

# Методы оптимизации, ФКН ВШЭ, зима 2017

## Семинар 7: Квазиньютоновские методы

21 февраля 2017 г.

## 1 Квазиньютоновские методы

### 1.1 Мотивация

Рассмотрим стандартную задачу гладкой безусловной оптимизации:

$$\min_{x \in \mathbb{R}^n} f(x), \quad (1)$$

где  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  — дважды непрерывно дифференцируемая функция.

Напомним, что методы спуска для решения задачи (1) итеративно выполняют обновления следующего вида:

$$x_{k+1} := x_k + \alpha_k d_k,$$

где  $d_k \in \mathbb{R}^n$  — направление спуска для функции  $f$  в точке  $x_k$ , а  $\alpha_k \geq 0$  — длина шага, настраиваемая с помощью линейного поиска.

Случай  $d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$  соответствует<sup>1</sup> *методу Ньютона*. Как известно, метод Ньютона имеет хорошую скорость сходимости (по крайней мере, на финальной стадии своей работы), однако является слишком дорогостоящим — на каждой итерации метода нужно:

1. вычислять матрицу-гессиан  $\nabla^2 f(x_k)$ ;
2. решать систему линейных уравнений  $\nabla^2 f(x_k) d = -\nabla f(x_k)$  (для матриц  $\nabla^2 f(x_k)$  общего вида сложность этой операции составляет  $O(n^3)$ ).

В зависимости от функции  $f$ , наиболее дорогостоящей операцией из двух перечисленных может оказаться либо само вычисление гессиана (например, для функции логистической регрессии), либо решение соответствующей системы линейных уравнений (например, для квадратичной функции). В любом случае, сложность итерации метода Ньютона для многих функций  $f$  составляет как минимум  $O(n^3)$ . Из-за этого метод Ньютона невозможно применять для многих задач больших (или даже средних) размеров.

Хотелось бы иметь методы, которые сочетают в себе, с одной стороны, высокую скорость сходимости метода Ньютона и, с другой стороны, имеют менее дорогостоящие итерации. Согласно вышесказанному, для таких методов естественно ввести следующие два ограничения:

1. ни на одной итерации нельзя вычислять матрицу-гессиан  $\nabla^2 f(x_k)$ ;
2. сложность итерации должна быть максимум  $O(n^2)$  (т. е. не должно быть никаких обращений матриц или решений систем линейных уравнений общего вида).

Такие методы были разработаны еще в 60–70-х годах прошлого века и известны под названием *квазиньютоновских методов* или *методов переменной метрики*.

---

<sup>1</sup>Здесь предполагается, что гессиан  $\nabla^2 f(x_k)$  является положительно определенной матрицей, и поэтому  $d_k$ , действительно, задает направление спуска. Если гессиан не является положительно определенным, то вместо него в этой формуле должна стоять его модифицированная версия.

## 1.2 Основная идея

В квазиньютоновских методах направление спуска  $d_k$  полагается равным  $-H_k \nabla f(x_k)$ , где  $H_k \in \mathbb{S}_{++}^n$  — некоторая матрица. Если бы матрица  $H_k$  в точности равнялась обратному гессиану  $[\nabla^2 f(x_k)]^{-1}$ , то получился бы метод Ньютона. Вместо этого в квазиньютоновских методах матрица  $H_k$  всего лишь «аппроксимирует» обратный гессиан — отсюда и название *квази-ньютоновские* методы.

Матрицы  $H_k$  строятся таким образом, чтобы в пределе обеспечить (по крайней мере, в хороших случаях) аппроксимацию истинного обратного гессиана:

$$H_k - [\nabla^2 f(x_k)]^{-1} \rightarrow 0 \quad \text{при } k \rightarrow \infty.$$

Таким образом, квазиньютоновские методы асимптотически приближаются к методу Ньютона, что гарантирует их сверхлинейную сходимость.

