

# Градиент.

Любую функцию от  $n$  переменных можно считать функцией из  $\mathbb{R}^n$  в  $\mathbb{R}$ . Пусть у нас есть величина  $q(\mathbf{u})$ , зависящая от вектора  $\mathbf{u} \in \mathbb{R}^n$ . Будем говорить, что  $q(\mathbf{u}) = o(\mathbf{u})$  (читается " $q(\mathbf{u})$  есть о маленькое от  $\mathbf{u}$ "), если для любой последовательности векторов  $\mathbf{u}_1, \mathbf{u}_2, \dots \in \mathbb{R}$  такой, что  $\|\mathbf{u}_i\| \rightarrow 0$ , последовательность  $q(\mathbf{u}_i)$  стремится к нулю.

Функция  $f$  от  $n$  переменных **дифференцируема** в точке  $\mathbf{x}$ , если существует такой вектор  $\nabla f(\mathbf{x}) \in \mathbb{R}^n$ , что для любого  $\mathbf{v} \in \mathbb{R}^n$

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \langle \mathbf{v}, \nabla f(\mathbf{x}) \rangle + o(\mathbf{v})$$

Вектор  $\nabla f(\mathbf{x})$  называется **градиентом** функции  $f$  в точке  $\mathbf{x}$ .

**Задача 1.** Убедитесь, что для  $n = 1$  новое определение дифференцируемости совпадает с уже знакомым определением дифференцируемости функции из  $\mathbb{R}$  в  $\mathbb{R}$ .

**Задача 2.** Пусть  $f$  и  $g$  – дифференцируемые функции из  $\mathbb{R}$  в  $\mathbb{R}$  и из  $\mathbb{R}^n$  в  $\mathbb{R}$  соответственно. Докажите, что  $f(g(\mathbf{x}))$  – дифференцируемая функция из  $\mathbb{R}^n$  в  $\mathbb{R}$ .

**Задача 3.** Пусть  $f$  и  $g$  – дифференцируемые функции из  $\mathbb{R}^n$  в  $\mathbb{R}$ . Докажите, что следующие функции дифференцируемы (и выразите их градиенты, через градиенты  $f$  и  $g$ ):

а)  $f + g$

б)  $f \cdot g$

в)  $f/g$  (в точках  $\mathbf{x}$ , где  $g(\mathbf{x}) \neq 0$ )

**Задача 4.** Докажите, что  $f(g_1(\mathbf{x}), \dots, g_m(\mathbf{x}))$  – всюду дифференцируемая функция из  $\mathbb{R}^n$  в  $\mathbb{R}$ , если  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$  и  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  всюду дифференцируемы.

**Задача 5.** Докажите, что следующие функции  $\mathbb{R}^n \rightarrow \mathbb{R}$  дифференцируемы по  $\mathbf{x}$ :

а)  $x_1 \cdot \dots \cdot x_n$

б)  $\sin(x_1 + \dots + x_n)$

в)  $\log(\sigma(\langle \mathbf{x}, \mathbf{w} \rangle))$ , где  $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$ .

г) И эта функция  $\mathbb{R}^n \rightarrow \mathbb{R}^n$ :  $\text{softmax}(\mathbf{x}) = \frac{(e^{x_1}, \dots, e^{x_n})}{e^{x_1} + \dots + e^{x_n}}$

**Частная производная** функции  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  по  $i$ -той переменной в точке  $\mathbf{x}$  это:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = \lim_{\varepsilon \rightarrow 0} \frac{f(x_1, \dots, x_i + \varepsilon, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{\varepsilon}$$

Иными словами мы фиксируем все переменные кроме  $i$ -той, рассматриваем  $f$  как функцию от одной переменной и берем ее производную в точке  $x_i$ .

**Задача 6.** Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  дифференцируема в точке  $\mathbf{x}$ . Докажите что:

$$\nabla f(\mathbf{x}) = \left( \frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right).$$

**Задача 7.** Пусть мы находимся в точке  $\mathbf{a}$ ,  $\nabla f(\mathbf{a}) = \mathbf{g} \neq \vec{0}$  и  $\|\mathbf{v}\|$ . Предположим  $\delta \rightarrow 0$ .

а) Убедитесь что:  $f(\mathbf{a} + \mathbf{v}) = f(\mathbf{a}) + \langle \mathbf{v}, \mathbf{g} \rangle + o(\delta)$

б) Для любого вектора  $\mathbf{v}$  обозначим через  $\mathbf{v}_\delta$  вектор  $\delta \cdot \frac{\mathbf{g}}{\|\mathbf{g}\|}$  (шаг в направлении  $\mathbf{v}$  длины  $\delta$ ). Попробуем сдвинуться на  $\delta$  так, чтобы функция  $f$  возросла как можно больше. Докажите, что для любого направления  $\mathbf{v}$ , непропорционального градиенту  $\mathbf{g}$ , для достаточно маленького  $\delta$   $f(\mathbf{a} + \mathbf{v}_\delta) < f(\mathbf{a} + \mathbf{g}_\delta)$ . Таким образом градиент указывает направление вдоль которого функция возрастает максимально быстро.

**Задача 8.** Докажите, что в локальном оптимуме градиент равен  $\vec{0}$ .

**Определение 1. Градиентный спуск** – это процесс "жадной" минимизации функции, действующий по правилу вида  $\mathbf{x}_{t+1} = \mathbf{x}_t - \lambda \cdot \nabla f(\mathbf{x}_t)$ . Здесь  $\lambda$  – это параметр называемый **learning rate**. Он как правило меняется от точки к точке (в зависимости от конкретной реализации).

**Задача 9\*.** Докажите, что  $f$  дифференцируема в точке  $\mathbf{x}$ , если ее частные производные определены в некоторой окрестности  $\mathbf{x}$  и непрерывны в  $\mathbf{x}$ .