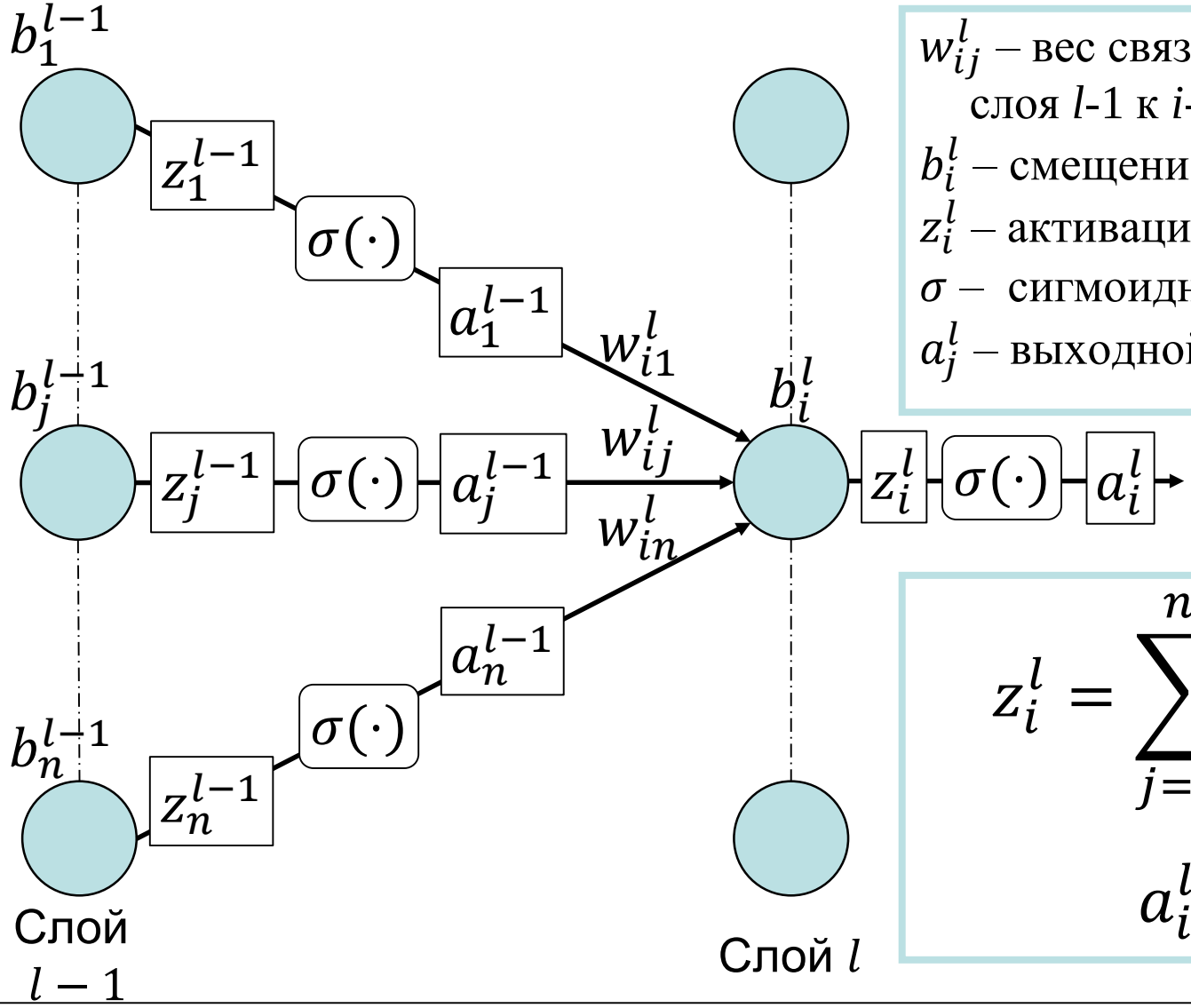
Машинное обучение

Метод обратного распространения ошибки (Error Back Propagation)

