Методы оптимизации
Лекция 10: Метод Ньютона. Квазиньютоновские методы

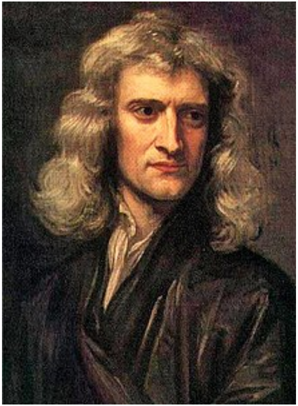
Александр Катруца

Физтех-школа прикладной математики и информатики
Московский физико-технический институт

16 ноября 2020 г.

Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$



Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- Метод *второго* порядка

Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- Метод *второго* порядка
- Квадратичная аппроксимация

$$\hat{f}(\mathbf{h}) = f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h}$$

Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- Метод *второго* порядка
- Квадратичная аппроксимация

$$\hat{f}(\mathbf{h}) = f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h}$$

- Пусть $f''(\mathbf{x}) \succ 0$, тогда

$$\hat{f}(\mathbf{h}) \to \min_{\mathbf{h}}$$

выпукла

Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- Метод *второго* порядка
- Квадратичная аппроксимация

$$\hat{f}(\mathbf{h}) = f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h}$$

- Пусть $f''(\mathbf{x}) \succ 0$, тогда

$$\hat{f}(\mathbf{h}) \to \min_{\mathbf{h}}$$

  выпукла
- Из условия первого порядка

$$f'(\mathbf{x}) + f''(\mathbf{x})\mathbf{h} = 0 \quad \Rightarrow \quad \mathbf{h}^* = -f''(\mathbf{x})^{-1} f'(\mathbf{x})$$

Метод Ньютона

$$\min_{\mathbf{x}} f(\mathbf{x})$$

- Метод *второго* порядка
- Квадратичная аппроксимация

$$\hat{f}(\mathbf{h}) = f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h}$$

- Пусть $f''(\mathbf{x}) \succ 0$, тогда

$$\hat{f}(\mathbf{h}) \to \min_{\mathbf{h}}$$

  выпукла

- Из условия первого порядка

$$f'(\mathbf{x}) + f''(\mathbf{x})\mathbf{h} = 0 \quad \Rightarrow \quad \mathbf{h}^* = -f''(\mathbf{x})^{-1} f'(\mathbf{x})$$

- Метод Ньютона

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)}$$

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

$$G(\mathbf{x}) = 0, \quad G : \mathbb{R}^n \to \mathbb{R}^n$$

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

$$G(\mathbf{x}) = 0, \quad G : \mathbb{R}^n \to \mathbb{R}^n$$

- Линейное приближение

$$G(\mathbf{x}_k + \Delta\mathbf{x}) \approx G(\mathbf{x}_k) + G'(\mathbf{x}_k)\Delta\mathbf{x} = 0,$$

где $G'(\mathbf{x})$ – матрица Якоби

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

$$G(\mathbf{x}) = 0, \quad G : \mathbb{R}^n \to \mathbb{R}^n$$

- Линейное приближение

$$G(\mathbf{x}_k + \Delta\mathbf{x}) \approx G(\mathbf{x}_k) + G'(\mathbf{x}_k)\Delta\mathbf{x} = 0,$$

  где $G'(\mathbf{x})$ – матрица Якоби
- Если $G'(\mathbf{x})$ обратима, то

$$\Delta\mathbf{x} = -G'(\mathbf{x}_k)^{-1}G(\mathbf{x}_k)$$

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

$$G(\mathbf{x}) = 0, \quad G : \mathbb{R}^n \to \mathbb{R}^n$$

- Линейное приближение

$$G(\mathbf{x}_k + \Delta\mathbf{x}) \approx G(\mathbf{x}_k) + G'(\mathbf{x}_k)\Delta\mathbf{x} = 0,$$

  где $G'(\mathbf{x})$ – матрица Якоби
- Если $G'(\mathbf{x})$ обратима, то

$$\Delta\mathbf{x} = -G'(\mathbf{x}_k)^{-1}G(\mathbf{x}_k)$$

- Метод Ньютона

$$\mathbf{x}_{k+1} = \mathbf{x}_k - G'(\mathbf{x}_k)^{-1}G(\mathbf{x}_k)$$

Связь с оптимизацией

- Пусть целевая функция $f(\mathbf{x})$ в задаче

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{1}$$

выпукла

Связь с оптимизацией

- Пусть целевая функция $f(\mathbf{x})$ в задаче

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{1}$$

выпукла
- Условие оптимальности первого порядка

$$f'(\mathbf{x}^*) = G(\mathbf{x}) = 0$$

Связь с оптимизацией

- Пусть целевая функция $f(\mathbf{x})$ в задаче

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{1}$$

  выпукла
- Условие оптимальности первого порядка

$$f'(\mathbf{x}^*) = G(\mathbf{x}) = 0$$

- Система для поиска направления $\mathbf{h}$

$$f'(\mathbf{x}) + f''(\mathbf{x})\mathbf{h} = 0$$

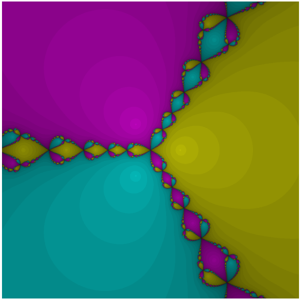
  эквивалентна системе в методе Ньютона для решения задачи (1)

Сравнение подходов к получению метода Ньютона

- Метод Ньютона для решения уравнений более общий, чем для решения задачи минимизации
  Q: Почему?

