PyTorch

什么是梯度

主讲人：龙良曲

Clarification

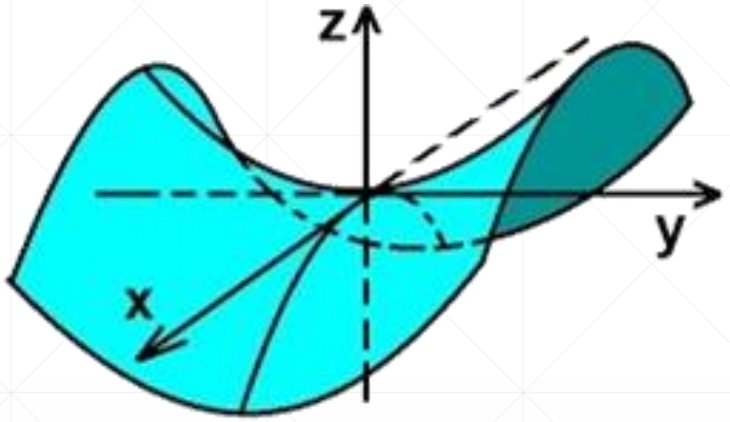
- 导数, derivate
- 偏微分, partial derivate
- 梯度, gradient

$$\nabla f = \left(\frac{\partial f}{\partial x_1}; \frac{\partial f}{\partial x_2}; \ldots; \frac{\partial f}{\partial x_n} \right)$$

