

# Введение в анализ данных

Лекция 5

Линейная регрессия и градиентный спуск

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2020

# Обучение линейной регрессии

- Можно посчитать градиент MSE:

$$\nabla \frac{1}{\ell} \|Xw - y\|^2 = \frac{2}{\ell} X^T(Xw - y)$$

- Приравниваем нулю и решаем систему линейных уравнений:

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

# Аналитическое решение

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

- Если матрица  $X^T X$  вырожденная, то будут проблемы
- Даже если она почти вырожденная, всё равно будут проблемы
- Если признаков много, то придётся долго ждать

# Регуляризация

- Регуляризованный функционал

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- Аналитическое решение:

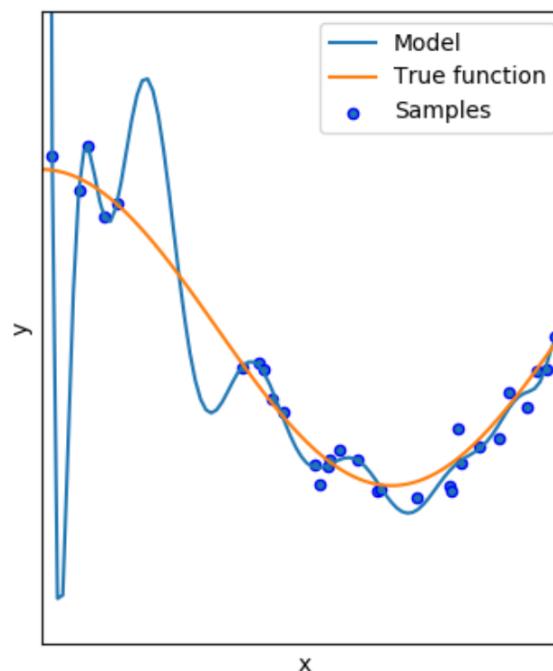
$$w = (X^T X + \lambda I)^{-1} X^T y$$

- Гребневая регрессия (Ridge regression)

# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

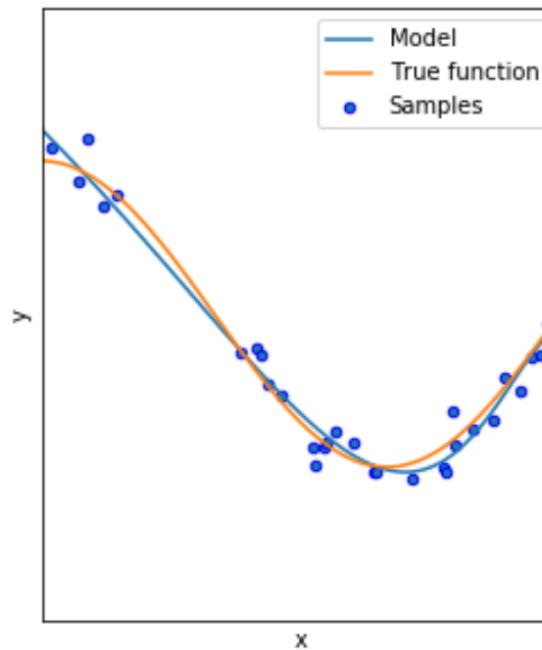
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

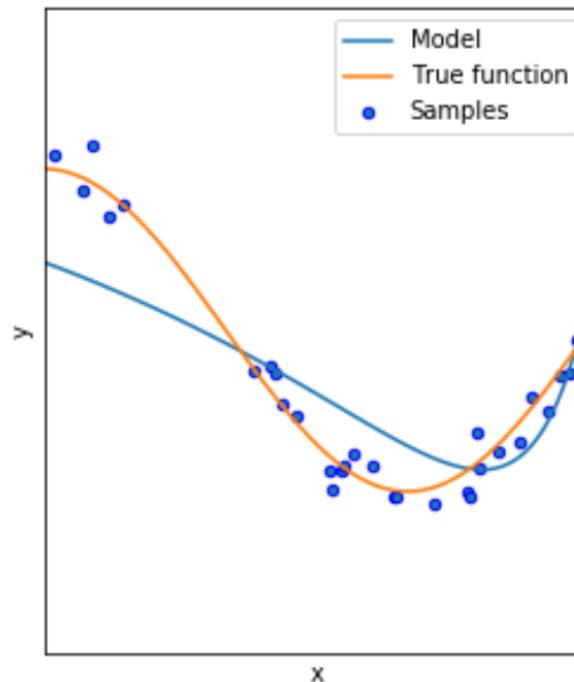
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{0.01} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

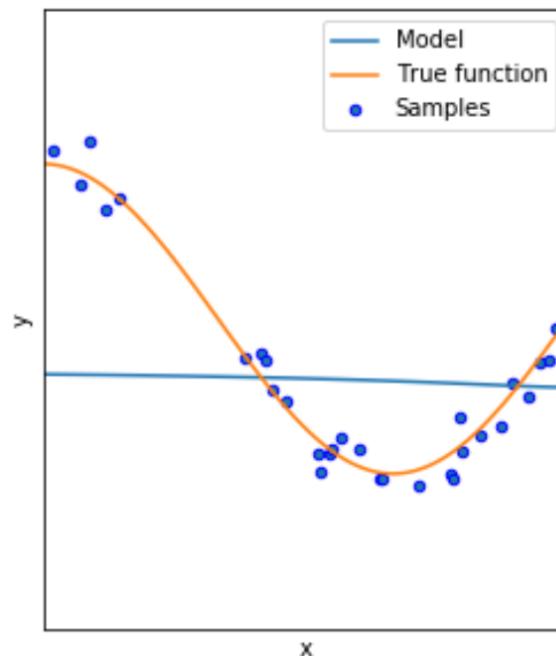
$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{1} \|w\|^2 \rightarrow \min_w$$



# Эффект регуляризации

$$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + \cdots + w_{15} x^{15}$$

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2 + \textcolor{red}{100} \|w\|^2 \rightarrow \min_w$$



# Градиент и его свойства

# Среднеквадратичная ошибка

- MSE для линейной регрессии:

$$Q(w_1, \dots, w_d) = \sum_{i=1}^{\ell} (\textcolor{red}{w_1}x_1 + \dots + \textcolor{red}{w_d}x_d - y_i)^2$$

# Градиент

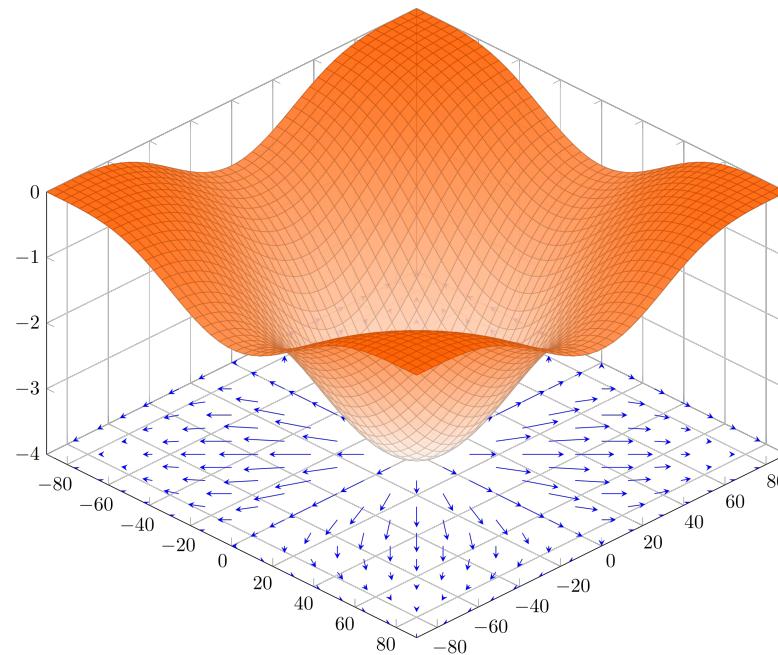
- Градиент — вектор частных производных

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d} \right)$$

- У градиента есть важное свойство!

# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?



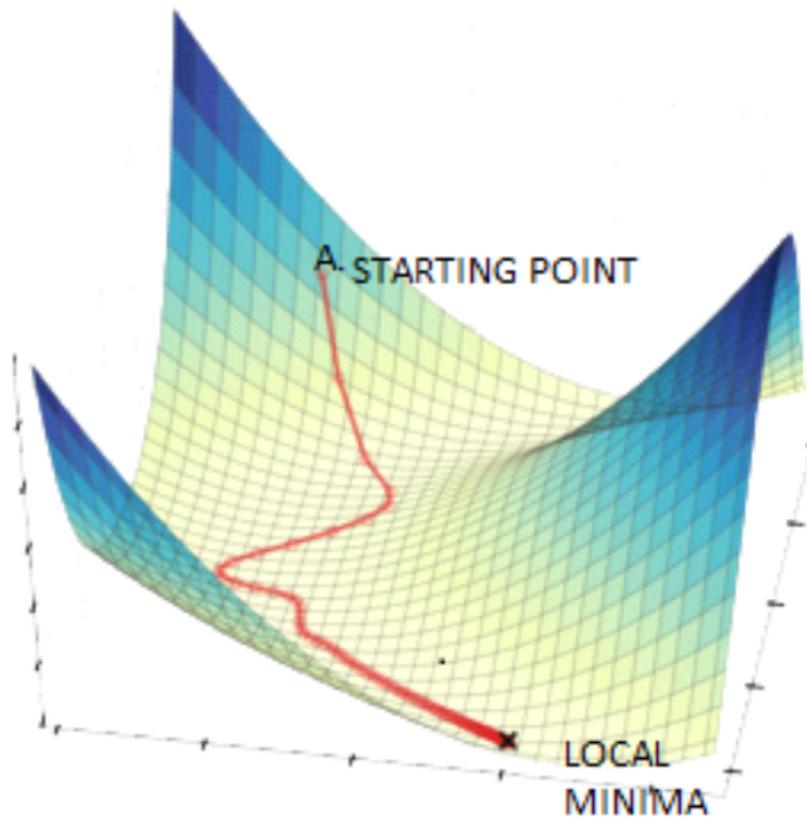
# Важное свойство

- Зафиксируем точку  $x_0$
- В какую сторону функция быстрее всего растёт?
- В направлении градиента!
- А быстрее всего убывает в сторону антиградиента

Как это пригодится?



Как это пригодится?



# Градиентный спуск

# Градиентный спуск

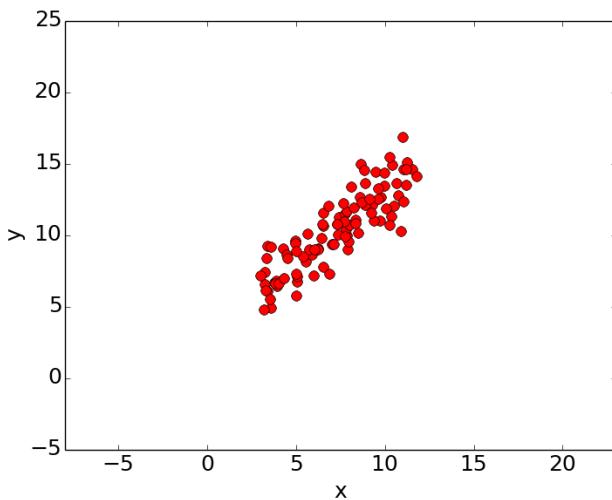
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

# Парная регрессия

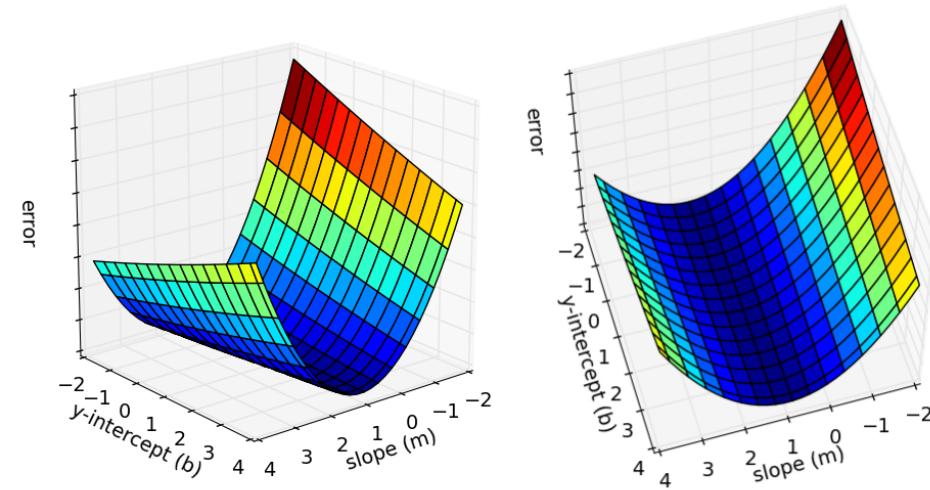
- Простейший случай: один признак
- Модель:  $a(x) = w_1 x + w_0$
- Два параметра:  $w_1$  и  $w_0$
- Функционал:

$$Q(w_0, w_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

# Парная регрессия



Выборка



Функционал ошибки

# Парная регрессия

$$Q(w_0, w_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (w_1 x_i + w_0 - y_i)$
- $\frac{\partial Q}{\partial w_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)$
- $\nabla Q(w) = \left( \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (w_1 x_i + w_0 - y_i), \frac{2}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i) \right)$