Лекция 3

Обозначения



w_{ij}^l – вес связи от j-того нейрона слоя l-1 к i-тому нейрону слоя l

b_i^l – смещение

z_i^l – активационный потенциал

σ – сигмоидная функция активации

a_j^l – выходной сигнал

$$z_i^l = \sum_{j=1}^{n} w_{ij}^l a_j^{l-1} + b_i^l$$

$$a_i^l = \sigma\left(z_i^l\right)$$

Обозначения

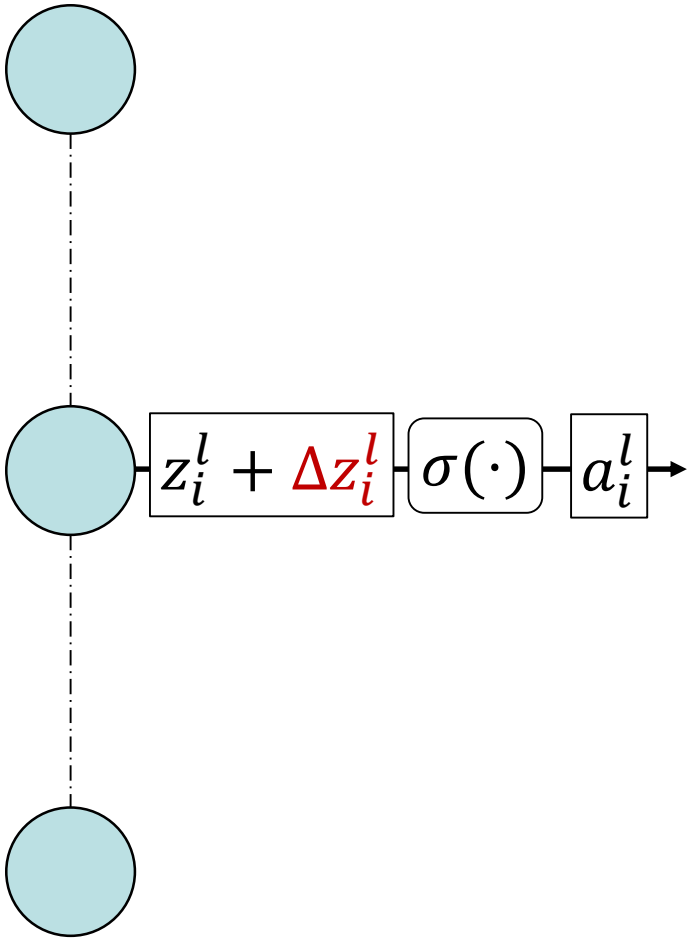
Везде далее

L – количество уровней нейронной сети

$$\vec{a}^L = \vec{\alpha}(\vec{x})$$

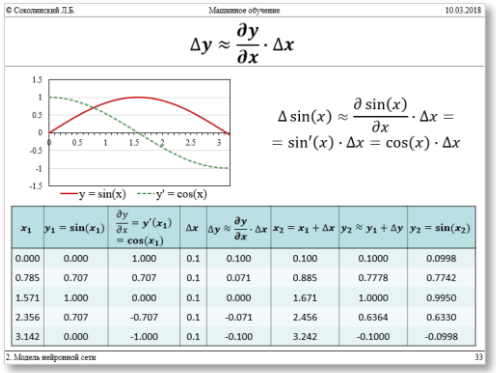
$$C = \mathbb{C}_{(\vec{x},\vec{y})} = \frac{\|\vec{a}^L - \vec{y}\|^2}{2}$$

Мера влияния нейрона на величину ошибки



$$C = \frac{\|\vec{a}^L - \vec{y}\|^2}{2}$$

$$\Delta C \approx \frac{\partial C}{\partial z_i^l} \Delta z_i^l$$

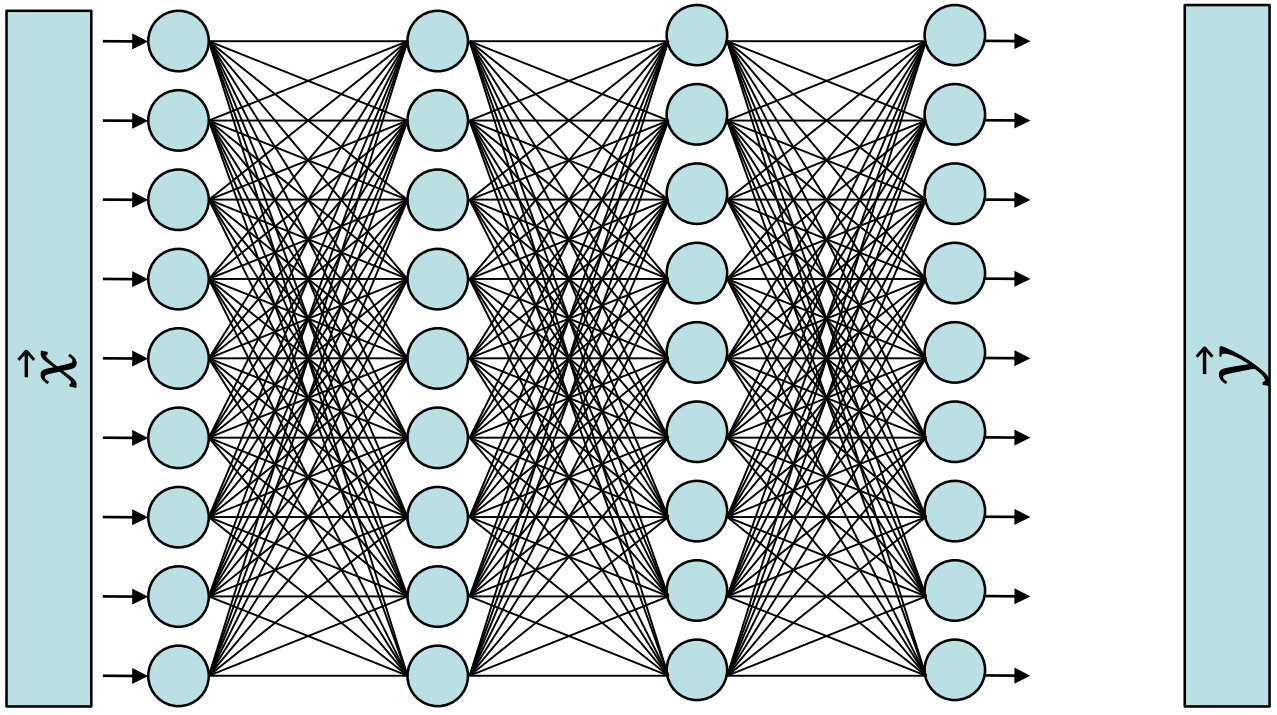


x_1	$y_1 = \sin(x_1)$	$\frac{\partial y}{\partial x} = y'(x_1) = \cos(x_1)$	Δx	$\Delta y \approx \frac{\partial y}{\partial x} \cdot \Delta x$	$x_2 = x_1 + \Delta x$	$y_2 \approx y_1 + \Delta y$	$y_2 = \sin(x_2)$
0.000	0.000	1.000	0.1	0.100	0.100	0.1000	0.0998
0.785	0.707	0.707	0.1	0.071	0.885	0.7778	0.7742
1.571	1.000	0.000	0.1	0.000	1.671	1.0000	0.9950
2.356	0.707	-0.707	0.1	-0.071	2.456	0.6364	0.6330
3.142	0.000	-1.000	0.1	-0.100	3.242	-0.1000	-0.0998