## 1.3 Правила обновления матриц

Матрицы  $H_k$  в квазиньютоновских методах строятся следующим образом. На самой первой итерации выбирается некоторая начальная матрица  $H_0 \in \mathbb{S}_{++}^n$  (обычно  $H_0 = I_n$ ). Далее матрицы  $H_k$  обновляются в итерациях — на каждой итерации метода

1. вычисляется новая точка  $x_{k+1} := x_k - \alpha_k H_k \nabla f(x_k)$ ;
2. выполняется обновление матрицы:  $H_k \rightarrow H_{k+1}$ .

Все квазиньютоновские методы отличаются между собой лишь способом обновления матрицы  $H_k$ . Напомним, что, согласно наложенным выше ограничениям, процедура обновления матрицы  $H_k$  должна быть достаточно эффективной: сложность этой процедуры должна составлять  $O(n^2)$ , и при этом не должно быть никаких вычислений гессиана. Таким образом, процедура обновления матрицы  $H_k$  должна каким-то эффективным образом «подгрузить» новую информацию о гессиане, при этом не вычисляя сам гессиан. Для этого используются градиенты  $\nabla f(x_k)$  и  $\nabla f(x_{k+1})$ , а также следующее правило.

### Квазиньютоновское правило

Выбрать новую матрицу  $H_{k+1}$  таким образом, чтобы выполнялось *уравнение секущей*:

$$H_{k+1}[\nabla f(x_{k+1}) - \nabla f(x_k)] = x_{k+1} - x_k. \quad (2)$$

Смысл уравнения секущей состоит в следующем. Если  $f$  — квадратичная функция  $f(x) := (1/2)\langle Ax, x \rangle - \langle b, x \rangle$ , где  $A \in \mathbb{S}_{++}^n$ ,  $b \in \mathbb{R}^n$ , тогда  $\nabla f(x) = Ax - b$ , и для всех  $x, y \in \mathbb{R}^n$  выполняется  $\nabla f(x) - \nabla f(y) = A(x - y)$ , т. е. обратный гессиан  $A^{-1}$  удовлетворяет уравнению секущей. Для произвольной функции  $f$  (не обязательно квадратичной) справедлива формула Ньютона–Лейбница:

$$\nabla f(x) - \nabla f(y) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x) dt = \left[ \int_0^1 \nabla^2 f(x + t(y - x)) dt \right] (y - x),$$

для всех  $x, y \in \mathbb{R}^n$ . Значит, уравнению секущей (2) удовлетворяет обратный средний гессиан  $[\int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k)) dt]^{-1}$  на отрезке между  $x_k$  и  $x_{k+1}$ .

Таким образом, некоторый гессиан (например, средний) уравнению секущей удовлетворяет. Тем не менее, само по себе уравнение секущей (2) описывает матрицу  $H_{k+1}$  однозначно лишь в одномерном случае  $n = 1$ . Действительно, уравнение (2) задает систему линейных уравнений относительно элементов матрицы  $H_{k+1}$ . Количество неизвестных в этой системе равно  $n(n+1)/2$  (поскольку матрица симметричная), а число уравнений равно  $n$ . Таким образом, при  $n > 1$  система является недоопределенной и имеет бесконечно много решений. (Даже если дополнительно наложить требование положительной определенности, то по-прежнему матрица  $H_{k+1}$  определяется неоднозначно.)

Итак, различных способов обновления матрицы  $H_k$ , обеспечивающих выполнение уравнения секущей (2) существует бесконечно много. Приведем несколько стандартных схем из литературы по квазиньютоновским методам, которые получены из различных «естественных» соображений и обычно считаются наиболее эффективными.

**Наиболее популярные схемы обновления квазиньютоновских матриц**

Обозначим  $s_k := x_{k+1} - x_k$  и  $y_k := \nabla f(x_{k+1}) - \nabla f(x_k)$ . Любая из следующих схем обновления матрицы  $H_k$  гарантирует выполнение уравнения секущей (2).

