

Основы машинного обучения

Лекция 6

Линейная регрессия и градиентный спуск

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2023

Интерпретация линейных моделей

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 10 * (\text{площадь в кв. см.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

$$\begin{aligned} a(x) = & 100.000 * (\text{площадь в кв. м.}) \\ & + 500.000 * (\text{число магазинов рядом}) \\ & + 100 * (\text{средний доход жильцов дома}) \end{aligned}$$

- Чем больше вес, тем важнее признак?
- Только если признаки масштабированы!

Масштабирование признаков

- Отмасштабируем j -й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

Масштабирование признаков

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

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

Регуляризация

- Если модель переобучается, то веса используются для запоминания обучающей выборки
- Правильнее масштабировать признаки и регуляризовать модель перед изучением весов

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

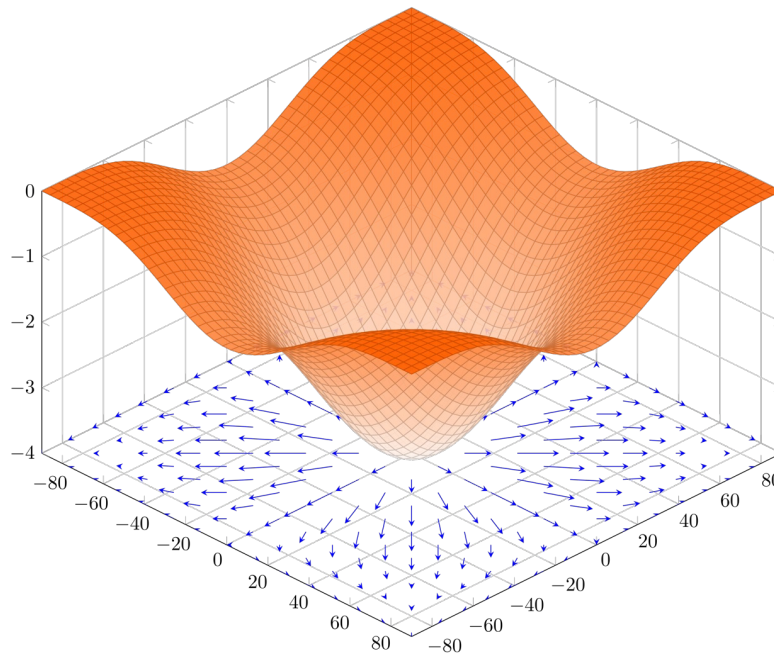
- Градиент — вектор частных производных

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

- У градиента есть важное свойство!

Важное свойство

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



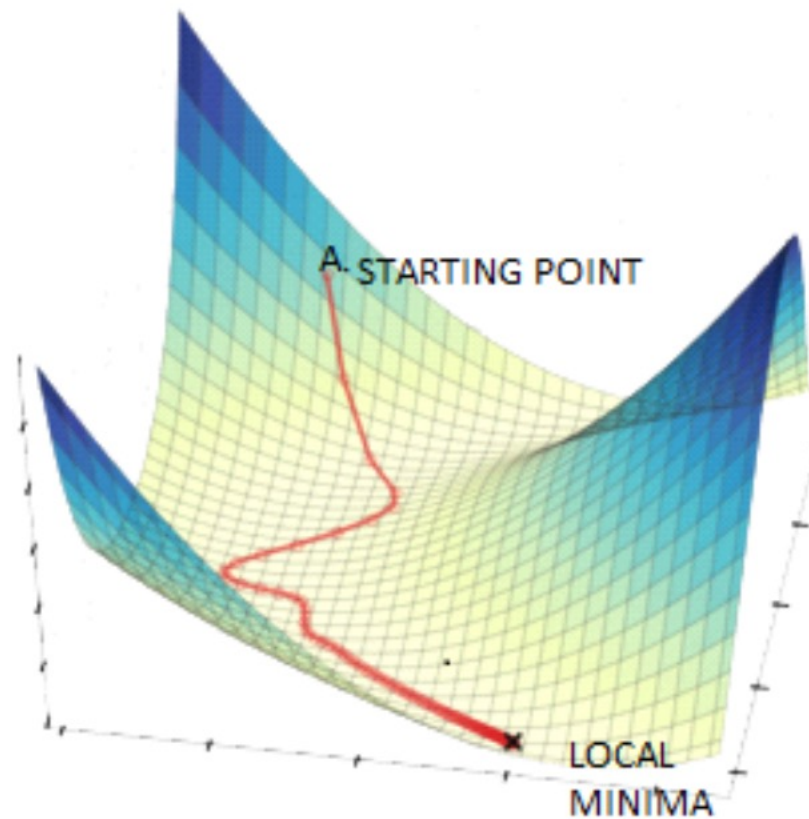
Важное свойство

- 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

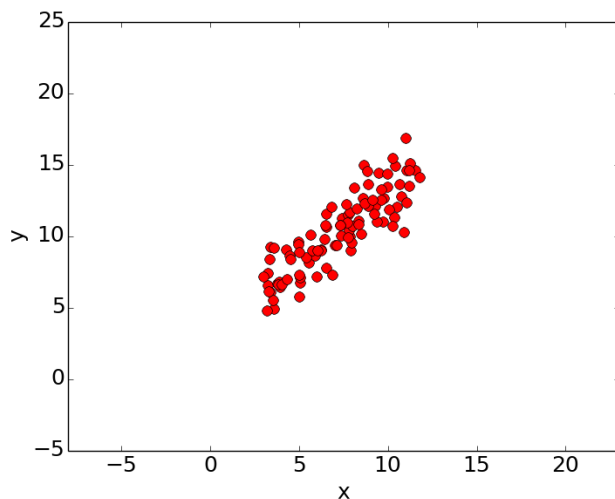
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

Парная регрессия

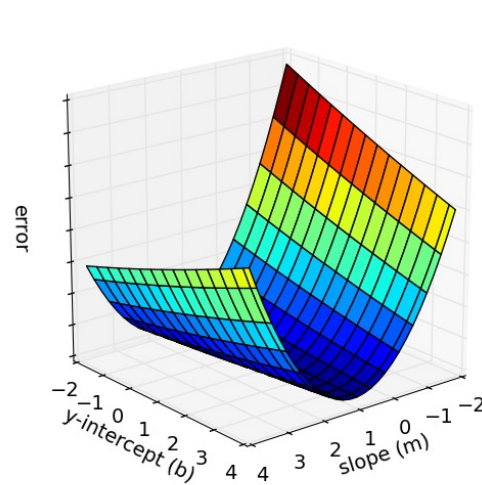
- Простейший случай: один признак
- Модель: $a(x) = w_1x + w_0$
- Два параметра: w_1 и w_0
- Функционал:

$$Q(w_0, w_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1x_i + w_0 - y_i)^2$$

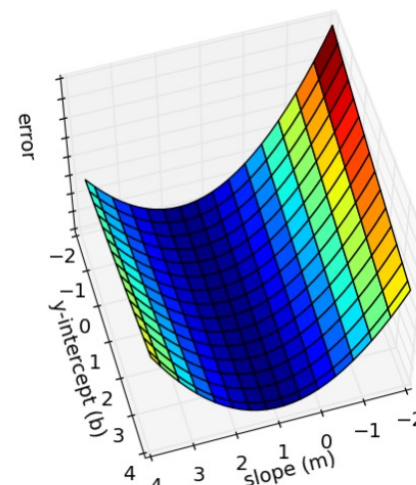
Парная регрессия



Выборка



Функционал ошибки



Парная регрессия

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

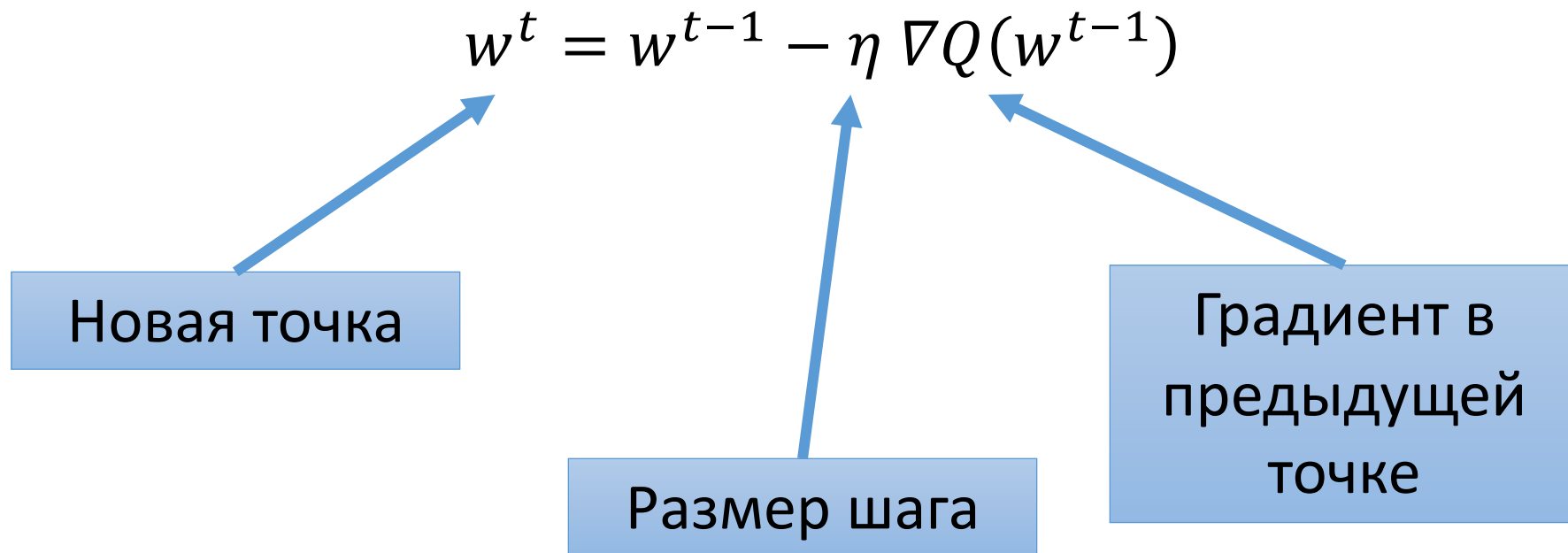
- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i (w_1 x_i + w_0 - y_i)$
- $\frac{\partial Q}{\partial w_0} = \frac{2}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)$
- $\nabla Q(w) = \left(\frac{2}{\ell} \sum_{i=1}^{\ell} x_i (w_1 x_i + w_0 - y_i), \frac{2}{\ell} \sum_{i=1}^{\ell}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