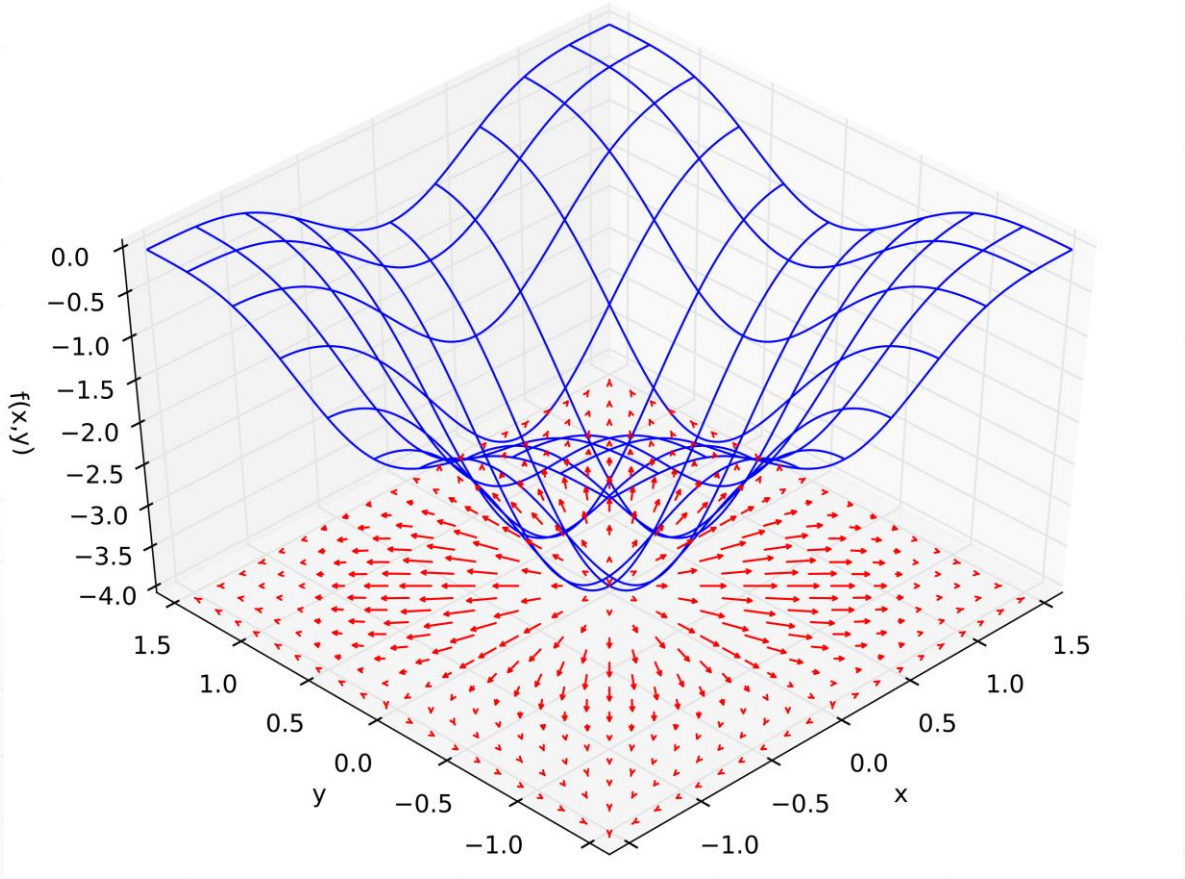
$$z = y^2 - x^2$$

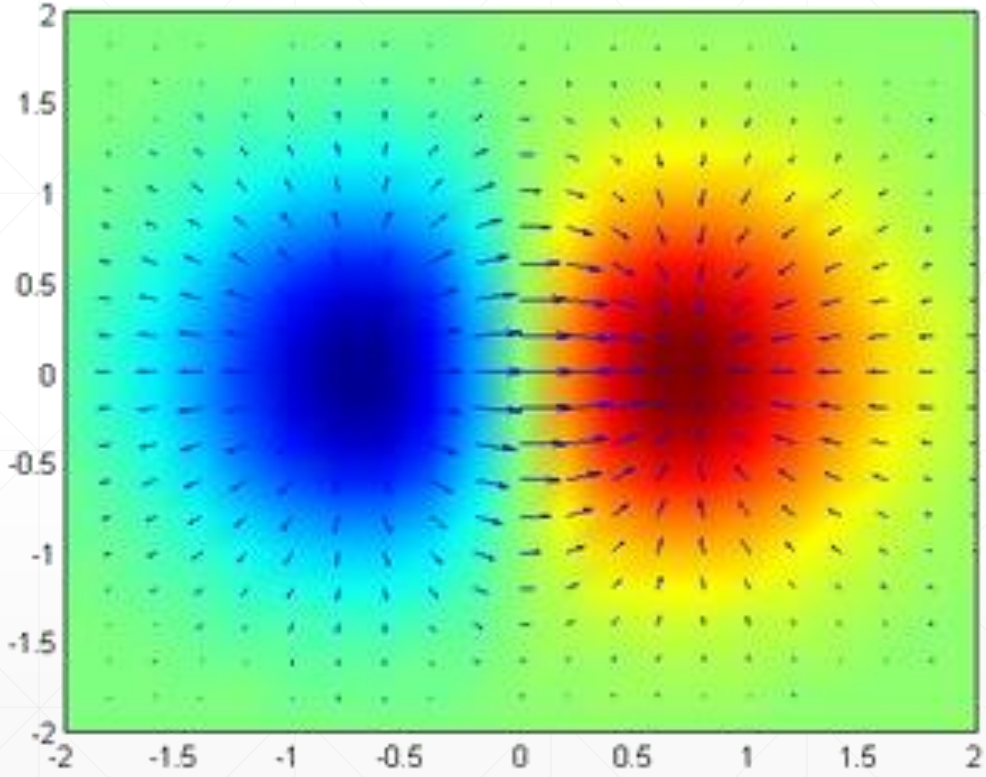
$$\frac{\partial z}{\partial x} = -2x$$

$$\frac{\partial z}{\partial y} = 2y$$



What does grad mean?





How to search for minima?

$$\theta_{t+1} = \theta_t - \alpha_t \nabla f(\theta_t)\,.$$

Function:

$$J(\theta_1, \theta_2) = {\theta_1}^2 + {\theta_2}^2$$

Objective:

$$\min_{\theta_1, \theta_2} J(\theta_1, \theta_2)$$

Update rules:

$$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1, \theta_2)$$

$$\theta_2 := \theta_2 - \alpha \frac{d}{d\theta_2} J(\theta_1, \theta_2)$$

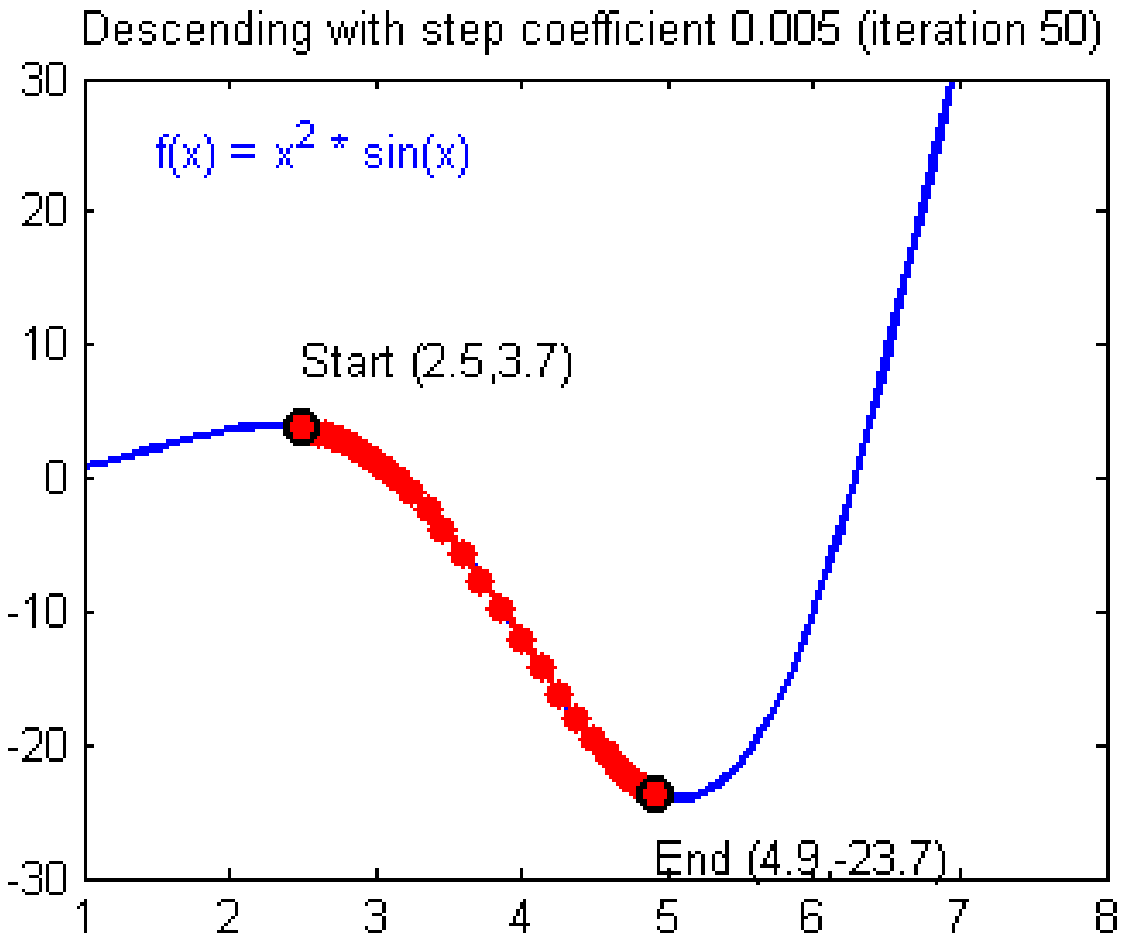
Derivatives:

$$\frac{d}{d\theta_1} J(\theta_1, \theta_2) = \frac{d}{d\theta_1} {\theta_1}^2 + \frac{d}{d\theta_1} {\theta_2}^2 = 2\theta_1$$