1. *Симметричная коррекция ранга 1 (SR1):*

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{\langle s_k - H_k y_k, y_k \rangle}.$$

2. *Схема Давидона–Флетчера–Пауэлла (DFP):*

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{\langle H_k y_k, y_k \rangle} + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}.$$

3. *Схема Бroyдена–Флетчера–Гольдфарба–Шанно (BFGS):*

$$H_{k+1} = \left( I_n - \frac{s_k y_k^T}{\langle y_k, s_k \rangle} \right) H_k \left( I_n - \frac{y_k s_k^T}{\langle y_k, s_k \rangle} \right) + \frac{s_k s_k^T}{\langle y_k, s_k \rangle}.$$

Наиболее эффективной и стабильной схемой считается BFGS.

Отметим, что процедура обновления SR1 из-за присутствия в знаменателе множителя  $\langle s_k - H_k y_k, y_k \rangle$ , вообще говоря, не сохраняет положительную определенность матрицы  $H_k$ : если матрица  $H_k$  положительно определенная, то новая матрица  $H_{k+1}$  может таковой не оказаться. Положительную определенность сохраняют процедуры обновления DFP и BFGS (при некоторых небольших предположениях).

Иногда помимо матриц  $H_k$ , аппроксимирующих обратный гессиан, бывает полезно иметь матрицы  $B_k := H_k^{-1}$ , аппроксимирующие сам гессиан. Формулы пересчета этих матриц  $B_k$  можно получить из приведенных выше формул пересчета обратных матриц  $H_k$  с помощью следующего полезного утверждения.

**Утверждение 1** (Формула Шермана–Моррисона). Пусть  $A \in \mathbb{R}^{n \times n}$  — квадратная матрица, и пусть  $u, v \in \mathbb{R}^n$  — векторы. Предположим, что матрица  $A$  обратима, и  $1 + \langle A^{-1}u, v \rangle \neq 0$ . Тогда матрица  $(A + uv^T)$  также обратима, и

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + \langle A^{-1}u, v \rangle}.$$

## 2 Задачи

Начнем с вывода формулы обновления Бroyдена, которая используется в одном из возможных подходов к получению формулы обновления BFGS.

**Задача 1.** Пусть  $A \in \mathbb{R}^{n \times n}$  — матрица (не обязательно симметричная). Пусть  $u, v \in \mathbb{R}^n$  — векторы, причем  $u \neq 0$ . Рассмотрим следующую матричную задачу оптимизации:

$$\min_{X \in \mathbb{R}^{n \times n}} \{ \|X - A\|_F : Xu = v \}.$$

Найдите решение этой задачи в явном виде.

**Решение.** Перейдем от исходной негладкой задачи к эквивалентной ей гладкой; также для удобства добавим множитель  $1/2$ :

$$\min_{X \in \mathbb{R}^{n \times n}} \left\{ \frac{1}{2} \|X - A\|_F^2 : Xu = v \right\}.$$

Эта задача является условной задачей оптимизации с ограничениями вида аффинных равенств, и может быть решена с помощью *правила множителей Лагранжа*.

Запишем функцию Лагранжа  $\mathcal{L} : \mathbb{R}^{n \times n} \times \mathbb{R}^n \rightarrow \mathbb{R}$  для рассматриваемой задачи:

$$\mathcal{L}(X; \mu) := \frac{1}{2} \|X - A\|_F^2 - \langle \mu, Xu - v \rangle.$$

Согласно правилу множителей Лагранжа, матрица  $X \in \mathbb{R}^{n \times n}$  является решением задачи тогда и только тогда, когда она является допустимой, т. е.  $Xu = v$ , и существует  $\mu \in \mathbb{R}^n$ , такое, что  $\nabla \mathcal{L}(X; \mu) = 0$ .

Вычислим градиент функции Лагранжа. Для этого сперва найдем дифференциал:

$$d\mathcal{L}(X; \mu) = \frac{1}{2} d\|X - A\|_F^2 - \langle \mu, (dX)u \rangle = \{d\|X\|_F^2 = 2\langle X, dX \rangle\} = \langle X - A, dX \rangle - \langle \mu u^T, dX \rangle.$$

Отсюда градиент равен

$$\nabla \mathcal{L}(X; \mu) = X - A - \mu u^T.$$

Из условия  $\nabla \mathcal{L}(X; \mu) = 0$  выразим  $X$  через  $\mu$ :

$$X = A + \mu u^T.$$

Осталось найти  $\mu$ . Для этого подставим полученное выражение для  $X$  в условие допустимости  $Xu = v$ :

$$Au + \|u\|_2^2 \mu = v.$$

Отсюда

$$\mu = \frac{v - Au}{\|u\|_2^2}.$$

Подставляя эту формулу в полученное выше выражение для  $X$ , окончательно получаем

$$X = A + \frac{(v - Au)u^T}{\|u\|_2^2}.$$

В литературе эта формула известна как *формула обновления Бройдена*.

