

Введение в анализ данных

Лекция 4

Математический анализ и анализ данных

Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

Напоминание

- \mathbb{X} — пространство объектов, \mathbb{Y} — пространство ответов
- $x = (x^1, \dots, x^d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал качества алгоритма a на выборке X
- Обучение: $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

Напоминание

- \mathbb{X} — пространство объектов, \mathbb{Y} — пространство ответов
- $x = (x^1, \dots, x^d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал качества алгоритма a на выборке X
- Обучение: $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

Линейная регрессия

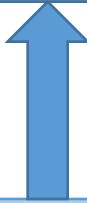
- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- Линейная модель: $a(x) = w_1x^1 + \dots + w_dx^d = \langle w, x \rangle$
- Обучение:

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

Линейная регрессия

- Задача регрессии: $\mathbb{Y} = \mathbb{R}$
- Линейная модель: $a(x) = w_1x^1 + \dots + w_dx^d = \langle w, x \rangle$
- Обучение:

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



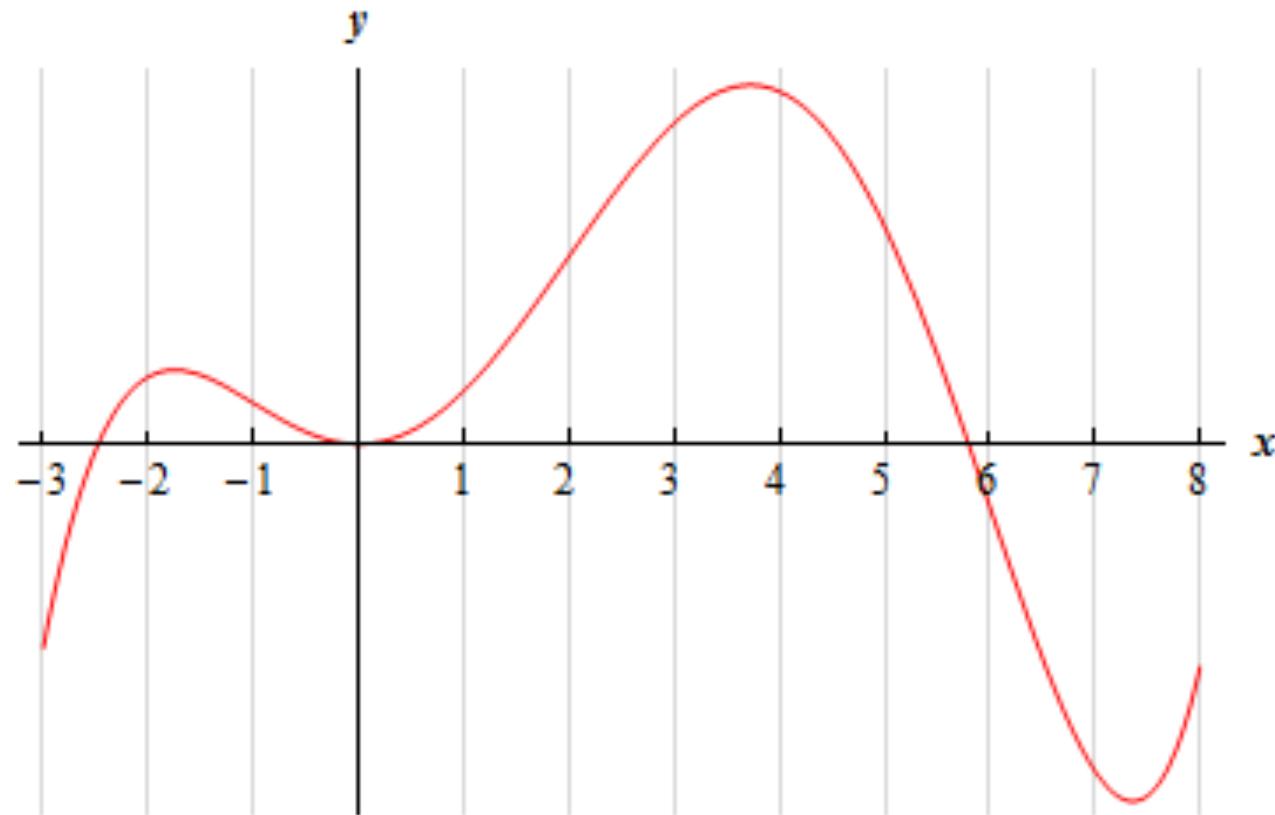
Функция с d аргументами

Вопросы на сегодня

- Что такое минимум и максимум функции? Какие они бывают?
- Как определить, является ли точка оптимумом?
- Как искать оптимальную точку?

Функции одной переменной

ФУНКЦИЯ



Предел последовательности

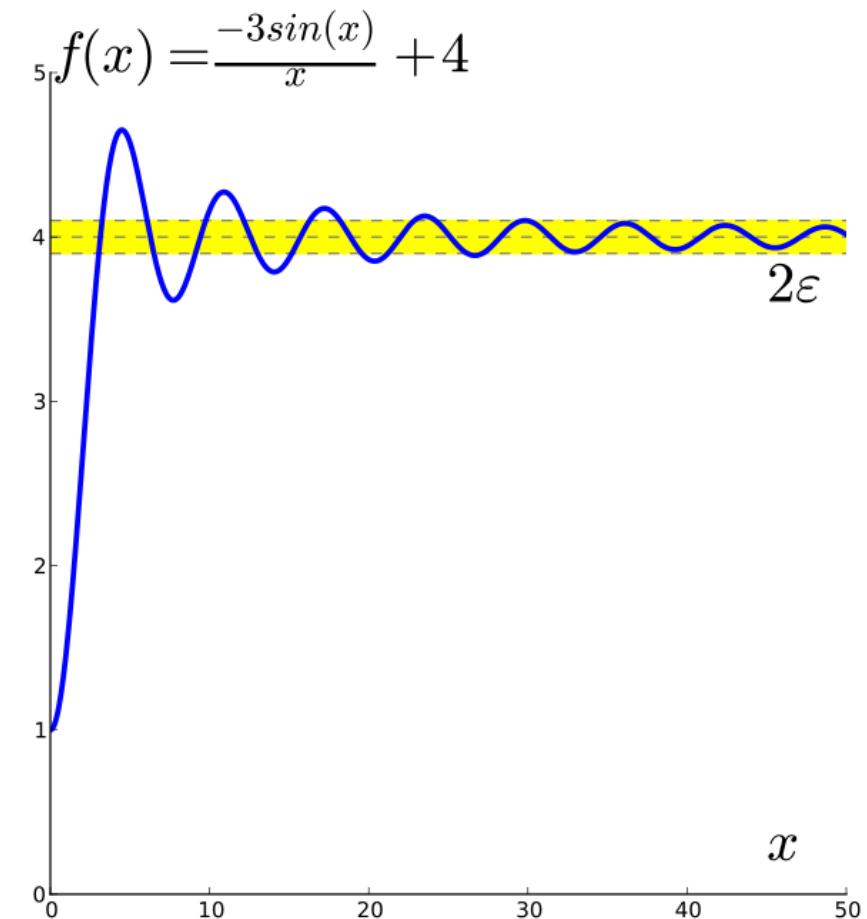
- $x_1, x_2, x_3, x_4, \dots$ — последовательность
- Предел последовательности: $\lim_{n \rightarrow \infty} x_n = L$
 - Чем больше n , тем ближе x_n к L
- Пример: $x_n = \frac{1}{n}$
- $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$

Предел последовательности

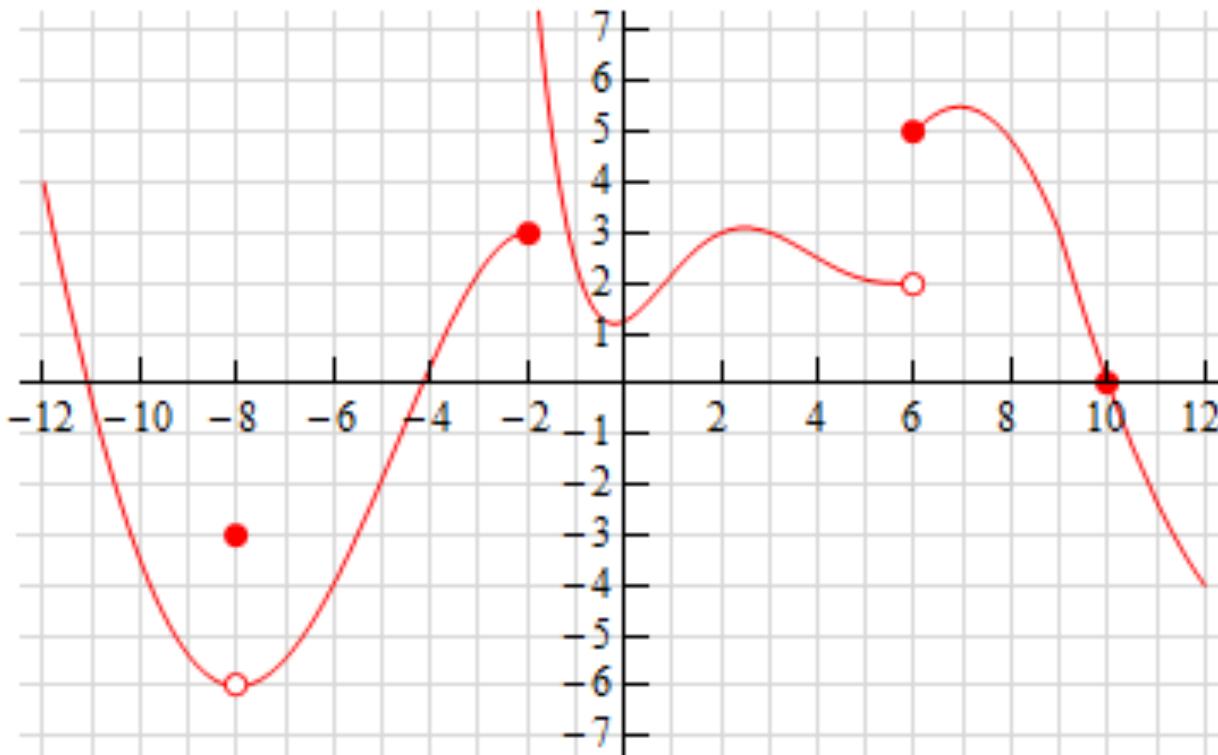
- Пример: $x_n = \left(1 + \frac{1}{n}\right)^n$
- $x_{10} = 2.594 \dots$
- $x_{100} = 2.705 \dots$
- $x_{1000} = 2.717 \dots$
- $x_{10000} = 2.718 \dots$
- $x_{100000} = 2.718 \dots$
- $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e = 2.718268 \dots$

Предел функции

- $\lim_{x \rightarrow x_0} f(x) = L$
- Если $x_n \rightarrow x_0$, то $f(x_n) \rightarrow L$
- Функция непрерывна в точке x_0 , если $\lim_{x \rightarrow x_0} f(x) = f(x_0)$