Сравнение подходов к получению метода Ньютона

- Метод Ньютона для решения уравнений более общий, чем для решения задачи минимизации
  Q: Почему?
- Анализ сходимости метода Ньютона в общем случае весьма нетривиален
- Фракталы Ньютона



Сходимость

Предположение $f''(\mathbf{x}) \succ 0$:

- если $f''(\mathbf{x}) \nsucc 0$, метод не работает
- модификации метода Ньютона для этого случая

Сходимость

Предположение $f''(\mathbf{x}) \succ 0$:

- если $f''(\mathbf{x}) \nsucc 0$, метод не работает
- модификации метода Ньютона для этого случая

Локальная сходимость: в зависимости от выбора $\mathbf{x}_0$ метод может

- сходиться
- расходиться
- осциллировать

Сходимость

Предположение $f''(\mathbf{x}) \succ 0$:

- если $f''(\mathbf{x}) \not\succ 0$, метод не работает
- модификации метода Ньютона для этого случая

Локальная сходимость: в зависимости от выбора $\mathbf{x}_0$ метод может

- сходиться
- расходиться
- осциллировать

Демпфированный метод Ньютона

$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)}$$

- Выбор шага по аналогии с градиентным спуском
- Введение шага расширяет область сходимости

Локальная сверхлинейная сходимость

- Пусть $\mathbf{x}^*$ – локальный минимум, тогда

$$f'(\mathbf{x}^*) = 0, \quad f''(\mathbf{x}^*) \succ 0$$

Локальная сверхлинейная сходимость

- Пусть $\mathbf{x}^*$ – локальный минимум, тогда

$$f'(\mathbf{x}^*) = 0, \quad f''(\mathbf{x}^*) \succ 0$$

- Ряд Тейлора

$$0 = f'(\mathbf{x}^*) = f'(\mathbf{x}_k) + f''(\mathbf{x}_k)(\mathbf{x}^* - \mathbf{x}_k) + o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

Локальная сверхлинейная сходимость

- Пусть $\mathbf{x}^*$ – локальный минимум, тогда

$$f'(\mathbf{x}^*) = 0, \quad f''(\mathbf{x}^*) \succ 0$$

- Ряд Тейлора

$$0 = f'(\mathbf{x}^*) = f'(\mathbf{x}_k) + f''(\mathbf{x}_k)(\mathbf{x}^* - \mathbf{x}_k) + o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- После умножения на $f''(\mathbf{x}_k)^{-1}$

$$\mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

Локальная сверхлинейная сходимость

- Пусть $\mathbf{x}^*$ – локальный минимум, тогда

$$f'(\mathbf{x}^*) = 0, \quad f''(\mathbf{x}^*) \succ 0$$

- Ряд Тейлора

$$0 = f'(\mathbf{x}^*) = f'(\mathbf{x}_k) + f''(\mathbf{x}_k)(\mathbf{x}^* - \mathbf{x}_k) + o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- После умножения на $f''(\mathbf{x}_k)^{-1}$

$$\mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- Итерация метода Ньютона $\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$, поэтому

$$\mathbf{x}_{k+1} - \mathbf{x}^* = o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

Локальная сверхлинейная сходимость

- Пусть $\mathbf{x}^*$ – локальный минимум, тогда

$$f'(\mathbf{x}^*) = 0, \quad f''(\mathbf{x}^*) \succ 0$$

- Ряд Тейлора

$$0 = f'(\mathbf{x}^*) = f'(\mathbf{x}_k) + f''(\mathbf{x}_k)(\mathbf{x}^* - \mathbf{x}_k) + o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- После умножения на $f''(\mathbf{x}_k)^{-1}$

$$\mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- Итерация метода Ньютона $\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$, поэтому

$$\mathbf{x}_{k+1} - \mathbf{x}^* = o(\|\mathbf{x}^* - \mathbf{x}_k\|)$$

- Локальная сверхлинейная сходимость $(\mathbf{x}_k \neq \mathbf{x}^*)$

$$\lim_{k \to \infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \lim_{k \to \infty} \frac{o(\|\mathbf{x}_k - \mathbf{x}^*\|)}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$

Локальная квадратичная сходимость

Теорема

Пусть

- $f(\mathbf{x})$ *локально сильно выпукла с константой* μ*:*
  $\exists\ \mathbf{x}^* :\ f''(\mathbf{x}^*) \succeq \mu \mathbf{I}$

Локальная квадратичная сходимость

Теорема

Пусть

- $f(\mathbf{x})$ *локально сильно выпукла с константой* μ*:*
  $\exists\ \mathbf{x}^* :\ f''(\mathbf{x}^*) \succeq \mu \mathbf{I}$
- *гессиан Липшицев:* $\|f''(\mathbf{x}) - f''(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|$

