@nbagiyan

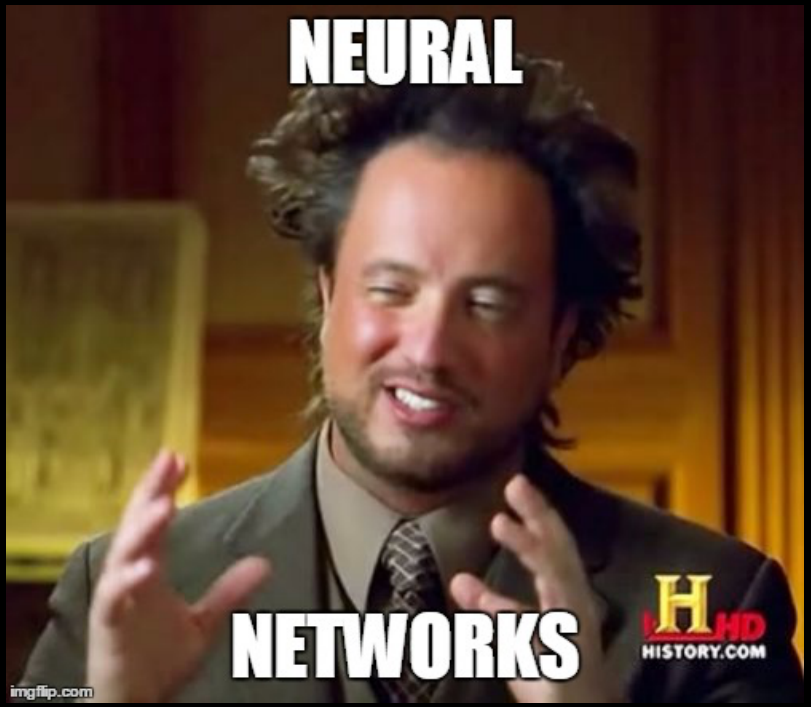
Нерсес

Deep learning ДПО ФКН

Занятие #1

Topics in this course:

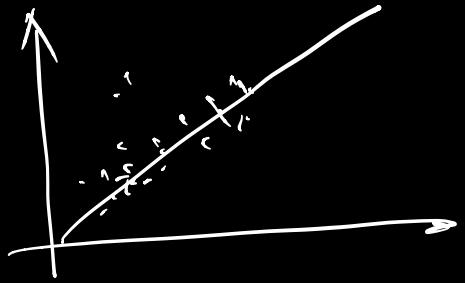
- Neural networks
- Neural networks
- Neural networks
- Neural networks



Recap: linear regression

$a(x) = \langle w, x \rangle$ $(x_1, \dots, x_d)$

$D = \{x_i, y_i\}_{i=1}^{\ell}$



$$L(a, y) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - a(x_i))^2 \to \min_w$$

MSE

$w - ?$

Recap: gradient descent

$$L(a, y) \to \min_w; \quad a(x) = \langle w, x \rangle = \langle w, x \rangle + b$$

$$x = (x_1, \ldots, x_d)$$

$$\uparrow$$

$$1$$

$$y = kx + b$$

Градиентный спуск:

$$w_{k+1} = w_k - \eta \nabla_w L$$

(антиградиент, градиент)

$$\nabla_w L = \left(\frac{\partial L}{\partial w_1}, \ldots, \frac{\partial L}{\partial w_d} \right)$$

Example:

$$L(y, a) = \frac{1}{e} \sum (y_i - a(x_i))^2 \to \min$$

$$\frac{1}{e} \sum_{i=1}^{e} (y_i - \langle w, x_i \rangle)^2$$

$$\frac{1}{e} \text{sum}((Y - Xw))^2.$$

$$\frac{1}{e} X^T (Y - Xw) = \frac{\partial L}{\partial w}$$

SGD

$\frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_w Q$ → градиент в каждом x

$\nabla_w L \approx \nabla_w Q_i$ — выбираем случайный индекс

$\nabla_w Q(w, x, y)$ → $(x_1, \dots, x_d)$

SGD:

$$w_{k+1} = w_k - \eta \nabla_w Q_i$$

SGD: Mini-Batches

$$\frac{1}{e}\sum_{i=1}^{e}\nabla_w q_i \approx \nabla_w q_{i\_rand} \approx \frac{1}{n}\sum_{i=1}^{n}\nabla_w q_i, \quad e \ll n$$

