

# Введение в анализ данных

## Лекция 5

### Линейная регрессия и методы оптимизации

Евгений Соколов

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

НИУ ВШЭ, 2018

# Градиент

- Градиент — вектор из частных производных:

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

- И зачем нам этот вектор?
- У градиента есть очень важное свойство!

# Градиент

- Зафиксируем точку  $x_0$
- В каком направлении функция быстрее всего растёт?

$$f'_v(x_0) \rightarrow \max_v$$

Угол между градиентом и направлением

- Связь производной по направлению и градиента:

$$f'_v(x_0) = \langle \nabla f(x_0), v \rangle = \|\nabla f(x_0)\| * \|v\| * \cos \varphi$$

# Градиент

- Произвольная по направлению максимальна, если направление совпадает с градиентом!
- **Градиент — направление наискорейшего роста функции**
- Антиградиент — направление наискорейшего убывания

# Условие оптимальности

- Как понять, является ли точка  $x_0$  экстремумом?
- Обобщение теоремы Ферма: если точка  $x_0$  — экстремум, и в ней существует градиент, то  $\nabla f(x_0) = 0$
- Если функция везде имеет градиент: решаем  $\nabla f(x) = 0$
- Если с градиентом проблемы: не повезло

Методы оптимизации

# Поиск минимума

- Функционал качества линейной регрессии:

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

- Как искать минимум?

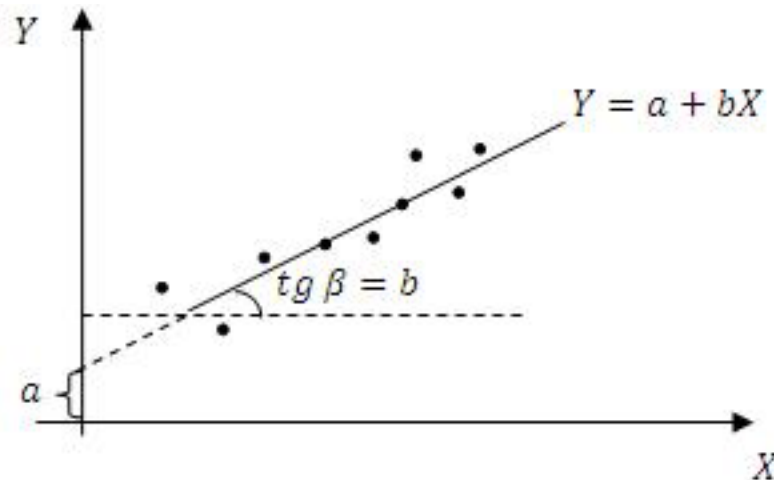
# Поиск минимума

- Можно решать уравнение:  $\nabla Q(w) = 0$
- А если уравнение сложное, и аналитически решить нельзя?
- Нужна численная оптимизация

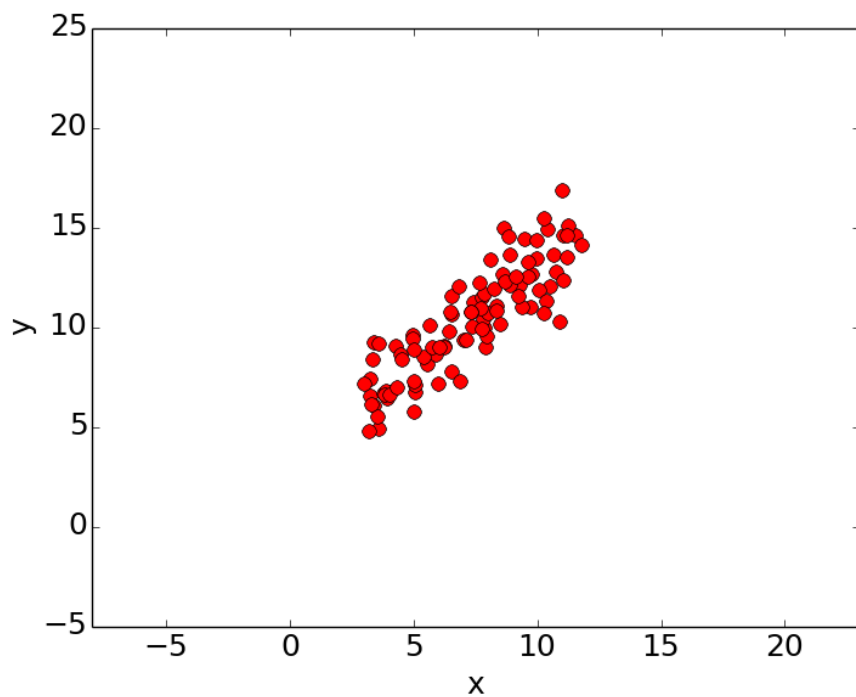


# Парная регрессия

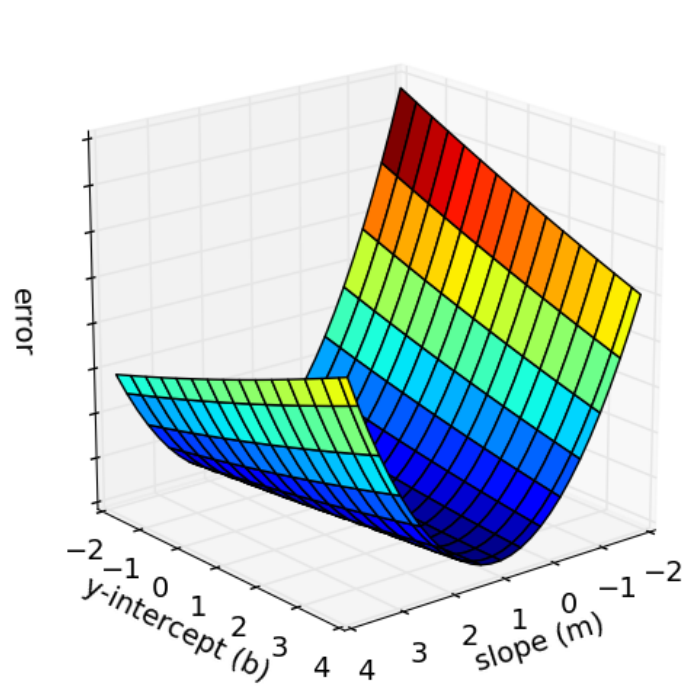
- Простейший случай: один признак
- Модель:  $a(x) = w_1 x + w_0$
- Два параметра:  $w_1$  и  $w_0$
- Функционал:  $Q(w_0, w_1) = \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$



