Лекция 5:
Методы оптимизации для задач машинного обучения

Задача оптимизации

Безусловная:

$$\min f(x)$$
$$x \in \mathbb{R}^n - \text{переменные}$$
$$f : \mathbb{R}^n \to \mathbb{R}^1 - \text{целевая функция}$$

$$\mathrm{x}^* \in \operatorname*{argmin}_X f(x) - \text{точка минимума},$$
$$f^* = f(x^*) - \text{значение минимума}$$

Условная:

$$\min f(x)$$
$$\boldsymbol{x} \in \boldsymbol{X} - \textbf{множество ограничений}$$
$$f : Y \to \mathbb{R}^1, X, Y \subseteq \mathbb{R}^n, X \subseteq Y$$

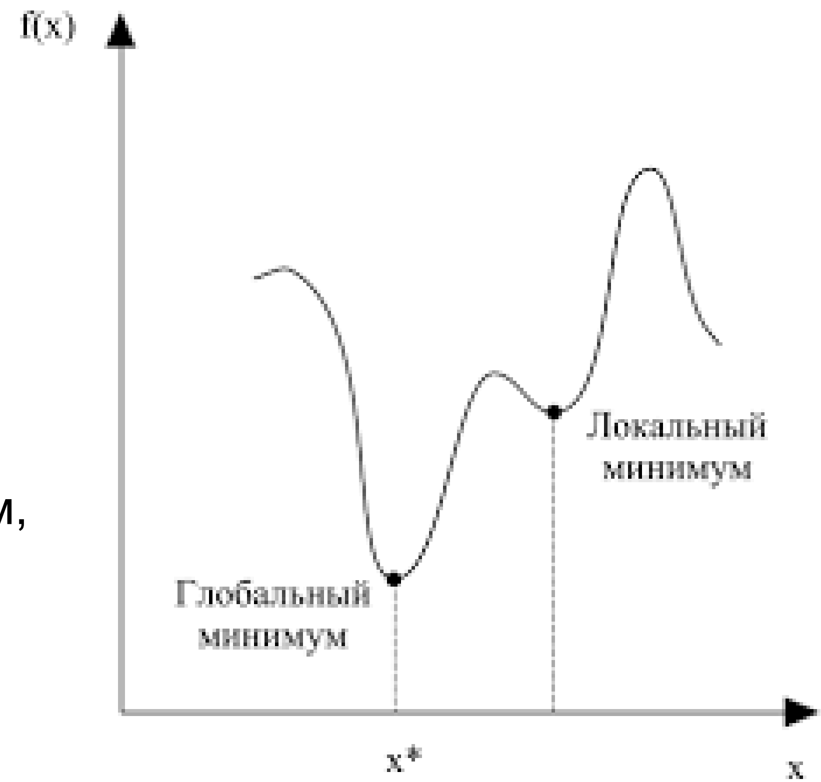
Определение минимума функции

Точка x^* называется точкой *локального* минимума функции $f(x)$, если существует $\varepsilon > 0$ такое, что $f(x) \geq f(x^*)$ для всех x: $\|x - x^*\| \leq \varepsilon$

Точка x^* называется точкой *глобального* минимума функции $f(x)$, если $f(x) \geq f(x^*)$ для всех $x \in \mathbb{R}^n$

Глобальное решение является локальным, обратное неверно

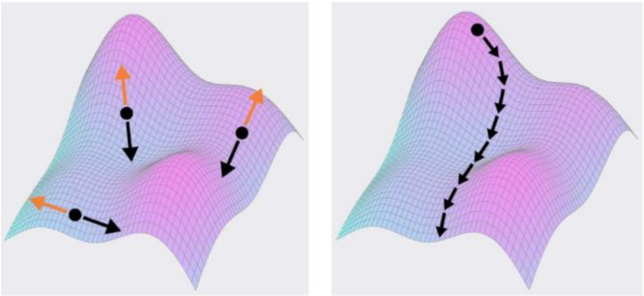
Может быть много локальных решений, глобальное решение единственно



Некоторые важные определения

Градиент $f: \mathbb{R}^n \to \mathbb{R}$ (вектор первых производных) – направление наибольшего роста функции в точке (можно доказать через разложение в ряд Тейлора и неравенство Коши-Буняковского):

$$f'(x) = \nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \dots, \frac{\partial f(x)}{\partial x_n}\right)$$



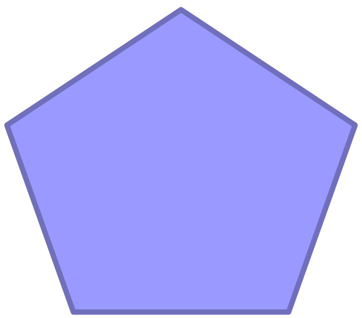
Матрица вторых производных, матрица Гессе, гессиан:

$$f''(x) = \nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{i,j=\overline{1,n}}$$

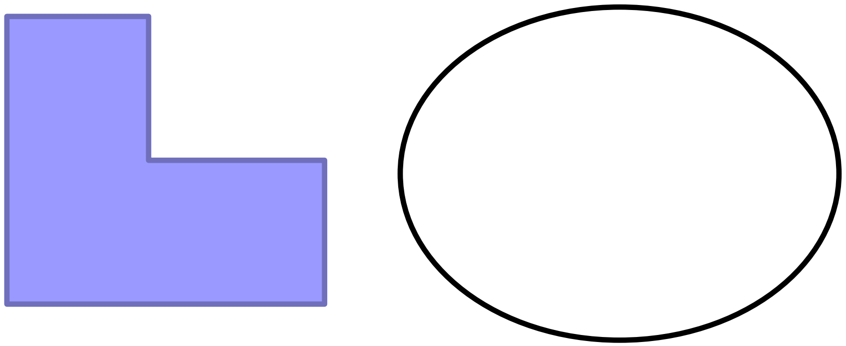
Выпуклые множества

Определение. Непустое множество $X \subseteq \mathbb{R}^n$ называется выпуклым, если для любых $x, y \in X, \lambda \in [0,1]$ выполняется $\lambda x + (1-\lambda)y \in X$

