Дисциплина

Основы машинного обучения и нейронные сети

Лекция 4
Градиентный спуск

Градиент и его свойства

Функционал качества

Квадратичная ошибка SSE для линейной регрессии:

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

Среднеквадратичная ошибка MSE для линейной регрессии:

$$Q_{MSE}(w_1, \dots, w_d) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, \mathbf{x} \rangle - y_i)^2$$

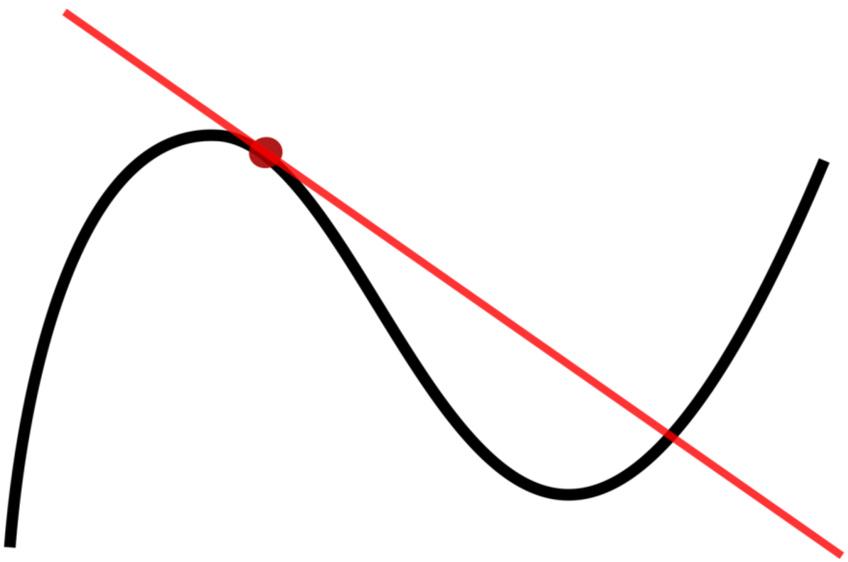
Задача:

$$Q(w_1, \dots, w_d) \to \min_{\mathbf{w}}$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} Q(\mathbf{w})$$

Производная (1/3)

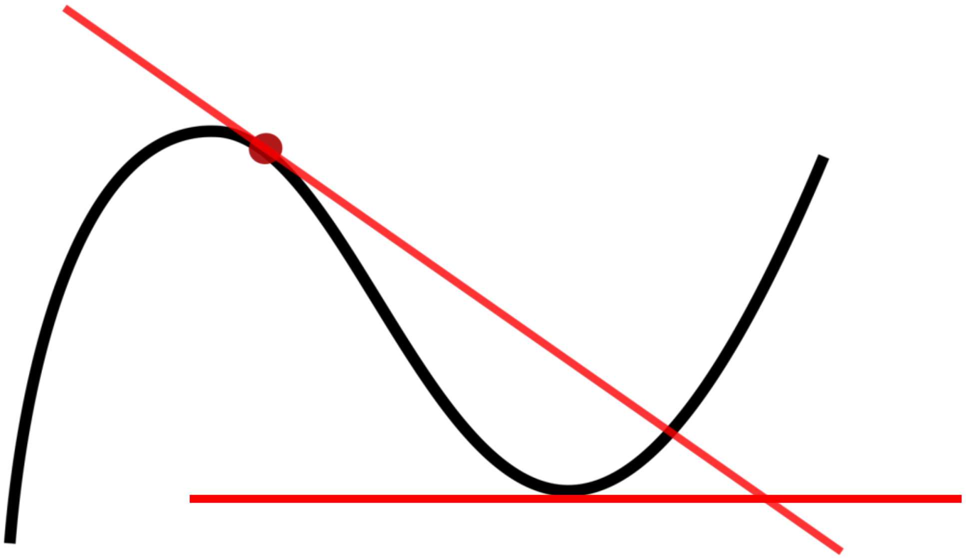
$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$



Производная (2/3)

Если точка x_0 — экстремум и в ней существует производная, то

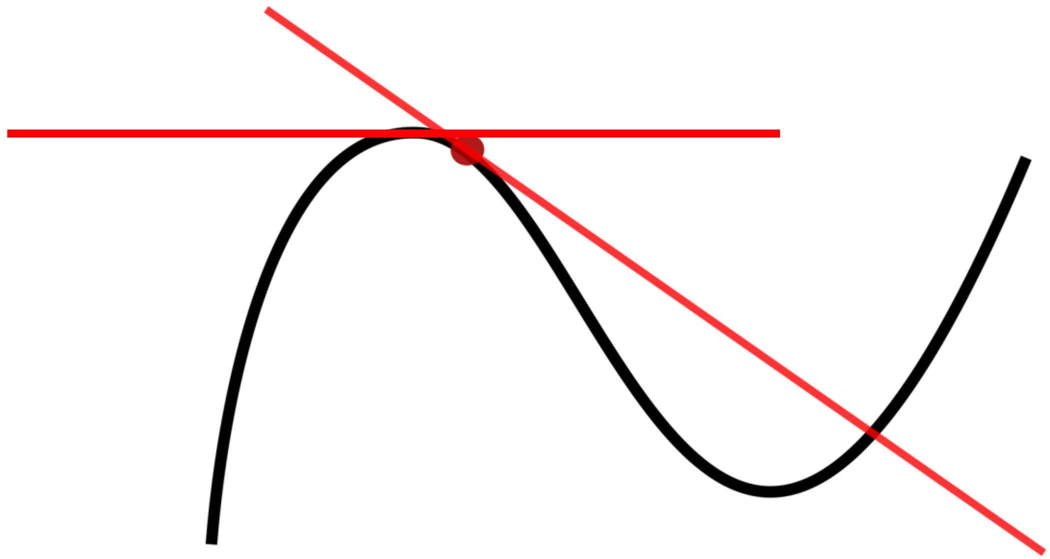
$$f'(x_0) = 0$$



Производная (3/3)

Если точка x_0 — экстремум и в ней существует производная, то

$$f'(x_0) = 0$$

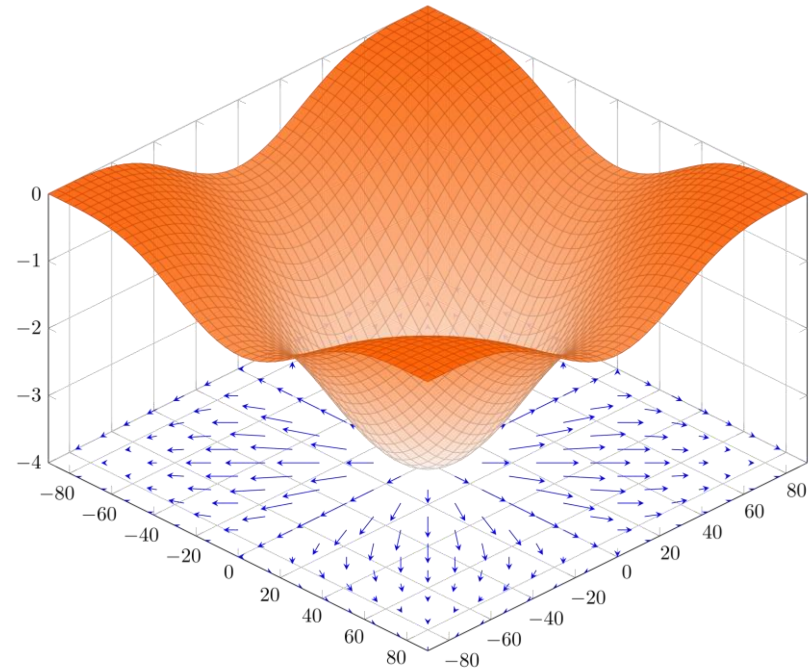


Градиент

Градиент — вектор, своим направлением указывающий направление наискорейшего роста некоторой скалярной величины.

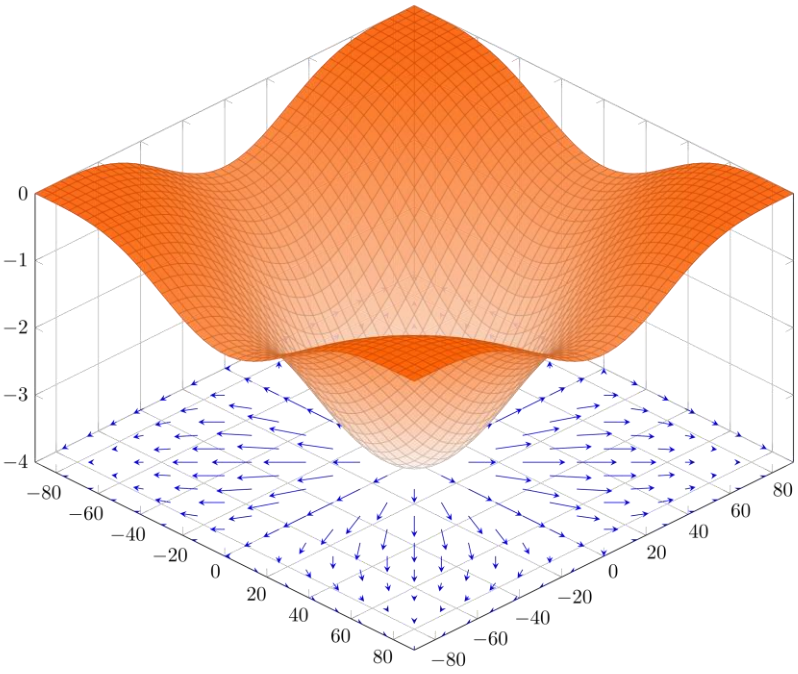
Компоненты - частные производные

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

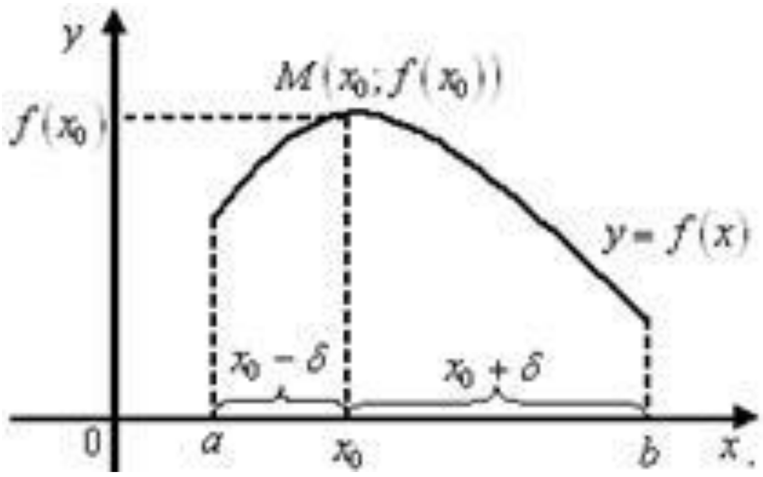


Важное свойство градиента

- В точке x_0 функция быстрее всего растёт в направлении градиента
- Если градиент равен нулю, то это экстремум



Определение экстремума
вспоминаем математический анализ



<u>Теорема 3.1.</u> Для того, чтобы дифференцируемая на (a, b) функция $f(x)$ не убывала (не возрастала) на этом интервале, необходимо и достаточно, чтобы $\nabla f(x_0) \geq 0$ ($\nabla f(x_0) \leq 0$) $\forall x \in (a, b)$.

Если $\nabla f(x_0) > 0$, то $f(x)$ возрастает.
Если $\nabla f(x_0) < 0$, то $f(x)$ убывает.

Точка x_0 - точка локального экстремума:
$f(x) - f(x_0) < 0$ – локальный максимум,
$f(x) - f(x_0) > 0$ – локальный минимум.
Наименьшее и наибольшее значения функции $f(x)$ на $[a, b]$ - *абсолютные* минимум и максимум или *глобальные* экстремумы функции $f(x)$.

Необходимое условие существования локального экстремума (1/2)

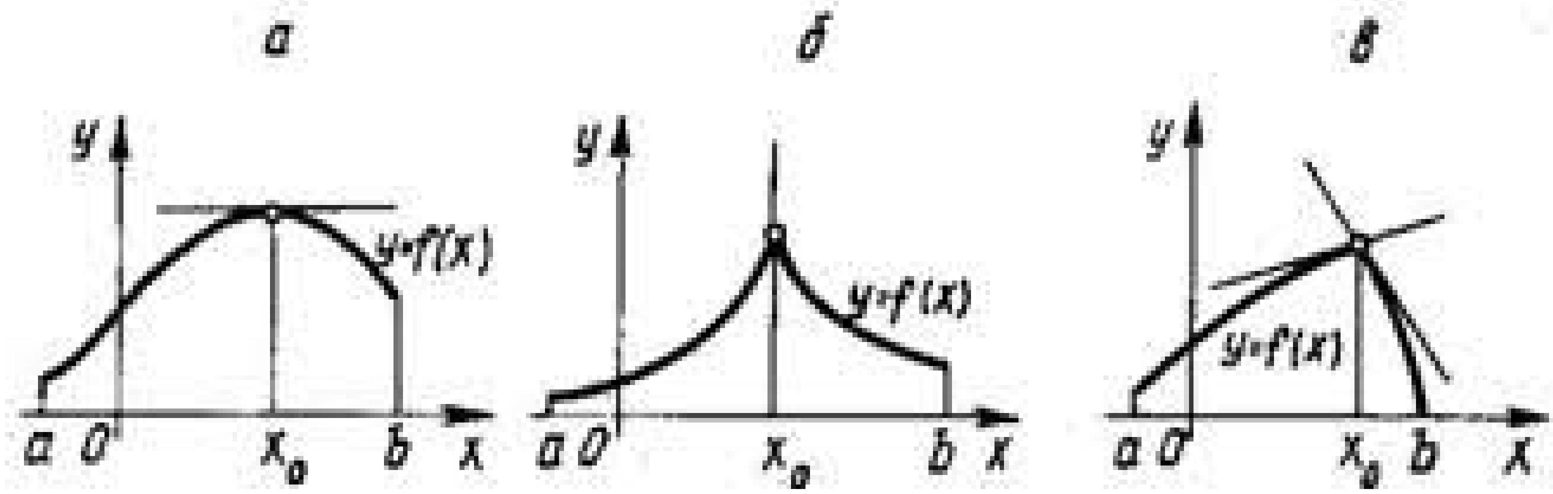
вспоминаем математический анализ

<u>Теорема 3.2.</u> Если в точке x_0 функция $f(x)$ имеет экстремум, то её производная в ней либо равна нулю $\nabla f(x_0) = 0$, либо не существует.