Начальное приближение

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

Градиентный спуск

- Повторять до сходимости:



СХОДИМОСТЬ

- Останавливаем процесс, если

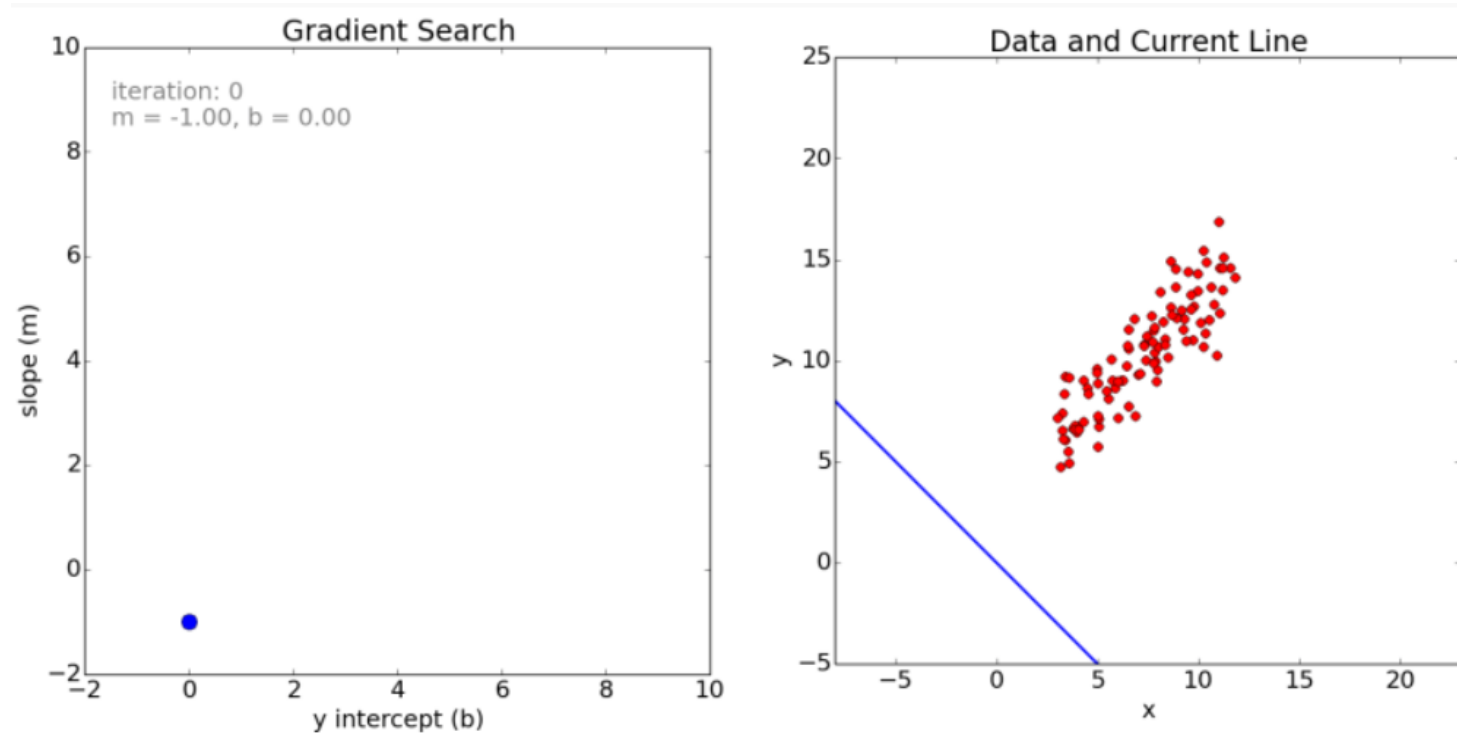
$$\|w^t - w^{t-1}\| < \varepsilon$$

- Другой вариант:

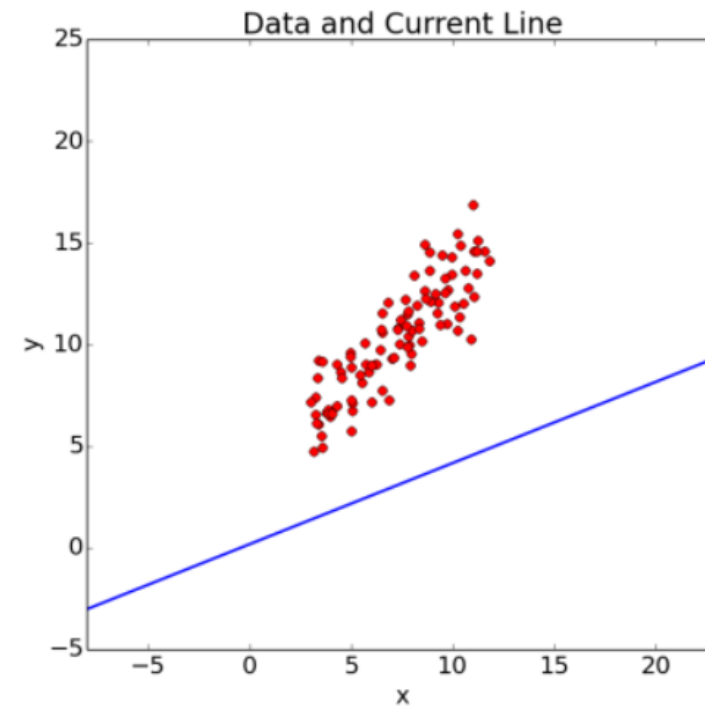
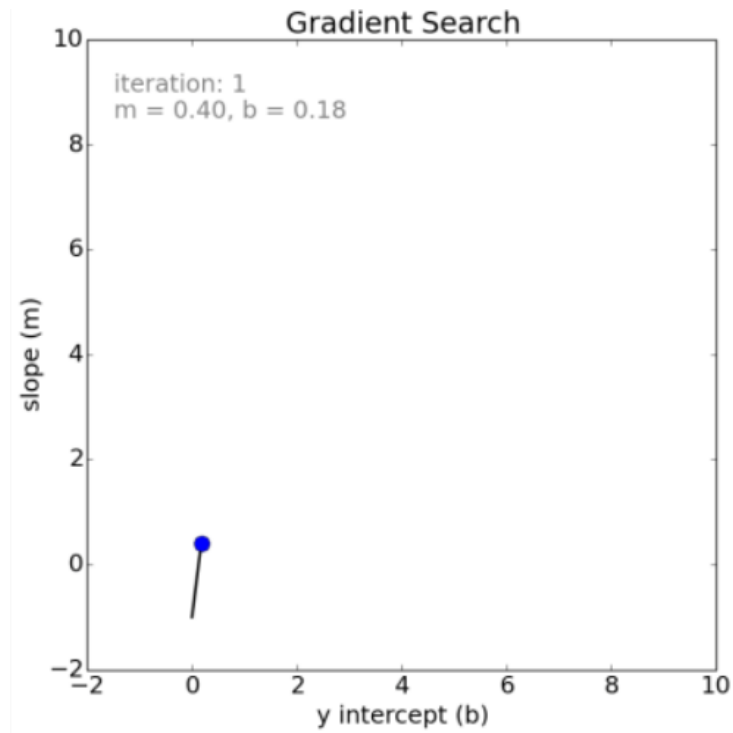
$$\|\nabla Q(w^t)\| < \varepsilon$$

- Или пока ошибка на отложенной выборке уменьшается

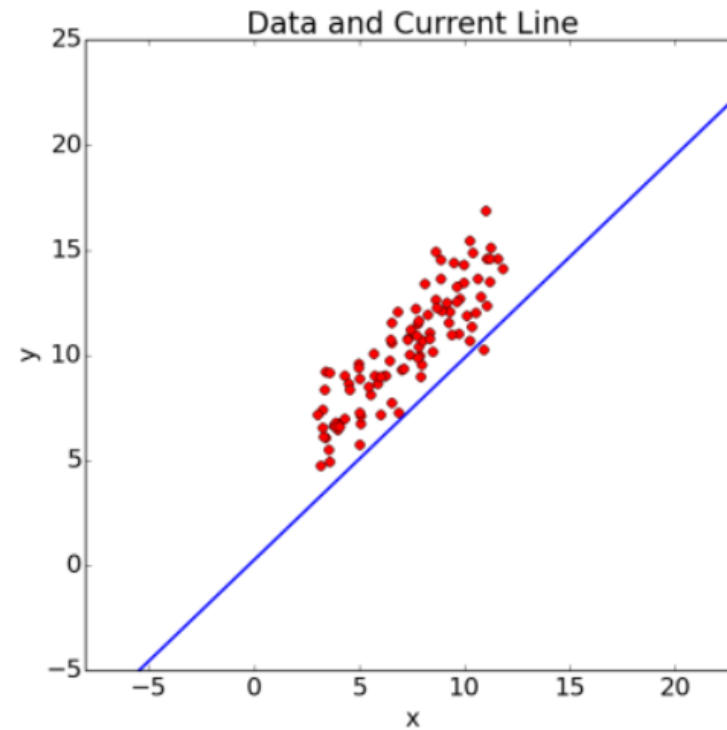
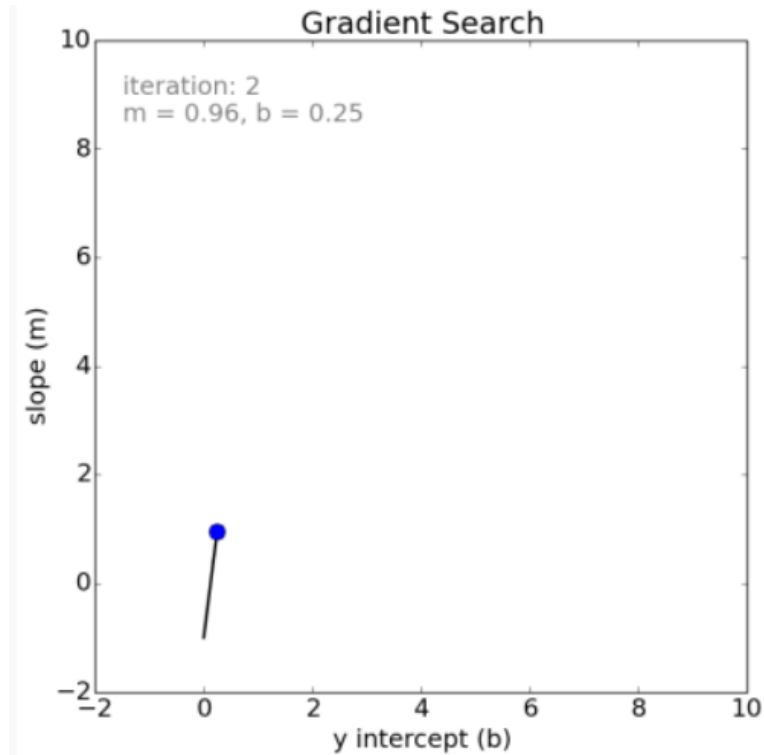
Парная регрессия



