Основы машинного обучения

Лекция 6

Линейная регрессия и градиентный спуск

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2024

Интерпретация линейных моделей

Предсказание стоимости квартиры

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

Предсказание стоимости квартиры

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

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

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

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

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

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

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

- Чем больше вес, тем важнее признак?

Предсказание стоимости квартиры

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

- 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:

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

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

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:

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

Градиент

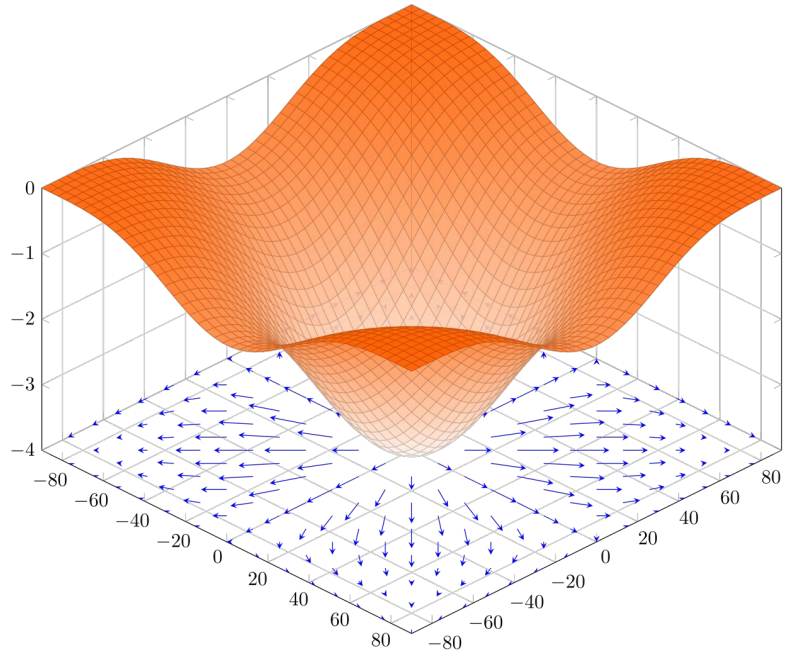
- Градиент — вектор частных производных

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

- У градиента есть важное свойство!

Важное свойство

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

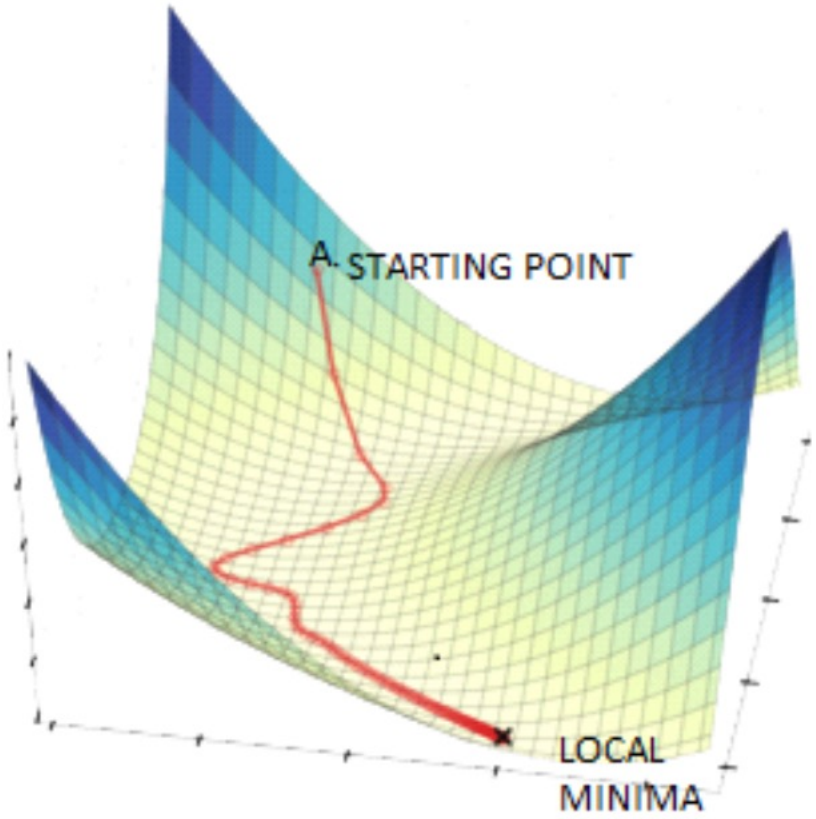


Важное свойство

- 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

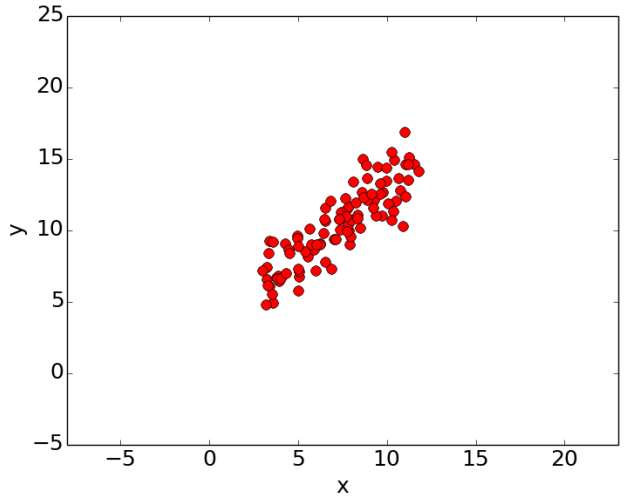
- Стартуем из случайной точки
- Сдвигаемся по антиградиенту
- Повторяем, пока не окажемся в точке минимума

