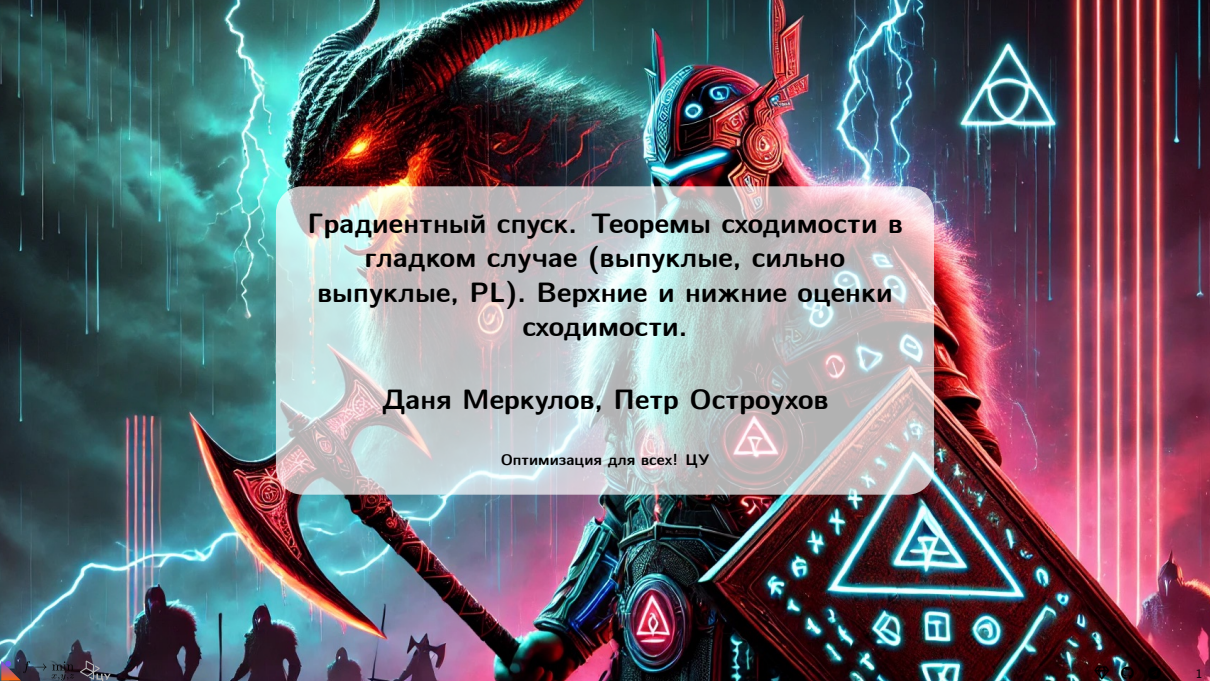


Градиентный спуск. Теоремы сходимости в гладком случае (выпуклые, сильно выпуклые, PL). Верхние и нижние оценки сходимости.

Даня Меркулов, Петр Остроухов

Оптимизация для всех! ЦУ

Когда стопать?