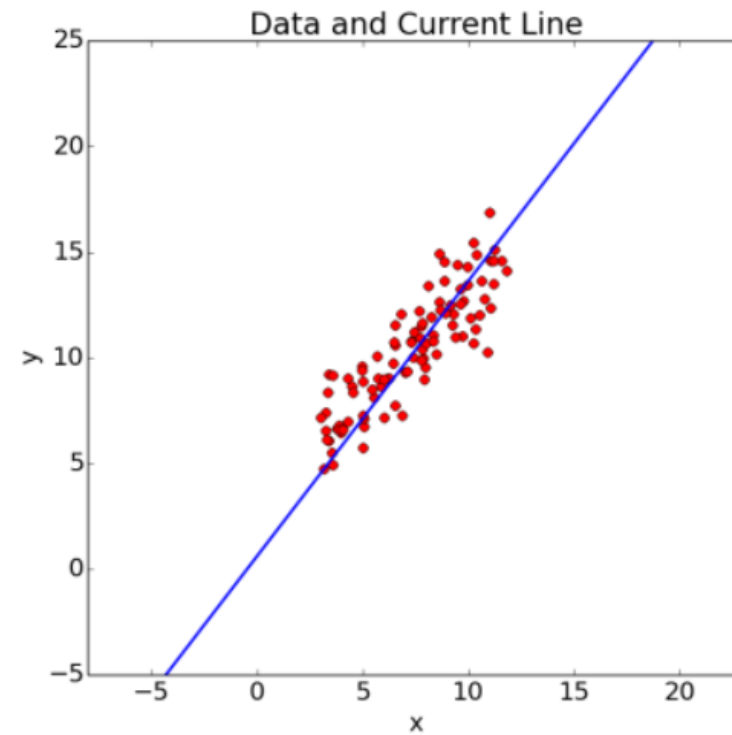
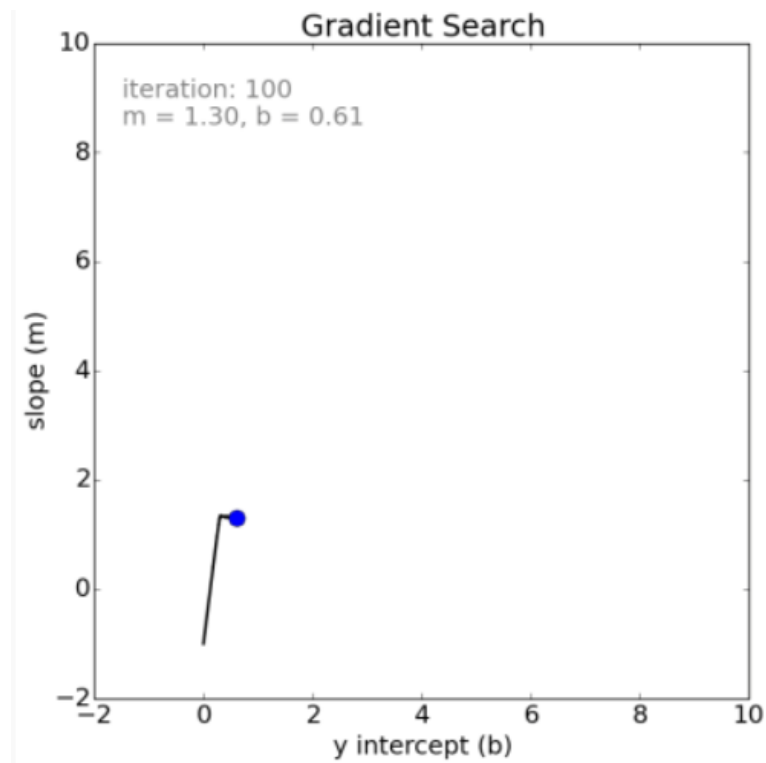
Парная регрессия



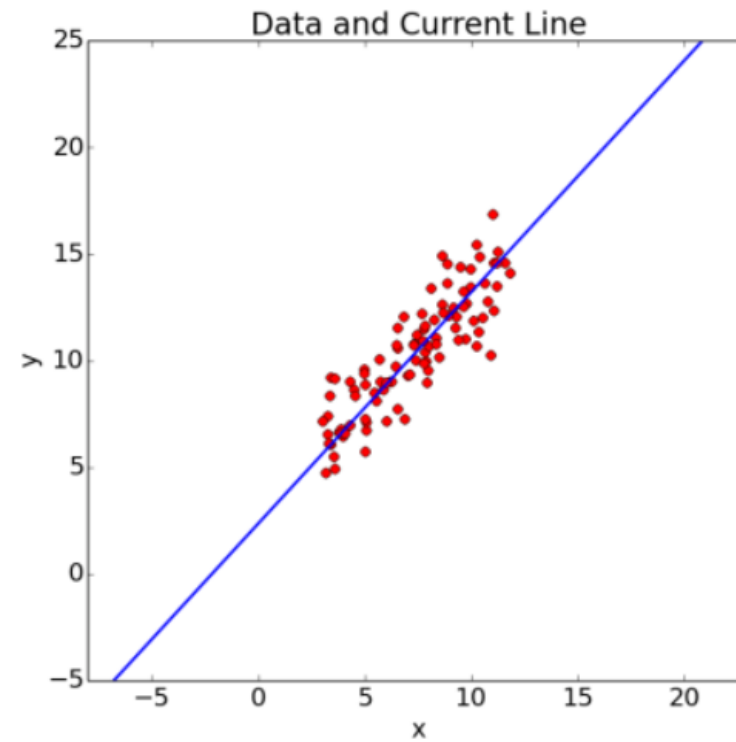
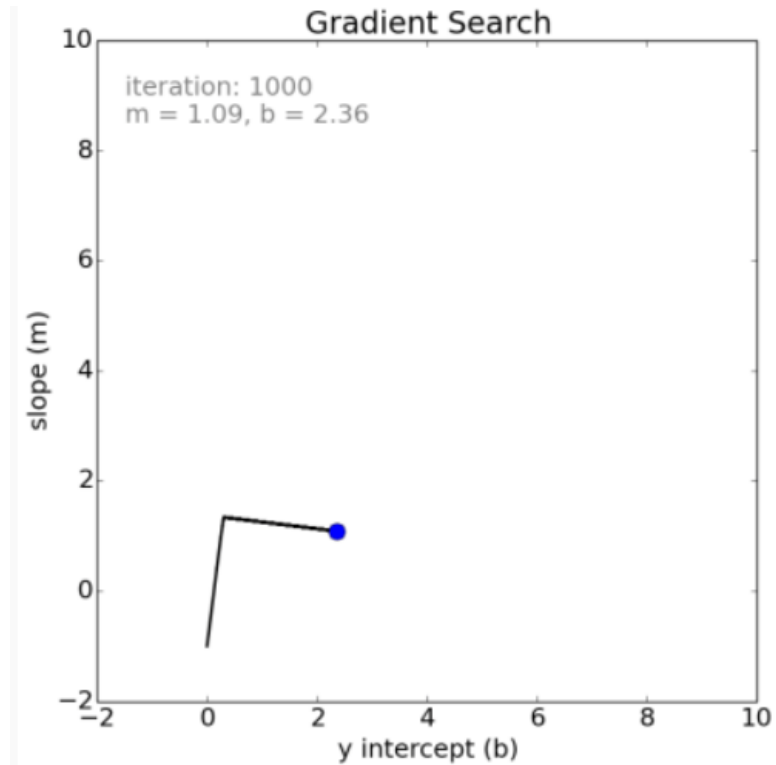
Парная регрессия



