Глубокое обучение и вообще

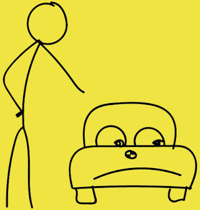
Ульянкин Филипп

Посиделка 4: Алгоритм обратного распространения ошибки

Agenda

- Алгоритм обратного распространения ошибки

Как обучить нейросеть?



Ты необучаем!

Нейросеть — сложная функция

- Прямое распространение ошибки (forward propagation):

$$X \Rightarrow X \cdot W_1 \Rightarrow f(X \cdot W_1) \Rightarrow f(X \cdot W_1) \cdot W_2 \Rightarrow ... \Rightarrow \hat{y}$$

- Считаем потери:

$$Loss = \frac{1}{2}(y - \hat{y})^2$$

- Все слои обычно дифференцируемы, поэтому можно посчитать производные по всем параметрам

- Для обучения можно использовать градиентный спуск

Как обучить нейросеть?

$$L(W_1, W_2) = \frac{1}{2} \cdot (y - f(X \cdot W_1) \cdot W_2)^2$$

Секрет успеха в умении брать производную

Как обучить нейросеть?

$$L(W_1, W_2) = \frac{1}{2} \cdot (y - f(X \cdot W_1) \cdot W_2)^2$$

Секрет успеха в умении брать производную

$$\boxed{f(g(x))' = f'(g(x)) \cdot g'(x)}$$

Как обучить нейросеть?

$$L(W_1, W_2) = \frac{1}{2} \cdot (y - f(X \cdot W_1) \cdot W_2)^2$$

Секрет успеха в умении брать производную

$$\boxed{f(g(x))' = f'(g(x)) \cdot g'(x)}$$

$$\frac{\partial L}{\partial W_2} = -(y - f(X \cdot W_1) \cdot W_2) \cdot f(X \cdot W_1)$$
$$\frac{\partial L}{\partial W_1} = -(y - f(X \cdot W_1) \cdot W_2) \cdot W_2 \cdot f'(X \cdot W_1) \cdot X$$

Как обучить нейросеть?

$$L(W_1, W_2) = \frac{1}{2} \cdot (y - f(X \cdot W_1) \cdot W_2)^2$$

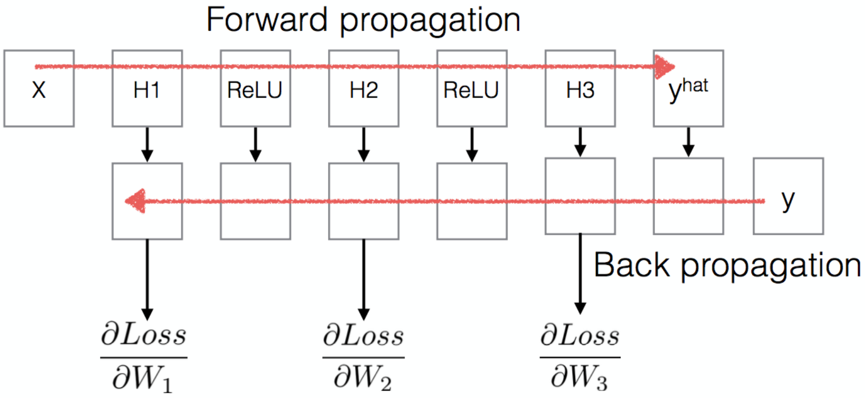
Секрет успеха в умении брать производную

$$\boxed{f(g(x))' = f'(g(x)) \cdot g'(x)}$$

$$\frac{\partial L}{\partial W_2} = -(y - f(X \cdot W_1) \cdot W_2) \cdot f(X \cdot W_1)$$
$$\frac{\partial L}{\partial W_1} = -(y - f(X \cdot W_1) \cdot W_2) \cdot W_2 \cdot f'(X \cdot W_1) \cdot X$$

Дважды ищем одно и то же $\Rightarrow$ оптимизация поиска производных даст нам алгоритм обратного распространения ошибки (back-propagation)

Back-propagation



Цепное правило

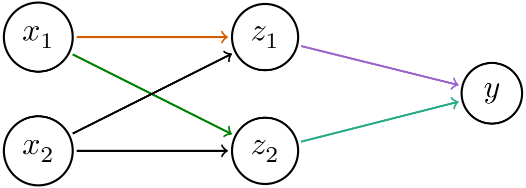
- Возьмём сложную функцию:

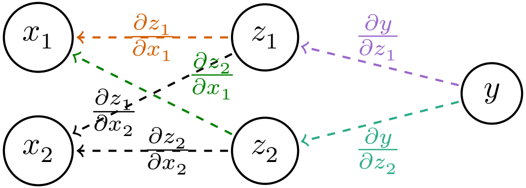
$$
\begin{aligned}
z_1 &= z_1(x_1, x_2) \\
z_2 &= z_2(x_1, x_2) \\
y &= y(z_1, z_2)
\end{aligned}
$$

- Производную такой функции можно найти по цепному правилу:

$$
\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1}
$$

Как считать производные?



