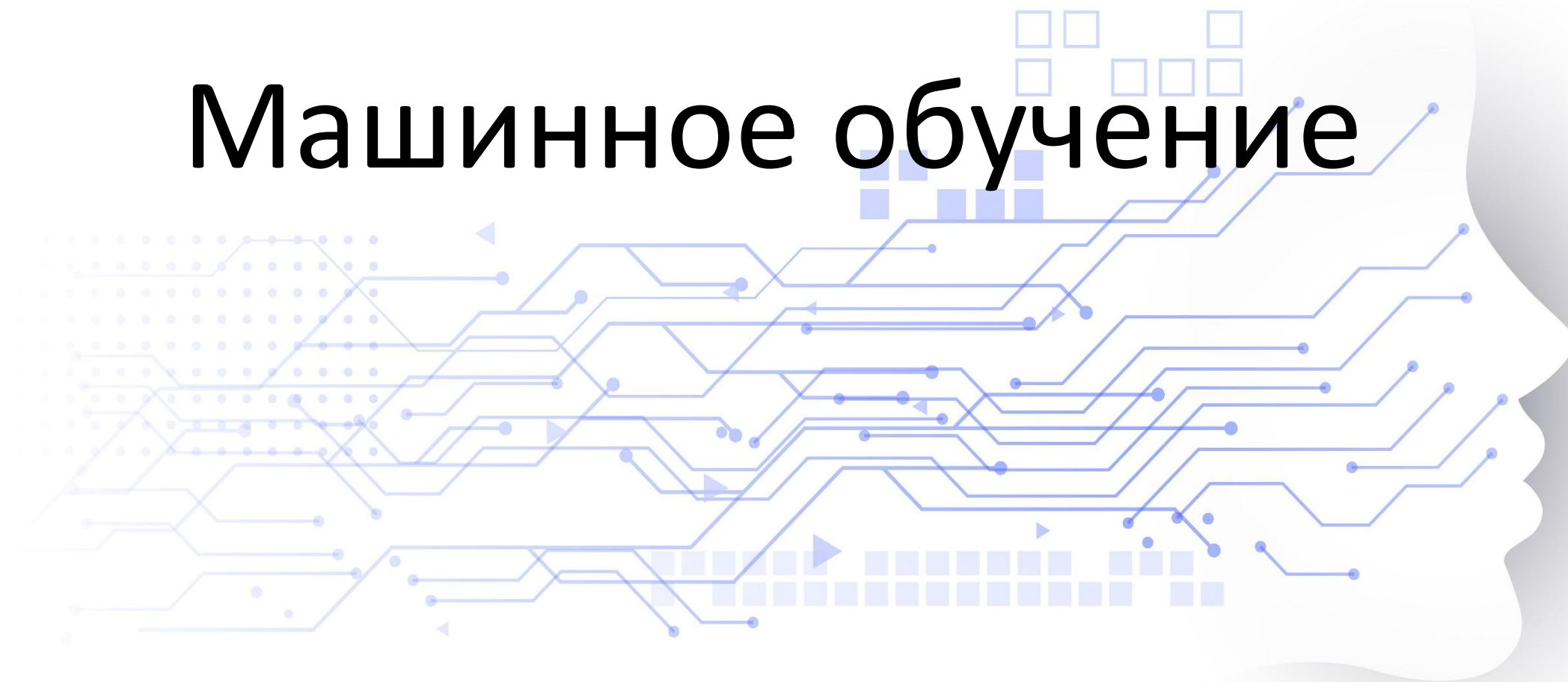


# Машинное обучение



## Общие понятия градиентного спуска

# Градиентный спуск (Gradient descent)

Модель:

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_d x_d = \langle \theta, x \rangle$$

Функционал ошибки

$$Q(a, \mathbb{X}) = \frac{1}{\ell} \|\mathbf{X}\theta - \mathbf{y}\|^2 \rightarrow \min_{\theta}$$

$$\theta = (X^T X)^{-1} X^T y$$

# Минимизация функции

Задачи минимизации можно решать тремя способами:

*1- подбором параметров вручную*

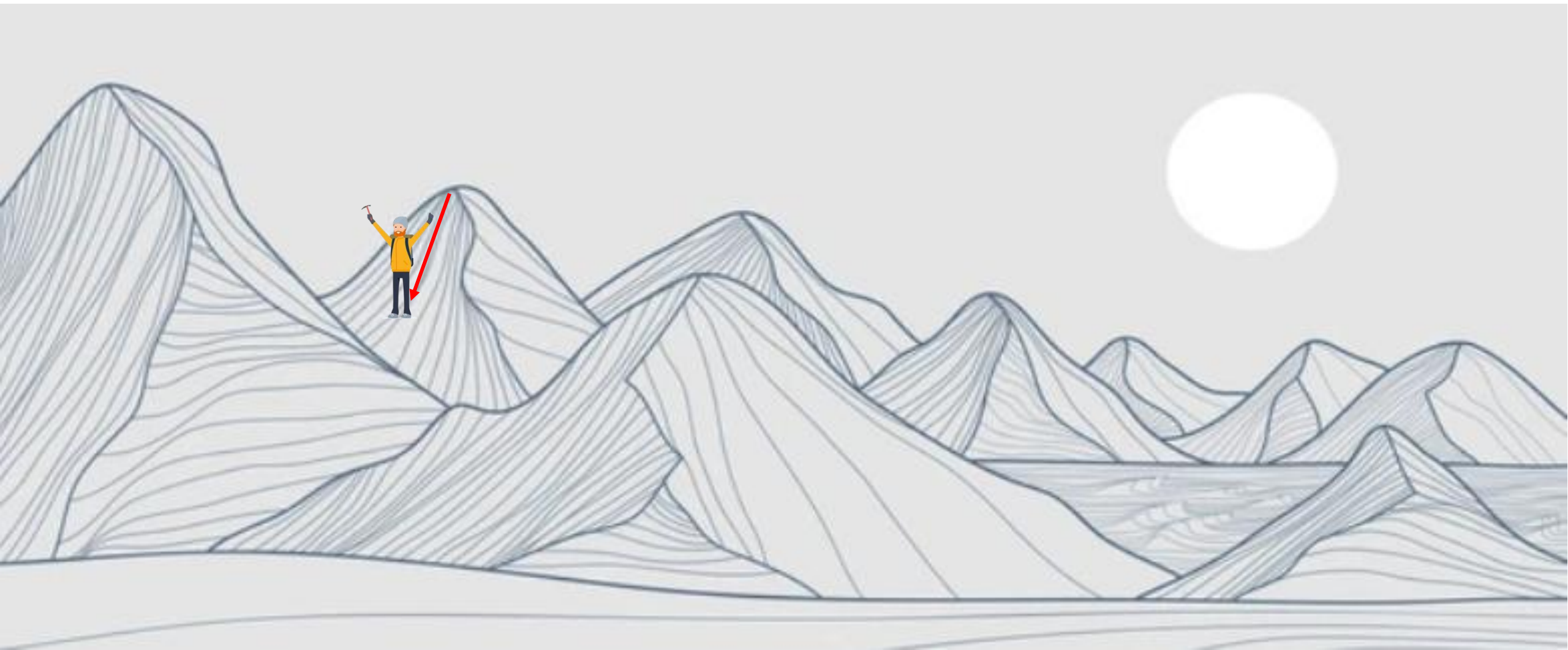
*2- аналитически*

*3- численными методами*

## Градиентный спуск (Gradient descent)



## Градиентный спуск (Gradient descent)

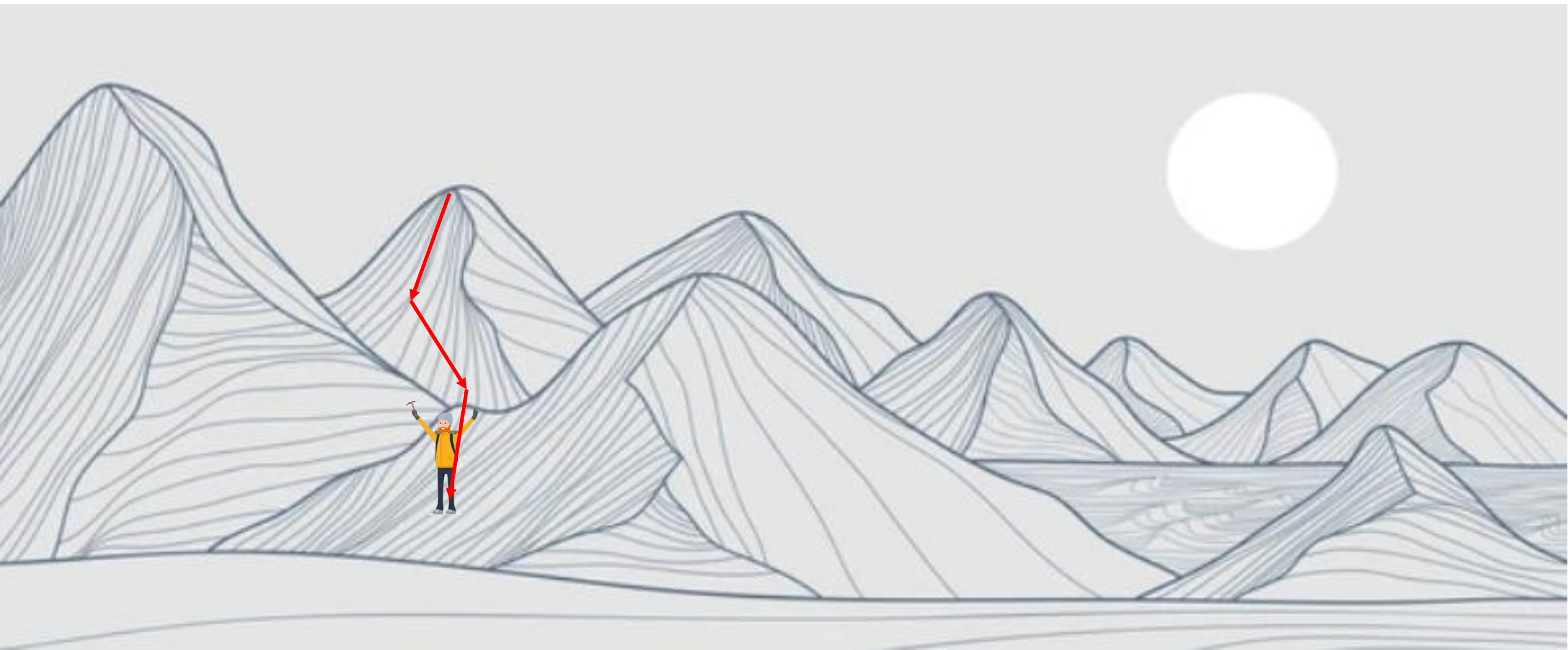




## Градиентный спуск (Gradient descent)

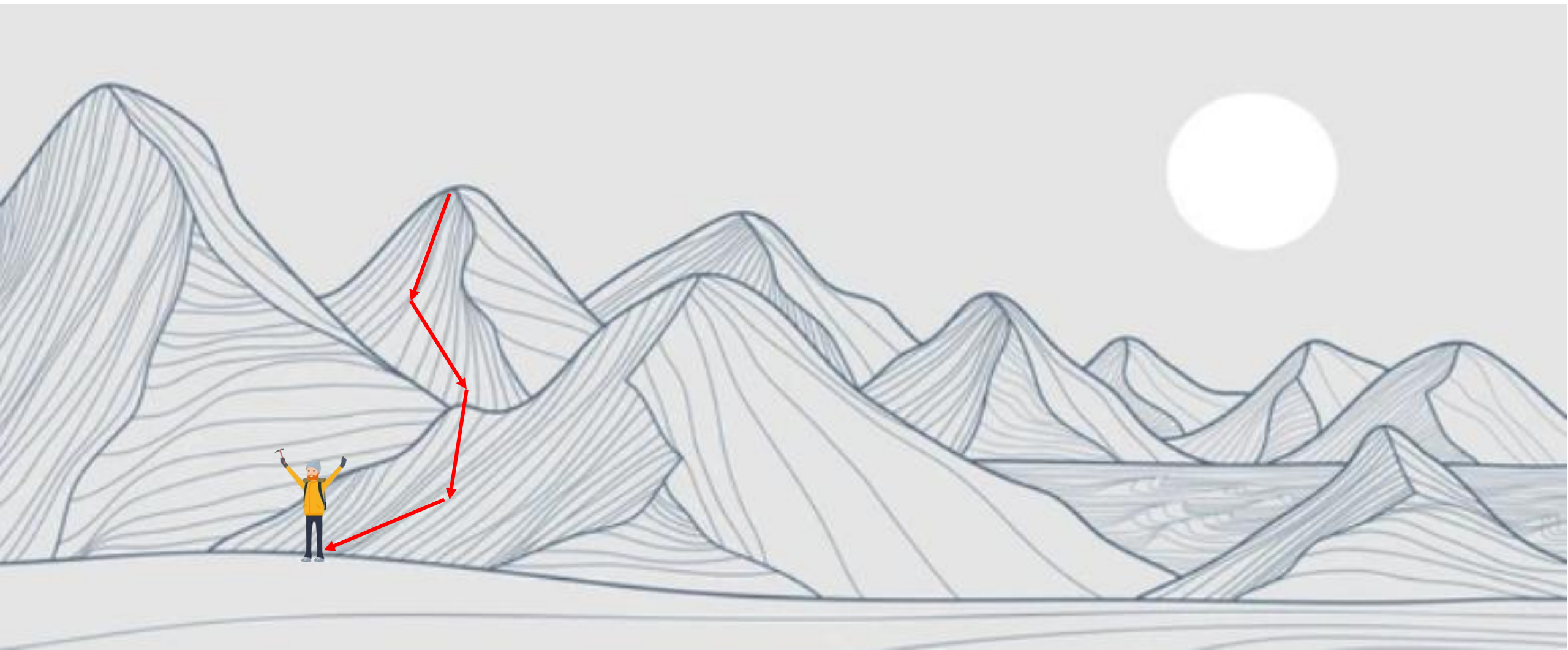


## Градиентный спуск (Gradient descent)

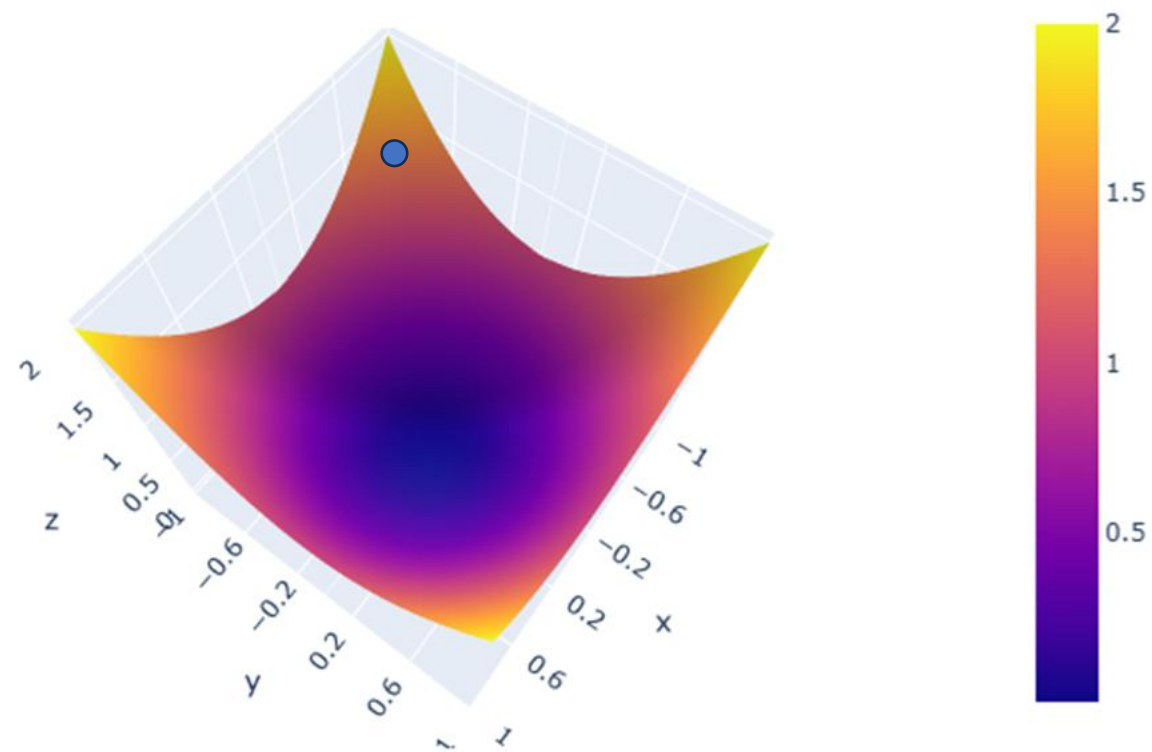
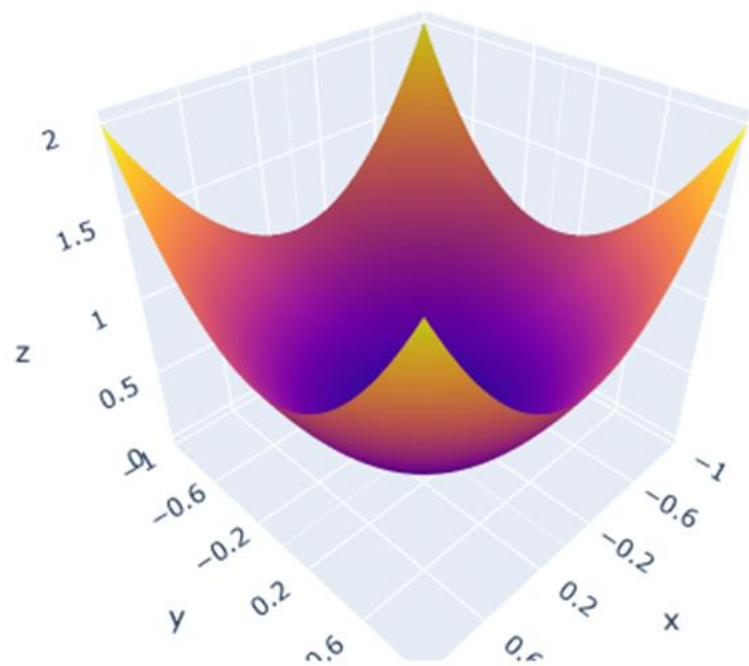




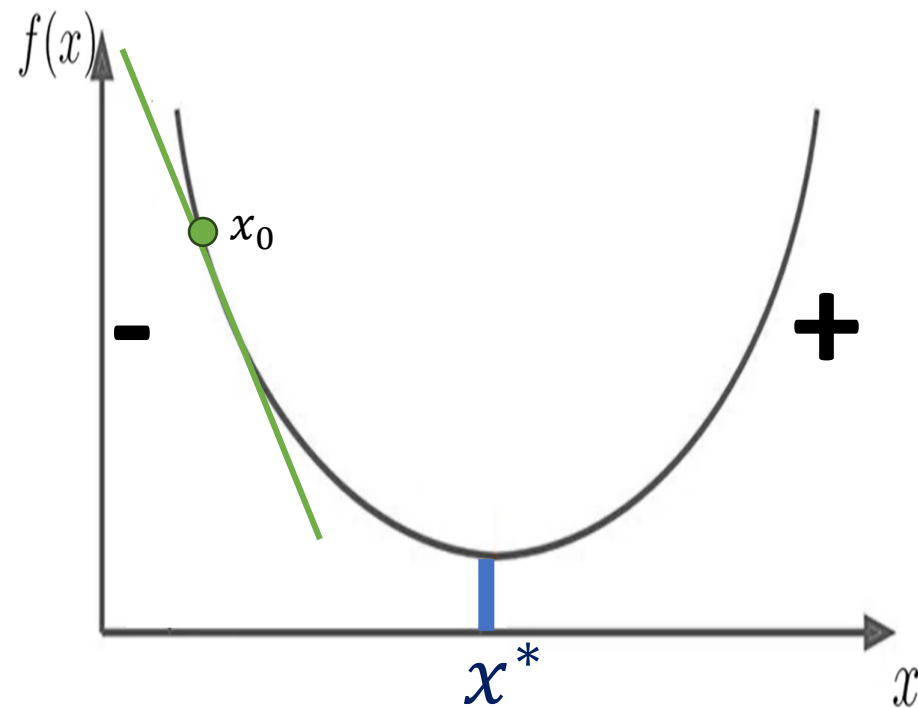
## Градиентный спуск (Gradient descent)



## Градиентный спуск (Gradient descent)

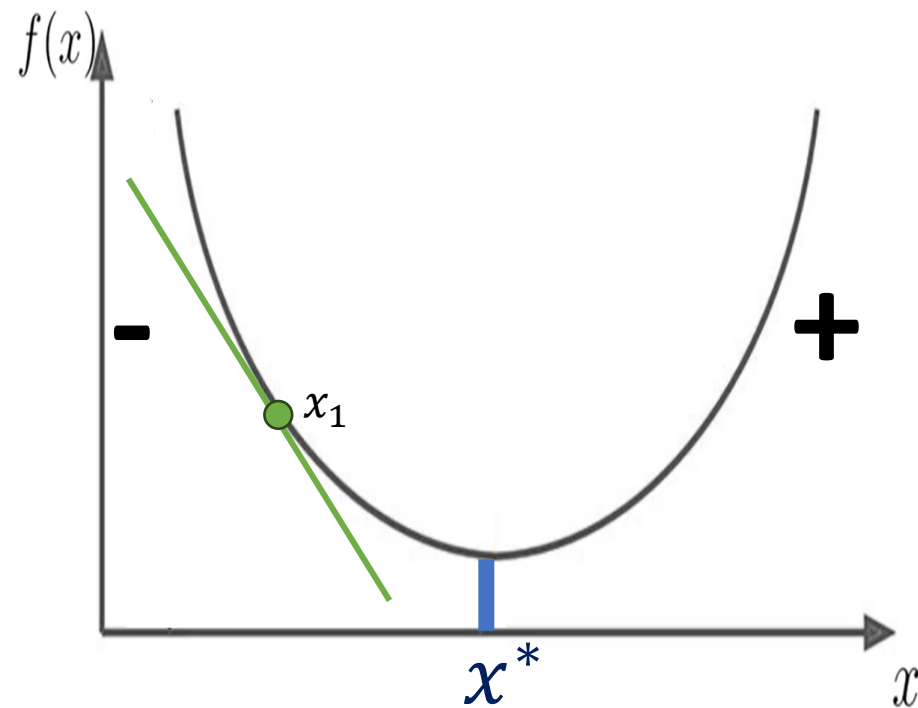


## Эвристика градиентного спуска



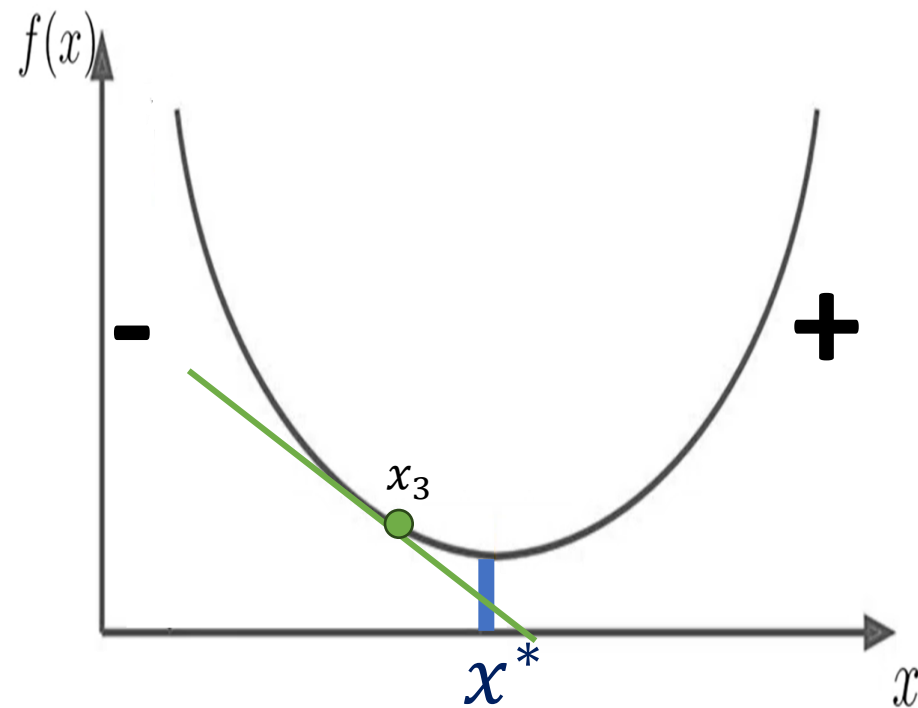
$$x_{n+1} = x_n - \alpha \frac{\partial f(x)}{\partial x}, n = 0, 1, 2, 3, \dots$$

## Эвристика градиентного спуска



$$x_{n+1} = x_n - \alpha \frac{\partial f(x)}{\partial x}, n = 0, 1, 2, 3, \dots$$

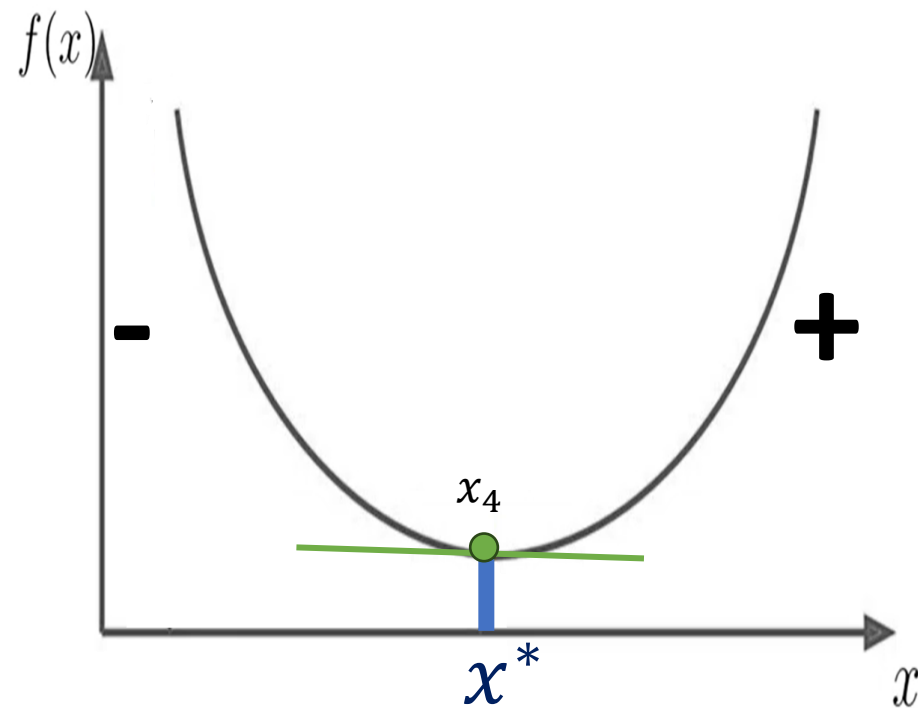
## Эвристика градиентного спуска



$$x_{n+1} = x_n - \alpha \frac{\partial f(x)}{\partial x}, n = 0, 1, 2, 3, \dots$$



## Эвристика градиентного спуска



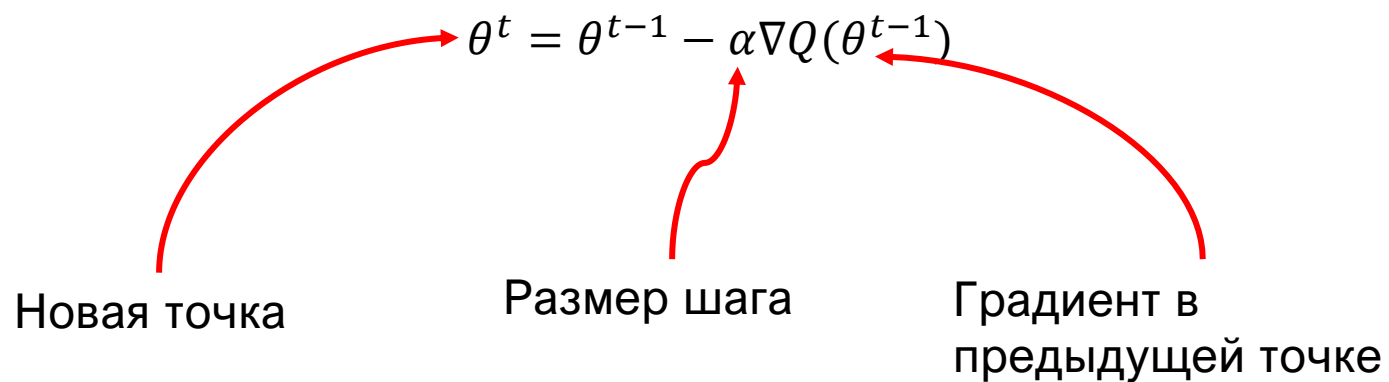
$$x_{n+1} = x_n - \alpha \frac{\partial f(x)}{\partial x}, n = 0, 1, 2, 3, \dots$$

# Градиентный спуск (Gradient descent)

Стартуем из случайной точки

Сдвигаемся по антиградиенту

Повторяем, пока не окажемся в точке минимума



The diagram illustrates the gradient descent update formula:  $\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$ . Three red curved arrows point from labels below to components of the formula: one from 'Новая точка' to  $\theta^t$ , one from 'Размер шага' to  $\alpha$ , and one from 'Градиент в предыдущей точке' to  $\nabla Q(\theta^{t-1})$ .

Новая точка

Размер шага

Градиент в предыдущей точке

$$\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$$

Применение градиентного спуска в линейных задачах

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$



# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

$$\nabla Q(\theta) = ?$$

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Градиент

$$\frac{\partial Q}{\partial \theta_0} = ?$$

$$\frac{\partial Q}{\partial \theta_1} = ?$$

# Цепное правило в математике

это способ посчитать производную сложной функции, которая состоит из нескольких вложенных функций.