Локальная квадратичная сходимость

Теорема

Пусть

- $f(\mathbf{x})$ *локально сильно выпукла с константой* μ*:*
  $\exists\ \mathbf{x}^* :\ f''(\mathbf{x}^*) \succeq \mu \mathbf{I}$
- *гессиан Липшицев:* $\|f''(\mathbf{x}) - f''(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$
- *начальная точка* $\mathbf{x}_0$ *достаточно близка к* $\mathbf{x}^*$*:*
  $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \frac{2\mu}{3M}$

Локальная квадратичная сходимость

Теорема

Пусть

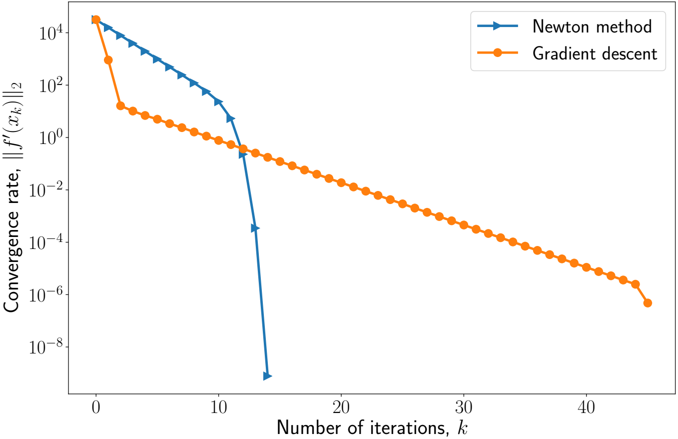
- $f(\mathbf{x})$ *локально сильно выпукла с константой* μ*:*
  $\exists\ \mathbf{x}^* : \ f''(\mathbf{x}^*) \succeq \mu \mathbf{I}$
- *гессиан Липшицев:* $\|f''(\mathbf{x}) - f''(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$
- *начальная точка* $\mathbf{x}_0$ *достаточно близка к* $\mathbf{x}^*$*:*
  $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq \frac{2\mu}{3M}$

тогда метод Ньютона сходится ***квадратично***

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \frac{M\|\mathbf{x}_k - \mathbf{x}^*\|^2}{2(\mu - M\|\mathbf{x}_k - \mathbf{x}^*\|)}$$

Пример

$$
-\sum_{i=1}^{m} \log(1 - \mathbf{a}_i^\top \mathbf{x}) - \sum_{i=1}^{n} \log(1 - x_i^2) \to \min_{\mathbf{x} \in \mathbb{R}^n}
$$



Доказательство в 9 шагов

Доказательство в 9 шагов

1. $\mathbf{r}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = \mathbf{r}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$

Доказательство в 9 шагов

1. $\mathbf{r}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = \mathbf{r}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$
2. Известный факт из анализа

$$\phi(b) - \phi(a) = \int_0^1 \phi'(a + t(b-a))(b-a)dt$$

Доказательство в 9 шагов

1. $\mathbf{r}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = \mathbf{r}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$
2. Известный факт из анализа

$$\phi(b) - \phi(a) = \int_0^1 \phi'(a + t(b-a))(b-a)dt$$

3. Для градиентов

$$f'(\mathbf{x}_k) = f'(\mathbf{x}_k) - f'(\mathbf{x}^*) = \int_0^1 f''(\mathbf{x}^* + t\mathbf{r}_k)\mathbf{r}_k dt$$

Доказательство в 9 шагов

1. $\mathbf{r}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = \mathbf{r}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$
2. Известный факт из анализа
$$\phi(b) - \phi(a) = \int_0^1 \phi'(a + t(b-a))(b-a)dt$$
3. Для градиентов
$$f'(\mathbf{x}_k) = f'(\mathbf{x}_k) - f'(\mathbf{x}^*) = \int_0^1 f''(\mathbf{x}^* + t\mathbf{r}_k)\mathbf{r}_k dt$$
4. Подставляем в первый шаг и группируем
$$\mathbf{r}_{k+1} = \underbrace{\left(\mathbf{I} - f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}^* + t\mathbf{r}_k)]dt\right)}_{\mathbf{G}_k} \mathbf{r}_k$$

Доказательство в 9 шагов

1. $\mathbf{r}_{k+1} = \mathbf{x}_{k+1} - \mathbf{x}^* = \mathbf{x}_k - \mathbf{x}^* - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k) = \mathbf{r}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$
2. Известный факт из анализа
$$\phi(b) - \phi(a) = \int_0^1 \phi'(a + t(b-a))(b-a)dt$$
3. Для градиентов
$$f'(\mathbf{x}_k) = f'(\mathbf{x}_k) - f'(\mathbf{x}^*) = \int_0^1 f''(\mathbf{x}^* + t\mathbf{r}_k)\mathbf{r}_k dt$$
4. Подставляем в первый шаг и группируем
$$\mathbf{r}_{k+1} = \underbrace{\left(\mathbf{I} - f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}^* + t\mathbf{r}_k)]dt\right)}_{\mathbf{G}_k} \mathbf{r}_k$$
5. $\|\mathbf{r}_{k+1}\| \leq \|\mathbf{G}_k\| \|\mathbf{r}_k\|$

6. Используем Липшицевость гессиана

$$
\begin{aligned}
\mathbf{G}_k &= f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)]dt \\
\|\mathbf{G}_k\| &\leq \|f''(\mathbf{x}_k)^{-1}\| \int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt
\end{aligned}
$$