Точка x_0 - *критическая* точка, точка возможного экстремума.

Необходимое условие существования локального экстремума (2/2)

вспоминаем математический анализ



касательная параллельна оси абсцисс,

$\nabla f(x_0) = 0$,
x_0 - *стационарная* точка

касательная параллельна оси ординат,
$\nabla f(x_0) = \infty$,
x_0 - точка *возврата*

существуют не совпадающие левая и правая касательные,

$\nabla f(x_0 - 0) \neq \nabla f(x_0 + 0)$,
x_0 - *угловая* точка

Достаточные условия (признаки) экстремума (1/3)

вспоминаем математический анализ

Теорема 3.3 (первый достаточный признак существования экстремума функции).

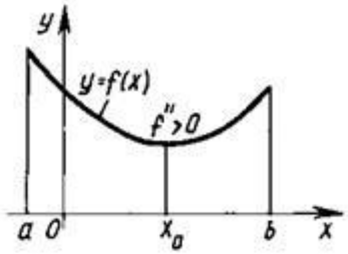
Если производная $\nabla f(x)$ непрерывной функции $f(x)$ при переходе через критическую точку x_0 меняет знак с «+» на «–», то x_0 - точка локального максимума,
с «-» на «+», то x_0 - точка локального минимума,
не меняет знак, то x_0 - не точка локального экстремума.

Достаточные условия (признаки) экстремума (2/3)

вспоминаем математический анализ

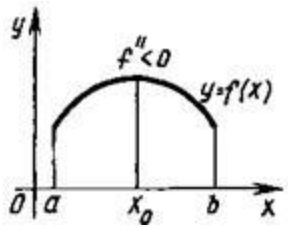
<u>Теорема 3.4</u> (второй достаточный признак существования экстремума функции).

Стационарная точка x_0 функции $f(x)$, дважды дифференцируемой в её δ-окрестности $[x_0 - \delta, x_0 + \delta]$, является



точкой локального минимума, если $f''(x_0) > 0$,

точкой локального максимума, если $f''(x_0) < 0$.



Достаточные условия (признаки) экстремума (3/3)

вспоминаем математический анализ

Теорема 3.5 (третий достаточный признак существования экстремума функции).

Пусть функция $f(x)$ n раз непрерывно дифференцируема в точке x_0 и в этой точке
$f'(x_0)= f''(x_0)=\ldots= f^{(n-1)}(x_0) = 0$, и $f^{(n)}(x_0) \neq 0$.

Если n чётное и $f^{(n)}(x_0) < 0$, то x_0 - точка локального максимума,
если n чётное и $f^{(n)}(x_0) > 0$, то x_0 - точка локального минимума,
если n нечётное, то x_0 - не точка локального экстремума.

Среднеквадратичная ошибка

- MSE для линейной регрессии:

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

Градиент

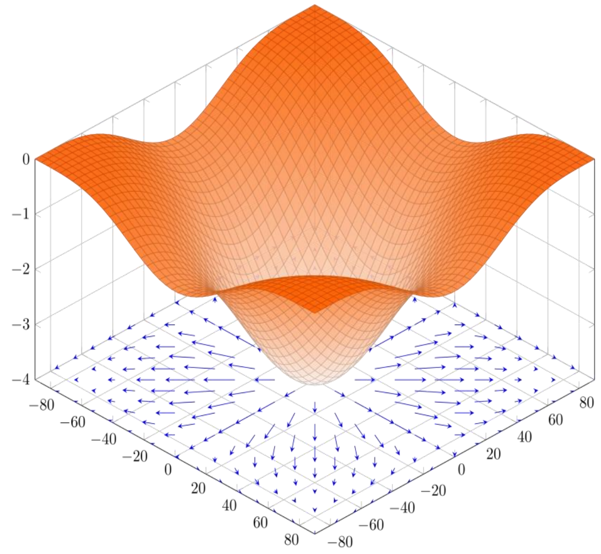
- Градиент — вектор частных производных

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

- У градиента есть важное свойство!

Важное свойство

- Зафиксируем точку x_0
- Градиент показывает, в какую сторону функция быстрее всего растёт и с какой скоростью.

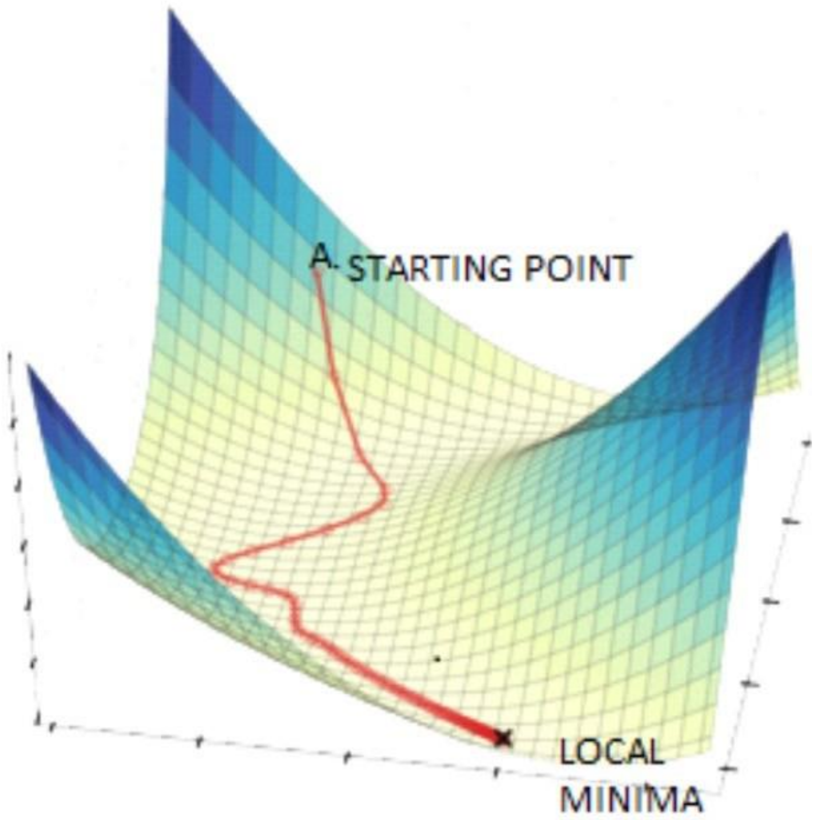


Важное свойство

- Зафиксируем точку x_0
- В какую сторону функция быстрее всего растёт?
- Быстрее всего растёт в направлении градиента.
- Быстрее всего *убывает* в направлении антиградиента – градиентный *спуск*.

Как это пригодится?

Как это пригодится?



Градиентный спуск
(Gradient Descent, GD)

Градиентный спуск

Градиентный спуск – алгоритм оптимизации, используемый для разнообразного обучения модели машинного обучения. Подходит для задач, в которых имеется большое число функций и большая обучающая выборка, широко используется во многих моделях МО, включая линейную регрессию, логистическую регрессию и нейронные сети.

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

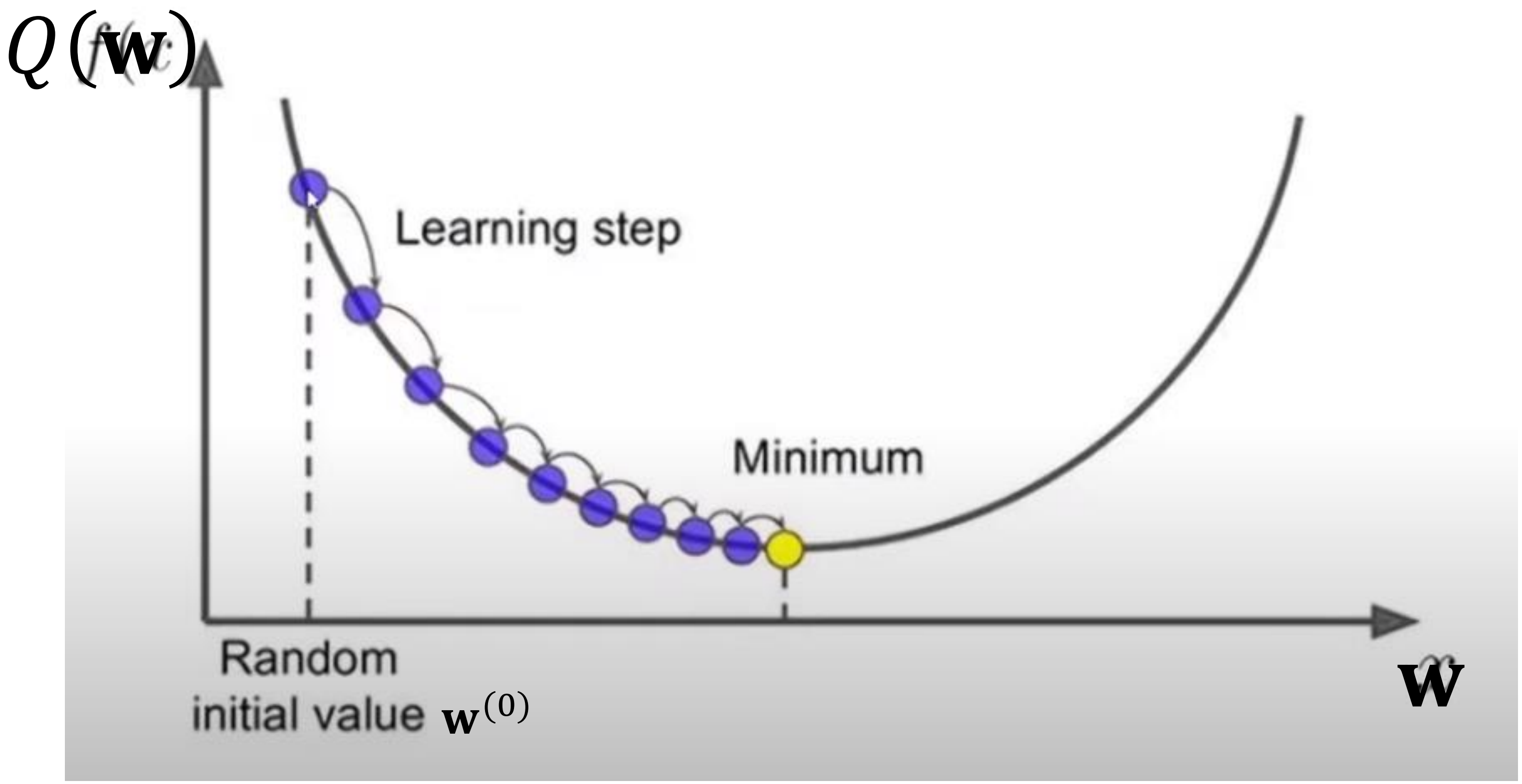
Градиентный спуск для парной регрессии (1/4)

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

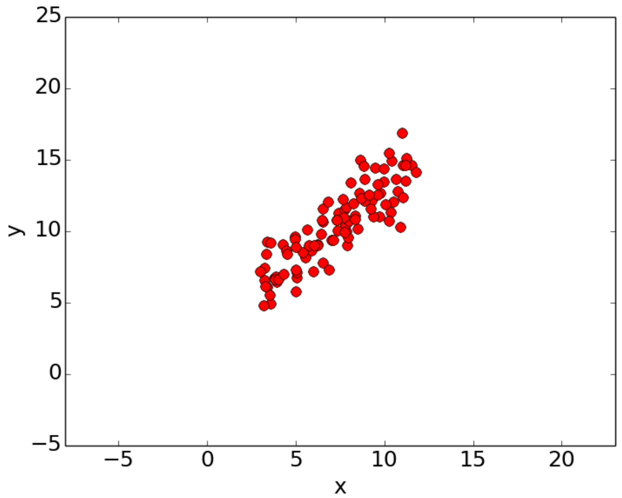
$$Q(w_0, w_1) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

Градиентный спуск для парной регрессии (2/4)

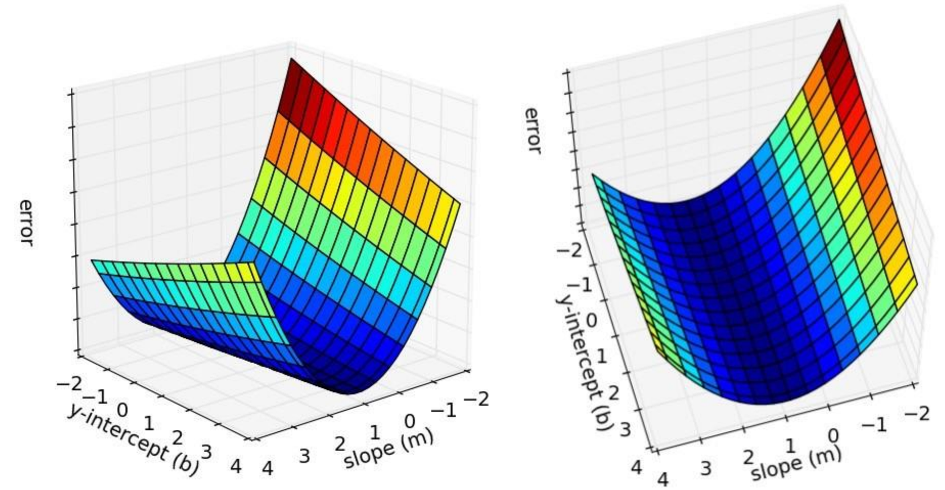
Ищем решение задачи оптимизации: $\mathbf{w}^* = \arg\min\limits_{\mathbf{w}\in\mathbb{R}^d} Q(\mathbf{w})$