$$\|\nabla f(x_k)\| \le \varepsilon$$

$$\min_{x \in \mathbb{R}^n} f(x)$$

$$f(x) = x_1 + x_2 - x_3$$

$$x \in \mathbb{R}^3$$

$$|f(x_k) - f(x^*)| < \varepsilon$$

Повторение

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

$n \times 1$ $\quad$ $n \times 1$ $\quad$ 1×1 $\quad$ $n \times 1$

Виды выпуклости

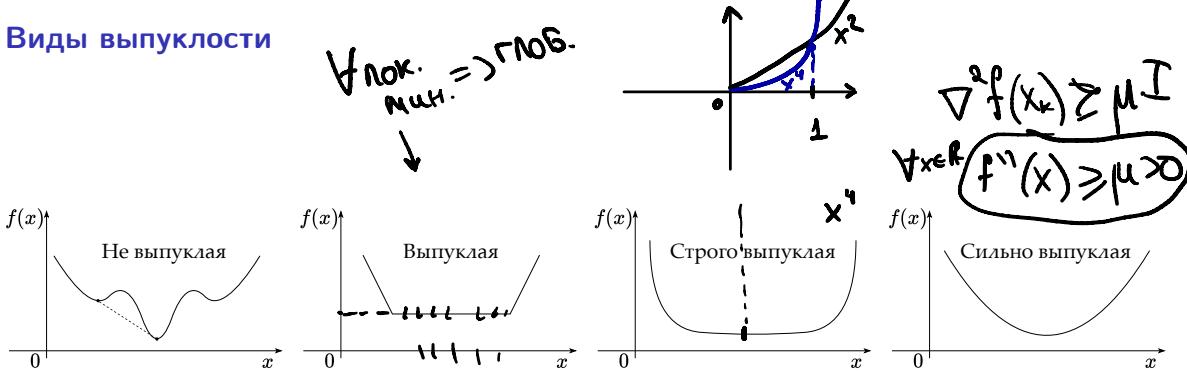


Рис. 1: Примеры выпуклых функций

Гладкость

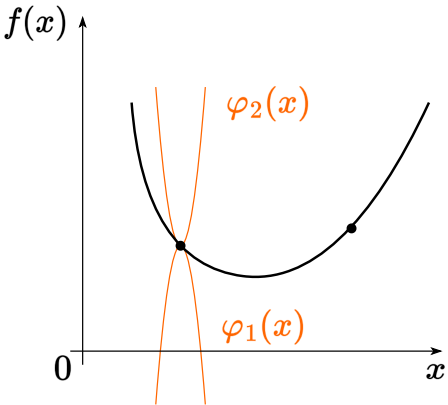


Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

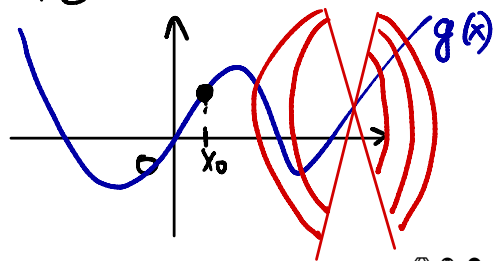
Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$

Липшицевость $g(x)$:

$\forall x, y$:

$$|g(x) - g(y)| \leq L \cdot \|x - y\|$$



Гладкость

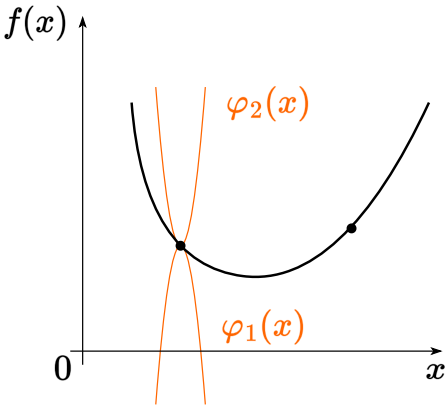
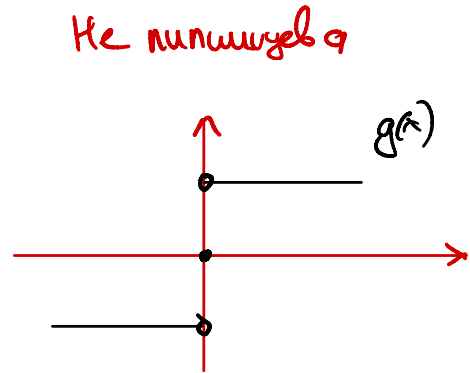


Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$



Гладкость

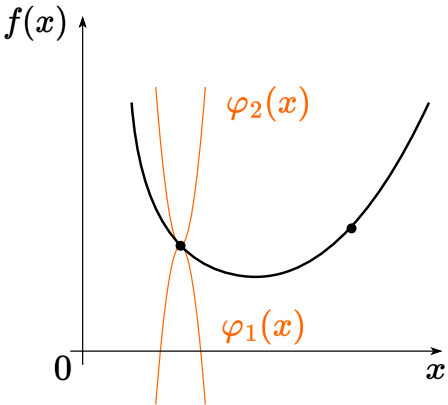
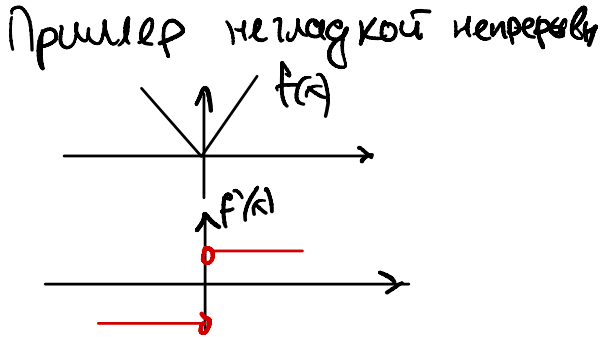


Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$



Гладкость

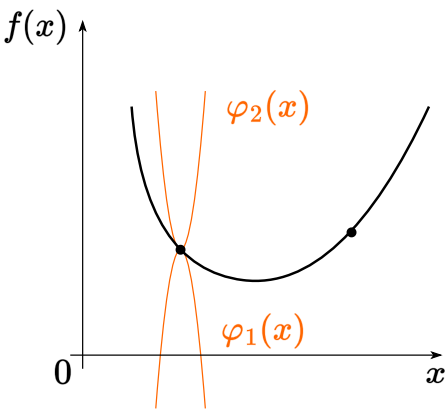


Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$

Обратим внимание, что значение константы гладкости (Липшицевости градиента) зависит от выбора нормы.
Если $f : \mathbb{R}^n \to \mathbb{R}$ - непрерывно дифференцируема и градиент Липшицев с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$\|f(y) - f(x) - \langle \nabla f(x), y - x \rangle\| \leqslant \frac{L}{2}\|y - x\|^2$$

Гладкость

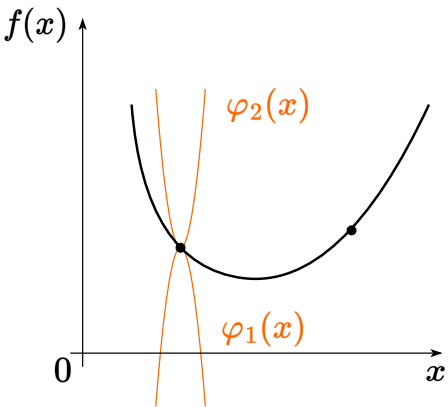


Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$

Обратим внимание, что значение константы гладкости (Липшицевости градиента) зависит от выбора нормы.
Если $f : \mathbb{R}^n \to \mathbb{R}$ - непрерывно дифференцируема и градиент Липшицев с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$\|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\| \leqslant \frac{L}{2}\|y - x\|^2$$

Если зафиксируем $x_0 \in \mathbb{R}^n$, то:

$$\varphi_1(x) = f(x_0) + \langle f(x_0), x - x_0\rangle - \frac{L}{2}\|x - x_0\|^2$$

$$\varphi_2(x) = f(x_0) + \langle f(x_0), x - x_0\rangle + \frac{L}{2}\|x - x_0\|^2$$

Гладкость

Рис. 2: Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Определение: Будем говорить, что функция $f : \mathbb{R}^n \to \mathbb{R}$ является L -гладкой, если $\forall x, y \in \mathbb{R}^n$ выполнено:

$$\|\nabla f(y) - \nabla f(x)\| \leqslant L\|y - x\|$$

Обратим внимание, что значение константы гладкости (Липшицевости градиента) зависит от выбора нормы.
Если $f : \mathbb{R}^n \to \mathbb{R}$ - непрерывно дифференцируема и градиент Липшицев с константой L, то $\forall x, y \in \mathbb{R}^n$:

$$\|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\| \leqslant \frac{L}{2}\|y - x\|^2$$

Если зафиксируем $x_0 \in \mathbb{R}^n$, то:

