权重的初始化

最原始的初始化：全部为固定值

$$W_{ij} = 0.1$$

稍好些的初始化：服从固定方差的独立高斯分布

$$W \sim G(0, \alpha^2)$$

Xavier初始化：服从动态方差的独立高斯分布

$$W \sim G(0, \sqrt{\frac{1}{n_{in}}}^2)$$

MSRA初始化：服从动态方差的独立高斯分布

$$W \sim G(0, \sqrt{\frac{2}{n_{in}}}^2)$$

optimizer

- GD（Gradient Descent）
  - 使用全部数据计算梯度

$$w = w - \eta \frac{1}{m} \sum_{i=1}^{m} \Delta w_i$$

- SGD（Stochastic Gradient Descent）
  - 使用一条数据计算梯度，或者
  - 使用batch_size条数据

- Momentum SGD

$$m_t = \mu * m_{t-1} + \eta \Delta w$$

$$w = w - m_t$$

- Nesterov Momentum

$$m_t = \mu * m_{t-1} + \eta \Delta w(w - \mu * m_{t-1})$$

$$w = w - m_t$$

- RMSprop

$$E[(\Delta w)^2]_t = 0.9E[(\Delta w)^2]_{t-1} + 0.1(\Delta w)_t^2$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{E[(\Delta w)^2]_t + \epsilon}} \odot \Delta w_t$$

