Глубокие нейронные сети

Функция потерь на основе перекрестной энтропии

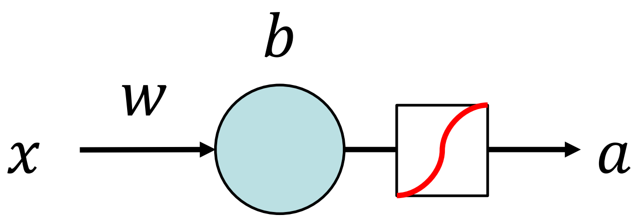
(Cross-entropy cost function)

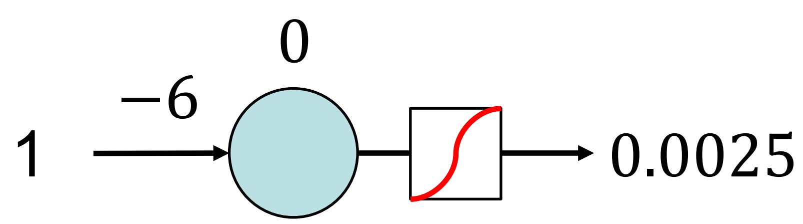
Лекция 5

Проблема с замедлением обучения (исчезающий градиент)

$$\mathrm{V} = \{(1; 0)\}$$

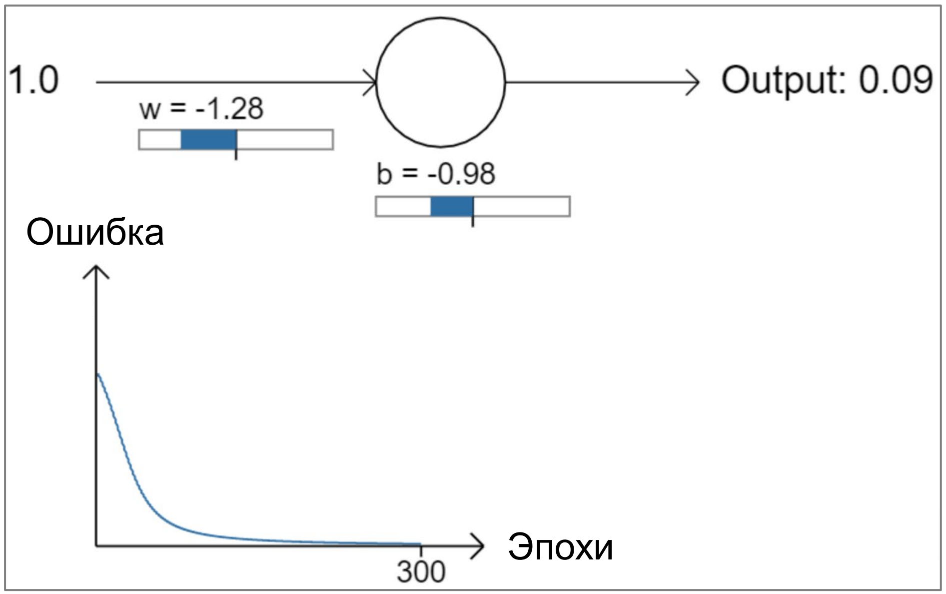
Обученный нейрон

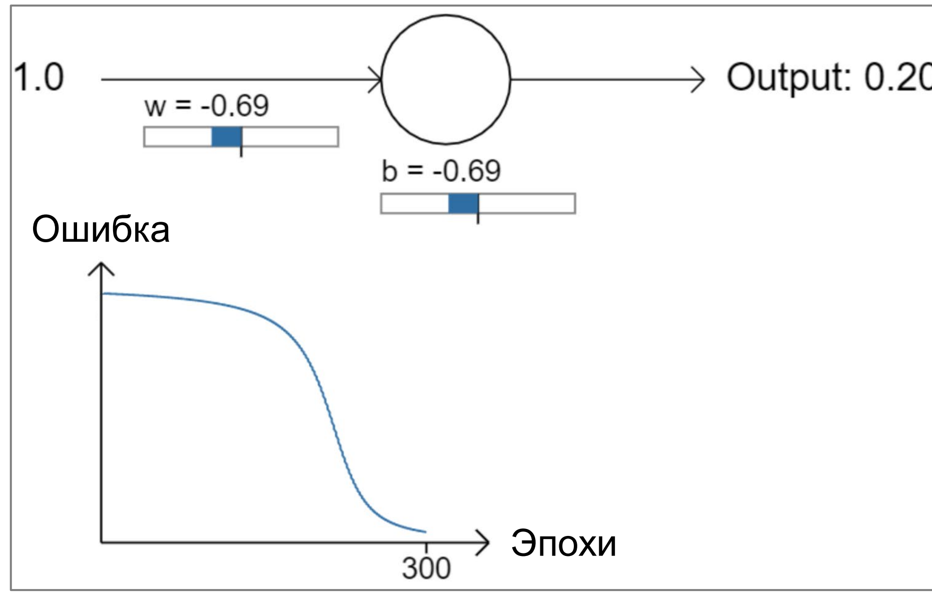




Зависимость скорости обучения от начальных значений w и b

$\eta = 0.15$ Правильный ответ: 0.0025





Начальные значения:
$w = 0.6$
$b = 0.9$

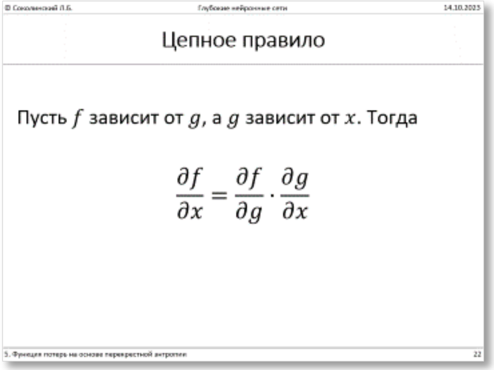
Начальные значения:
$w = 2$
$b = 2$

Анализ проблемы

$$\left.\begin{array}{c} C=\dfrac{(y-a)^2}{2} \\ a=\sigma(z);\ z=wx+b;\ \boldsymbol{x=1};\ \boldsymbol{y=0}\end{array}\right\}\Rightarrow$$