Следующие три задачи направлены на исследование вопроса о сохранении положительной определенности в процедурах обновления DFP и BFGS.

Заметим, что из уравнения текущей (2) следует следующее *необходимое* условие положительной определенности: если матрица  $H_{k+1}$  положительно определенная, то  $\langle y_k, s_k \rangle > 0$ . Таким образом, если на некоторой итерации оказалось, что  $\langle y_k, s_k \rangle \leq 0$ , то матрица  $H_{k+1}$  никак положительно определенной быть не может. Покажем, что при определенных предположениях такая ситуация никогда не произойдет.

**Задача 2.** Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  — непрерывно дифференцируемая функция. Покажите, что если функция  $f$  является строго выпуклой, то ее градиент является *строго монотонным оператором*, т. е.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0 \tag{3}$$

для всех  $x, y \in \mathbb{R}^n$ ,  $x \neq y$ .

*Подсказка.* Для непрерывно-дифференцируемой функции  $f$  строгая выпуклость эквивалентна тому, что  $f$  может быть глобально ограничена снизу касательной, построенной к графику функции в любой точке, т. е.

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle \tag{4}$$

для всех  $x, y \in \mathbb{R}^n$ ,  $x \neq y$ .

**Решение.** Пусть  $x, y \in \mathbb{R}^n$  и  $x \neq y$ . Запишем неравенство (4) для пары  $(x, y)$ , а затем для пары  $(y, x)$ :

$$\begin{aligned} f(y) &> f(x) + \langle \nabla f(x), y - x \rangle, \\ f(x) &> f(y) + \langle \nabla f(y), x - y \rangle. \end{aligned}$$

Складывая эти два неравенства, получаем

$$0 > \langle \nabla f(x), y - x \rangle + \langle \nabla f(y), x - y \rangle.$$

Переупорядочивая, получаем в точности неравенство (3).

В предыдущей задаче было показано, что для строго выпуклых функций необходимое условие положительной определенности  $\langle y_k, s_k \rangle > 0$  всегда выполняется, причем абсолютно не важно, как были получены точки  $x_k$  и  $x_{k+1}$ , определяющие векторы  $y_k$  и  $s_k$ . Если функция  $f$  не является строго выпуклой, тогда условие  $\langle y_k, s_k \rangle > 0$  можно обеспечить за счет использования «правильного» линейного поиска.

**Задача 3.** Пусть  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  — непрерывно-дифференцируемая функция (возможно, невыпуклая). Пусть  $x \in \mathbb{R}^n$  — точка, и  $d \in \mathbb{R}^n$  — направление спуска для функции  $f$  в точке  $x$ . Рассмотрим итерацию

$$x_+ := x + \bar{\alpha}d,$$

где  $\bar{\alpha} > 0$  — длина шага. Покажите, что если  $\bar{\alpha}$  удовлетворяет условиям Вульфа (сильным или слабым), то

$$\langle \nabla f(x_+) - \nabla f(x), x_+ - x \rangle > 0.$$

*Подсказка.* Условия Вульфа (сильные или слабые) обеспечивают выполнение неравенства

$$\phi'(\bar{\alpha}) \geq c_2 \phi'(0), \tag{5}$$

где  $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$  — функция  $\phi(\alpha) := f(x + \alpha d)$ , и  $0 < c_2 < 1$ .

