Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(государственный университет)»

Физтех-школа прикладной математики и информатики

Факультет управления и прикладной математики

Кафедра информатики и вычислительной математики

ИССЛЕДОВАНИЕ ПОВЕДЕНИЯ ГРАДИЕНТНЫХ МЕТОДОВ И ИХ ВАРИАЦИЙ НА ШТИФЕЛЕВОМ МНОГООБРАЗИИ НА ПРИМЕРЕ КВАДРАТИЧНОЙ ФУНКЦИИ

Выпускная квалификационная работа
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:
студент 476 группы _____ Лукашевич Александр Леонидович

Научный руководитель:
д. т. н., г. н. с. в ИПУ РАН _____ Поляк Борис Теодорович

Москва 2018

Оглавление

1. Введение 3
   1.1. Основы теории 3
   1.2. Численные методы 5
   1.3. Минимизация на единичной сфере 6
2. Однородная квадратичная функция 8
   2.1. Теоретические вопросы 8
   2.2. Метод решения. Модифицированный метод проекции градиента 11
   2.3. Численные эксперименты 24
3. Общий случай 25
   3.1. Теоретические вопросы 25
   3.2. Метод решения. Тригонометрическое представление геодезической 26
   3.3. Выбора шага. "Сферическое правило Армихо" . . . 30
   3.4. Выбор шага. Явный вид 37
4. Квадратичная функция с линейным членом 43
   4.1. Теоретические вопросы 43
   4.2. Численные эксперименты 57
5. Выводы 59

Список литературы 60

1. Введение

1.1. Основы теории

Рассматриваемая задача является задачей с ограничениями типа равенств, которая является частным случаем общей задачи математического программирования:

$$\begin{aligned} &\min_{x\mathbb{R}^n} f(x) \\ &\quad g_i(x) = 0,\ i = 1, \dots, l, \\ &\quad h_j(x) \leq 0,\ j = 1, \dots, k. \end{aligned}$$

Задача с ограничениям типа равенства:

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} f(x) \\ &\quad g_i(x) = 0,\ i = 1, \dots, m. \end{aligned} \tag{1}$$

Задачи классифицируются на *выпуклые, невыпуклые, линейные* и *нелинейные*. Выпуклые задачи весьма просты. В них целевая функция $f(x)$ и множество, задаваемое ограничениями $g_i(x)$ выпуклы, что сильно упрощает их анализ и получение решений, чего нельзя сказать о невыпуклых, в которых возникает много тонкостей.

Важным инструментом при исследовании задач такого типа является *правило множителей Лагранжа*[5]. Оно оперирует понятием регулярной точки минимума.

Определение 1.1. Говорят, что x^* – регулярная точка минимума, если $f(x),\ g_i(x)$ являются непрерывно дифференцируемыми функциями в ее окрестности и $g_i'(x^*)$ линейно независимы $\forall\ i = 1, \dots, m$.

Теорема 1.1.1. *Если x^* – регулярная точка минимума, то найдутся*

$y_1^*, \dots y_m^*$ *такие, что*

$$f'(x^*) + \sum_{i=1}^{m} y_i^* g_i'(x^*) = 0.$$

Числа $y_1^*, \dots, y_m^*$ называются множителями Лагранжа. Правило множителей Лагранжа позволяет определить стационарные точки задачи. Далее возникает вопрос о том, являются ли данные стационарные точки точками локального минимума или нет. Для ответа на этот вопрос этого есть необходимое и достаточное условия второго порядка:

Теорема 1.1.2. *Пусть x^* – регулярная точка минимума в задаче (1), $f(x)$ и $g_i(x)$ дважды непрерывно дифференцируемы в окрестности x^*, а y_i^*, $i = 1, \dots m$ – множители Лагранжа. Тогда*

$$\left\langle L''_{xx}(x^*, y^*)s, s \right\rangle = \left\langle (f''(x^*) + \sum_{i=1}^{n} y_i^* g_i''(x^*))s, s) \right\rangle \geq 0 \tag{2}$$

для всех $s \in \mathcal{S} = \{s : \ \left\langle g_i'(x^*), s \right\rangle = 0, \ i = 1, \dots, m\}$.

Иными словами, матрица $L''_{xx}(x^*, y^*)$ неотрицательно определена на касательном подпространстве $\mathcal{S}$.

Теорема 1.1.3. *Пусть $g_i(x^*) = 0$, $i = 1, \dots, m$, функции $f(x)$, $g_i(x)$ дважды непрерывно дифференцируемы в окрестности x^*, пусть $g_i'(x^*)$, $i = 1, \dots, m$ линейно независимы, выполнено необходимое условие минимума (2) и*

$$\left\langle L''_{xx}(x^*, y^*)s, s \right\rangle > 0$$

при всех $s \in \mathcal{S}$. Тогда x^* – точка локального минимума в задаче 1.

Стоит отметить, что в обычных курсах математического анализа исследование задачи с ограничениями типа равенств заканчивается вы-

водом условий экстремума, которые позволят найти решение. Для академических примеров, которые специально составлены в упражнениях и заданиях, это так. Но на деле все гораздо сложнее. Зачастую система уравнений, которую задает правило множителей Лагранжа оказывается нелинейной, а возможности явно выписать решения нет.

Подлинное значение условий экстремума заключается в другом. С их помощью строятся численные методы отыскания решения, они дают возможность судить о его типе.

1.2. Численные методы

Выделяют *методы нулевого, первого* и *второго порядков* в зависимости от порядка используемых производных. Далее методы делятся на *прямые* (в которых итерации ведутся в пространстве прямых переменных x) и *двойственные* (которые существенно используют двойственные переменные y). Во многих методах на каждом шаге решается вспомогательная задача. Это может быть задача безусловной минимизации, линейной функции при квадратичных ограничениях и многие другие. В данной работе будем иметь дело с прямыми методами первого порядка, в которых вычисляются градиенты $f'(x),\ g'(x)$.

Приведем примеры стандартных методов минимизации для задачи (1). Введем обозначение $Q = \{x:\ g_i(x) = 0,\ i = 1, \dots, m\}$.

1. Метод штрафных функций

$$x_k = \operatorname*{argmin}_{x \in Q} f_k(x),\ f_k(x) = f(x) + \frac{1}{2} K_k \|g(x)\|^2,\ K_k \to \infty.$$

2. Метод проекции градиента

$$x_{k+1} = P_Q(x_k - \gamma f'(x_k)).$$

Приведенные в пример методы кардинально отличаются в следующем смысле: последовательность точек, которая генерируется методом штрафных функций, может не находиться в допустимом множестве Q, в то время как последовательность, генерируемая методом проекции градиента всегда находится в Q. Методы, предложенные в этой работе будут схожи с методом проекции градиента, в смысле, описанном выше.

1.3. Минимизация на единичной сфере

Сформулируем постановку задачи в общем случае, однородном квадратичном и неоднородном:

$$\min_{\text{s.t. } x \in \mathbb{S}} f(x), \tag{3}$$

$$\min_{\text{s.t. } x \in \mathbb{S}} \frac{1}{2}\langle Ax, x\rangle, \tag{4}$$

$$\begin{gathered} \min_{\text{s.t. } x \in \mathbb{S}} \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \\ \mathbb{S} = \{x \in \mathbb{R}^n : \|x\| = 1\}. \end{gathered} \tag{5}$$

Целевую функцию $f(x)$ считаем гладкой по Липшицу с константой L. Множество $\mathbb{S}$ является частным случаем Штифелева многообразия: $V_k(\mathbb{R}^n) = \{X \in \mathbb{M}^{k\times n} \ : \ X^T X = I_k\}$ – множества всех ортогональных матриц размера n на k. Поговорим о том, что известно из литературы. В книге Фаддеевых [6], практически, рассмотрен случай (4), но не совсем эквивалентный. Рассматривается задача

$$\max \frac{\langle Ax, x\rangle}{\langle x, x\rangle}.$$

Для решения применяется градиентный спуск, производится его анализ.

Стоит отметить, что целевая функция весьма необычная:

$$\frac{\langle Ax, x\rangle}{\langle x, x\rangle} = \frac{\langle Ax, x\rangle}{\|x\| \cdot \|x\|} = \Big\langle A\frac{x}{\|x\|}, \frac{x}{\|x\|}\Big\rangle.$$

Отсюда видно, что на целой прямой целевая функция принимает одно и то же значение. Такой подход оказался не слишком удобен для (20) и совершенно непригоден для (13).

Стоит обратить внимание на то, что решение задачи (4), как мы увидим дальше, собственный вектор единичной длины, соответствующий минимальному собственному значению. Если же вместо единичной сферы рассмотреть $V_k(\mathbb{R}^n)$ – Штифелево многообразие – все ортогональные матрицы размера $n \times k$, и взять скалярное произведение в этом многообразии, то решением задачи будут первые (по величине собственных значений) k собственных векторов.

Случай неоднородной квадратичной целевой функции подробно рассмотрен в [2]. Как часть статьи, рассматривается неоднородная квадратичная функция на единичной сфере, проводится серьезное преобразование исходной задачи и сведение к более простой эквивалентной, методом множителей Лагранжа задача сводится к решению однородной уравнения.

Случай, когда целевая функция не имеет конкретного вида, в основном, рассматривается уже не на единичной сфере, а на её обобщении – Штифелевом многообразии [3], [7]. В статье Эдельмана большое внимание уделяется самому многообразию, геодезическим на нем, касательным и нормальным подпространствам в точках. Рассматриваются несколько методом оптимизации. В статье Биртея рассматривается скорейший градиентный спуск на Штифелевых многообразиях, описан алгоритм и приведены численные результаты.

О структуре данной работы: На примере общего случае в секции 3 рассмотрен один из методов оптимизации, который так же хорошо ложится на частные случаи (4), (20). В секции 3 рассмотрен и проанализирован отдельный метод с тригонометрическим представлением геодезической (дуга большого круга для сферы), в секции 2 рассмотрена однородная квадратичная функция, и предложен еще один метод, немного отличающийся от простого метода проекции градиента, и, наконец, в секции 4 рассмотрен теоретический аспект задачи с неоднородной квадратичной функцией.

2. Однородная квадратичная функция

Напомним формулировку:

$$\begin{aligned} &\min_{\text{s.t. } x \in \mathbb{S}} \frac{1}{2}\langle Ax, x\rangle, \\ &\mathbb{S} = \{x \in \mathbb{R}^n : \|x\| = 1\}. \end{aligned} \tag{6}$$

2.1. Теоретические вопросы

Докажем ряд лемм, которые дадут представление о происходящем. Для этого нам понадобится функция Лагранжа

$$L = \frac{1}{2}\langle Ax, x\rangle - \frac{\lambda}{2}\Big(\|x\|^2 - 1\Big). \tag{7}$$

Лемма 2.1.1. *Стационарные точки в задаче* (4) *- собственные векторы матрицы A единичной длины.*

Доказательство. С помощью необходимого условия минимума первого

порядка, определим стационарные точки:

$$\nabla L = \begin{pmatrix} Ax - \lambda x \\ \|x\|^2 - 1 \end{pmatrix} = 0$$

Отсюда видно, что стационарные точки - собственные векторы матрицы A, соответствующие собственным значениям λ, по норме равные 1:

$$\begin{aligned} Ae_i &= \lambda_i e_i \\ \|e_i\| &= 1 \end{aligned} \tag{8}$$

□

Покажем, что решение (4) - собственный вектор, соответствующий минимальному собственному значению. Упорядочим собственные значение по возрастанию: $\lambda_1 < \lambda_2 < \cdots < \lambda_n$.

С помощью достаточного условия минимума второго порядка, убедимся в том, что наименьший собственный вектор - глобальный минимум.

Теорема 2.1.2. *Решение задачи (4) - собственный вектор A, соответствующий наименьшему собственному значению.*

Доказательство.

$$L_{xx} = A - \lambda I.$$

Достаточное условие минимума второго порядка говорит о том, что эта матрица должна быть положительно определена на касательном подпространстве к стационарной точке x^*. Обозначим его за $\mathcal{S}$:

$$\mathcal{S} = \{s : (s, 2x^*) = 0\}. \tag{9}$$

Для начала рассмотрим стационарную точку (x^*, λ^*). Поскольку x^* - собственный вектор матрицы A, например с номером k, то в разложении s по собственным векторам A k-го собственного вектора не будет. Пусть $\{e_i\}_{i=1}^n$ - собственные векторы A, соответствующие собственным значениям λ_i. Тогда

$$s = \sum_{i=1,\, i\neq k}^{n} \alpha_i e_i.$$

Рассмотрим квадратичную форму из достаточного условия минимума второго порядка:

$$\begin{aligned}
\Big\langle (A-\lambda^* I)s, s\Big\rangle &= \Big\langle (A-\lambda^* I)\sum_{i=1,\, i\neq k}^{n} \alpha_i e_i, \sum_{i=1,\, i\neq k}^{n} \alpha_i e_i\Big\rangle = \\
&= \Big\langle \sum_{i=1,\, i\neq k}^{n} \alpha_i \lambda_i e_i - \lambda^* \alpha_i e_i, \sum_{i=1,\, i\neq k}^{n} \alpha_i e_i\Big\rangle = \\
&= \sum_{i=1,\, i\neq k}^{n} \alpha_i^2 (\lambda_i - \lambda^*).
\end{aligned}$$

Напомним, что $\lambda^* = \lambda_k$ - какое-то собственное значение. Таким образом, в силу произвольности α_i, последнее выражение больше нуля, когда все $(\lambda_i - \lambda_k) \geq 0 \iff k = 1$. Стационарная точка, соответвующая этому λ_1 - наименьший собственный вектор. □

Замечание 2.1.1. *Говоря более точно, в задаче* (4) *два локальных минимума –* $\pm e_1$. *Поскольку однородная квадратичная функция является четной, то* $\pm e_1$ *дают одинаковые значения целевой функции.*

Замечание 2.1.2. *Стационарные точки – собственные векторы – с номерами больше* 1 *не являются локальными минимумами.*

Доказательство. Действительно, необходимое условие второго порядка говорит о том, что $L_{xx} \succeq 0$ на касательном подпространстве $\mathcal{S}$. Повторая