$$\frac{\partial C}{\partial w}=\frac{\partial C}{\partial a}\cdot\frac{\partial\sigma(z)}{\partial z}\cdot\frac{\partial z}{\partial w}=(a-y)\sigma'(z)x=a\sigma'(z)1$$

$$\frac{\partial C}{\partial b}=\frac{\partial C}{\partial a}\cdot\frac{\partial\sigma(z)}{\partial z}\cdot\frac{\partial z}{\partial b}=(a-y)\sigma'(z)=a\sigma'(z)$$


© Соколинский Л.Б. Глубокие нейронные сети 14.10.2023

Цепное правило

Пусть f зависит от g, а g зависит от x. Тогда

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial g}\cdot\frac{\partial g}{\partial x}$$

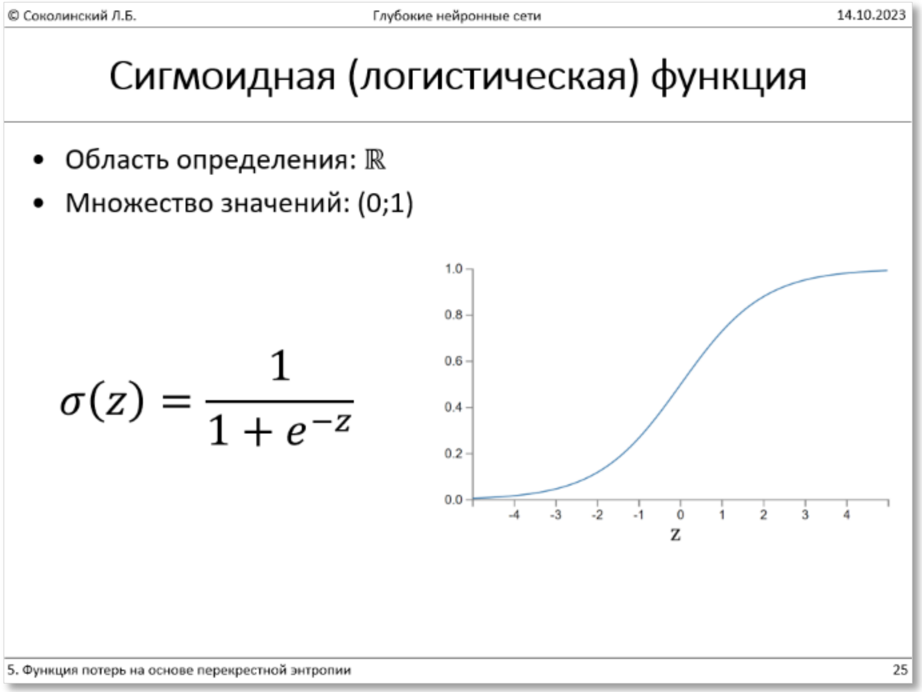
5. Функция потерь на основе перекрестной энтропии 22

