Лекция 3
Алгоритмы безусловной нелинейной оптимизации. Методы первого и второго порядка

Анализ и разработка алгоритмов

УНИВЕРСИТЕТ ИТМО

Содержание

1. Термины
2. Градиентный спуск
3. (Нелинейный) метод сопряженных градиентов
4. Метод Ньютона
5. Алгоритм Левенберга-Марквардта

Термины

Проблема

$f : \mathbb{R}^n \to \mathbb{R}$ – выпуклая; $f = f(\mathbf{x})$, где $\mathbf{x} = (x_1, \ldots, x_n)^\mathsf{T}$ – вектор-столбец
Решить проблему оптимизации $f(\mathbf{x}) \to \min_{\mathbf{x} \in Q}$ означает найти $\mathbf{x}^* \in Q$, где Q – область допустимых значений, такое, что f достигает минимального значения в $\mathbf{x}^*$. Обозначение: $\mathbf{x}^* = \arg\min_{\mathbf{x} \in Q} f(\mathbf{x})$.

Замечание. Погрешность приближения задана через $\varepsilon > 0$. В приведенных ниже итерационных алгоритмах остановка происходит, если $\|\mathbf{a}_n - \mathbf{a}_{n-1}\| < \varepsilon$; при этом полагаем, что $\mathbf{x}^* \approx \frac{1}{2}(\mathbf{a}_n + \mathbf{a}_{n-1})$ с погрешностью ε.

Вспомним:

- Производные первого и второго порядка функции одного переменного
- Градиент
- Гессиан
- Ряд Тейлора

Найдите производные первого и второго порядка для функций

$$x,\ x^3,\ \sin x,\ \ln x,\ |x|.$$

Градиент

Градиент – это обобщение первой производной на случай функции многих переменных

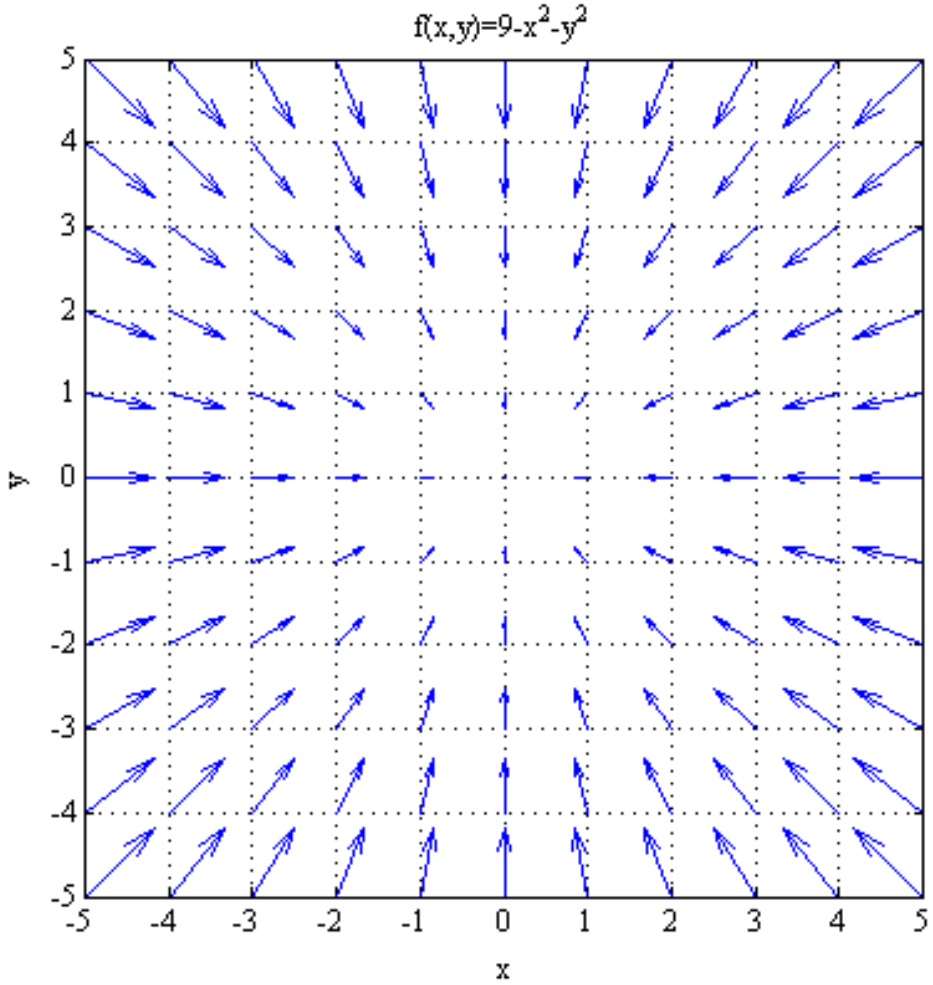
Градиент дифференцируемой функции $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ в точке $\mathbf{a}$ – это вектор-**столбец** (-строка), элементами которого являются частные производные функции f в точке $\mathbf{a}$:

$$\nabla_{\mathbf{a}} f = \left(\frac{\partial f}{\partial x_i}\bigg|_{\mathbf{a}} \right)_{i=1}^{n}.$$

Пример: $f(\mathbf{x}) = 2x_1 + 3x_2^2$, $\nabla f = (2 \quad 6x_2)^{\mathsf{T}}$, $\nabla_{\mathbf{a}} f = (2 \quad 6)^{\mathsf{T}}$ при $\mathbf{a} = (0, 1)$.

Если в точке $\mathbf{a}$ градиент функции не является нулевым вектором, то он указывает направление **наибольшего возрастания** этой функции в точке $\mathbf{a}$.

Градиент



Гессиан и ряд Тейлора

Матрица Гессе или **Гессиан** – это квадратная матрица, элементами которой являются вторые частные производные, которая описывает локальную кривизну функции и является обобщением второй производной на функции нескольких переменных.

Если все вторые частные производные $f = f(\mathbf{x})$ существуют и непрерывны, то Гессиан $\mathbf{H}_{\mathbf{a}} f$ функции f в точке $\mathbf{a}$ – это матрица размера $n \times n$ с элементами

$$\mathbf{H}_{i,j} = \left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right|_{\mathbf{a}}, \qquad i, j = 1, \ldots, n.$$

Ряд (разложение) Тейлора (бесконечно) дифференцируемой функции $f : \mathbb{R} \to \mathbb{R}$ в точке a имеет вид

$$T_f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots.$$

Обобщение на функцию многих переменных $f : \mathbb{R}^n \to \mathbb{R}$ в точке $\mathbf{a}$ имеет вид

$$T_f(\mathbf{x}) = f(\mathbf{a}) + (\mathbf{x}-\mathbf{a})^{\mathsf{T}} \nabla_{\mathbf{a}} f + \frac{1}{2!}(\mathbf{x}-\mathbf{a})^{\mathsf{T}} \mathbf{H}_{\mathbf{a}} f \, (\mathbf{x}-\mathbf{a}) + \ldots.$$

Напомним, что $\mathbf{x}$, $\mathbf{a}$ и $\nabla_{\mathbf{a}} f$ являются векторами-столбцами по определению.

Градиентный спуск (Наискорейший спуск)

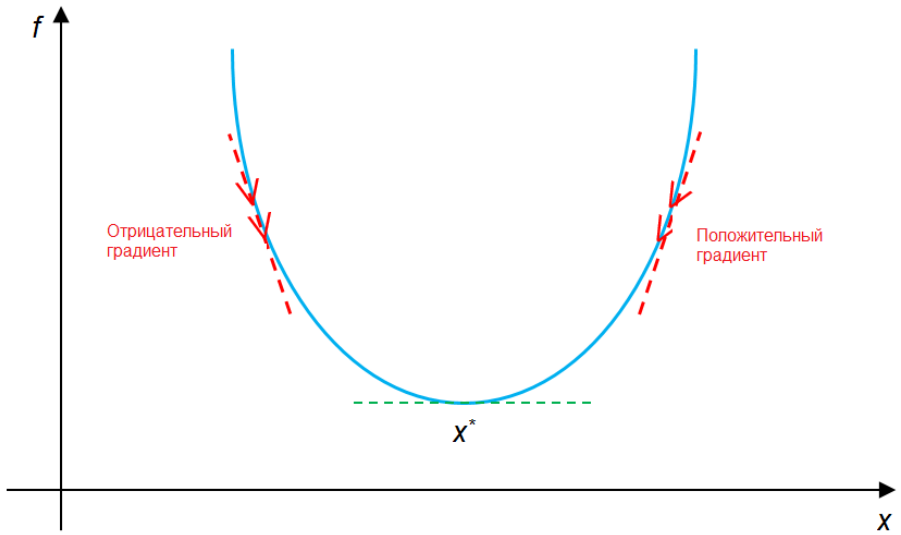
Градиентный спуск основан на том, что если $f(\mathbf{x})$ определена и дифференцируема в точке $\mathbf{a}$, то $f(\mathbf{x})$ быстрее всего убывает в окрестности точки $\mathbf{a}$ в направлении $-\nabla f(\mathbf{a})$. Получается следующая формула:

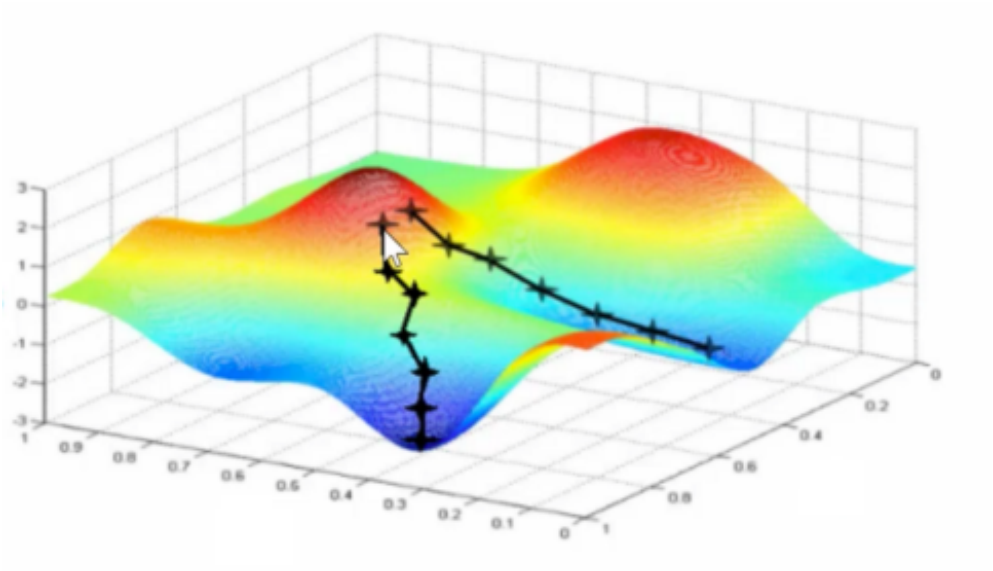
$$\mathbf{a}_{n+1} = \mathbf{a}_n - \gamma \nabla f(\mathbf{a}_n)$$

для $\gamma \in \mathbb{R}_+$ достаточно малой для того, чтобы $f(\mathbf{a}_n) \geq f(\mathbf{a}_{n+1})$. Используя это и начальное приближение $\mathbf{a}_0$ для локального минимума f, строим последовательность $\{\mathbf{a}_n\}$ такую, что

$$\mathbf{a}_{n+1} = \mathbf{a}_n - \gamma_n \nabla F(\mathbf{a}_n), \ n \geq 0,$$

где значение γ_n может быть не фиксирован и может меняться на каждой итерации для достижения сходимости (существует много способов выбора).


f
Отрицательный
градиент
Положительный
градиент
x*
x



(Нелинейный) метод сопряженных градиентов

Для заданной функции $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$, и начального приближения $\mathbf{a}_0$, начинаем как в методе градиентного спуска:

$$\Delta\mathbf{a}_0 = -\nabla_{\mathbf{a_0}} f.$$

Находим шаг $\alpha_0 := \arg\min_\alpha f(\mathbf{a}_0 + \alpha\Delta\mathbf{a}_0)$ и следующую точку $\mathbf{a}_1 = \mathbf{a}_0 + \alpha_0\Delta\mathbf{a}_0$. После этой итерации следующие шаги образуют итерацию для движения в направлении сопряженного градиента s_n, где $s_0 = \Delta\mathbf{a}_0$:

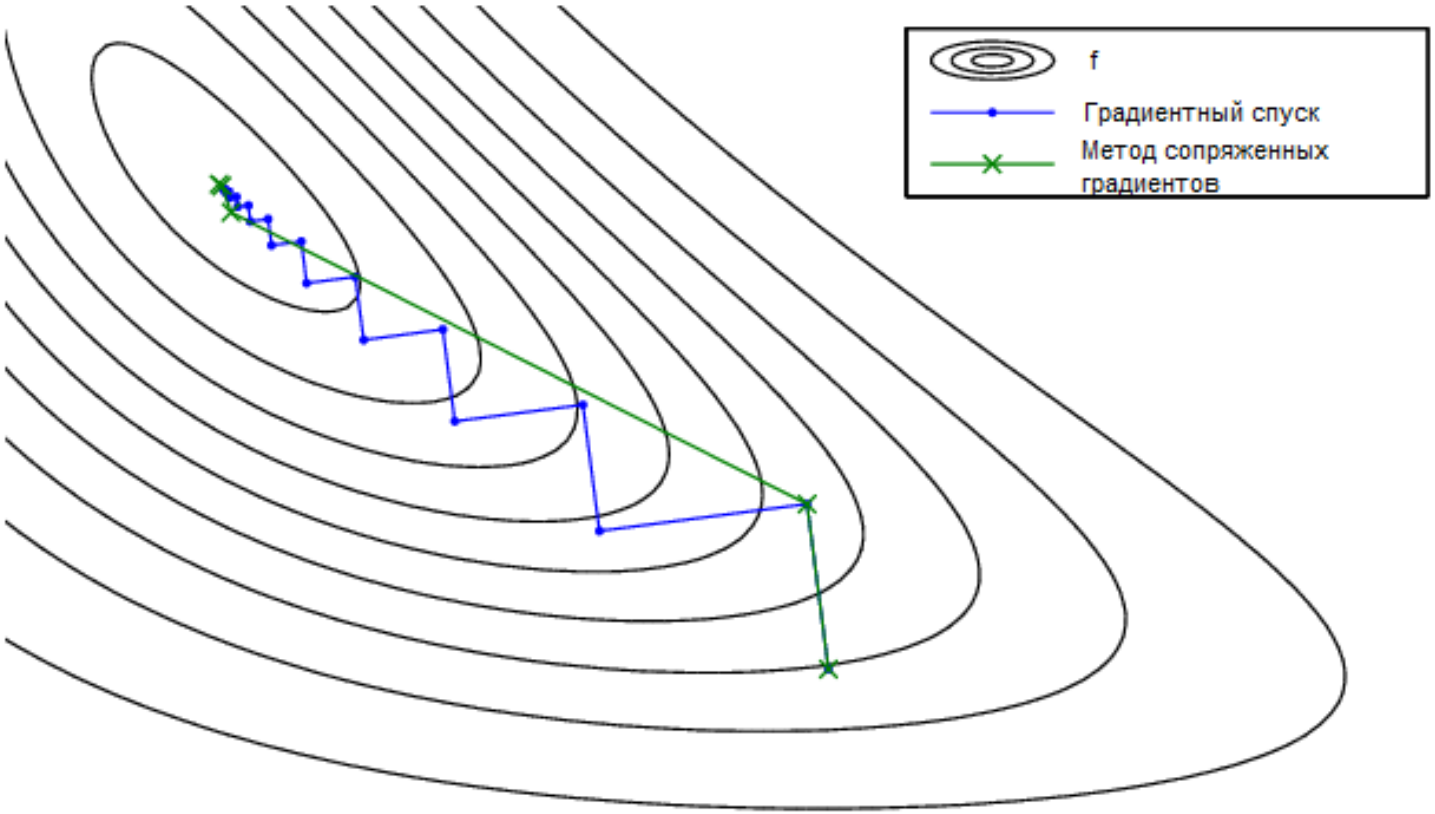
- Вычисляем направление антиградиента $\Delta\mathbf{a}_n = -\nabla_{\mathbf{a}_n} f$.
- Вычисляем β_n по опеределенным формулам (см. ниже).
- Обновляем направление движения $s_n = \Delta\mathbf{a}_n + \beta_n s_{n-1}$.
- Находим $\alpha_n = \arg\min_\alpha f(\mathbf{a}_n + \alpha s_n)$.
- Обновляем точку $\mathbf{a}_{n+1} = \mathbf{a}_n + \alpha_n s_n$.

