Математическое моделирование и методы оптимизации
Часть 3.

Тремба Андрей Александрович
Москва, к.ф.-м.н., с.н.с.
Институт проблем управления РАН

21 июня 2019 г., Иннополис

Третья часть

Методы второго порядка
и промежуточные методы

- Вторые производные
- Метод Ньютона и квази-ньютоновские методы
- Метод тяжёлого шарика
- Метод сопряжённых градиентов
- Ускоренные методы (Нестерова)

Было

1. Выпуклость
2. Направление + шаг
3. Выбор направлений: от покоординатного спуска до антиградиентного
4. Субградиенты
5. Необходимые и достаточные условия оптимальности

Производные высшего порядка

Дважды дифференцируемые функции

Если для любых $z \in \mathbb{R}^n$ выполняется:

$$f(x+z) = f(x) + (\nabla f(x), z) + \frac{1}{2}(Hz, z) + o(\|z\|^2)$$

То $H(x)$ обозначается как $\nabla^2 f(x)$,

и называется «второй производной» f.

Для конечномерного случая: матрица Гессе.

$$f(x+z) = f(x) + (\nabla f(x))^T z + \frac{1}{2} z^T H z + o(\|z\|^2)$$

Производные векторных функций

$$P(x), \quad P : \mathbb{R}^n \to \mathbb{R}^m$$

В конечномерном случае производная является матрицей Якоби ($J_P(x)$ – для P!)

$$P'(x) = \begin{bmatrix} \frac{\partial P_1(x)}{\partial x_1} & \frac{\partial P_1(x)}{\partial x_2} & \dots & \frac{\partial P_1(x)}{\partial x_n} \\ \frac{\partial P_2(x)}{\partial x_1} & \frac{\partial P_2(x)}{\partial x_2} & \dots & \frac{\partial P_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial P_m(x)}{\partial x_1} & \frac{\partial P_m(x)}{\partial x_2} & \dots & \frac{\partial P_m(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

с разложением в ряд Тейлора

$$P(x+z) = P(x) + P'(x)z + \overline{o}(\|z\|) \in \mathbb{R}^n$$

Градиент как векторная функция

Матрица Якоби градиента функции

$$\left(\nabla f(x)\right)' = \begin{bmatrix} \dfrac{\partial^2 f(x)}{\partial x_1^2} & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(x)}{\partial x_n \partial x_1} & \dfrac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

Совпадает с матрицей Гессе этой же функции (доказывается)

$$\left(\nabla f(x)\right)' = \nabla^2 f(x)$$

Повторная формула Ньютона-Лейбница

Для градиента

$$\nabla f(x+z) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\tau z) z \, d\tau$$

Повторно - для функции

$$f(x+z) = f(x) + (\nabla f(x), x) + \int_0^1 \int_0^t \left(\nabla^2 f(x+\tau z) z, z\right) d\tau dt$$

Теорема о среднем значении

$\exists \hat{\tau} \in [0,1] : f(x+z) =$
$f(x) + (\nabla f(x), z) + \left(\nabla^2 f(x+\hat{\tau} z) z, z\right)$

Свойства производных (включая векторные функции)

Правила вычисления производных

- «линейность»

$$\Big(aP(x)\Big)' = aP'(x), \quad a \in \mathbb{R}$$

$$\Big(P(x) + S(x)\Big)' = P'(x) + S'(x)$$

Следствие:

$$\nabla^2(af(x)) = a\nabla^2 f(x)$$

$$\nabla^2(f(x) + g(x)) = \nabla^2 f(x) + \nabla^2 g(x))$$

Вычисление производных сложных функций

- Скалярные функции $g : \mathbb{R} \to \mathbb{R}, \quad h : \mathbb{R} \to \mathbb{R}$

$$h\big(g(x)\big)'_x = h'\big(g(x)\big) \cdot g'(x)$$

- Внутренняя векторная функция, внешняя - скалярная

  $g : \mathbb{R}^n \to \mathbb{R}^m, \quad h : \mathbb{R}^m \to \mathbb{R}$

$$\nabla_x h\big(g(x)\big) = \big(g'(x)\big)^T \nabla_y h\big(g(x)\big)$$

- Composition rule (vector)

