



ТИНЬКОФФ

Лекция 4

# Оптимизация: градиентный спуск

Иван Карпухин

Ведущий исследователь-разработчик





# Карпухин Иван

Ведущий исследователь-разработчик

@ [i.a.karpukhin@tinkoff.ru](mailto:i.a.karpukhin@tinkoff.ru)

Профессионально занимаюсь машинным обучением  
более 8 лет. Опыт:

- Исследования в области computer vision
- Голосовые технологии
- Распознавание лиц и текстов
- Говорящие головы
- Оптимизация NN

# Цели вебинара

01

Вспомнить, что такое  
дифференцируемые  
функции

02

Научиться искать  
локальные оптимумы  
функций одной  
переменной

03

Научиться искать  
локальные оптимумы  
функций многих  
переменных

# Маршрут вебинара

01



Одномерный  
случай

02

Многомерный  
случай

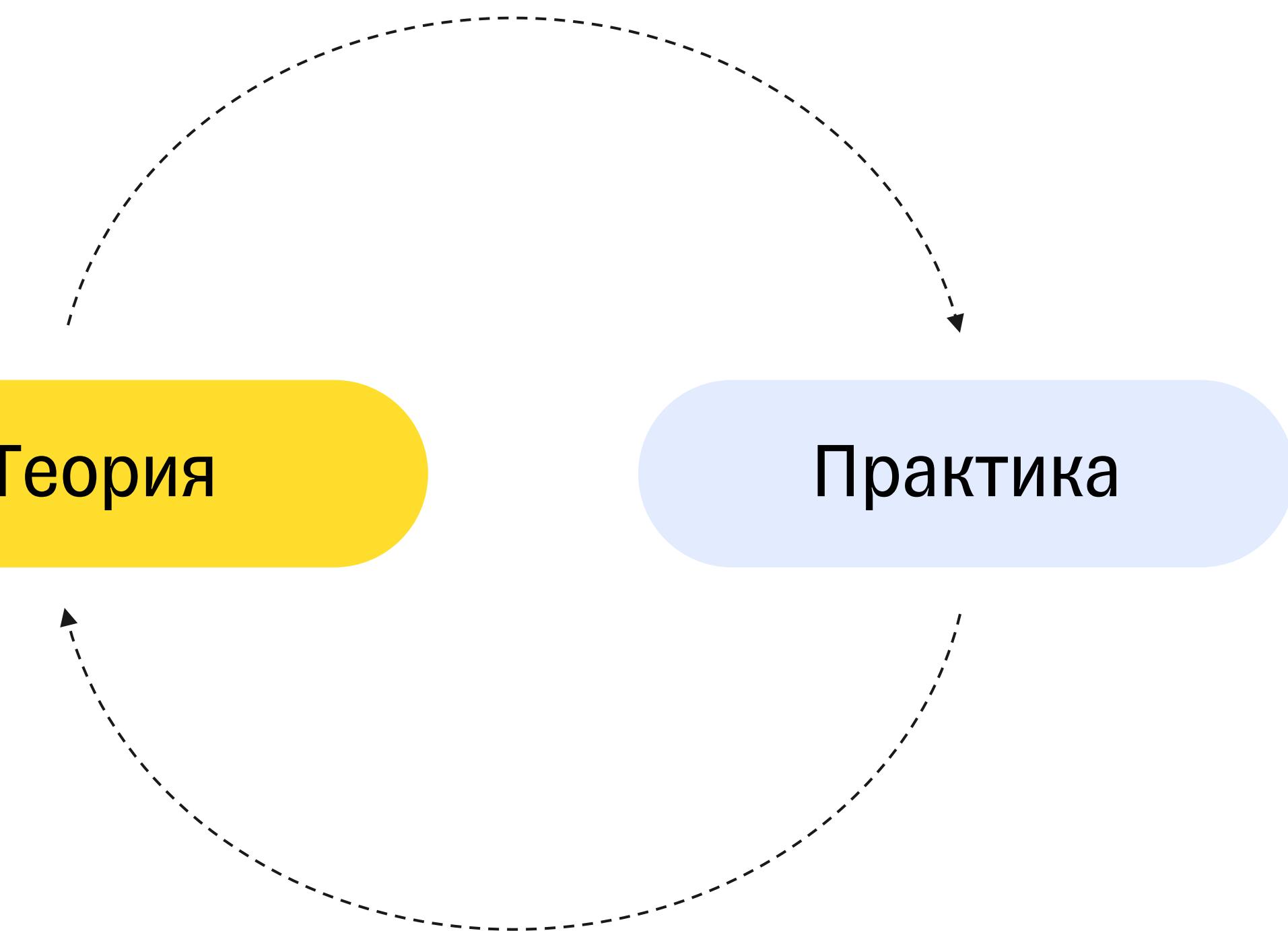
03

Продвинутые  
техники

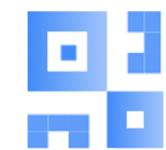
# формат

Теория

Практика



# формат: практика



Тесты по QR или ссылкам



Задания “на бумаге”



Задания по программированию



**ТИНЬКОФФ**

# Одномерный случай

# Задание

Тест по основам анализа



Ссылка в чате



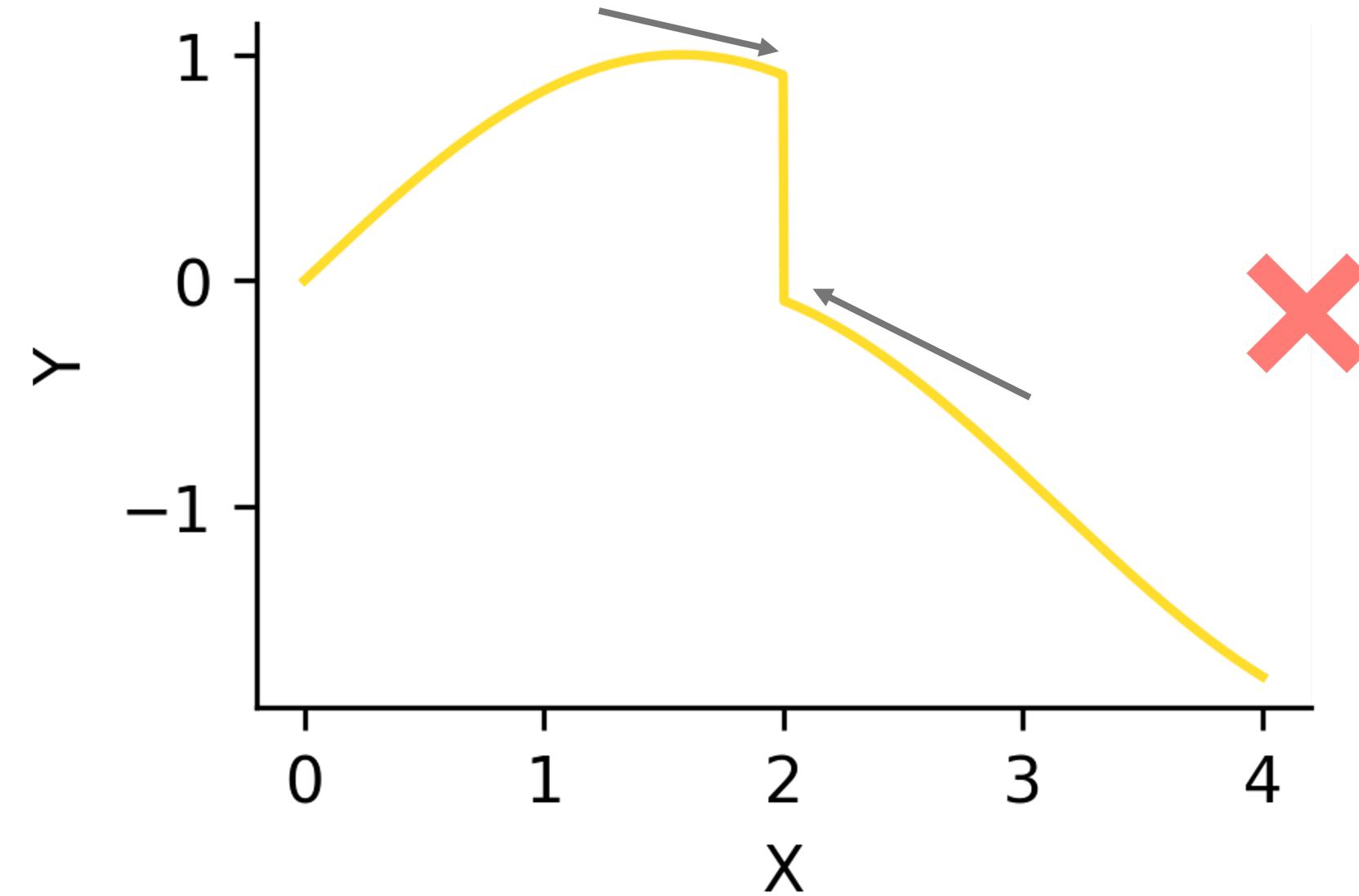
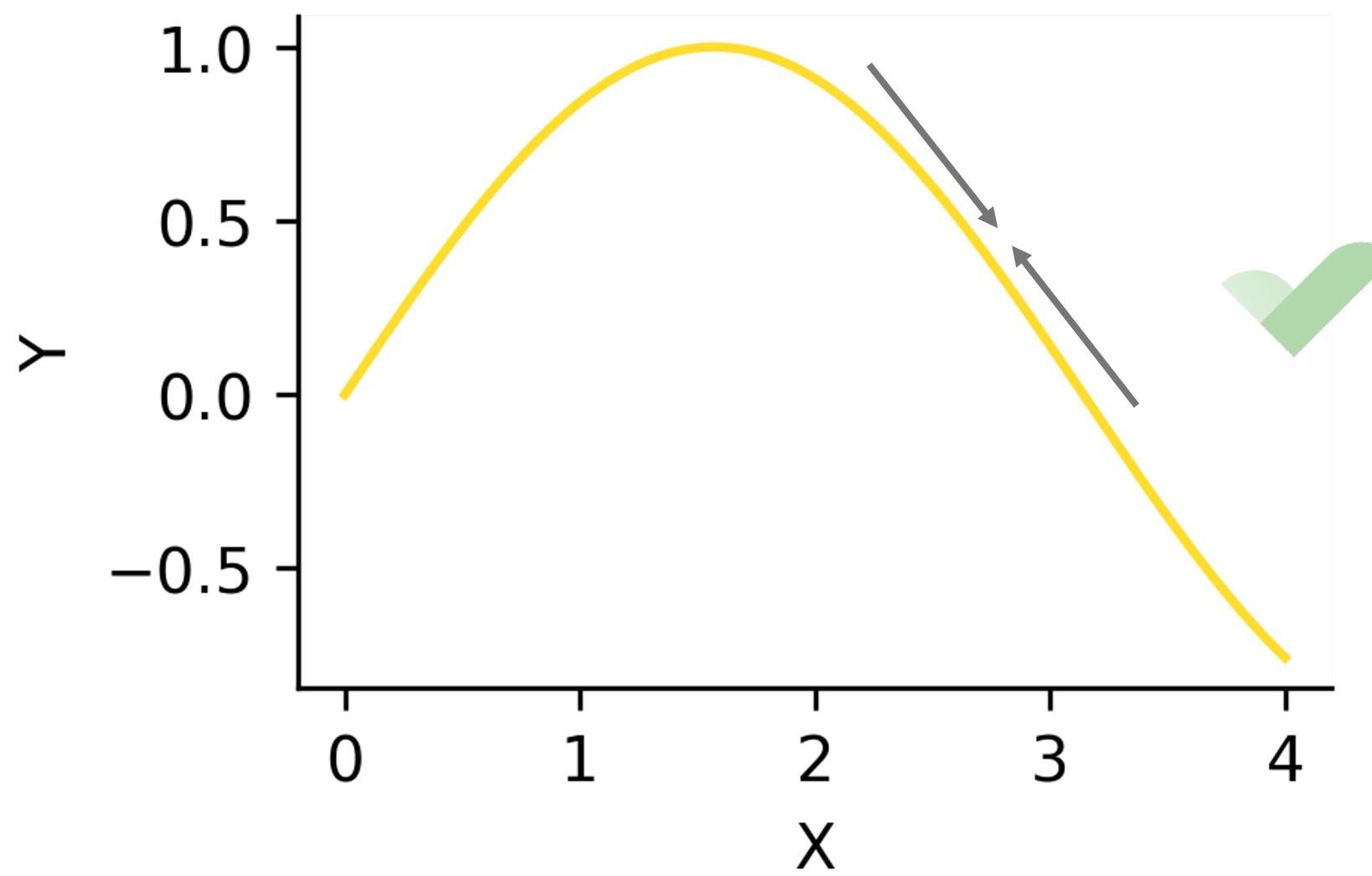
5 минут



Анонимно

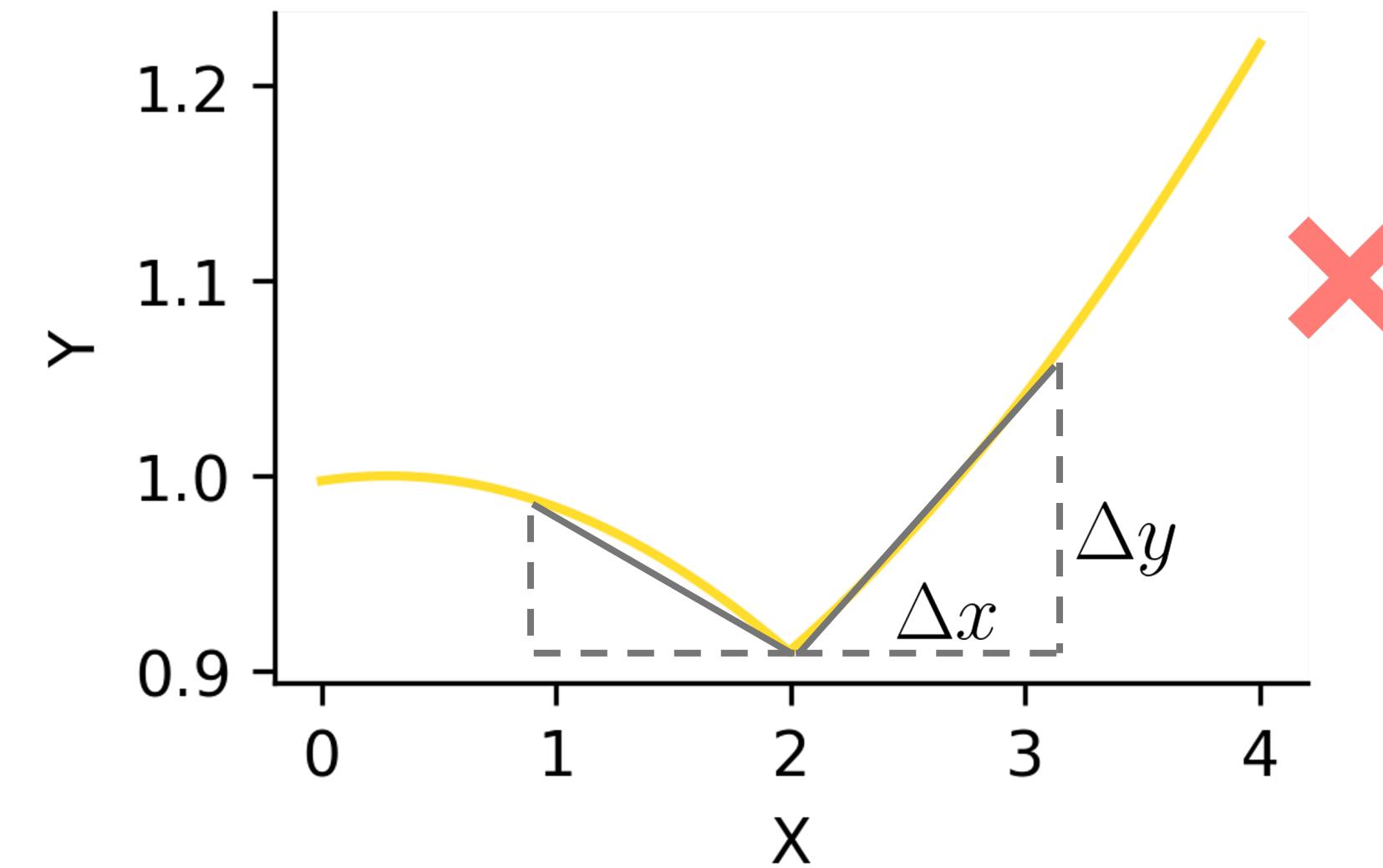
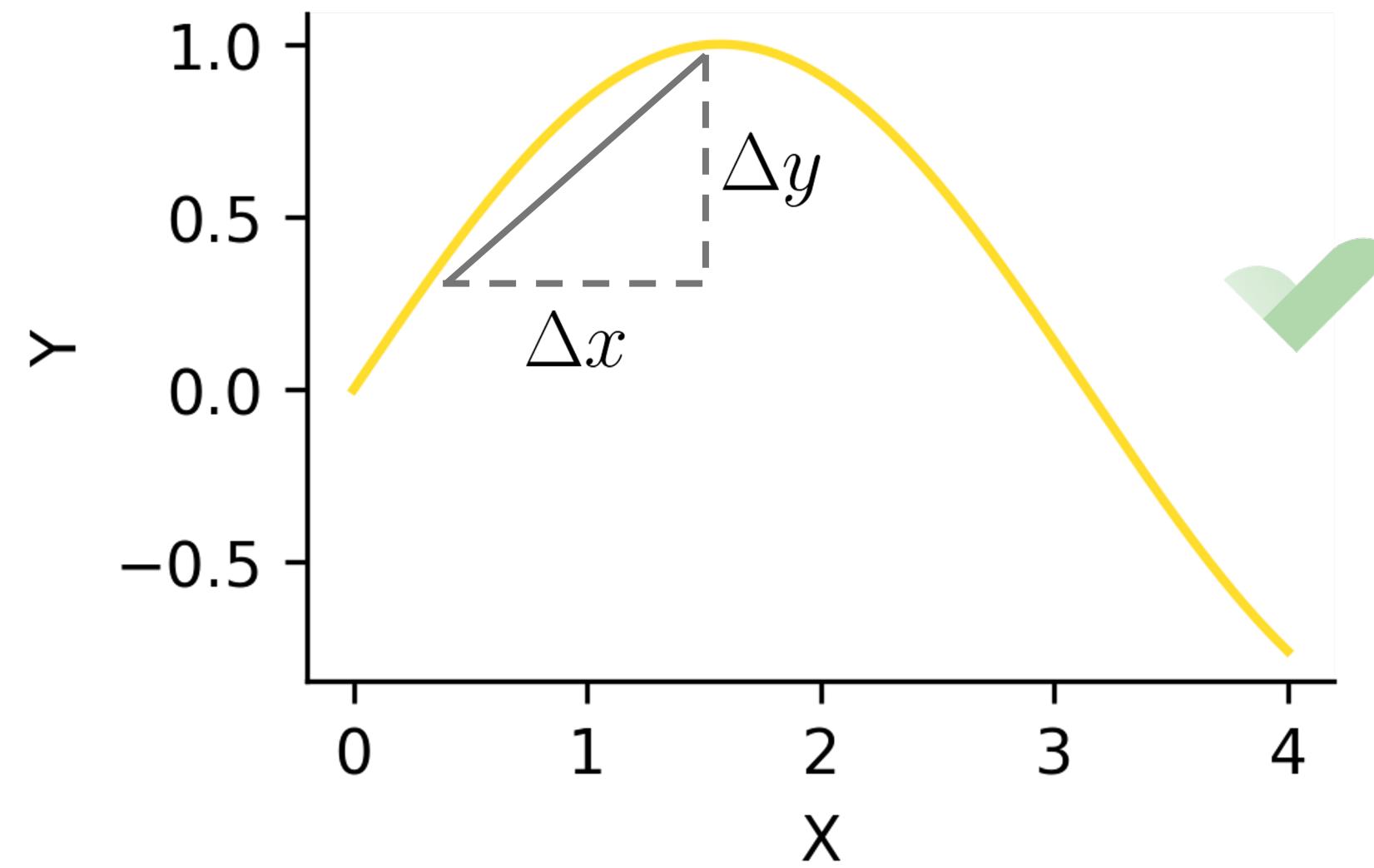
# Непрерывность