(e — количество выборок)

$$w_{k+1} = w_k - \frac{\eta}{n}\sum_{i=1}^{n}\nabla_w q_i$$

Вопрос:

$$w_{k+1} = w_k - \eta\sum_{i=1}^{n}\nabla_w q_i$$

```
dw = -2 * X_T[rand_ind]@(y[rand_int] - X[rand_int]@w)
```

SGD: Momentum

$L \to \min_w$

GD

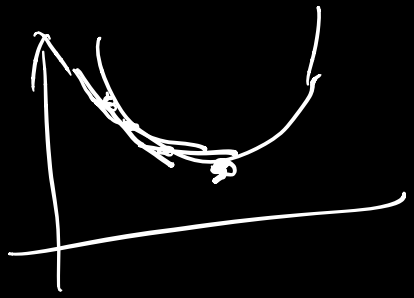
SGD

SGD - mini - batches

$m_0 = 0$

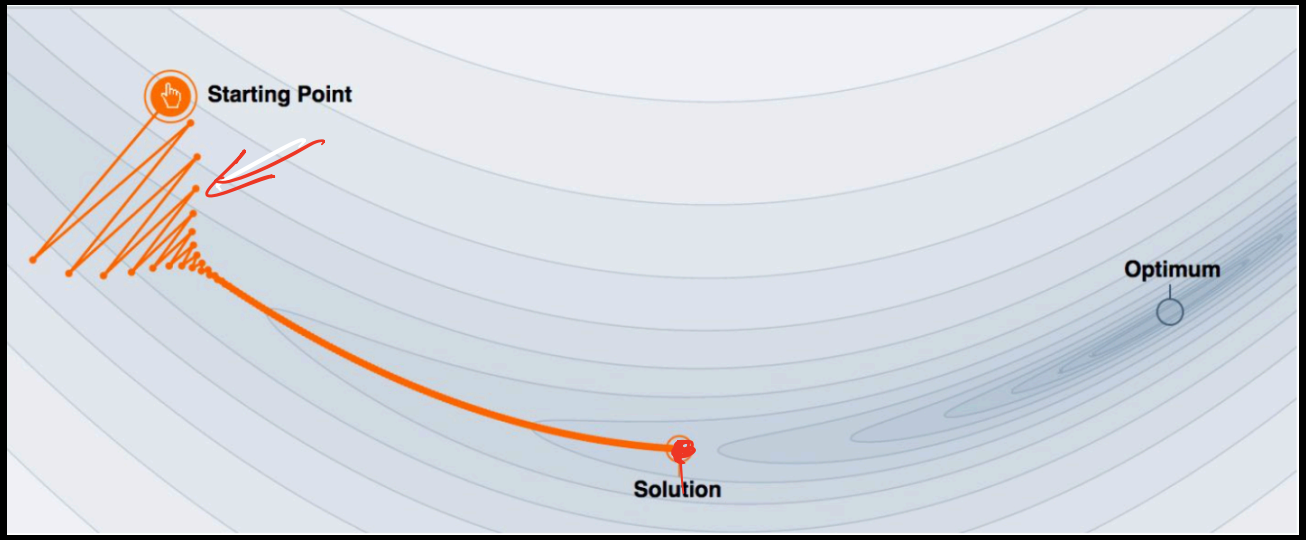
$m_{k+1} = \alpha m_k - \eta \nabla_w L$