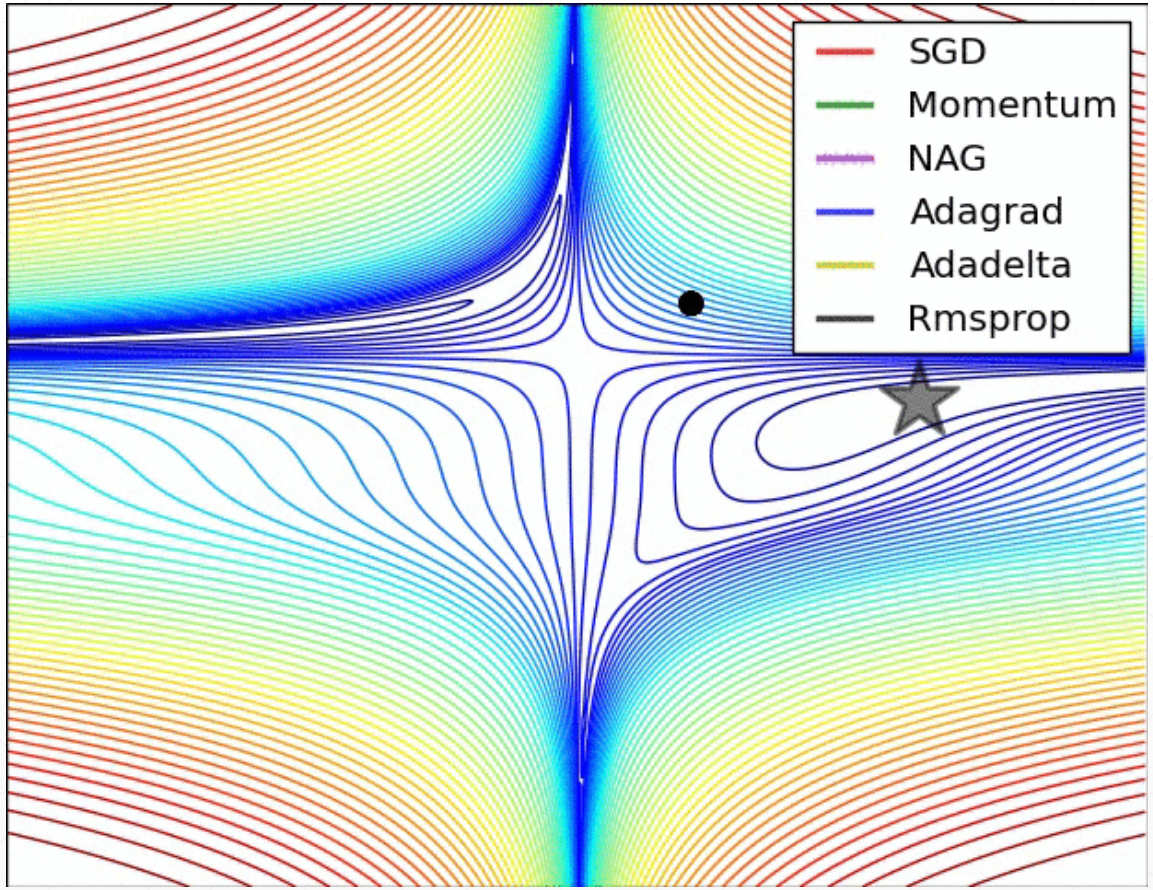
$$\frac{d}{d\theta_2} J(\theta_1, \theta_2) = \frac{d}{d\theta_2} {\theta_1}^2 + \frac{d}{d\theta_2} {\theta_2}^2 = 2\theta_2$$