6. Используем Липшицевость гессиана

$$
\mathbf{G}_k = f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)]dt
$$

$$
\|\mathbf{G}_k\| \leq \|f''(\mathbf{x}_k)^{-1}\| \int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt
$$

7. Оценим интеграл

$$
\int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt \leq \int_0^1 M\|\mathbf{r}_k - t\mathbf{r}_k\|dt = \frac{M\|\mathbf{r}_k\|}{2}
$$

6. Используем Липшицевость гессиана

$$
\mathbf{G}_k = f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)]dt
$$

$$
\|\mathbf{G}_k\| \leq \|f''(\mathbf{x}_k)^{-1}\| \int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt
$$

7. Оценим интеграл

$$
\int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt \leq \int_0^1 M\|\mathbf{r}_k - t\mathbf{r}_k\|dt = \frac{M\|\mathbf{r}_k\|}{2}
$$

8. Следствие Липшицевости гессиана и сильной выпуклости f в $\mathbf{x}^*$

$$
f''(\mathbf{x}_k) \succeq f''(\mathbf{x}^*) - M\|\mathbf{r}_k\|\mathbf{I} \succeq (\mu - M\|\mathbf{r}_k\|)\mathbf{I}
$$

6. Используем Липшицевость гессиана

$$\mathbf{G}_k = f''(\mathbf{x}_k)^{-1} \int_0^1 [f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)]dt$$

$$\|\mathbf{G}_k\| \le \|f''(\mathbf{x}_k)^{-1}\| \int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt$$

7. Оценим интеграл

$$\int_0^1 \|f''(\mathbf{x}_k) - f''(\mathbf{x}^* + t\mathbf{r}_k)\|dt \le \int_0^1 M\|\mathbf{r}_k - t\mathbf{r}_k\|dt = \frac{M\|\mathbf{r}_k\|}{2}$$

8. Следствие Липшицевости гессиана и сильной выпуклости f в $\mathbf{x}^*$

$$f''(\mathbf{x}_k) \succeq f''(\mathbf{x}^*) - M\|\mathbf{r}_k\|\mathbf{I} \succeq (\mu - M\|\mathbf{r}_k\|)\mathbf{I}$$

9. Оценим норму обратного гессиана

$$\|f''(\mathbf{x}_k)^{-1}\| \le \frac{1}{\mu - M\|\mathbf{r}_k\|}$$

Pro & Contra

Pro & Contra

Pro

- Квадратичная сходимость
- Высокая точность решения
- Аффинная инвариантность

Pro & Contra

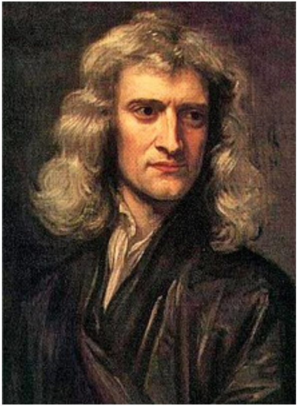
Pro

- Квадратичная сходимость
- Высокая точность решения
- Аффинная инвариантность

Contra

- Хранение гессиана: $O(n^2)$ памяти
- Необходимо решать линейные системы: $O(n^3)$ операций в общем случае
- Гессиан может оказаться вырожденным

Что объединяет градиентный спуск и метод Ньютона?



Что объединяет градиентный спуск и метод Ньютона?

Пусть градиент $f'(\mathbf{x})$ липшицев с константой L

- Градиентный спуск

$$f(\mathbf{x}+\mathbf{h}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2\alpha}\mathbf{h}^\top \mathbf{I}\mathbf{h} \equiv f_g(\mathbf{h}), \quad \alpha \in (0, 1/L]$$

$$\min_{\mathbf{h}} f_g(\mathbf{h}) \Rightarrow \mathbf{h}^* = -\alpha f'(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k f'(\mathbf{x}_k)$$

Что объединяет градиентный спуск и метод Ньютона?

Пусть градиент $f'(\mathbf{x})$ липшицев с константой L

- Градиентный спуск

$$f(\mathbf{x}+\mathbf{h}) \le f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2\alpha}\mathbf{h}^\top \mathbf{I}\mathbf{h} \equiv f_g(\mathbf{h}), \quad \alpha \in (0, 1/L]$$

$$\min_{\mathbf{h}} f_g(\mathbf{h}) \Rightarrow \mathbf{h}^* = -\alpha f'(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k f'(\mathbf{x}_k)$$

- Метод Ньютона

$$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h} \equiv f_N(\mathbf{h})$$

$$\min_{\mathbf{h}} f_N(\mathbf{h}) \Rightarrow \mathbf{h}^* = -(f''(\mathbf{x}))^{-1} f'(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$$

Что объединяет градиентный спуск и метод Ньютона?