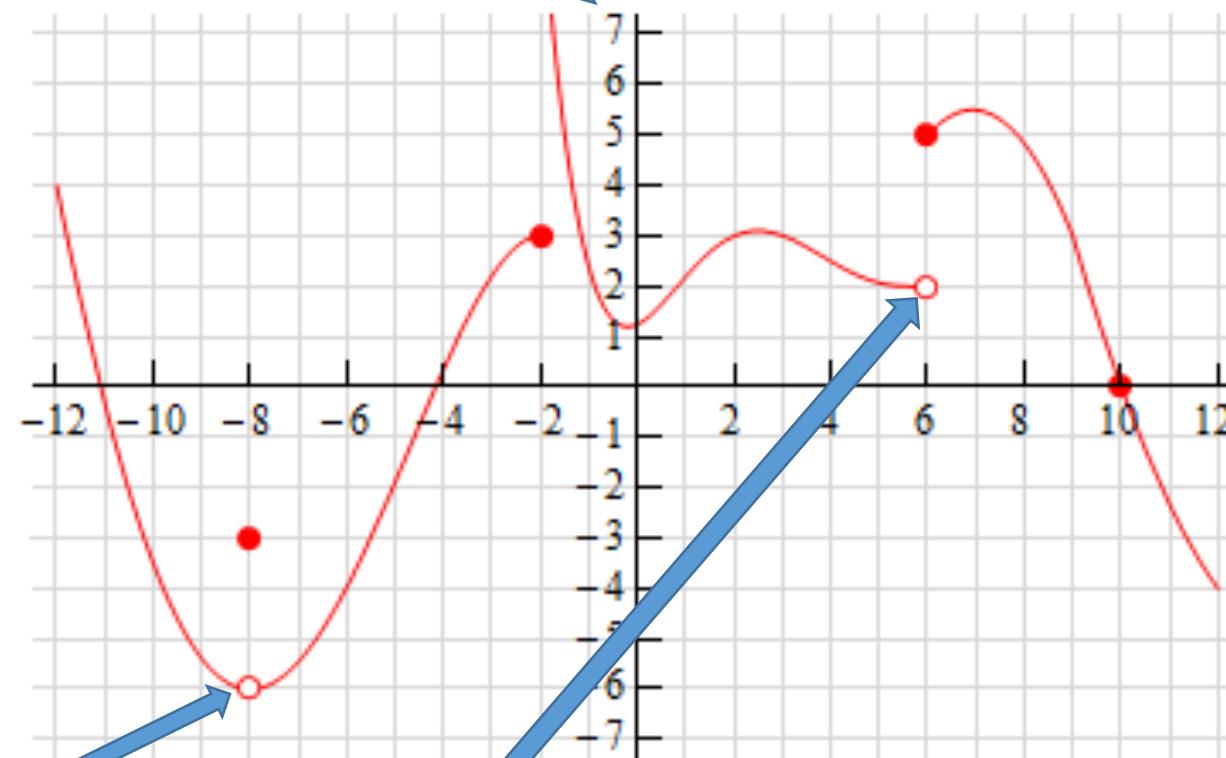


Непрерывность



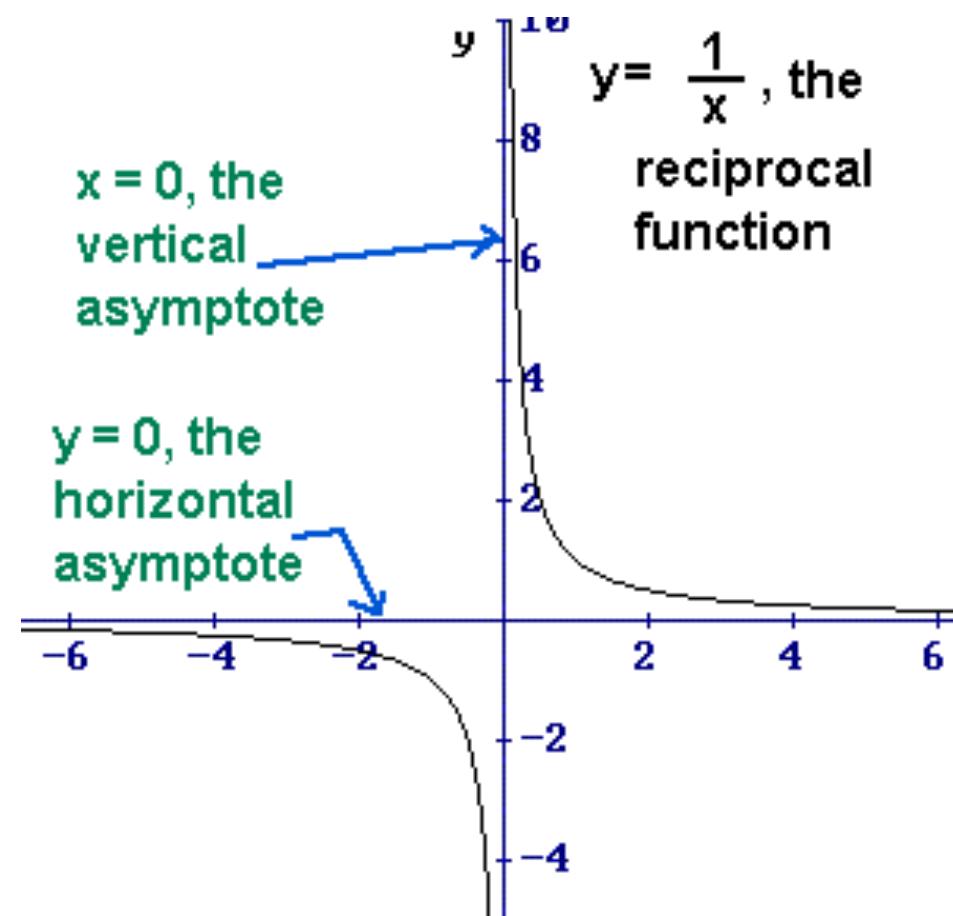
Непрерывность

Разрыв второго рода

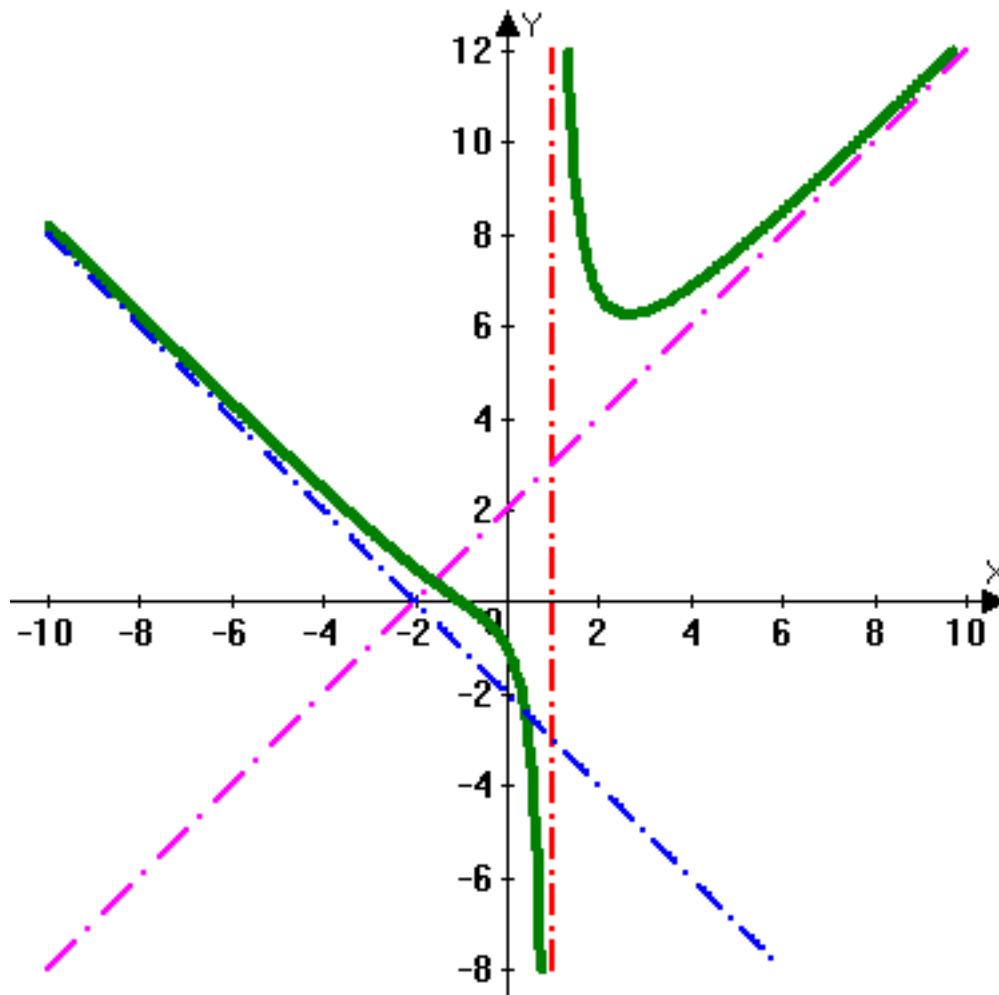


Разрыв первого рода

Асимптоты



Асимптоты



Скорость роста

- Численность населения:

1950	1960	1970	1980	1990	2000
2,525,778,669	3,026,002,942	3,691,172,616	4,449,048,798	5,320,816,667	6,127,700,428

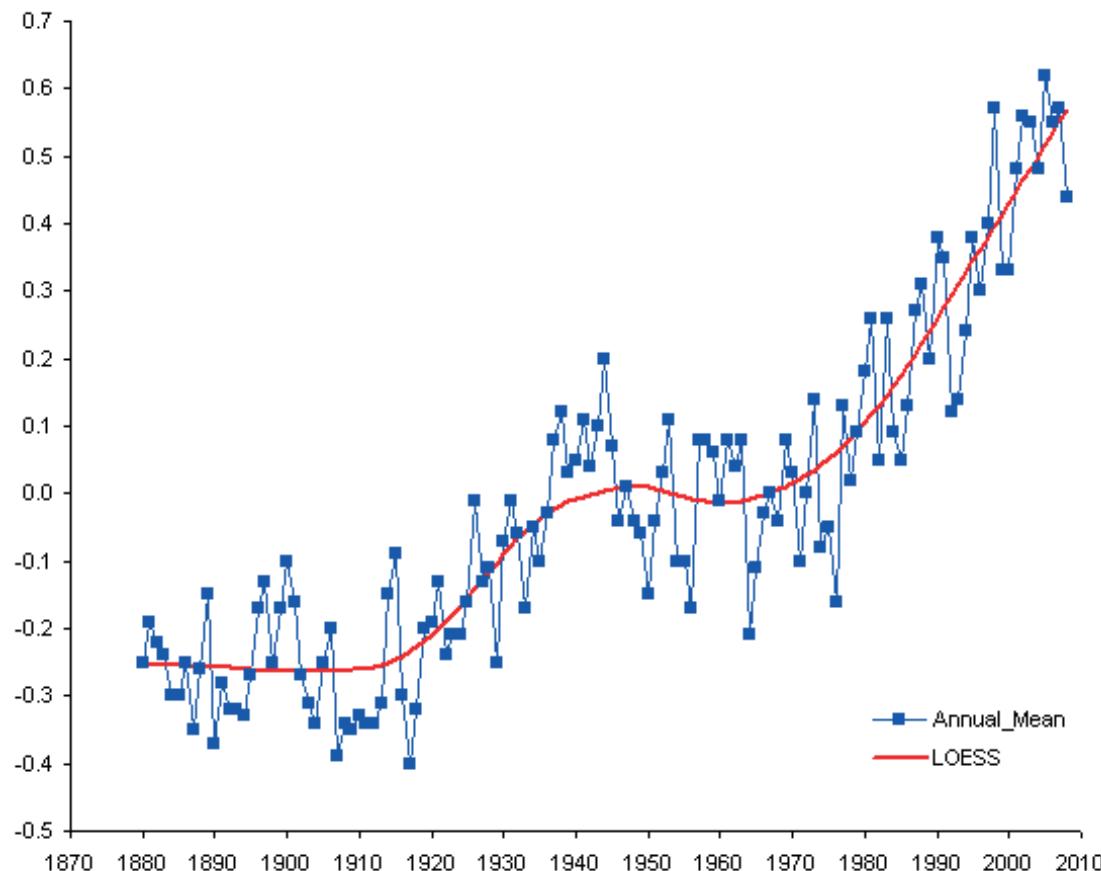
- Скорость роста между 1990 и 2000:

$$\frac{6127700428 - 5320816667}{10} = 80,688,376$$

- Дискретная величина

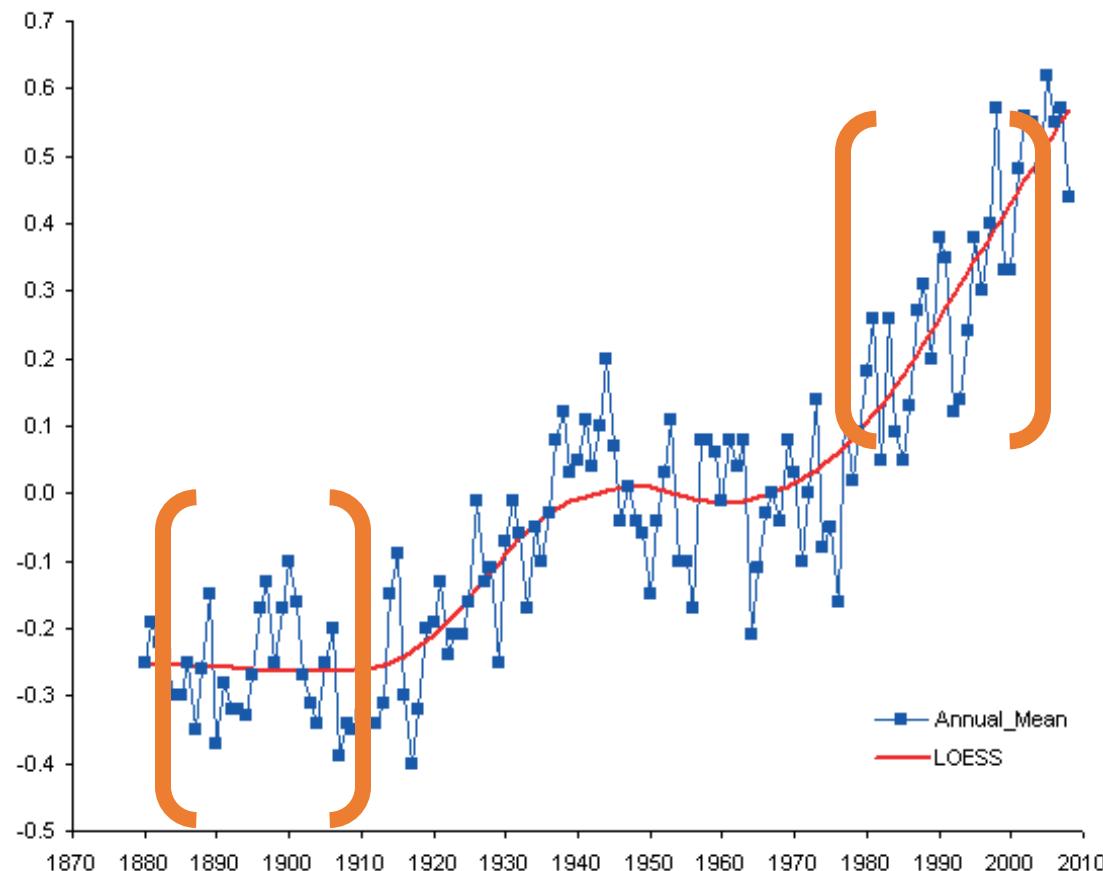
Скорость роста

- Отклонение температуры от нормы (непрерывная величина):



Скорость роста

- Отклонение температуры от нормы:



Скорость роста

- Можем измерить скорость на интервале $[x_0, x]$:

$$\frac{f(x) - f(x_0)}{x - x_0}$$

- Как измерить мгновенную скорость в конкретный момент x_0 ?
- Устремим x к x_0 !

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

Скорость роста

- Можем измерить скорость на интервале $[x_0, x]$:

$$\frac{f(x) - f(x_0)}{x - x_0}$$

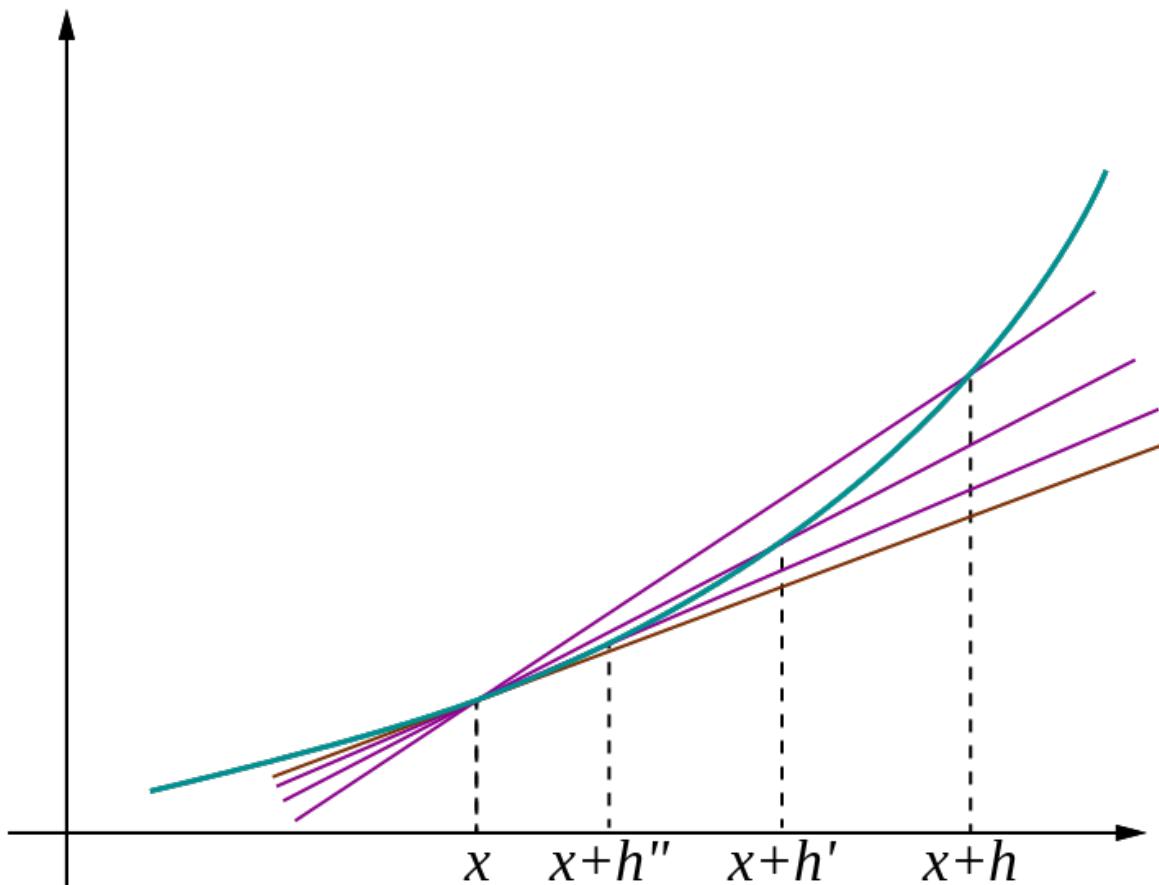
- Как измерить мгновенную скорость в конкретный момент x_0 ?
- Устремим x к x_0 !