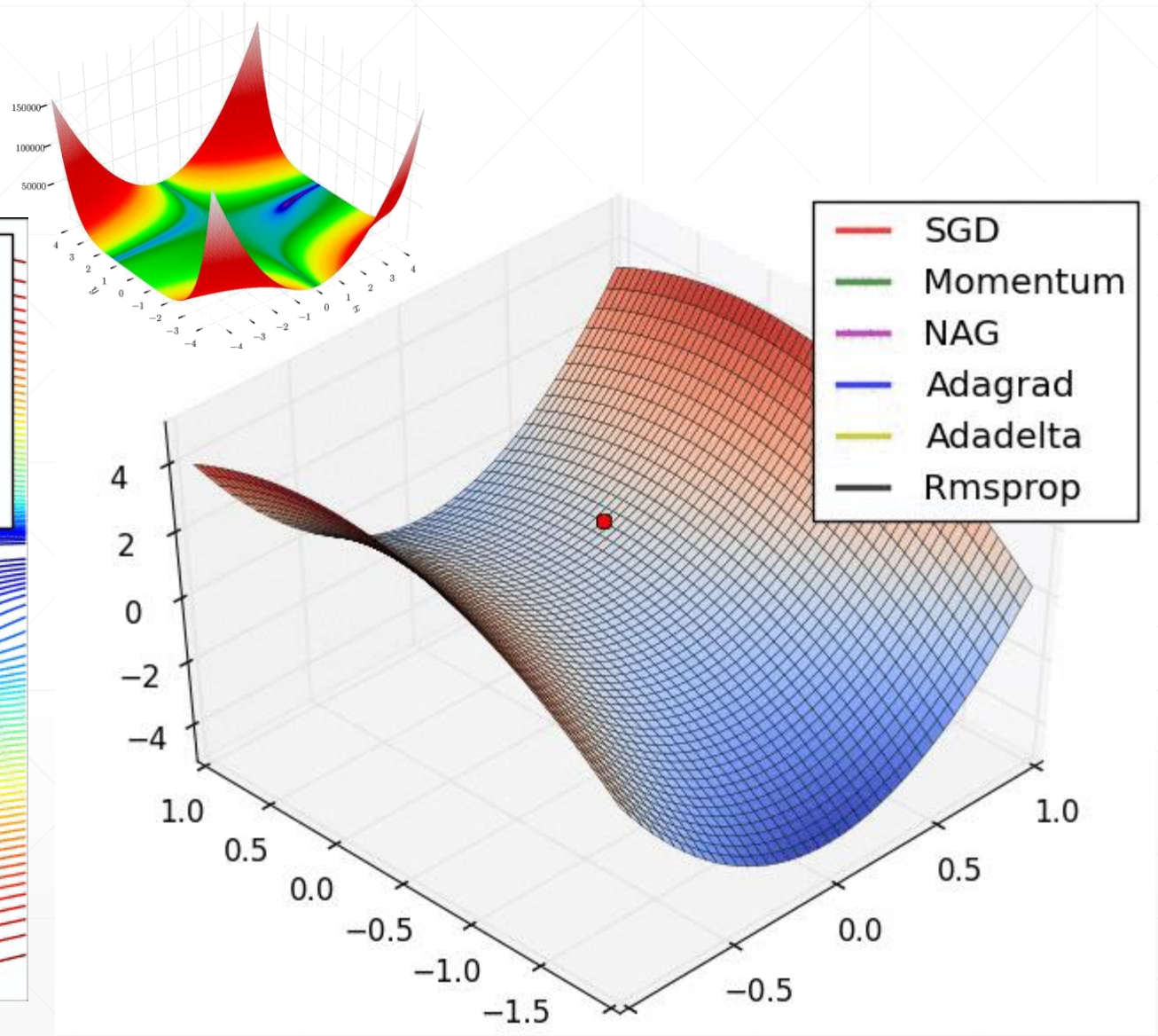
http://mccormickml.com/2014/03/04/gradient-descent-derivation/