# Начальное приближение

- $w^0$  — инициализация весов
- Например, из стандартного нормального распределения

# Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

Новая точка

Размер шага

Градиент в  
предыдущей  
точке

# Сходимость

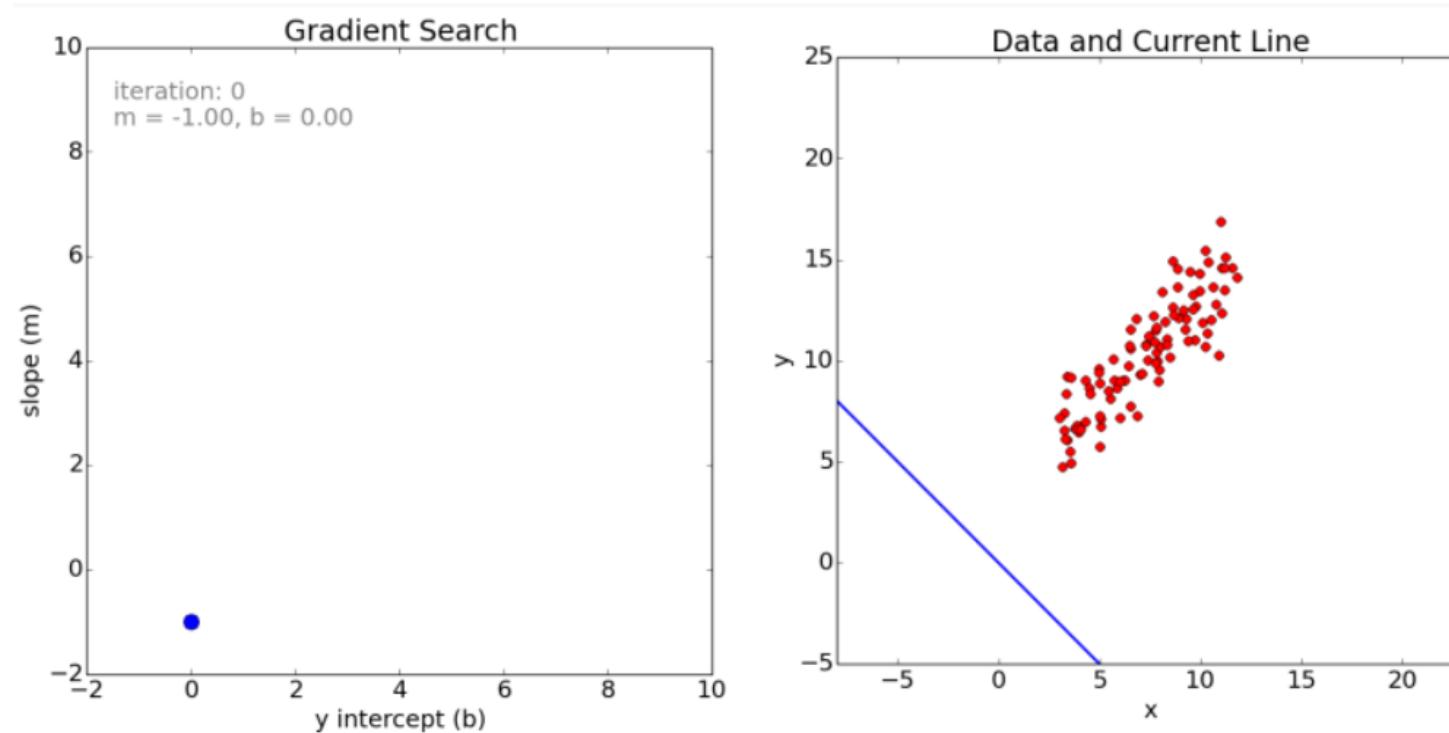
- Останавливаем процесс, если

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

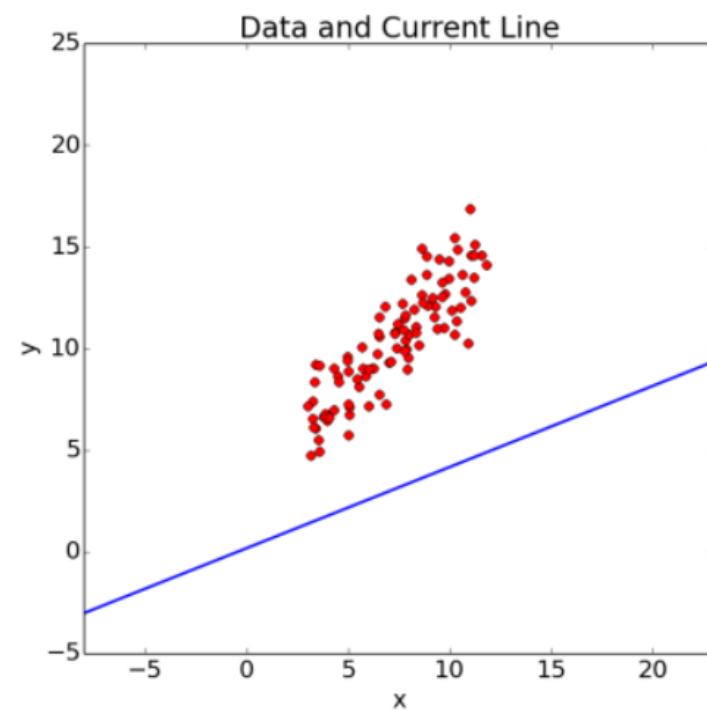
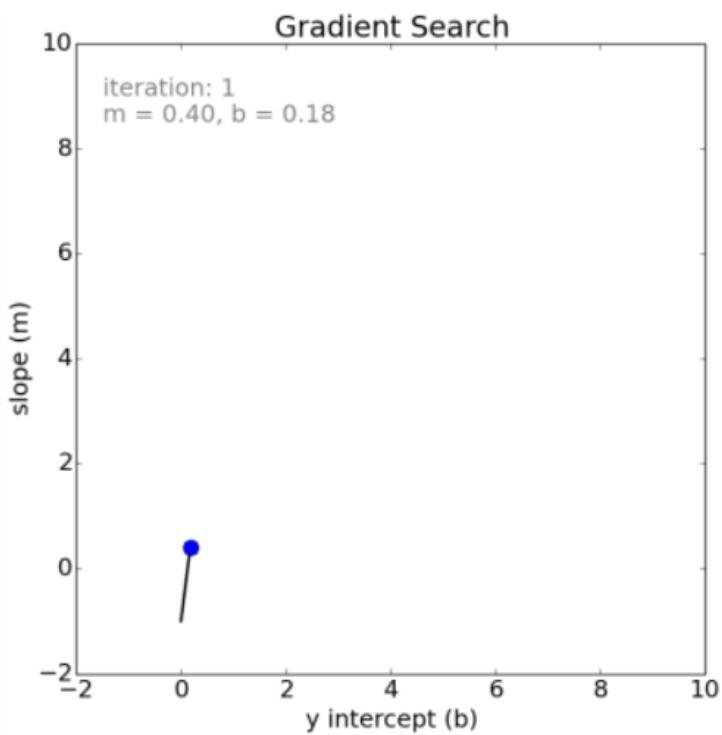
- Другой вариант:

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

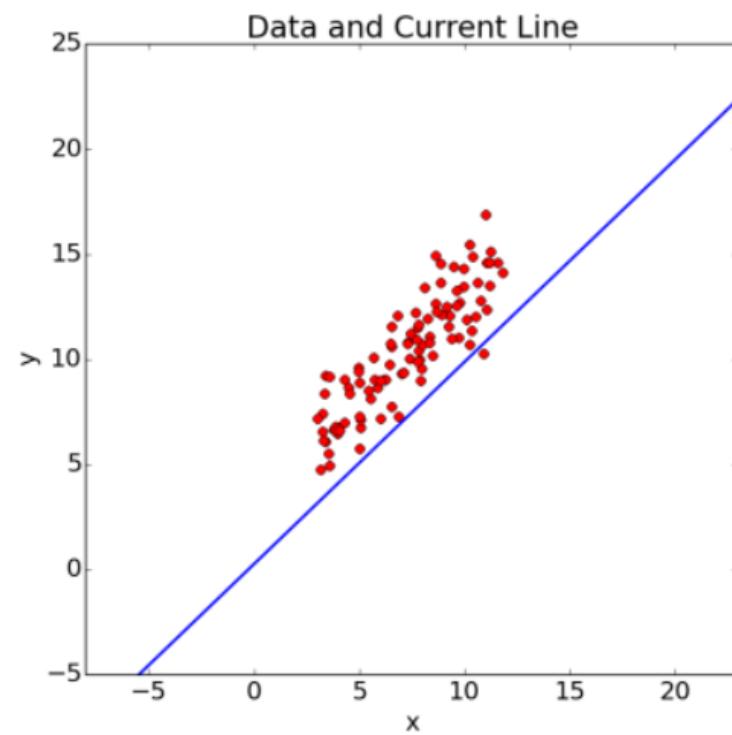
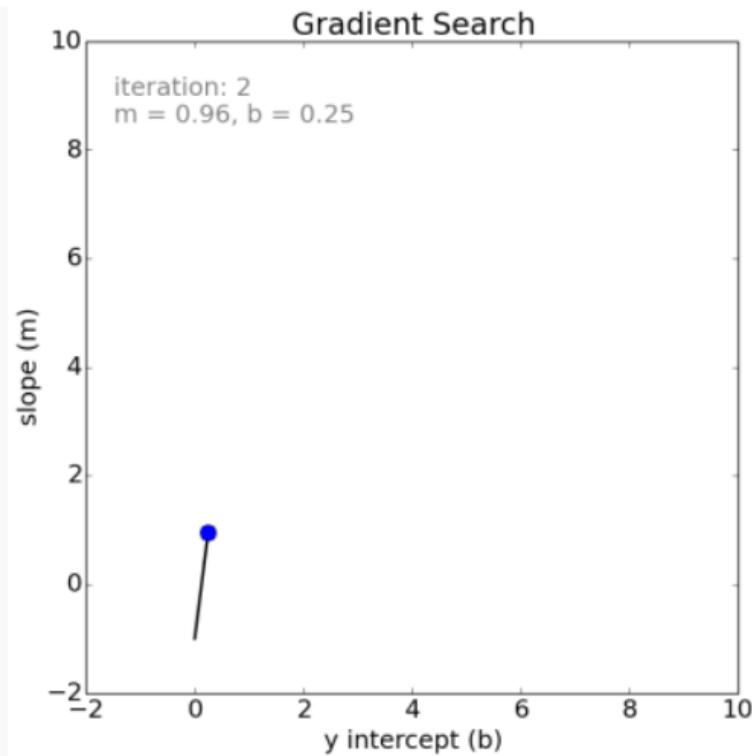
# Парная регрессия