Продолжение анализа проблемы

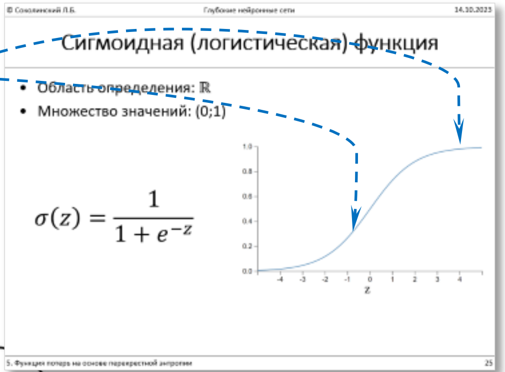
$$\left.\begin{aligned} w &:= w - \eta \frac{\partial C}{\partial w} \\ b &:= b - \eta \frac{\partial C}{\partial b} \end{aligned}\right\} \Rightarrow \begin{cases} w := w - \eta a \sigma'(z) \\ b := b - \eta a \sigma'(z) \end{cases}$$

$$\lim_{z \to +\infty} \sigma'(z) = 0$$

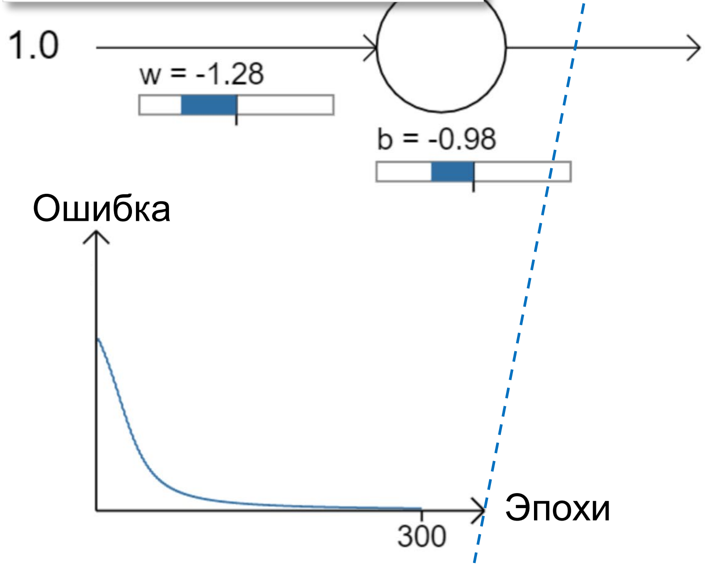
$$\lim_{z \to -\infty} \sigma'(z) = 0$$

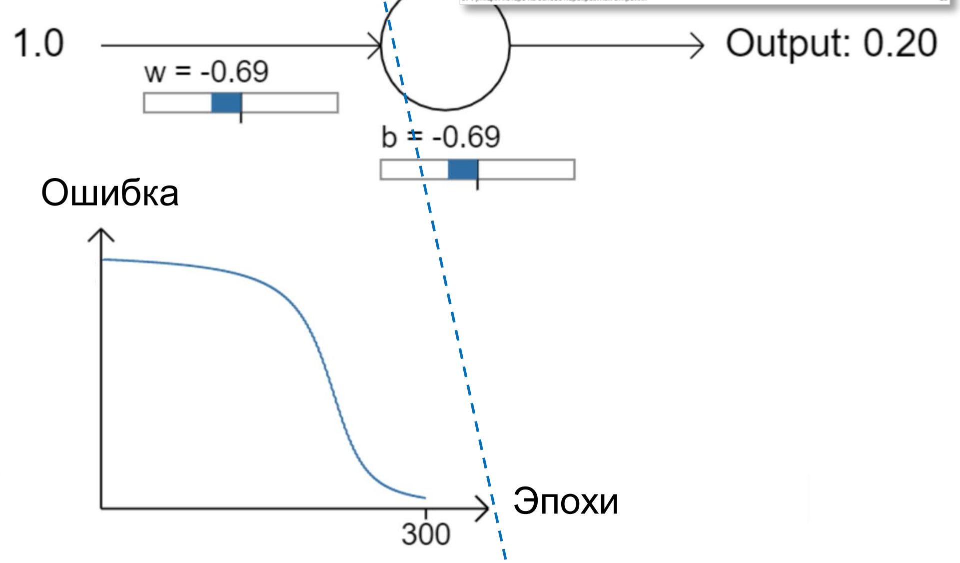


Зависимость скорости обучения от начальных значений w и b



$$z = 1.5 \leftarrow z = wx + b \rightarrow z = 4$$





Начальные значения:
$w = 0.6$
$b = 0.9$

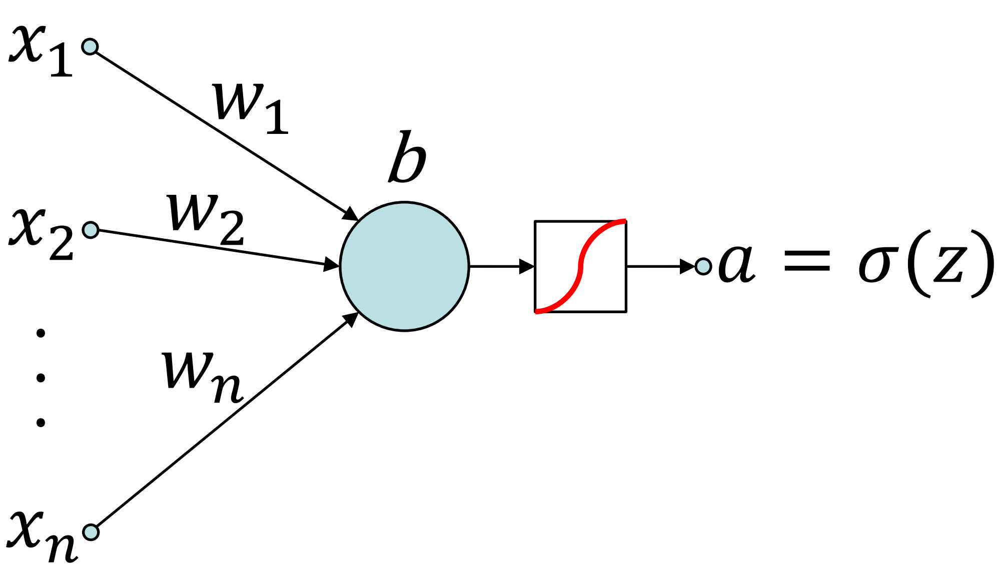
Начальные значения:
$w = 2$
$b = 2$

Выводы

- Если выходной сигнал нейрона близок к нулю или единице, то значение производной для сигмоидной функции близко к нулю
- Следовательно, в этих случаях будет наблюдаться замедление обучения
- Что делать?
- **Решение**: заменить среднеквадратичную функцию потерь на функцию *перекрестной энтропии* (при условии, что выходными сигналами могут быть только значения из множества {0,1})

Использование функции перекрестной энтропии в качестве функции потерь