Парная регрессия

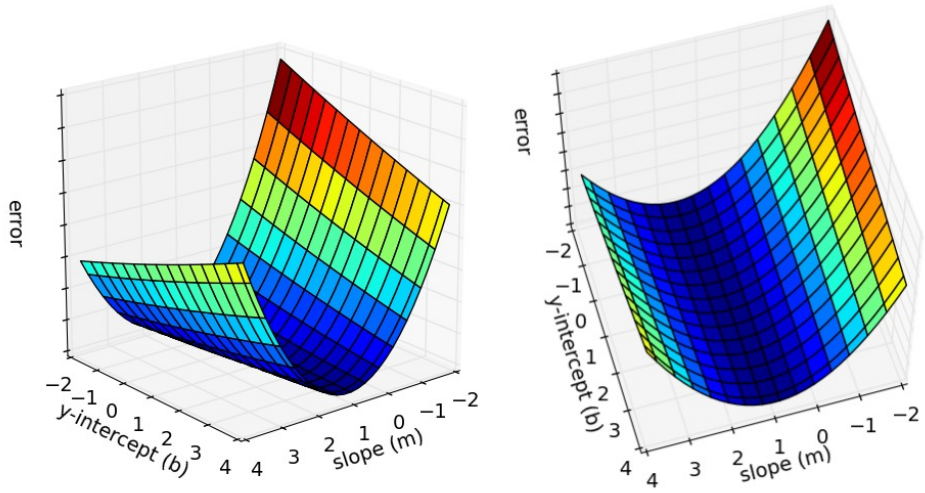
- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра: w_1 и w_0
- Функционал:

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

Парная регрессия



Выборка



Функционал ошибки

https://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/

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

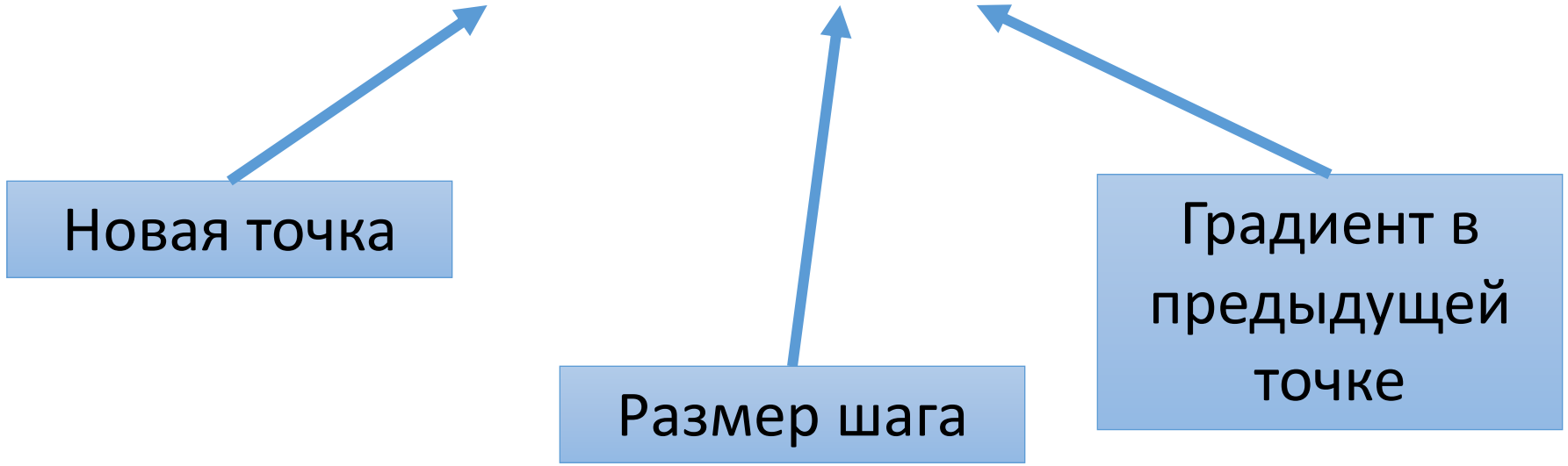
- w^0 — инициализация весов

- Например, из стандартного нормального распределения

Градиентный спуск

- Повторять до сходимости:

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



Сходимость

- Останавливаем процесс, если

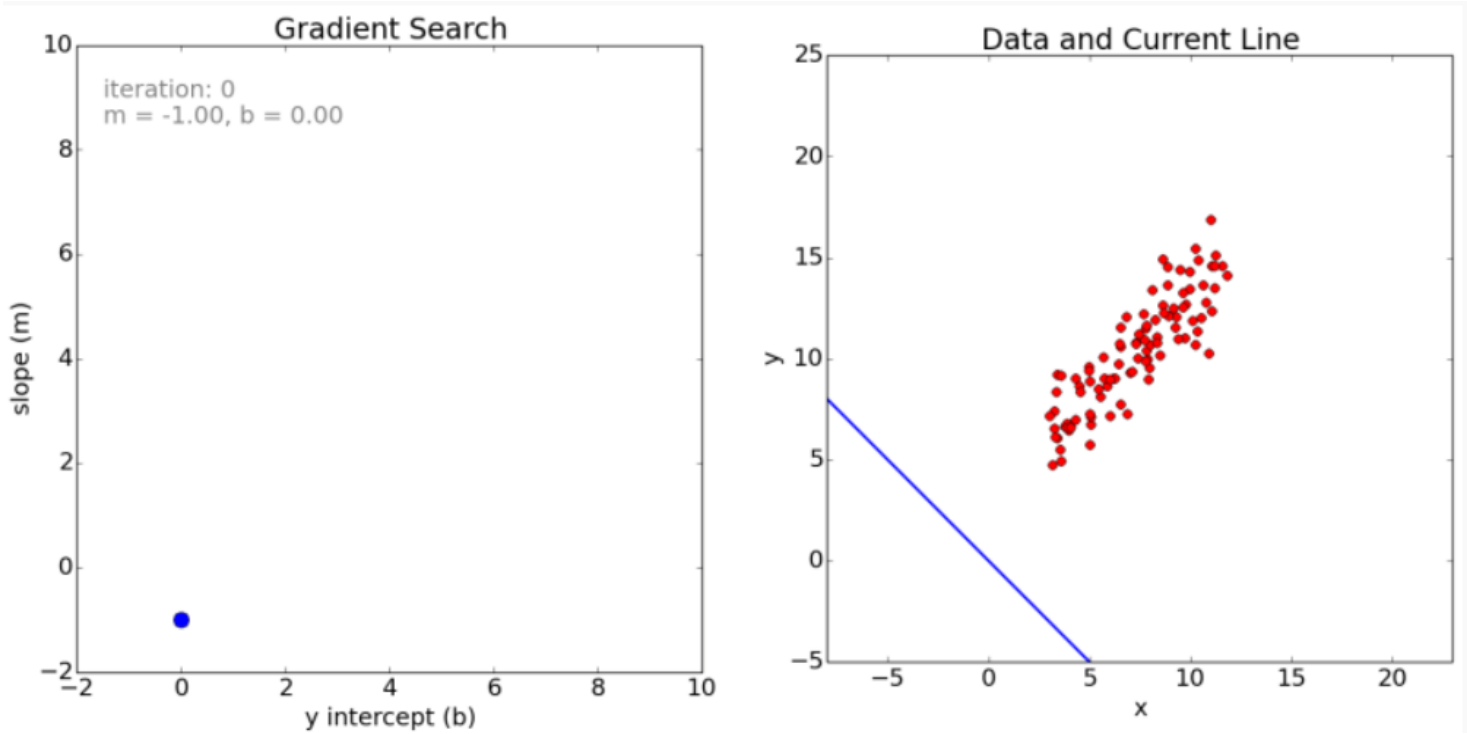
$$\|w^t - w^{t-1}\| < \varepsilon$$

- Другой вариант:

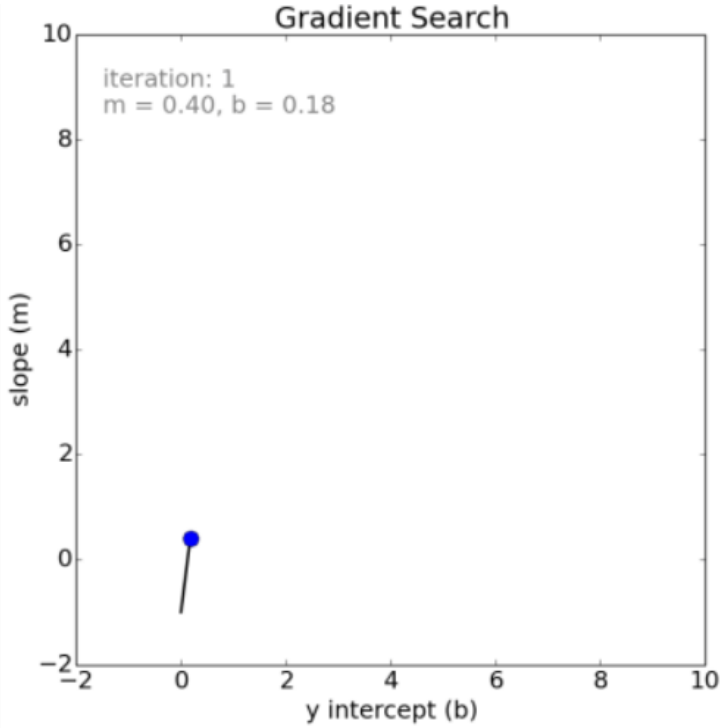
$$\|\nabla Q(w^t)\| < \varepsilon$$

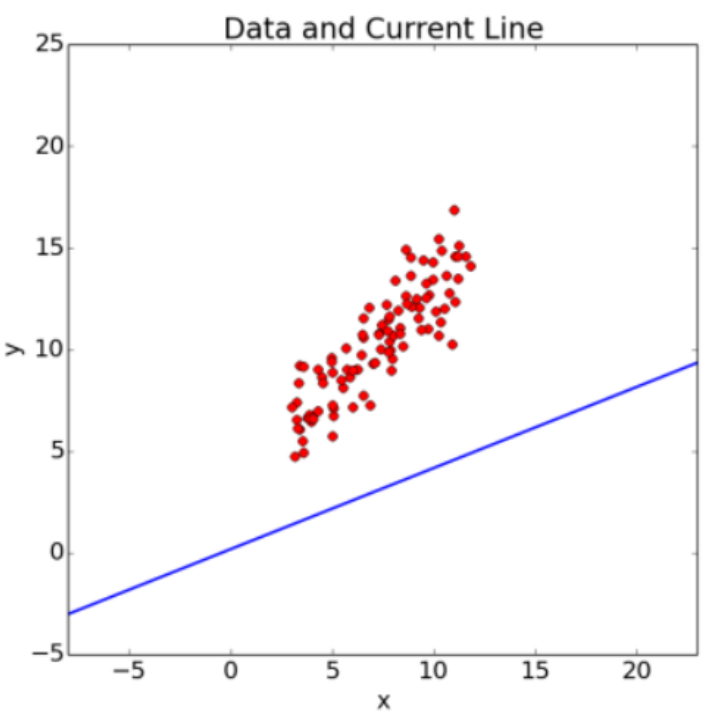
- Или пока ошибка на отложенной выборке уменьшается