Малое изменение  $x$  ведет к малому изменению  $y$



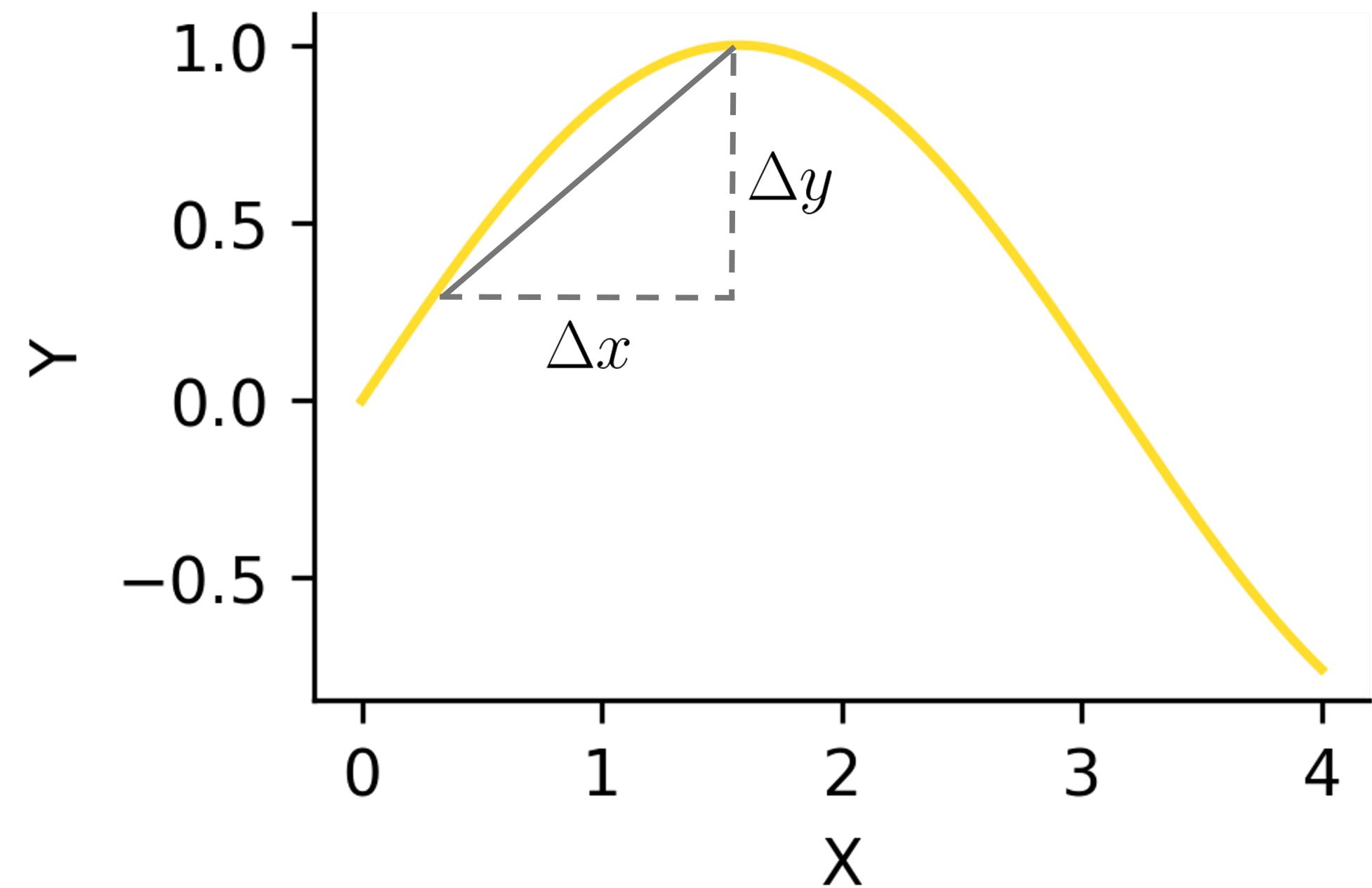
# Дифференцируемость

Производная: отношение  $\Delta y$  к  $\Delta x$  в некоторой точке при малом  $\Delta x$



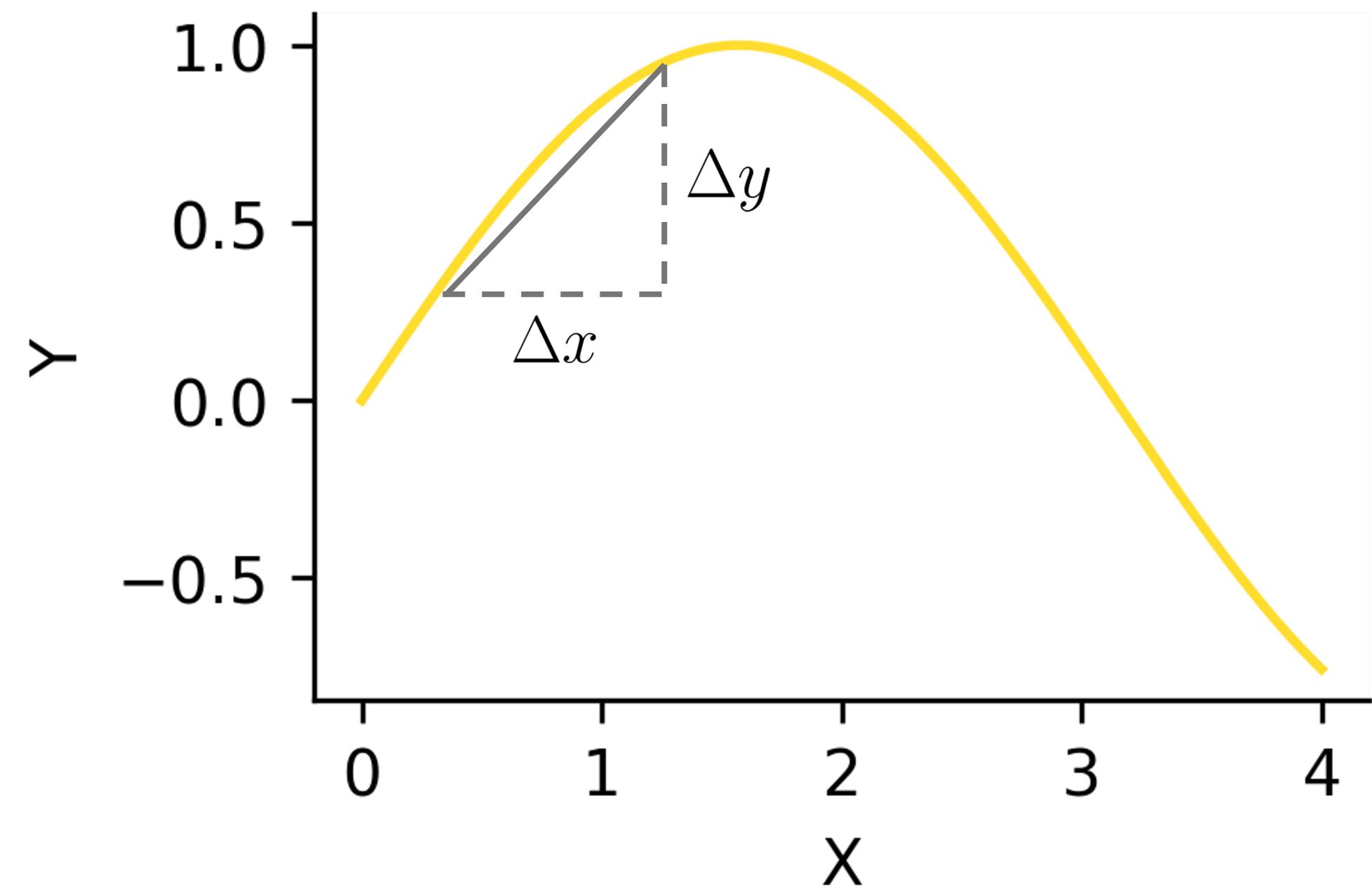
# Дифференцируемость

Тангенс угла наклона касательной



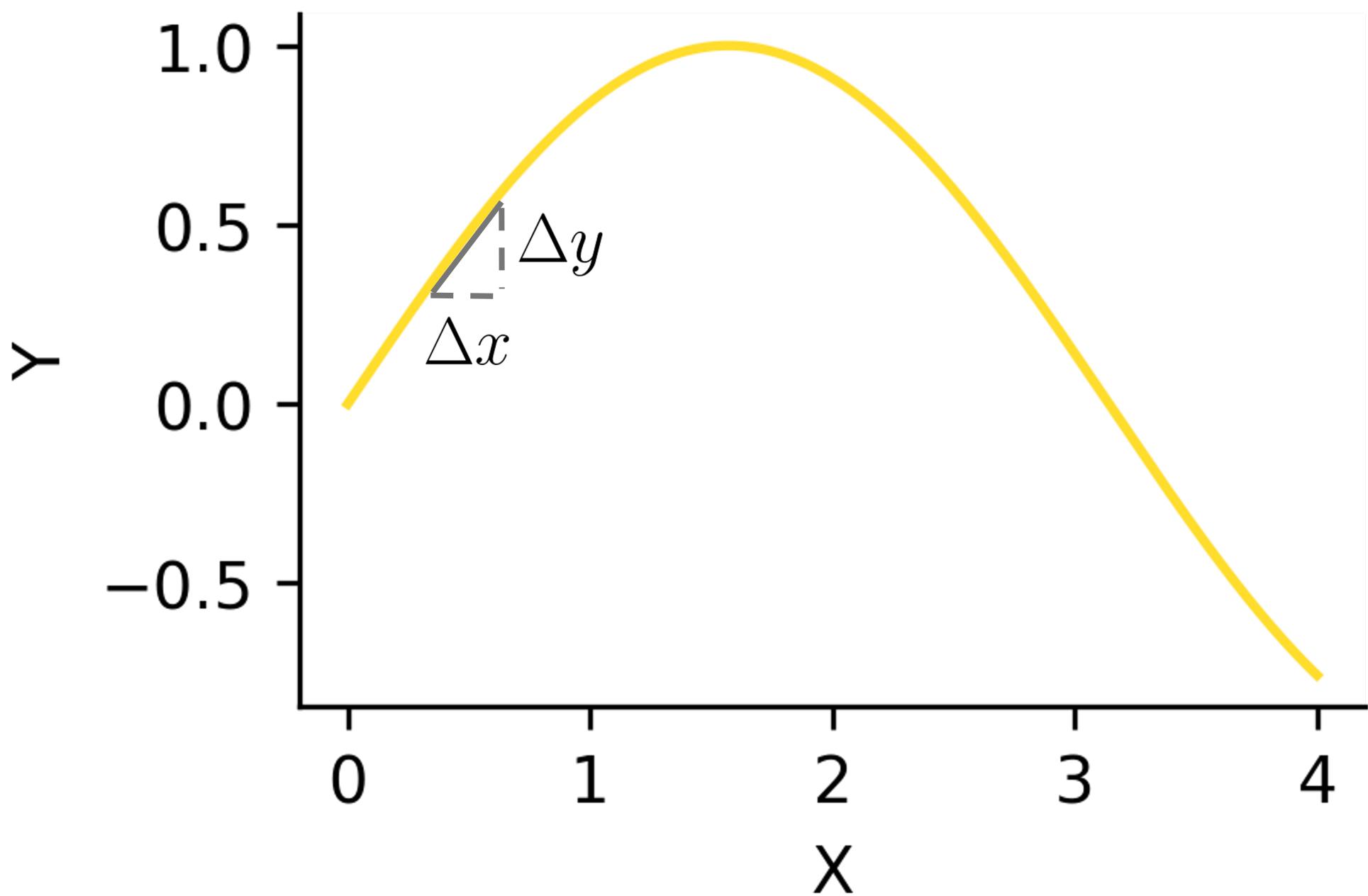
# Дифференцируемость

Тангенс угла наклона касательной



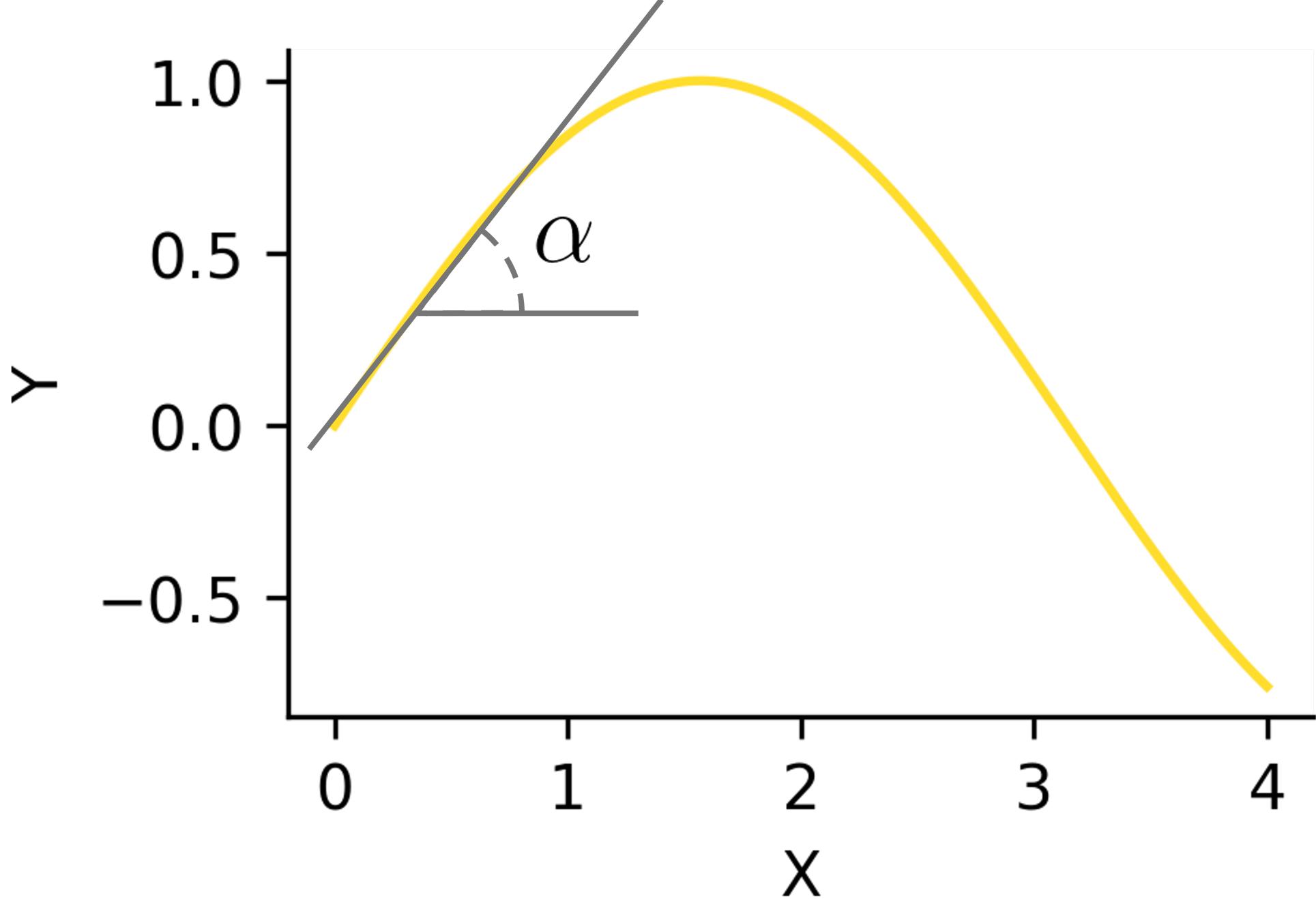
# Дифференцируемость

Тангенс угла наклона касательной



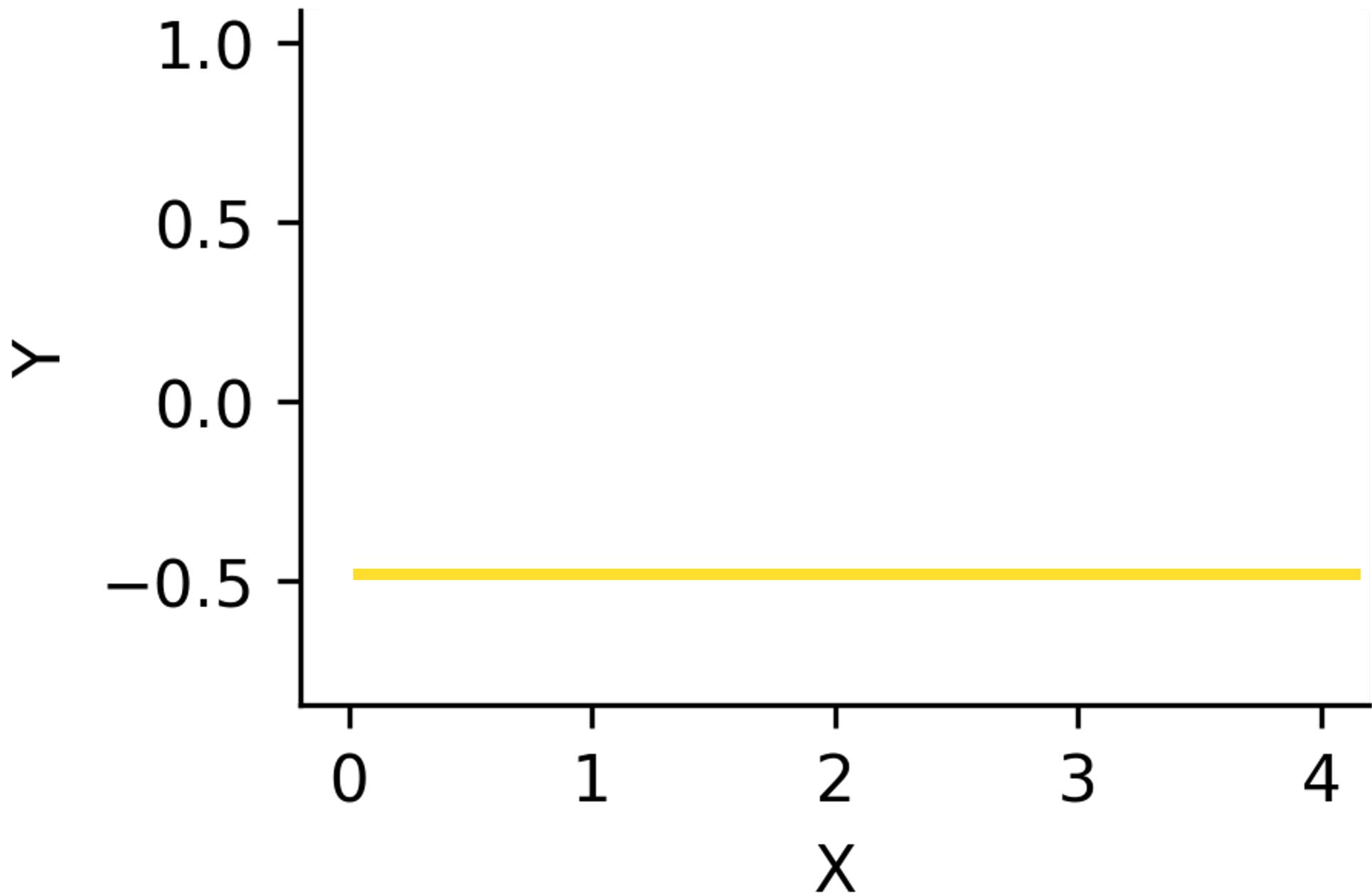
# Дифференцируемость

Тангенс угла наклона касательной



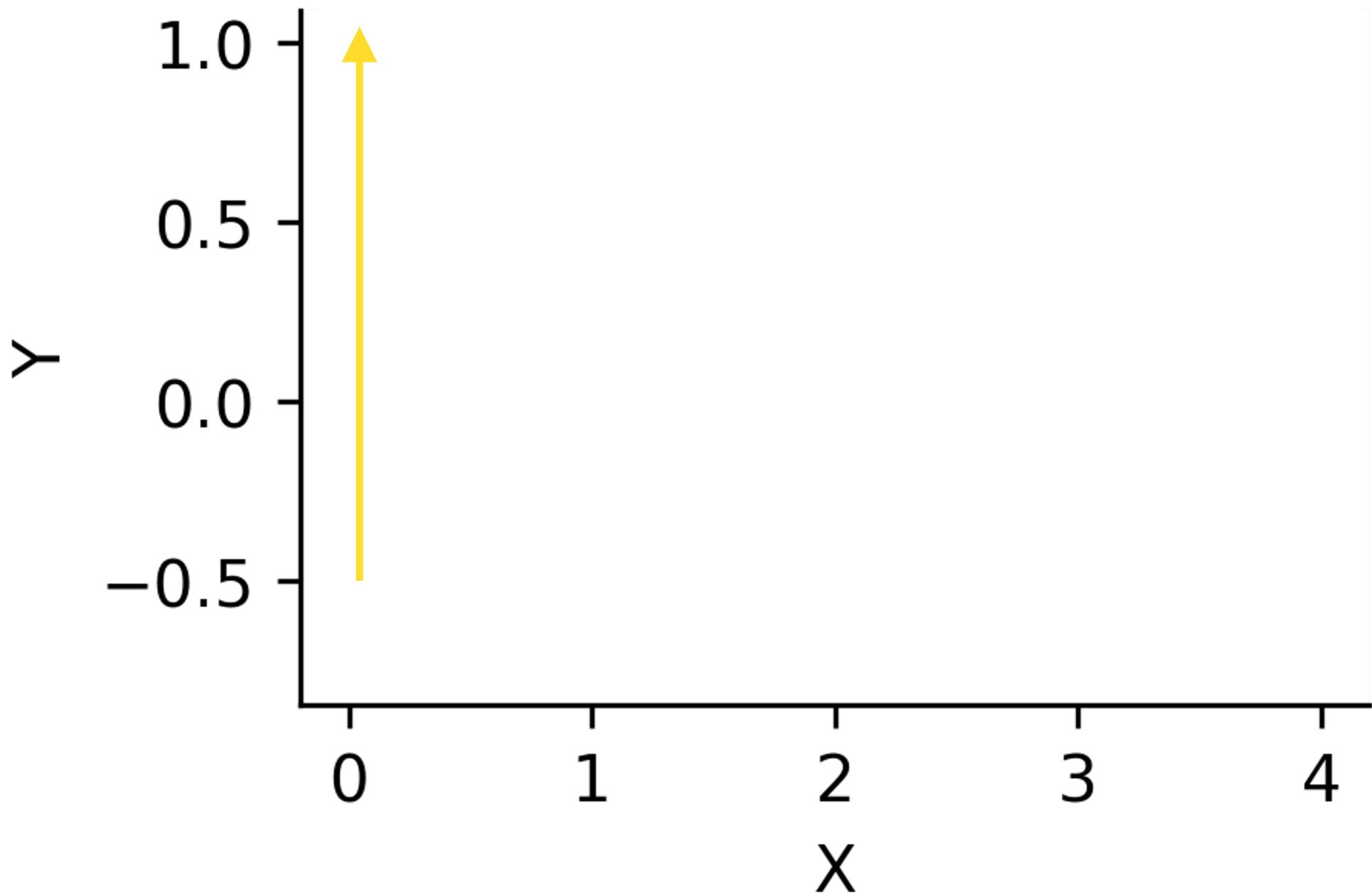
# Чему равна производная?

