Методы оптимизации
Лекция 11: Введение в проксимальные методы

Александр Катруца

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

13 ноября 2018 г.

На прошлой лекции

- Метод Ньютона

На прошлой лекции

- Метод Ньютона
- Квазиньютоновские методы

На прошлой лекции

- Метод Ньютона
- Квазиньютоновские методы
- Метод Barzilai-Borwein

На прошлой лекции

- Метод Ньютона
- Квазиньютоновские методы
- Метод Barzilai-Borwein
- Квазиньютоновские методы с ограниченной памятью

От градиентного спуска к методу проекции градиента

- Градиентный спуск

$$x_{k+1} = \arg\min_x \left(\langle f'(x_k), x - x_k \rangle + \frac{L}{2} \|x - x_k\|_2^2 \right)$$
$$= \arg\min_x \left\| x - \left(x_k - \frac{1}{L} f'(x_k) \right) \right\|_2^2$$

- Пусть есть ограничение вида $x \in C$, тогда

$$x_{k+1} = \arg\min_{x \in C} \left\| x - \left(x_k - \frac{1}{L} f'(x_k) \right) \right\|_2^2 = \pi_C(x_k - \frac{1}{L} f'(x_k))$$

- Получили метод проекции градиента

Проекция и её свойства

Определение

Для данной точки $y \in \mathbb{R}^n$ решение следующей задачи

$$\min_{x \in C} \|x - y\|_2$$

называется проекцией точки y на множество C

Теорема

Если множество C выпукло и замкнуто, то проекция существует и единственна

Нерастягивающий оператор

Определение

Оператор f называется нерастягивающим, если

$$\|f(x) - f(y)\| \leq \|x - y\|,$$

то есть константа Липшица равна 1

Q: как называется оператор с константой Липшица меньше 1?

Теорема

Оператор проекции является нерастягивающим.

Доказательство

1. По свойству проекции, для любой точки y_1

$$\langle y_1 - \pi_C(y_1), x - \pi_C(y_1)\rangle \le 0, \quad \forall x \in C$$

2. Пусть $x = \pi_C(y_2)$, тогда

$$\begin{aligned}\langle y_1 - \pi_C(y_1), \pi_C(y_2) - \pi_C(y_1)\rangle \le 0\\ \langle \pi_C(y_2) - y_2, \pi_C(y_2) - \pi_C(y_1)\rangle \le 0\end{aligned}$$

3. Сложим

$$\langle \pi_C(y_2) - \pi_C(y_1), \pi_C(y_2) - \pi_C(y_1)\rangle \le \langle \pi_C(y_2) - \pi_C(y_1), y_2 - y_1\rangle$$

4. По неравенству КБШ

$$\|\pi_C(y_2) - \pi_C(y_1)\| \le \|y_2 - y_1\|$$

Firmly non-expansiveness

Определение

Оператор f называется firmly non-expansive, если

$$\|f(x) - f(y)\|_2^2 \leq \langle f(x) - f(y), x - y\rangle$$

Теорема

Оператор проекции является firmly non-expansive:

$$\|\pi_C(x) - \pi_C(y)\|_2^2 \leq \langle \pi_C(x) - \pi_C(y), x - y\rangle$$

Сходимость

Лемма

Если x^* решение задачи, то $x^* = \pi_C(x^* - \alpha f'(x^*))$

Доказательство

1. $g^* \equiv f'(x^*)$
2. Общее условие оптимальности $\langle g^*, x - x^* \rangle \geq 0, \ \forall x \in C$
3. Свойство проекции $z = \pi_C(y)$

$$\langle z - y, x - z \rangle \geq 0, \ \forall x \in C$$

4. Подставим $z = x^*$ и $y = x^* - \alpha g^*$ и разделим на $\alpha > 0$

$$\langle g^*, x - x^* \rangle \geq 0$$

- $\|x_{k+1} - x^*\|_2^2 = \|\pi_C(x_k - \alpha f'(x_k)) - \pi_C(x^* - \alpha f'(x^*))\|_2^2 \leq$ $\|x_k - x^* - \alpha(f'(x_k) - f'(x^*))\|_2^2 =$ $\|x_k - x^*\|_2^2 + \alpha^2\|f'(x_k) - f'(x^*)\|_2^2 - 2\alpha\langle f'(x_k) - f'(x^*), x_k - x^*\rangle$
- Если функция выпукла и дифференцируема, а также градиент Липшицев с константой L тогда

$$\langle f'(x) - f'(y), x - y\rangle \geq \frac{1}{L}\|f'(x) - f'(y)\|_2^2$$

- $\|x_{k+1} - x^*\|_2^2 \leq \|x_k - x^*\|_2^2 + \alpha(\alpha - \frac{2}{L})\|f'(x_k) - f'(x^*)\|_2^2$
- $\alpha \leq \frac{2}{L} \to r_{k+1} \leq r_k \leq r_0$

