

Оптимизация - алгоритмы, использующие для решения. Использование нап-мов с учетом ускорение расчетов можно.

Adagrad - оптимизатор, использует скорость обучения  $\eta$  для каждого нап-ма на каждом шаге. Рассматриваем для функ.  $\varphi$ -уши ошибки.

$$g_t = \nabla_{\theta} \varphi(\theta_t) \text{ - прямой ф-зии}$$

$$\theta_{t+1} = \theta_t + g_t^2$$

$$\theta_{t+1} = \theta_t - \frac{1}{\sqrt{\theta_t^2}} \cdot g_t \text{ - обновление нап-мов}$$

$\eta$  - скорость обучения, как. изменение гл. зиг. нап-ра  $\theta$  в данный момент времени на основе предыдущих угадываний, рассчитанного нап-ма.

Adadelta - расширение Adagrad. Ограничено размером промежутка угадываний по нек. фикс. размеру, вместо него, гладко

хранить все фс. Используем эксп.

смогущая среднее, а не сумма градиентов.

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2$$

$$RMS[g]_t = \sqrt{E[g^2]_t + \epsilon}$$

$$\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} \cdot g_t$$

Adam - работает с импульсами 1 и 2 порядков. Тогда учитывается в качестве скорости во внесение просматриваемых импульсов. В дополнение к хранению экспоненц. смогущих средних (или Babbelts) сохр. экспоненц. смогущ. среднее.

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

Характеристики  
Gradient Descent

$$\theta = \theta - \alpha \nabla J(\theta)$$

Регулирование

Минимум в лок. мин.  
Быстро сход. мал. шаг  
но если шаг. слишком  
шаг не сойдет

Stochastic Gradient Descent

$$\theta = \theta - \eta \nabla J(\theta; x_i, y_i)$$

Быстр. градиент разн. шаг  
один шагом на весь шаг  
использовать небольшое значение  
шага для сходимости

Mini-Batch Gradient Descent

$$\theta = \theta - \eta \nabla J(\theta; \mathcal{B}_j)$$

Минимум в лок. мин.  
При небольш. батчах шаг. одн. шаг  
некотор. итерациях. где шаг  
шаг. сокр.

Adagrad

$$g_t = \nabla_{\theta} J(\theta_t)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\sum_{i=1}^t g_i^2}} \cdot g_t$$

Быстр. градиент (стаб. шаг).  
Зависит нап. от а)  
Несколько медленно

Adadelta

$$RMS[g]_t = \sqrt{\mathbb{E}[g_t^2]^{\frac{1}{2}}}$$

Быстр. градиент

$$\theta_{t+1} = \theta_t - \frac{\eta}{RMS[g]_t} \cdot g_t$$

RMSProp

$$\theta_{t+1} = \theta_t - \frac{\eta}{RMS[g]_t} \cdot g_t$$

Быстр. градиент

Adam

$$m_t = \frac{m_{t-1}}{1-\beta_1} + \eta g_t = \frac{\eta g_t}{1-\beta_1}$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\frac{\eta^2}{1-\beta_1^2} + \epsilon}} \cdot m_t$$

Быстр. градиент

Nesterov Accelerated Gradient

$$\theta = \gamma \cdot \nabla \theta + \eta \nabla J(\theta - \gamma \nabla \theta)$$
$$\theta \leftarrow \theta - \nabla \theta$$

Низкая сход. и  
однородность