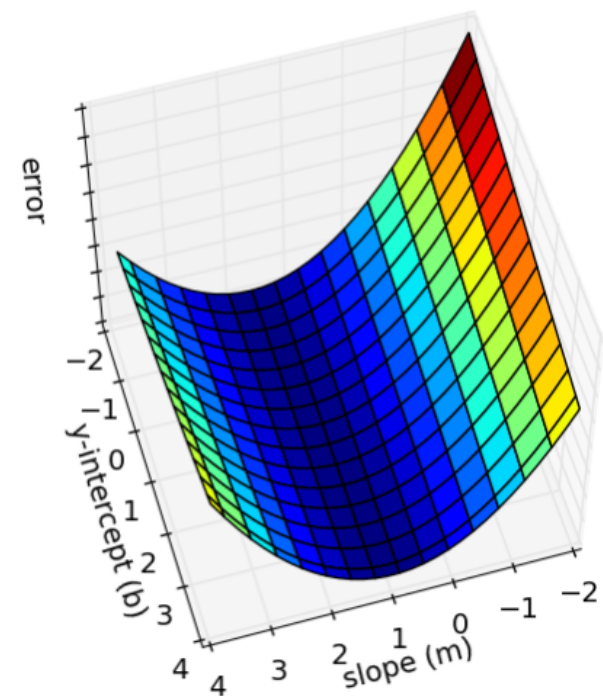
# Парная регрессия



Выборка

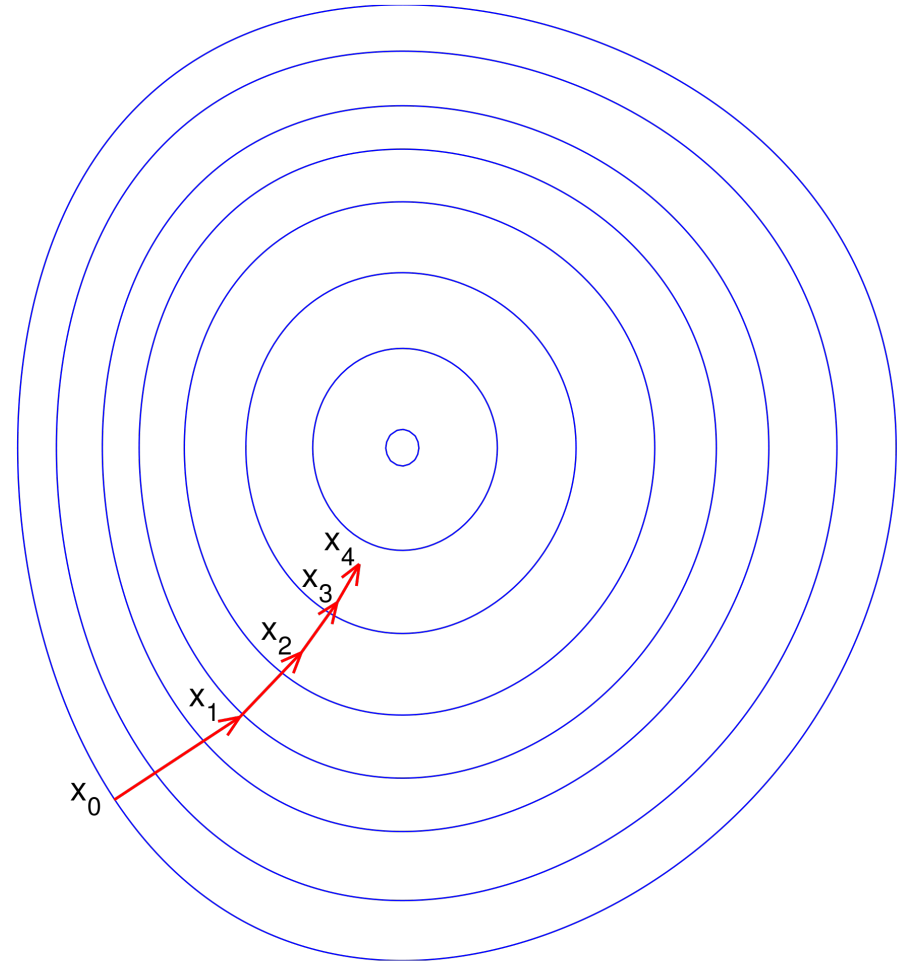


Функционал качества



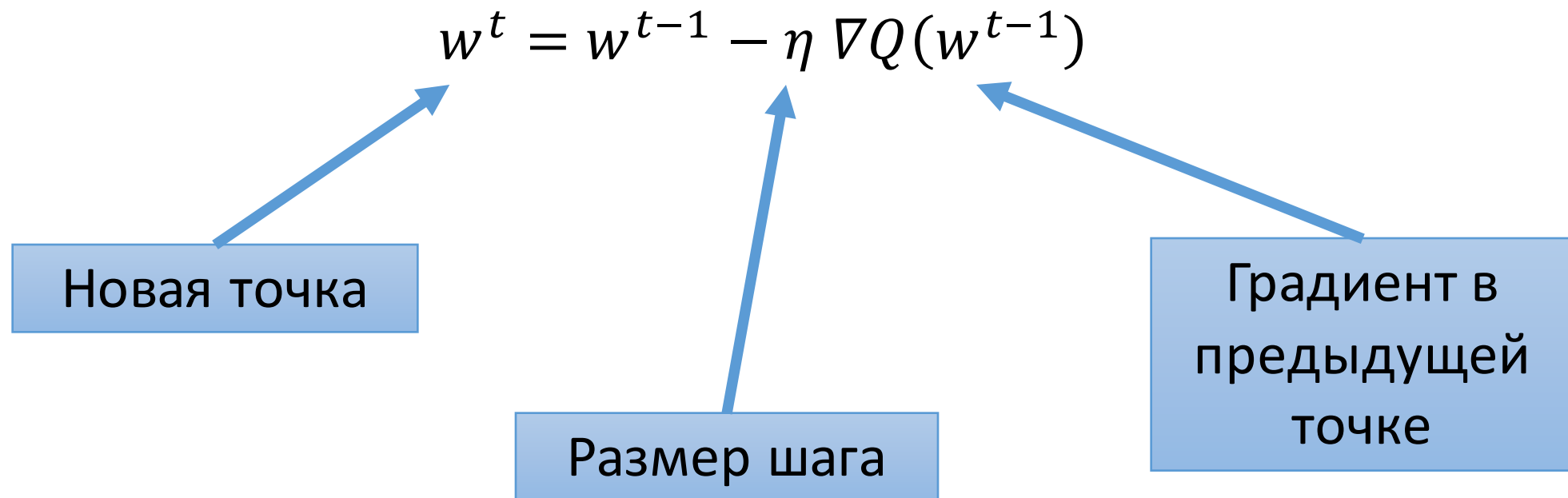
# Градиентный спуск

- Допустим, мы выбрали начальное приближение  $w^0 = (w_0^0, w_1^0)$
- Как его улучшить?
- Шагнуть в сторону наискорейшего убывания
- То есть в сторону антиградиента!



# Градиентный спуск

- Повторять до сходимости:



# Градиентный спуск

- Повторять до сходимости:

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

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

# Градиент для парной регрессии

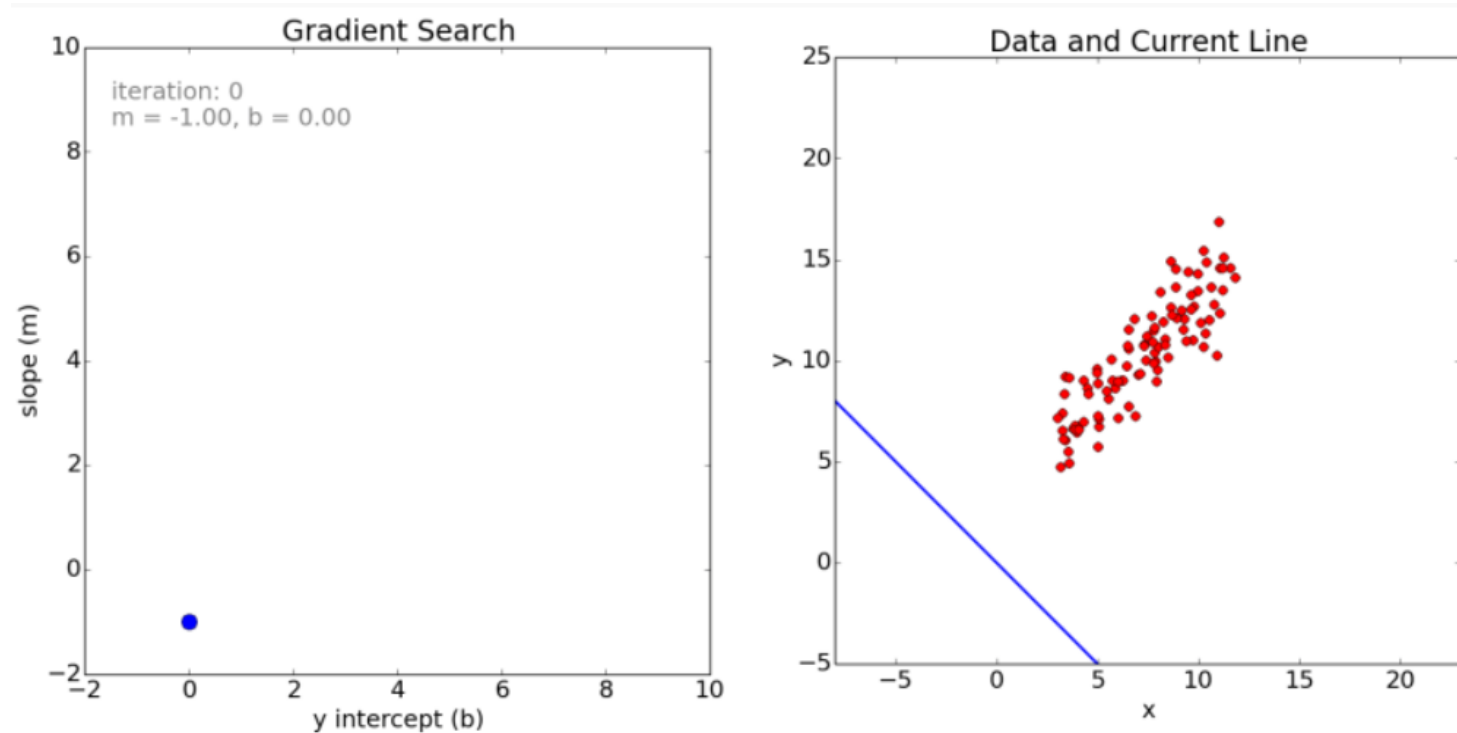
$$Q(w_0, w_1) = \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

- Частные производные:

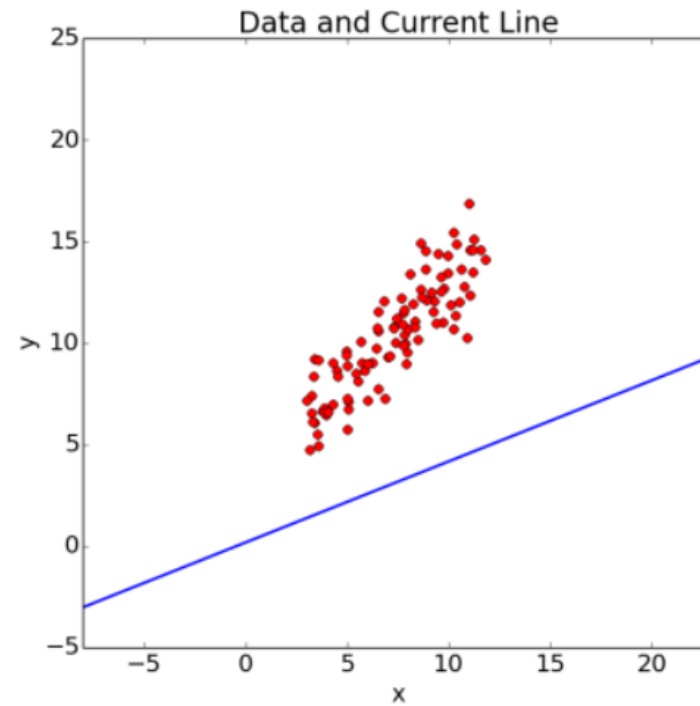
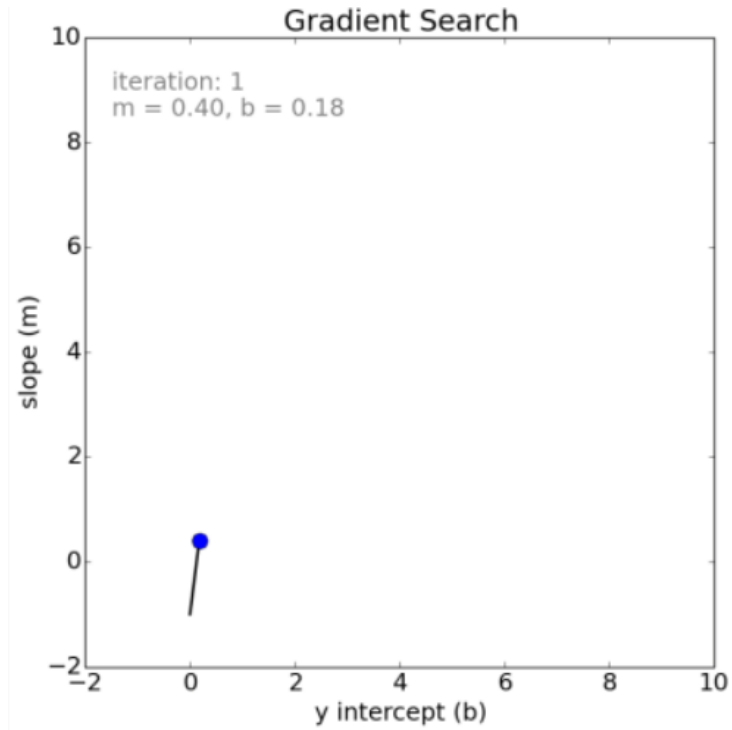
$$\frac{\partial Q}{\partial w_1} = 2 \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i) x_i$$

