Методы оптимизации
Метод Ньютона. Квазиньютоновские методы

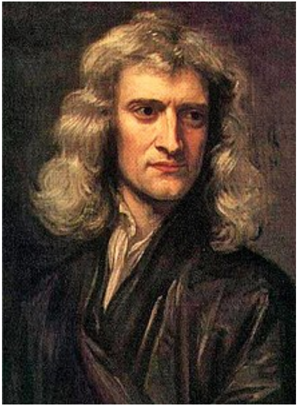
Александр Катруца

Московский физико-технический институт

20 апреля 2020 г.

Метод Ньютона

$$\min_x f(x)$$



Метод Ньютона

$$\min_x f(x)$$

- Метод *второго* порядка

Метод Ньютона

$$\min_x f(x)$$

- Метод *второго* порядка
- Квадратичная аппроксимация

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

Метод Ньютона

$$\min_x f(x)$$

- Метод *второго* порядка
- Квадратичная аппроксимация

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

- Пусть $f''(x) \succ 0$, тогда

$$\hat{f}(h) \to \min_h$$

выпукла

Метод Ньютона

$$\min_x f(x)$$

- Метод *второго* порядка
- Квадратичная аппроксимация

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

- Пусть $f''(x) \succ 0$, тогда

$$\hat{f}(h) \to \min_h$$

  выпукла
- Из условия первого порядка

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

Метод Ньютона

$$\min_x f(x)$$

- Метод *второго* порядка
- Квадратичная аппроксимация

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

- Пусть $f''(x) \succ 0$, тогда

$$\hat{f}(h) \to \min_h$$

  выпукла
- Из условия первого порядка

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

- Метод Ньютона

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

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

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

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

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

- Линейное приближение

$$G(x_k + \Delta x) \approx G(x_k) + G'(x_k)\Delta x = 0,$$

где $G'(x)$ – матрица Якоби

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

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

- Линейное приближение

$$G(x_k + \Delta x) \approx G(x_k) + G'(x_k)\Delta x = 0,$$

где $G'(x)$ – матрица Якоби

- Если $G'(x)$ обратима, то

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

Метод Ньютона для систем нелинейных уравнений

- Система нелинейных уравнений

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

- Линейное приближение

$$G(x_k + \Delta x) \approx G(x_k) + G'(x_k)\Delta x = 0,$$

где $G'(x)$ – матрица Якоби

- Если $G'(x)$ обратима, то

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

- Метод Ньютона

$$x_{k+1} = x_k - G'(x_k)^{-1} G(x_k)$$

Связь с оптимизацией

- Пусть целевая функция $f(x)$ в задаче

$$\min_x f(x) \tag{1}$$

выпукла

Связь с оптимизацией

- Пусть целевая функция $f(x)$ в задаче

$$\min_x f(x) \tag{1}$$

  выпукла
- Условие оптимальности первого порядка

$$f'(x^*) = G(x) = 0$$

Связь с оптимизацией

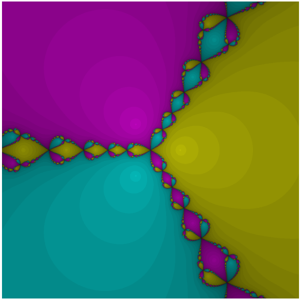
- Пусть целевая функция $f(x)$ в задаче
$$\min_x f(x) \tag{1}$$
выпукла
- Условие оптимальности первого порядка
$$f'(x^*) = G(x) = 0$$
- Система для поиска направления h
$$f'(x) + f''(x)h = 0$$
эквивалентна системе в методе Ньютона для решения задачи (1)

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 $f''(x) \succ 0$:

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

Сходимость

Предположение $f''(x) \succ 0$:

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

Локальная сходимость: в зависимости от выбора x_0 метод может

- сходиться
- расходиться
- осциллировать

Сходимость

Предположение $f''(x) \succ 0$:

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

Локальная сходимость: в зависимости от выбора x_0 метод может

- сходиться
- расходиться
- осциллировать

Демпфированный метод Ньютона

$$\boxed{x_{k+1} = x_k - {\color{red}\alpha_k} f''(x_k)^{-1} f'(x_k)}$$

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

Локальная сверхлинейная сходимость

- Пусть x^* – локальный минимум, тогда

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

Локальная сверхлинейная сходимость

- Пусть x^* – локальный минимум, тогда

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

- Ряд Тейлора

$$0 = f'(x^*) = f'(x_k) + f''(x_k)(x^* - x_k) + o(\|x^* - x^k\|)$$

Локальная сверхлинейная сходимость

- Пусть x^* – локальный минимум, тогда

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

- Ряд Тейлора

$$0 = f'(x^*) = f'(x_k) + f''(x_k)(x^* - x_k) + o(\|x^* - x^k\|)$$

- После умножения на $f''(x_k)^{-1}$

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

Локальная сверхлинейная сходимость

- Пусть x^* – локальный минимум, тогда

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

- Ряд Тейлора

$$0 = f'(x^*) = f'(x_k) + f''(x_k)(x^* - x_k) + o(\|x^* - x^k\|)$$

- После умножения на $f''(x_k)^{-1}$

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

- Итерация метода Ньютона $x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$, поэтому

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

Локальная сверхлинейная сходимость

- Пусть x^* – локальный минимум, тогда

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

- Ряд Тейлора

$$0 = f'(x^*) = f'(x_k) + f''(x_k)(x^* - x_k) + o(\|x^* - x^k\|)$$

- После умножения на $f''(x_k)^{-1}$

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

- Итерация метода Ньютона $x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$, поэтому

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

- Локальная сверхлинейная сходимость $(x_k \neq x^*)$

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

Локальная квадратичная сходимость

Теорема

Пусть

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

Локальная квадратичная сходимость

Теорема

Пусть

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

Локальная квадратичная сходимость

Теорема

Пусть

- $f(x)$ *локально сильно выпукла с константой* μ*:*
  $\exists\ x^* :\ f''(x^*) \succeq \mu I$
- *гессиан Липшицев:* $\|f''(x) - f''(y)\| \leq M\|x - y\|$
- *начальная точка* x_0 *достаточно близка к* x^**:*
  $\|x_0 - x^*\| \leq \frac{2\mu}{3M}$

Локальная квадратичная сходимость

Теорема

Пусть

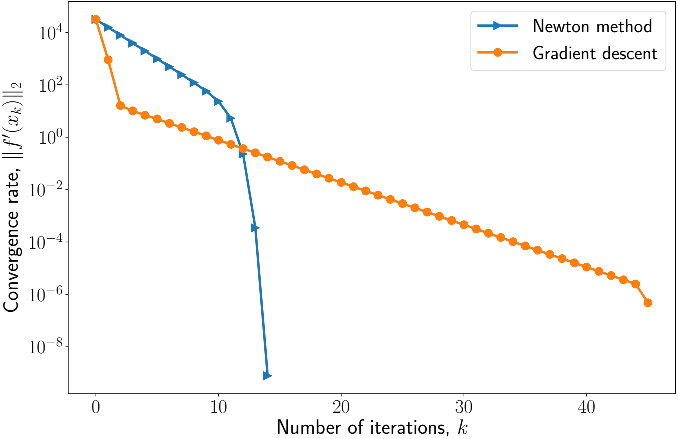
- $f(x)$ *локально сильно выпукла с константой* μ*:*
  $\exists\ x^* : \ f''(x^*) \succeq \mu I$
- *гессиан Липшицев:* $\|f''(x) - f''(y)\| \leq M\|x - y\|$
- *начальная точка* x_0 *достаточно близка к* x^**:*
  $\|x_0 - x^*\| \leq \frac{2\mu}{3M}$

тогда метод Ньютона сходится ***квадратично***

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

Пример

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



Доказательство в 9 шагов

Доказательство в 9 шагов

1. $r_{k+1} = x_{k+1} - x^* = x_k - x^* - f''(x_k)^{-1} f'(x_k) = r_k - f''(x_k)^{-1} f'(x_k)$