  $g : \mathbb{R}^n \to \mathbb{R}^m, \quad h : \mathbb{R}^m \to \mathbb{R}^s$

$$h\big(g(x)\big)'_x = h'\big(g(x)\big) g'(x)$$

«Правило» для производной сложной функции

В случае стандартного скалярного произведения $(a, b) = a^T b$

$$f(x) = h\big(g(x)\big)$$

$$J_f(x) = J_h\big(g(x)\big) \cdot J_g(x)$$

Для функции, зависящей от нескольких переменных $p : \mathbb{R}^n \to \mathbb{R}$

$$J_p(x) \equiv \big(\nabla p(x)\big)^T \in \mathbb{R}^{1 \times n}$$

Где используется матрица Гессе (вторая производная)

Условия оптимальности второго порядка

Необходимое и достаточное условие строгого локального минимума: Если в x^* выполняются

$$\nabla f(x^*) = 0$$

и

$$\nabla^2 f(x^*) \succ 0$$

то x^* – строгий (изолированный) локальный минимум.

Условия оптимальности второго порядка

Необходимое и достаточное условие строгого локального минимума: Если в x^* выполняются

$$\nabla f(x^*) = 0$$

и

$$\nabla^2 f(x^*) \succ 0$$

то x^* – строгий (изолированный) локальный минимум.

Верно и для невыпуклых функций.

Алгоритм решения задач оптимизации

для дифференцируемых функций

1. Решить $\nabla f(x) = 0$.
2. Выбрать подходящие $x : \nabla^2 f(x^*) \succ 0$

Выпуклые функции (эквивалентные определения)

1. $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$
2. $f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2}$
3. $f(y) \geq f(x) + (\nabla f(x), y - x)$
4. $(\nabla f(x) - \nabla f(y), x - y) \geq 0$
5. $\nabla^2 f(x) \succcurlyeq 0$

Сильная выпуклость

Константа $\mu > 0$.

1. $f(\alpha x + (1-\alpha)y) \leq$
   $\alpha f(x) + (1-\alpha) f(y) - \mu \frac{\alpha(1-\alpha)}{2} \|x - y\|^2$
2. $f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2} - \frac{\mu}{8} \|x - y\|^2$
3. $f(y) \geq f(x) + (\nabla f(x), y - x) + \frac{\mu}{2} \|x - y\|^2$
4. $(\nabla f(x) - \nabla f(y), x - y) \geq \mu \|x - y\|^2$
5. $\nabla^2 f(x) \succcurlyeq \mu I$

Аналогии с константой Липшица градиента

1.
2.
3.
4. $f(y) \leq f(x) + (\nabla f(x), z - x) + \frac{L}{2}\|x - y\|^2$
5. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
6. $\nabla^2 f(x) \preccurlyeq LI$

Метод Ньютона

Метод Ньютона: решение уравнений и в оптимизации

Идея: необходимое условие оптимальности $\rightarrow$ решить $\nabla f(x) = 0$.

Линеаризация и итерации

$$\nabla f(x^{k+1}) \approx \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)$$

Метод Ньютона: решение уравнений и в оптимизации

Идея: необходимое условие оптимальности $\to$ решить $\nabla f(x) = 0$.

Линеаризация и итерации

$$\nabla f(x^{k+1}) \approx \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0$$

Метод Ньютона: решение уравнений и в оптимизации

Идея: необходимое условие оптимальности $\to$ решить $\nabla f(x) = 0$.

Линеаризация и итерации

$$\nabla f(x^{k+1}) \approx \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0$$

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

Условия – невырожденность $\nabla^2 f(x^*)$).

Квадратичная, но локальная сходимость.