$$z = b + \sum_{i=1}^{n} w_i x_i$$

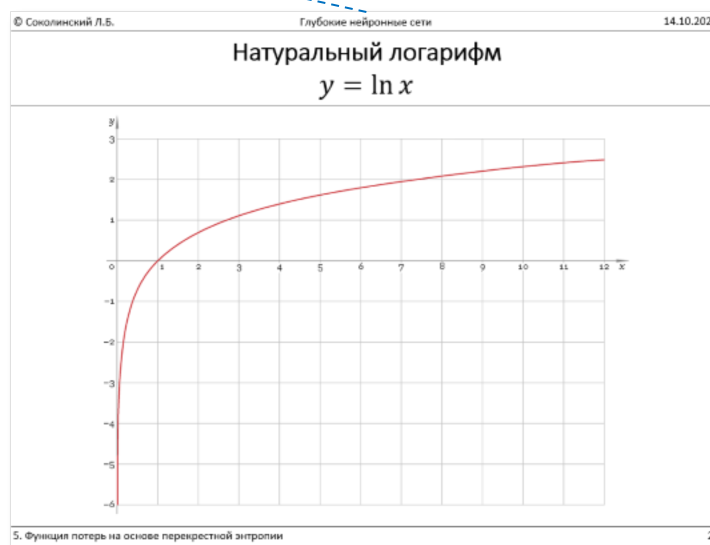


$$C = -(y \ln a + (1 - y) \ln(1 - a))$$

Проверка ограничения $C \geq 0$

$$C = y(-\ln a) + (1 - y)(-\ln(1 - a))$$

$$\left.\begin{array}{l} a = \sigma(z) \Rightarrow 0 < a < 1 \Rightarrow \ln a < 0 \wedge \ln(1 - a) < 0 \\ 0 \leq y \leq 1 \end{array}\right\} \Rightarrow C \geq 0$$



Свойства функции потерь на основе перекрестной энтропии

$$\boxed{C = -(y \ln a + (1-y)\ln(1-a))}$$

$$\lim_{a \to y} C = 0 \qquad (1)$$

Если выходной сигнал близок желаемому, величина функции потерь близка к нулю

$$\frac{\partial C}{\partial w_i} = x_i \cdot (\sigma(z) - y) \qquad (2)$$

$$\frac{\partial C}{\partial b} = \sigma(z) - y \qquad (3)$$

Нет замедления обучения при $\sigma'(z) \to 0$

Доказательство $\lim\limits_{a \to y} C = 0$ (1)

$$C = -(y \ln a + (1 - y) \ln(1 - a))$$

$$y = 0 \approx a \Rightarrow \begin{cases} y \ln a = 0 \\ (1 - y) \ln(1 - a) = \ln(1 - a) \approx \ln 1 = 0 \end{cases} \Rightarrow$$

$$\Rightarrow C \approx 0$$

$$y = 1 \approx a \Rightarrow \begin{cases} y \ln a = \ln a \approx \ln 1 = 0 \\ (1 - y) \ln(1 - a) = 0 \end{cases}$$

