ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

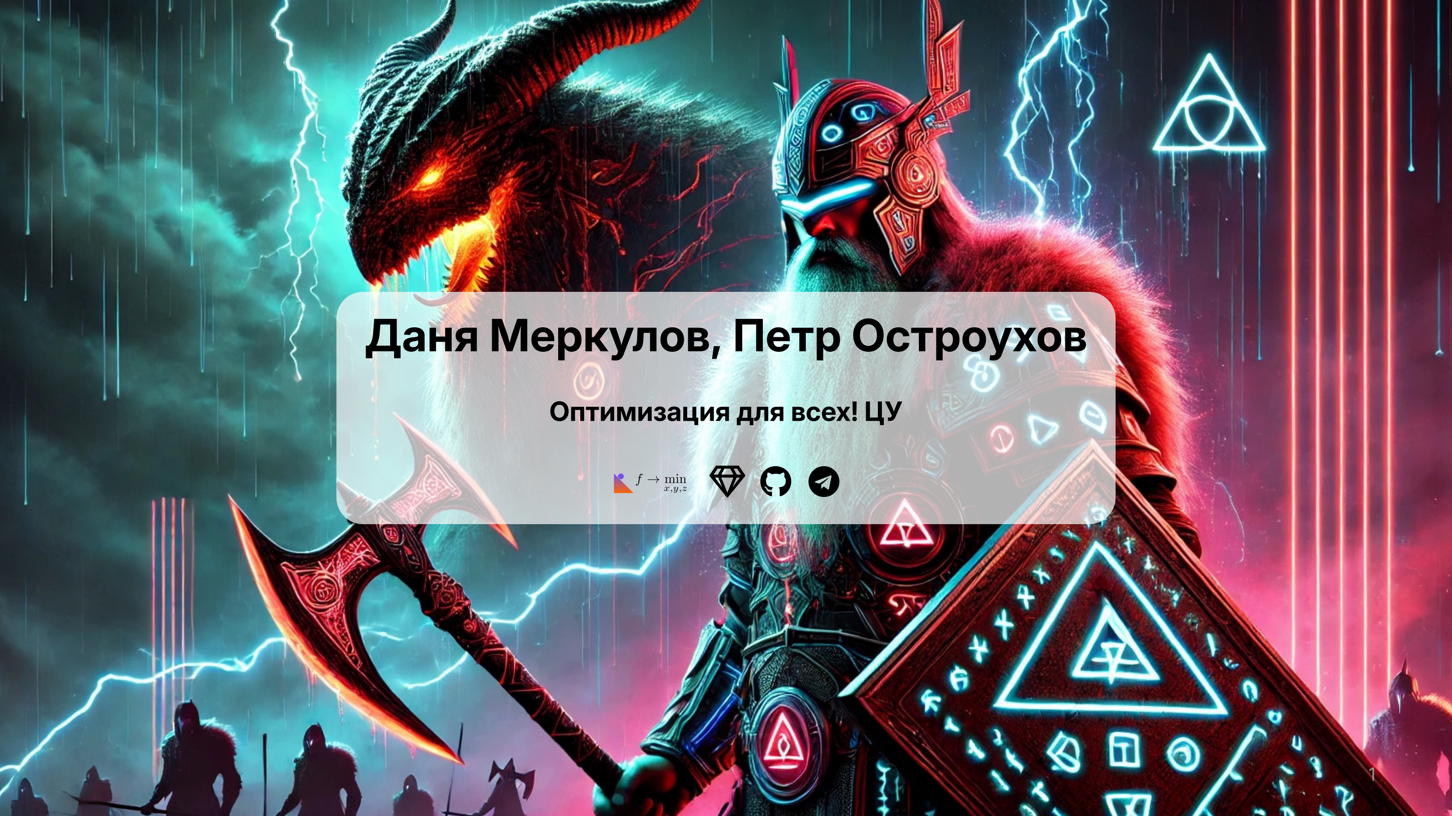
Градиентный спуск

МЕТОДЫ ВЫПУКЛОЙ ОПТИМИЗАЦИИ

НЕДЕЛЯ 7

Даня Меркулов

Пётр Остроухов



Даня Меркулов, Петр Остроухов

Оптимизация для всех! ЦУ

Повторение

Виды выпуклости

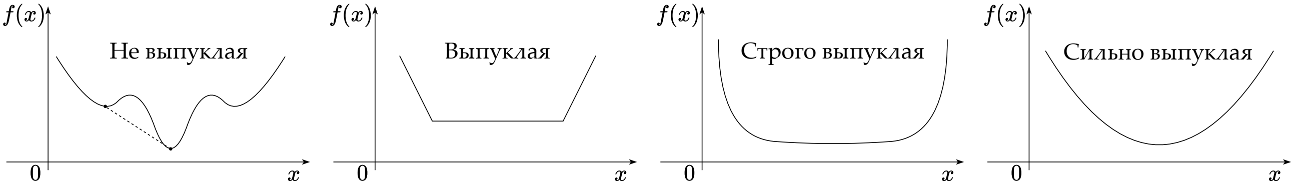


Рисунок 1. Примеры выпуклых функций

Гладкость

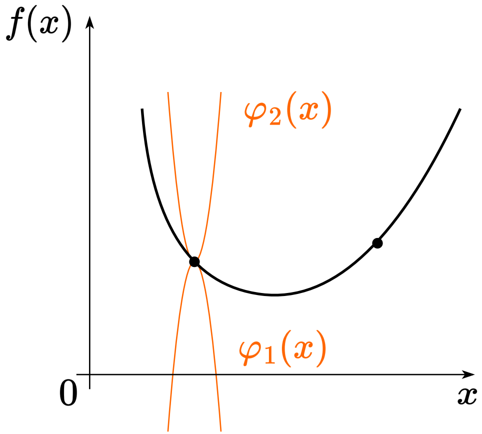


Рисунок 2. Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

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

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 $\forall x, y \in \mathbb{R}^n$:

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