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

# Парная регрессия

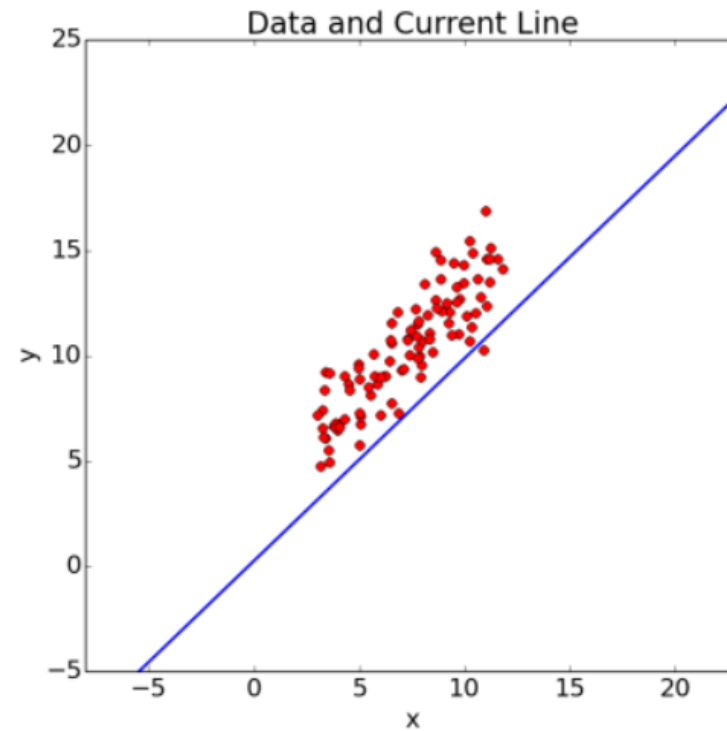
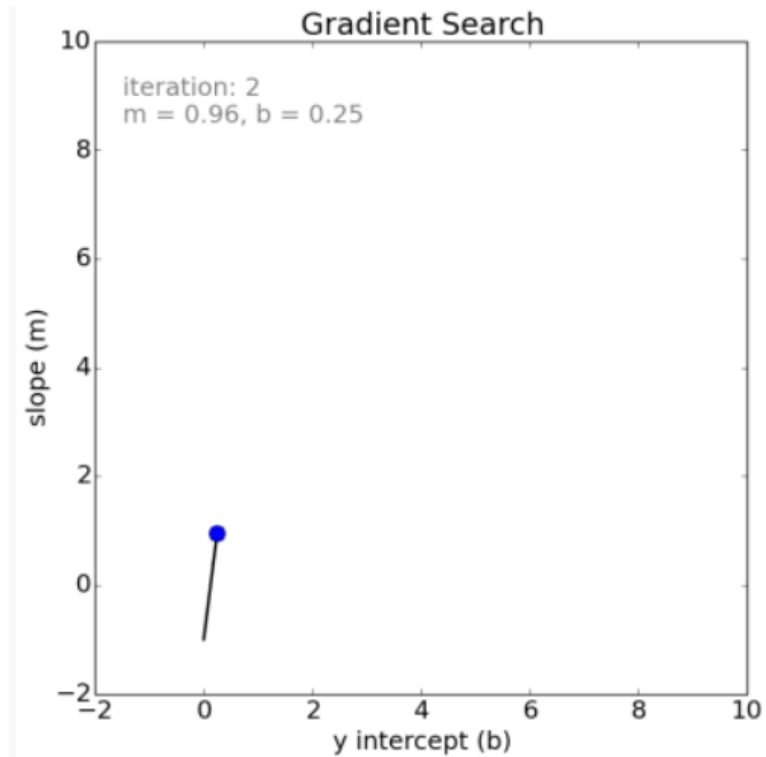


# Парная регрессия

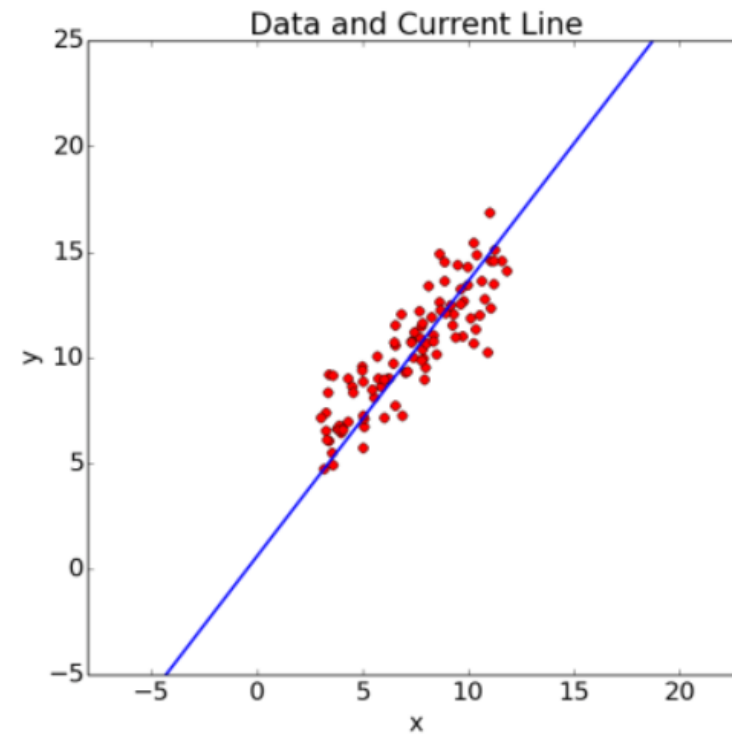
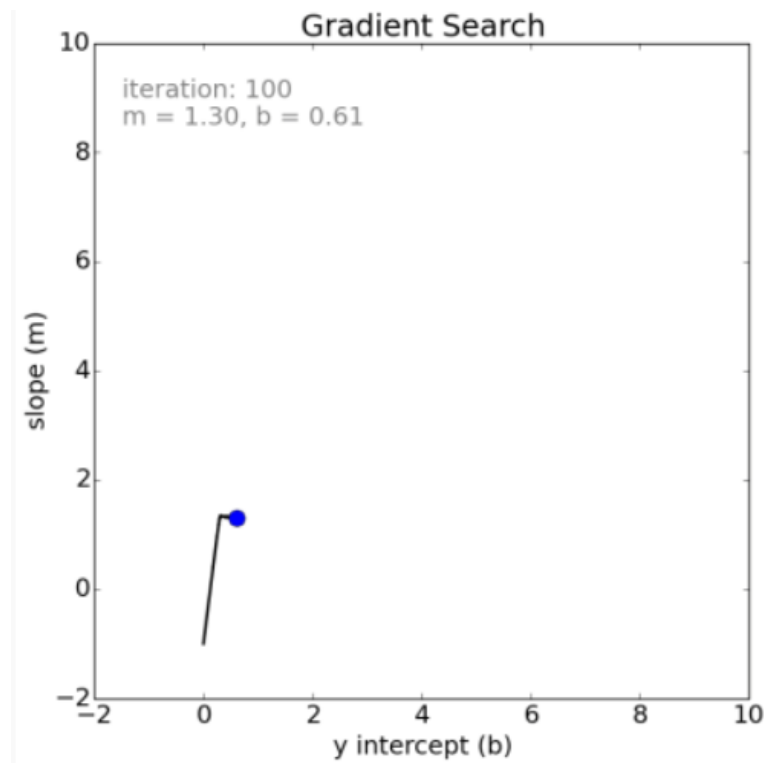




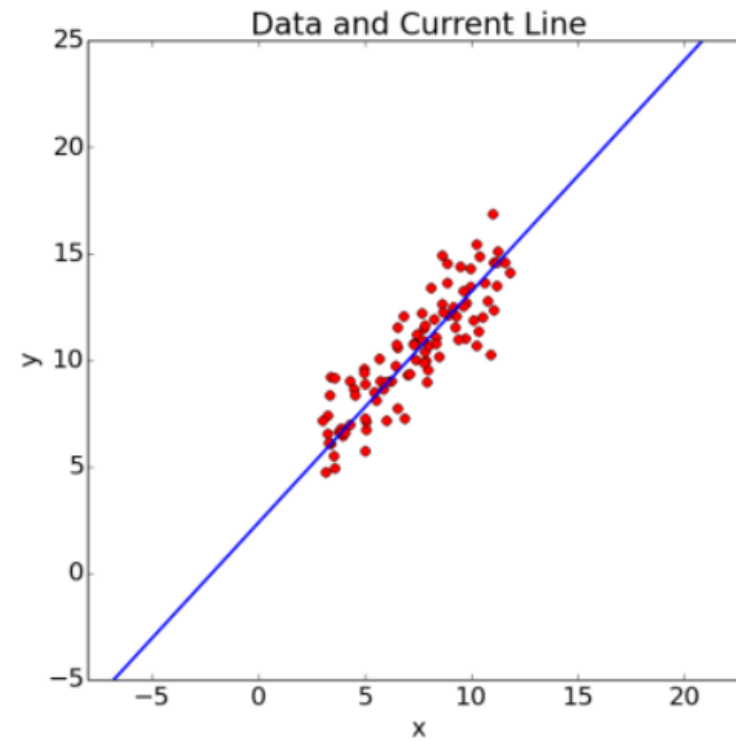
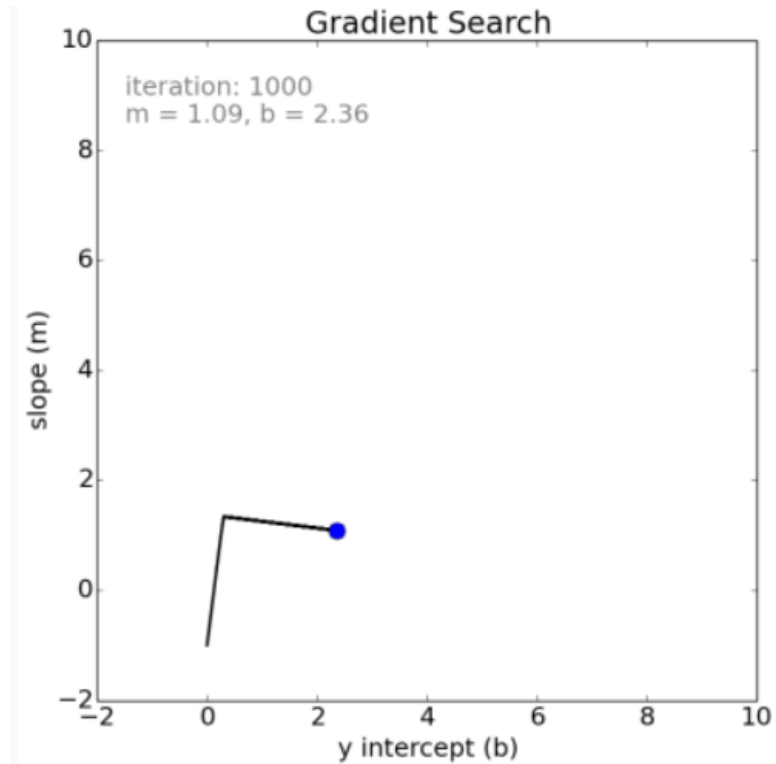
# Парная регрессия