Формула:

$$\frac{\partial}{\partial x} f(g(x)) = f'(g(x)) \cdot g'(x)$$

Функция стоимости  $Q$  не зависит напрямую от параметров  $\theta_0, \theta_1$

$$\theta_0, \theta_1 \rightarrow a_{\theta}(x) = \theta_0 + \theta_1 x \quad a_{\theta}(x) \rightarrow u = a_{\theta}(x) - y \quad u \rightarrow Q = \frac{1}{\ell} \sum u^2$$

Это **цепочка функций**. Поэтому, чтобы найти  $\frac{\partial J}{\partial \theta_0}$ , мы применяем цепное правило:

- сначала берём производную  $Q$  по  $u$ ,
- потом  $u$  по  $a_{\theta}$
- потом  $a_{\theta}$  по  $\theta_0$

$$\frac{\partial Q}{\partial \theta_0} = \frac{\partial Q}{\partial u} \cdot \frac{\partial u}{\partial a_{\theta}} \cdot \frac{\partial a_{\theta}}{\partial \theta_0}$$

$$\begin{cases} Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^n (\mathbf{a}_{\theta}(x^{(i)}) - y^{(i)})^2 \\ \mathbf{a}_{\theta}(x) = \theta_0 + \theta_1 x_i \end{cases} \rightarrow \min_{\theta_0, \theta_1} Q$$

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^n \mathbf{u}^2 \rightarrow \begin{cases} \frac{\partial J}{\partial \theta_0} = ? \\ \frac{\partial J}{\partial \theta_1} = ? \end{cases}$$

$$\frac{\partial Q}{\partial u} = \frac{1}{\ell} \sum_{i=1}^n 2\mathbf{u} = \frac{2}{\ell} \sum_{i=1}^n u = \frac{2}{\ell} \sum_{i=1}^n (a_{\theta}(x^{(i)}) - y^{(i)})$$

$$\frac{\partial Q}{\partial \theta_0} = \frac{\partial Q}{\partial u} \cdot \frac{\partial u}{\partial a_{\theta}} \cdot \frac{\partial a_{\theta}}{\partial \theta_0} = \frac{2}{\ell} \sum_{i=1}^n (a_{\theta}(x^{(i)}) - y^{(i)}) \cdot 1 \cdot 1$$

$$\frac{\partial u}{\partial a_{\theta}} = \frac{\partial a_{\theta}(x^{(i)}) - y^{(i)}}{\partial a_{\theta}} = 1$$

$$\frac{\partial a_{\theta}}{\partial \theta_0} = \frac{\partial \theta_0 + \theta_1 x_i}{\partial \theta_0} = 1$$

$$\frac{\partial a_{\theta}}{\partial \theta_1} = \frac{\partial \theta_0 + \theta_1 x_i}{\partial \theta_1} = x_i$$



$$\frac{\partial Q}{\partial \theta_1} = \frac{\partial Q}{\partial u} \cdot \frac{\partial u}{\partial a_{\theta}} \cdot \frac{\partial a_{\theta}}{\partial \theta_1} = \frac{2}{\ell} \sum_{i=1}^n (a_{\theta}(x^{(i)}) - y^{(i)}) \cdot 1 \cdot x_i$$



# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Градиент

$$\frac{\partial Q}{\partial \theta_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (\theta_1 x_1 + \theta_0 - y_i)$$

$$\frac{\partial Q}{\partial \theta_1} = ?$$

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Градиент

$$\frac{\partial Q}{\partial \theta_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (\theta_1 x_1 + \theta_0 - y_i)$$

$$\frac{\partial Q}{\partial \theta_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (\theta_1 x_1 + \theta_0 - y_i)$$

# Парная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta_0, \theta_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Градиент

$$\frac{\partial Q}{\partial \theta_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (\theta_1 x_1 + \theta_0 - y_i)$$

$$\frac{\partial Q}{\partial \theta_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (\theta_1 x_1 + \theta_0 - y_i)$$

$$\nabla Q(\theta) = \left( \frac{2}{\ell} \sum_{i=1}^{\ell} (\theta_1 x_1 + \theta_0 - y_i), \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (\theta_1 x_1 + \theta_0 - y_i) \right)$$

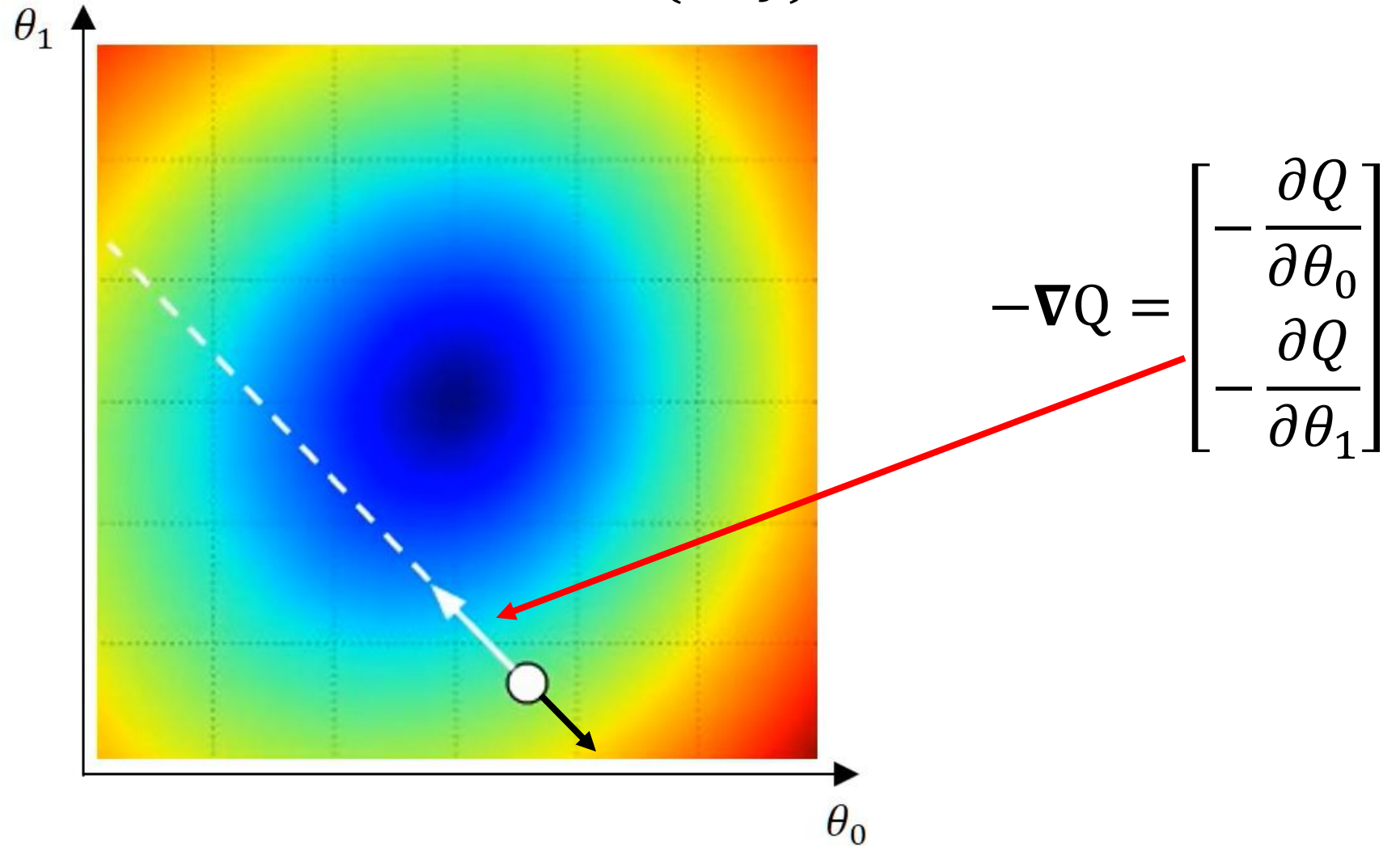
# Градиентный спуск (Gradient descent)

$$Q(\theta_0, \theta_1) = J(\theta)$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\nabla Q = \begin{bmatrix} \frac{\partial Q}{\partial \theta_0} \\ \frac{\partial Q}{\partial \theta_1} \end{bmatrix}$$

$$\theta^{(new)} = \theta^{(old)} + \alpha(-\nabla J)$$



# Градиентный спуск (Gradient descent)

1. Стартуем из случайной точки

$\theta$  — Инициализация весов

2. Сдвигаемся по антиградиенту

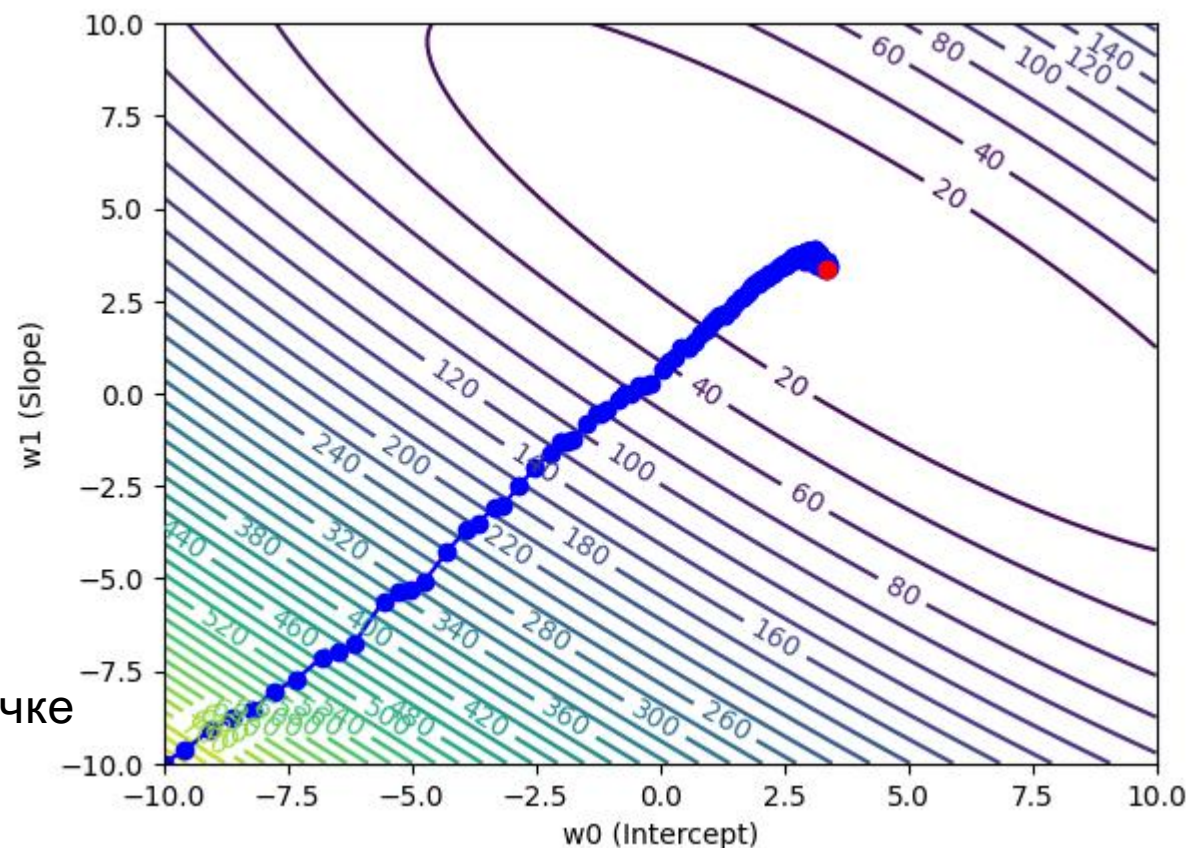
$$\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$$

Новая точка

Размер шага

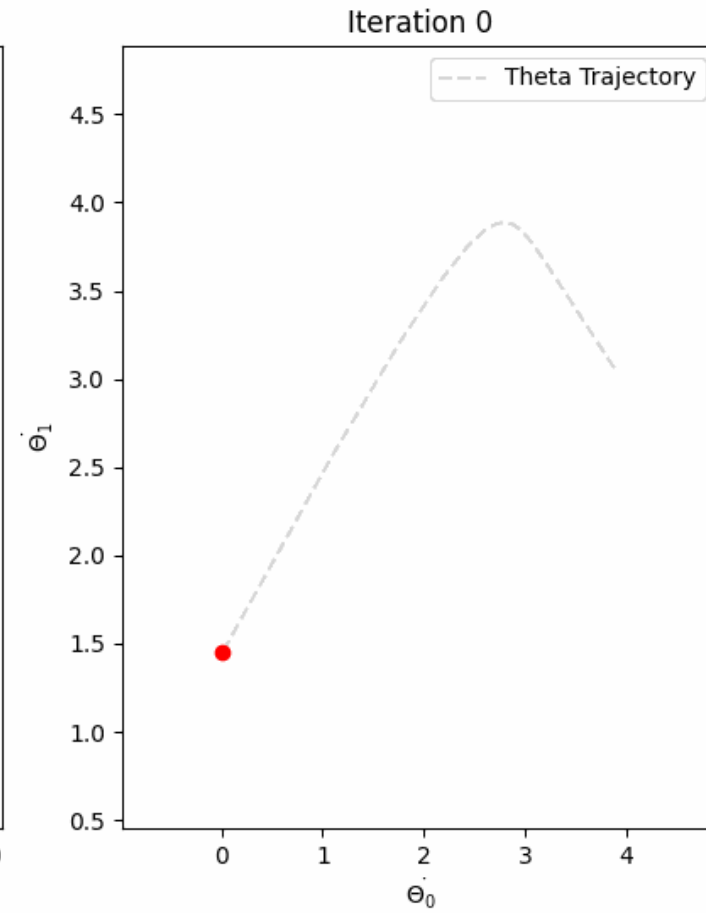
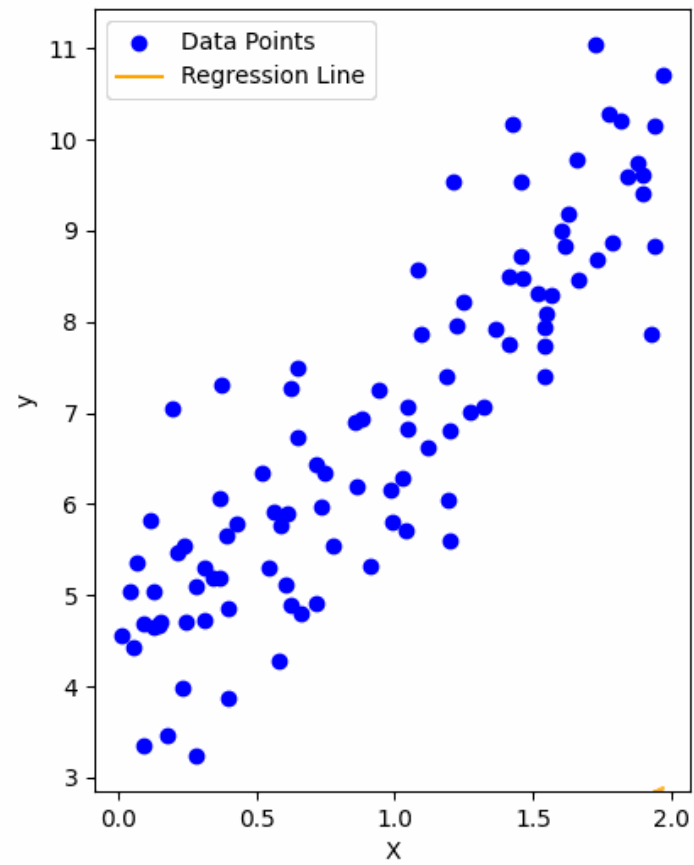
Градиент в  
предыдущей точке

3. Повторяем до сходимости





# Градиентный спуск



## Градиентный спуск для поиска минимума функции многих переменных

$$Q(\theta_1, \theta_1, \dots, \theta_d) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Чтобы применять метод градиентного спуска, необходимо вычислять градиент функции в точке:

$$\nabla f(\theta_1, \theta_1, \dots, \theta_d) = \left( \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}, \dots, \frac{\partial f}{\partial \theta_d} \right)$$

На каждом шаге будем менять все переменные, от которых зависит функция:

$$\begin{aligned} \theta_1 &= \theta_1 - \alpha \frac{\partial f}{\partial \theta_1} \\ &\dots \\ \theta_d &= \theta_d - \alpha \frac{\partial f}{\partial \theta_n} \end{aligned}$$

Повторяем пока изменение не будет достаточно маленьким, пройдет много итераций или до сходимости

# Градиентный спуск в векторном виде

Модель:

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_d x_d = \langle \theta, x \rangle$$

# Градиентный спуск в векторном виде

Модель:

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_d x_d = \langle \theta, x \rangle$$

Функционал ошибки

$$Q(\theta) = \frac{1}{\ell} \|X\theta - y\|^2$$

Градиент

$$\nabla_{\theta} Q(\theta) = \frac{2}{\ell} X^T (X\theta - y)$$

Правило обновления

$$\theta: \theta - \alpha \cdot \nabla_{\theta} Q(\theta)$$

# Градиентный спуск в векторном виде

Градиент

$$\nabla Q(\theta) = \frac{2}{\ell} X^T (X\theta - y)$$

1.  $\theta$  — Инициализация весов

2.  $\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$

3. Повторяем до сходимости

# Критерии останова

1. Количество итераций

2. Останавливаем процесс, если

$$\|\theta^t - \theta^{t-1}\| < \varepsilon$$

3. Другой вариант

$$\|\nabla Q(\theta^t)\| < \varepsilon$$

4. Другой вариант

Размер шага или скорость обучения

# Градиентный шаг (learning rate)

Модель:

$$a(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_d x_d = \langle \theta, x \rangle$$