Градиентный спуск для парной регрессии (3/4)



Выборка



Функционал ошибки

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

Градиентный спуск для парной регрессии (4/4)

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

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

Начальное приближение

- $\mathbf{w}^{(0)}$ — инициализация весов (начальное приближение)
- Например, из стандартного нормального распределения

Градиентный спуск

- Повторять до сходимости:

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

Новая точка

Размер шага

Градиент в предыдущей точке

Сходимость

- Критерий 1 останова:

$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\right\| < \varepsilon$$

- Критерий 2 останова:

$$\left\|\nabla Q\left(\mathbf{w}^{(t)}\right) =\right\| < \varepsilon$$

- Критерий 3 останова: когда ошибка на отложенной выборке перестаёт уменьшаться

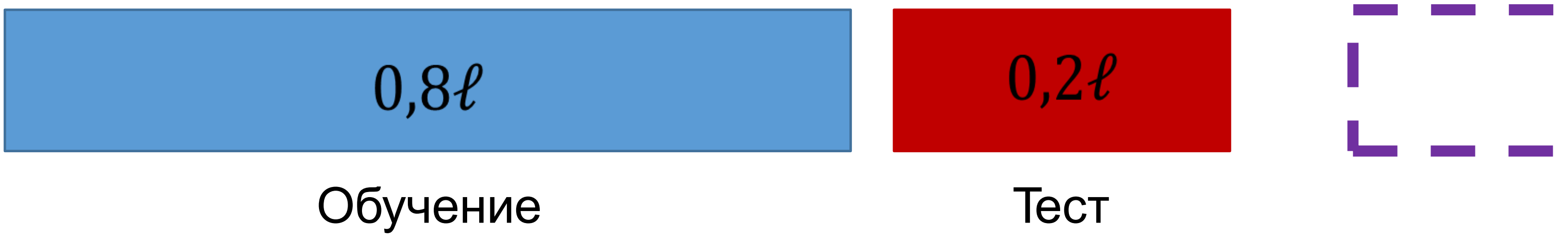
Отложенная выборка (1/2)

$$X^{\ell} = (\mathbf{x}_i, y_i)_{i=1}^{\ell}$$

тестовая выборка может быть недоступна разработчику модели, тогда тестирование проводится на валидационной выборке

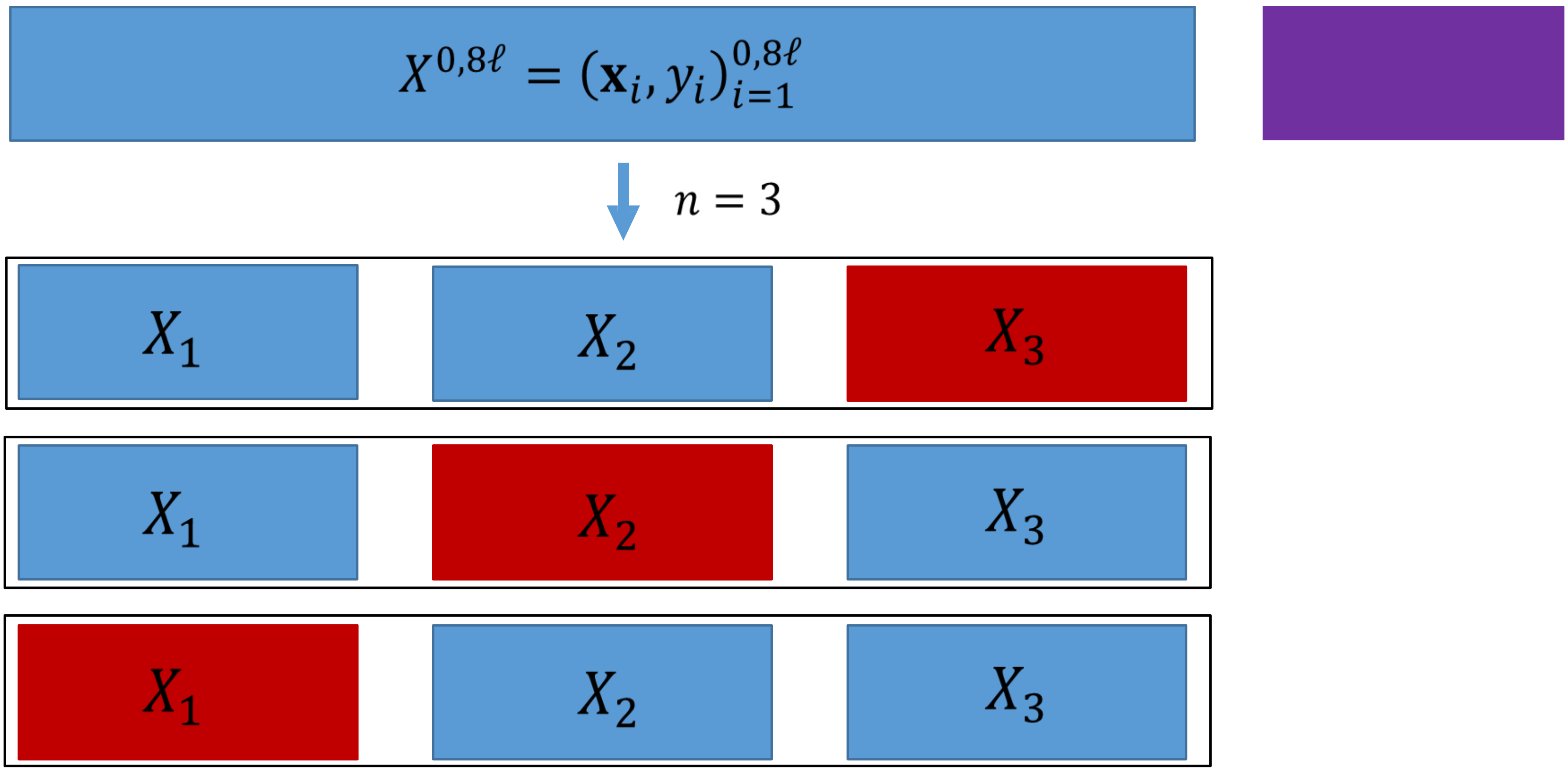
отложенная выборка - для валидации модели и финальной оценки модели

Отложенная выборка (2/2)

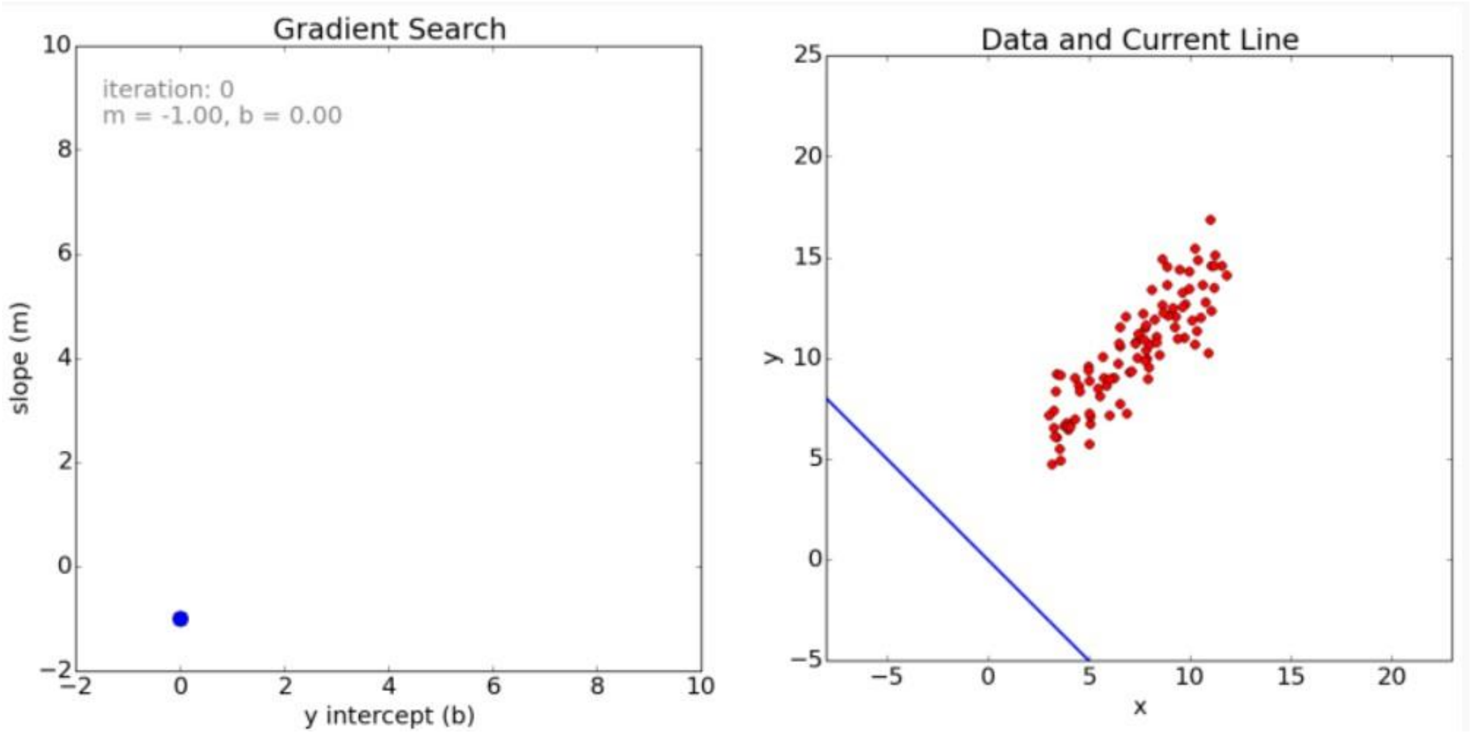


- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не успеет обучиться
- Обычно: 70/30 или 80/20