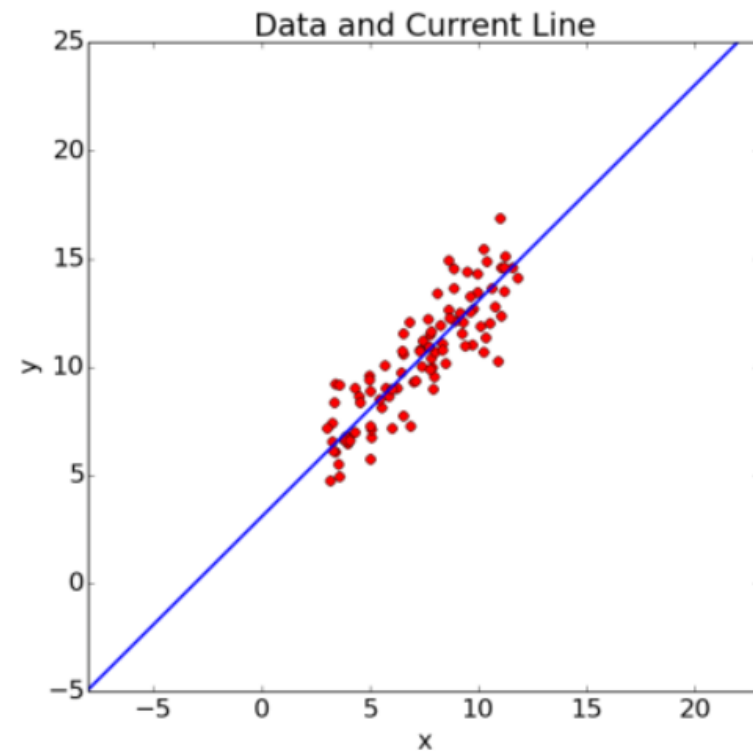
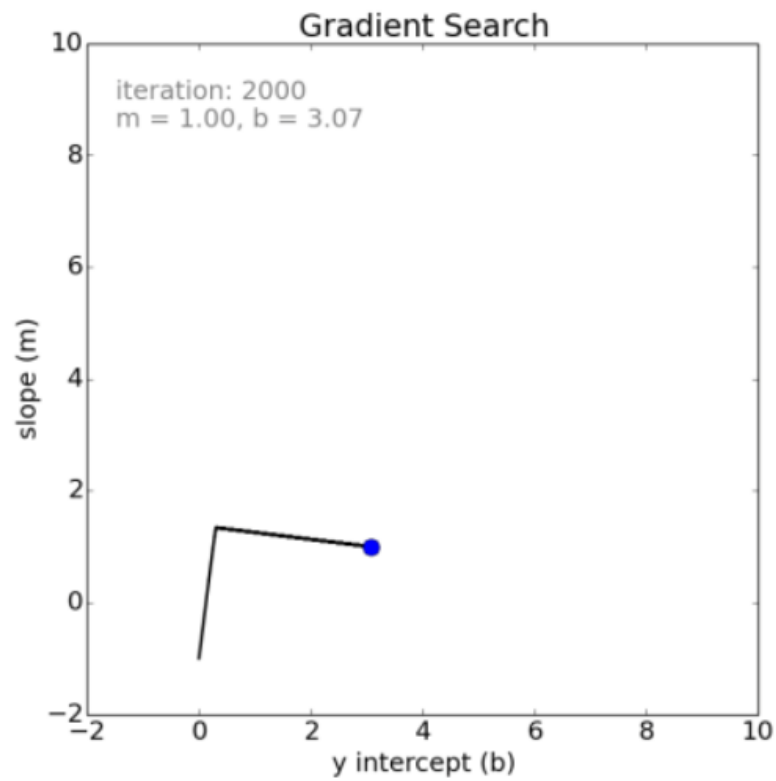
# Парная регрессия