Функционал ошибки

$$Q(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \theta, x \rangle - y_i)^2$$

Парная регрессия

Градиент

$$\nabla Q(\theta) = \frac{2}{\ell} X^T (X\theta - y)$$

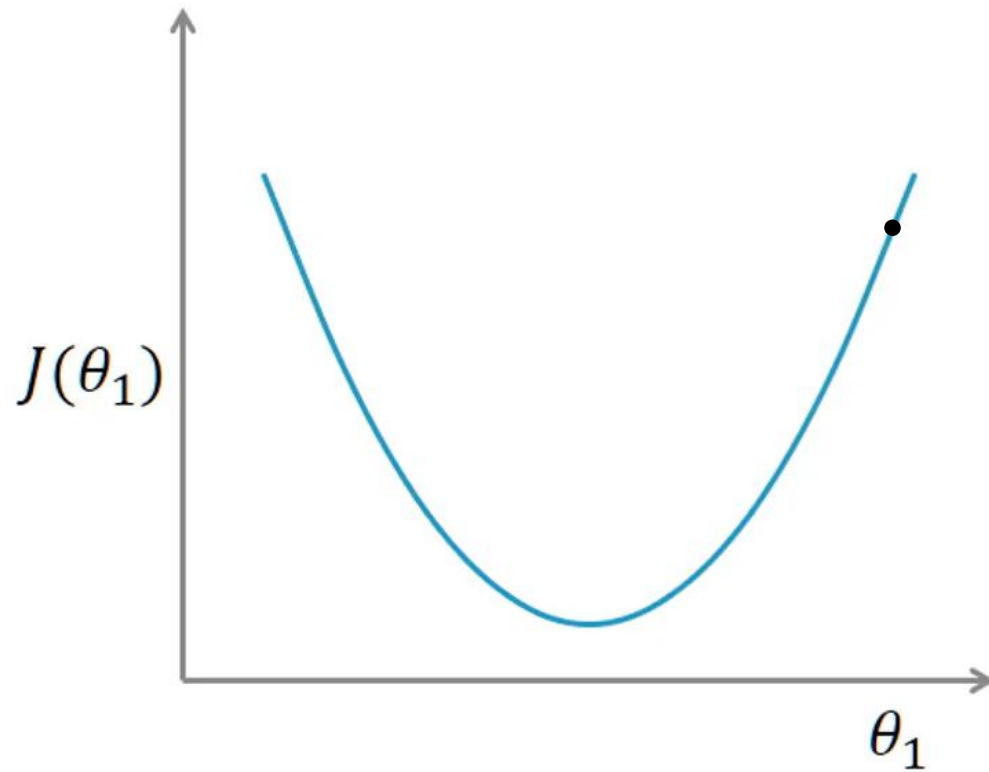
$$\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$$

Размер шага



# Размер шага (learning rate)

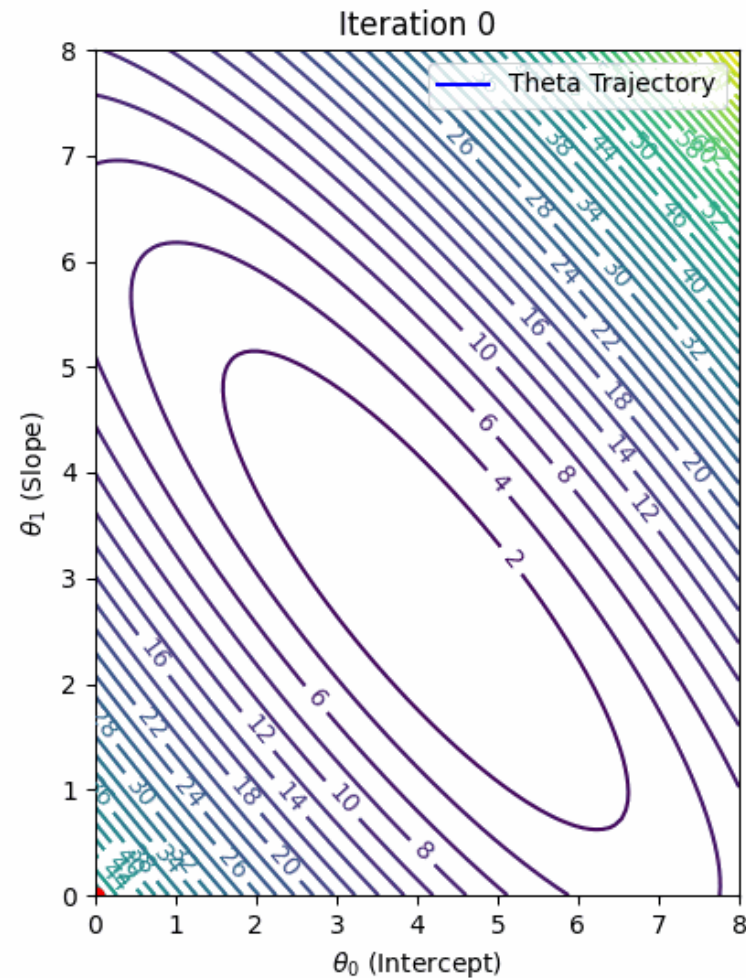
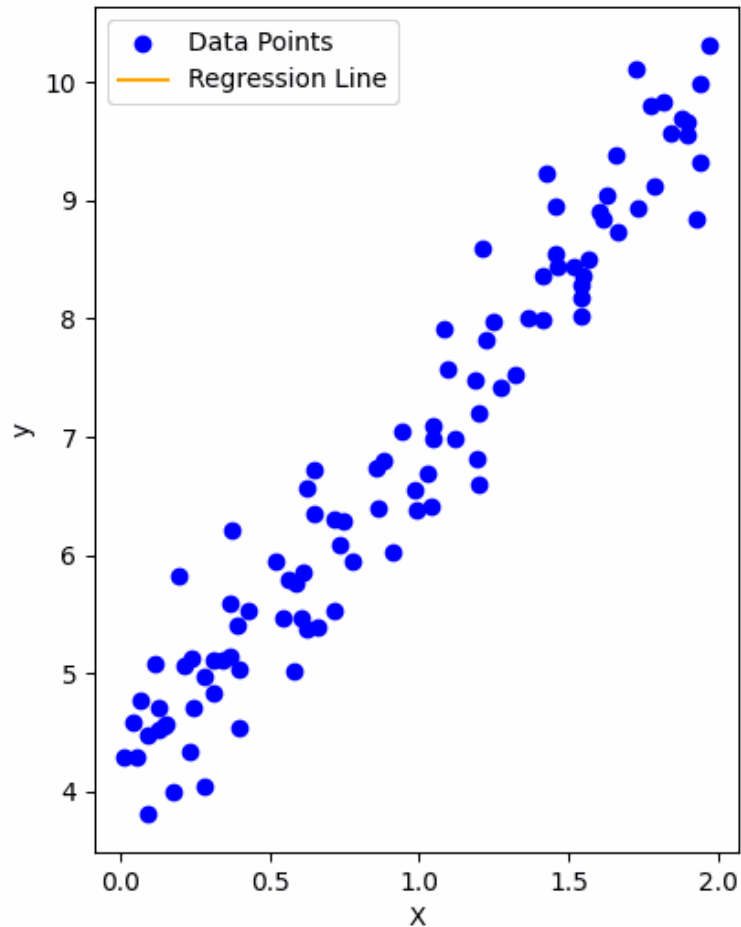
Если размер шага (learning rate) слишком мал, градиентный спуск сходится слишком медленно или не достигает до точки минимума



$$\theta^t = \theta^{t-1} - \alpha \nabla Q(\theta^{t-1})$$

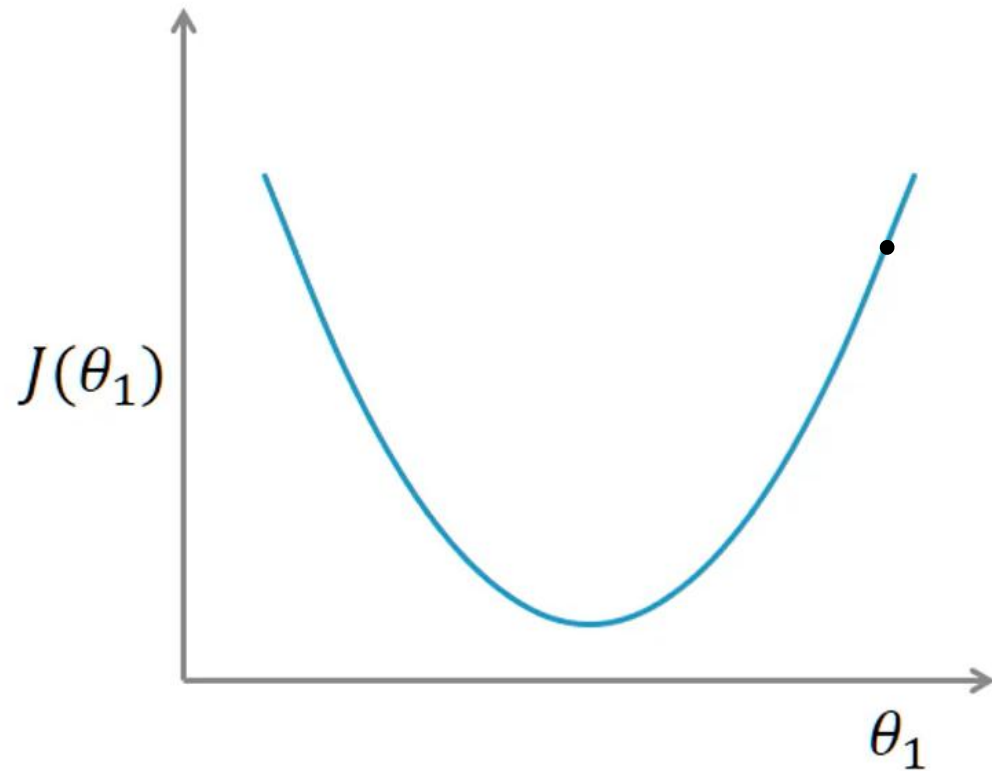
# Размер шага (learning rate)

Если размер шага (learning rate) слишком мал, градиентный спуск сходится слишком медленно или не доходит до точки минимума



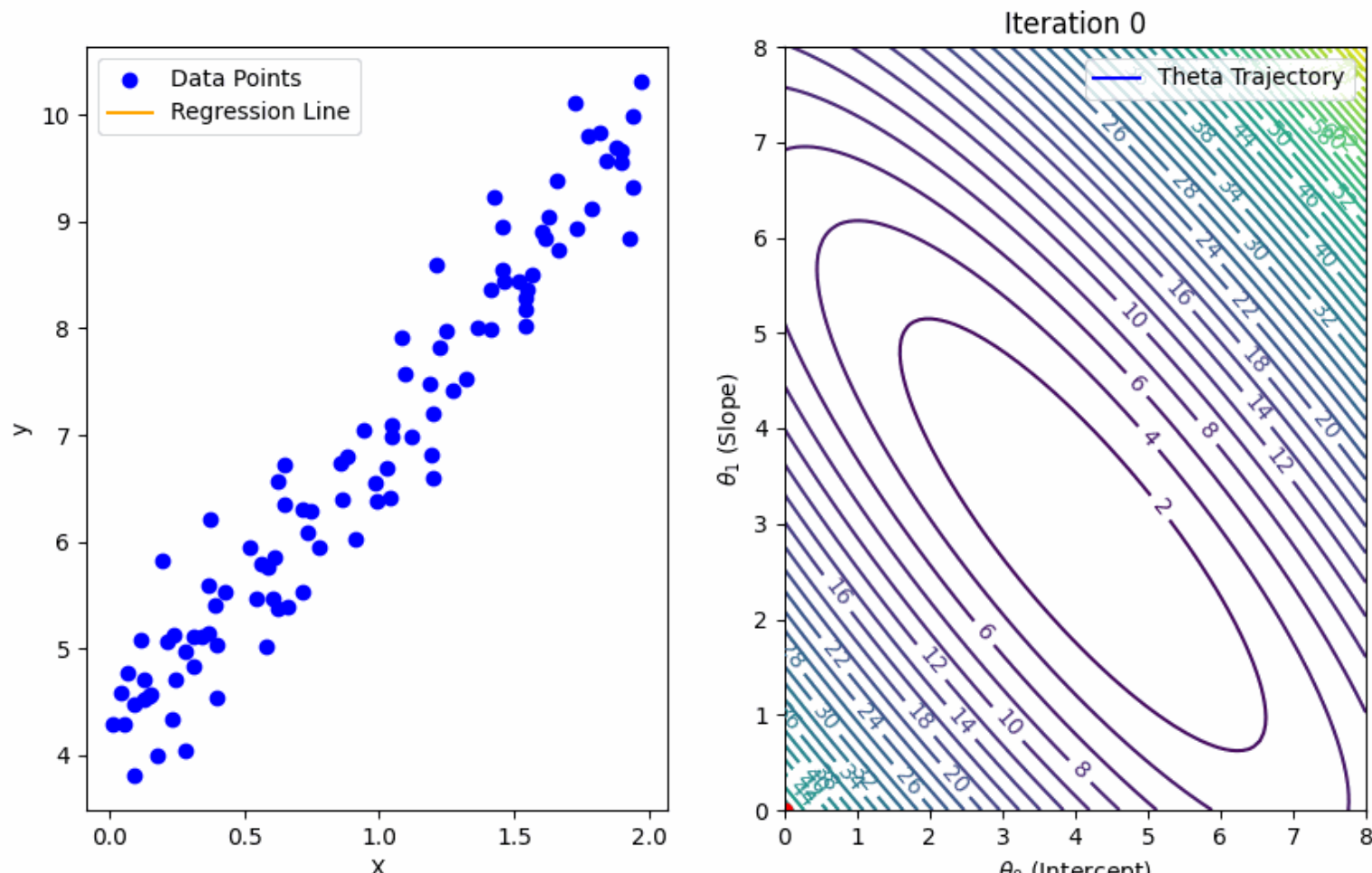
# Градиентный спуск (Gradient descent)

Если размер шага (learning rate) большой, градиентный спуск сходится слишком медленно или вообще не сходится



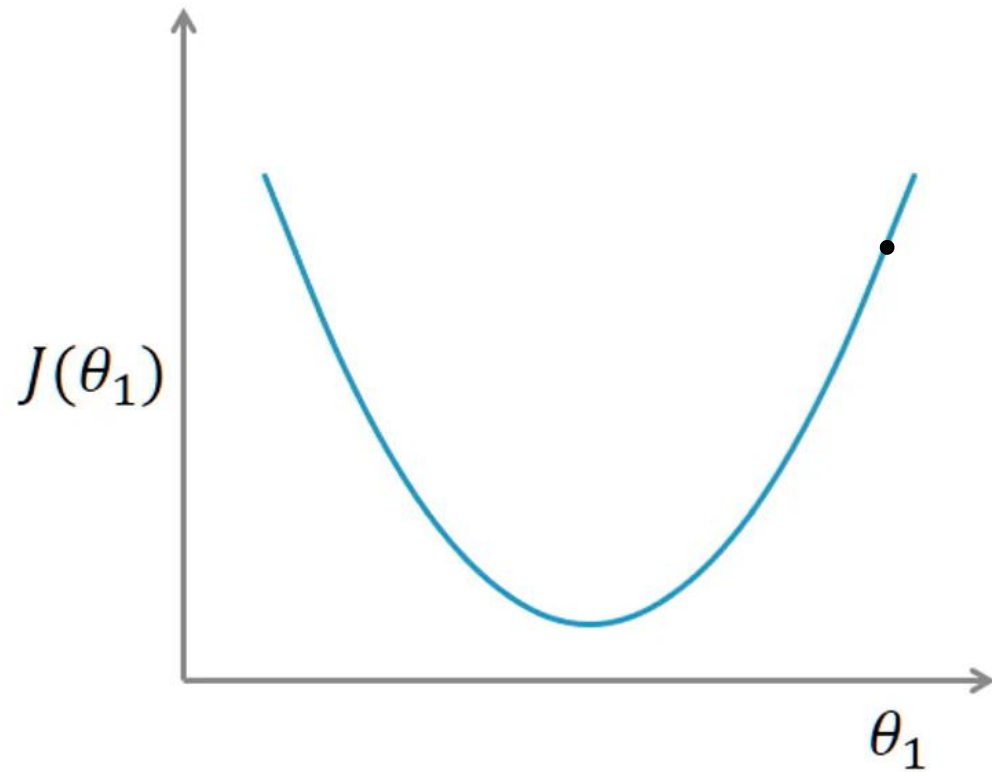
# Градиентный спуск (Gradient descent)