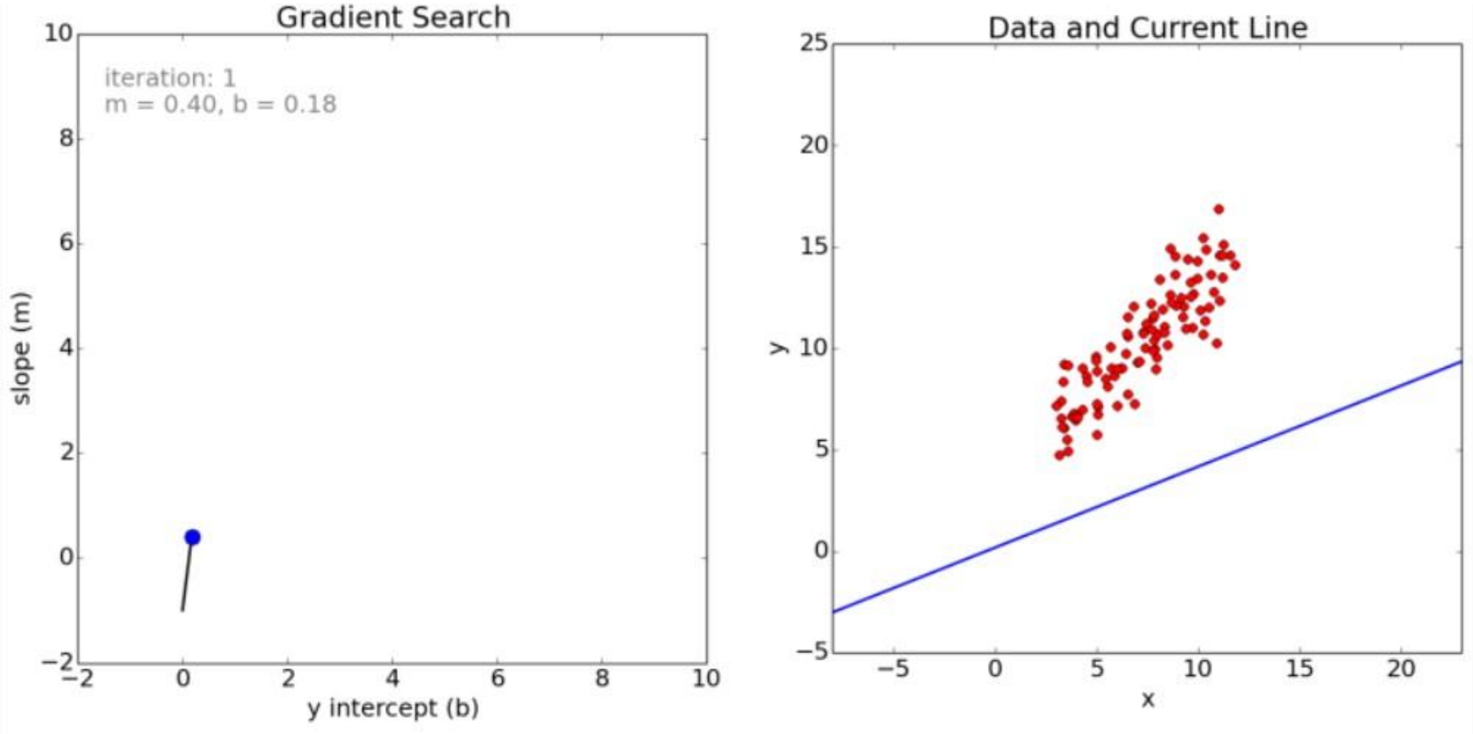
Кросс-валидация



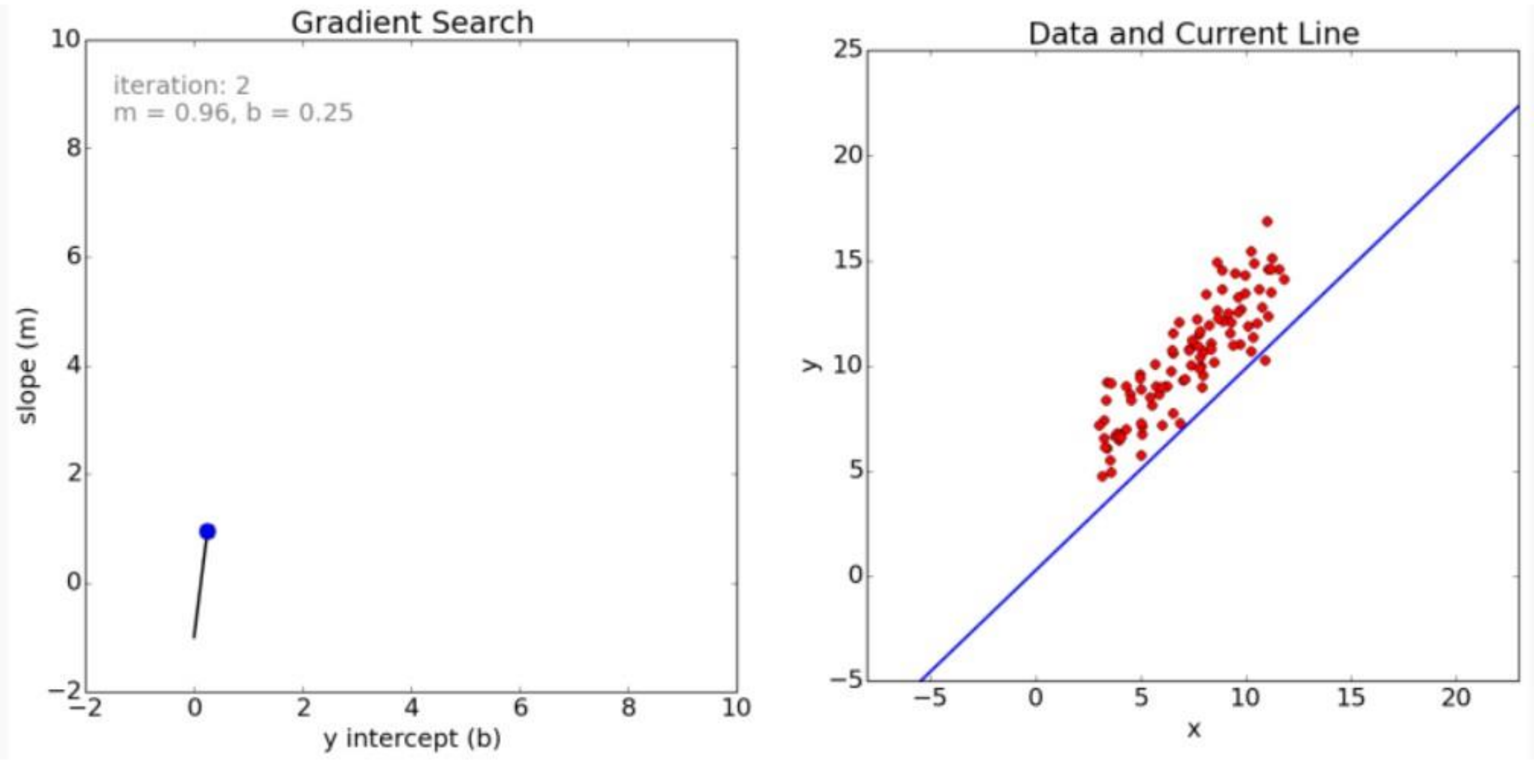
Парная регрессия (1/6)



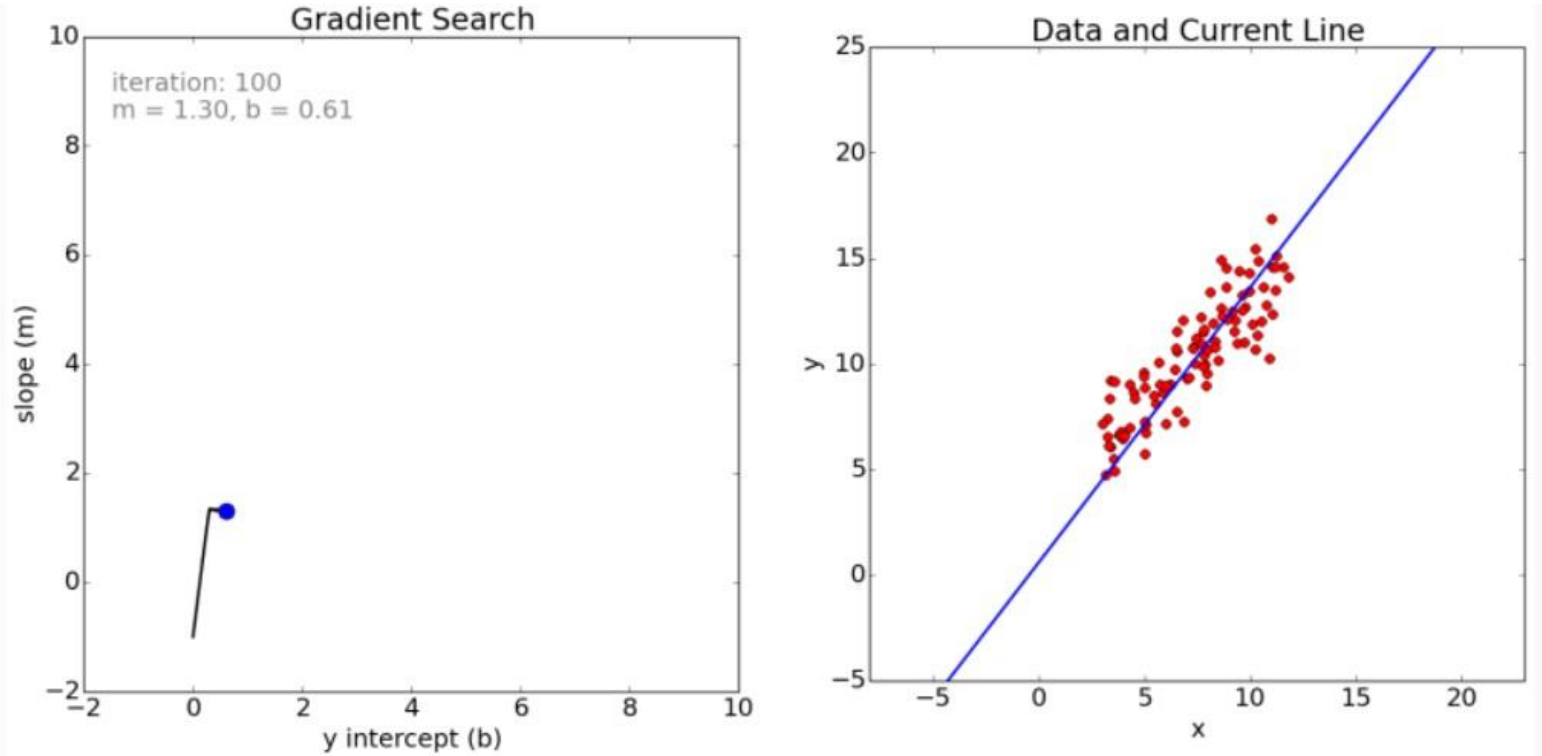
Парная регрессия (2/6)



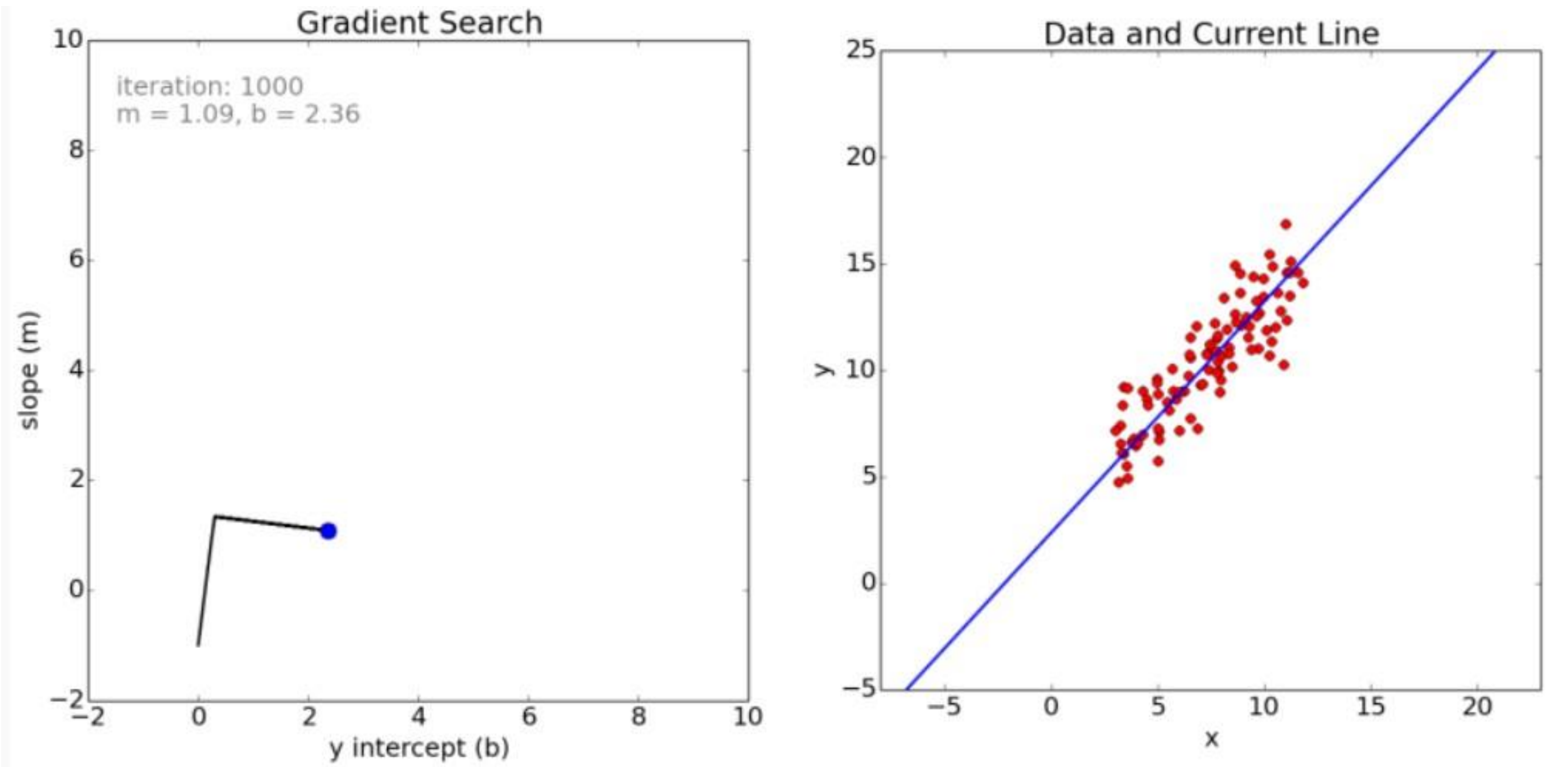
Парная регрессия (3/6)



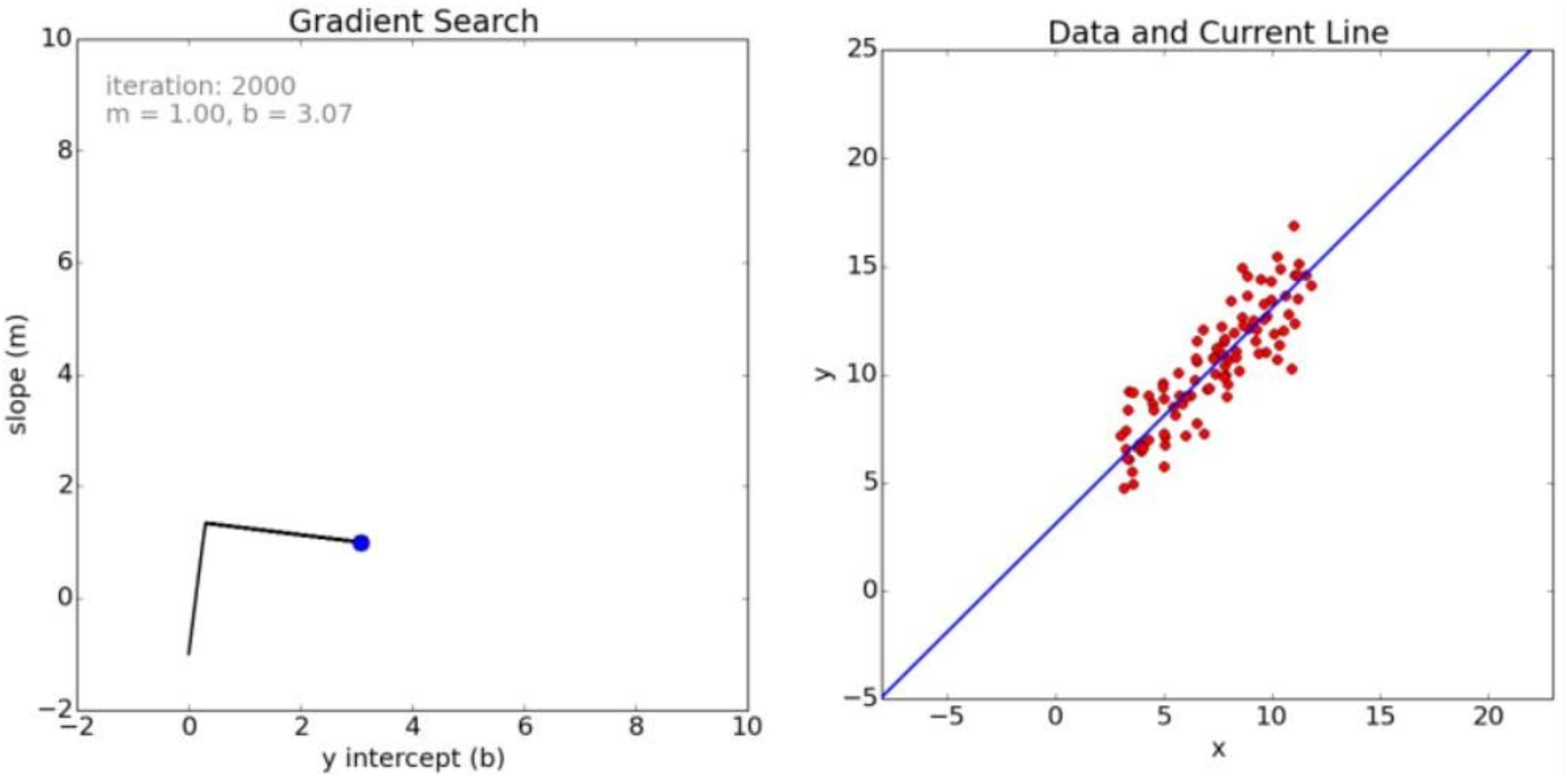
Парная регрессия (4/6)