Выпуклое



Невыпуклые



Свойства выпуклых множеств:

- Пересечение выпуклых множеств выпукло
- Линейная комбинация выпуклых множеств выпукло

Выпуклые функции

Функция $f\colon \mathbb{R}^n \to \mathbb{R}$, определенная на выпуклом множестве $X \subset \mathbb{R}^n$, называется *выпуклой*, если для любых $x, y \in X, \lambda \in [0,1]$:

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$$

строго выпуклой

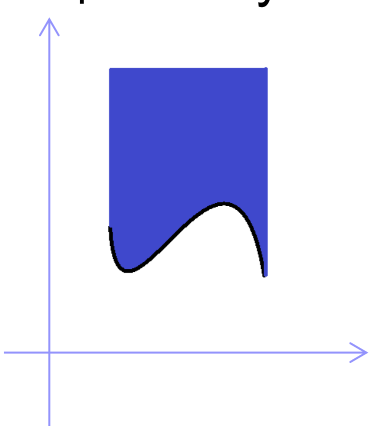
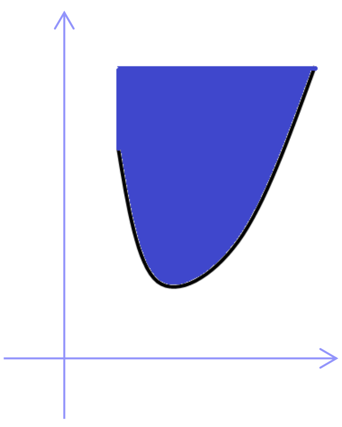
$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

сильно выпуклой

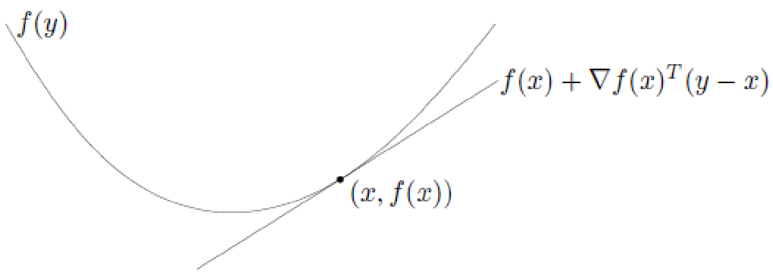
$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) - \boldsymbol{\alpha\lambda(1-\lambda)\|x-y\|^2}$$

Свойства:

- Линейная комбинация выпуклых функций выпукла
- Надграфики выпуклой функции является выпуклым множеством

Свойства выпуклых функций для задач оптимизации



Выпуклая функция достигает максимума и минимума в **единственной точке**

Неравенство Йенсена:

Пусть $f(x)$– выпуклая функция на выпуклом множестве X, тогда:

$$f\left(\sum_{i=1}^{m} \lambda_i x^i\right) \leq \sum_{i=1}^{m} \lambda_i f(x^i),$$

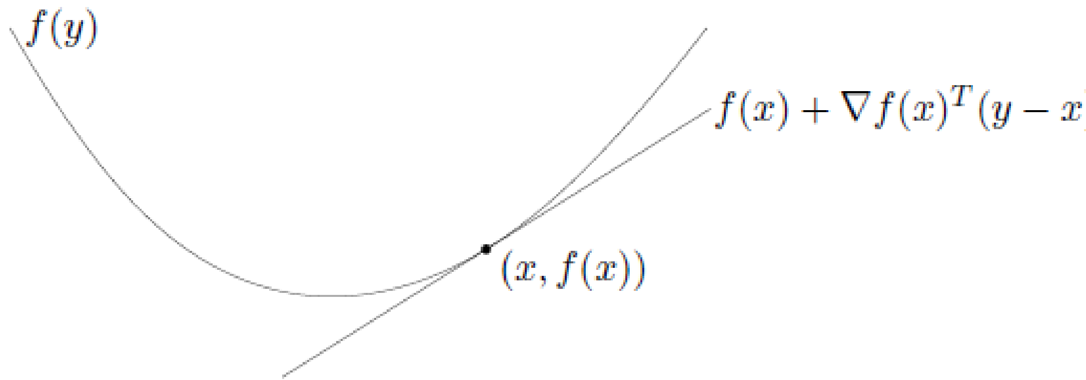
$$\forall m = 1,2\ldots, x^i \in X,$$

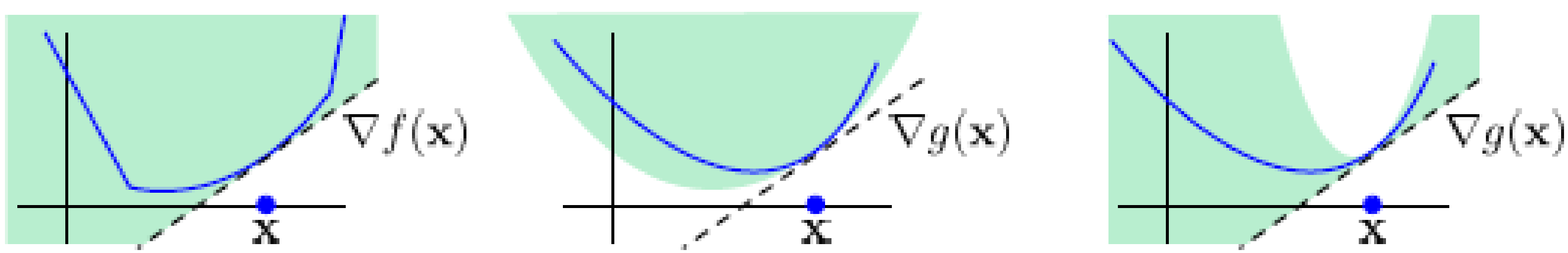
$$\lambda_i \geq 0, i = \overline{1,m}, \sum_{i=1}^{m} \lambda_i = 1$$