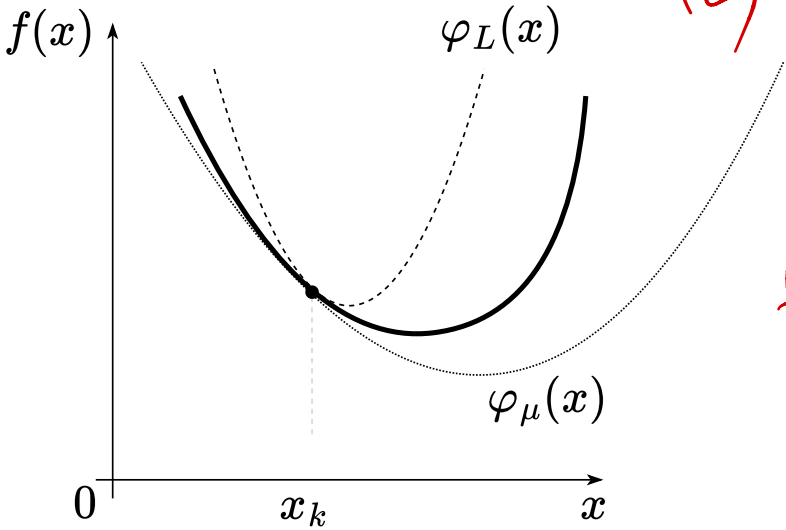
$$\varphi_1(x) = f(x_0) + \langle \nabla f(x_0), x - x_0\rangle - \frac{L}{2}\|x - x_0\|^2$$

$$\varphi_2(x) = f(x_0) + \langle \nabla f(x_0), x - x_0\rangle + \frac{L}{2}\|x - x_0\|^2$$

Это две параболы, и для них верно, что

$$\varphi_1(x) \leqslant f(x) \leqslant \varphi_2(x) \ \forall x$$

Гладкость и сильная выпуклость

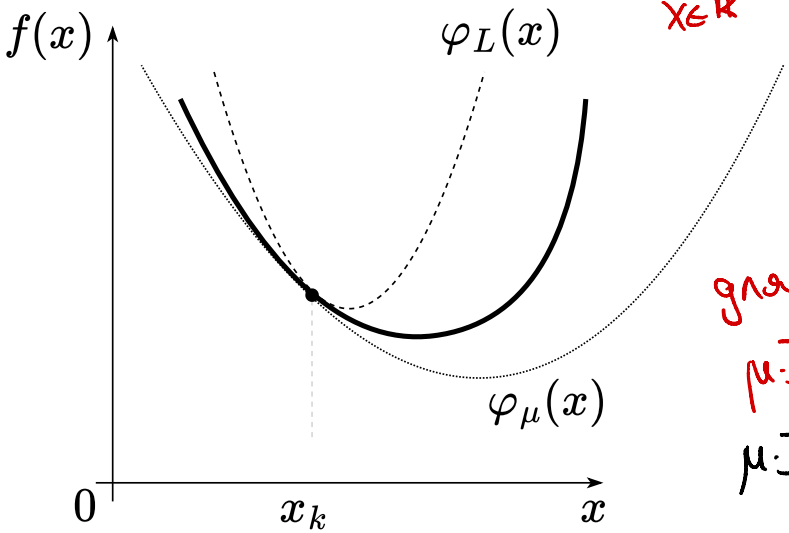


оценка сверху (L)

оценка снизу (μ)

f(x) μ; L

Гладкость и сильная выпуклость



$A \in S^n_{++}$

$x \in \mathbb{R}^n$

Пример:

$f(x) = \frac{1}{2} x^T A x$

$\nabla f = Ax$

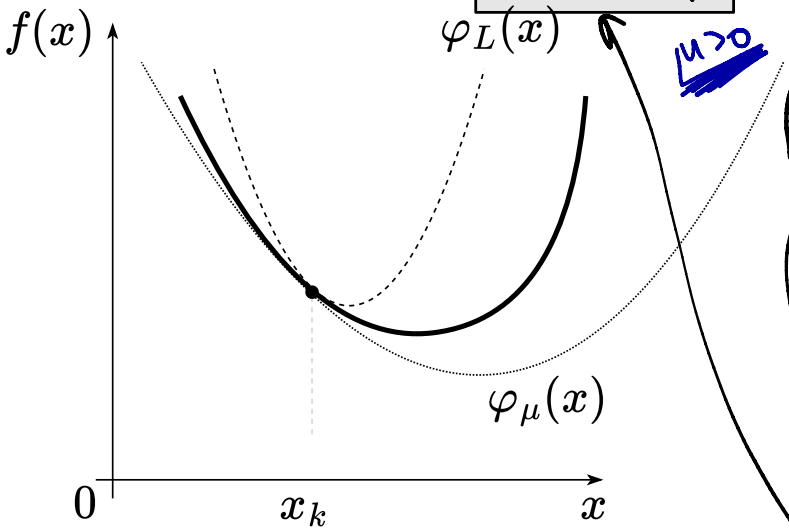
$\nabla^2 f = A$

для сильн. вып.

$\mu \cdot I \leq \nabla^2 f \leq L \cdot I$

$\mu \cdot I \leq A \leq L \cdot I$

Гладкость и сильная выпуклость



$\lambda_{min}(A) \geq \mu$

$\mu > 0$

$\mu \cdot I \leq A$

$A - \mu \cdot I \geq 0$

$\begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} - \mu \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \geq 0$

$\begin{pmatrix} \cdot - \mu & \cdot \\ \cdot & \cdot - \mu \end{pmatrix} \geq 0$

$\Updownarrow$

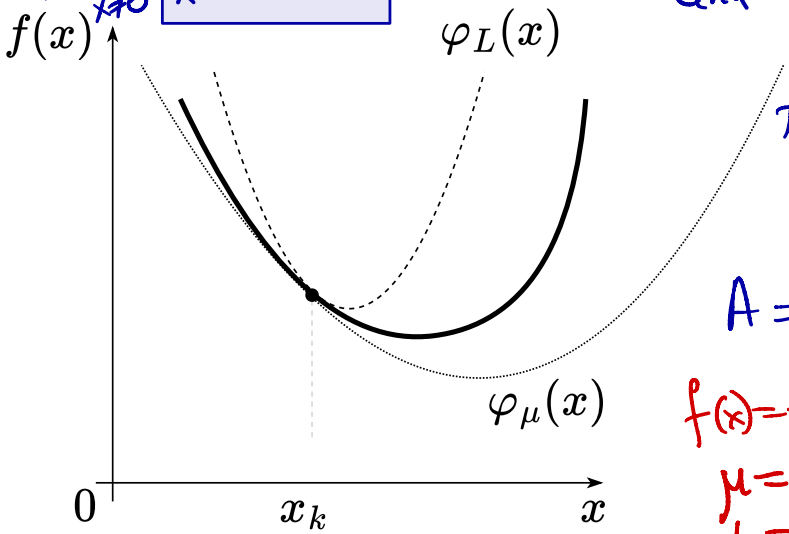
$\lambda(\cdots) \geq 0$

$\Updownarrow$

$\lambda_{min}(\cdots) \geq 0$

Гладкость и сильная выпуклость

$\forall x \in \mathbb{R}^n, x \neq 0$ $\quad x^T A x \geq 0$



для гладкости

если $\lambda_{max}(A) \leq L$

то $f(x) = \frac{1}{2} x^T A x$ – L-гладкая

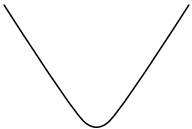
$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ $\lambda_{min} = 1$ $\lambda_{max} = 2$

$f(x) = \frac{1}{2} x^T A x$

$\mu = 1$

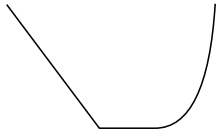
$L = 2$

Гладкость и сильная выпуклость

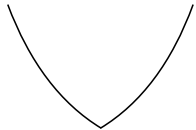


Гладкая
Выпуклая

Гладкая
μ - сильно выпуклая



Негладкая
Выпуклая



Негладкая
μ - сильно выпуклая

Градиентный спуск

Направление локального наискорейшего спуска

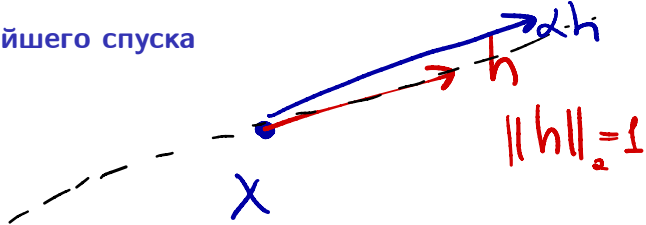
Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha\langle\nabla f(x), h\rangle + o(\alpha)$$