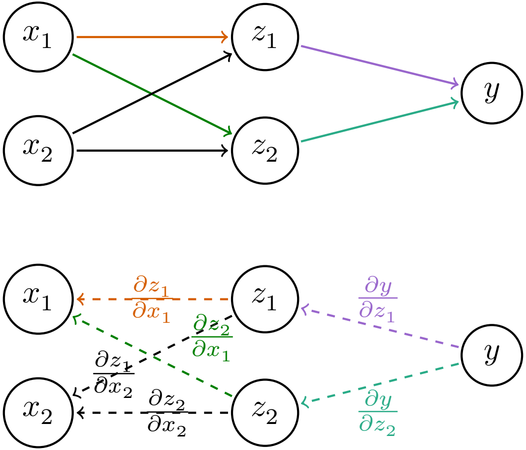


Граф вычислений:

$$z_1 = z_1(x_1, x_2)$$
$$z_2 = z_2(x_1, x_2)$$
$$y = y(z_1, z_2)$$

Из него можно построить граф производных, каждому ребру будет приписана производная

Как считать производные?



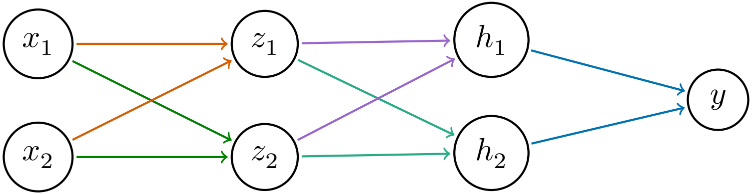
Граф вычислений:

$$z_1 = z_1(x_1, x_2)$$
$$z_2 = z_2(x_1, x_2)$$
$$y = y(z_1, z_2)$$

Можно догадаться как работает цепное правило:

$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1}$$

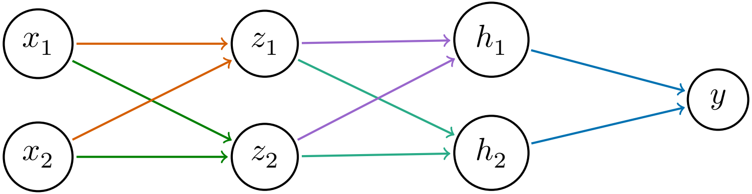
Пойдём глубже



$$z_1 = z_1(x_1, x_2) \quad h_1 = h_1(z_1, z_2) \quad y = y(h_1, h_2)$$

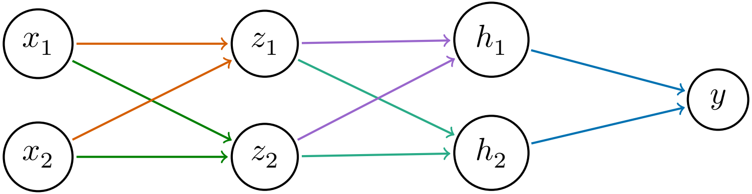
$$z_2 = z_2(x_1, x_2) \quad h_2 = h_2(z_1, z_2)$$

Пойдём глубже



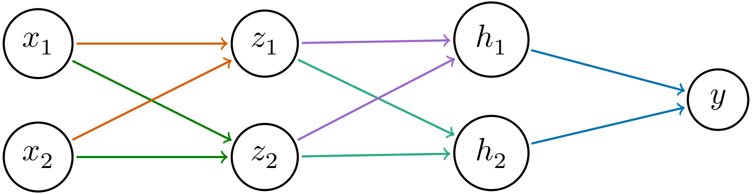
$$\frac{\partial y}{\partial x_1} = ?$$

Пойдём глубже



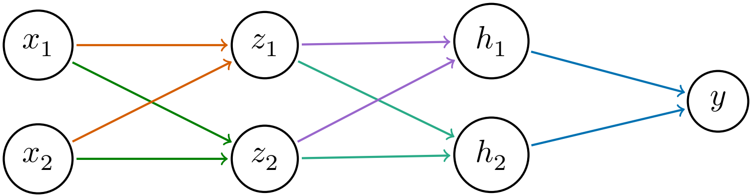
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial x_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial x_1}$$

Пойдём глубже



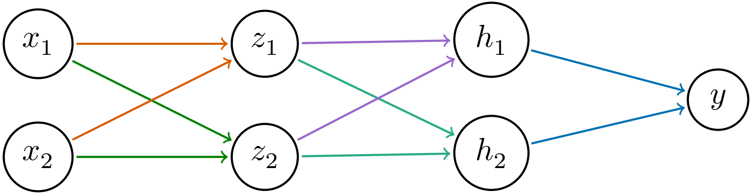
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1} \cdot \boxed{\frac{\partial h_1}{\partial x_1}} + \frac{\partial y}{\partial h_2} \cdot \boxed{\frac{\partial h_2}{\partial x_1}}$$

Пойдём глубже



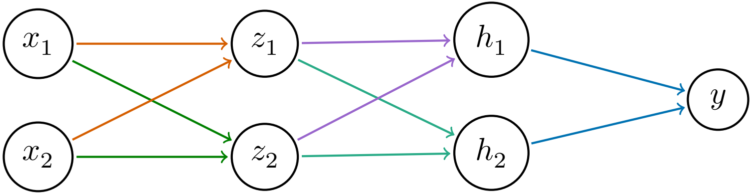
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1} \cdot \underbrace{\boxed{\frac{\partial h_1}{\partial x_1}}}_{\frac{\partial h_1}{\partial z_1}\cdot\frac{\partial z_1}{\partial x_1}+\frac{\partial h_1}{\partial z_2}\cdot\frac{\partial z_2}{\partial x_1}} + \frac{\partial y}{\partial h_2} \cdot \underbrace{\boxed{\frac{\partial h_2}{\partial x_1}}}_{\frac{\partial h_2}{\partial z_1}\cdot\frac{\partial z_1}{\partial x_1}+\frac{\partial h_2}{\partial z_2}\cdot\frac{\partial z_2}{\partial x_1}}$$