Если размер шага (learning rate) большой, градиентный спуск сходится слишком медленно



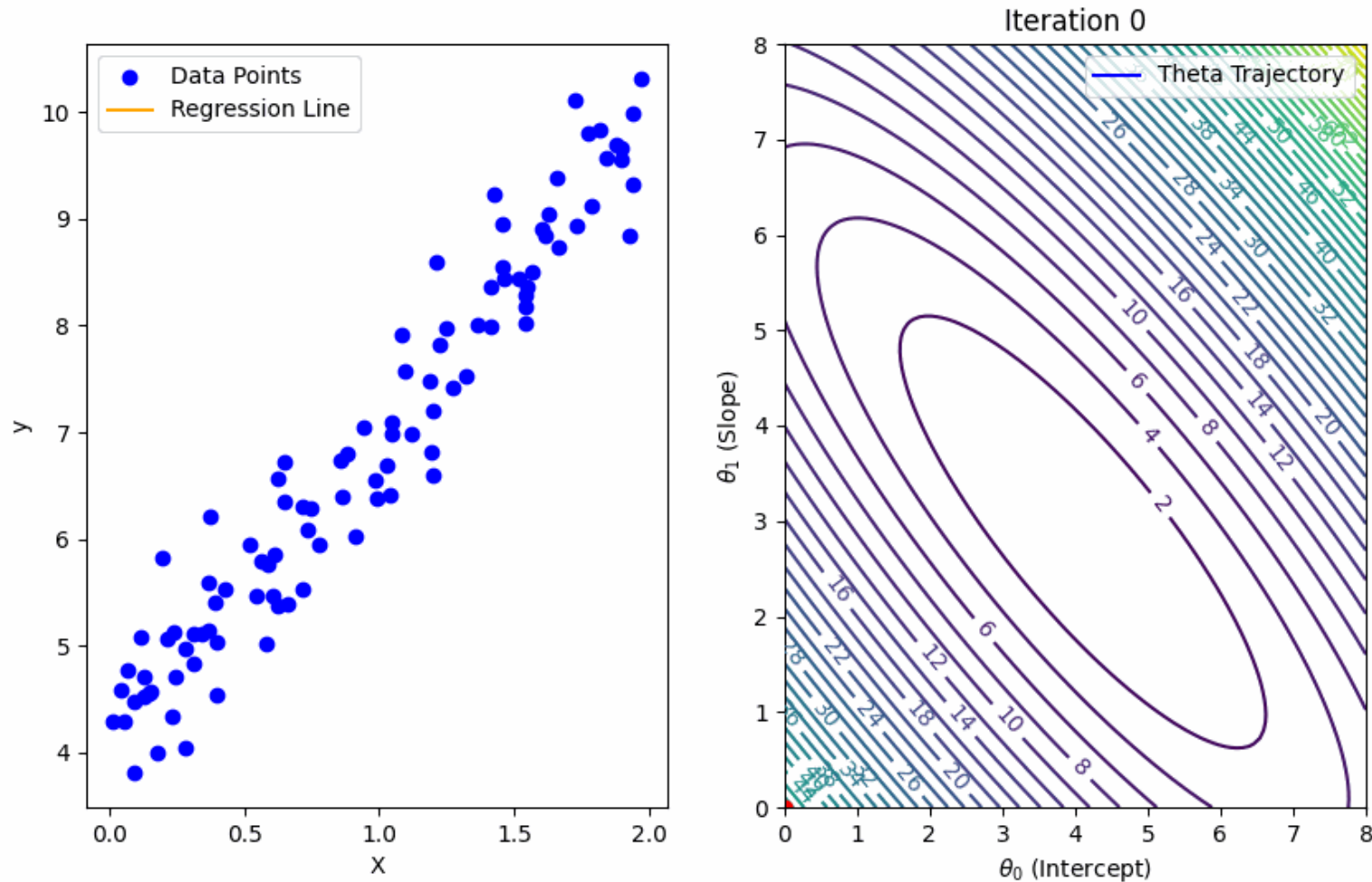
# Градиентный спуск (Gradient descent)

Если размер шага (learning rate) слишком большой, градиентный спуск вообще не сходится



# Градиентный спуск (Gradient descent)

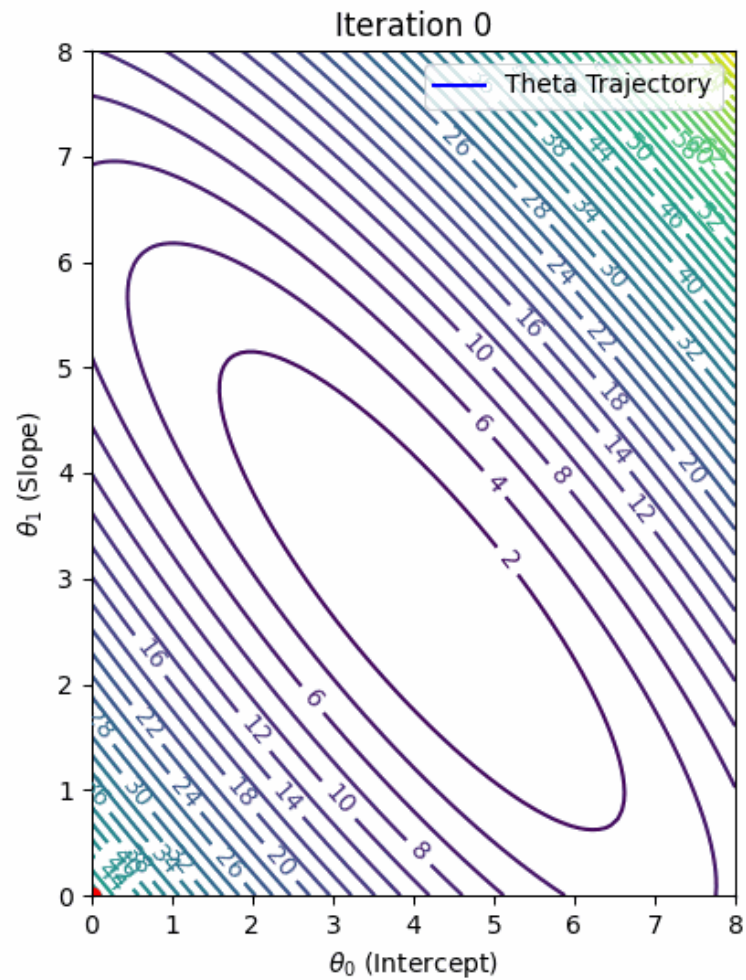
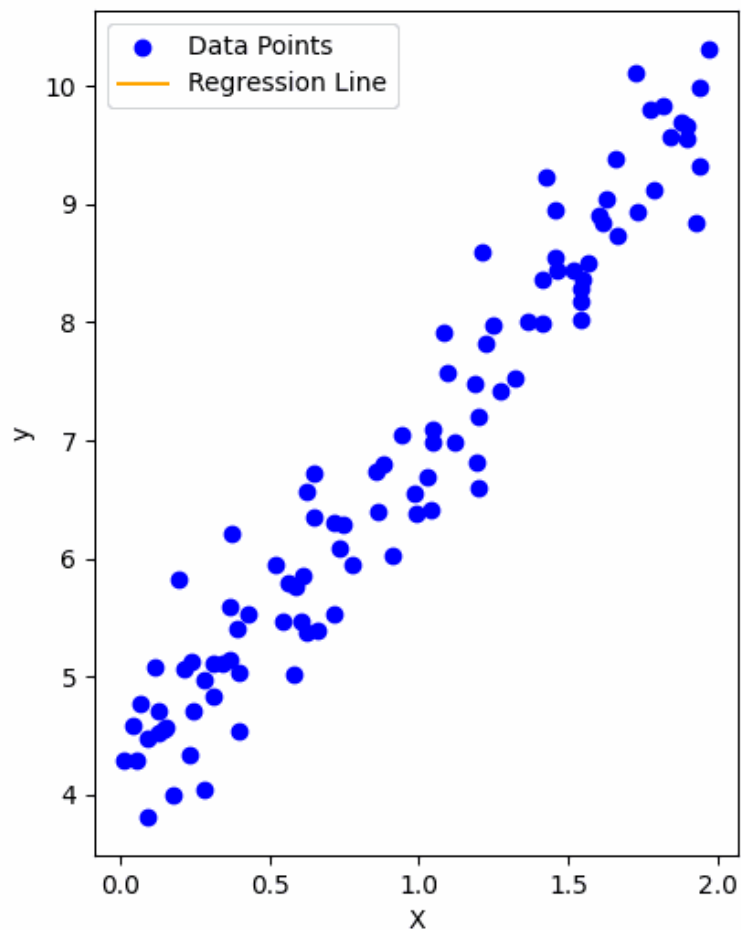
Если размер шага (learning rate) слишком большой, градиентный спуск вообще не сходится.



# Градиентный спуск (Gradient descent)

Длину шага можно менять в зависимости от шага

$$\theta^t = \theta^{t-1} - \alpha \nabla J(\theta^{t-1}) \quad \alpha_t = \frac{1}{t}$$

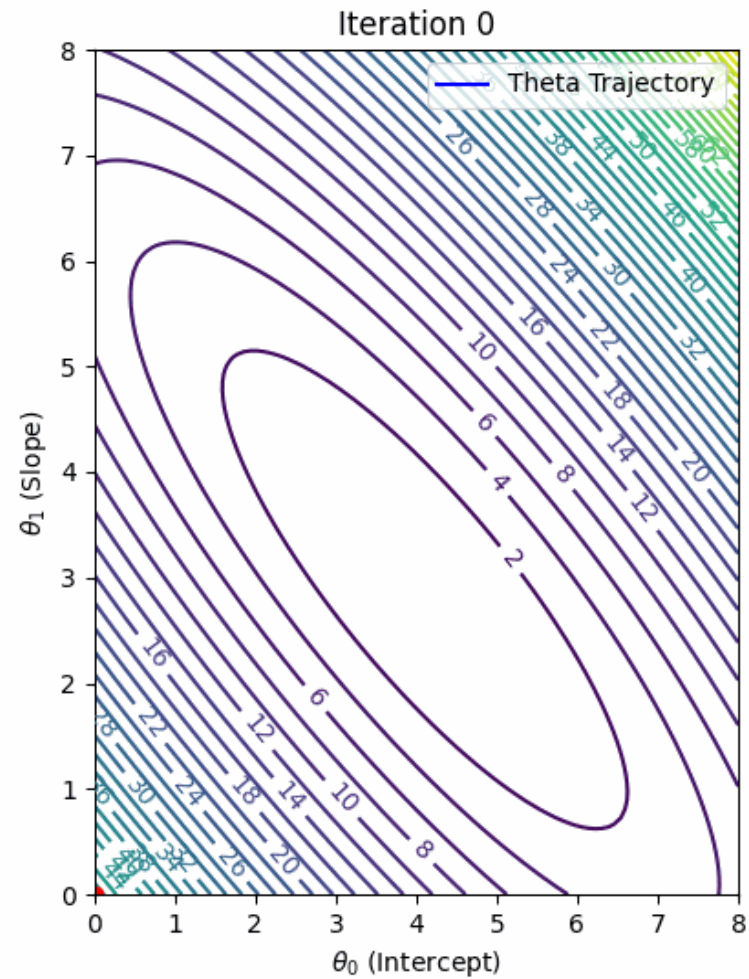
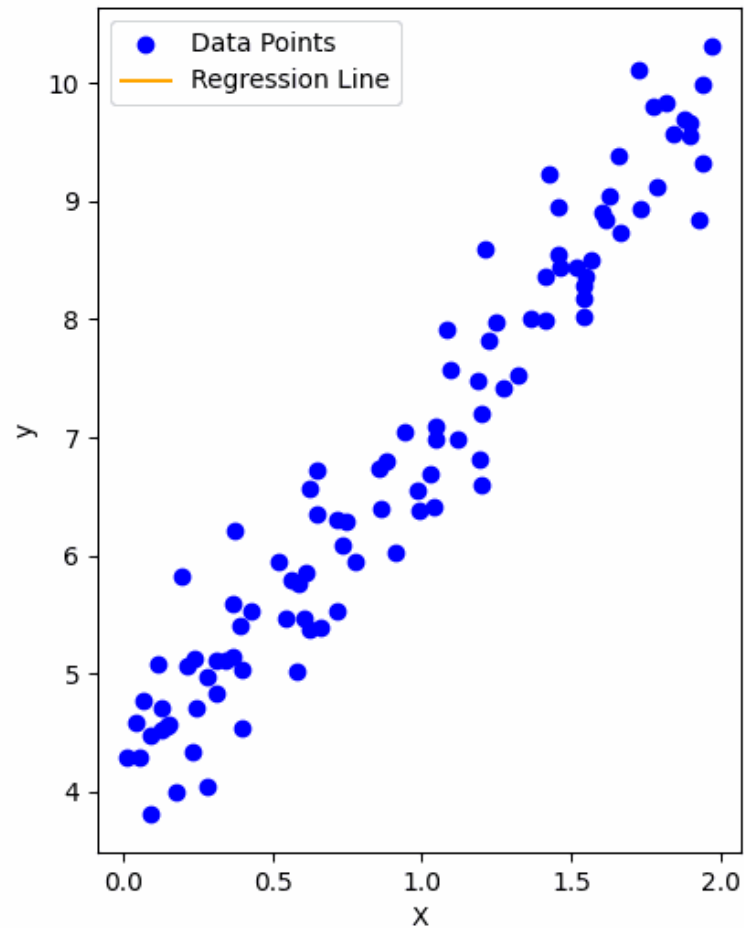




# Градиентный спуск (Gradient descent)

Еще вариант

$$\theta^t = \theta^{t-1} - \alpha \nabla J(\theta^{t-1}) \quad \alpha_t = \frac{0.1}{t\beta}$$





## Стохастический градиентный спуск

# Линейная регрессия

Модель:

$$a(x) = \theta_1 x_1 + \theta_0$$

Функционал ошибки

$$Q(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \theta, x \rangle - y_i)^2$$

Градиент

$$\nabla Q(\theta) = \frac{2}{\ell} X^T (X\theta - y)$$

# Стохастический градиентный спуск (Stochastic gradient descent)

Для вычисления градиента, как правило, надо просуммировать что-то по всем объектам для одного маленького шага

Градиент:

$$\nabla Q(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} \nabla L(y_i, a(x_i))$$