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



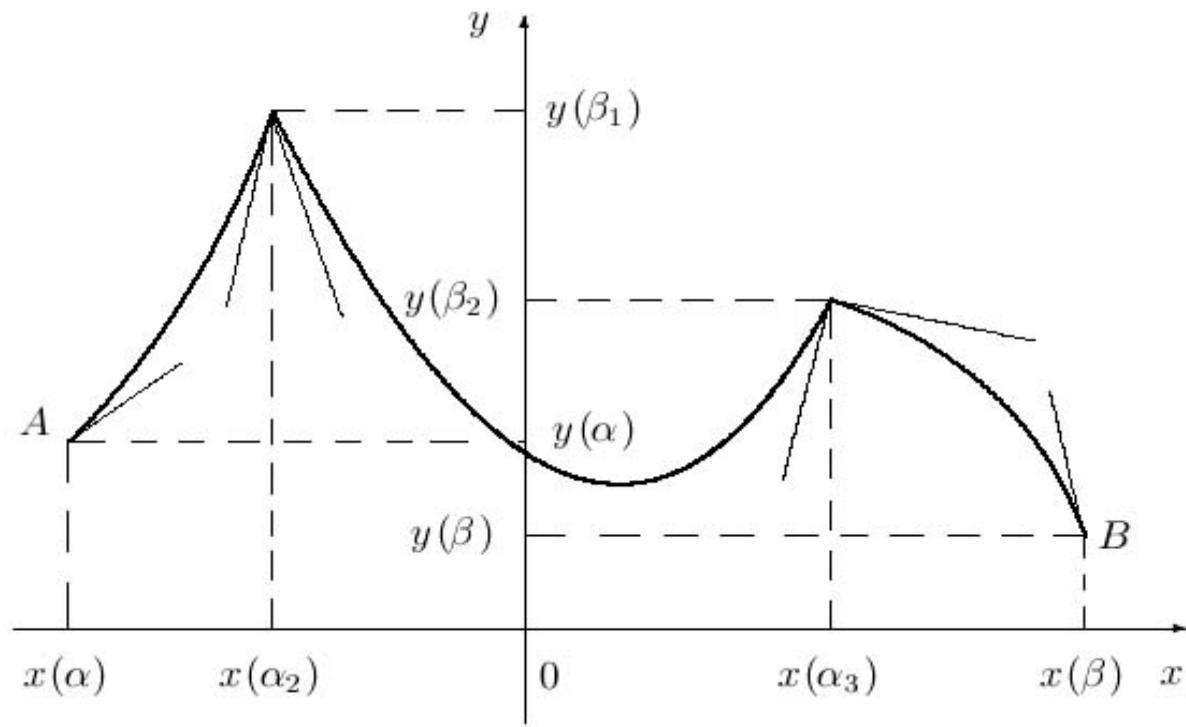
Производная

Производная



Гладкость

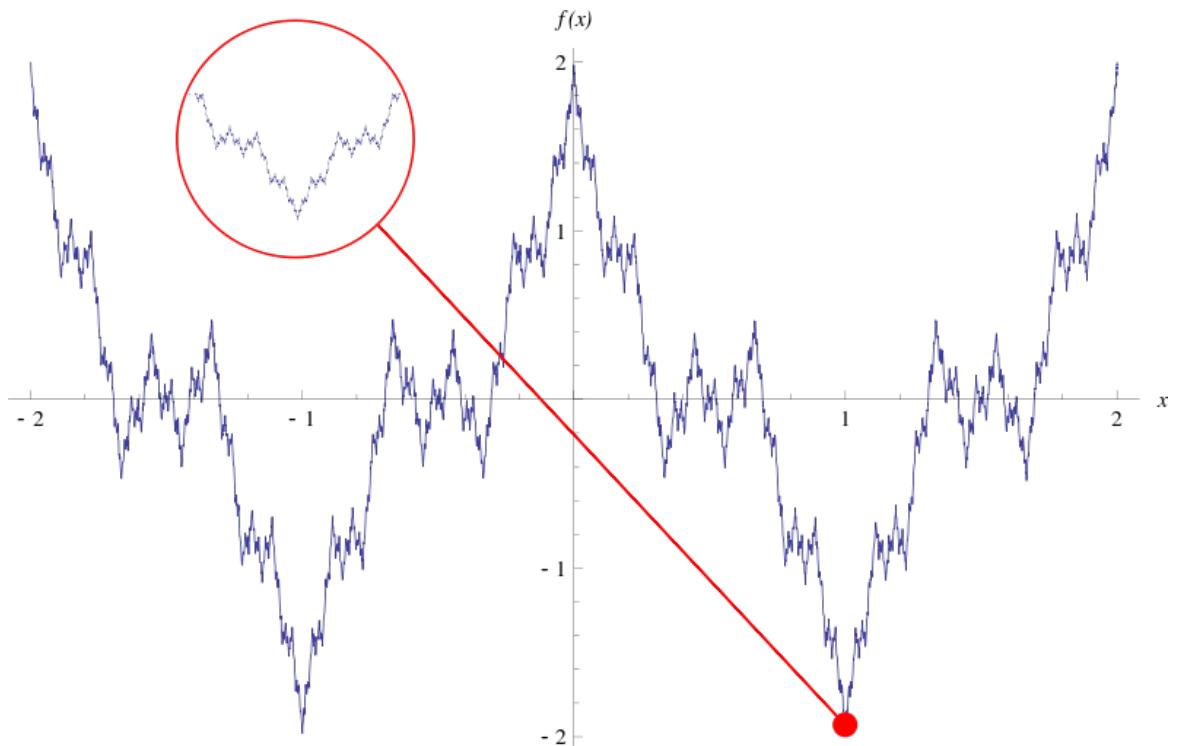
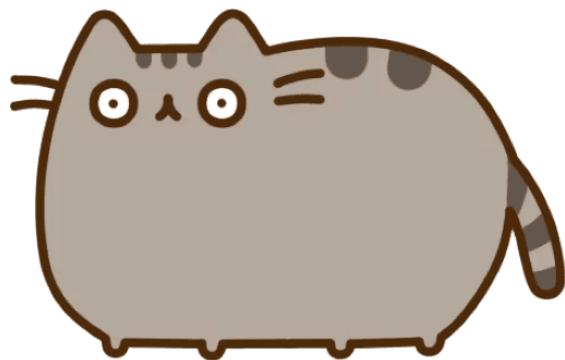
- Функция гладкая в точке, если там непрерывная производная
- Неформально: у графика нет углов



Гладкость

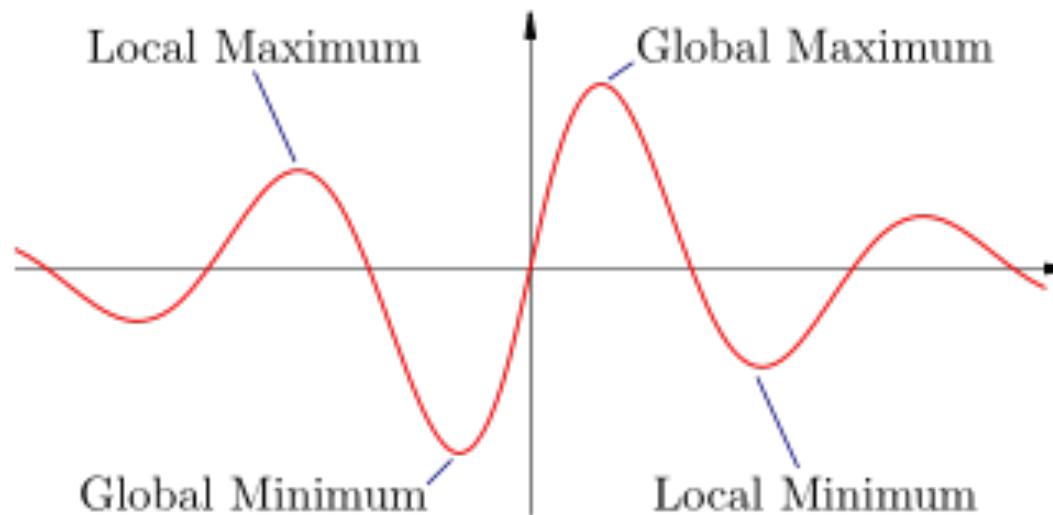
- Пример нигде не гладкой функции (функция Вейерштрасса):

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$



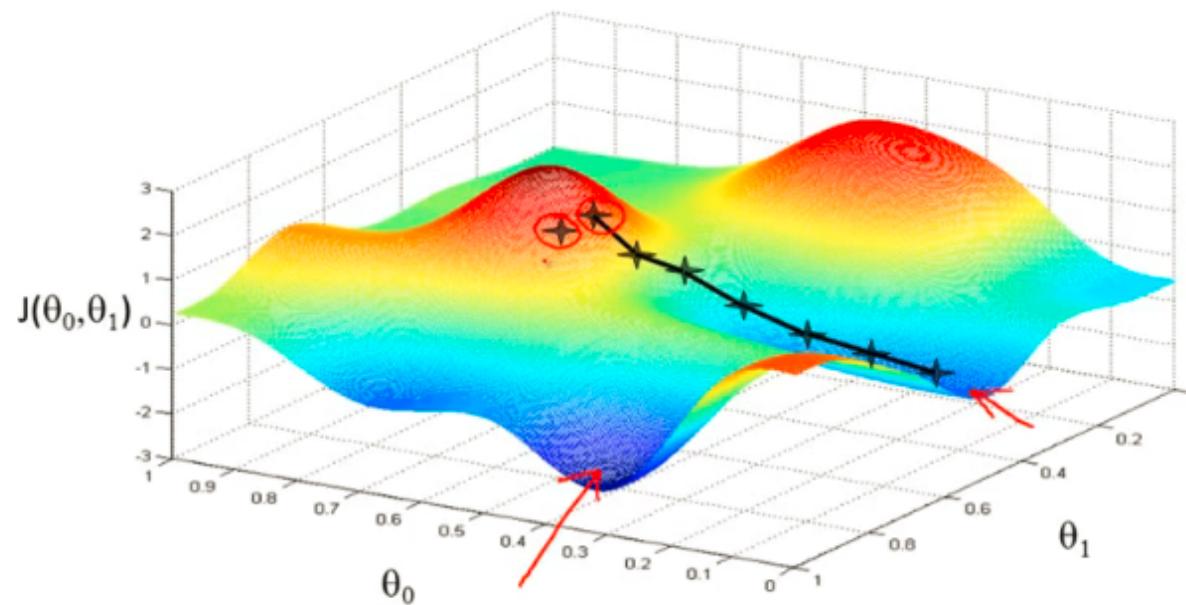
Экстремумы

- Экстремум — минимум или максимум
- Локальный минимум — меньше всех значений в некоторой окрестности
- Глобальный минимум — меньше всех значений



Экстремумы

- Локальные минимумы — одна из главных проблем в машинном обучении

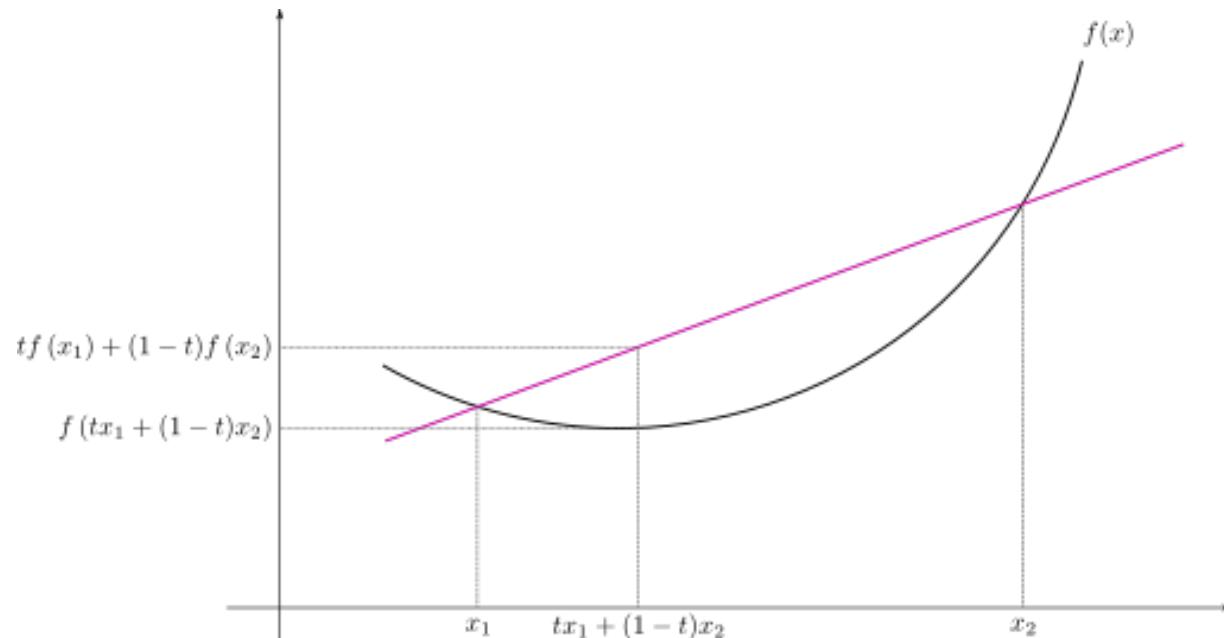


Условие оптимальности

- Как понять, является ли точка x_0 экстремумом?
- Теорема Ферма: если точка x_0 — экстремум, и в ней существует производная, то $f'(x_0) = 0$
- Если функция везде имеет производную: решаем $f'(x) = 0$
- Если с производной проблемы: не повезло
- Даже если производная есть, то что делать с локальными экстремумами?

Выпуклые функции

- Функция выпуклая, если ее график лежит ниже любого отрезка, соединяющего две точки

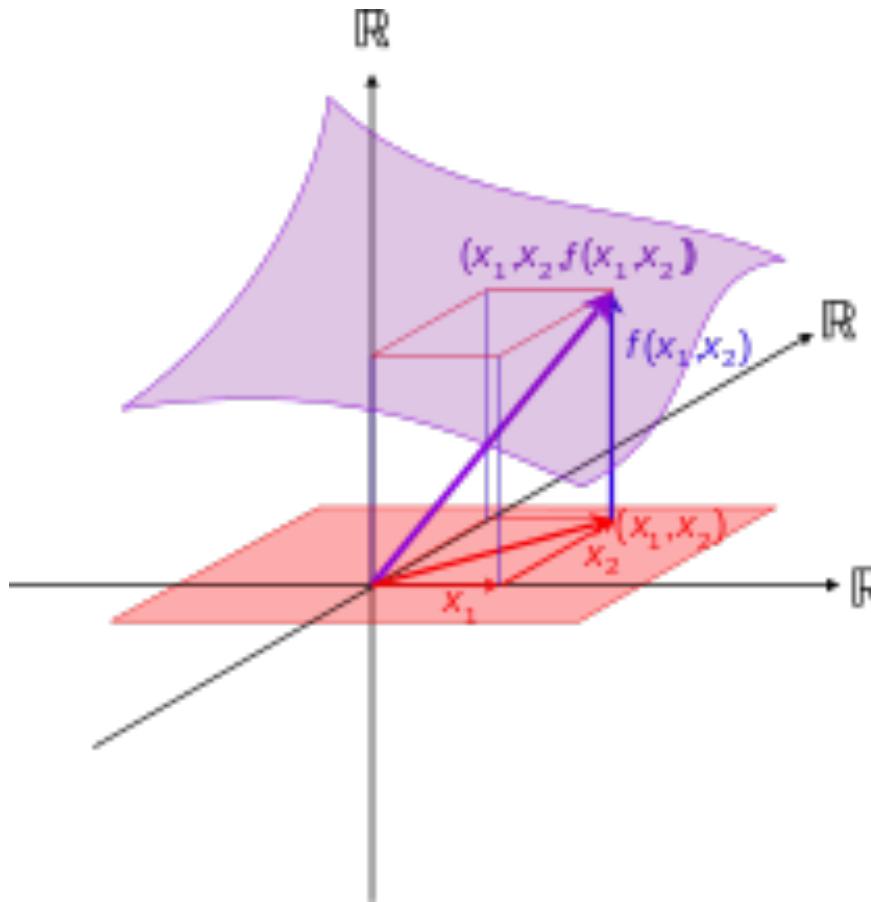


Выпуклые функции

- Функция выпуклая, если во всех точках $f''(x) \geq 0$
- Важное свойство: любой локальный экстремум выпуклой функции является глобальным
- Решая уравнение $f'(x) = 0$, получим глобальные экстремумы
- Вывод: будем стараться выбирать выпуклые функционалы!

