

# Метод Ньютона и квазиньютоновские методы

МЕТОДЫ ВЫПУКЛОЙ ОПТИМИЗАЦИИ

НЕДЕЛЯ 10

Даня Меркулов



# **Метод Ньютона и квазиньютоновские методы**

## **Семинар**

**Оптимизация для всех! ЦУ**



# **Воспоминания с лекции**

# Идея метода Ньютона для нахождения корней функции



Рассмотрим функцию  $\varphi(x) : \mathbb{R} \rightarrow \mathbb{R}$ . Мы хотим найти корень уравнения  $\varphi(x) = 0$ .

Основная идея заключается в том, чтобы построить линейное приближение в точке  $x_k$  и найти его корень, который будет новой точкой итерации:

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

Мы получаем итерационную схему:

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

Теперь, если мы рассмотрим  $\varphi(x) \equiv \nabla f(x)$ , это станет методом оптимизации Ньютона:

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

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

1. Найдем производную:

$$\varphi'(t) = -\frac{t^2}{(1+t^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+t^2}}$$

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

1. Найдем производную:

$$\varphi'(t) = -\frac{t^2}{(1+t^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+t^2}}$$

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

1. Найдем производную:

$$\varphi'(t) = -\frac{t^2}{(1+t^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+t^2}}$$

2. Тогда итерация метода принимает вид:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)} = x_k - x_k(x_k^2 + 1) = -x_k^3$$

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

1. Найдем производную:

$$\varphi'(t) = -\frac{t^2}{(1+t^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+t^2}}$$

2. Тогда итерация метода принимает вид:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)} = x_k - x_k(x_k^2 + 1) = -x_k^3$$

# Пример метода линеаризации Ньютона

## Question

Примените метод Ньютона для нахождения корня уравнения  $\varphi(t) = 0$  и определите область сходимости:

$$\varphi(t) = \frac{t}{\sqrt{1+t^2}}$$

1. Найдем производную:

$$\varphi'(t) = -\frac{t^2}{(1+t^2)^{\frac{3}{2}}} + \frac{1}{\sqrt{1+t^2}}$$

2. Тогда итерация метода принимает вид:

$$x_{k+1} = x_k - \frac{\varphi(x_k)}{\varphi'(x_k)} = x_k - x_k(x_k^2 + 1) = -x_k^3$$

Легко видеть, что метод сходится только если  $|x_0| < 1$ , подчеркивая **локальный** характер метода Ньютона.

# 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  $x_k$ :

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

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

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \right\}$$

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

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

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

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

# 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  $x_k$ :

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

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

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \right\}$$

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

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

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

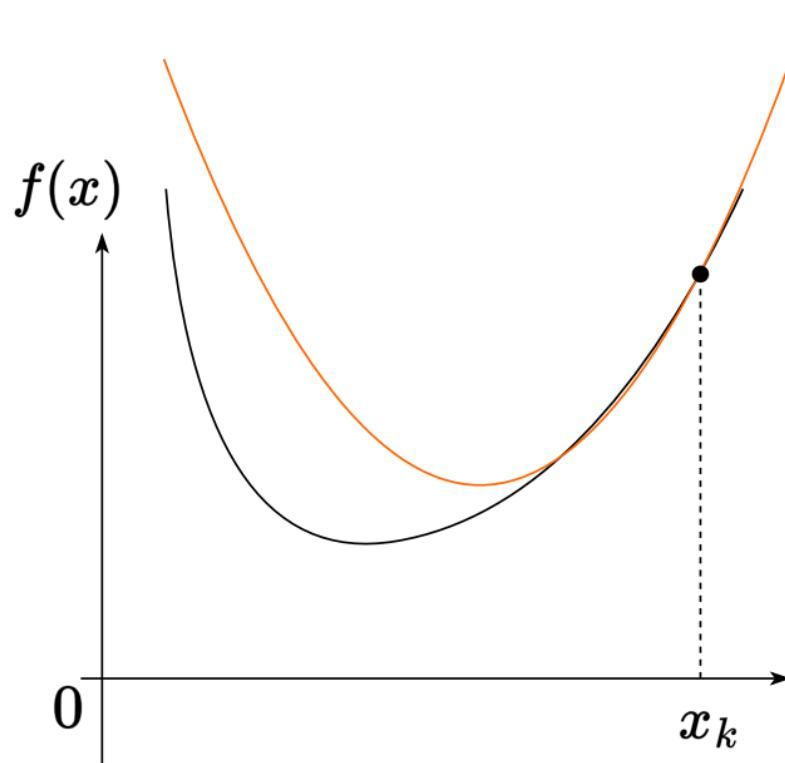
$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

Обратите внимание на ограничения, связанные с необходимостью невырожденности (для существования метода) и положительной определенности (для гарантии сходимости) гессиана.