Пойдём глубже



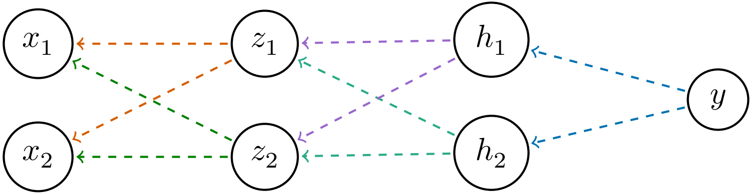
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1} \cdot \left(\frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial h_1}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1} \right) + \frac{\partial y}{\partial h_2} \cdot \left(\frac{\partial h_2}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1} \right)$$

Пойдём глубже



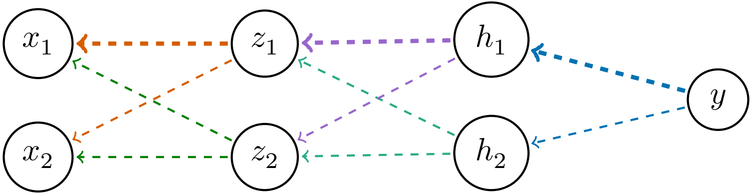
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

Пойдём глубже



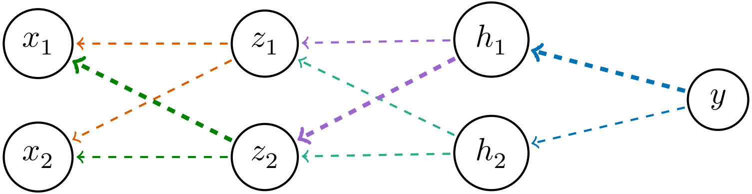
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

Пойдём глубже



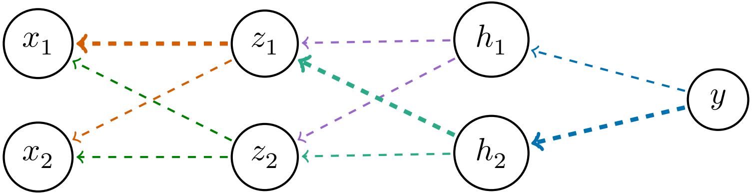
$$\frac{\partial y}{\partial x_1} = \boxed{\frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1}} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

Пойдём глубже



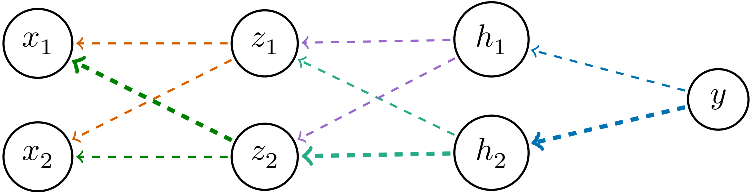
$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \boxed{\frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1}} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

Пойдём глубже



$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \boxed{\frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1}} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

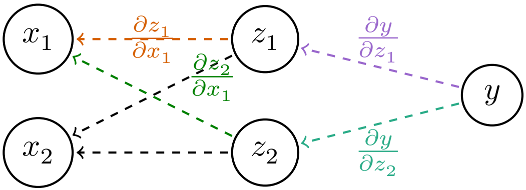
Пойдём глубже



$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \boxed{\frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}}$$

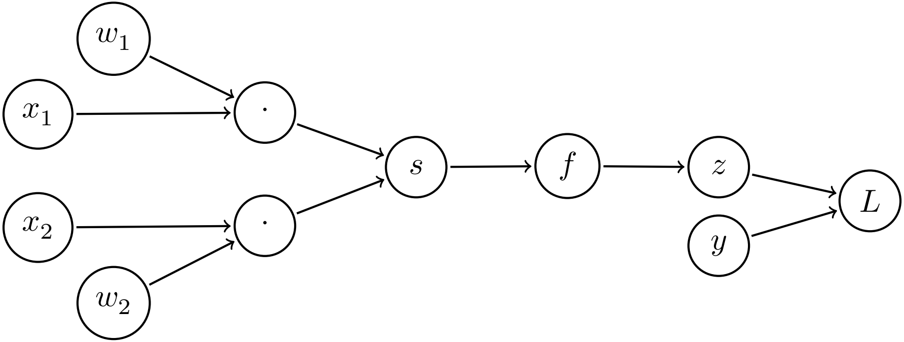
Алгоритм поиска производной в графе

- Как посчитать производную a по b?
- Находим непосещённый путь из a в b
- Перемножаем значения на рёбрах пути
- Добавляем в сумму



$$\frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1}$$

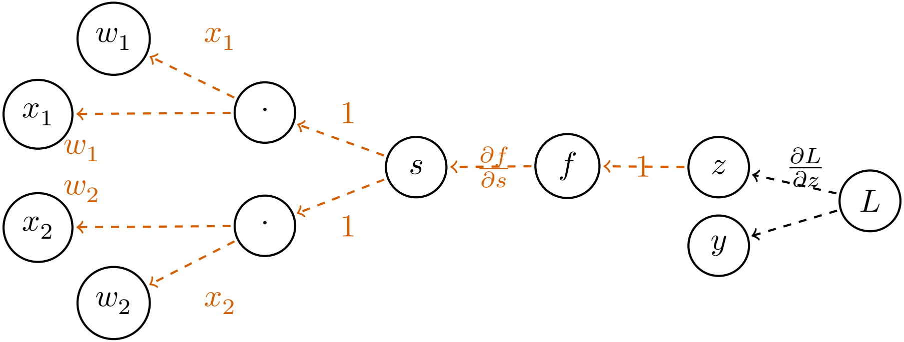
На примере одного нейрона



$$z = f(s) = f(w_1 \cdot x_1 + w_2 \cdot x_2)$$

Для SGD нам нужны $\frac{\partial L}{\partial w_1}$ и $\frac{\partial L}{\partial w_2}$