- $f(x_{k+1}) \leq f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle + \frac{1}{2}\|x_{k+1} - x_k\|_2^2 = f(x_k) + \langle f'(x_k), \pi_C(x_k - \alpha f'(x_k)) - \pi_C(x_k) \rangle + \frac{1}{2}\|x_{k+1} - x_k\|_2^2$
- $\|\pi_C(x_k - \alpha f'(x_k)) - \pi_C(x_k)\|_2^2 \leq \langle \pi_C(x_k - \alpha f'(x_k)) - \pi_C(x_k), -\alpha f'(x_k) \rangle$
- $-\alpha^{-1}\|x_{k+1} - x_k\|_2^2 \geq \langle f'(x_k), \pi_C(x_k - \alpha f'(x_k)) - \pi_C(x_k) \rangle$
- $f(x_{k+1}) \leq f(x_k) + \left(\frac{L}{2} - \frac{1}{\alpha}\right)\|x_{k+1} - x_k\|_2^2 = f(x_k) - \frac{L}{2}\|\pi_C(x_k - \alpha f'(x_k)) - x\|_2^2$
- $f_k \geq f_{k+1} + \frac{L}{2}\|\pi_C(x_k - \alpha f'(x_k)) - x_k\|_2^2$
- Складываем

$$f_0 - f^* \geq f_{k+1} - f^* + \frac{L}{2}\sum_{i=0}^{k}\|\pi_C(x_i - \alpha f'(x_i)) - x_i\|_2^2$$

- Левая часть конечна, значит ряд сходится и

$$\lim_{k \to \infty}\|\pi_C(x_k - \alpha f'(x_k)) - x_k\| = 0$$

- Получили сходимость!
- Если использовать неравенства-следствия выпуклости f, то можно показать сходимость вида $\mathcal{O}(1/k)$

Примеры множеств и проекции на них

- $C = \{x \mid a_i \leq x_i \leq b_i\} \to \pi_C(y) = \texttt{clip}_{[a_i, b_i]}(y)$
- $C = \{x \mid Ax = b\} \to \pi_C(y) = y - A^\top (AA^\top)^{-1}(Ay - b)$
- $C = \{x \mid \|x\|_2 \leq 1\} \to \pi_C(y) = y / \|y\|_2$
- $C = \mathbb{S}^n_+ \to \pi_C(V) = \sum\limits_{i=1}^n \max(0, \lambda_i) u_i u_i^\top$

Pro & Contra

Pro

- часто можно аналитически вычислить проекцию
- сходимость аналогична градиентному спуску в безусловной оптимизации
- обобщается на негладкий случай

Contra

- при больших n аналитическое вычисление проекции может быть слишком затратно
- при обновлении градиента может теряться структура задачи

Проксимальный метод

После дискретизации ОДУ

$$\frac{dx}{dt} = -f'(x(t))$$

вида

$$\frac{x_{k+1} - x_k}{\alpha} = -f'(x_{k+1}),$$

получим, что

$$\left(\frac{1}{2\alpha}\|u - x_k\|_2^2 + f(u)\right)'(x_{k+1}) = 0$$

$$x_{k+1} = \arg\min_u \left(f(u) + \frac{1}{2\alpha}\|u - x_k\|_2^2\right) = prox_{\alpha f}(x_k)$$

- Сильно выпуклая функция
- Единственный минимум

Неподвижная точка проксимального оператора

Теорема

Точка x^* является решением задачи тогда и только тогда, когда

$$x^* = prox_f(x^*)$$

Доказательство

1. Если x^* решение задачи, то $f(x) \geq f(x^*)$, тогда
$$f(x) + \frac{1}{2}\|x - x^*\|_2^2 \geq f(x^*) = f(x^*) + \frac{1}{2}\|x^* - x^*\|_2^2$$
2. x^* точка минимума функции $f(x) + \frac{1}{2}\|x - x^*\|_2^2$
3. Следовательно $x^* = prox_f(x^*)$
4. Из критерия первого порядка $\bar{x}$ минимизирует $f(x) + \frac{1}{2}\|x - v\|_2^2$ iff $f'(\bar{x}) + (x - v) = 0$
5. $\bar{x} = v = x^* \to f'(x^*) = 0$

Интерпретация

Пусть f дважды дифференцируемы и сильно выпукла.

- Градиентный спуск как аппроксимация проксимального метода

$$x_{k+1} = prox_{\alpha f}(x_k) = (I+\alpha f')^{-1}(x_k) \approx x_k - \alpha f'(x_k) + o(\alpha), \ \alpha \to 0$$

- Проксимальный метод для аппроксимации второго порядка $\hat{f}(v) = f(x) + \langle f'(x), v - x\rangle + \frac{1}{2}\langle v - x, f''(x)(v - x)\rangle$

$$prox_{\alpha \hat{f}}(x_k) = x_k - (f''(x_k) + (1/\alpha)I)^{-1} f'(x_k)$$

Случай разделения переменных

Если f имеет вид

$$f(x) = \sum_{i=1}^{n} f_i(x_i)$$

тогда

$$prox_f(v)_i = prox_{f_i}(v_i)$$

- Параллельное вычисление компонент
- Консенсусная форма постановки задачи

$$\min \sum_{i=1}^{n} f_i(x) \to \min \sum_{i=1}^{n} f_i(x_i), \text{ s.t. } x_1 = \ldots = x_n$$