$$\Rightarrow C \approx 0$$

Доказательство $\dfrac{\partial C}{\partial w_i} = x_i \cdot (\sigma(z) - y)$ (2)

$$\boxed{C = -(y \ln a + (1-y)\ln(1-a))}$$

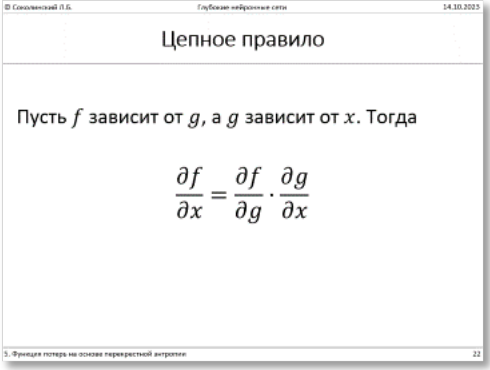
$$\frac{\partial C}{\partial w_i} = \frac{\partial C}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_i} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right) \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_i} =$$

$$= -\left(\frac{y}{\sigma(z)} - \frac{1-y}{1-\sigma(z)}\right) \cdot \sigma'(z) \cdot \frac{\partial z}{\partial w_i} =$$

$$= -\left(\frac{y}{\sigma(z)} - \frac{1-y}{1-\sigma(z)}\right) \cdot \sigma'(z) \cdot x_i =$$

$$= -\frac{\sigma'(z) \cdot x_i}{\sigma(z)\big(1-\sigma(z)\big)} \big(y - \sigma(z)\big) \cdot x_i =$$

$$= -x_i \cdot \big(y - \sigma(z)\big) = x_i \cdot \big(\sigma(z) - y\big)$$

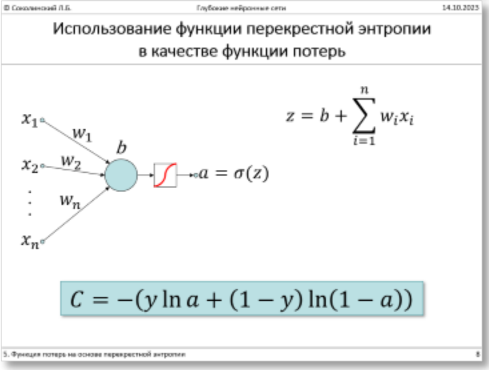

© Соколинский Л.Б. Глубокие нейронные сети 14.10.2023

Цепное правило

Пусть f зависит от g, а g зависит от x. Тогда

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

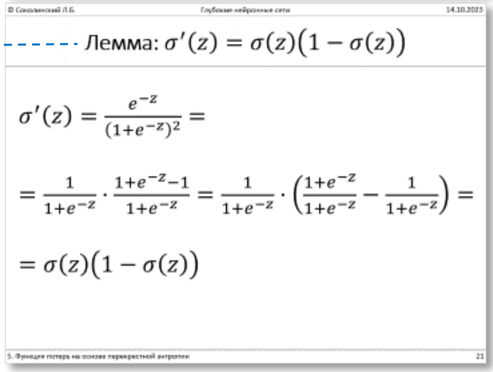
5. Функция потерь на основе перекрестной энтропии 22


© Соколинский Л.Б. Глубокие нейронные сети 14.10.2023

Использование функции перекрестной энтропии в качестве функции потерь

$$C = -(y \ln a + (1-y)\ln(1-a))$$

5. Функция потерь на основе перекрестной энтропии 8


© Соколинский Л.Б. Глубокие нейронные сети 14.10.2023

Лемма: $\sigma'(z) = \sigma(z)\big(1-\sigma(z)\big)$

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} =$$

$$= \frac{1}{1+e^{-z}} \cdot \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{1}{1+e^{-z}} \cdot \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) =$$

$$= \sigma(z)\big(1-\sigma(z)\big)$$

5. Функция потерь на основе перекрестной энтропии 21

Доказательство $\dfrac{\partial C}{\partial b} = \sigma(z) - y$ (3)

$$\boxed{C = -(y \ln a + (1-y)\ln(1-a))}$$

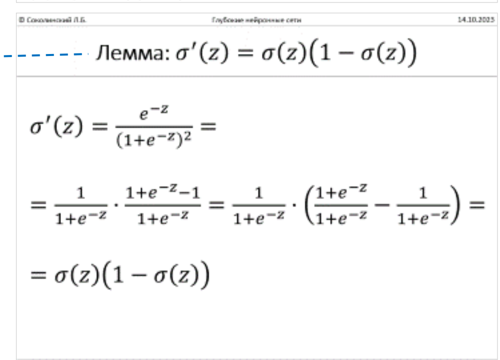
$$\frac{\partial C}{\partial b} = \frac{\partial C}{\partial a}\cdot\frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial b} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right)\cdot\frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial b} =$$

$$= -\left(\frac{y}{\sigma(z)} - \frac{1-y}{1-\sigma(z)}\right)\cdot\sigma'(z)\cdot\frac{\partial z}{\partial b} =$$

$$= -\left(\frac{y}{\sigma(z)} - \frac{1-y}{1-\sigma(z)}\right)\cdot\sigma'(z)\cdot 1 =$$