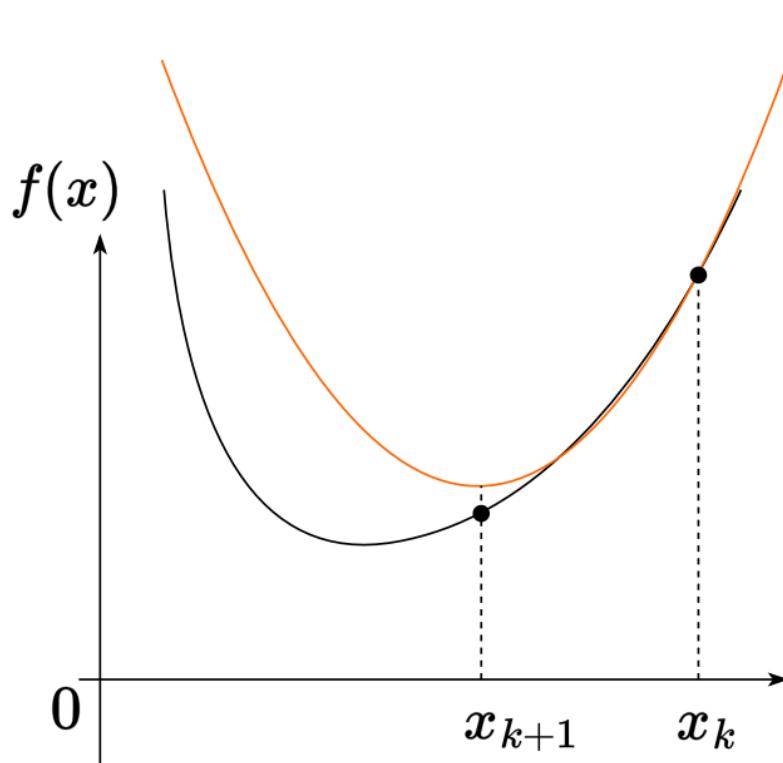
# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



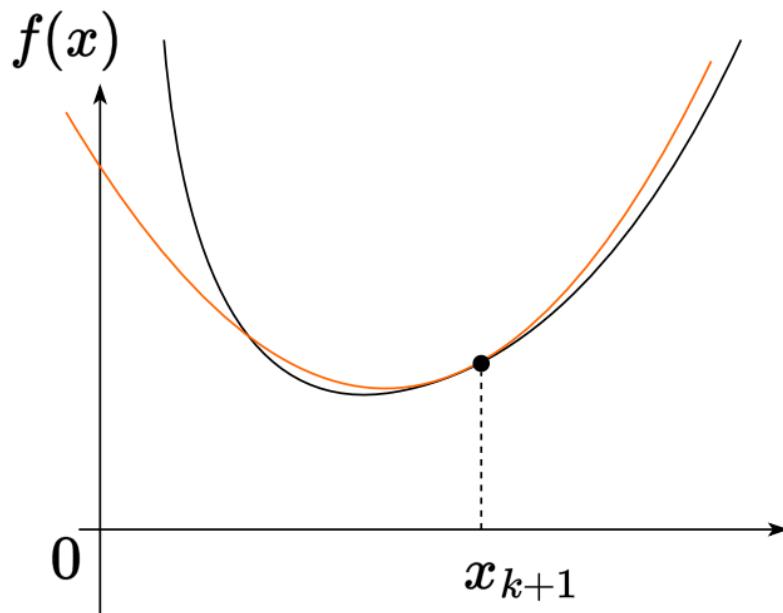
# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