Выбор β_n согласно Флетчеру-Ривсу:

$$\beta_n^{FR} = \frac{\Delta\mathbf{a}_n^T \Delta\mathbf{a}_n}{\Delta\mathbf{a}_{n-1}^T \Delta\mathbf{a}_{n-1}}.$$

Выбор β_n согласно Полаку-Рибьере:

$$\beta_n^{PR} = \frac{\Delta\mathbf{a}_n^T (\Delta\mathbf{a}_n - \Delta\mathbf{a}_{n-1})}{\Delta\mathbf{a}_{n-1}^T \Delta\mathbf{a}_{n-1}}.$$


f
Градиентный спуск
Метод сопряженных
градиентов

Метод Ньютона. Случай одной переменной

Пусть $f : \mathbb{R} \to \mathbb{R}$ – выпуклая и дважды дифференцируемая функция. Найдем нули функции f' путем построения последовательности a_n из начального приближения a_0 такую, что $a_n \to x^*$ при $n \to \infty$, где $f'(x^*) = 0$, т.е. x^* является **стациональной** точкой f.

Из разложения Тейлора для f в окрестности a_n (считаем, что $x^* \approx a_n + \Delta a$),

$$f(a_n + \Delta a) \approx T_f(\Delta a) := f(a_n) + f'(a_n)\Delta a + \frac{1}{2} f''(a_n)(\Delta a)^2.$$

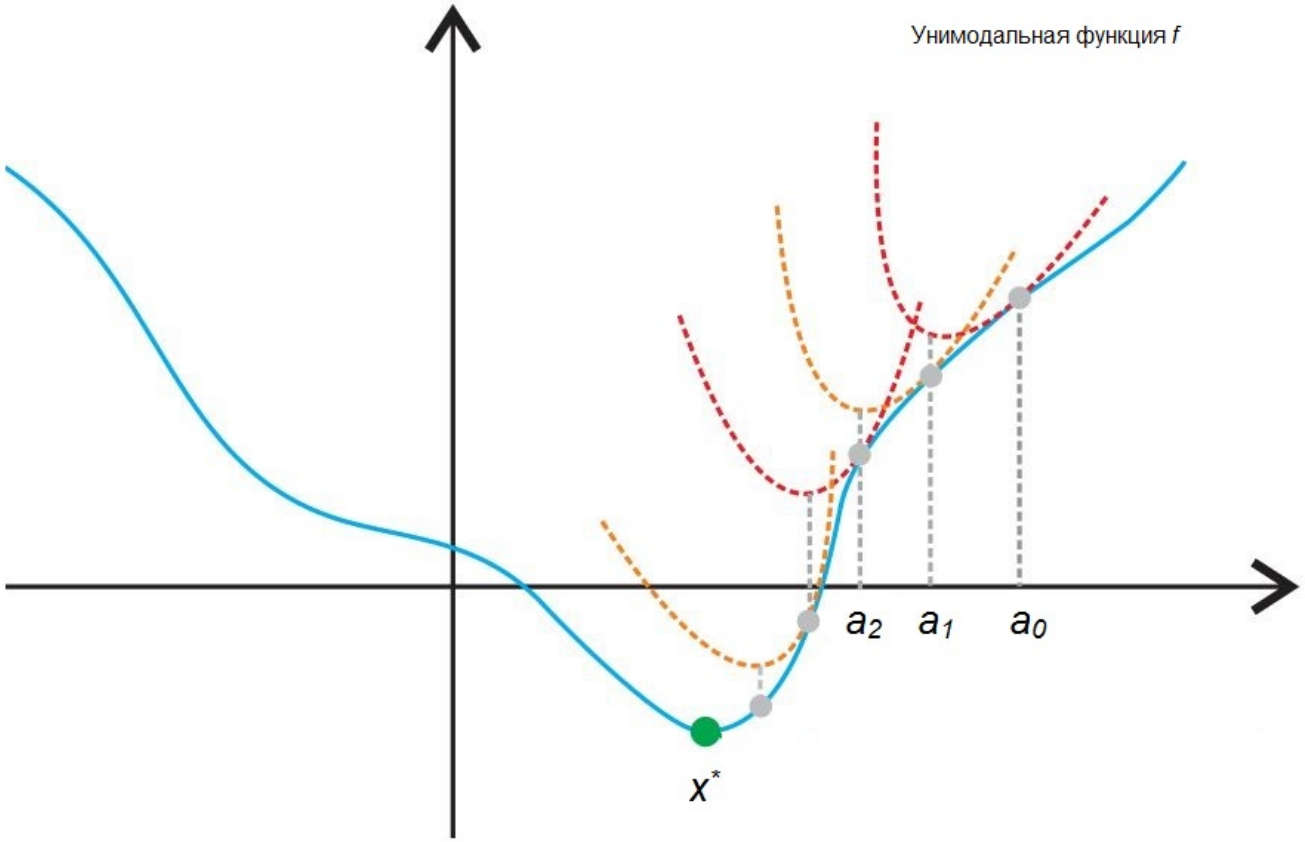
Используем эту квадратичную функцию как аппроксимацию f в окрестности a_n и найдем ее точки минимума (учтем, что $f''(x) \geqslant 0$ (*почему?*)):