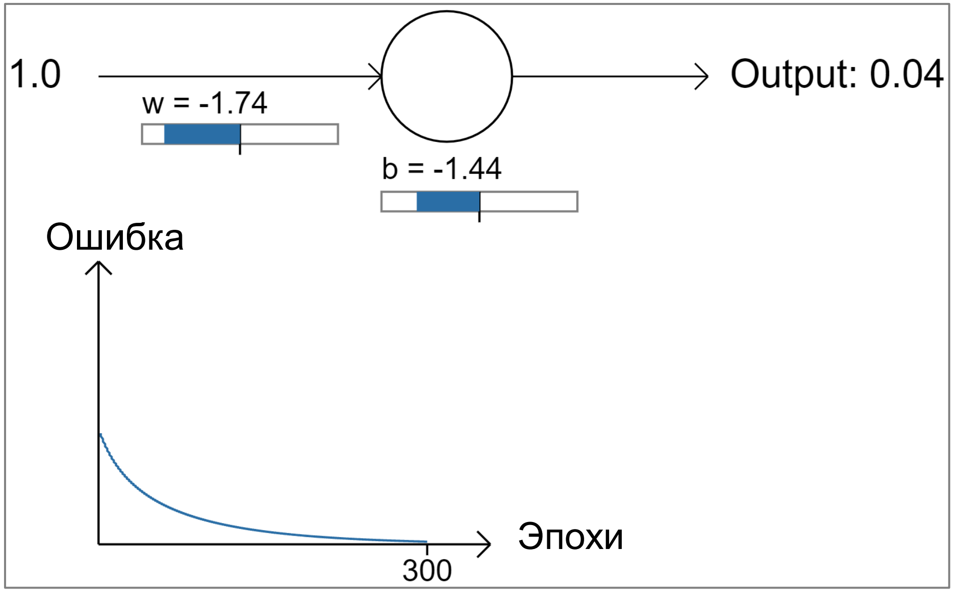
$$= -\frac{\sigma'(z)}{\sigma(z)\big(1-\sigma(z)\big)}\big(y-\sigma(z)\big) =$$

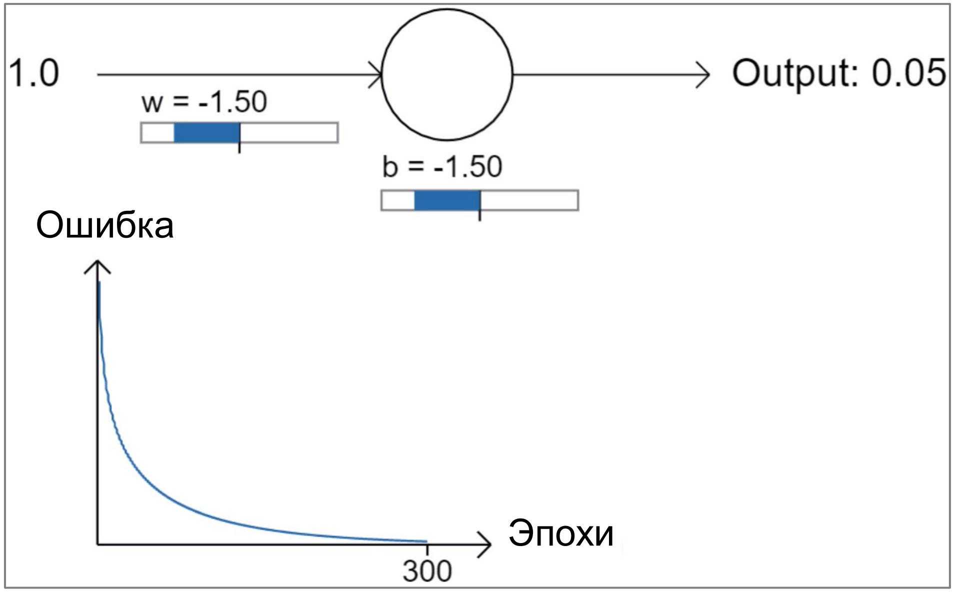
$$= -\big(y-\sigma(z)\big) = \sigma(z) - y$$



Эффект от применения функции перекрестной энтропии

$\eta = 0.005$ Правильный ответ: 0.0025





Начальные значения:
$w = 0.6$
$b = 0.9$

Начальные значения:
$w = 2$
$b = 2$

Обобщение функции потерь для нейронной сети с несколькими выходами

$$C = -\sum_j \left(y_j \ln a_j^{(L)} + (1 - y_j) \ln \left(1 - a_j^{(L)} \right) \right)$$

