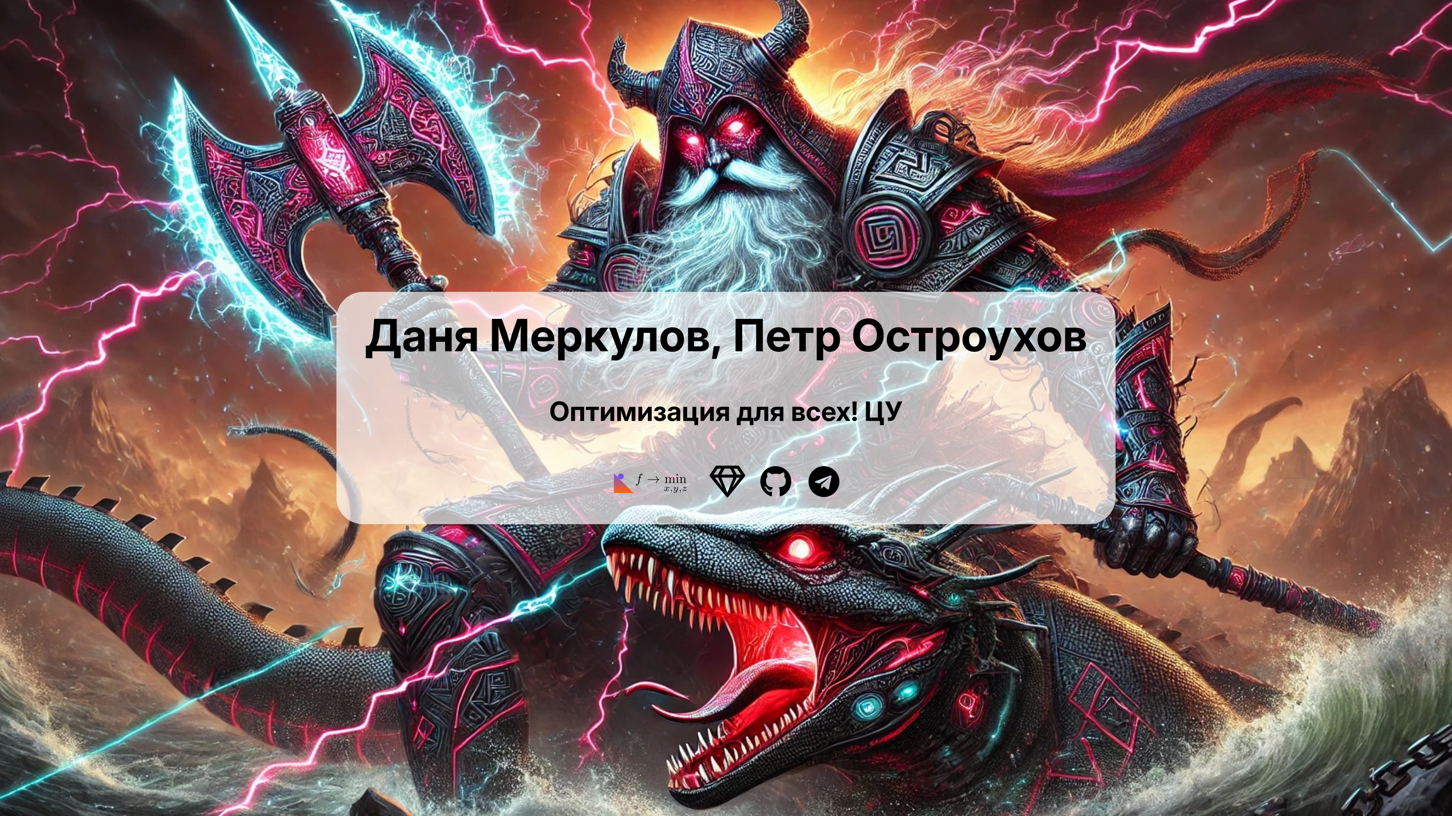
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Метод Ньютона и квазиньютоновские методы

МЕТОДЫ ВЫПУКЛОЙ ОПТИМИЗАЦИИ

НЕДЕЛЯ 10

Даня Меркулов

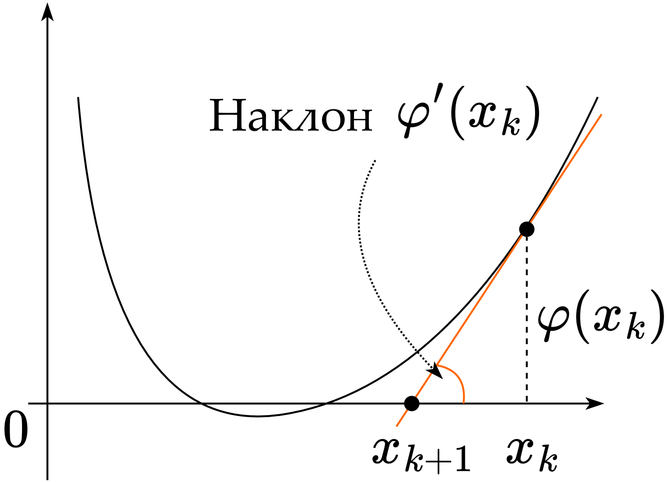


Даня Меркулов, Петр Остроухов

Оптимизация для всех! ЦУ

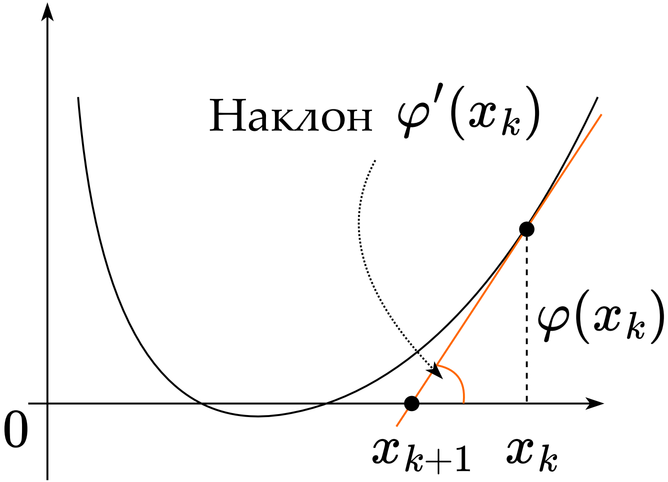
Метод Ньютона

Идея метода Ньютона для нахождения корней функции



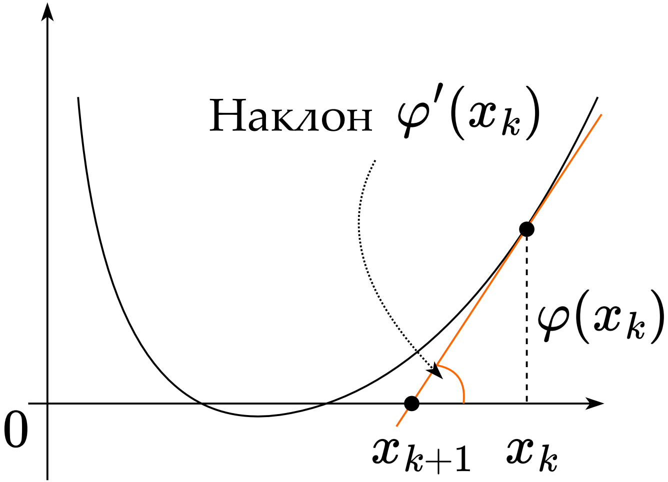
Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.

Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

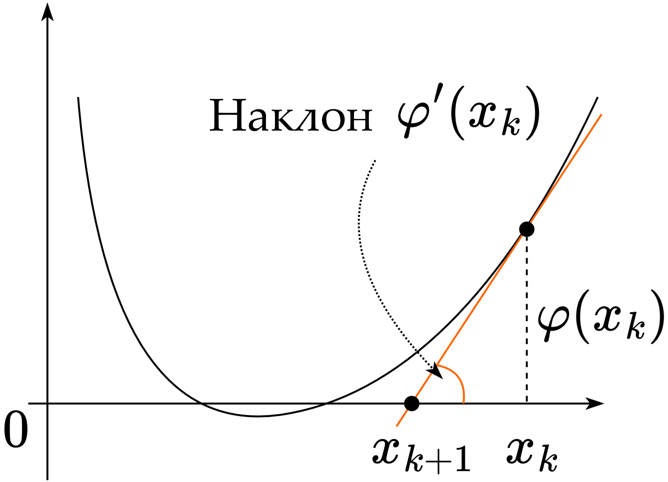
Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

$$\varphi'(x_k) = \frac{\varphi(x_k)}{x_k - x_{k+1}}$$

Идея метода Ньютона для нахождения корней функции

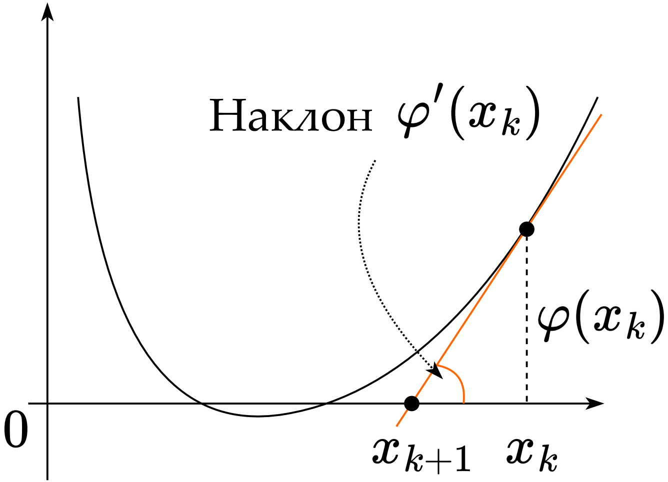


Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

$$\varphi'(x_k) = \frac{\varphi(x_k)}{x_k - x_{k+1}}$$

Мы получаем итерационную схему:

Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

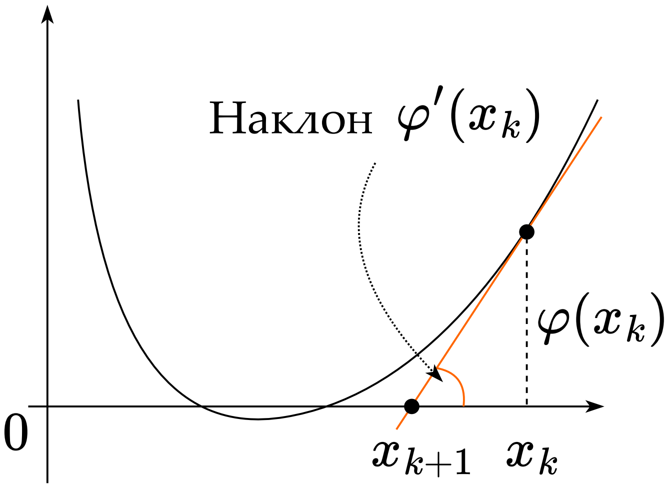
$$\varphi'(x_k) = \frac{\varphi(x_k)}{x_k - x_{k+1}}$$

Мы получаем итерационную схему:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)}.$$

[1] Мы фактически решаем задачу нахождения стационарных точек $\nabla f(x) = 0$

Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

$$\varphi'(x_k) = \frac{\varphi(x_k)}{x_k - x_{k+1}}$$

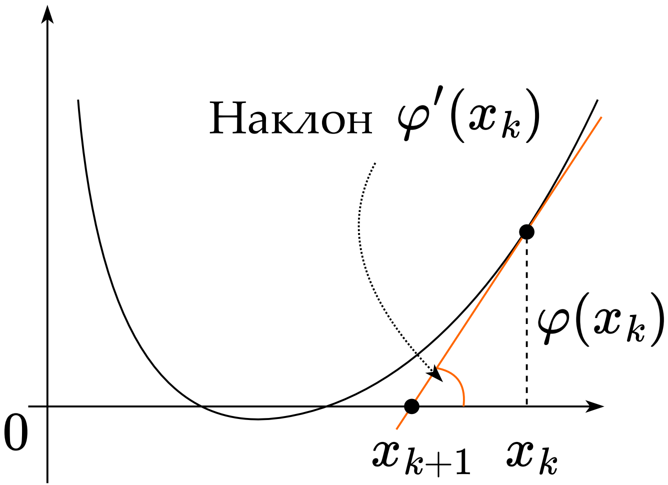
Мы получаем итерационную схему:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)}.$$

Этот метод станет методом оптимизации Ньютона в случае $f'(x) = \varphi(x)$[1]:

[1]Мы фактически решаем задачу нахождения стационарных точек $\nabla f(x) = 0$

Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию $\varphi(x) : \mathbb{R} \to \mathbb{R}$.
Основная идея заключается в том, чтобы построить линейное приближение в точке x_k и найти его корень, который будет новой точкой итерации:

$$\varphi'(x_k) = \frac{\varphi(x_k)}{x_k - x_{k+1}}$$

Мы получаем итерационную схему:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)}.$$

Этот метод станет методом оптимизации Ньютона в случае $f'(x) = \varphi(x)$[1]:

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

[1]Мы фактически решаем задачу нахождения стационарных точек $\nabla f(x) = 0$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f_{x_k}^{II}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f^{II}_{x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f^{II}_{x_k}(x)$, т.е. $\nabla f^{II}_{x_k}(x_{k+1}) = 0$.