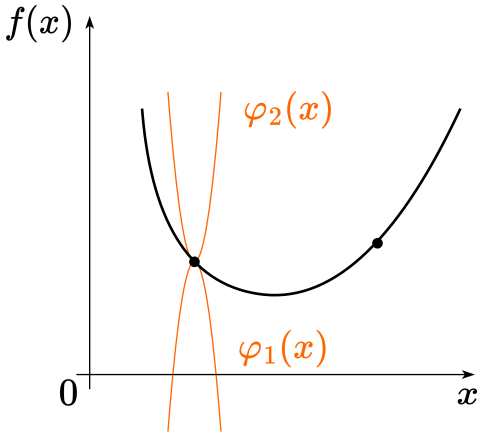


Рисунок 2. Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

Гладкость

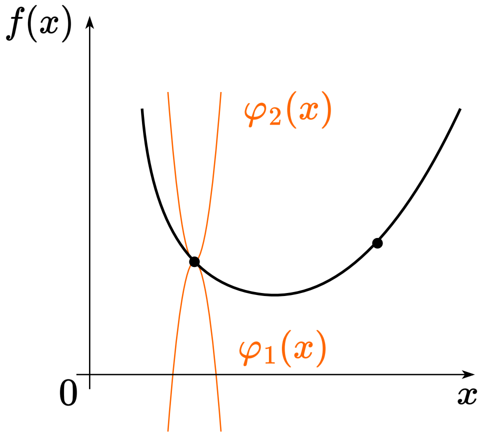


Рисунок 2. Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

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

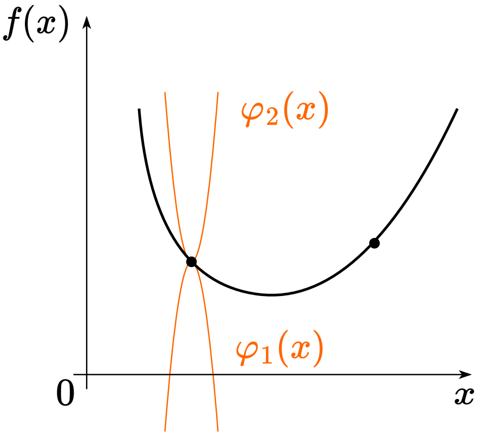


Рисунок 2. Иллюстрация Липшицевых парабол, между которыми зажата гладкая функция. Чаще нас интересует мажорирующая из них.

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:

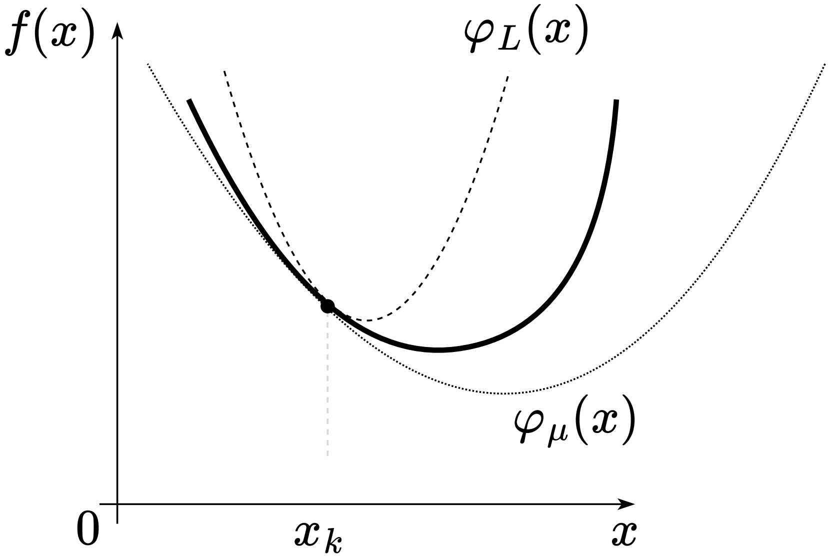
$$\varphi_1(x) = f(x_0) + \langle f(x_0), x - x_0\rangle - \frac{L}{2}\|x - x_0\|^2$$