Функции многих переменных

ФУНКЦИЯ



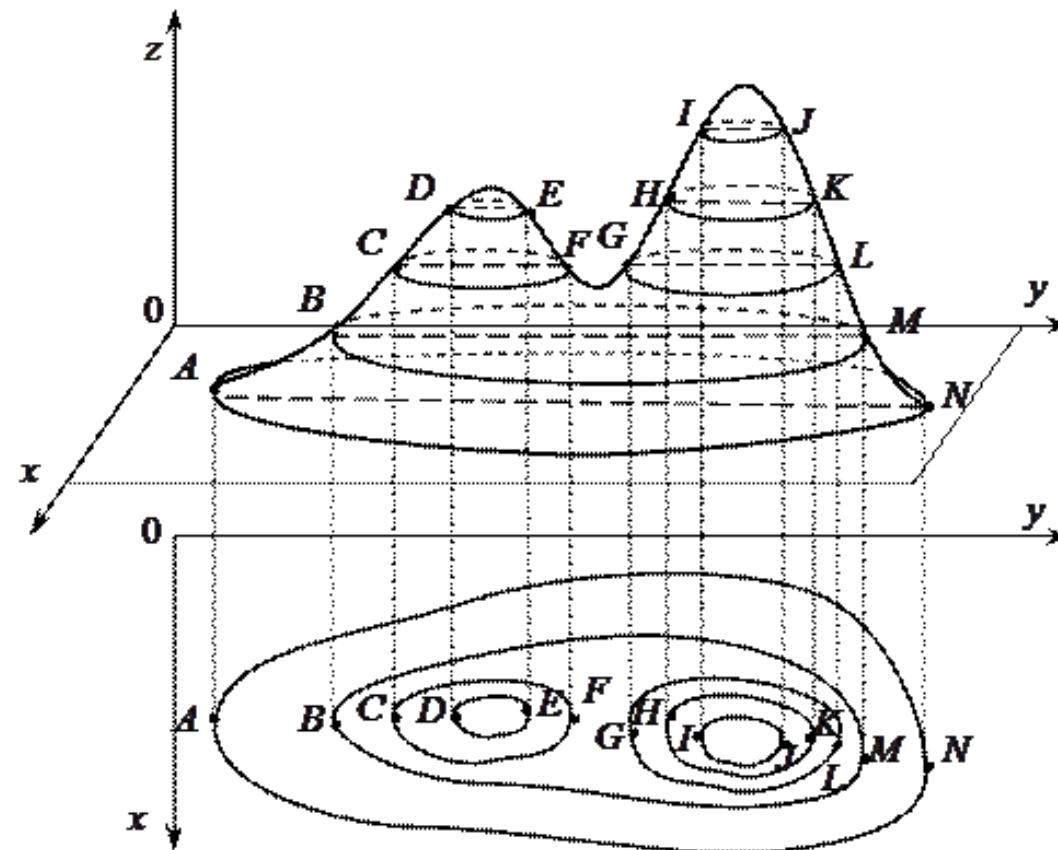
Пример

- Функционал качества линейной регрессии:

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

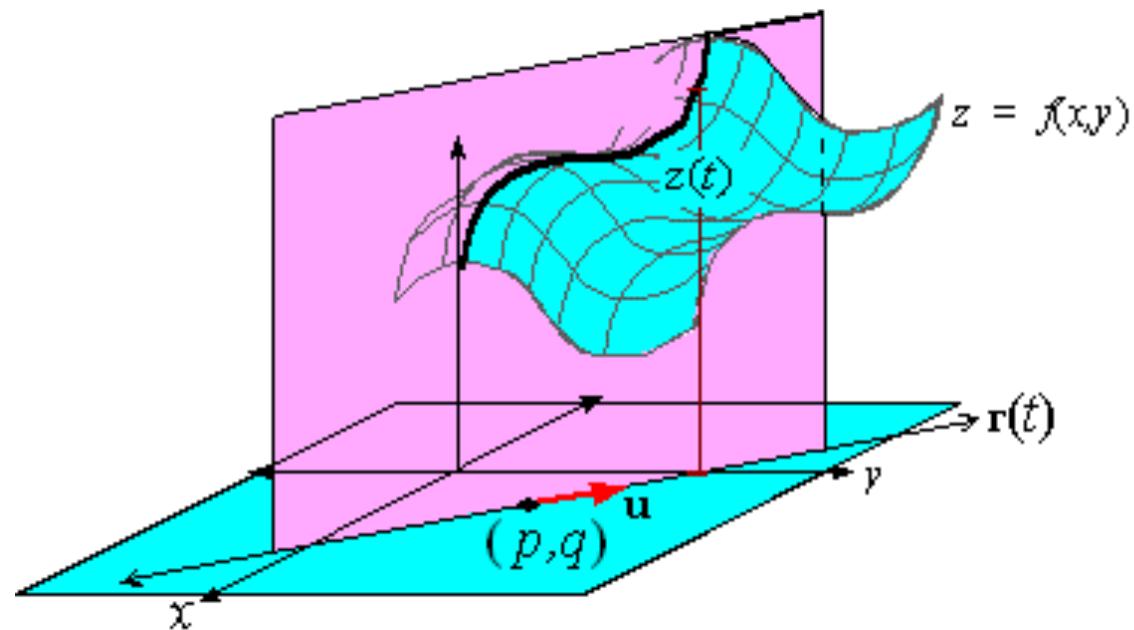
- Как искать ее минимум?

Линии уровня



Производная по направлению

- С какой скоростью растет функция в конкретном направлении?



Производная по направлению

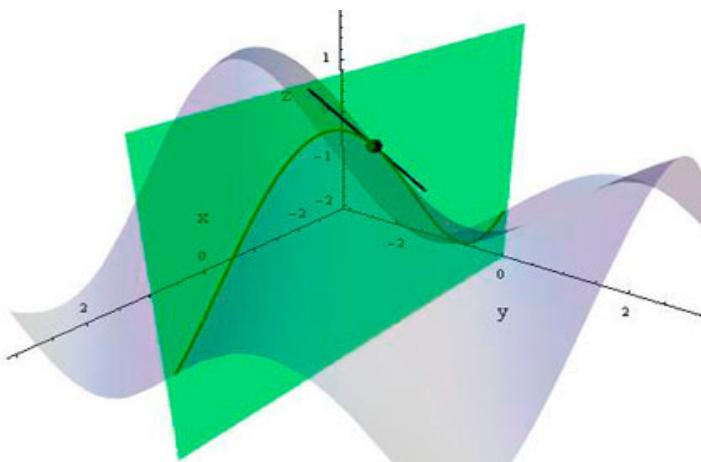
- Направление: ν , причем $\|\nu\| = 1$
- Производная:

$$f'_\nu(x_0) = \lim_{t \rightarrow 0} \frac{f(x_0 + t\nu) - f(x_0)}{t}$$

Частные производные

- С какой скоростью функция меняется вдоль переменной x_i ?
- Частная производная по x_i :

$$\frac{\partial f}{\partial x_i} = \lim_{t \rightarrow 0} \frac{f(x_1, \dots, x_i + t, \dots, x_d) - f(x_1, \dots, x_i, \dots, x_d)}{t}$$



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 растет?

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

Угол между градиентом и
направлением

- Связь производной по направлению и градиента:

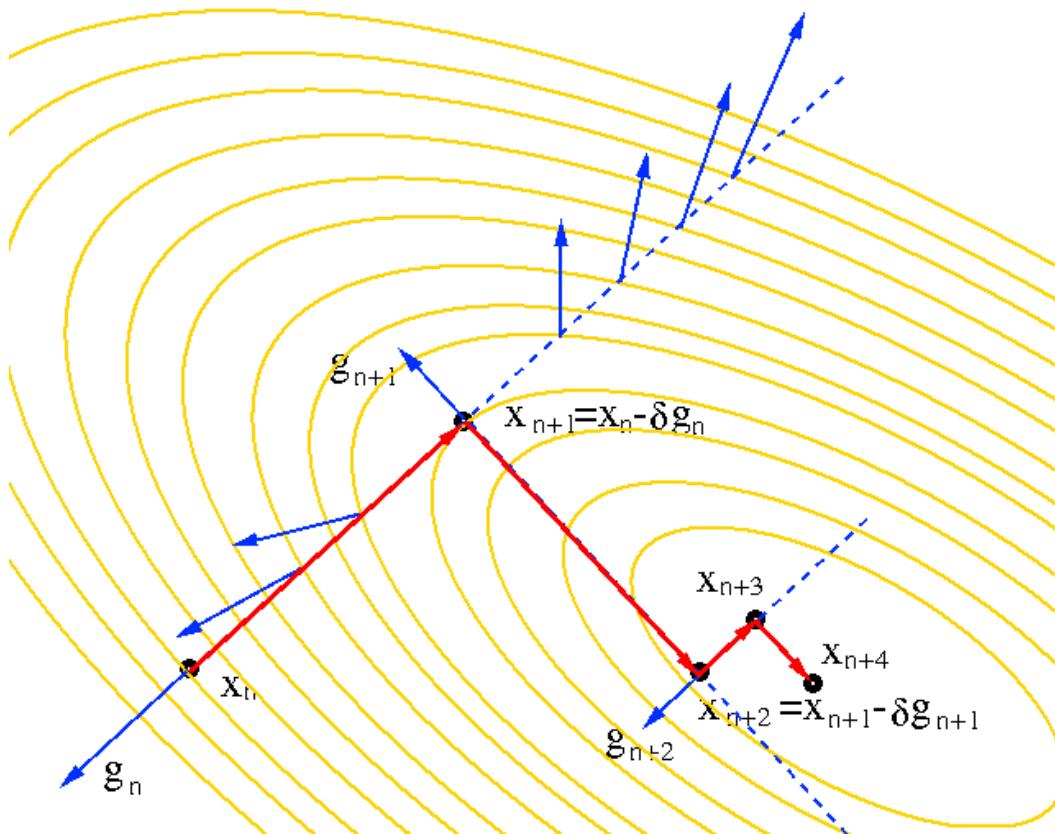
$$f'_v(x_0) = \langle \nabla f(x_0), v \rangle = \|\nabla f(x_0)\| * \|v\| * \cos \varphi$$

Градиент

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

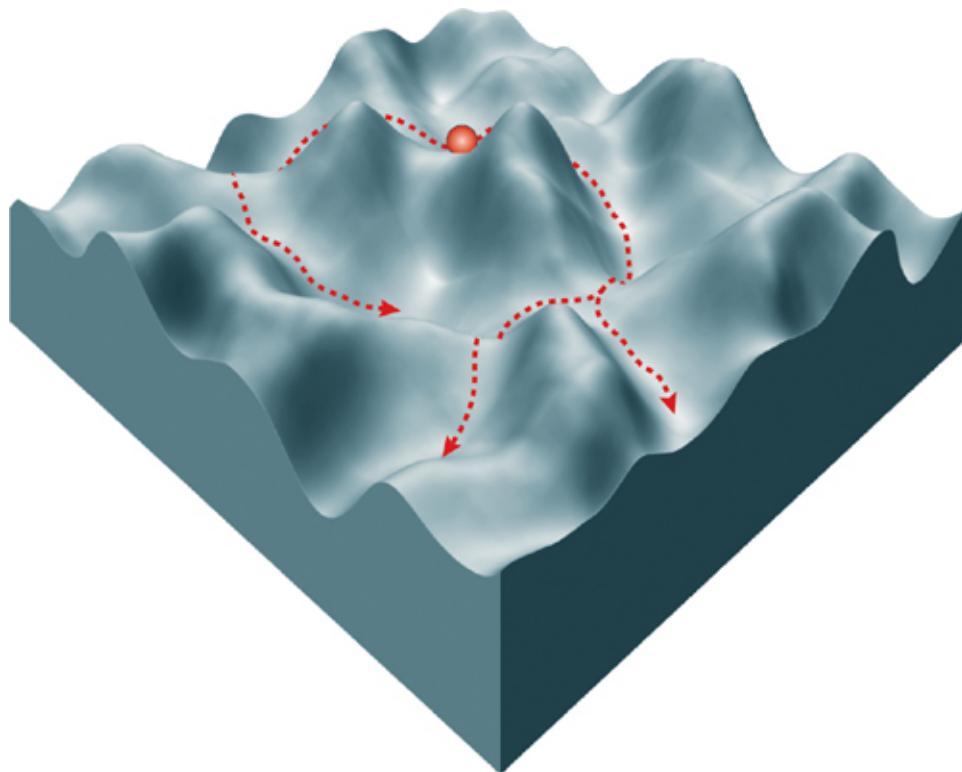
Градиент

- Еще одно свойство: градиент ортогонален линии уровня



Экстремумы

- Проблема с локальными экстремумами все еще актуальна

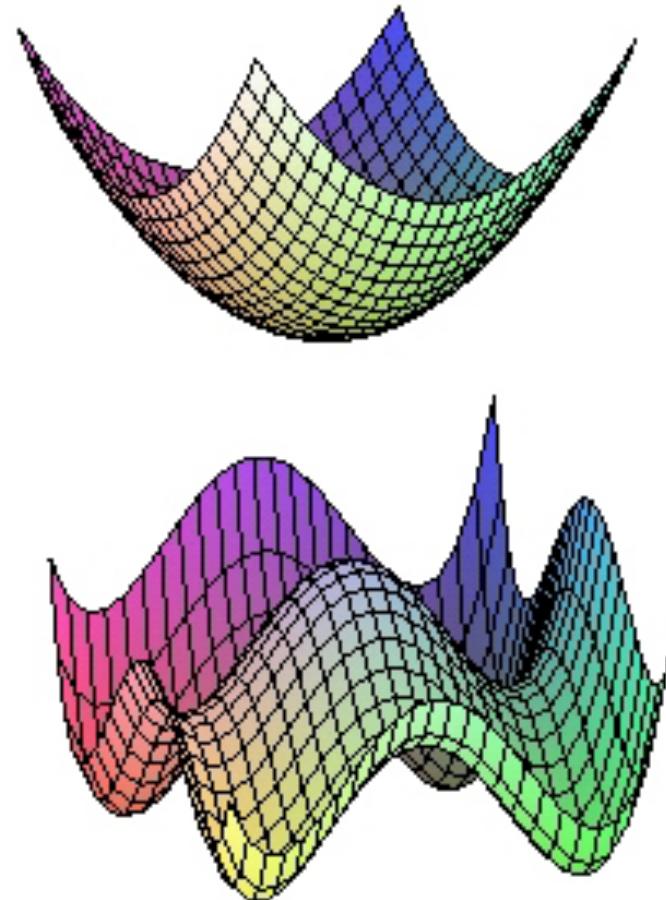


Условие оптимальности

- 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

Выпуклые функции

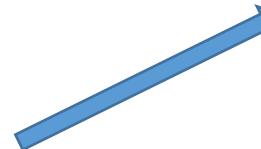
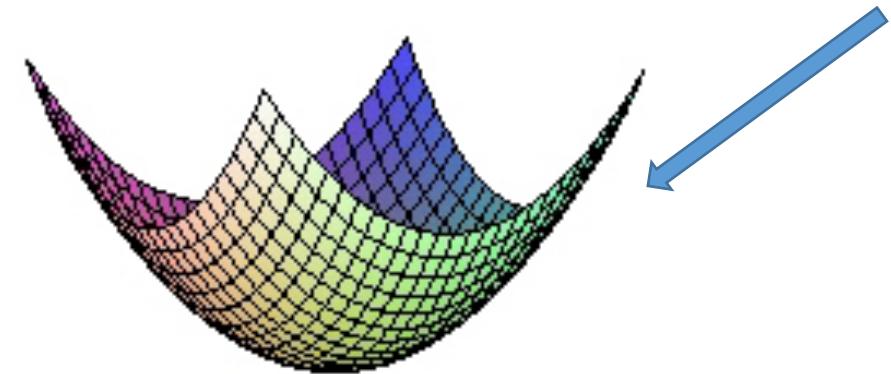
- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки



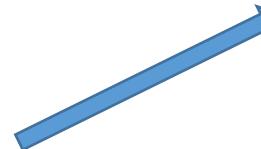
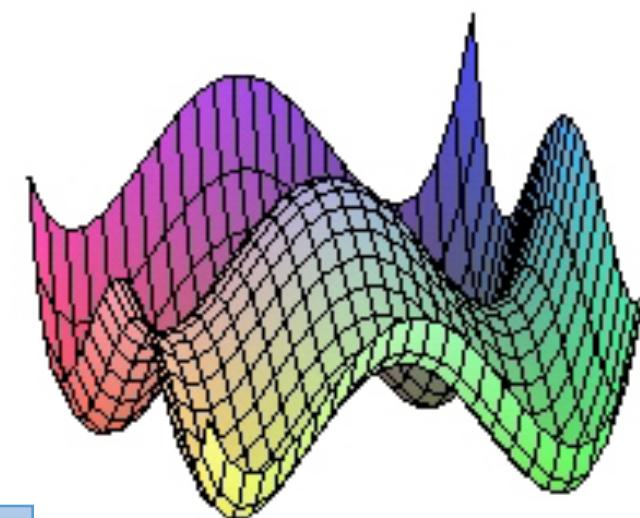
Выпуклые функции

- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки

Выпуклая функция



Невыпуклая функция



Выпуклые функции

- Функция выпуклая, если матрица Гессе (матрица вторых производных) H неотрицательно определена в любой точке:

$$x^T H x \geq 0 \quad \forall x$$

- Важное свойство: любой локальный экстремум выпуклой функции является глобальным
- Вывод: будем стараться выбирать выпуклые функционалы!

Методы оптимизации

Поиск минимума

- Функционал качества линейной регрессии:

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

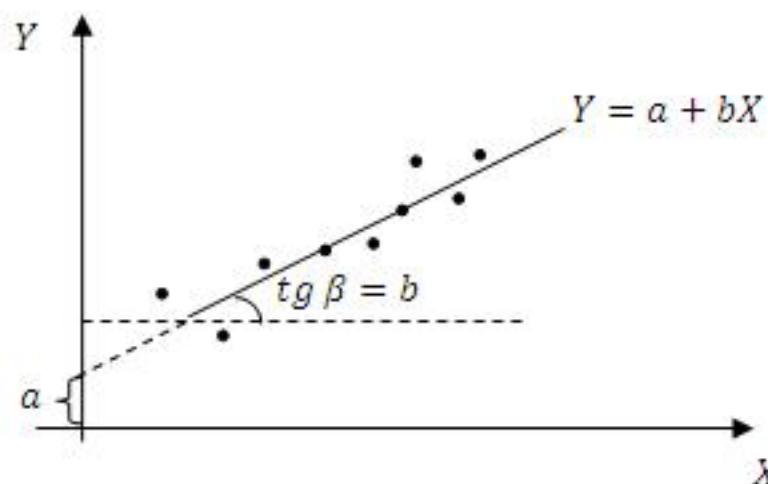
- Как искать минимум?

Поиск минимума

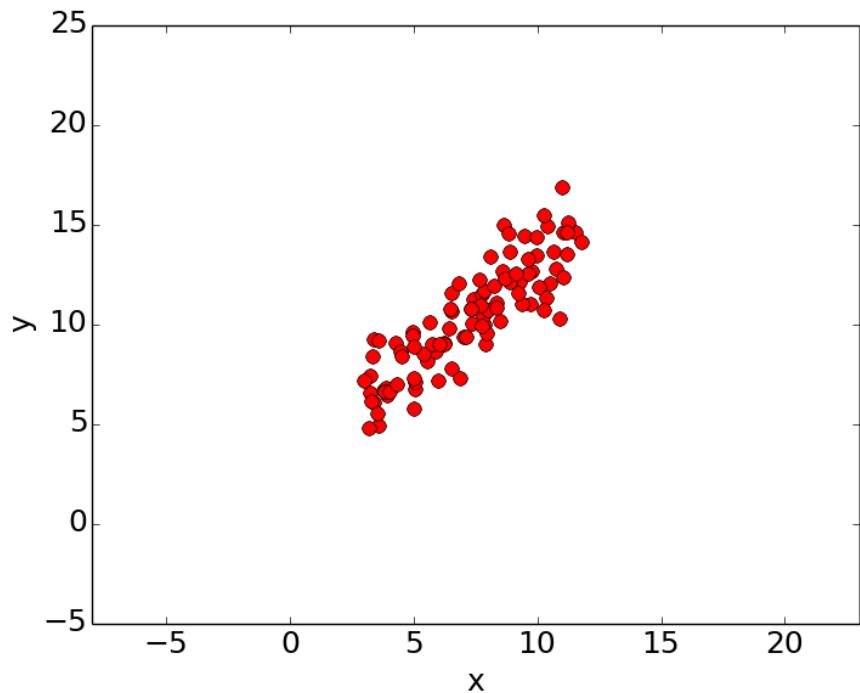
- Можно решать уравнение: $\nabla Q(w) = 0$
- А если уравнение сложное, и аналитически решить нельзя?
- Нужна численная оптимизация

Парная регрессия

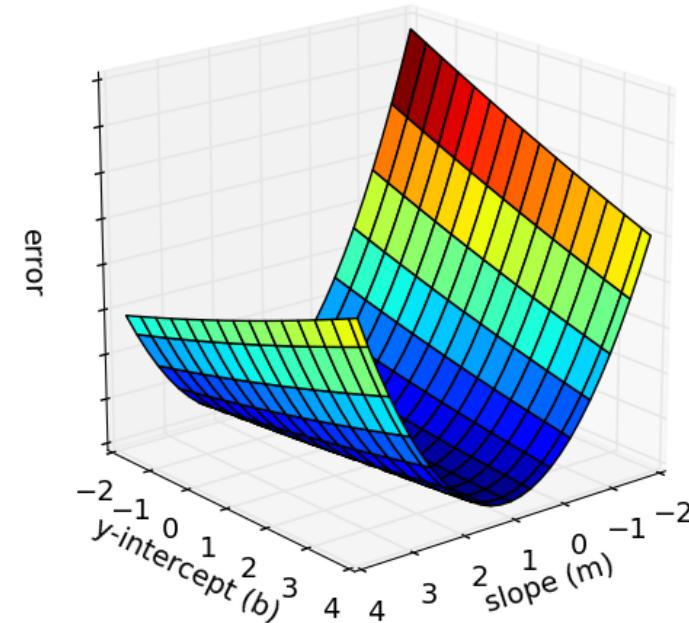
- Простейший случай: один признак
- Модель: $a(x) = w_1x + w_0$
- Два параметра: w_1 и w_0
- Функционал: $Q(w_0, w_1) = \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$



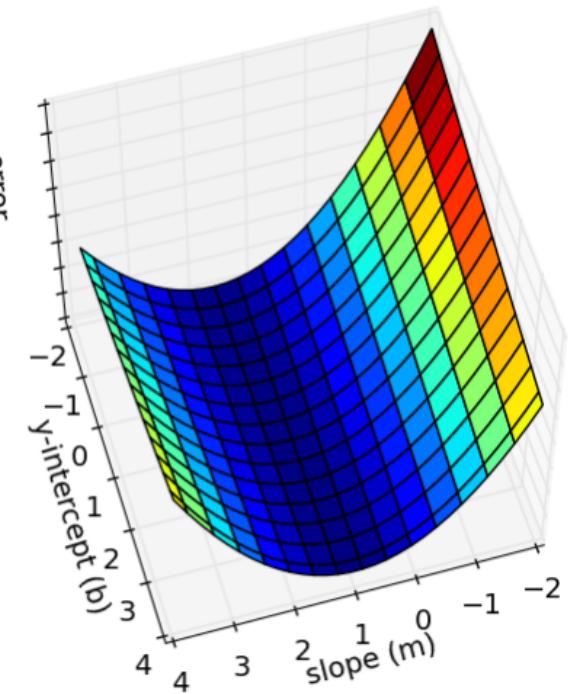
Парная регрессия



Выборка

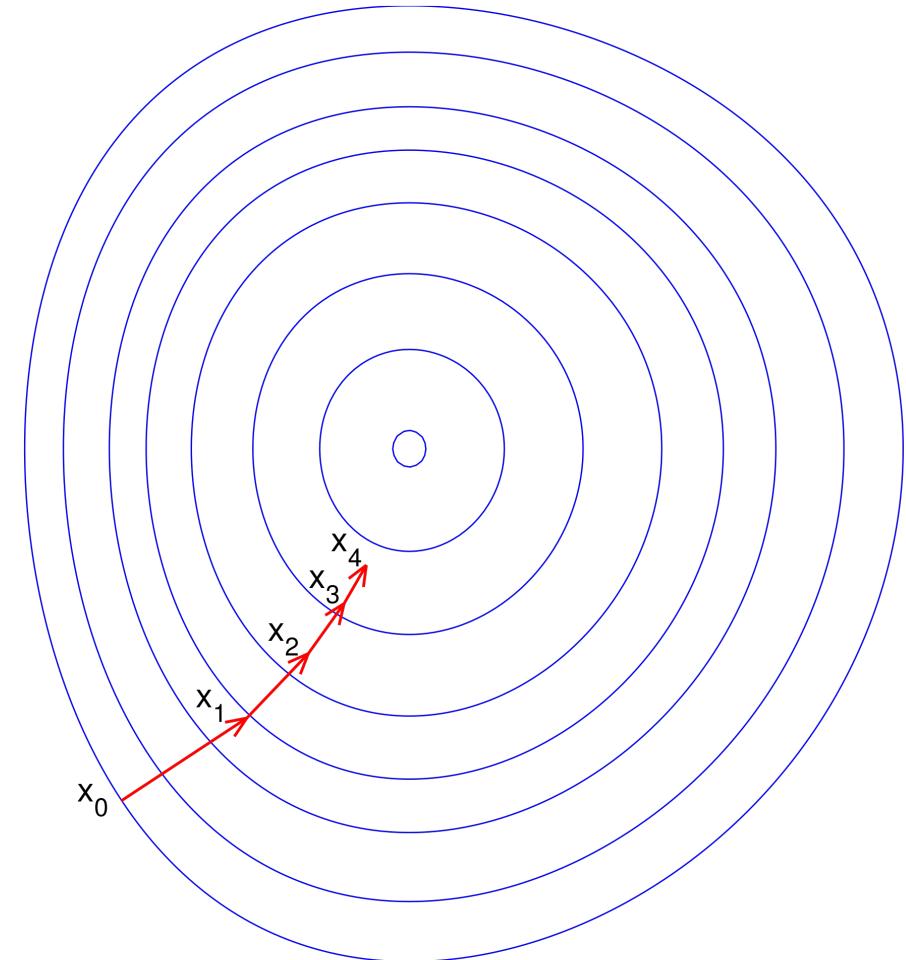


Функционал качества



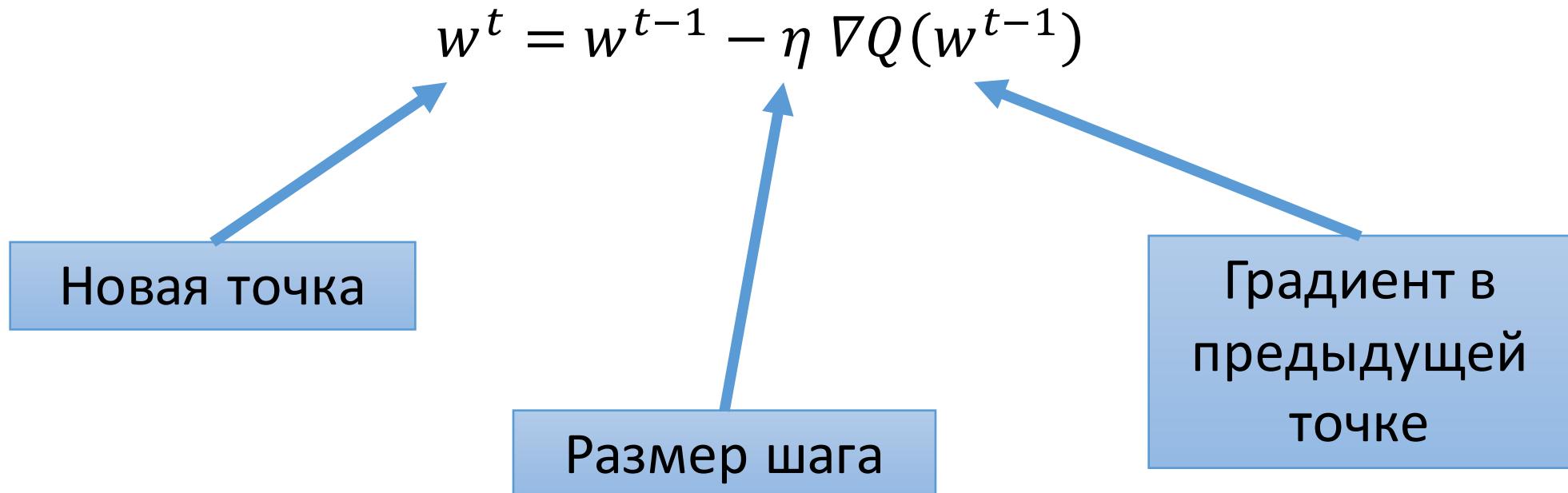
Градиентный спуск

- 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

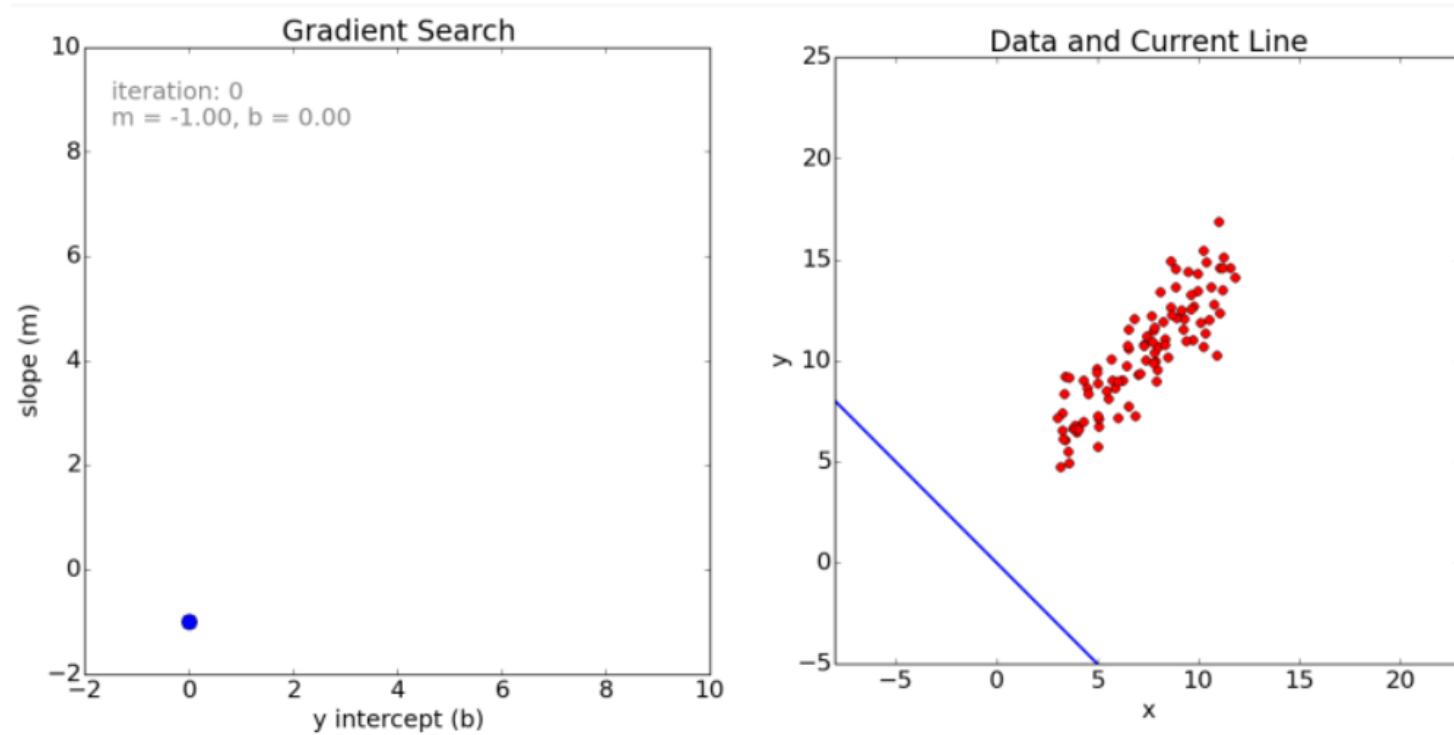
$$Q(w_0, w_1) = \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