Условие выпуклости первого порядка:

- График функции лежит не ниже касательной гиперплоскости к нему в точке
  $$\forall x, y \in X, f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle$$
- Строго выпукла, если строгое неравенство
- Сильно выпукла, если есть «отступ»
  $\exists \lambda$:
  $$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle + \boldsymbol{\lambda} \|\boldsymbol{y}\|^2/2$$

Примеры выпуклости





— $f : \mathbb{R}^d \to \mathbb{R}$ — $g : \mathbb{R}^d \to \mathbb{R}$ — $g : \mathbb{R}^d \to \mathbb{R}$

выпуклая строго сильно

Необходимые и достаточные условия экстремума

Условия первого порядка:

- **Теорема**. (*необходимое условие*) Если x^* - точка минимума дифференцируемой в точке x^* функции $f(x)$. Тогда $\nabla f(x^*) = 0$.
- **Теорема** (*достаточное условие*). Пусть $f(x)$ – **выпуклая** функция, дифференцируемая в точке x^* , и $\nabla f(x^*) = 0$. Тогда x^* - точка глобального минимума функции $f(x)$

Условия второго порядка:

- **Теорема.** (*необходимое условие*) Пусть x^* точка минимума функции $f(x)$, дважды дифференцируема в точке x^*. Тогда гессиан $\nabla^2 f(x^*)$ неотрицательно определен:$\forall h \langle \nabla^2 f(x^*)h, h \rangle \geq 0$
- **Теорема.** (*достаточное условие*) Пусть $f(x)$ дважды дифференцируема в точке x^*, $\nabla f(x^*) = 0$ и $\nabla^2 f(x^*)$ неотрицательно определен. Тогда x^* - точка локального минимума.

Задача безусловной оптимизации

Чтобы найти $x^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$ необходимо решить $\nabla f(x) = 0$

Если $f(x)$ выпуклая $\Rightarrow$ единственный минимум

Если $f(x)$ невыпуклая:

- решения может не быть (функция вогнута, не имеет минимума)
- решений может быть много $\Rightarrow$ необходимо определить, является ли оно экстремумом, если да, то максимум или минимум.

Если нет аналитического решения $\Rightarrow$ **итерационная процедура**, определяющее решение приближенно: $|f(x) - f(x*)| \leq \varepsilon$

Поиск приближенного решения

Метод решения задачи оптимизации – построение приближения к решению исходной задачи (точке минимума *x**) на основе информации о характеристиках целевой функции.

- *Методы нулевого порядка* используют только значения функции
- *Методы первого порядка* – первые производные
- *Методы второго порядка* – вторые производные

$$y = f(\mathbf{x} + \Delta\mathbf{x}) \approx \boxed{f(\mathbf{x})} + \boxed{J(\mathbf{x})}\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^{\mathrm{T}}\boxed{H(\mathbf{x})}\Delta\mathbf{x}$$

Общий вид метода оптимизации

Многие методы оптимизации имеют вид

$$x^{k+1} = x^k + \eta_k d^k, d^k \in \mathbb{R}^n, \eta_k \in \mathbb{R}, k = 0,1 \dots$$

x^0 - начальное приближение

d^k – направление минимизации, определяется характеристиками минимизируемой функции и выбранной процедуры

η_k – длина шага (в МО называют скорость обучения)

- Если точное решение находится за конечное число шагов, метод называется *конечношаговым,* иначе – *бесконечношаговым*
- Бесконечношаговый сходится, если $x^k \to x^*, k \to \infty$
- Скорость сходимости (число шагов, не количество вычислений):
  - Линейная $\|x^{k+1} - x^*\| \le q\|x^k - x^*\|, q \in (0,1)$
  - Сверхлинейная $\|x^{k+1} - x^*\| \le q_k\|x^k - x^*\|, q_k \to 0$
  - Квадратичная $\|x^{k+1} - x^*\| \le q\|x^k - x^*\|^2$
  - и т.д.

Методы спуска

Определение. Вектор $d \in \mathbb{R}^n$ называется направлением убывания функции $f: \mathbb{R}^n \to \mathbb{R}$ в точке $x \in \mathbb{R}^n$, если для малого η:

$$f(x + \eta d) < f(x)$$

$D_f(x)$ – множество (конус) всех направлений убывания в точке (все вектора, чье скалярное произведение с антиградиентом положительно)

$$\forall d \in D_f(x), \langle \nabla f(x), d \rangle \leq 0$$

Методы спуска:

$$x^{k+1} = x^k + \eta_k d^k, d \in D_f\left(x^k\right), k = 0,1, \dots,$$

$$\eta_k: \{f(x^k)\} \text{ убывает}$$

Каждый метод – свой подход к выбору **шага** и выбору **направления** убывания

Критерии остановки: число шагов, малое изменение целевой функции, близость нормы градиента к 0