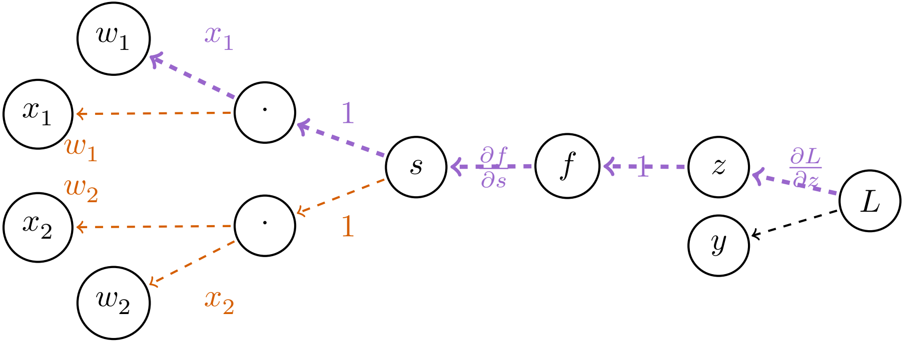
Граф производных



$$z = f(s) = f(w_1 \cdot x_1 + w_2 \cdot x_2)$$

Для SGD нам нужны $\frac{\partial L}{\partial w_1}$ и $\frac{\partial L}{\partial w_2}$

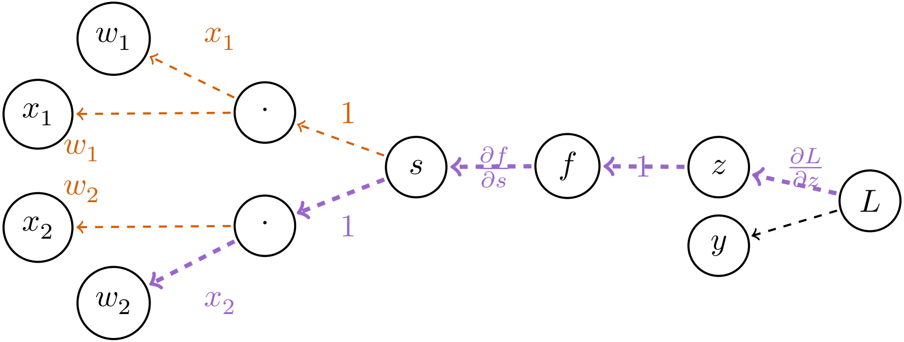
Граф производных



$$z = f(s) = f(w_1 \cdot x_1 + w_2 \cdot x_2)$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z} \cdot \frac{\partial f}{\partial s} \cdot x_1$$

Граф производных



$$z = f(s) = f(w_1 \cdot x_1 + w_2 \cdot x_2)$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z} \cdot \frac{\partial f}{\partial s} \cdot x_1 \qquad \frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z} \cdot \frac{\partial f}{\partial s} \cdot x_2$$

Цепное правило и граф производных

- Теперь у нас есть алгоритм для подсчета производных для любых дифференцируемых графов вычислений
- Осталось делать вычисления быстро

Обратное распространение ошибки

Мы хотим поменять параметры нейрона в рамках SGD

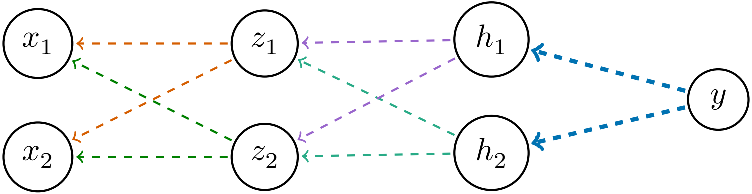
$$h_2 = f(w_0 + w_1 z_1 + w_2 z_2)$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial w_1} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_1}$$

$$w_1^t = w_1^{t-1} - \gamma \cdot \frac{\partial L}{\partial w_1}(w_1^{t-1})$$

Обратное распространение ошибки

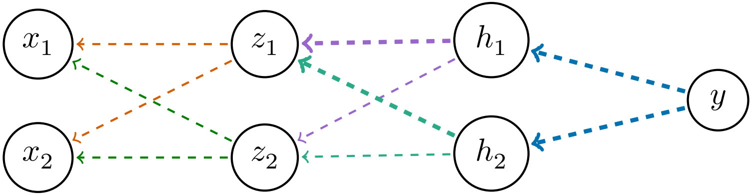
$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$



Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$

$$2: \quad \frac{\partial y}{\partial z_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_1} \qquad \frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

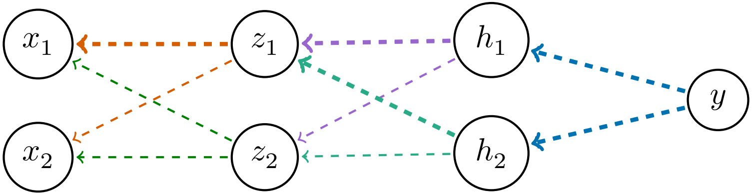


Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$

$$2: \quad \frac{\partial y}{\partial z_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_1} \qquad \frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

$$1: \quad \frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial h_1}\frac{\partial h_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} + \frac{\partial y}{\partial h_2}\frac{\partial h_2}{\partial z_2}\frac{\partial z_2}{\partial x_1}$$