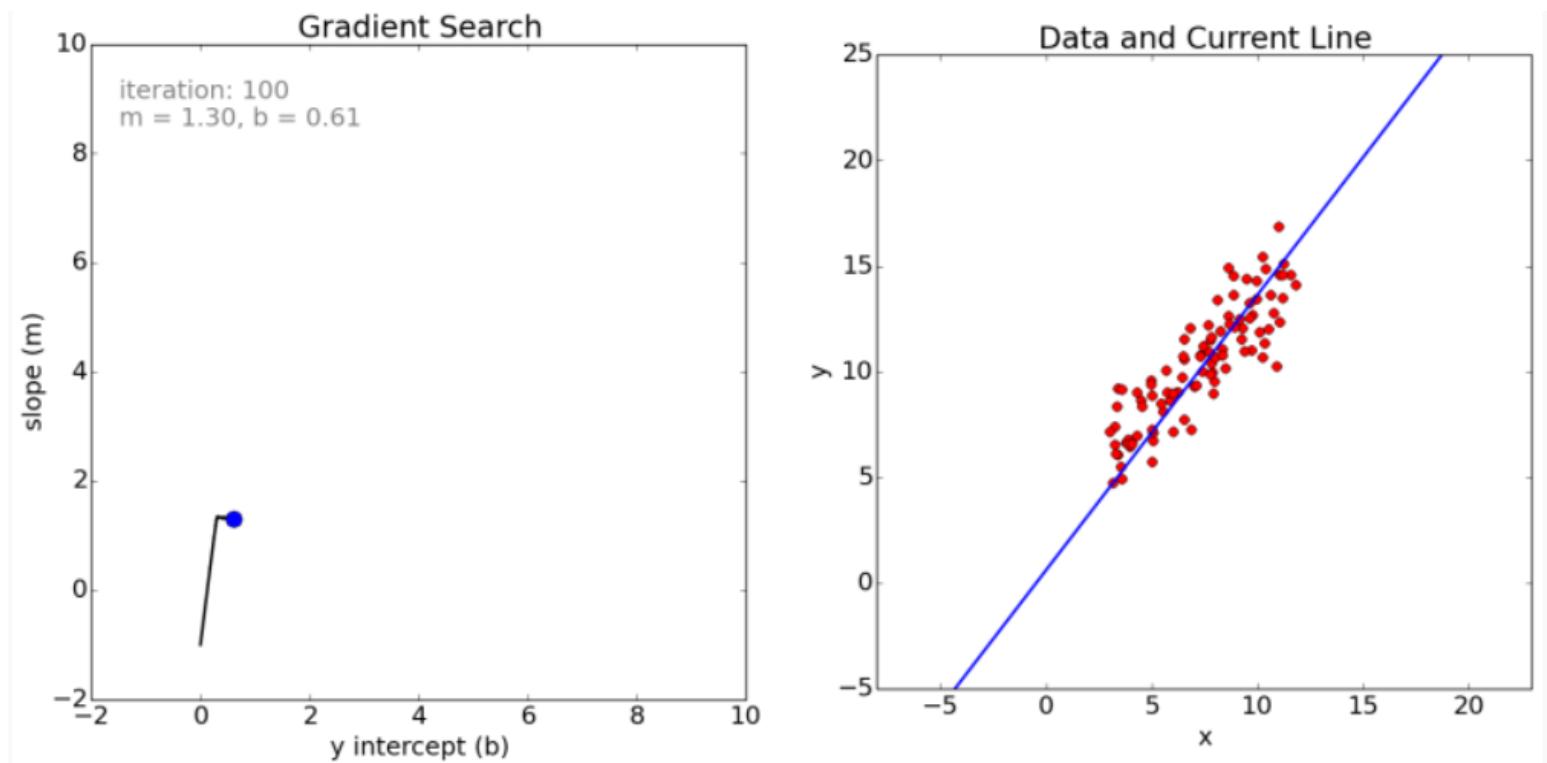
# Парная регрессия



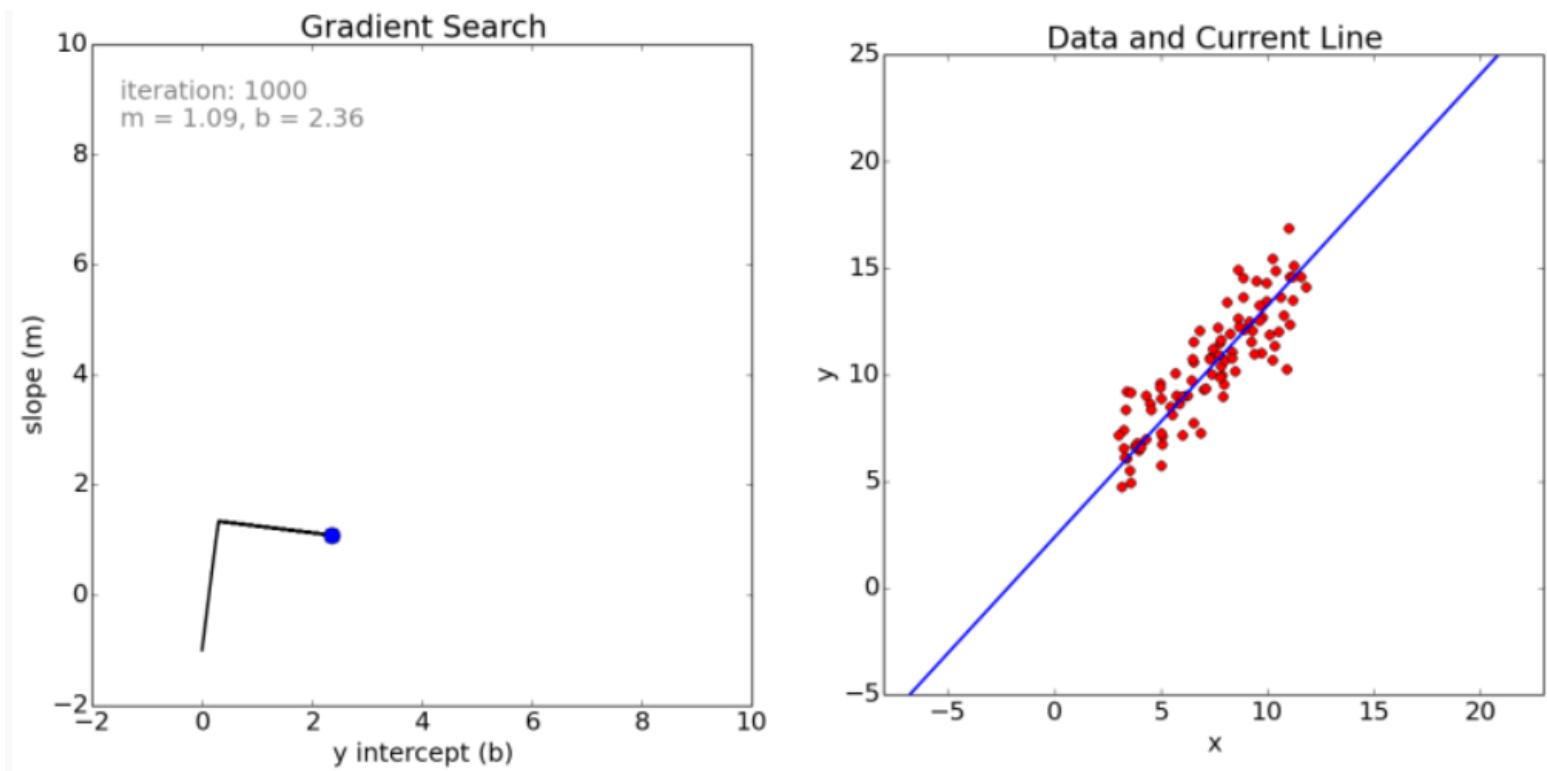
# Парная регрессия



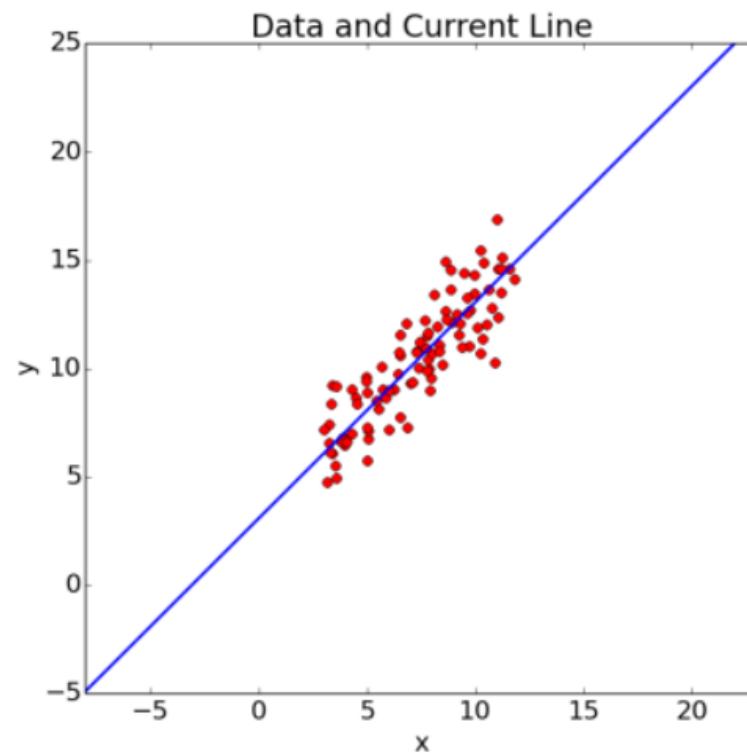
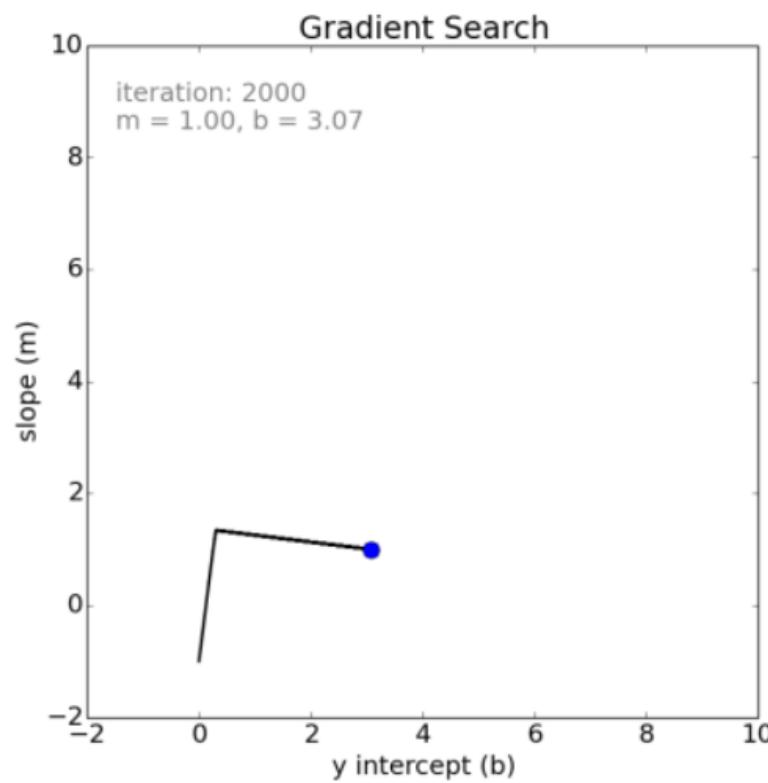
# Парная регрессия