Демпфированный метод Ньютона

Ньютоновское направление

$$-\left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

(Демпфированный) метод Ньютона

$$x^{k+1} = x^k - \gamma_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

Методы скорейшего спуска и Ньютона

Градиентный метод

$$x^{k+1} = x^k - \gamma_k \quad \nabla f(x^k)$$

(Демпфированный) метод Ньютона

$$x^{k+1} = x^k - \gamma_k \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)$$

Скорости сходимости

Гарантированная* скорость

1. Сублинейная
   - $\varepsilon = O(\frac{1}{\sqrt{N}})$ - субградиентный метод
   - $\varepsilon = O(\frac{1}{N})$ - градиентный метод, одномерная минимизация
2. Линейная $\varepsilon = O(q^N)$ - градиентный метод для сильно выпуклых функций, одномерная минимизация для выпуклых функций
3. квадратичная $\varepsilon = O(q^{(2^N)})$ - метод Ньютона

Минимизация квадратичной функции

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}(Ax, x) + (b, x) = \frac{1}{2}x^T Ax + b^T x$$

$$A = A^T,\ \mu = \lambda_1(A),\ L = \lambda_n(A)$$

$$x^{k+1} = x^k - \gamma_k(Ax^k + b)$$

Скорость сходимости

$$\|x^{N+1} - x^*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^N \|x^0 - x^*\|$$

Метод Ньютона

$$x^{k+1} = x^k - \gamma_k A^{-1}b$$

Промежуточные заметки

- Метод Ньютона инвариантен к смене координат
- В некоторых координатах градиентный метод совпадает с методом Ньютона

Другая точка зрения:

- Переменная метрика (скалярное произведение)

Скалярные произведения ($R \succ 0$)

$$(x, y)_R \doteq (Rx, y) = x^T Ry$$

$$\begin{aligned} f(x+z) &= f(x) + (\nabla_{(R)} f(x), z)_R + o(\|z\|) \\ &= f(x) + (R\nabla_{(R)} f(x), z) + o(\|z\|) \end{aligned}$$

или

$$\begin{aligned} f(x+z) &= f(x) + (\nabla f(x), z) + o(\|z\|) \\ &= f(x) + (H^{-1}\nabla f(x), z)_H + o(\|z\|) \end{aligned}$$

Градиентные методы:

$$x^{k+1} = x^k - \gamma \nabla f(x)$$

$$x^{k+1} = x^k - \gamma \nabla_{(H)} f(x)$$

Скалярные произведения ($R \succ 0$)

$$(x, y)_R \doteq (Rx, y) = x^T Ry$$

$$\begin{aligned} f(x+z) &= f(x) + (\nabla_{(R)} f(x), z)_R + o(\|z\|) \\ &= f(x) + (R\nabla_{(R)} f(x), z) + o(\|z\|) \end{aligned}$$

или

$$\begin{aligned} f(x+z) &= f(x) + (\nabla f(x), z) + o(\|z\|) \\ &= f(x) + (H^{-1}\nabla f(x), z)_H + o(\|z\|) \end{aligned}$$

Градиентные методы:

$$x^{k+1} = x^k - \gamma \nabla f(x)$$

$$x^{k+1} = x^k - \gamma \nabla_{(H)} f(x) = x^k - \gamma H^{-1} \nabla f(x)$$

Препятствия и идеи

Сложности реализации (демпфированного метода Ньютона):

- матрица Гессе недоступна
- сложности с вычислением/хранением

Препятствия и идеи

Сложности реализации (демпфированного метода Ньютона):

- матрица Гессе недоступна
- сложности с вычислением/хранением

Идеи:

- Иметь приближение H матрицы $\nabla^2 f(x)$
- Приближённые вычисления

Замечание: в Википедии обозначения инвертированы! $\cancel{B \approx \nabla^2 f(x), H \approx (\nabla^2 f(x))^{-1}}$

Методы переменной метрики (квазиньютоновские методы)

$$x^{k+1} = x^k - \gamma \nabla^2 (f(x^k))^{-1} \nabla f(x^k)$$

Аппроксимировать $\nabla^2 f(x^k)$ матрицей H_k:

$$x^{k+1} = x^k - \gamma H_k^{-1} \nabla f(x^k)$$

(То же, что выбрать специальную метрику $\| \cdot \|_H = \sqrt{(\cdot, \cdot)_{H_k}}$)