$(1-\alpha)$

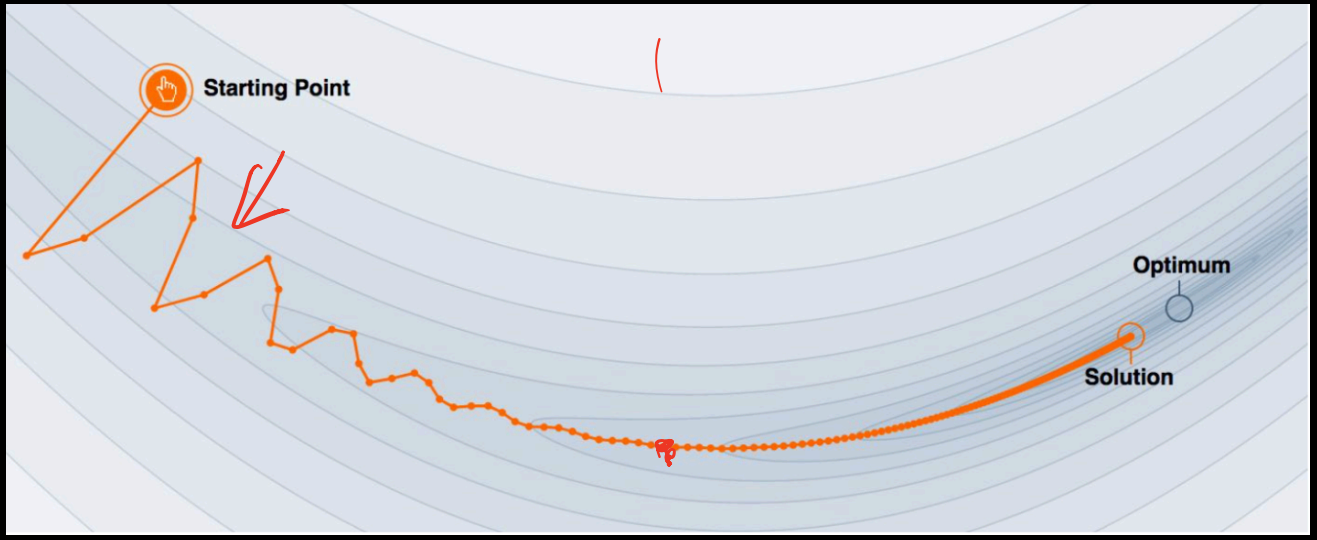


$w_{k+1} = w_k + m_{k+1}$

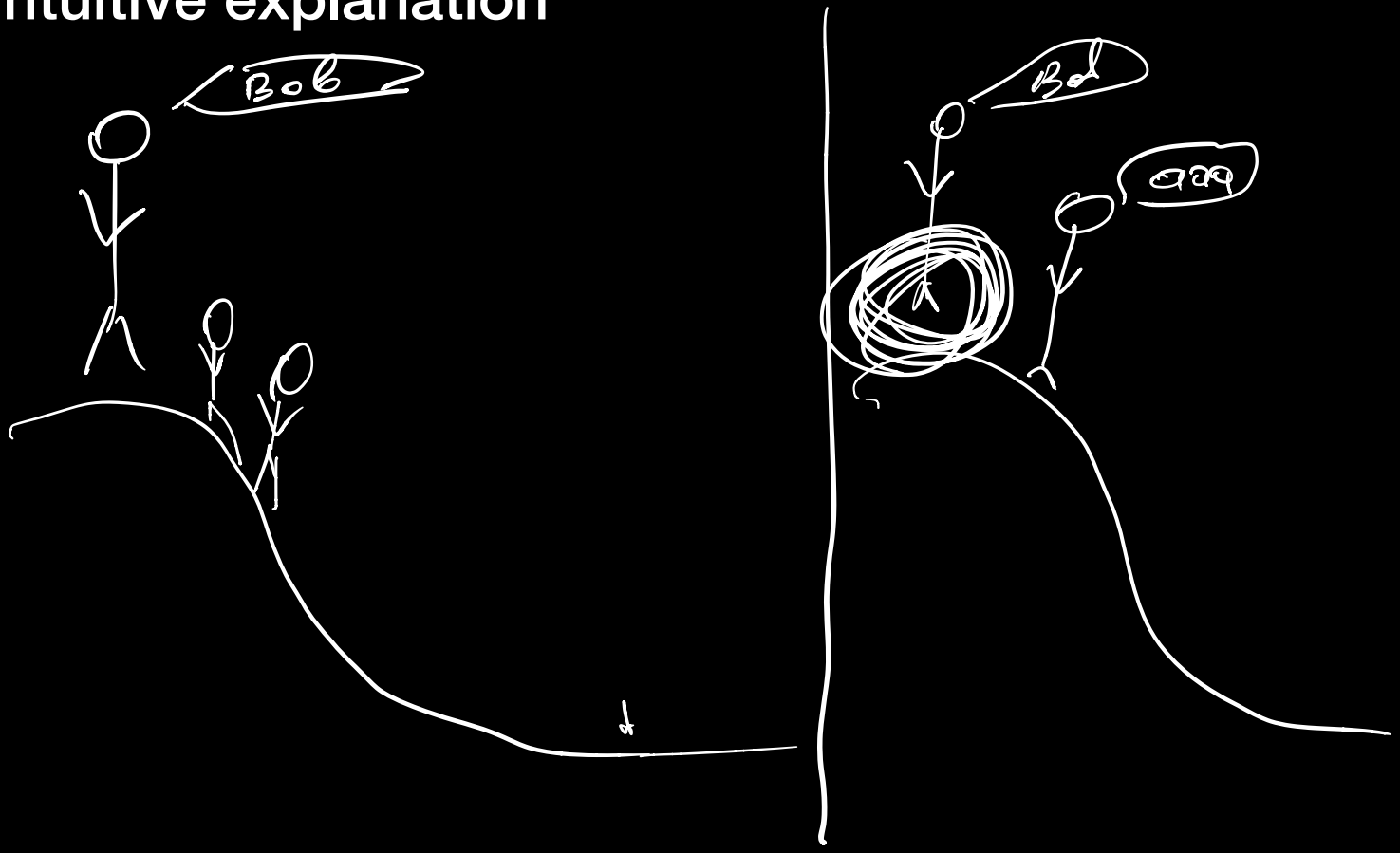
Example: No momentum



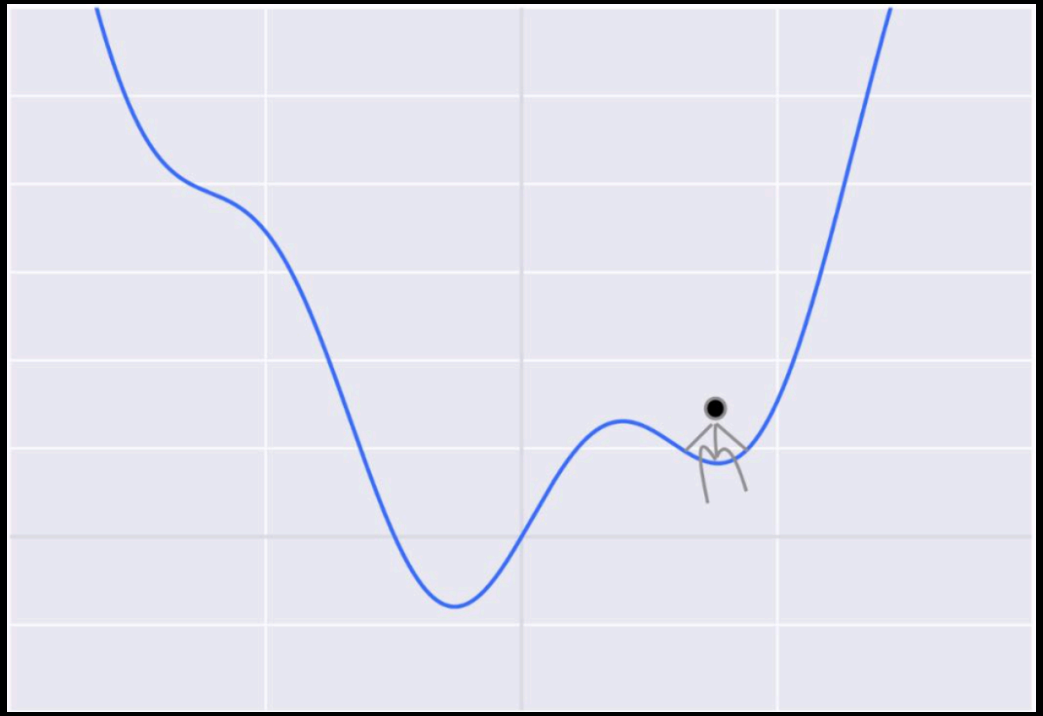
Example: With momentum

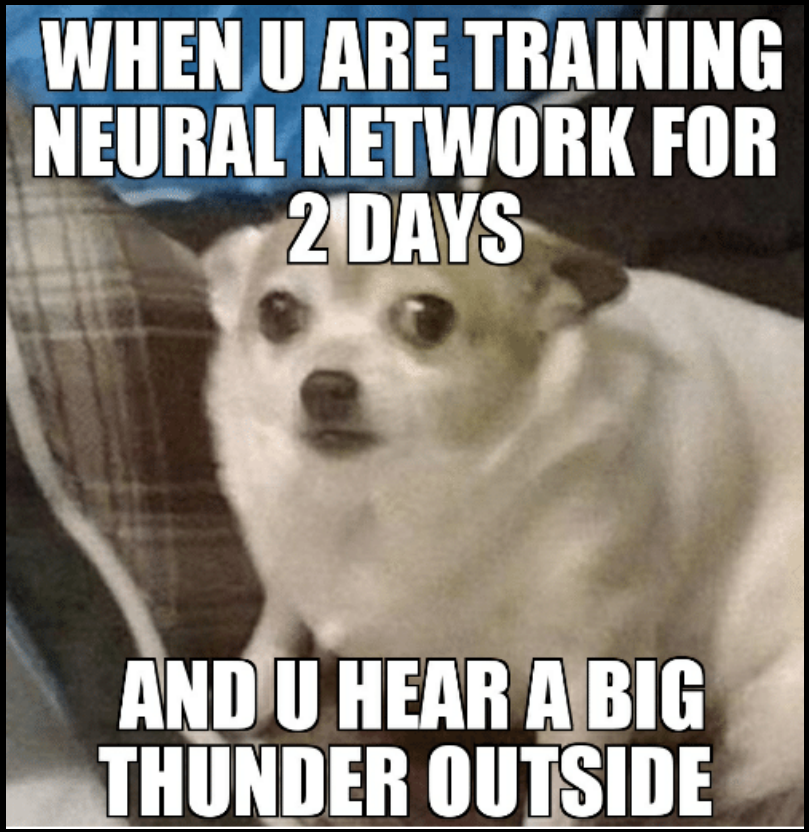


Intuitive explanation



Intuitive explanation

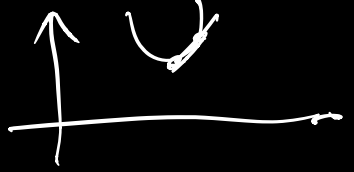



WHEN U ARE TRAINING NEURAL NETWORK FOR 2 DAYS
AND U HEAR A BIG THUNDER OUTSIDE

Second order approximation

$$f(x_0) + f'(x_0)(x - x_0) \to \min$$

$$x \in \mathbb{R}$$