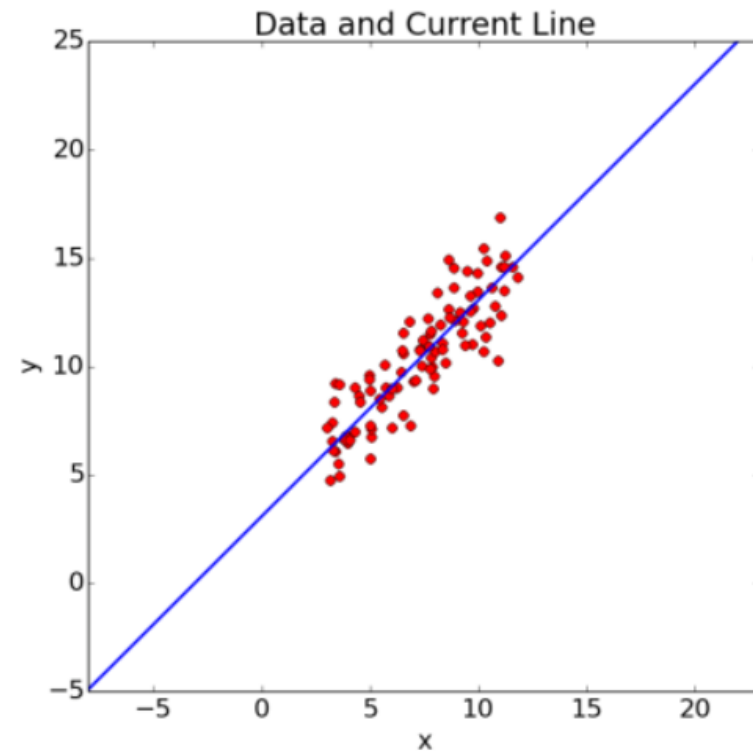
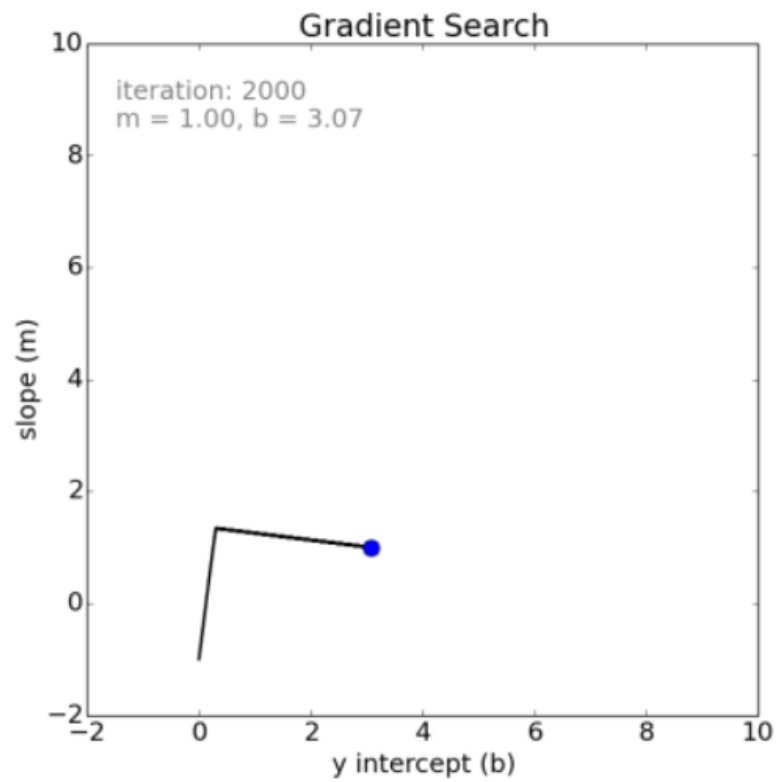
Парная регрессия

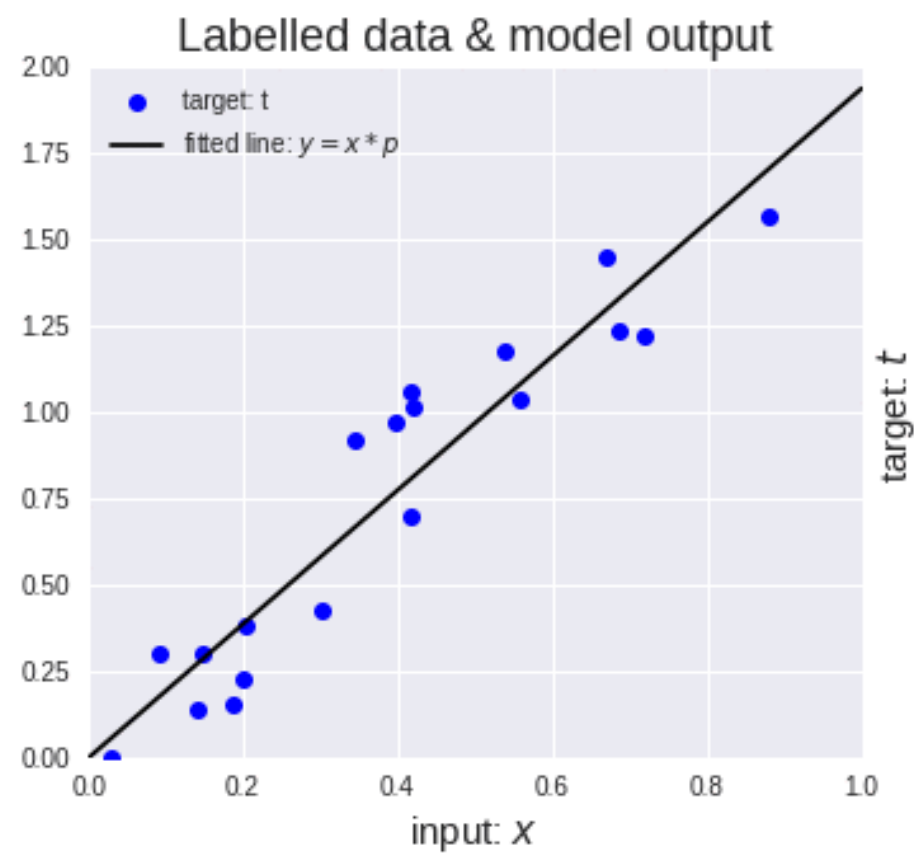
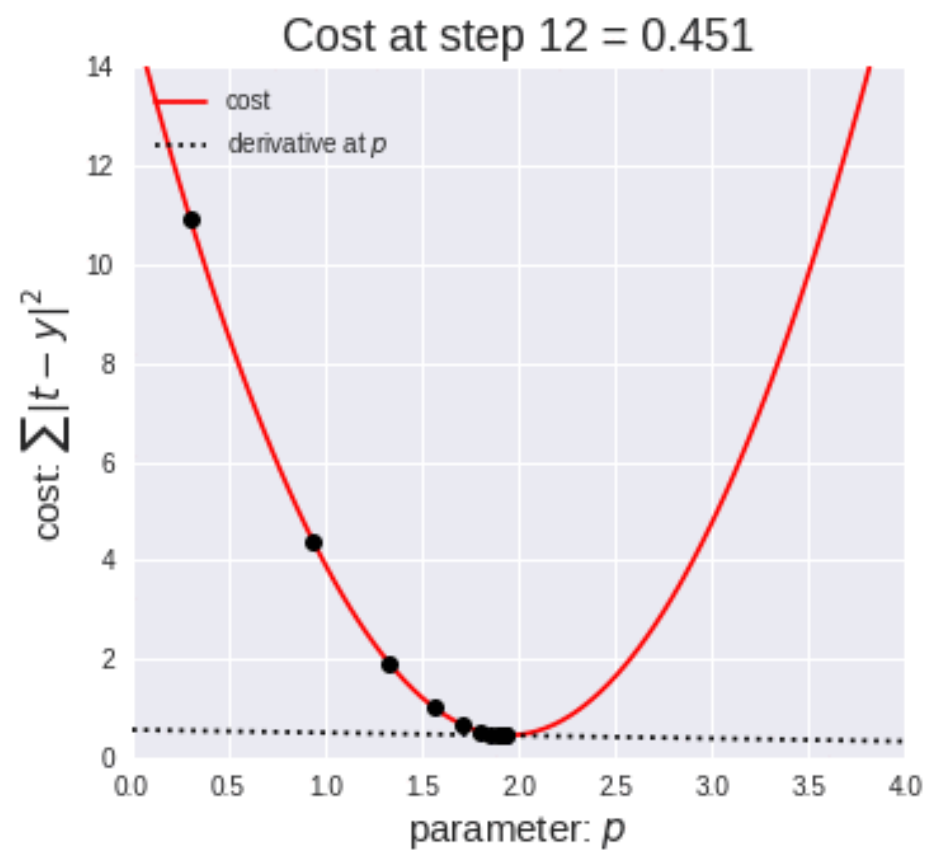


Парная регрессия

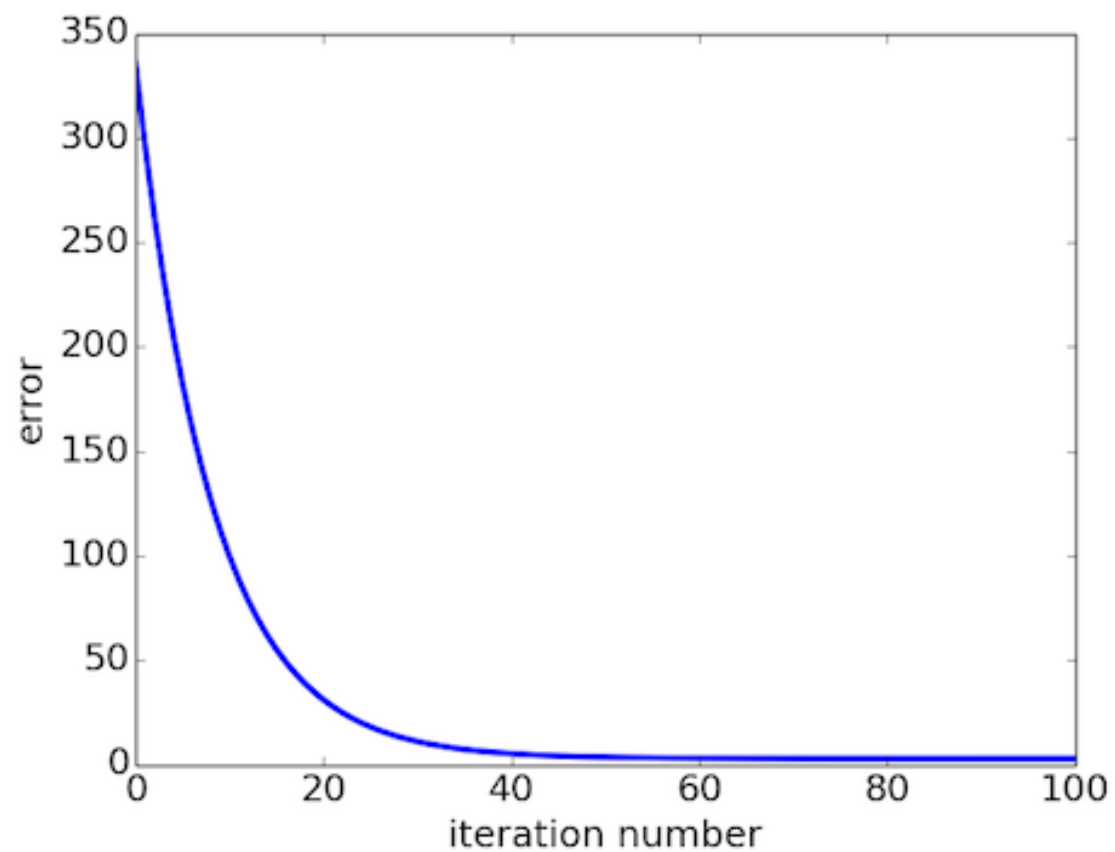


Парная регрессия





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:

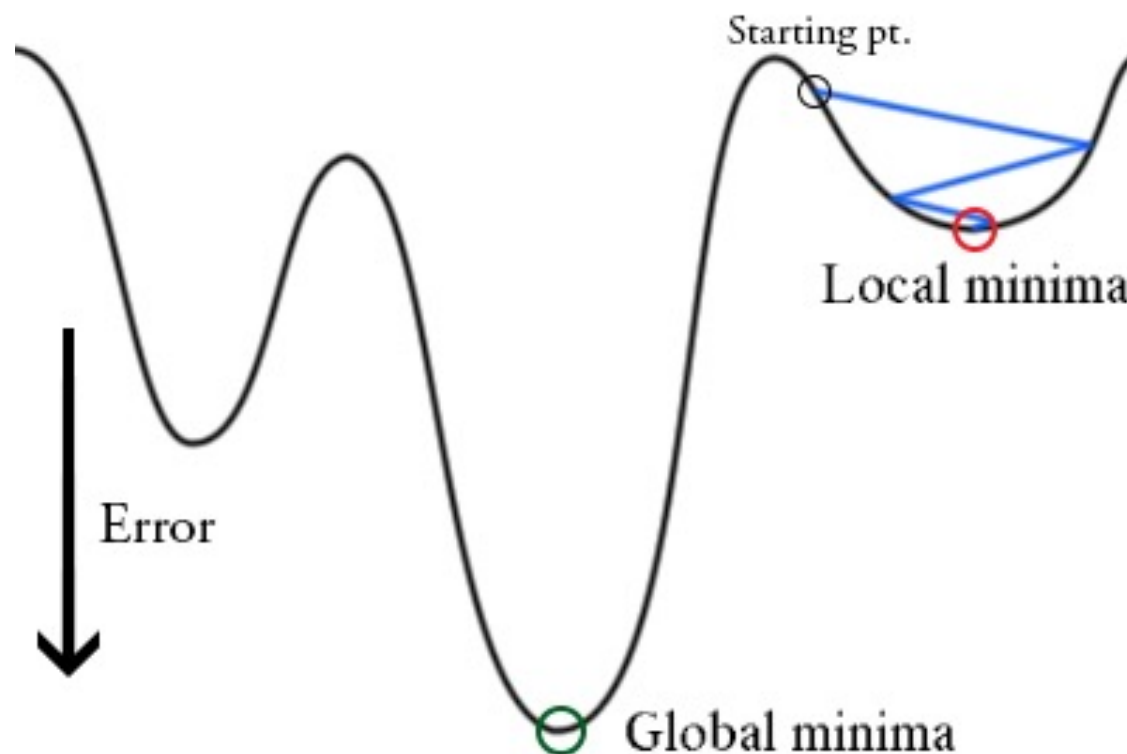
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