Либо сразу аппроксимировать обратную к матрице Гессе как B_k:

$$x^{k+1} = x^k - \gamma B_k \nabla f(x^k)$$

Квазиньютоновские методы: условие

$$H \approx \nabla^2 f(x), B \approx (\nabla^2 f(x))^{-1}$$

Итеративное вычисление B_k (или H_k)

$$\{x^k, B_k, \nabla f(x^k)\} \to \{x^{k+1}, \nabla f(x^{k+1})\} \to B_{k+1}$$

Квазиньютоновское условие (одна форма)

$$H_{k+1}(\nabla f(x^{k+1}) - \nabla f(x^k)) = x^{k+1} - x^k$$

Квазиньютоновское условие (другая форма)

$$\nabla f(x^{k+1}) - \nabla f(x^k) = B_{k+1}(x^{k+1} - x^k)$$

Например $B_{k+1} = B_k + \alpha u u^T + \beta v v^T$

DFP, BFGS ($H_{k+1}d = s$, $d = B_{k+1}s$)

$$s \doteq x^{k+1} - x^k, \quad d \doteq \nabla f(x^{k+1}) - \nabla f(x^k)$$

- Метод Давидона-Флетчера-Пауэлла

$$H_{k+1} = \left(I - \frac{sd^T}{(d, s)}\right) H_k \left(I - \frac{ds^T}{(d, s)}\right) + \frac{dd^T}{(d, s)}$$

$$B_{k+1} = B_k + \frac{dd^T}{(d, s)} - \frac{B_k s s^T B_k}{(B_k s, s)}$$

- Метод Бройдена-Флетчера-Гольдфарба-Шенно (BFGS)

$$H_{k+1} = H_k + \frac{dd^T}{(d, s)} - \frac{H_k s s^T H_k}{(H_k s, s)}$$

$$B_{k+1} = \left(I - \frac{ds^T}{(d, s)}\right) B_k \left(I - \frac{sd^T}{(d, s)}\right) + \frac{dd^T}{(d, s)}$$

Выводы по квазиньютоновским методам

- Для выпуклых функций $H_k \to \nabla^2 f(x^k)$
- H_k и B_k как «настройка метрики»
- «Подправленный» градиентный спуск
$$x^{k+1} = x^k - \gamma_k B_k \nabla f(x^k)$$
- L-BGFS
- Периодические рестарты: $H_k = I$ $(B_k = I)$

Хорошо, а можно совсем без матриц?
Что не так с градиентным методом?

Анализ градиентного метода

Эффективность градиентного метода $\gamma = \frac{2}{\mu+L}$

$$\|x^{k+1} - x^*\| \leq \left(\frac{L-\mu}{L+\mu}\right) \|x^k - x^*\|$$

Общая константа

$$\varkappa \doteq \frac{L}{\mu} = \frac{\sup_x eig_n(\nabla^2 f(x))}{\inf_x eig_1(\nabla^2 f(x))}$$

$\|x^k - x^*\| \leq c\left(\frac{\varkappa-1}{\varkappa+1}\right)^k$

К точке - число обусловленности

$$\varkappa_x = \frac{eig_n(\nabla^2 f(x))}{eig_1(\nabla^2 f(x))} = \frac{\|\nabla^2 f(x^k)\|}{\|\nabla^2 f(x^k)^{-1}\|}$$

Плохой и лучший случаи

Для квадратичной функции:

- $\varkappa \gg 1$: зиг-загообразный шаг
- «наилучшее» число обусловленности $\varkappa = 1$ для $\nabla^2 f(x) = I \Rightarrow$ «мгновенная» сходимость

Причины

- Градиентный спуск локален и ничего не «запоминает»
- Метод Ньютона обладает «дополнительной (глобальной) информацией»

Метод тяжёлого шарика

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})$$

- $0 \leq \beta < 1, \quad 0 < \alpha < \frac{2(1+\beta)}{L}$
- $\alpha = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}, \quad \beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{(\sqrt{L}+\sqrt{\mu}}\right)^2$

Скорость сходимости вблизи минимума

$$\|x^{k+1} - x^*\| \approx c \left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right)^k \|x^0 - x^*\|$$

Метод сопряжённых градиентов

$$p^k = -\nabla f(x^k) + \beta_k p^{k-1}$$
$$x^{k+1} = x^k + \alpha_k p^k$$

$$\alpha_k = \arg\min_{\alpha} f(x^k + \alpha p^k)$$

$$\beta_k = 0, \quad (k \bmod n) = 0$$

Скрость сходимости: n шагов СГ $\approx$ 1 шаг метода Ньютона

Варианты метода сопряжённых градиентов

$$p^k = -\nabla f(x^k) + \beta_k p^{k-1}$$

- Флетчер-Ривз

$$\beta_k = \frac{\left(\nabla f(x^k), \nabla f(x^k)\right)}{\left(\nabla f(x^{k-1}), \nabla f(x^{k-1})\right)} = \frac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^{k-1})\|^2}$$