Доказательство в 9 шагов

1. $r_{k+1} = x_{k+1} - x^* = x_k - x^* - f''(x_k)^{-1}f'(x_k) = r_k - f''(x_k)^{-1}f'(x_k)$
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. $r_{k+1} = x_{k+1} - x^* = x_k - x^* - f''(x_k)^{-1} f'(x_k) =$
   $r_k - f''(x_k)^{-1} f'(x_k)$
2. Известный факт из анализа
$$\phi(b) - \phi(a) = \int_0^1 \phi'(a + t(b-a))(b-a)dt$$
3. Для градиентов
$$f'(x_k) = f'(x_k) - f'(x^*) = \int_0^1 f''(x^* + tr_k)r_k dt$$

Доказательство в 9 шагов

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

Доказательство в 9 шагов

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

6. Используем Липшицевость гессиана

$$G_k = f''(x_k)^{-1} \int_0^1 [f''(x_k) - f''(x^* + tr_k)] dt$$

$$\|G_k\| \leq \|f''(x_k)^{-1}\| \int_0^1 \|f''(x_k) - f''(x^* + tr_k)\| dt$$

6. Используем Липшицевость гессиана

$$G_k = f''(x_k)^{-1} \int_0^1 [f''(x_k) - f''(x^* + tr_k)]dt$$

$$\|G_k\| \leq \|f''(x_k)^{-1}\| \int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt$$

7. Оценим интеграл

$$\int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt \leq \int_0^1 M\|r_k - tr_k\|dt = \frac{M\|r_k\|}{2}$$

6. Используем Липшицевость гессиана

$$G_k = f''(x_k)^{-1} \int_0^1 [f''(x_k) - f''(x^* + tr_k)]dt$$

$$\|G_k\| \leq \|f''(x_k)^{-1}\| \int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt$$

7. Оценим интеграл

$$\int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt \leq \int_0^1 M\|r_k - tr_k\|dt = \frac{M\|r_k\|}{2}$$

8. Следствие Липшицевости гессиана и сильной выпуклости f в x^*

$$f''(x_k) \succeq f''(x^*) - M\|r_k\|I \succeq (\mu - M\|r_k\|)I$$

6. Используем Липшицевость гессиана

$$G_k = f''(x_k)^{-1} \int_0^1 [f''(x_k) - f''(x^* + tr_k)]dt$$

$$\|G_k\| \leq \|f''(x_k)^{-1}\| \int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt$$

7. Оценим интеграл

$$\int_0^1 \|f''(x_k) - f''(x^* + tr_k)\|dt \leq \int_0^1 M\|r_k - tr_k\|dt = \frac{M\|r_k\|}{2}$$

8. Следствие Липшицевости гессиана и сильной выпуклости f в x^*

$$f''(x_k) \succeq f''(x^*) - M\|r_k\|I \succeq (\mu - M\|r_k\|)I$$

9. Оценим норму обратного гессиана

$$\|f''(x_k)^{-1}\| \leq \frac{1}{\mu - M\|r_k\|}$$

Pro & Contra

Pro & Contra

Pro

- Квадратичная сходимость
- Высокая точность решения
- Аффинная инвариантность

Pro & Contra

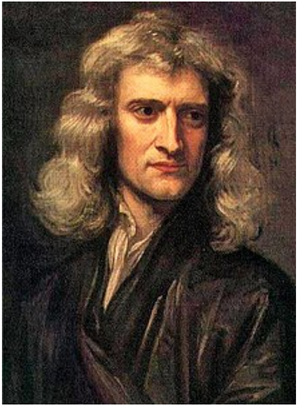
Pro

- 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 $f'(x)$ липшицев с константой L

- Градиентный спуск

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

$$\min_h f_g(h) \Rightarrow h^* = -\alpha f'(x)$$

$$x_{k+1} = x_k - \alpha_k f'(x_k)$$

Что объединяет градиентный спуск и метод Ньютона?

Пусть градиент $f'(x)$ липшицев с константой L

- Градиентный спуск

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

$$\min_h f_g(h) \Rightarrow h^* = -\alpha f'(x)$$

$$x_{k+1} = x_k - \alpha_k f'(x_k)$$

- Метод Ньютона

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

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

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

Что объединяет градиентный спуск и метод Ньютона?