рассуждения из предыдущей теоремы,

$$\sum_{i=1,\, i\neq k}^{n} \alpha_i^2(\lambda_i - \lambda_k) \geq 0. \tag{10}$$

Допустим, что $k > 1$. Тогда положим $\alpha_l = 0\ \forall l \neq 1$, а $\alpha_1 = 1$. Тогда (10) будет выглядеть следующим образом

$$\lambda_1 - \lambda_k < 0.$$

Что означает, что необходимое условие минимума второго порядка не выполнено и $e_k,\ k > 1$, не может быть локальным минимумом. □

2.2. Метод решения. Модифицированный метод проекции градиента

Немного улучшим, в некотором смысле, метод проекции градиента. Напомним его вид:

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

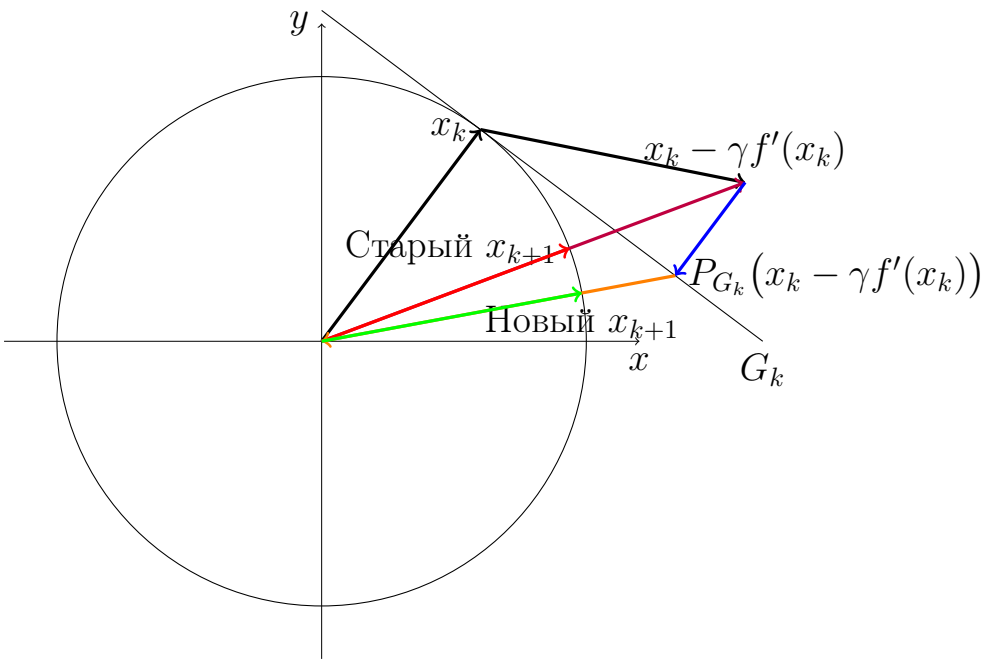
Для начала посмотрим на рисунок и объясним смысл происходящего:

Пусть дана точка x_k, градиент целевой функции в этой точке $f'(x_k)$ и шаг γ. Вместо того, чтобы просто проецировать $x_k - \gamma f'(x_k)$ на сферу (красный вектор), будем сначала проецировать $x_k - \gamma f'(x_k)$ на G_k – касательную к сфере, содержащую x_k и $f'(x_k)$, в точке x_k – получим $P_{G_k}\big(x_k - \gamma f'(x_k)\big)$. А потом уже на сферу:

$$x_{k+1} = \frac{P_{G_k}\big(x_k - \gamma f'(x_k)\big)}{\|P_{G_k}\big(x_k - \gamma f'(x_k)\big)\ |}.$$

Получим вид G_k. Лемму докажем для произвольных неколлинеарных векторов, но оставим обозначения градиентов и точек, чтобы не

Рис. 1. Модифицированный метод проекции градиента



розниться с рисунком.

Лемма 2.2.1. *Пусть* $x_k \in \mathbb{S}$, $\mathbb{S} = \{s \in \mathbb{R}^n : \ \|s\| = 1\}$, $f'(x_k)$ *- вектор из* $\mathbb{R}^n$ *Тогда любой элемент* $G_k = x_k + \nabla\mathbb{S} \cap Span(x_k, f'(x_k)) = \{x_k + x : (y, x) = 0, \ y \in Span(x_k, f'(x_k))\}$ *описывается следующим выражением:*

$$x_k + \beta\big(f'(x_k) - \big\langle x_k, f'(x_k)\big\rangle x_k\big), \ \beta \in \mathbb{R}. \tag{11}$$

Доказательство. Основной вопрос - получить вид $\nabla\mathbb{S} \cap Span(x_k, f'(x_k))$. Множество $\nabla\mathbb{S} \cap Span(x_k, f'(x_k))$ является пересечением гиперплоскости, содержащей все вектора, ортогональные x_k, с гиперплоскостью, натянутой на x_k и $f'(x_k)$. Формально:

$\nabla\mathbb{S} \cap Span(x_k, f'(x_k)) = \{y \ : \ y = \alpha x_k + \beta f'(x_k), \ \big\langle y, x_k\big\rangle = 0, \ \alpha, \ \beta \in \mathbb{R}^n\}$.

Остается только уточнить α и β.

$$0 = \langle y, x_k \rangle = \alpha \underbrace{\langle x_k, x_k \rangle}_{=1} + \beta \langle f'(x_k), x_k \rangle \Longrightarrow$$

$$\Longrightarrow \alpha = -\langle x_k, f'(x_k) \rangle \beta.$$

Итак, мы получили, что $\nabla\mathbb{S} \cap Span(x_k, f'(x_k)) = \{y \ : \ y = \beta(f'(x_k) - \langle x_k, f'(x_k)x_k \rangle), \ \beta \ \in \mathbb{R}^n\}$. Остается к этому множеству прибавить x_k и получить требуемое утверждение. □

Теперь получим выражение для $P_{G_k}(x_k - \gamma f'(x_k))$. Для начала отметим, что этот вектор принадлежит G_k, то есть он имеет вид (11). Здесь нужно обратить внимание на то, что вектор $\mathbf{a} = x_k - \gamma f'(x_k) - P_{G_k}(x_k - \gamma f'(x_k))$ – синий вектор на рисунке 1 – ортогонален $\nabla\mathbb{S} \cap Span(x_k, f'(x_k))$. Отсюда и найдем число β, которое задает позицию на G_k.

Лемма 2.2.2. *В условиях леммы* (2.2.1)

$$P_{G_k}(x_k - \gamma f'(x_k)) = x_k - \gamma f'(x_k) + \gamma \langle x_k, f'(x_k) \rangle x_k.$$

Доказательство. Будем следовать идее, описанной прямо перед формулировкой леммы.

$\mathbf{a} = \cancel{x_k} + \beta f'(x_k) - \beta \langle x_k, f'(x_k) \rangle x_k - \cancel{x_k} + \gamma f'(x_k) = \gamma f'(x_k) + \beta f'(x_k) - \beta \langle x_k, f'(x_k) \rangle x_k$.

Теперь потребуем, чтобы вектор $\mathbf{a}$ был отрогонален $\nabla\mathbb{S} \cap Span(x_k, f'(x_k))$, то есть любому вектору, сонаправленному с $f'(x_k) - \langle x_k, f'(x_k) \rangle x_k$. Получим:

$$\Big\langle f'(x_k)(\gamma+\beta)-\beta(x_k,f'(x_k))x_k, f'(x_k)-(x_k,f'(x_k))x_k\Big\rangle =$$
$$= \Big\langle f'(x_k), f'(x_k)\Big\rangle(\gamma+\beta)-$$
$$-\beta\Big\langle x_k, f'(x_k)\Big\rangle^2 - \Big\langle f'(x_k), x_k\Big\rangle^2(\gamma+\beta)+\beta\Big\langle x_k, f'(x_k)\Big\rangle^2 = 0.$$

Отсюда видно, что $\gamma = -\beta$. Таким образом

$$P_{G_k}(x_k-\gamma f'(x_k)) = x_k - \gamma f'(x_k) + \gamma\Big\langle x_k, f'(x_k)\Big\rangle x_k =$$
$$= x_k(1+\gamma\Big\langle x_k, f'(x_K)\Big\rangle x_k) - \gamma f'(x_k).$$

$\square$

Перед тем, как записать метод, обратим внимание на то, что метод с постоянным шагом неуместен. Мы покажем это чуть дальше. Схему метода записать нужно, она понадобится для анализа.

Алгоритм 1 Модифицированный метод проекции градиента

1: $\varepsilon \leftarrow$ точность
2: $x_0 \leftarrow s \in \mathbb{S}$
3: $k \leftarrow 1$
4: Вычислить $f'(x_{k-1})$
5: Выбрать шаг γ_k
6: $x_k \leftarrow \frac{x_{k-1}\left(1+\gamma_k\left\langle x_{k-1}, f'(x_{k-1})\right\rangle\right)-\gamma_k f'(x_{k-1})}{\|x_{k-1}\left(1+\gamma_k\left\langle x_{k-1}, f'(x_{k-1})\right\rangle\right)-\gamma_k f'(x_{k-1})\|}$
7: **До тех пока** $\|x_{k-1} - x_k\|^2 > \varepsilon$ **выполнять**
8: $\quad k \leftarrow k+1$
9: $\quad$ Вычислить $f'(x_{k-1})$
10: $\quad$ Выбрать шаг γ_k
11: $\quad x_k \leftarrow \frac{x_{k-1}\left(1+\gamma_k\left\langle x_{k-1}, f'(x_{k-1})\right\rangle\right)-\gamma_k f'(x_{k-1})}{\|x_{k-1}\left(1+\gamma_k\left\langle x_{k-1}, f'(x_{k-1})\right\rangle\right)-\gamma_k f'(x_{k-1})\|}$
12: **Конец цикла**
$\quad$ **Возвратить** x_k

Применим данный подход к $f(x) = \frac{1}{2}\left\langle Ax, x\right\rangle$. Введем обозначения

$$a_i^k = \left\langle A^i x_k, x_k\right\rangle,$$

заметим, что $a_1^k = f(x_k)$. В Алгоритме 1 шаг 11 выглядит следующим образом:

$$x_{k+1} = \frac{x_k(1+\gamma_k a_1^k) - \gamma_k A x_k}{\|x_k(1+\gamma_k a_1^k) - \gamma_k A x_k\|}.$$

Рассмотрим $f(x_{k+1}) - f(x_k) = \Delta f_k$.

$$\Delta f_k = \frac{\Big[\big\langle A(x_k(1+\gamma_k a_1^k - \gamma_k A_x k)), (x_k(1+\gamma_k a_1^k - \gamma_k A_x k)\big\rangle\Big]}{\|x_k(1+\gamma_k a_1^k) - \gamma_k A x_k\|^2} - a_1^k =$$

$$= \frac{\gamma_k^2 \overbrace{(a_3^k + 2(a_1^k)^3 - 3a_1^k a_2^k)}^{\alpha_k} - 2\gamma_k \overbrace{(a_2^k - a_1^k)}^{\beta_k}}{\gamma_k^2(a_2^k - a_1^k) + 1} =$$

$$= \frac{\gamma_k^2 \alpha_k - 2\gamma_k \beta_k}{\gamma_k^2 \beta_k + 1} = \frac{\gamma_k(\gamma_k \alpha_k - 2\beta_k)}{\gamma_k^2 \beta_k + 1}.$$

Обратим внимание на α_k и β_k. Величина α_k может принимать как положительные, так и отрицательные значения. Величина β_k неотрицательна, в силу неравенства Коши-Буняковского:

$$(a_1^k)^2 = \big\langle Ax_k, x_k\big\rangle^2 \le \big\langle Ax_k, Ax_k\big\rangle \cdot \big\langle x_k, x_k\big\rangle = a_2^k.$$

Поговорим о поведении Δf_k в зависимости от γ_k. и знака α_k.

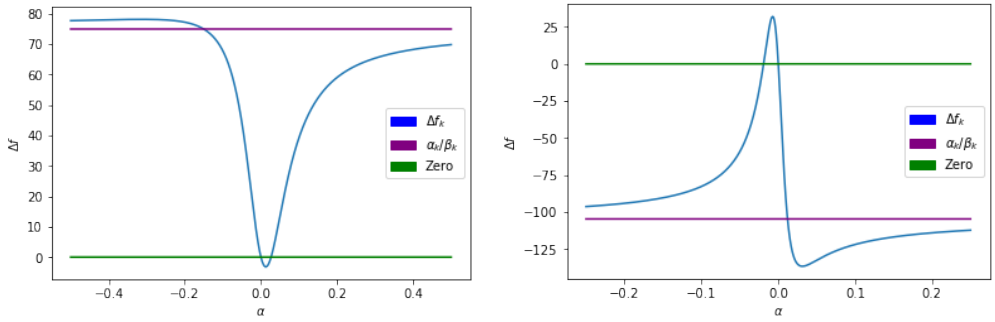
- $\alpha_k > 0,\ \ \Delta f_k < 0$ при $\gamma_k < \frac{2\beta_k}{\alpha_k}$.
- $\alpha_k < 0,\ \ \Delta f_k <$ при любом положительном γ_k.

Построим графики поведения Δf_k в зависимости от γ_k на примере функции с матрицей

$$A = \begin{pmatrix} 100 & 0 & 0 \\ 0 & -70 & 0 \\ 0 & 0 & -15 \end{pmatrix}.$$

Легко видеть, как из рисунков, так и из выражения для Δf_k, что при отрицательном α_k шаг можно брать любой положительный. Так же видно, что в обоих случаях имеет место шаг, дающий минимальное значение Δf_k.

Рис. 2. $\alpha_k > 0,\ \alpha_k < 0$



Лемма 2.2.3. *Пусть при решении задачи* (4) *применяется Алгоритм 1. Тогда шаг, минимизирующий* Δf_k *- разность значений функции в двух соседних точках метода, дается выражением*

$$\gamma_k = \frac{2\beta_k}{\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}.$$

Доказательство. Вычислим стационарные точки Δf_k

$$\begin{aligned}
\Delta f_k' &= \frac{(2\gamma_k\alpha_k - 2\beta_k)(\gamma_k^2\beta_k + 1) - (2\gamma_k\beta_k)(\gamma_k^2\alpha_k - 2\gamma_k\beta_k)}{(\gamma_k^2\beta_k + 1)^2} = \\
&= \frac{2\gamma_k^3\alpha_k\beta_k + 2\gamma_k\alpha_k - 2\beta_k^2\gamma_k^2 - 2\beta_k - 2\gamma_k^3\beta_k\alpha_k + 4\gamma_k^2\beta_k^2}{(\gamma_k^2\beta_k + 1)^2} = \\
&= \frac{2\gamma_k^2\beta_k^2 + 2\gamma_k\alpha_k - 2\beta_k}{(\gamma_k^2\beta_k + 1)^2} = \\
&= 2\frac{\left(\gamma_k - \frac{-\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}{2\beta_k^2}\right)\left(\gamma_k - \frac{-\alpha_k - \sqrt{\alpha_k^2 + 4\beta_k^3}}{2\beta_k^2}\right)}{(\gamma_k^2\beta_k + 1)^2} = 0.
\end{aligned}$$

Отсюда очевидно, что нам нужен корень $\frac{-\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}{2\beta_k^2}$, поскольку он положителен, он больше другого корня, между ними $\Delta f_k'$ отрицательна, а после этого корня она положительна. Значит $\frac{-\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}{2\beta_k^2}$ доставляет минимум Δf_k.