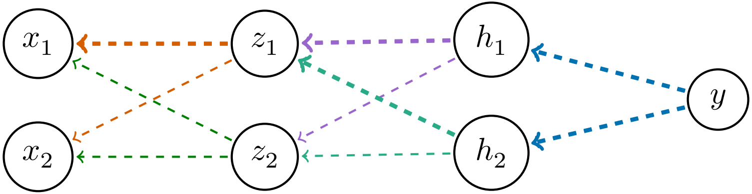


Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$

$$2: \quad \frac{\partial y}{\partial z_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_1} \qquad \frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

$$1: \quad \frac{\partial y}{\partial x_1} = \left(\frac{\partial y}{\partial h_1} \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \frac{\partial h_2}{\partial z_1} \right) \cdot \frac{\partial z_1}{\partial x_1} + \left(\frac{\partial y}{\partial h_1} \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \frac{\partial h_2}{\partial z_2} \right) \cdot \frac{\partial z_2}{\partial x_1}$$

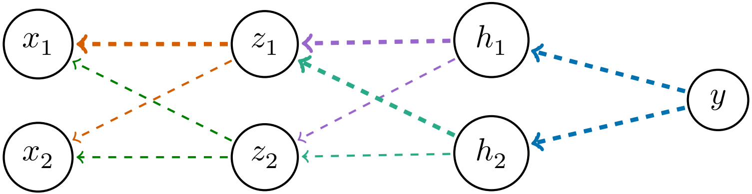


Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$

$$2: \quad \frac{\partial y}{\partial z_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_1} \qquad \frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

$$1: \quad \frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1}$$

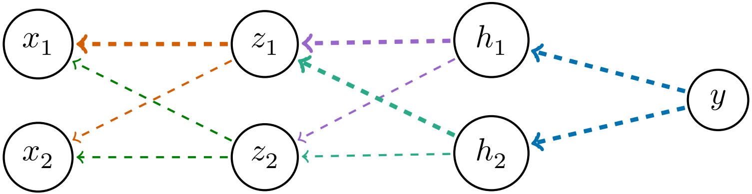


Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \frac{\partial y}{\partial h_1}$$

$$2: \quad \frac{\partial y}{\partial z_1} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_1} \qquad \frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

$$1: \quad \frac{\partial y}{\partial x_1} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1} \qquad \frac{\partial y}{\partial x_2} = \frac{\partial y}{\partial z_1} \cdot \frac{\partial z_1}{\partial x_2} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_2}$$

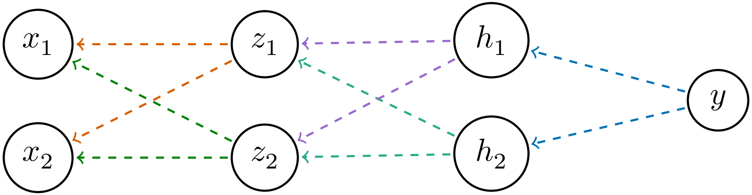


Обратное распространение ошибки

$$3: \quad \frac{\partial y}{\partial h_2} \qquad \boxed{\frac{\partial y}{\partial h_1}}$$

$$2: \quad \boxed{\frac{\partial y}{\partial z_1}} = \boxed{\frac{\partial y}{\partial h_1}} \cdot \frac{\partial h_1}{\partial z_1} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \qquad \frac{\partial y}{\partial z_2} = \boxed{\frac{\partial y}{\partial h_1}} \cdot \frac{\partial h_1}{\partial z_2} + \frac{\partial y}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2}$$

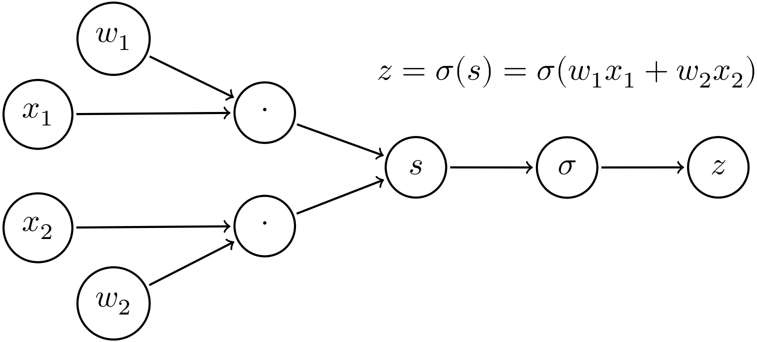
$$1: \quad \frac{\partial y}{\partial x_1} = \boxed{\frac{\partial y}{\partial z_1}} \cdot \frac{\partial z_1}{\partial x_1} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_1} \qquad \frac{\partial y}{\partial x_2} = \boxed{\frac{\partial y}{\partial z_1}} \cdot \frac{\partial z_1}{\partial x_2} + \frac{\partial y}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_2}$$



Обратное распространение ошибки

- Это называется reverse-mode дифференцирование, в теории нейросетей это называют back-propagation (обратное распространение ошибки)
- Работает быстро, потому что переиспользует вычисленные ранее значения
- На самом деле, по каждому ребру пройдемся всего раз, то есть сложность линейна по количеству ребер (т.е. параметров)