$$\delta_i^l = \frac{\partial C}{\partial z_i^l}$$

Идея метода обратного распространения ошибки

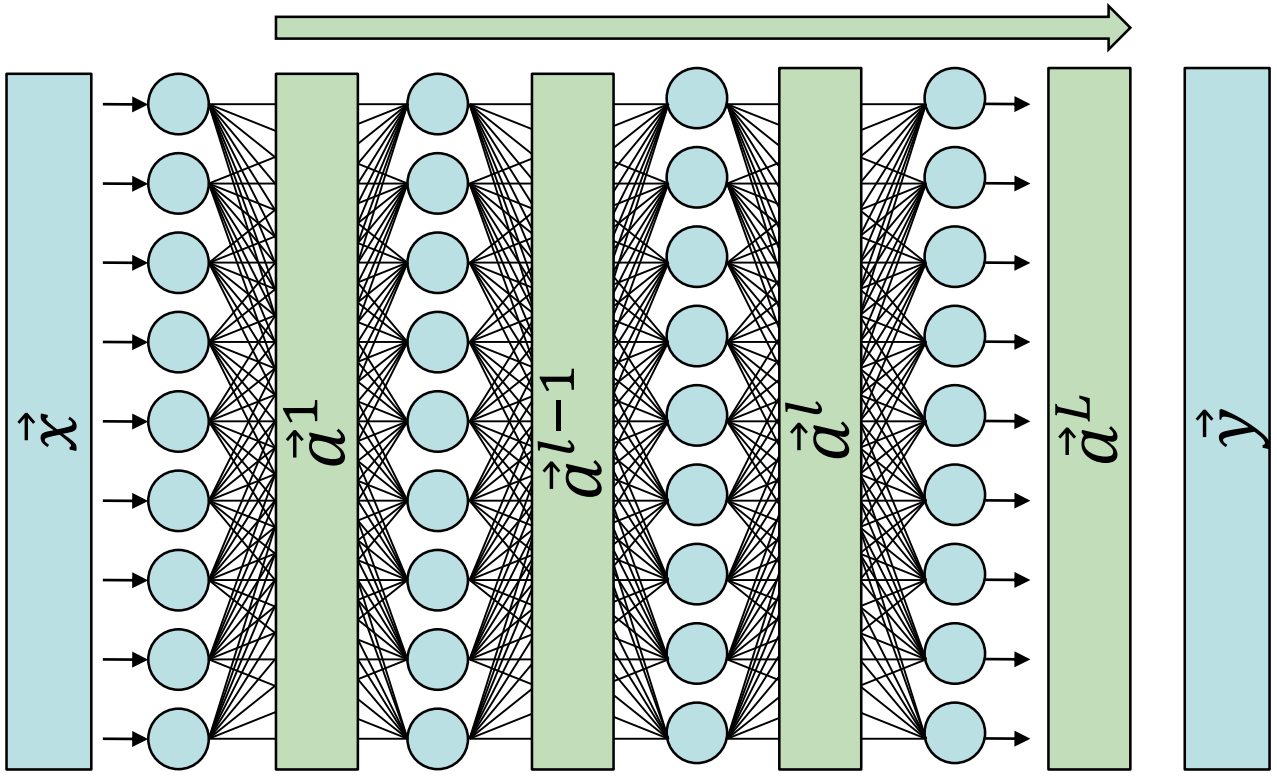
Шаг 1. Взять образец $(\vec{x}, \vec{y})$ и подать сигнал $\vec{x}$ на вход нейронной сети



Идея метода обратного распространения ошибки

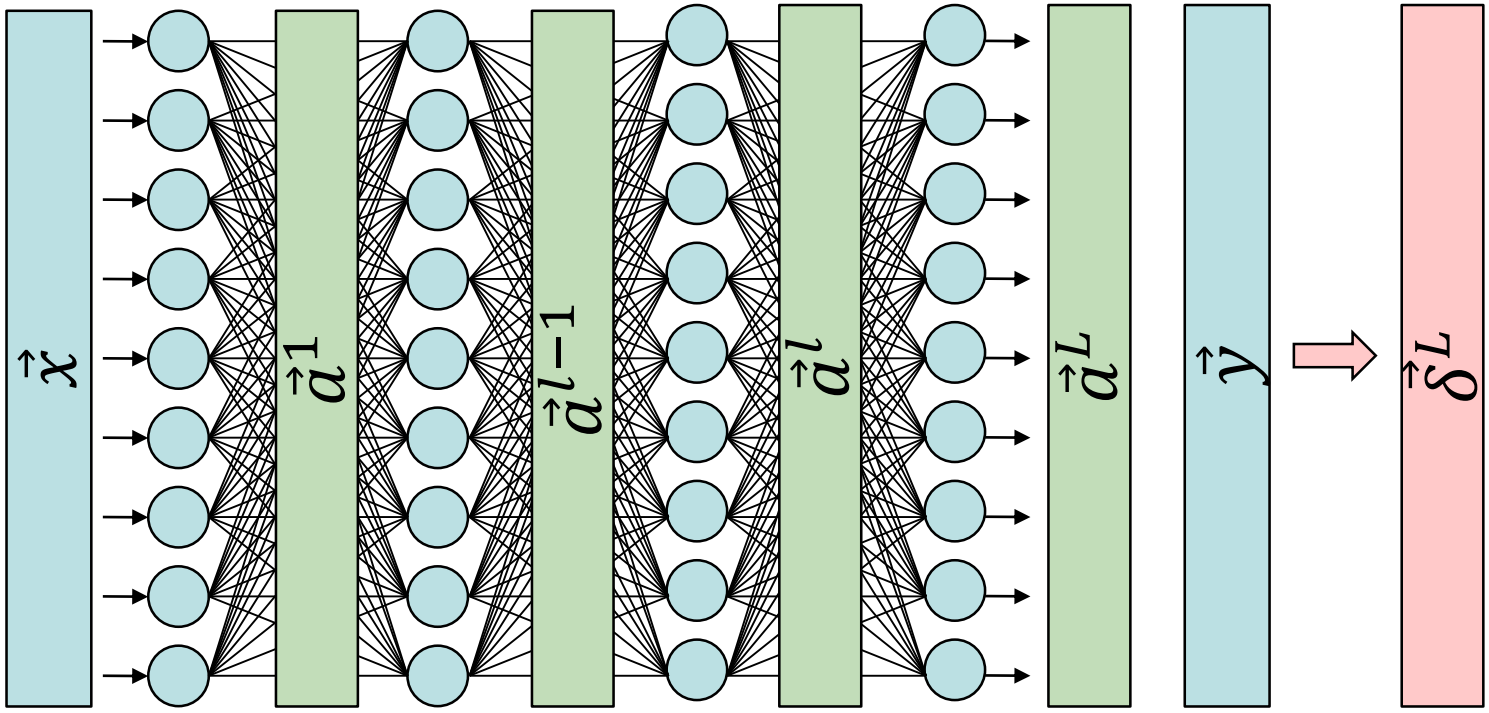
Шаг 2. Последовательно вычислить выходные сигналы $\vec{a}^l$ для каждого слоя:

$$\vec{z}^l = W^l \cdot \vec{a}^{l-1^T} + \vec{b}^l$$



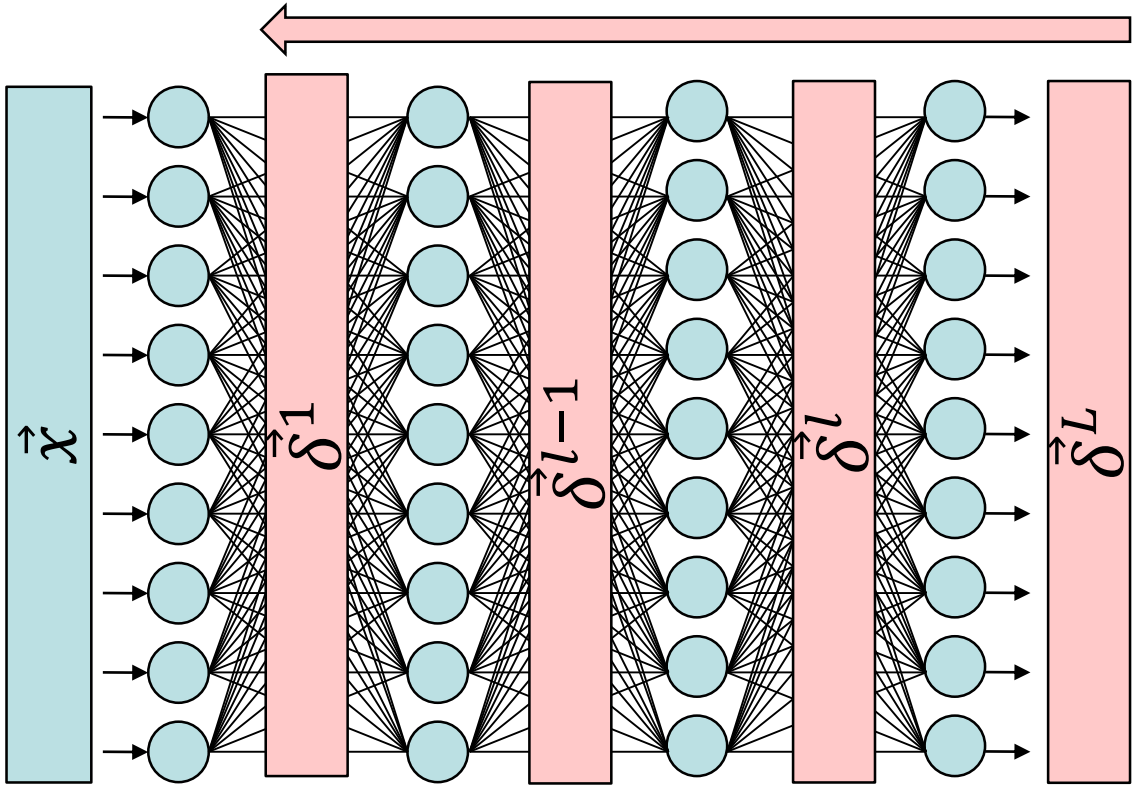
Идея метода обратного распространения ошибки

Шаг 3. Вычислить меру влияния $\vec{\delta}^L$ нейронов выходного слоя L на ошибку C



Идея метода обратного распространения ошибки

Шаг 4. Вычислить в обратном порядке меру влияния $\vec{\delta}^l$ на ошибку для каждого слоя



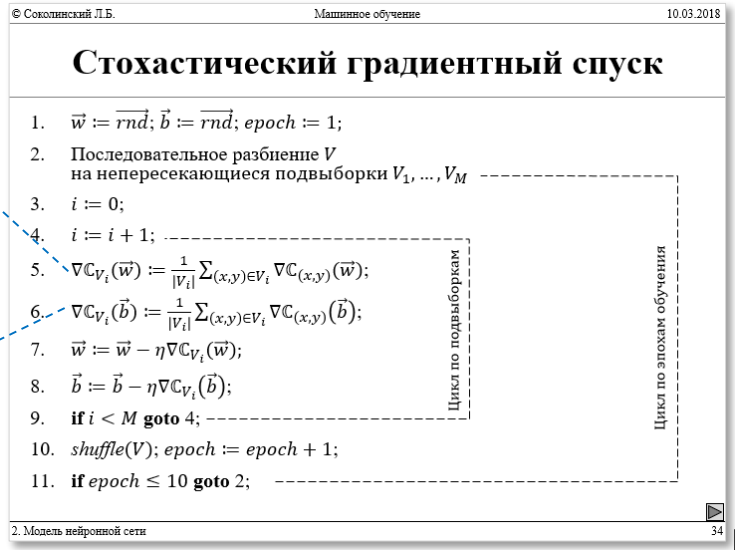
Идея метода обратного распространения ошибки

Шаг 5. Используя $\vec{\delta}^l$, вычислить $\nabla C(W^l)$ и $\nabla C(\vec{b}^l)$ для каждого слоя l

Шаг 6. Выполнив шаги 1-5 для всей подвыборки V_i, вычислить

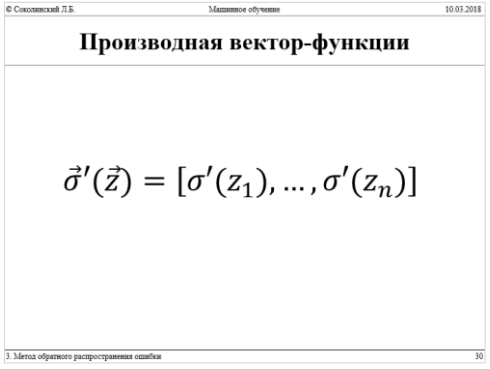
$$\nabla \mathbb{C}_{V_i}(\vec{w}) := \frac{1}{|V_i|} \sum_{(x,y)\in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{w})$$

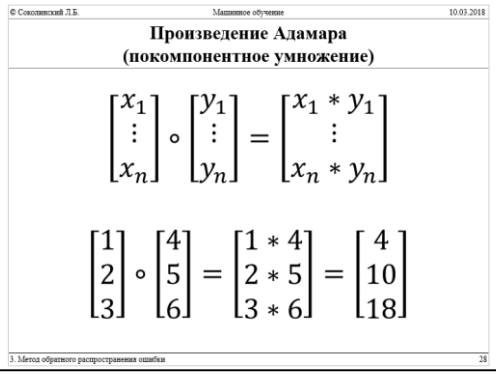
$$\nabla \mathbb{C}_{V_i}(\vec{b}) := \frac{1}{|V_i|} \sum_{(x,y)\in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{b})$$