- Частные производные:

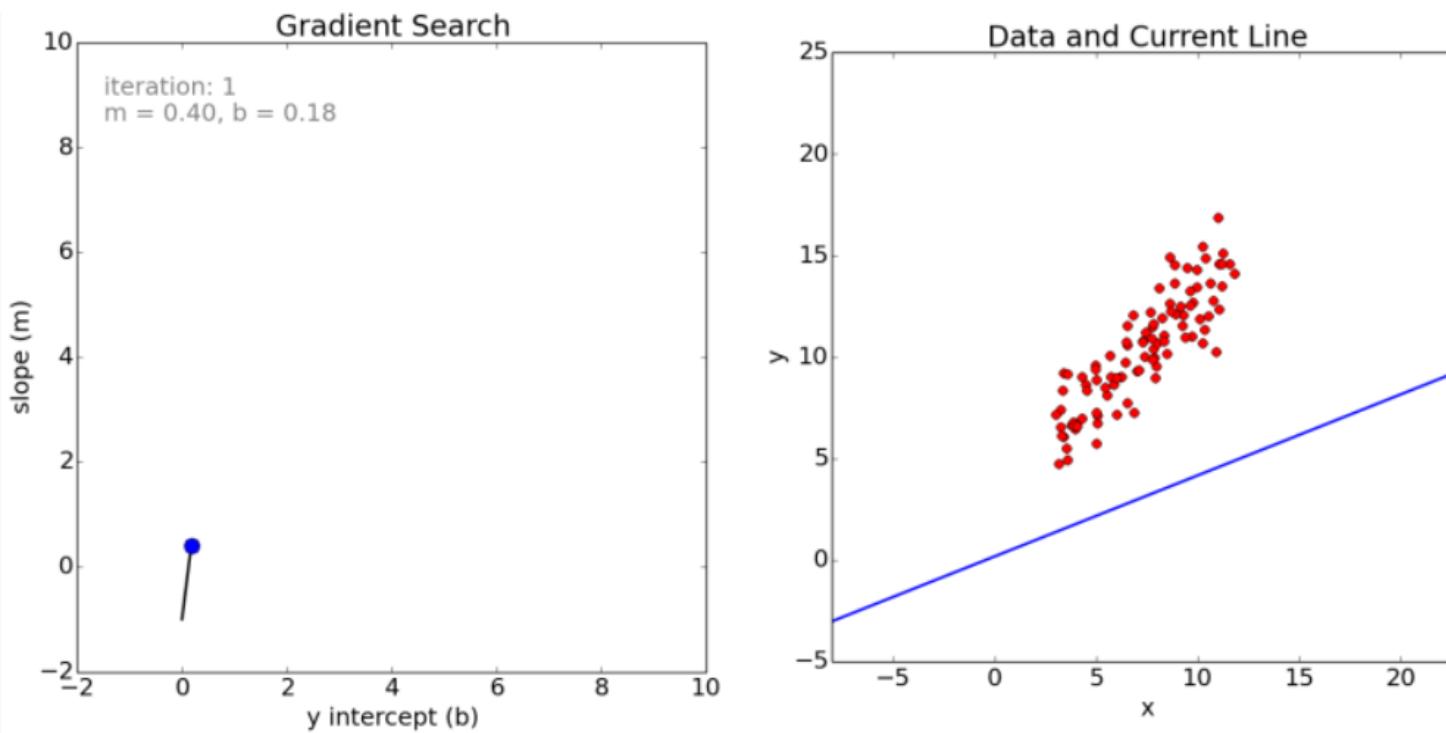
$$\frac{\partial Q}{\partial w_1} = 2 \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i) x_i$$

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

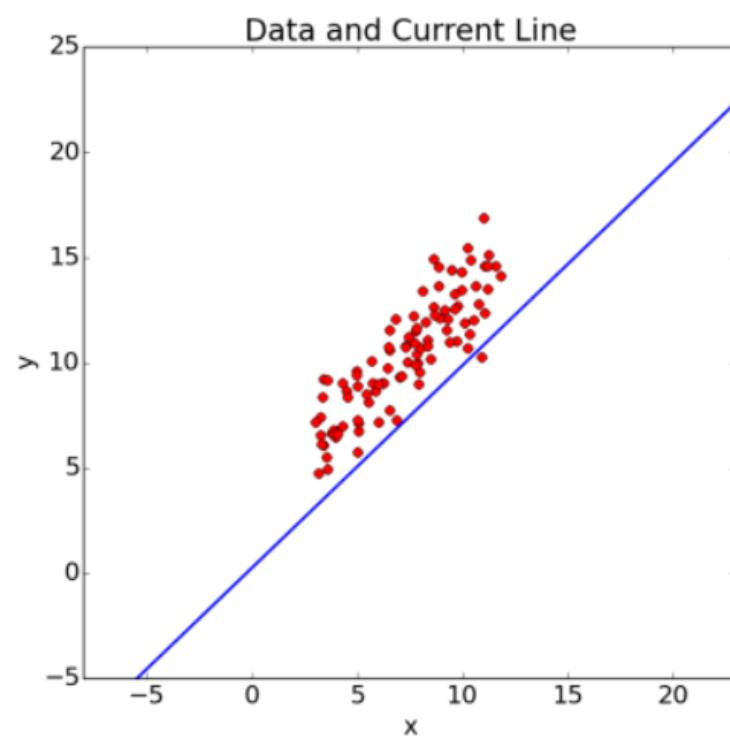
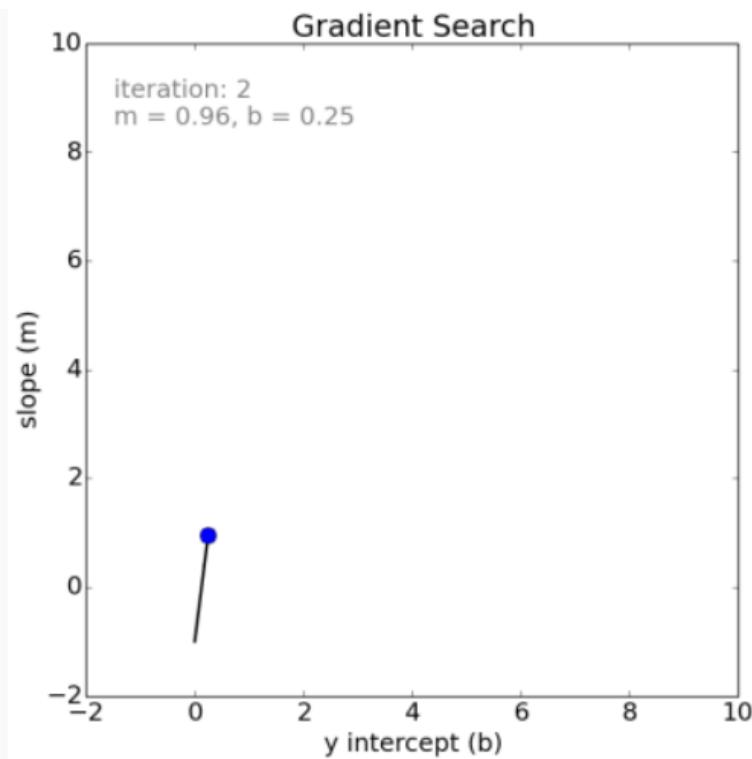
Парная регрессия