$f_x'(x+\alpha h)$



Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha\langle\nabla f(x), h\rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) - f(x) < 0$$

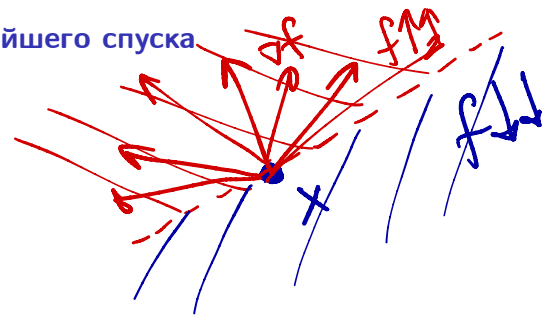
$$\alpha\langle\nabla f(x), h\rangle + o(\alpha) < 0$$

убывание

$\alpha : \alpha \to 0$

$\cdot\ \langle\nabla f(x), h\rangle$

Направление локального наискорейшего спуска



Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) - f(x) < 0$$

$$\alpha \langle \nabla f(x), h \rangle + o(\alpha) < 0$$

Переходя к пределу при $\alpha \to 0$:

$$\langle \nabla f(x), h \rangle < 0$$

если h сонаправлено с ∇f $f \uparrow\uparrow$

$\langle \nabla f, h \rangle \geqslant 0$

если h противонаправлено ∇f $f \downarrow\downarrow$

$\langle \nabla f, h \rangle \leq 0$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha\langle\nabla f(x), h\rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) - f(x) < 0$$

$$\alpha\langle\nabla f(x), h\rangle + o(\alpha) < 0$$

Переходя к пределу при $\alpha \to 0$:

$$\langle\nabla f(x), h\rangle < 0$$

Более того, мы хотим, чтобы разница $f(x) - f(x + \alpha h)$ была максимальна:

$$h = \arg\max_h (-\langle\nabla f(x), h\rangle) = \arg\min_h \langle\nabla f(x), h\rangle.$$

Также из неравенства Коши–Буняковского получаем:

$$|\langle\nabla f(x), h\rangle| \le \|\nabla f(x)\|_2 \|h\|_2$$

3·1

$$\langle\nabla f(x), h\rangle \ge -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2$$

$$-3 \le \langle\nabla f(x), h\rangle \le 3$$

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) - f(x) < 0$$

$$\alpha \langle \nabla f(x), h \rangle + o(\alpha) < 0$$

Переходя к пределу при $\alpha \to 0$:

$$\langle \nabla f(x), h \rangle < 0$$

Более того, мы хотим, чтобы разница $f(x) - f(x + \alpha h)$ была максимальна:

$$h = \arg\max_h \left(-\langle \nabla f(x), h \rangle\right) = \arg\min_h \langle \nabla f(x), h \rangle.$$

Также из неравенства Коши–Буняковского получаем:

$$|\langle \nabla f(x), h \rangle| \leq \|\nabla f(x)\|_2 \|h\|_2$$

$$\langle \nabla f(x), h \rangle \geq -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2$$

Таким образом, направление антиградиента

$$h = \arg\min_h \langle \nabla f(x), h \rangle = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

представляет собой направление **наискорейшего локального** убывания функции f.