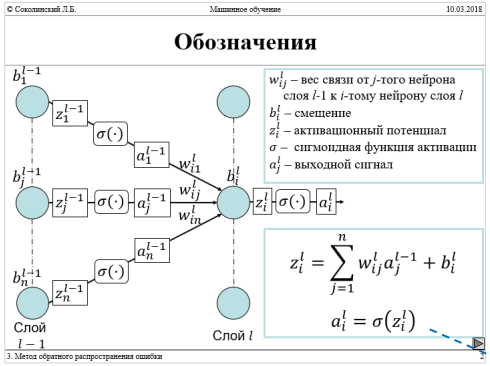
Мера влияния нейронов выходного слоя L на величину ошибки

$$\vec{\delta}^L = (\vec{a}^L - \vec{y}) \circ \vec{\sigma}'(\vec{z}^L)$$ BP1'





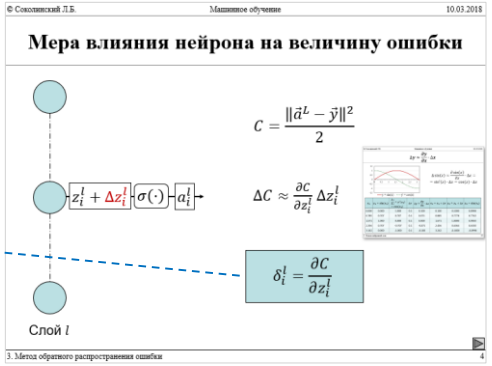
Доказательство BP1'

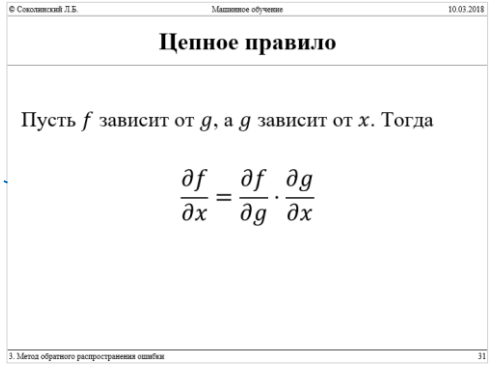


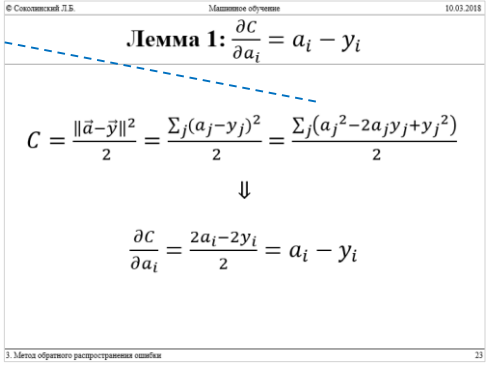
$$\delta_i^L = \frac{\partial C}{\partial z_i^L} =$$

$$= \frac{\partial C}{\partial a_i^L} \cdot \frac{\partial a_i^L}{\partial z_i^L} =$$

$$= (a_i^L - y_i) \cdot \sigma'(z_i^L)$$





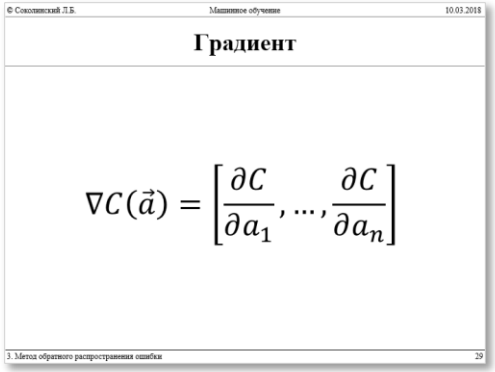


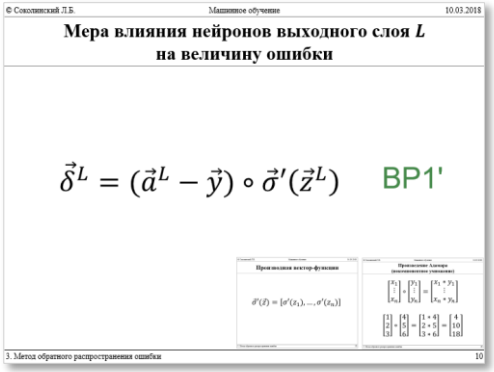
То же самое в векторной форме

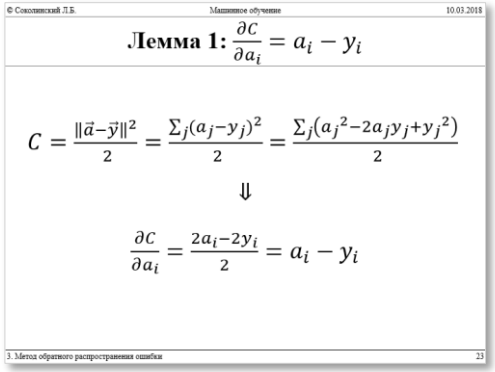
$$\vec{\delta}^L = (\vec{a}^L - \vec{y}) \circ \vec{\sigma}'(\vec{z}^L)$$

Обобщение BP1'

$$\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L)$$ BP1



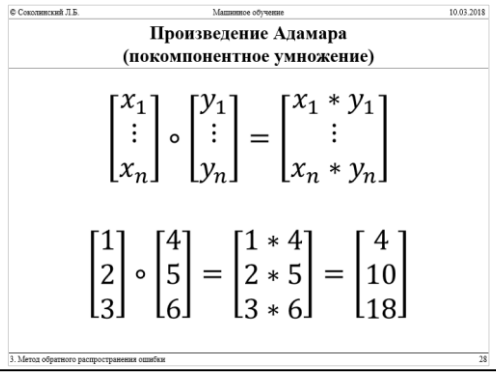




Мера влияния нейронов слоя l на величину ошибки

$$\vec{\delta}^{l} = \left(\left(W^{l+1} \right)^{T} \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'\left(\vec{z}^{l} \right)$$