Производная: отношение  $\Delta y$  к  $\Delta x$   
в некоторой точке при малом  $\Delta x$



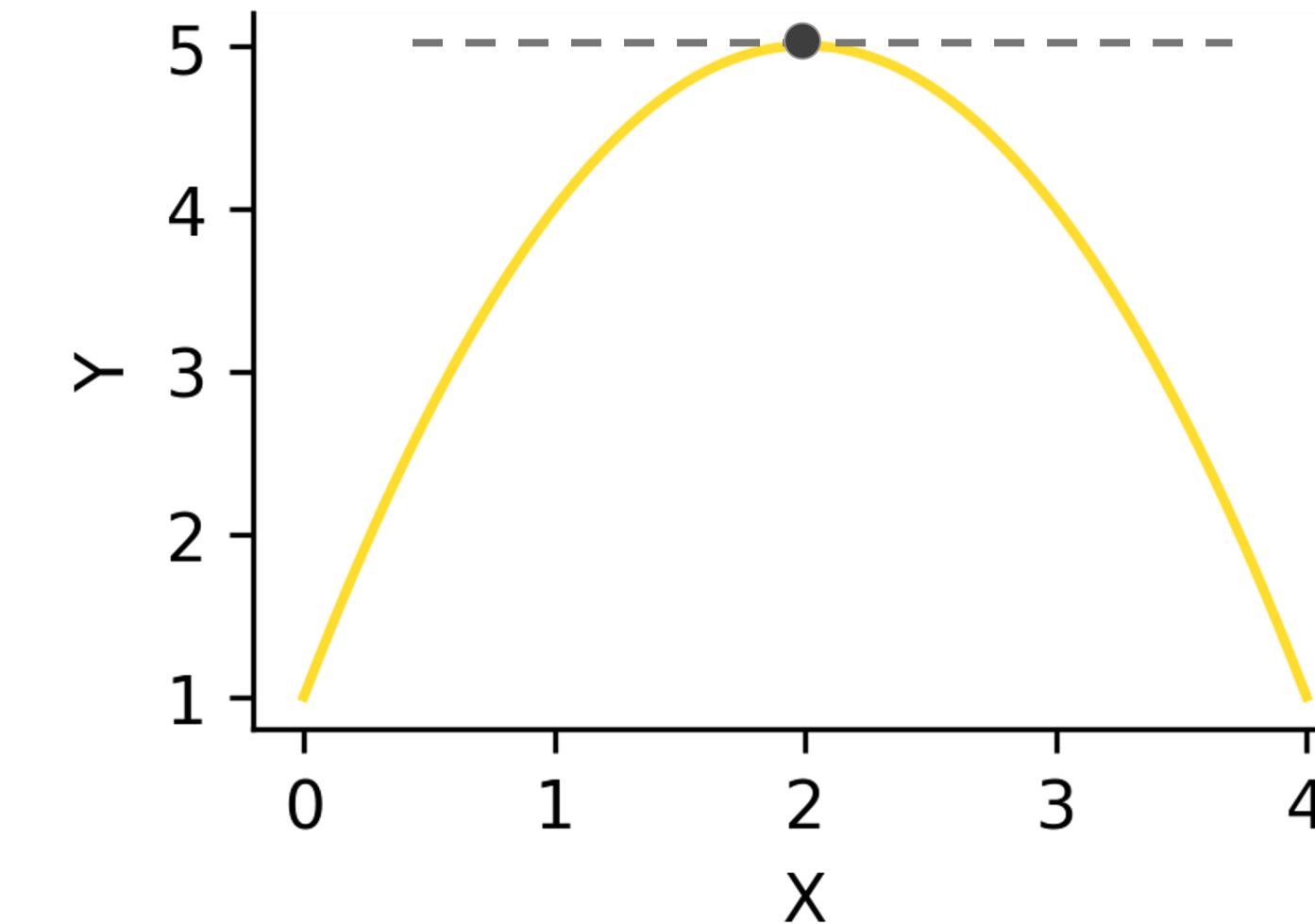
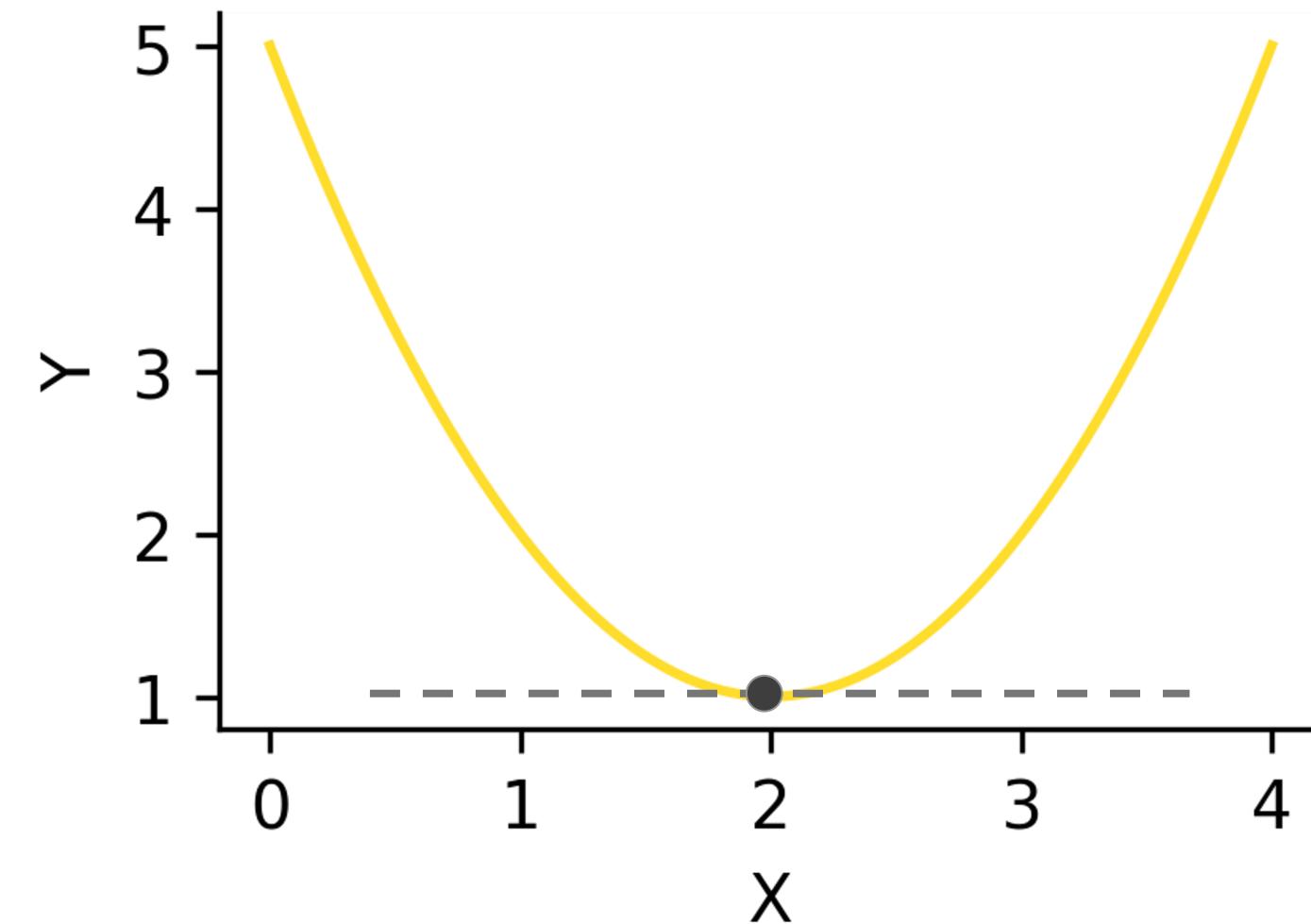
# Чему равна производная?

Производная: отношение  $\Delta y$  к  $\Delta x$   
в некоторой точке при малом  $\Delta x$



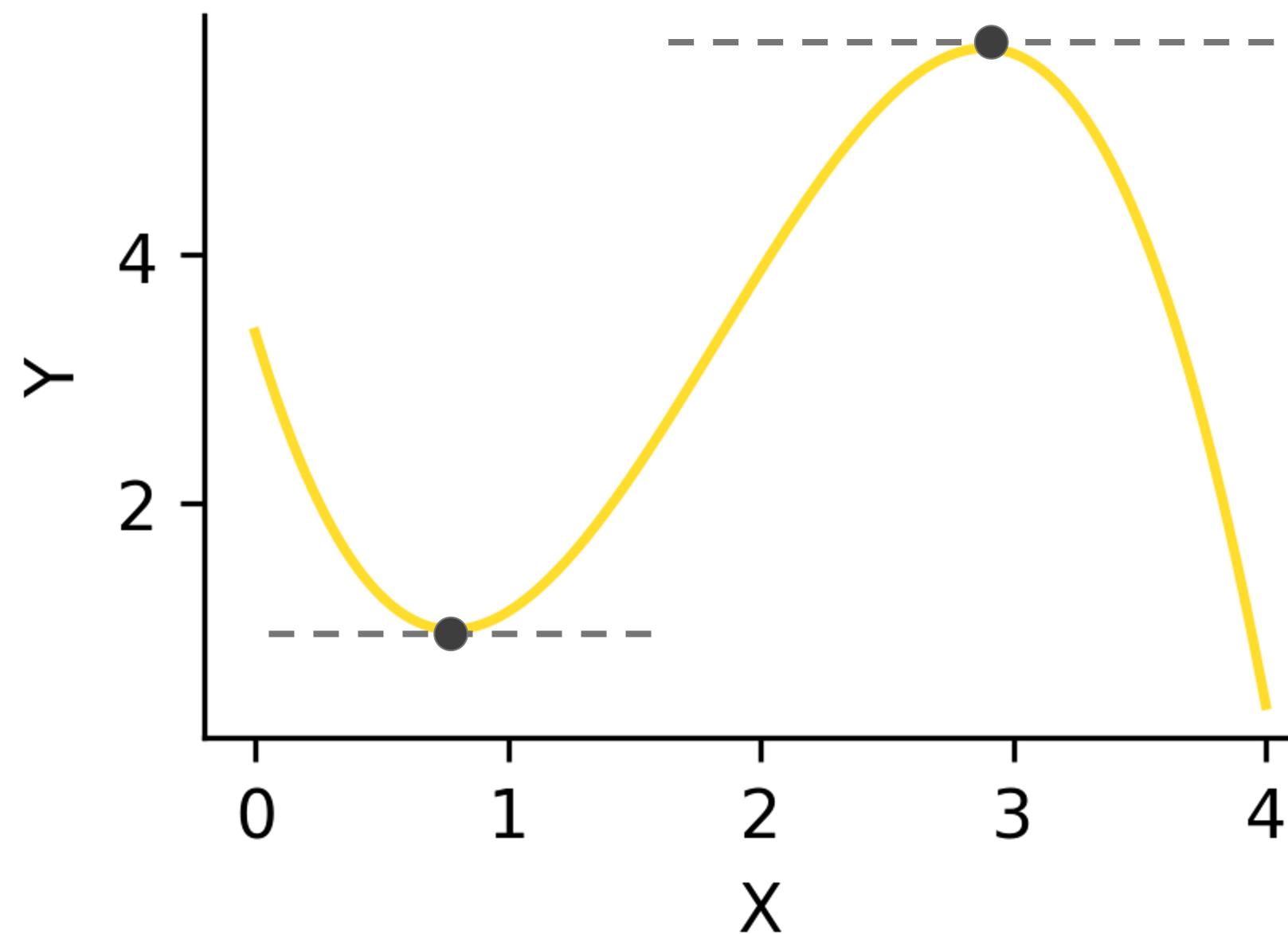
# Производная и локальные оптимумы

Если производная равна нулю в точке, то в её окрестности функция почти не растет и почти не убывает



# Производная и локальные оптимумы

Если производная равна нулю в точке, то в её окрестности функция почти не растет и почти не убывает



# Вопрос

Может ли производная  
не существовать в точке минимума?

[Приведите пример](#)