Градиентный метод

Направление спуска противоположно градиенту: $d^k = -\nabla f(x)$

- Выбираем начальное приближение $x^0 \in \mathbb{R}^n$
- Каждое следующее приближение определяется по правилу:

$$x^{k+1} = x^k - \eta_k \nabla f(x^k), k = 0,1,\dots$$

Правила выбора шага:

- априорное задание $\{\eta_k\}_{k=0}^{\infty}$ в виде константы или правил пересчета, например:

$$\eta_k = \eta, \eta_k = \frac{\eta}{1 + \beta k}$$

- полная релаксация - *метод наискорейшего спуска*

$$\eta_k = \arg\min_{\eta \geq 0} f(x^k - \eta \nabla f(x^k))$$

- правило Армихо, Вульфа и другие

Траектория для квадратичной функции

Правило Армихо (линейный поиск)

Фиксируем параметры: дробления θ (например, пополам), допуск $\varepsilon \in (0,1)$, начальный шаг $\eta_0 > 0$

Путем дробления $\eta \coloneqq \eta\theta$ добиваемся выполнения неравенства

$$f(x^k + \eta d^k) \le f\left(x^k\right) + \varepsilon\eta\left\langle \nabla f\left(x^k\right), d^k \right\rangle$$

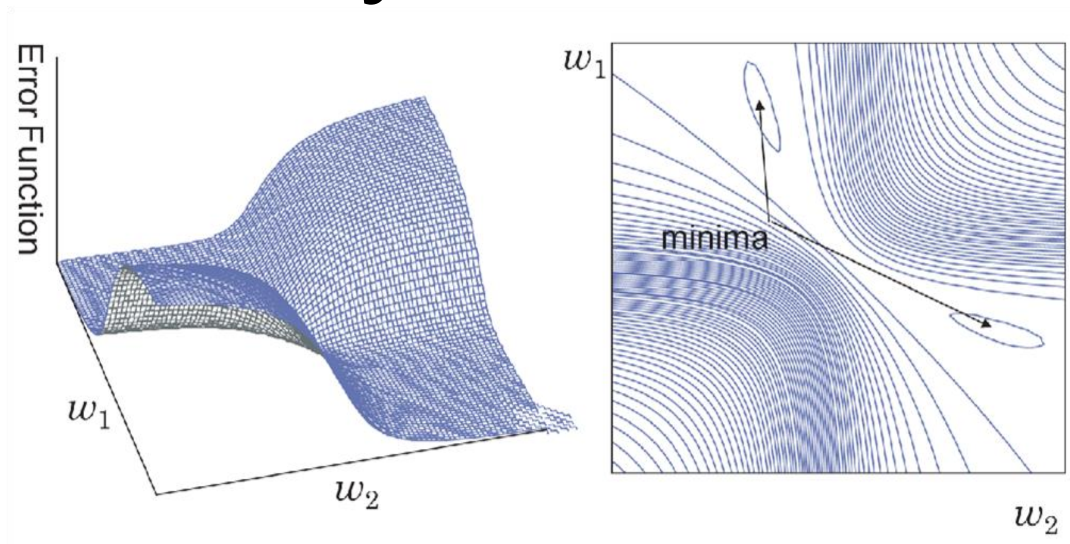
Как только неравенство выполнено, полагаем $\eta_k \coloneqq \eta$

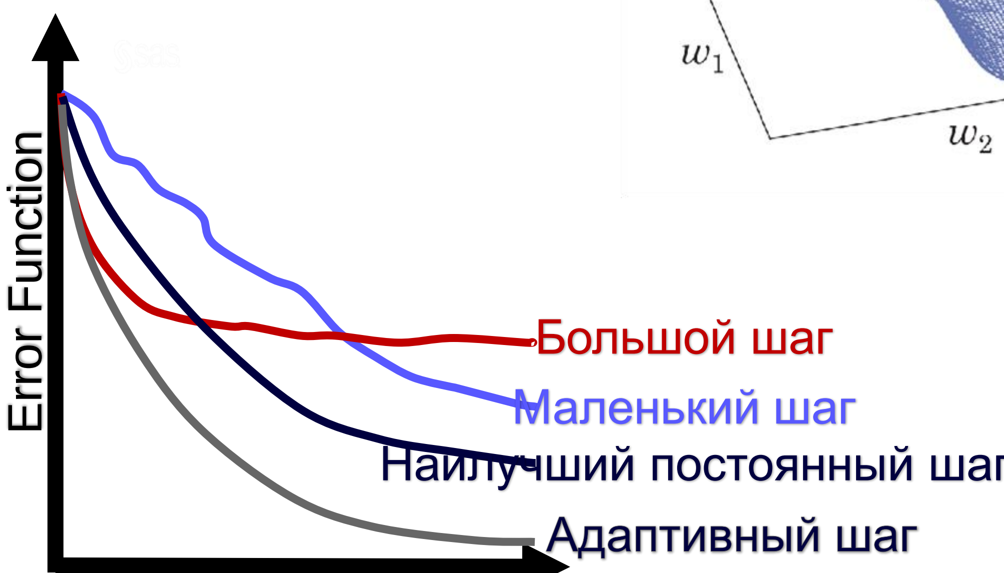
Полезные свойства, доказанные для случая выбора шага по правилу Армихо:

- Сходимость к стационарной точке (минимуму если выпуклая функция), если градиент удовлетворяет условию Липшица
$$\forall\ x, y\colon \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$$
- Линейная скорость сходимости, если функция дважды дифференцируема и выполнено ограничение на гессиан:
$$\forall\ x, h\colon \left\langle \nabla^2 f(x)h, h \right\rangle \le D\|h\|^2$$