BP2



Доказательство BP2

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} =$$

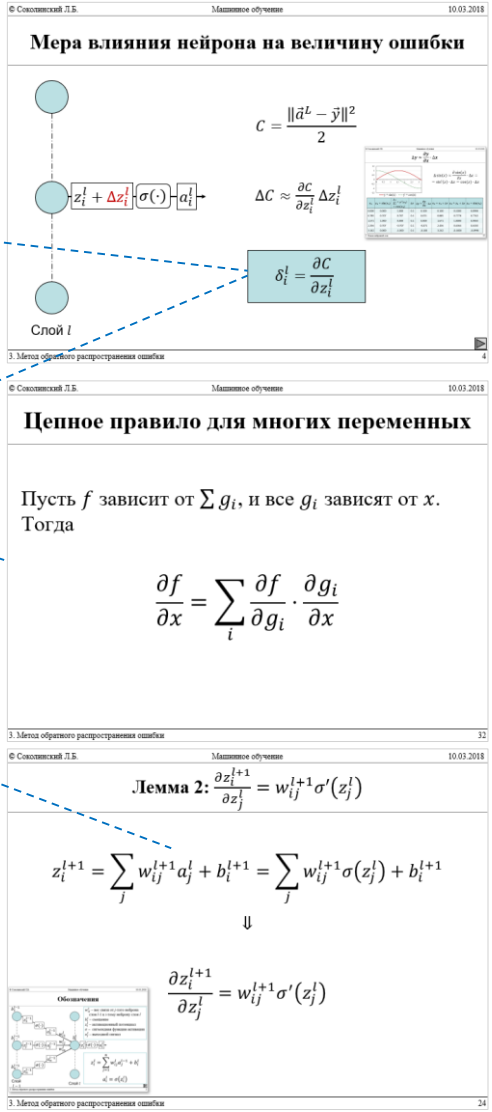
$$= \sum_i \frac{\partial C}{\partial z_i^{l+1}} \cdot \frac{\partial z_i^{l+1}}{\partial z_j^l} =$$

$$= \sum_i \frac{\partial z_i^{l+1}}{\partial z_j^l} \cdot \frac{\partial C}{\partial z_i^{l+1}} = \sum_i w_{ij}^{l+1} \sigma'(z_j^l) \cdot \delta_i^{l+1} =$$

$$= \sum_i w_{ij}^{l+1} \delta_i^{l+1} \sigma'(z_j^l)$$

То же самое в векторной форме

$$\vec{\delta}^l = \left((W^{l+1})^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'(\vec{z}^l)$$



Формула для вычисления градиента по смещению

$$\nabla C\left(\vec{b}^{\,l}\right) = \vec{\delta}^{\,l}$$

BP3

Доказательство ВРЗ

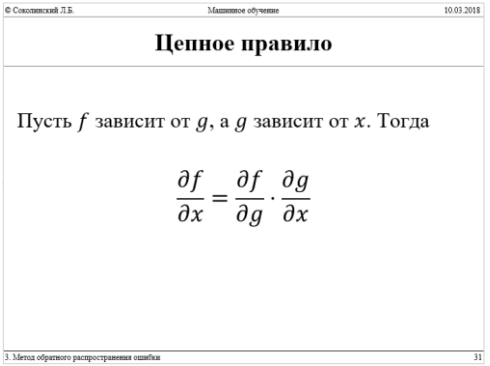
$$\frac{\partial C}{\partial b_i^l} = \frac{\partial C}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial b_i^l} =$$

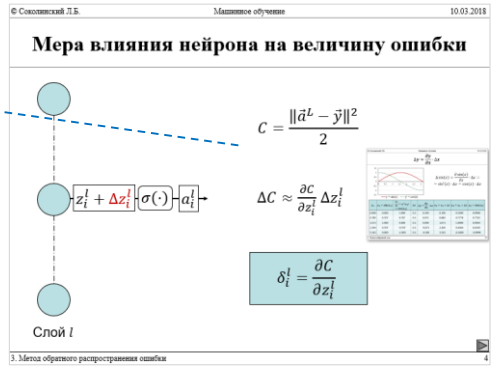
$$= \delta_i^l \cdot \frac{\partial z_i^l}{\partial b_i^l} =$$

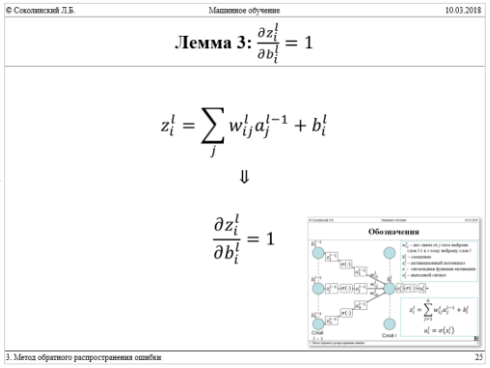
$$= \delta_i^l$$

То же самое в векторной форме

$$\nabla C(\vec{b}) = \vec{\delta}^l$$







Формула для вычисления градиента по весам

$$\nabla C(W^l) = \vec{\delta}^{l\,T} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix}$$

BP4

Доказательство BP4

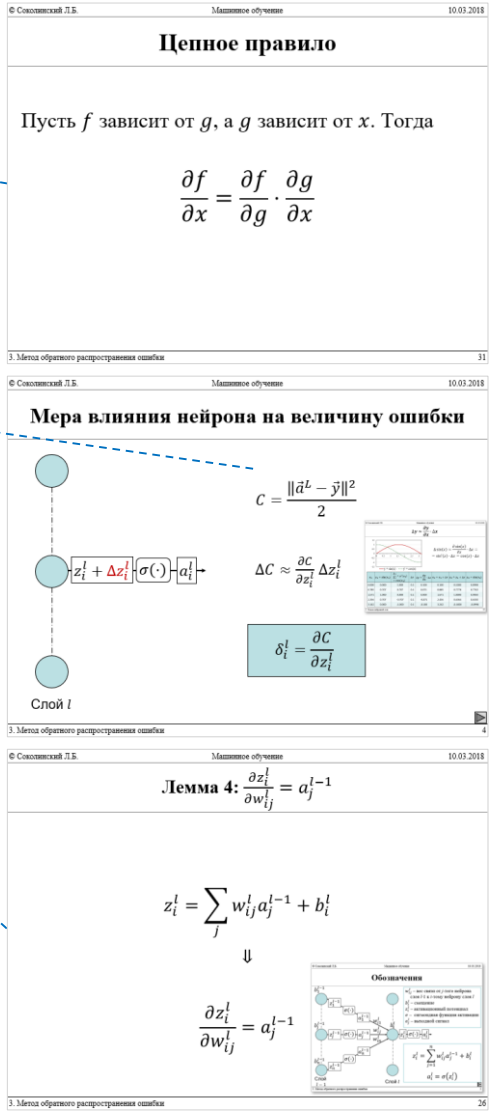
$$\frac{\partial C}{\partial w_{ij}^l} = \frac{\partial C}{\partial z_i^l} \cdot \frac{\partial z_i^l}{\partial w_{ij}^l} =$$