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f_{x_k}^{II}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f_{x_k}^{II}(x)$, т.е. $\nabla f_{x_k}^{II}(x_{k+1}) = 0$.

$$\nabla f_{x_k}^{II}(x_{k+1}) = \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f^{II}_{x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k\rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f^{II}_{x_k}(x)$, т.е. $\nabla f^{II}_{x_k}(x_{k+1}) = 0$.

$$\begin{aligned} \nabla f^{II}_{x_k}(x_{k+1}) &= \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\ \nabla^2 f(x_k)(x_{k+1} - x_k) &= -\nabla f(x_k) \end{aligned}$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f^{II}_{x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k\rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f^{II}_{x_k}(x)$, т.е. $\nabla f^{II}_{x_k}(x_{k+1}) = 0$.

$$\begin{aligned}
\nabla f^{II}_{x_k}(x_{k+1}) &= \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\
\nabla^2 f(x_k)(x_{k+1} - x_k) &= -\nabla f(x_k) \\
\left[\nabla^2 f(x_k)\right]^{-1} \nabla^2 f(x_k)(x_{k+1} - x_k) &= -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)
\end{aligned}$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f_{x_k}^{II}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f_{x_k}^{II}(x)$, т.е. $\nabla f_{x_k}^{II}(x_{k+1}) = 0$.

$$\begin{aligned}
\nabla f_{x_k}^{II}(x_{k+1}) &= \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\
\nabla^2 f(x_k)(x_{k+1} - x_k) &= -\nabla f(x_k) \\
\left[\nabla^2 f(x_k)\right]^{-1} \nabla^2 f(x_k)(x_{k+1} - x_k) &= -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) \\
x_{k+1} &= x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k).
\end{aligned}$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

$$f_{x_k}^{II}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f_{x_k}^{II}(x)$, т.е. $\nabla f_{x_k}^{II}(x_{k+1}) = 0$.

$$\begin{aligned}
\nabla f_{x_k}^{II}(x_{k+1}) &= \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\
\nabla^2 f(x_k)(x_{k+1} - x_k) &= -\nabla f(x_k) \\
\left[\nabla^2 f(x_k)\right]^{-1} \nabla^2 f(x_k)(x_{k+1} - x_k) &= -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) \\
x_{k+1} &= x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k).
\end{aligned}$$

Метод Ньютона как оптимизация локальной квадратичной аппроксимации

Пусть у нас есть функция $f(x)$ и некоторая точка x_k. Рассмотрим квадратичное приближение этой функции в окрестности x_k:

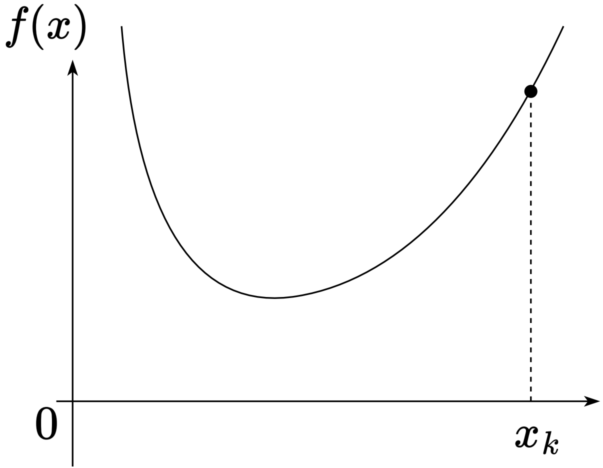
$$f^{II}_{x_k}(x) = f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k\rangle.$$

Идея метода заключается в том, чтобы найти точку x_{k+1}, которая минимизирует функцию $f^{II}_{x_k}(x)$, т.е. $\nabla f^{II}_{x_k}(x_{k+1}) = 0$.

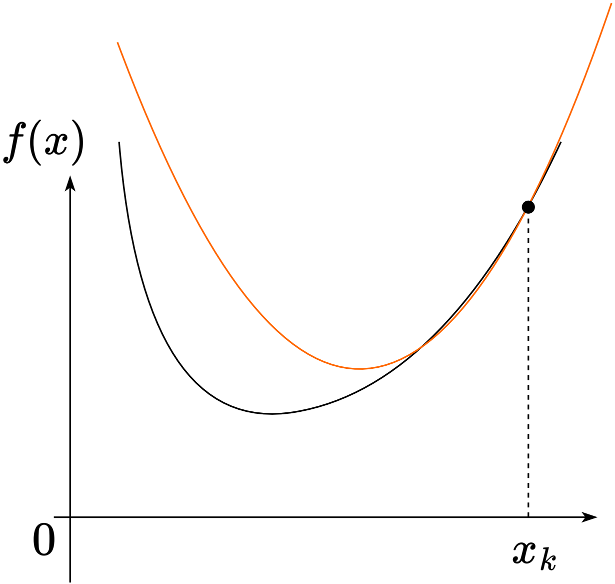
$$\begin{aligned}
\nabla f^{II}_{x_k}(x_{k+1}) &= \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0 \\
\nabla^2 f(x_k)(x_{k+1} - x_k) &= -\nabla f(x_k) \\
\left[\nabla^2 f(x_k)\right]^{-1} \nabla^2 f(x_k)(x_{k+1} - x_k) &= -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) \\
x_{k+1} &= x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k).
\end{aligned}$$

Необходимо отметить ограничения, связанные с необходимостью невырожденности (для существования метода) и положительной определенности (для гарантии сходимости) гессиана.

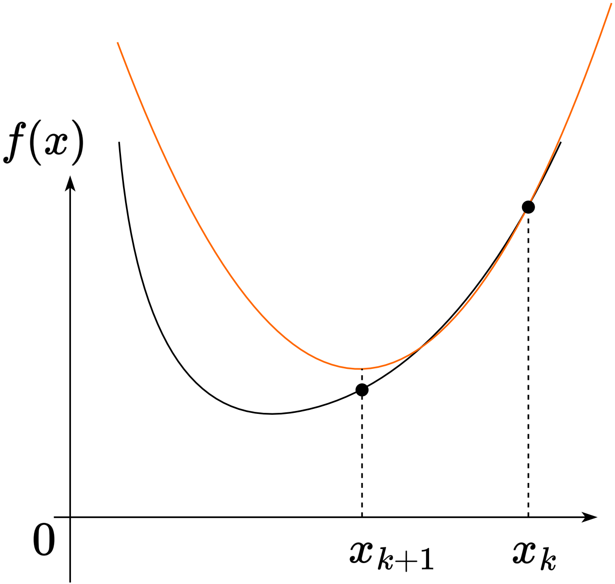
Метод Ньютона как оптимизация локальной квадратичной аппроксимации



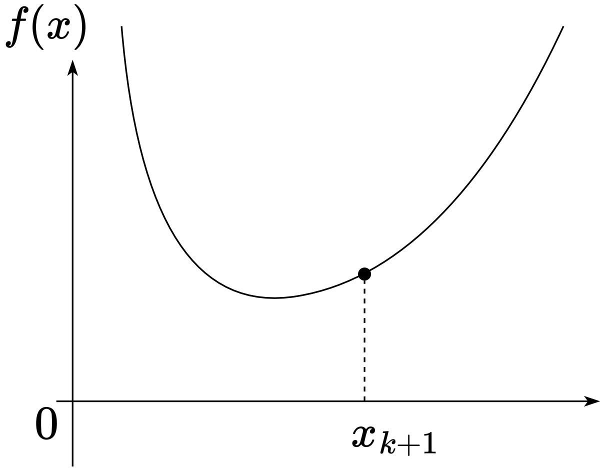
Метод Ньютона как оптимизация локальной квадратичной аппроксимации



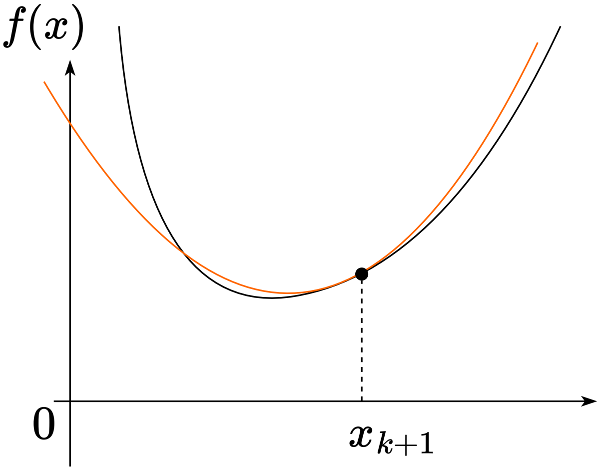
Метод Ньютона как оптимизация локальной квадратичной аппроксимации



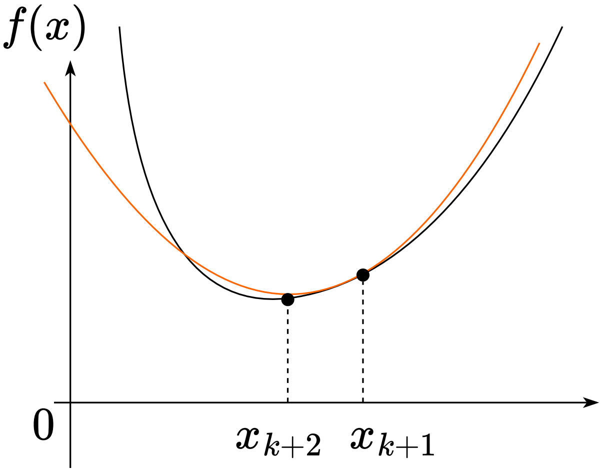
Метод Ньютона как оптимизация локальной квадратичной аппроксимации



Метод Ньютона как оптимизация локальной квадратичной аппроксимации



Метод Ньютона как оптимизация локальной квадратичной аппроксимации



Свойства метода Ньютона

Квадратичная сходимость метода Ньютона

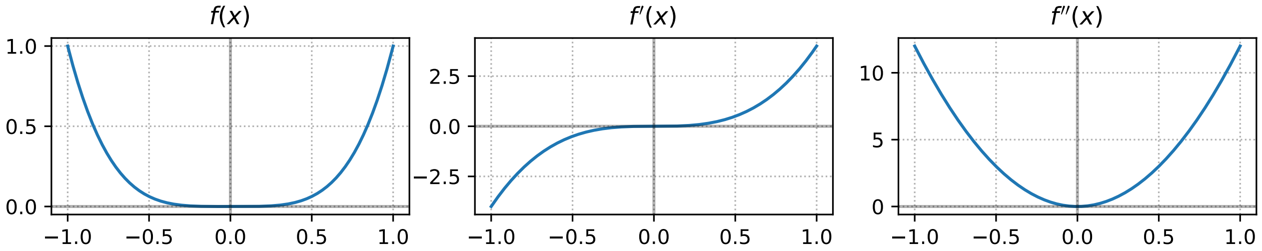
Theorem

Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:

$$\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2$$

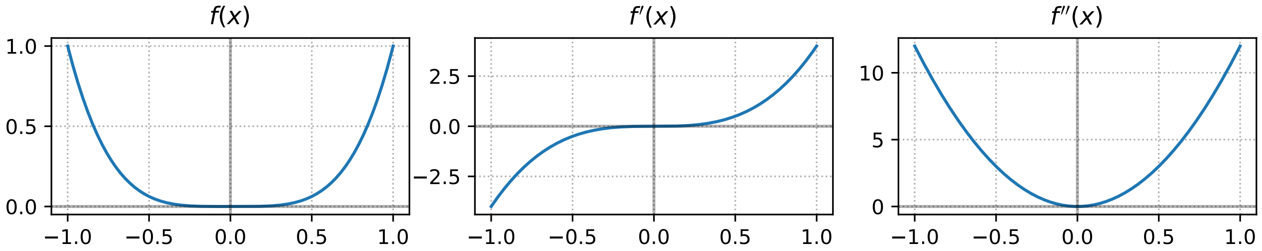
Отсутствие квадратичной сходимости, если некоторые предположения нарушаются

$$f(x) = x^4 \qquad f'(x) = 4x^3 \qquad f''(x) = 12x^2$$



Отсутствие квадратичной сходимости, если некоторые предположения нарушаются

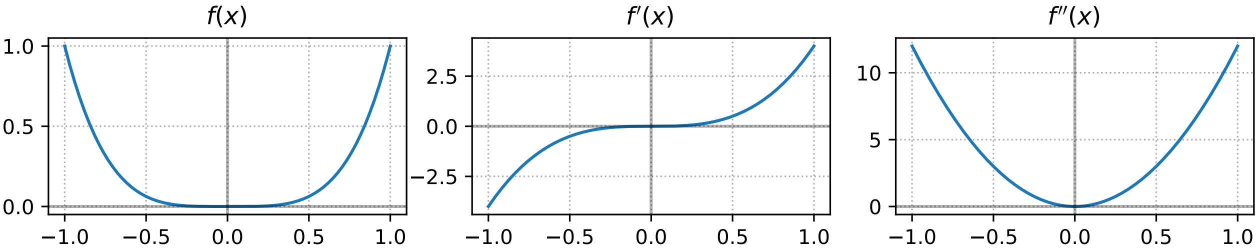
$$f(x) = x^4 \qquad f'(x) = 4x^3 \qquad f''(x) = 12x^2$$



$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - \frac{4x_k^3}{12x_k^2} = x_k - \frac{1}{3}x_k = \frac{2}{3}x_k,$$

Отсутствие квадратичной сходимости, если некоторые предположения нарушаются

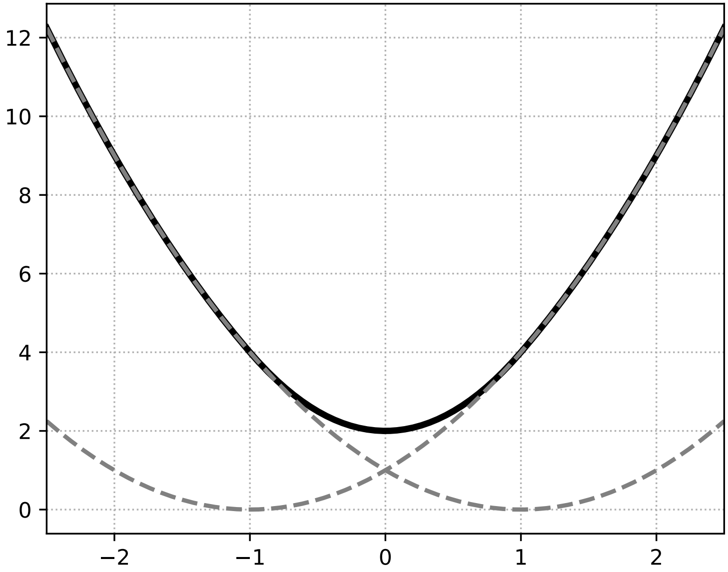
$$f(x) = x^4 \qquad f'(x) = 4x^3 \qquad f''(x) = 12x^2$$



$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - \frac{4x_k^3}{12x_k^2} = x_k - \frac{1}{3}x_k = \frac{2}{3}x_k,$$

сходится линейно к 0, единственному решению задачи, с линейной скоростью.