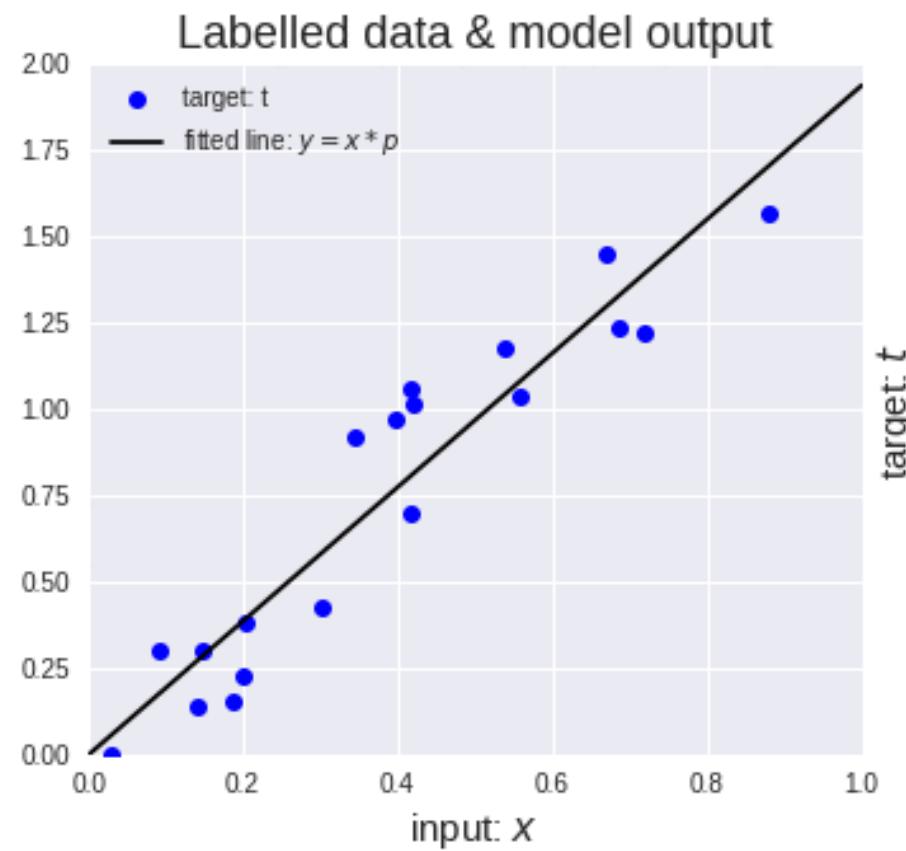
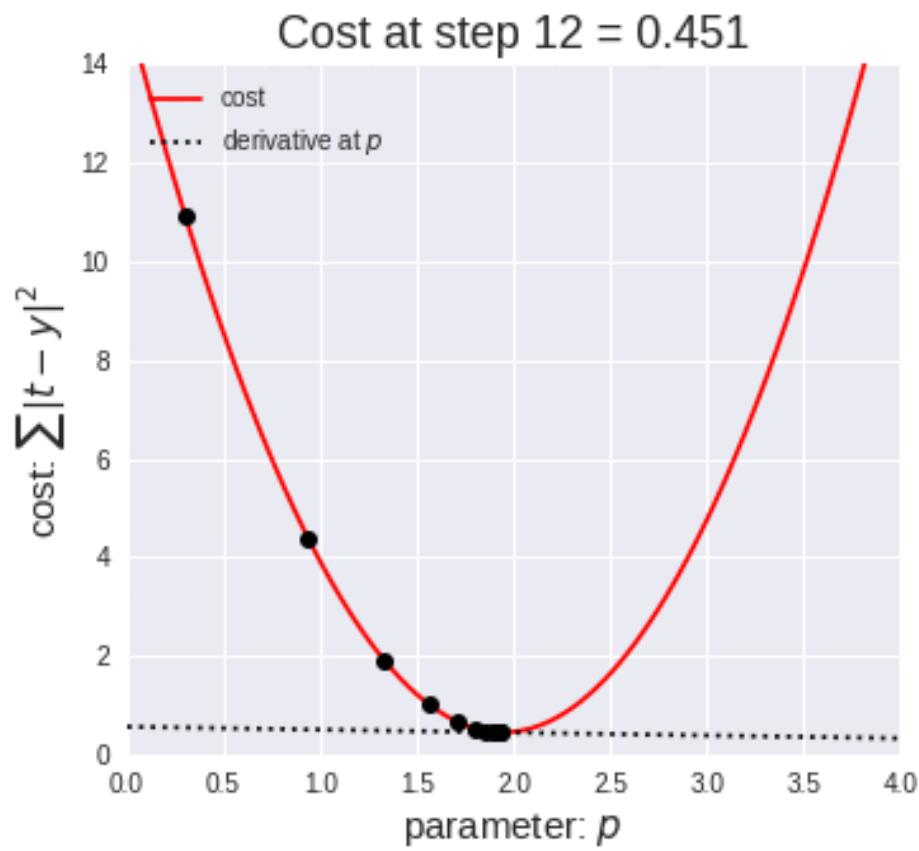


# Парная регрессия

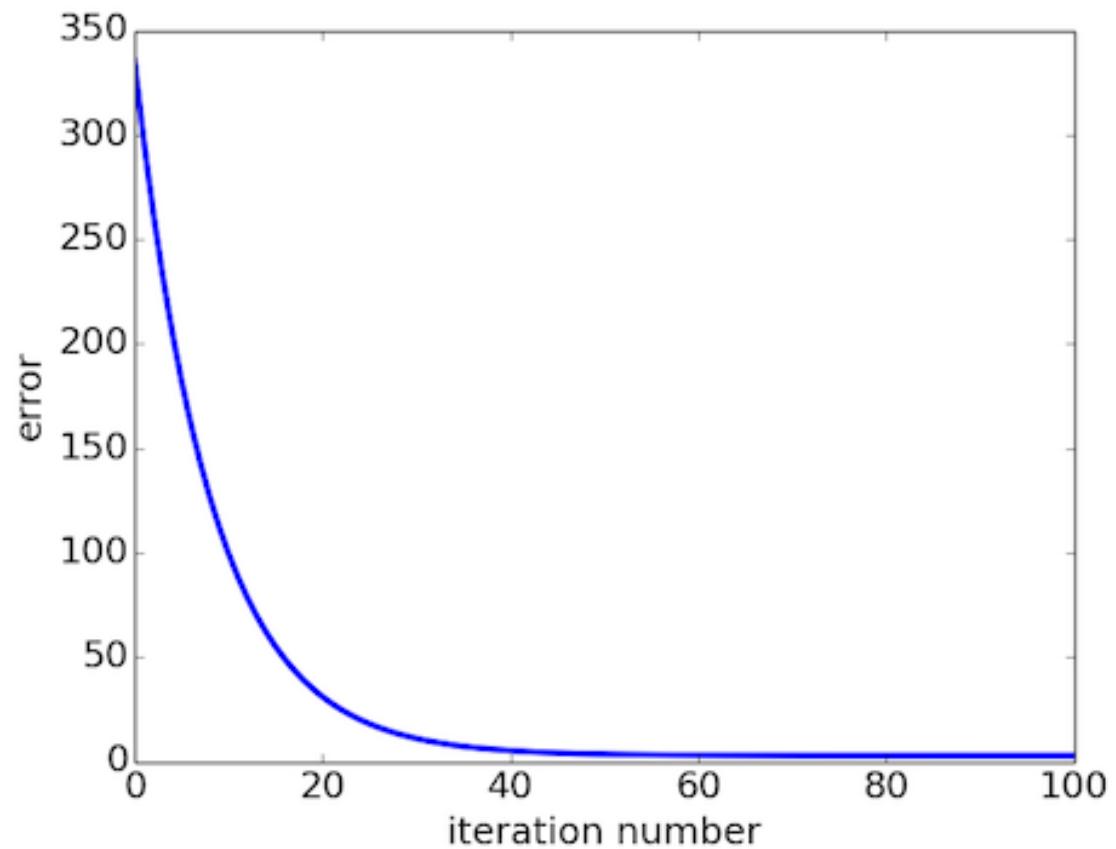


# Парная регрессия





# Функционал ошибки



# Линейная регрессия

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

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{i1} (\langle w, x \rangle - y_i)$
- ...
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{id} (\langle w, x \rangle - y_i)$
- $\nabla Q(w) = \frac{2}{\ell} X^T (Xw - y)$

# Градиентный спуск

1. Начальное приближение:  $w^0$

2. Повторять:

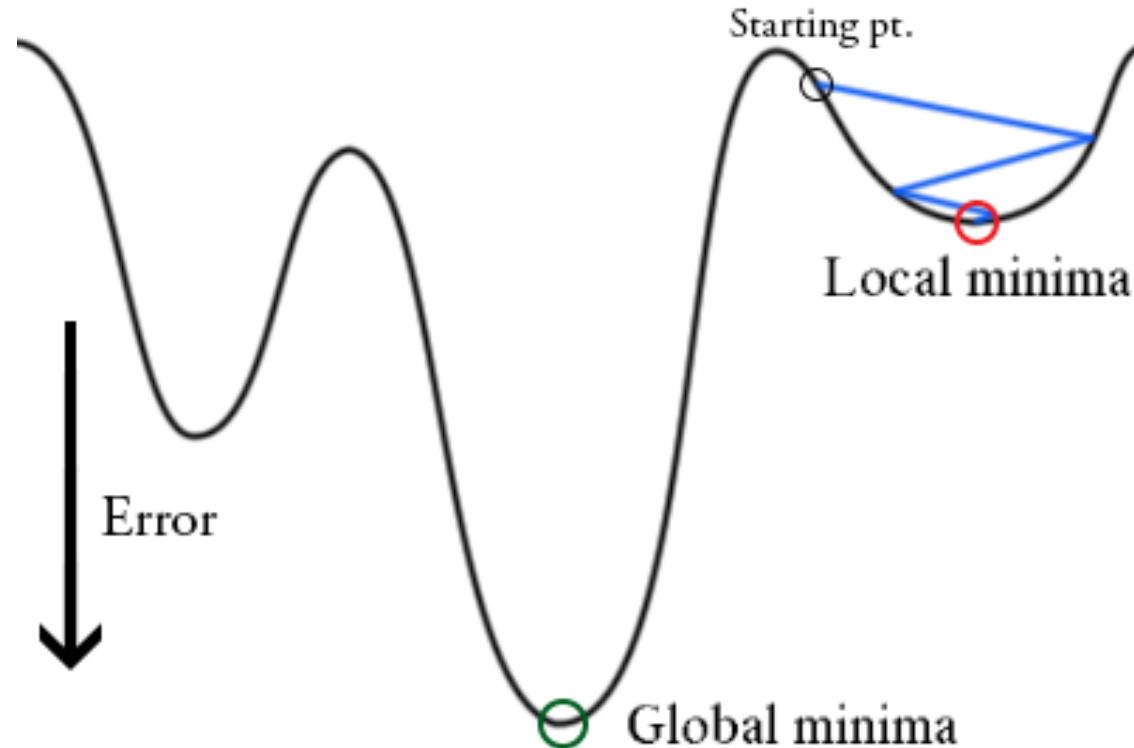
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