$$0 = \frac{dT_f(\Delta a)}{d\Delta a} = f'(a_n) + f''(a_n)\Delta a \quad \Rightarrow \quad \Delta a = -\frac{f'(a_n)}{f''(a_n)}.$$

Изменением a_n на Δa получаем точку ближе к x^*:

$$a_{n+1} = a_n + \Delta a = a_n - \frac{f'(a_n)}{f''(a_n)}.$$

Доказано, что для выбранного класса функций f имеем $a_n \to x^*$ при $n \to \infty$.



Вопрос: в чем проблема с этой иллюстрацией из интернета?

Метод Ньютона. Случай функций многих переменных

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ является выпуклой и матрица Гессе $H_{\mathbf{x}} f$ обратима при всех $\mathbf{x} \in \mathbb{R}^n$. Одномерная схема может быть обобщена на случай функций многих переменных заменой производной на градиент, ∇f, и второй производной на матрицу Гессе, $\mathbf{H} f$:

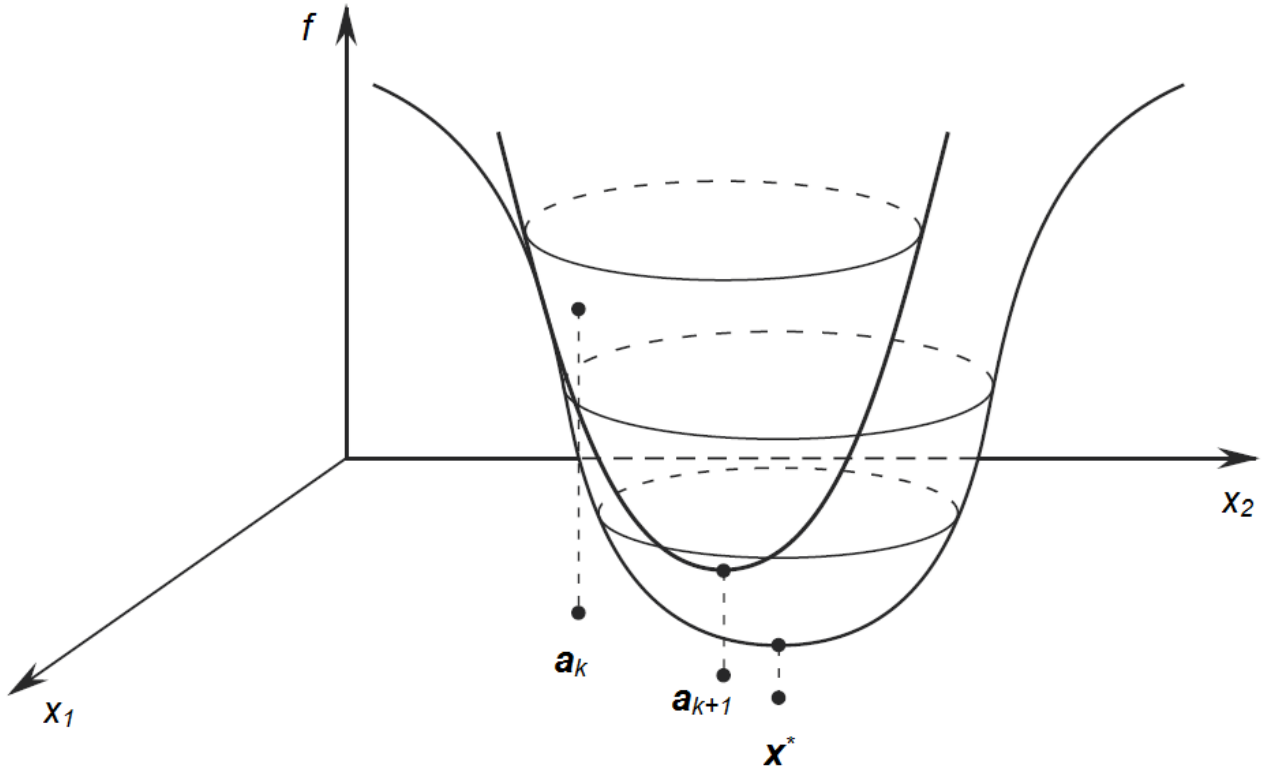
$$\mathbf{a}_{n+1} = \mathbf{a}_n - [\mathbf{H}_{\mathbf{a}_n} f]^{-1} \nabla_{\mathbf{a}_n} f, \ n \geq 0.$$