Проксимальный градиентный метод (PGM)

Рассмотрим выпуклую функцию $f(x)$ такую что

$$f(x) = h(x) + g(x),$$

- $h(x)$ выпуклая и дифференцируемая
- $g(x)$ выпуклая и может принимать бесконечные значения $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$

Тогда один шаг проксимального градиентного метода

$$x_{k+1} = prox_{\alpha_k g}(x_k - \alpha_k h'(x_k))$$

- Скорость сходимости $\mathcal{O}(1/k)$ для шага $\alpha_k \equiv \alpha \in (0, 1/L]$, где L константа Липшица для градиента f'
- Возможен адаптивный подбор L

Почему это работает?

$$
\begin{aligned}
x_{k+1} &= \arg\min_x \left(g(x) + \frac{1}{2\alpha_k} \|x - (x_k - \alpha_k h'(x_k))\|_2^2 \right) \\
&= \arg\min_x \left(g(x) + h(x_k) + \langle h'(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right)
\end{aligned}
$$

- Аналог градиентного спуска, но для композитной целевой функции
- Для негладкой функции f использование гладкой компоненты h повышает скорость сходимости

Частные случаи

- Если $g(x)$ индикаторная функция некоторого выпуклого множества G

$$g(x) = \begin{cases} 0 & x \in G \\ +\infty & x \notin G \end{cases}$$

  PGM – метод проекции градиента

$$prox_g(y) = \arg\min_{u \in G}(\|y - u\|_2^2) \equiv \pi_G(y)$$

- Если $h \equiv 0$, тогда PGM – проксимальный метод
- Если $g \equiv 0$, тогда PGM – градиентный спуск

Ускоренный PGM (FISTA)

- Сходимость PGM аналогична сходимости градиентного спуска
- Известно, что такую сходимость можно ускорить
- Ускоренный PGM
    1. $y_1 = x_0$, $t_1 = 1$, $k = 1$
    2. $x_k = prox_{\alpha_k g}(y_k - \alpha_k h'(y_k))$
    3. $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
    4. $y_{k+1} = x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1})$
- Сходимость $\mathcal{O}(1/k^2)$
- Доказательство занимает страницу, см тут

Пример

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$$

- $h(x) = \frac{1}{2}\|Ax-y\|_2^2$
- $g(x) = \gamma\|x\|_1$

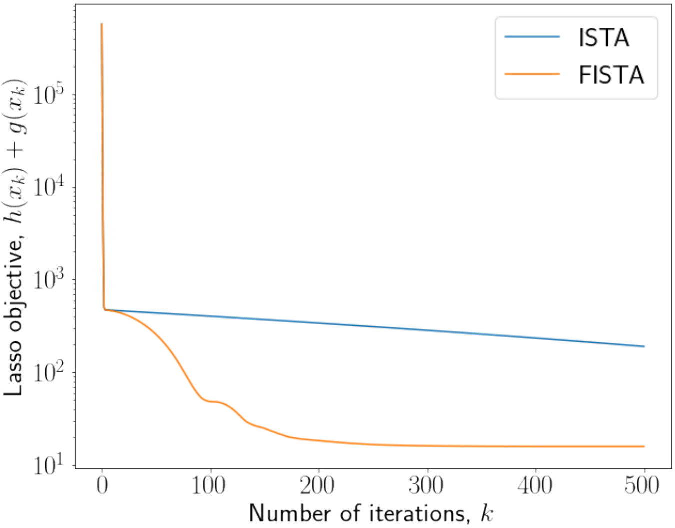
Soft thresholding

- $prox_{\alpha\|x\|_1}(x)_i = prox_{\alpha|\cdot|}(x_i)$
- $prox_{\alpha|\cdot|}(x_i) = \arg\min\limits_u(|u| + 1/(2\alpha)(x_i-u)^2)$
- Аналитическое решение

$$prox_{\alpha|\cdot|}(x_i) = \begin{cases} x_i - \alpha & x_i \geq \alpha \\ 0 & |x_i| \leq \alpha \\ x_i + \alpha & x_i \leq -\alpha \end{cases}$$

- Векторизованный ответ
  $prox_{\alpha\|x\|_1}(x) = \mathsf{sign}(x)(|x|-\alpha)_+,\ |x| = (|x_1|, \dots, |x_n|)$

Сравнение



ISTA
FISTA
10^5
10^4
10^3
10^2
10^1
Lasso objective, $h(x_k) + g(x_k)$
0
100
200
300
400
500
Number of iterations, k

Направления исследований

- Неевклидовы расстояния в качестве прокс-слагаемого. Зеркальный спуск
- Невыпуклые задачи определённой структуры
- Параллельные и распределённые реализации
- Различные способы дискретизаций ОДУ дают новые методы оптимизации, подробности тут