Глубокие нейронные сети

Градиентный спуск (Gradient descent)

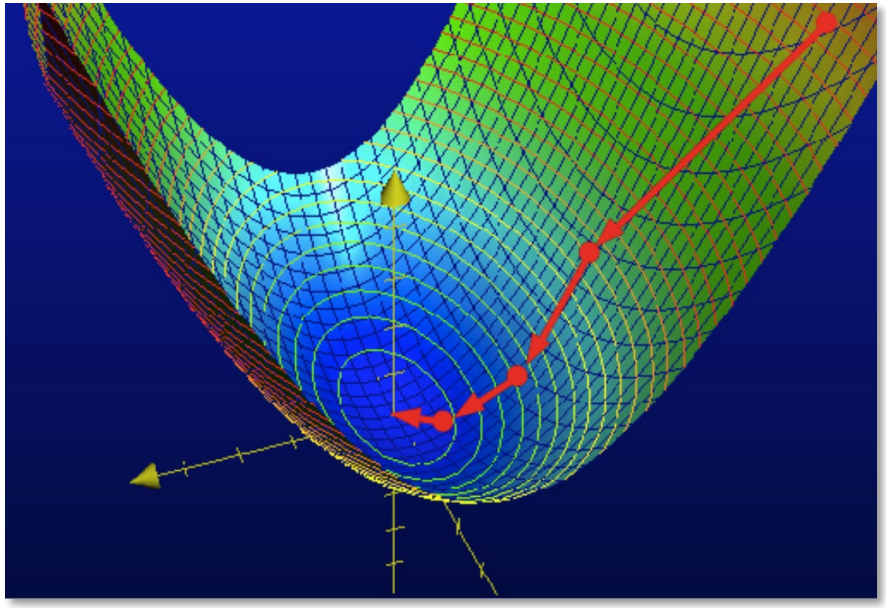
Лекция 3

Метод градиентного спуска (двумерный случай)

- Найти минимум функции $f(x_1, x_2)$: $\mathbb{R}^2 \to \mathbb{R}$
- Градиент:

$$\nabla f = \left(\frac{\partial f(x_1, x_2)}{\partial x_1}; \frac{\partial f(x_1, x_2)}{\partial x_2} \right)$$

- $\Delta \boldsymbol{x} := -\eta \nabla f$,
  где $\eta > 0$ – малый параметр (*скорость обучения*)



∇ – символ «набла»