Можно найти градиент по одному слагаемому

$$\nabla J(\theta) \cong \nabla L(y_i, a(x_i))$$

# Стохастический градиентный спуск (Stochastic gradient descent)

1. Стартуем из случайной точки

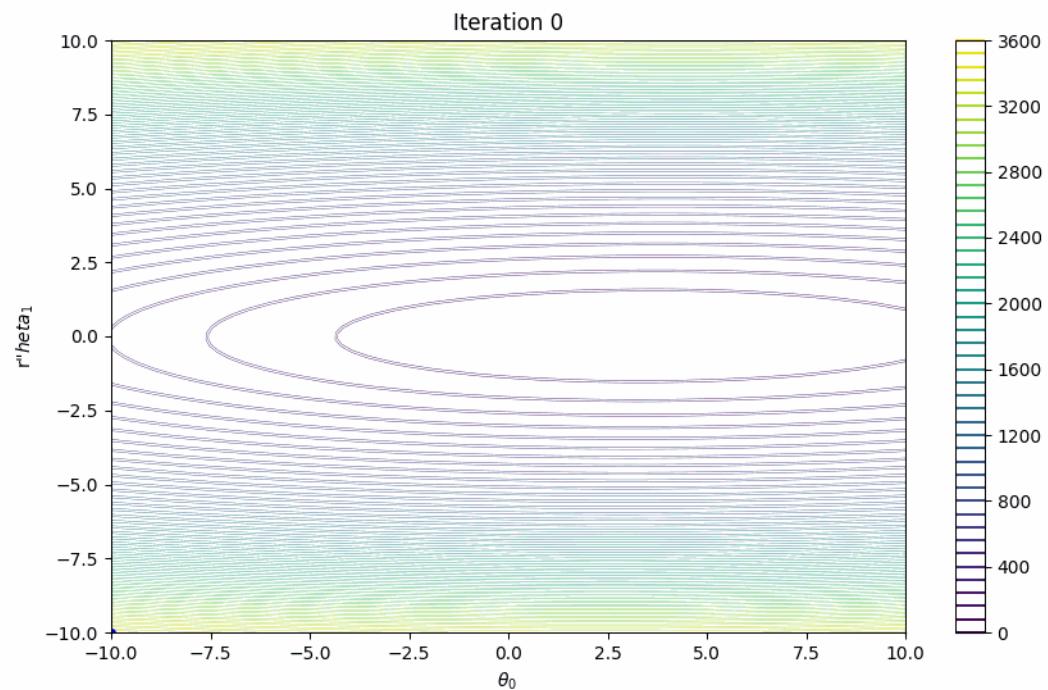
$\theta$  — Инициализация весов

2. Сдвигаемся по антиградиенту, каждый раз выбираем случайный объект  $i_t$

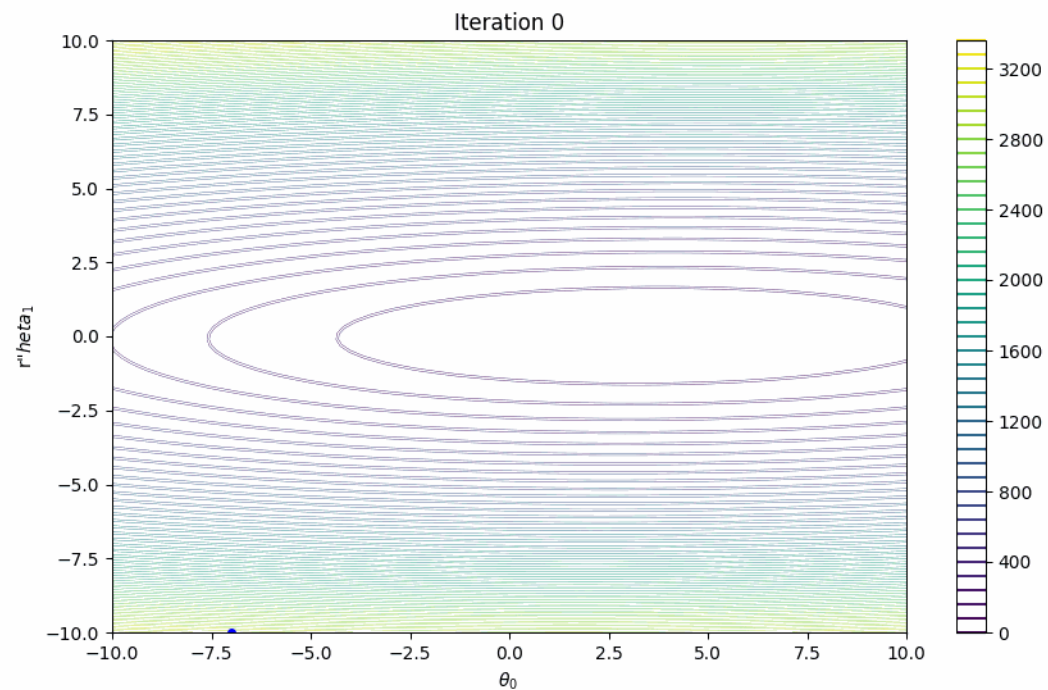
$$\theta^t = \theta^{t-1} - \alpha \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Повторяем пока не окажемся в точке минимума

# Стохастический градиентный спуск (Stochastic gradient descent)

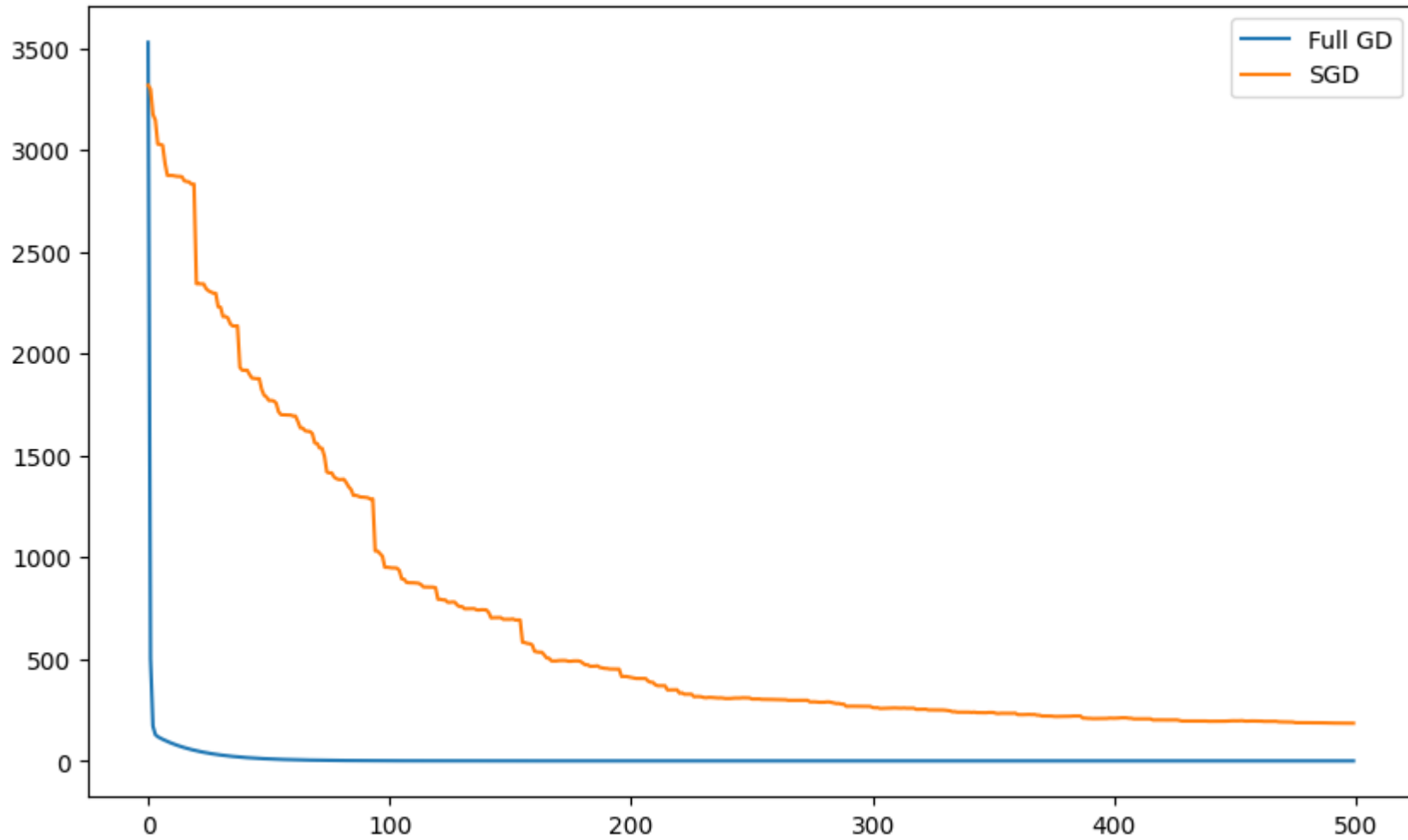


Stochastic gradient descent

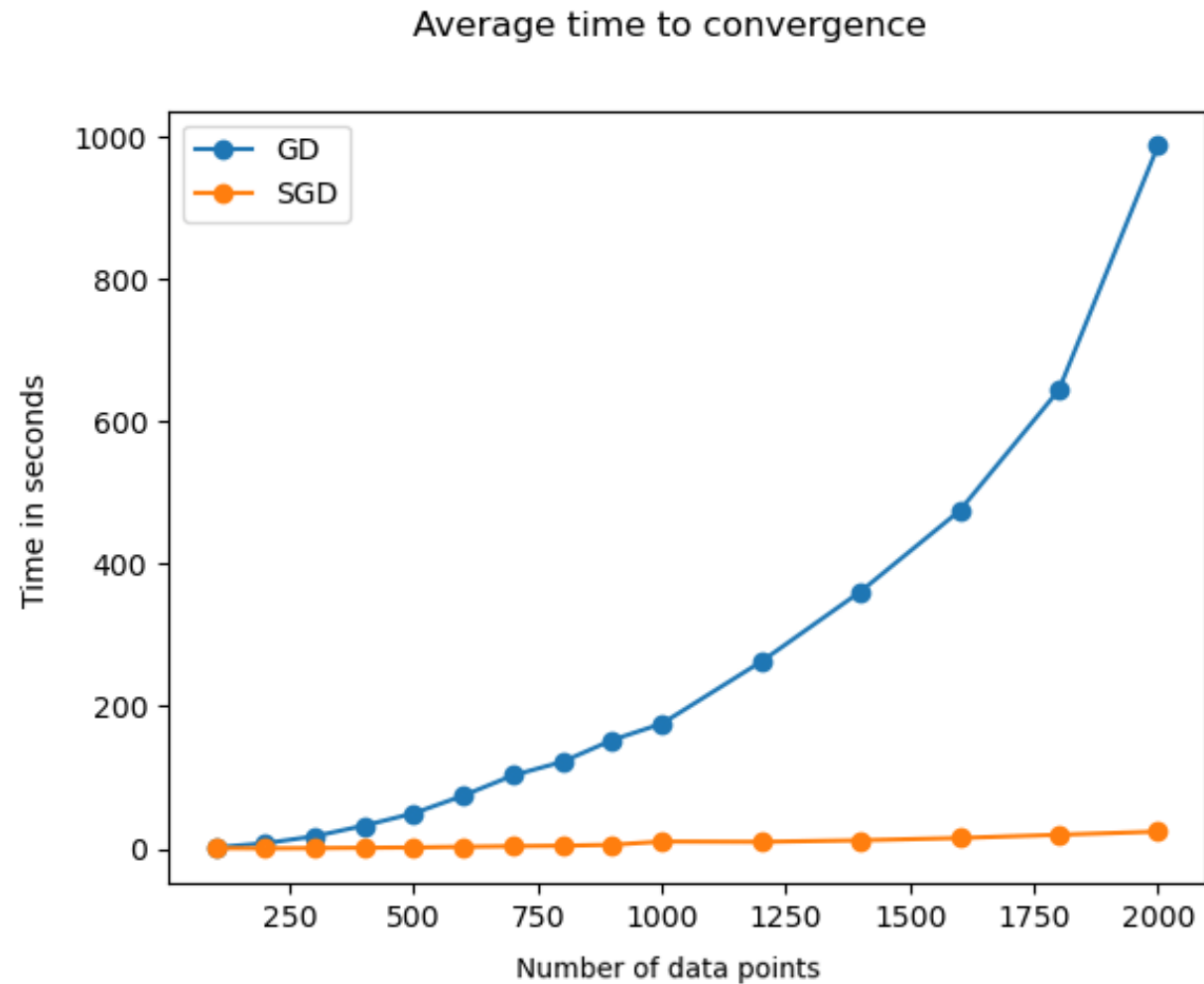


gradient descent

# Стохастический градиентный спуск (Stochastic gradient descent)



# Стохастический градиентный спуск (Stochastic gradient descent)



# Стохастический градиентный спуск (Stochastic gradient descent)

## Mini-batch

1. Стартуем из случайной точки

$\theta$  — Инициализация весов

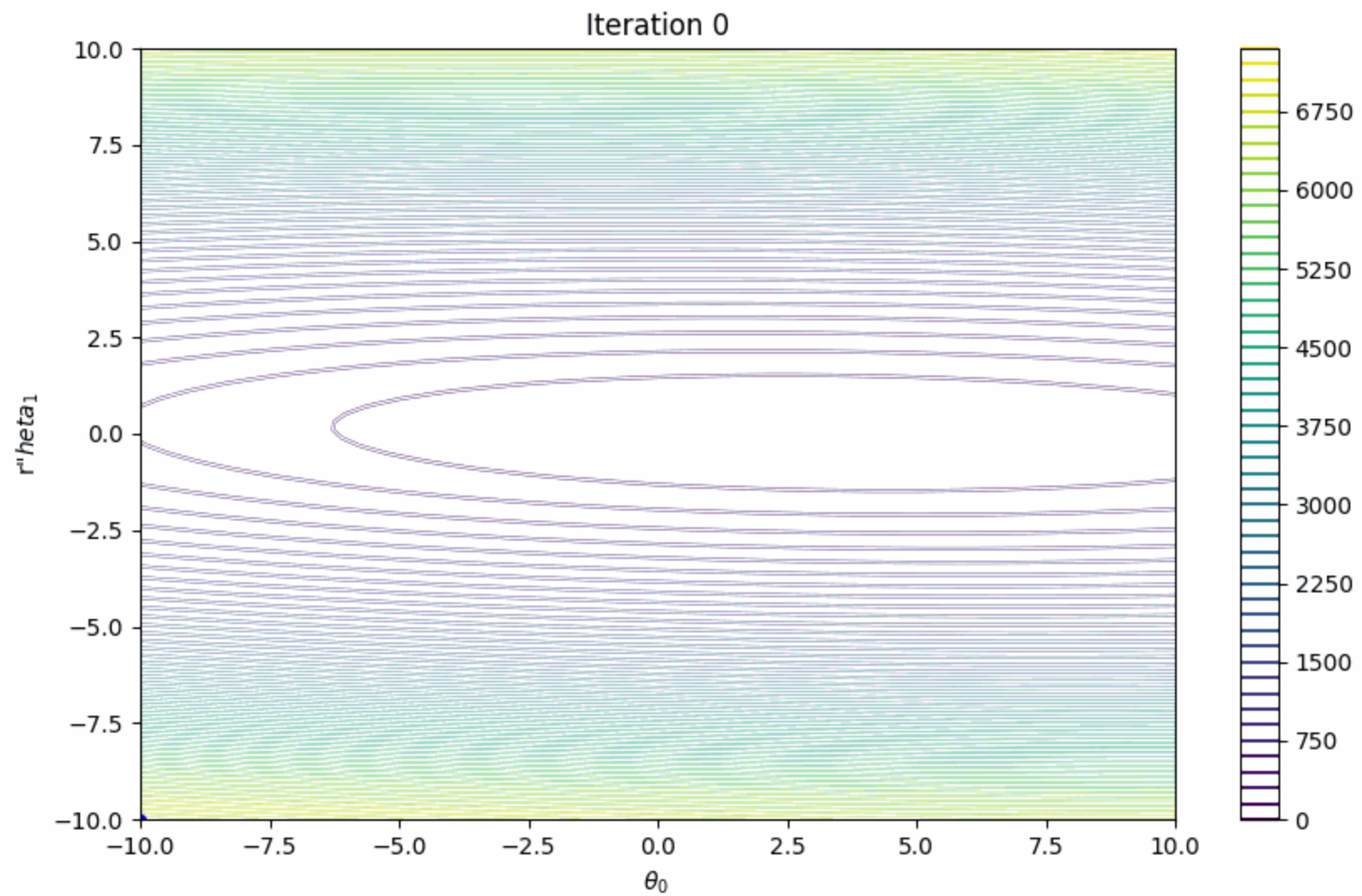
2. Повторяем и сдвигаемся по антиградиенту, каждый раз выбираем  $m$  случайных объектов  $i_1, \dots, i_m$

$$\theta^t = \theta^{t-1} - \alpha \frac{1}{\ell} \sum_{j=1}^{\ell} \nabla L(y_{i_j}, a(x_{i_j}))$$