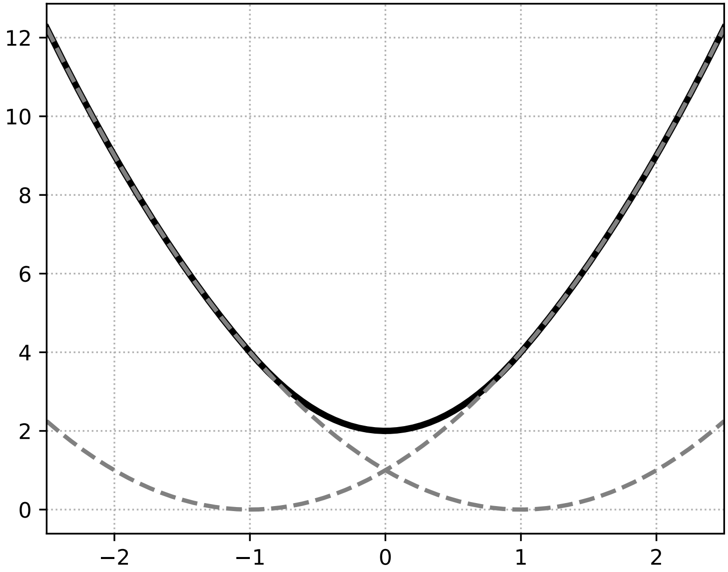
Локальная сходимость метода Ньютона для гладкой сильно выпуклой $f(x)$



$$f(x) = \begin{cases} (x-1)^2, & x \leq -1 \\ 2x^2 + 2, & -1 < x < 1 \\ (x+1)^2, & x \geq 1 \end{cases}$$

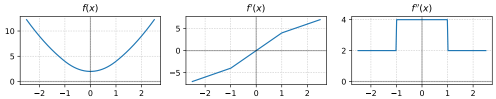
Эта функция сильно выпукла, но вторая производная не является липшицевой.

Локальная сходимость метода Ньютона для гладкой сильно выпуклой $f(x)$

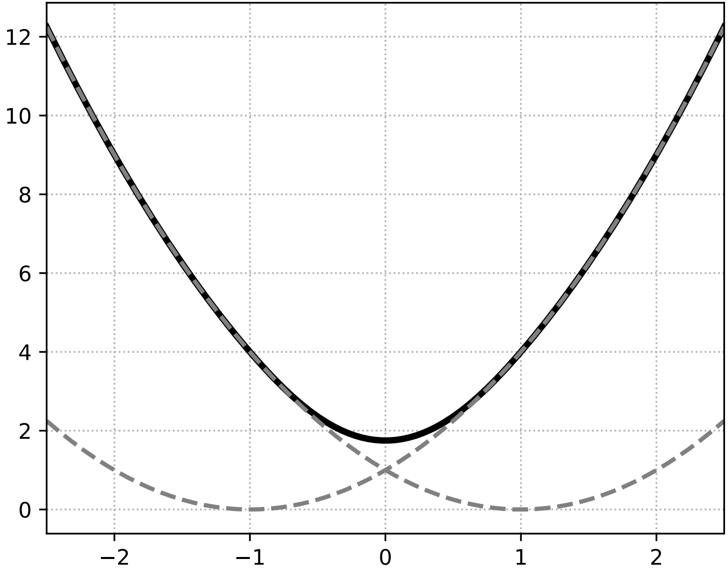


$$f(x) = \begin{cases} (x-1)^2, & x \leq -1 \\ 2x^2 + 2, & -1 < x < 1 \\ (x+1)^2, & x \geq 1 \end{cases}$$

Эта функция сильно выпукла, но вторая производная не является липшицевой.

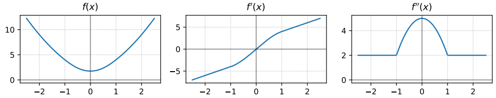


Локальная сходимость метода Ньютона даже если $\nabla^2 f$ липшицев

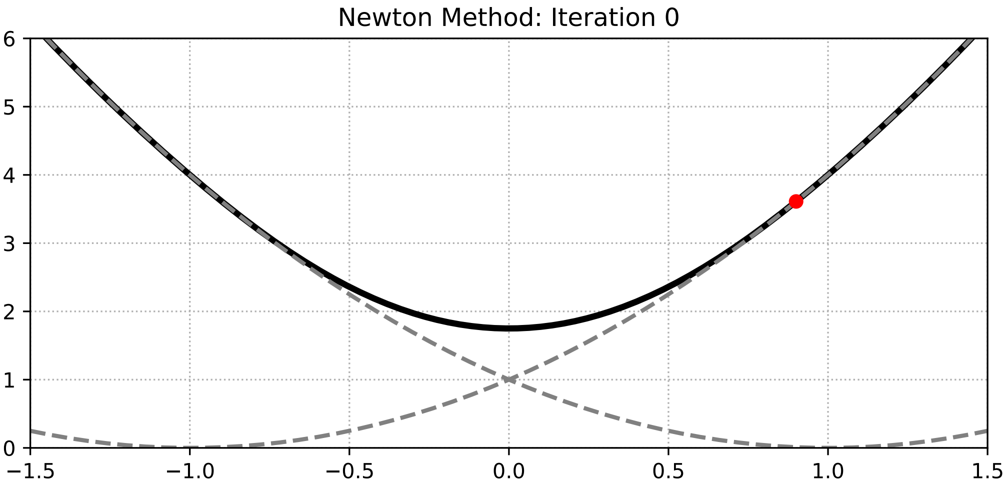


$$f(x) = \begin{cases} (x-1)^2, & x \leq -1 \\ -\frac{1}{4}x^4 + \frac{5}{2}x^2 + \frac{7}{4}, & -1 < x < 1 \\ (x+1)^2, & x \geq 1 \end{cases}$$

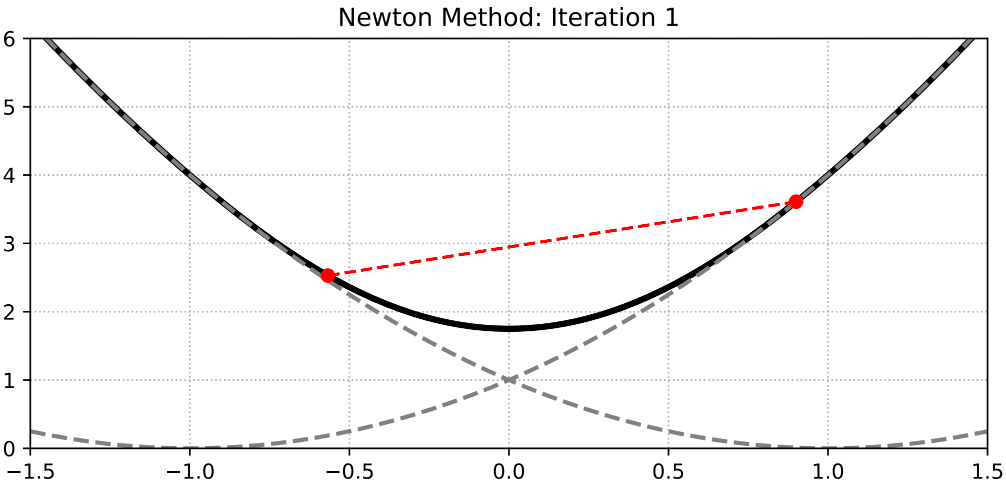
Эта функция сильно выпукла и вторая производная является липшицевой.



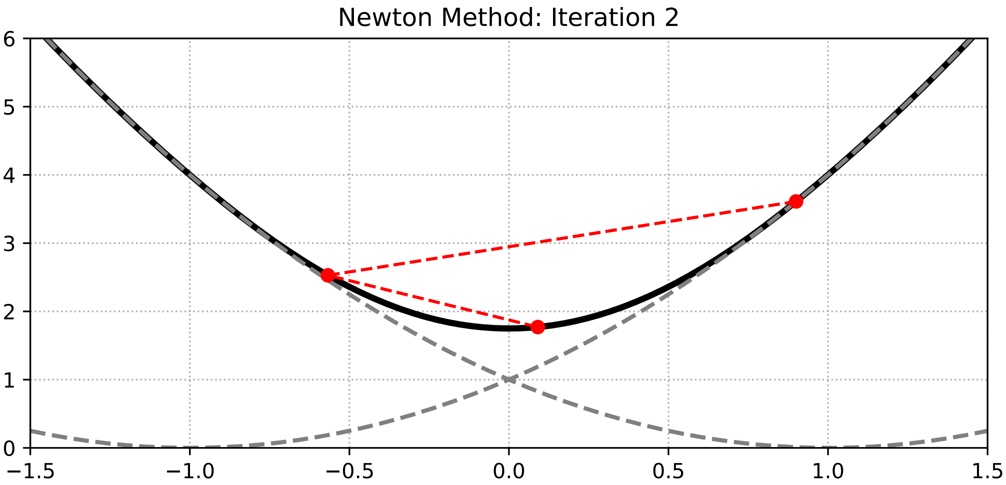
Локальная сходимость метода Ньютона. Хорошая инициализация



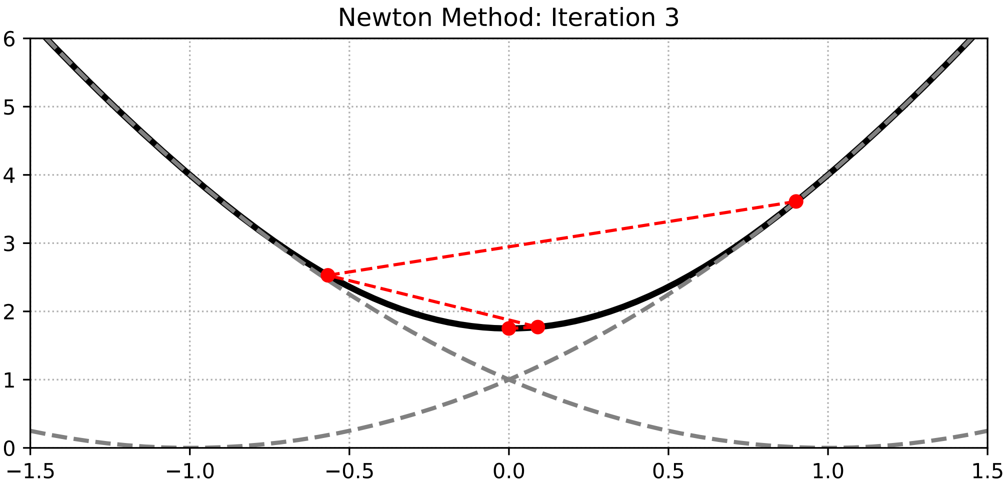
Локальная сходимость метода Ньютона. Хорошая инициализация



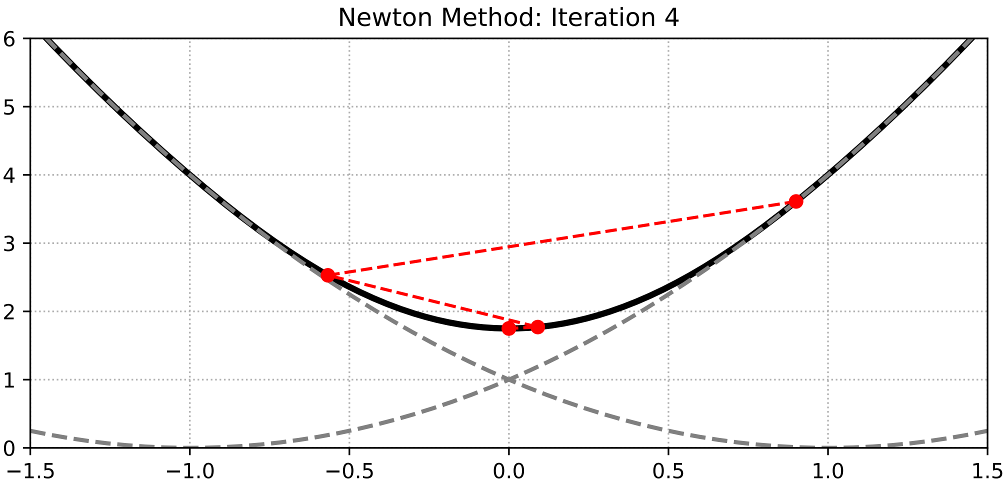
Локальная сходимость метода Ньютона. Хорошая инициализация