Направление локального наискорейшего спуска

Рассмотрим линейное приближение дифференцируемой функции f вдоль направления h, где $\|h\|_2 = 1$:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha)$$

Хотим, чтобы h было направлением убывания:

$$f(x + \alpha h) - f(x) < 0$$

$$\alpha \langle \nabla f(x), h \rangle + o(\alpha) < 0$$

Переходя к пределу при $\alpha \to 0$:

$$\langle \nabla f(x), h \rangle < 0$$

Более того, мы хотим, чтобы разница $f(x) - f(x + \alpha h)$ была максимальна:

$$h = \arg\max_h (-\langle \nabla f(x), h \rangle) = \arg\min_h \langle \nabla f(x), h \rangle.$$

Также из неравенства Коши–Буняковского получаем:

$$|\langle \nabla f(x), h \rangle| \le \|\nabla f(x)\|_2 \|h\|_2$$
$$\langle \nabla f(x), h \rangle \ge -\|\nabla f(x)\|_2 \|h\|_2 = -\|\nabla f(x)\|_2$$

Таким образом, направление антиградиента

$$h = \arg\min_h \langle \nabla f(x), h \rangle = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

представляет собой направление **наискорейшего локального** убывания функции f.

Итерация метода имеет вид:

$$x^{k+1} = x^k - \alpha \nabla f(x^k)$$

размер шага

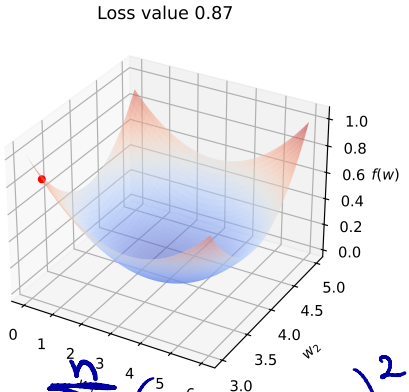
learning rate

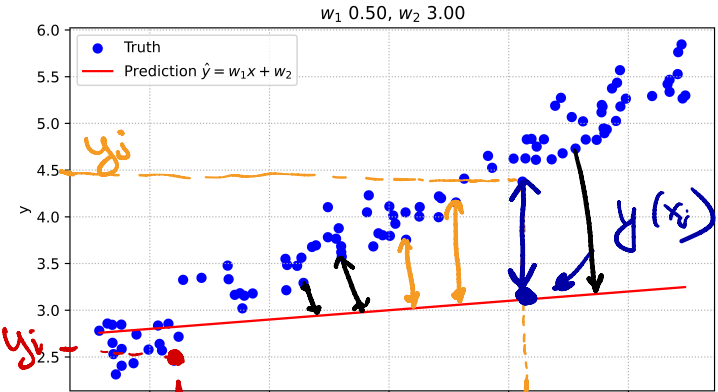
Сходимость алгоритма градиентного спуска

$y = w_0 + w_1 \cdot x$

$y(x_i) = ?$

Код для построения анимации ниже. Сходимость существенно зависит от выбора шага α:





$$\sum_{i=1}^{n} \left(y(x_i) - y_i \right)^2 \to \min_{w_0, w_1}$$

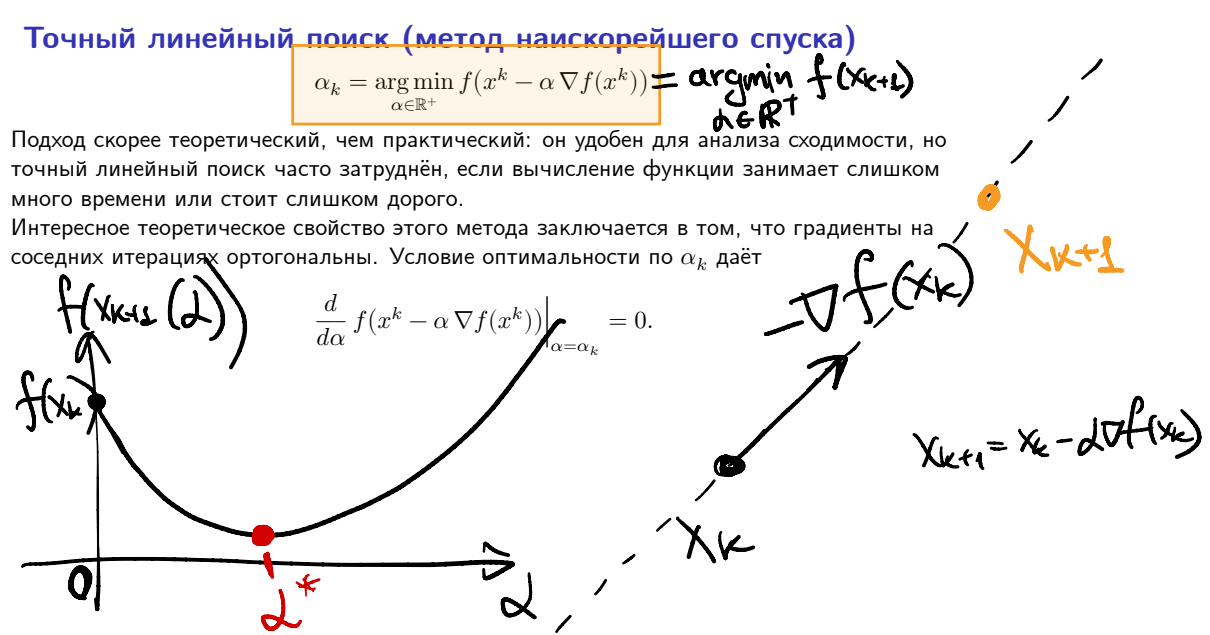
Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k))\Big|_{\alpha=\alpha_k} = 0.$$



Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k))\Big|_{\alpha=\alpha_k} = 0.$$

$$\frac{\partial x_{k+1}}{\partial \alpha} = \frac{\partial (x_k - \alpha \nabla f(x_k))}{\partial \alpha} = -\nabla f(x_k)$$

$$\frac{\partial f}{\partial x}^T \cdot \frac{\partial x_{k+1}}{\partial \alpha} = 0$$

$$-\left(\nabla f(x_{k+1})\right) \cdot \nabla f(x_k) = 0$$

Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k))\Big|_{\alpha=\alpha_k} = 0.$$

Условия оптимальности:

Точный линейный поиск (метод наискорейшего спуска)

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}^+} f(x^k - \alpha \nabla f(x^k))$$

Подход скорее теоретический, чем практический: он удобен для анализа сходимости, но точный линейный поиск часто затруднён, если вычисление функции занимает слишком много времени или стоит слишком дорого.

Интересное теоретическое свойство этого метода заключается в том, что градиенты на соседних итерациях ортогональны. Условие оптимальности по α_k даёт

$$\frac{d}{d\alpha} f(x^k - \alpha \nabla f(x^k))\Big|_{\alpha=\alpha_k} = 0.$$

Условия оптимальности:

$$\nabla f(x^{k+1})^\top \nabla f(x^k) = 0$$

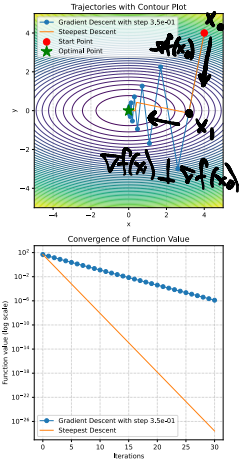


Рис. 3: Наискорейший спуск

Открыть в Colab ♣

$f(x) = \frac{1}{2} x^T A x$

$x_{k+1} = x_k - \alpha \nabla f(x_k)$

$A = A^T$

$\nabla f(x) = Ax$

$: \nabla f(x_{k+1})^T \nabla f(x_k) = 0$

$(Ax_{k+1})^T \cdot Ax_k = 0$

Сильно выпуклые квадратичные функции

$(A(x_k - \alpha A x_k))^T A x_k = 0$

$(Ax_k - \alpha A A x_k)^T A x_k = 0$

$\nabla f(x_k) = g_k$

$(x_k^T A^T - \alpha x_k^T A^T A^T) A x_k = 0$

$x_k^T A^T A x_k - \alpha x_k^T A^T A^T A x_k = 0$

$g_k^T g_k - \alpha g_k^T A^T g_k = 0$

$$\alpha = \frac{g_k^T g_k}{g_k^T A g_k}$$