Пусть градиент $f'(x)$ липшицев с константой L

- Градиентный спуск

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

$$\min_h f_g(h) \Rightarrow h^* = -\alpha f'(x)$$

$$x_{k+1} = x_k - \alpha_k f'(x_k)$$

- Метод Ньютона

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

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

$$x_{k+1} = x_k - f''(x_k)^{-1} f'(x_k)$$

- Лучше чем $f_g(x)$, но быстрее, чем $f_N(x)$?

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$

- Минимум $f_q(h)$ достигается в точке

$$h_k = -B_k^{-1} f'(x_k)$$

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$

- Минимум $f_q(h)$ достигается в точке

$$h_k = -B_k^{-1} f'(x_k)$$

- Квазиньютоновский метод

$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h\rangle + \frac{1}{2}h^\top B_k h, \quad B_k \succ 0$$

- Минимум $f_q(h)$ достигается в точке

$$h_k = -B_k^{-1} f'(x_k)$$

- Квазиньютоновский метод

$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Требования к оценке гессиана B_k

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$
$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$
- Минимум $f_q(h)$ достигается в точке
$$h_k = -B_k^{-1} f'(x_k)$$
- Квазиньютоновский метод
$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Требования к оценке гессиана B_k

- Быстрое обновление $B_k \to B_{k+1}$, **доступны только градиенты**

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$

- Минимум $f_q(h)$ достигается в точке

$$h_k = -B_k^{-1} f'(x_k)$$

- Квазиньютоновский метод

$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Требования к оценке гессиана B_k

- Быстрое обновление $B_k \to B_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления h_k

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$
$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$
- Минимум $f_q(h)$ достигается в точке
$$h_k = -B_k^{-1} f'(x_k)$$
- Квазиньютоновский метод
$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Требования к оценке гессиана B_k

- Быстрое обновление $B_k \to B_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления h_k
- Компактное хранение B_k

Квазиньютоновские методы

- Квадратичная оценка $f(x_{k+1})$

$$f_q(h) = f(x_k) + \langle f'(x_k), h \rangle + \frac{1}{2} h^\top B_k h, \quad B_k \succ 0$$

- Минимум $f_q(h)$ достигается в точке

$$h_k = -B_k^{-1} f'(x_k)$$

- Квазиньютоновский метод

$$x_{k+1} = x_k - \alpha_k B_k^{-1} f'(x_k) = x_k - \alpha_k H_k f'(x_k)$$

Требования к оценке гессиана B_k

- Быстрое обновление $B_k \to B_{k+1}$, **доступны только градиенты**
- Быстрый поиск направления h_k
- Компактное хранение 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 B_k?

Как обновлять B_k?

Правило двух градиентов

- $f_q'(-\alpha_k h_k) = f'(x_k) \Rightarrow f'(x_{k+1}) - \alpha_k B_{k+1} h_k = f'(x_k)$
- $f_q'(0) = f'(x_{k+1})$ – выполнено по построению

Как обновлять B_k?

Правило двух градиентов

- $f'_q(-\alpha_k h_k) = f'(x_k) \Rightarrow f'(x_{k+1}) - \alpha_k B_{k+1} h_k = f'(x_k)$
- $f'_q(0) = f'(x_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $s_k = x_{k+1} - x_k$
- $y_k = f'(x_{k+1}) - f'(x_k)$

$$B_{k+1} s_k = y_k,$$

Как обновлять B_k?

Правило двух градиентов

- $f_q'(-\alpha_k h_k) = f'(x_k) \Rightarrow f'(x_{k+1}) - \alpha_k B_{k+1} h_k = f'(x_k)$
- $f_q'(0) = f'(x_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $s_k = x_{k+1} - x_k$
- $y_k = f'(x_{k+1}) - f'(x_k)$

$$B_{k+1} s_k = y_k,$$

Q: всегда ли это уравнение имеет решение?

Q: единственно ли оно?

Как обновлять B_k?

