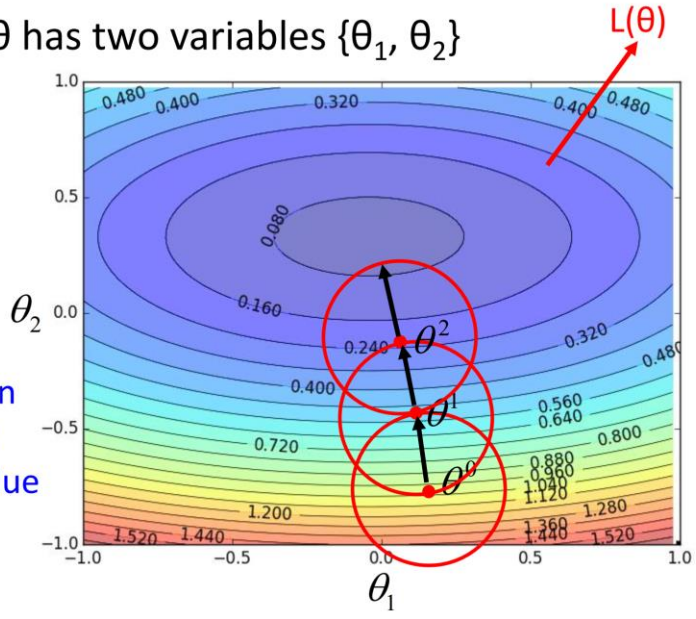
Gradient Descent

Theory

Formal Derivation

- Suppose that θ has two variables $\{\theta_1, \theta_2\}$



Given a point, we can easily find the point with the smallest value nearby. How?

Back to Formal Derivation

Based on Taylor Series:

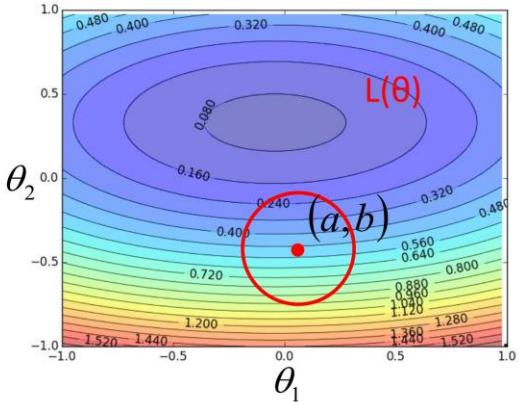
If the red circle is ***small enough***, in the red circle

$$L(\theta) \approx L(a,b) + \frac{\partial L(a,b)}{\partial \theta_1}(\theta_1 - a) + \frac{\partial L(a,b)}{\partial \theta_2}(\theta_2 - b)$$

$$s = L(a,b)$$

$$u = \frac{\partial L(a,b)}{\partial \theta_1}, v = \frac{\partial L(a,b)}{\partial \theta_2}$$

$$L(\theta) \approx s + u(\theta_1 - a) + v(\theta_2 - b)$$



Back to Formal Derivation

Based on Taylor Series:

If the red circle is ***small enough***, in the red circle

constant

$$s = L(a,b)$$

$$u = \frac{\partial L(a,b)}{\partial \theta_1}, v = \frac{\partial L(a,b)}{\partial \theta_2}$$

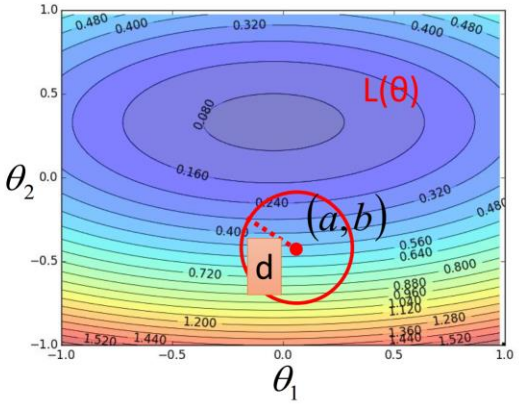
$$L(\theta) \approx s + u(\theta_1 - a) + v(\theta_2 - b)$$

Find θ_1 and θ_2 in the red circle

minimizing L(θ)

$$(\theta_1 - a)^2 + (\theta_2 - b)^2 \le d^2$$

Simple, right?

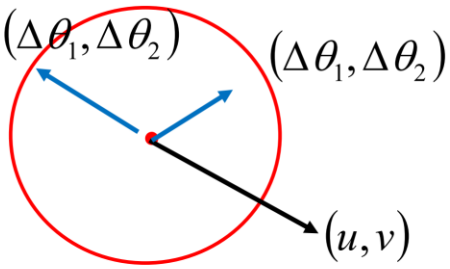


Gradient descent – two variables

Red Circle: (If the radius is small)

$$L(\theta) \approx s + u(\theta_1 - a) + v(\theta_2 - b)$$

$$\Delta\theta_1 \qquad \Delta\theta_2$$



Find θ_1 and θ_2 in the red circle ***minimizing*** L(θ)

$$(\theta_1 - a)^2 + (\theta_2 - b)^2 \le d^2$$

$$\Delta\theta_1 \qquad \Delta\theta_2$$

To minimize L(θ)

$$\begin{bmatrix}\Delta\theta_1 \\ \Delta\theta_2\end{bmatrix} = -\eta\begin{bmatrix}u \\ v\end{bmatrix} \Rightarrow \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} = \begin{bmatrix}a \\ b\end{bmatrix} - \eta\begin{bmatrix}u \\ v\end{bmatrix}$$

Back to Formal Derivation

Based on Taylor Series:

If the red circle is ***small enough***, in the red circle

$$L(\theta) \approx s + u(\theta_1 - a) + v(\theta_2 - b)$$

constant

$$s = L(a,b)$$

$$u = \frac{\partial L(a,b)}{\partial \theta_1}, v = \frac{\partial L(a,b)}{\partial \theta_2}$$

Find θ_1 and θ_2 yielding the smallest value of $L(\theta)$ in the circle

$$\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} = \begin{bmatrix}a \\ b\end{bmatrix} - \eta\begin{bmatrix}u \\ v\end{bmatrix} = \begin{bmatrix}a \\ b\end{bmatrix} - \eta\begin{bmatrix}\frac{\partial L(a,b)}{\partial \theta_1} \\ \frac{\partial L(a,b)}{\partial \theta_2}\end{bmatrix}$$

This is gradient descent.

Not satisfied if the red circle (learning rate) is not small enough

You can consider the second order term, e.g. Newton's method.