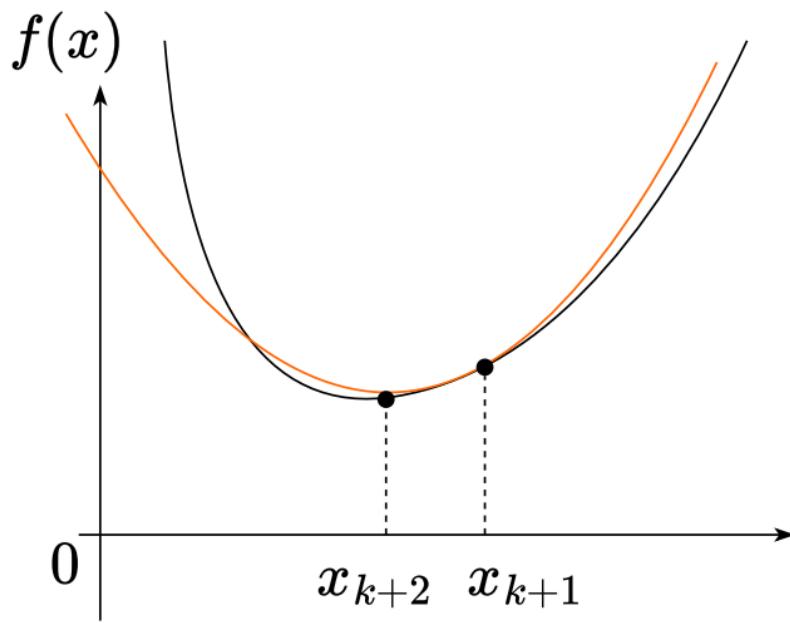
# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



# Метод Ньютона как оптимизация локальной квадратичной аппроксимации



# Метод Ньютона vs градиентный спуск



Рисунок 7. Функция потерь изображена черным, аппроксимация в виде пунктирной красной линии

Градиентный спуск  $\equiv$  линейное приближение

Метод Ньютона  $\equiv$  квадратичное приближение

# Сходимость

## Theorem

Пусть  $f(x)$  — сильно выпуклая дважды непрерывно дифференцируемая функция на  $\mathbb{R}^n$ , для второй производной которой выполняются неравенства:  $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$ . Тогда метод Ньютона с постоянным шагом

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

локально сходится к решению с суперлинейной скоростью. Если, в дополнение, гессиан является  $M$ -липшицевым, то этот метод локально сходится к  $x^*$  с квадратичной скоростью:

$$\|x_{k+1} - x^*\|_2 \leq \frac{M \|x_k - x^*\|_2^2}{2(\mu - M \|x_k - x^*\|_2)}$$

# Сходимость

## Theorem

Пусть  $f(x)$  — сильно выпуклая дважды непрерывно дифференцируемая функция на  $\mathbb{R}^n$ , для второй производной которой выполняются неравенства:  $\mu I_n \preceq \nabla^2 f(x) \preceq L I_n$ . Тогда метод Ньютона с постоянным шагом

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

локально сходится к решению с суперлинейной скоростью. Если, в дополнение, гессиан является  $M$ -липшицевым, то этот метод локально сходится к  $x^*$  с квадратичной скоростью:

$$\|x_{k+1} - x^*\|_2 \leq \frac{M \|x_k - x^*\|_2^2}{2(\mu - M \|x_k - x^*\|_2)}$$

“Локальная сходимость” означает, что скорость сходимости, описанная выше, гарантируется только если начальная точка достаточно близка к точке минимума, в частности  $\|x_0 - x^*\| < \frac{2\mu}{3M}$

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .
2. Рассмотрим квадратичное приближение:

$$\begin{aligned} g(y + u) &\approx g(y) + \langle g'(y), u \rangle + \frac{1}{2} u^\top g''(y) u \rightarrow \min_u \\ u^* &= - (g''(y))^{-1} g'(y) \quad y_{k+1} = y_k - (g''(y_k))^{-1} g'(y_k) \end{aligned}$$

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .
2. Рассмотрим квадратичное приближение:

$$\begin{aligned} g(y + u) &\approx g(y) + \langle g'(y), u \rangle + \frac{1}{2} u^\top g''(y) u \rightarrow \min_u \\ u^* &= - (g''(y))^{-1} g'(y) \quad y_{k+1} = y_k - (g''(y_k))^{-1} g'(y_k) \end{aligned}$$