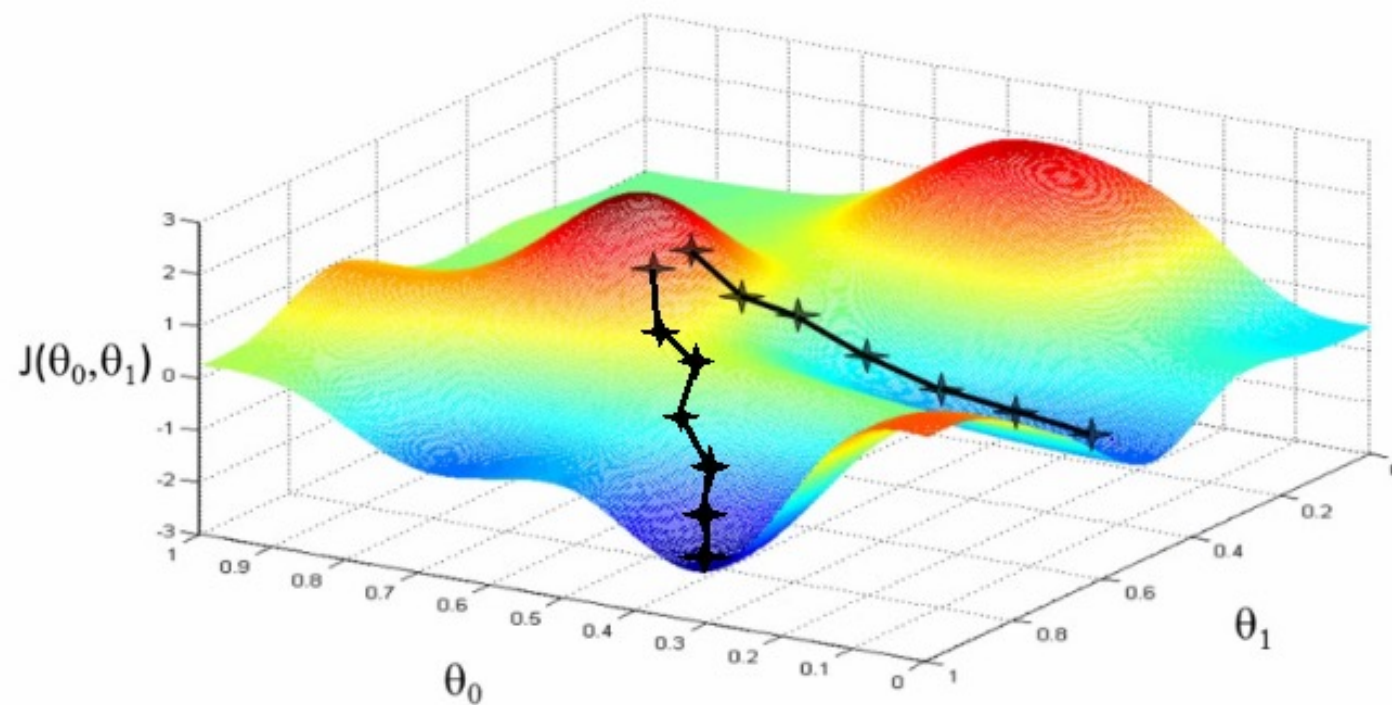
$$\|w^t - w^{t-1}\| < \varepsilon$$

Локальные минимумы

- Градиентный спуск находит только локальные минимумы



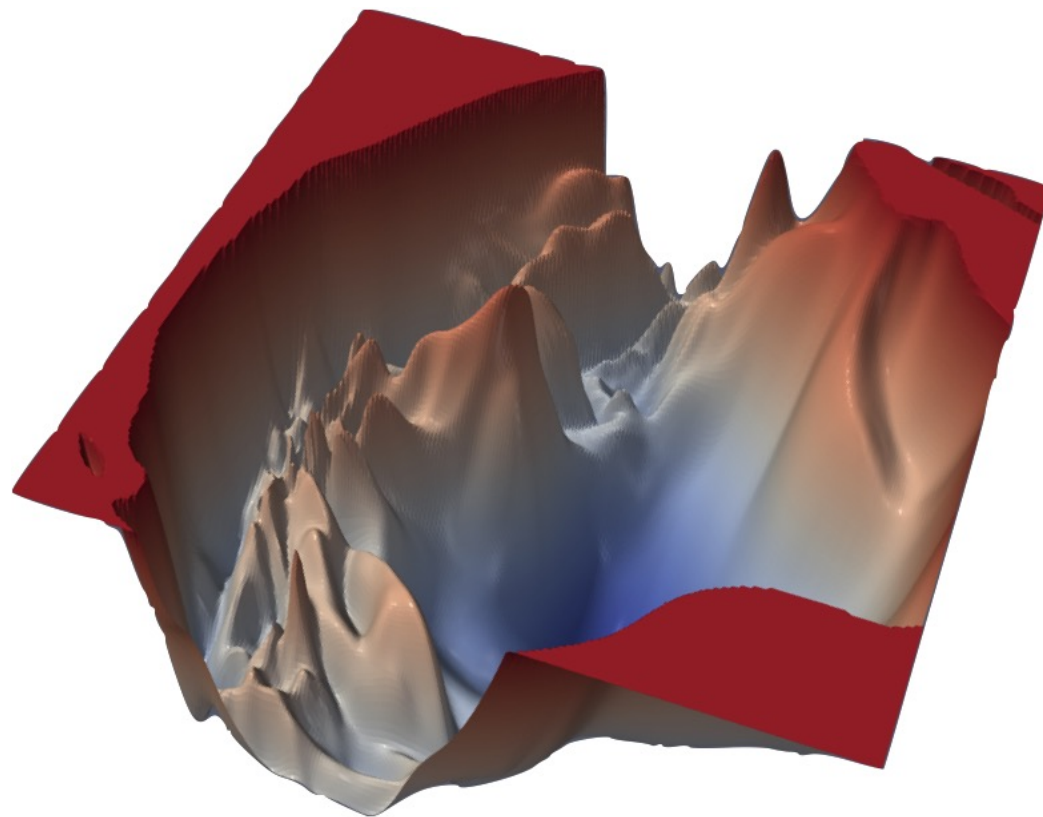
Локальные минимумы



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

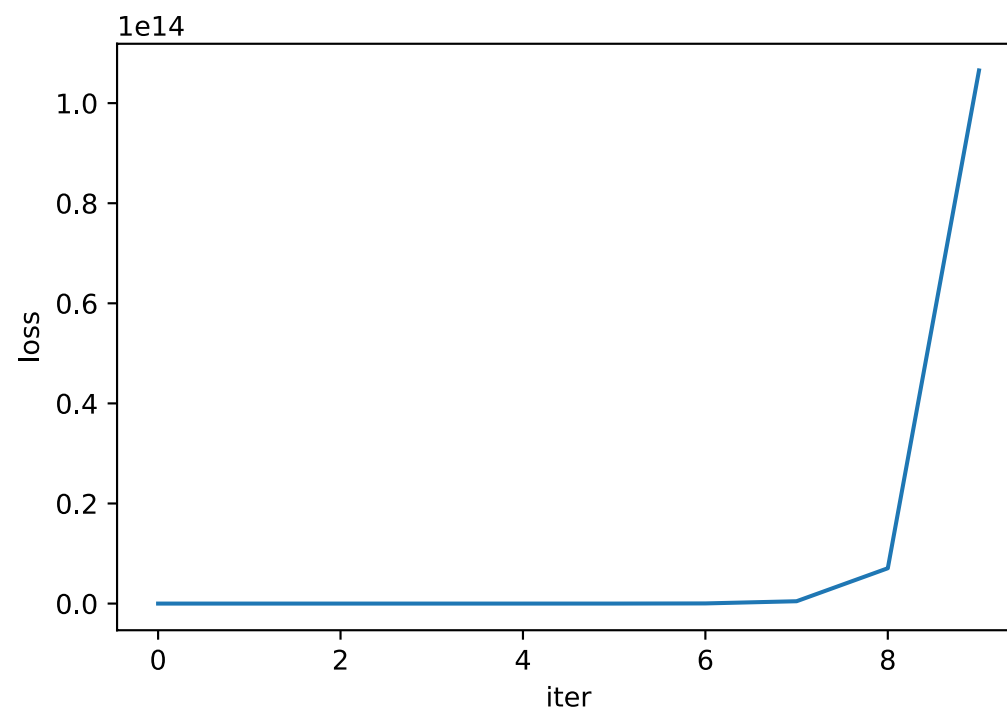
Длина шага

```
[[ 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:

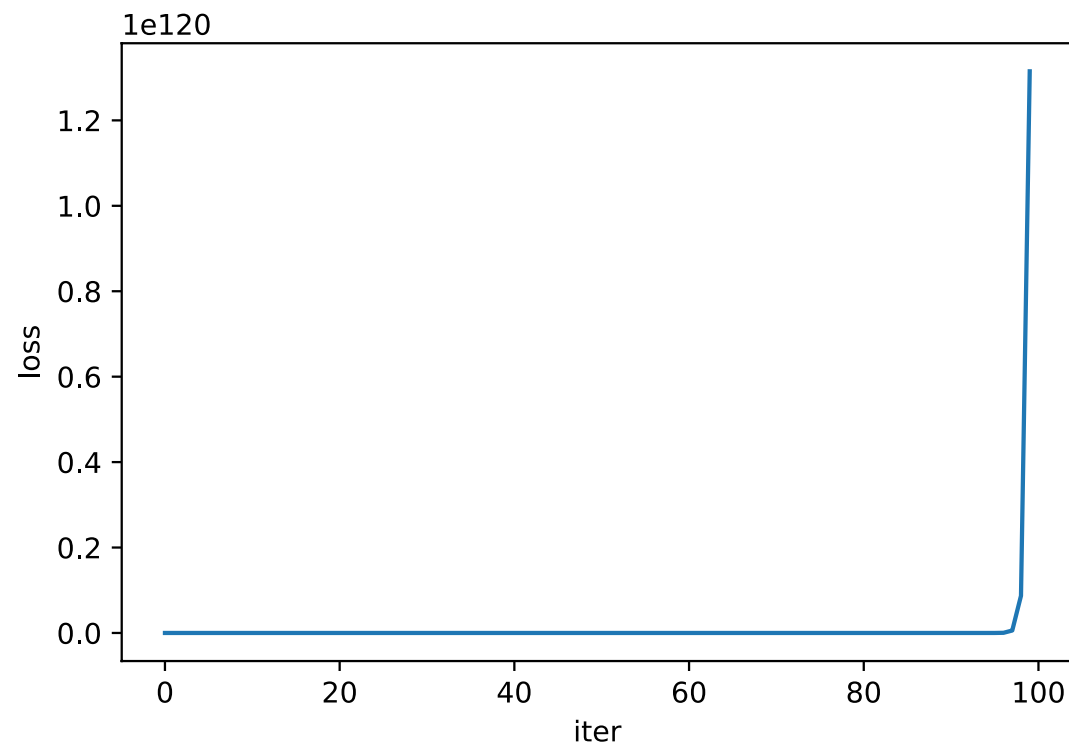
[26.52, 564.80, 682.90, 5097.71, 12110.87]



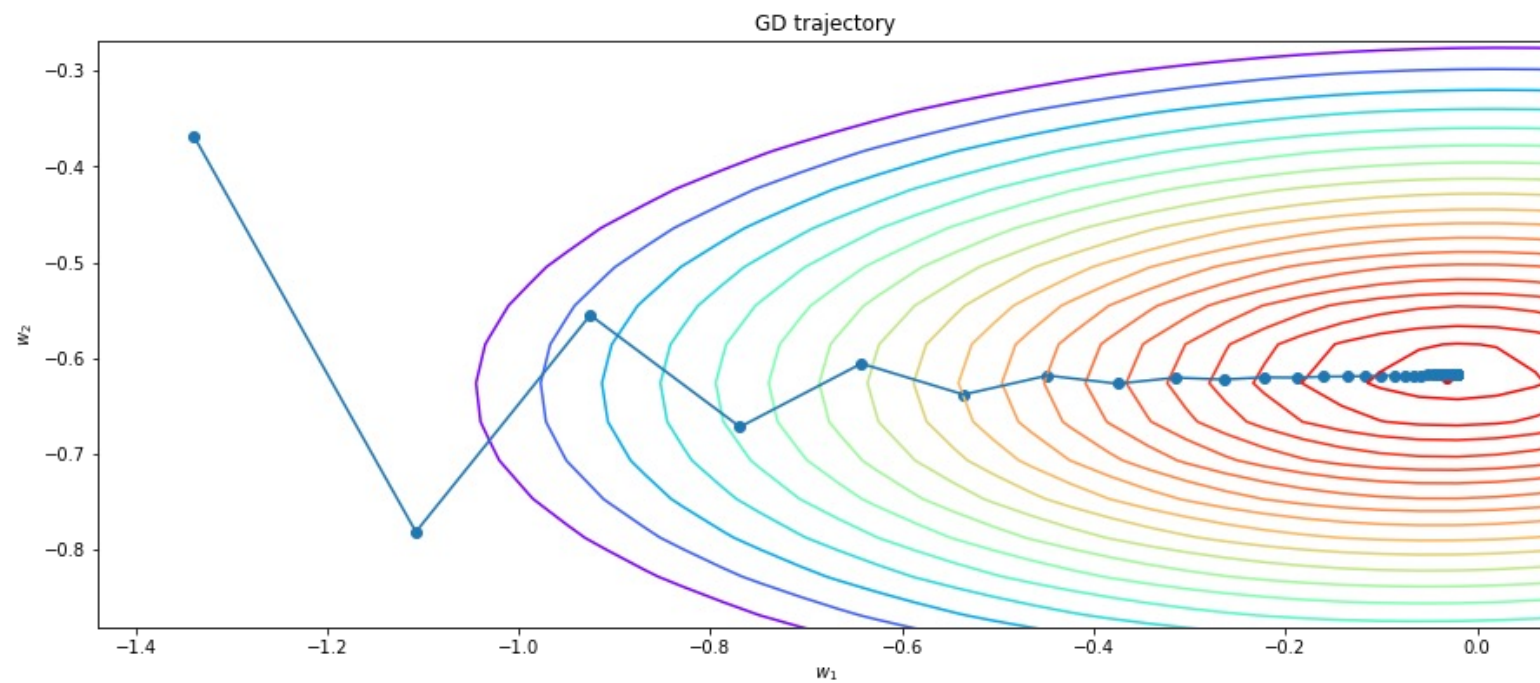
Длина шага

Градиент на первом шаге:

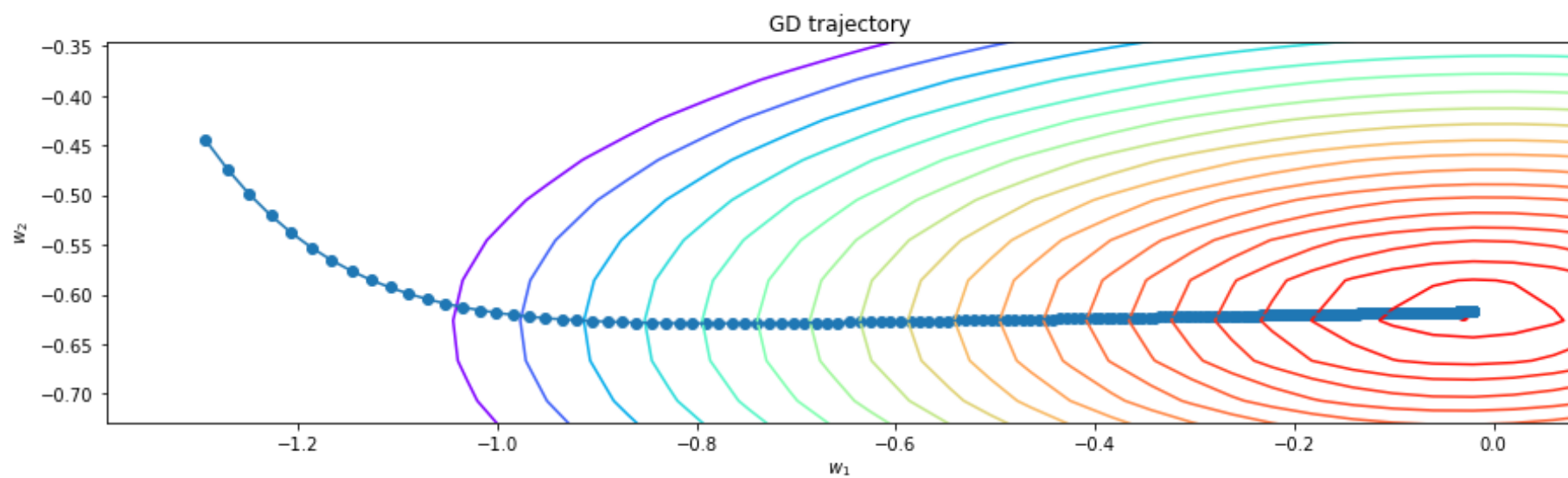
[26.52, 564.80, 682.90, 5097.71, 12110.87]



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

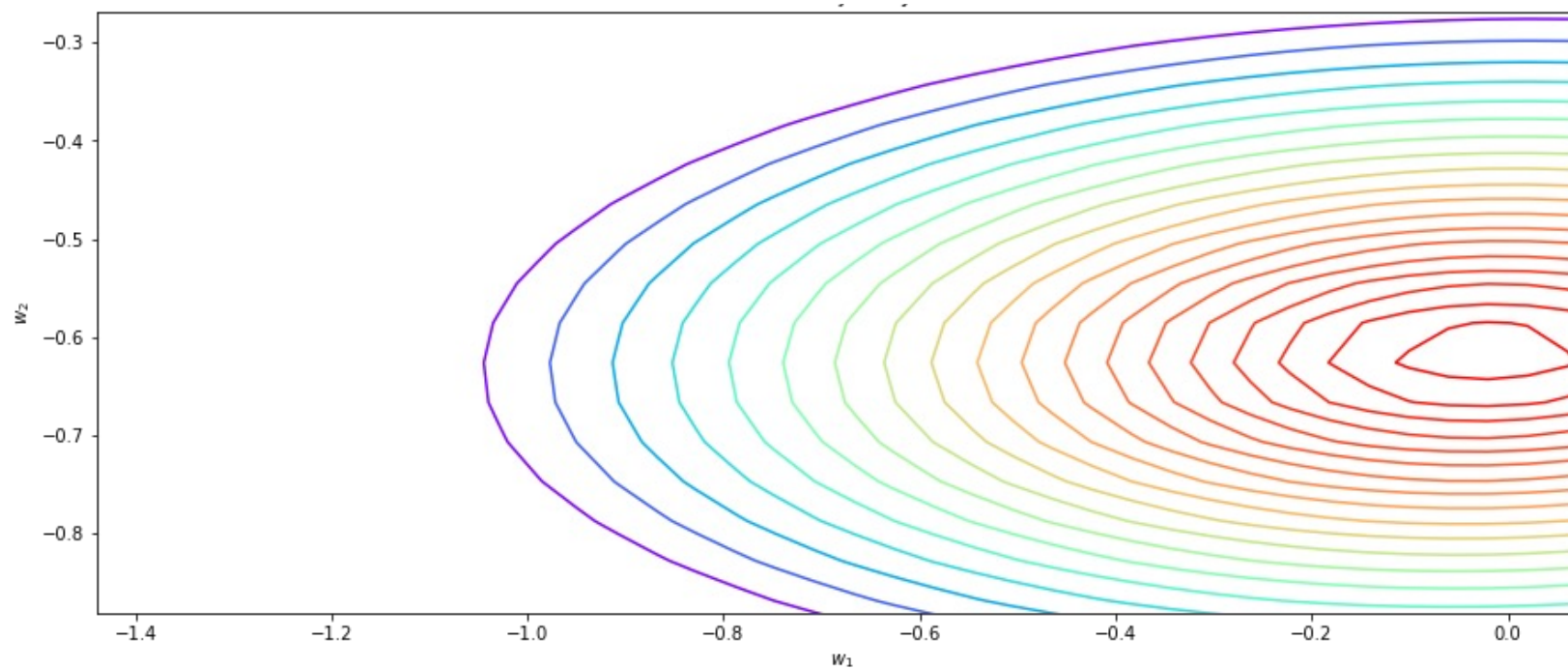
$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1})$$

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

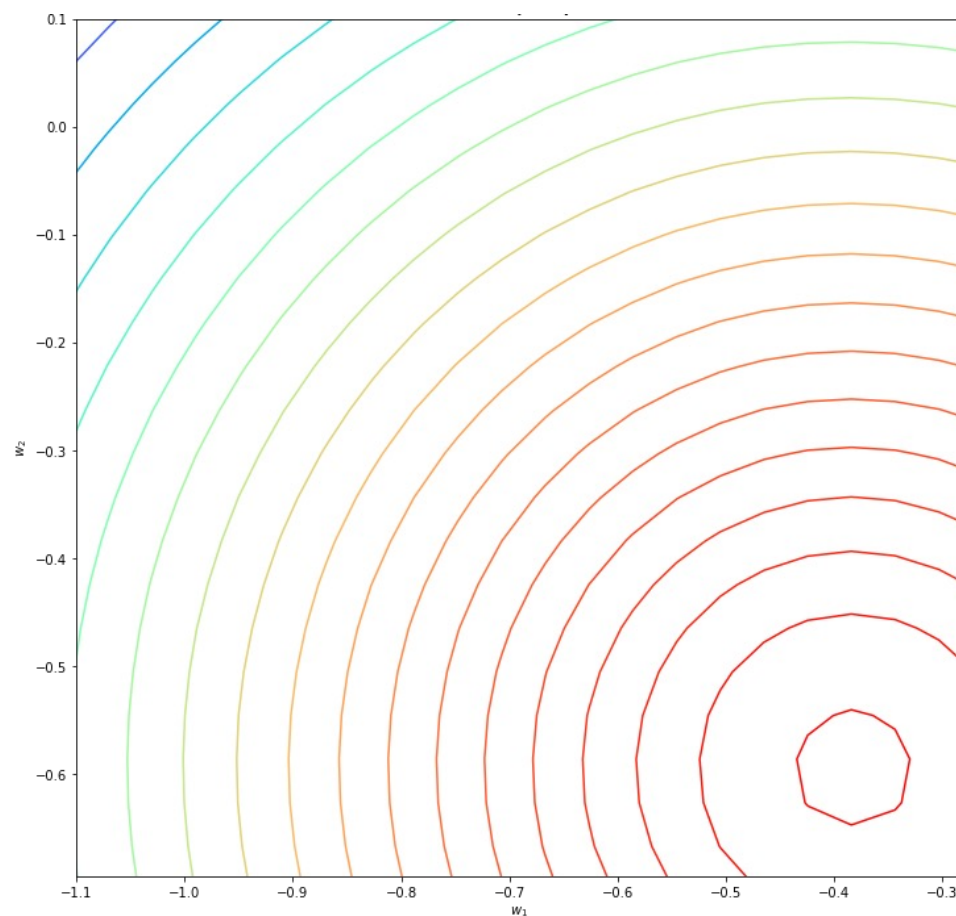
Масштабирование признаков