Градиентный метод в машинном обучении

«Поверхность ошибки» и контурная проекция - невыпуклая





Зависимость от выбора шага для невыпуклой функции

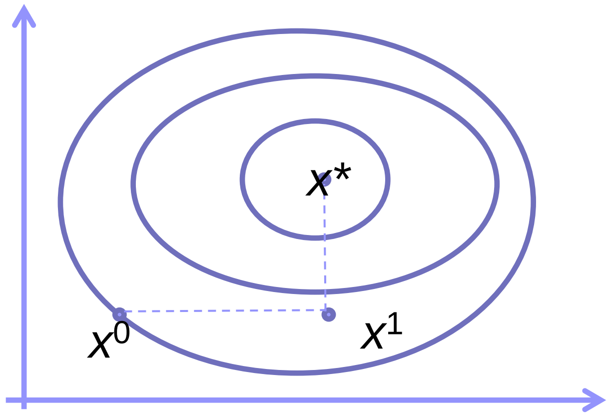
Покоординатный спуск

Суть метода:

- спуск, но по одной или нескольким координатам, их компоненты в d^k ненулевые, остальные равны 0
- Выбор шага обычно – наискорейший спуск
- Метод ведет себя разумно только для гладких функций

Выбор ненулевых направлений:

- Последовательный
- Блочный
- Случайный



Метод Ньютона

Пусть $f(x)$ – выпуклая дважды дифференцируемая функция

По определению дважды дифференцируемой функции (разложим в ряд Тейлора):

$$f(x) - f(x^k) = \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + o(\|x - x^k\|^2)$$

Минимизируем квадратичную часть:

$$\langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle \to \min$$

Она выпукла, так как ее гессиан $\nabla^2 f(x^k) \geq 0$

Необходимое и достаточное условие минимума:

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$$

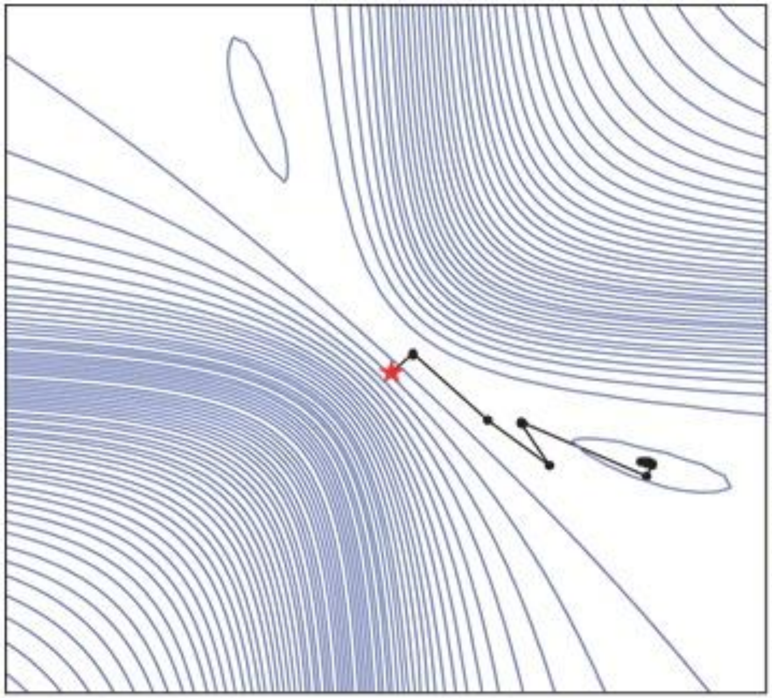
Отсюда получаем метод Ньютона:

$$x^{k+1} = x^k + h^k, h^k = -(\boldsymbol{\nabla^2 f(x^k)})^{-1} \nabla f(x^k)$$