Теперь слегка преобразуем его

$$\gamma_k = \frac{-\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}{2\beta_k^2} = \frac{(-\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3})(\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3})}{2\beta_k^2(\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3})} =$$
$$= \frac{2\beta_k}{\alpha_k + \sqrt{\alpha_k^2 + 4\beta_k^3}}.$$

□

Теперь зададим начальное приближение $x_0 \in \mathbb{S}$, не являющееся стационарной точкой – собственным вектором матрицы A. Нормированные на единицу собственные векторы матрицы A обозначим как $\{e_i\}_{i=1}^n$, λ_i - соответствующие им собственные значения. Индексы упорядочим по возрастанию собственных значений. Будем считать, что

$$x_0 = \sum_{i=1}^n c_i^0 e_i, \ c_i^0 \neq 0 \ \forall i = 1, \ldots, n$$

то есть в разложении x_0 по ОНБ из собственных векторов присутствуют все собственные вектора. Можно от этого отказаться, на ход последующих рассуждений это не повлияет. Если отказаться от этого предположения, то вместо всех собственных векторов будем работать лишь с теми, которые входят в инвариантное подпространство, порожденное x_0. То есть собственные вектора, присутствующие в разложении x_0 по ним же.

Сперва покажем, что по мере работы метода, метод "не разворачивает" x-ы по направлению e_1. Более того, проекция на этот вектор в нуль не обращается, если в x_0 не была таковой.

Лемма 2.2.4. *Пусть* $x_0 = \sum_{i=1}^n c_i^0 e_i$, $c_1^0 > 0$. *Тогда для последовательности, генерируемой Алгоритмом 1, выполнено* $c_1^k > 0 \ \forall \ k > 0$.

Доказательство. Пусть мы находимся на итерации $k+1$. Посмотрим

на c_1^{k+1} :

$$c_1^{k+1} = (x_{k+1}, e_1) = \frac{\left\langle x_k(1+\gamma_k a_1^k) - \gamma_k A x_k, e_1 \right\rangle}{\|x_k(1+\gamma_k a_1^k) - \gamma_k A x_k\|} =$$
$$= \frac{(c_1^k(1+\gamma_k a_1^k) - \gamma_k \overbrace{(Ax_k, e_1)}^{=(x_k, Ae_1)=\lambda_1 c_1^k})}{\|x_k(1+\gamma_k a_1^k) - \gamma_k A x_k\|} =$$
$$= \frac{1}{\|x_k(1+\gamma_k a_1^k) - \gamma_k A x_k\|}(1+\gamma_k(a_1^k - \lambda_1))c_1^k = \cdots =$$
$$\prod_{i=0}^{k} \frac{1}{\|x_i(1+\gamma_k a_1^i) - \gamma_k A x_i\|}(1+\gamma_k(a_1^i - \lambda_1))c_1^0 > 0,$$

ибо $c_1^0 > 0$ по условию, а $a_1^i - \lambda_1 \geq 0$, поскольку a_1^i - значение целевой функции на i-й итерации, которое не может быть меньше оптимального – λ_1. □

Заметим, что требование на знак c_1^0 не умаляет общности. Теперь покажем, что имеет место сходимость метода к наименьшему собственному вектору в двух теоремах.

Теорема 2.2.5. *Если в Алгоритме 1 выбирать шаг как в Лемме 2.2.3, то последовательность $\{x_k\}$, генерируемая Алгоритмом 1, сходится к собственному вектору.*

Доказательство. Перепишем шаг 11 в алгоритме следующим образом:

$$x_{k+1} = \frac{x_k - \gamma_k \xi_k}{\|x_k - \gamma_k \xi_k\|},$$
$$\xi_k = -a_1^k x_k + A x_k.$$

Рассмотрим квадрат нормы этого вектора – $\|\xi_k\|^2$:

$$\left\langle \xi_k, \xi_k \right\rangle = (a_1^k)^2 + a_2^k - 2(a_1^k)^2 = a_2^k - (a_1^k)^2 = \beta_k.$$

С другой стороны, если разложить x_k по собственным векторам,

$$x_k = \sum_{i=1}^{n} c_i^k e_i,$$

то $\|\xi_k\|^2$ можно представить в следующем виде:

$$\begin{aligned}
\langle \xi_k, \xi_k \rangle &= \Big\| - a_1^k \sum_{i=1}^{n} c_i^k e_i + \sum_{i=1}^{n} A c_i^k e_i \Big\|^2 = \\
&= (a_1^k)^2 \sum_{i=1}^{n} (c_i^k)^2 + (c_i^k)^2 \lambda_i^2 - 2 \sum_{i=1}^{n} (c_i^k)^2 \lambda_i a_1^k = \\
&= \sum_{i=1}^{n} (c_i^k)^2 \big[(a_1^k)^2 - 2\lambda_i a_1^k + \lambda_i^2 \big] = \\
&= \sum_{i=1}^{n} (c_i^k)^2 \big[a_1^k - \lambda_i \big]^2.
\end{aligned} \tag{12}$$

Еще раз обратим внимание на то, что $a_1^k = f(x_k)$.

Тут остановимся и подумаем, что было бы, если $< \xi_k, \xi_k > \to 0$? Согласно аксиомам нормы равенство нулю достигается только и только тогда, когда $\xi_k = 0$. Рассмотрим вид ξ_k и обратим его в нуль:

$$\begin{aligned}
\xi_k = \gamma_k A x_k - \langle A x_k, x_k \rangle x_k = 0 \Longrightarrow \\
\Longrightarrow A x_k = \frac{\langle A x_k, x_k \rangle}{\gamma_k} x_k.
\end{aligned}$$

Это очень похоже на условие стационарности – градиент в точке пропорционален самой точке. То есть в последнем выражении в (12) просходило бы следующее: поскольку в сумме все слагаемые неотрицательны, то все они должны были бы стремиться к нулю. Либо $\big[a_1^k - \lambda_i\big]^2 \to 0$, либо $(c_i^k)^2 \to 0$. Лишь с помощью всех $(c_i^k)^2$ мы не смогли бы удовлетворить стремление всех слагаемых к нулю. Действительно, ведь c_i^k - коэффи-

циенты разложения x_k по собственным векторам, сумма их квадратов должна давать единицу:

$$\sum_{i=1}^{n}(c_i^k)^2 = 1 \ \forall k.$$

Есть степень свободы: число a_1^k . Для какого-то одного слагаемого $a_1^k \to \lambda_i$ (пока не знаем к какому именно λ_i), а во всех остальных слагаемых $(c_i^k)^2 \to 0$.

Покажем, что $\langle \xi_k, \xi_k \rangle = \beta_k \to 0$. Для этого рассмотрим $\Delta f_k = f(x_{k+1}) - f(x_k)$ и просуммируем от 0 до s:

$$f(x_s) - f(x_0) = \sum_{i=1}^{s} \frac{\gamma_k^2 \alpha_k - 2\gamma_k \beta_k}{\gamma_k^2 \beta_k^2 + 1}.$$

Устремим $s \to \infty$, получим

$$f(z) - f(x_0) = \sum_{i=1}^{\infty} \frac{\gamma_k^2 \alpha_k - 2\gamma_k \beta_k}{\gamma_k^2 \beta_k^2 + 1}.$$

При этом $z \in \mathbb{S}$, поскольку эта точка сгенерирована Алгоритмом 1, который не уводит точки со сферы. Заметим, что слева стоит число. Действительно, ведь $f(x) = \frac{1}{2} < Ax, x >$ - непрерывная функция, а $\mathbb{S}$ - компакт. Это значит, что функция ограничена и слева действительно конечное число.

Отсюда следуюет сходимость ряда, что стоит справа, по определению. По необходимому условию сходимости ряда, $\frac{\gamma_k^2 \alpha_k - 2\gamma_k \beta_k}{\gamma_k^2 \beta_k^2 + 1} \to 0$. То же

верно и для его модуля. Рассмотрим его подробнее:

$$
\begin{aligned}
|\Delta f_k(\gamma_k)| &= \left|\frac{\gamma_k^2\alpha_k - 2\gamma_k\beta_k}{\gamma_k^2\beta_k^2+1}\right| = \\
&= \left|\frac{\frac{4\beta_k^2\alpha_k}{(\alpha_k+\sqrt{\alpha_k^2+4\beta_k^3})^2} - \frac{4\beta_k^2}{(\alpha_k+\sqrt{\alpha_k^2+4\beta_k^3})}}{\frac{4\beta_k^3\alpha_k}{(\alpha_k+\sqrt{\alpha_k^2+4\beta_k^3})^2}+1}\right| = \\
&= \frac{2\beta_k^2\sqrt{\alpha_k^2+4\beta_k^3}}{4\beta_k^3+\alpha_k^2+\alpha_k\sqrt{\alpha_k^2+4\beta_k^3}}.
\end{aligned}
$$

По необходимому условию сходимости ряда и теореме о двух милиционерах (последовательность поджата снизу нулем, а сверху последовательностью, сходящейся туда же), выражение сверху сходится к нулю. За счет чего оно может стремиться к нулю? Только засчет числителя, поскольку знаменатель ограничен сверху. Это возможно тогда, когда $\beta_k^2 \to 0$, либо выражение под корнем сходится к нулю. Для второго $\beta_k^3 \to 0$ и $\alpha_k^2 \to 0$, поскольку обе величины положительны. В каждом случае предположим противное:

- $\beta_k \to \beta \neq 0,\ \alpha_k \to 0$:

  В таком случае $|\Delta f_k(\gamma_k)| \to \frac{1}{\beta_k} \not\to 0$ – противоречение со сходимостью $|\Delta f_k|$ к нулю.

- $\beta_k \to \beta \neq 0,\ \alpha_k \to \alpha \neq 0$:

  В таком случае $|\Delta f_k(\gamma_k)| \to \frac{1}{\beta_k} \to \frac{2\beta^2\sqrt{\alpha^2+4\beta^3}}{4\beta^3+\alpha^2+\alpha\sqrt{\alpha^2+4\beta^3}} \not\to 0$ – противоречение со сходимостью $|\Delta f_k|$ к нулю.

В любом случае получено противоречие, значит $\beta_k \to 0$. Что нам и нужно. Итак, мы выяснили, что $\beta_k \to 0$. То есть $\langle \xi_k, \xi_k \rangle \to 0$. Значит все рассуждения о сходимости $a_1^k = f(x_k)$ к собственному значению справедливы. Это означает, что имеет место сходимость к стационарной точке – собственному вектору. $\square$

Финальный шаг:

Теорема 2.2.6. *В условиях Теоремы (2.2.5) имеет место сходимость к наименьшему собственному вектору.*

Доказательство. Докажем от противного. Пусть имеет место сходимость не к наименьшему собственному вектору, а к j-му. Тогда верно, что

$$\left|\frac{c_1^{k+1}}{c_j^{k+1}}\right| \to 0,\ j > 1.$$

Рассмотрим это отношение подробнее:

$$\begin{aligned}\left|\frac{c_1^{k+1}}{c_j^{k+1}}\right| &= \left|\frac{(1+\gamma_k(a_1^k-\lambda_1))}{(1+\gamma_k(a_1^k-\lambda_j))}\right|\left|\frac{c_1^k}{c_j^k}\right| = \\ &= \left|\frac{(1+\gamma(a_1^k-\lambda_j)+\gamma_k(\lambda_j-\lambda_1))}{(1+\gamma(a_1^k-\lambda_j))}\right|\left|\frac{c_1^k}{c_j^k}\right| = \\ &= \left|1+\frac{\gamma_k(\lambda_j-\lambda_1)}{(1+\gamma_k(a_1^k-\lambda_j))}\right|\left|\frac{c_1^k}{c_j^k}\right| > \left|\frac{c_1^{k-1}}{c_j^{k-1}}\right|.\end{aligned}$$

Левый множитель в последнем элементе цепочки равенств больше единицы, поскольку дробь положительна. Действительно, $\lambda_j > \lambda_1$, а $a_1^k - \lambda_j > 0$ по свойству шага – он всегда уменьшает значение функции $a_1^k = f(x_k)$, значит мы подбираемся к оптимальному значению сверху. В итоге, получено противоречие со сходмостью $\left|\frac{c_1^{k+1}}{c_j^{k+1}}\right| \to 0,\ j > 1.$ □

Замечание 2.2.1. *Из полученных результатов видно, что постоянный шаг брать нельзя. Ведь шаг должен удовлетворять*

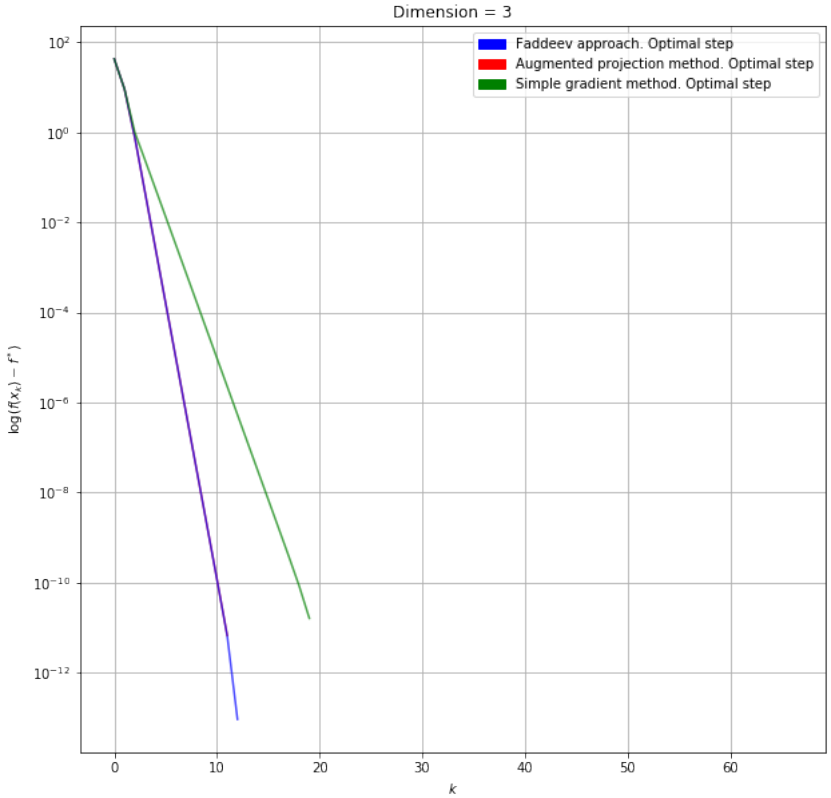
$$\begin{aligned}0 <&\gamma_k < \frac{2\beta_k}{\alpha_k},\ \alpha_k > 0,\\ 0 <&\gamma_k,\ \alpha_k < 0\end{aligned}$$

на каждом шаге.