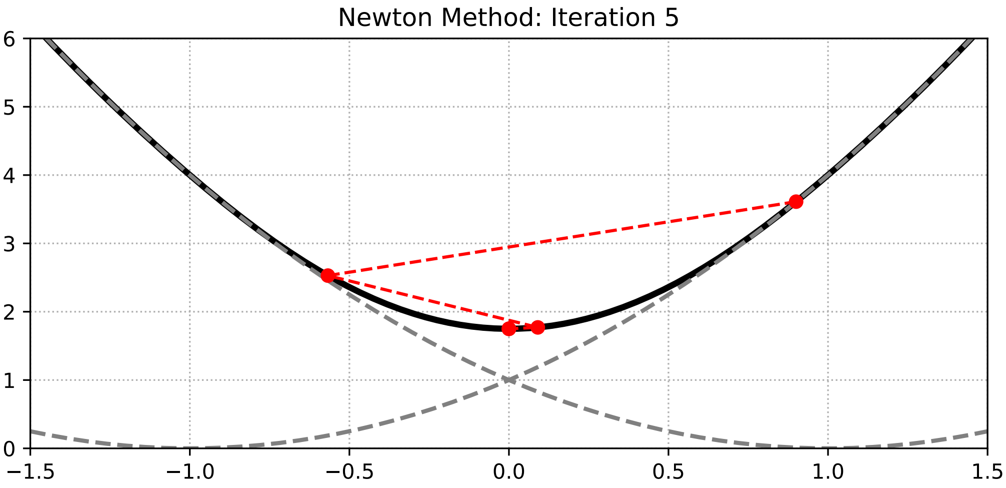
Локальная сходимость метода Ньютона. Хорошая инициализация



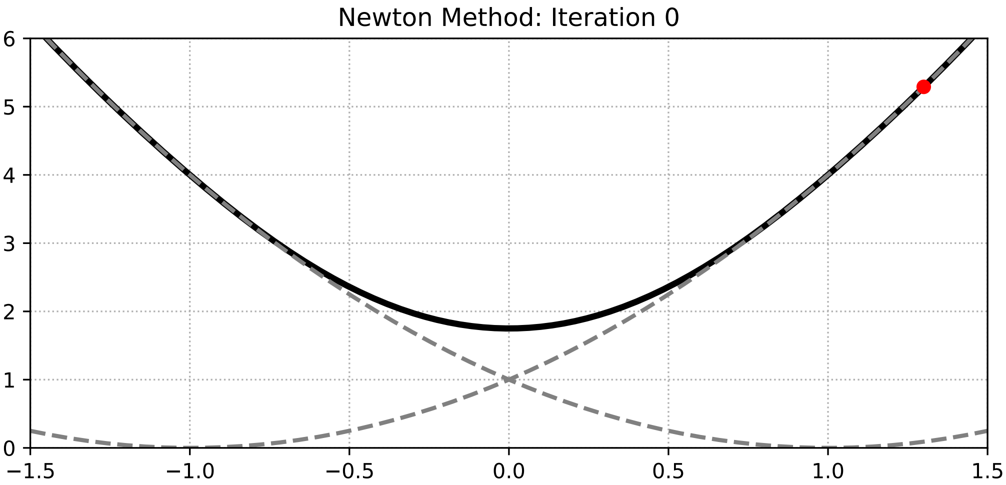
Локальная сходимость метода Ньютона. Хорошая инициализация



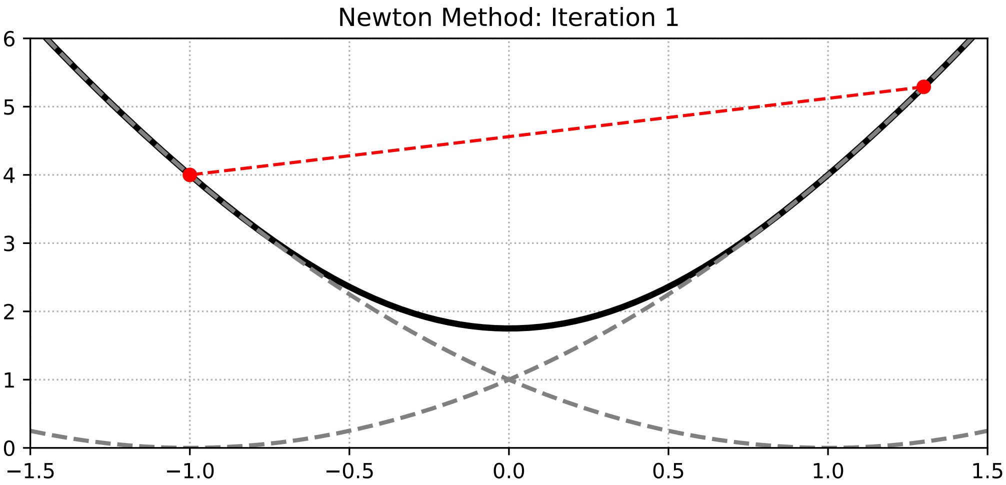
Локальная сходимость метода Ньютона. Хорошая инициализация



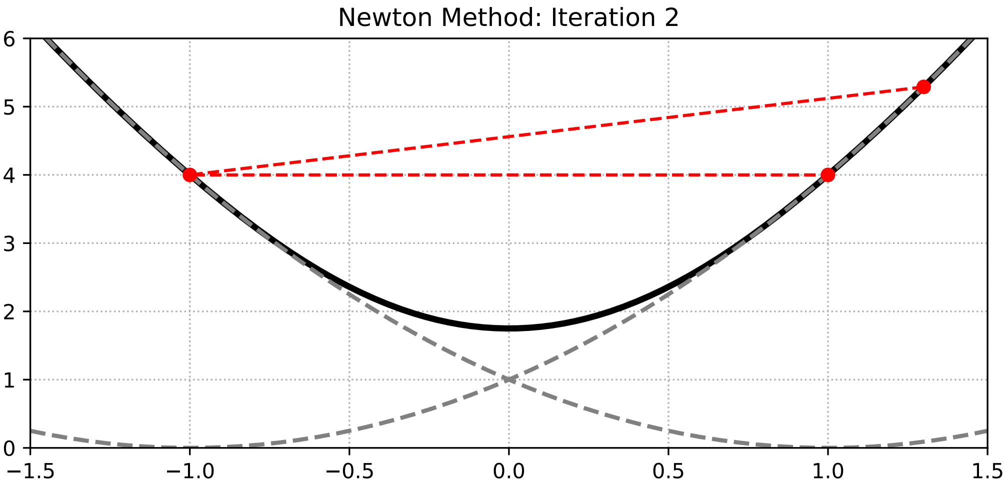
Локальная сходимость метода Ньютона. Плохая инициализация



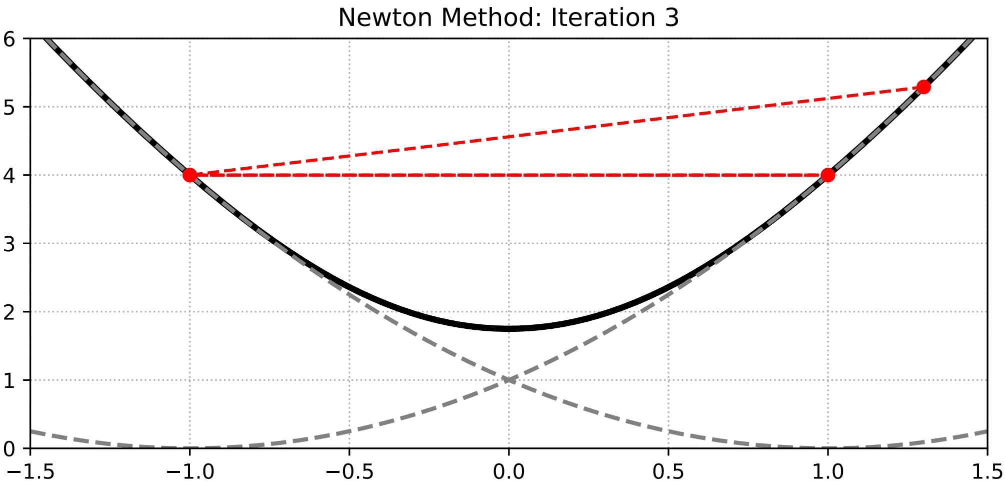
Локальная сходимость метода Ньютона. Плохая инициализация



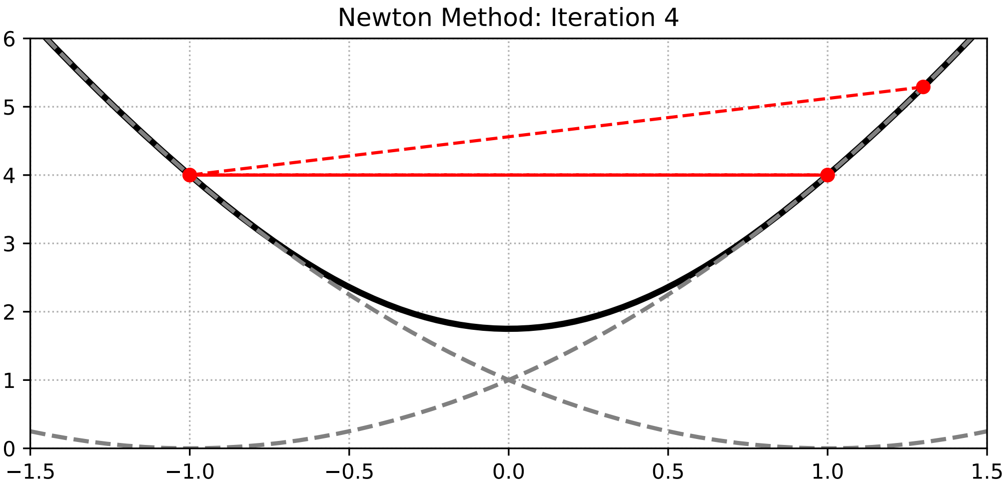
Локальная сходимость метода Ньютона. Плохая инициализация



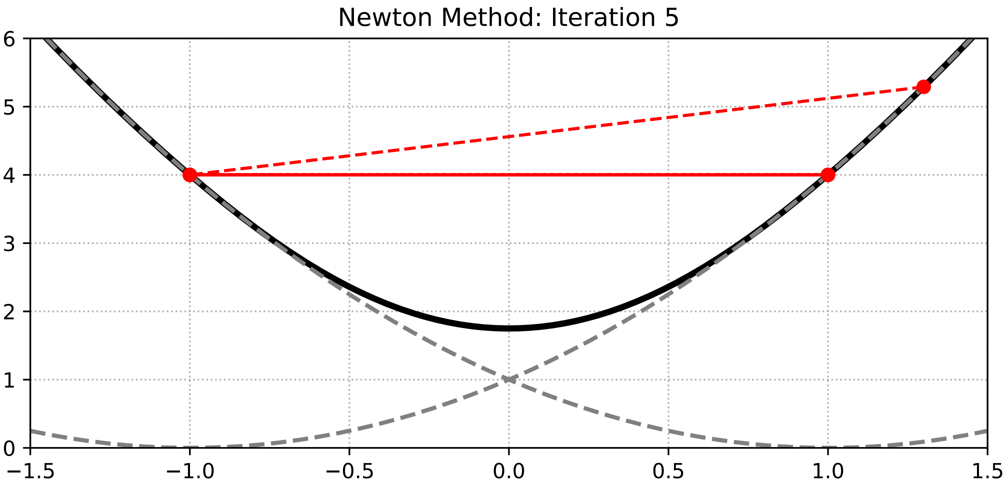
Локальная сходимость метода Ньютона. Плохая инициализация



Локальная сходимость метода Ньютона. Плохая инициализация



Локальная сходимость метода Ньютона. Плохая инициализация



Проблемы метода Ньютона

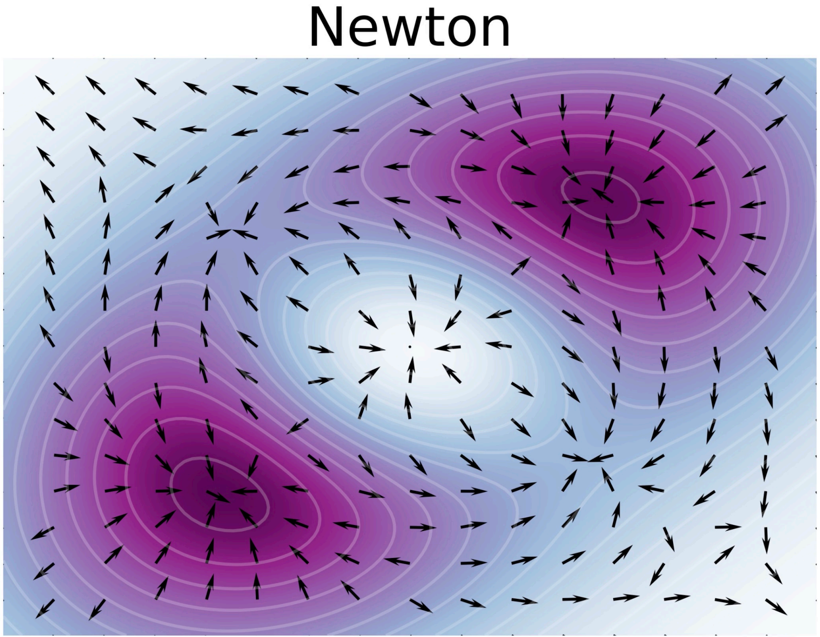


Рисунок 1. Animation

Проблемы метода Ньютона

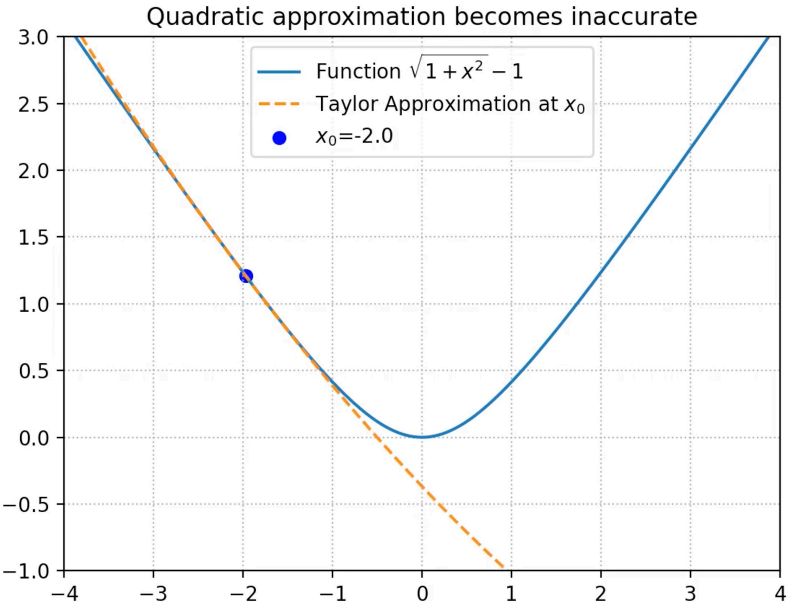


Рисунок 2. Animation

Метод Ньютона для квадратичной задачи (линейной регрессии)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top A x - b^\top x, \qquad A \in \mathbb{R}^{n\times n}, \quad \lambda(A) \in [\mu; L].$$