Обобщение функции потерь на всю обучающую выборку

$$\mathbb{C} = -\frac{1}{|V|} \sum_{(x,y)\in V} \sum_{j} \left(y_j \ln a_j^{(L)} + (1 - y_j) \ln \left(1 - a_j^{(L)} \right) \right)$$

Вычисление меры $\delta_i^{(L)}$ влияния нейрона на величину ошибки для выходного слоя

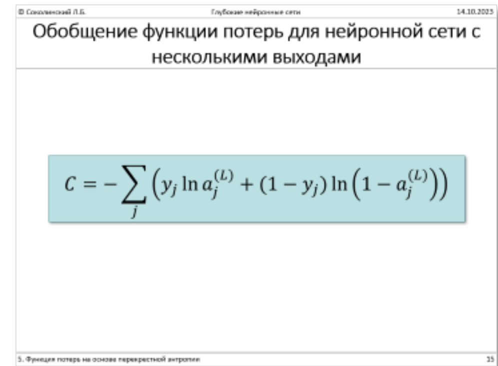
$$\delta_i^{(L)} = \frac{\partial C}{\partial z_i^{(L)}} = \frac{\partial C}{\partial a_i^L} \cdot \frac{\partial a_i^{(L)}}{\partial z_i^L} = -\left(\frac{y_i}{a_i^L} - \frac{1-y_i}{1-a_i^L}\right) \cdot \frac{\partial a_i^{(L)}}{\partial z_i^{(L)}} =$$

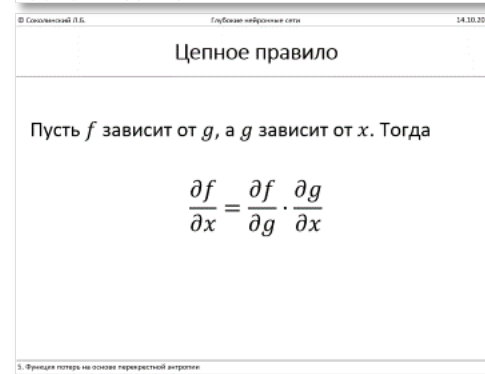
$$= -\left(\frac{y_i}{\sigma\left(z_i^{(L)}\right)} - \frac{1-y_i}{1-\sigma\left(z_i^{(L)}\right)}\right) \cdot \sigma'\left(z_i^{(L)}\right) =$$

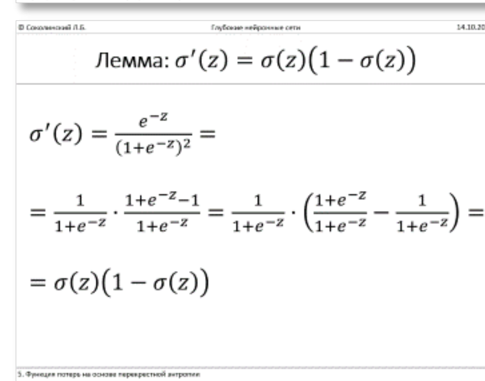
$$= -\frac{y_i-\sigma\left(z_i^{(L)}\right)}{\sigma\left(z_i^{(L)}\right)\cdot\left(1-\sigma\left(z_i^{(L)}\right)\right)} \cdot \sigma'\left(z_i^{(L)}\right) =$$

$$= -\frac{y_i-\sigma\left(z_i^{(L)}\right)}{\sigma\left(z_i^{(L)}\right)\cdot\left(1-\sigma\left(z_i^{(L)}\right)\right)} \cdot \sigma\left(z_i^{(L)}\right)\left(1-\sigma\left(z_i^{(L)}\right)\right) =$$

$$= -\left(y_i - \sigma\left(z_i^{(L)}\right)\right) = \sigma\left(z_i^{(L)}\right) - y_i = a_i^{(L)} - y_i$$







Формулы обратного распространения ошибки

Среднеквадратичная ошибка	Перекрестная энтропия	
$\delta_i^{(L)} = \left(a_i^{(L)} - y_i\right) \cdot \sigma'\left(z_i^{(L)}\right)$	$\delta_i^{(L)} = a_i^{(L)} - y_i$	**BP1**
$\delta_j^{(l)} = \sigma'\left(z_j^{(l)}\right) \sum_i w_{ji}^{(l+1)} \delta_i^{(l+1)}$		**BP2**
$\dfrac{\partial C}{\partial b_i^{(l)}} = \delta_i^{(l)}$		**BP3**
$\dfrac{\partial C}{\partial w_{ji}^{(l)}} = \delta_i^{(l)} \cdot a_j^{(l-1)}$		**BP4**