# Парная регрессия

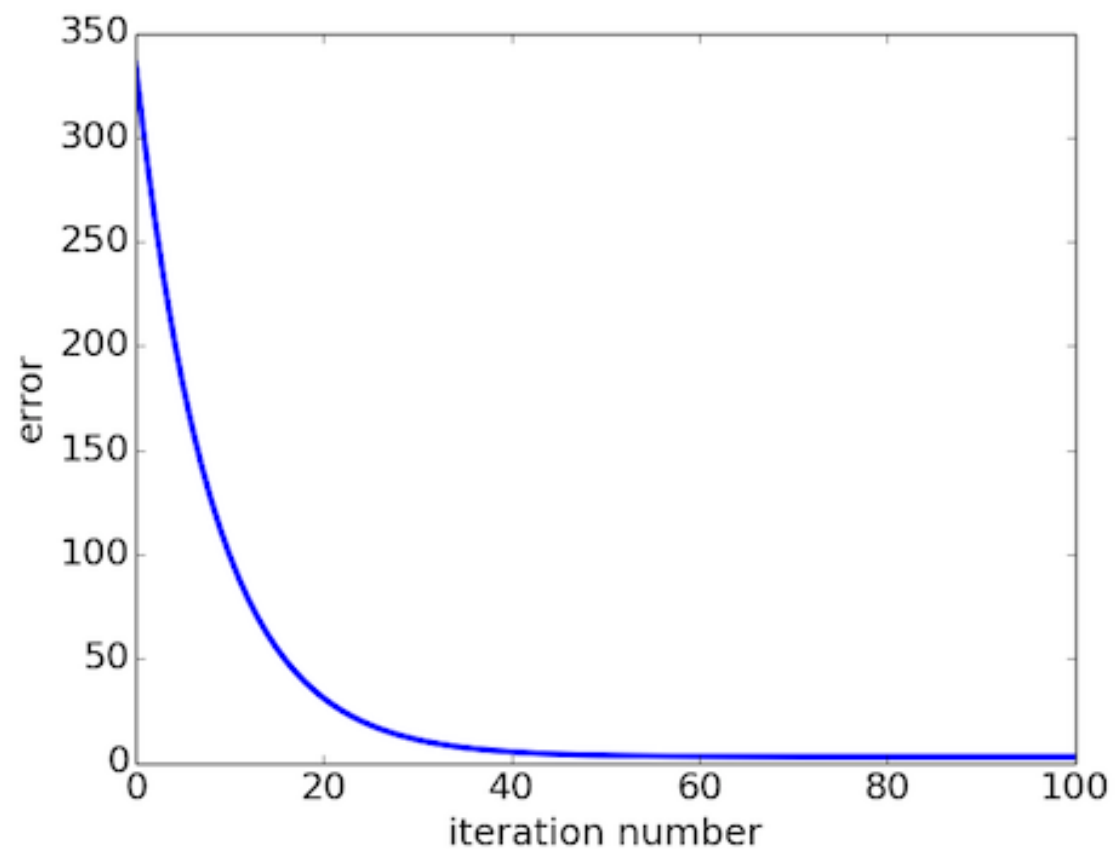


# Парная регрессия



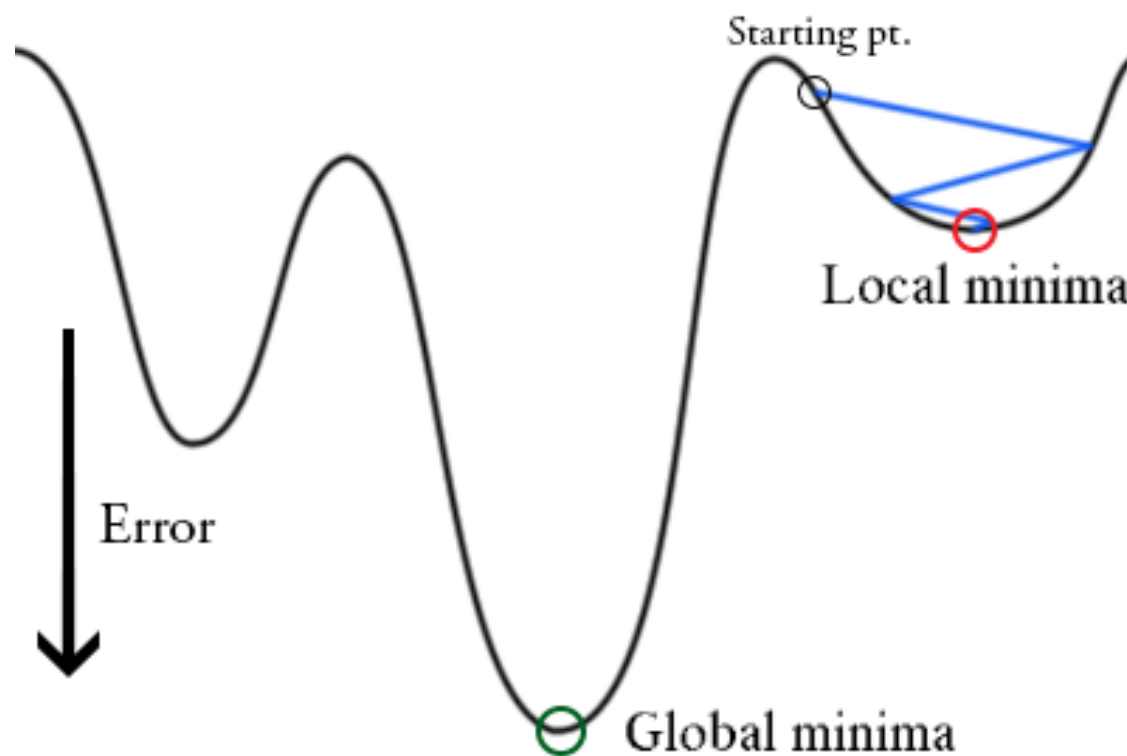


# Функционал качества



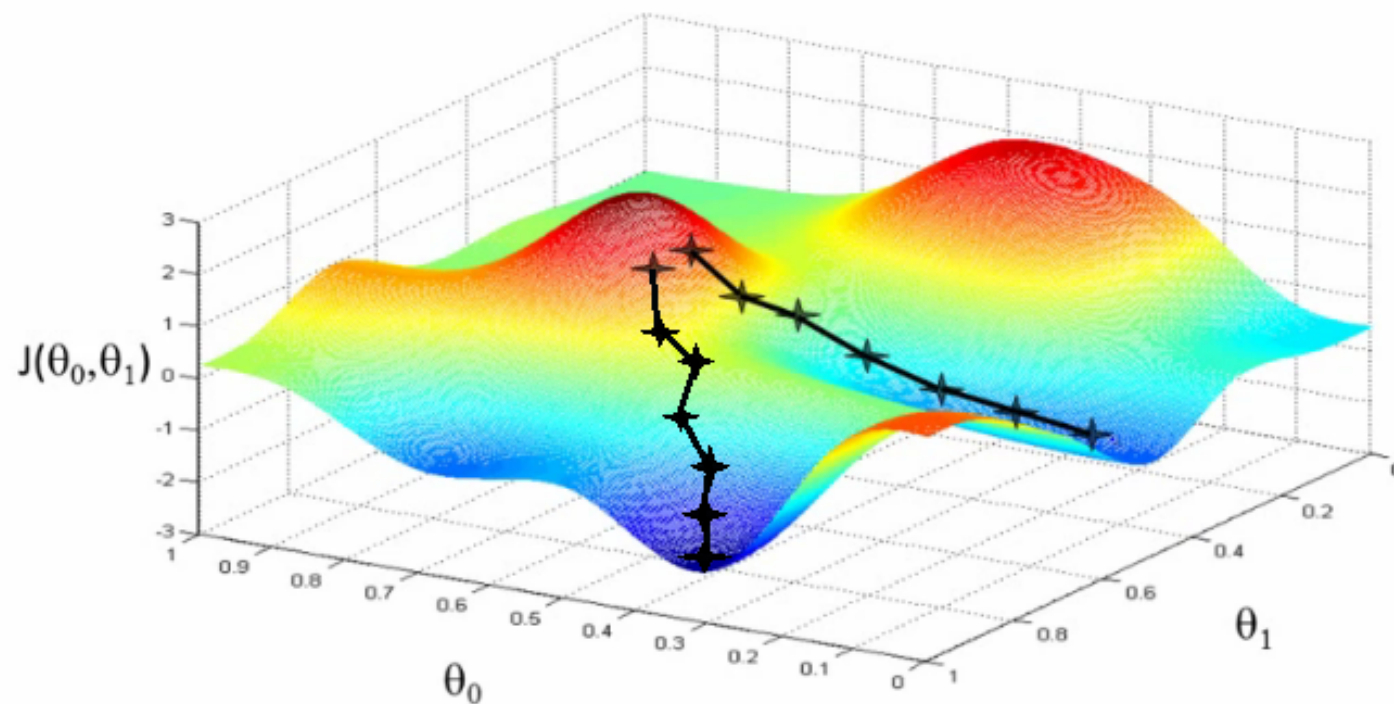
# Локальные минимумы

- Градиентный спуск находит только локальные минимумы



# Локальные минимумы

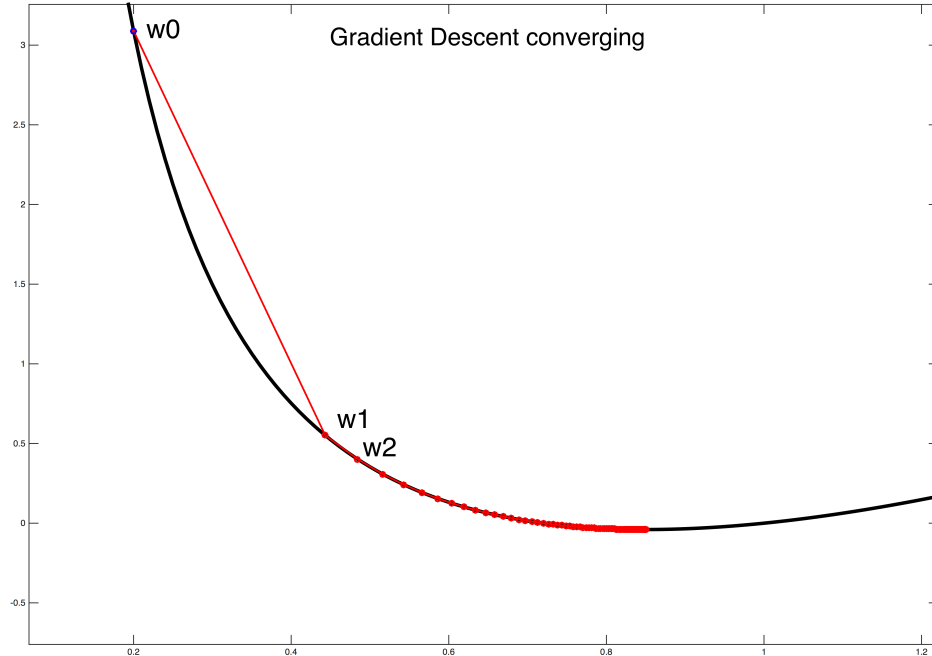
- Результат зависит от начального приближения
- Мультистарт



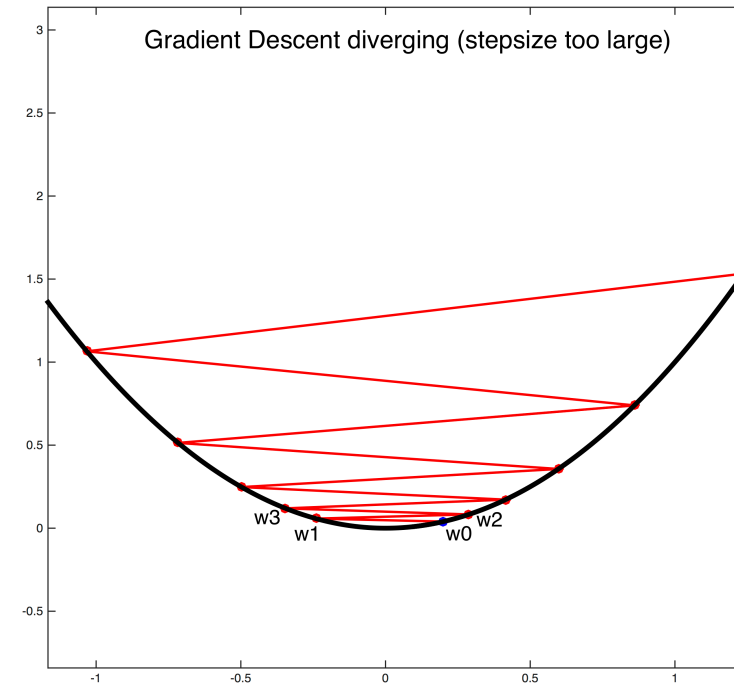


# Размер шага

- Выбор размера шага  $\eta$  — искусство



Маленький шаг



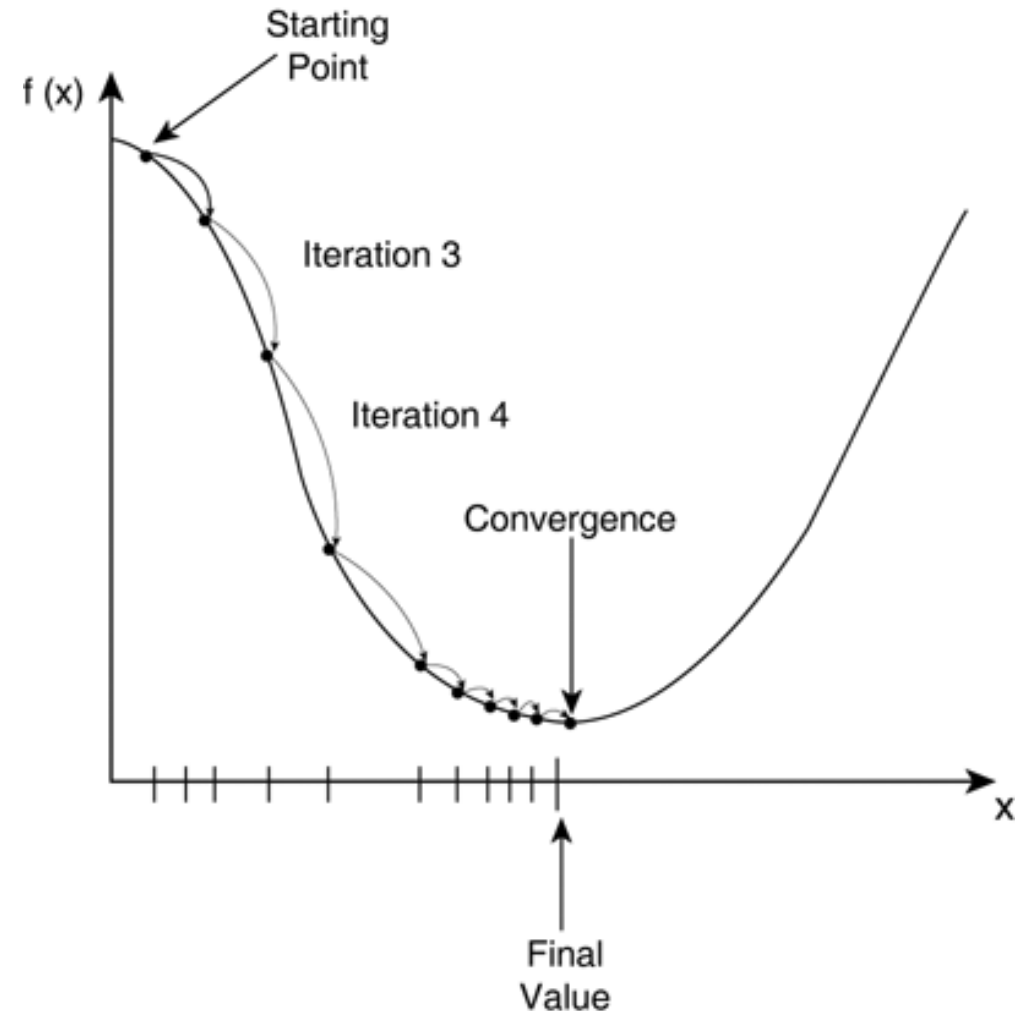
Большой шаг

# Размер шага

- Маленький шаг — больше шансов на сходимость, но требуется больше итераций
- Большой шаг — есть риск отсутствия сходимости
- Наискорейший градиентный спуск:
$$\eta_t = \arg \min_{\eta} Q(w^{t-1} - \eta \nabla Q(w^{t-1}))$$
- Нужно делать одномерный поиск на каждой итерации

# Размер шага

- Обычно пользуются эвристиками
- Чем ближе к минимуму, тем меньше надо шагать
- Неплохо работает:  $\eta_t = \frac{1}{t}$
- Еще лучше:  $\eta_t = \lambda \left( \frac{s}{s+t} \right)^p$ , где  $\lambda, s, p$  — параметры



# Системы линейных уравнений

- $Xw = y$
- Можно решать градиентным спуском следующую задачу:
$$\|Xw - y\|^2 \rightarrow \min_w$$
- Функционал выпуклый
- Если решение есть — минимальное значение равно нулю
- Если решения нет — найдем наилучшее приближение

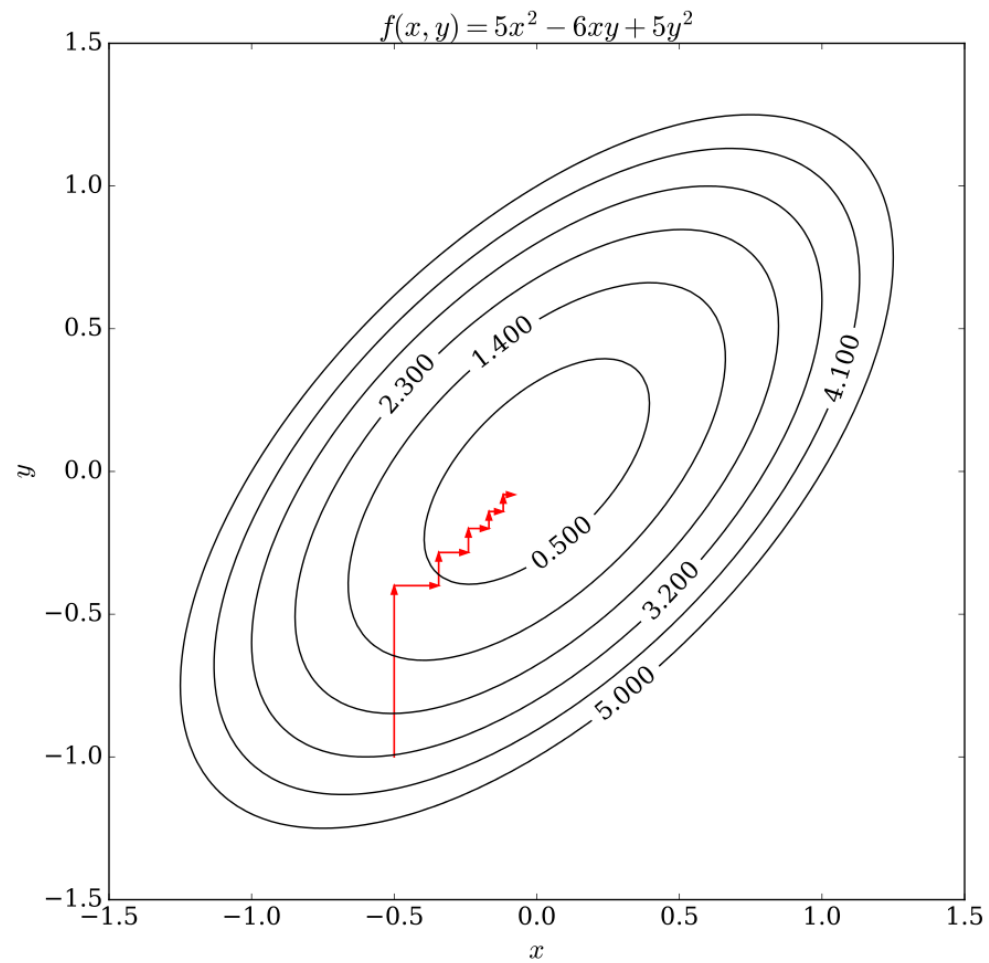
# Другие методы оптимизации

- Методы первого порядка — используют первые производные
  - Градиентный спуск
  - Стохастический градиентный спуск
  - Квазиньютоновские методы, BFGS
  - Stochastic Average Gradient, Nesterov momentum, ...
- Методы второго порядка — используют вторые производные
  - Метод Ньютона
- Методы нулевого порядка — без производных
  - Покоординатный спуск
  - Стохастическая оптимизация