Back-propagation на одном нейроне



$$z = \sigma(s) = \sigma(w_1 x_1 + w_2 x_2)$$

Данные текут сквозь нейрон:

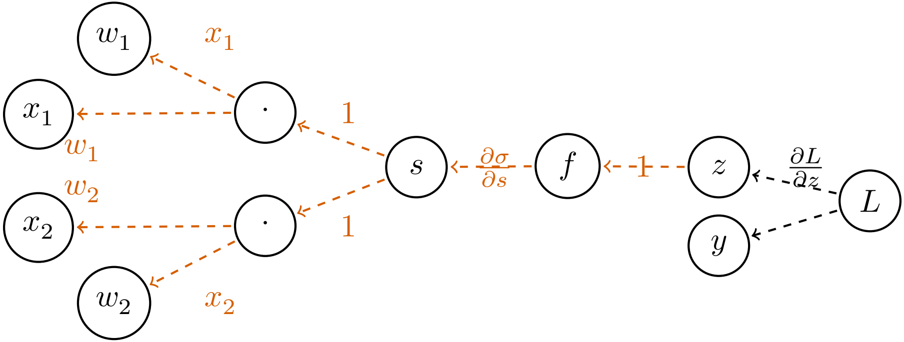
$$\boxed{X} \Rightarrow \boxed{s = X \cdot W} \Rightarrow \boxed{z = \sigma(s)} \Rightarrow \boxed{L(z, y) = (y - z)^2}$$

Back-propagation на одном нейроне

Forward pass:

$$\boxed{X} \Rightarrow \boxed{s = X \cdot W} \Rightarrow \boxed{z = \sigma(s)} \Rightarrow \boxed{L(z, y) = (y - z)^2}$$

Backward pass:



Back-propagation на одном нейроне

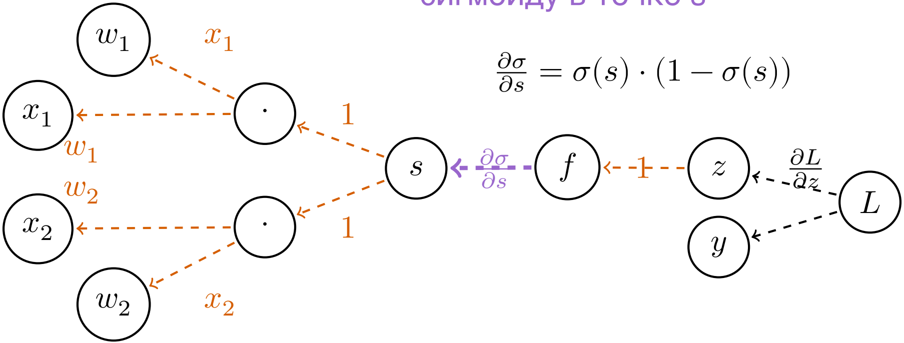
Forward pass:

$$\boxed{X} \Rightarrow \boxed{s = X \cdot W} \Rightarrow \boxed{z = \sigma(s)} \Rightarrow \boxed{L(z, y) = (y - z)^2}$$

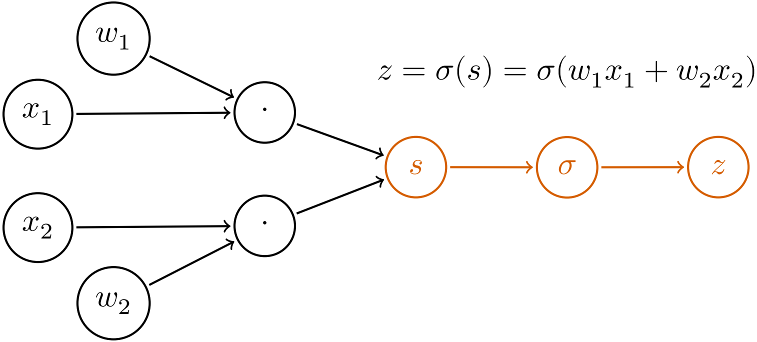
Backward pass:

Нам нужно вычислить сигмоиду в точке s

$$\frac{\partial \sigma}{\partial s} = \sigma(s) \cdot (1 - \sigma(s))$$



Сигмоида: прямой проход (forward pass)



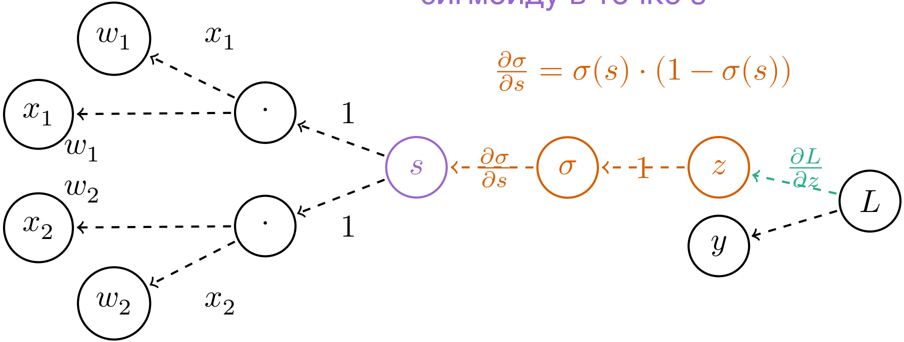
$$z = \sigma(s) = \sigma(w_1 x_1 + w_2 x_2)$$

```
def forward_pass(s):
    return  1/(1 + np.exp(-s))
```

Сигмоида: обратный проход (backward pass)