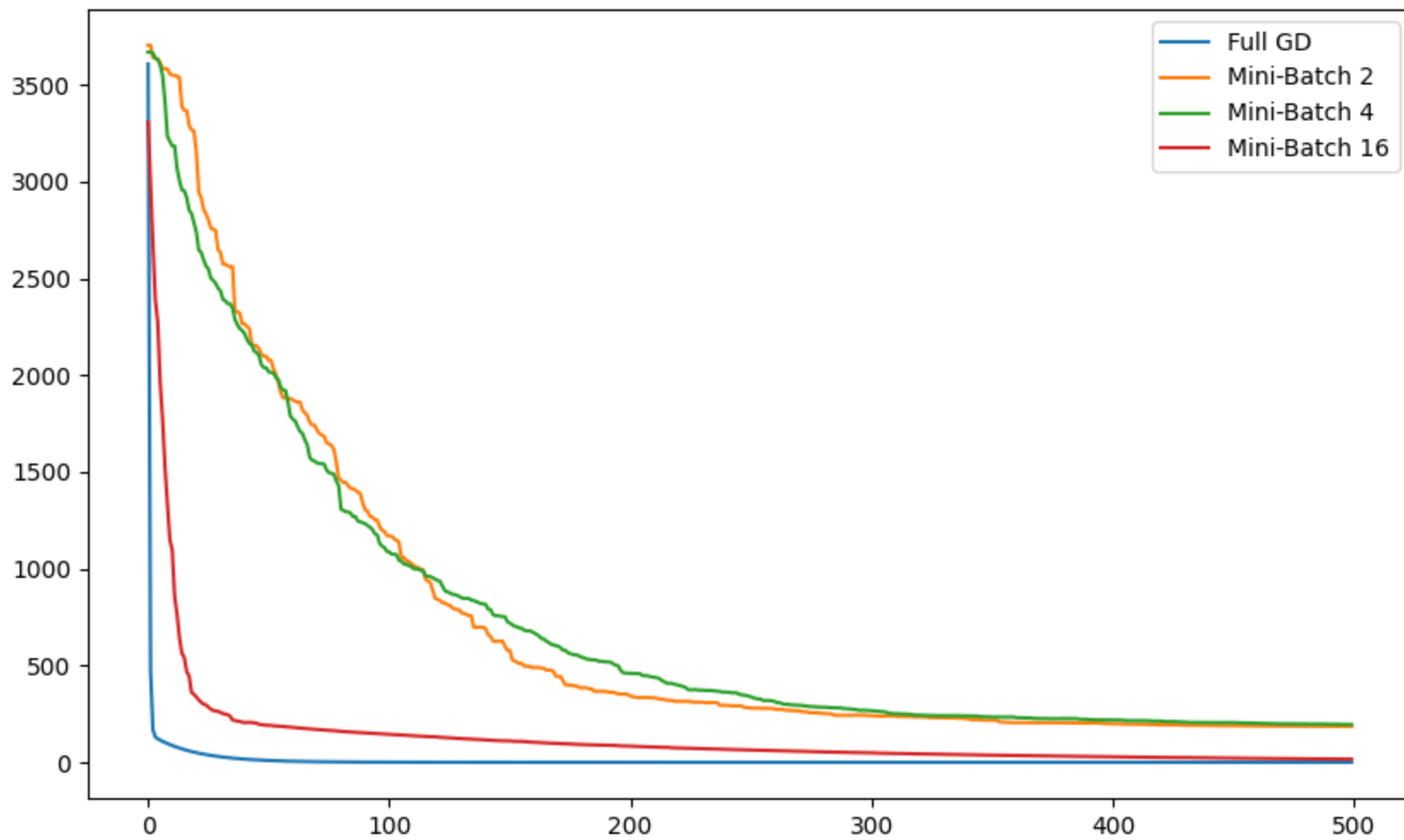
3. Повторяем пока не окажемся в точке минимума



## Mini-batch



# Mini-batch



# Масштабирование данных

## Масштабирование данных (Data Scaling)

Площадь квартиры ( $x_1$ )	Этаж квартиры ( $x_2$ )	Количество комнат ( $x_3$ )	Расстояние от метро ( $x_4$ )	Стоимость квартиры ( $Y$ )
460	2	6	1800	195
230	7	4	4500	130
315	1	3	3200	140
178	3	4	2740	80
...	...	...	...	....

$x_1$ : Площадь квартиры ( 1 – 2000 )

$x_2$ : Этаж квартиры ( 1 – 10 )

$x_3$ : Количество комнат ( 1 – 10 )

$x_4$ : Расстояние от метро ( 1 – 10000 )

зависящие

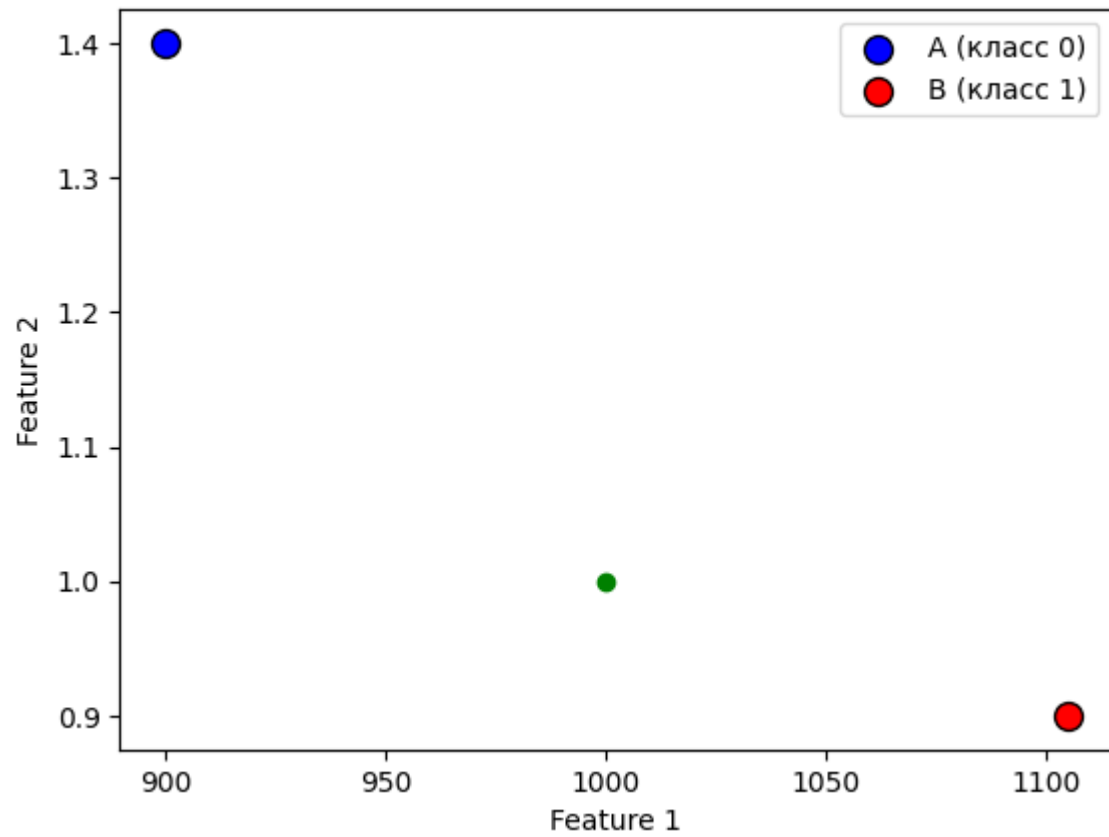
- Метод ближайших соседей
- Линейная регрессия
- Логистическая регрессия
- Метод опорных векторов (SVM)
- Нейронные сети
- Некоторые алгоритмы кластеризации (K-means)
- Анализ главных компонент (Principal Component Analysis, PCA)

Не зависящие

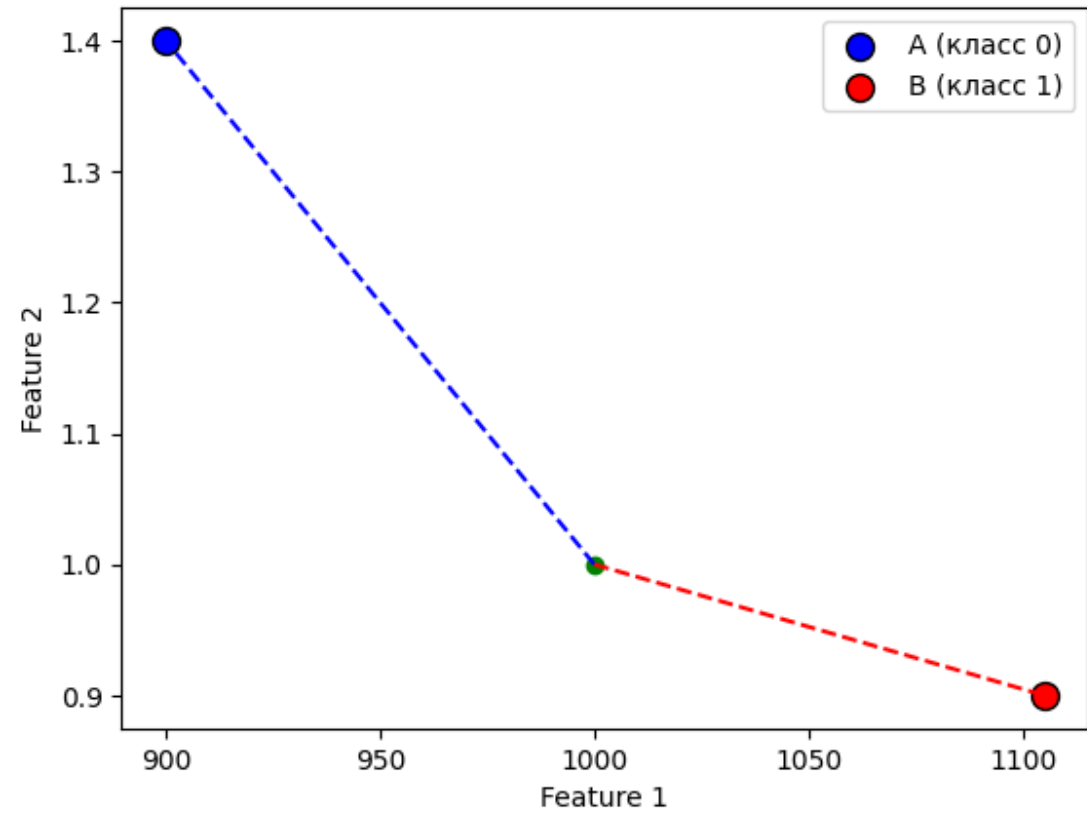
- Случайный лес
- Градиентный бустинг

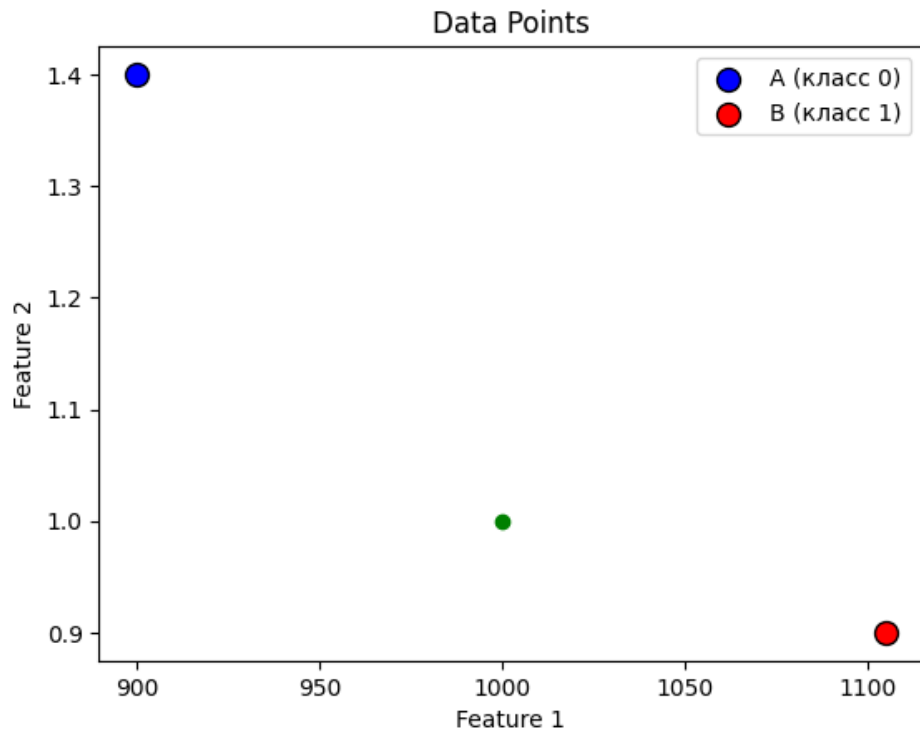
# Knn

Data Points



Data Points





### Без масштабирования

$$d(q, A) = \sqrt{(1000 - 900)^2 + (1.0 - 1.4)^2} = \sqrt{10000 + 0.16} \approx \mathbf{100.0008}$$

$$d(q, B) = \sqrt{(1000 - 1105)^2 + (1.0 - 0.9)^2} = \sqrt{11025 + 0.01} \approx \mathbf{105}$$

### После масштабирования

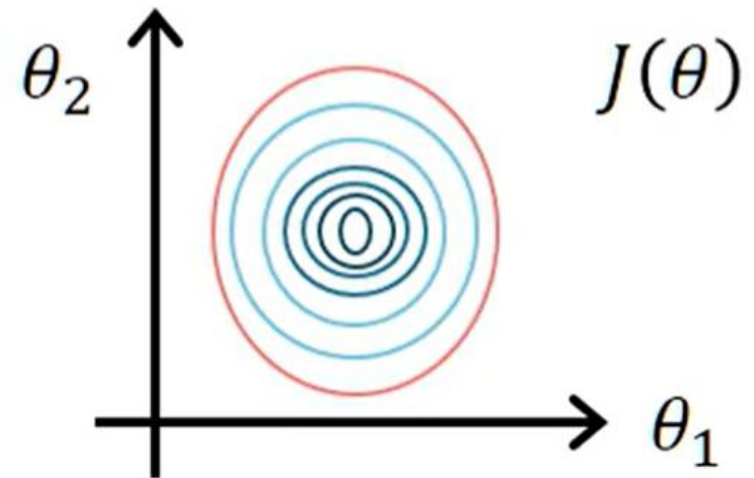
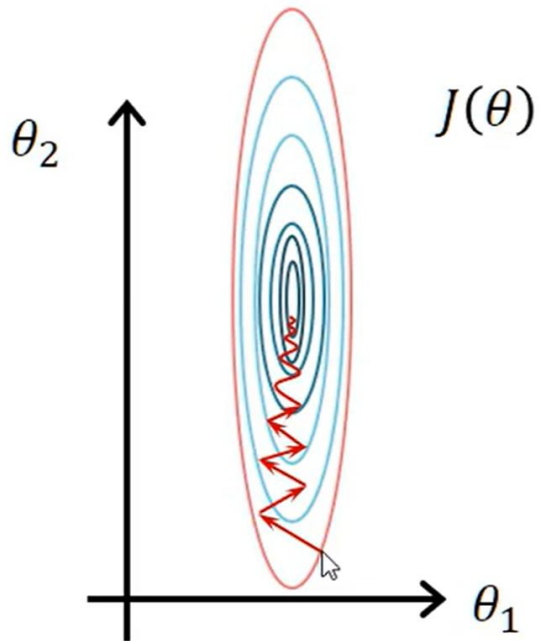
$$d(q, A) = 1.8740$$

$$d(q, B) = \mathbf{1.0997}$$

После приведения масштабов **Feature2** перестаёт быть «незаметным» и реально влияет на расстояние ближайший сосед меняется, следовательно, меняется предсказанный класс (k=1).