```
[[ 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

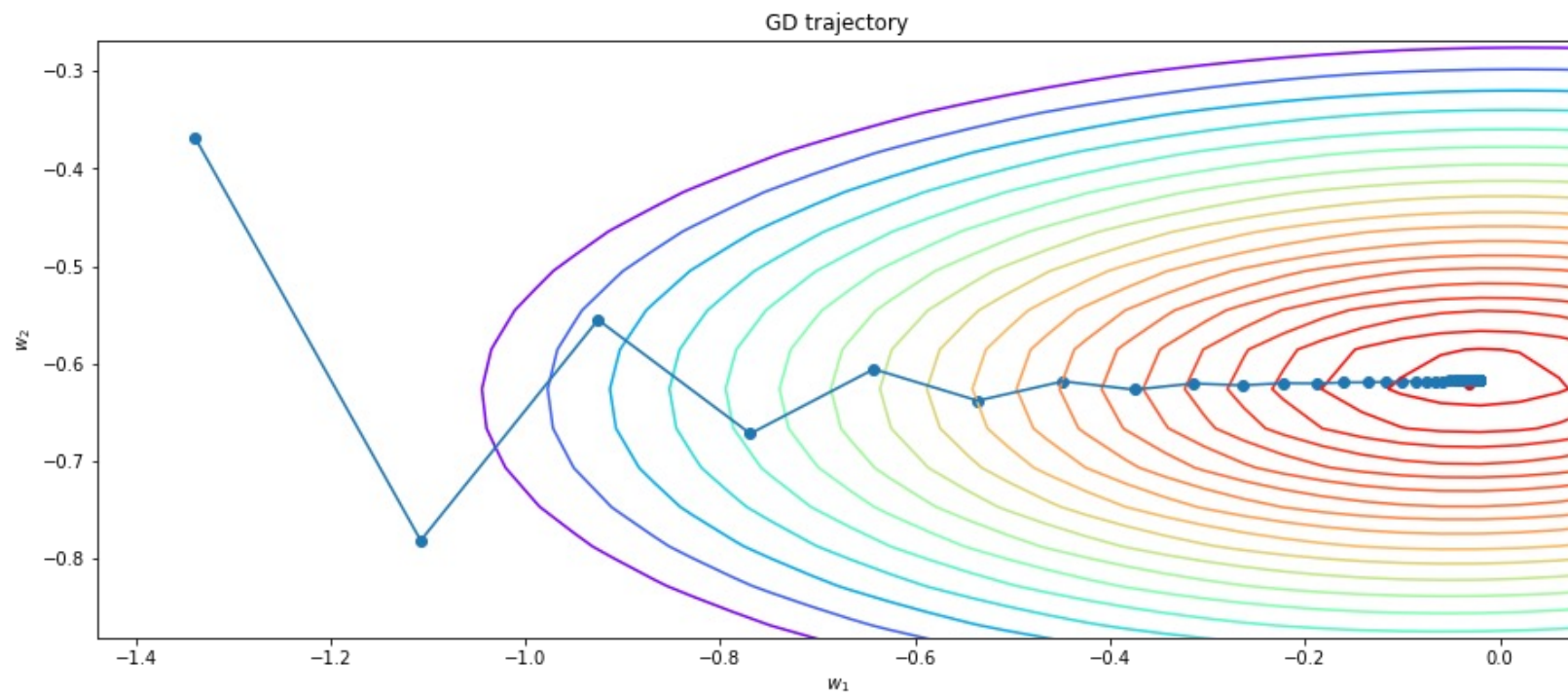
1. Начальное приближение: w^0
2. Повторять, каждый раз выбирая случайный объект i_t :