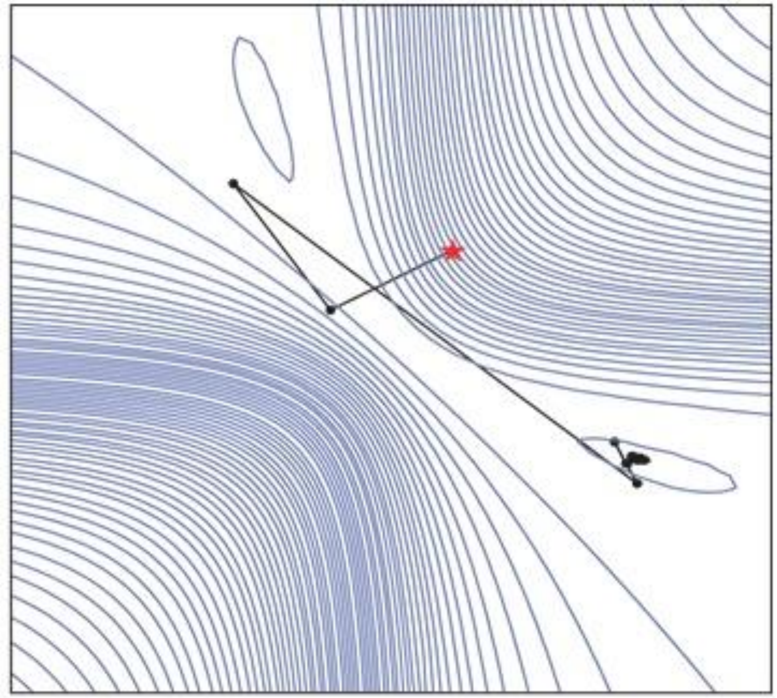
Особенности метода Ньютона

- Нет проблемы **выбора шага**
- **Сходимость:**
  - Если целевая функция дважды дифференцируема, сильно выпуклая, тогда получаемая в методе Ньютона последовательность приближений сходится к точке минимума с квадратичной скоростью.
  - Для невыпуклых функций сходимость гарантирована только при условии близости начальной точки к точке минимума.
- Необходимость выбора **начального приближения**:
  - в окрестности решения, условие близости трудно проверить
- **Вычислительная** трудоемкость, а иногда и нестабильность:
  - нужно вычислять на каждом шаге матрицу вторых производных и затем обратную к ней.
- Желание сохранить быструю скорость сходимости и устранить недостатки привели к разработке множества **модификаций** метода Ньютона и **квазиньютоновских** методов

Пример



38 итераций



57 итераций

Не нужно задавать длину шага, мало итераций, но много вычислений на каждой – нужно считать матрицу вторых производных и обращать ее

Простые модификации метода Ньютона

Метод Ньютона с регулировкой шага:

$$x^{k+1} = x^k + \eta_k h^k, \eta_k > 0, h^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

- Шаг выбирается либо из условия минимизации функции вдоль заданного направления (наискорейший спуск), либо по правилу дробления шага (линейный поиск).
- Скорость сходимости сверхлинейная или квадратичная
- Сократить вычисления можно, пересчитывая гессиан один раз в несколько шагов. Шаг пересчета подбирается эмпирически.

Быстрое обратное распространение ошибки:

- Приближаем целевую функцию «параболой», вычисляем диагональ Гессиана «приближенной» функции:

$$x^{k+1} = x^k - [diag(\widetilde{H})]^{-1} \nabla f(x^k), \widetilde{H} \approx \nabla^2 f\left(x^k\right)$$

Регуляризация:

$$x^{k+1} = x^k - [\nabla^2 f\left(x^k\right) + \gamma I]^{-1} \nabla f(x^k)$$

Общая схема квазиньютоновских методов

Суть подхода:

$$x^{k+1} = x^k + \eta_k d^k, d^k = -H_k \nabla f\left(x^k\right), k = 0,1, \ldots$$

$H_k \in \mathbb{R}(n,n)$ - некоторая симметричная положительно определенная матрица,

тогда d^k - направление спуска, поскольку

$$\left\langle \nabla f\ (x^k), d^k \right\rangle = -\left\langle H_k \nabla f\left(x^k\right), \nabla f\left(x^k\right) \right\rangle < 0$$

Если $H_k \equiv I$– градиентный метод, если $H_k = (\nabla^2 f(x^k))^{-1}$ и $\eta_k = 1$ – метод Ньютона

Квазиньютоновские методы аппроксимируют:

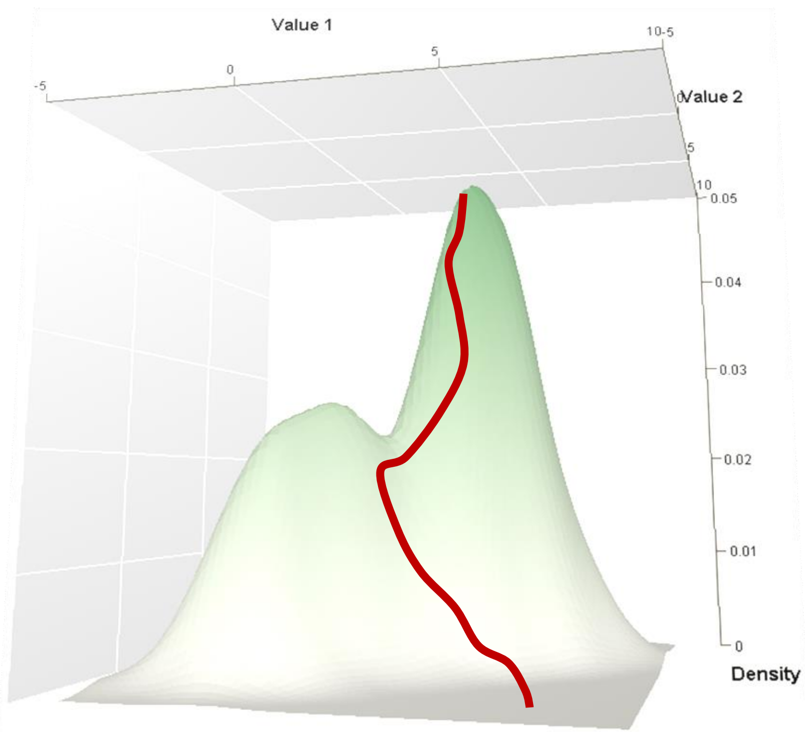
$$H_k \approx \left[\nabla^2 f\left(x^k\right)\right]^{-1}$$

Например, метод Бройдена-Флетчера-Голдфарба-Шэнно (БФГШ) (Broyden-Fletcher-Goldfarb-Shanno):

$$H_{k+1} = H_k + \frac{(r^k - H_k s^k)(r^k)^T + r^k (r^k - H s^k)^T}{\left\langle r^k, s^k \right\rangle} - \frac{\left\langle r^k - H_k s^k, s^k \right\rangle r^k (r^k)^T}{\left\langle r^k, s^k \right\rangle^2}$$

$$r^k = x^{k+1} - x^k, s^k = \nabla f(x^{k+1}) - \nabla f(x^k)$$

Классические методы оптимизации для задачи машинного обучения



- Целевая функция – эмпирический риск, усредненная сумма значений функций риска для всех наблюдений
- Используют все наблюдения выборки для расчета эмпирического риска, его градиента или приближения гессиана
- Результат – гладкая траектория оптимизации
- Но долго по времени и много по памяти (вся выборка)