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .
2. Рассмотрим квадратичное приближение:

$$\begin{aligned} g(y + u) &\approx g(y) + \langle g'(y), u \rangle + \frac{1}{2} u^\top g''(y) u \rightarrow \min_u \\ u^* &= - (g''(y))^{-1} g'(y) \quad y_{k+1} = y_k - (g''(y_k))^{-1} g'(y_k) \end{aligned}$$

3. Подставим явные выражения для  $g''(y_k)$ ,  $g'(y_k)$ :

$$y_{k+1} = y_k - (A^\top f''(Ay_k) A)^{-1} A^\top f'(Ay_k) = y_k - A^{-1} (f''(Ay_k))^{-1} f'(Ay_k)$$

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

1. Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .
2. Рассмотрим квадратичное приближение:

$$\begin{aligned} g(y + u) &\approx g(y) + \langle g'(y), u \rangle + \frac{1}{2} u^\top g''(y) u \rightarrow \min_u \\ u^* &= - (g''(y))^{-1} g'(y) \quad y_{k+1} = y_k - (g''(y_k))^{-1} g'(y_k) \end{aligned}$$

3. Подставим явные выражения для  $g''(y_k)$ ,  $g'(y_k)$ :

$$y_{k+1} = y_k - (A^\top f''(Ay_k) A)^{-1} A^\top f'(Ay_k) = y_k - A^{-1} (f''(Ay_k))^{-1} f'(Ay_k)$$

# Аффинная инвариантность

## Question

Рассмотрим функцию  $f(x)$  и преобразование с обратимой матрицей  $A$ . Давайте выясним, как изменится итерационный шаг метода Ньютона после применения преобразования.

- Пусть  $x = Ay$  и  $g(y) = f(Ay)$ .
- Рассмотрим квадратичное приближение:

$$\begin{aligned} g(y + u) &\approx g(y) + \langle g'(y), u \rangle + \frac{1}{2}u^\top g''(y)u \rightarrow \min_u \\ u^* &= -(g''(y))^{-1} g'(y) \quad y_{k+1} = y_k - (g''(y_k))^{-1} g'(y_k) \end{aligned}$$

- Подставим явные выражения для  $g''(y_k)$ ,  $g'(y_k)$ :

$$y_{k+1} = y_k - (A^\top f''(Ay_k) A)^{-1} A^\top f'(Ay_k) = y_k - A^{-1} (f''(Ay_k))^{-1} f'(Ay_k)$$

- Таким образом, шаг метода преобразуется линейным преобразованием **таким же образом**, как и координаты:

$$Ay_{k+1} = Ay_k - (f''(Ay_k))^{-1} f'(Ay_k) \quad x_{k+1} = x_k - (f''(x_k))^{-1} f'(x_k)$$

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости
- необходимо хранить гессиан на каждой итерации:  $\mathcal{O}(n^2)$  памяти

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

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

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости
- необходимо хранить гессиан на каждой итерации:  $\mathcal{O}(n^2)$  памяти
- необходимо решать линейные системы:  $\mathcal{O}(n^3)$  операций
- гессиан может быть вырожден

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости
- необходимо хранить гессиан на каждой итерации:  $\mathcal{O}(n^2)$  памяти
- необходимо решать линейные системы:  $\mathcal{O}(n^3)$  операций
- гессиан может быть вырожден
- гессиан может не быть положительно определен → направление  $-(f''(x))^{-1} f'(x)$  может не быть убывающим 

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости
- необходимо хранить гессиан на каждой итерации:  $\mathcal{O}(n^2)$  памяти
- необходимо решать линейные системы:  $\mathcal{O}(n^3)$  операций
- гессиан может быть вырожден
- гессиан может не быть положительно определен → направление  $-(f''(x))^{-1} f'(x)$  может не быть убывающим

# Сводка метода Ньютона

## Pros

- квадратичная сходимость вблизи решения
- высокая точность полученного решения
- аффинная инвариантность

## Cons

- отсутствие глобальной сходимости
- необходимо хранить гессиан на каждой итерации:  $\mathcal{O}(n^2)$  памяти
- необходимо решать линейные системы:  $\mathcal{O}(n^3)$  операций
- гессиан может быть вырожден
- гессиан может не быть положительно определен → направление  $-(f''(x))^{-1} f'(x)$  может не быть убывающим ⚡

Метод кубической регуляризации Ньютона и квазиньютоновские методы частично решают эти проблемы!