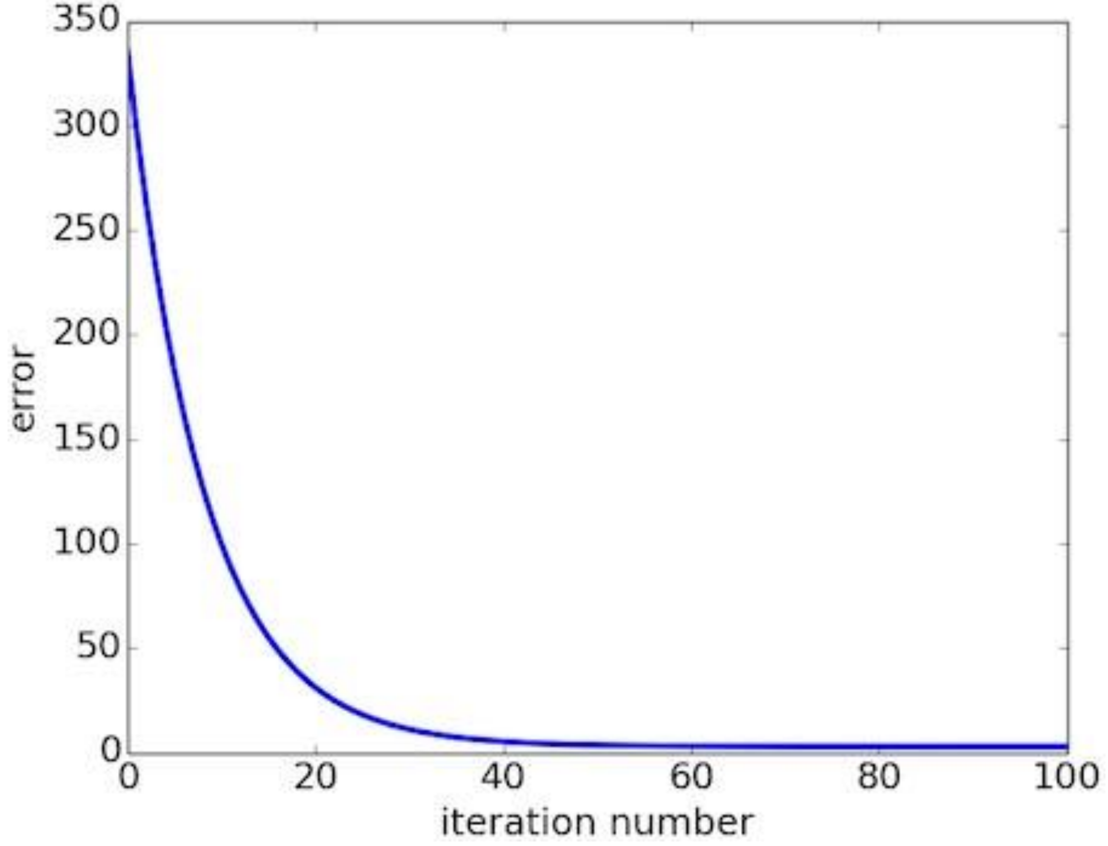
Парная регрессия (5/6)



Парная регрессия (6/6)



Функционал ошибки



Линейная регрессия, d >2

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

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell}\sum_{i=1}^{\ell} x_{i_1}(\langle \mathbf{w}, \mathbf{x}\rangle - y_i)$
- …
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell}\sum_{i=1}^{\ell} x_{i_d}(\langle \mathbf{w}, \mathbf{x}\rangle - y_i)$
- $\nabla Q(\mathbf{w}) = \frac{2}{\ell}X^T(\mathbf{Xw} - \mathbf{y})$

Градиентный спуск

1. Начальное приближение: $\mathbf{w}^{(0)}$

2. Повторять:

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

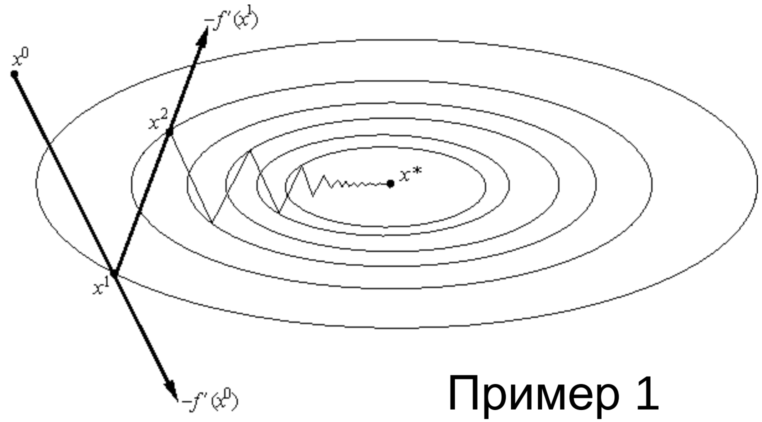
3. Останавливаемся, если

$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\right\| < \varepsilon$$

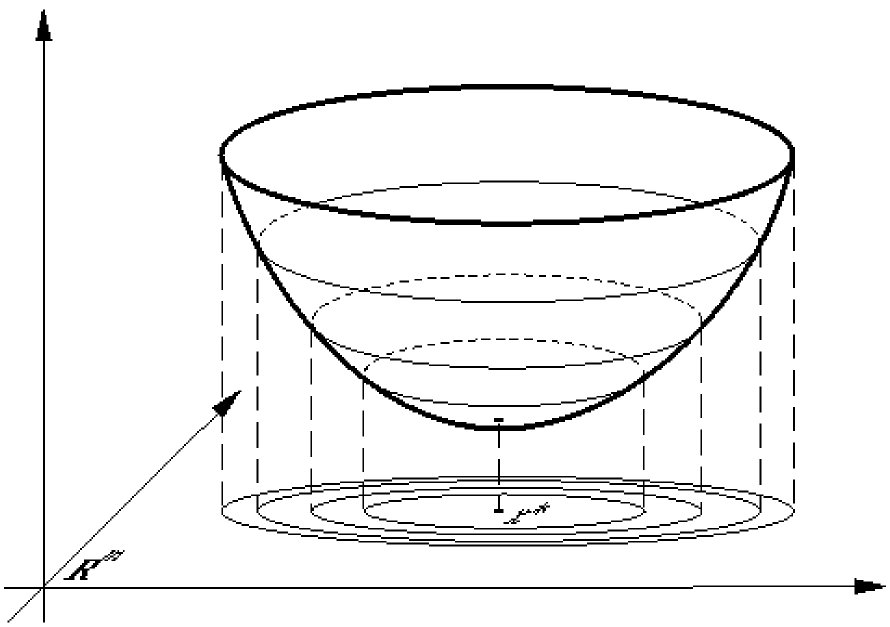
Примеры траекторий градиентного спуска

Ищем решение задачи оптимизации:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x})$$



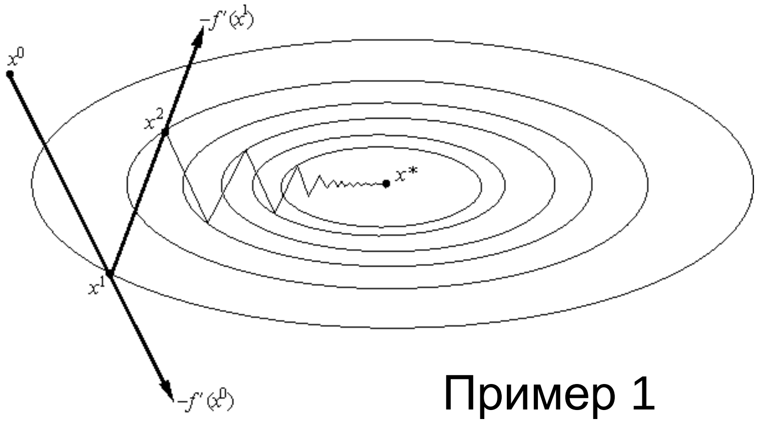
Пример 1



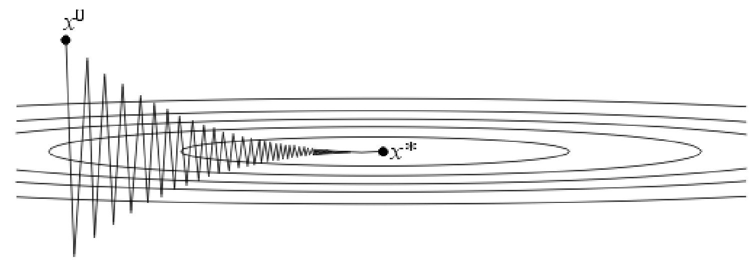
Примеры траекторий градиентного спуска

Ищем решение задачи оптимизации:

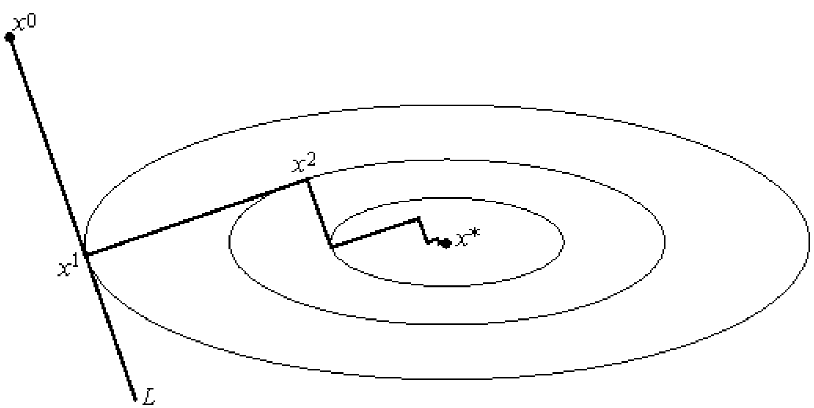
$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} Q(\mathbf{x})$$



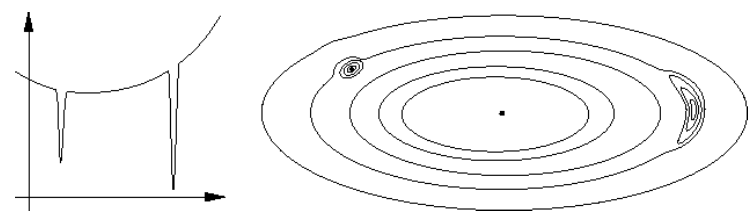
Пример 1



Пример 3



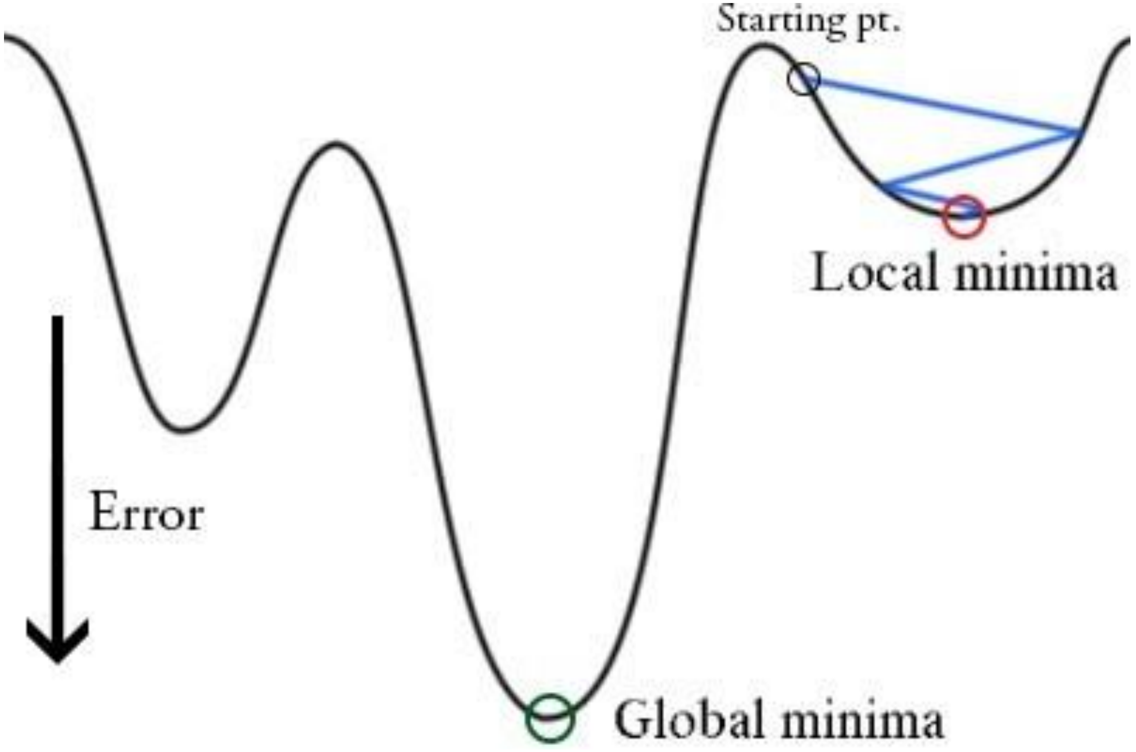
Пример 2



Пример 4

Локальные минимумы (1/3)