Градиентный спуск для задачи машинного обучения

- Дана «размеченная» выборка :

$$Z = \{(x_i, y_i)\}_{i=1}^{l} \in X \times Y$$

- Фиксируем параметрическое семейство алгоритмов (моделей)

$$A = \{g(x, w) \mid w \in \mathrm{W}\}, g: X \times \mathrm{W} \to Y$$

- Минимизируем:

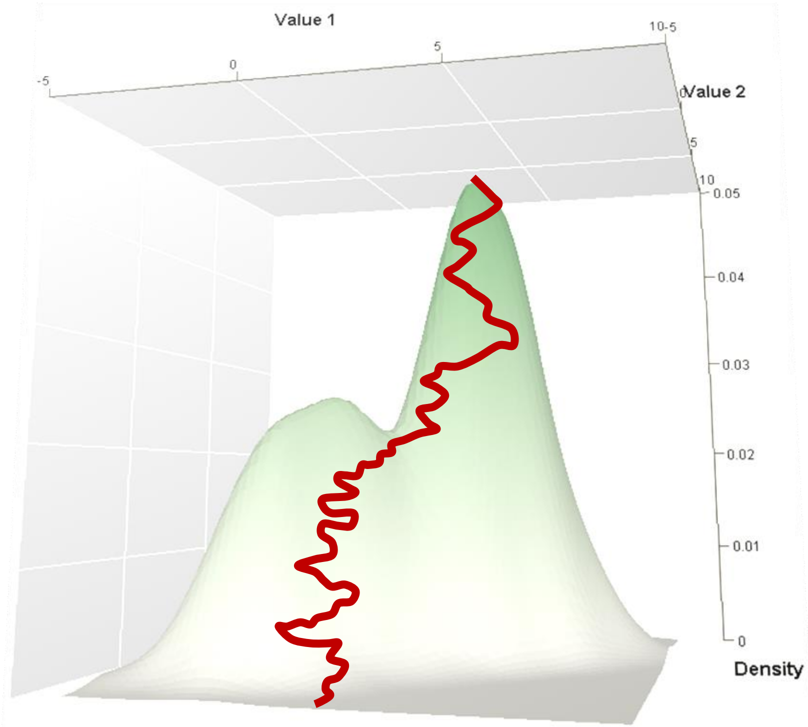
$$Q(w, Z) = \frac{1}{l} \sum_{Z} L\big(y_i, g(x_i, w)\big) \to min$$

- Градиентный спуск:
  - Выбираем начальное приближение $w^{(0)}$
  - В цикле до достижения условий сходимости считаем

$$w^{(t+1)} = w^{(t)} - \eta \nabla Q\big(w^{(t)}\big) = w^{(t)} - \eta \frac{1}{l} \sum_{\boldsymbol{Z}} \boldsymbol{\nabla L}\left(\boldsymbol{y_i}, \boldsymbol{g}\big(\boldsymbol{x_i}, \boldsymbol{w}^{(t)}\big)\right)$$

- А что если считать ∇Q не по всей выборке?

Стохастические и пакетные методы обучения



- Аппроксимируют вычисление градиента эмпирического риска (или приближение гессиана) только по части выборки
- Стохастические методы используют только одно наблюдение
- Пакетные – часть наблюдений
- Результат – «хаотическая» траектория (чем больше пакет тем меньше «хаоса»)
- Зато быстро считать и не нужно все брать из памяти – MPP!!!

Стохастический градиентный спуск для задачи машинного обучения

- Параметры: η – скорость обучения, λ – скорость забывания
- Инициализация $w^{(0)}$ и $Q^{(0)}$-средний по случайному подмножеству Z
- Цикл стохастического градиентного спуска
  - Выбираем x_i из Z
  - Вычисляем функцию потерь на нем $\varepsilon_i = L\left(y_i, g\left(x_i, w^{(t)}\right)\right)$
  - Делаем шаг градиента $w^{(t+1)} = w^{(t)} - \eta \nabla L\left(y_i, g\left(x_i, w^{(t)}\right)\right)$
  - Оцениваем $Q^{(t)} = \lambda \varepsilon_i + (1-\lambda) Q^{(t-1)}$
- Приближение $Q^{(t)}$:
  - Среднее арифметическое $Q^{(t)} = \frac{1}{t}\varepsilon_t + \frac{1}{t}\varepsilon_{t-1}$ … или $Q^{(t)} = \frac{1}{t}\varepsilon_t + \left(1 - \frac{1}{t}\right) Q^{(t-1)}$
  - Экспоненциальное сглаживание $Q^{(t)} = \lambda \varepsilon_t + (1-\lambda) Q^{(t-1)}$, λ порядка $\frac{1}{t}$
- Выбор объектов для обучения:
  - Случайный или на основе ошибки – чем хуже тем лучше

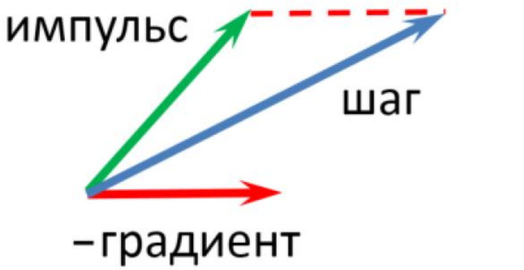
Важные модификации градиентного метода

- Учет инерции («метод моментов», или импульса) – «сгладить» траекторию за счет предыдущих направлений (импульса), α задает «важность» старого направления

$$w^{k+1} = w^k + v^k, v^k = -(1-\alpha)\eta_k \nabla Q\left(w^k\right) + \boldsymbol{\alpha v^{k-1}}$$