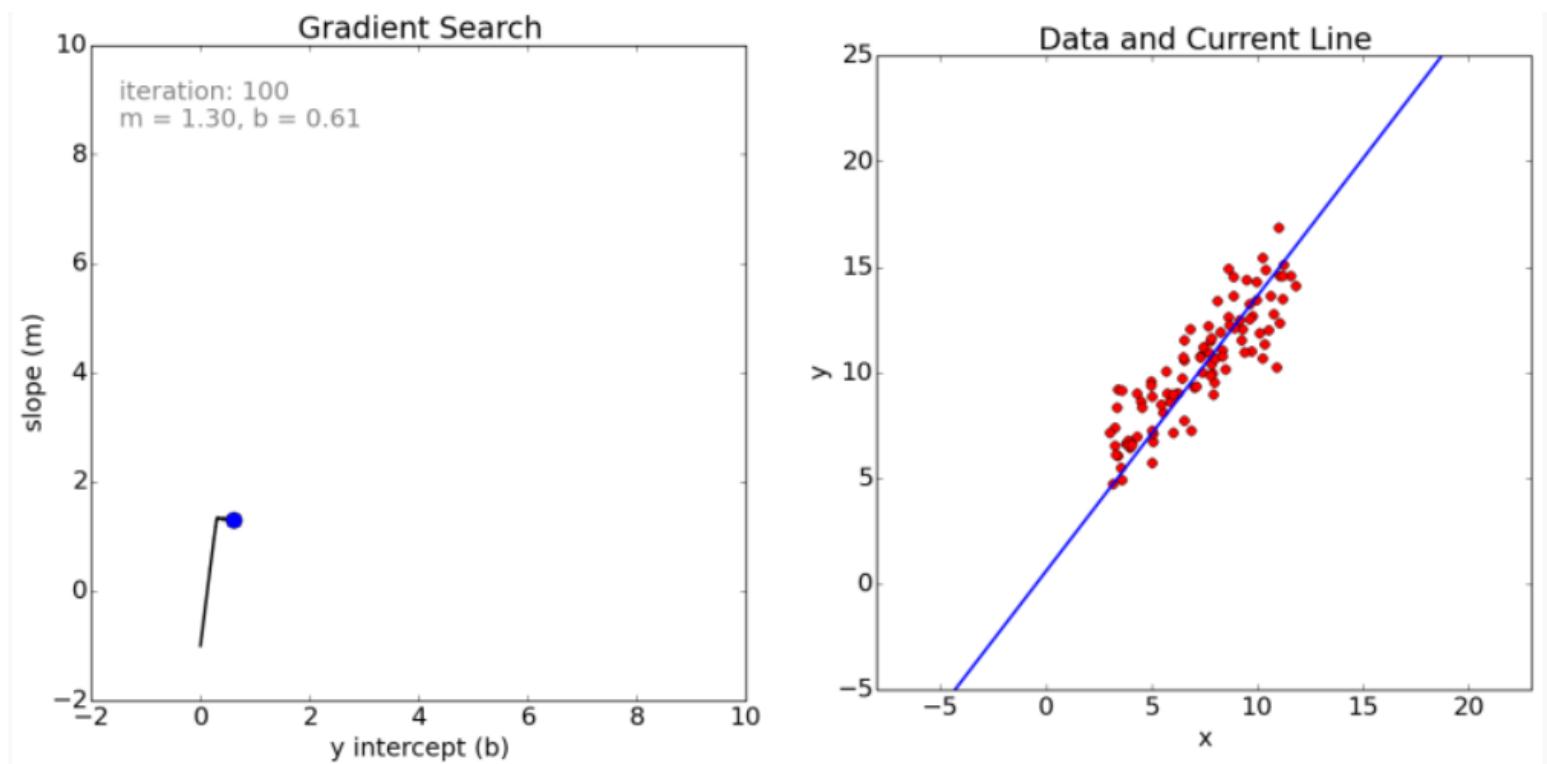
Парная регрессия



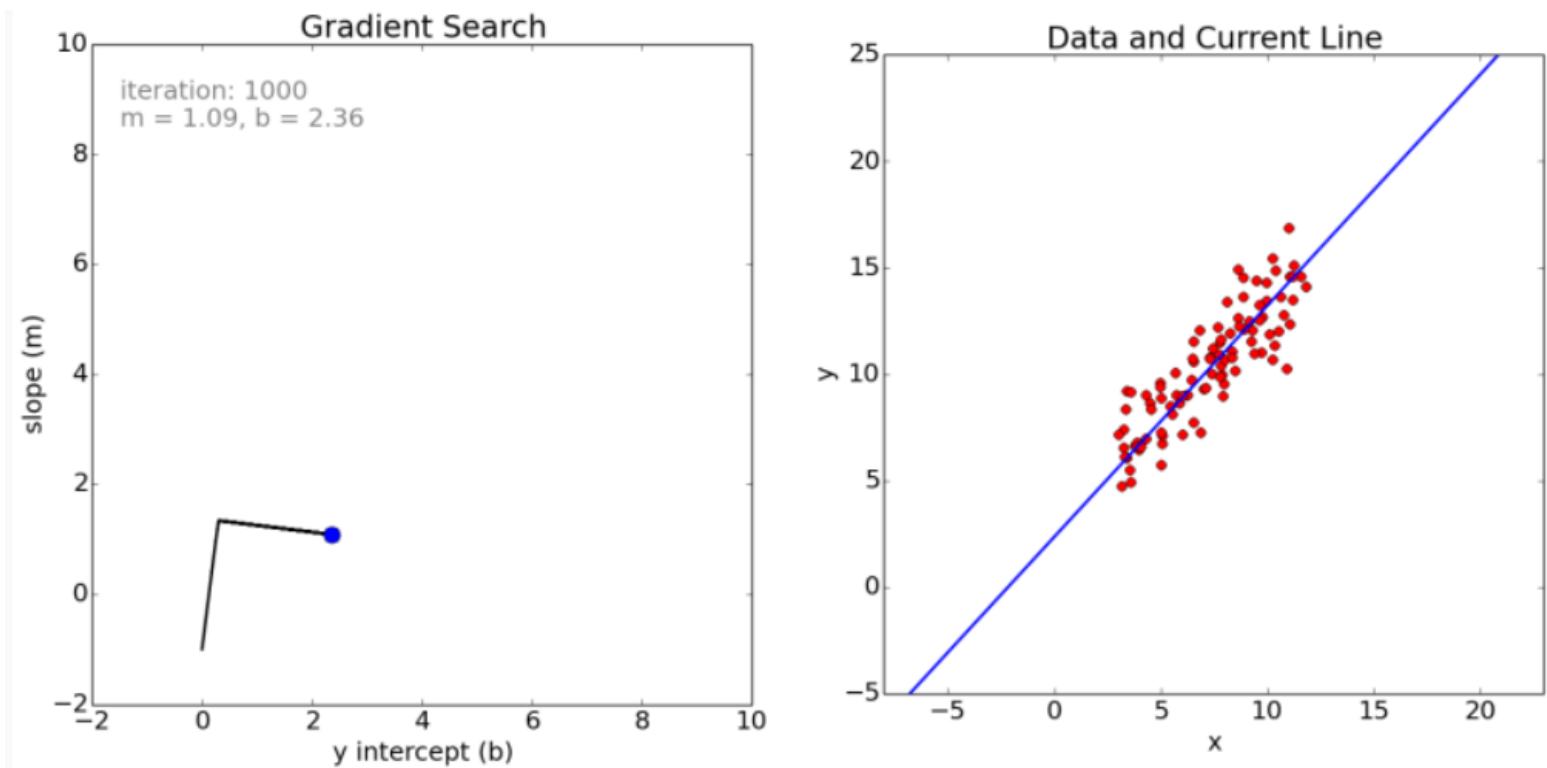
Парная регрессия



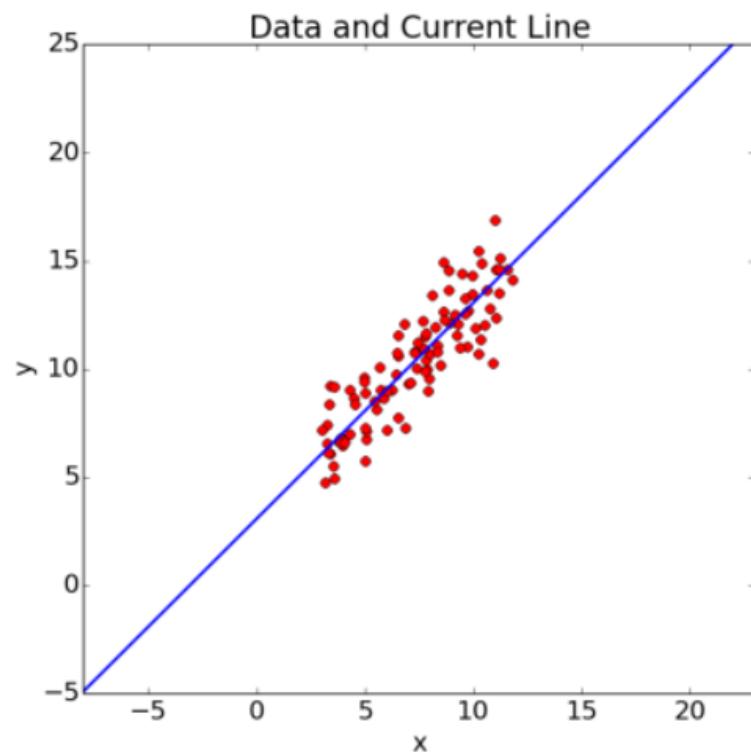
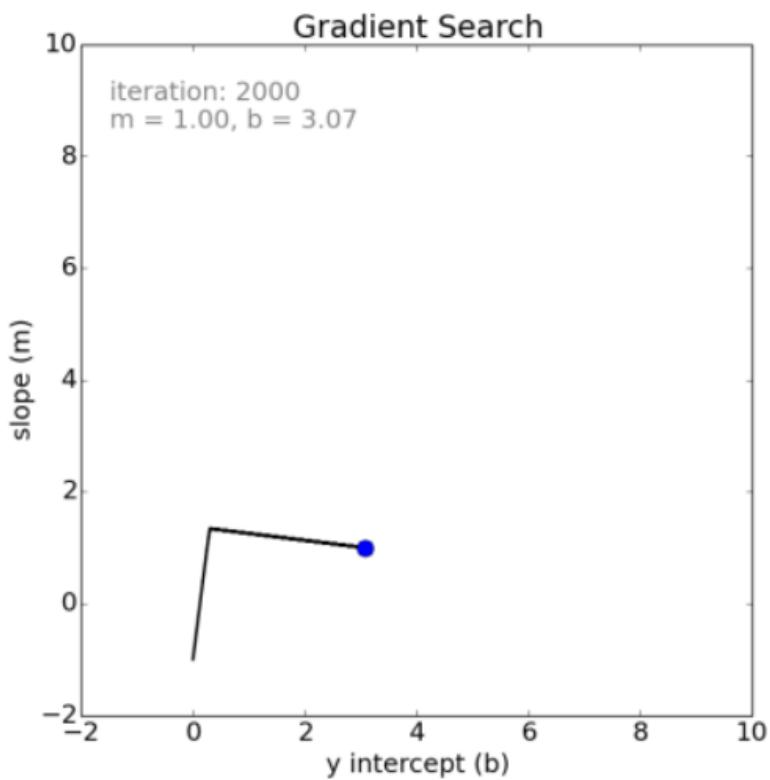
Парная регрессия