Learning process-1



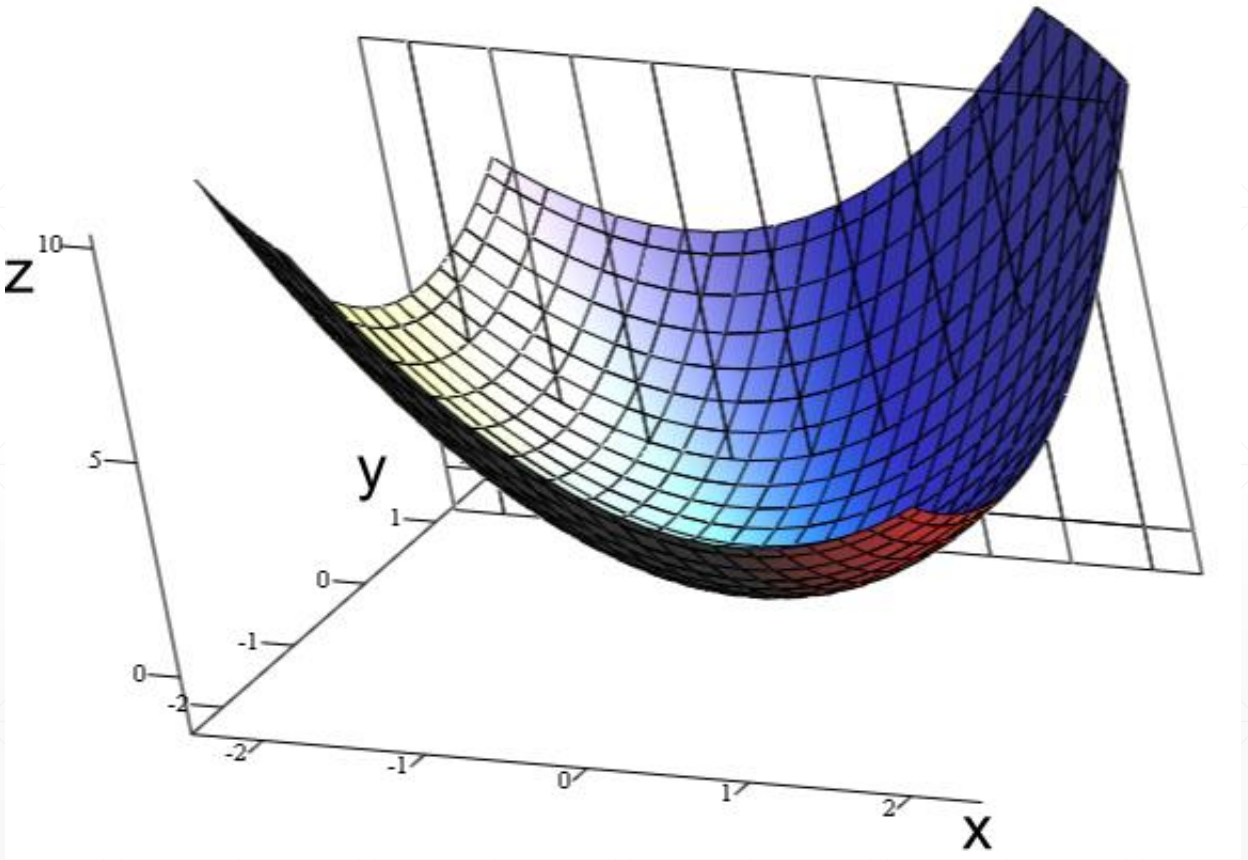
Learning process-2



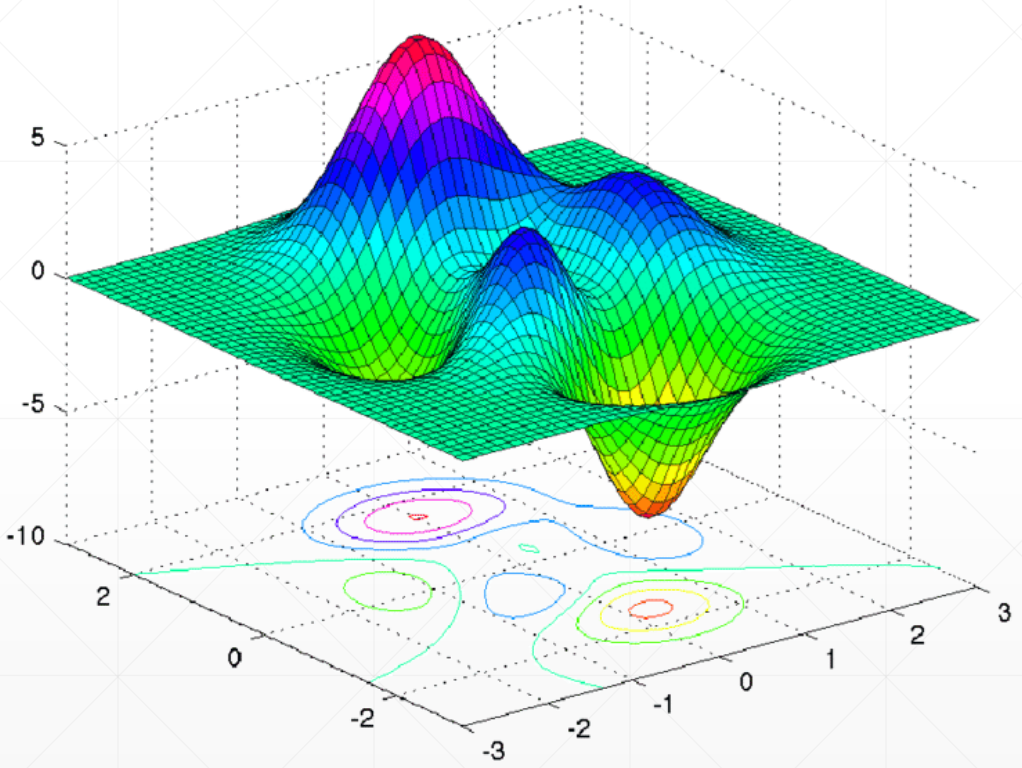


http://ruder.io/optimizing-gradient-descent/

Convex function



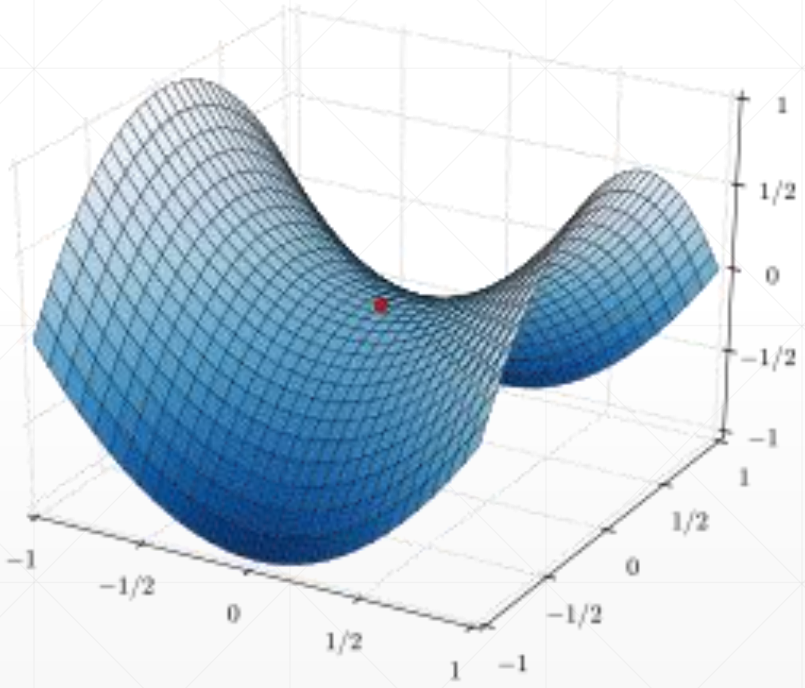
Local Minima