$$= \delta_i^l \cdot \frac{\partial z_i^l}{\partial w_{ij}^l} =$$

$$= \delta_i^l \cdot a_j^{l-1}$$

То же самое в векторной форме

$$\nabla C(W^l) = \vec{\delta}^{l\,T} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix}$$



Формулы обратного распространения ошибки

$$\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L) \quad \text{(BP1)}$$

$$\vec{\delta}^l = \left(\left(W^{l+1} \right)^T \vec{\delta}^{l+1} \right) \circ \vec{\sigma}'\left(\vec{z}^l \right) \quad \text{(BP2)}$$

$$\nabla C\left(\vec{b}^l \right) = \vec{\delta}^l \quad \text{(BP3)}$$

$$\nabla C\left(W^l \right) = \vec{\delta}^{l^T} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix} \quad \text{(BP4)}$$

Алгоритм обратного распространения ошибки

1. **Вход** x: Установить соответствующие значения активации a^1 для входного уровня
2. **Прямое распространение**: Для $l = 2, \dots, L$ последовательно вычислить $\vec{z}^l = W^l\left(\vec{a}^{l-1}\right) + \vec{b}^l$ и $\vec{a}^l = \vec{\sigma}\left(\vec{z}^l\right)$
3. **Вычислить**: $\vec{\delta}^L = \nabla C(\vec{a}) \circ \vec{\sigma}'(\vec{z}^L)$
4. **Обратное распространение**: Для $l = L - 1, \dots, 2$ последовательно вычислить $\vec{\delta}^l = \left(\left(W^{l+1}\right)^T \vec{\delta}^{l+1}\right) \circ \vec{\sigma}'\left(\vec{z}^l\right)$
5. **Выход**: Для $l = 2, \dots, L$ вычислить

$$\nabla \mathbb{C}_{(x,y)}\left(W^l\right) = \vec{\delta}^{l^T} \circ \begin{pmatrix} \vec{a}^{l-1} \\ \vdots \\ \vec{a}^{l-1} \end{pmatrix}$$

$$\nabla \mathbb{C}_{(x,y)}\left(\vec{b}^l\right) = \vec{\delta}^l$$

Конец лекции 3

Вспомогательные слайды

Лемма 1: $\frac{\partial C}{\partial a_i} = a_i - y_i$

$$C = \frac{\|\vec{a}-\vec{y}\|^2}{2} = \frac{\sum_j (a_j - y_j)^2}{2} = \frac{\sum_j \left(a_j{}^2 - 2a_j y_j + y_j{}^2\right)}{2}$$

$$\Downarrow$$

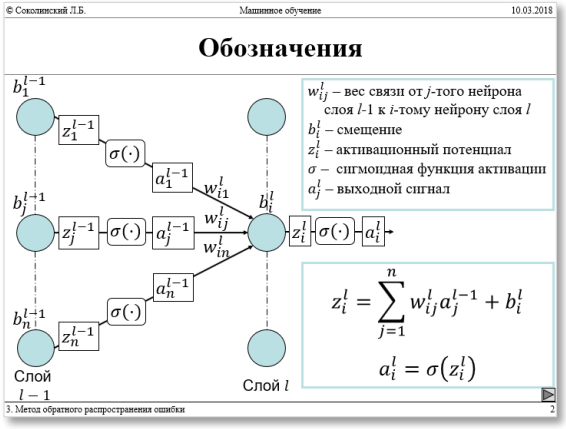
$$\frac{\partial C}{\partial a_i} = \frac{2a_i - 2y_i}{2} = a_i - y_i$$

Лемма 2: $\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1}\sigma'\left(z_j^l\right)$

$$z_i^{l+1} = \sum_j w_{ij}^{l+1} a_j^l + b_i^{l+1} = \sum_j w_{ij}^{l+1}\sigma\left(z_j^l\right) + b_i^{l+1}$$

$$\Downarrow$$

$$\frac{\partial z_i^{l+1}}{\partial z_j^l} = w_{ij}^{l+1}\sigma'\left(z_j^l\right)$$