# Покоординатный спуск

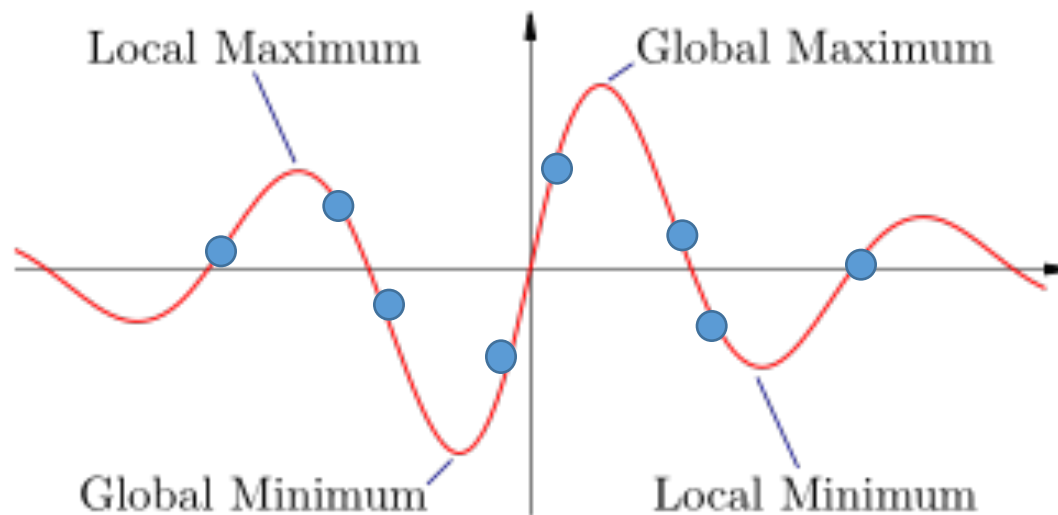
- По очереди меняем каждую координату
- Шаг по каждой координате — случайный, наискорейший, эвристический...
- Быстрые итерации, но может медленно сходиться
- Используется в методе опорных векторов (один из линейных)

# Покоординатный спуск



# Стохастическая оптимизация

- Простейший алгоритм:
  - Генерируем  $N$  раз случайную точку
  - Выбираем ту, на которой значение функционала наименьшее
- Не самый лучший подход
- Нужно более направленное движение

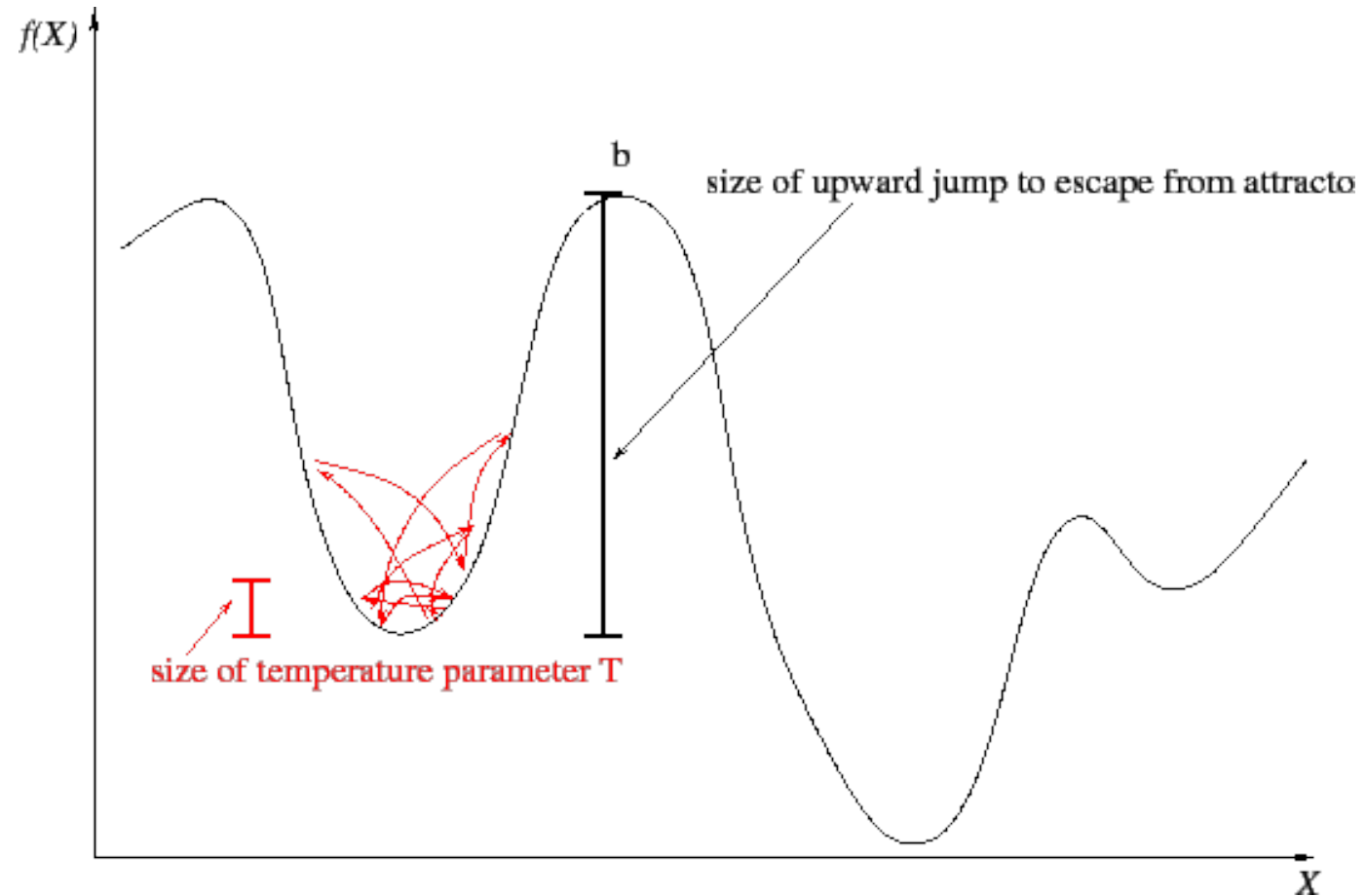




# Метод имитации отжига

- $w^t$  — текущее приближение
- Генерируем кандидата  $w$
- Если  $Q(w) < Q(w^t)$ 
  - то переходим:  $w^{t+1} = w$
- Если  $Q(w) > Q(w^t)$ 
  - то переходим с вероятностью  $\exp\left(-\frac{Q(w)-Q(w^t)}{C_t}\right)$

# Метод имитации отжига



# Обучение линейной регрессии

# Задача оптимизации

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

- Гладкая функция
- Выпуклая функция
- Единственный минимум (не всегда)

# Градиентный спуск

- Повторять до сходимости:

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

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

# Градиент

$$\nabla Q(w, X) = \left( \frac{\partial Q}{\partial w_1}, \dots, \frac{\partial Q}{\partial w_d} \right)$$

Производные:

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

# Нюансы

- Выбор длины шага  $\eta$  — пробуем разные значения
- Выборка должна быть масштабирована
- Признаки не должны коррелировать

# Аналитическое решение

- Векторная запись MSE:

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

- Условие минимума:

$$\nabla Q(w, X) = 0$$

- Что, если попробуем решить эту систему уравнений?



# Аналитическое решение

- Она решается аналитически!

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

- Но обращение матрицы — очень сложная операция
- Градиентный спуск гораздо быстрее

Мультиколлинеарность

# Объекты-признаки

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Задача предсказания прибыли магазина в следующем месяце
- Рассмотрим в качестве векторов столбцы матрицы (признаки)

# Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- Первый и второй признаки:  $x_2 = 1000x_1$
- Первый — общий вес товаров в тоннах, второй — в килограммах

# Подозрительные зависимости

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

- $x_5 = 0.5x_3 + 0.5x_4$
- Пятый — средняя прибыль за последние два месяца
- Третий и четвертый — прибыль в прошлом и позапрошлом месяце

# Линейная зависимость

— один из векторов равен сумме с весами остальных векторов

- Это плохо:
  - Избыточная информация
  - Лишние затраты на хранение данных
  - Вредит некоторым методам машинного обучения

# Линейная зависимость

- Пусть дан набор векторов  $x_1, \dots, x_n$
- Они линейно зависимы, если
  - существуют такие числа  $\beta_1, \dots, \beta_n$ ,
  - хотя бы одно из которых не равно нулю,
  - что сумма векторов с такими коэффициентами равна нулю

$$\beta_1 x_1 + \dots + \beta_n x_n = 0$$

# Мультиколлинеарность

- Наличие зависимостей между признаками
- Приводит к тому, что решений бесконечное число
- Далеко не все из них имеют хорошую обобщающую способность



# Линейная зависимость

- Худший случай — линейно зависимые признаки
- Существуют такие  $\alpha = (\alpha_1, \dots, \alpha_d)$ , что для любого объекта:

$$\alpha_1 x^1 + \dots + \alpha_d x^d = \langle \alpha, x \rangle = 0$$

# Линейная зависимость

- Допустим, мы нашли решение  $w_*$
- Модифицируем:  $w_1 = w_* + t\alpha$
- ( $t$  — число)
- Ответ нового алгоритма на любом объекте:

$$\langle w_1, x \rangle = \langle w_* + t\alpha, x \rangle = \langle w_*, x \rangle + t\langle \alpha, x \rangle = \langle w_*, x \rangle$$

- $w_1$  — тоже решение!