Пусть градиент $f'(\mathbf{x})$ липшицев с константой L

- Градиентный спуск

$$f(\mathbf{x}+\mathbf{h}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2\alpha}\mathbf{h}^\top \mathbf{I}\mathbf{h} \equiv f_g(\mathbf{h}), \quad \alpha \in (0, 1/L]$$

$$\min_{\mathbf{h}} f_g(\mathbf{h}) \Rightarrow \mathbf{h}^* = -\alpha f'(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k f'(\mathbf{x}_k)$$

- Метод Ньютона

$$f(\mathbf{x}+\mathbf{h}) \approx f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{h}\rangle + \frac{1}{2}\mathbf{h}^\top f''(\mathbf{x})\mathbf{h} \equiv f_N(\mathbf{h})$$

$$\min_{\mathbf{h}} f_N(\mathbf{h}) \Rightarrow \mathbf{h}^* = -(f''(\mathbf{x}))^{-1} f'(\mathbf{x})$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - f''(\mathbf{x}_k)^{-1} f'(\mathbf{x}_k)$$

- Лучше чем $f_g(\mathbf{x})$, но быстрее, чем $f_N(\mathbf{x})$?

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Требования к оценке гессиана $\mathbf{B}_k$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Требования к оценке гессиана $\mathbf{B}_k$

- Быстрое обновление $\mathbf{B}_k \to \mathbf{B}_{k+1}$, **доступны только градиенты**

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Требования к оценке гессиана $\mathbf{B}_k$

- Быстрое обновление $\mathbf{B}_k \to \mathbf{B}_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления $\mathbf{h}_k$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h}\rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Требования к оценке гессиана $\mathbf{B}_k$

- Быстрое обновление $\mathbf{B}_k \to \mathbf{B}_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления $\mathbf{h}_k$
- Компактное хранение $\mathbf{B}_k$

Квазиньютоновские методы

- Квадратичная оценка $f(\mathbf{x}_{k+1})$

$$f_q(\mathbf{h}) = f(\mathbf{x}_k) + \langle f'(\mathbf{x}_k), \mathbf{h} \rangle + \frac{1}{2}\mathbf{h}^\top \mathbf{B}_k \mathbf{h}, \quad \mathbf{B}_k \succ 0$$

- Минимум $f_q(\mathbf{h})$ достигается в точке

$$\mathbf{h}_k = -\mathbf{B}_k^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновский метод

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{B}_k^{-1} f'(\mathbf{x}_k) = \mathbf{x}_k - \alpha_k \mathbf{H}_k f'(\mathbf{x}_k)$$

Требования к оценке гессиана $\mathbf{B}_k$

- Быстрое обновление $\mathbf{B}_k \to \mathbf{B}_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления $\mathbf{h}_k$
- Компактное хранение $\mathbf{B}_k$
- Сверхлинейная сходимость

Немного истории

- Первый квазиньютоновский метод придумал физик William Davidon в середине 1950-х
- Статью не приняли к публикации в Journal of Mathematics and Physics, и она оставалась препринтом более 30 лет
- Опубликована в 1991 году в первом выпуске SIAM Journal on Optimization

Как обновлять $\mathbf{B}_k$?

Как обновлять $\mathbf{B}_k$?

Правило двух градиентов

- $f'_q(-\alpha_k \mathbf{h}_k) = f'(\mathbf{x}_k) \Rightarrow f'(\mathbf{x}_{k+1}) - \alpha_k \mathbf{B}_{k+1} \mathbf{h}_k = f'(\mathbf{x}_k)$
- $f'_q(0) = f'(\mathbf{x}_{k+1})$ – выполнено по построению

Как обновлять $\mathbf{B}_k$?

Правило двух градиентов

- $f'_q(-\alpha_k \mathbf{h}_k) = f'(\mathbf{x}_k) \Rightarrow f'(\mathbf{x}_{k+1}) - \alpha_k \mathbf{B}_{k+1} \mathbf{h}_k = f'(\mathbf{x}_k)$
- $f'_q(0) = f'(\mathbf{x}_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$
- $\mathbf{y}_k = f'(\mathbf{x}_{k+1}) - f'(\mathbf{x}_k)$

$$\mathbf{B}_{k+1} \mathbf{s}_k = \mathbf{y}_k,$$

Как обновлять $\mathbf{B}_k$?

Правило двух градиентов

- $f'_q(-\alpha_k \mathbf{h}_k) = f'(\mathbf{x}_k) \Rightarrow f'(\mathbf{x}_{k+1}) - \alpha_k \mathbf{B}_{k+1}\mathbf{h}_k = f'(\mathbf{x}_k)$
- $f'_q(0) = f'(\mathbf{x}_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$
- $\mathbf{y}_k = f'(\mathbf{x}_{k+1}) - f'(\mathbf{x}_k)$

$$\mathbf{B}_{k+1}\mathbf{s}_k = \mathbf{y}_k,$$

Q: всегда ли это уравнение имеет решение?

Q: единственно ли оно?

Как обновлять $\mathbf{B}_k$?

Правило двух градиентов

- $f'_q(-\alpha_k \mathbf{h}_k) = f'(\mathbf{x}_k) \Rightarrow f'(\mathbf{x}_{k+1}) - \alpha_k \mathbf{B}_{k+1}\mathbf{h}_k = f'(\mathbf{x}_k)$
- $f'_q(0) = f'(\mathbf{x}_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$
- $\mathbf{y}_k = f'(\mathbf{x}_{k+1}) - f'(\mathbf{x}_k)$

$$\mathbf{B}_{k+1}\mathbf{s}_k = \mathbf{y}_k,$$

Q: всегда ли это уравнение имеет решение?