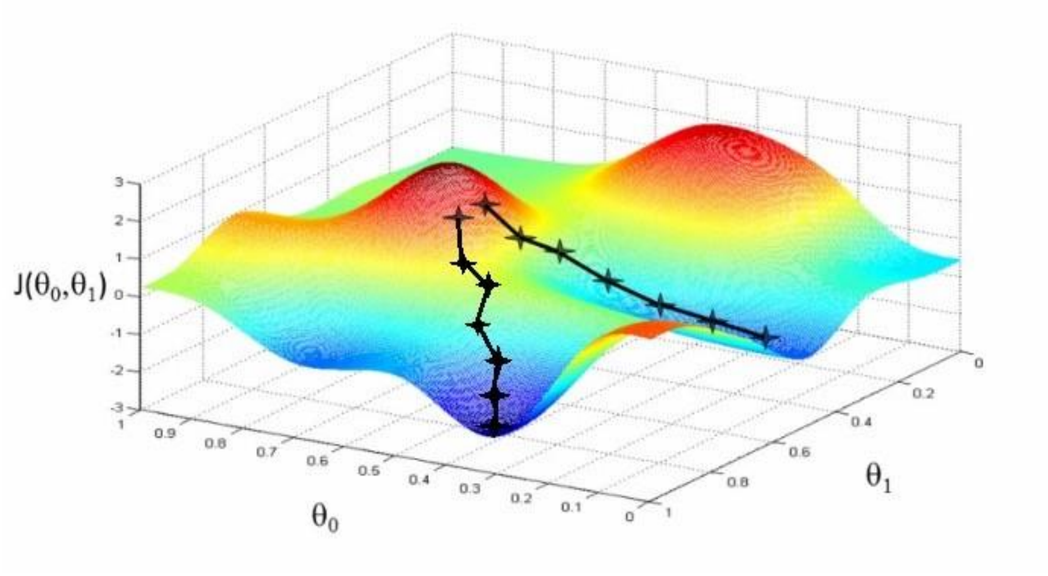
- Градиентный спуск находит только локальные минимумы



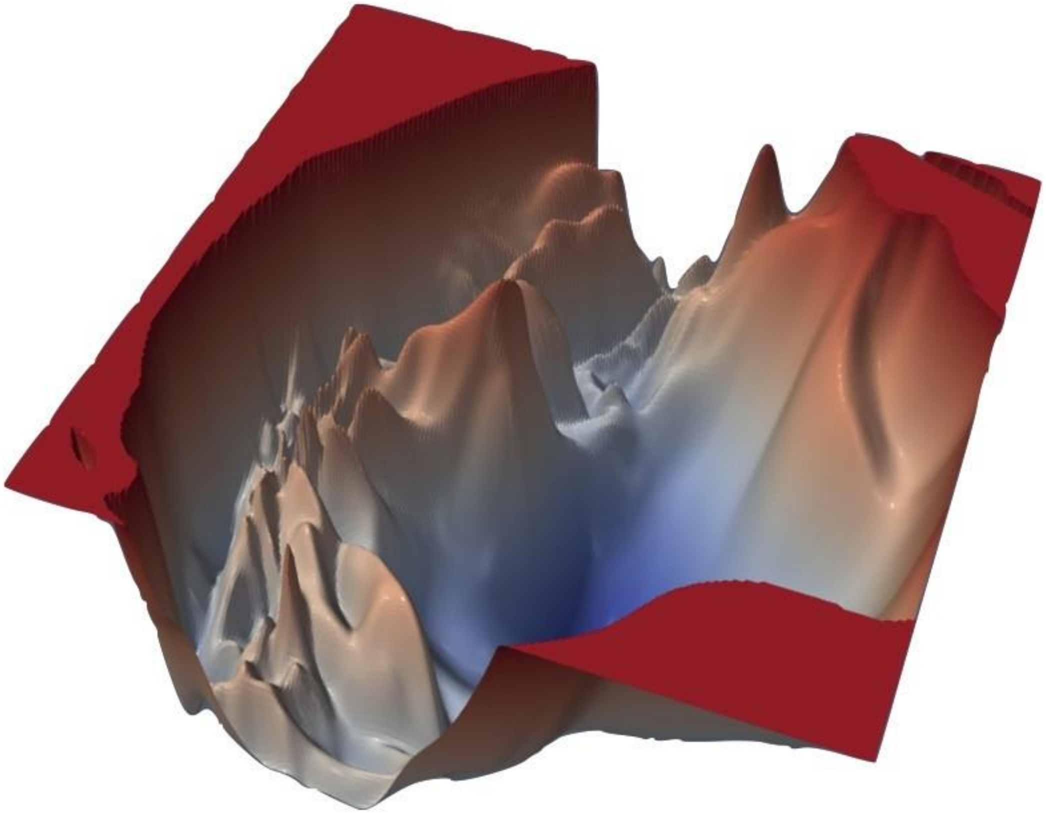
Локальные минимумы (2/3)

- Градиентный спуск находит **локальный минимум**

- *Мультистарт* (запуск градиентного спуска из разных начальных точек) может улучшить результат



Локальные минимумы (3/3)



Длина шага (1/7)

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

- Позволяет контролировать скорость обучения

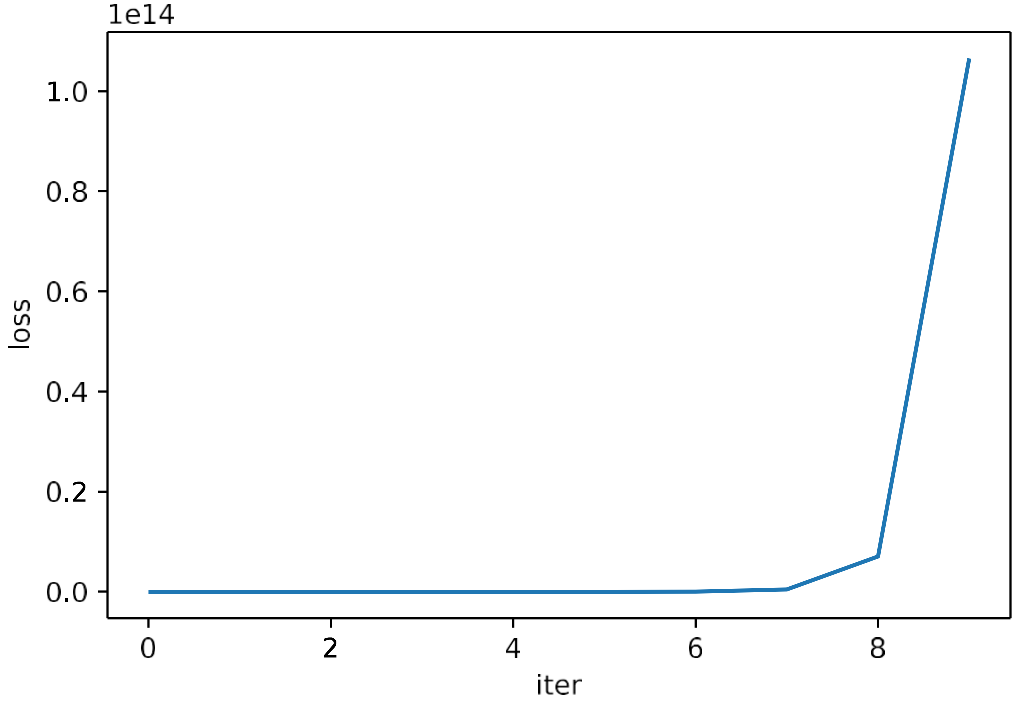
Длина шага (2/7)

```
[[  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 (3/7)

Градиент на первом шаге:

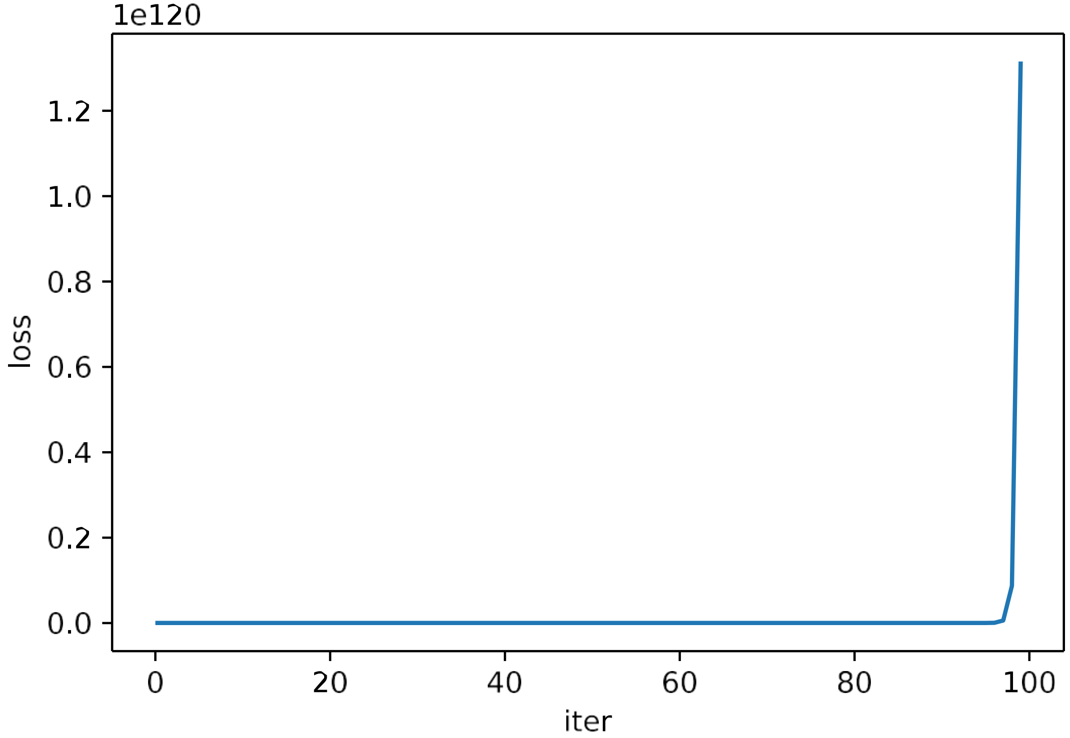
[26.52, 564.80, 682.90, 5097.71, 12110.87]



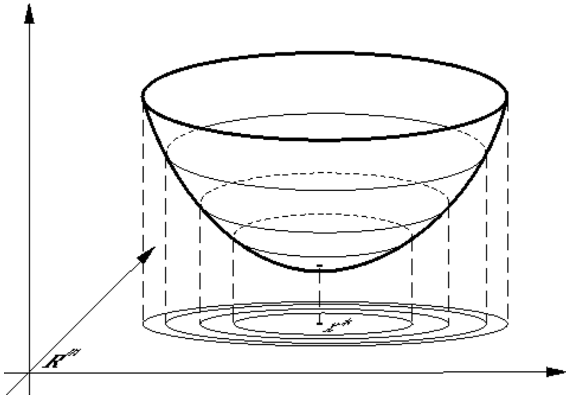