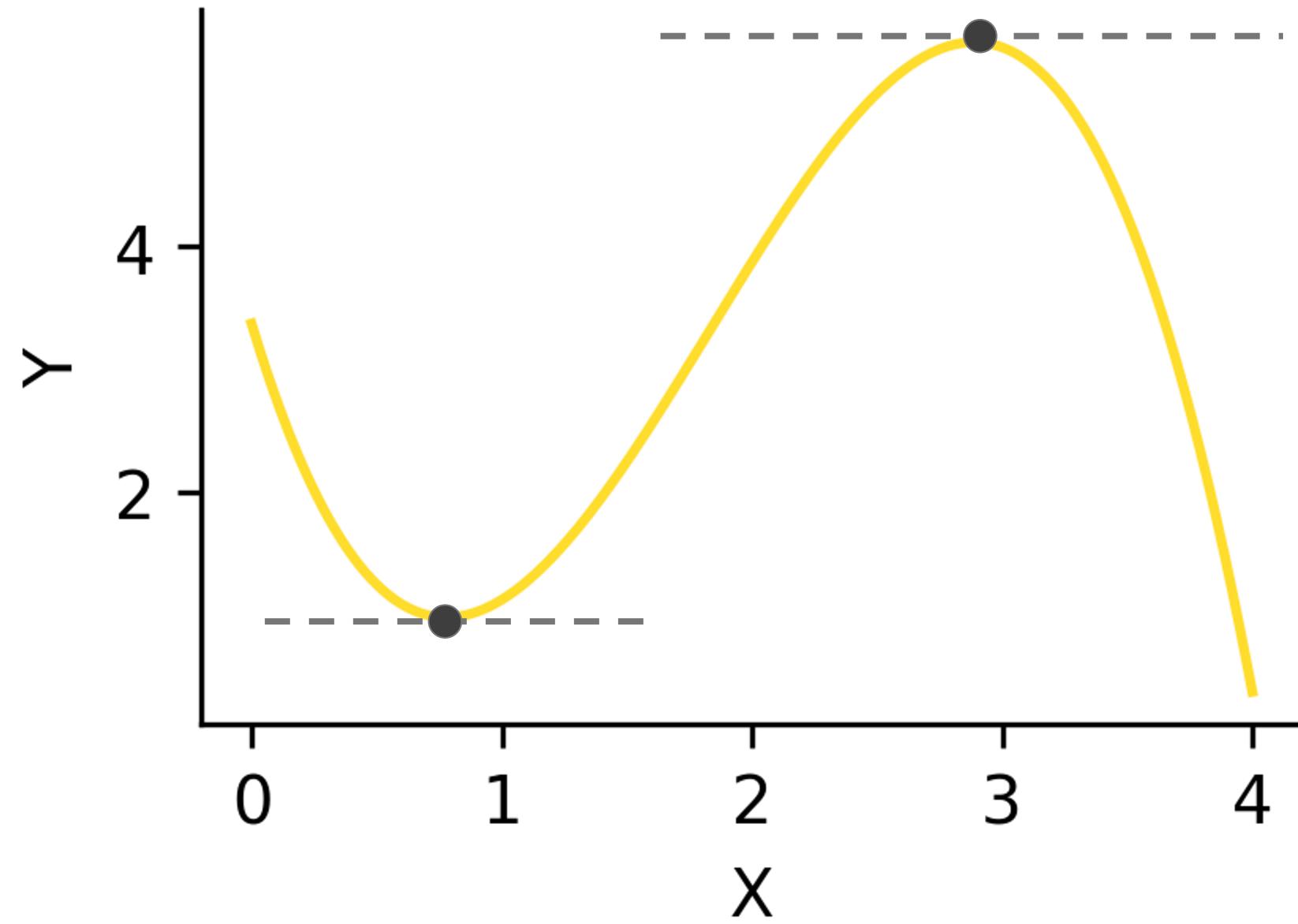


# Использование производных в оптимизации



Цель –

искать точки с производной,  
равной нулю



# Производные простых функций

$$(ax + b)' = a$$

$$(x^n)' = nx^{n-1}$$

$$(e^x)' = e^x$$

$$(\log x)' = \frac{1}{x}$$

# Производные простых функций

$$(ax + b)' = a$$

$$(x^n)' = nx^{n-1}$$

$$(e^x)' = e^x$$

$$(\log x)' = \frac{1}{x}$$

$$(f(x) + g(x))' = f'(x) + g'(x)$$

$$(f(x) * g(x))' = f'(x)g(x) + f(x)g'(x)$$

$$f(g(x))' = f'(g(x))g'(x)$$

# Задание

Вычислить производные  
и найти оптимумы



Ссылка в чате

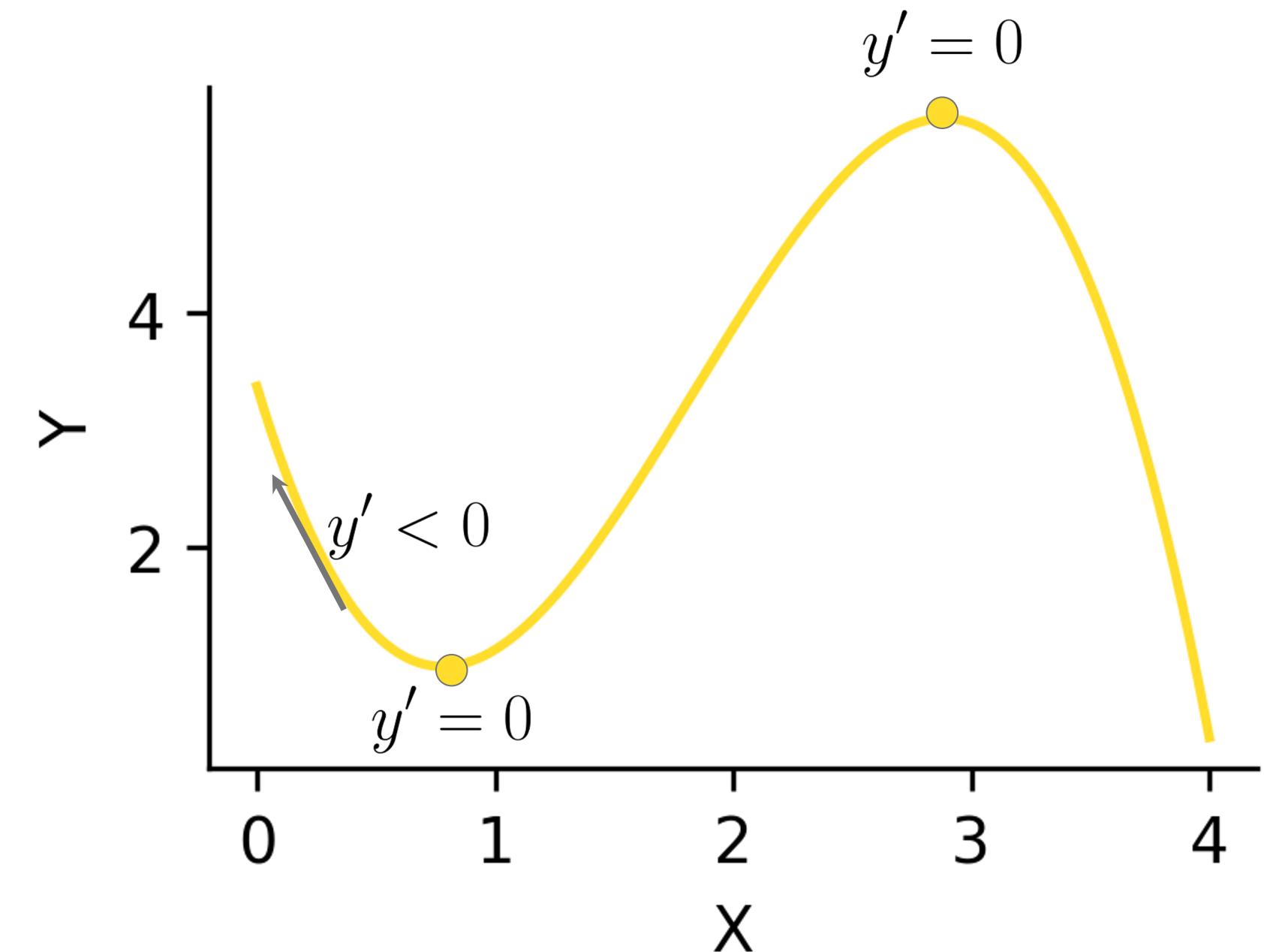


5 минут

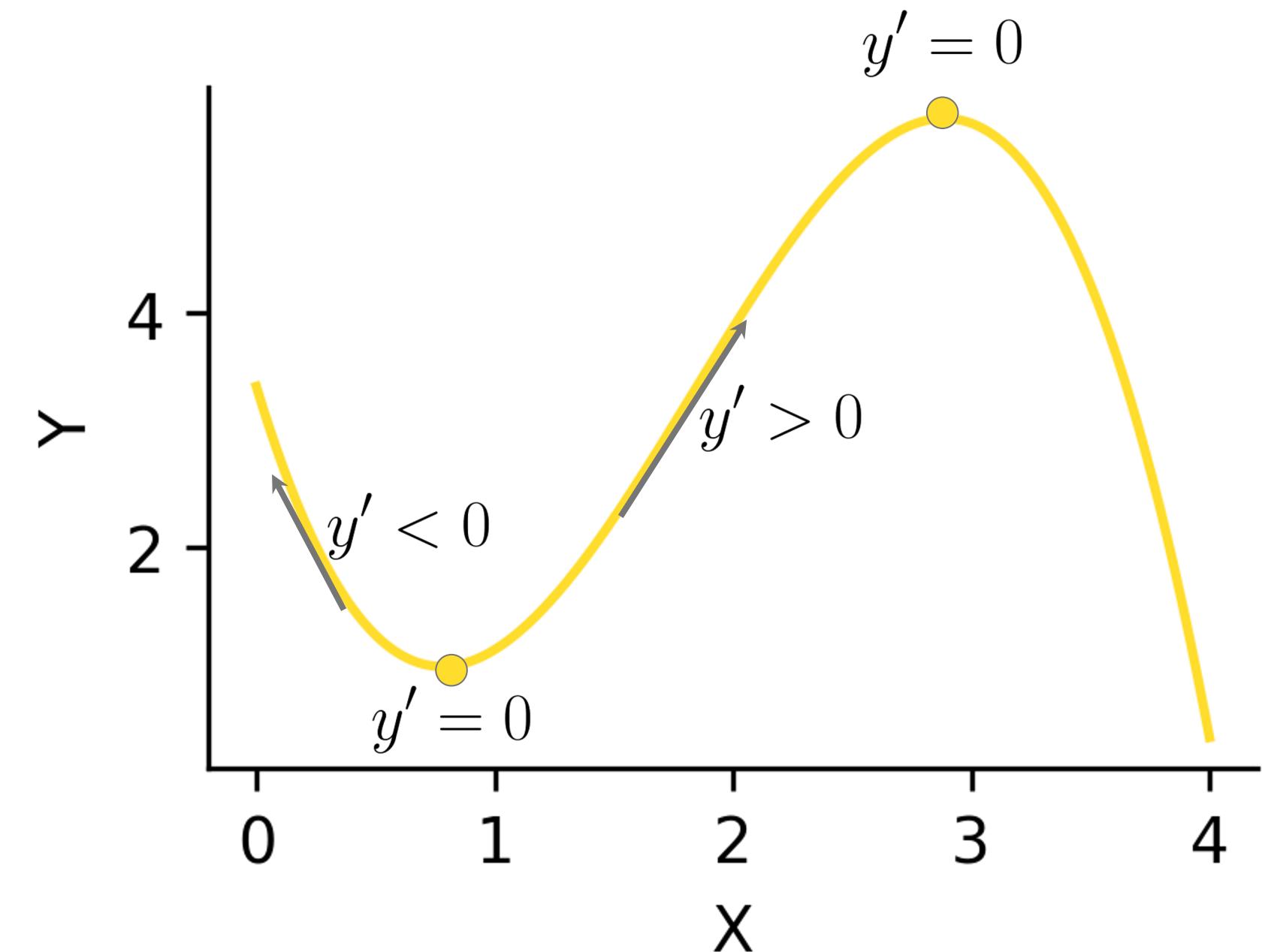


Анонимно

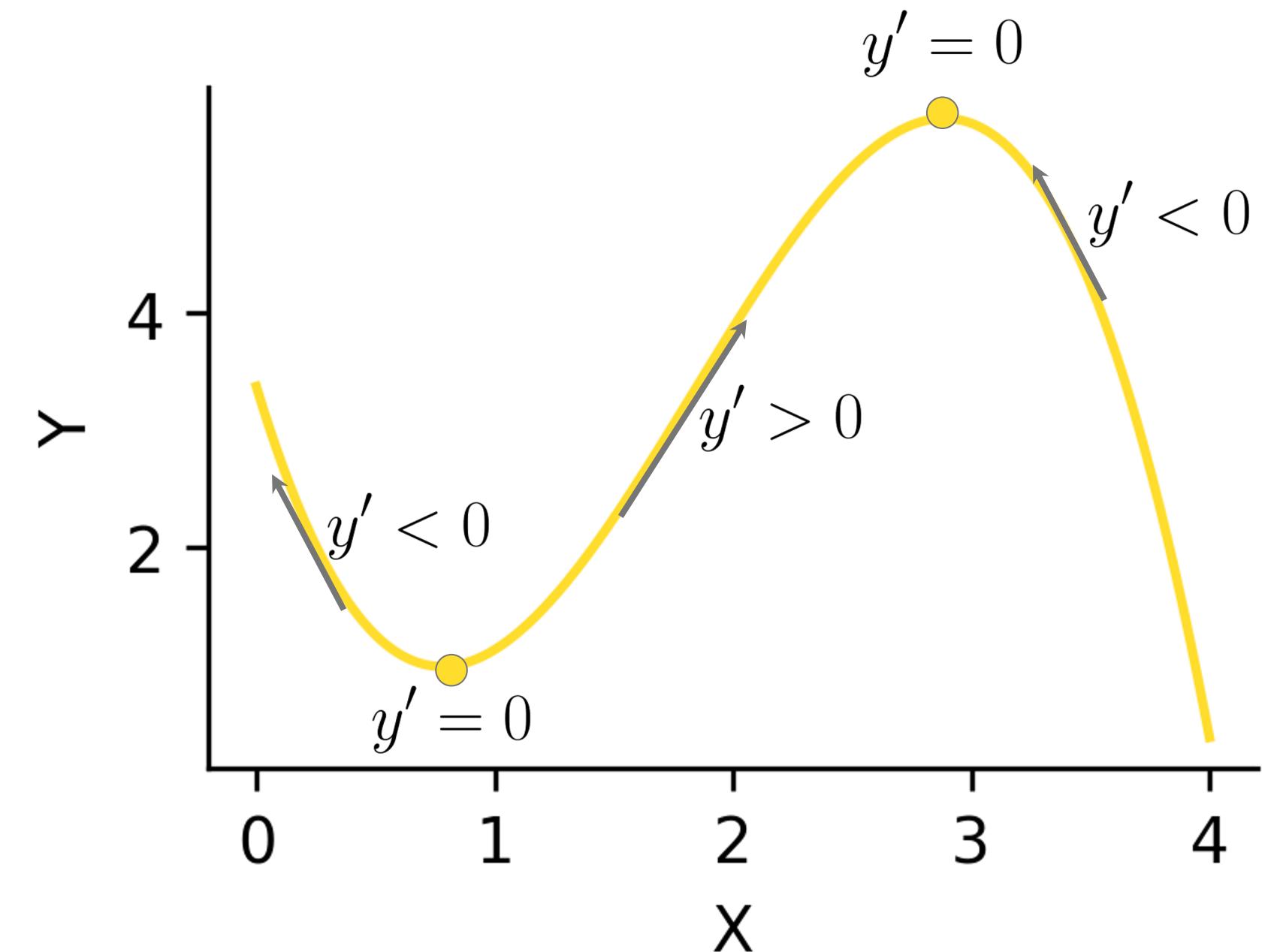
# Производная в окрестности локального оптимума



# Производная в окрестности локального оптимума



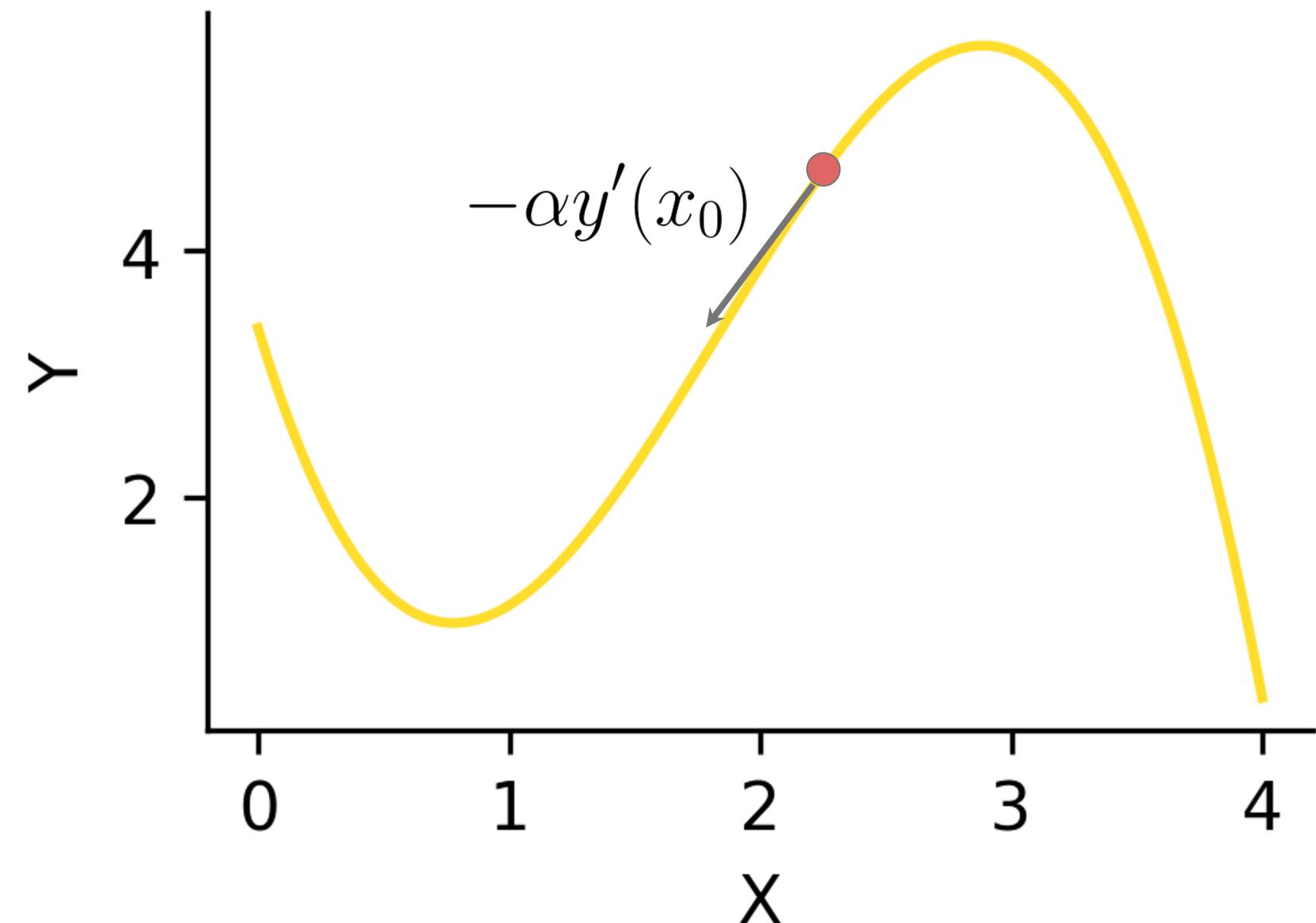
# Производная в окрестности локального оптимума



