

Введение в АД

Лекция 8

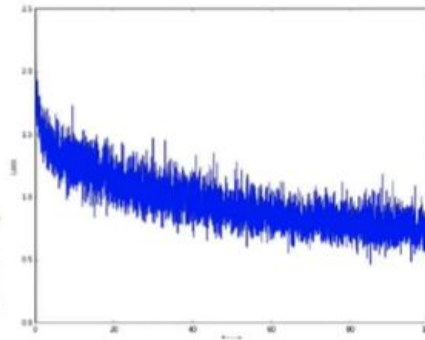
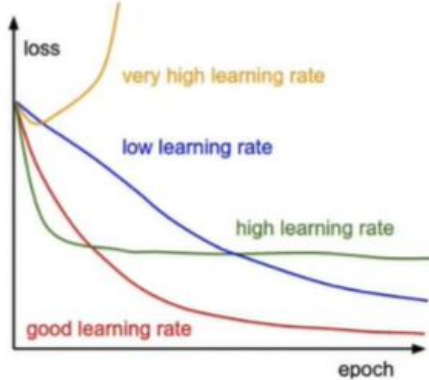
Оптимизация и регуляризация в DL



SGD: стохастический градиентный спуск

Stochastic gradient descent is used to optimize NN parameters.

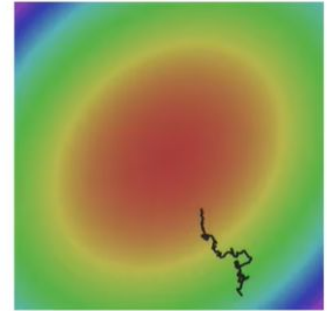
$$x_{t+1} = x_t - \text{learning rate} \cdot dx$$



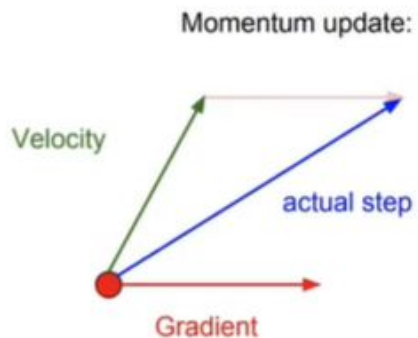
$$L(W) = \frac{1}{N} \sum_{i=1}^N L_i(x_i, y_i, W)$$

$$\nabla_W L(W) = \frac{1}{N} \sum_{i=1}^N \nabla_W L_i(x_i, y_i, W)$$

Averaging over mini batches => noisy gradient

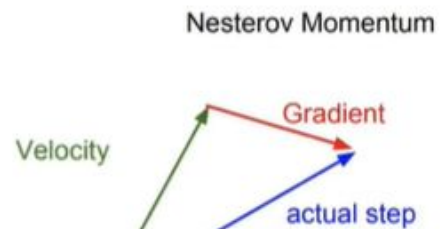


SGD с импульсом



$$v_{t+1} = \rho v_t + \nabla f(x_t)$$

$$x_{t+1} = x_t - \alpha v_{t+1}$$



$$v_{t+1} = \rho v_t - \alpha \nabla f(\boxed{x_t + \rho v_t})$$

$$x_{t+1} = x_t + v_{t+1}$$

Учет анизотропности

Adagrad: SGD with cache

$$\text{cache}_{t+1} = \text{cache}_t + (\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$



RMSProp: SGD with cache with exp. Smoothing

$$\text{cache}_{t+1} = \beta \text{cache}_t + (1 - \beta)(\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$

ADAM: комбинация двух идей

Let's combine the momentum idea and RMSProp normalization:

$$v_{t+1} = \gamma v_t + (1 - \gamma) \nabla f(x_t)$$

$$\text{cache}_{t+1} = \beta \text{cache}_t + (1 - \beta) (\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{v_{t+1}}{\text{cache}_{t+1} + \varepsilon}$$



L1/L2 регуляризация

Добавляем специальный член в функцию ошибок

$$L = \frac{1}{N} \sum_{i=1}^N \sum_{j \neq y_i} \max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1) + \lambda R(W)$$

Adding some extra term to the loss function.

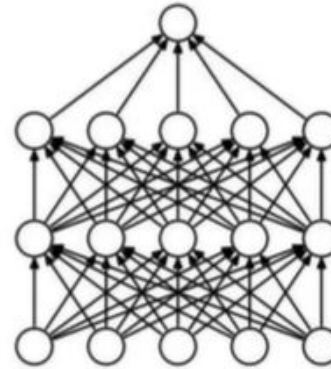
Common cases:

- L2 regularization: $R(W) = \|W\|_2^2$
- L1 regularization: $R(W) = \|W\|_1$
- Elastic Net (L1 + L2): $R(W) = \beta \|W\|_2^2 + \|W\|_1$

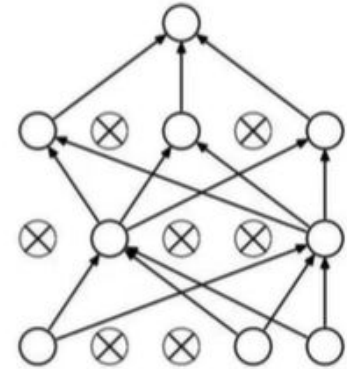


Dropout

На каждом шаге обучения случайно выключаем нейроны. Например с вероятностью 0.5 умножаем выход нейрона на 0. Проводим оптимизацию, потом на тестовой выборке включаем все нейроны.



(a) Standard Neural Net



(b) After applying dropout.



Data augmentation

Расширяем выборку путем каких-либо преобразований над исходными данными.

