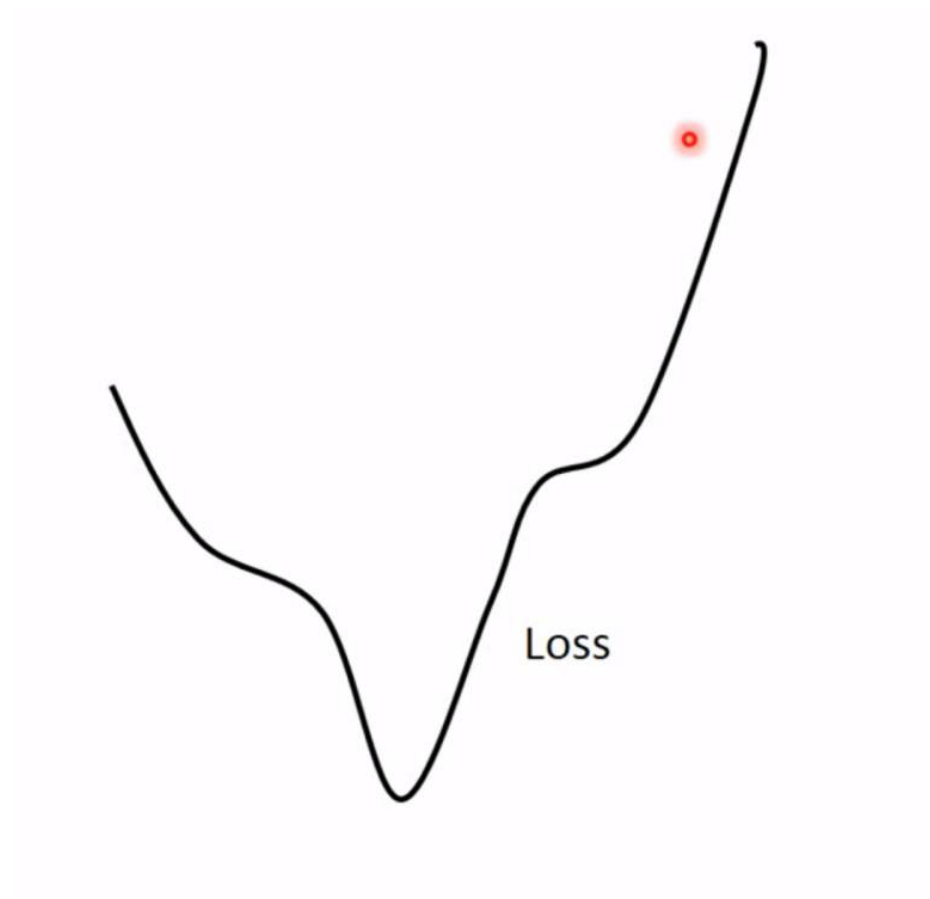


Gradient Descent



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

Randomly start at $\theta^0 = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix}$

$$\nabla L(\theta) = \begin{bmatrix} \partial C(\theta_1)/\partial \theta_1 \\ \partial C(\theta_2)/\partial \theta_2 \end{bmatrix}$$