Длина шага (4/7)

Градиент на первом шаге:

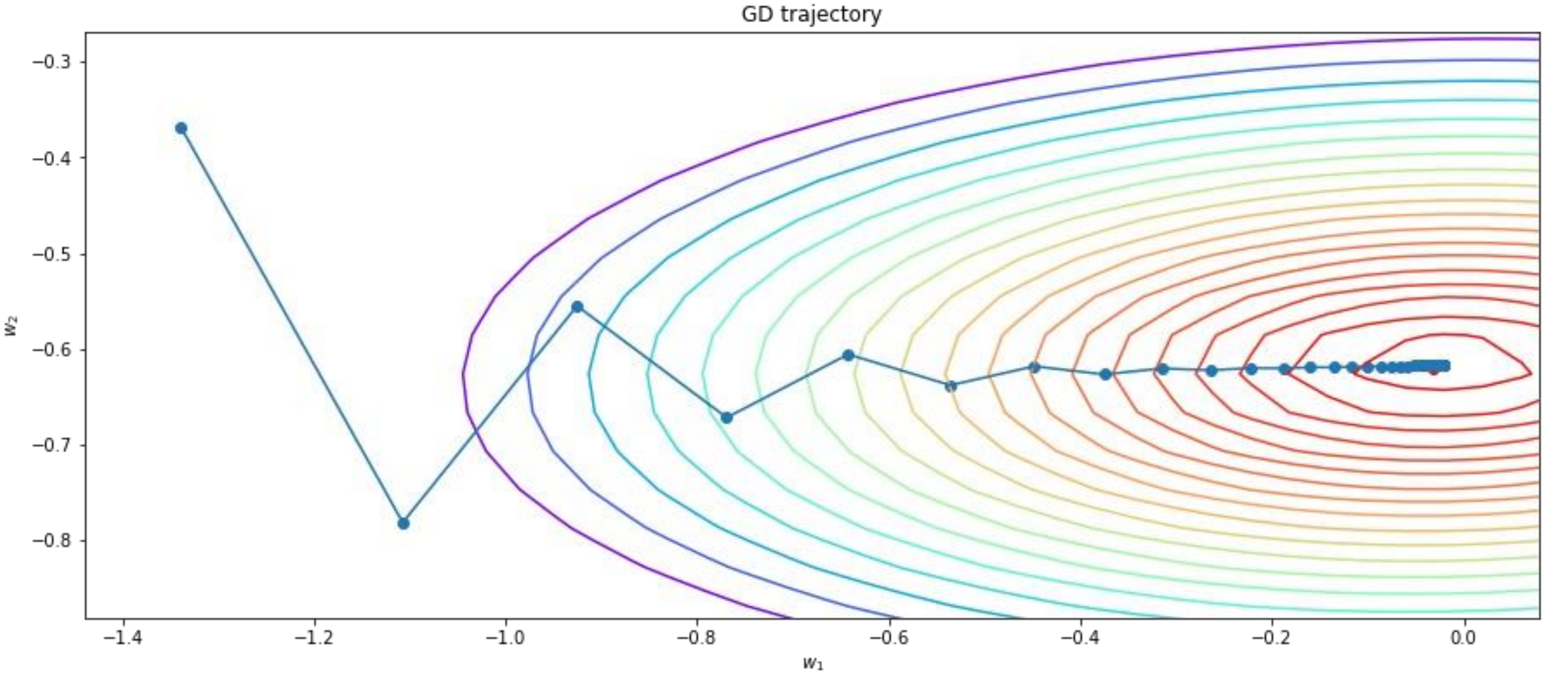
[26.52, 564.80, 682.90, 5097.71, 12110.87]



Длина шага (5/7)

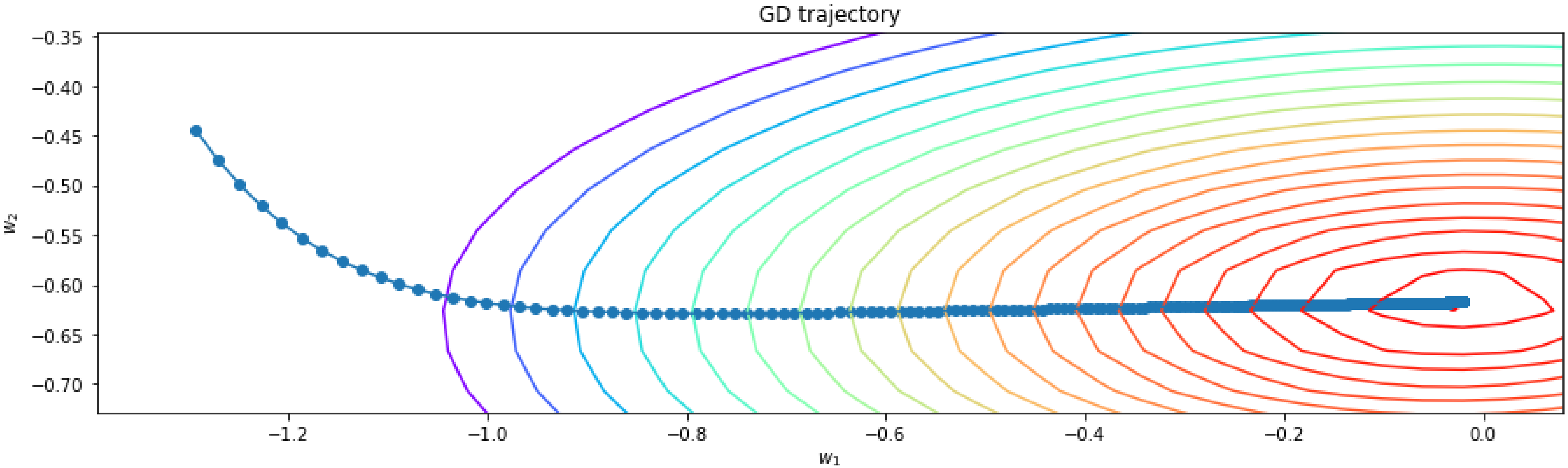


Более длинный шаг – ломаная.



Длина шага (6/7)

Менее длинный шаг – гладкая кривая.



Длина шага (7/7)

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

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

Переменная длина шага

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

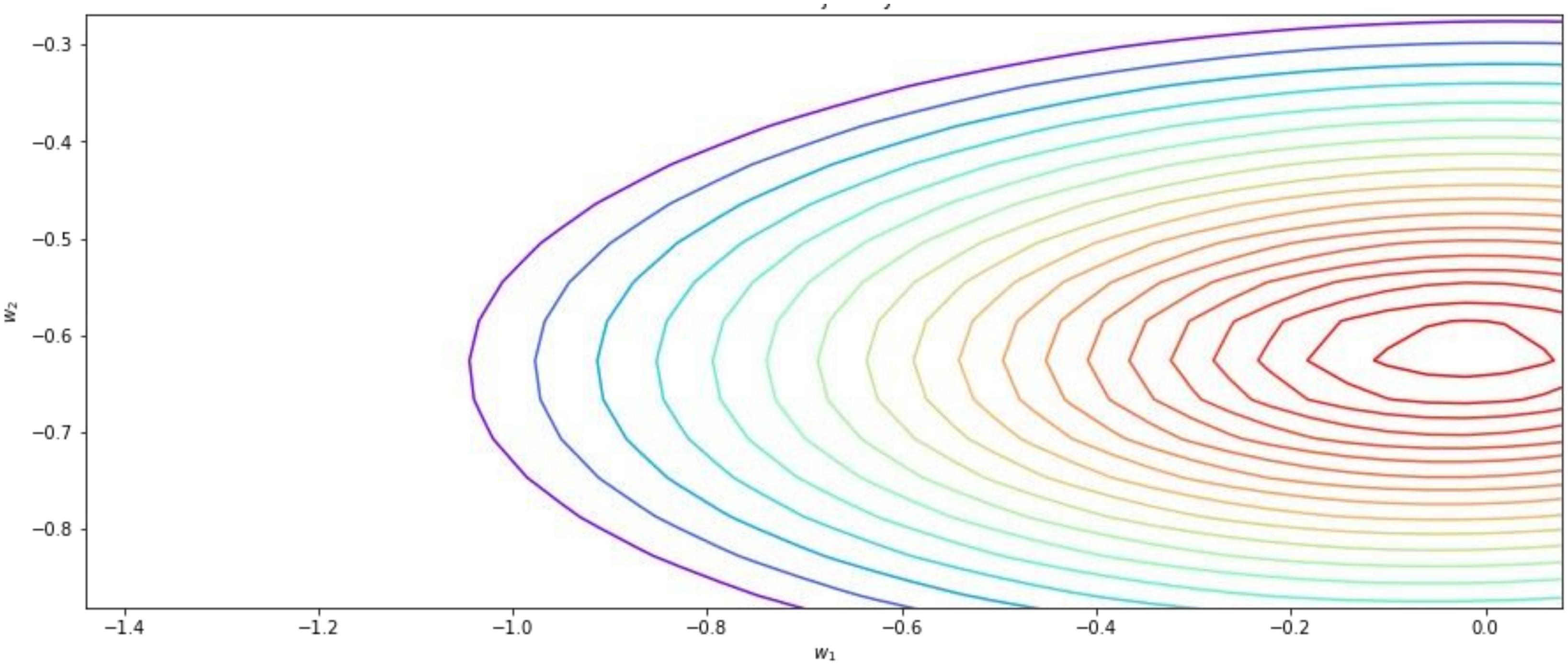
- Длину шага можно менять в зависимости от шага
- Например: $\eta_t = \frac{1}{t}$
- Ещё вариант: $\eta_t = \lambda \left(\frac{s}{s+t}\right)^p$

Масштабирование признаков (1/4)