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

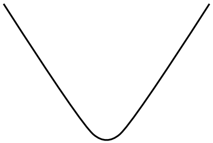
Это две параболы, и для них верно, что

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

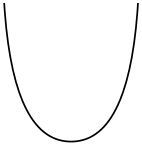
Гладкость и сильная выпуклость



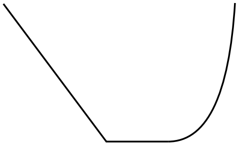
Гладкость и сильная выпуклость



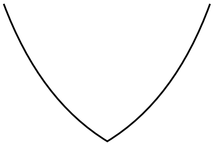
Гладкая
Выпуклая



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 $\|h\|_2 = 1$:

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

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:

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

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

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 $\alpha \to 0$:

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

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:

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

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:

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

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

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

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:

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

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

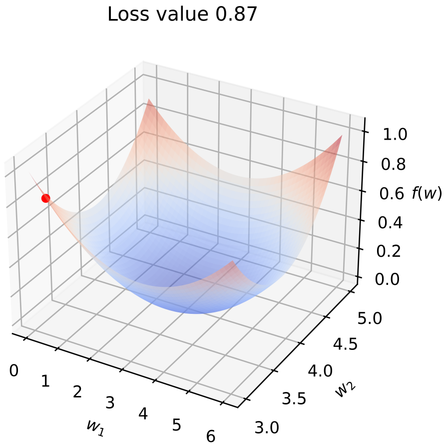
$$h = \arg\min_h \langle \nabla f(x), h\rangle = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

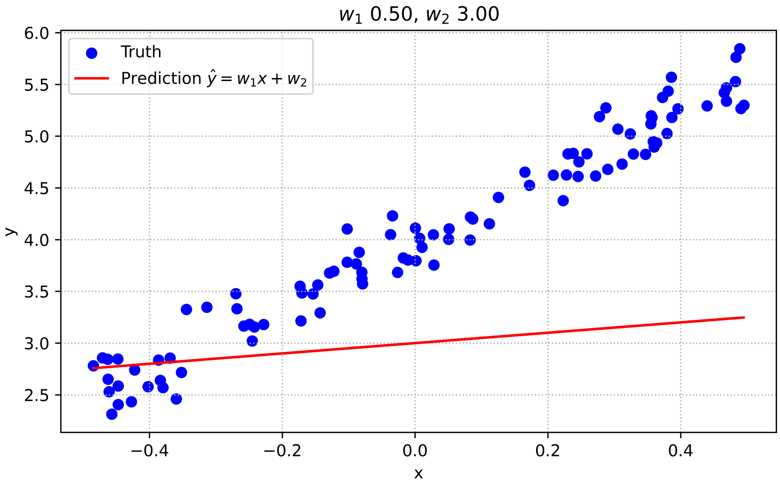
представляет собой направление **наискорейшего локального** убывания функции f.
Итерация метода имеет вид:

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

Сходимость алгоритма градиентного спуска

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





Точный линейный поиск (метод наискорейшего спуска)

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

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)

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

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)

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

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:

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

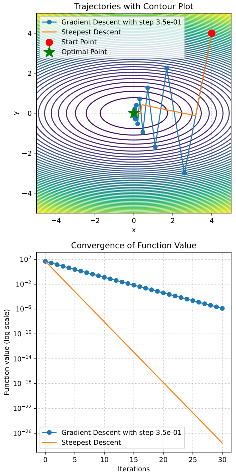


Рисунок 3. Наискорейший спуск

Открыть в Colab ♣

Экстра: Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \tag{GF}$$

Экстра: Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \tag{GF}$$

Дискретизируем его на равномерной сетке с шагом α:

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

Экстра: Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \tag{GF}$$

Дискретизируем его на равномерной сетке с шагом α:

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

где $x^k \equiv x(t_k)$ и $\alpha = t_{k+1} - t_k$ — шаг сетки.
Отсюда получаем выражение для x^{k+1}:

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

являющееся точной формулой обновления градиентного спуска.
Открыть в Colab ♣

Экстра: Дифференциальное уравнение градиентного потока

Рассмотрим дифференциальное уравнение градиентного потока:

$$\frac{dx}{dt} = -\nabla f(x(t)). \tag{GF}$$

Дискретизируем его на равномерной сетке с шагом α:

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

где $x^k \equiv x(t_k)$ и $\alpha = t_{k+1} - t_k$ — шаг сетки.
Отсюда получаем выражение для x^{k+1}:

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

являющееся точной формулой обновления градиентного спуска.
Открыть в Colab ♣

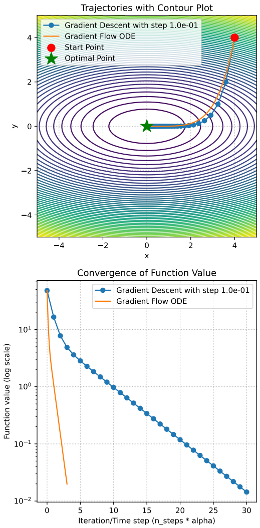


Рисунок 4. Траектория градиентного потока