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

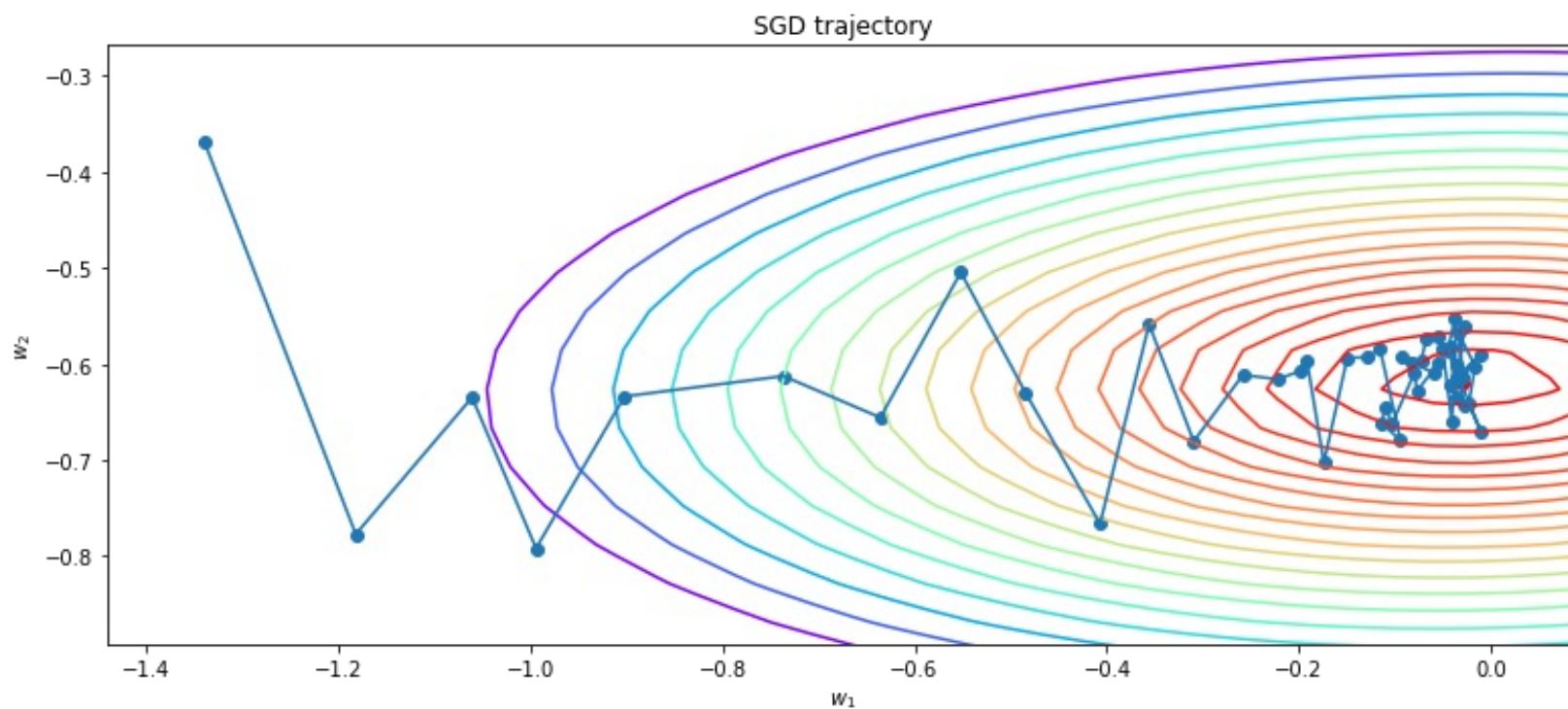
3. Останавливаемся, если

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

Градиентный спуск



Стохастический градиентный спуск



Стохастический градиентный спуск

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

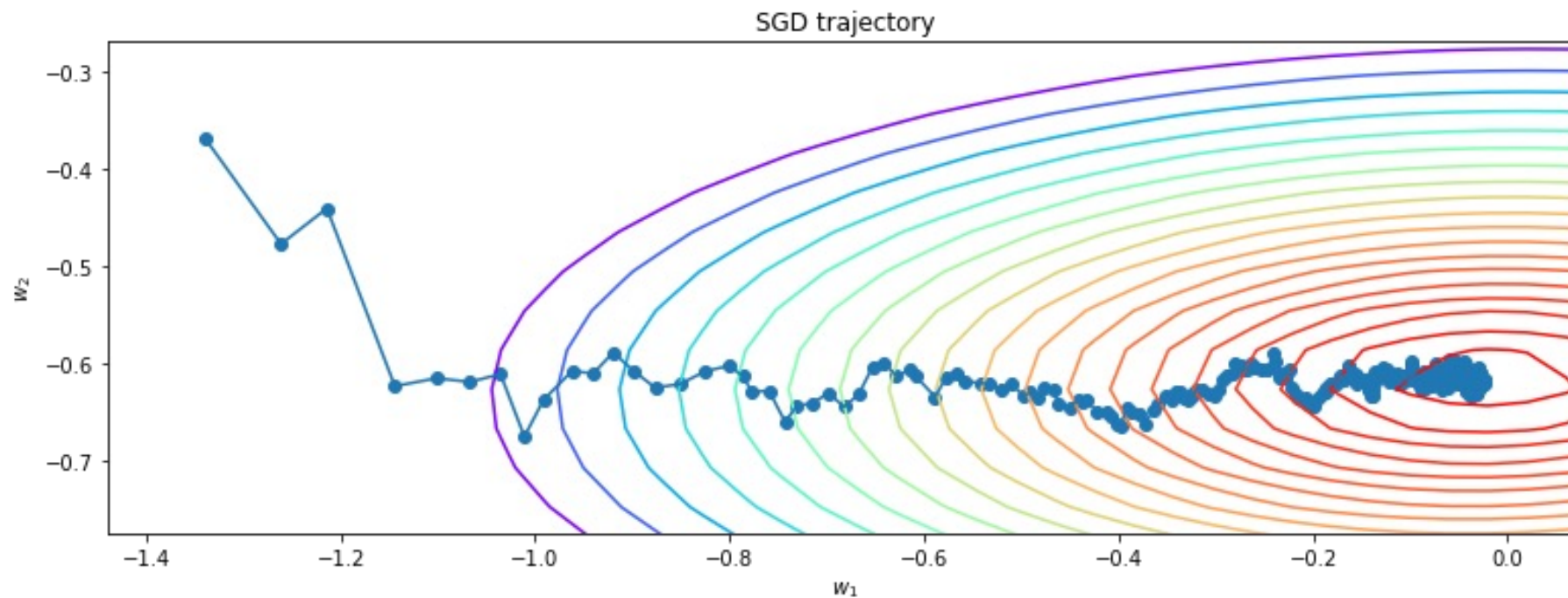
$$w^t = w^{t-1} - \eta_t \nabla L(y_{i_t}, a(x_{i_t}))$$

3. Останавливаемся, если

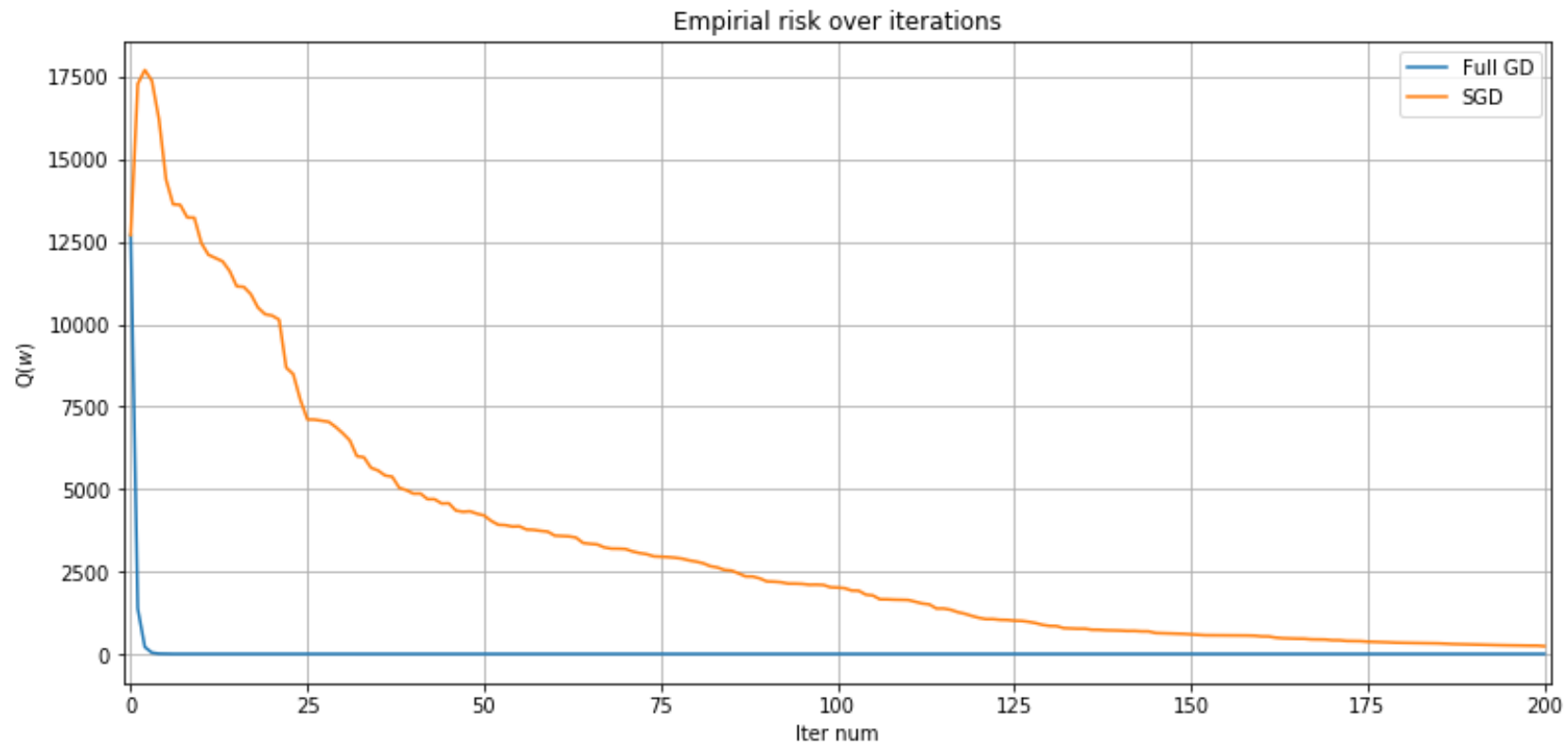
$$\|w^t - w^{t-1}\| < \varepsilon$$

Стохастический градиентный спуск

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



Стохастический градиентный спуск



Mini-batch

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

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

3. Останавливаемся, если

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