2.3. Численные эксперименты

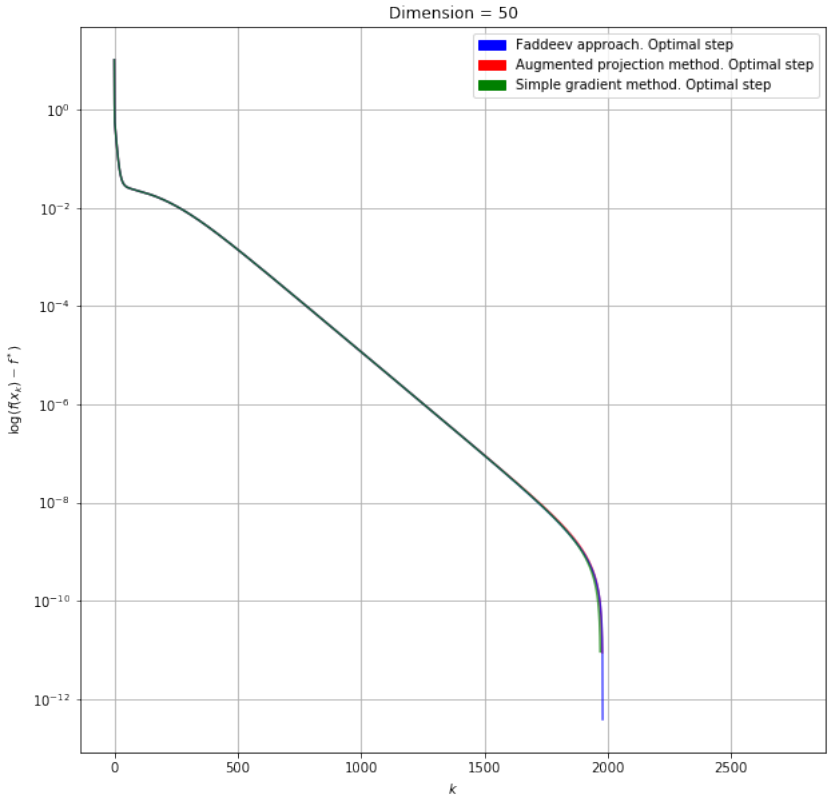
Проведем сравнение Алгоритма 1 с подходом Фаддеевых[6] и простым методом проекции градиента. Сравнение проведем на матрице $A \in \mathbb{M}^{3\times 3}$, $A = A^T$ сгенирированной случайно, так же как и начальное приближение на сфере. Точность выбрана $\varepsilon = 10^{-6}$. Рисунок 3. Теперь по-

Рис. 3. Модифицированный метод проекции градиента, метод проекции градиента и подход Фаддеевых. Размерность 3.



пробуем на матрице $A \in \mathbb{M}^{50\times 50}$ Результат на Рисунке 4. Методы, в целом, ведут себя идентично. Приведем данные о времени работы методов в Таблице 1.

Рис. 4. Модифицированный метод проекции градиента, метод проекции градиента и подход Фаддеевых. Размерность 50.



Метод	Размерность	Время (сек)
Подх. Фаддеевых	3/50	0.009/2.09
ММПГ	3/50	0.001/0.15
МПГ	3/50	0.01/1.8

Таблица 1. Время работы

3. Общий случай

Откажемся от квадратичной функции и будем рассматривать произвольную функцию $f(x)$ с липшицевым градиентом.

$$\min_{x \in \mathbb{S}} f(x). \tag{13}$$

3.1. Теоретические вопросы

Основной факт, который понадобится отразим в лемме:

Лемма 3.1.1. *Пусть $f(x)$ - дифференцируемая функция на единичной сфере $\mathbb{S} = \{s : \|s\|^2 = 1\}$. Тогда стационарные точки в задаче (13) – точки, коллинеарные с градиентом в этой точке:*

$$x = \pm \frac{f'(x)}{\|f'(x)\|}.$$

Доказательство. Запишем функцию Лагранжа:

$$L = f(x) - \frac{\lambda}{2}(\|x\|^2 - 1).$$

Условие стационарности:

$$L' = \begin{pmatrix} f'(x) - \lambda x \\ \|x\|^2 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Отсюда $x = \pm\frac{1}{\lambda} f'(x) = \pm\frac{f'(x)}{\|f'(x)\|}$. □

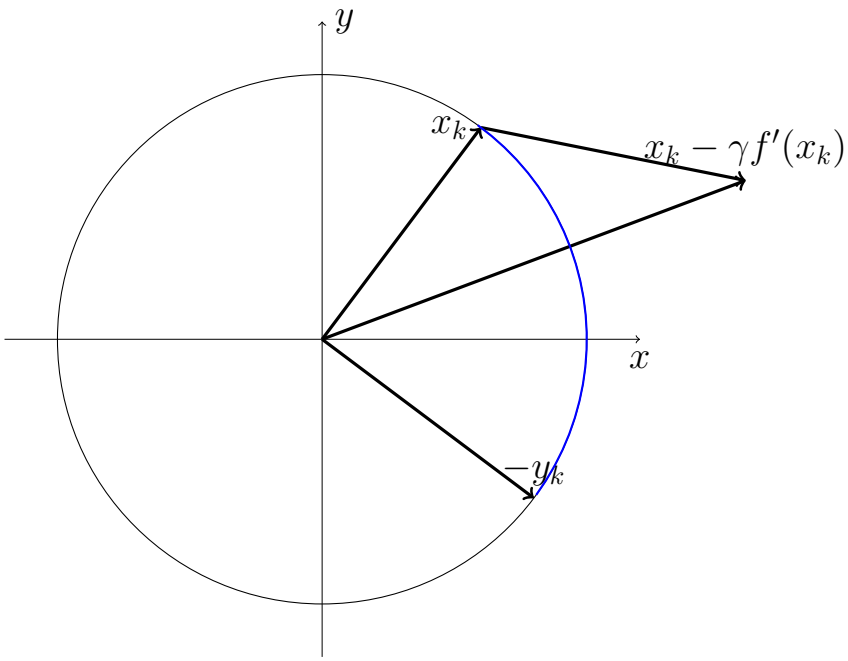
3.2. Метод решения. Тригонометрическое представление геодезической

Обратим внимание на Рисунок 5. Пусть даны точка x_k на сфере, градиент $f'(x_k)$ в этой точке. К точке x_k построим $-y_k$ - ортогональную точку к x_k, сонаправленную с антиградиентом. Точка y_k лежит в плоскости, натянутой на x_k и $f'(x_k)$. Рассмотрим кривую (синий цвет на рисунке)

$$z_k(t) = x_k \cos t - y_k \sin t, \ t \in \left[0, \frac{\pi}{2}\right]$$

И на этой кривой будем искать следующую точку метода. При этом t задает положение на дуге. Например, при $t = 0$ $z_k(0) = x_k$, а при $t = \frac{\pi}{2}$ имеем $z_k(\pi/2) = -y_k$.

Рис. 5. Тригонометрическое представление геодезической



Рассчитаем точку y_k. В следующей лемме сохраним обозначения, оставленные на рисунке, но будем иметь ввиду, что x_k и $f'(x_k)$ можно воспринимать как любые неколлинераные векторы.

Лемма 3.2.1. *Пусть $x_k \in \mathbb{S} = \{s : \|s\| = 1\}$, $f(x)$ - дифференцируемая функция на единичной сфере, градиент функции $f(x)$ в точке x_k не коллинеарен x_k:*

$$\left\langle x_k, f'(x_k) \right\rangle \neq 0.$$

Тогда y_k - точка на сфере, ортогональная x_k, лежащая в линейной оболочке x_k и $f'(x_k)$: $\left\langle f'(x_k), y_k \right\rangle > 0$ – сонаправленная с градиентом в точке – задается следующим выражением:

$$y_k = \frac{(I - x_k x_k^T) f'(x_k)}{\|(I - x_k x_k^T) f'(x_k)\|}. \tag{14}$$

Доказательство. Будем искать y_k в виде

$$y_k = \alpha x_k + \beta f'(x_k),$$

то есть в линейной оболочке x_k и $f'(x_k)$. Потребуем от y_k следующее:

- Норма – $\|y_k\| = 1$
- Ортогональность – $\langle x_k, y_k \rangle = 0$

Этого будет чуть более, чем достаточно. Не стоит забывать, что $\|x_k\| = 1$. Реализуем эти условия:

- $0 = \langle x_k, y_k \rangle = \alpha \underbrace{\langle x_k, x_k \rangle}_{=1} + \beta \langle x_k, f'(x_k) \rangle \implies \alpha = -\beta \langle x_k, f'(x_k) \rangle,$

$$y_k = \beta(f'(x_k) - x_k^T f'(x_k) x_k) = \beta(I - x_k x_k^T) f'(x_k)$$

- $1 = \|y_k\| = \beta \|(I - x_k x_k^T) f'(x_k)\| \implies \beta = \frac{1}{\|(I - x_k x_k^T) f'(x_k)\|}.$

Итак,

$$y_k = \frac{(I - x_k x_k^T) f'(x_k)}{\|(I - x_k x_k^T) f'(x_k)\|} = \\ = \frac{(f'(x_k) - \langle x_k, f'(x_k) \rangle x_k)}{\|(f'(x_k) - \langle x_k, f'(x_k) \rangle x_k)\|}.$$

□

Теперь запишем метод, который базируется на рассуждениях выше (Алгоритм 2).

Алгоритм 2 Тригонометрическое представление геодезической

$\varepsilon \leftarrow$ точность
$x_0 \leftarrow s \in \mathbb{S}$
$k \leftarrow 0$
Вычислить $f'(x_k)$
Вычислить $y_k = \dfrac{\left(f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k\right)}{\left\|\left(f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k\right)\right\|}$
Выбрать шаг t
$x_{k+1} \leftarrow x_k \cos t - y_k \sin t$
$k \leftarrow k + 1$
Повторять
    Вычислить $f'(x_k)$
    Вычислить $y_k = \dfrac{\left(f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k\right)}{\left\|\left(f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k\right)\right\|}$
    Выбрать шаг t
    $x_{k+1} \leftarrow x_k \cos t - y_k \sin t$
    $k \leftarrow k + 1$
Пока выполняется $\|x_{k+1} - x_k\|^2 > \varepsilon$
    Возвратить x_k

3.3. Выбора шага. "Сферическое правило Армихо"

Теперь займемся вопросом выбора шага t. Создадим правило шага на основании правила Армихо. Напомним его вид:

$$f(x_k + \alpha s_k) - f(x_k) \leq \varepsilon\alpha\big\langle f'(x_k), s_k\big\rangle,\ 0 < \varepsilon < 1.$$

Здесь направление $s_k :\ \big\langle f'(x_k), s_k\big\rangle < 0$. Вспомним простой градиентный метод для безусловной задачи. Выбирая шаг из правила Армихо, двигаясь по направлению антиградиента, можно, в конце концов, показать, что $\|f'(x_k)\| \overset{k\to\infty}{\longrightarrow} 0$. То есть в итоге выполнится условие стационарности для задачи безусловной минимизации. В нашем случае условие стационарноести выглядит иначе:

$$f'(x) = \|f'(x)\|x.$$

На основании выражения для условия стационарности, сформулируем "сферическое правило Армихо":

Алгоритм 3 "Сферическое правило Армихо"

1: $\varepsilon_0 \leftarrow\ \epsilon \in (0,1)$
2: $\vartheta \leftarrow \vartheta_0 \in (0,1)$
3: $t^0 \leftarrow t > 0$
4: $i \leftarrow 0$
5: **Если** $f(x_{k+1}) - f(x_k) \leq -t^i\varepsilon_0\big\langle f'(x_k), y_k\big\rangle$ **тогда**
    Возвратить t^i
6: **иначе** $t^{i+1} = \vartheta t^i$
    $i \leftarrow i + 1$
    go to 5
7: **Конец условия**

Ясно, что $\Big(\|f'(x_k)\|^2 - (x_k, f'(x_k))^2\Big) \geq 0$, ведь $\|x_k\| = 1$. В нуль это

выражение обращается, когда $f'(x_k) = \lambda x_k$. Действительно:

$$\left(\lambda^2\|x_k\|^2 - \lambda^2\|x_k\|^2\right) = 0.$$

Убедимся в том, что это правило может выполняться.

