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 ненулевых векторов $\mathbf{u}_1, \mathbf{u}_2, \ldots \in \mathbb{R}^n$ такой, что $\|\mathbf{u}_i\| \to 0$, последовательность $q(\mathbf{u}_i)/\|\mathbf{u}_i\|$ стремится к нулю.

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

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

Вектор $\mathbf{a}$ называется ***градиентом*** *функции f в точке $\mathbf{x}$ и обозначается*[1] $\nabla f(\mathbf{x})$.

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. Докажите, что в локальном оптимуме градиент равен $\overrightarrow{\mathbf{0}}$.

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

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

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 3. Пусть $f : \mathbb{R}^n \to \mathbb{R}$ дифференцируема в точке $\mathbf{x}$. Докажите что:

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

Задача 4*. Докажите, что f дифференцируема в точке $\mathbf{x}$, если ее частные производные определены в некоторой окрестности $\mathbf{x}$ и непрерывны в $\mathbf{x}$.

Задача 5. Пусть мы находимся в точке $\mathbf{a}$, $\nabla f(\mathbf{a}) = \mathbf{g} \neq \overrightarrow{\mathbf{0}}$. Для любого вектора $\mathbf{v}$ обозначим через $\mathbf{v}_\delta$ вектор $\delta \cdot \frac{\mathbf{v}}{\|\mathbf{v}\|}$ (шаг в направлении $\mathbf{v}$ длины δ). Попробуем сдвинуться на δ так, чтобы функция f возросла как можно больше. Докажите, что для любого направления $\mathbf{v}$, непропорционального градиенту $\mathbf{g}$ выполнено $f(\mathbf{a} + \mathbf{v}_\delta) < f(\mathbf{a} + \mathbf{g}_\delta)$ при достаточно маленьком δ. Таким образом, градиент указывает направление вдоль которого функция возрастает максимально быстро.

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 λ – это параметр, который в машинном обучении называют* ***learning rate****. Он как правило меняется в процессе (в зависимости от конкретной реализации).*

Задача 6. 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 7. 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 8. Докажите, что $f(g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))$ – всюду дифференцируемая функция из $\mathbb{R}^n$ в $\mathbb{R}$, если $g_i : \mathbb{R}^n \to \mathbb{R}$ и $f : \mathbb{R}^m \to \mathbb{R}$ всюду дифференцируемы.

Задача 9. Докажите, что следующие функции $\mathbb{R}^n \to \mathbb{R}$ дифференцируемы по $\mathbf{x}$:

а) $x_1 \cdot \ldots \cdot x_n$

б) $\sin(x_1 + \ldots + x_n)$

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 $\mathbb{R}^n \to \mathbb{R}^n$: $\text{softmax}(\mathbf{x}) = \frac{(e^{x_1}, \ldots, e^{x_n})}{e^{x_1} + \ldots + e^{x_n}}$

[1] Комментарий к обозначению: в записи $\nabla f(\mathbf{x})$ опущены скобки. Более подробная запись такая: $(\nabla f)(\mathbf{x})$, где ∇f – это функция из $\mathbb{R}^n$ в $\mathbb{R}^n$, которая по точке $\mathbf{x} \in \mathbb{R}^n$ выдает градиент функции f в точке $\mathbf{x}$.