Методы оптимизации
Лекция 10: Метод Ньютона. Квазиньютоновские методы

Александр Катруца

Факультет инноваций и высоких технологий
Физтех-школа прикладной математики и информатики

24 августа 2018 г.

На прошлой лекции

- Стохастические методы первого порядка

На прошлой лекции

- Стохастические методы первого порядка
  - Стохастический градиентный спуск
  - AdaGrad
  - AdaDelta
  - RMSprop
  - Adam

На прошлой лекции

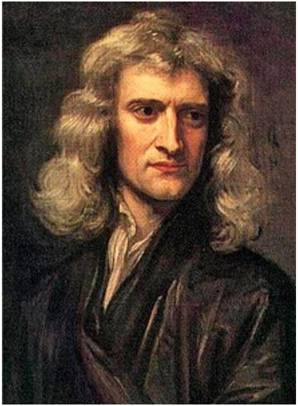
- Стохастические методы первого порядка
  - Стохастический градиентный спуск
  - AdaGrad
  - AdaDelta
  - RMSprop
  - Adam
- Теоретические оценки сходимости

На прошлой лекции

- Стохастические методы первого порядка
  - Стохастический градиентный спуск
  - AdaGrad
  - AdaDelta
  - RMSprop
  - Adam
- Теоретические оценки сходимости
- Стохастические модификации других методов первого порядка

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

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

- 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:* $\|f''(x) - f''(y)\| \le M\|x - y\|$
- *начальная точка* x_0 *достаточно близка к* x^**:*
  $\|x_0 - x^*\| \le \frac{2\mu}{3M}$

Локальная квадратичная сходимость

Теорема

Пусть

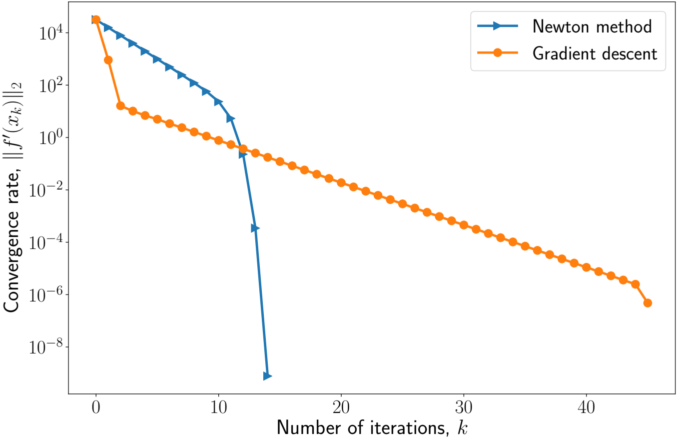
- $f(x)$ *локально сильно выпукла с константой* μ*:*
  $\exists\ x^* :\ f''(x^*) \succeq \mu I$
- *гессиан Липшицев:* $\|f''(x) - f''(y)\| \leq M\|x - y\|$
- *начальная точка* x_0 *достаточно близка к* x^**:*
  $\|x_0 - x^*\| \leq \frac{2\mu}{3M}$

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

Pro & Contra

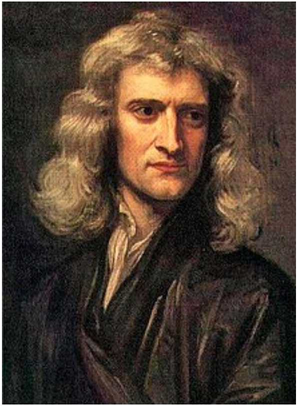
Pro

- 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 $f(x_{k+1})$

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

- Минимум $f_q(h)$ достигается в точке

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

- Квазиньютоновский метод

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

Правила обновления оценки гессиана

- Barzilai-Borwein
- DFP
- BFGS

Метод Barzilai-Borwein

J. Barzilai

J. Borwein

Метод DFP

Метод BFGS

Квазиньютоновские методы с ограниченной памятью

Метод L-BFGS

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

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