Правило двух градиентов

- $f'_q(-\alpha_k h_k) = f'(x_k) \Rightarrow f'(x_{k+1}) - \alpha_k B_{k+1} h_k = f'(x_k)$
- $f'_q(0) = f'(x_{k+1})$ – выполнено по построению

Квазиньютоновское уравнение (Secant equation)

- $s_k = x_{k+1} - x_k$
- $y_k = f'(x_{k+1}) - f'(x_k)$

$$B_{k+1} s_k = y_k,$$

Q: всегда ли это уравнение имеет решение?
Q: единственно ли оно?

- Новая оценка гессиана должна быть близка к текущей

Параметры

- Необходимо задать B_0, обычно $B_0 = \gamma I$ для некоторого γ

Параметры

- Необходимо задать B_0, обычно $B_0 = \gamma I$ для некоторого γ
- Параметры в процедуре поиска шага

Параметры

- Необходимо задать B_0, обычно $B_0 = \gamma I$ для некоторого γ
- Параметры в процедуре поиска шага
- Все вычисления необходимо организовать так, чтобы не было операций сложностью $O(n^3)$

Параметры

- Необходимо задать B_0, обычно $B_0 = \gamma I$ 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:

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

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

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

- Квазиньютоновское уравнение

$$\alpha_k^{-1} s_{k-1} \approx y_{k-1}$$

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

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

- Квазиньютоновское уравнение

$$\alpha_k^{-1} s_{k-1} \approx y_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|s_{k-1} - \alpha_k y_{k-1}\|_2 \Rightarrow \alpha_k = \frac{s_{k-1}^\top y_{k-1}}{y_{k-1}^\top y_{k-1}}$$

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

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

- Квазиньютоновское уравнение

$$\alpha_k^{-1} s_{k-1} \approx y_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|s_{k-1} - \alpha_k y_{k-1}\|_2 \Rightarrow \alpha_k = \frac{s_{k-1}^\top y_{k-1}}{y_{k-1}^\top y_{k-1}}$$

- Можно ставить другие задачи для поиска α_k

Метод Barzilai-Borwein

- Аппроксимация гессиана диагональной матрицей:

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

- Квазиньютоновское уравнение

$$\alpha_k^{-1} s_{k-1} \approx y_{k-1}$$

- Задача и решение

$$\min_{\alpha_k} \|s_{k-1} - \alpha_k y_{k-1}\|_2 \Rightarrow \alpha_k = \frac{s_{k-1}^\top y_{k-1}}{y_{k-1}^\top y_{k-1}}$$

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

Метод DFP

- Задача поиска B_{k+1}

$$\begin{aligned} &\min_B \|B_k - B\| \\ \text{s.t. } &B = B^\top \\ &Bs_k = y_k \end{aligned}$$

- Решение

$$B_{k+1} = (I - \rho_k y_k s_k^\top) B_k (I - \rho_k s_k y_k^\top) + \rho_k y_k y_k^\top,$$

где $\rho_k = \dfrac{1}{y_k^\top s_k}$

- По формуле ШВМ

$$B_{k+1}^{-1} = H_{k+1} = H_k - \frac{H_k y_k y_k^\top H_k}{y_k^\top H_k y_k} + \frac{s_k s_k^\top}{y_k^\top s_k}$$

Метод BFGS

Broyden, Fletcher, Goldfarb, Shanno

Метод BFGS

- Задача

$$
\begin{aligned}
&\min_{H} \|H_k - H\| \\
\text{s.t. } &H = H^\top \\
&H y_k = s_k
\end{aligned}
$$

Метод BFGS

- Задача

$$
\begin{aligned}
&\min_H \|H_k - H\| \\
\text{s.t. } &H = H^\top \\
&H y_k = s_k
\end{aligned}
$$

- Решение

$$
H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top,
$$

где $\rho_k = \dfrac{1}{y_k^\top s_k}$

Метод BFGS

- Задача

$$\begin{aligned} &\min_H \|H_k - H\| \\ \text{s.t. } &H = H^\top \\ &H y_k = s_k \end{aligned}$$