$$f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \to \min_x$$

$$x = x_0 - \frac{f'(x_0) -}{f''(x_0) -}$$

Second order approximation

$x \in \mathbb{R}^d$ - ???

$$f(x) = f(x_0) + (x - x_0)^T \nabla f(x_0) +$$

$$+ \frac{1}{2} (x - x_0)^T \nabla^2 f(x_0) (x - x_0)$$

$$H_2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}$$

$$x = x_0 - (\nabla^2 f)^{-1}$$

$$X = X_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$$

$d \times d$

Метод Ньютона

$$w_{k+1} = w_k - \gamma (\nabla^2 L)^{-1} \nabla L$$

обычно мало

AdaGrad

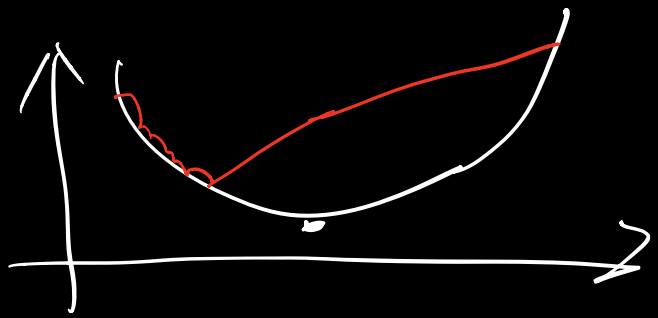
$x = (x^1, \dots, x^j, \dots, x^d)$

$$G_{kj} = G_{k-1,j} + (\nabla_w L)_j^2 \rightarrow$$

G_{kj}: k — итерация, j — компонента

$$w_j^{(k)} = w_j^{(k-1)} - \frac{\eta}{\sqrt{G_{kj} + \varepsilon}} \left(\nabla_w L(w^{(k-1)})\right)_j$$

$w_j^{(k)}$: k — итерация, j — компонента



RMSprop

$$G_{kj} = \alpha\, G_{k-1,j} + (1-\alpha)\left(\nabla_w L_S(w^{(k-1)})\right)_j^2$$

Material & Links

- https://distill.pub/2017/momentum/
- https://ru.coursera.org/learn/intro-to-deep-learning