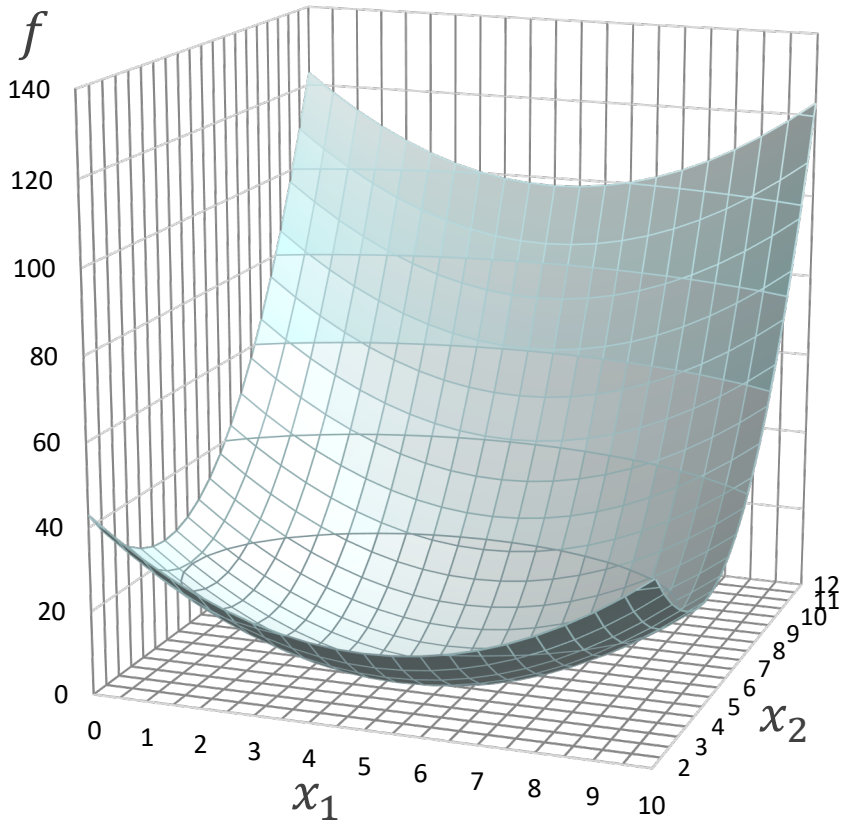
Пример вычисления градиента

$$f(x_1, x_2) = {x_1}^2 - 10x_1 + 2{x_2}^2 - 20x_2 + 75$$

$$\frac{\partial f(x_1,x_2)}{\partial x_1} = \ ?$$

$$\frac{\partial f(x_1,x_2)}{\partial x_2} = \ ?$$

$$\nabla f = ?$$



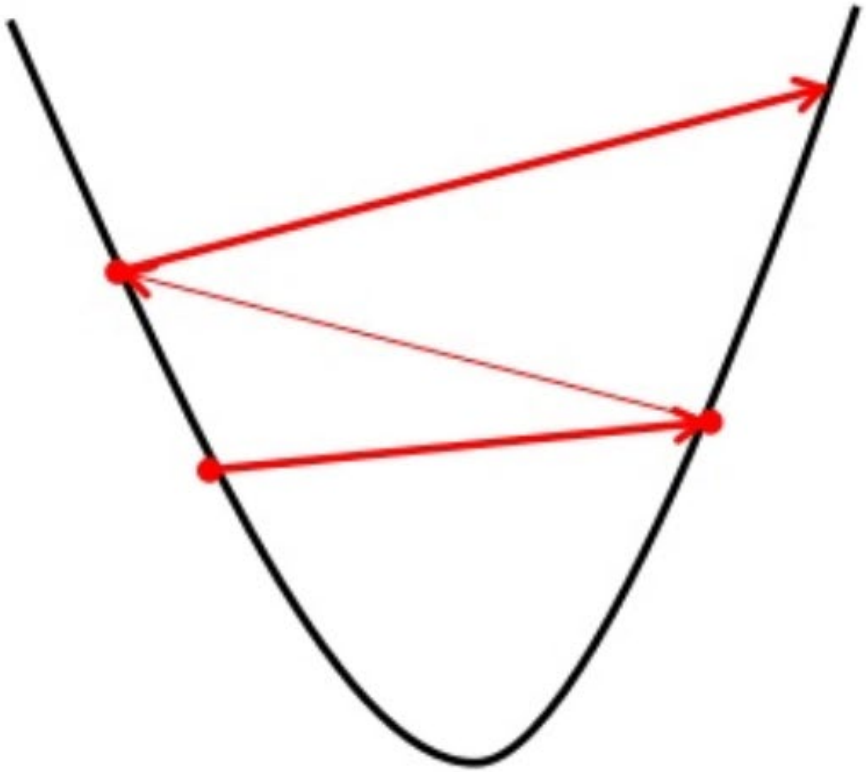
Нахождение минимума функции f методом градиентного спуска

$$\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}, \boldsymbol{x}^{(n+1)} \ldots$$