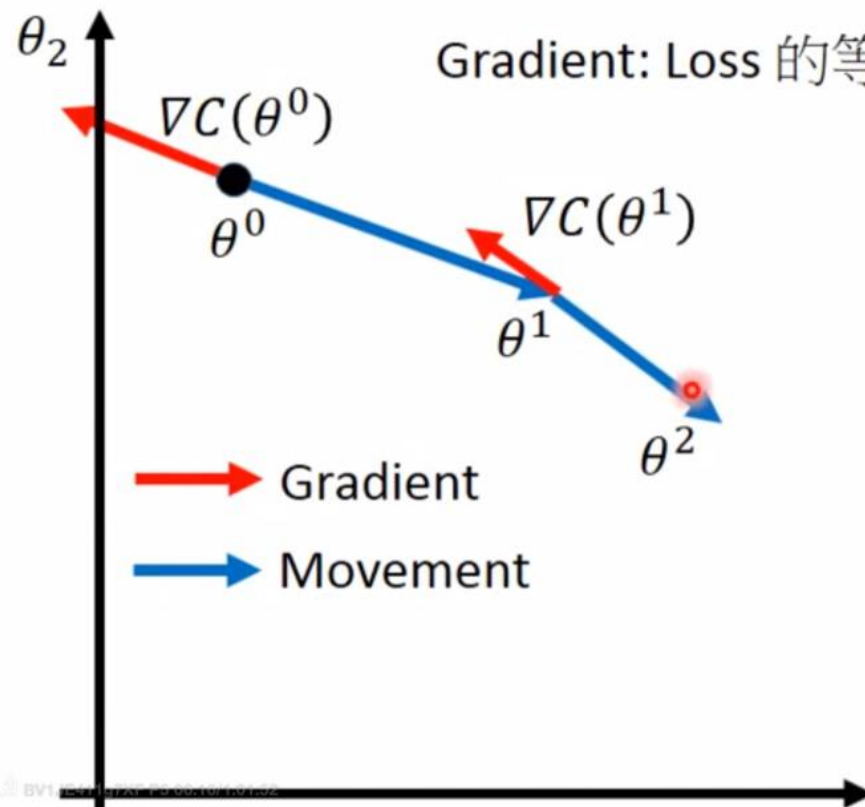
$$\begin{bmatrix} \theta_1^1 \\ \theta_2^1 \end{bmatrix} = \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \end{bmatrix} - \eta \begin{bmatrix} \partial L(\theta_1^0)/\partial \theta_1 \\ \partial L(\theta_2^0)/\partial \theta_2 \end{bmatrix} \Rightarrow \theta^1 = \theta^0 - \eta \nabla L(\theta^0)$$

$$\begin{bmatrix} \theta_1^2 \\ \theta_2^2 \end{bmatrix} = \begin{bmatrix} \theta_1^1 \\ \theta_2^1 \end{bmatrix} - \eta \begin{bmatrix} \partial L(\theta_1^1)/\partial \theta_1 \\ \partial L(\theta_2^1)/\partial \theta_2 \end{bmatrix} \Rightarrow \theta^2 = \theta^1 - \eta \nabla L(\theta^1)$$