# Коррелирующие признаки

- Тоже плохо
- Сначала разберёмся с корреляцией

# Коэффициент корреляции

$$\rho(\xi, \eta) = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{D}\xi \mathbb{D}\eta}}$$

Выборочная корреляция:

$$\rho(x, z) = \frac{\sum_{i=1}^{\ell} (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell} (x_i - \bar{x})^2 \sum_{i=1}^{\ell} (z_i - \bar{z})^2}}$$

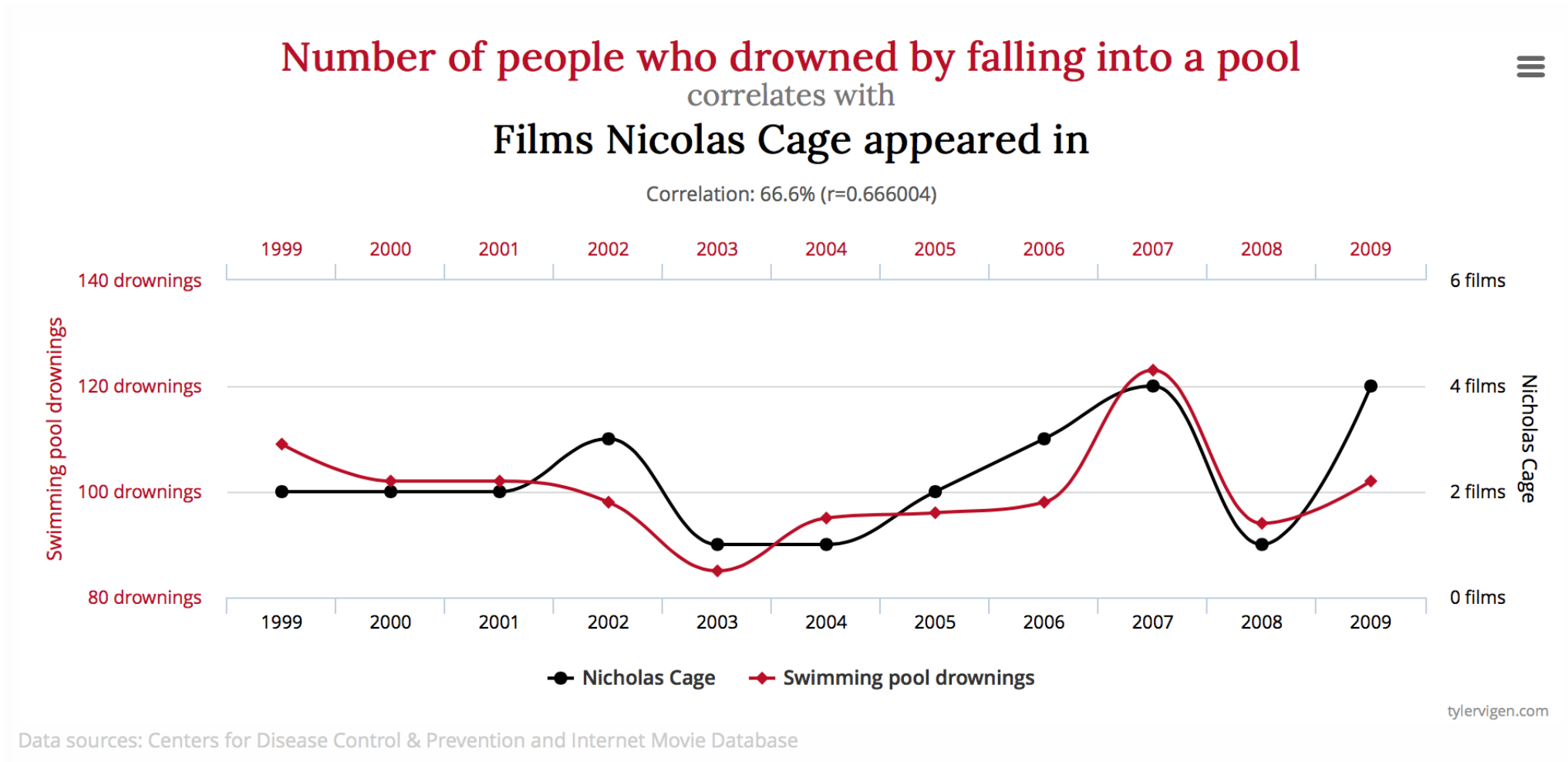
$$\bar{x} = \frac{1}{\ell} \sum_{j=1}^{\ell} x_j; \quad \bar{z} = \frac{1}{\ell} \sum_{j=1}^{\ell} z_j$$

# Коэффициент корреляции

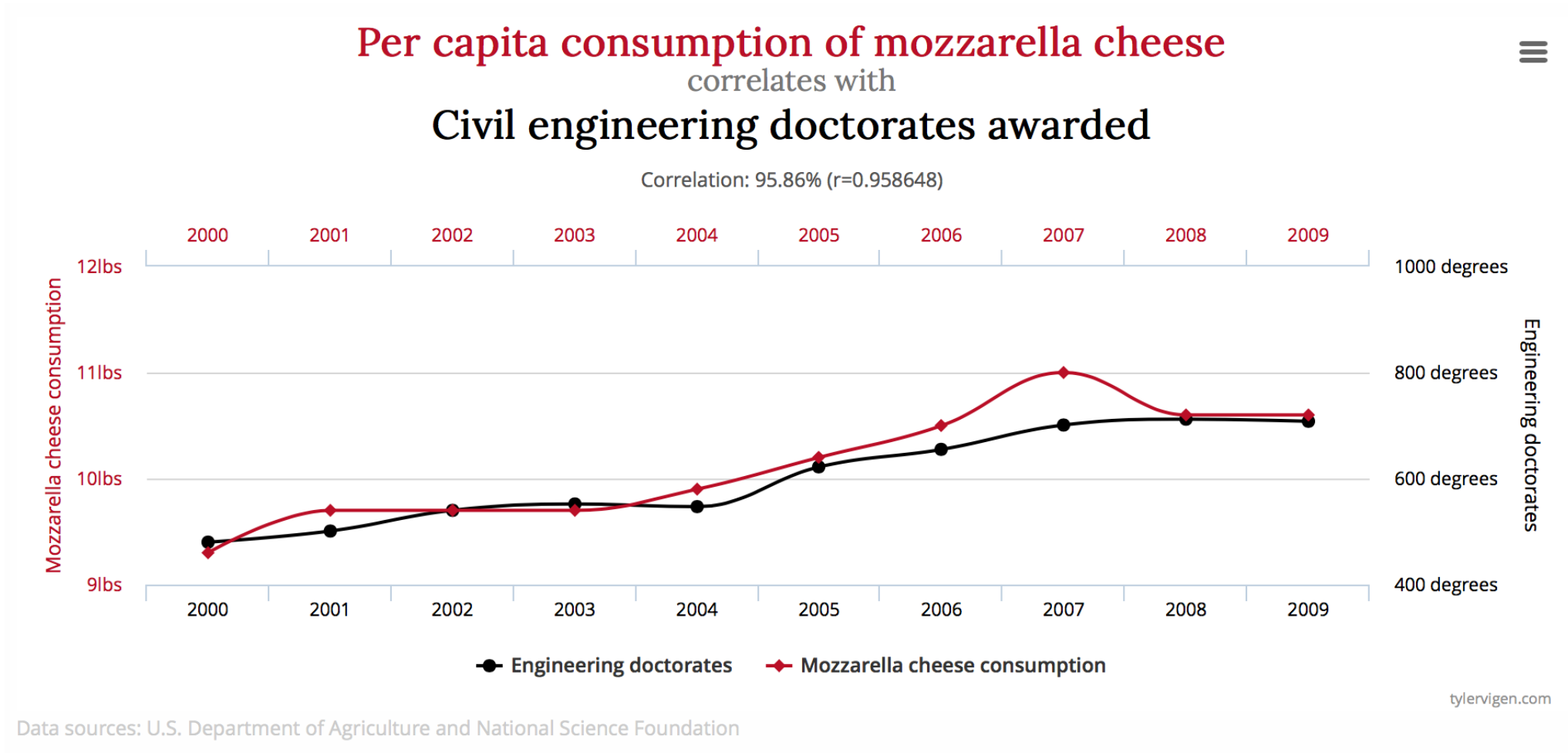
$$\rho(x, z) = \frac{\sum_{i=1}^{\ell} (x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{\ell} (x_i - \bar{x})^2 \sum_{i=1}^{\ell} (z_i - \bar{z})^2}}$$

- $\rho(x, z) \in [-1, +1]$
- Очень грубо: чем ближе к +1 или -1, тем точнее выполнено уравнение
$$x = az + b$$
- Мера линейной зависимости

# Пример



# Пример



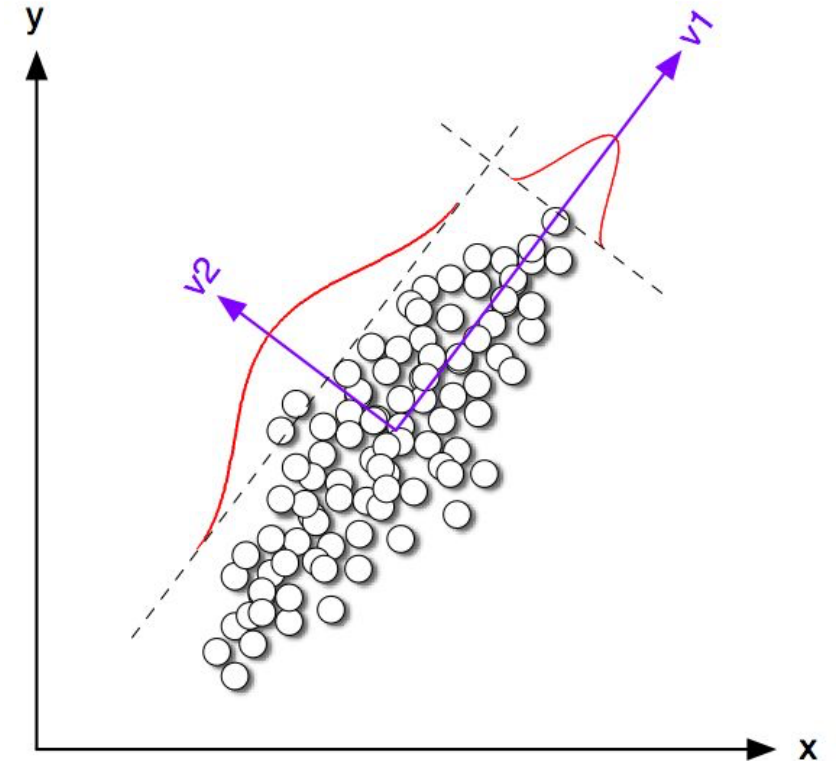
# Распространённое заблуждение

- Может показаться, что из корреляции следует причинно-следственная связь
  - Это не так!
  - Корреляция означает, что события часто происходят вместе
  - Но никак не следуют друг из друга
- 
- Больше примеров: <http://tylervigen.com/spurious-correlations>