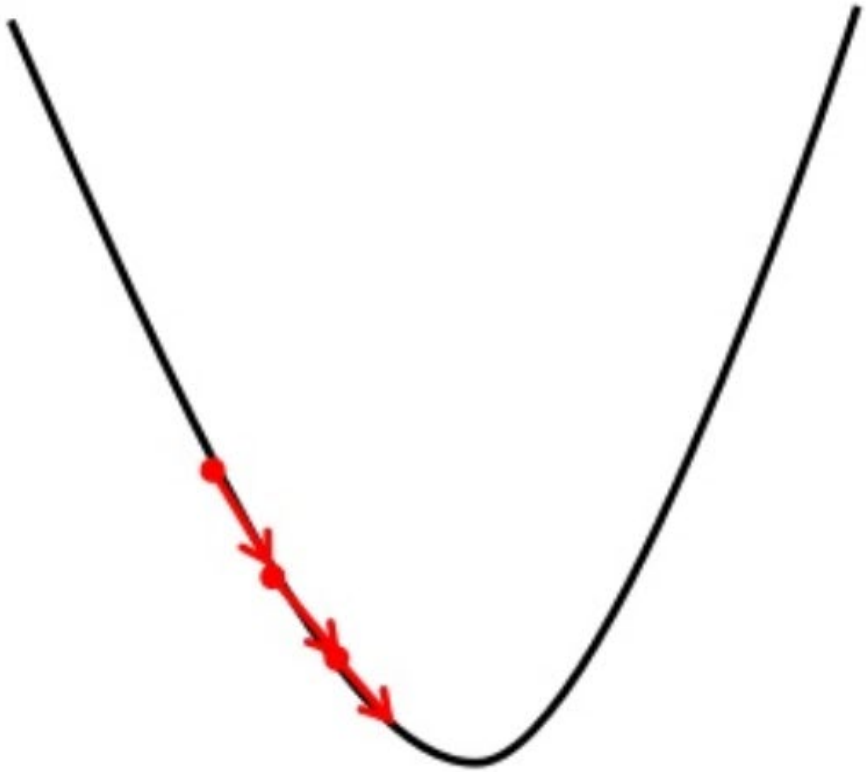
$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \eta \nabla f\left(\boldsymbol{x}^{(n)}\right)$$