Created with EverCam.
<http://www.camdemy.com>

Gradient :vector

Visualization



Start at position θ^0

Compute gradient at θ^0

Move to $\theta^1 = \theta^0 - \eta \nabla C(\theta^0)$

Compute gradient at θ^1

Move to $\theta^2 = \theta^1 - \eta \nabla C(\theta^1)$

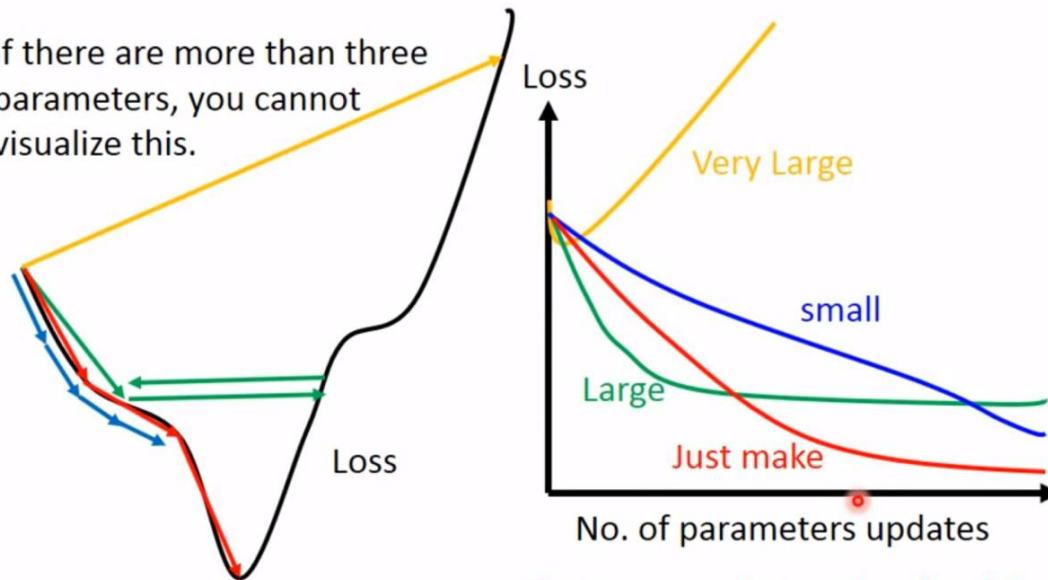
Gradient: vector, 红色箭头, 更新参数

Learning Rate

$$\theta^i = \theta^{i-1} - \eta \nabla C(\theta^{i-1})$$

Set the learning rate η carefully

If there are more than three parameters, you cannot visualize this.



But you can always visualize this.

Created with EverCam
<http://www.camdemy.com>

- E.g. 1/t decay: $\eta^t = \eta / \sqrt{t + 1}$

- Learning rate cannot be one-size-fits-all
 - Giving different parameters different learning rates

Adagrad

$$w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t$$

σ^t : *root mean square* of the previous derivatives of parameter w

Created with EverCam.
<http://www.camdemy.com>

$$\eta^t = \frac{\eta}{\sqrt{t+1}} \quad g^t = \frac{\partial \mathcal{C}(\theta^t)}{\partial w}$$

- Divide the learning rate of each parameter by the **root mean square of its previous derivatives**

The diagram illustrates the derivation of the RMS learning rate formula. It starts with a general update rule for a parameter w at time $t+1$:

$$w^{t+1} \leftarrow w^t - \frac{\eta^t}{\sigma^t} g^t$$

Here, η^t is the learning rate at time t , and σ^t is the root mean square of its previous derivatives. A blue arrow points from this equation to the simplified version below.

The learning rate η^t is defined as:

$$\eta^t = \frac{\eta}{\sqrt{t+1}} \quad \text{1/t decay}$$

The root mean square σ^t is defined as:

$$\sigma^t = \sqrt{\frac{1}{t+1} \sum_{i=0}^t (g^i)^2}$$

Substituting these into the update rule, we get the final simplified formula:

$$w^{t+1} \leftarrow w^t - \frac{\eta}{\sqrt{\sum_{i=0}^t (g^i)^2}} g^t$$

THERE ARE MANY OTHER METHODS...

Contradiction? $\eta^t = \frac{\eta}{\sqrt{t+1}}$ $g^t = \frac{\partial C(\theta^t)}{\partial w}$

Vanilla Gradient descent

$$w^{t+1} \leftarrow w^t - \eta^t \underline{g^t} \longrightarrow \text{Larger gradient, larger step}$$