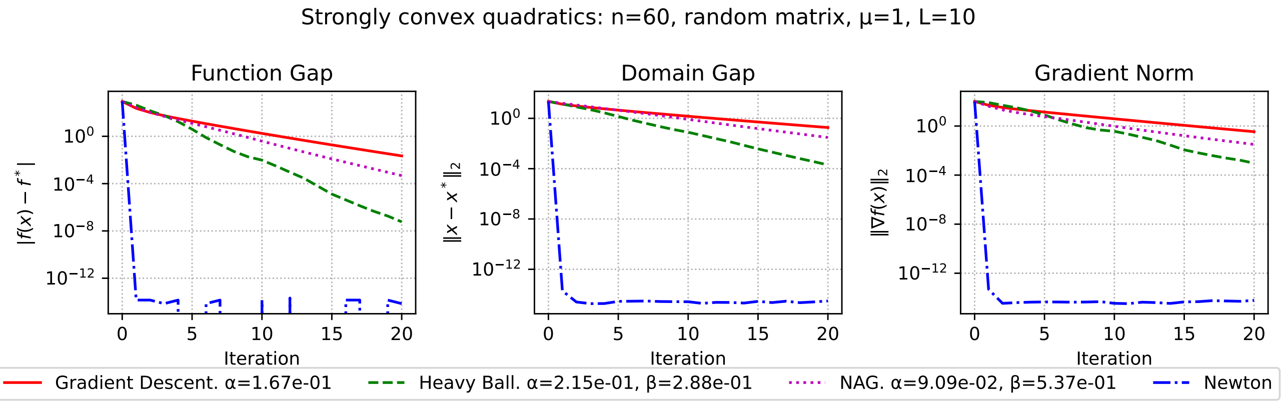


Рисунок 3. Так как задача - квадратичная, то метод Ньютона сходится за один шаг.

Метод Ньютона для квадратичной задачи (линейной регрессии)

$$
\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x, \qquad A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].
$$

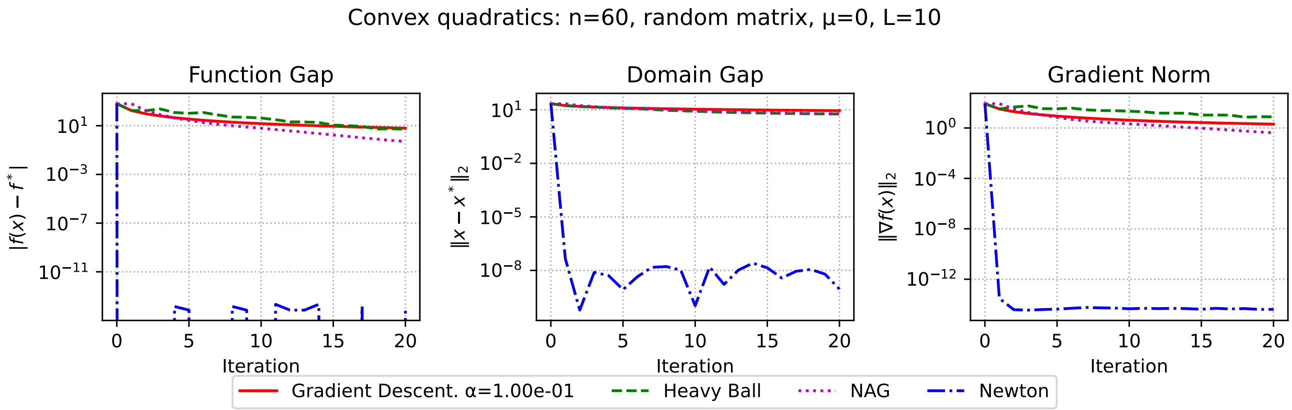


Рисунок 4. В этом случае метод Ньютона тоже крайне быстро сходится, однако, отметим, что это происходит благодаря тому, что минимальное собственное число гессиана не 0, а около 10^{-8}. Если применять метод Ньютона в наивной форме с обращением матрицы, то получится ошибка, так как матрица вырождена. На практике все равно можно использовать метод, если для направления

Метод Ньютона для квадратичной задачи (линейной регрессии)

$$
\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x, \qquad A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].
$$

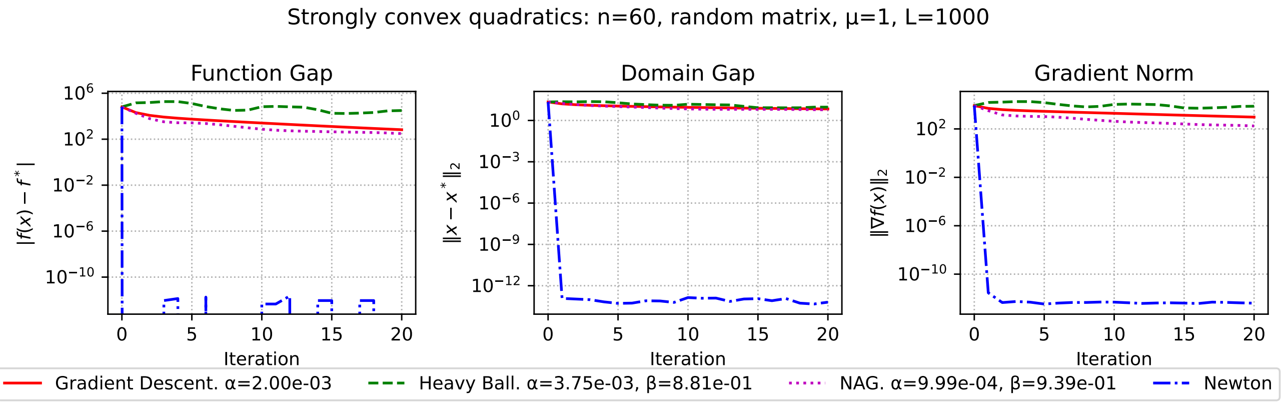


Рисунок 5. Здесь число обусловленности гессиана в 1000 раз больше, чем в предыдущем случае, и метод Ньютона сходится за 1 итерацию.

Метод Ньютона для задачи бинарной логистической регрессии

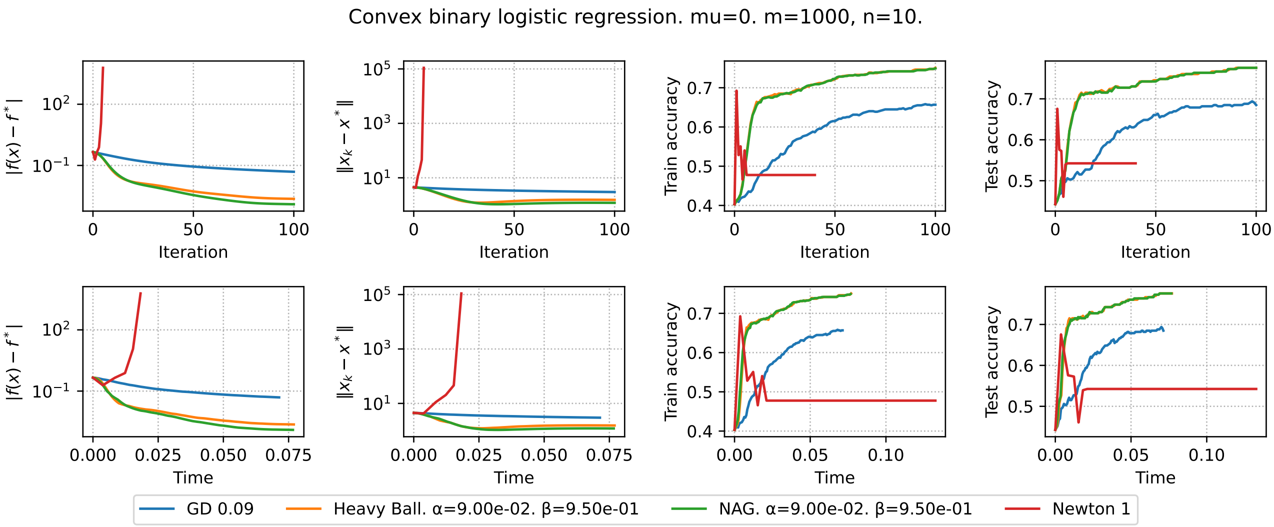


Рисунок 6. Наблюдается расходимость метода Ньютона. Сразу отметим, что в задаче нет регуляризации и гарантии сильной выпуклости. А также нет гарантий того, что мы инициализируем метод в окрестности решения.

Метод Ньютона для задачи бинарной логистической регрессии

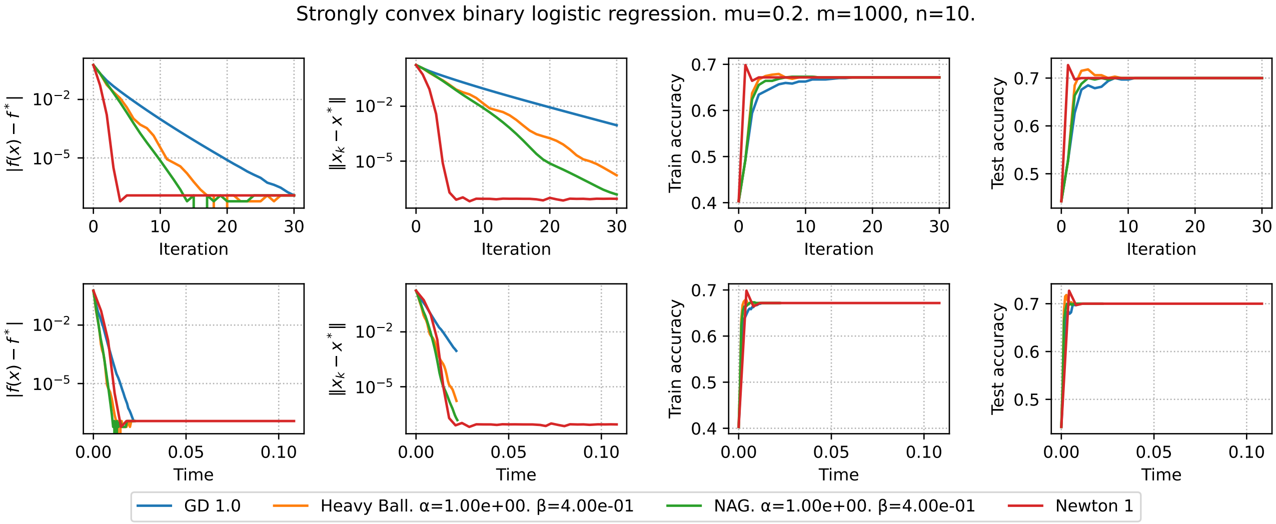


Рисунок 7. Добавление регуляризации гарантирует сильную выпуклость, наблюдается сходимость метода Ньютона.

