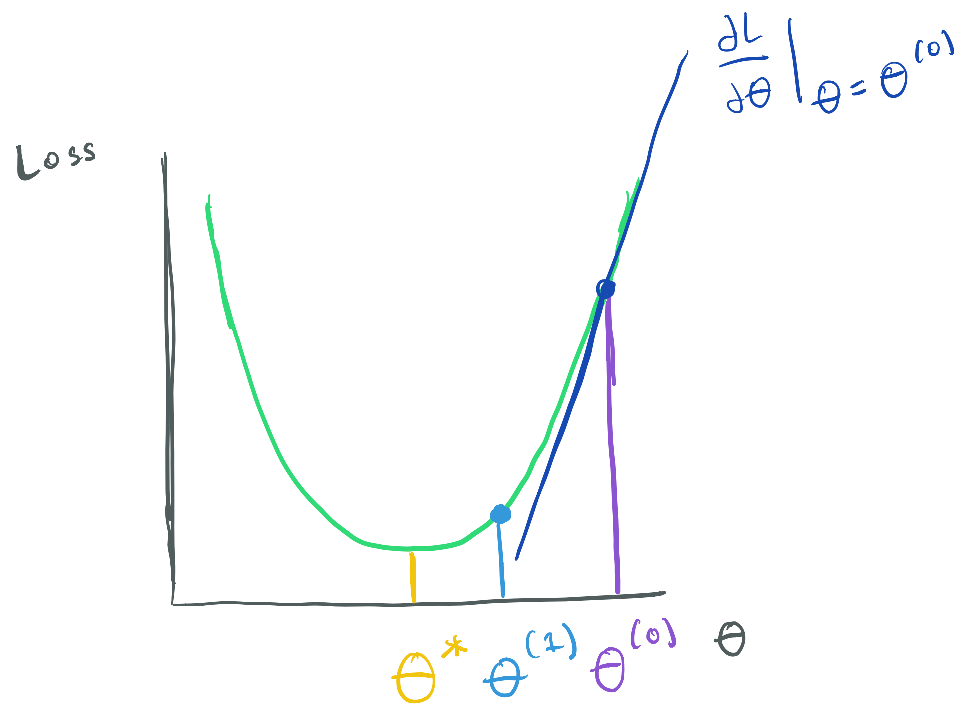
Additionally, we push back the estimate for θ^* by an amount proportional to $\frac{\partial L}{\partial \theta}$ because when $\frac{\partial L}{\partial \theta}$ is large, we know we are far from θ^*.

One last thing to take care of: what is the purpose / role of α?



When we push back $\theta^{(0)}$ in the graph above, depending on how large $\frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(0)}}$ is, we might push θ back to a value like θ', which overshoots θ^* on the other side.

To avoid this problem of overshooting θ^*, we multiply $\frac{\partial L}{\partial \theta}$ by a fraction α so we don't push our estimate by the full magnitude of $\frac{\partial L}{\partial \theta}$. Keep in mind $0 < \alpha < 1$ (think about why). With this fraction α, we push $\theta^{(0)}$ to a value like $\theta^{(1)}$.



All of what we've said so far happens each time θ is updated. But how do we know how many times to update θ?

We continue updating θ until $\theta^{(t+1)} = \theta^{(t)}$. This is because, looking at the update equation, the only way for $\theta^{(t+1)} = \theta^{(t)}$ is if $\frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(t)}} = 0$. But if $\frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(t)}} = 0$, then $\theta^{(t)}$ is the minimum! Thus, $\theta^{(t)} = \theta^{*}$ and we've found the optimal θ^{*}.