```
[[  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]]
```

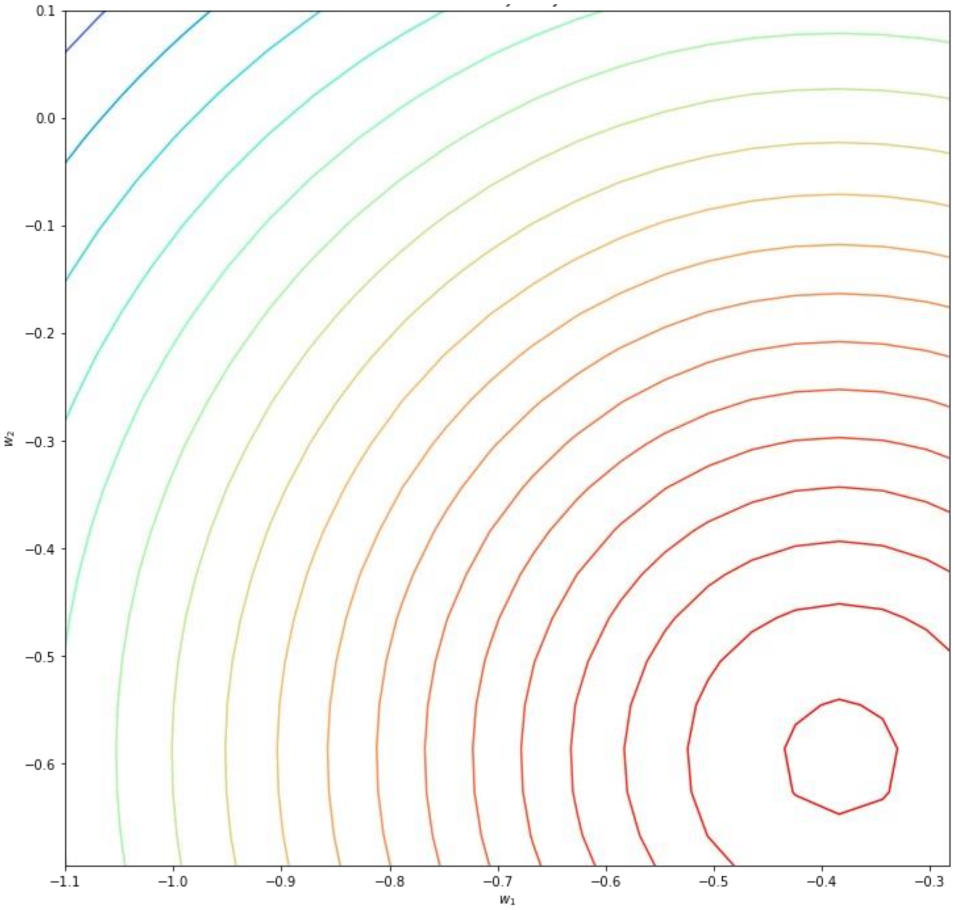
Масштабирование признаков (2/4)

Без масштабирования:



Масштабирование признаков (3/4)

После масштабирования:



Масштабирование признаков (1/2)

1. Для j-го признака, $j = 1, \dots, d$, вычисляем среднее значение и стандартное (среднеквадратическое) отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_{ij}$$

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

Масштабирование признаков (2/2)

2. Вычитаем из каждого значения признака среднее по выборке и делим на стандартное отклонение по выборке:

$$\hat{x}_{ij} := \frac{x_{ij} - \mu_j}{\sigma_j}, \qquad i = 1, \dots, \ell, \qquad j = 1, \dots, d$$

Стохастический градиентный спуск (Stochastic Gradient Descent, SGD)

Градиентный спуск

1. Начальное приближение: $\mathbf{w}^{(0)}$

2. Повторять:

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

3. Останавливаемся, если

$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\right\| < \varepsilon$$

Линейная регрессия

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

- $\frac{\partial Q}{\partial w_1} = \frac{2}{\ell}\sum_{i=1}^{\ell} x_{i_1}(\langle \mathbf{w}, \mathbf{x}\rangle - y_i)$
- $\dots$
- $\frac{\partial Q}{\partial w_d} = \frac{2}{\ell}\sum_{i=1}^{\ell} x_{i_d}(\langle \mathbf{w}, \mathbf{x}\rangle - y_i)$
- $\nabla Q(\mathbf{w}) = \frac{2}{\ell} X^T(\mathbf{Xw} - \mathbf{y})$

Сложности градиентного спуска

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

Оценка градиента

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

- Градиент:

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

- Может, оценить градиент одним слагаемым?

$$\nabla Q(\mathbf{w}) \approx \nabla L(a(\mathbf{x}_i), y_i)$$

Стохастический градиентный спуск
Stochastic Gradient Descent

1. Начальное приближение: $\mathbf{w}^{(0)}$

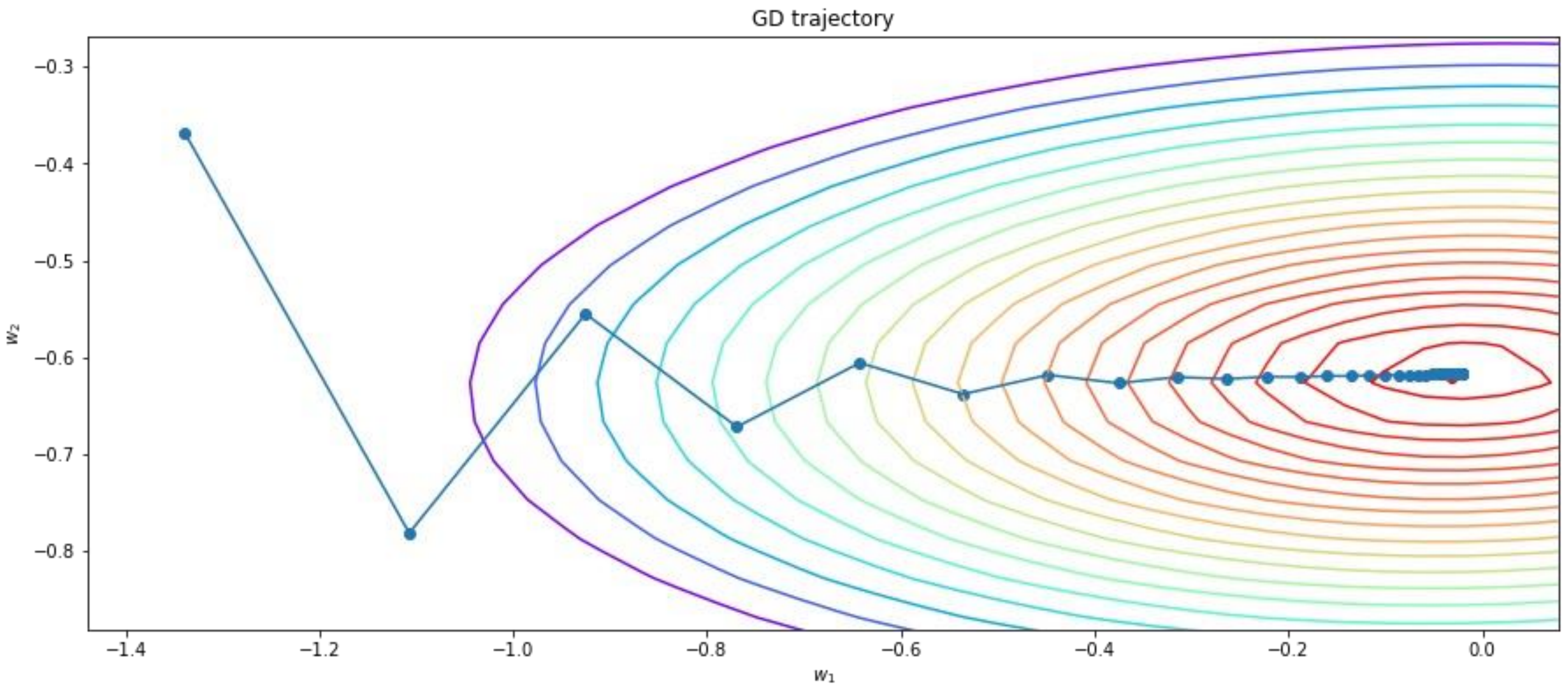
2. Повторять, каждый раз выбирая случайный объект i_t:

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta \nabla L\big(a(\mathbf{x}_{i_t}), y_{i_t}\big)$$

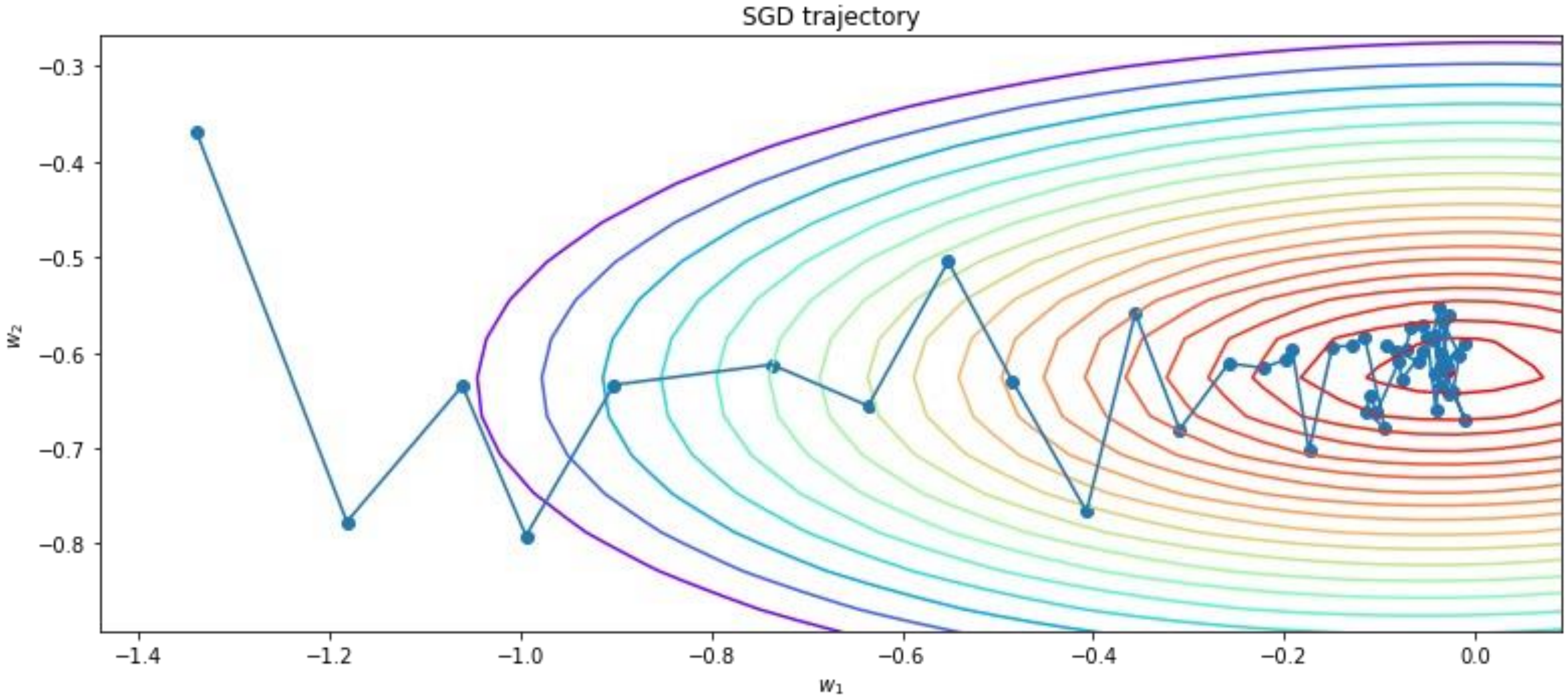
3. Останавливаемся, если

$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\right\| < \varepsilon$$

Градиентный спуск



Стохастический градиентный спуск



Стохастический градиентный спуск (1/3)

1. Начальное приближение: $\mathbf{w}^{(0)}$

2. Повторять, каждый раз выбирая случайный объект i_t:

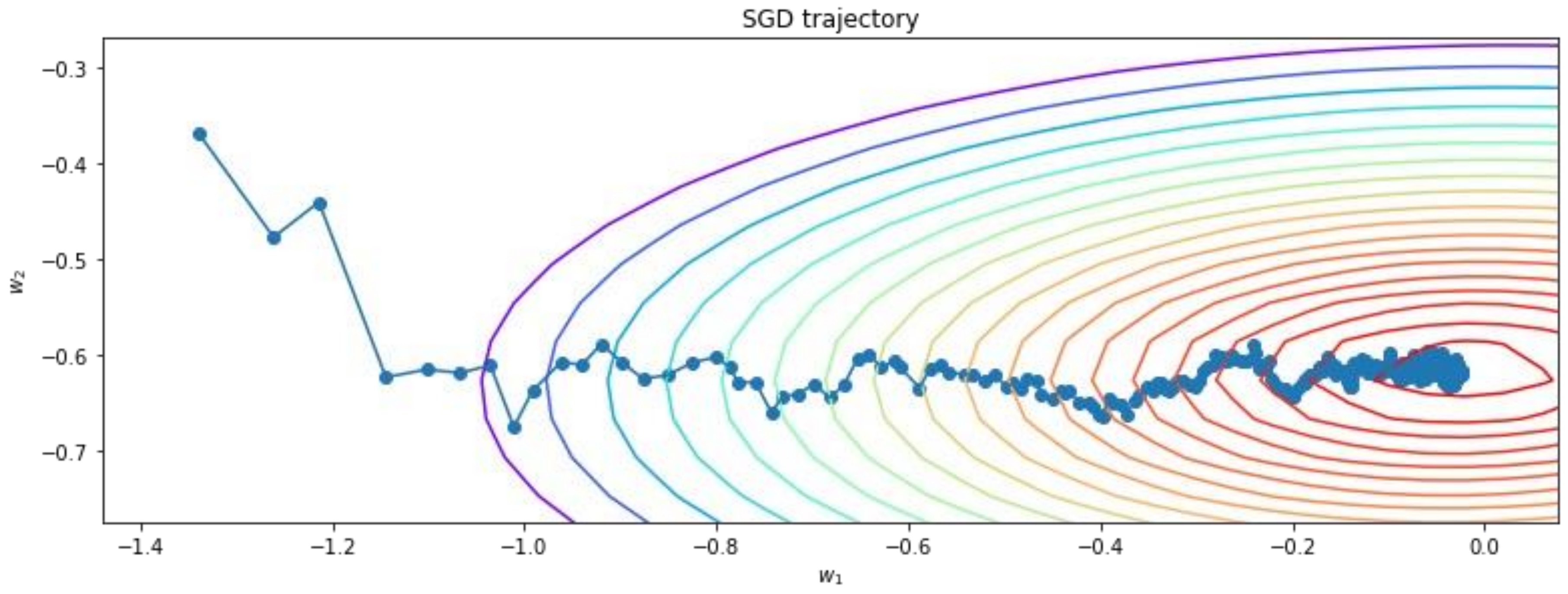
$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta_t \nabla L\left(a(\mathbf{x}_{i_t}), y_{i_t}\right)$$

3. Останавливаемся, если

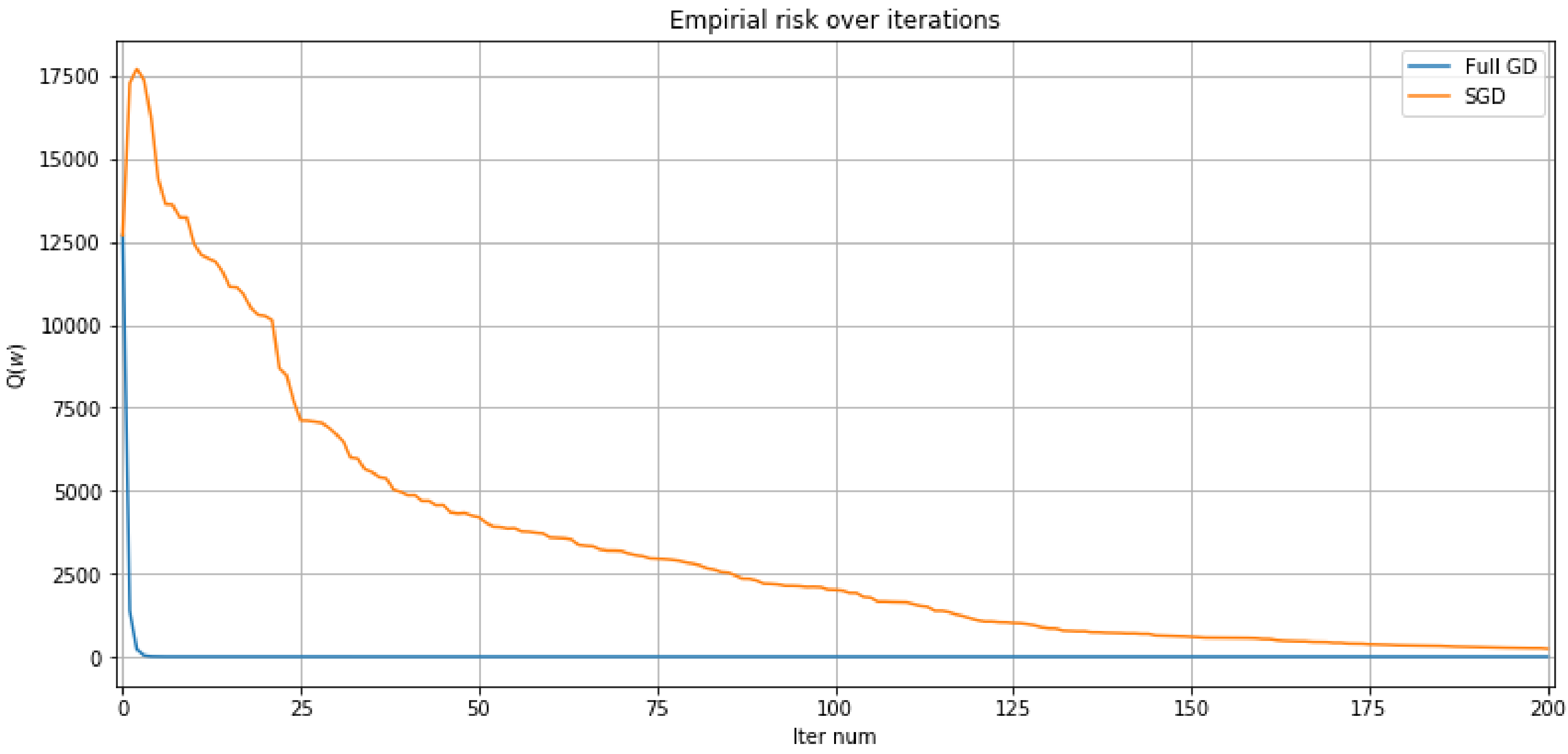
$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\right\| < \varepsilon$$

Стохастический градиентный спуск (2/3)

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



Стохастический градиентный спуск (3/3)



Мини-пакетный (mini-batch) градиентный спуск

1. Начальное приближение: $\mathbf{w}^{(0)}$

2. Повторять, каждый раз выбирая m случайных объектов $i_1, \dots, i_m$:

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} - \eta_t \frac{1}{m} \sum_{j=1}^{m} \nabla L\big(a(\mathbf{x}_{i_t}), y_{i_t}\big)$$

3. Останавливаемся, если

$$\left\| \mathbf{w}^{(t)} - \mathbf{w}^{(t-1)} \right\| < \varepsilon$$

Градиентный спуск: выводы

- Выбор скорости обучения

  Скорость обучения (размер шага) является важным гиперпараметром. Высокая скорость обучения может привести к превышению минимума, низкая - к медленной сходимости. Для получения хороших результатов важен тщательный выбор скорости обучения.

- Сходимость и переобучение

  Градиентный спуск может сходиться к минимуму функции потерь, но он также может «застрять» в локальном минимуме, что приведет к переобучению модели на обучающей выборке. Чтобы избежать переобучения, можно использовать методы регуляризации, такие как регуляризация L1 или L2.

- Градиентный спуск может быть дорогостоящим с точки зрения вычислений

  Вычисление градиента функции потерь и обновление параметров на каждой итерации градиентного спуска может быть дорогостоящим с точки зрения вычислений, особенно для больших наборов данных или сложных моделей. Чтобы сократить время вычислений, можно использовать такие методы, как параллельные вычисления и векторизация.