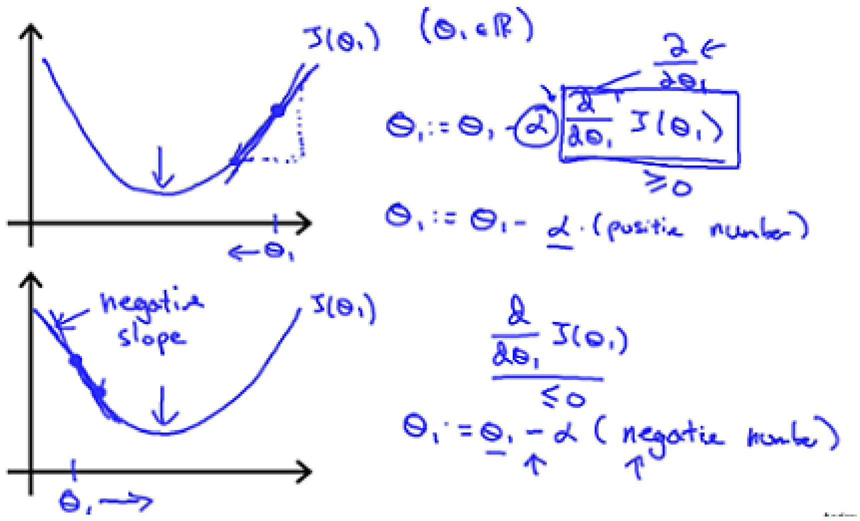
Gradient Descent Intuition

In lesson video we explored the scenario where we used one parameter θ_1 and plotted its cost function to implement a gradient descent. Our formula for a single parameter was:

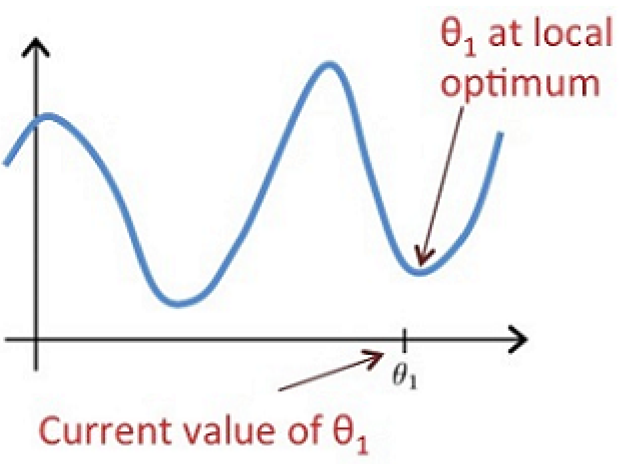
Repeat until convergence:

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

Regardless of the slope's sign for $\frac{\partial}{\partial \theta_1} J(\theta_1)$, θ_1 eventually converges to its minimum value. The following graph shows that when the slope is negative, the value of θ_1 increases and when it is positive, the value of θ_1 decreases.



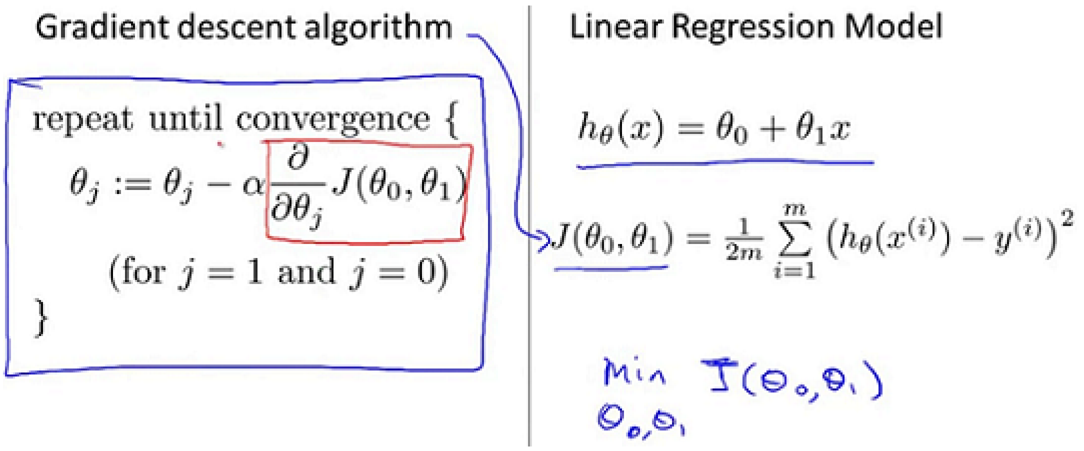
Video Question: Suppose θ_1 is at local optimum of $J(\theta_1)$, such as shown in the figure. What will one step of gradient descent $\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$ do?



Leave θ_1 unchanged

- Change θ_1 in a random direction
- Move θ_1 in the direction of the global minimum of $J(\theta_1)$
- Drecrease θ_1

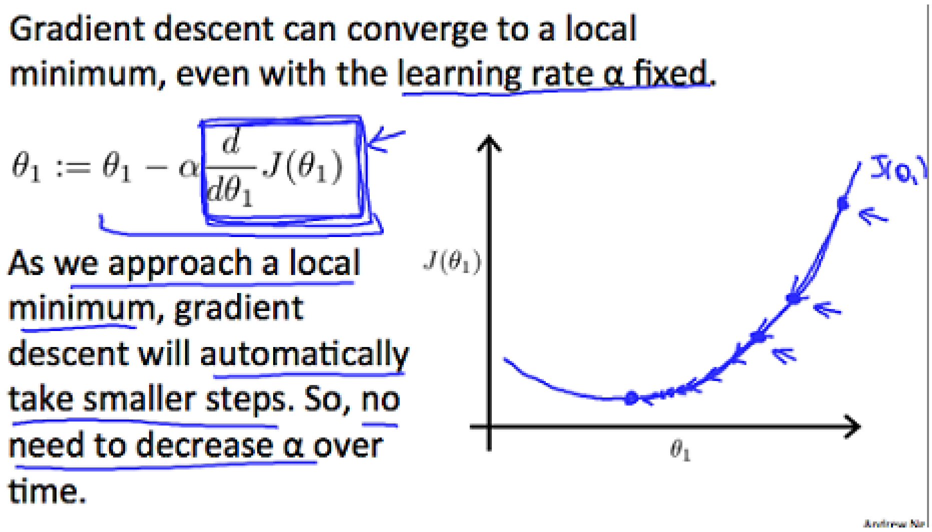
On a side note, we should adjust our parameter α to ensure that the gradient descent algorithm converges in a reasonable time. Failure to converge or too much time to obtain the minimum value imply that our step size is wrong.



How does gradient descent converge with a fixed step size α?

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

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



That's the gradient descent algorithm and we can use it to try to minimize any cost function J, not the cost function J that we defined for linear regression.