Парная регрессия

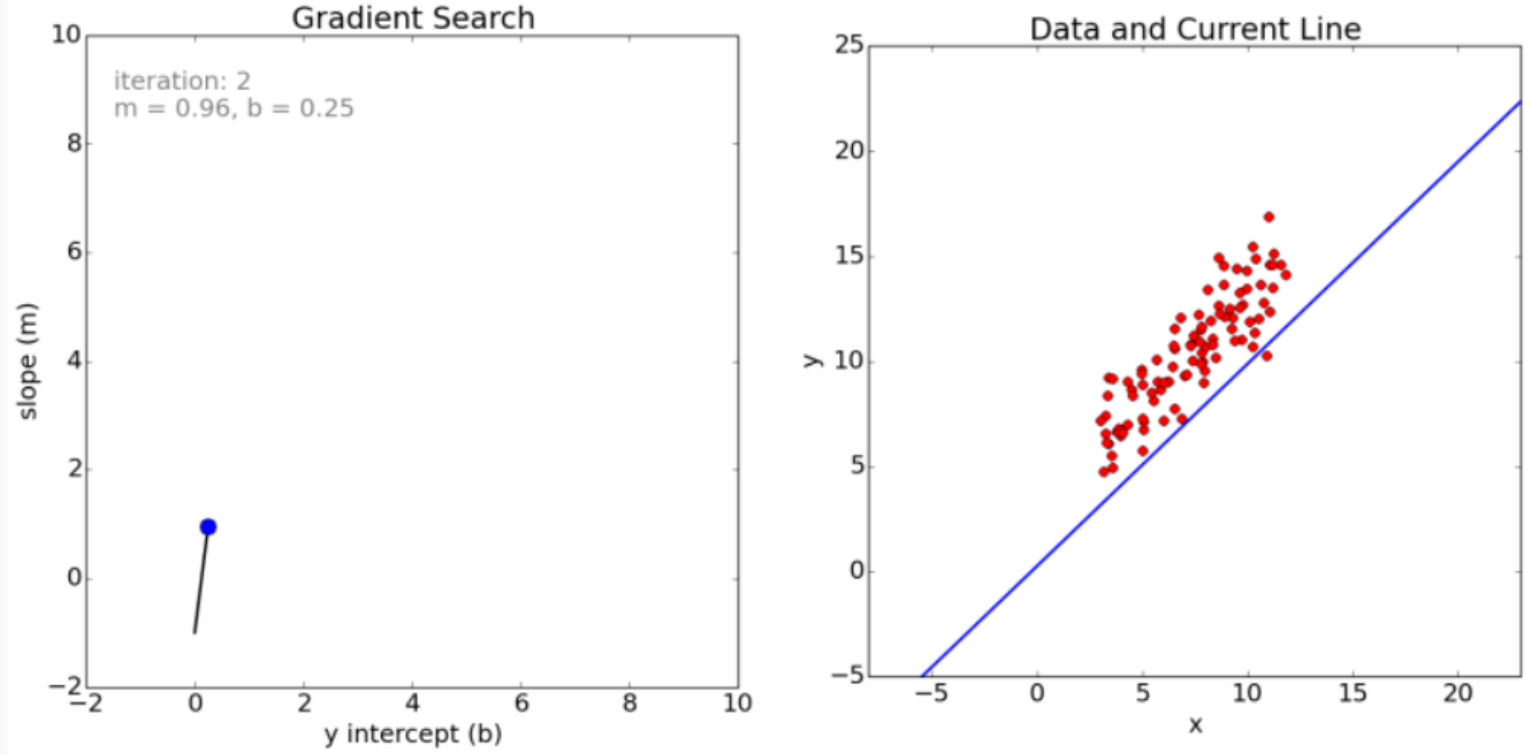


Парная регрессия

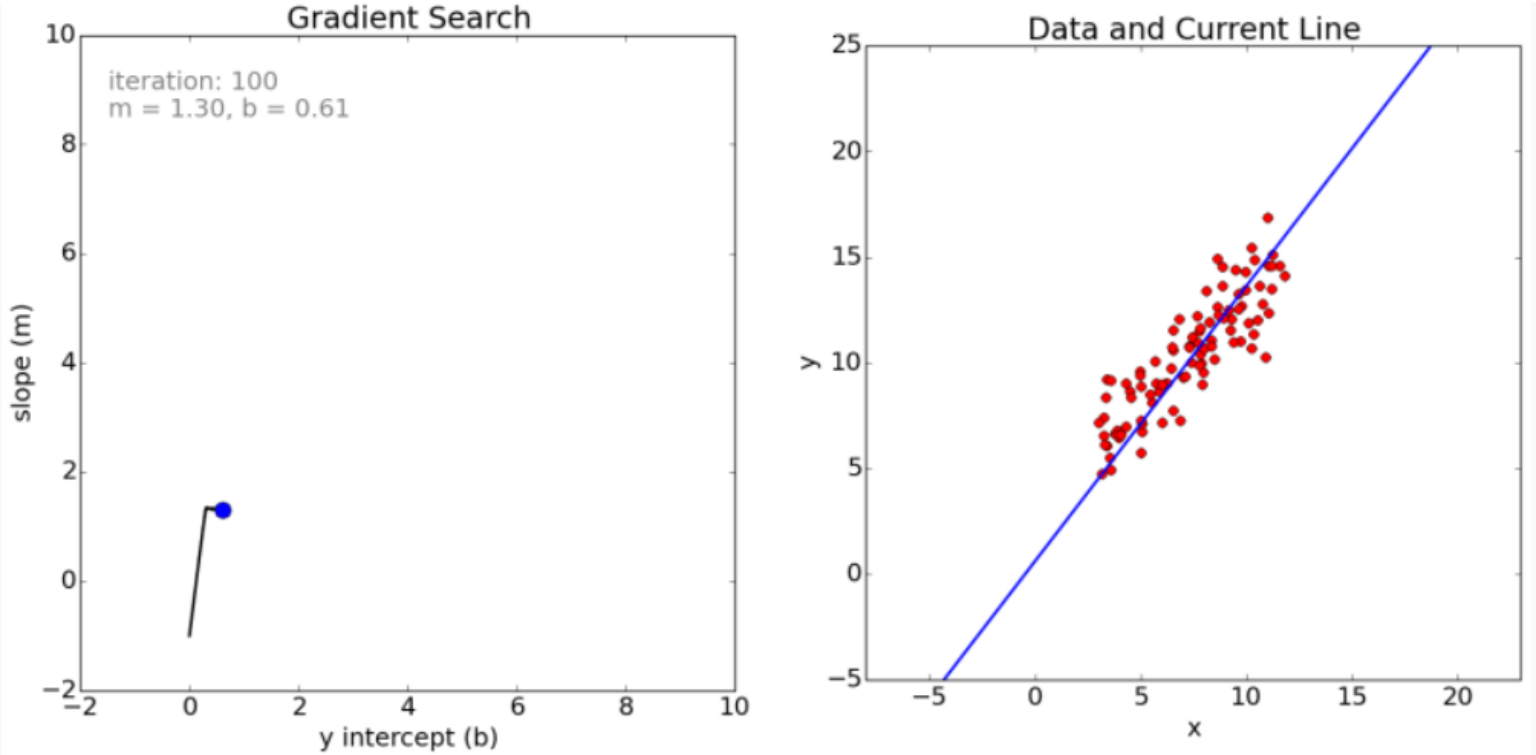




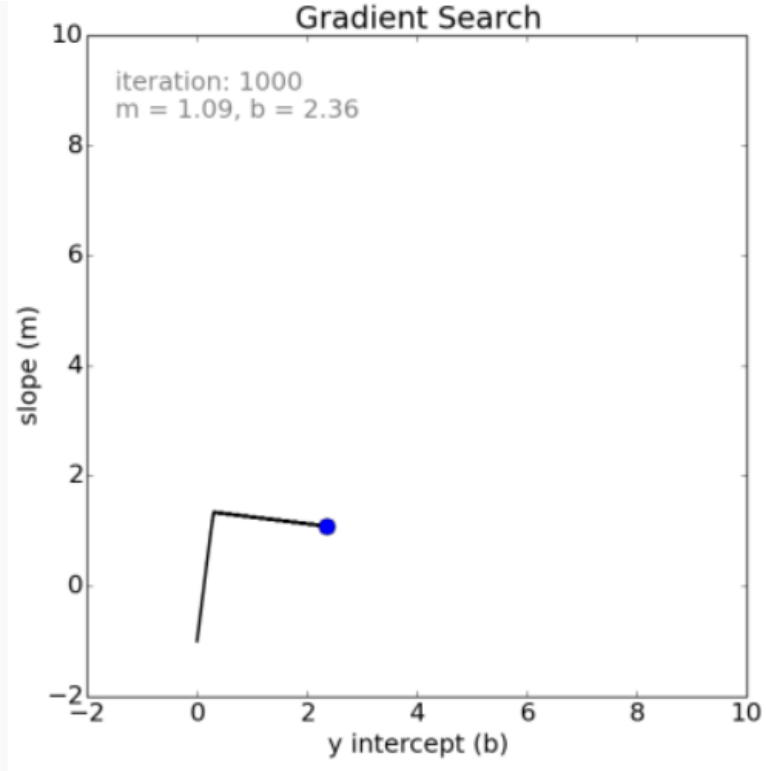
Парная регрессия

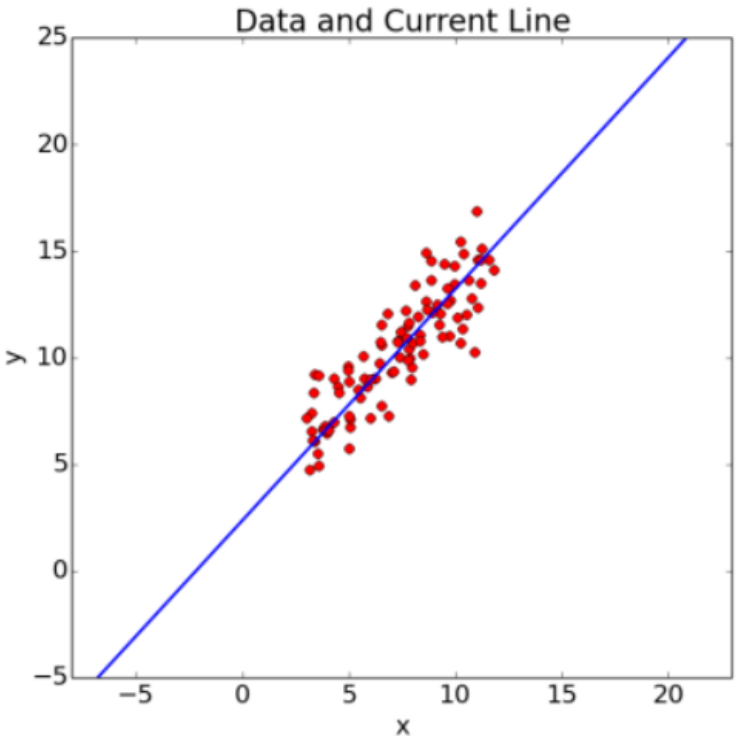