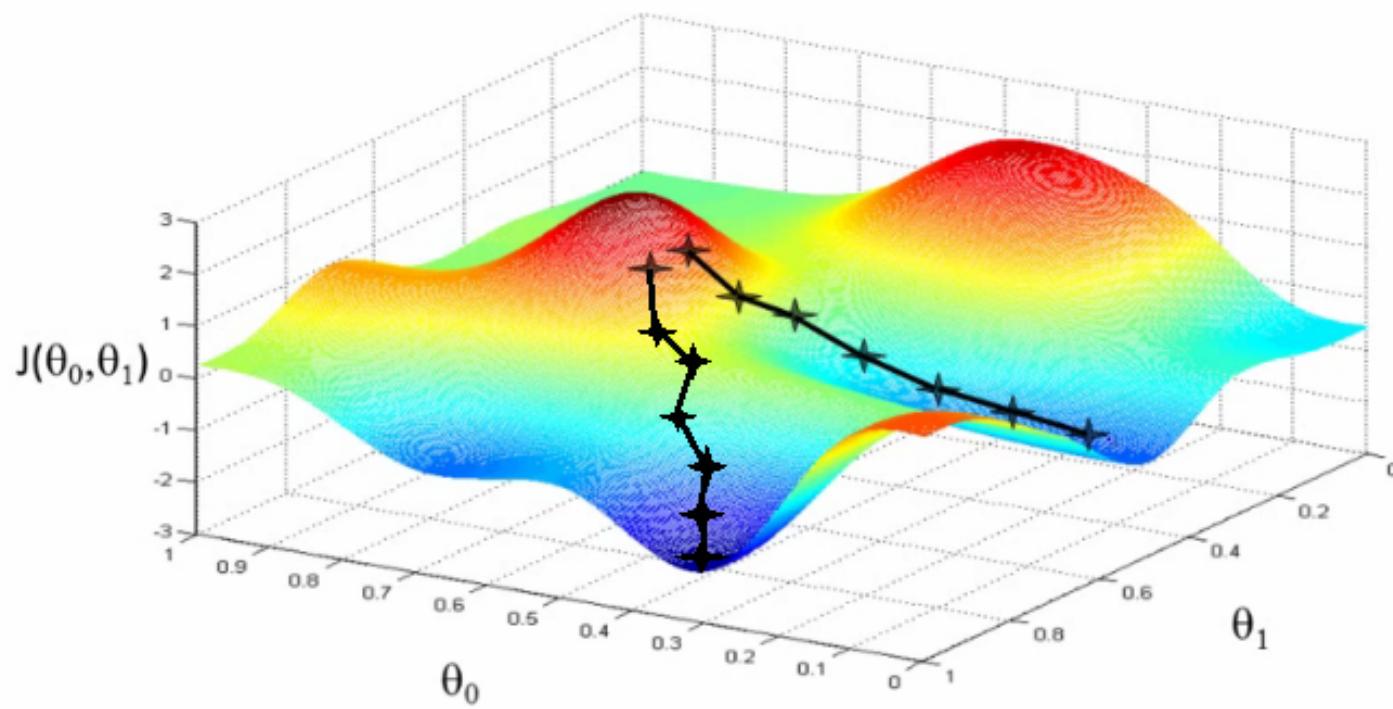
$$\|w^t - w^{t-1}\| < \varepsilon$$

# Локальные минимумы

- Градиентный спуск находит только локальные минимумы



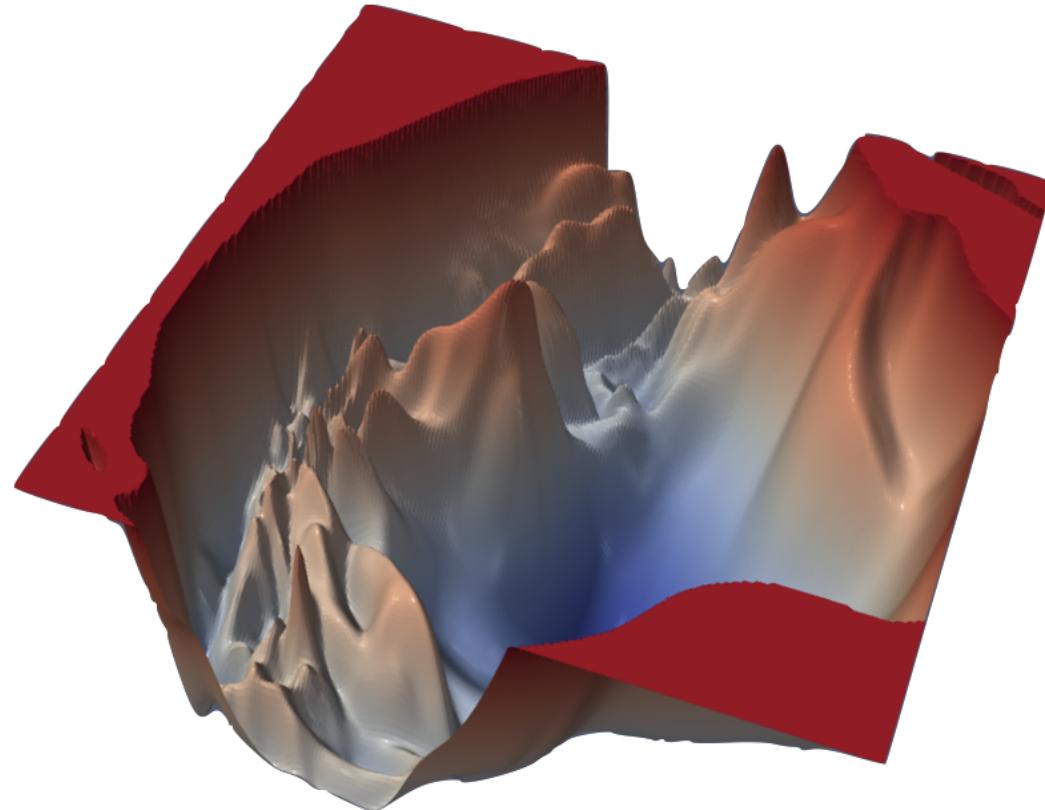
# Локальные минимумы



# Локальные минимумы

- Градиентный спуск находит **локальный минимум**
- Мультистарт — запуск градиентного спуска из разных начальных точек
- Может улучшить результат

# Локальные минимумы



# Длина шага

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Позволяет контролировать скорость обучения

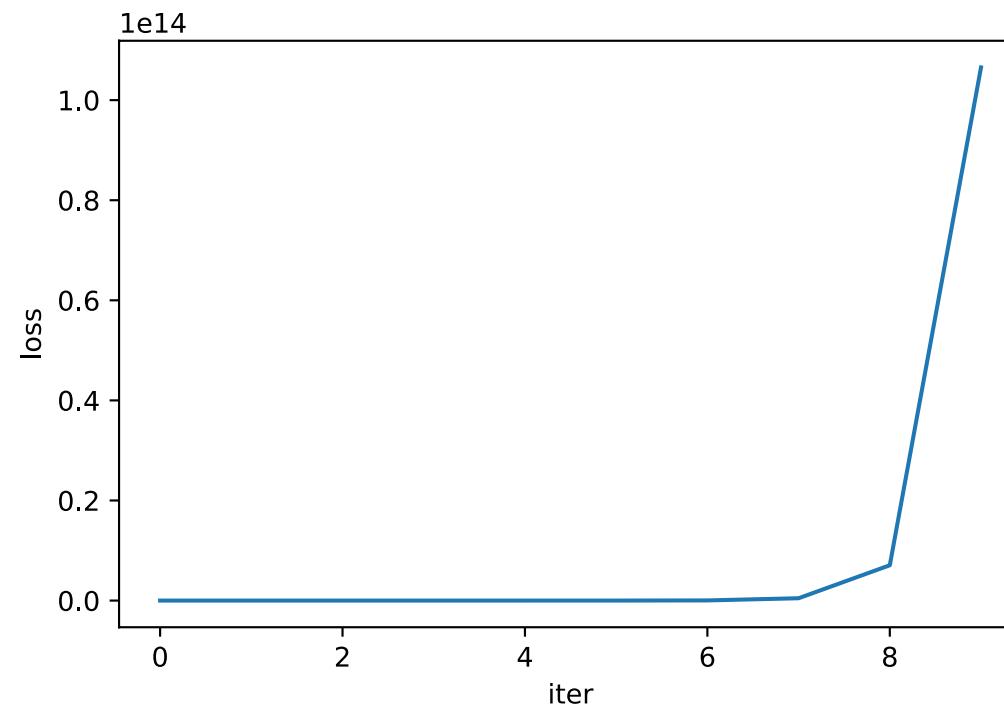
# Длина шага

```
[[ 0.8194022 -11.97609413 -34.41655678  0.98167246 -34.14405489]
 [ -2.83614512  17.19489715   3.29562399  63.8054227  39.70301275]
 [  3.10906179  11.26049837   0.51404712  22.64032379 -28.62078735]
 ...,
 [ -3.61976507  17.63933655  31.65890573  22.5124188 -75.6386039 ]
 [ -1.98472285  3.98588887  29.6135414 -11.11816  33.98746403]
 [ -3.34136103 -12.81955782 -19.5542601  12.62435442  50.24876879]]
```

# Длина шага

Градиент на первом шаге:

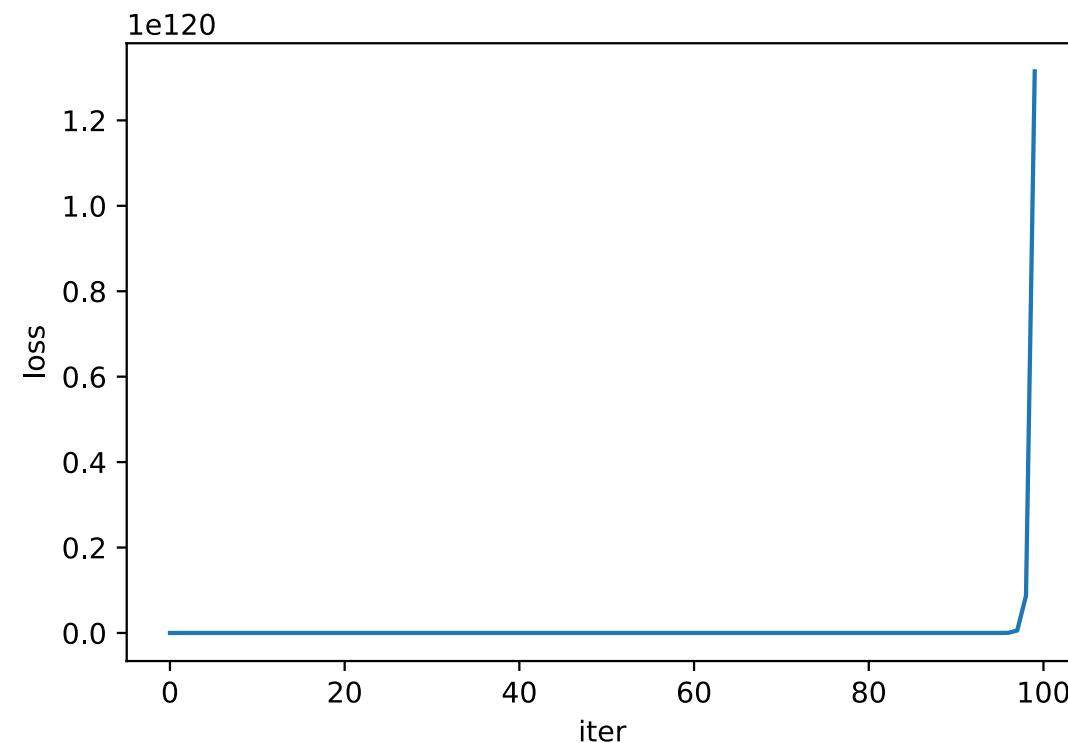
[ 26.52, 564.80, 682.90, 5097.71, 12110.87]



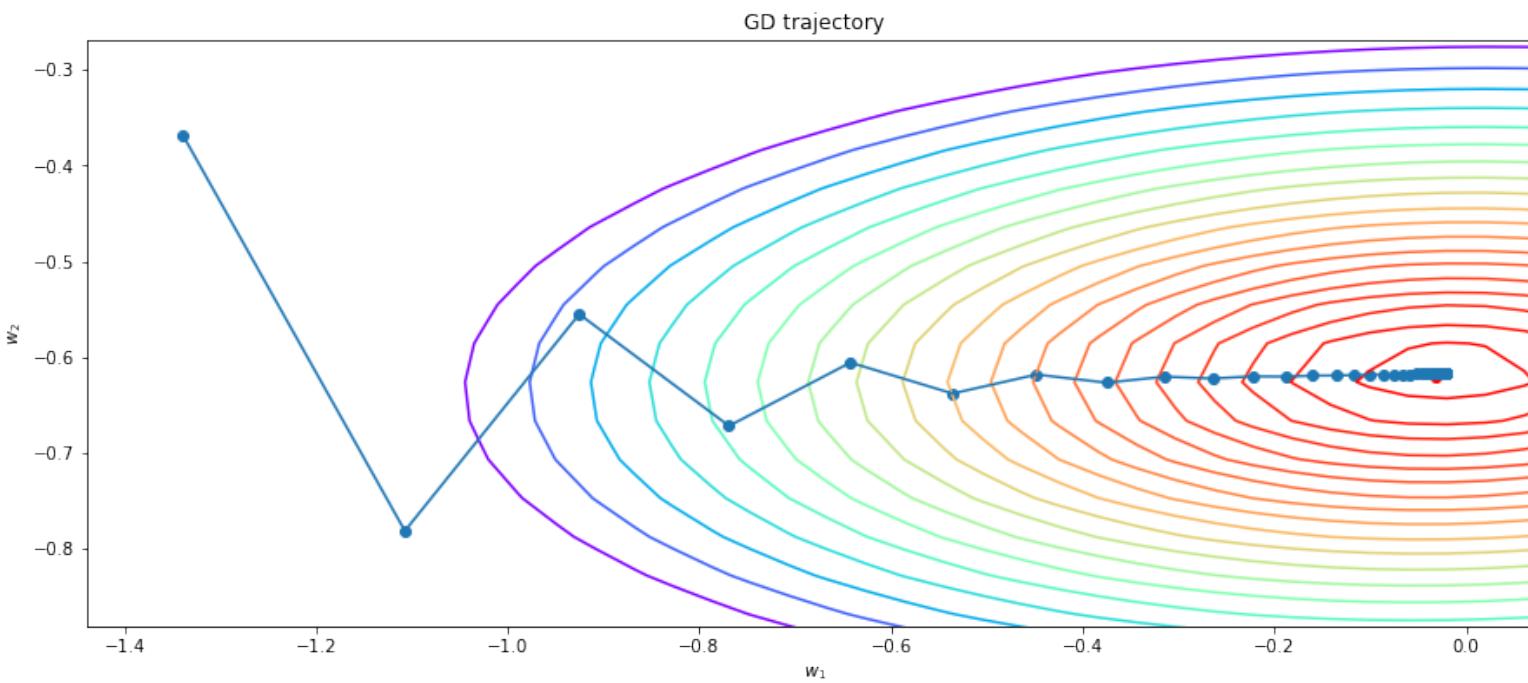
# Длина шага

Градиент на первом шаге:

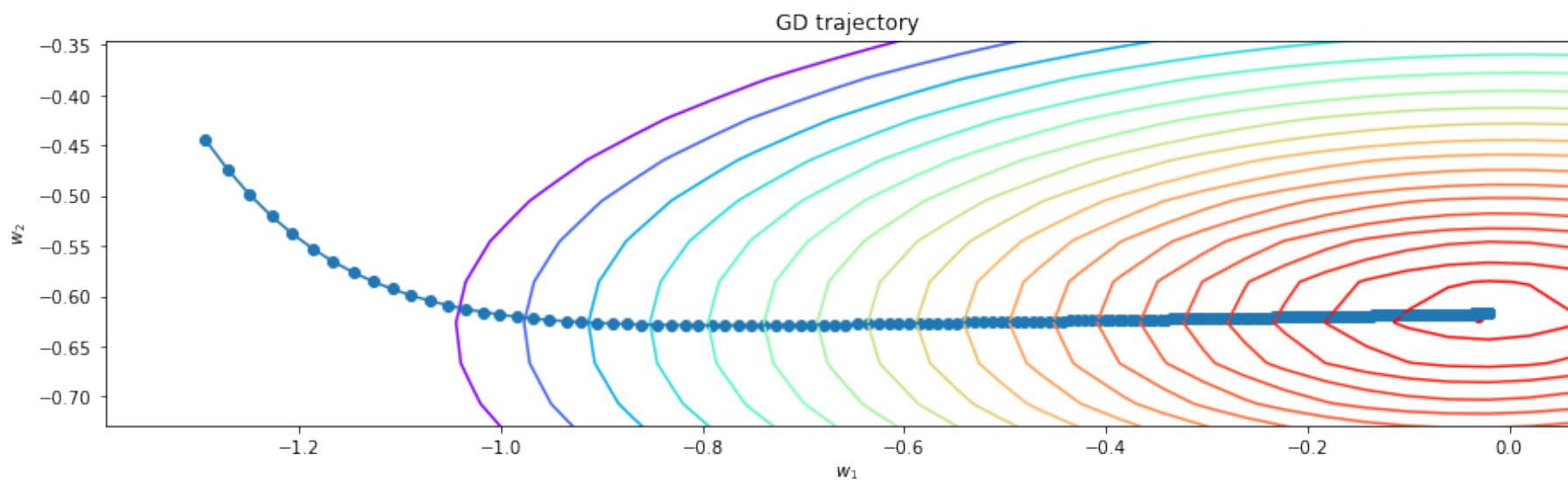
[ 26.52, 564.80, 682.90, 5097.71, 12110.87]



# Длина шага



# Длина шага



# Длина шага

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Позволяет контролировать скорость обучения
- Если сделать длину шага недостаточно маленькой, градиентный спуск может разойтись
- Длина шага — параметр, который нужно подбирать

# Переменная длина шага

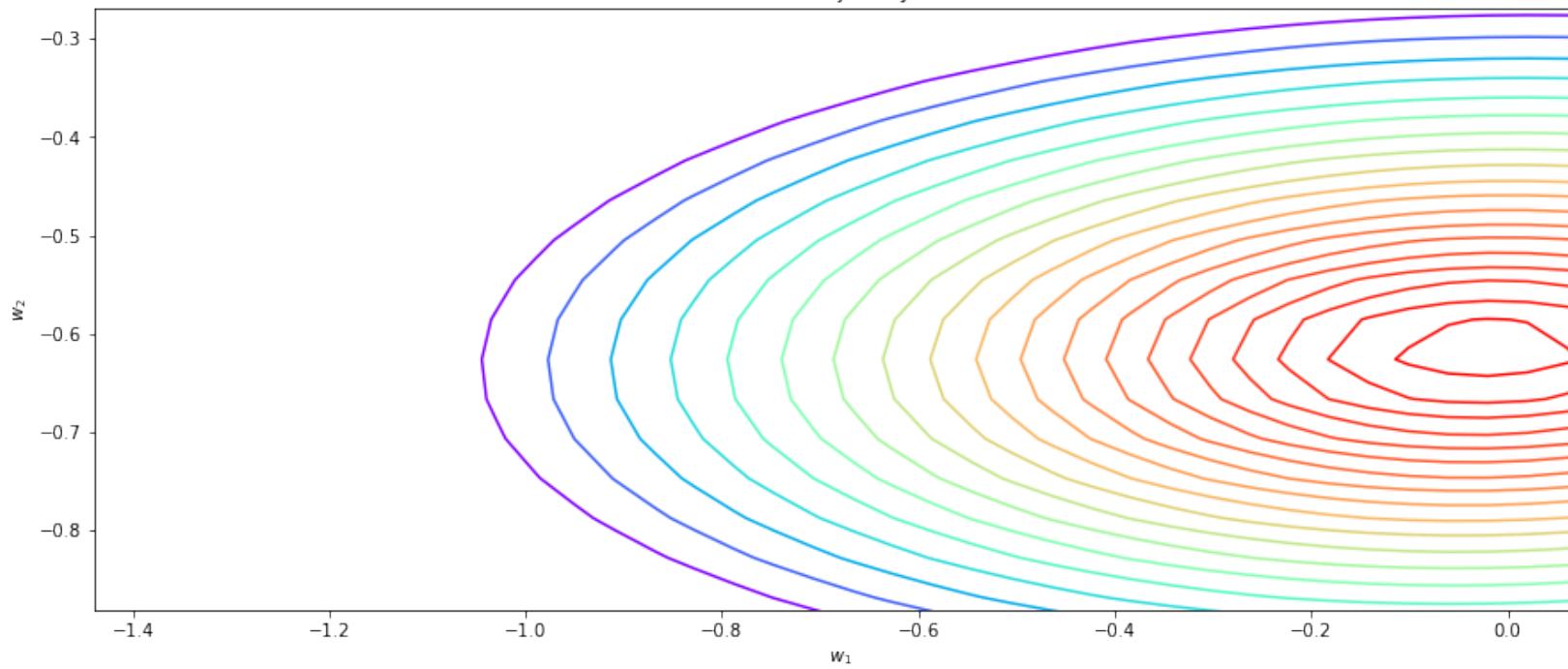
$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

- Длину шага можно менять в зависимости от шага
- Например:  $\eta_t = \frac{1}{t}$
- Ещё вариант:  $\eta_t = \lambda \left( \frac{s}{s+t} \right)^p$

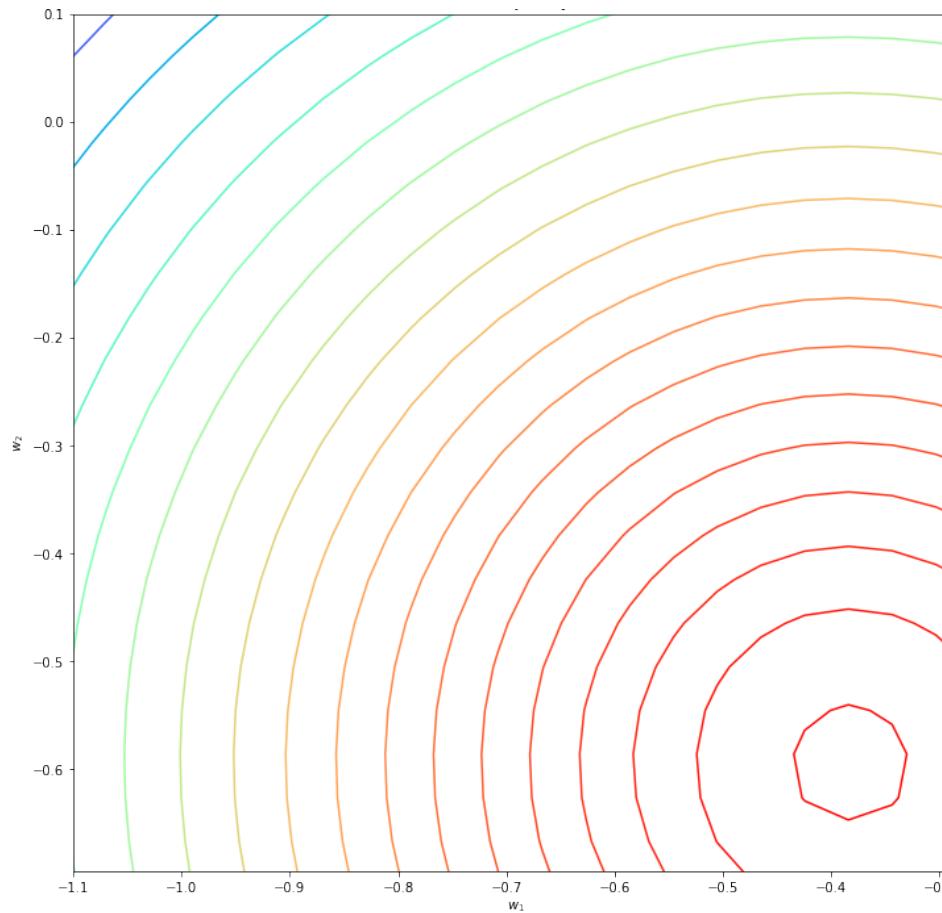
# Масштабирование признаков

```
[[  0.8194022 -11.97609413 -34.41655678   0.98167246 -34.14405489 ]
 [ -2.83614512  17.19489715   3.29562399  63.8054227  39.70301275 ]
 [  3.10906179  11.26049837   0.51404712  22.64032379 -28.62078735 ]
 [ ...,
 [ -3.61976507  17.63933655  31.65890573  22.5124188 -75.6386039  ]
 [ -1.98472285  3.98588887  29.6135414   -11.11816  33.98746403 ]
 [ -3.34136103 -12.81955782 -19.5542601   12.62435442  50.24876879 ]]
```

# Масштабирование признаков



# Масштабирование признаков



# Масштабирование признаков

- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

# Стохастический градиентный спуск

# Градиентный спуск

1. Начальное приближение:  $w^0$

2. Повторять:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

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

# Линейная регрессия

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

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{i1} (\langle w, x \rangle - y_i)$
- ...
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_{id} (\langle w, x \rangle - y_i)$
- $\nabla Q(w) = \frac{2}{\ell} X^T (Xw - y)$

# Сложности градиентного спуска

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

# Оценка градиента

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

- Градиент:

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

- Может, оценить градиент одним слагаемым?

$$\nabla Q(w) \approx \nabla L(y_i, a(x_i))$$

# Стохастический градиентный спуск

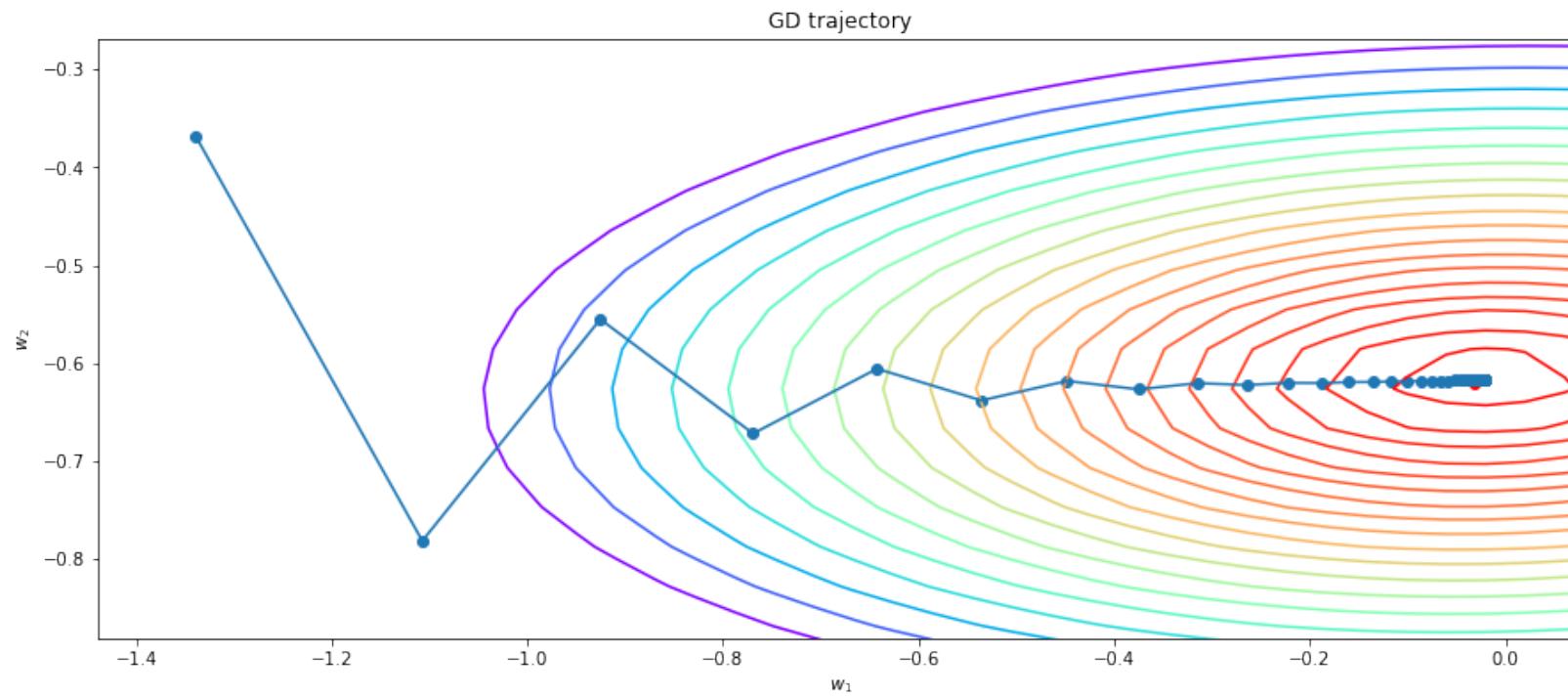
1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