Лемма 3.3.1. *Пусть f - дифференцируемая функция на единичной сфере $\mathbb{S}$. Тогда в методе, который задается Алгоритмом 2, на каждой итерации можно выбрать достаточно малый шаг t_k, чтобы "сферическое правило Армихо" выполнялось.*

Доказательство. Рассмотрим разность соседних точек в методе:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(z(t_k)) - f(x_k) = \\ &= f(x_k \cos t_k - y_k \sin t_k) - f(x_k) = \\ &= f(x_k + x_k(\cos t_k - 1) - y_k \sin t_k) - f(x_k). \end{aligned}$$

Воспользуемся формулой Тейлора для функции $f(\cdot)$. Разложение проведем до первого порядка:

$$\begin{aligned} &f(x_k + x_k(\cos t_k - 1) - y_k \sin t_k) - f(x_k) = \\ &= \cancel{f(x_k)} + \left\langle f'(x_k), x_k(\cos t_k - 1) - y_k \sin t_k \right\rangle + \\ &+ o(\|x_k(\cos t_k - 1) - y_k \sin t_k\|) - \cancel{f(x_k)}. \end{aligned}$$

Теперь разложим синус с косинусом – с тем же порядком.

$$\begin{aligned} \left\langle f'(x_k), x_k(\cos t_k - 1) - y_k \sin t_k \right\rangle + o(\|x_k(\cos t_k - 1) - y_k \sin t_k\|) &= \\ &= \left\langle f'(x_k), \cancel{x_k(1-1)} - y_k t_k \right\rangle + o(t) = \\ &= -t_k \left\langle f'(x_k), y_k \right\rangle + o(t_k). \end{aligned}$$

Вспомним, что $y_k = \frac{(I - x_k x_k^T) f'(x_k)}{\|(I - x_k x_k^T) f'(x_k)\|}$:

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &= -t_k \big\langle f'(x_k), y_k \big\rangle + o(t) = \\
&= -t_k \big\langle f'(x_k), \frac{(I - x_k x_k^T) f'(x_k)}{\|(I - x_k x_k^T) f'(x_k)\|} \big\rangle + o(t) = \\
&= -t_k \frac{\Big(\|f'(x_k)\|^2 - (x_k, f'(x_k))^2 \Big)}{\|(I - x_k x_k^T) f'(x_k)\|} + o(t_k).
\end{aligned}
$$

Избавимся от $o(t_k)$:

$$
\begin{aligned}
&f(x_{k+1}) - f(x_k) = -\frac{t_k \Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k) \big\rangle^2 \Big)}{\|(I - x_k x_k^T) f'(x_k)\|} + o(t_k) \overset{\varepsilon_0 \le 1}{=} \\
&= -\frac{t_k \varepsilon_0 \Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k) \big\rangle^2 \Big)}{\|(I - x_k x_k^T) f'(x_k)\|} - \\
&- t_k (1 - \varepsilon_0) \left(\frac{\Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k) \big\rangle^2 \Big)}{\|(I - x_k x_k^T) f'(x_k)\|} + \frac{o(t_k)}{t_k} \right) \underset{t_k > 0}{\le} \\
&\underset{t_k > 0}{\le} -t_k \frac{\varepsilon_0 \Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k) \big\rangle^2 \Big)}{\|(I - x_k x_k^T) f'(x_k)\|}.
\end{aligned}
$$

Последний переход справедлив при достаточно малых t_k. □

Итак, мы убедились, что "сферическое правило Армихо" выполняется при малых t_k. Выясним, дает ли такой подход сходимость к стационарной точке.

Лемма 3.3.2. *Пусть $f(x)$ – дифференцируемая функция на единичной сфере, градиент Липшицев, с константой L:*

$$\|f'(x) - f'(y)\| \le L\|x - y\|.$$

Тогда выполнение условия "сферического правила Армихо" достигается

за конечное число дроблений. То есть $\forall\ k\ \exists\ \tau_k :\ \ \forall t \in (0, \tau_k] :$ *правило выполнено. Конкретно,*

$$\tau_k = \frac{-1/2(L - \langle f'(x_k), x_k\rangle)}{1/3\langle f'(x_k), y_k\rangle} + \\ \frac{\sqrt{1/4\big(L - \langle f'(x_k), x_k\rangle\big)^2 + 4/6\langle f'(x_k), y_k\rangle^2(\frac{\varepsilon_0}{\varepsilon_1} - 1)}}{1/3\langle f'(x_k), y_k\rangle}.$$

Доказательство. Поскольку $f'(x)$ липшицев:

$$|f(x_{k+1}) - f(x_k) - \langle f'(x_k), x_{k+1} - x_k)\rangle| \le \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

В частности,

$$f(x_{k+1}) - f(x_k) \le \langle f'(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

Рассмотрим правую часть неравенства более подробно:

- Первое слагаемое:

$$\begin{aligned}\langle f'(x_k), x_{k+1} - x_k\rangle = \\ &= \langle f'(x_k), x_k\rangle(\cos t_k - 1) - \langle f'(x_k), y_k\rangle \sin t_k = \\ &= -\langle f'(x_k), x_k\rangle\frac{t_k^2}{2} - \langle f'(x_k), y_k\rangle(t_k - \frac{t_k^3}{6}) + o(t_k^3).\end{aligned}$$

- Второе слагаемое:

$$\begin{aligned}\|x_k \cos t_k - y_k \sin t_k - x_k\|^2 &= 1 - 2\cos t_k + 1 = \\ &= 4\sin^2\frac{t_k}{2} = t_k^2 + o(t_k^3).\end{aligned}$$

Поэтому

$$f(x_{k+1}) - f(x_k) \leq$$
$$\leq \frac{t_k^3}{6}\langle f'(x_k), y_k\rangle + \frac{t_k^2}{2}(L - \langle f'(x_k), x_k\rangle) - t_k\langle f'(x_k), y_k\rangle + o(t_k^3) \underset{\varepsilon_1<\varepsilon_0<1}{\leq}$$
$$\underset{\varepsilon_1<\varepsilon_0<1}{\leq} \varepsilon_1\left(\frac{t_k^3}{6}\langle f'(x_k), y_k\rangle + \frac{t_k^2}{2}(L - \langle f'(x_k), x_k\rangle) - t_k\langle f'(x_k), y_k\rangle\right). \tag{15}$$

От последнего выражения потребуем следующее:

$$\varepsilon_1 t_k\left(\frac{t_k^2}{6}\langle f'(x_k), y_k\rangle + \frac{t_k}{2}(L - \langle f'(x_k), x_k\rangle) - \langle f'(x_k), y_k\rangle\right) \leq -t_k\varepsilon_0\langle f'(x_k), y_k\rangle,$$

$$\frac{t_k^2}{6}\langle f'(x_k), y_k\rangle + \frac{t_k}{2}(L - \langle f'(x_k), x_k\rangle) - \langle f'(x_k), y_k\rangle(\frac{\varepsilon_0}{\varepsilon_1} - 1) \leq 0, \tag{16}$$

где ε_0 из Алгоритма 3.

Поэтому, находя корни квадратного многочлена,

$$\tau_{k_{1,2}} = \frac{-1/2(L - \langle f'(x_k), x_k\rangle)}{1/3\langle f'(x_k), y_k\rangle} \pm \frac{\sqrt{1/4\big(L - \langle f'(x_k), x_k\rangle\big)^2 + 4/6\langle f'(x_k), y_k\rangle^2(\frac{\varepsilon_0}{\varepsilon_1} - 1)}}{1/3\langle f'(x_k), y_k\rangle}$$

Итак, для удовлетворения (16) и сохранения смысла достаточно выбирать следующие t_k :

$$0 < t_k < \frac{-1/2(L - \langle f'(x_k), x_k\rangle)}{1/3\langle f'(x_k), y_k\rangle} + \frac{\sqrt{1/4\big(L - \langle f'(x_k), x_k\rangle\big)^2 + 4/6\langle f'(x_k), y_k\rangle^2(\frac{\varepsilon_0}{\varepsilon_1} - 1)}}{1/3\langle f'(x_k), y_k\rangle} = \tau_k.$$

□

Теорема 3.3.3. *Пусть f является дифференцируемой функцией на единичной сфере, градиент липшицев с константой L :*

$$\|f'(x) - f'(y)\| \leq L\|x - y\|.$$

Пусть шаг t_k в Алгоритме 2 выбирается на каждом шаге метода в соответствии с 3. Тогда $f(x_k)$ монотонно убывает и

$$\left\langle f'(x_k), y_k \right\rangle^2 \overset{k\to\infty}{\longrightarrow} 0,$$

$$y_k = \frac{(I - x_k x_k^T) f'(x_k)}{\|(I - x_k x_k^T) f'(x_k)\|}.$$

Доказательство. Из Леммы 3.3.2

$$f(x_{k+1}) - f(x_k) \leq -t_k \varepsilon_0 \left\langle f'(x_k), y_k \right\rangle < 0,$$

если t_k достаточно мал.

Поэтому, с такими t_k, последовательность $f(x_k)$ монотонно убывает. Целевая функция $f(x)$ ограничена снизу, поскольку она непрерывна и рассматривается на единичной сфере, которая является компактом, последовательность $f(x_k)$ сходится. Тогда

$$f(x_{k+1}) - f(x_k) \overset{k\to\infty}{\longrightarrow} 0.$$

Рассмотрим Алгоритм 3 подробнее: Из Леммы 3.3.2) имеем, что $t_k \geq$

$\vartheta\tau_k$, где ϑ – фактор дробления. Введем обозначения

$$\tau_k^{\mathrm{d}} = 1/3\big\langle f'(x_k), y_k\big\rangle,$$
$$\tau_k^{\mathrm{n}} = -1/2(L - \big\langle f'(x_k), x_k\big\rangle)+$$
$$\sqrt{1/4\big(L - \big\langle f'(x_k), x_k\big\rangle\big)^2 + 4/6\big\langle f'(x_k), y_k\big\rangle^2(\frac{\varepsilon_0}{\varepsilon_1} - 1)}.$$

То есть τ_k из Леммы 3.3.2 имеет вид $\tau_k = \frac{\tau_k^{\mathrm{n}}}{\tau_k^{\mathrm{d}}}$

$$f(x_{k+1}) - f(x_k) \le -t_k\varepsilon_0\big\langle f'(x_k), y_k\big\rangle$$
$$\big\langle f'(x_k), y_k\big\rangle \le \frac{f(x_k) - f(x_{k+1})}{t_k\varepsilon_0} \le \frac{(f(x_k) - f(x_{k+1}))\cdot\tau_k^{\mathrm{d}}}{\varepsilon_0\vartheta\tau_k^{\mathrm{n}}},$$
$$\varepsilon_0\vartheta\tau_k^{\mathrm{n}} \le \frac{1}{3}(f(x_k) - f(x_{k+1})) \overset{k\to\infty}{\to} 0.$$

В выражении для $\tau_k^{\mathrm{n}}, \tau_k^{\mathrm{d}}$ обозначим $a_k = -\frac{1}{2}\big(L - \big\langle f'(x_k), x_k\big\rangle\big)$ и $b_k = \frac{4}{6}\big\langle f'(x_k), y_k\big\rangle^2(\frac{\varepsilon_0}{\varepsilon_1} - 1)$. Заметим, что a_k не обязана сходиться к нулю. Пример такого поведения легко подбирается. Поэтому сходимость выражения слева может достигаться лишь сходимостью b_k к нулю. Предположим, что $b_k \not\to 0$, т. е. $b_k \to b^* \neq 0$. То же и для $a_k \to a^*$. Поэтому, рассматривая в пределе, пренебрегая постоянными множителями:

$$-a^* + \sqrt{(a^*)^2 + (b^*)^2} = 0 \Longrightarrow 2(a^*)^2 + (b^*)^2 = 2a^*\sqrt{(a^*)^2 + (b^*)^2} \Longrightarrow$$
$$\Longrightarrow \cancel{4(a^*)^4} + \cancel{4(a^*)^2(b^*)^2} + (b^*)^4 = \cancel{4(a^*)^2(b^*)^2} + \cancel{4(a^*)^4}.$$

Значит $(b^*)^4 = 0$, а это вытекает в $\big\langle f'(x_k), y_k\big\rangle^2 \overset{k\to\infty}{\longrightarrow} 0$. □

Итак, при выборе достаточно малых t_k метод сходится к стационарной точке. Эта достаточная малость возникла при разложении по формуле Тейлора. Рассмотрим другой подход к выбору шага, который

не базируется на достаточной малости.

3.4. Выбор шага. Явный вид

Можно явно указать выражение для шага: $t_k = a\big\langle f'(x_k), y_k\big\rangle$, где $a \in (0, 1/(2L + M))$. Здесь $M = \|f'(0)\|$ Убедимся, что он подходит. Заметим, что выражение для него основано на условии стационарности: чем ближе метод походит к стационарной точке, тем меньше $\big\langle f'(x_k), y_k\big\rangle$ – градиент все более коллинеарен точке, все более ортогонален y_k.

Сперва определим функцию $F_k(t) = f(z_k(t)) = f(x_k \cos t - y_k \sin t)$. Вычислим ее производную по t:

$$F_k'(t) = \big\langle f'(x_k), -x_k \sin t - y_k \cos t\big\rangle.$$

Обратим внимание на следующее:

1. Значение $F'(t)$ в нуле

$$\begin{aligned} F'(0) &= -\big\langle f'(x_k), y_k\big\rangle = \\ &= -\frac{\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k)\big\rangle^2}{\|f'(x_k) - \big\langle x_k, f'(x_k)\big\rangle x_k\|} = -d_k \end{aligned} \tag{17}$$

2. Квадрат значения $F'(0)$

$$\begin{aligned} d_k^2 &= \frac{\Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k)\big\rangle^2\Big)^2}{\Big(\|f'(x_k)\|^2 - \big\langle x_k, f'(x_k)\big\rangle^2\Big)} = \\ &= \|f'(x_k)\|^2 - \big\langle x_k, f'(x_k)\big\rangle^2 \end{aligned} \tag{18}$$

Докажем, что $F_k(t)$ имеет липшицев градиент с констатной $4L+2M$, где L – константа Липшица для градиента функции $f(x)$.

Лемма 3.4.1. *Пусть f – дифференцируемая функция на единичной сфере, градиент этой функции непрерывен по Липшицу с константой L :*

$$\|f'(x_1) - f'(x_2)\| \leq L\|x_1 - x_2\|.$$

Помимо этого, норма градиента $f'(x)$ конечна в точке $x = 0$:

$$\|f'(0)\| = M \in \mathbb{R}.$$

Тогда функция одной переменной

$$F_k(t) = f(z_k(t)) = f(x_k \cos t - y_k \sin t)$$

имеет липшицев градиент с константой $4L + 2M$.