Парная регрессия

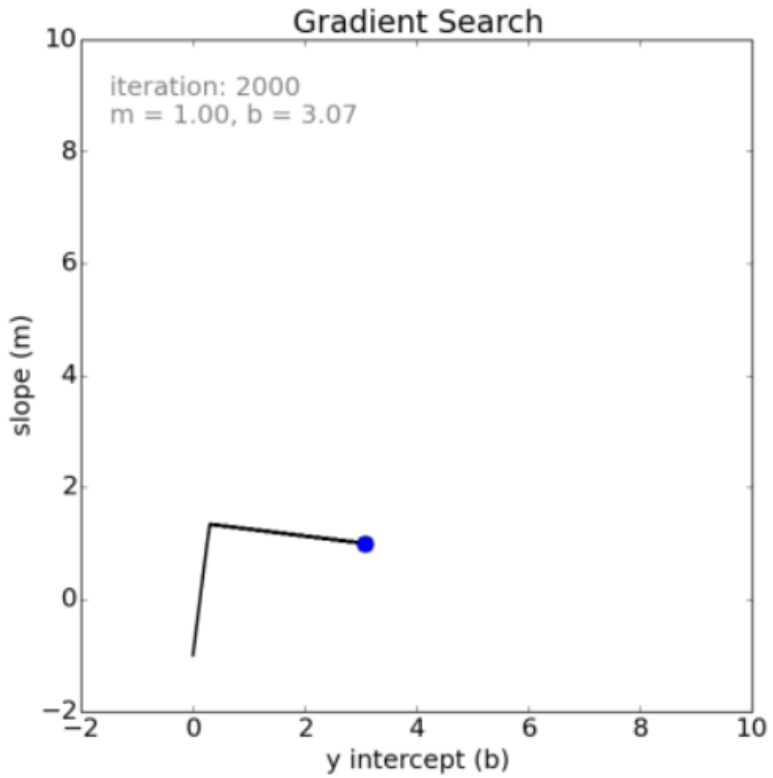


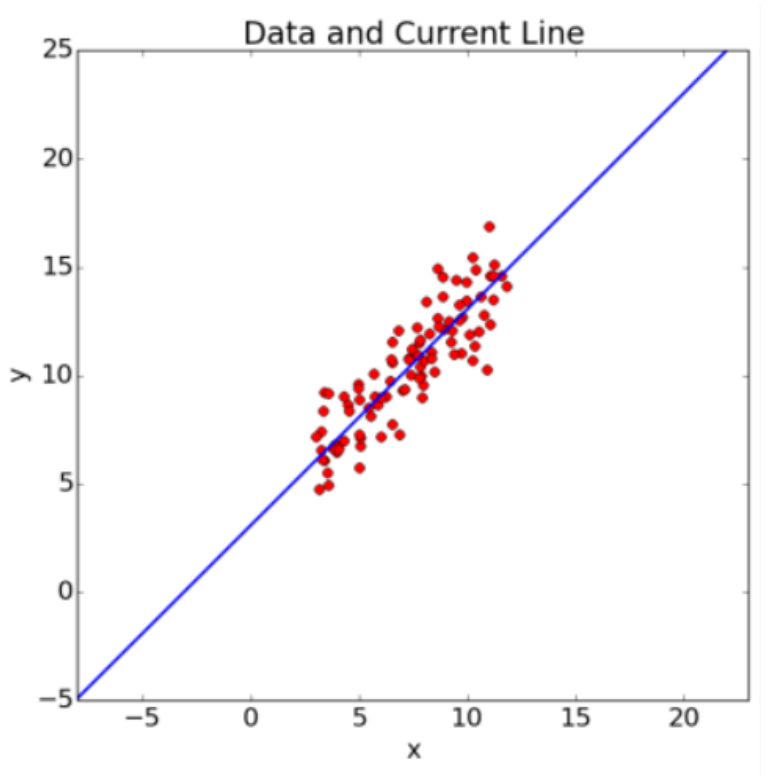
Парная регрессия

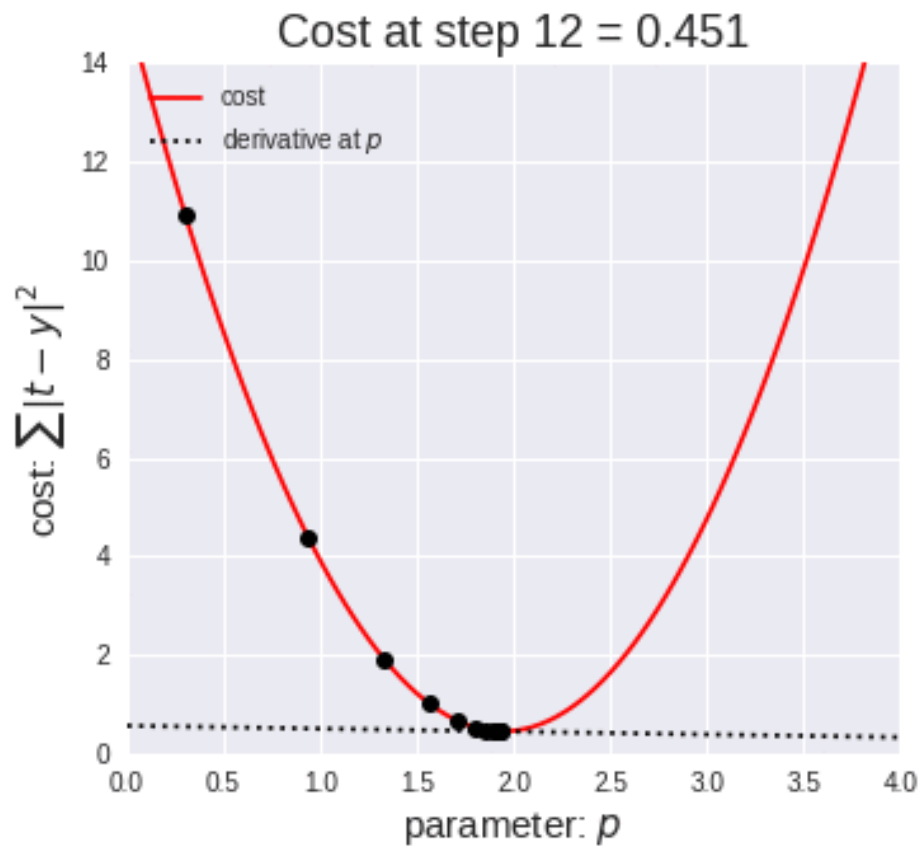


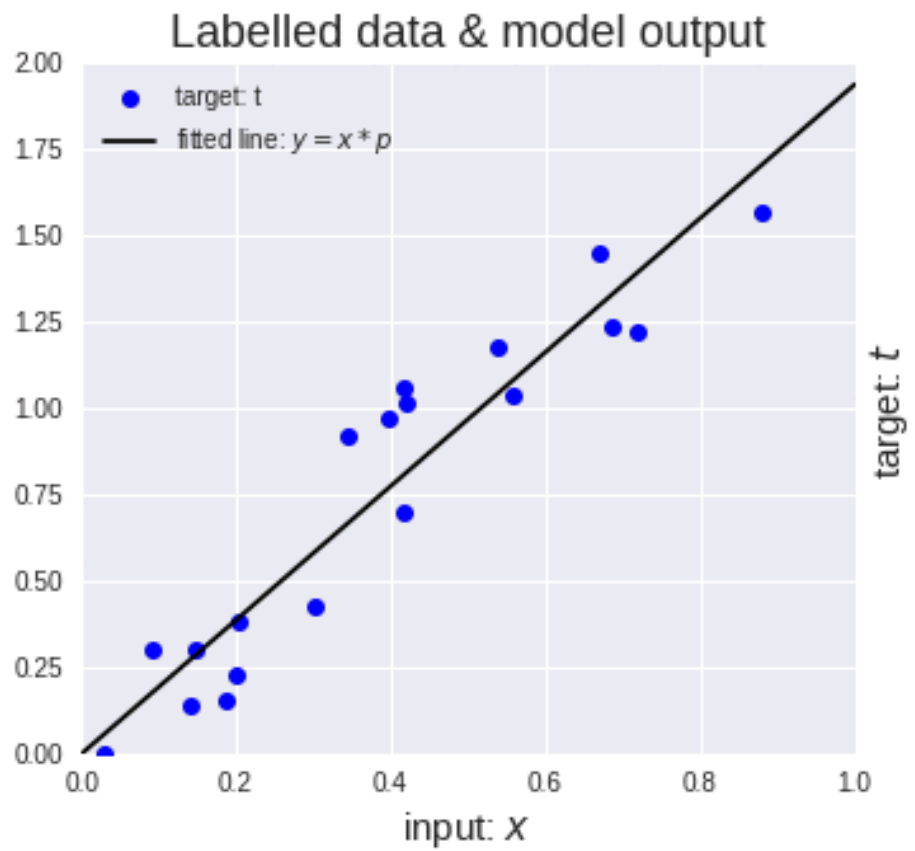