Отличительные свойства перекрестной энтропии

1. Ориентирована на нейронные сети с выходными сигналами из множества {0,1}
2. Не приводит к замедлению обучения при $\boldsymbol{\sigma}\left(z_i^{(L)}\right) \to 0$
3. Не приводит к замедлению обучения при $\boldsymbol{\sigma}\left(z_i^{(L)}\right) \to 1$

Конец лекции 5

Лемма: $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} =$$

$$= \frac{1}{1+e^{-z}} \cdot \frac{1+e^{-z}-1}{1+e^{-z}} = \frac{1}{1+e^{-z}} \cdot \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}}\right) =$$

$$= \sigma(z)\big(1 - \sigma(z)\big)$$

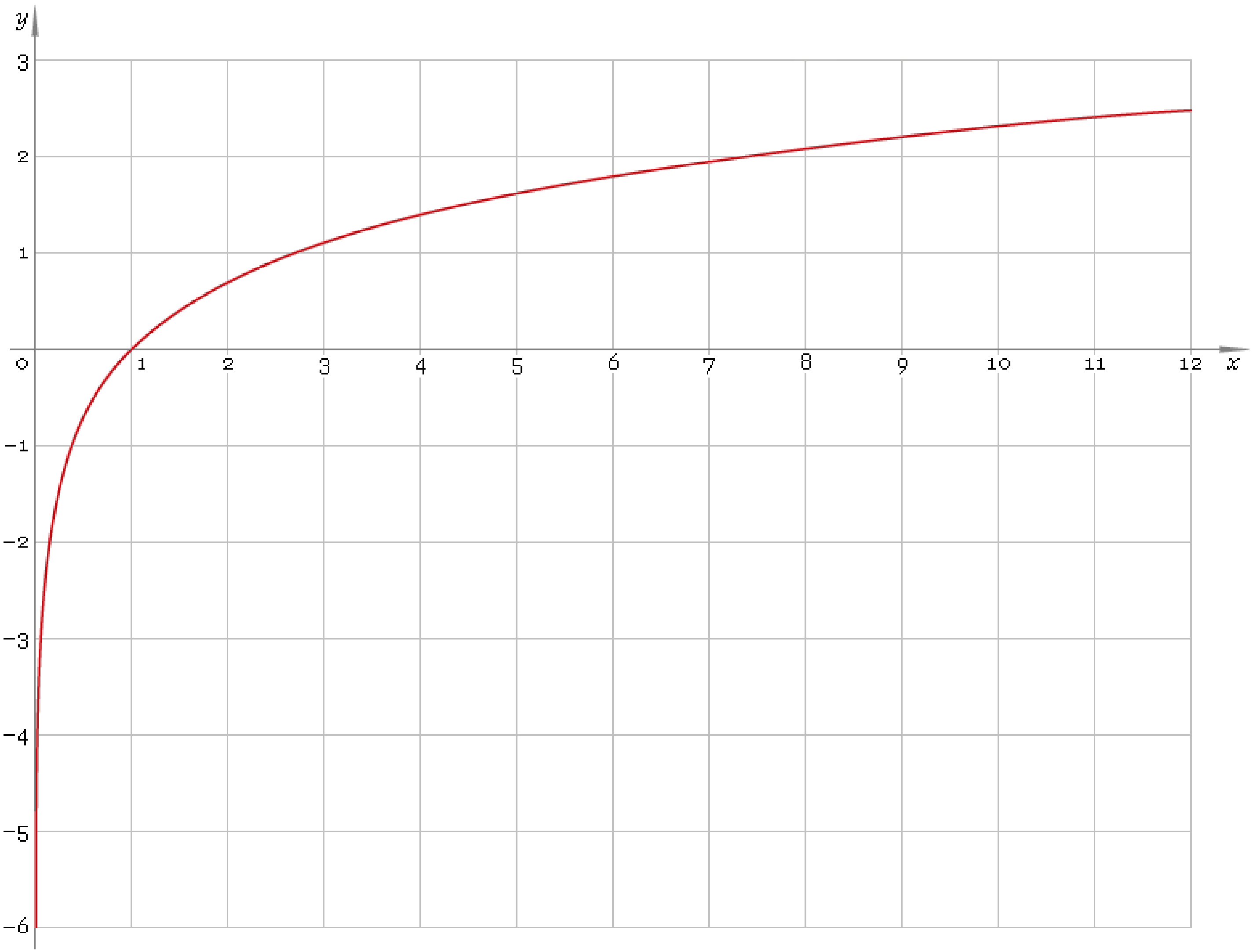
Цепное правило

Пусть f зависит от g, а g зависит от x. Тогда

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

Натуральный логарифм

$$y = \ln x$$



Примерно равно: ≈

$$x \approx h \Rightarrow y \approx g$$

$$\Updownarrow$$

$$\lim_{x \to h} y = g$$

Сигмоидная (логистическая) функция

- Область определения: $\mathbb{R}$
- Множество значений: (0;1)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$