Доказательство. Напомним вид производной

$$F_k'(t) = \big\langle f'(x_k \cos t - y_k \sin t), -x_k \sin t - y_k \cos t \big\rangle.$$

Получим константу липшица

$$\Big| \big\langle f'(x_k \cos t - y_k \sin t), -x_k \sin t - y_k \cos t \big\rangle -$$

$$- \big\langle f'(x_k \cos \tau - y_k \sin \tau), -x_k \sin \tau - y_k \cos \tau \big\rangle \Big| =$$

$$= \Big| \big\langle f'(x_k \cos t - y_k \sin t) - f'(x_k \cos \tau - y_k \sin \tau), -x_k \sin t - y_k \cos t \big\rangle -$$

$$- \big\langle f'(x_k \cos \tau - y_k \sin \tau), -x_k \sin \tau - y_k \cos \tau + x_k \sin t + y_k \cos t \big\rangle \Big|.$$

Из липшицевости градиента

$$\|f'(x_1) - f'(x_2)\| \leq L\|x_1 - x_2\|$$

в случае $x_1 : \ \|x_1\| = 1$ и $x_2 = 0$ имеем

$$\|f'(x_1) - f'(0)\| \leq L \Longrightarrow \|f'(x_1)\| \leq L + M.$$

Помимо этого

$$\|x_k \cos t - y_k \sin t\| =$$
$$= \|x_k\|^2 \cos^2 t + \|y_k\|^2 \sin^2 t - < x_k, y_k > \sin t \cos t = \cos^2 t + \sin^2 t = 1.$$

Продолжим цепочку неравенств:

$$\left| F_k'(\tau) - F_k'(t) \right| \leq$$
$$L\|x_k(\cos t - \cos\tau) - y_k(\sin t - \sin\tau)\| \cdot \|x_k \sin t + y_k \cos t\| +$$
$$+ (L + M) \cdot \|x_k(\sin t - \sin\tau) + y_k(\cos t - \cos\tau)\| \leq$$
$$\leq (2L + M)|\cos t - \cos\tau| + (2L + M)|\sin t - \sin\tau| =$$
$$= (2L + M)\left| 2\sin\frac{t-\tau}{2}\sin\frac{t+\tau}{2} \right| + (2L + M)\left| 2\sin\frac{t-\tau}{2}\cos\frac{t+\tau}{2} \right| \leq$$

Здесь заметим, что $|\sin t| \leq |t|, \ \ |\sin t| \leq 1$ и $|\cos t| \leq 1$.

$$\leq (2L + M)|\tau - t| + (2L + M)|\tau - t| = 2(2L + M)|\tau - t|.$$

Поэтому $F_k(t)$ имеет липшицев градиент с константой $4L + 2M$. $\square$

Теперь докажем теорему, которая позволит уверенно пользоваться таким выбором шага.

Теорема 3.4.2. *Пусть f дифференцируема на единичной сфере, градиент липшицев с константой L :*

$$\|f'(x_1) - f'(x_2)\| \leq L\|x_1 - x_2\|.$$

Пусть шаг в Алгоритме 2 выбирается следующим образом:

$$t_k = ad_k, \ a \in (0, 1/(2L + M)).$$

Тогда

- *На каждом шаге метода:*

$$f(x_{k+1}) - f(x_k) \leq (2L + M)a^2 d_k^2 - ad_k^2 \leq 0 \tag{19}$$

- *Последовательность $f(x_k)$, где x_k сгенерированы Алгоритмом 2 монотонно убывает и*

$$\|f'(x_k)\|^2 - \left\langle x_k, f'(x_k)\right\rangle^2 \overset{k\to\infty}{\longrightarrow} 0.$$

Доказательство. Воспользуемся формулой Ньютона-Лейбница:

$$f(x_{k+1}) - f(x_k) = f(z_k(t_k)) - f(z_k(0)) =$$
$$= \int\limits_0^{ad_k} F_k'(\xi)d\xi = \int\limits_0^{ad_k} \Big(F_k'(\xi) - F_k'(0)\Big)d\xi + \int\limits_0^{ad_k} F_k'(0)d\xi.$$

Вспомним, что $F_k'(0) = -d_k$ и липшицевость $F_k'(t)$:

$$f(x_{k+1}) - f(x_k) \leq (4L + 2M) \int\limits_0^{ad_k} \xi d\xi - d_k a d_k = \frac{4L + 2M}{2} a^2 d_k^2 - ad_k^2.$$

Оценки на a получим из неравенства $(2L+M)a^2d_k^2 - ad_k^2 < 0$. Поэтому $a \in (0, 1/(2L+M))$. Значит (19) выполняется. Покажем сходимость к стационарной точке. Согласно (19):

- Последовательность $\{f(x_k)\}$ монотонно убывает. Вместе с фактом, что $f(x)$ непрерывна и сфера является компактом $\{f(x_k)\}$ сходится к некоторому числу $f(z)$.
- Неравенство $((2L+M)a^2 - a)d_k^2 \leq f(x_k) - f(x_{k+1})$.

Просуммируем последнее неравенство по всем $k \in \mathbb{N}$:

$$((2L+M)a^2 - a)\sum_{k=0}^{\infty} d_k^2 \leq f(x_0) - f(z).$$

В правой части неравенства стоит конечное число, поэтому ряд слега сходится. Из необходимого условия сходимости ряда:

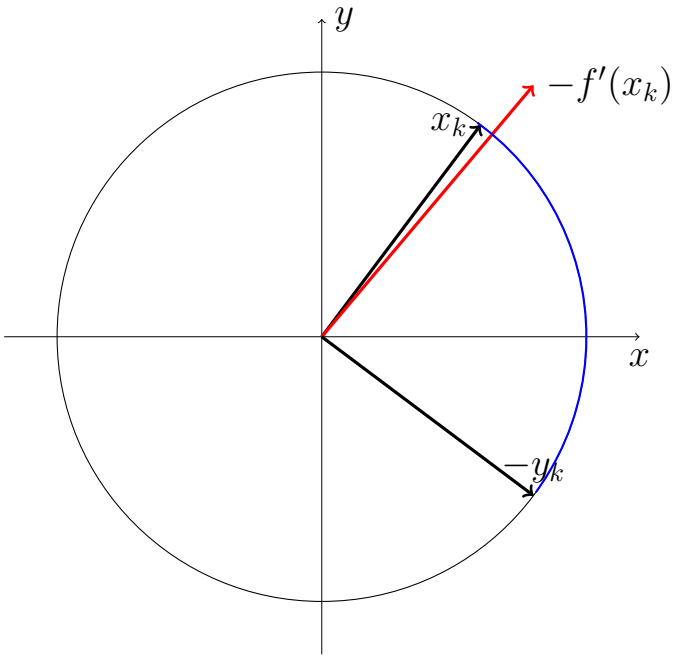
$$\|f'(x_k)\|^2 - \left\langle x_k, f'(x_k)\right\rangle^2 = d_k^2 \overset{k\to\infty}{\longrightarrow} 0.$$

Вспомним, что

$$d_k^2 = \left\langle f'(x_k), y_k\right\rangle^2 = \left\langle f'(x_k), \frac{f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k}{f'(x_k) - \left\langle x_k, f'(x_k)\right\rangle x_k}\right\rangle^2.$$

Физически, сходимость d_k^2 к нулю означает, что $f'(x_k)$ становится более ортогонально к y_k, а это приводит к коллинеарности с x_k.

Рис. 6. Близость к условию стационарности



Поэтому $\{x_k\}$ сходится к стационарной точке. $\square$

4. Квадратичная функция с линейным членом

4.1. Теоретические вопросы

В случае с линейным членом задача выглядит следующим образом:

$$\min_{x\in\mathbb{S}} \frac{1}{2}\big\langle Ax, x\big\rangle - \big\langle b, x\big\rangle. \tag{20}$$

Теперь обратимся к вопросу о стационарных точках. Здесь все не так легко, как в задаче (4), но все же можно воспользоваться спецификой целевой функции. Рассмотрим результат Уильяма Хагера[1]. Он сформулировал и доказал лемму, которая дает представление о стационарных точках в задаче. Введем обозначения, принятые в его лемме:

- $L = \frac{1}{2}\big\langle Ax, x\big\rangle - \big\langle b, x\big\rangle - \frac{\mu}{2}(\|x\|^2 - 1)$
- $\Lambda = \{\lambda_i\}_{i=1}^n : \ \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$
- $\mathbf{\Phi} = \{\phi_i\}_{i=1}^n : \ A\phi_i = \lambda_i\phi_i$
- $\mathcal{E}_1 = \{i \ : \ \lambda_i = \lambda_1\}$
- $\mathcal{E}_+ = \{i \ : \ \lambda_i > \lambda_1\}$
- $\beta_i = \big\langle b, \phi_i\big\rangle$

Теперь сформулируем лемму Хагера:

Лемма 4.1.1. ***(Hager)****Вектор* $\boldsymbol{\phi} = \sum_{i=1}^n c_i\phi_i$ *является решением задачи* (20) *тогда и только тогда, когда* $\mathbf{c}$ *выбирается следующим образом:*

- *Вырожденный случай: Если* $\beta_i = 0$ *для всех* $i \in \mathcal{E}_1$, *и*

$$\sum_{i\in\mathcal{E}_+} \frac{\beta_i^2}{(\lambda_i - \lambda_1)^2} \le 1, \tag{21}$$

тогда $\mu = -\lambda_1$, *и* $c_i = \beta_i/(\lambda_i - \lambda_1)$ *для* $i \in \mathcal{E}_+$; *оставшиеся* c_i *из* $i \in \mathcal{E}_1$ *выбираются произвольно так, чтобы*

$$\sum_{i\in\mathcal{E}_1} c_i^2 = 1 - \sum_{i\in\mathcal{E}_+} \frac{\beta_i^2}{(\lambda_i - \lambda_1)^2} \tag{22}$$

- *Невырожденный случай: Если предыдущее не выполнено, тогда* $c_i = \beta_i/(\lambda_i + \mu)$, $1 \le i \le n$, *где* $\mu > -\lambda_1$ *выбрано так, что*

$$\sum_{i=1}^{n} \frac{\beta_i^2}{(\lambda_i + \mu)^2} = 1 \tag{23}$$

Будем анализировать невырожденный случай. Сфокусируемся на стационарных точках. Для того, чтобы их получить, нужно численно решить (23). Проделаем это на примере со следующими матрицей A и вектором b:

$$A = \begin{pmatrix} 100 & 0 & 0 \\ 0 & -70 & 0 \\ 0 & 0 & -15 \end{pmatrix}, \; b = \begin{pmatrix} 50 \\ 20 \\ 10 \end{pmatrix}.$$

Обозначим

$$\Psi(\mu) = \sum_{i\in\mathcal{I}} \left(\frac{\beta_i}{\lambda_i - \mu} \right)^2,$$

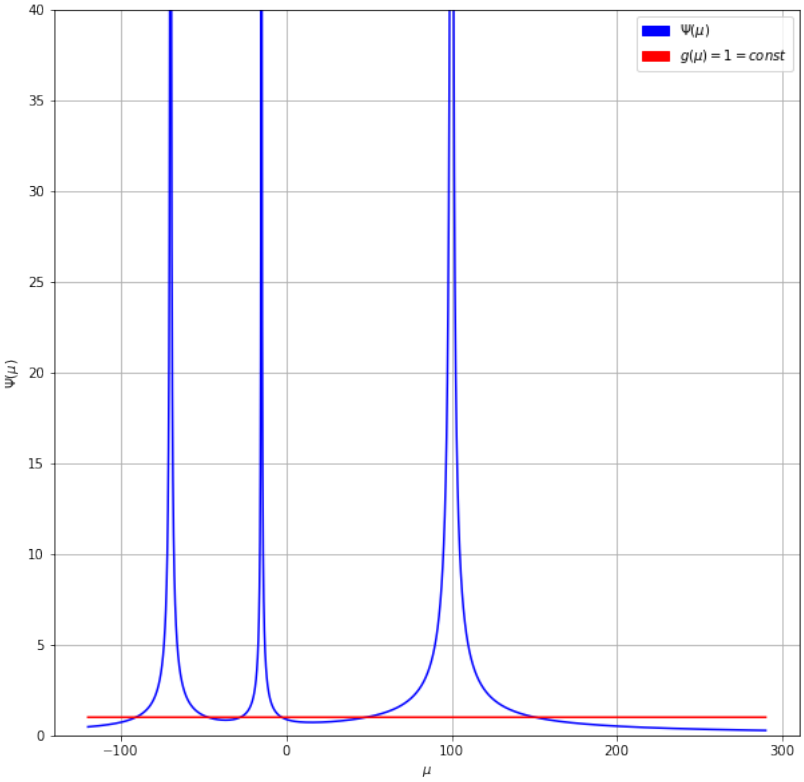
где множество индексов $\mathcal{I} = \{i : \beta_i \neq 0\}$.

Здесь μ, при которых $g(\mu) = 1$ пересекает $\Psi(\mu)$, являются теми множителями Лагранжа, при которых выполняется условие стационарности. Видно, что их несколько. А что если в этой задаче несколько локальных минимумов? Проверим это.

Вычислим эти μ, решив уравнение

$$\big(\Psi(\mu) - 1\big)^2 = 0.$$

Рис. 7. График функции $\Psi(\mu)$ и $g(\mu) = 1$.



Качественно с ситуацией можно ознакомиться на рисунке 7. Для вышеуказанных матрицы A и вектора b:

$$\mu^* = \begin{pmatrix} -90.92 \\ -47.50 \\ -27.67 \\ -2.83 \\ 48.62 \\ 150.30 \end{pmatrix} \tag{24}$$

Стационарные точки, в соответствующем порядке, представлены ниже:

$$\mathbf{x}^{*\mathbf{T}} = \begin{pmatrix} 0.26 & 0.96 & 0.13 \\ 0.34 & -0.89 & 0.30 \\ 0.39 & -0.47 & 0.79 \\ 0.47 & -0.30 & -0.82 \\ 0.97 & -0.17 & -0.16 \\ -0.99 & -0.09 & -0.06 \end{pmatrix}$$

Теперь проверим, единственен ли локальный минимум? Докажем вспомогательную лемму и теорему о касательном подпространстве в точке x. Теорема специфична именно для данной целевой функции – квадратичной с линейным членом.

Лемма 4.1.2. *Пусть вектор x: $\|x\| = 1$, вектор $z \in \mathbb{R}^n$ – неколлинеарен. Тогда проекция z на $\mathbb{S} = \{s \in \mathbb{R}^n \ \langle s, x \rangle = 0\}$ представляется в следующем виде:*

$$P_{\mathbb{S}}(z) = (I - xx^T)z = z - \langle z, x \rangle x. \tag{25}$$

Доказательство. По Лемме 2.2.2 для $x_k - \gamma f'(x_k) := z$, $x_k := x$, вычитая из результата x, получаем требуемое утверждение. □

Теперь сформулируем и докажем теорему, которая позволит легко проверять: является ли данная точка локальным минимумом?