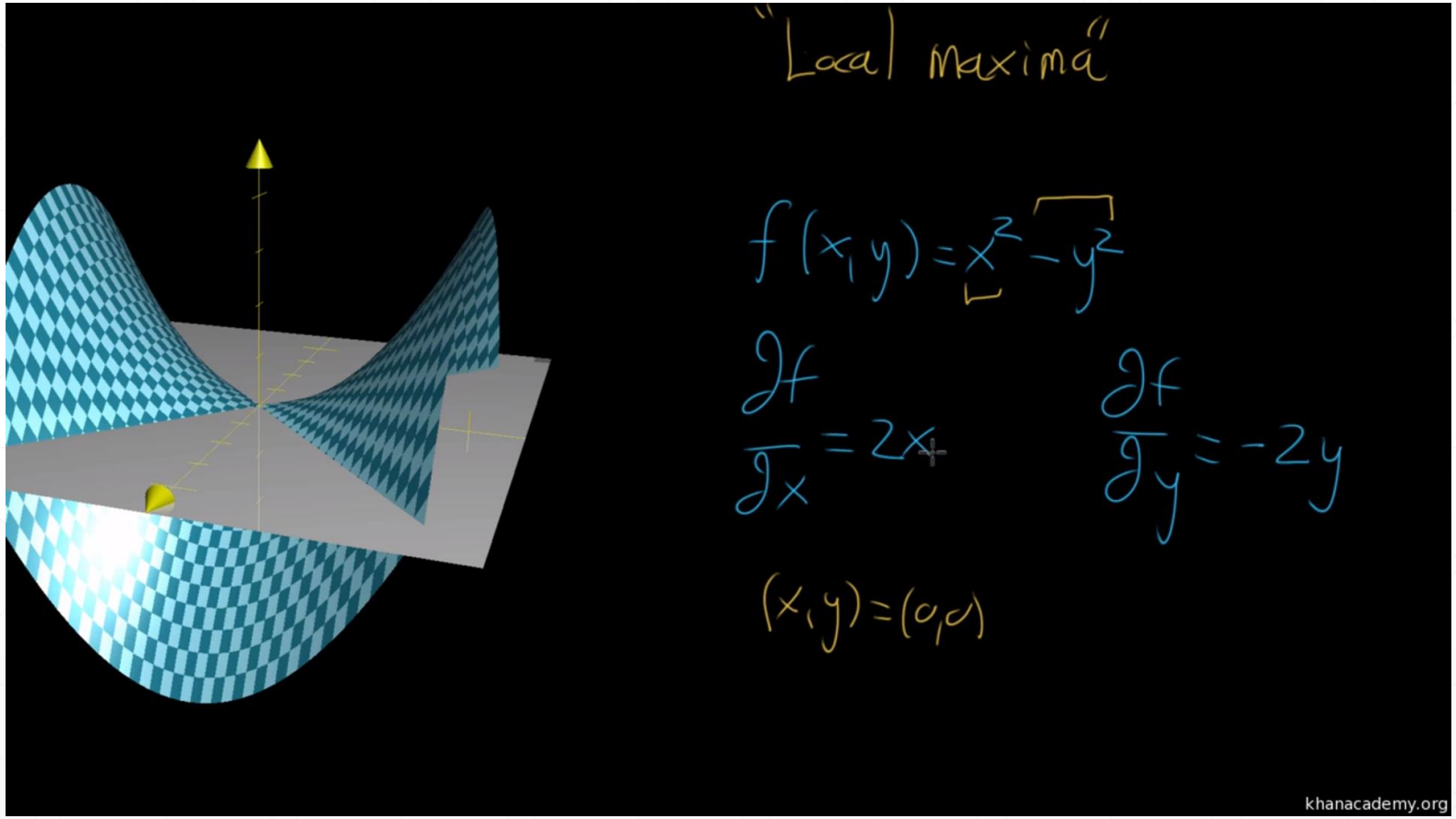


ResNet-56

https://github.com/tomgoldstein/loss-landscape

Saddle point



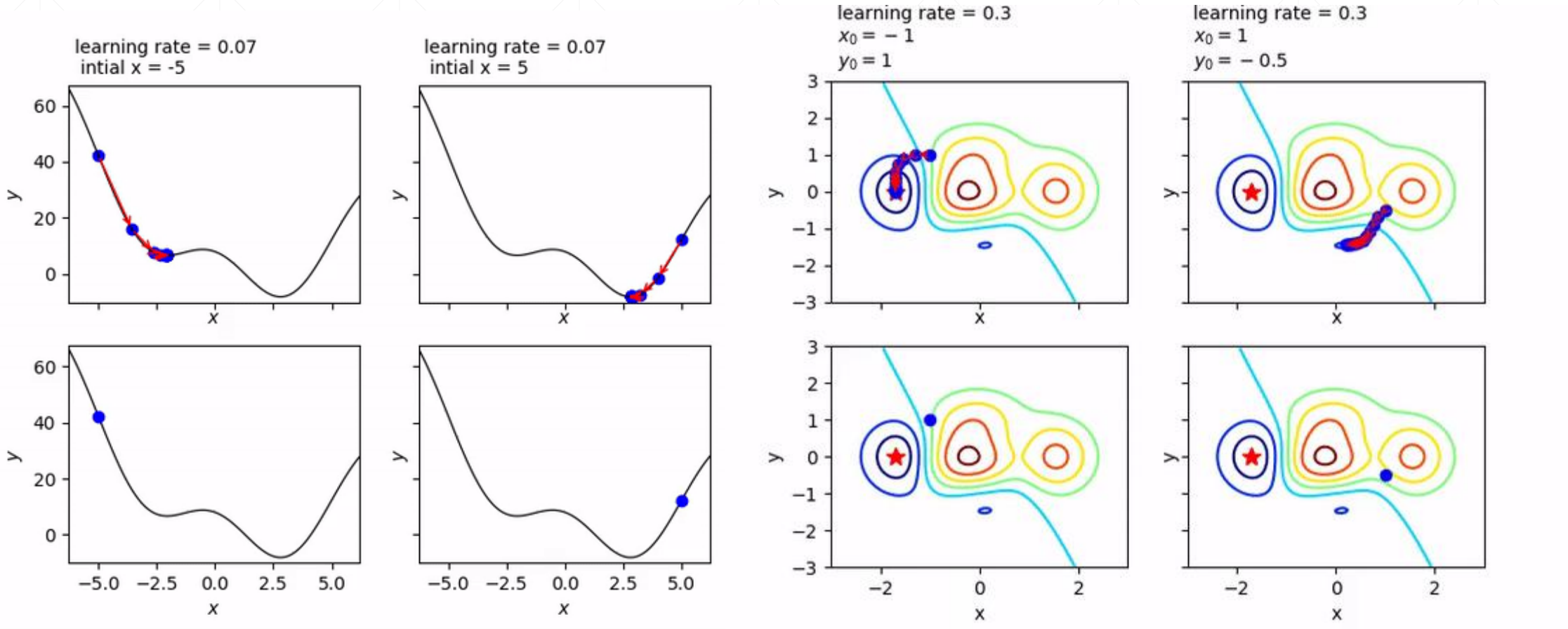


https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/optimizing-multivariable-functions-videos/v/saddle-points