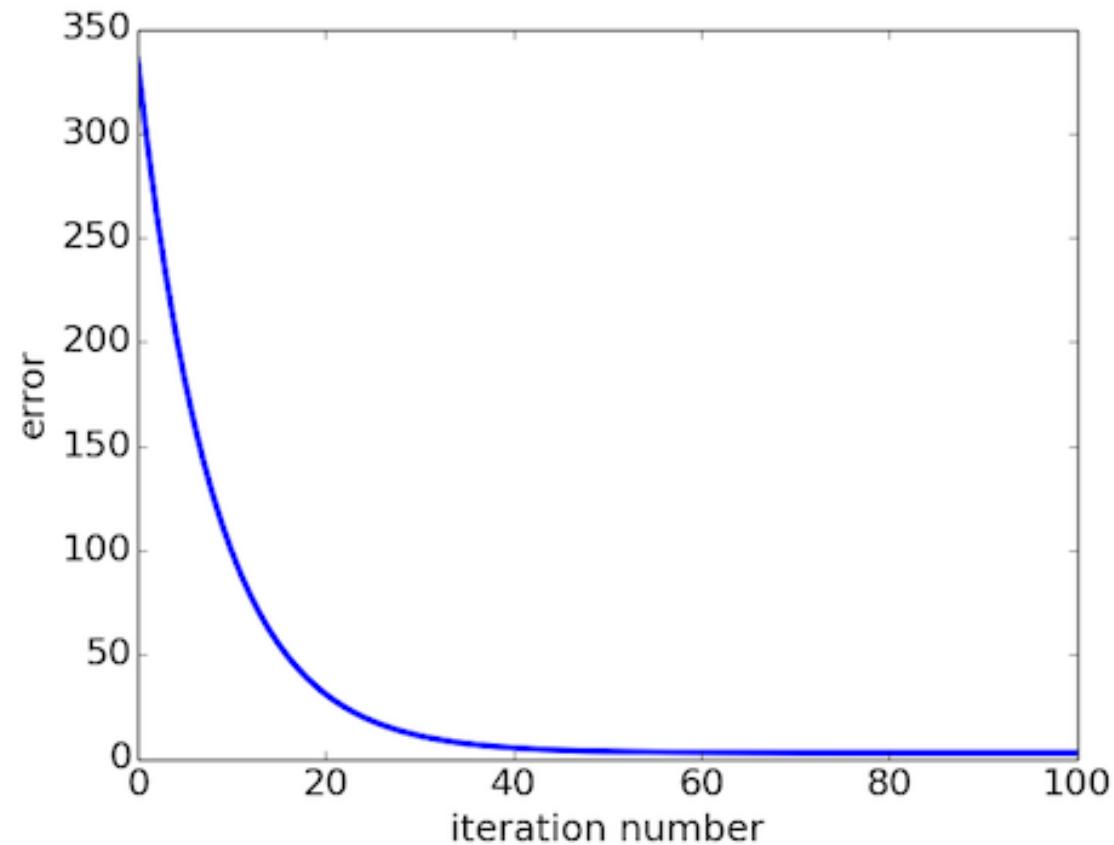
Парная регрессия



Парная регрессия

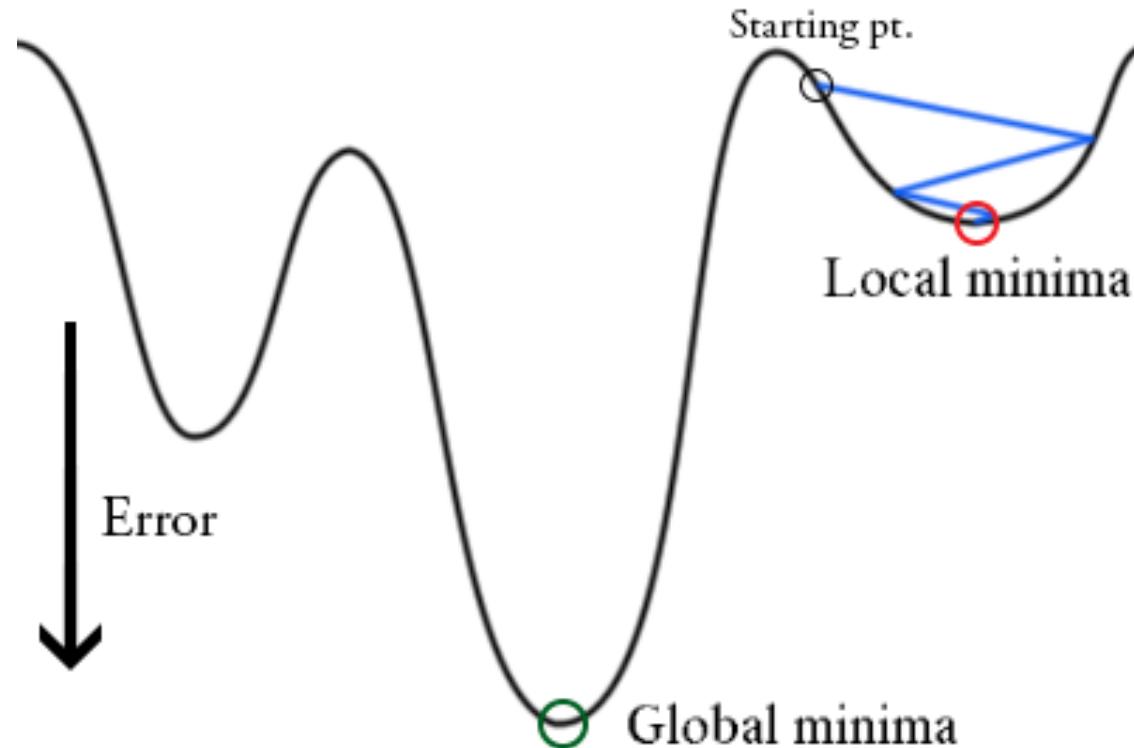


Функционал качества



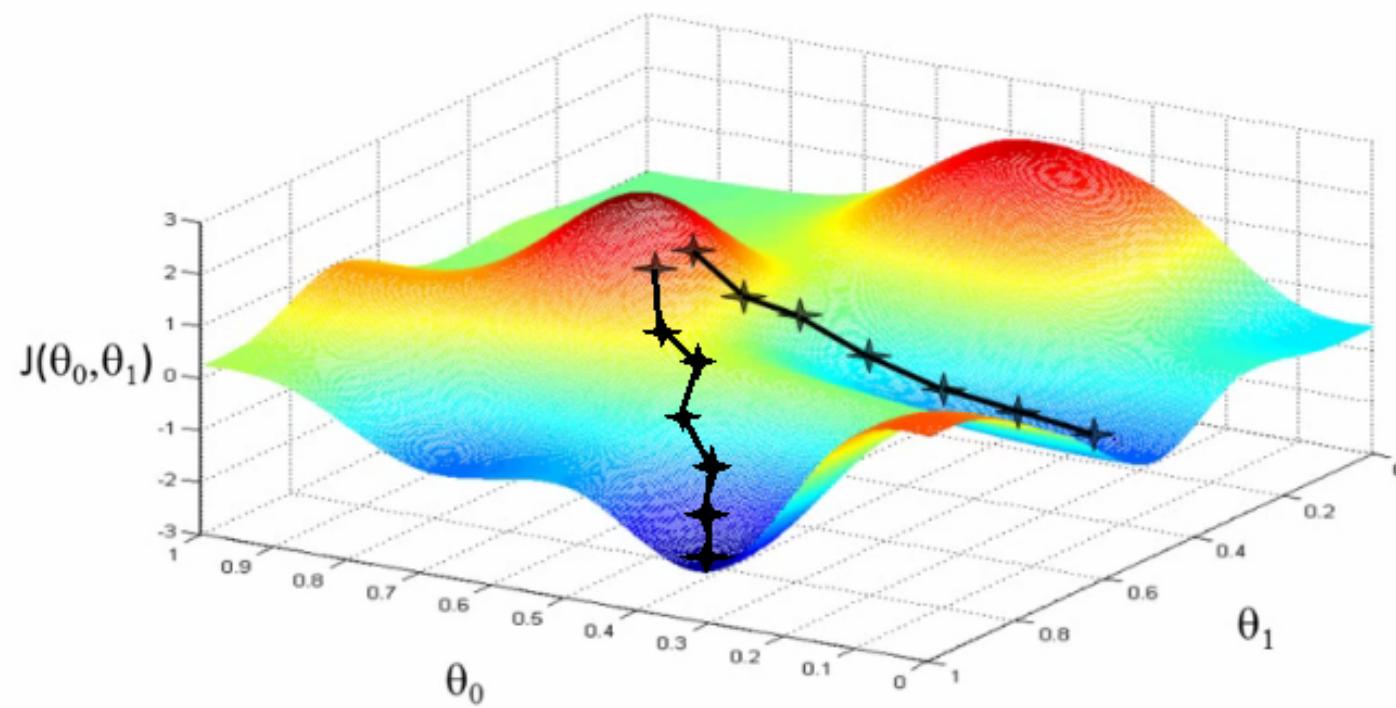
Локальные минимумы

- Градиентный спуск находит только локальные минимумы



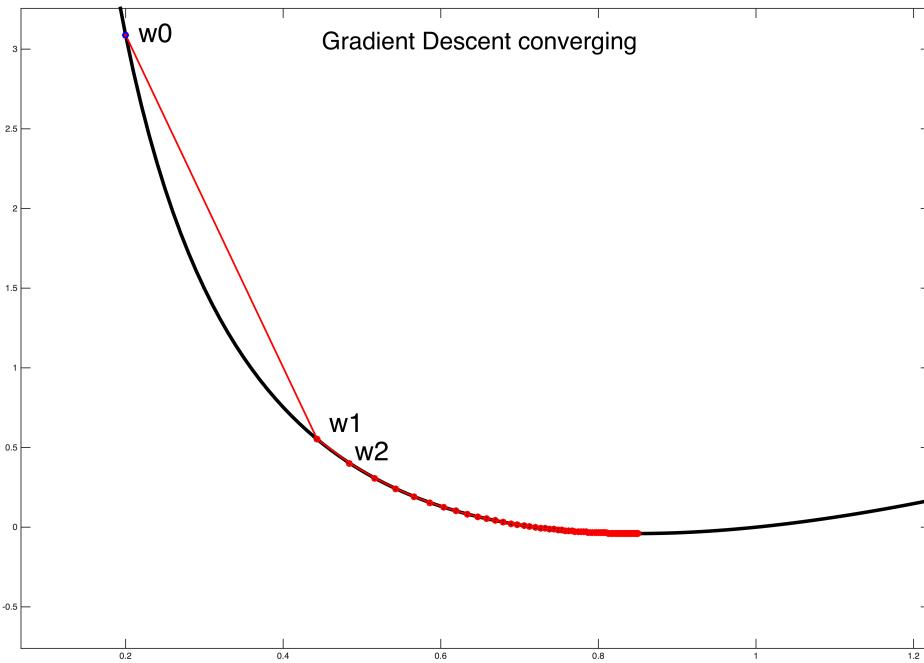
Локальные минимумы

- Результат зависит от начального приближения
- Мультистарт

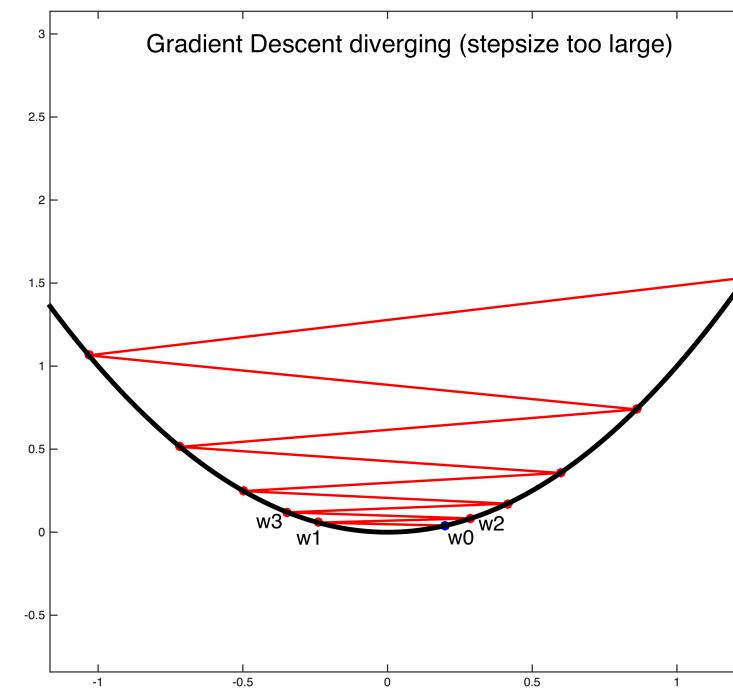


Размер шага

- Выбор размера шага η — искусство



Маленький шаг



Большой шаг

Размер шага

- Маленький шаг — больше шансов на сходимость, но требуется больше итераций
- Большой шаг — есть риск отсутствия сходимости

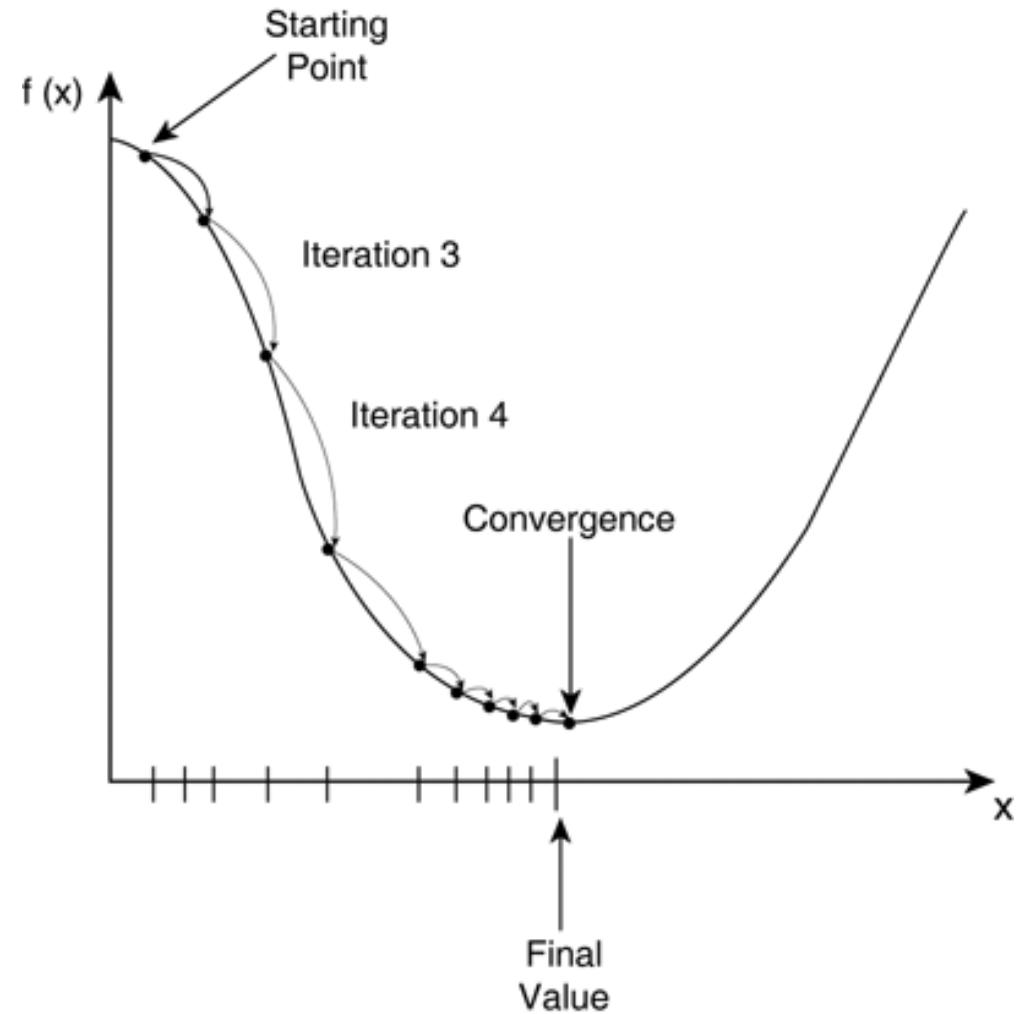
- Наискорейший градиентный спуск:

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

- 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 λ, 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

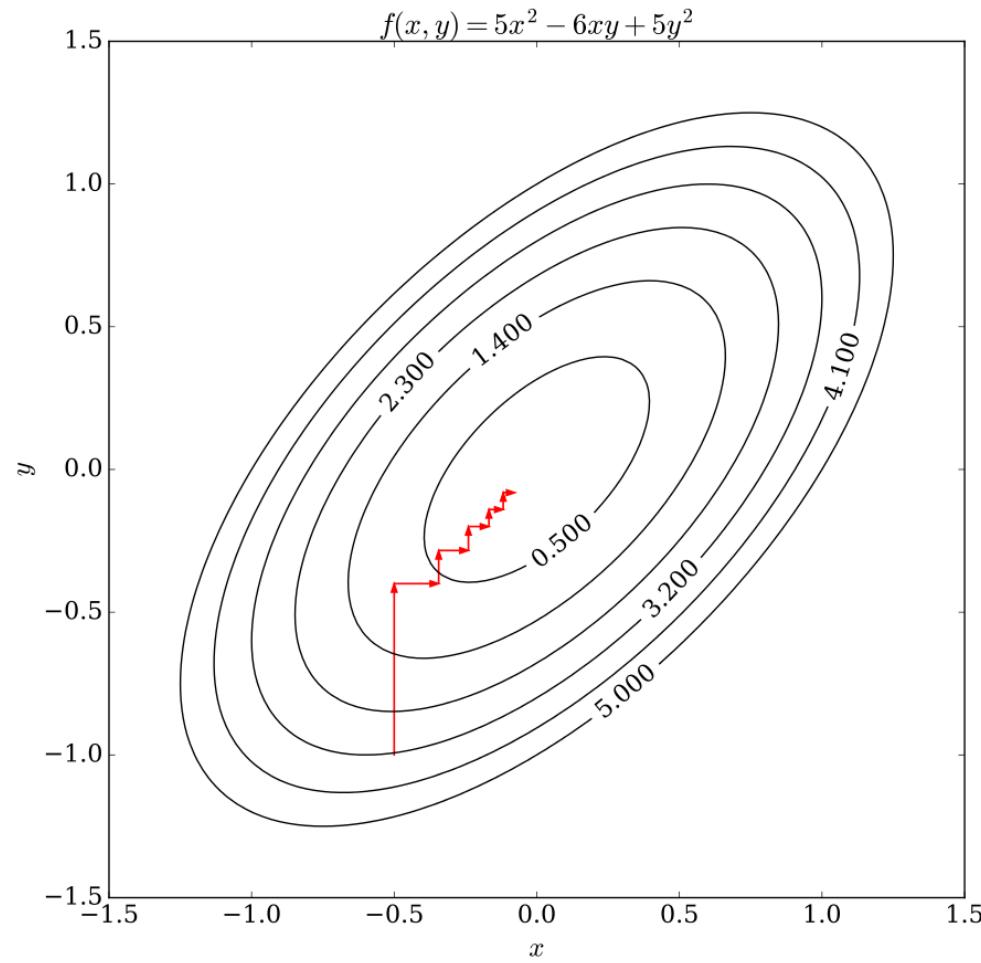
Другие методы оптимизации

- 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

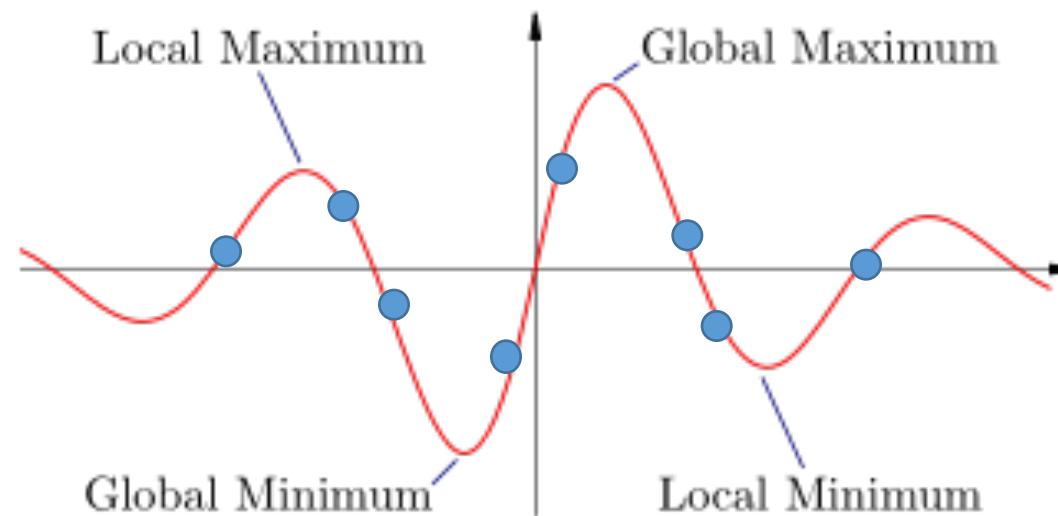
- По очереди меняем каждую координату
- Шаг по каждой координате — случайный, наискорейший, эвристический...
- Быстрые итерации, но может медленно сходиться
- Используется в методе опорных векторов (один из линейных)

Покоординатный спуск



Стохастическая оптимизация

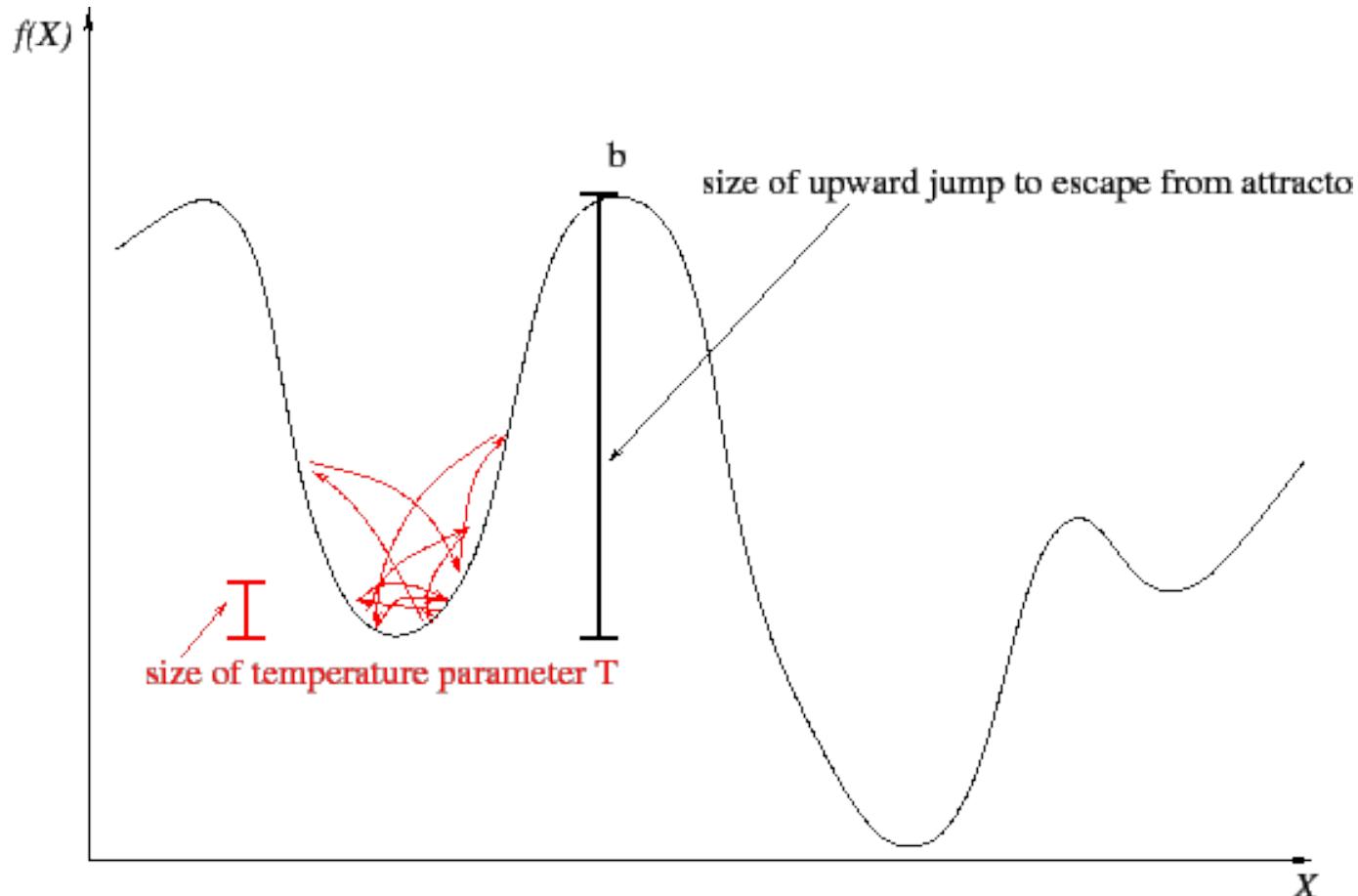
- Простейший алгоритм:
 - Генерируем N раз случайную точку
 - Выбираем ту, на которой значение функционала наименьшее
- Не самый лучший подход
- Нужно более направленной движение



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



Резюме

- Градиент позволяет определить точки минимума
- Выпуклые функции особенно удобны для оптимизации
- Градиентный спуск — движение в сторону скорейшего убывания
- Есть методы оптимизации, не требующие производных

На следующей лекции

- Случайные величины
- Распределения и их использование в машинном обучении
- Базовые статистики для работы с данными