**Решение.** Вспомним, как выражается производная функции  $\phi$  через градиент функции  $f$ :

$$\phi'(\alpha) = \langle \nabla f(x + \alpha d), d \rangle.$$

Используя эту формулу и определение точки  $x_+$ , неравенство (5) можно переписать в следующем виде:

$$\langle \nabla f(x_+), d \rangle \geq c_2 \langle \nabla f(x), d \rangle.$$

Поскольку  $d$  является направлением спуска для функции  $f$  в точке  $x$ , то  $\langle \nabla f(x), d \rangle < 0$ . Учитывая, что  $c_2 < 1$ , получаем оценку

$$c_2 \langle \nabla f(x), d \rangle > \langle \nabla f(x), d \rangle.$$

Значит,

$$\langle \nabla f(x_+), d \rangle > \langle \nabla f(x), d \rangle.$$

Осталось домножить обе части неравенства на  $\bar{\alpha}$  и воспользоваться тождеством  $\bar{\alpha}d = x_+ - x$ .

Итак, тем или иным образом можно добиться выполнения необходимого условия положительной определенности  $\langle y_k, s_k \rangle > 0$ . Оказывается, что выполнение одного только этого неравенства полностью достаточно для того, чтобы процедуры обновления DFP и BFGS сохраняли положительную определенность. Покажем это, например, для схемы DFP.

**Задача 4.** Рассмотрим формулу обновления матрицы  $H_k$  в методе DFP. Пусть  $\langle y_k, s_k \rangle > 0$ . Докажите, что если  $H_k$  является положительно определенной матрицей, то  $H_{k+1}$  также будет положительно определенной.

*Подсказка.* Воспользуйтесь определением положительной определенности ( $\langle H_{k+1}u, u \rangle > 0$  для всех  $u \in \mathbb{R}^n \setminus \{0\}$ ) и неравенством Коши–Буняковского.

**Решение.** Для упрощения всех формул опустим всюду индекс  $k$ , а вместо индекса  $k+1$  будем писать символ «+».

Пусть  $u \in \mathbb{R}^n \setminus \{0\}$ . Рассмотрим соответствующую квадратичную форму

$$\langle H_+u, u \rangle = \langle Hu, u \rangle - \frac{\langle Hy, u \rangle^2}{\langle Hy, y \rangle} + \frac{\langle s, u \rangle^2}{\langle y, s \rangle}.$$

Покажем, что эта квадратичная форма всюду положительная.

Заметим, что последнее слагаемое в правой части выписанного равенства является неотрицательным. Проверим, что разность первых двух слагаемых также является неотрицательной:

$$\langle Hu, u \rangle - \frac{\langle Hy, u \rangle^2}{\langle Hy, y \rangle} = \frac{\langle Hu, u \rangle \langle Hy, y \rangle - \langle Hy, u \rangle^2}{\langle Hy, y \rangle}.$$

Напомним, что  $H \in \mathbb{S}_{++}^n$ , а, значит,  $H$  задает в пространстве  $\mathbb{R}^n$  скалярное произведение  $\langle x, y \rangle_H := \langle Hx, y \rangle^{1/2}$  и порожденную этим скалярным произведением норму  $\|x\|_H := \langle x, x \rangle_H^{1/2} = \langle Hx, x \rangle^{1/2}$ . Согласно неравенству Коши–Буняковского, получаем

$$\langle Hy, u \rangle^2 \leq \langle Hy, y \rangle \langle Hu, u \rangle. \quad (6)$$

Значит, числитель в последней дроби неотрицательный. Итак,  $\langle H_+u, u \rangle \geq 0$ .

Осталось показать, что  $\langle H_+u, u \rangle$  не может быть равно нулю. Если  $\langle s, u \rangle > 0$ , тогда, в силу доказанного выше,  $\langle H_+u, u \rangle > 0$ . Пусть теперь  $\langle s, u \rangle = 0$ . Тогда  $\langle H_+u, u \rangle$  может быть равно нулю лишь в том случае, когда в неравенстве Коши–Буняковского (6) достигается равенство. Последнее возможно лишь в том случае, когда векторы  $y$  и  $u$  коллинеарны, т. е.  $u = \alpha y$  для некоторого  $\alpha \in \mathbb{R}$ . Заметим, что это противоречит условию положительной определенности  $\langle y, s \rangle > 0$ , поскольку по предположению  $\langle y, s \rangle = \alpha \langle u, s \rangle = 0$ . Итак,  $\langle H_+u, u \rangle > 0$ .