Optimizer Performance

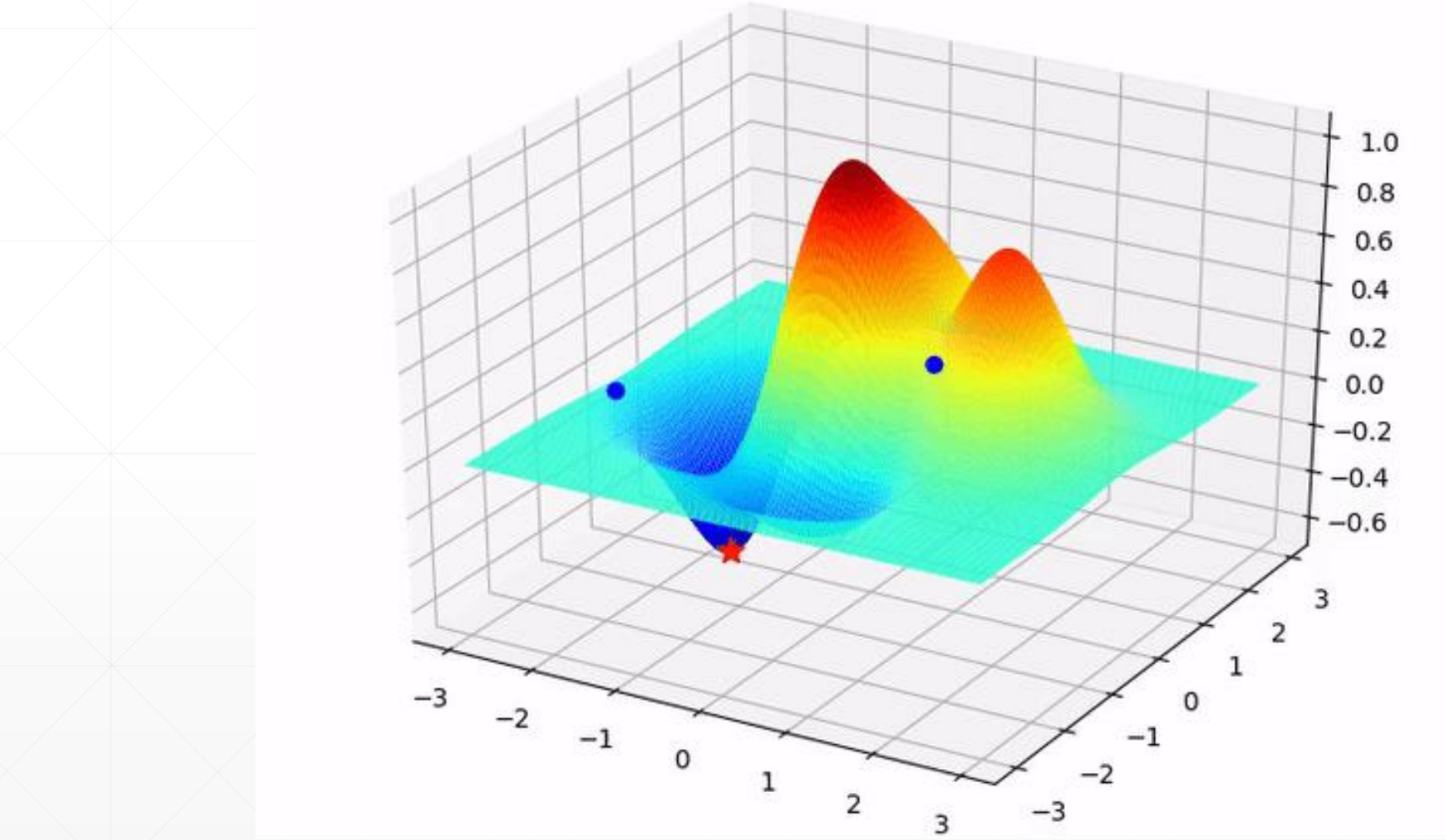
- initialization status
- learning rate
- momentum
- etc.