# Градиентный спуск

$$x_{n+1} = x_n - \alpha y'(x_n)$$

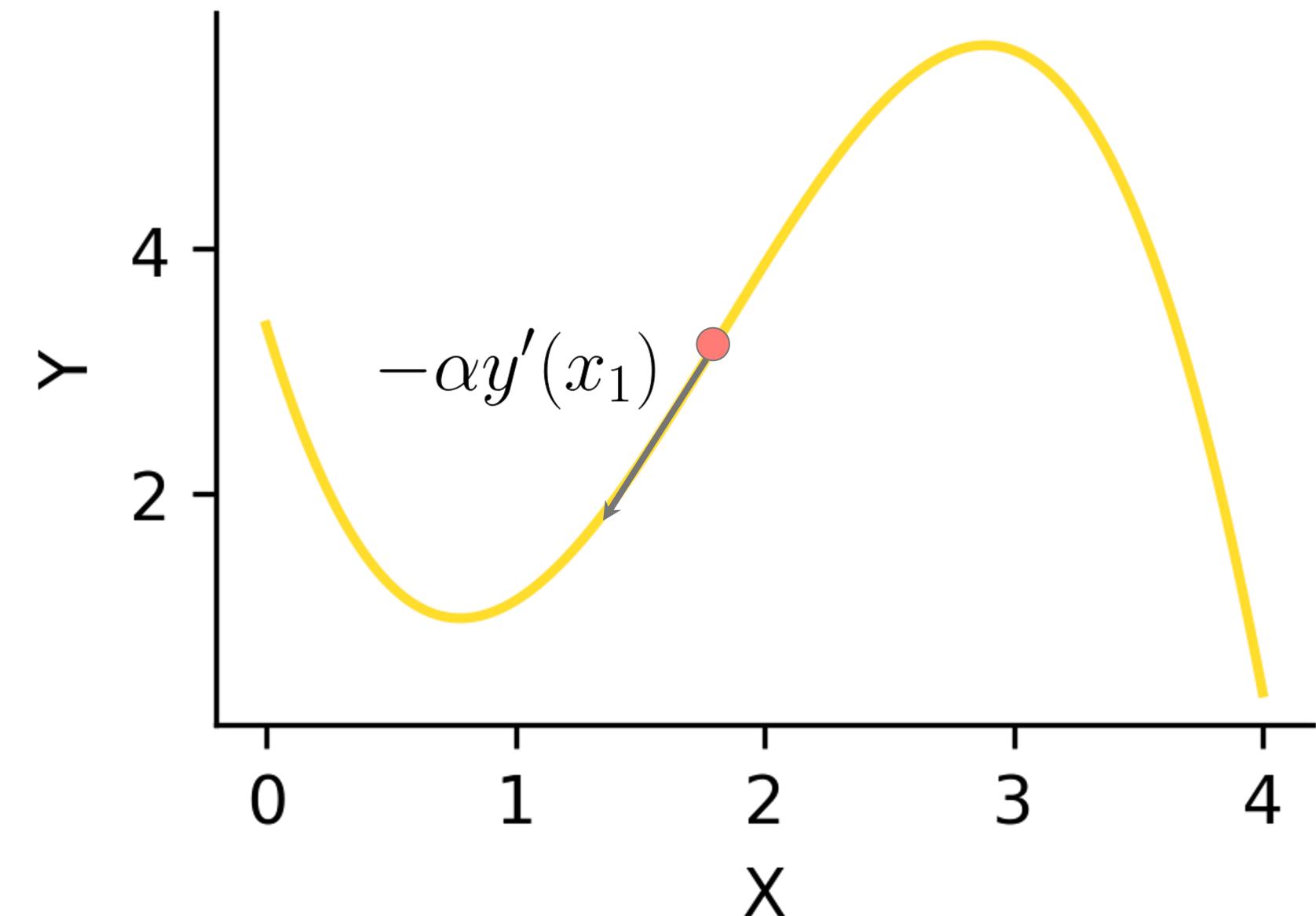
$x_0$  – случайное



# Градиентный спуск

$$x_{n+1} = x_n - \alpha y'(x_n)$$

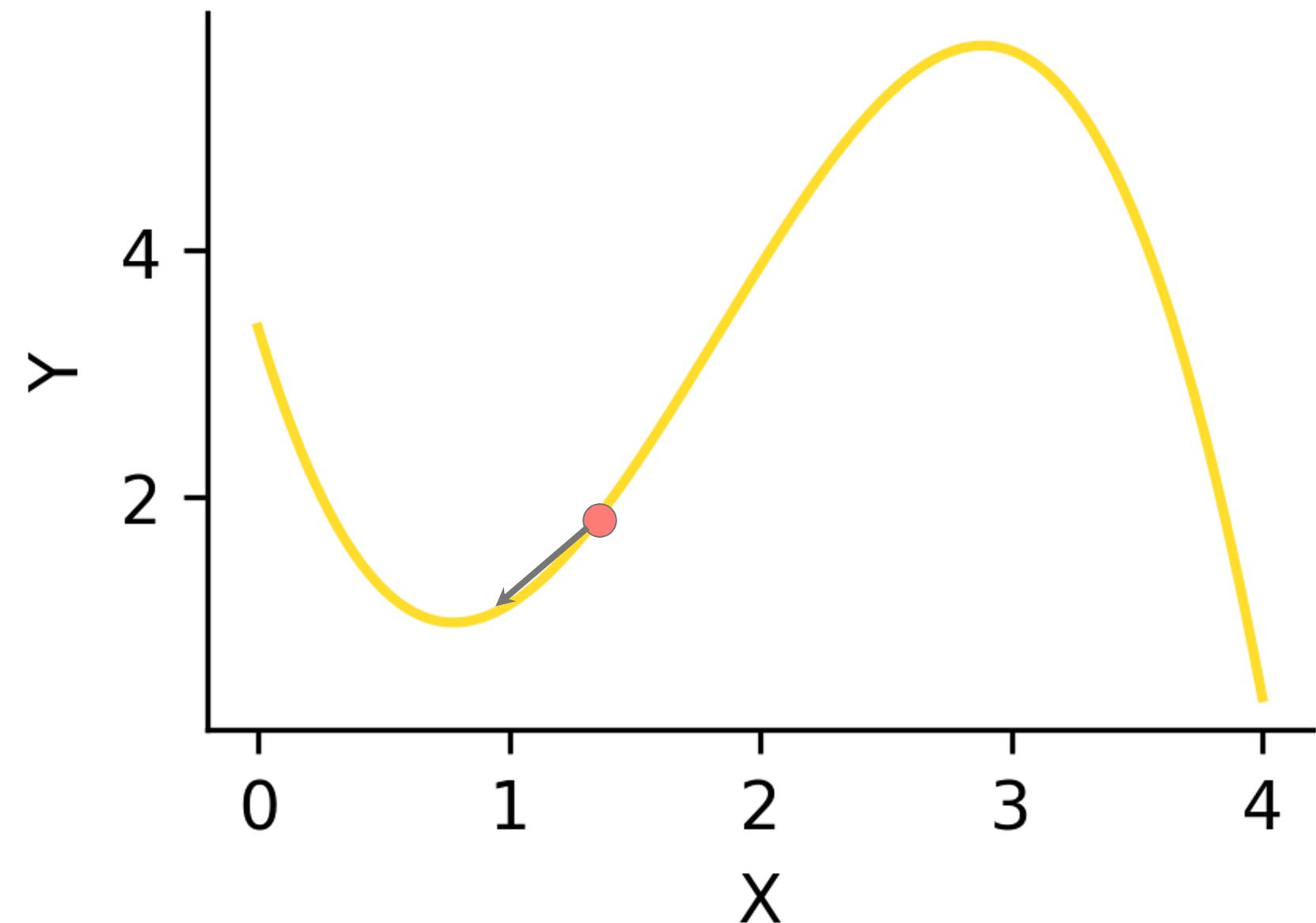
$x_0$  – случайное



# Градиентный спуск

$$x_{n+1} = x_n - \alpha y'(x_n)$$

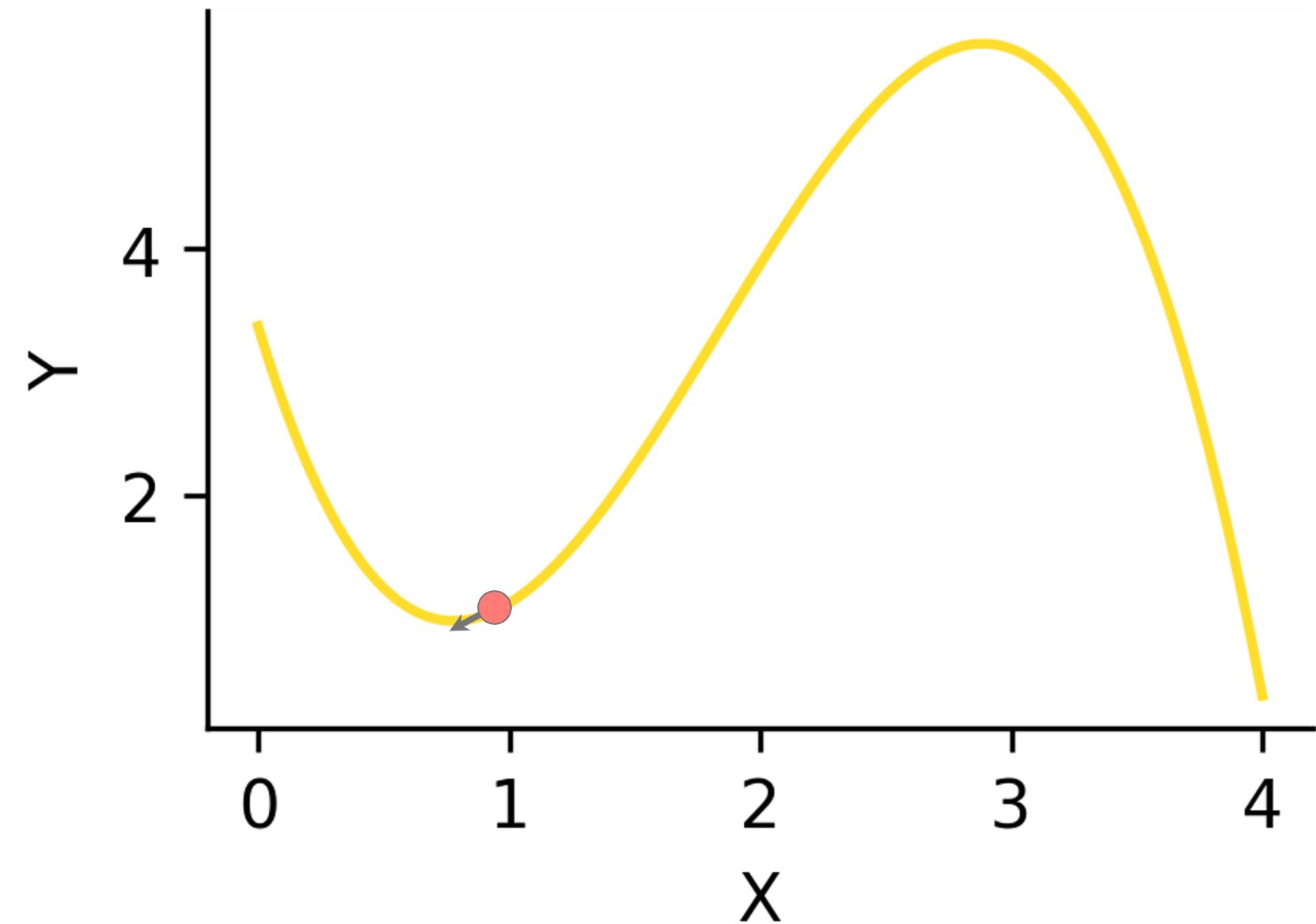
$x_0$  – случайное



# Градиентный спуск

$$x_{n+1} = x_n - \alpha y'(x_n)$$

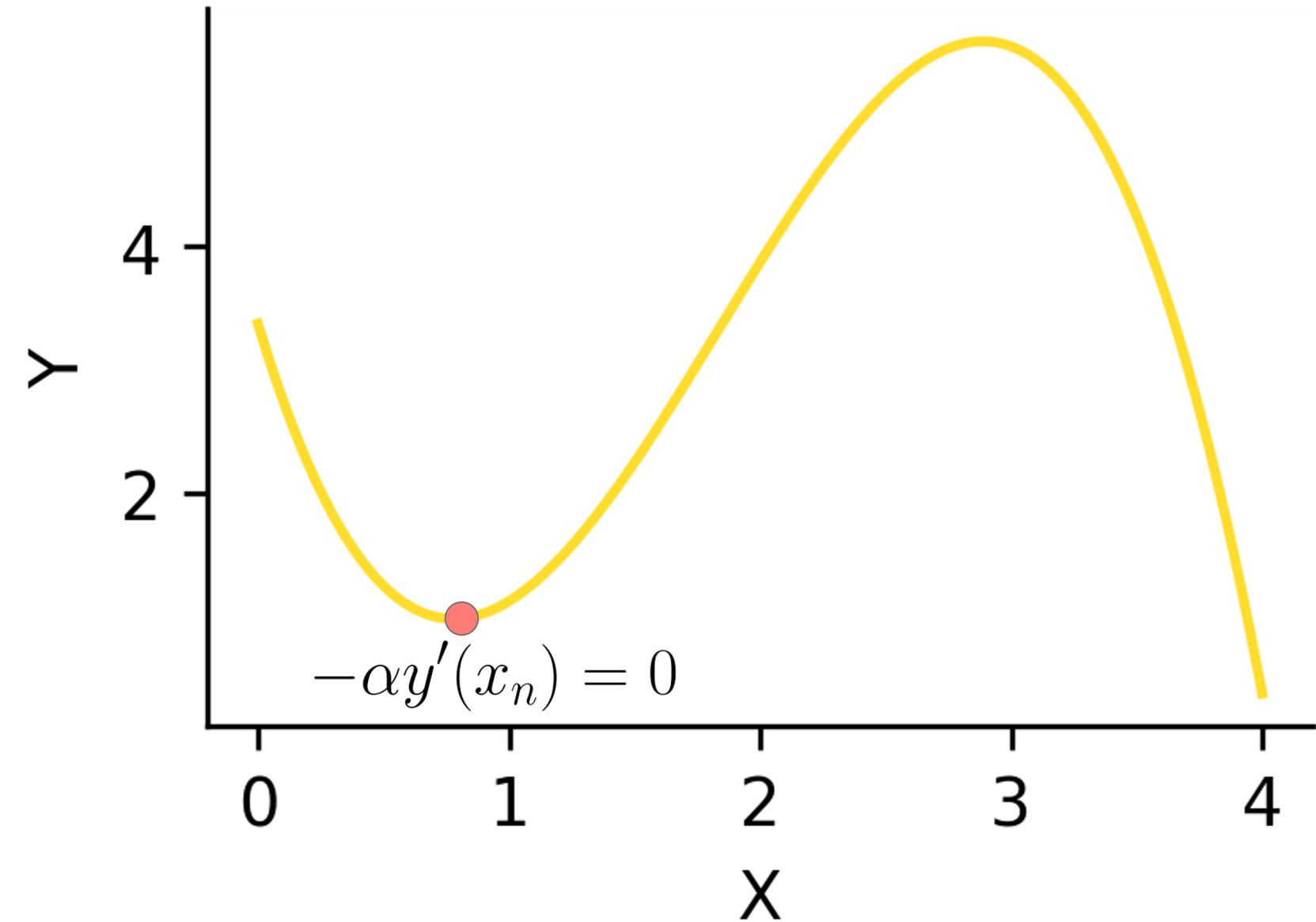
$x_0$  – случайное



# Градиентный спуск

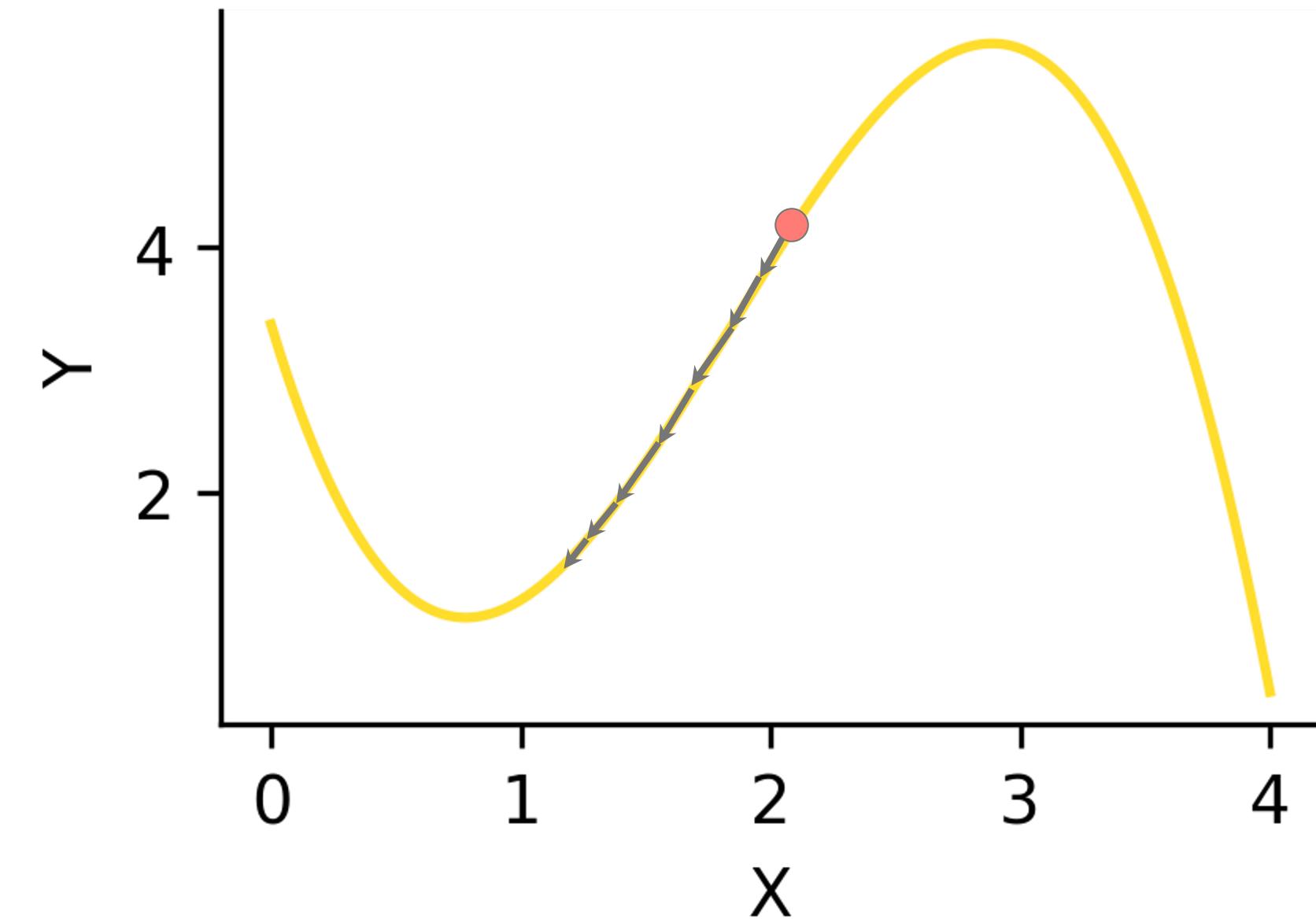
$$x_{n+1} = x_n - \alpha y'(x_n)$$

$x_0$  – случайное



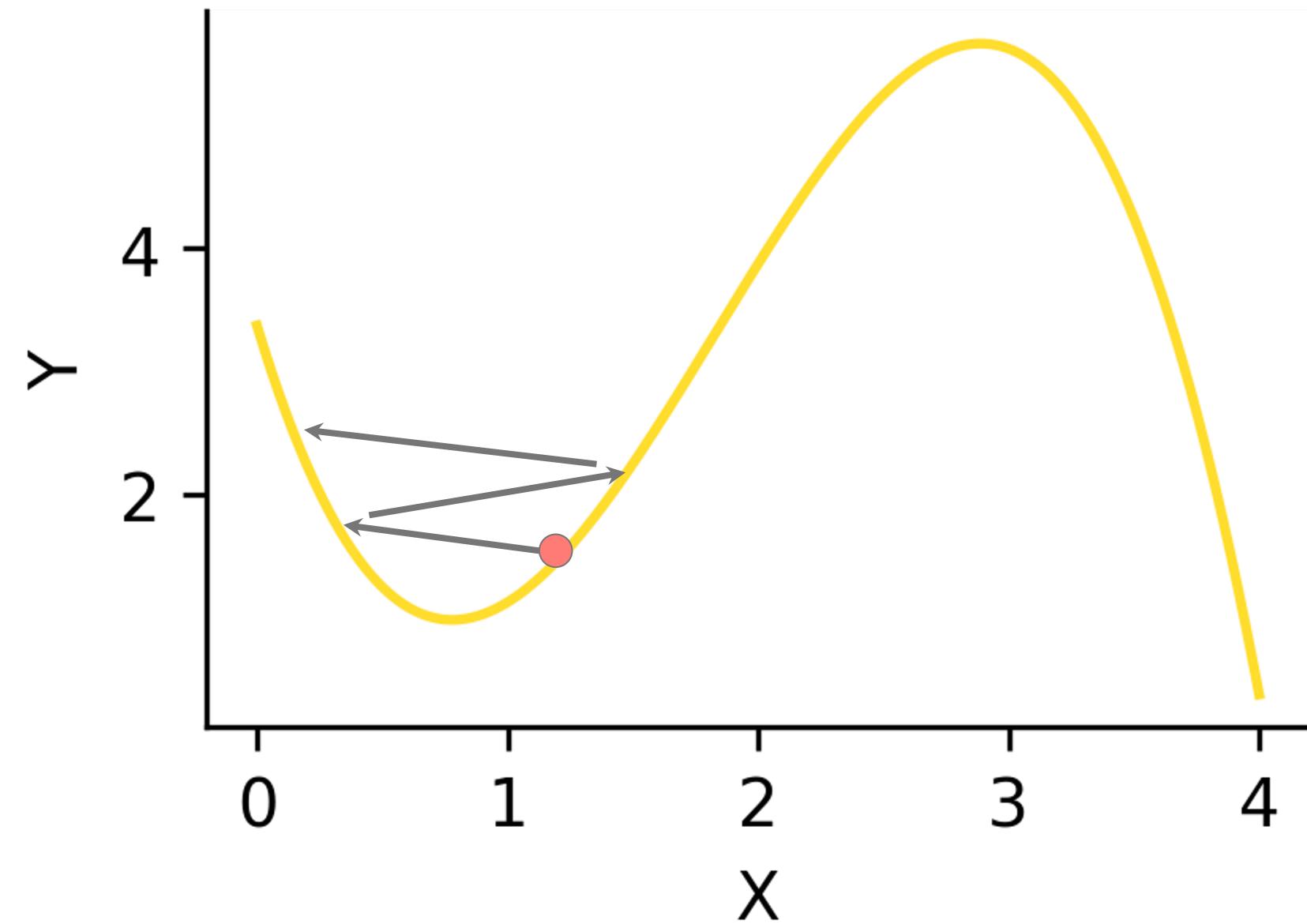
# Learning rate

Если слишком маленький →  
долгая сходимость



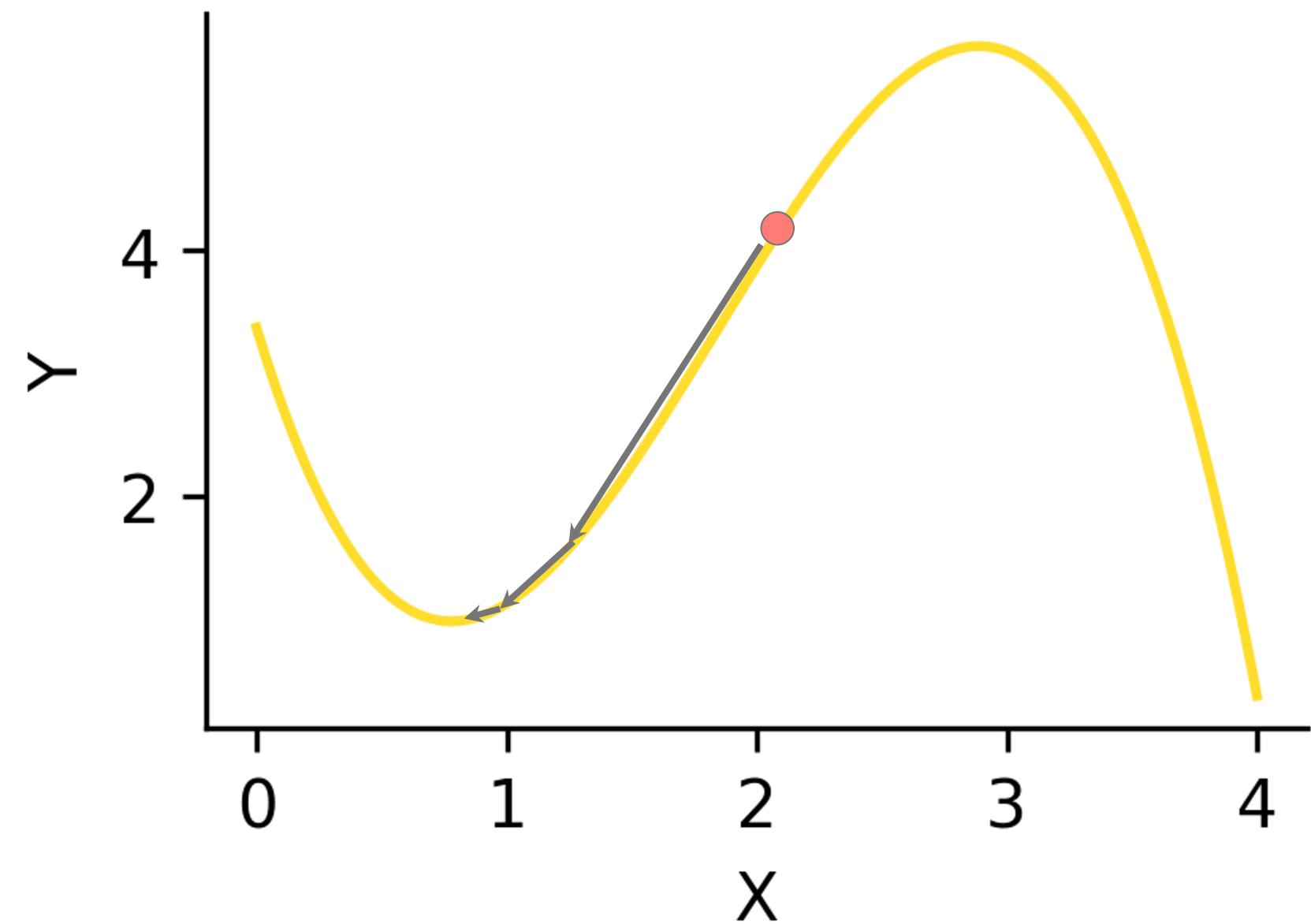
# Learning rate

Если слишком большой →  
отсутствие сходимости



# Learning rate

Если угадать → сходимость  
за оптимальное число шагов



# Задание

Реализовать gradient descent



**Ссылка на ноутбук в чате  
(1-gradient-descent-1D)**



**20 минут**



**Можно шарить экран  
и задавать вопросы  
(Zoom комнаты?)**

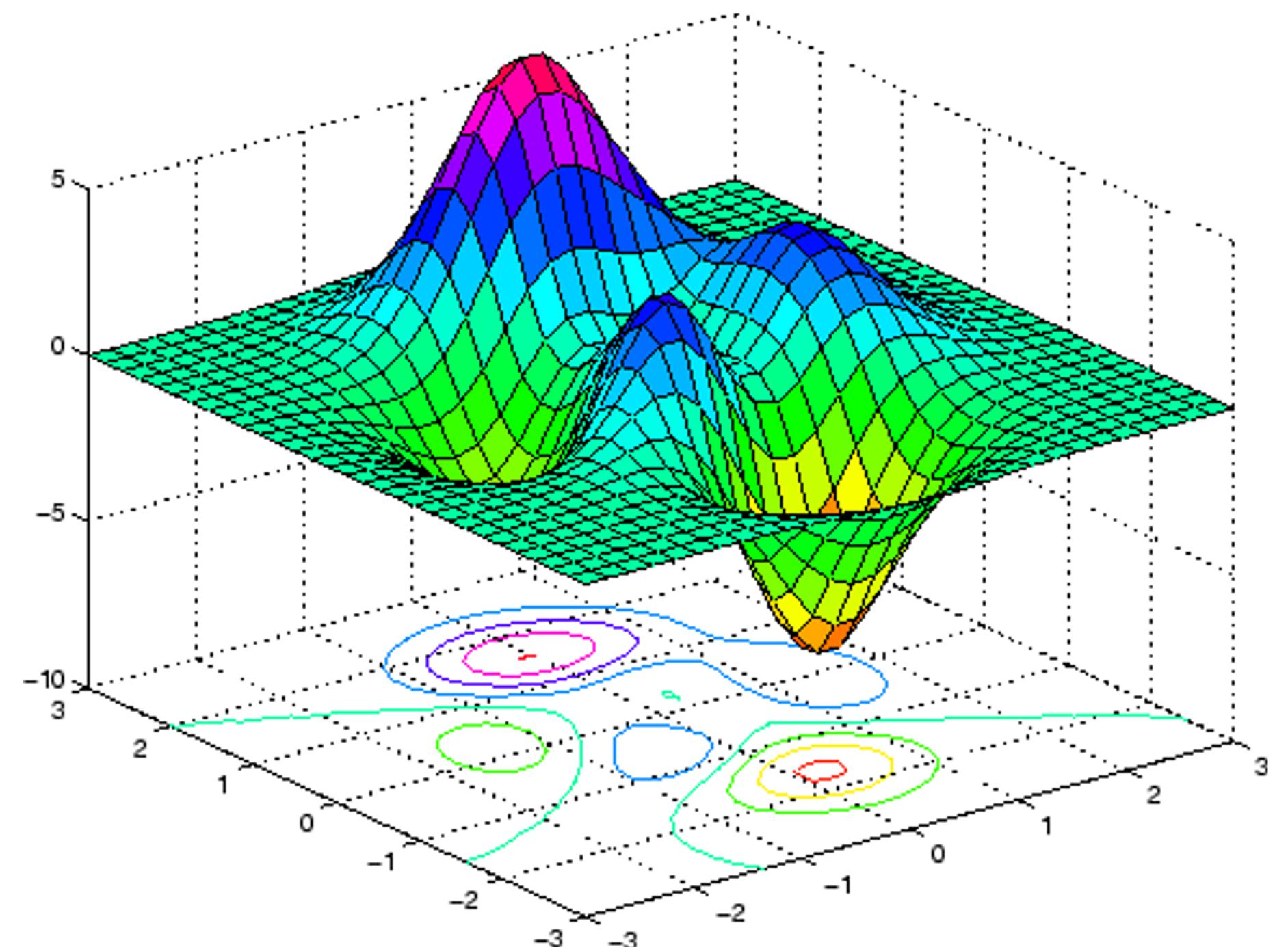


**ТИНЬКОФФ**

# Многомерный случай

# ФУНКЦИЯ НЕСКОЛЬКИХ ПЕРЕМЕННЫХ

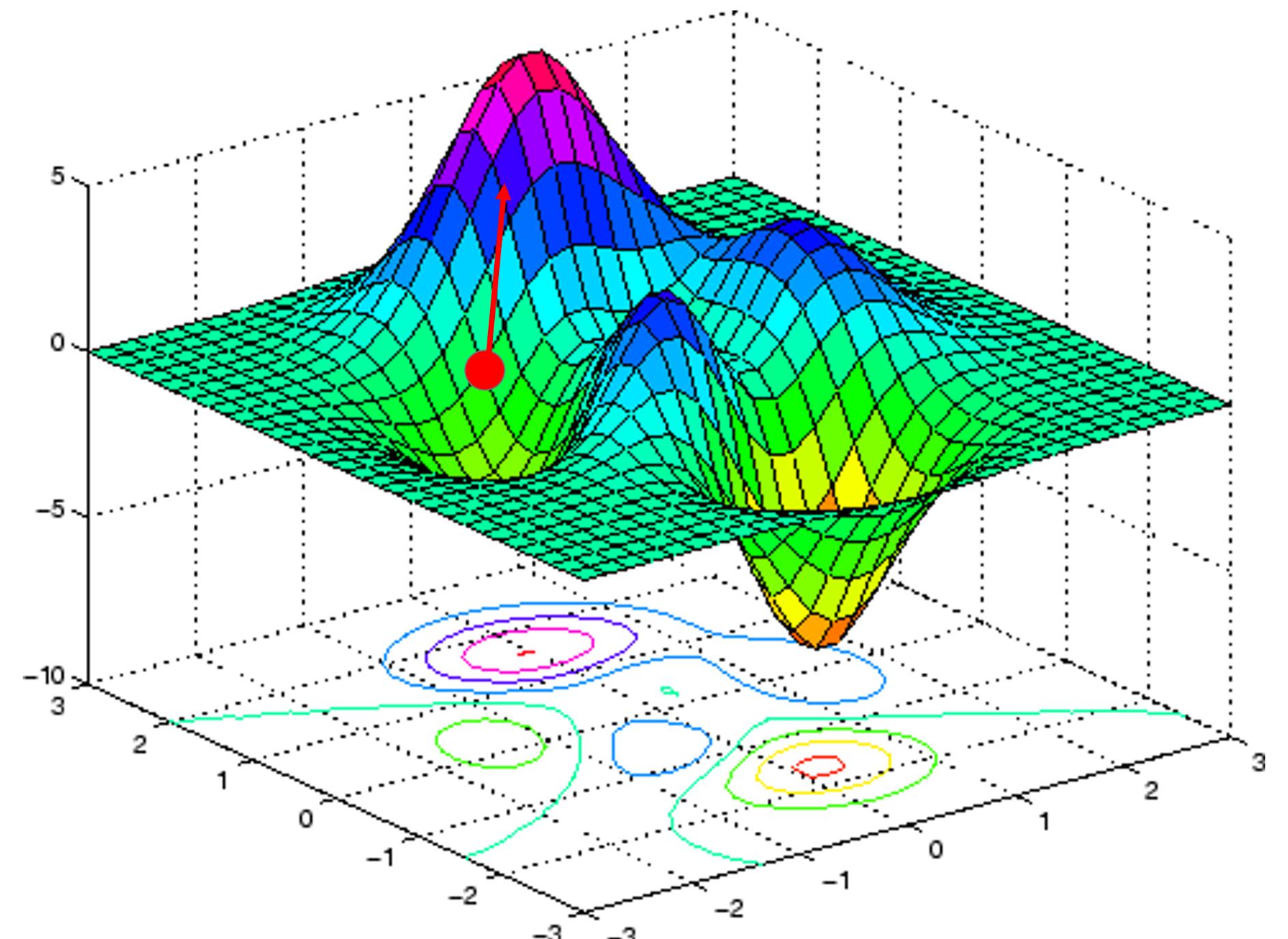
$$y = f(x_1, x_2, \dots, x_n)$$



# Градиент

$$y = f(x_1, x_2, \dots, x_n)$$

$$\nabla y = \begin{pmatrix} f'_{x_1}(x_1, x_2, \dots, x_n) \\ f'_{x_2}(x_1, x_2, \dots, x_n) \\ \vdots \\ f'_{x_n}(x_1, x_2, \dots, x_n) \end{pmatrix}$$