Нам нужно вычислить
сигмоиду в точке s

$$\frac{\partial \sigma}{\partial s} = \sigma(s) \cdot (1 - \sigma(s))$$




```
def backward_pass(s, incoming_gradient):
        sigm = 1/(1 + np.exp(-s)))
    return  sigm * (1 - sigm)  * incoming_gradient
```

$$\frac{\partial L}{\partial s} = \frac{\partial \sigma}{\partial s} \cdot \frac{\partial L}{\partial \sigma}$$

Полносвязный слой: прямой проход (forward pass)

- Два нейрона с тремя входами:

$$z_1 = x_1 w_{11} + x_2 w_{21} + x_3 w_{31}$$
$$z_2 = x_1 w_{12} + x_2 w_{22} + x_3 w_{32}$$

- Матричная запись:

$$\begin{pmatrix} z_1 & z_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}$$

$$z = xW$$

Полносвязный слой: обратный проход (backward pass)

- Матричная запись:

$$\begin{pmatrix} z_1 & z_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \cdot \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{32} \end{pmatrix}$$

$$Z = XW$$

- Для обратного прохода нам нужна $\frac{\partial L}{\partial W}$:

$$W_t = W_{t-1} - \eta_t \cdot \left.\frac{\partial L}{\partial W}\right|_{W_{t-1}}$$

Полносвязный слой в numpy

Прямой проход:

```
def forward_pass(X, W):
      return  X.dot(W))
```

$$Z = XW$$

Обратный проход:

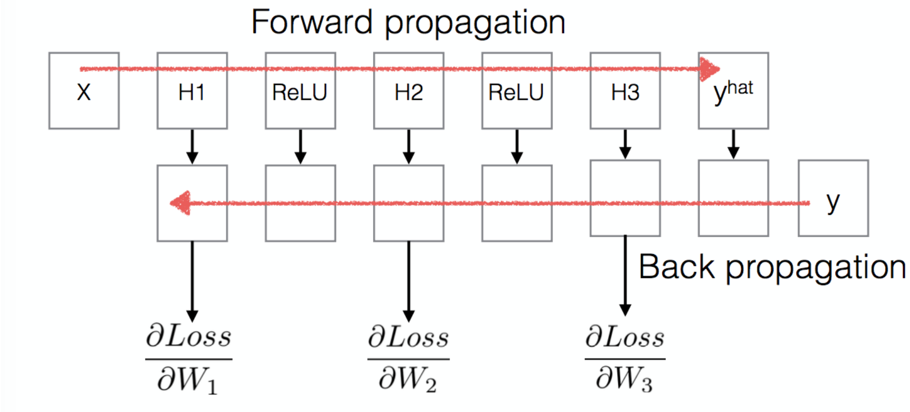
```
def forward_pass(X, W, in_grad):
      dX = in_grad.dot(W.T)
      dW = X.T.dot(in_grad)
      return  dX, dW
```

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Z} \cdot W^T$$

$$\frac{\partial L}{\partial W} = X^T \cdot \frac{\partial L}{\partial Z}$$

Эти формулы мы получим на семинаре

Back-propagation



Back-propagation

Forward pass:

$$X \xrightarrow{W_1} H_1 \xrightarrow{f} O_1 \xrightarrow{W_2} H_2 \xrightarrow{f} O_2 \xrightarrow{W_3} \hat{y} \longrightarrow MSE$$

Backward pass:

$$X \xleftarrow{\frac{\partial H_1}{\partial X}} H_1 \xleftarrow{\frac{\partial O_1}{\partial H_1}} O_1 \xleftarrow{\frac{\partial H_2}{\partial O_1}} H_2 \xleftarrow{\frac{\partial O_2}{\partial H_2}} O_2 \xleftarrow{\frac{\partial \hat{y}}{\partial O_2}} \hat{y} \xleftarrow{\frac{\partial MSE}{\partial \hat{y}}} MSE$$

$$\frac{\partial H_1}{\partial W_1} = X^T \qquad \frac{\partial H_2}{\partial W_2} = O_1^T \qquad \frac{\partial \hat{y}}{\partial W_3} = O_2^T$$

Back-propagation

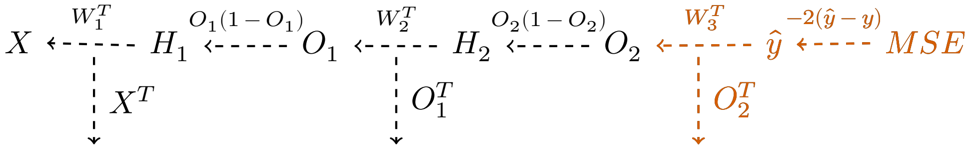
Forward pass:

$$X \xrightarrow{W_1} H_1 \xrightarrow{\sigma} O_1 \xrightarrow{W_2} H_2 \xrightarrow{\sigma} O_2 \xrightarrow{W_3} \hat{y} \longrightarrow MSE$$

Backward pass:

$$\begin{tikzcd}
X & {} \arrow[l, dashed] \arrow[d, dashed, "X^T"] & H_1 \arrow[l, dashed, "W_1^T"'] & O_1 \arrow[l, dashed, "O_1(1-O_1)"'] & {} \arrow[l, dashed] \arrow[d, dashed, "O_1^T"] & H_2 \arrow[l, dashed, "W_2^T"'] & O_2 \arrow[l, dashed, "O_2(1-O_2)"'] & {} \arrow[l, dashed] \arrow[d, dashed, "O_2^T"] & \hat{y} \arrow[l, dashed, "W_3^T"'] & MSE \arrow[l, dashed, "-2(\hat{y}-y)"'] \\
 & {} & & & {} & & & {} & &
\end{tikzcd}$$