Метод Ньютона для задачи бинарной логистической регрессии

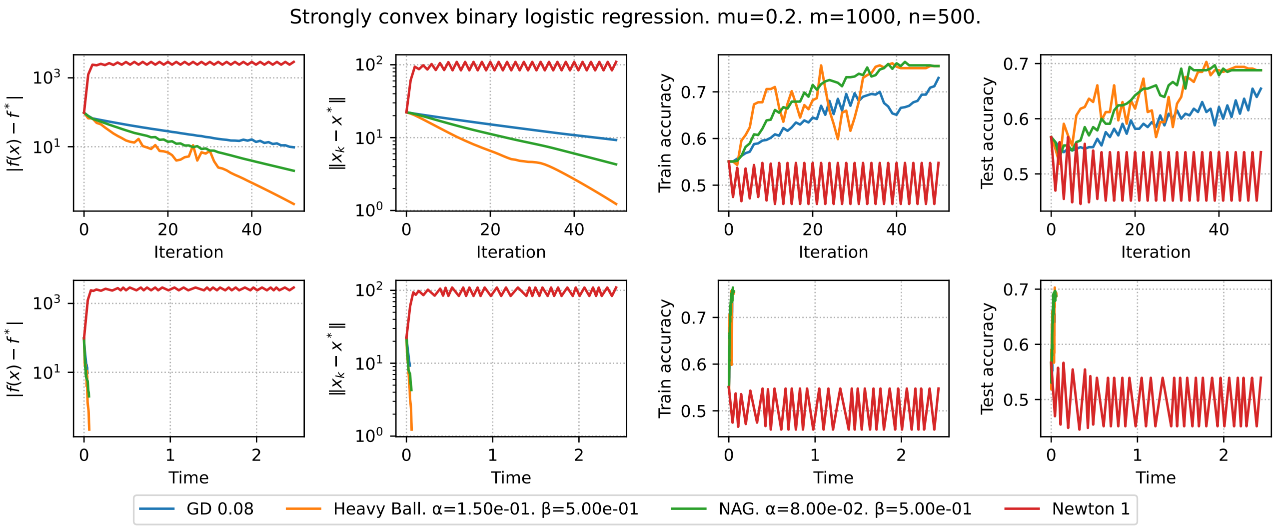


Рисунок 8. Увеличим размерность в 50 раз и наблюдаем расходимость метода Ньютона. Это можно связать с тем, что мы инициализируем метод в точке, далекой от решения

Метод Ньютона для задачи бинарной логистической регрессии

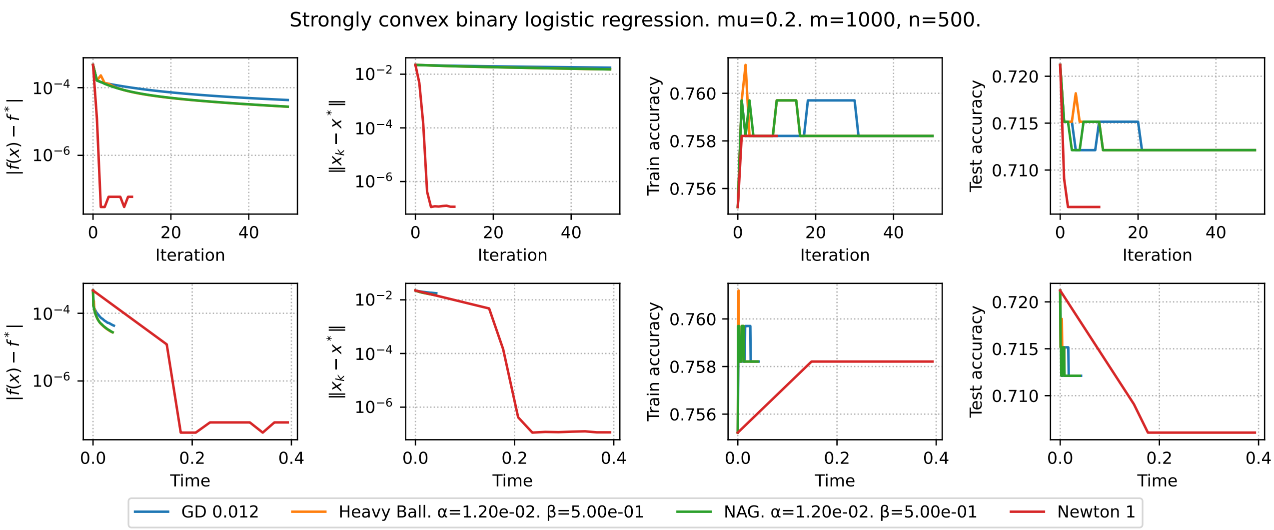


Рисунок 9. Не меняя задачу, изменим начальную точку и наблюдаем квадратичную сходимость метода Ньютона. Однако, обратите внимание на время работы. Уже при небольшой размерности, метод Ньютона работает значительно дольше, чем градиентные методы.

Задача нахождения аналитического центра многогранника

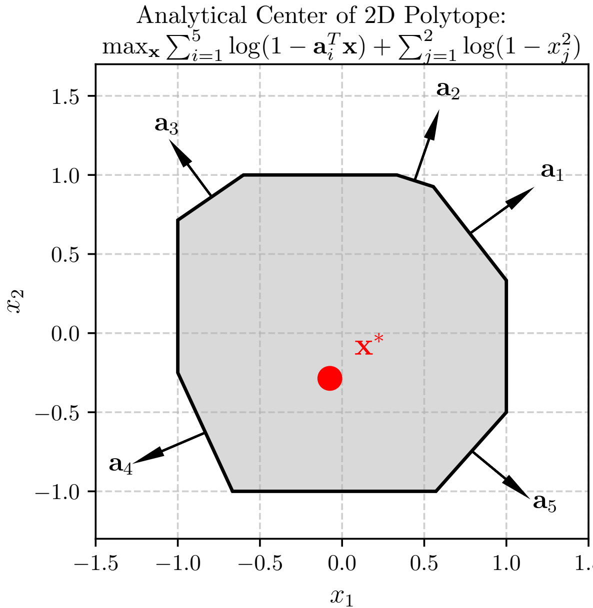
Найти точку $x \in \mathbb{R}^n$, которая максимизирует сумму логарифмов расстояний до границ многогранника:

$$\max_x \sum_{i=1}^{m} \log(1 - a_i^T x) + \sum_{j=1}^{n} \log(1 - x_j^2)$$

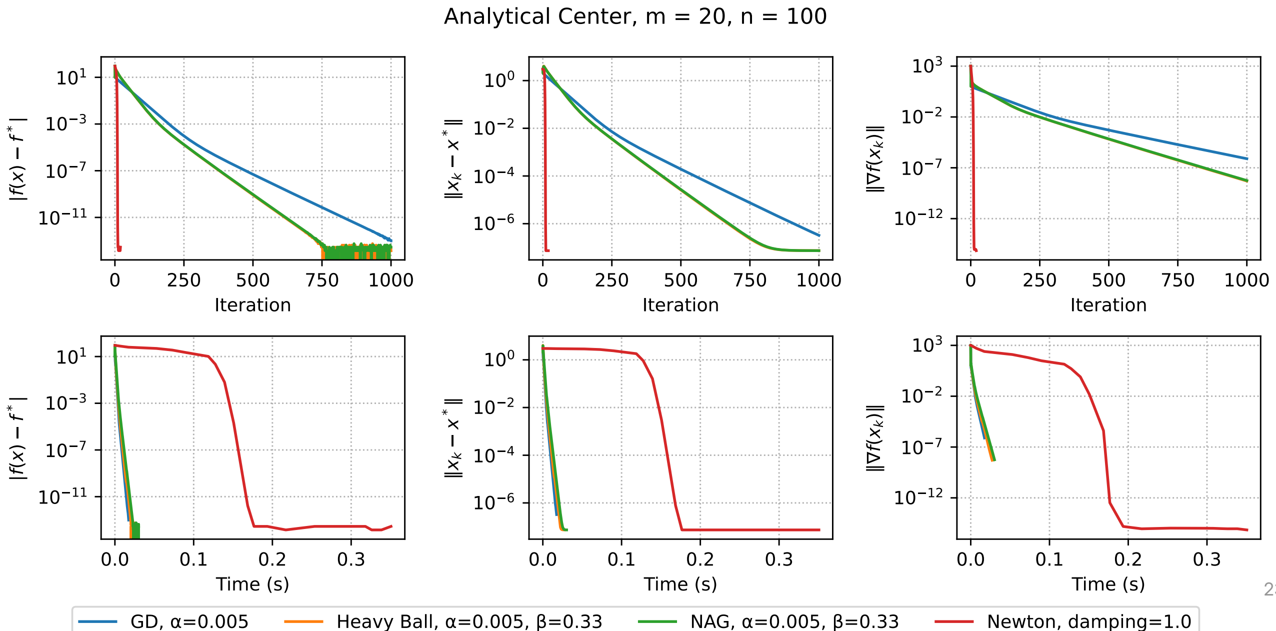
или, эквивалентно, минимизирует:

$$\min_x - \sum_{i=1}^{m} \log(1 - a_i^T x) - \sum_{j=1}^{n} \log(1 - x_j^2)$$

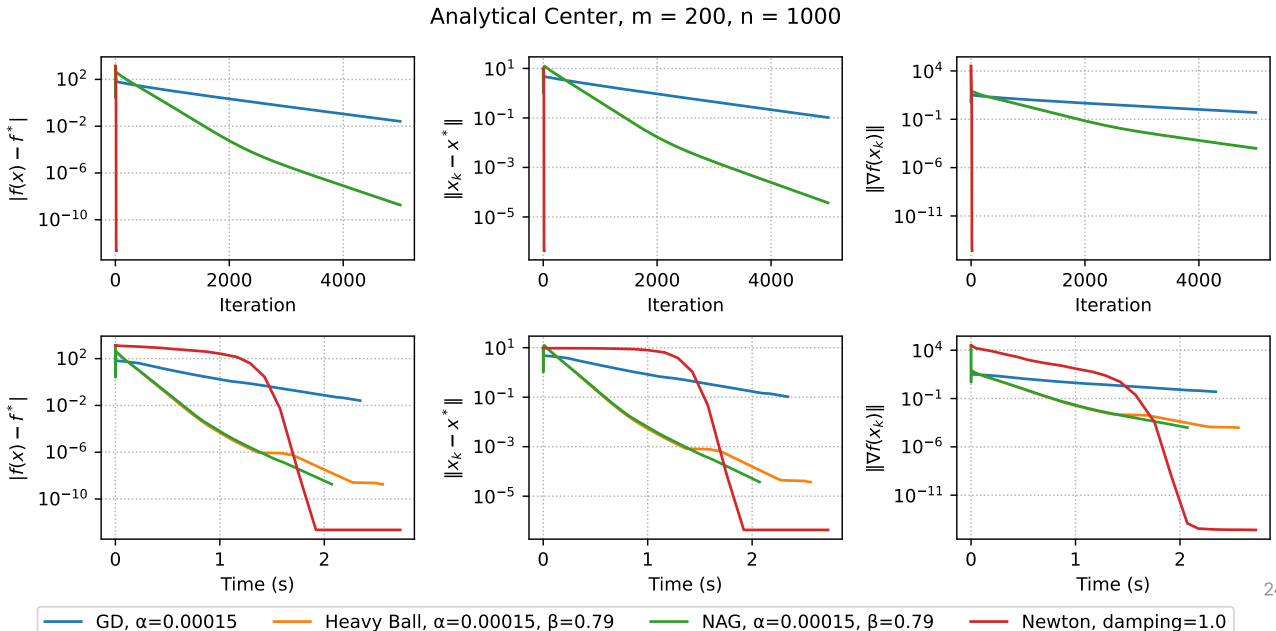
при ограничениях: - $a_i^T x < 1$ для всех $i = 1, ..., m$, где a_i - строки матрицы A^T - $|x_j| < 1$ для всех $j = 1, ..., n$
Аналитический центр многогранника - это точка, которая максимально удалена от всех границ многогранника в смысле логарифмического барьера. Эта концепция широко используется в методах внутренней точки для выпуклой оптимизации.



Задача нахождения аналитического центра многогранника



Задача нахождения аналитического центра многогранника



Аффинная инвариантность метода Ньютона

Важным свойством метода Ньютона является **аффинная инвариантность**. Пусть дана функция f и невырожденная матрица $A \in \mathbb{R}^{n \times n}$, пусть $x = Ay$, и пусть $g(y) = f(Ay)$. Заметим, что $\nabla g(y) = A^T \nabla f(x)$ и $\nabla^2 g(y) = A^T \nabla^2 f(x) A$. Шаги Ньютона на g выражаются как:

$$y_{k+1} = y_k - \left(\nabla^2 g(y_k)\right)^{-1} \nabla g(y_k)$$

Аффинная инвариантность метода Ньютона

Важным свойством метода Ньютона является **аффинная инвариантность**. Пусть дана функция f и невырожденная матрица $A \in \mathbb{R}^{n\times n}$, пусть $x = Ay$, и пусть $g(y) = f(Ay)$. Заметим, что $\nabla g(y) = A^T \nabla f(x)$ и $\nabla^2 g(y) = A^T \nabla^2 f(x) A$. Шаги Ньютона на g выражаются как:

$$y_{k+1} = y_k - \left(\nabla^2 g(y_k)\right)^{-1} \nabla g(y_k)$$

Раскрывая это, мы получаем:

$$y_{k+1} = y_k - \left(A^T \nabla^2 f(Ay_k) A\right)^{-1} A^T \nabla f(Ay_k)$$

Аффинная инвариантность метода Ньютона

Важным свойством метода Ньютона является **аффинная инвариантность**. Пусть дана функция f и невырожденная матрица $A \in \mathbb{R}^{n \times n}$, пусть $x = Ay$, и пусть $g(y) = f(Ay)$. Заметим, что $\nabla g(y) = A^T \nabla f(x)$ и $\nabla^2 g(y) = A^T \nabla^2 f(x) A$. Шаги Ньютона на g выражаются как:

$$y_{k+1} = y_k - \left(\nabla^2 g(y_k)\right)^{-1} \nabla g(y_k)$$

Раскрывая это, мы получаем:

$$y_{k+1} = y_k - \left(A^T \nabla^2 f(Ay_k) A\right)^{-1} A^T \nabla f(Ay_k)$$

Используя свойство обратной матрицы $(AB)^{-1} = B^{-1} A^{-1}$, это упрощается до:

$$y_{k+1} = y_k - A^{-1} \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$
$$Ay_{k+1} = Ay_k - \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$

Аффинная инвариантность метода Ньютона

Важным свойством метода Ньютона является **аффинная инвариантность**. Пусть дана функция f и невырожденная матрица $A \in \mathbb{R}^{n \times n}$, пусть $x = Ay$, и пусть $g(y) = f(Ay)$. Заметим, что $\nabla g(y) = A^T \nabla f(x)$ и $\nabla^2 g(y) = A^T \nabla^2 f(x) A$. Шаги Ньютона на g выражаются как:

$$y_{k+1} = y_k - \left(\nabla^2 g(y_k)\right)^{-1} \nabla g(y_k)$$

Раскрывая это, мы получаем:

$$y_{k+1} = y_k - \left(A^T \nabla^2 f(Ay_k) A\right)^{-1} A^T \nabla f(Ay_k)$$

Используя свойство обратной матрицы $(AB)^{-1} = B^{-1} A^{-1}$, это упрощается до:

$$y_{k+1} = y_k - A^{-1} \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$
$$Ay_{k+1} = Ay_k - \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$

Таким образом, правило обновления для x выглядит так:

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

Аффинная инвариантность метода Ньютона

Важным свойством метода Ньютона является **аффинная инвариантность**. Пусть дана функция f и невырожденная матрица $A \in \mathbb{R}^{n \times n}$, пусть $x = Ay$, и пусть $g(y) = f(Ay)$. Заметим, что $\nabla g(y) = A^T \nabla f(x)$ и $\nabla^2 g(y) = A^T \nabla^2 f(x) A$. Шаги Ньютона на g выражаются как:

$$y_{k+1} = y_k - \left(\nabla^2 g(y_k)\right)^{-1} \nabla g(y_k)$$

Раскрывая это, мы получаем:

$$y_{k+1} = y_k - \left(A^T \nabla^2 f(Ay_k) A\right)^{-1} A^T \nabla f(Ay_k)$$

Используя свойство обратной матрицы $(AB)^{-1} = B^{-1}A^{-1}$, это упрощается до:

$$y_{k+1} = y_k - A^{-1} \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$
$$Ay_{k+1} = Ay_k - \left(\nabla^2 f(Ay_k)\right)^{-1} \nabla f(Ay_k)$$

Таким образом, правило обновления для x выглядит так:

$$x_{k+1} = x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$$

Это показывает, что итерация метода Ньютона, не зависит от масштаба задачи. У градиентного спуска такого свойства нет!