# Пример

$$y = x_1^2 - x_2$$

# Пример

$$y = x_1^2 - x_2$$

$$y'_{x_1} = 2x_1$$

$$y'_{x_2} = -1$$

# Пример

$$y = x_1^2 - x_2$$

$$y'_{x_1} = 2x_1$$

$$y'_{x_2} = -1$$

$$\nabla y(1, 2) = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

# Задание



**Вычислите градиент  
функции**

$$f(x_1, x_2) = e^{x_1} x_2^2$$

**в точке (0, 1)**



**Ответ в чат**

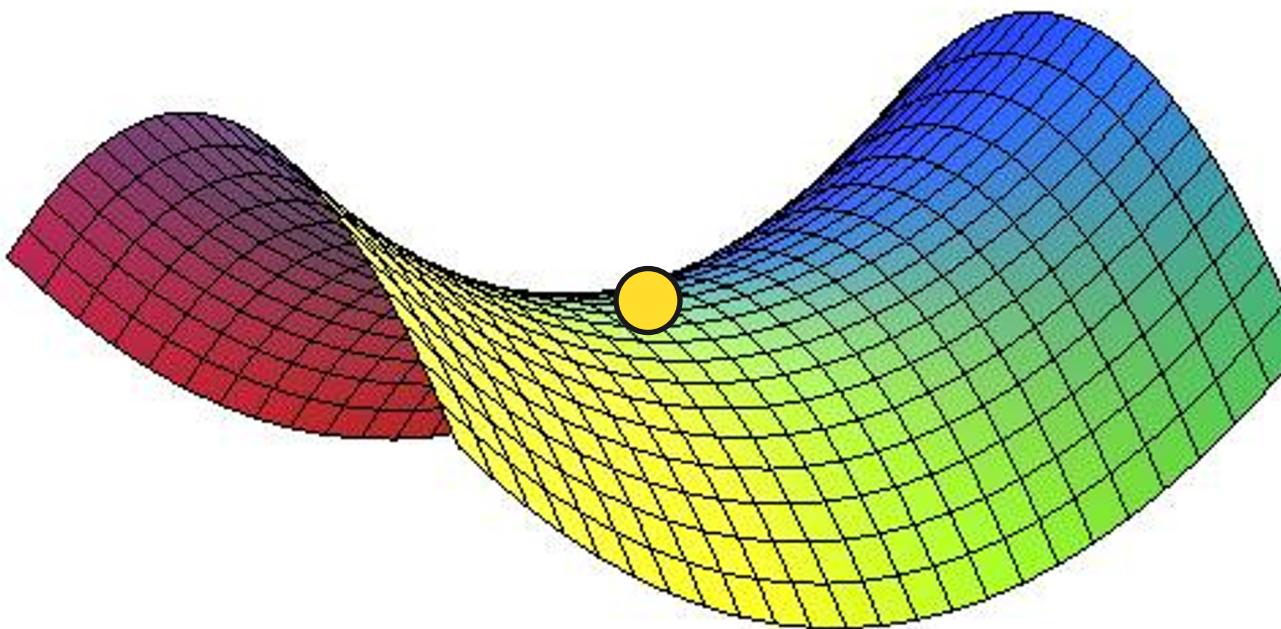
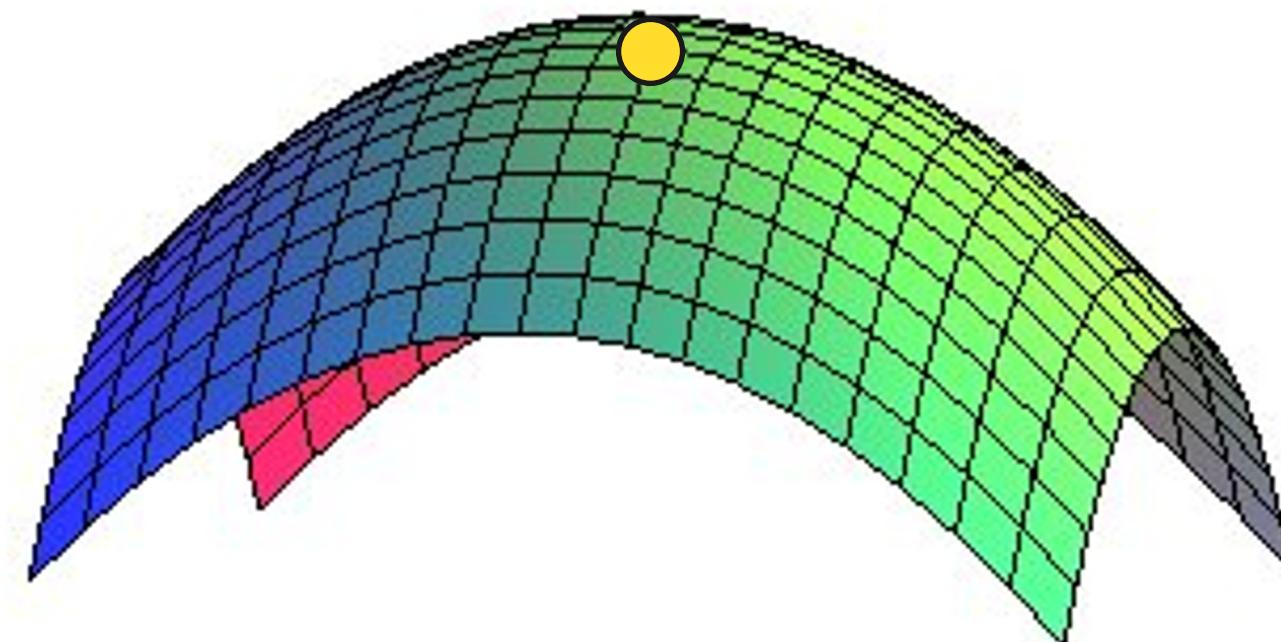
# Градиент и локальные оптимумы



Градиент –  
направление наискорейшего роста

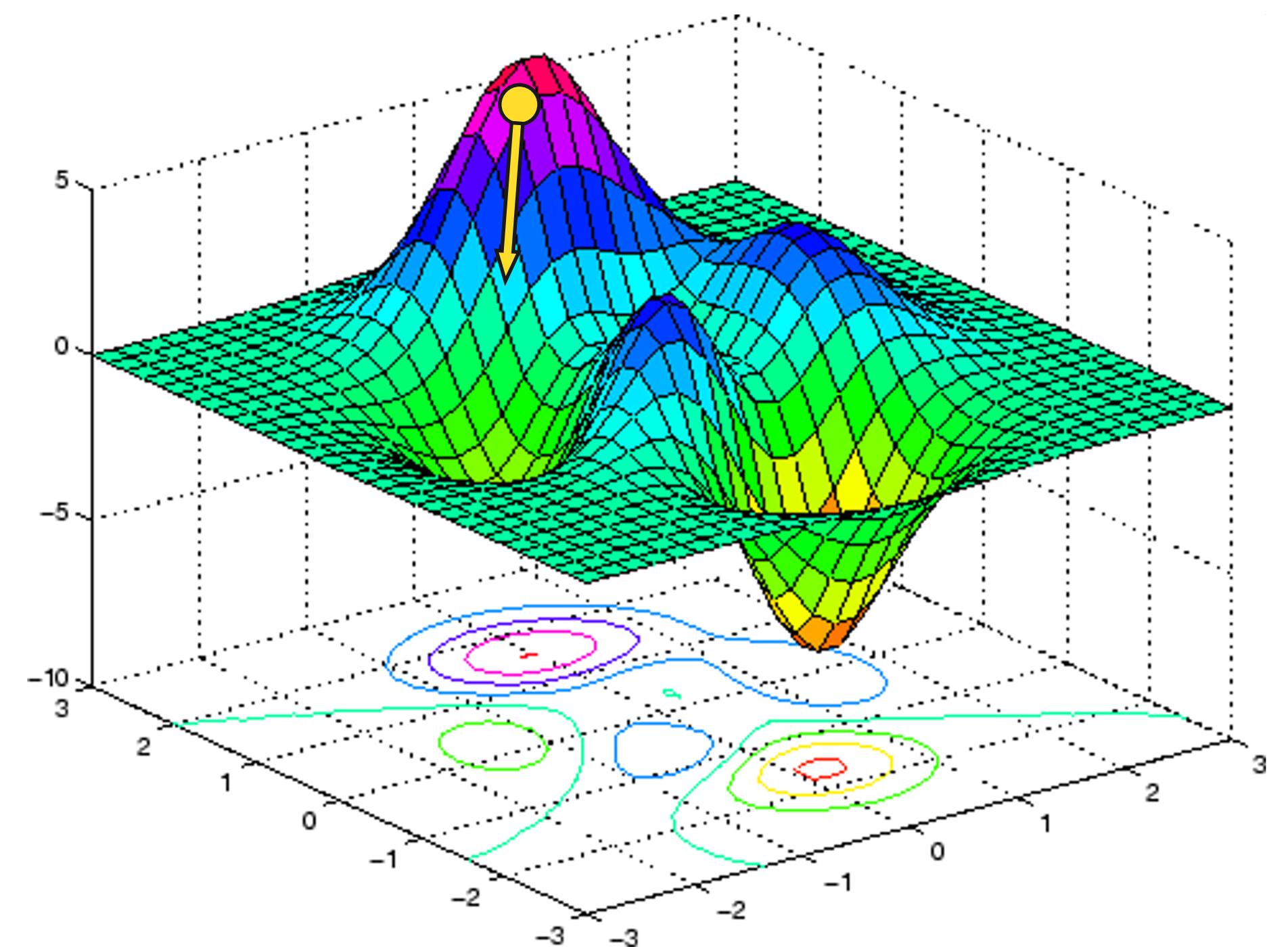


Если градиент равен нулевому вектору →  
локальный оптимум или седловая точка



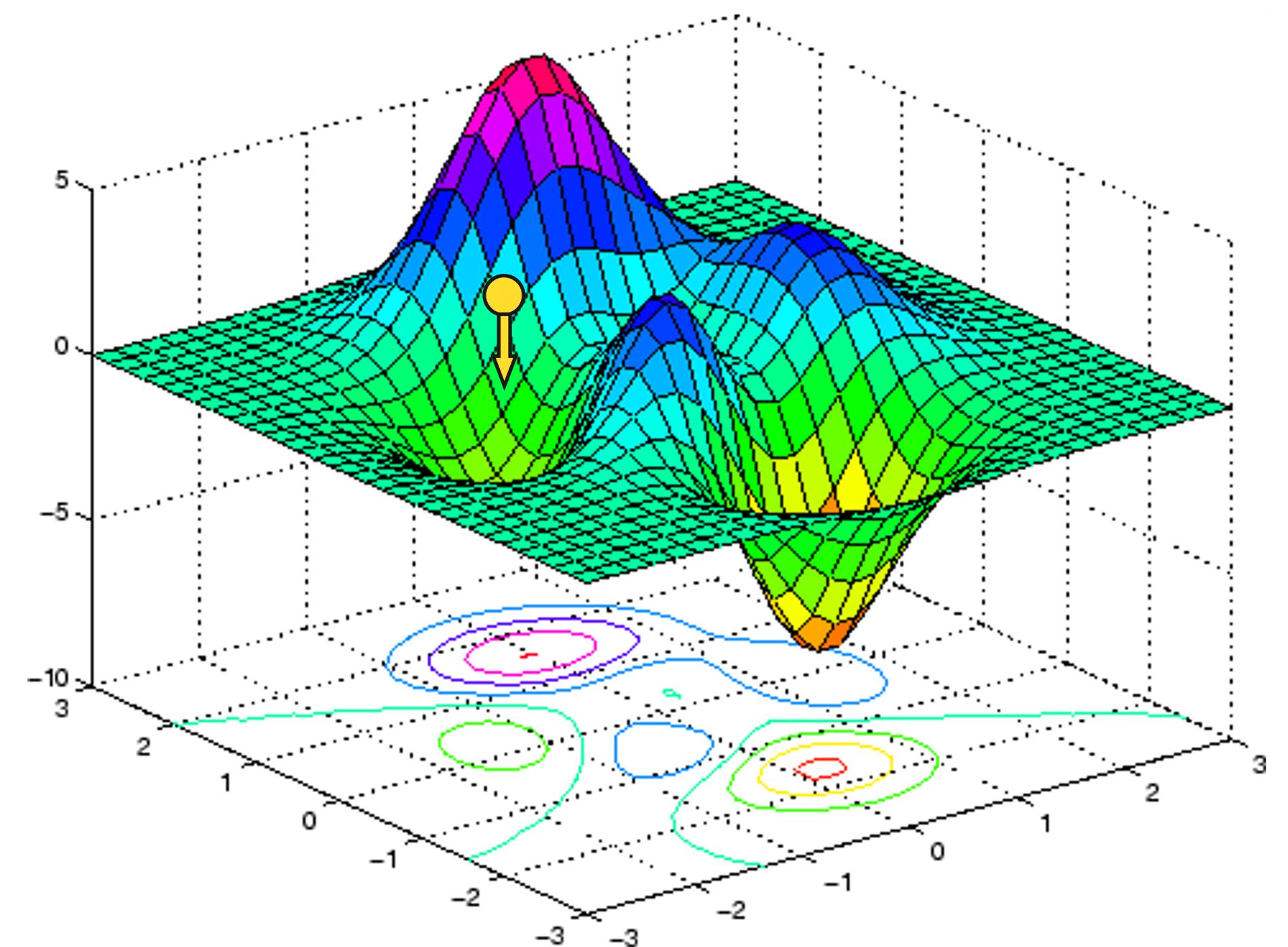
# Градиентный спуск

Шаг в направлении,  
противоположном градиенту  
функции в текущей точке



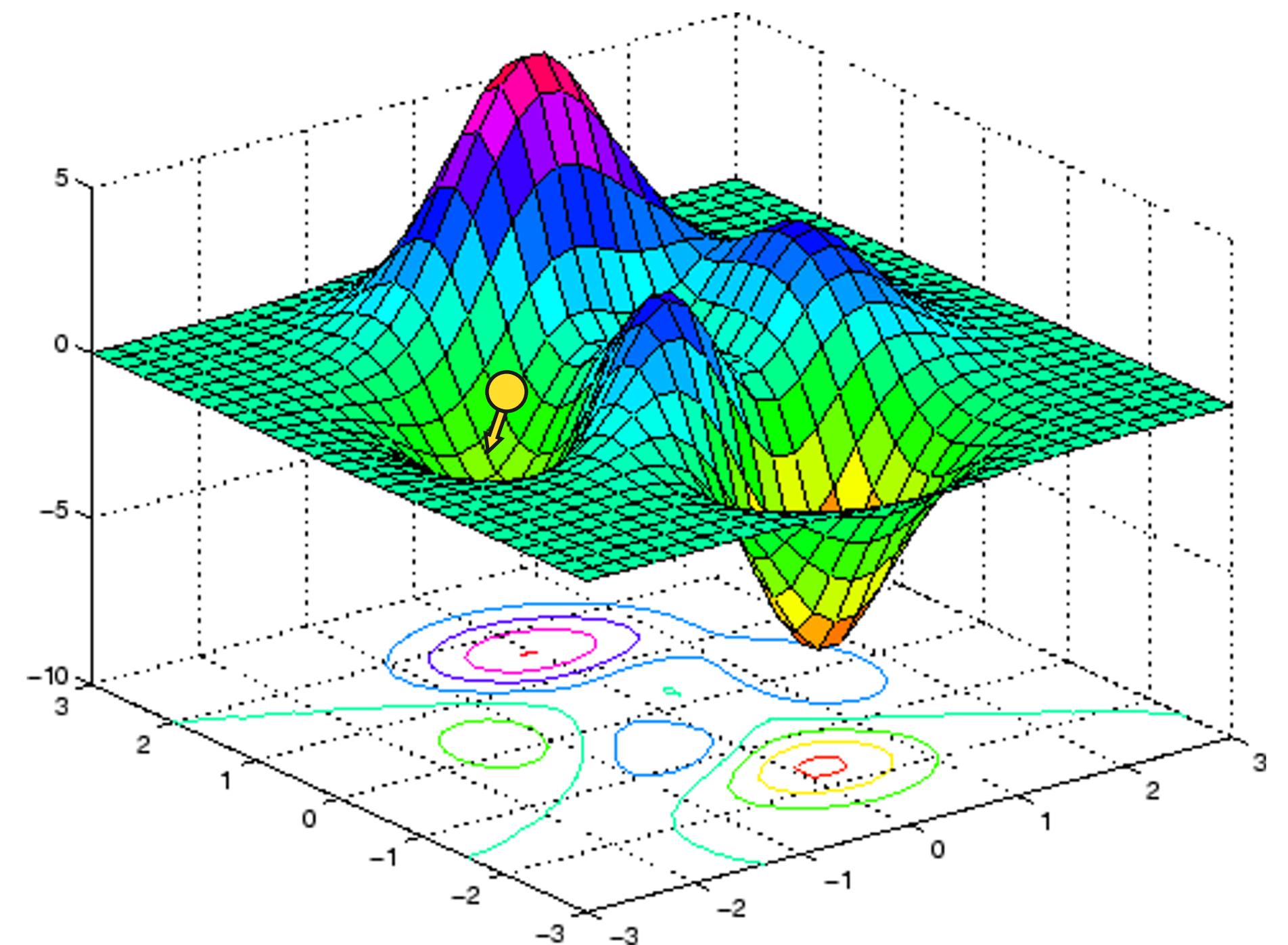
# Градиентный спуск

Шаг в направлении,  
противоположном градиенту  
функции в текущей точке



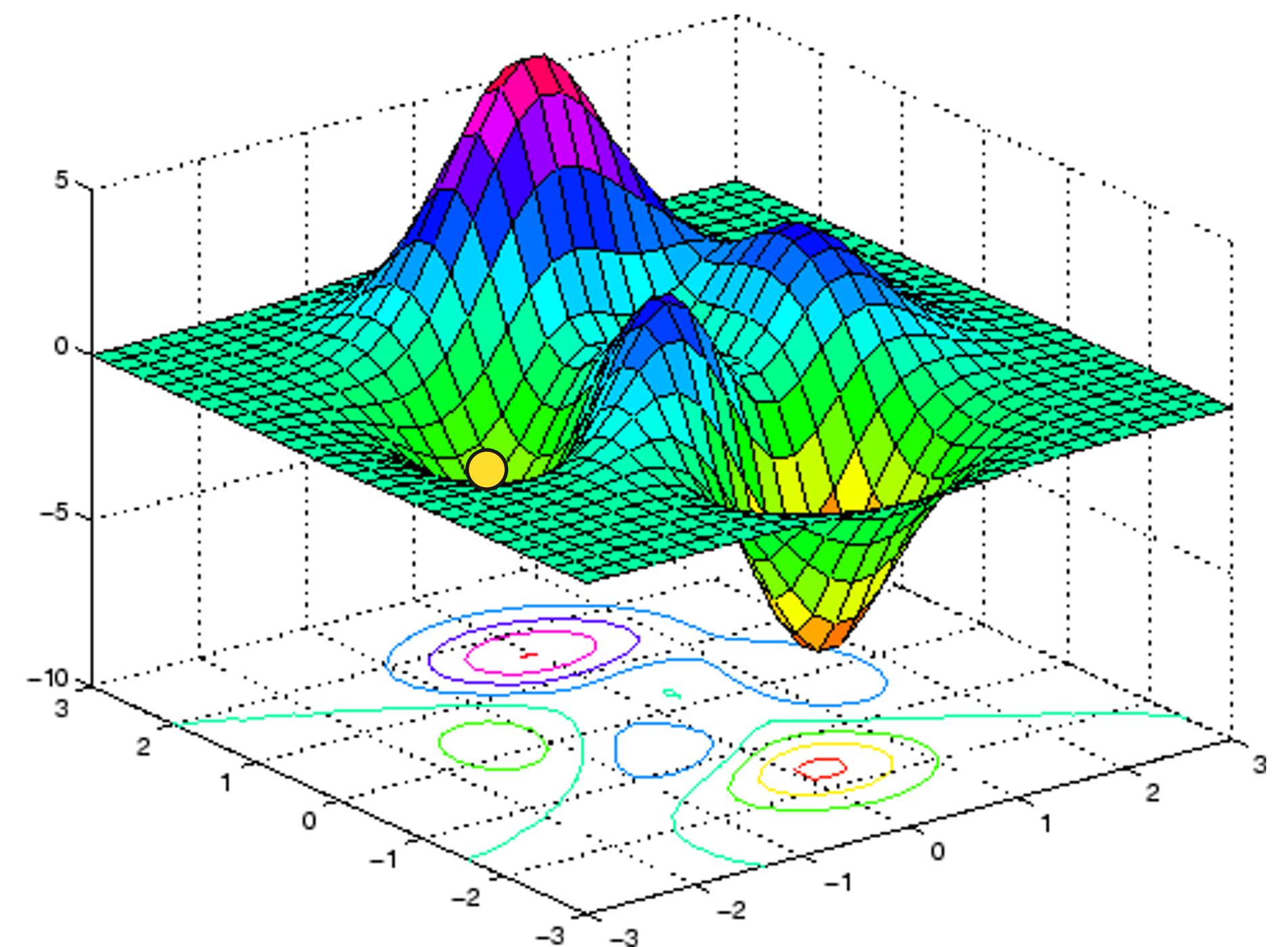
# Градиентный спуск

Шаг в направлении,  
противоположном градиенту  
функции в текущей точке



# Градиентный спуск

Шаг в направлении,  
противоположном градиенту  
функции в текущей точке





# Перерыв ;)

**После:** вычисление  
градиентов сложных функций

# Вычисление градиентов сложных функций

$$y = f(x_1, g(x_2, x_3))$$

$$\nabla y = ?$$

# Вычисление градиентов сложных функций

$$y = f(x_1, g(x_2, x_3))$$

$$\nabla y = ?$$

➡ Какой размерности  
вектор?