- Метод Нестерова – «сглаживать» после применения старого шага:

$$v^k = -(1-\alpha)\eta_k \nabla Q\left(w^k + \boldsymbol{\alpha v^{k-1}}\right)$$



- Регуляризация (штраф за сложность):

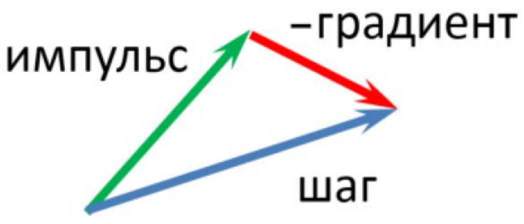
$$\min Q\,(w) + \boldsymbol{\gamma R}(\boldsymbol{w})$$

$R(.)$ – гладкая сильно выпуклая функция,

например, регуляризация L_p

γ-константа регуляризации

$$v^k = -\eta_k[\nabla Q\left(w^k\right) + \gamma \nabla R\left(w^k\right)]$$



Регуляризация

- В оптимизации – «улучшение» целевой функции

$$\min Q\,(w) + \gamma R(w)$$

$R(.)$ – гладкая сильно выпуклая функция, γ-константа регуляризации

есть методы меняющие γ или $R(.)$ «на ходу»

- В машинном обучении – контроль сложности модели для борьбы с переобучением.
  - Обычно γ и $R(.)$ фиксированы и $R(w) = L_p(w)$, тогда градиенты легко модифицируются, но необходимо «нормализовывать» параметры
  - LASSO регуляризация $L_1(w) = \sum|w|$, градиент $\nabla Q(w) + \gamma sign(w)$
  - Ridge («квадратичная» или «гребневая») регуляризация»

$$L_2(w) = \sum w^2,\ \text{градиент } \nabla Q(w) + 2\gamma w$$

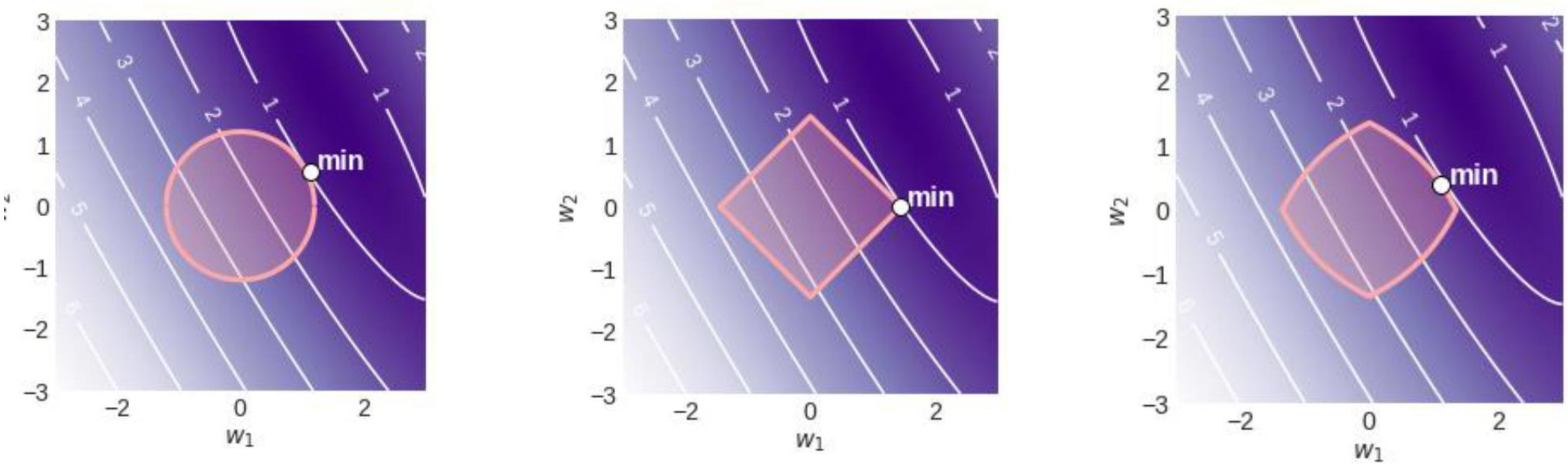
  - Elastic Net= LASSO+Ridge

$$L_1(w) = \gamma_1 \sum|w| + \gamma_2 \sum w^2,\ \text{градиент } \nabla Q(w) + 2\gamma_2 w + \gamma_1 sign(w)$$

Регуляризация для выпуклых целевых функций

- Выпуклые целевые функции – достаточно часто встречаются в задачах машинного обучения (линейные регрессии с регуляризацией, SVM, обобщенные линейные модели и др.)
- Для выпуклых целевых функций, можно показать, что:

$$\min Q(w) + \gamma L_p(w) \Leftrightarrow \begin{cases} \min Q(w) \\ L_p \le C \end{cases}$$



- Чем меньше p тем более модель настроена на «отбор» признаков

Другие популярные модификации SGD

- RProp – устойчивый к угасанию градиента (в многослойных нейросетях), учитывает знак градиента, но не учитывает его величину (шаг постоянный)
- Adam (Adaptive Moment Estimation) – экспоненциальное сглаживание градиентов
- AdaGrad и RMSProp – индивидуальные скорости обучения для каждого параметра «по ситуации»
- …