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*

Минусы:

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность

Минусы:

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода

Минусы:

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода
- Сходимость можно сделать глобальной, если использовать демпфированный метод Ньютона (добавить процедуру линейного поиска и шага метода)

Минусы:

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода
- Сходимость можно сделать глобальной, если использовать демпфированный метод Ньютона (добавить процедуру линейного поиска и шага метода)

Минусы:

- Необходимо хранить (обратный) гессиан на каждой итерации: $\mathcal{O}(n^2)$ памяти

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода
- Сходимость можно сделать глобальной, если использовать демпфированный метод Ньютона (добавить процедуру линейного поиска и шага метода)

Минусы:

- Необходимо хранить (обратный) гессиан на каждой итерации: $\mathcal{O}(n^2)$ памяти
- Необходимо решать линейные системы: $\mathcal{O}(n^3)$ операций

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода
- Сходимость можно сделать глобальной, если использовать демпфированный метод Ньютона (добавить процедуру линейного поиска и шага метода)

Минусы:

- Необходимо хранить (обратный) гессиан на каждой итерации: $\mathcal{O}(n^2)$ памяти
- Необходимо решать линейные системы: $\mathcal{O}(n^3)$ операций
- Гессиан может быть вырожденным в x^*

Резюме

Плюсы:

- Квадратичная сходимость вблизи решения x^*
- Аффинная инвариантность
- Отсутствие параметров у метода
- Сходимость можно сделать глобальной, если использовать демпфированный метод Ньютона (добавить процедуру линейного поиска и шага метода)

Минусы:

- Необходимо хранить (обратный) гессиан на каждой итерации: $\mathcal{O}(n^2)$ памяти
- Необходимо решать линейные системы: $\mathcal{O}(n^3)$ операций
- Гессиан может быть вырожденным в x^*
- Гессиан может не быть положительно определенным $\to$ направление $-(f''(x))^{-1}f'(x)$ может не быть направлением спуска

Квазиньютоновские методы

Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации $f(x) \to \min\limits_{x \in \mathbb{R}^n}$ общий алгоритм итерационного метода записывается как:

$$x_{k+1} = x_k + \alpha_k d_k$$

Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации $f(x) \to \min\limits_{x \in \mathbb{R}^n}$ общий алгоритм итерационного метода записывается как:

$$x_{k+1} = x_k + \alpha_k d_k$$

В методе Ньютона направление d_k (направление Ньютона) устанавливается решением линейной системы на каждом шаге:

$$B_k d_k = -\nabla f(x_k), \quad B_k = \nabla^2 f(x_k)$$

Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации $f(x) \to \min\limits_{x \in \mathbb{R}^n}$ общий алгоритм итерационного метода записывается как:

$$x_{k+1} = x_k + \alpha_k d_k$$

В методе Ньютона направление d_k (направление Ньютона) устанавливается решением линейной системы на каждом шаге:

$$B_k d_k = -\nabla f(x_k), \quad B_k = \nabla^2 f(x_k)$$

т.е. на каждой итерации необходимо **вычислить** гессиан и градиент и **решить** линейную систему.

Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации $f(x) \to \min\limits_{x \in \mathbb{R}^n}$ общий алгоритм итерационного метода записывается как:

$$x_{k+1} = x_k + \alpha_k d_k$$

В методе Ньютона направление d_k (направление Ньютона) устанавливается решением линейной системы на каждом шаге:

$$B_k d_k = -\nabla f(x_k), \quad B_k = \nabla^2 f(x_k)$$

т.е. на каждой итерации необходимо **вычислить** гессиан и градиент и **решить** линейную систему.

Обратите внимание, что если мы возьмем единичную матрицу $B_k = I_n$ в качестве B_k на каждом шаге, мы получим точно метод градиентного спуска.

Общий алгоритм квазиньютоновских методов основан на выборе матрицы B_k так, чтобы она в некотором смысле стремилась к истинному значению гессиана $\nabla^2 f(x_k)$ при $k \to \infty$.

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Основная идея: Поскольку B_k уже содержит информацию о гессиане, используем подходящее обновление матрицы для формирования B_{k+1}.

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Основная идея: Поскольку B_k уже содержит информацию о гессиане, используем подходящее обновление матрицы для формирования B_{k+1}.

Разумное требование для B_{k+1} (вдохновленное методом секущих):

$$
\begin{aligned}
\nabla f(x_{k+1}) - \nabla f(x_k) &= B_{k+1}(x_{k+1} - x_k) \\
g_k &= B_{k+1} s_k
\end{aligned}
$$

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, \ldots$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Основная идея: Поскольку B_k уже содержит информацию о гессиане, используем подходящее обновление матрицы для формирования B_{k+1}.

Разумное требование для B_{k+1} (вдохновленное методом секущих):

$$
\begin{gathered}
\nabla f(x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k) \\
g_k = B_{k+1} s_k
\end{gathered}
$$

Помимо уравнения секущей, мы хотим:

- B_{k+1} симметричная

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Основная идея: Поскольку B_k уже содержит информацию о гессиане, используем подходящее обновление матрицы для формирования B_{k+1}.

Разумное требование для B_{k+1} (вдохновленное методом секущих):

$$
\begin{gathered}
\nabla f(x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k) \\
g_k = B_{k+1} s_k
\end{gathered}
$$

Помимо уравнения секущей, мы хотим:

- B_{k+1} симметричная
- B_{k+1} близка к B_k

Шаблон квазиньютоновского метода

Пусть $x_0 \in \mathbb{R}^n$, $B_0 \succ 0$. Для $k = 1, 2, 3, ...$, повторяем:

1. Решить $B_k d_k = -\nabla f(x_k)$
2. Обновить $x_{k+1} = x_k + \alpha_k d_k$
3. Вычислить B_{k+1} из B_k

Разные квазиньютоновские методы реализуют шаг 3 по-разному. Мы скоро увидим, что обычно мы можем вычислить $(B_{k+1})^{-1}$ из $(B_k)^{-1}$.

Основная идея: Поскольку B_k уже содержит информацию о гессиане, используем подходящее обновление матрицы для формирования B_{k+1}.

Разумное требование для B_{k+1} (вдохновленное методом секущих):

$$
\begin{aligned}
\nabla f(x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k) \\
g_k = B_{k+1} s_k
\end{aligned}
$$

Помимо уравнения секущей, мы хотим:

- B_{k+1} симметричная
- B_{k+1} близка к B_k
- $B_k \succ 0 \Rightarrow B_{k+1} \succ 0$

Симметричное одноранговое обновление

Попробуем обновление вида:

$$B_{k+1} = B_k + auu^T$$

Симметричное одноранговое обновление

Попробуем обновление вида:

$$B_{k+1} = B_k + auu^T$$

Уравнение секущей $B_{k+1}s_k = g_k$ дает:

$$(au^T s_k)u = g_k - B_k s_k$$

Симметричное одноранговое обновление

Попробуем обновление вида:

$$B_{k+1} = B_k + auu^T$$

Уравнение секущей $B_{k+1}s_k = g_k$ дает:

$$(au^T s_k)u = g_k - B_k s_k$$

Это верно только если u является параллельным $g_k - B_k s_k$. Положив $u = g_k - B_k s_k$, мы решаем уравнение,

$$a = \frac{1}{(g_k - B_k s_k)^T s_k},$$

Симметричное одноранговое обновление

Попробуем обновление вида:

$$B_{k+1} = B_k + auu^T$$

Уравнение секущей $B_{k+1}s_k = g_k$ дает:

$$(au^T s_k)u = g_k - B_k s_k$$

Это верно только если u является параллельным $g_k - B_k s_k$. Положив $u = g_k - B_k s_k$, мы решаем уравнение,

$$a = \frac{1}{(g_k - B_k s_k)^T s_k},$$

что приводит к

$$B_{k+1} = B_k + \frac{(g_k - B_k s_k)(g_k - B_k s_k)^T}{(g_k - B_k s_k)^T s_k}$$

Это называется симметричным одноранговым (SR1) обновлением или методом Бройдена.

Симметричное одноранговое обновление с инверсией

Как мы можем решить

$$B_{k+1} d_{k+1} = -\nabla f(x_{k+1}),$$

чтобы сделать следующий шаг? Помимо распространения B_k на B_{k+1}, давайте распространим инверсии, т.е. $C_k = B_k^{-1}$ на $C_{k+1} = (B_{k+1})^{-1}$.

Формула Шермана-Моррисона:

Формула Шермана-Моррисона утверждает:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$$

Таким образом, для SR1 обновления, обратная матрица также легко обновляется:

$$C_{k+1} = C_k + \frac{(s_k - C_k g_k)(s_k - C_k g_k)^T}{(s_k - C_k g_k)^T g_k}$$

В общем, SR1 прост и дешев, но у него есть ключевой недостаток: он не сохраняет положительную определенность.

Обновление Давидона-Флетчера-Пауэлла

Мы могли бы продолжить ту же идею для обновления обратной матрицы C:

$$C_{k+1} = C_k + auu^T + bvv^T.$$

Обновление Давидона-Флетчера-Пауэлла

Мы могли бы продолжить ту же идею для обновления обратной матрицы C:

$$C_{k+1} = C_k + auu^T + bvv^T.$$

Умножая на g_k, используя уравнение секущей $s_k = C_k g_k$ и решая для a, b, получаем:

$$C_{k+1} = C_k - \frac{C_k g_k g_k^T C_k}{g_k^T C_k g_k} + \frac{s_k s_k^T}{g_k^T s_k}$$

Это обновление Давидона-Флетчера-Пауэлла (DFP). Также дешево: $O(n^2)$, но сохраняет положительную определенность. Не так популярно, как BFGS.

Обновление Бройдена-Флетчера-Гольдштейна-Шенно

Попробуем теперь двухранговое обновление:

$$B_{k+1} = B_k + auu^T + bvv^T.$$

Обновление Бройдена-Флетчера-Гольдштейна-Шенно

Попробуем теперь двухранговое обновление:

$$B_{k+1} = B_k + auu^T + bvv^T.$$

Уравнение секущей $g_k = B_{k+1}s_k$ дает:

$$g_k - B_k s_k = (au^T s_k)u + (bv^T s_k)v$$

Обновление Бройдена-Флетчера-Гольдштейна-Шенно

Попробуем теперь двухранговое обновление:

$$B_{k+1} = B_k + auu^T + bvv^T.$$

Уравнение секущей $g_k = B_{k+1}s_k$ дает:

$$g_k - B_k s_k = (au^T s_k)u + (bv^T s_k)v$$

Положив $u = g_k$, $v = B_k s_k$ и решая для a, b, получаем:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{g_k g_k^T}{s_k^T g_k}$$

Это обновление Бройдена-Флетчера-Гольдштейна-Шенно (BFGS).