Метод Ньютона может быть модифицирован введением переменного шага $\gamma_n \in (0, 1)$:

$$\mathbf{a}_{n+1} = \mathbf{a}_n - \gamma_n [\mathbf{H}_{\mathbf{a}_n} f]^{-1} \nabla_{\mathbf{a}_n} f, \ n \geq 0,$$

для достижения сходимости.

Замечание. Вычилять гессиан зачастую затруднительно. Методы, называемые **квази-ньютоновскими**, предлагают использовать аппрокимацию гессиана для упрощения метода
$\to$ Мы рассмотрим некоторые из этих методов в рамках курсовых проектов.



Демонстрация для методов Ньютона и градиентного спуска

Алгоритм Левенберга-Марквардта (LMA)

Основное приложение LMA – решение задачи приближения функции **методом наименьших квадратов**: для данного набора данных $(x_i, y_i)_{i=1}^m$ найти вектор-столбец параметров β в модели $f(x, \beta)$ так, что сумма квадратов отклонений $S(\beta)$ была минимальной:

$$\arg\min_{\beta} S(\beta) \equiv \arg\min_{\beta} \sum_{i=1}^{m} \left[y_i - f\left(x_i, \beta\right)\right]^2 .$$

Начнем с начального приближения β. На каждой итерации вектор *beta* заменяется новым вектором $\beta + \Delta\beta$. Для определения $\Delta\beta$ функция $f\left(x_i, \beta + \Delta\beta\right)$ аппроксируется ее линейной частью:

$$f(x_i, \beta + \Delta\beta) \approx f(x_i, \beta) + J_i \Delta\beta, \qquad \mathbf{J}_i = \left(\nabla f(x_i, \beta)\right)^T .$$

Сумма $S(\beta)$ достигает минимального значения в точке нулевого градиента относительно β. Указанная выше линейная аппроксимация функции $f(x_i, \beta + \Delta\beta)$ дает

$$S(\beta + \Delta\beta) \approx \sum_{i=1}^{m} \left[y_i - f\left(x_i, \beta\right) - \mathbf{J}_i \Delta\beta\right]^2 ,$$

или в векторной форме

$$S(\beta + \Delta\beta) \approx [\mathbf{y} - \mathbf{f}(\beta)]^{\mathrm{T}} [\mathbf{y} - \mathbf{f}(\beta)] - 2 [\mathbf{y} - \mathbf{f}(\beta)]^{\mathrm{T}} \mathbf{J}\Delta\beta + \Delta\beta^{\mathrm{T}} \mathbf{J}^{\mathrm{T}} \mathbf{J}\Delta\beta,$$

где $\mathbf{J}$ — якобиан (матрица Якоби), чьи i-е строки содержат $\mathbf{J}_i$, и где $\mathbf{f}(\beta)$ и $\mathbf{y}$ – векторы с i-ми компонентами $f(x_i, \beta)$ и y_i соответственно.
Дифференцирование $S(\beta + \Delta\beta)$ по $\Delta\beta$ и приравнивание полученной производной к нулю приводит к равенству