# Метод кубической регуляризации Ньютона

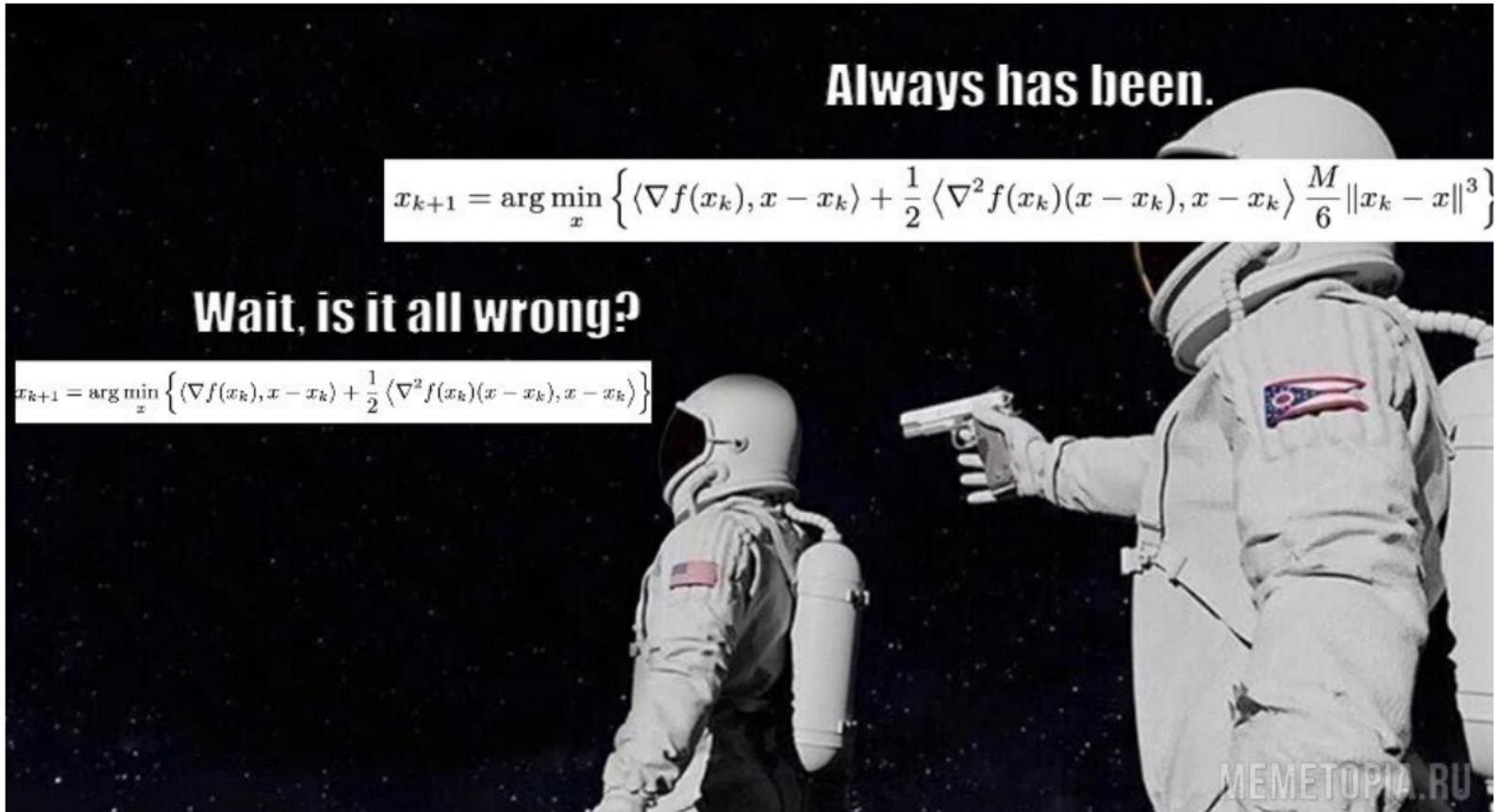
# Небольшой базовый материал.