Adagrad

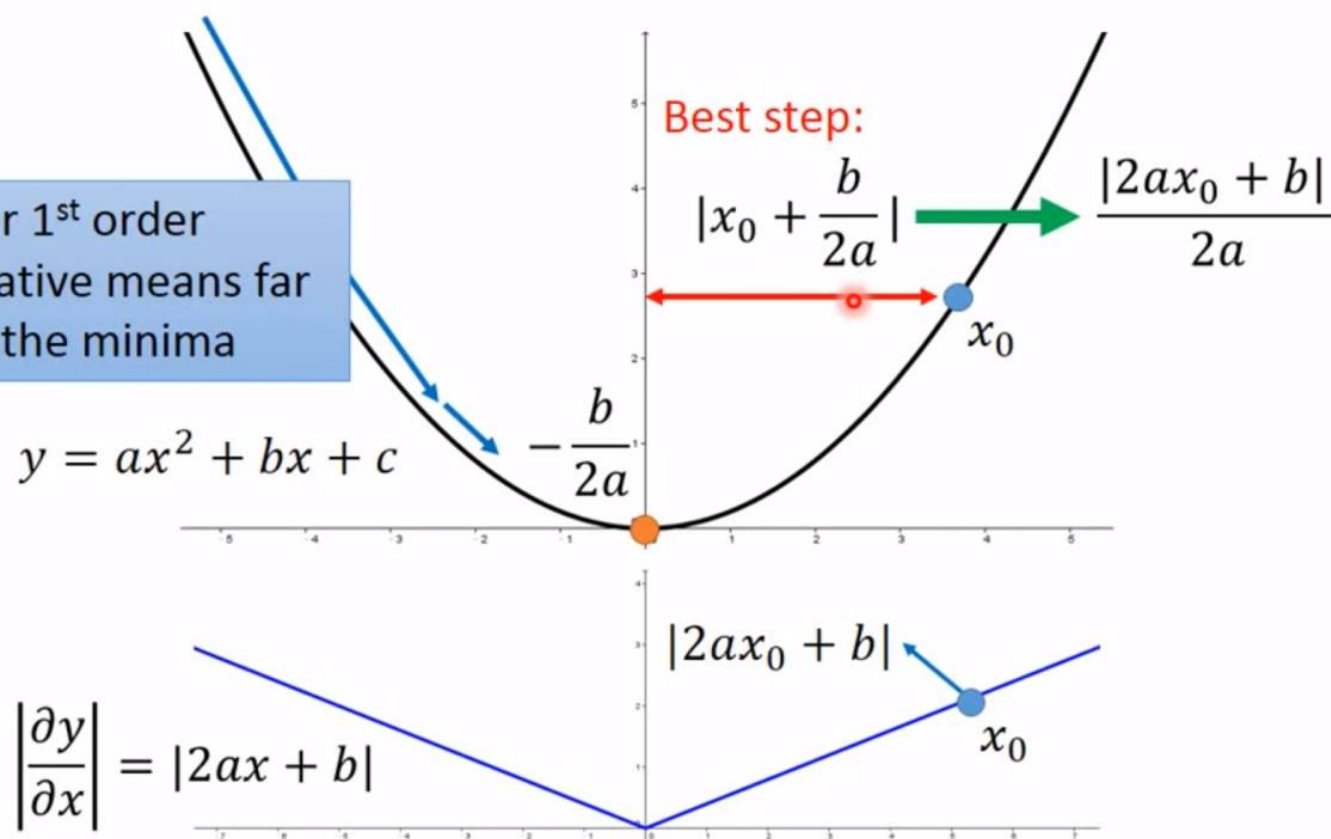
$$w^{t+1} \leftarrow w^t - \frac{\eta}{\sqrt{\sum_{i=0}^t (g^i)^2}} \underline{g^t}$$