- Полак-Рибьер-Поляк

$$\beta_k = \frac{\left(\nabla f(x^k), \nabla f(x^k) - \nabla f(x^{k-1})\right)}{\left(\nabla f(x^{k-1}), \nabla f(x^{k-1})\right)}$$

Многошаговые градиентные методы

- Метод тяжёлого шарика

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta (x^k - x^{k-1})$$

- Метод сопряжённых градиентов

$$\begin{aligned} p^k &= -\nabla f(x^k) + \beta_k p^{k-1} \\ x^{k+1} &= x^k + \gamma_k p^k \end{aligned}$$

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k) + \gamma_k \frac{\beta_k}{\gamma_{k-1}} (x^k - x^{k-1})$$

с периодическими рестартами

- Инерционные алгоритмы

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Быстрый градиентный метод (идеи)

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k(x^k - x^{k-1})$$

- оптимизировать по α_k, β_k
- задать α_k, β_k как функции от k
- придумать что-то ещё

Быстрый градиентный метод (Нестеров)

Экстраградиентная форма

$$y^0 = x^0$$

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$

$$y^{k+1} = x^{k+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x^{k+1} - x^k)$$

Скорость сходимости:

$$\|x^{k+1} - x^*\| \approx c\left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right)^k \|x^0 - x^*\|$$

Быстрый градиентный метод

$$y^0 = x^0, t_0 = 1$$

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

$$y^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}}(x^{k+1} - x^k)$$

Быстрый градиентный метод («подобные треугольники»)

$$v^0 = x^0$$
$$y^k = (1-\theta_k)x^k + \theta_k v^k$$
$$v^{k+1} = v^k - \frac{1}{L\theta_k}\nabla f(y^k)$$
$$x^{k+1} = (1-\theta_k)x^k + \theta_k v^{k+1}$$

$$\theta_k = \frac{2}{k+2}$$

В чём «быстрый» градиентный метод быстрее?

«Оптимистическая» скорость сходимости

- Субградиентный метод (шаг Поляка и $\operatorname{int} X^* \neq 0$): $f(x^k) - f^* \leq 0, k \geq k^*$
$$N(\varepsilon) = O\left(1\right)$$
- Градиентный метод (сильная выпуклость):
$$f(x^k) - f^* \leq O(\left(\frac{\varkappa - 1}{\varkappa + 1}\right)^k)$$
- Быстрый градиентный метод:
$$f(x^k) - f^* \leq O(\left(\frac{\sqrt{\varkappa} - 1}{\sqrt{\varkappa} + 1}\right)^k)$$

«Пессимистическая» скорость сходимости

- Субградиентный метод (негладкий случай): $f(x^k) - f^* \leq O(\frac{1}{\sqrt{k}})$
$$N(\varepsilon) = O\left(\frac{1}{\varepsilon^2}\right)$$
- Градиентный метод: $f(x^k) - f^* \leq O(\frac{1}{k})$
$$N(\varepsilon) = O\left(\frac{1}{\varepsilon}\right)$$
- Быстрый градиентный метод $f(x^k) - f^* \leq O(\frac{1}{k^2})$
$$N(\varepsilon) = O\left(\frac{1}{\sqrt{\varepsilon}}\right)$$

Другие подходы

- Методы с подстройкой расстояния (зеркальный спуск)
- Адаптивные методы (Adagrad, Adam, ...)
- Стохастические методы (SGD)
- Структурированные функции: гладкая + негладкая и т.п.
- Техники глобальной оптимизации
  - Метод ветвей и границ
  - Рандомизация
  - Мультистарт
  - Частицы, «стаи»

Выводы по безусловной оптимизации

- Чем больше информации, тем лучше.

- Метод золотого сечения, метод Нилдера-мида, покоординатный спуск, градиентный метод, случайный поиск, метод Ньютона, метод тяжёлого шарика, метод сопряжённых градиентов, квази-ньютоновские методы, быстрые градиентные методы.