Теорема 4.1.3. *Пусть x^* – стационарная точка в задаче (4), соответствующий множитель Лагранжа μ^*. Тогда точка x^* является локальным минимумом задачи, если матрица*

$$(I - x^* x^{*T})(A - \mu^* I)(I - x^* x^{*T})$$

положительно определена.

Доказательство. Основной инструмент – достаточное условие локального минимума второго порядка. Согласно ему, точка x^* является локальным минимумом, если

$$\left\langle L''_{xx}s, s\right\rangle > 0 \ \forall \ s \in \mathcal{S} = \{s \ : \ \left\langle s, x^*\right\rangle = 0\}. \tag{26}$$

То есть L''_{xx} – положительно определена на касательном подпространстве к точке x^*.

Посмотрим подробнее на это условие. Для начала посчитаем L''_{xx}:

$$\begin{aligned} L &= \frac{1}{2}\left\langle Ax, x\right\rangle - \left\langle b, x\right\rangle - \frac{\mu}{2}(\|x\|^2 - 1), \\ L'_x &= Ax - b - \mu x, \\ L''_{xx} &= A - \mu I. \end{aligned}$$

Применим к (26) результат леммы (4.1.2).

$$\left\langle (A - \mu^* I)s, s\right\rangle$$

Здесь $s \in \mathcal{S}$. С помощью вышедоказанной леммы, s можно представить в виде

$$s = (I - x^* x^{*T})z, \ z \in \mathbb{R}^n.$$

Заметим, что $\left(I - x^* x^{*T}\right)^T = \left(I - (x^* x^{*T})^T\right) = \left(I - x^* x^{*T}\right)$, то есть это симметричная матрица.

Тогда

$$\begin{aligned}\big\langle (A-\mu^* I)s, s\big\rangle = \\ &= \big\langle (A-\mu^* I)(I - x^* x^{*T})z, (I - x^* x^{*T})z\big\rangle = \\ &= \big\langle (I - x^* x^{*T})(A-\mu^* I)(I - x^* x^{*T})z, z\big\rangle.\end{aligned}$$

А условие (26) перепишется в виде

$$\big\langle (I - x^* x^{*T})(A-\mu^* I)(I - x^* x^{*T})z, z\big\rangle > 0 \ \forall\ z \in \mathbb{R}^n \setminus \{0\}.$$

То есть, $(I - x^* x^{*T})(A-\mu^* I)(I - x^* x^{*T})$ – положительно определена. $\square$

Итак, с помощью этой теоремы можно исследовать, какие стационарные точки являются локальными минимумами. Используем Алгоритм 2 для нашей целевой функции

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle,$$

с матрицей

$$A = \begin{pmatrix} 100 & 0 & 0 \\ 0 & -70 & 0 \\ 0 & 0 & -15 \end{pmatrix},$$

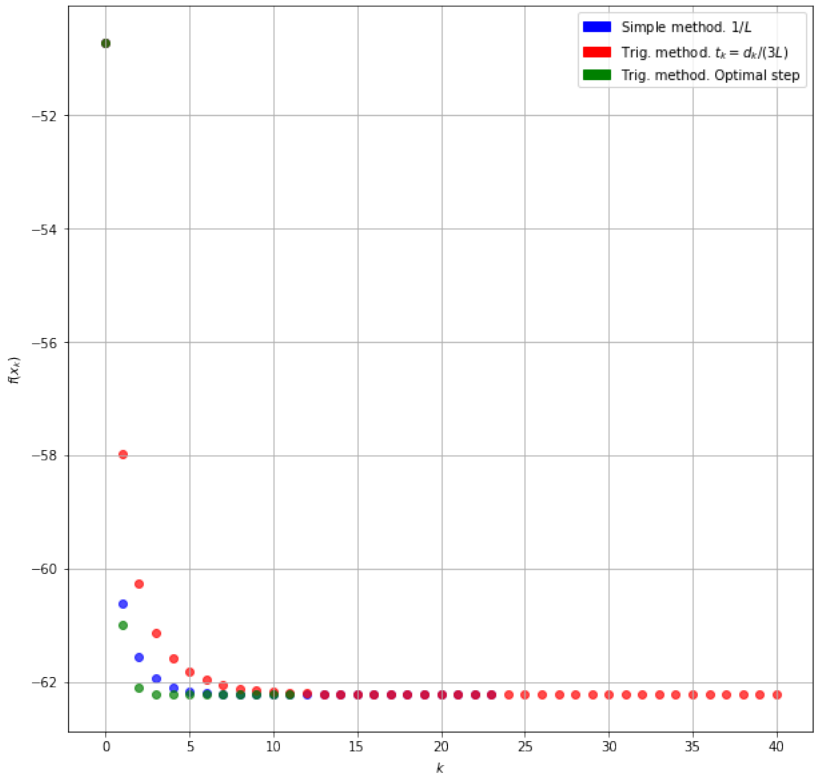
и вектором

$$b = \begin{pmatrix} 50 \\ 20 \\ 10 \end{pmatrix}.$$

Заодно сравним работу метода с простым методом проекции градиента. На рисунке 8 представлены результаты работы методов проек-

ции градиента, тригонометрического (Алгоритм 2) с шагами из Теоремы 3.4.2 и выбор шага методом наискорейшего спуска.

Рис. 8. Методы проекции градиента, тригонометрический с явным шагом и с шагом наискорейшего спуска



Здесь методы были запущены из точки

$$x_0 = \begin{pmatrix} -0.027 \\ 0.899 \\ 0.437 \end{pmatrix}. \tag{27}$$

Все сошлись в точку

$$x^* = \begin{pmatrix} 0.262 \\ 0.956 \\ 0.132 \end{pmatrix}$$

За 22, 39, 10 итераций, соответственно. Значение целевой функции $f^* = -62.226$. Теперь запустим методы из другой начальной точки:

$$x_0 = \begin{pmatrix} -0.009 \\ -0.854 \\ -0.521 \end{pmatrix}. \tag{28}$$

Все сошлись в другую точку

$$x^* = \begin{pmatrix} 0.339 \\ -0.889 \\ 0.308 \end{pmatrix}.$$

за 54, 113, 9 итераций, значение целевой функции $f^* = -24.874$.

Возможно, это два локальных минимума. Воспользуемся Теоремой 4.1.3. С таким инструментом, нам нужно просто проверить матрицу на положительную определенность. Для этого можно ее просто диагонализовать и посмотреть на диагональные элементы (собственные значения).

Сначала рассмотрим решение, которое дает меньшее значение целевой функции:

$$x_1^* = \begin{pmatrix} 0.262 \\ 0.956 \\ 0.132 \end{pmatrix}.$$

Нужная нам для проверки матрица имеет следующий вид:

$$(I - x_1^* x_1^{*T})(A - \mu_1^* I)(I - x_1^* x_1^{*T}).$$

До этого, мы вычислили μ^*, соответствующие каждой стационарной точке – (24). Нас интересует $\mu_1^* = -90.92$. Диагонализуем, получаем, что

$$(I - x_1^* x_1^{*T})(A - \mu_1^* I)(I - x_1^* x_1^{*T}) = U_1^{-1} diag(446.784, 102.158, 26.573) U_1,$$

где U_1 – матриц перехода в базис из собственных векторов матрицы слева. Как видно, все собственные значения положительны, значит, по критерию Сильвестра матрица определена положительно, а x_1 локальный минимум.

Что же в случае с

$$x_2^* = \begin{pmatrix} 0.339 \\ -0.889 \\ 0.308 \end{pmatrix},$$

$$\mu_2^* = -47.50?$$

$$(I - x_2^* x_2^{*T})(A - \mu_2^* I)(I - x_2^* x_2^{*T}) = U_2^{-1} diag(-32.177, 44.258, 302.947) U_2,$$

где U_2 – матрица перехода в базис из собственных векторов матрицы слева. Не все собственные значения положительны, не все главные миноры положительны, значит, по критерию Сильвестра, матрица не является положительно определенной.

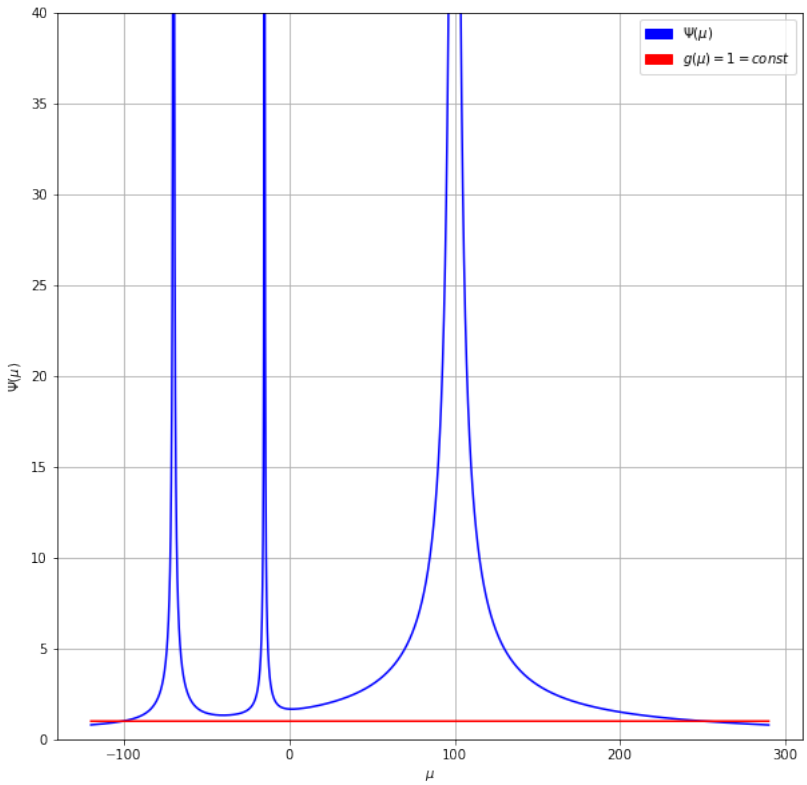
Полученный результат говорит о том, что в такой задаче (с такой матрицей A и вектором b) метод может сойтись в стационарную точку. Это важный вывод. Посмотрим на график $\Psi(\mu)$ при другой матрице и другом векторе. Например, при той же матрице, но с другим вектором:

$$A = \begin{pmatrix} 100 & 0 & 0 \\ 0 & -70 & 0 \\ 0 & 0 & -15 \end{pmatrix},$$

$$b = \begin{pmatrix} 150 \\ 20 \\ 10 \end{pmatrix}.$$

График функции $\Psi(\mu)$ можно увидеть на рисунке 9

Рис. 9. $\Psi(\mu)$ с двумя пересечениями с $g(x) = const = 1$



Здесь всего два пересечения, то есть всего две стационарные точки. Это напоминает ситуацию, когда целевая функция не квадратичная, а линейная.

То есть решается задача

$$\min_{x \in \mathbb{S}} \langle b, x \rangle.$$

Функция Ланранжа задачи:

$$L = \langle b, x \rangle - \frac{\mu}{2}(\|x\|^2 - 1).$$

Необходимое условие минимума первого порядка:

$$L' = \begin{pmatrix} b - \mu x \\ \|x\|^2 - 1 \end{pmatrix} = 0.$$

Отсюда $x = b/\mu$, а $\mu = \pm\|b\|$. Достаточное условие локального минимума второго порядка: $L'' = -\mu I \succ 0$ на касательном подпространстве. Отсюда $\mu = -\|b\|$. Если бы искали максимум – то $\mu = \|b\|$. Две стационарные точки: одна минимум, другая максимум.

Так на рисунке 9 похожая ситуация. Возможно, квадратичная часть не так сильно "искривила" линейную функцию и свойства линейной части доминируют. Может можно указать что-то вроде соотношения между нормами матрицы A и вектора b, при котором сохраняется линейная тенденция?

Рассмотрим целевую функцию с числовым параметром τ

$$f(x, \tau) = \frac{\tau}{2}\langle Ax, x \rangle - \langle b, x \rangle,$$

будем решать задачу

$$\min_{\text{s.t. } x \in \mathbb{S}} f(x, \tau) = \frac{\tau}{2}\langle Ax, x \rangle - \langle b, x \rangle. \tag{29}$$

при различных τ. Сперва заметим, что каждое собственное значение λ перешло в $\tau\lambda$. Попробуем выяснить, при каких τ есть только 2 стационарные точки. То есть нужно, чтобы $\Psi_\tau(\mu)$ пересекала $g(x) = const = 1$

только в левой краней точке и в правой.

Взглянем на рисунок 9 и на вид $\Psi_\tau(\mu)$ еще раз. Рисунок даст нам качественное понимание картины.

$$\Psi_\tau(\mu) = \sum_{i\in\mathcal{I}} \left(\frac{\beta_i}{\tau\lambda_i - \mu}\right)^2.$$

Далее нужно будет рассматривать лишь $i \in \mathcal{I}$. Для удобства будем обозначать их $i_1, i_2, \ldots, i_{|\mathcal{I}|}$. Заметим, что функция обращается в бесконечность в точках $\mu = \tau\lambda_i$, то есть в собственных значениях. В промежутках между ними функция убывает, далее, при μ, приближающихся к следующему собственному значению, функция возрастает. Ясно, что потребовав $\Psi(\mu) > 1,\ \mu \in (\tau\lambda_{i_j}, \tau\lambda_{i_{j+1}})\ \forall i_j \in \mathcal{I} \setminus \max\{\mathcal{I}\}$, будет всего два пересечения $\Psi(\mu)$ и $g(x) = const = 1$. То есть будут всего две стационарные точки. Сформиулируем лемму.