Back-propagation



Шаг 1:

$$d = -2(\hat{y} - y)$$

$$\frac{\partial MSE}{\partial W_3} = O_2^T \cdot d$$

Back-propagation

$$
\begin{tikzcd}
X & & H_1 \arrow[ll, dashed, "W_1^T"'] & O_1 \arrow[l, dashed, "O_1(1-O_1)"'] & & H_2 \arrow[ll, dashed, "W_2^T"'] & O_2 \arrow[l, dashed, "O_2(1-O_2)"'] & & \hat{y} \arrow[ll, dashed, "W_3^T"'] & MSE \arrow[l, dashed, "-2(\hat{y}-y)"'] \\
& {} \arrow[from=1-2, d, dashed, "X^T"] & & & {} \arrow[from=1-5, d, dashed, "O_1^T"] & & & {} \arrow[from=1-8, d, dashed, "O_2^T"] & & \\
& {} & & & {} & & & {} & &
\end{tikzcd}
$$

Шаг 1:

$$d = -2(\hat{y} - y)$$

$$\frac{\partial MSE}{\partial W_3} = O_2^T \cdot d$$

Шаг 2:

$$d = d \cdot W_3^T * O_2 * (1 - O_2)$$

$$\frac{\partial MSE}{\partial W_2} = O_1^T \cdot d$$

Back-propagation

$$
\begin{tikzcd}
X & {} \arrow[l, dashed, "W_1^T"'] \arrow[d, dashed, "X^T"] & H_1 \arrow[l, dashed, no head] & O_1 \arrow[l, dashed, "O_1(1-O_1)"'] & {} \arrow[l, dashed, "W_2^T"'] \arrow[d, dashed, "O_1^T"] & H_2 \arrow[l, dashed, no head] & O_2 \arrow[l, dashed, "O_2(1-O_2)"'] & {} \arrow[l, dashed, "W_3^T"'] \arrow[d, dashed, "O_2^T"] & \hat{y} \arrow[l, dashed, no head] & MSE \arrow[l, dashed, "-2(\hat{y}-y)"'] \\
 & {} & & & {} & & & {} & &
\end{tikzcd}
$$

Шаг 1:

$$d = -2(\hat{y} - y)$$

$$\frac{\partial MSE}{\partial W_3} = O_2^T \cdot d$$

Шаг 2:

$$d = d \cdot W_3^T * O_2 * (1 - O_2)$$

$$\frac{\partial MSE}{\partial W_2} = O_1^T \cdot d$$

Шаг 3:

$$d = d \cdot W_2^T * O_1 * (1 - O_1)$$

$$\frac{\partial MSE}{\partial W_1} = X^T \cdot d$$

Back-propagation

$$
\begin{tikzcd}
X & {} \arrow[l, dashed, "W_1^T"'] \arrow[d, dashed, "X^T"] & H_1 \arrow[l, dashed, no head] & O_1 \arrow[l, dashed, "O_1(1-O_1)"'] & {} \arrow[l, dashed, "W_2^T"'] \arrow[d, dashed, "O_1^T"] & H_2 \arrow[l, dashed, no head] & O_2 \arrow[l, dashed, "O_2(1-O_2)"'] & {} \arrow[l, dashed, "W_3^T"'] \arrow[d, dashed, "O_2^T"] & \hat{y} \arrow[l, dashed, no head] & MSE \arrow[l, dashed, "-2(\hat{y}-y)"'] \\
 & {} & & & {} & & & {} & &
\end{tikzcd}
$$

Шаг 1:

$$d = -2(\hat{y} - y)$$

$$\frac{\partial MSE}{\partial W_3} = O_2^T \cdot d$$

Шаг 2:

$$d = d \cdot W_3^T * O_2 * (1 - O_2)$$

$$\frac{\partial MSE}{\partial W_2} = O_1^T \cdot d$$

Шаг 3:

$$d = d \cdot W_2^T * O_1 * (1 - O_1)$$

$$\frac{\partial MSE}{\partial W_1} = X^T \cdot d$$

Шаг SGD:

$$W_3^t = W_3^{t-1} - \eta \cdot \frac{\partial MSE}{\partial W_3} \qquad W_2^t = W_2^{t-1} - \eta \cdot \frac{\partial MSE}{\partial W_2} \qquad W_1^t = W_1^{t-1} - \eta \cdot \frac{\partial MSE}{\partial W_1}$$

Численная оценка производной

- Способа считать производную быстрее — нет!
- Можно попробовать посчитать численную оценку

$$\frac{f(x, w+\varepsilon) - f(x, w)}{\varepsilon}$$

- При больших ε результат будет очень неточным, при малых ε начнутся численные приколы с точностью вычислений
- На практике лучше работает формула

$$\frac{f(x, w+\varepsilon) - f(x, w-+\varepsilon)}{2 \cdot \varepsilon}$$

Что такое слой в нейронной сети?

- Любой слой - это какая-то абстракция, которая умеет делать прямой шаг и обратный шаг

- Для всех слоёв, которые мы дальше будем изучать, мы всегда будем смотреть на то как выглядят эти два шага

А мне точно надо понимать backprop?

- Да, точно!
- "Backprop – leaky abstraction!"
- Почему сеть не обучается?
- Почему сеть обучается слишком медленно?
- Какие проблемы могут возникать в обучении из-за плохой архитектуры?