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

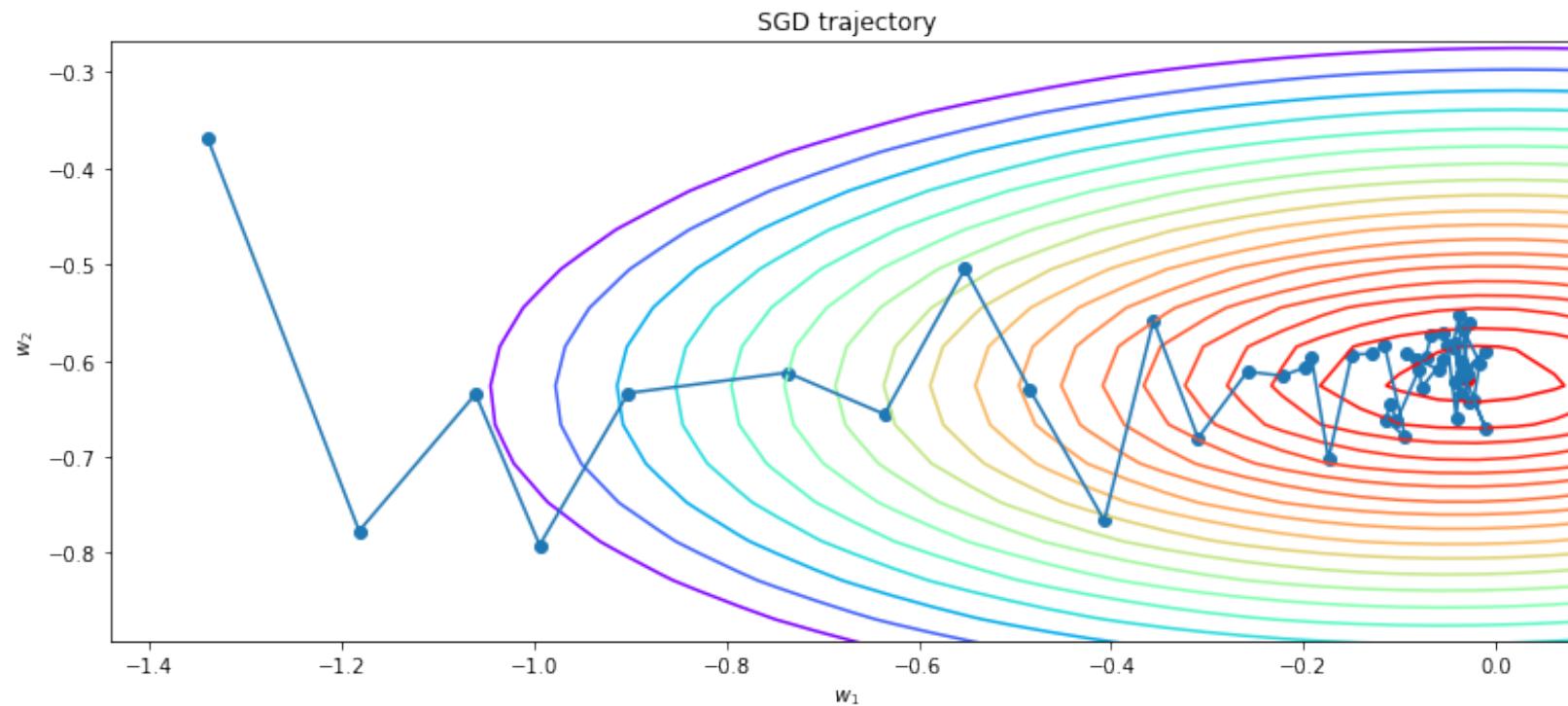
3. Останавливаемся, если

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

# Градиентный спуск



# Стochastic gradient descent



# Стохастический градиентный спуск

1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая случайный объект  $i_t$ :

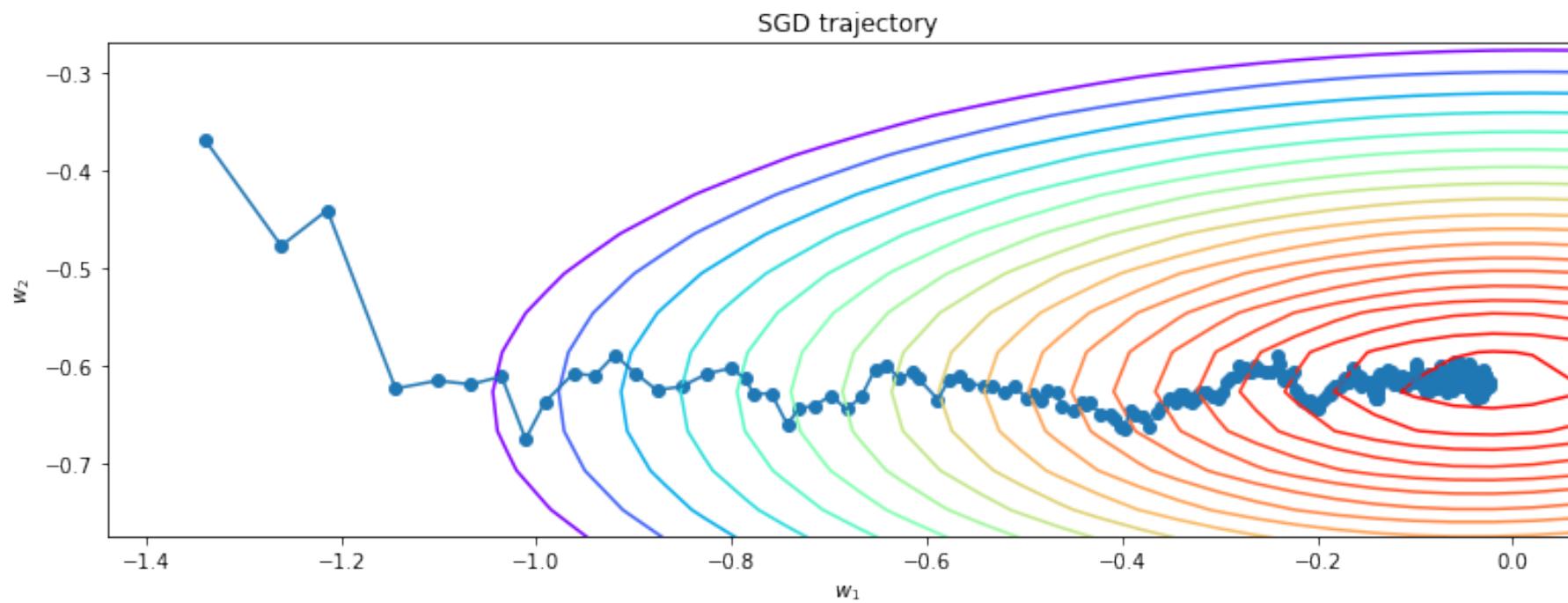
$$w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

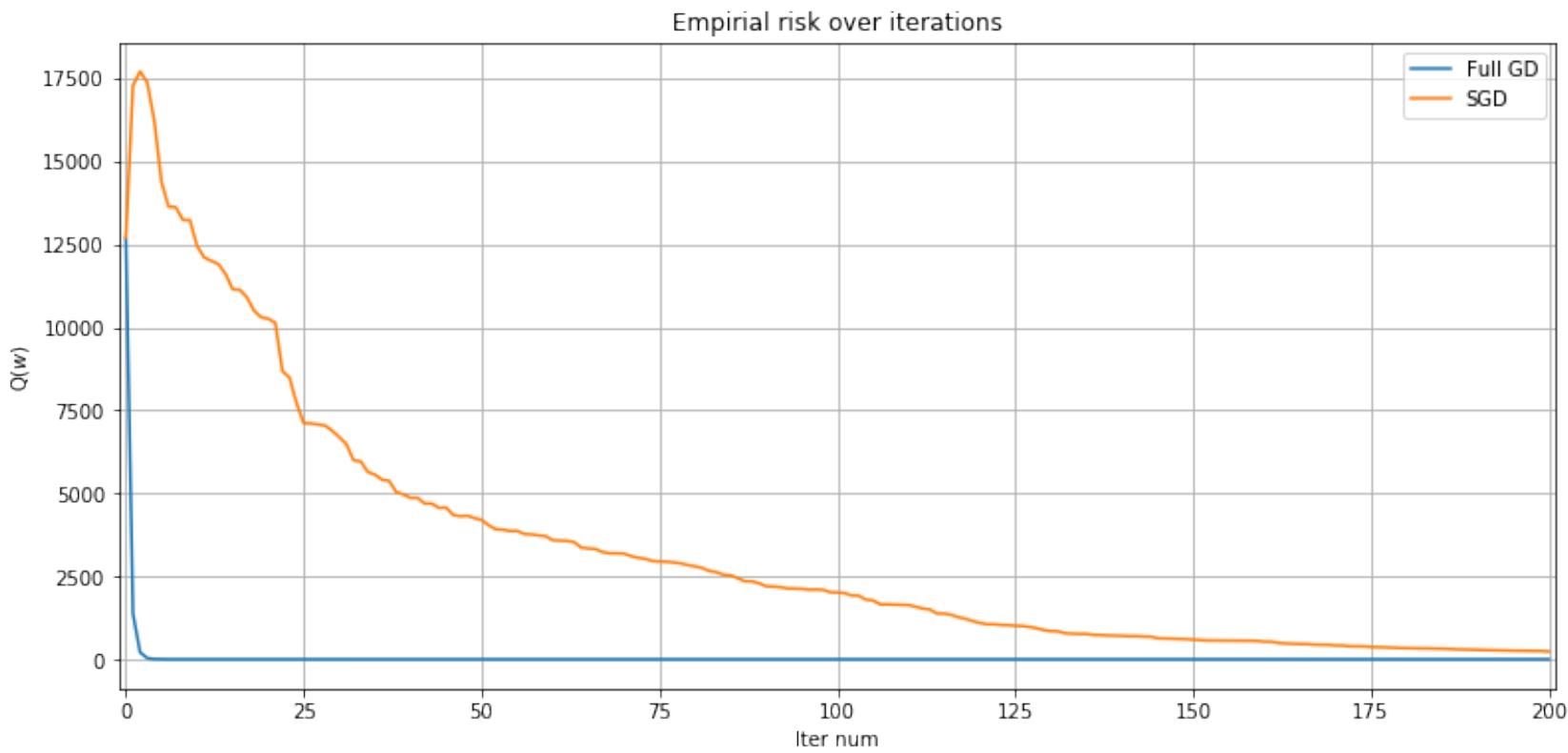
$$\|w^t - w^{t-1}\| < \varepsilon$$

# Стochastic gradient descent

$$\eta_t = \frac{0.1}{t^{0.3}}$$



# Стochastic gradient descent



# Mini-batch

1. Начальное приближение:  $w^0$
2. Повторять, каждый раз выбирая  $m$  случайных объектов  $i_1, \dots, i_m$ :

$$w^t = w^{t-1} - \eta_t \frac{1}{m} \sum_{j=1}^m \nabla L \left( y_{i_j}, a(x_{i_j}) \right)$$