Лемма 4.1.4. *В задаче* (29) *существует достаточное* τ_0*: при* $\tau < \tau_0$ *задача имеет две стационарные точки. Величина* τ_0 *имеет вид*

$$\tau_0 = \sqrt{\frac{1}{2\|A\|} \min_{i\in\mathcal{I}\setminus\max\{\mathcal{I}\}} \left\{\frac{\beta_{i_j}^2}{\theta_{0i_j}^2} + \frac{\beta_{i_{j+1}}^2}{(1-\theta_{0i_j})^2}\right\}}$$

Доказательство. Рассмотрим промежуток для μ между двумя соседними собственными значениями:

$$\mu \in (\lambda_{i_j}, \lambda_{i_{j+1}}),\ i_j \in \mathcal{I} \setminus \max\{\mathcal{I}\}.$$

Параметризуем переменную:

$$\mu = \tau\big(\lambda_{i_j} + \theta(\lambda_{i_{j+1}} - \lambda_{i_j})\big), \theta \in (0, 1).$$

Разобьем $\Psi_\tau(\mu)$ на слагаемые с $\lambda_{i_j},\ \lambda_{i_{j+1}}$ и без:

$$\Psi_\tau(\mu) = \sum_{k\in\mathcal{I}\setminus\{i_j,i_{j+1}\}} \frac{\beta_k^2}{(\tau\lambda_k-\mu)^2} +$$
$$+ \frac{\beta_{i_j}^2}{\cancel{\tau\lambda_i} - \cancel{\tau\lambda_i}\theta(\tau\lambda_{i+1}-\tau\lambda_i)^2} + \frac{\beta_{ij+1}^2}{\tau\lambda_{i+1}-\tau\lambda_i\theta(\tau\lambda_{i+1}-\tau\lambda_i)^2} =$$
$$= \sum_{k\in\mathcal{I}\setminus\{i_j,i_{j+1}\}} \frac{\beta_k^2}{(\tau\lambda_k-\mu)^2} + \frac{\beta_{i_j}^2}{\tau^2\theta^2(\lambda_{i+1}-\lambda_i)^2} + \frac{\beta_{i_{j+1}}^2}{\tau^2(1-\theta)^2(\lambda_{i+1}-\lambda_i)^2}.$$

Здесь ясно видно, что уменьшая τ можно добиться того, чтобы $\Psi_\tau(\mu)$ превосходило единицу, поскольку сумма при рассматриваемых $\mu\in(\tau\lambda_i,\tau\lambda_{i_{j+1}})$ – конечна, так же, как и два остальных слагаемых.

Из вида функции $\Psi_\tau(\mu)$ на рассматриваем промежутке (качественно на рисунке 9) видно, что у нее есть точка минимума. Действительно, при $\mu=\tau\lambda_{i_j},\ \tau\lambda_{i_{j+1}}$, функция обращается в $+\infty$, а так как она непрерывна, то минимум находится где-то в середине.

Здесь сделаем сильное упрощение. Пренебрежем суммой, будем рассматривать только два слагаемых с $\tau\lambda_{i_j},\ \tau\lambda_{i_{j+1}}$. Для наших целей это достаточно.

$$\hat{\Psi}_{tau}(\theta) = \frac{\beta_{i_j}^2}{\tau^2\theta^2(\lambda_{i_{j+1}}-\lambda_{i_j})^2} + \frac{\beta_{i_{j+1}}^2}{\tau^2(1-\theta)^2(\lambda_{i_{j+1}}-\lambda_{i_j})^2}.$$

Найдем стационарные точки $\hat{\Psi}_\tau(\mu)$ на $(\tau\lambda_{i_j},\tau\lambda_{i_{j+1}})$. Напомним, что мы приняли

$$\mu = \tau\big(\lambda_{i_j}+\theta(\lambda_{i_{j+1}}-\lambda_{i_j})\big),$$

поэтому производную будем брать по θ.

$$\hat{\Psi}'_{tau}(\theta) = -2\frac{\beta_{i_j}^2}{\tau^2\theta^3(\lambda_{i_{j+1}} - \lambda_{i_j})^2} + 2\frac{\beta_{i_{j+1}}^2}{\tau^2(\lambda_{i_{j+1}} - \lambda_{i_j})^2(1-\theta)^3} = 0.$$

$$\frac{\beta_{i_{j+1}}^2}{\theta^3} - \frac{\beta_{i_j}^2}{(1-\theta)^3} = 0.$$

Отсюда видно, что равентсво имеет место в пределе – при $\theta \to \pm\infty$, но мы договорились, что $\theta \in (0, 1)$.

Поэтому нам подходит

$$\theta_{0i_j} = \frac{\beta_{i_{j+1}}^{2/3}}{\beta_{i_{j+1}}^{2/3} + \beta_{i_j}^{2/3}}.$$

Это число между нулем и единицей. Действительно, как мы отмечали ранее, когда только ввели функцию $\Psi(\mu)$, все β_{i_j} не нулевые, а $\beta_{i_j}^{2/3} > 0 \ \forall \ i_j \in \mathcal{I}$. Это так же распространяется и на $\hat{\Psi}_\tau(\mu)$

Мы грубо нашли положение нижней точки функции $\Psi_\tau(\mu)$:

$$\mu = \tau\left(\lambda_{i_j} + \frac{\beta_{i_{j+1}}^{2/3}}{\beta_{i_{j+1}}^{2/3} + \beta_{i_j}^{2/3}}(\lambda_{i_{j+1}} - \lambda_{i_j})\right)$$

Если мы потребуем, чтобы в этой точке $\hat{\Psi}_\tau(\mu) > 1$, то гарантированно это будет выполнено и для всей $\Psi_\tau(\mu)$, ибо все слагаемые положительны. Сделаем последний шаг.

$$\hat{\Psi}_\tau(\mu) = \frac{\beta_{i_j}^2}{\tau^2\theta_{0i_j}^2(\lambda_{i_{j+1}} - \lambda_{i_j})^2} + \frac{\beta_{i_{j+1}}^2}{\tau^2(1-\theta_{0i_j})^2(\lambda_{i_{j+1}} - \lambda_{i_j})^2} > 1 \Longrightarrow$$

$$\tau^2 < \frac{\beta_{i_j}^2}{\theta_{0i_j}^2(\lambda_{i_{j+1}} - \lambda_{i_j})^2} + \frac{\beta_{i_{j+1}}^2}{(1-\theta_{0i_j})^2(\lambda_{i_{j+1}} - \lambda_{i_j})^2}.$$

Это должно быть выполнено для всех интервалов $(\lambda_{i_j}, \lambda_{i_{j+1}}),\ i_j \in \mathcal{I}$. Отсюда

$$\tau^2 < \min_{i \in \mathcal{I} \setminus \max\{\mathcal{I}\}} \left\{ \frac{\beta_{i_j}^2}{\theta_{0i_j}^2 (\lambda_{i_{j+1}} - \lambda_{i_j})^2} + \frac{\beta_{i_{j+1}}^2}{(1-\theta_{0i_j})^2 (\lambda_{i_{j+1}} - \lambda_{i_j})^2} \right\}.$$

Неравенство можно упростить еще, если принять во внимание факт $|\lambda_{i_{j+1}} - \lambda_{i_j}| < 2\|A\|$ и потребовать от τ^2 еще более жесткое условие, но более простое:

$$\begin{aligned} \tau^2 &< \frac{1}{2\|A\|} \min_{i \in \mathcal{I} \setminus \max\{\mathcal{I}\}} \left\{ \frac{\beta_{i_j}^2}{\theta_{0i_j}^2} + \frac{\beta_{i_{j+1}}^2}{(1-\theta_{0i_j})^2} \right\} < \\ &< \min_{i \in \mathcal{I} \setminus \max\{\mathcal{I}\}} \left\{ \frac{\beta_{i_j}^2}{\theta_{0i_j}^2 (\lambda_{i_{j+1}} - \lambda_{i_j})^2} + \frac{\beta_{i_{j+1}}^2}{(1-\theta_{0i_j})^2 (\lambda_{i_{j+1}} - \lambda_{i_j})^2} \right\}. \end{aligned}$$

□

Таким образом, анализ сводится к рассмотрению компонент вектора b. Помимо этого, нужно знать, хотя бы оценку сверху, нормы матрицы A.

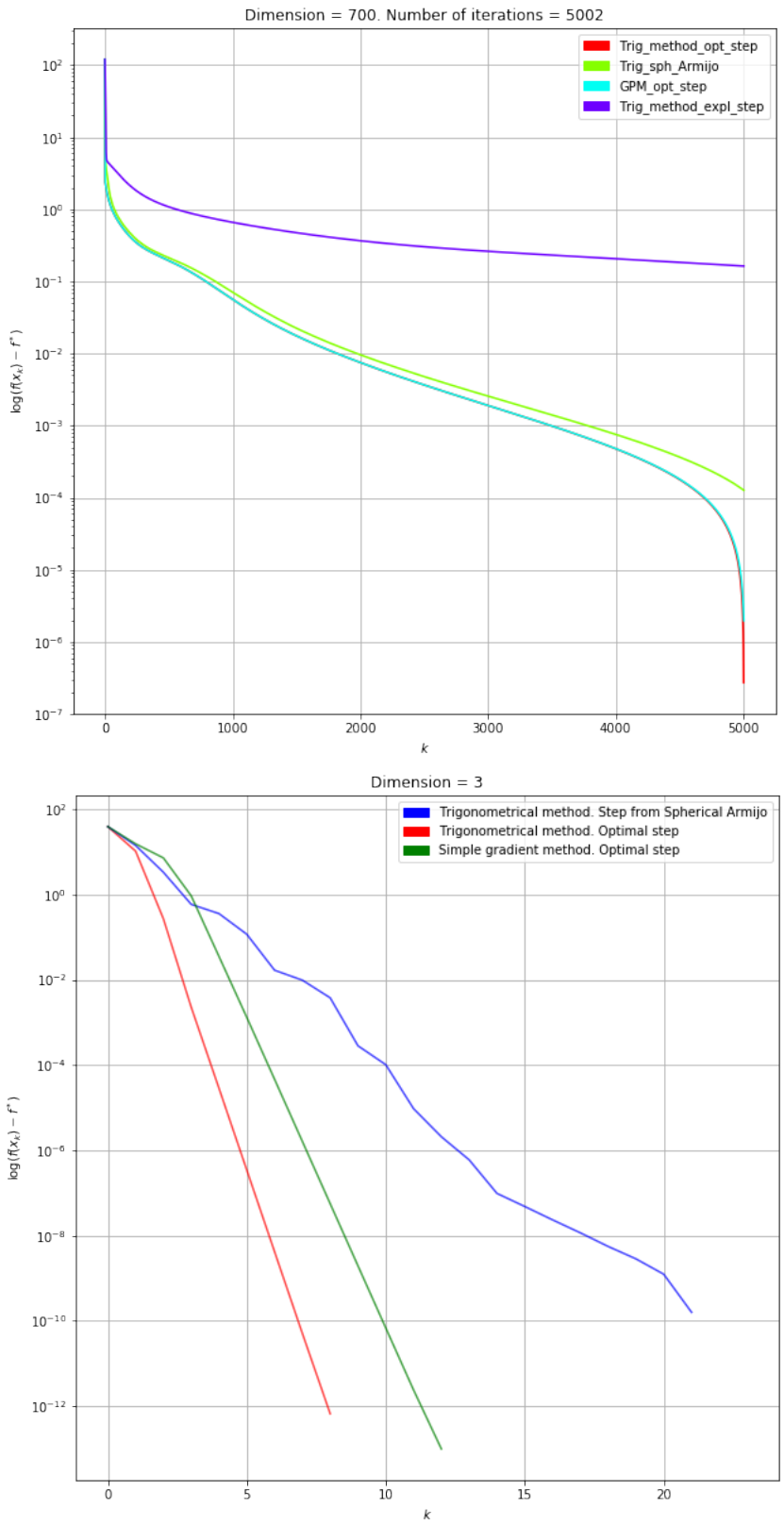
4.2. Численные эксперименты

Применим метод из секции 3 к квадратичной функции с линейным членом. Сравнивать будем метод проекции градиента с шагом, возникающим при одномерной минимизации и с шагом из сферического правила Армихо. Рисунок 10. Здесь стоит обратить внимание на время работы методов при размерности 700. Оно кардинально различается от метода к методу:

Метод	**Точн.**	**Время(с)**
Триг. м. опт. шаг Красный	10^{-7}	18.6
Триг. м. явный шаг Синий	10^{-1}	4.4
Триг. м. Армихо Зеленый	10^{-4}	53.4
МПГ опт. шаг Голубой	10^{-6}	72.7

Таблица 2. Времена работы

Рис. 10. Тригонометрические методы и метод проекции градиента



5. Выводы

Исследованная задача имеет много тонкостей как в теоритечских вопросах, так и в численных. Рассмотрен простой и важный случай однородной квадратичной функции, предложен метод решения, схожий с методом проекции градиента, доказана его сходимость к решению задачи для данной целевой функции. Для общего случая построен метод с тригонометрическим представлением геодезической сферы, несколько правил выбора шага, доказана сходимость к стационарной точке в обоих случаях. Случай с неоднородной квадратичной функцией показывает, что при определенных условиях на вектор b задача сильно упрощается в смысле количества стационарных точек. Для предложенных методов было показано, что имеет место сходимость в стационарную точку.В будущем стоит ответить на вопрос о скорости сходимости методов, обобщить мысли о количестве стационарных точек на случай неквадратичной функции.

Список литературы

1. Hager W. (2001), MINIMIZING A QUADRATIC OVER A SPHERE, SIAM J. OPTIM. Vol. 12, No. 1, pp. 188-208
2. Quadratic Programming Over Ellipsoids with Applications to Constrained Linear Regression and Tensor Decomposition, Anh-Huy Phan · Masao Yamagishi · Danilo Mandic · Andrzej Cichocki (2017), arXiv:1711.04401 [math.OC]
3. THE GEOMETRY OF ALGORITHMS WITH ORTHOGONALITY CONSTRAINTS, ALAN EDELMAN, T.A. ARIAS AND STEVEN T. SMITH (1998), SIAM J. Matrix Anal. Appl
4. Nesterov, Y. (2013). Introductory lectures on convex optimization: A basic course(Vol. 87). Springer Science & Business Media
5. Поляк Б. Т. (1987). Введение в оптимизацию
6. Фаддеев Д. К., Фаддеева В. Н. (1975), стр: 480-487
7. Petre Birtea, Ioan Caşu and Dan Comănescu (2017). Steepest descent algorithm on orthogonal Stiefel manifolds, arXiv:1709.06295v1 [math.OC] 19 Sep 2017
8. Linear convergence of first order methods for non-strongly convex optimization (2016) I. Necoara · Yu. Nesterov · F. Glineur arXiv:1504.06298 [math.OC]
9. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak- Lojasiewicz Condition Hamed Karimi, Julie Nutini, and Mark Schmidt Department of Computer Science, University of British Columbia Vancouver, British Columbia, Canada