$\eta > 0$ – *скорость обучения*

Выбор скорости обучения η

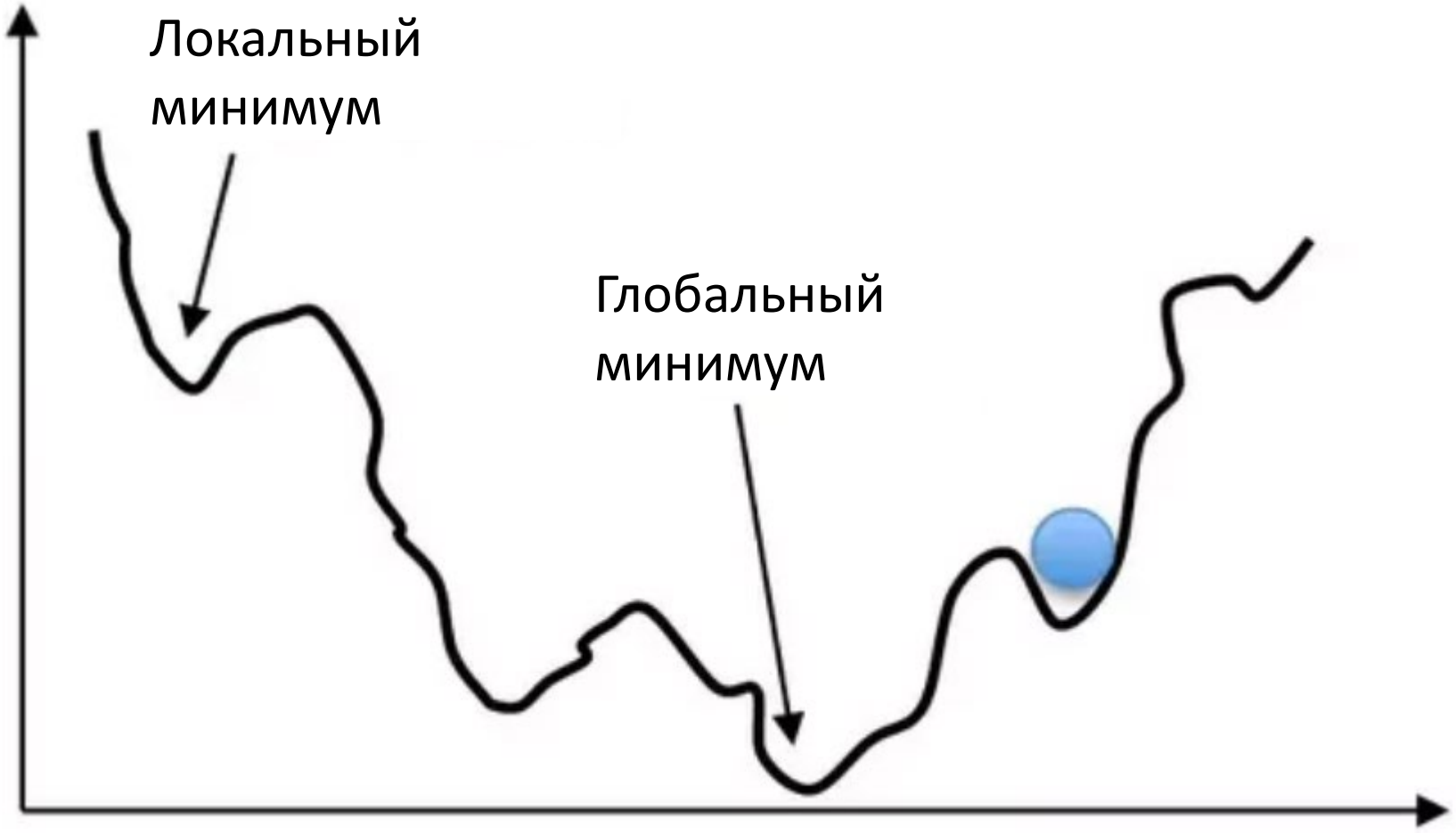


Большая скорость обучения



Малая скорость обучения

Высокая скорость обучения предотвращает остановку в локальном минимуме



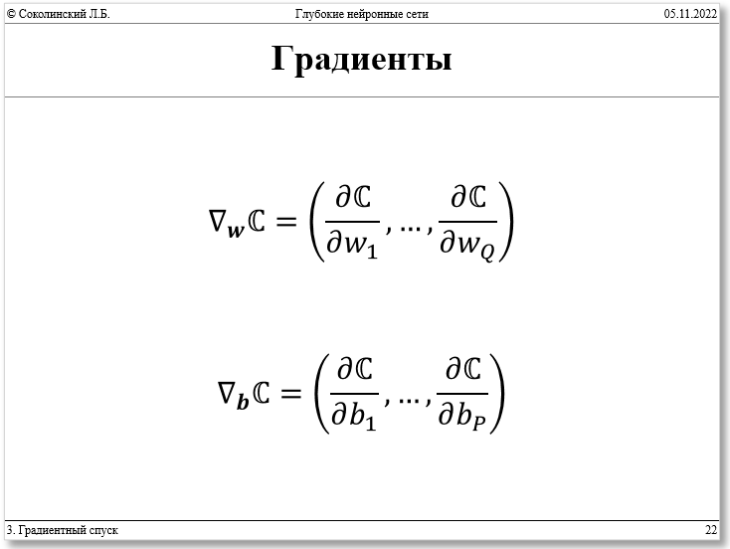
Обучение нейронной сети методом градиентного спуска

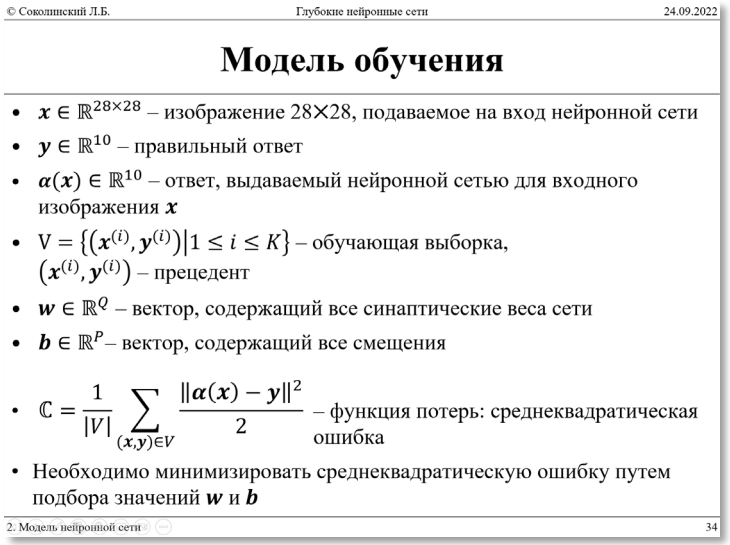
Необходимо минимизировать среднеквадратичную ошибку:

$$\mathbb{C} = \frac{1}{|V|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in V} \frac{\|\boldsymbol{\alpha}(\boldsymbol{x}) - \boldsymbol{y}\|^2}{2}$$

$$\boldsymbol{w} := \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} \mathbb{C}$$

$$\boldsymbol{b} := \boldsymbol{b} - \eta \nabla_{\boldsymbol{b}} \mathbb{C}$$





Проблема

Высокая вычислительная сложность процесса обучения

60 000 прецедентов

10 000 000 весов

100 000 смещений

$$\mathbb{C} = \frac{1}{|V|} \sum_{(x,y) \in V} \frac{\|\boldsymbol{\alpha}(x) - y\|^2}{2}$$

$$\boldsymbol{w} := \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} \mathbb{C}$$

$$\boldsymbol{b} := \boldsymbol{b} - \eta \nabla_{\boldsymbol{b}} \mathbb{C}$$

Переход к векторно-матричным операциям

m – количество нейронов в слое l

n – количество нейронов в слое l-1

$$W^{(l)} = \begin{bmatrix} w_{11}^{(l)} & \cdots & w_{1m}^{(l)} \\ \vdots & \ddots & \vdots \\ w_{n1}^{(l)} & \cdots & w_{nm}^{(l)} \end{bmatrix}$$

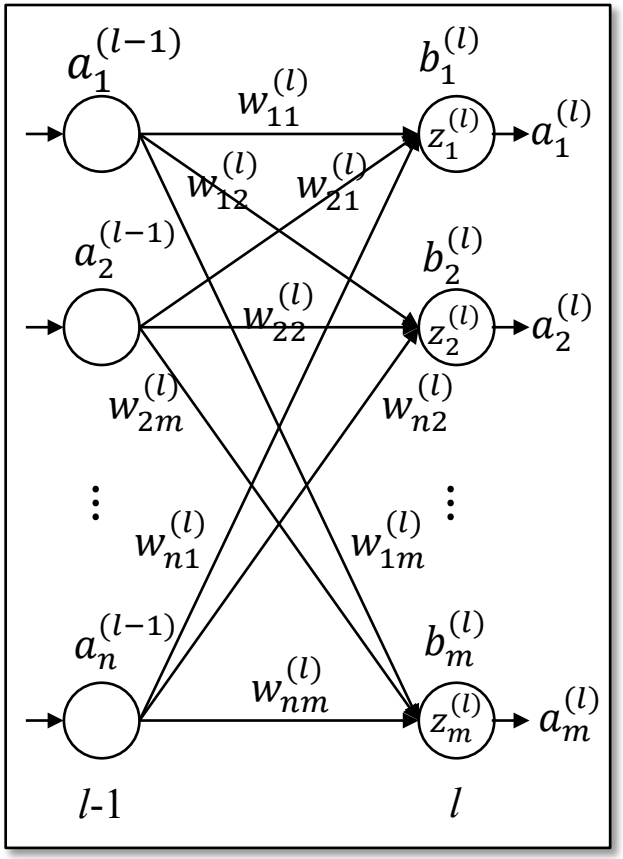
$$\boldsymbol{b}^{(l)} = \begin{bmatrix} b_1^{(l)} & \cdots & b_m^{(l)} \end{bmatrix}^\top \qquad \boldsymbol{z}^{(l)} = \begin{bmatrix} z_1^{(l)} & \cdots & z_m^{(l)} \end{bmatrix}^\top$$