# Коррелирующие признаки

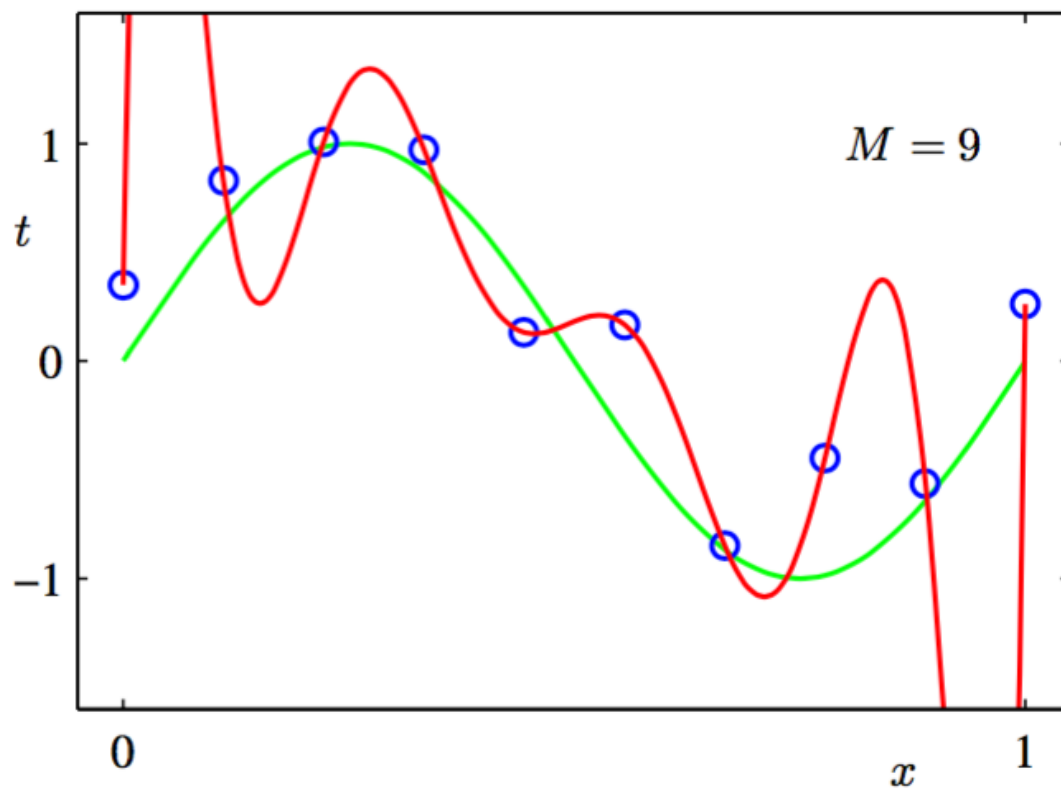
- Плохо, если есть коррелирующие признаки
- Решение: отбор признаков или их декорреляция
- В следующих лекциях



Переобучение и регуляризация

# Пример

- Один признак  $x$
- $a(x) = w_0 + w_1x + w_2x^2 + \dots + w_9x^9$



# Пример

- Коэффициенты:

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots + 2740x^9$$

- Большие коэффициенты — симптом переобучения
- (эмпирическое наблюдение)

# Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу
  - $a(x) = 698x - 41714$
- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже на правильную зависимость

# Регуляризация

- Будем штрафовать за большие веса!
- Функционал:

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

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^d w_j^2$$

# Регуляризация

- Регуляризованный функционал ошибки:

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

- Всё ещё гладкий и выпуклый

# Коэффициент регуляризации

- $\lambda$  — новый параметр, надо подбирать
- Высокий  $\lambda$  — простые модели
- Низкий  $\lambda$  — риск переобучения
- Нужно балансировать
- Подбор  $\lambda$  — с помощью кросс-валидации



# Смысл регуляризации

- Минимизация регуляризованного функционала равносильна решению условной задачи:

$$\begin{cases} \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w \\ \|w\|^2 \leq C \end{cases}$$

# $L_1$ -регуляризация

- $L_1$ -регуляризатор:

$$\|w\|_1 = \sum_{j=1}^d |w_j|$$

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|_1 \rightarrow \min_w$$

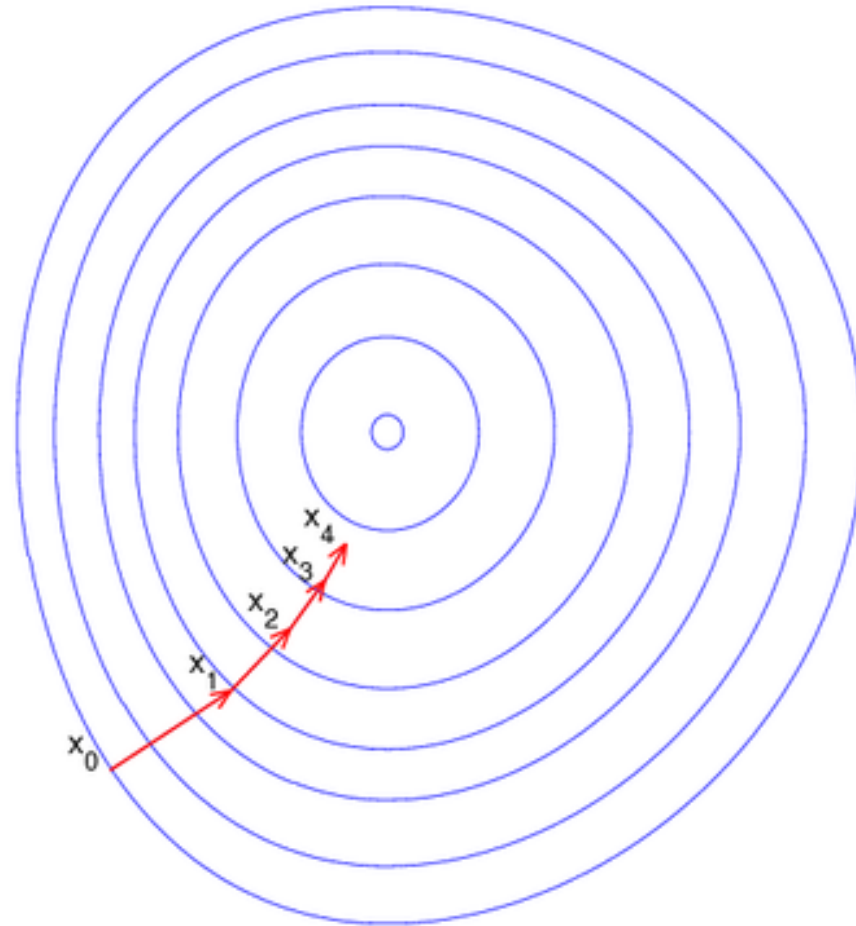
# $L_1$ -регуляризация

- Функционал становится негладким
- Сложнее оптимизировать
- Зато производится отбор признаков
- Часть весов в решении будут нулевыми

Масштабирование признаков

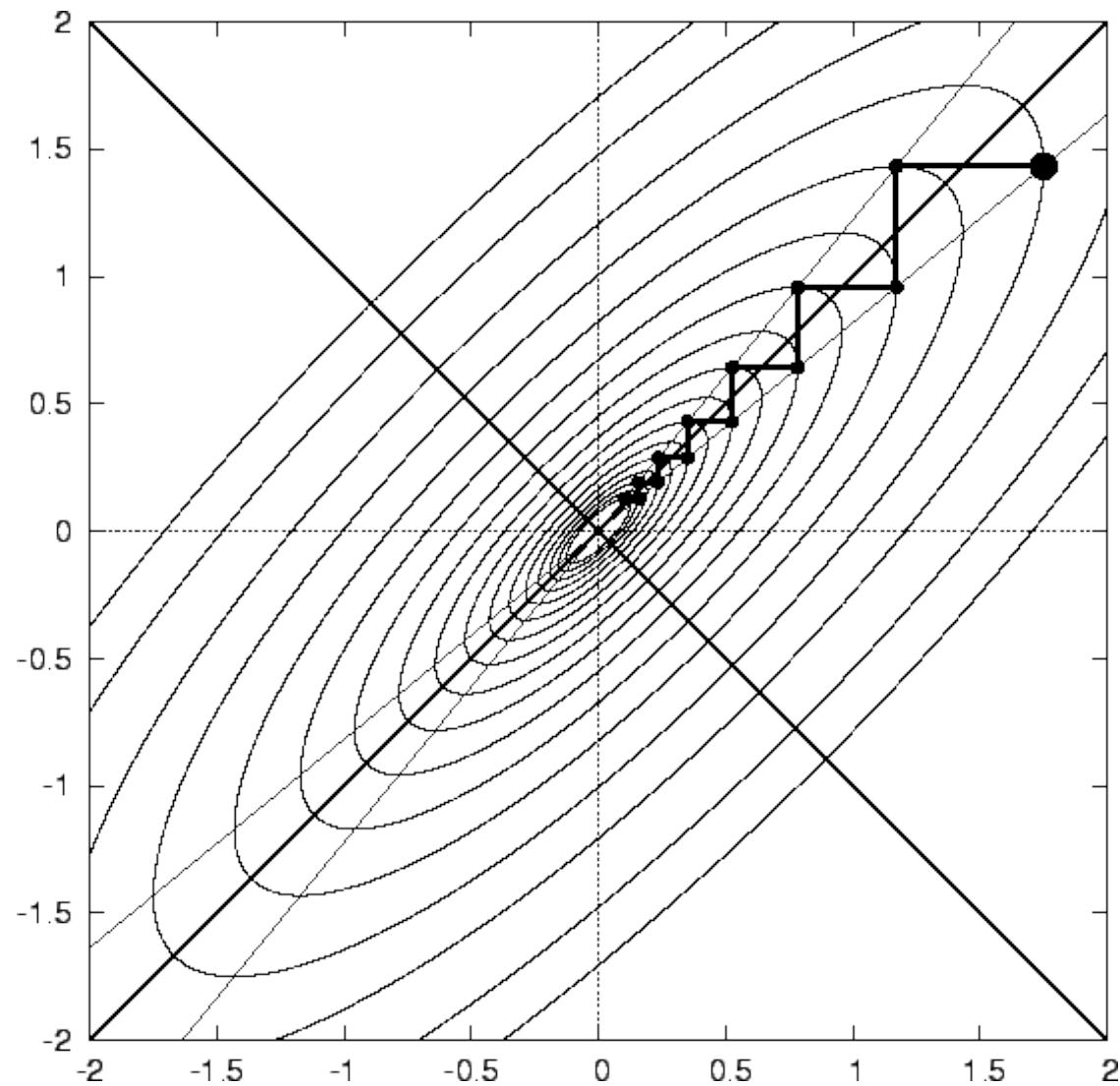
# Масштабирование выборки

Хороший случай



# Масштабирование выборки

Плохой случай



# Масштабирование выборки

- Задача: одобряют ли заявку на грант?
- 1-й признак: сколько успешных заявок было до этого у заявителя
- 2-й признак: год рождения заявителя
- Масштаб: единицы и тысячи
- Все признаки должны иметь одинаковый масштаб

# Масштабирование выборки

- 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 выборки

- Отмасштабируем  $j$ -й признак
- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

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

# Резюме

- Градиентный спуск — универсальный инструмент обучения дифференцируемых моделей
- Линейные зависимости и корреляции в признаках приводят к проблемам при обучении
- Регуляризация — способ борьбы с переобучением
- Масштабирование помогает улучшить сходимость градиентных методов