Q: единственно ли оно?

- Новая оценка гессиана должна быть близка к текущей

Параметры

- Необходимо задать $\mathbf{B}_0$, обычно $\mathbf{B}_0 = \gamma \mathbf{I}$ для некоторого γ

Параметры

- Необходимо задать $\mathbf{B}_0$, обычно $\mathbf{B}_0 = \gamma \mathbf{I}$ для некоторого γ
- Параметры в процедуре поиска шага

Параметры

- Необходимо задать $\mathbf{B}_0$, обычно $\mathbf{B}_0 = \gamma \mathbf{I}$ для некоторого γ
- Параметры в процедуре поиска шага
- Все вычисления необходимо организовать так, чтобы не было операций сложностью $O(n^3)$

Параметры

- Необходимо задать $\mathbf{B}_0$, обычно $\mathbf{B}_0 = \gamma \mathbf{I}$ для некоторого γ
- Параметры в процедуре поиска шага
- Все вычисления необходимо организовать так, чтобы не было операций сложностью $O(n^3)$

Примеры квазиньютоновских методов

- Barzilai-Borwein
- DFP
- BFGS

Метод Barzilai-Borwein

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

$$\alpha_k f'(\mathbf{x}_k) = \alpha_k \mathbf{I} f'(\mathbf{x}_k) = \left(\frac{1}{\alpha_k}\mathbf{I}\right)^{-1} f'(\mathbf{x}_k) \approx f''(\mathbf{x}_k))^{-1} f'(\mathbf{x}_k)$$

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

$$\alpha_k f'(\mathbf{x}_k) = \alpha_k \mathbf{I} f'(\mathbf{x}_k) = \left(\frac{1}{\alpha_k} \mathbf{I} \right)^{-1} f'(\mathbf{x}_k) \approx f''(\mathbf{x}_k))^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновское уравнение

$$\alpha_k^{-1} \mathbf{s}_{k-1} \approx \mathbf{y}_{k-1}$$

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

$$\alpha_k f'(\mathbf{x}_k) = \alpha_k \mathbf{I} f'(\mathbf{x}_k) = \left(\frac{1}{\alpha_k} \mathbf{I} \right)^{-1} f'(\mathbf{x}_k) \approx f''(\mathbf{x}_k))^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновское уравнение

$$\alpha_k^{-1} \mathbf{s}_{k-1} \approx \mathbf{y}_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|\mathbf{s}_{k-1} - \alpha_k \mathbf{y}_{k-1}\|_2 \Rightarrow \alpha_k = \frac{\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^\top \mathbf{y}_{k-1}}$$

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

$$\alpha_k f'(\mathbf{x}_k) = \alpha_k \mathbf{I} f'(\mathbf{x}_k) = \left(\frac{1}{\alpha_k}\mathbf{I}\right)^{-1} f'(\mathbf{x}_k) \approx f''(\mathbf{x}_k))^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновское уравнение

$$\alpha_k^{-1}\mathbf{s}_{k-1} \approx \mathbf{y}_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|\mathbf{s}_{k-1} - \alpha_k \mathbf{y}_{k-1}\|_2 \Rightarrow \alpha_k = \frac{\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^\top \mathbf{y}_{k-1}}$$

- Можно ставить другие задачи для поиска α_k

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

$$\alpha_k f'(\mathbf{x}_k) = \alpha_k \mathbf{I} f'(\mathbf{x}_k) = \left(\frac{1}{\alpha_k}\mathbf{I}\right)^{-1} f'(\mathbf{x}_k) \approx f''(\mathbf{x}_k))^{-1} f'(\mathbf{x}_k)$$

- Квазиньютоновское уравнение

$$\alpha_k^{-1}\mathbf{s}_{k-1} \approx \mathbf{y}_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|\mathbf{s}_{k-1} - \alpha_k \mathbf{y}_{k-1}\|_2 \Rightarrow \alpha_k = \frac{\mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}}{\mathbf{y}_{k-1}^\top \mathbf{y}_{k-1}}$$

- Можно ставить другие задачи для поиска α_k
- Имеет стохастическую модификацию, статья на NIPS 2016

Метод DFP

- Задача поиска $\mathbf{B}_{k+1}$

$$
\begin{aligned}
&\min_{\mathbf{B}} \|\mathbf{B}_k - \mathbf{B}\| \\
\text{s.t. } & \mathbf{B} = \mathbf{B}^\top \\
& \mathbf{B}\mathbf{s}_k = \mathbf{y}_k
\end{aligned}
$$

- Решение

$$
\mathbf{B}_{k+1} = (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top)\mathbf{B}_k(\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top) + \rho_k \mathbf{y}_k \mathbf{y}_k^\top,
$$

где $\rho_k = \dfrac{1}{\mathbf{y}_k^\top \mathbf{s}_k}$

- По формуле Шермана-Моррисона-Вудбери

$$
\mathbf{B}_{k+1}^{-1} = \mathbf{H}_{k+1} = \mathbf{H}_k - \frac{\mathbf{H}_k \mathbf{y}_k \mathbf{y}_k^\top \mathbf{H}_k}{\mathbf{y}_k^\top \mathbf{H}_k \mathbf{y}_k} + \frac{\mathbf{s}_k \mathbf{s}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k}
$$