# Вычисление градиентов сложных функций

Матричные операции

Линейная аппроксимация  
(матрица Якоби)

Градиент сложной функции

# Матрицы

**Двумерный массив чисел:**

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{pmatrix}$$

# Матрицы

**Двумерный массив чисел:**

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{pmatrix}$$

★ Пример:

$$A = \begin{pmatrix} 0.2 & -3 & 2 \\ 0 & 0 & 5 \end{pmatrix}$$

$$A + B = \begin{pmatrix} a_{1,1} + b_{1,1} & \dots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & \dots & a_{2,n} + b_{2,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} + b_{m,1} & \dots & a_{m,n} + b_{m,n} \end{pmatrix}$$

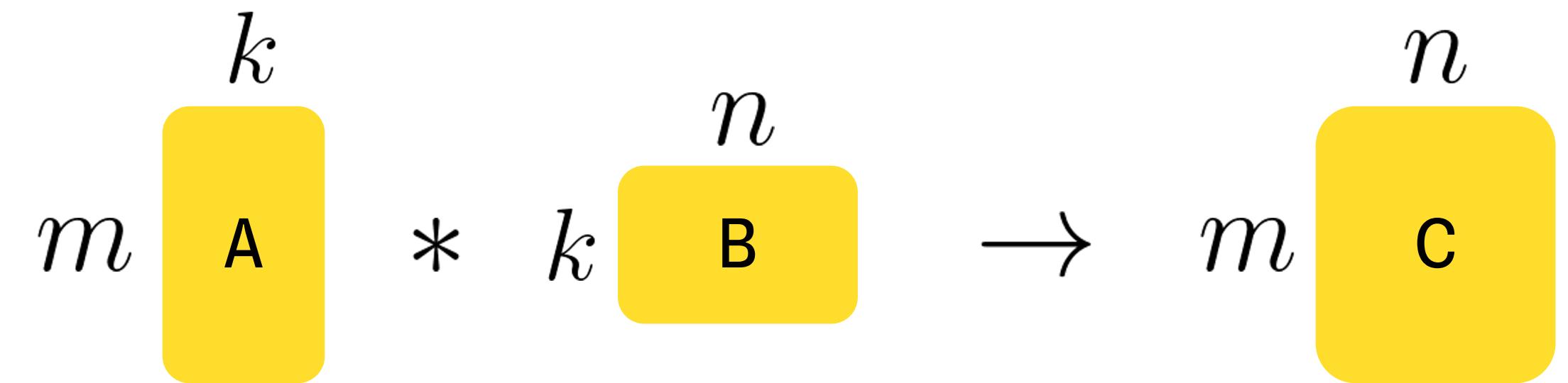
# Сложение матриц

$$A + B = \begin{pmatrix} a_{1,1} + b_{1,1} & \dots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & \dots & a_{2,n} + b_{2,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} + b_{m,1} & \dots & a_{m,n} + b_{m,n} \end{pmatrix}$$

# Сложение матриц

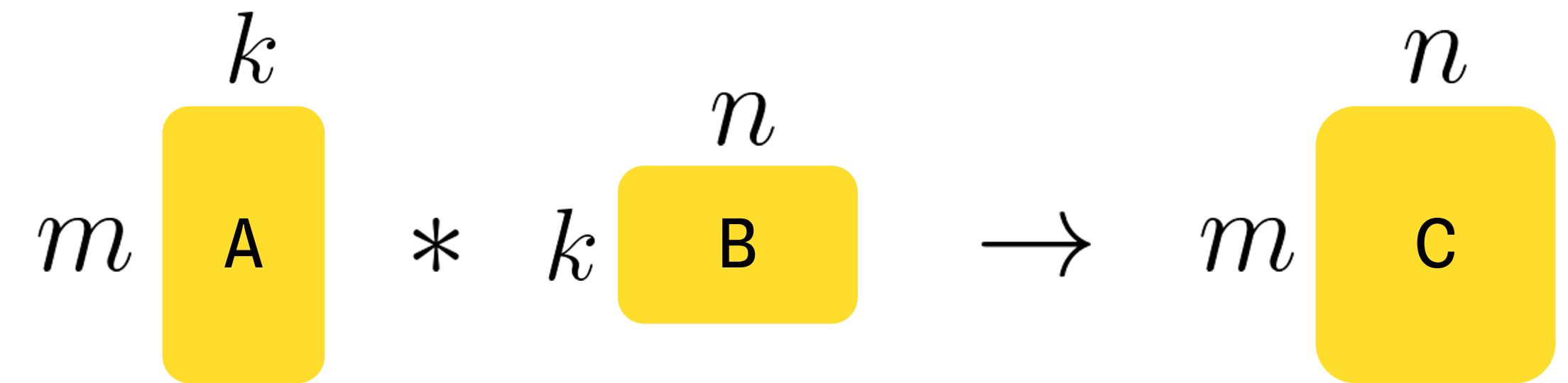
★ Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1.2 \end{pmatrix}$$



# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

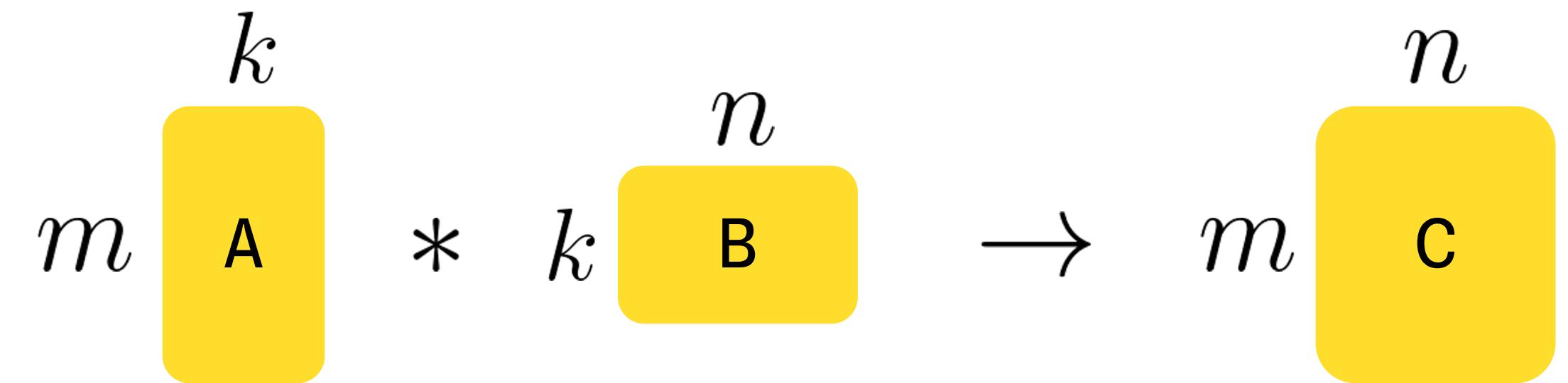


# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

$$AB \neq BA$$

---

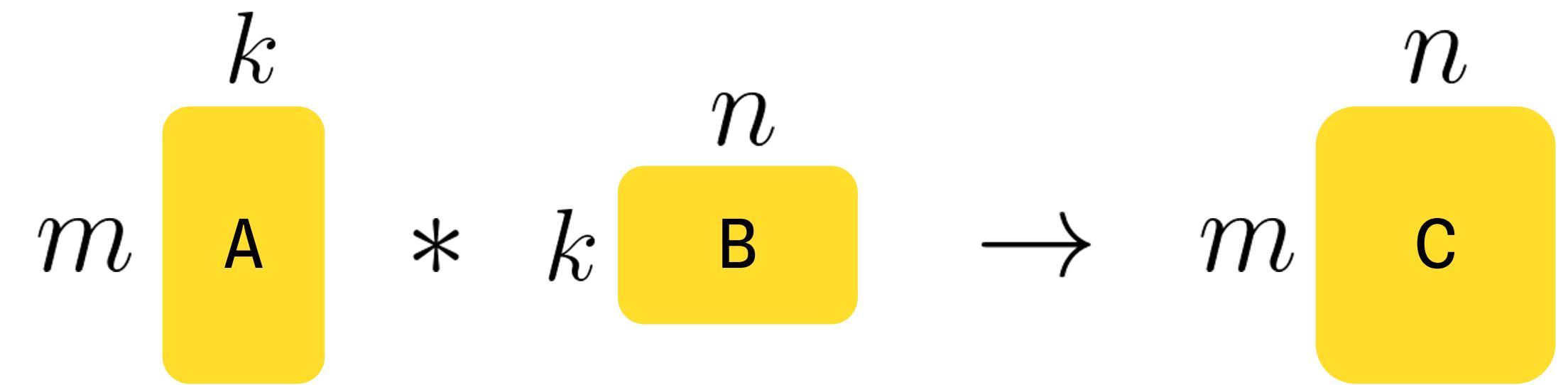


# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

 Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} & \\ & \end{pmatrix}$$

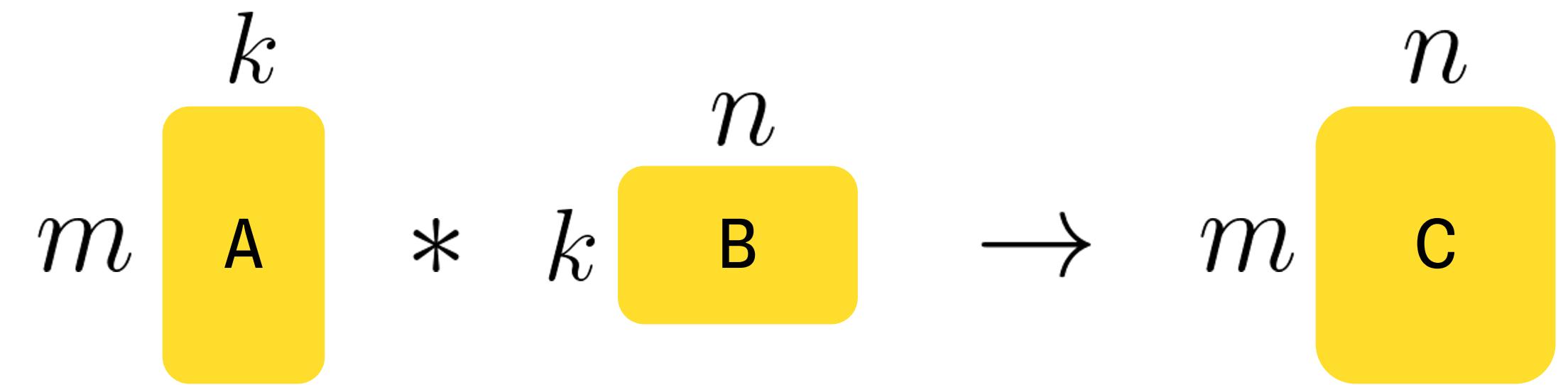


# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & ? \\ ? & ? \end{pmatrix}$$

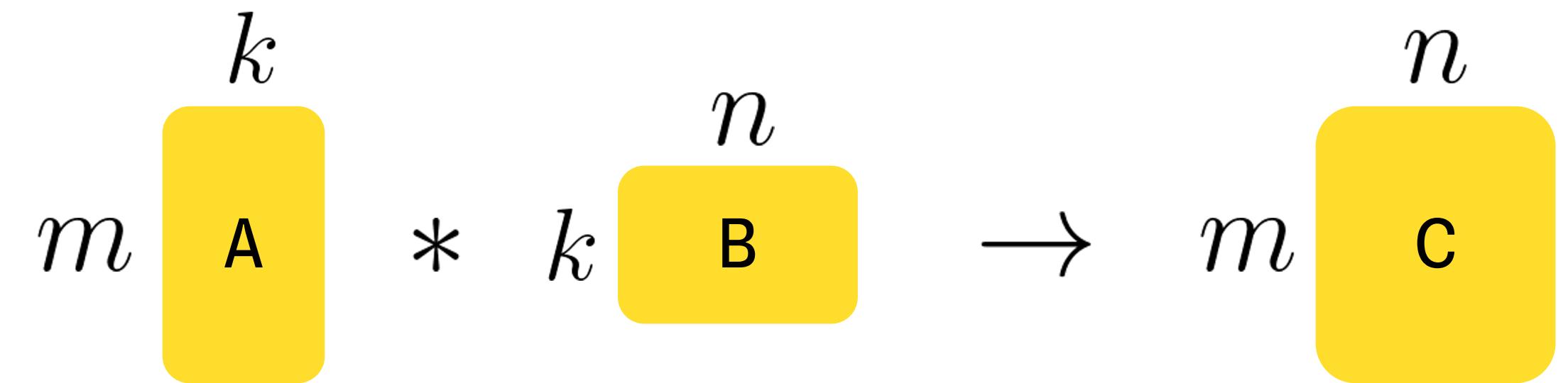


# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

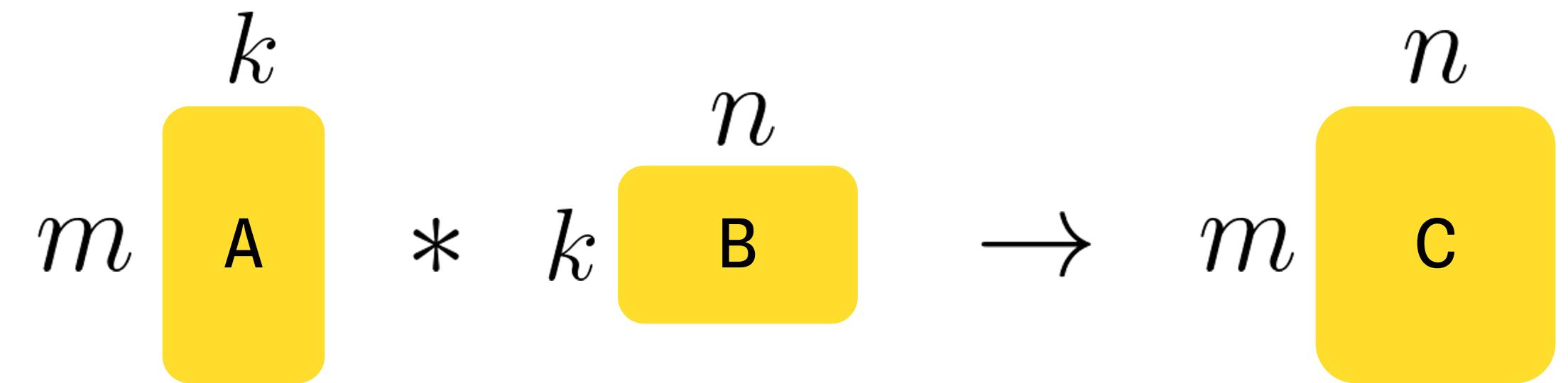


# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$



# Умножение матриц

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

Пример:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0.2 \end{pmatrix} * \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}$$

# Задание



Вычислите произведение  
матриц

$$\begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 2 & -2 \\ 1 & -1 \end{pmatrix} = \left( \begin{array}{c} ? \end{array} \right)$$

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$



Ответ в чат

(четыре числа слева направо, сверху вниз)

# Умножение матрицы на вектор

$$\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} * \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

$$C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

$$\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix}^T = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}$$

# Транспонирование

$$\begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix}^T = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix} \quad C_{i,j} = \sum_{r=1}^k A_{i,r} B_{r,j}$$

# Транспонирование

**Умножение вектора на матрицу**

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}^T * \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} = (2 \ 1) * \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} = (1 \ 4)$$

# **Вопросы**



# Вычисление градиентов сложных функций

Матричные операции

Линейная аппроксимация  
(матрица Якоби)

Градиент сложной функции

# Векторные функции

Вектор X на входе, вектор Y на выходе

$$y_i = f_i(x_1, x_2, \dots, x_n), i = \overline{1, m}$$

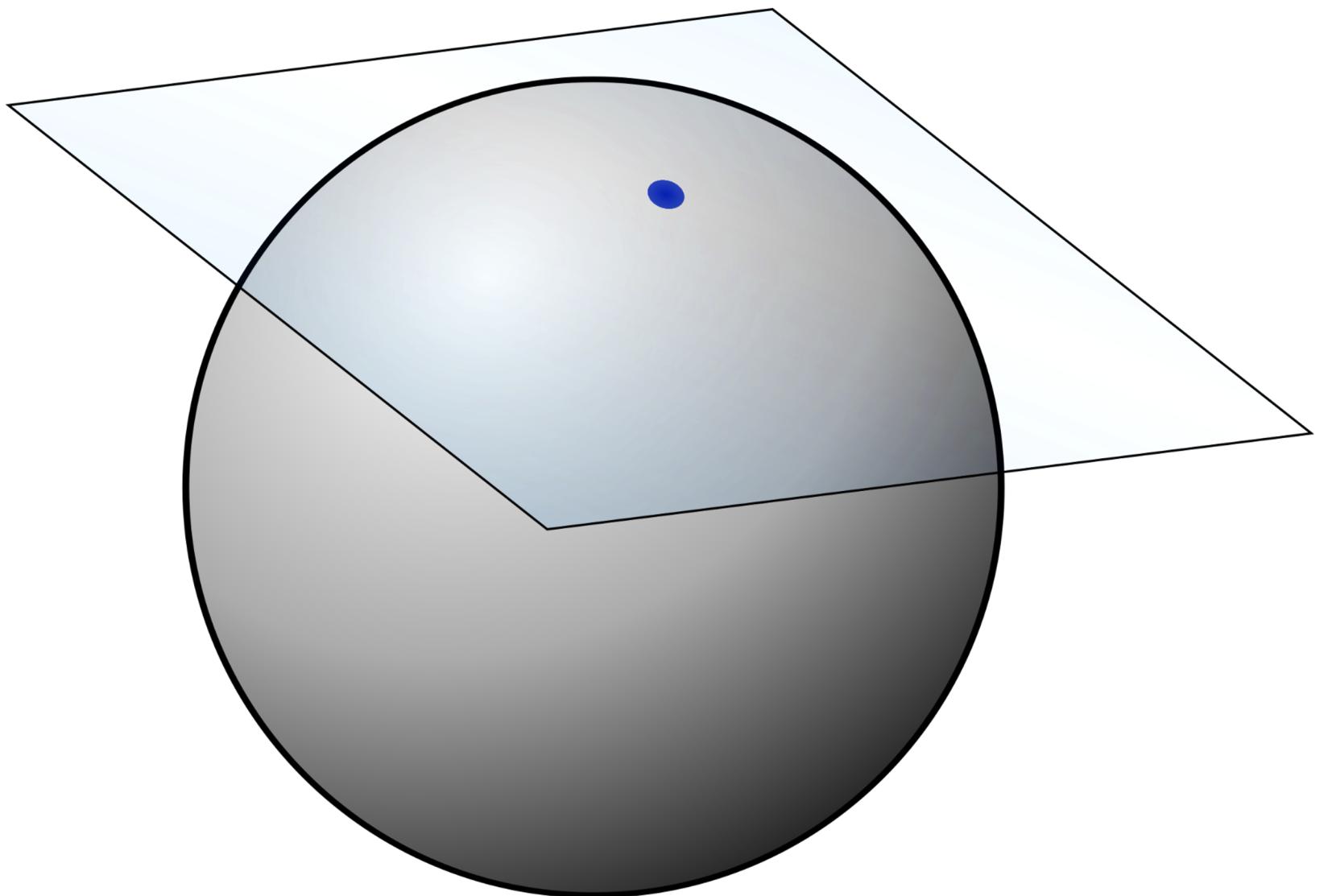
# Матрица Якоби

В окрестности точки дифференцируемая функция приближается линейной

$$y_i = f_i(x_1, \dots, x_n)$$

$$y(x + \Delta x) = y(x) + A\Delta x$$

$$A_{i,j} = (y_i)'_{x_j}$$



# Пример

$$y_1 = 5x_1^2 + 4x_2 + 1$$

$$y_2 = -x_2$$

$$A = \begin{pmatrix} 10 & 4 \\ 0 & -1 \end{pmatrix}$$

$$A_{i,j} = (y_i)'_{x_j}$$



# Вычисление градиентов сложных функций

Матричные операции

Линейная аппроксимация  
(матрица Якоби)