Larger gradient, larger step

Larger gradient, smaller step

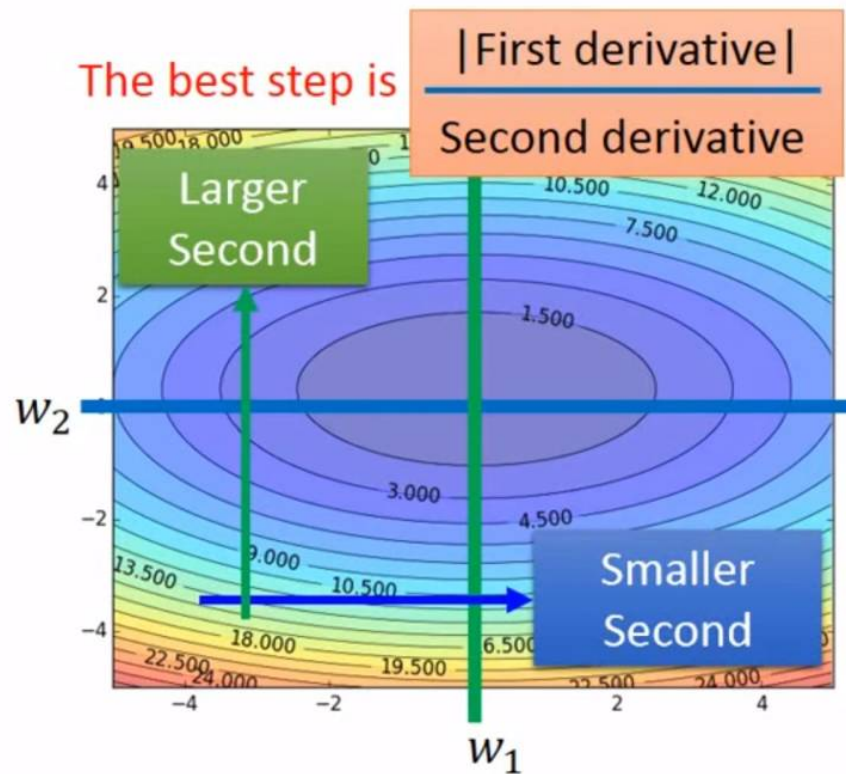
Larger gradient, larger steps?

Larger 1st order derivative means far from the minima



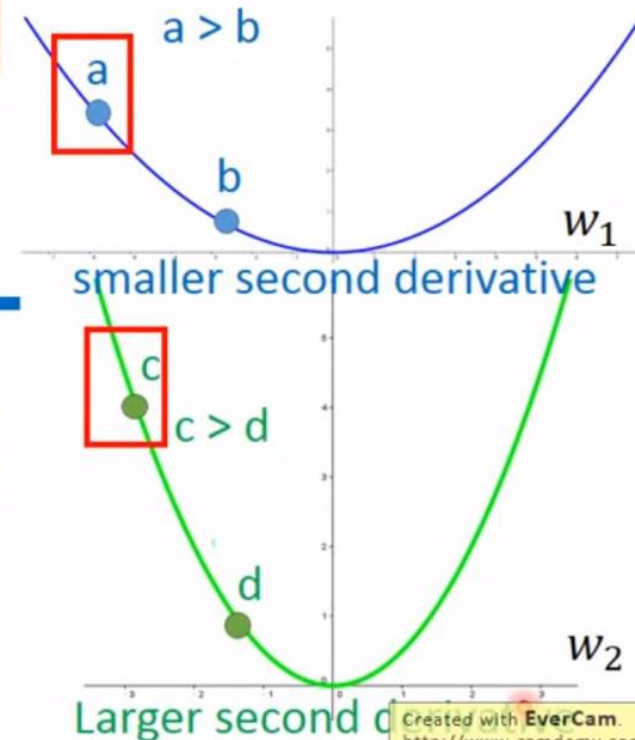
只考虑一个参数时：
距离与微分的大小成
正比

Comparison between different parameters



~~Larger 1st order derivative means far from the minima~~

Do not cross parameters



Stochastic Gradient Descent (SGD)

——faster

$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i^n \right) \right)^2$$

Loss is the summation over all training examples

◆ **Gradient Descent** $\theta^i = \theta^{i-1} - \eta \nabla L(\theta^{i-1})$

◆ **Stochastic Gradient Descent**

Pick an example x^n

$$L^n = \left(\hat{y}^n - \left(b + \sum w_i x_i^n \right) \right)^2 \quad \theta^i = \theta^{i-1} - \eta \nabla L^n(\theta^{i-1})$$