- Решение

$$H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top,$$

где $\rho_k = \dfrac{1}{y_k^\top s_k}$

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 H_k можно также получить как решение задачи

$$\min_H \operatorname{trace}(H_k^\top H^{-1}) - \log\det(H_k H^{-1}) - n$$

$$\text{s.t. } Hy_k = s_k$$

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

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 $f'(x)$

Квазиньютоновские методы с ограниченной памятью

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

Идея

Использовать последние $m \ll n$ значений (s, y) и корректировать $H_{m,0}$ 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 $f'(x)$
- Значения y и s на первых итерациях могут портить оценку B или H на более поздних итерациях

Идея

Использовать последние $m \ll n$ значений (s, y) и корректировать $H_{m,0}$ 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 $f'(x)$
- Значения y и s на первых итерациях могут портить оценку B или H на более поздних итерациях

Идея

Использовать последние $m \ll n$ значений (s, y) и корректировать $H_{m,0}$ 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

$$H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top, \quad V_k = I - \rho_k y_k s_k^\top$$

Метод L-BFGS

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

$$H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top, \quad V_k = I - \rho_k y_k s_k^\top$$

- Развернём m шагов рекурсии

$$\begin{aligned}
&H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top \\
&= V_k^\top V_{k-1}^\top H_{k-1} V_{k-1} V_k + \rho_{k-1} V_k^\top V_{k-1}^\top s_{k-1} s_{k-1}^\top V_{k-1} V_k + \rho_k s_k s_k^\top \\
&= V_k^\top \dots V_{k-m+1}^\top H_{m,0} V_{k-m+1} \dots V_k \\
&+ \rho_{k-m+1} V_k^\top \dots V_{k-m+2}^\top s_{k-m+1} s_{k-m+1}^\top V_{k-m+2} \dots V_k \\
&+ \dots + \rho_k s_k s_k^\top
\end{aligned}$$

Метод L-BFGS

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

$$H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top, \quad V_k = I - \rho_k y_k s_k^\top$$

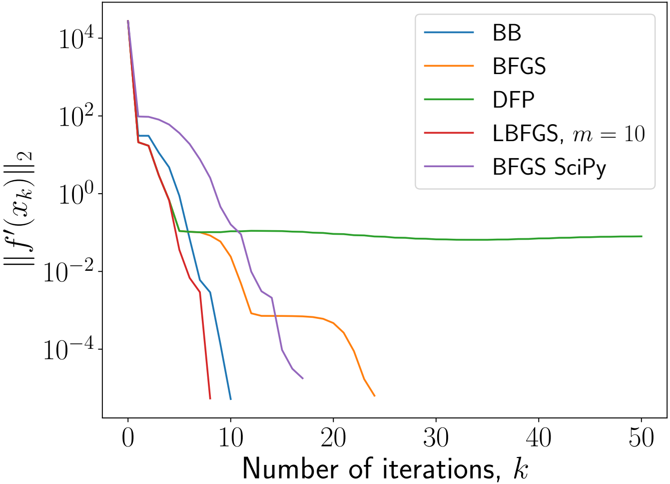
- Развернём m шагов рекурсии

$$\begin{aligned}
&H_{k+1} = V_k^\top H_k V_k + \rho_k s_k s_k^\top \\
&= V_k^\top V_{k-1}^\top H_{k-1} V_{k-1} V_k + \rho_{k-1} V_k^\top V_{k-1}^\top s_{k-1} s_{k-1}^\top V_{k-1} V_k + \rho_k s_k s_k^\top \\
&= V_k^\top \dots V_{k-m+1}^\top H_{m,0} V_{k-m+1} \dots V_k \\
&+ \rho_{k-m+1} V_k^\top \dots V_{k-m+2}^\top s_{k-m+1} s_{k-m+1}^\top V_{k-m+2} \dots V_k \\
&+ \dots + \rho_k s_k s_k^\top
\end{aligned}$$

- Эффективное вычисление $H_k f'(x)$ без явного формирования H_k

Пример

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



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 B_0 или H_0
- Нет разработанной теории сходимости и оптимальности
- Не любой способ выбора шага гарантирует выполнение условия кривизны $y_k^\top s_k > 0$