В заключение рассмотрим, каким образом по представленным формулам пересчета обратных матриц можно получить формулы пересчета прямых матриц. Наиболее просто это делается для процедуры обновления SR1.

**Задача 5.** По формуле пересчета обратной матрицы  $H_k$  в методе SR1 получите формулу пересчета прямой матрицы  $B_k$ .

**Решение.** Запишем формулу пересчета обратной матрицы (для упрощения обозначений соответствующие индексы опустим):

$$H_+ = H + \frac{(s - Hy)(s - Hy)^T}{\langle s - Hy, y \rangle}.$$

Чтобы выписать соответствующее обновление для прямой матрицы  $B := H^{-1}$ , воспользуемся формулой Шермана–Моррисона (утверждение 1). Обозначим

$$u := \frac{s - Hy}{\langle s - Hy, y \rangle}, \quad v := s - Hy.$$

Тогда, используя то, что  $B$  является обратной матрицей к  $H$ , т. е.  $BH = I_n$ , получаем

$$\begin{aligned} B_+ &= B - \frac{B \frac{s-Hy}{\langle s-Hy, y \rangle} (s-Hy)^T B}{1 + \left\langle s-Hy, B \frac{s-Hy}{\langle s-Hy, y \rangle} \right\rangle} = B - \frac{(Bs-y)(Bs-y)^T}{\langle s-Hy, y \rangle + \langle s-Hy, Bs-y \rangle} \\ &= B - \frac{(Bs-y)(Bs-y)^T}{\langle s-Hy, Bs \rangle} = B - \frac{(Bs-y)(Bs-y)^T}{\langle Bs-y, s \rangle} = B + \frac{(y-Bs)(y-Bs)^T}{\langle y-Bs, s \rangle}. \end{aligned}$$

(Для строгости здесь нужно отметить, что приведенное рассуждение корректно, когда обновления прямой и обратной матриц корректно определены, т. е. когда  $\langle s-Hy, y \rangle \neq 0$  и  $\langle y-Bs, s \rangle \neq 0$ .)

Итак, формула обновления прямой матрицы  $B_k$  в методе SR1 выглядит следующим образом:

$$B_{k+1} = B_k - \frac{(B_k s_k - y_k)(B_k s_k - y_k)^T}{\langle B_k s_k - y_k, s_k \rangle} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{\langle y_k - B_k s_k, s_k \rangle}.$$

Интересно отметить, что эта формула является *двойственной* к соответствующей формуле обновления обратной матрицы в том смысле, что любая из этих формул может быть получена из другой формальной заменой  $y_k \leftrightarrow s_k$  и  $B_k \leftrightarrow H_k$ .

Аналогичным образом, применяя формулу Шермана–Моррисона дважды, можно получить формулы пересчета прямых матриц  $B_k$  по формулам пересчета обратных матриц  $H_k$  для процедур DFP и BFGS. Опустим соответствующие вычисления в силу их громоздкости и приведем сразу итоговый результат:

1. (DFP)	$B_{k+1} = \left( I_n - \frac{y_k s_k^T}{\langle y_k, s_k \rangle} \right) B_k \left( I_n - \frac{s_k y_k^T}{\langle y_k, s_k \rangle} \right) + \frac{y_k y_k^T}{\langle y_k, s_k \rangle}.$
2. (BFGS)	$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{\langle B_k s_k, s_k \rangle} + \frac{y_k y_k^T}{\langle y_k, s_k \rangle}.$

Отсюда видно, что формулы обновления DFP и BFGS являются *двойственными* в том смысле, что формула обновления прямой/обратной матрицы в методе DFP совпадает с формулой обновления обратной/прямой матрицы в методе BFGS при формальной замене  $y_k \leftrightarrow s_k$  и  $B_k \leftrightarrow H_k$ .