Обновление Бройдена-Флетчера-Гольдштейна-Шенно с инверсией

Формула Вудбери

Формула Вудбери, обобщение формулы Шермана-Моррисона, дается как:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Обновление Бройдена-Флетчера-Гольдштейна-Шенно с инверсией

Формула Вудбери

Формула Вудбери, обобщение формулы Шермана-Моррисона, дается как:

$$
(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}
$$

Примененная к нашему случаю, мы получаем двухранговое обновление на обратной матрице C:

$$
C_{k+1} = C_k + \frac{(s_k - C_k g_k)s_k^T}{g_k^T s_k} + \frac{s_k(s_k - C_k g_k)^T}{g_k^T s_k} - \frac{(s_k - C_k g_k)^T g_k}{(g_k^T s_k)^2} s_k s_k^T
$$

$$
C_{k+1} = \left(I - \frac{s_k g_k^T}{g_k^T s_k}\right) C_k \left(I - \frac{g_k s_k^T}{g_k^T s_k}\right) + \frac{s_k s_k^T}{g_k^T s_k}
$$

Эта формулировка обеспечивает, что обновление BFGS, оставаясь достаточно общим, сохраняет вычислительную эффективность и требует $O(n^2)$ операций. Важно, что обновление BFGS сохраняет положительную определенность: $B_k \succ 0 \Rightarrow B_{k+1} \succ 0.$ Эквивалентно, $C_k \succ 0 \Rightarrow C_{k+1} \succ 0$

Код

- Открыть в Colab

Код

- Открыть в Colab
- Сравнение квазиньютоновских методов

Код

- Открыть в Colab
- Сравнение квазиньютоновских методов
- Некоторые практические замечания о методе Ньютона

Бонус: доказательства

Квадратичная сходимость метода Ньютона

Theorem

Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:

$$
\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2
$$

Квадратичная сходимость метода Ньютона

Theorem

Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:

$$\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2$$

Доказательство

Квадратичная сходимость метода Ньютона

> **Theorem**
>
> Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:
>
> $$\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2$$

Доказательство

1. Мы будем использовать формулу Ньютона-Лейбница

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau$$

Квадратичная сходимость метода Ньютона

Theorem

Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:

$$\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu} \|x_k - x^*\|^2$$

Доказательство

1. Мы будем использовать формулу Ньютона-Лейбница

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*) d\tau$$

2. Мы будем отслеживать расстояние до решения

Квадратичная сходимость метода Ньютона

Theorem

Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:

$$
\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2
$$

Доказательство

1. Мы будем использовать формулу Ньютона-Лейбница

$$
\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*)d\tau
$$

2. Мы будем отслеживать расстояние до решения

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

Квадратичная сходимость метода Ньютона

> **Theorem**
>
> Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция на $\mathbb{R}^n$, для второй производной которой выполняются неравенства: $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$. Пусть также гессиан функции M-липшицев. Тогда метод Ньютона сходится локально к решению с квадратичной скоростью, т.е. при $\|x_0 - x^*\| < \frac{2\mu}{3M}$:
>
> $$\|x_{k+1} - x^*\| \leq \frac{3M}{2\mu}\|x_k - x^*\|^2$$

Доказательство

1. Мы будем использовать формулу Ньютона-Лейбница

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*)d\tau$$

2. Мы будем отслеживать расстояние до решения

$$x_{k+1} - x^* = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) - x^* = x_k - x^* - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) =$$
$$= x_k - x^* - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))(x_k - x^*)d\tau$$

Сходимость

3.

$$= \left(I - \left[\nabla^2 f(x_k) \right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

Сходимость

3.

$$= \left(I - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

$$= \left[\nabla^2 f(x_k)\right]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) =$$

Сходимость

3.

$$
\begin{aligned}
&= \left(I - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)(x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)(x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)\right)(x_k - x^*) =
\end{aligned}
$$

Сходимость

3.

$$
\begin{aligned}
&= \left(I - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)(x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)(x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*))d\tau\right)\right)(x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} G_k(x_k - x^*)
\end{aligned}
$$

Сходимость

3.

$$
\begin{aligned}
&= \left(I - \left[\nabla^2 f(x_k)\right]^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\nabla^2 f(x_k) - \int_0^1 \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) (x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left(\int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) \right) (x_k - x^*) = \\
&= \left[\nabla^2 f(x_k)\right]^{-1} G_k (x_k - x^*)
\end{aligned}
$$

4. Введём:

$$
G_k = \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right).
$$

Сходимость

5. Попробуем оценить размер G_k с помощью $r_k = \|x_k - x^*\|$:

Сходимость

5. Попробуем оценить размер G_k с помощью $r_k = \|x_k - x^*\|$:

$$\|G_k\| = \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) \right\| \leq$$

Сходимость

5. Попробуем оценить размер G_k с помощью $r_k = \|x_k - x^*\|$:

$$\|G_k\| = \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) \right\| \leq$$

$$\leq \int_0^1 \left\| \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) \right\| d\tau \leq$$ (Липшицевость гессиана)

Сходимость

5. Попробуем оценить размер G_k с помощью $r_k = \|x_k - x^*\|$:

$$\|G_k\| = \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau\right) \right\| \leq$$

$$\leq \int_0^1 \left\| \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) \right\| d\tau \leq \qquad \text{(Липшицевость гессиана)}$$

$$\leq \int_0^1 M\|x_k - x^* - \tau(x_k - x^*)\| d\tau = \int_0^1 M\|x_k - x^*\|(1-\tau) d\tau = \frac{r_k}{2} M,$$

Сходимость

5. Попробуем оценить размер G_k с помощью $r_k = \|x_k - x^*\|$:

$$\|G_k\| = \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) d\tau \right) \right\| \leq$$

$$\leq \int_0^1 \left\| \nabla^2 f(x_k) - \nabla^2 f(x^* + \tau(x_k - x^*)) \right\| d\tau \leq \qquad \text{(Липшицевость гессиана)}$$

$$\leq \int_0^1 M \|x_k - x^* - \tau(x_k - x^*)\| d\tau = \int_0^1 M \|x_k - x^*\| (1 - \tau) d\tau = \frac{r_k}{2} M,$$

6. Получаем:

$$r_{k+1} \leq \left\| \left[\nabla^2 f(x_k) \right]^{-1} \right\| \cdot \frac{r_k}{2} M \cdot r_k$$

   и нам нужно оценить норму обратного гессиана

Сходимость

7. Из липшицевости и симметричности гессиана:

Сходимость

7. Из липшицевости и симметричности гессиана:

$$\nabla^2 f(x_k) - \nabla^2 f(x^*) \succeq -Mr_k I_n$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$
\begin{aligned}
\nabla^2 f(x_k) - \nabla^2 f(x^*) &\succeq -Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - Mr_k I_n
\end{aligned}
$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$
\begin{aligned}
\nabla^2 f(x_k) - \nabla^2 f(x^*) &\succeq -Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \mu I_n - Mr_k I_n
\end{aligned}
$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$
\begin{aligned}
\nabla^2 f(x_k) - \nabla^2 f(x^*) &\succeq -Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \mu I_n - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq (\mu - Mr_k) I_n
\end{aligned}
$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$
\begin{aligned}
\nabla^2 f(x_k) - \nabla^2 f(x^*) &\succeq -Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \mu I_n - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq (\mu - Mr_k) I_n
\end{aligned}
$$

8. Из сильной выпуклости следует, что $\nabla^2 f(x_k) \succ 0$, т.е. $r_k < \frac{\mu}{M}$.

$$
\begin{aligned}
\left\| \left[\nabla^2 f(x_k) \right]^{-1} \right\| &\leq (\mu - Mr_k)^{-1} \\
r_{k+1} &\leq \frac{r_k^2 M}{2(\mu - Mr_k)}
\end{aligned}
$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$\nabla^2 f(x_k) - \nabla^2 f(x^*) \succeq -Mr_k I_n$$
$$\nabla^2 f(x_k) \succeq \nabla^2 f(x^*) - Mr_k I_n$$
$$\nabla^2 f(x_k) \succeq \mu I_n - Mr_k I_n$$
$$\nabla^2 f(x_k) \succeq (\mu - Mr_k) I_n$$

8. Из сильной выпуклости следует, что $\nabla^2 f(x_k) \succ 0$, т.е. $r_k < \frac{\mu}{M}$.

$$\left\| \left[\nabla^2 f(x_k) \right]^{-1} \right\| \leq (\mu - Mr_k)^{-1}$$
$$r_{k+1} \leq \frac{r_k^2 M}{2(\mu - Mr_k)}$$

9. Потребуем, чтобы верхняя оценка на r_{k+1} была меньше r_k, учитывая, что $0 < r_k < \frac{\mu}{M}$:

$$\frac{r_k^2 M}{2(\mu - Mr_k)} < r_k$$
$$\frac{M}{2(\mu - Mr_k)} r_k < 1$$
$$Mr_k < 2(\mu - Mr_k)$$
$$3Mr_k < 2\mu$$
$$r_k < \frac{2\mu}{3M}$$

Сходимость

7. Из липшицевости и симметричности гессиана:

$$
\begin{aligned}
\nabla^2 f(x_k) - \nabla^2 f(x^*) &\succeq -Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \nabla^2 f(x^*) - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq \mu I_n - Mr_k I_n \\
\nabla^2 f(x_k) &\succeq (\mu - Mr_k) I_n
\end{aligned}
$$

8. Из сильной выпуклости следует, что $\nabla^2 f(x_k) \succ 0$, т.е. $r_k < \frac{\mu}{M}$.

$$
\left\| \left[\nabla^2 f(x_k) \right]^{-1} \right\| \leq (\mu - Mr_k)^{-1}
$$

$$
r_{k+1} \leq \frac{r_k^2 M}{2(\mu - Mr_k)}
$$

9. Потребуем, чтобы верхняя оценка на r_{k+1} была меньше r_k, учитывая, что $0 < r_k < \frac{\mu}{M}$:

$$
\begin{aligned}
\frac{r_k^2 M}{2(\mu - Mr_k)} &< r_k \\
\frac{M}{2(\mu - Mr_k)} r_k &< 1 \\
Mr_k &< 2(\mu - Mr_k) \\
3Mr_k &< 2\mu \\
r_k &< \frac{2\mu}{3M}
\end{aligned}
$$

10. Возвращаясь к оценке невязки на $k+1$-ой итерации, получаем:

$$
r_{k+1} \leq \frac{r_k^2 M}{2(\mu - Mr_k)} < \frac{3Mr_k^2}{2\mu}
$$

Таким образом, мы получили важный результат: метод Ньютона для функции с липшицевым положительно определённым гессианом сходится **квадратично** вблизи решения.

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg}\min_{x \in B_\varepsilon(x_0)} f(x)$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \arg\min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$
x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)
$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$
s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}
$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$
d(x, x_0) = (x - x_0)^\top A(x - x_0)
$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$d(x, x_0) = (x - x_0)^\top A (x - x_0)$$

Далее рассмотрим первый порядок аппроксимации функции $f(x)$ в окрестности точки x_0:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f(x_0)^\top \delta x \tag{1}$$

Теперь мы можем сформулировать задачу нахождения s, как это было сказано выше.

$$\begin{aligned} &\min_{\delta x \in \mathbb{R}^{\ltimes}} f(x_0 + \delta x) \\ \text{s.t. } &\delta x^\top A \delta x = \varepsilon^2 \end{aligned}$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$d(x, x_0) = (x - x_0)^\top A (x - x_0)$$

Далее рассмотрим первый порядок аппроксимации функции $f(x)$ в окрестности точки x_0:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f(x_0)^\top \delta x \tag{1}$$

Теперь мы можем сформулировать задачу нахождения s, как это было сказано выше.

$$\begin{aligned} &\min_{\delta x \in \mathbb{R}^n} f(x_0 + \delta x) \\ &\text{s.t. } \delta x^\top A \delta x = \varepsilon^2 \end{aligned}$$

Используя уравнение 1, получаем:

$$\begin{aligned} &\min_{\delta x \in \mathbb{R}^n} \nabla f(x_0)^\top \delta x \\ &\text{s.t. } \delta x^\top A \delta x = \varepsilon^2 \end{aligned}$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$d(x, x_0) = (x - x_0)^\top A (x - x_0)$$

Далее рассмотрим первый порядок аппроксимации функции $f(x)$ в окрестности точки x_0:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f(x_0)^\top \delta x \qquad (1)$$

Теперь мы можем сформулировать задачу нахождения s, как это было сказано выше.

$$\min_{\delta x \in \mathbb{R}^\ltimes} f(x_0 + \delta x)$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя уравнение 1, получаем:

$$\min_{\delta x \in \mathbb{R}^\ltimes} \nabla f(x_0)^\top \delta x$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя метод множителей Лагранжа:

$$\delta x = -\frac{2\varepsilon^2}{\nabla f(x_0)^\top A^{-1} \nabla f(x_0)} A^{-1} \nabla f$$

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$d(x, x_0) = (x - x_0)^\top A (x - x_0)$$

Далее рассмотрим первый порядок аппроксимации функции $f(x)$ в окрестности точки x_0:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f(x_0)^\top \delta x \tag{1}$$

Теперь мы можем сформулировать задачу нахождения s, как это было сказано выше.

$$\min_{\delta x \in \mathbb{R}^{\ltimes}} f(x_0 + \delta x)$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя уравнение 1, получаем:

$$\min_{\delta x \in \mathbb{R}^{\ltimes}} \nabla f(x_0)^\top \delta x$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя метод множителей Лагранжа:

$$\delta x = -\frac{2\varepsilon^2}{\nabla f(x_0)^\top A^{-1} \nabla f(x_0)} A^{-1} \nabla f$$

Новое направление наискорейшего спуска : $-A^{-1} \nabla f(x_0)$.

Бонус: Идея методов адаптивной метрики

Пусть дана функция $f(x)$ и точка x_0. Определим $B_\varepsilon(x_0) = \{x \in \mathbb{R}^n : d(x, x_0) = \varepsilon^2\}$ как множество точек с расстоянием ε до x_0. Здесь мы предполагаем существование функции расстояния $d(x, x_0)$.

$$x^* = \text{arg} \min_{x \in B_\varepsilon(x_0)} f(x)$$

Далее, мы можем определить другое направление *наискорейшего спуска* в терминах минимизатора функции на сфере:

$$s = \lim_{\varepsilon \to 0} \frac{x^* - x_0}{\varepsilon}$$

Предположим, что расстояние локально определяется некоторой метрикой A:

$$d(x, x_0) = (x - x_0)^\top A (x - x_0)$$

Далее рассмотрим первый порядок аппроксимации функции $f(x)$ в окрестности точки x_0:

$$f(x_0 + \delta x) \approx f(x_0) + \nabla f(x_0)^\top \delta x \tag{1}$$

Теперь мы можем сформулировать задачу нахождения s, как это было сказано выше.

$$\min_{\delta x \in \mathbb{R}^{\ltimes}} f(x_0 + \delta x)$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя уравнение 1, получаем:

$$\min_{\delta x \in \mathbb{R}^{\ltimes}} \nabla f(x_0)^\top \delta x$$
$$\text{s.t. } \delta x^\top A \delta x = \varepsilon^2$$

Используя метод множителей Лагранжа:

$$\delta x = -\frac{2\varepsilon^2}{\nabla f(x_0)^\top A^{-1} \nabla f(x_0)} A^{-1} \nabla f$$

Новое направление наискорейшего спуска : $-A^{-1}\nabla f(x_0)$. Действительно, если пространство изотропно и $A = I$, мы сразу получаем формулу градиентного спуска, в то время как метод Ньютона использует локальный гессиан как матрицу метрик.