Always has been.

$$x_{k+1} = \arg \min_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \frac{M}{6} \|x_k - x\|^3 \right\}$$

Wait, is it all wrong?

$$x_{k+1} = \arg \min_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle \right\}$$



# Интуитивно о том, как улучшить метод Ньютона

## 💡 Gradient Descent recap

Пусть  $f$  имеет  $L$ -липшицевый градиент, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Тогда каждый шаг градиентного спуска для функции  $f$  с  $L$ -липшицевым градиентом является минимизацией мажорирующей параболоиды:

$$\begin{aligned} x_{k+1} &= \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 \right\} \\ &= x_k - \frac{1}{L} \nabla f(x_k). \end{aligned}$$

# Интуитивно о том, как улучшить метод Ньютона

## Gradient Descent recap

Пусть  $f$  имеет  $L$ -липшицевый градиент, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Тогда каждый шаг градиентного спуска для функции  $f$  с  $L$ -липшицевым градиентом является минимизацией мажорирующей параболоиды:

$$\begin{aligned} x_{k+1} &= \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|^2 \right\} \\ &= x_k - \frac{1}{L} \nabla f(x_k). \end{aligned}$$

Но если функция  $f$  имеет  $M$ -липшицевый гессиан, то легко показать, что

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

**Что если мы используем ту же логику, что и в градиентном спуске для функции с  $M$ -липшицевым гессианом?**

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

**Вопрос**

Какие проблемы вы видите в (1)?

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

- Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

1. Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.
2. Подзадача внутри (1) может быть невыпуклой.

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

1. Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.
2. Подзадача внутри (1) может быть невыпуклой.

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

1. Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.
2. Подзадача внутри (1) может быть невыпуклой.

## 💡 Solutions

1. Мы можем использовать численные методы с быстрой сходимостью

---

<sup>a</sup>Nesterov, Y. (2018). Lectures on convex optimization. Springer.

<sup>b</sup>Nesterov, Y. (2021). Implementable tensor methods in unconstrained convex optimization. Mathematical Programming.

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

1. Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.
2. Подзадача внутри (1) может быть невыпуклой.

## 💡 Solutions

1. Мы можем использовать численные методы с быстрой сходимостью
2. Подзадача эквивалентна задаче одномерной оптимизации с выпуклыми ограничениями.<sup>a</sup>

---

<sup>a</sup>Nesterov, Y. (2018). Lectures on convex optimization. Springer.

<sup>b</sup>Nesterov, Y. (2021). Implementable tensor methods in unconstrained convex optimization. Mathematical Programming.

# Метод кубической регуляризации Ньютона

Пусть  $f$  имеет  $M$ -липшицевый гессиан, тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \frac{M}{6} \|y - x\|^3.$$

Минимизируя правую часть этого неравенства, мы приходим к методу кубической регуляризации Ньютона

$$x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M}{6} \|x - x_k\|^3 \right\}. \quad (1)$$

## ! Challenges

1. Мы не можем получить явные выражения для  $x_{k+1}$  (без  $\operatorname{argmin}$ ) из (1) как мы могли в градиентном спуске.
2. Подзадача внутри (1) может быть невыпуклой.

## 💡 Solutions

1. Мы можем использовать численные методы с быстрой сходимостью
2. Подзадача эквивалентна задаче одномерной оптимизации с выпуклыми ограничениями.<sup>a</sup>
3. Подзадачу можно сделать выпуклой с помощью правильного коэффициента регуляризации.<sup>b</sup>

<sup>a</sup>Nesterov, Y. (2018). Lectures on convex optimization. Springer.

<sup>b</sup>Nesterov, Y. (2021). Implementable tensor methods in unconstrained convex optimization. Mathematical Programming.

# Сходимость<sup>1</sup>

## Theorem

Пусть  $f(x)$  —  $\mu$ -сильно выпуклая функция с  $M$ -липшицевым гессианом. Тогда, метод кубической регуляризации Ньютона (1) сходится глобально суперлинейно как

$$f(x_{k+1}) - f^* \leq \gamma_k(f(x_k) - f^*), \quad \gamma_k \rightarrow 0.$$

---

<sup>1</sup>Kamzolov, D., et al. (2024). Optami: Global superlinear convergence of high-order methods. Accepted to ICLR 2025.

# Квазиньютоновские методы

# Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации  $f(x) \rightarrow \min_{x \in \mathbb{R}^n}$  общая схема итерационного метода записывается как:

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

# Интуиция квазиньютоновских методов

Для классической задачи безусловной оптимизации  $f(x) \rightarrow \min_{x \in \mathbb{R}^n}$  общий схема 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  $f(x) \rightarrow \min_{x \in \mathbb{R}^n}$  общий схема 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  $f(x) \rightarrow \min_{x \in \mathbb{R}^n}$  общая схема 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 одну матрицу  $B_k = I_n$  как  $B_k$  на каждом шаге, мы точно получим метод градиентного спуска.