$$\boldsymbol{a}^{(l-1)} = \begin{bmatrix} a_1^{(l-1)} & \cdots & a_n^{(l-1)} \end{bmatrix}^\top \quad \boldsymbol{a}^{(l)} = \begin{bmatrix} a_1^{(l)} & \cdots & a_m^{(l)} \end{bmatrix}^\top$$

$$\boldsymbol{\sigma}\left(\boldsymbol{z}^{(l)}\right) = \begin{bmatrix} \sigma\left(z_1^{(l)}\right) & \cdots & \sigma\left(z_m^{(l)}\right) \end{bmatrix}^\top$$

$$\boldsymbol{z}^{(l)} = \boldsymbol{W}^{(l)\top} \boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}$$

$$\boldsymbol{a}^{(l)} = \boldsymbol{\sigma}\left(\boldsymbol{z}^{(l)}\right)$$



Преобразование формулы среднеквадратичной ошибки

Имеем $$\mathbb{C} = \frac{1}{|V|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in V} \frac{\|\boldsymbol{\alpha}(\boldsymbol{x}) - \boldsymbol{y}\|^2}{2}$$

Обозначим $$C_{(\boldsymbol{x},\boldsymbol{y})} = \frac{\|\boldsymbol{\alpha}(\boldsymbol{x}) - \boldsymbol{y}\|^2}{2}$$

Тогда $$\mathbb{C} = \frac{1}{|V|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in V} C_{(\boldsymbol{x},\boldsymbol{y})}$$

Следовательно

$$\boldsymbol{\nabla_w}\mathbb{C} = \frac{1}{|V|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in V} \boldsymbol{\nabla_w} C_{(\boldsymbol{x},\boldsymbol{y})}$$

$$\boldsymbol{\nabla_b}\mathbb{C} = \frac{1}{|V|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in V} \boldsymbol{\nabla_b} C_{(\boldsymbol{x},\boldsymbol{y})}$$

Стохастический градиентный спуск (SGD)

1. $\boldsymbol{w} \coloneqq \boldsymbol{rnd}$; $\boldsymbol{b} \coloneqq \boldsymbol{rnd}$ // Присваиваем случайные значения
2. **for** $epoch = 1 \ldots 10$ **do**
3. $V \to V_1, \ldots, V_M$ // Последовательно разбиваем V на подвыборки
4. **for** $i = 1 \ldots M$ **do**
5. $\nabla_{\boldsymbol{w}} \mathbb{C}_{V_i} \coloneqq \frac{1}{|V_i|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in V_i} \nabla_{\boldsymbol{w}} C_{(\boldsymbol{x},\boldsymbol{y})}$
6. $\nabla_{\boldsymbol{b}} \mathbb{C}_{V_i} \coloneqq \frac{1}{|V_i|} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in V_i} \nabla_{\boldsymbol{b}} C_{(\boldsymbol{x},\boldsymbol{y})}$
7. $\boldsymbol{w} \coloneqq \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} \mathbb{C}_{V_i}$
8. $\boldsymbol{b} \coloneqq \boldsymbol{b} - \eta \nabla_{\boldsymbol{b}} \mathbb{C}_{V_i}$
9. **end for**
10. $shuffle(V)$ // Перемешиваем V
11. **end for**

Цикл по подвыборкам

Цикл по эпохам обучения

Градиенты

$$\nabla_{\boldsymbol{w}}\mathbb{C} = \left(\frac{\partial\mathbb{C}}{\partial w_1}, \ldots, \frac{\partial\mathbb{C}}{\partial w_Q}\right)$$

$$\nabla_{\boldsymbol{b}}\mathbb{C} = \left(\frac{\partial\mathbb{C}}{\partial b_1}, \ldots, \frac{\partial\mathbb{C}}{\partial b_P}\right)$$

Проблема

- Необходим эффективный алгоритм вычисления градиентов

$$\nabla_{\boldsymbol{w}} C_{(\boldsymbol{x},\boldsymbol{y})}$$

$$\nabla_{\boldsymbol{b}} C_{(\boldsymbol{x},\boldsymbol{y})}$$

Конец лекции 3