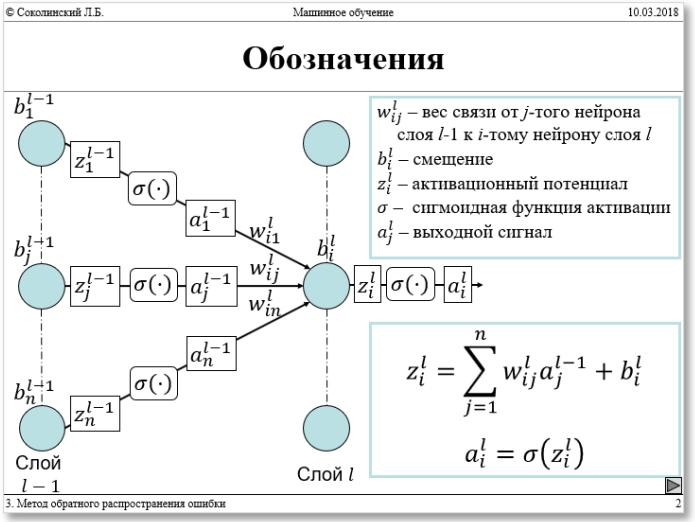


Лемма 3: $\frac{\partial z_i^l}{\partial b_i^l} = 1$

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l$$

$$\Downarrow$$

$$\frac{\partial z_i^l}{\partial b_i^l} = 1$$

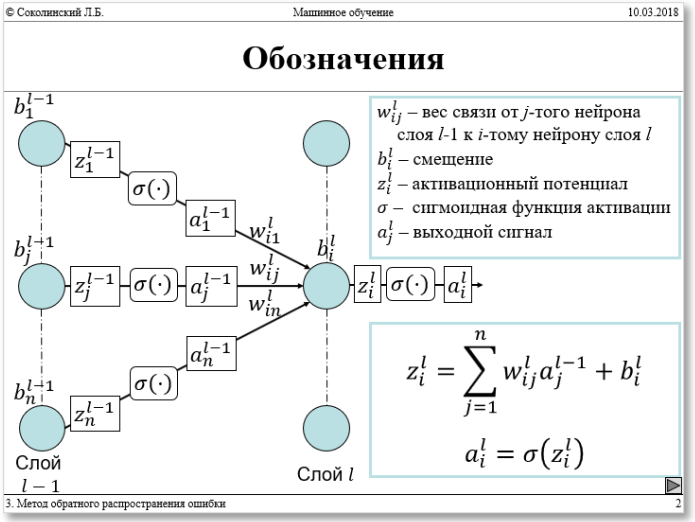


Лемма 4: $\frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l$$

$$\Downarrow$$

$$\frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$$



Транспонирование

$$[x_1 \dots x_n]^T = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$\begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}^T = \begin{bmatrix} w_{11} & \cdots & w_{m1} \\ \vdots & \ddots & \vdots \\ w_{1n} & \cdots & w_{mn} \end{bmatrix}$$

Произведение Адамара (покомпонентное умножение)

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \circ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 * y_1 \\ \vdots \\ x_n * y_n \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \circ \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 * 4 \\ 2 * 5 \\ 3 * 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}$$

Градиент

$$\nabla C(\vec{a}) = \left[\frac{\partial C}{\partial a_1}, \ldots, \frac{\partial C}{\partial a_n}\right]$$

Производная вектор-функции

$$\vec{\sigma}'(\vec{z}) = [\sigma'(z_1), \dots, \sigma'(z_n)]$$

Цепное правило

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

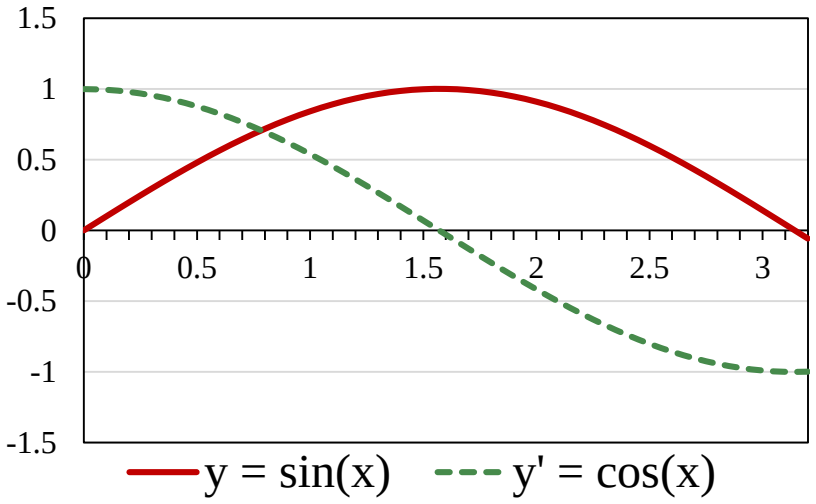
$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial g} \cdot \frac{\partial g}{\partial x}$$

Цепное правило для многих переменных

Пусть f зависит от $\sum g_i$, и все g_i зависят от x. Тогда

$$\frac{\partial f}{\partial x} = \sum_i \frac{\partial f}{\partial g_i} \cdot \frac{\partial g_i}{\partial x}$$

$$\Delta y \approx \frac{\partial y}{\partial x} \cdot \Delta x$$



$$\Delta \sin(x) \approx \frac{\partial \sin(x)}{\partial x} \cdot \Delta x = \sin'(x) \cdot \Delta x = \cos(x) \cdot \Delta x$$

x_1	$y_1 = \sin(x_1)$	$\frac{\partial y}{\partial x} = y'(x_1) = \cos(x_1)$	Δx	$\Delta y \approx \frac{\partial y}{\partial x} \cdot \Delta x$	$x_2 = x_1 + \Delta x$	$y_2 \approx y_1 + \Delta y$	$y_2 = \sin(x_2)$
0.000	0.000	1.000	0.1	0.100	0.100	0.1000	0.0998
0.785	0.707	0.707	0.1	0.071	0.885	0.7778	0.7742
1.571	1.000	0.000	0.1	0.000	1.671	1.0000	0.9950
2.356	0.707	-0.707	0.1	-0.071	2.456	0.6364	0.6330
3.142	0.000	-1.000	0.1	-0.100	3.242	-0.1000	-0.0998

Стохастический градиентный спуск

1. $\vec{w} \coloneqq \overrightarrow{rnd}; \vec{b} \coloneqq \overrightarrow{rnd}; epoch \coloneqq 1;$
2. Последовательное разбиение V на непересекающиеся подвыборки $V_1, \dots, V_M$
3. $i \coloneqq 0;$
4. $i \coloneqq i + 1;$
5. $\nabla \mathbb{C}_{V_i}(\vec{w}) \coloneqq \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{w});$
6. $\nabla \mathbb{C}_{V_i}(\vec{b}) \coloneqq \frac{1}{|V_i|} \sum_{(x,y) \in V_i} \nabla \mathbb{C}_{(x,y)}(\vec{b});$
7. $\vec{w} \coloneqq \vec{w} - \eta \nabla \mathbb{C}_{V_i}(\vec{w});$
8. $\vec{b} \coloneqq \vec{b} - \eta \nabla \mathbb{C}_{V_i}(\vec{b});$
9. **if** $i < M$ **goto** 4;
10. $shuffle(V); epoch \coloneqq epoch + 1;$
11. **if** $epoch \leq 10$ **goto** 2;

Цикл по подвыборкам

Цикл по эпохам обучения