3. Останавливаемся, если

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

# Функции потерь в задачах регрессии

# Среднеквадратичная ошибка

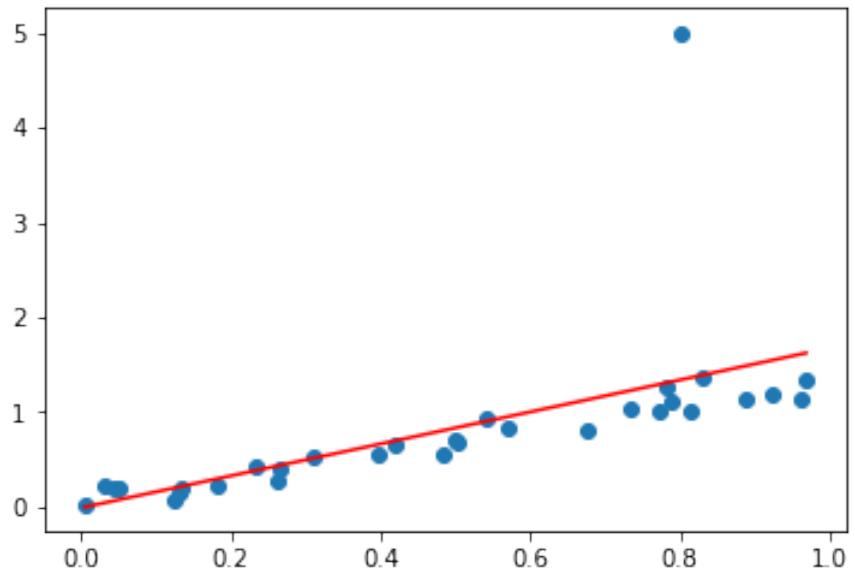
- Частый выбор — квадратичная функция потерь

$$L(y, a) = (a - y)^2$$

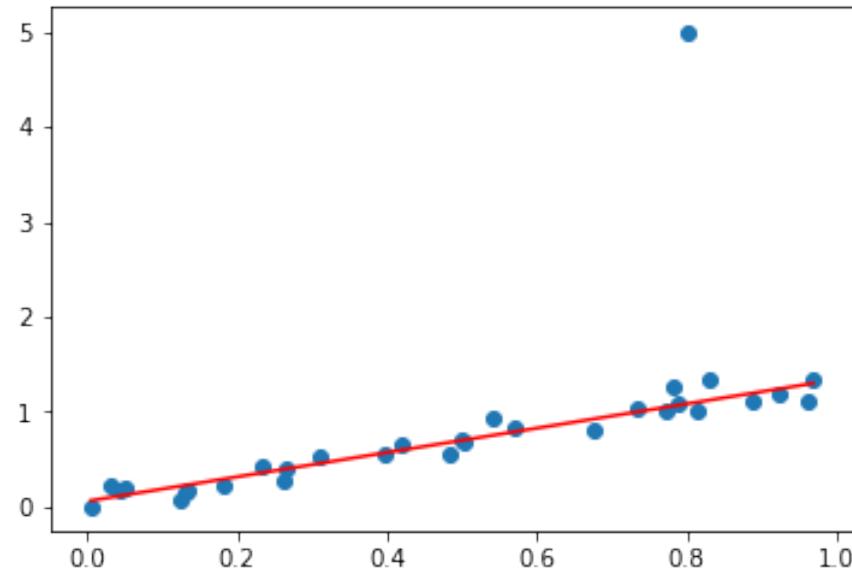
- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

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

# Выбросы



С учётом выброса



Без учёта выброса

Обучение на среднеквадратичную ошибку

# Выбросы

$a(x)$	$y$	$(a(x) - y)^2$
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	8649
6	7	1

$$MSE \approx 1236$$

# Выбросы

$a(x)$	$y$	$(a(x) - y)^2$
4	1	9
5	2	9
6	3	9
7	4	9
8	5	9
10	100	8100
10	7	9

$$MSE \approx 1164$$

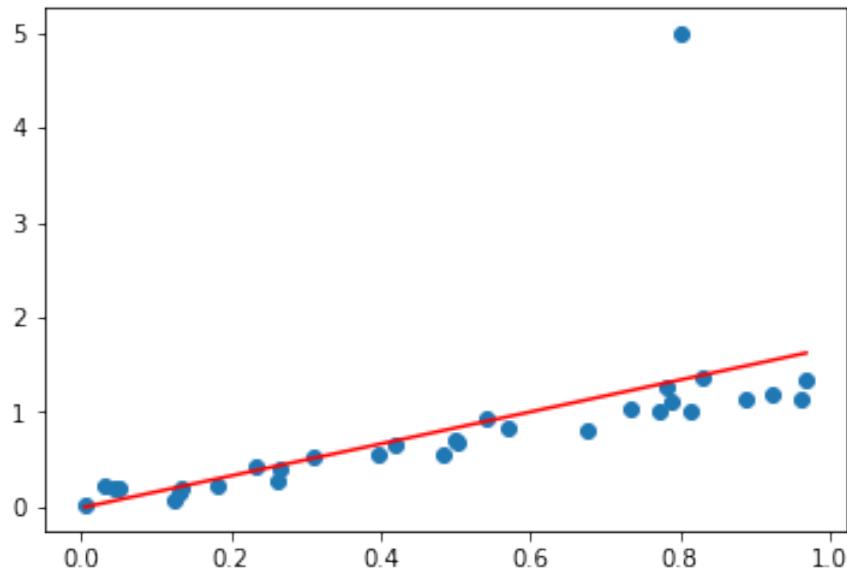
# Средняя абсолютная ошибка

$$L(y, a) = |a - y|$$

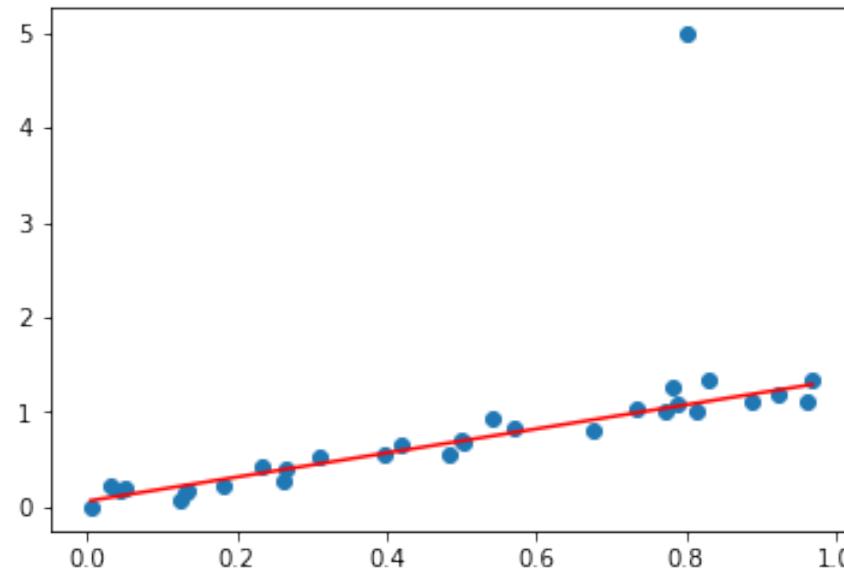
- Функционал ошибки — средняя абсолютная ошибка (mean absolute error, MAE)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

# Выбросы



Обучение на MSE



Обучение на MAE

# Выбросы

$a(x)$	$y$	$ a(x) - y $
2	1	1
1	2	1
2	3	1
5	4	1
6	5	1
7	100	93
6	7	1

$$MAE \approx 14.14$$

# Выбросы

$a(x)$	$y$	$ a(x) - y $
4	1	3
5	2	3
6	3	3
7	4	3
8	5	3
10	100	90
10	7	3

$$MAE \approx 15.43$$

# ФУНКЦИЯ ПОТЕРЬ ХУБЕРА

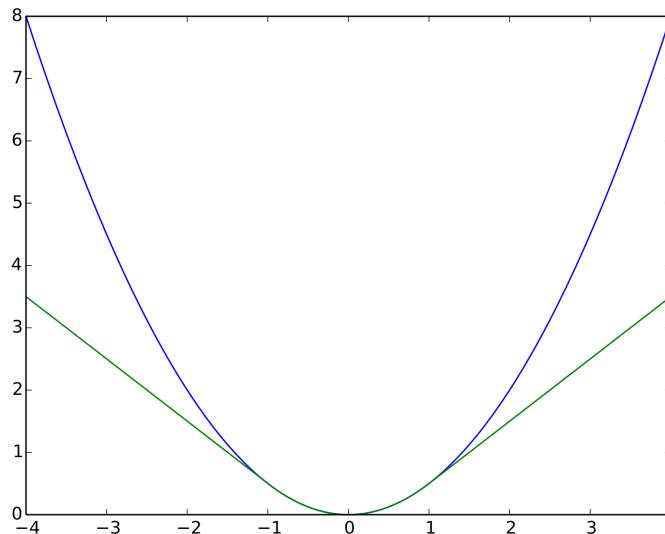
$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left( |y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$

- ФУНКЦИОНАЛ ОШИБКИ:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} L_H(y_i, a(x_i))$$

# Функция потерь Хубера

$$L_H(y, a) = \begin{cases} \frac{1}{2}(y - a)^2, & |y - a| < \delta \\ \delta \left( |y - a| - \frac{1}{2}\delta \right), & |y - a| \geq \delta \end{cases}$$



# MAPE

- Mean Absolute Percentage Error (средний модуль относительной ошибки)

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

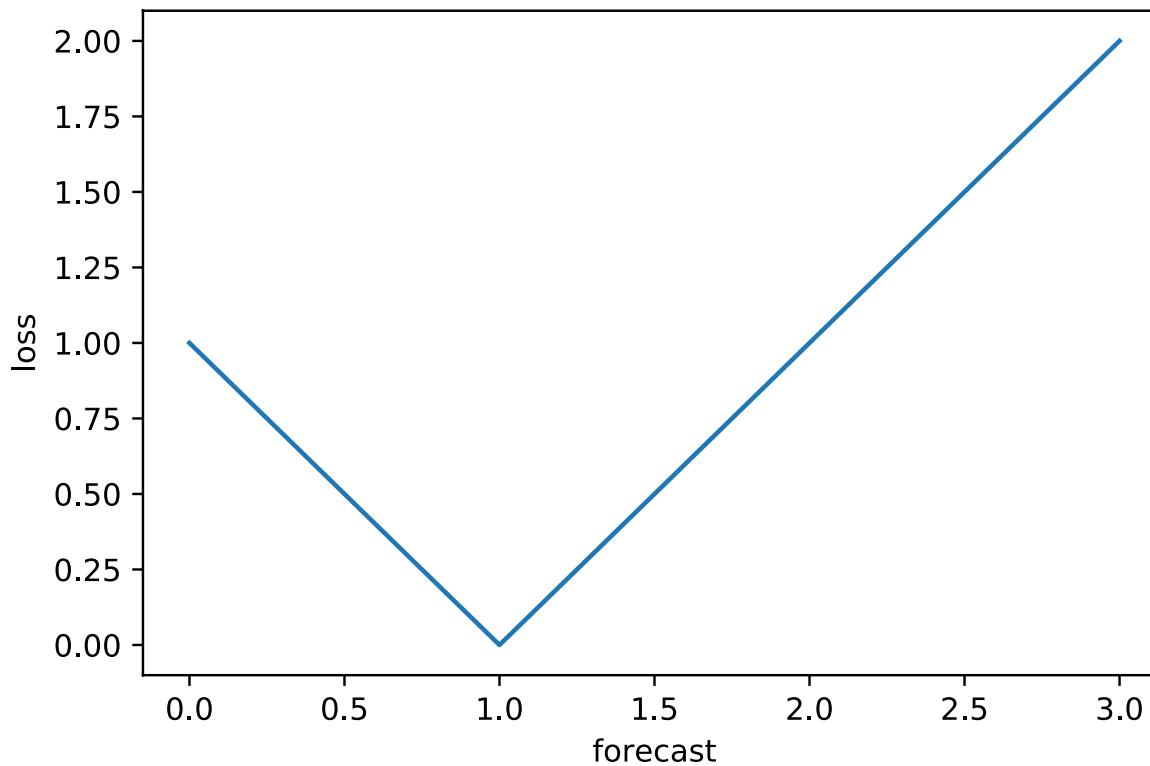
# MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$

- Особенности (при  $a \geq 0$ ):
  - Недопрогноз штрафуется максимум на единицу
  - Перепрогноз может быть оштрафован любым числом
  - Несимметричная функция потерь (отдаёт предпочтение недопрогнозу)

# MAPE

$$L(y, a) = \left| \frac{y - a}{y} \right|$$



# SMAPE

- Symmetric Mean Absolute Percentage Error (симметричный средний модуль относительной ошибки)

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$

$$Q(a, X) = \frac{100\%}{\ell} \sum_{i=1}^{\ell} \frac{|y_i - a(x_i)|}{(|y_i| + |a(x_i)|)/2}$$

# SMAPE

$$L(y, a) = \frac{|y - a|}{(|y| + |a|)/2}$$