Общая схема квазиньютоновских методов основана на выборе матрицы  $B_k$  так, чтобы она в некотором смысле стремилась к истинному значению гессиана  $\nabla^2 f(x_k)$  при  $k \rightarrow \infty$ .

# Шаблон квазиньютоновского метода

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

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$

# Шаблон квазиньютоновского метода

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

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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  $k = 1, 2, 3, \dots$ , повторяем:

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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  $k = 1, 2, 3, \dots$ , повторяем:

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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  $k = 1, 2, 3, \dots$ , повторяем:

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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  $k = 1, 2, 3, \dots$ , повторяем:

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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  $k = 1, 2, 3, \dots$ , повторяем:

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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):

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

# Шаблон квазиньютоновского метода

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

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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):

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

Помимо уравнения секущей, мы хотим:

- $B_{k+1}$  должна быть симметричной

# Шаблон квазиньютоновского метода

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

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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):

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

Помимо уравнения секущей, мы хотим:

- $B_{k+1}$  должна быть симметричной
- $B_{k+1}$  должна быть "близка" к  $B_k$

# Шаблон квазиньютоновского метода

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

1. Найти  $d_k : B_k d_k = -\nabla f(x_k)$
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. Как мы увидим, часто 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):

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

Помимо уравнения секущей, мы хотим:

- $B_{k+1}$  должна быть симметричной
- $B_{k+1}$  должна быть "близка" к  $B_k$
- $B_k \succ 0 \Rightarrow B_{k+1} \succ 0$

# Задача 1: Симметричное одноранговое (SR1) обновление

Попробуем обновление с матрицей единичного ранга:

$$B_{k+1} = B_k + a u u^T$$

 Question

Какие  $a$  и  $u$  мы можем выбрать? Как будет выглядеть обновление  $B_{k+1}$ ?

# Задача 1: Симметричное одноранговое (SR1) обновление

Попробуем обновление с матрицей единичного ранга:

$$B_{k+1} = B_k + a u u^T$$

 Question

Какие  $a$  и  $u$  мы можем выбрать? Как будет выглядеть обновление  $B_{k+1}$ ?

# Сходимость SR1

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}$$

называется симметричным одноранговым (SR1) обновлением или методом Бройдена.

## Theorem

Пусть

- $f$  — дважды непрерывно дифференцируемая функция, имеет единственную стационарную точку  $x^*$ ,

Тогда в SR1  $x_k \rightarrow x^*$  суперлинейно.

# Сходимость SR1

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}$$

называется симметричным одноранговым (SR1) обновлением или методом Бройдена.

## Theorem

Пусть

- $f$  — дважды непрерывно дифференцируемая функция, имеет единственную стационарную точку  $x^*$ ,
- $0 \succ \nabla^2 f(x^2), \nabla^2 f(x)$  — липшицева в окрестности  $x^*$ ,

Тогда в SR1  $x_k \rightarrow x^*$  суперлинейно.

# Сходимость SR1

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}$$

называется симметричным одноранговым (SR1) обновлением или методом Бройдена.

## Theorem

Пусть

- $f$  — дважды непрерывно дифференцируемая функция, имеет единственную стационарную точку  $x^*$ ,
- $0 \succ \nabla^2 f(x^2), \nabla^2 f(x)$  — липшицева в окрестности  $x^*$ ,
- последовательность матриц  $\{B_k\}$  ограничена в норме,

Тогда в SR1  $x_k \rightarrow x^*$  суперлинейно.

# Сходимость SR1

$$B_{k+1} = B_k + \frac{(\Delta y_k - B_k d_k)(\Delta y_k - B_k d_k)^T}{(\Delta y_k - B_k d_k)^T d_k}$$

называется симметричным одноранговым (SR1) обновлением или методом Бройдена.