Метод BFGS

Broyden, Fletcher, Goldfarb, Shanno

Метод BFGS

- Задача

$$\min_{\mathbf{H}} \|\mathbf{H}_k - \mathbf{H}\|$$
$$\text{s.t. } \mathbf{H} = \mathbf{H}^\top$$
$$\mathbf{H}\mathbf{y}_k = \mathbf{s}_k$$

Метод BFGS

- Задача

$$
\begin{aligned}
&\min_{\mathbf{H}} \|\mathbf{H}_k - \mathbf{H}\| \\
\text{s.t. } &\mathbf{H} = \mathbf{H}^\top \\
&\mathbf{H}\mathbf{y}_k = \mathbf{s}_k
\end{aligned}
$$

- Решение

$$
\mathbf{H}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top)\mathbf{H}_k(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top) + \rho_k \mathbf{s}_k \mathbf{s}_k^\top,
$$

где $\rho_k = \dfrac{1}{\mathbf{y}_k^\top \mathbf{s}_k}$

Метод BFGS

- Задача

$$\min_{\mathbf{H}} \|\mathbf{H}_k - \mathbf{H}\|$$
$$\text{s.t. } \mathbf{H} = \mathbf{H}^\top$$
$$\mathbf{H}\mathbf{y}_k = \mathbf{s}_k$$

- Решение

$$\mathbf{H}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top)\mathbf{H}_k(\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top) + \rho_k \mathbf{s}_k \mathbf{s}_k^\top,$$

где $\rho_k = \dfrac{1}{\mathbf{y}_k^\top \mathbf{s}_k}$

Теорема (почти)

Пусть f сильно выпукла с Липшицевым гессианом. Тогда при некоторых дополнительных технических условиях BFGS сходится сверхлинейно.

Ещё немного про BFGS

- Очень хорошо работает на практике

Ещё немного про BFGS

- Очень хорошо работает на практике
- Обладает свойством самокоррекции

Ещё немного про BFGS

- Очень хорошо работает на практике
- Обладает свойством самокоррекции
- Формулу обновления $\mathbf{H}_k$ можно также получить как решение задачи

$$\min_{\mathbf{H}} \operatorname{trace}(\mathbf{H}_k^\top \mathbf{H}^{-1}) - \log\det(\mathbf{H}_k \mathbf{H}^{-1}) - n$$

$$\text{s.t. } \mathbf{H}\mathbf{y}_k = \mathbf{s}_k$$

Целевая функция $\equiv$ дивергенции Кульбака-Лейблере между распределениями $\mathcal{N}(0, \mathbf{H}^{-1})$ и $\mathcal{N}(0, \mathbf{H}_k^{-1})$

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$
- Необходима не сама матрица, а **эффективная** процедура умножения её на вектор $f'(\mathbf{x})$

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$
- Необходима не сама матрица, а **эффективная** процедура умножения её на вектор $f'(\mathbf{x})$
- Значения $\mathbf{y}$ и $\mathbf{s}$ на первых итерациях могут портить оценку $\mathbf{B}$ или $\mathbf{H}$ на более поздних итерациях

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$
- Необходима не сама матрица, а **эффективная** процедура умножения её на вектор $f'(\mathbf{x})$
- Значения $\mathbf{y}$ и $\mathbf{s}$ на первых итерациях могут портить оценку $\mathbf{B}$ или $\mathbf{H}$ на более поздних итерациях

Идея

Использовать последние $m \ll n$ значений $(\mathbf{s}, \mathbf{y})$ и корректировать $\mathbf{H}_{m,0}$ для каждой итерации

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$
- Необходима не сама матрица, а **эффективная** процедура умножения её на вектор $f'(\mathbf{x})$
- Значения $\mathbf{y}$ и $\mathbf{s}$ на первых итерациях могут портить оценку $\mathbf{B}$ или $\mathbf{H}$ на более поздних итерациях

Идея

Использовать последние $m \ll n$ значений $(\mathbf{s}, \mathbf{y})$ и корректировать $\mathbf{H}_{m,0}$ для каждой итерации

- Сложность стала $O(mn)$

Квазиньютоновские методы с ограниченной памятью

- Сложность хранения и обновления гессиана $O(n^2)$
- Необходима не сама матрица, а **эффективная** процедура умножения её на вектор $f'(\mathbf{x})$
- Значения $\mathbf{y}$ и $\mathbf{s}$ на первых итерациях могут портить оценку $\mathbf{B}$ или $\mathbf{H}$ на более поздних итерациях

Идея

Использовать последние $m \ll n$ значений $(\mathbf{s}, \mathbf{y})$ и корректировать $\mathbf{H}_{m,0}$ для каждой итерации

- Сложность стала $O(mn)$

Q: как на каждой итерации поддерживать хранение последних m пар?