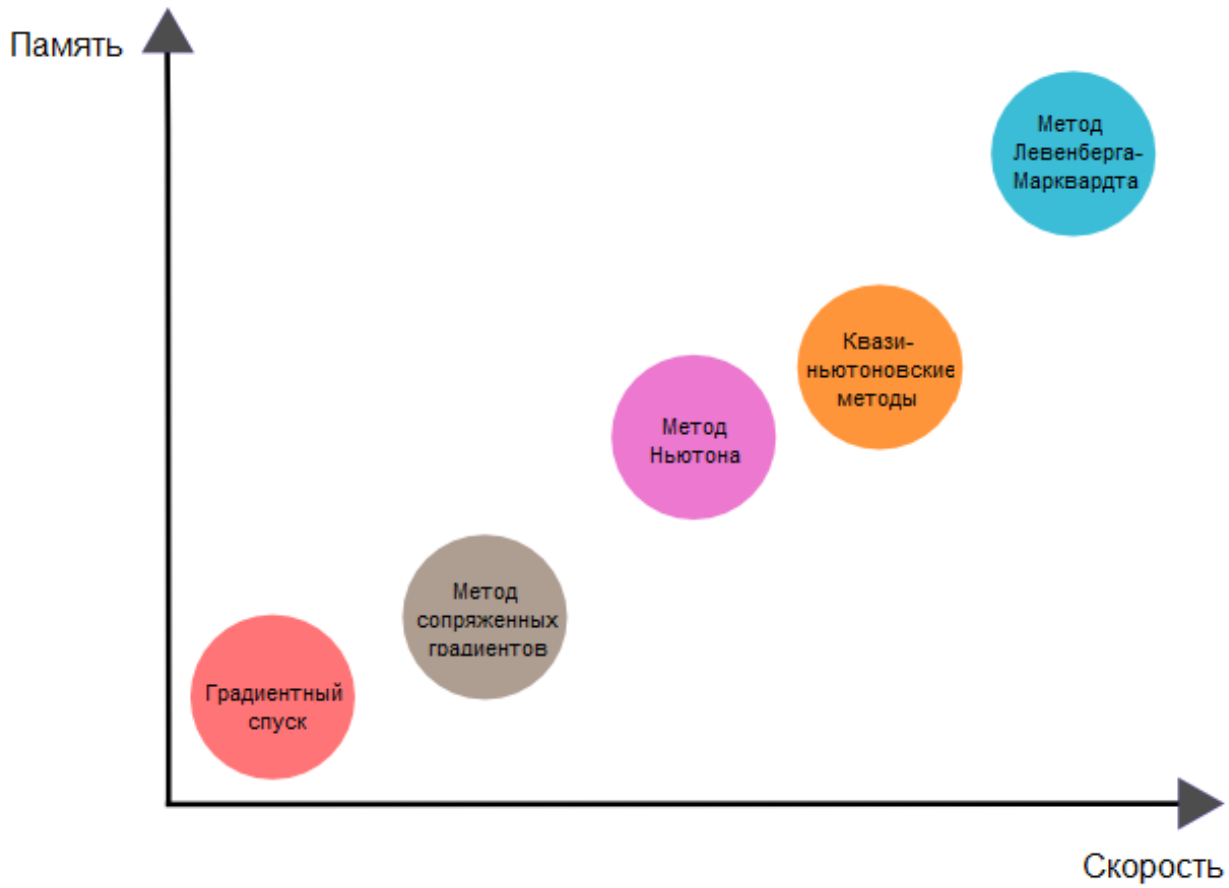
$$\left(\mathbf{J}^{\mathrm{T}} \mathbf{J}\right) \Delta\beta = \mathbf{J}^{\mathrm{T}} [\mathbf{y} - \mathbf{f}(\beta)],$$

которое по сути имеет вид системы линейных уравнений относительно $\Delta\beta$.
Указанное равенство может быть заменено на следующее:

$$\left(\mathbf{J}^{\mathrm{T}} \mathbf{J} + \lambda \mathbf{I}\right) \Delta\beta = \mathbf{J}^{\mathrm{T}} [\mathbf{y} - \mathbf{f}(\beta)],$$

где $\mathbf{I}$ – единичная матрица, дающая приращение $\Delta\beta$ вектору β.

Демонстрация



Блог по машинному обучению: 5 алгоритмов для обучения нейронных сетей

Спасибо за внимание!