## Theorem

Пусть

- $f$  — дважды непрерывно дифференцируемая функция, имеет единственную стационарную точку  $x^*$ ,
- $0 \succ \nabla^2 f(x^2), \nabla^2 f(x)$  — липшицева в окрестности  $x^*$ ,
- последовательность матриц  $\{B_k\}$  ограничена в норме,
- $|(\Delta y_k - B_k d_k)^T d_k| \geq r \|d_k\| \|\Delta y_k - B_k d_k\|$ ,  $0 < r \ll 1$ .

Тогда в SR1  $x_k \rightarrow x^*$  суперлинейно.

# SR1 с обратным обновлением

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 обратные, т.е.  $C_k = B_k^{-1}$  на  $C_{k+1} = (B_{k+1})^{-1}$ .

[Формула Шермана-Моррисона:](#)

Формула Шермана-Моррисона утверждает:

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

# SR1 с обратным обновлением

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 обратные, т.е.  $C_k = B_k^{-1}$  на  $C_{k+1} = (B_{k+1})^{-1}$ .

[Формула Шермана-Моррисона:](#)

Формула Шермана-Моррисона утверждает:

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

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

$$C_{k+1} = C_k + \frac{(d_k - C_k \Delta y_k)(d_k - C_k \Delta y_k)^T}{(d_k - C_k \Delta y_k)^T \Delta y_k}$$

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

## Задача 2: Обновление

### Broyden-Fletcher-Goldfarb-Shanno (BFGS)

Попробуем теперь двухранговое обновление:

$$B_{k+1} = B_k + a u u^T + b v v^T.$$

#### Question

Какие  $a$ ,  $u$ ,  $b$  и  $v$  мы можем выбрать? Как будет выглядеть обновление  $B_{k+1}$ ?

# Сходимость BFGS

$$B_{k+1} = B_k - \frac{B_k d_k d_k^T B_k}{d_k^T B_k d_k} + \frac{\Delta y_k \Delta y_k^T}{d_k^T \Delta y_k}$$

называется обновлением Бройдена-Флетчера-Гольдфарба-Шанно (BFGS).

## Theorem

Пусть  $f(x)$  — дважды непрерывно дифференцируемая функция, имеет липшицевый гессиан в  $x^*$  и дополнительно  $\sum_{k=1}^{\infty} \|x_k - x^*\| \leq \infty$ . Тогда в BFGS  $x_k \rightarrow x^*$  суперлинейно.

# BFGS обновление с инверсией

## Формула Вудбери

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

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

# BFGS обновление с инверсией

## Формула Вудбери

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

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

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

$$\begin{aligned} C_{k+1} &= C_k + \frac{(d_k - C_k \Delta y_k) d_k^T}{\Delta y_k^T d_k} + \frac{d_k (d_k - C_k \Delta y_k)^T}{\Delta y_k^T d_k} - \frac{(d_k - C_k \Delta y_k)^T \Delta y_k}{(\Delta y_k^T d_k)^2} d_k d_k^T \\ C_{k+1} &= \left( I - \frac{d_k \Delta y_k^T}{\Delta y_k^T d_k} \right) C_k \left( I - \frac{\Delta y_k d_k^T}{\Delta y_k^T d_k} \right) + \frac{d_k d_k^T}{\Delta y_k^T d_k} \end{aligned}$$

Эта формулировка гарантирует, что обновление BFGS, хоть и объемное, остается вычислительно эффективным, требуя  $O(n^2)$  операций. Важно, что обновление BFGS сохраняет положительную определенность. Помните, это означает  $B_k \succ 0 \Rightarrow B_{k+1} \succ 0$ . Эквивалентно,  $C_k \succ 0 \Rightarrow C_{k+1} \succ 0$

# Основная идея L-BFGS

- L-BFGS не хранит полную матрицу  $B_k$  ( $C_k$ ), вместо этого она хранит две последовательности векторов длины  $m : m < n$

# Основная идея L-BFGS

- L-BFGS не хранит полную матрицу  $B_k$  ( $C_k$ ), вместо этого она хранит две последовательности векторов длины  $m : m < n$
- память уменьшается с  $O(n^2)$  до  $O(mn)$ , делая его более подходящим для высокоразмерных задач

# Вычислительные эксперименты

# Вычислительные эксперименты

- Вычислительные эксперименты для квазиньютоновских методов, CG и GD 

# Вычислительные эксперименты

- Вычислительные эксперименты для квазиньютоновских методов, CG и GD 
- Вычислительные эксперименты для методов Ньютона и квазиньютоновских методов .