Градиент сложной функции

# Матрица Якоби сложной функции

$$l(x) = f(g(k(x)))$$

$$A_l = A_f A_g A_k$$

# Матрица Якоби сложной функции

$$l(x) = f(g(k(x)))$$

$$A_l = A_f A_g A_k$$

$$(A_l)_{1,i} = l'_{x_i}$$

# Пример

MSE для линейной функции

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$l(y_1, y_2) = (y_1 - 1)^2 + y_2^2$$

# Пример

MSE для линейной функции

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$l(y_1, y_2) = (y_1 - 1)^2 + y_2^2$$

$$A_y = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$$

$$A_l = (2(y_1 - 1) \ 2y_2)$$

# Пример

MSE для линейной функции

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$l(y_1, y_2) = (y_1 - 1)^2 + y_2^2$$

$$A_y = \begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix}$$

$$A_l = (2(y_1 - 1) \ 2y_2)$$

$$\nabla_x l = A_l A_y = (2(y_1 - 1), \ 4(y_1 - 1) - 2y_2)$$

# Вопрос



Чему равна матрица Якоби  
матричного умножения?

$$y = Ax$$

$$y = A * x$$

# Задание

Реализовать N-dim gradient descent



**Ссылка на ноутбук в чате  
(1-gradient-descent-1D)**



**30 минут**



**Можно шарить экран  
и задавать вопросы  
(Zoom комнаты?)**



**ТИНЬКОФФ**

# **Расширения градиентного спуска**

# Подбор параметров функции



**Было**

- Дано функция
- Нужно подобрать значение аргумента, минимизирующую функцию

$$\min_x f(x)$$



Стало

# Подбор параметров функции

- Данна функция, зависящая от параметров
- Дан набор данных (вход, выход)
- Нужно подобрать параметры функции, минимизирующие ошибку

$$\min_{\theta} \hat{\mathcal{L}}(\theta) = \min_{\theta} \sum_{x,y} \mathcal{L}(f(x, \theta), y)$$

$f(x, \theta)$  – параметрическая функция, например  $x_1\theta_1 + x_2\theta_2$

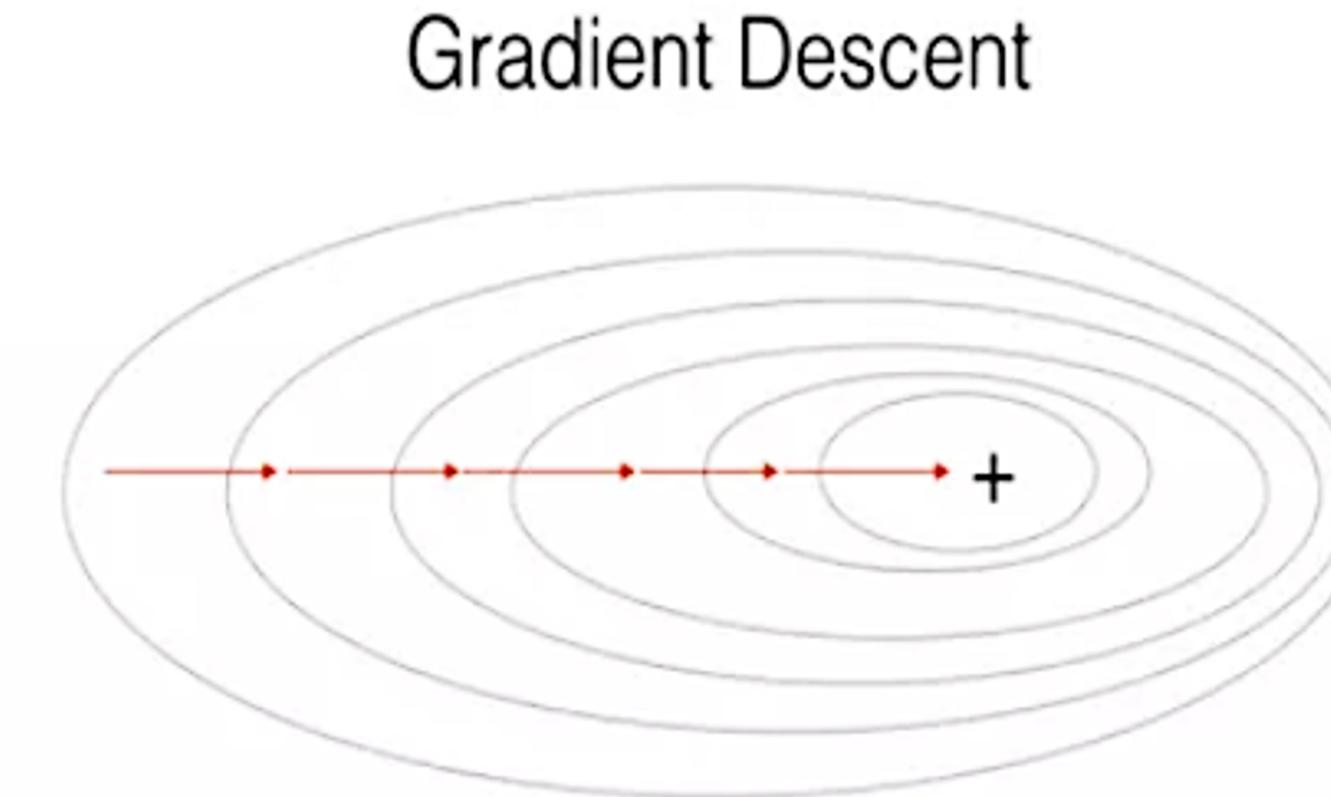
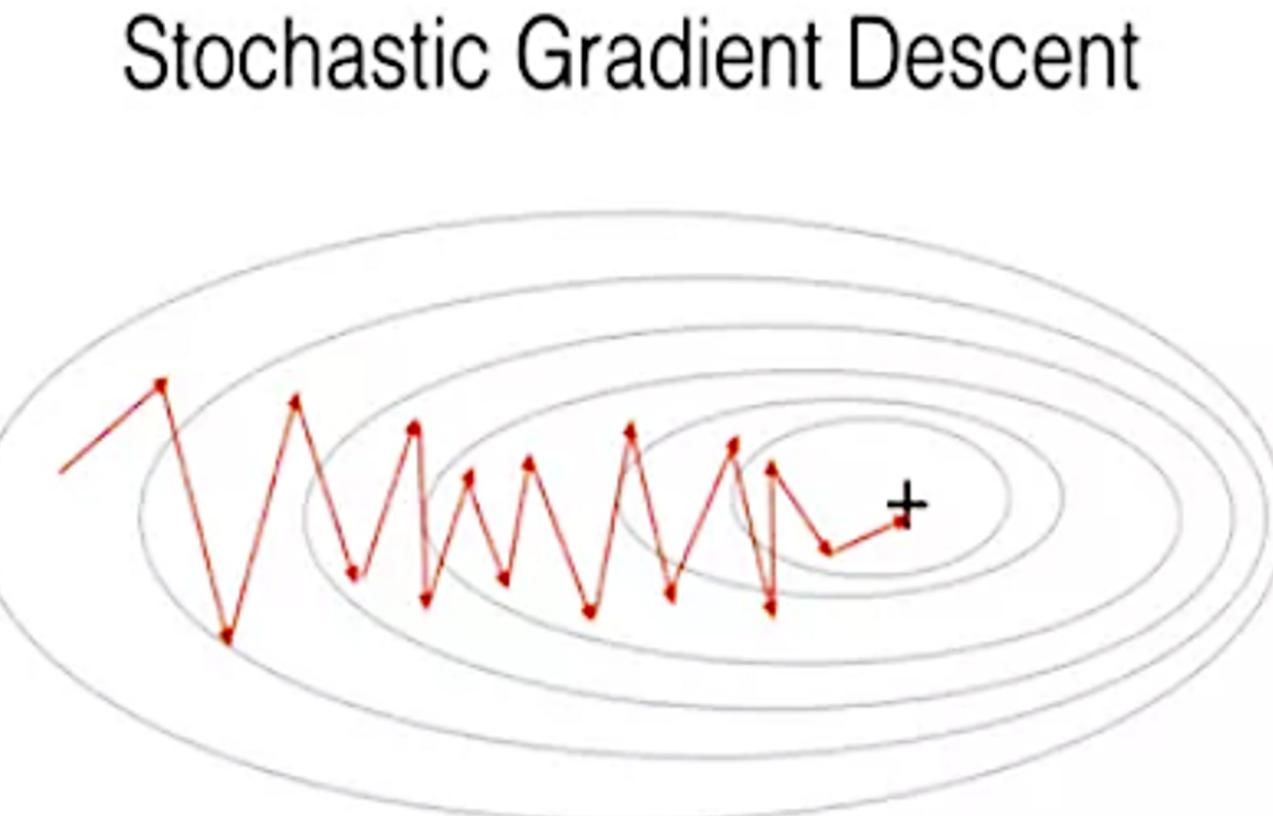
$\theta$  – вектор параметров модели, например  $(0, 1)$

$\mathcal{L}(y^*, y)$  – функция ошибки, например  $(y_1^* - y_1)^2 + (y_2^* - y_2)^2$

# Stochastic gradient descent

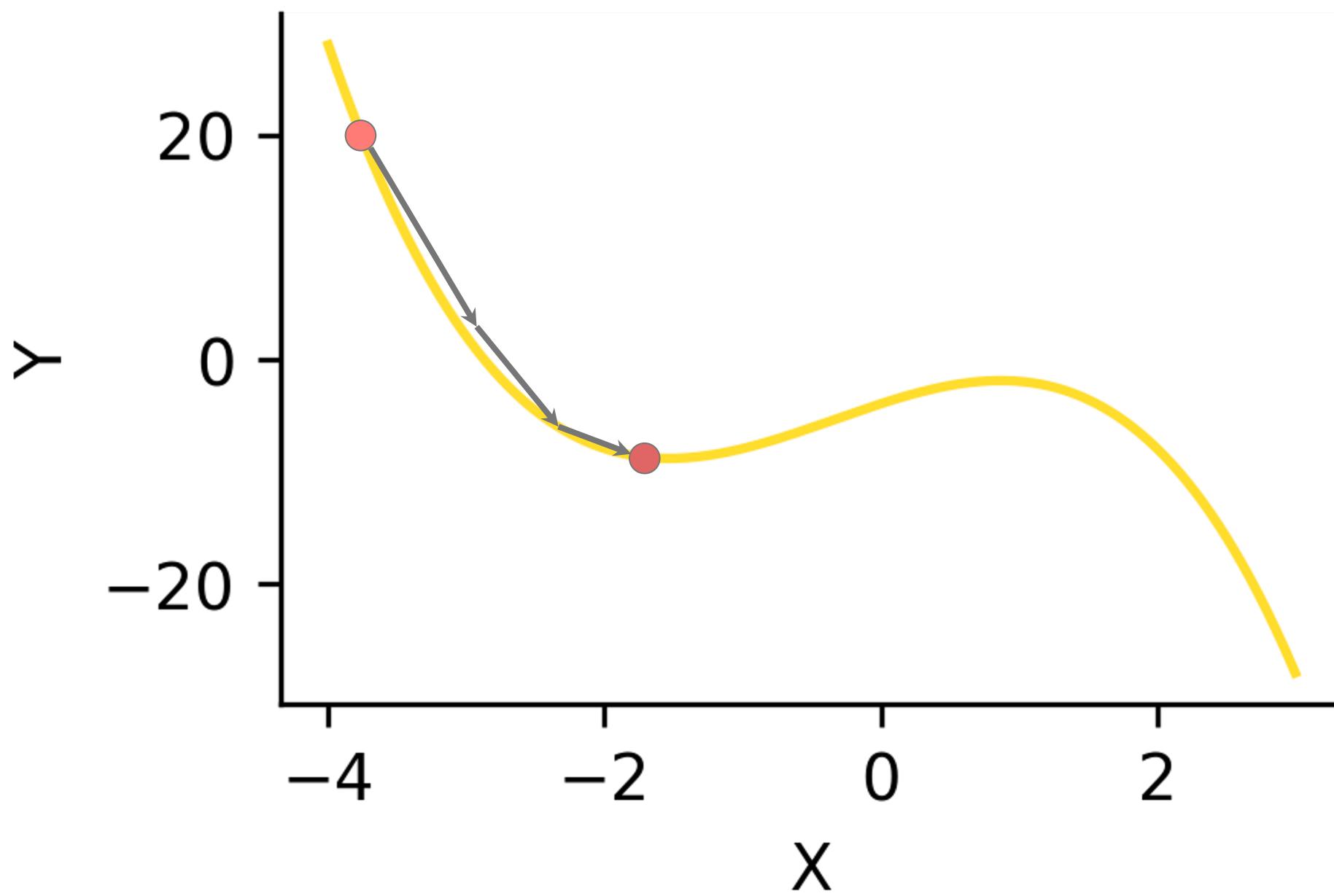
Набор данных  $(x_i, y_i)$  может быть очень большим

- Можно оценивать градиент по небольшому подмножеству  $(x_i, y_i)$
- Такое подмножество называется mini-batch



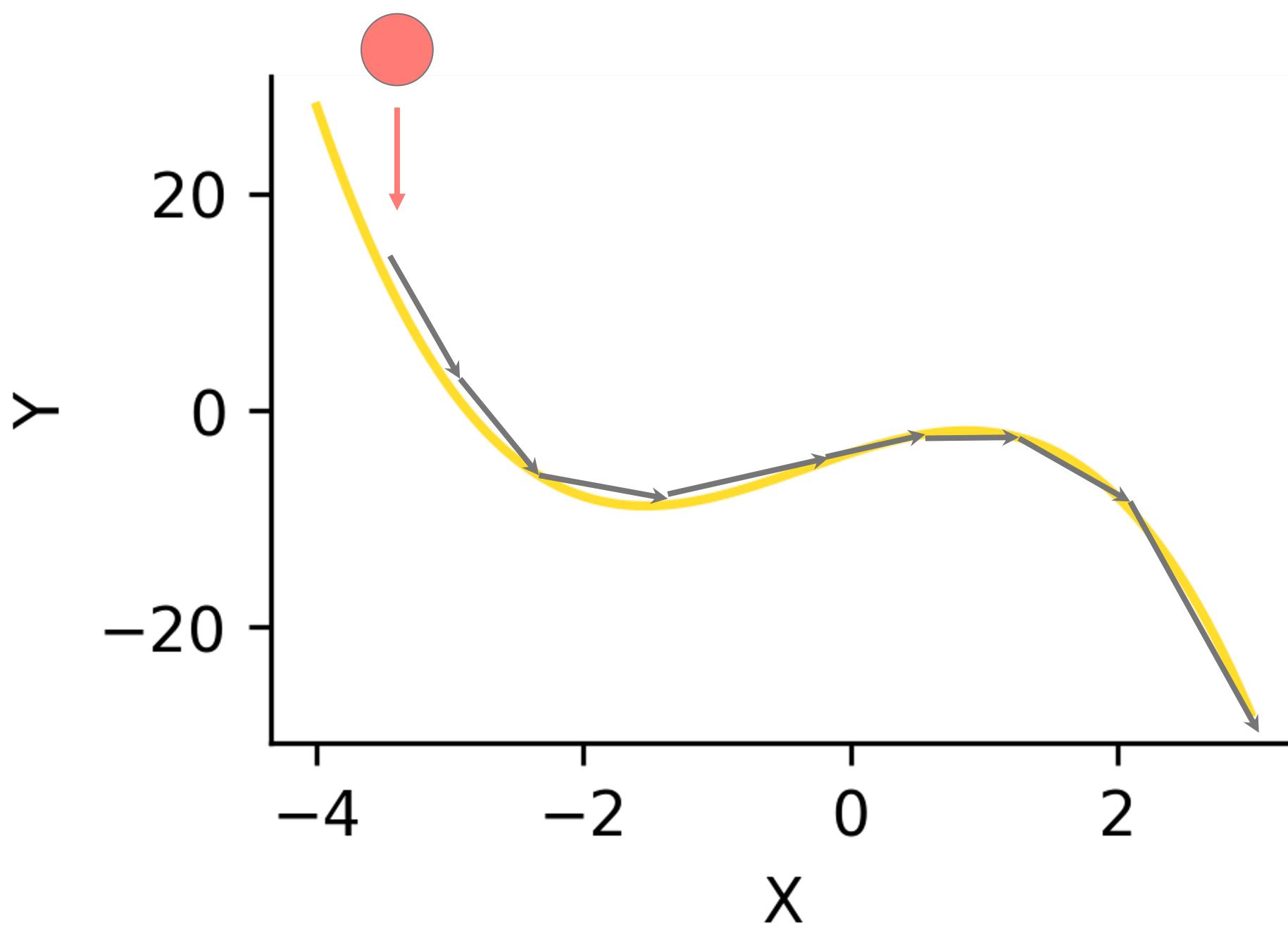
# Momentum

Мотивация



# Momentum

Мотивация



# Momentum



**Обычный GD:**

$$z = \nabla f(\theta)$$

$$\theta_{n+1} = \theta_n - \lambda z$$



**С моментом:**

$$z_{n+1} = \beta z_n + \nabla f(\theta)$$

$$\theta_{n+1} = \theta_n - \lambda z_{n+1}$$



**ТИНЬКОФФ**

# Резюме

# Что мы изучили



Принципы оптимизации  
первого порядка



Реализация метода  
градиентного спуска



Подбор оптимальных  
параметров функции



**ТИНЬКОФФ**

**Q & A**

# Домашнее задание

Подобрать методом GD параметры функции

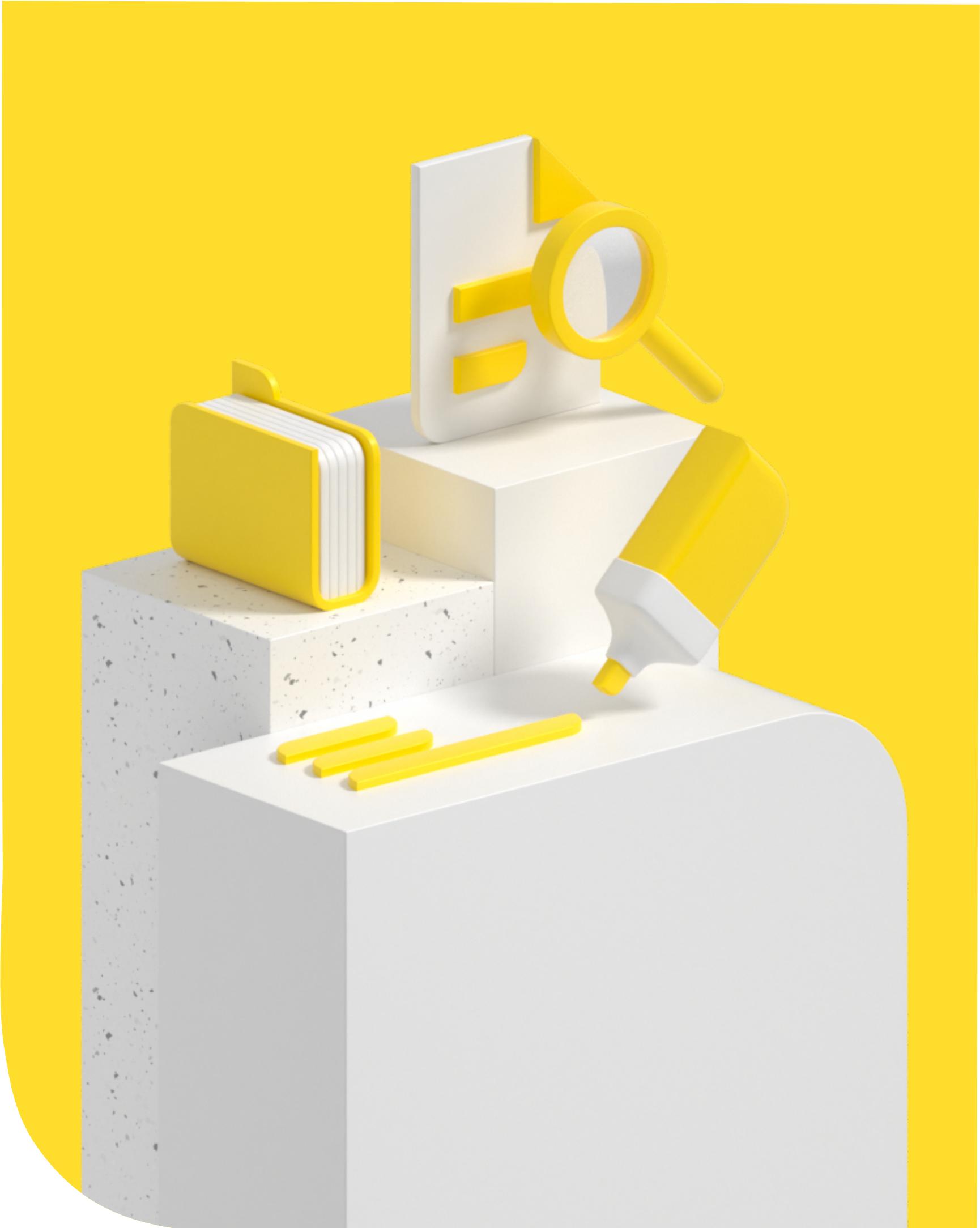
$$f(x, \theta) = x_1\theta_1 + x_2\theta_2 + \theta_3 \quad \theta_{start} = (1, 1, 0)$$

с функцией ошибки

$$\mathcal{L}(\theta) = 0.1 \|\theta\|^2 + \frac{1}{N} \sum_{i=1}^N \max(0, 1 - y_i f(x_i, \theta))$$

Для набора данных

X1	0	1	1	-0.5	0
X2	1	1	0	0.5	-0.5
Y	1	1	1	-1	-1



**Спасибо ;)**



**ТИНЬКОФФ**

**Он такой один**