# Масштабирование данных (Data Scaling)

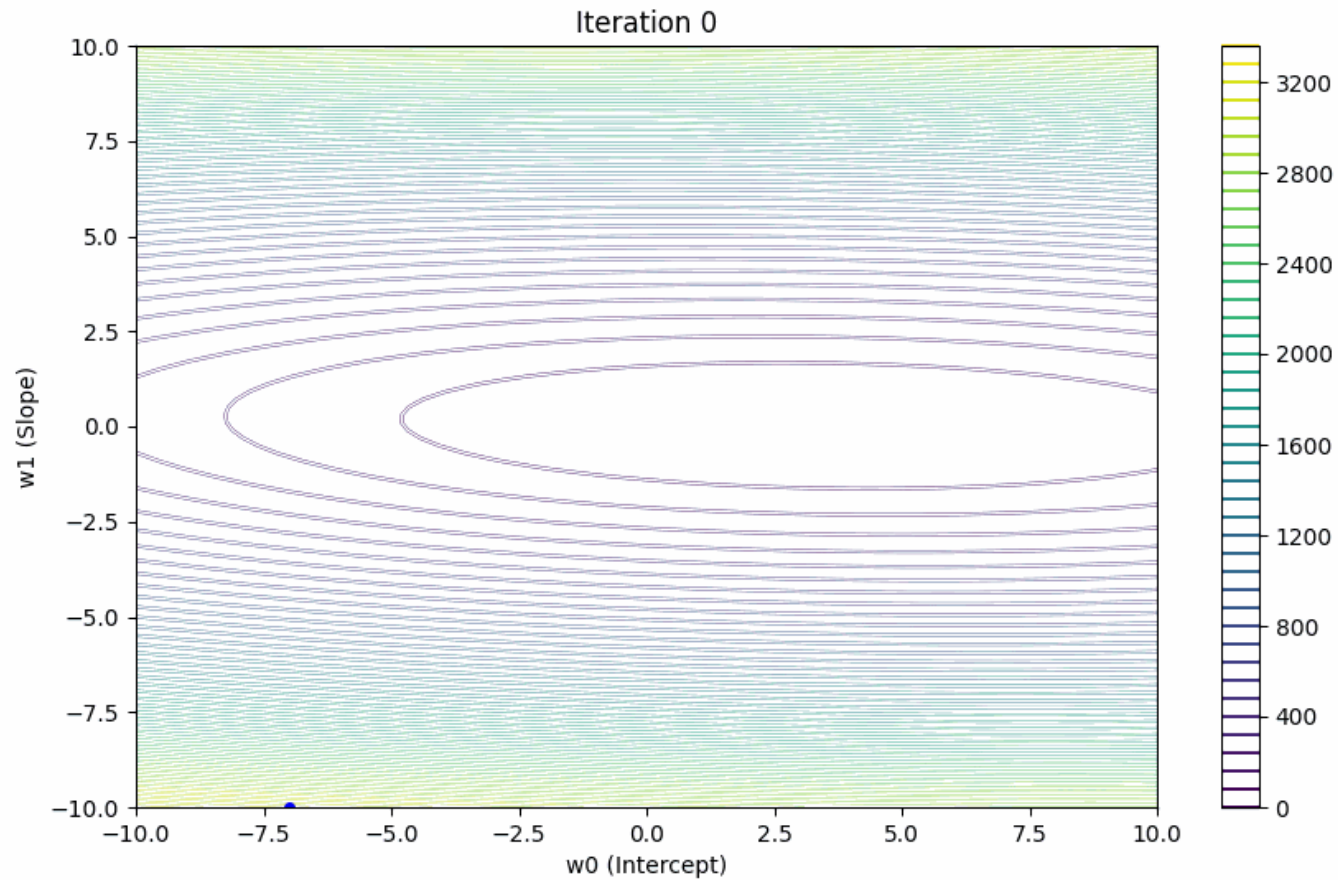
**Цель:** увеличение скорости сходимости в градиентном спуске.





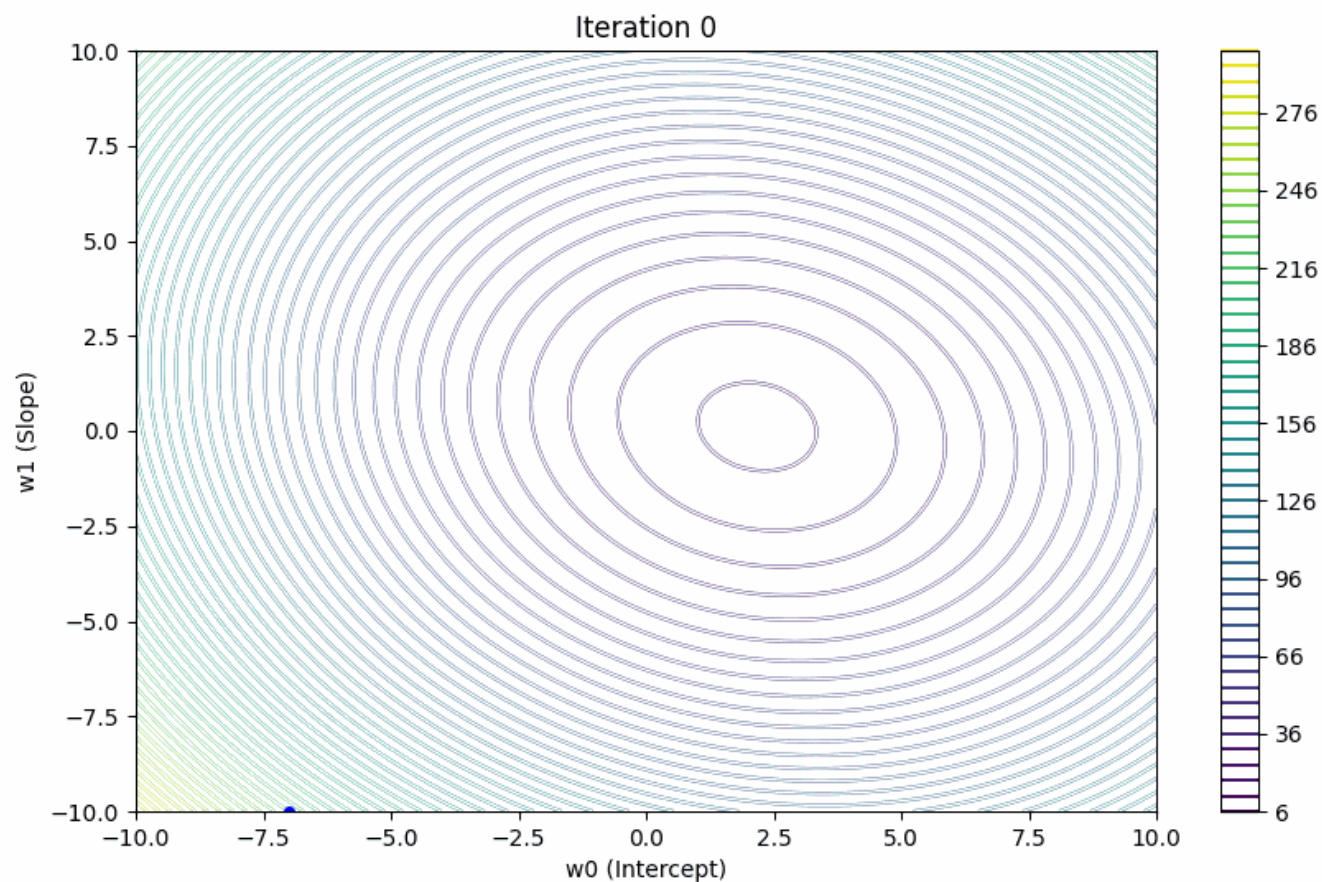
# Масштабирование данных (Data Scaling)

**Цель:** увеличение скорости сходимости в градиентном спуске.



# Масштабирование признаков

Процесс изменения данных происходит таким образом, чтобы они имели одинаковый масштаб



# Масштабирование признаков на основе Z-оценки

Вычтем из каждого значения признака среднее и поделим на стандартное отклонение

$x_i^{d_j}$  - признак

$\mu_j$ - mean(x) – среднее значение

$$x_i^{d_j} = \frac{x_i^{d_j} - \mu_j}{\sigma_j}$$

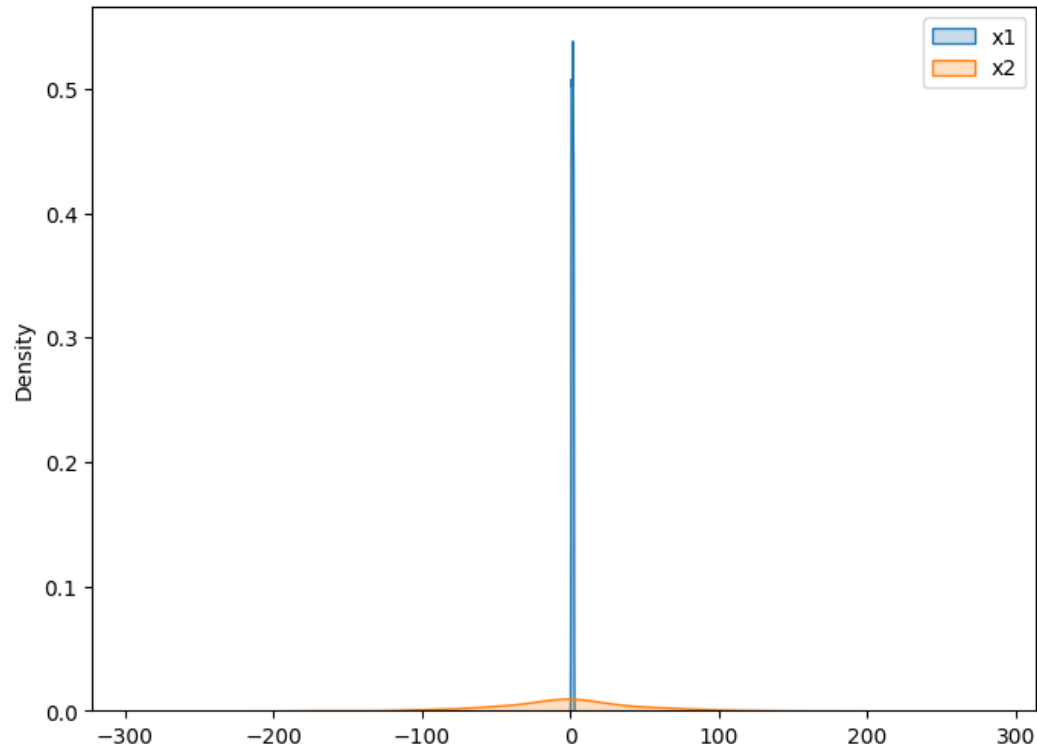
$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$\sigma_j$ - std(x) – стандартное отклонение

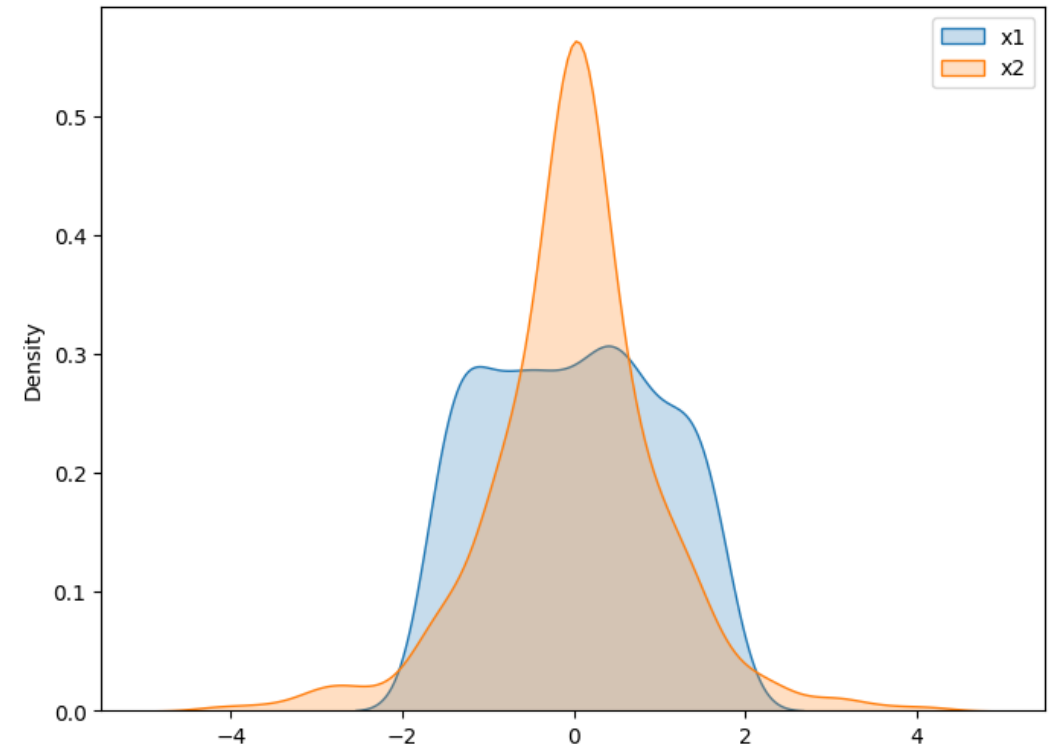
$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

# Масштабирование признаков на основе Z-оценки

До масштабирования



После масштабирования



- Среднее значение масштабированных данных становится равным 0.
- Стандартное отклонение масштабированных данных становится равным 1.
- Выбросы сохраняются

# Mean Normalization

$$x_i^{d_j} = \frac{x_i^{d_j} - \mu_j}{\max(x_i^{d_j}) - \min(x_i^{d_j})}$$

$x_i^{d_j}$  - признак

$\mu_j$  - mean(x) – среднее значение

- Среднее значение масштабированных данных становится равным 0.
- Максимальные и минимальные значения в диапазоне [-1;1]
- Выбросы сохраняются

# Mean Normalization

## Данные без выброса

$$X = [2, 4, 6, 8]$$

Среднее  $\mu=5$

Min=2, Max=8 , диапазон=6

Вычислим:

$$\frac{2 - 5}{6} = -0.5$$

$$X_{scale} = [-0.5, -0.17, 0.17, +0.5]$$

## Данные с выбросом

$$X = [2, 4, 6, 8, 100]$$

Среднее  $\mu=24$

Min=2, Max=8 , диапазон=98

Вычислим:

$$\frac{2 - 24}{98} = -0.22$$

$$X_{scale} = [-0.22, -0.20, -0.18, -0.16, +0.78]$$

# MinMax-масштабирование

$$x_i^{d_j} = \frac{x_i^{d_j} - \min(x_i^{d_j})}{\max(x_i^{d_j}) - \min(x_i^{d_j})}$$

$x_i^{d_j}$  - признак

- Среднее значение и среднеквадратичное отклонение может варьироваться.
- Максимальные и минимальные значения в диапазоне [0;1]
- Выбросы сохраняются

# Масштабирование по максимальному значению(MaxAbs Scaling)

$$x_i^{d_j} = \frac{x_i^{d_j}}{\max(|x_i^{d_j}|)}$$

$x_i^{d_j}$  - признак

- Среднее значение не центрируется.
- Максимальные и минимальные значения в диапазоне [-1;1]
- Среднеквадратичное отклонение не масштабируется.



$$x'_i = \frac{x_i - \text{Median}(x)}{IQR}$$

$\text{Median}(x)$  — медиана признака (центральное значение)

$IQR = Q3 - Q1$  межквартильный размах  
(разница между 75-м и 25-м перцентилями)

- Данные центрируются вокруг медианы (вместо среднего).
- Значения не ограничены диапазоном  $[-1;1]$  или  $[0;1]$
- Выбросы не влияют на масштаб