Парная регрессия

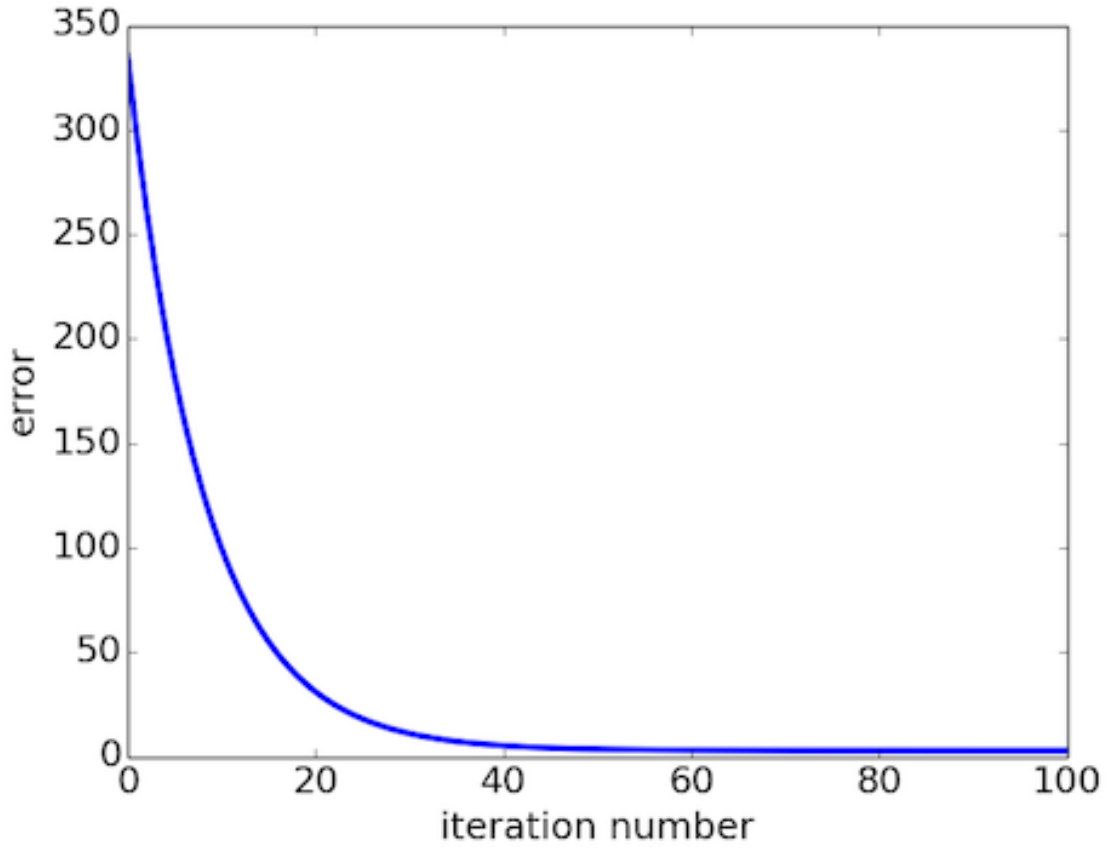





Cost at step 12 = 0.451
cost
derivative at p
cost: $\sum|t-y|^2$
parameter: p
0
2
4
6
8
10
12
14
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Labelled data & model output
target: t
fitted line: $y = x * p$
target: t
input: x
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0.0
0.2
0.4
0.6
0.8
1.0

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:

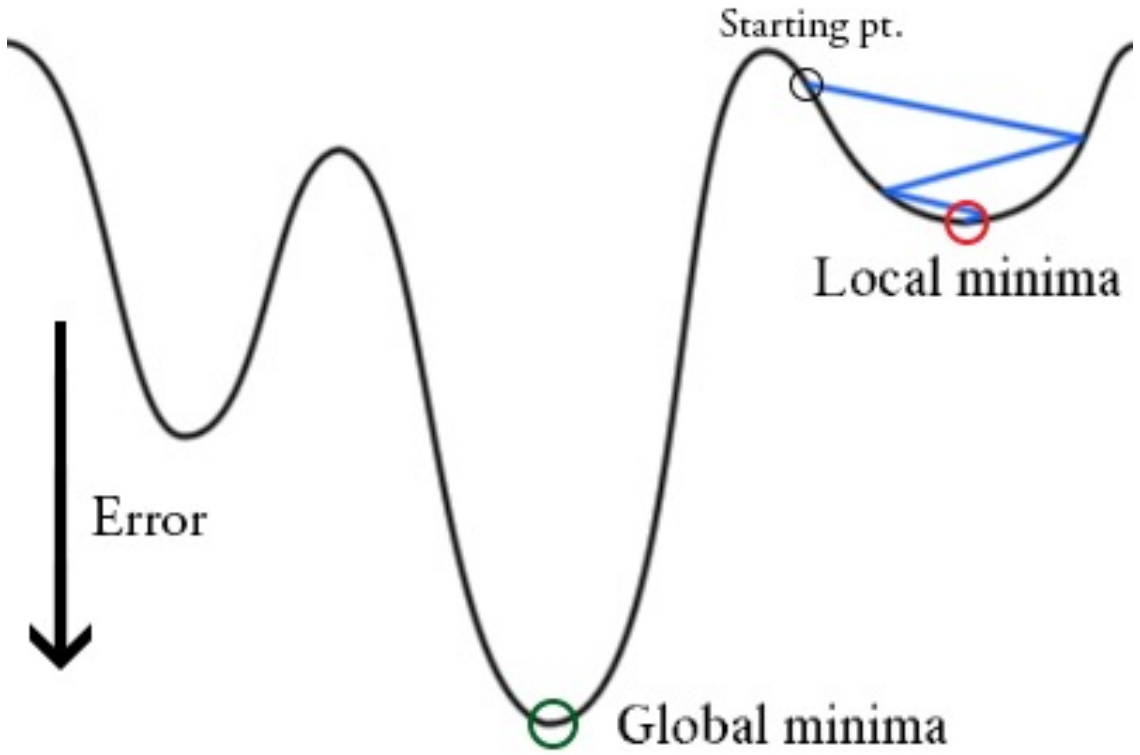
$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

3. Останавливаемся, если

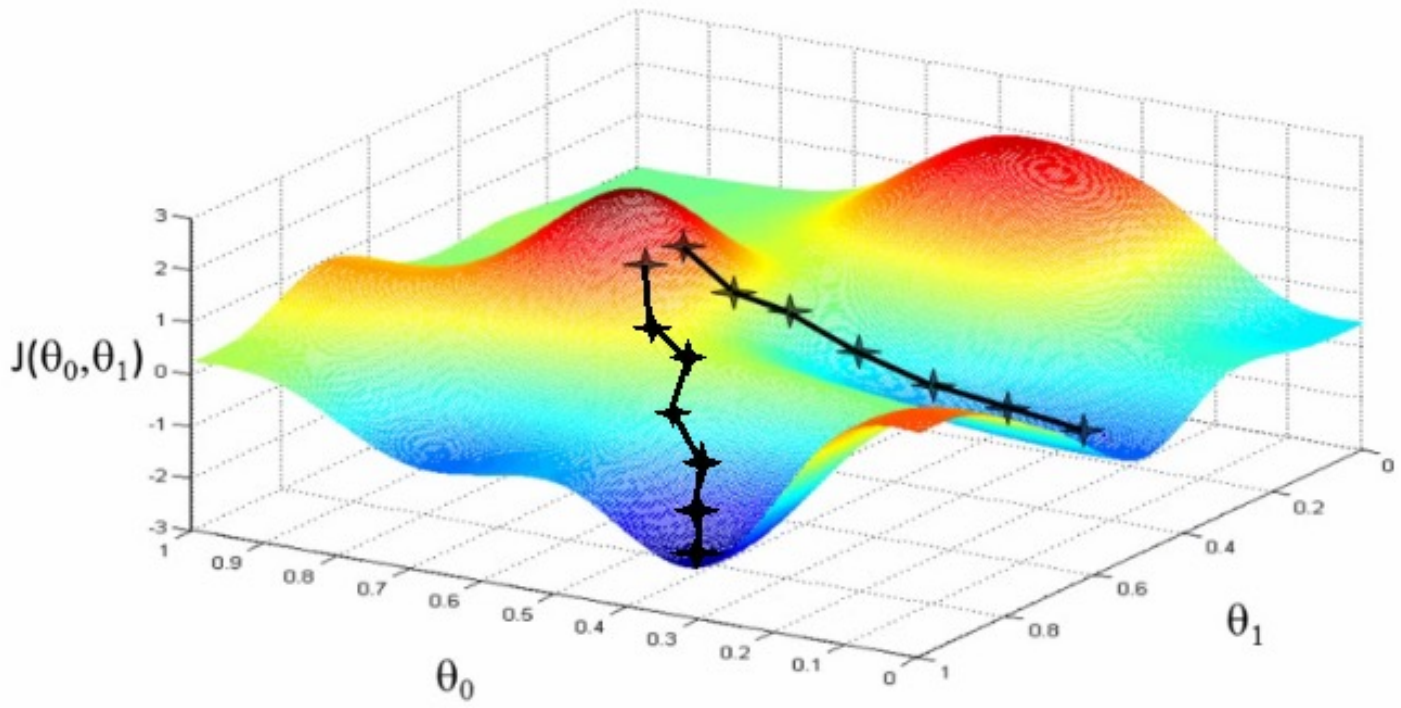
$$\|w^t - w^{t-1}\| < \varepsilon$$

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