Initialization

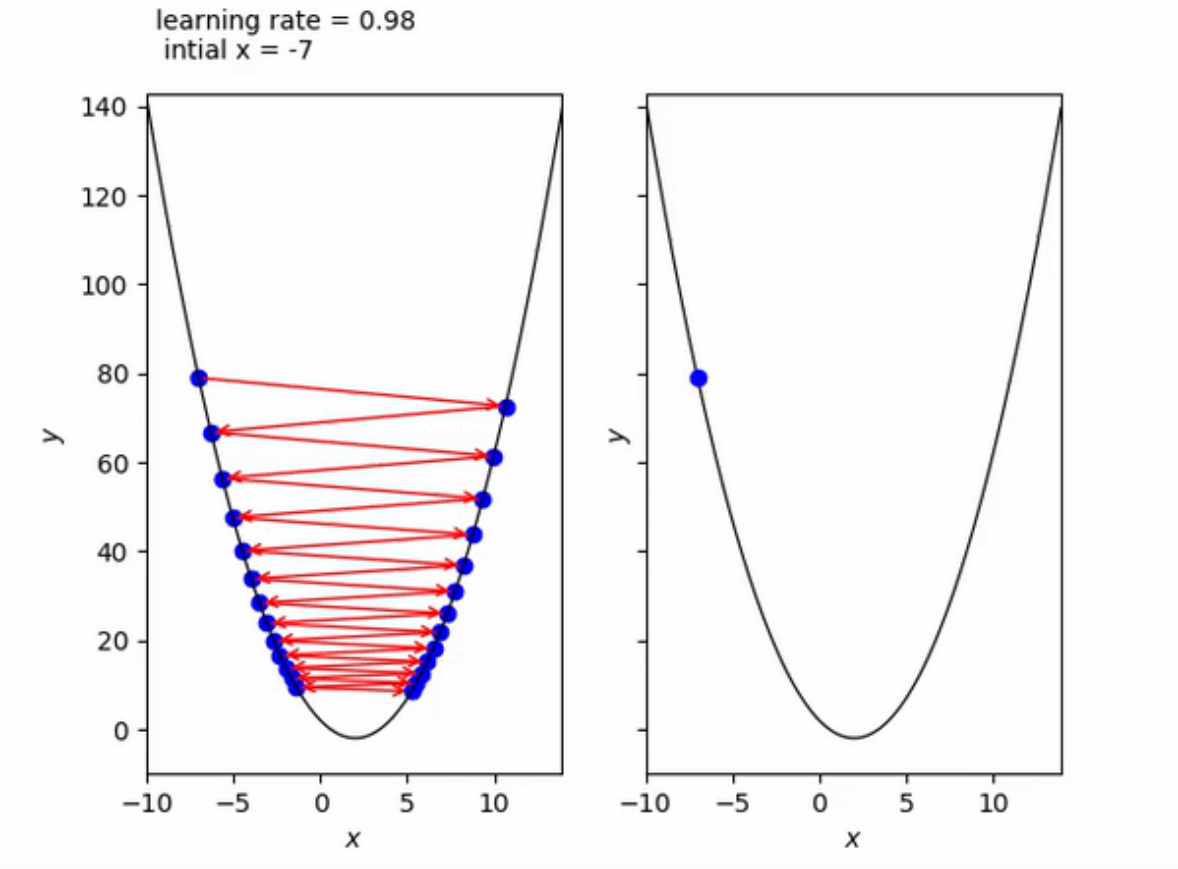


https://jed-ai.github.io/opt2_gradient_descent_1/

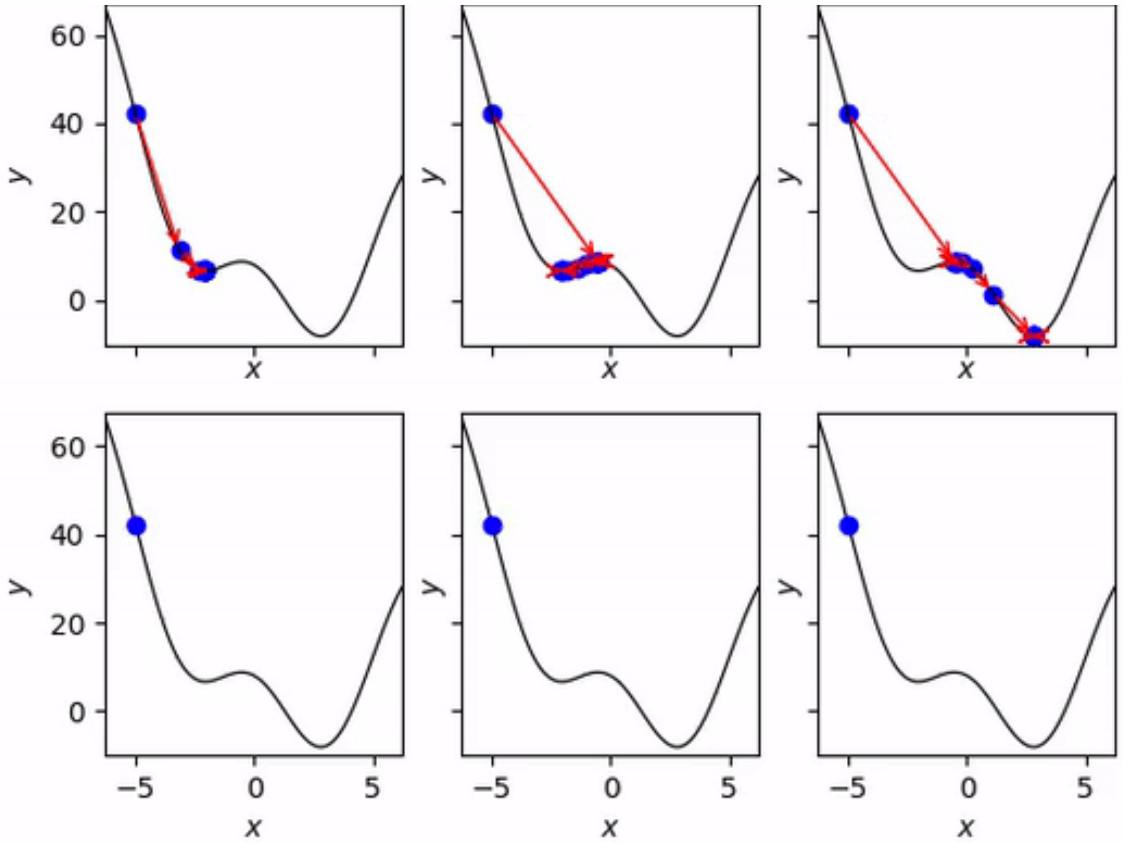
Initialization



Learning rate



Escape minima



下一课时

常见函数梯度

Thank You.