Loss for only one example

Created with EverCam.
<http://www.camdemy.com>

看完所有example，更新参数；每看到一个example，update一次参数，更新20次参数。

Feature Scaling

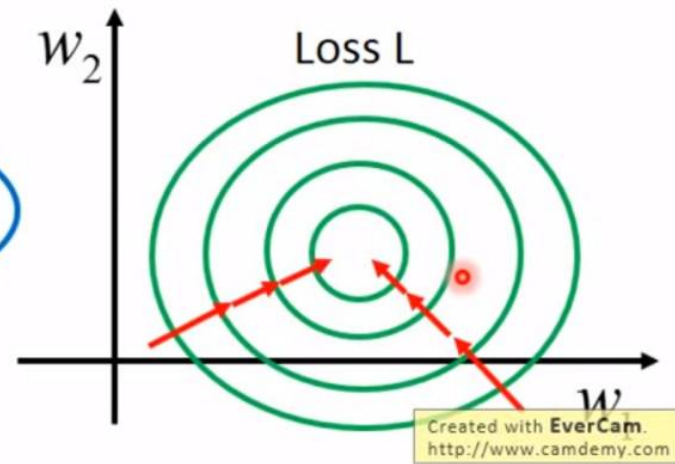
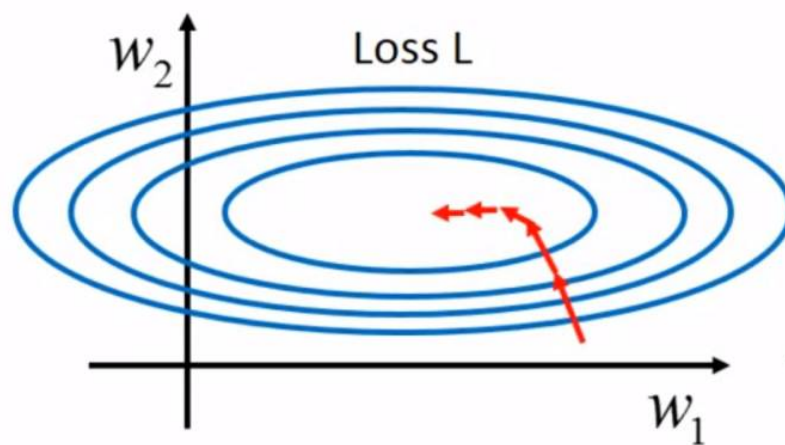
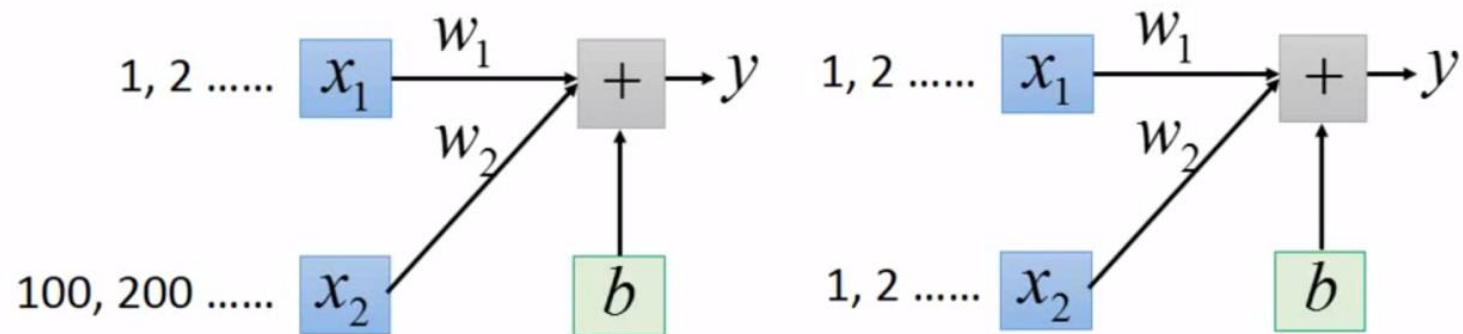
$$x_i^r \leftarrow \frac{x_i^r - m_i}{\sigma_i}$$

The means of all dimensions are 0,
and the variances are all 1

For each
dimension i:

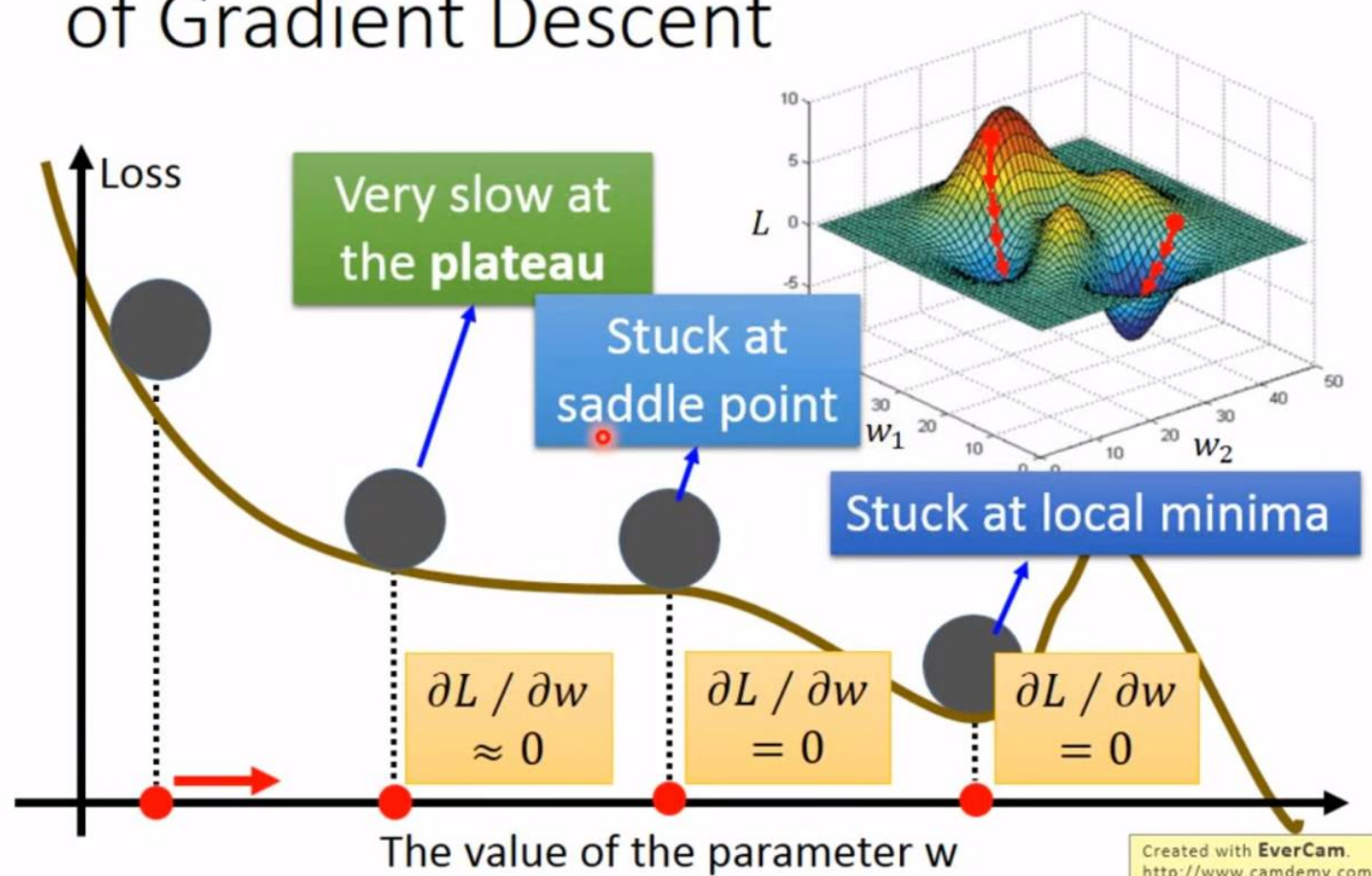
mean: m_i

standard
deviation: σ_i



Gradient Descent Theory — Taylor Series

More Limitation of Gradient Descent



Thanks!