Метод L-BFGS

- Лучше всего работает на практике

Метод L-BFGS

- Лучше всего работает на практике
- Нужно заранее определить m

Метод L-BFGS

- Лучше всего работает на практике
- Нужно заранее определить m
- BFGS обновляет H рекурсивно

$$\mathbf{H}_{k+1} = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^\top, \quad \mathbf{V}_k = \mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top$$

Метод L-BFGS

- Лучше всего работает на практике
- Нужно заранее определить m
- BFGS обновляет H рекурсивно

$$\mathbf{H}_{k+1} = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^\top, \quad \mathbf{V}_k = \mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top$$

- Развернём m шагов рекурсии

$$\begin{aligned}
&\mathbf{H}_{k+1} = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^\top \\
&= \mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \mathbf{H}_{k-1} \mathbf{V}_{k-1} \mathbf{V}_k + \rho_{k-1} \mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \mathbf{s}_{k-1} \mathbf{s}_{k-1}^\top \mathbf{V}_{k-1} \mathbf{V}_k + \rho_k \mathbf{s}_k \\
&= \mathbf{V}_k^\top \dots \mathbf{V}_{k-m+1}^\top \mathbf{H}_{m,0} \mathbf{V}_{k-m+1} \dots \mathbf{V}_k \\
&+ \rho_{k-m+1} \mathbf{V}_k^\top \dots \mathbf{V}_{k-m+2}^\top \mathbf{s}_{k-m+1} \mathbf{s}_{k-m+1}^\top \mathbf{V}_{k-m+2} \dots V_k \\
&+ \dots + \rho_k \mathbf{s}_k \mathbf{s}_k^\top
\end{aligned}$$

Метод L-BFGS

- Лучше всего работает на практике
- Нужно заранее определить m
- BFGS обновляет H рекурсивно

$$\mathbf{H}_{k+1} = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^\top, \quad \mathbf{V}_k = \mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top$$

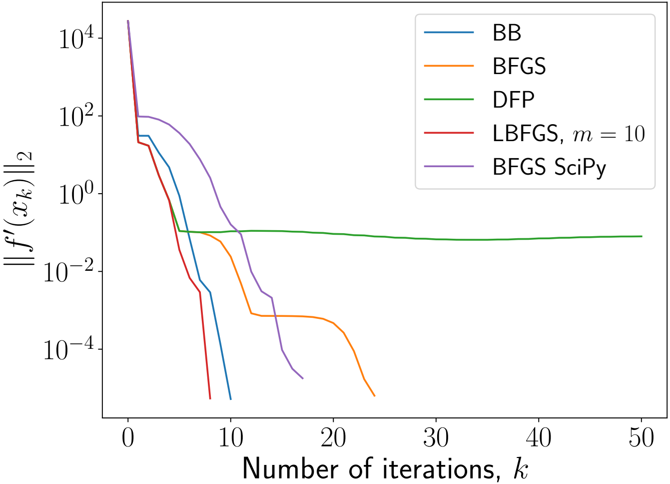
- Развернём m шагов рекурсии

$$\begin{aligned}
&\mathbf{H}_{k+1} = \mathbf{V}_k^\top \mathbf{H}_k \mathbf{V}_k + \rho_k \mathbf{s}_k \mathbf{s}_k^\top \\
&= \mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \mathbf{H}_{k-1} \mathbf{V}_{k-1} \mathbf{V}_k + \rho_{k-1} \mathbf{V}_k^\top \mathbf{V}_{k-1}^\top \mathbf{s}_{k-1} \mathbf{s}_{k-1}^\top \mathbf{V}_{k-1} \mathbf{V}_k + \rho_k \mathbf{s}_k \\
&= \mathbf{V}_k^\top \ldots \mathbf{V}_{k-m+1}^\top \mathbf{H}_{m,0} \mathbf{V}_{k-m+1} \ldots \mathbf{V}_k \\
&+ \rho_{k-m+1} \mathbf{V}_k^\top \ldots \mathbf{V}_{k-m+2}^\top \mathbf{s}_{k-m+1} \mathbf{s}_{k-m+1}^\top \mathbf{V}_{k-m+2} \ldots V_k \\
&+ \ldots + \rho_k \mathbf{s}_k \mathbf{s}_k^\top
\end{aligned}$$

- Эффективное вычисление $\mathbf{H}_k f'(\mathbf{x})$ без явного формирования $\mathbf{H}_k$

Пример

$$
-\sum_{i=1}^{m} \log(1 - \mathbf{a}_i^\top \mathbf{x}) - \sum_{i=1}^{n} \log(1 - x_i^2) \to \min_{\mathbf{x} \in \mathbb{R}^n}
$$



Pro & Contra

Pro & Contra

Pro

- Сложность одной итерации $O(n^2) + \ldots$ по сравнению с $O(n^3) + \ldots$ в методе Ньютона
- Для метода L-BFGS требуется линейное количество памяти по размерности задачи
- Самокоррекция метода BFGS
- Сверхлинейная сходимость к решению задачи

Pro & Contra

Pro

- Сложность одной итерации $O(n^2) + \ldots$ по сравнению с $O(n^3) + \ldots$ в методе Ньютона
- Для метода L-BFGS требуется линейное количество памяти по размерности задачи
- Самокоррекция метода BFGS
- Сверхлинейная сходимость к решению задачи

Contra

- Обобщение на стохастический случай не работает
- Выбор начального приближения $\mathbf{B}_0$ или $\mathbf{H}_0$
- Нет разработанной теории сходимости и оптимальности
- Не любой способ выбора шага гарантирует выполнение условия кривизны $\mathbf{y}_k^\top \mathbf{s}_k > 0$