$$\eta_r = 1e^{-3}$$

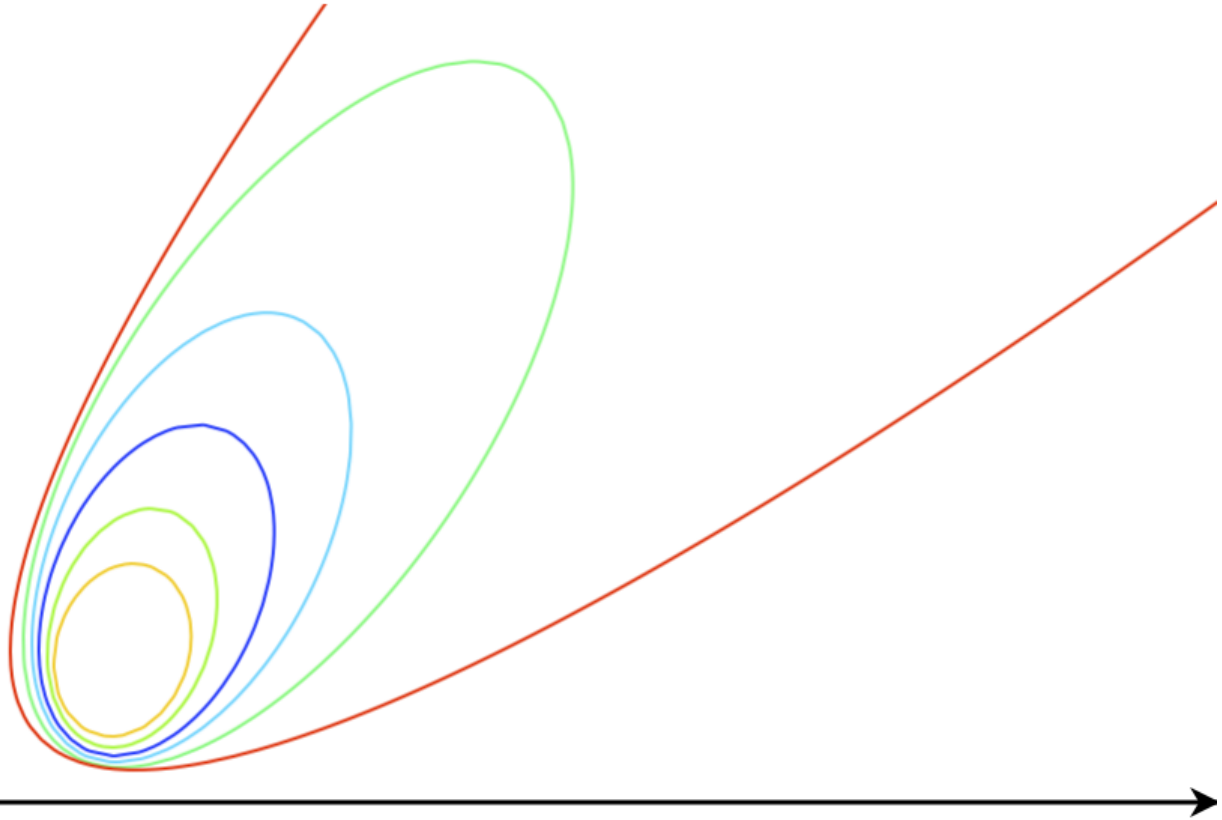
- Adam

$$m_t = \frac{\beta_1 m_{t-1} + (1 - \beta_1)\Delta w_t}{1 - \beta_1^t}$$

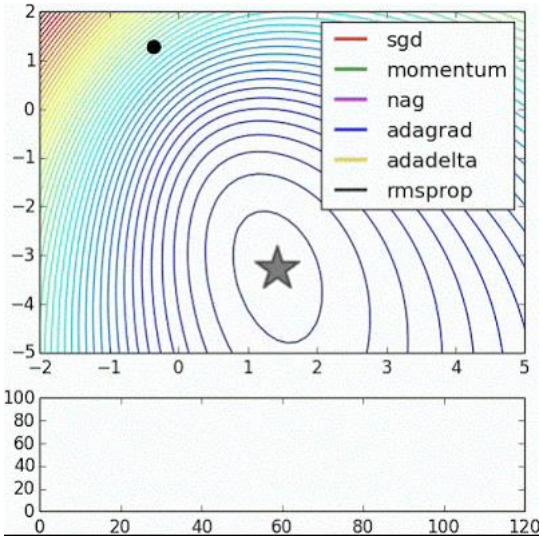
$$v_t = \frac{\beta_2 v_{t-1} + (1 - \beta_2)\Delta w_t^2}{1 - \beta_2^t}$$

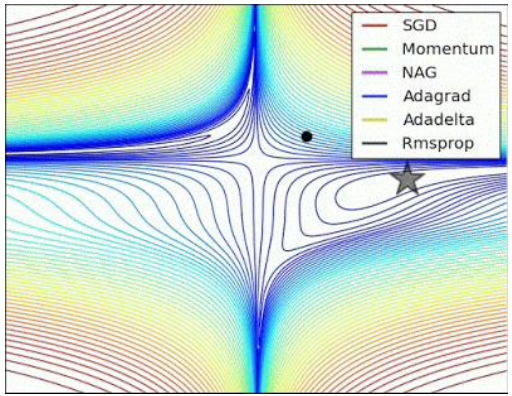
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t} + \epsilon} m_t$$

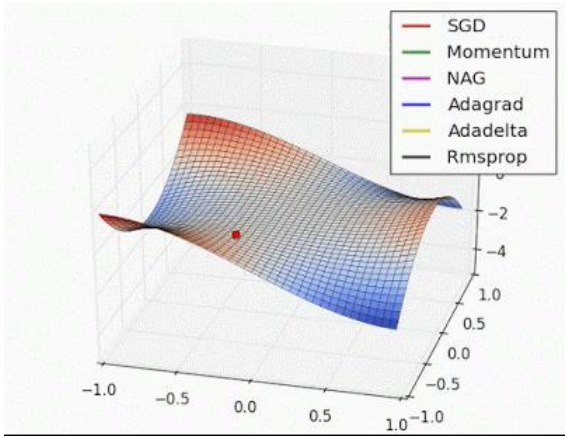
$$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$$

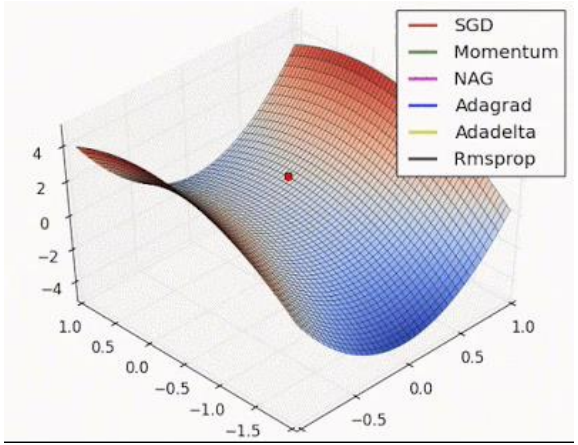


optimizer









http://www.denizyuret.com/2015/03/alec-radfords-animations-for.html

Learning rate decay

- 我是谁？
  为什么要衰减
- 我从哪里来？
  什么时机衰减
  - 通常是loss走平/震荡时
  - 或者一直衰减
- 我要到哪里去？
  衰减到多少
  - 1/10衰减
  - 1/3衰减
  - 0.94/0.87/0.74/0.575