Квадратичная функция на единичной сфере

Лукашевич Александр
МФТИ 476

1. Постановка задачи

Будем рассматривать квадратичную функцию на единичной сфере без линейной части:

$$\begin{gathered} \min_{x \in S} f(x) = \frac{1}{2} x^T A x \\ S = \{x \in \mathbb{R}^n \big| \|x\|^2 = 1\} \end{gathered} \tag{1.1}$$

Сразу же отметим, что в конечномерном евклидовом пространстве $\mathbb{R}^n$ единичная сфера является компактным множеством, что говорит о том, что решение задачи существует, поскольку целевая функция непрерывна. Прежде чем приступить к вопросам об условиях экстремума запишем функцию Лагранжа:

$$L(x, \lambda) = \frac{1}{2} x^T A x + \lambda\big(\|x\|^2 - 1\big). \tag{1.2}$$

2. Условия экстремума,стационарные точки и точки локального минимума

2.1. Необходимые условия

- *Необходимым условием (первого порядка)* того, что точка x_0 является точкой экстремума задачи (1.1), является существование вектора множителей Лагранжа λ_0, такого, что (x_0, λ_0) - стационарная точка функции Лагранжа (7.1):

$$\nabla L(x_0, \lambda_0) = \vec{0}. \tag{2.1}$$

  В случае квадратичной функции (2.1) выражается следующий образом:

$$\nabla L(x_0, \lambda_0) = \begin{pmatrix} Ax_0 + 2\lambda_0 x_0 \\ \|x_0\|^2 - 1 \end{pmatrix} = \begin{pmatrix} \vec{0} \\ 0 \end{pmatrix}$$

  Из этого условия явно видно, что выполняется условие допустимости.

- *Необходимым условием (второго порядка)* того, что точка x_0 является точкой экстремума задачи (1.1) является следующее неравенство:

$$\begin{gathered} \big(L_{xx}(x_0, \lambda_0)s, s\big) = \big((A + 2\lambda_0 E)s, s\big) \geq 0 \\ s \in \mathbb{S} = \{s \big| (2x, s) = 0\}, \end{gathered} \tag{2.2}$$

  где $2x$ суть градиент ограничения, отражающего принадлежность точки к сфере.

2.2. Достаточные условия

- *Достаточными условиями второго порядка* для того, чтобы точка x_0 была точкой локального минимума в задаче (1.1) является существование вектора множителей Лагранжа λ_0 такого, что выполнены следующие условия:

$$\begin{gathered} \nabla L(x_0, \lambda_0) = \vec{0}, \\ \big(L_{xx}(x_0, \lambda_0)s, s\big) = \big((A + 2\lambda_0 E)s, s\big) > 0, \end{gathered} \tag{2.3}$$

  где $s \in \mathbb{S}$, определенном в (2.2)

2.3. Стационарные точки и точки локального минимума

- Эти виды точек в данной задаче определяются условиями (2.1) и (2.3) соответственно. Рассмотрим этот вопрос подробнее.
  - Запишем функцию Лагранжа:

$$L = \frac{1}{2} x^T A x + \lambda\big(\|x\|^2 - 1\big). \tag{2.4}$$

  - Необходимое условие первого порядка:

$$\nabla L = \begin{pmatrix} Ax + 2\lambda x \\ \|x\|^2 - 1 \end{pmatrix} = \vec{0}. \tag{2.5}$$

– Вторая производная L по x:

$$L_{xx} = A + 2\lambda E. \tag{2.6}$$

– Касательное подпространство к $\mathbb{S}$:

$$E_t = \{s \in \mathbb{R}^n : (s, 2x) = 0\}. \tag{2.7}$$

– Условие минимума второго порядка:
Если x^* такова, что

$$\begin{gathered}\nabla L(x^*, \lambda^*) = \begin{pmatrix} Ax^* + 2\lambda^* x^* \\ \|x^*\|^2 - 1 \end{pmatrix} = \vec{0}. \\ \forall s \in E_t \hookrightarrow (L_{xx}(x^*, \lambda^*)s, s) > 0.\end{gathered} \tag{2.8}$$

Тогда x^* - точка локального минимума в задаче (1.1).

Рассмотрим (2.8) подробнее. Учитывая (2.6),

$$(L_{xx}(x^*, \lambda^*)s, s) = ((A + 2\lambda^* E)s, s) = s^T As + 2\lambda^* s^T s. \tag{2.9}$$

Рассмотрим A как матрицу некоторого оператора. Так как матрица симметрична, то все собственные значения вещественны, помимо этого, собственные векторы попарно ортогональны.

Из (4.20) следуюет, что

$$Ax^* = -2\lambda^* x^*, \tag{2.10}$$

То есть стационарные точки есть собственные векторы оператора A - $\{h_i\}_{i=1}^n$, соответствующие собственным значениям $\{-2\lambda_i\}_{i=1}^n$.

Поскольку при проверке (2.8) нас будут интересовать лишь $s \in E_t$, рассмотрим это множество подробнее: для фиксированной стационарной точки, E_t представляет собой ортогональное дополнение собственного вектора, соответствующего этой стационарной точке. Поскольку собственные векторы оператора A попарно ортогональны, то s можно разложить по остальным собственным векторам. Пусть стационарной точке соответствует собственный вектор с номером k. Тогда

$$s = \sum_{i=1}^{k-1} \alpha_i h_i + \sum_{i=k+1}^{n} \alpha_i h_i.$$

Чтобы не загромождать запись, будем считать, что $k = n$, что не умаляет общности.

Используя вышесказанное, перепишем (2.9):

$$s^T As + 2\lambda s^T s = \left(\sum_{i=1}^{n-1} \alpha_i h_i\right)^T A \left(\sum_{i=1}^{n-1} \alpha_i h_i\right) + 2\lambda^* \left(\sum_{i=1}^{n-1} \alpha_i h_i\right)^T \left(\sum_{i=1}^{n-1} \alpha_i h_i\right). \tag{2.11}$$

Поскольку каждый h_i - собственный вектор оператора A, то

$$Ah_i = -2\lambda_i h_i.$$

Итак, (2.11) принимает вид

$$\begin{gathered}\left(\sum_{i=1}^{n-1} \alpha_i h_i\right)^T \left(\sum_{i=1}^{n-1} (-2\lambda_i)\alpha_i h_i\right) + 2\lambda^* \left(\sum_{i=1}^{n-1} \alpha_i h_i\right)^T \left(\sum_{i=1}^{n-1} \alpha_i h_i\right) = \\ = \left(\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} (-2\lambda_i)\alpha_i h_i\right) + 2\lambda^* \left(\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right).\end{gathered} \tag{2.12}$$

Поскольку $\{h_i\}_{i=1}^n$ попарно ортогональны, то (2.12) примет вид

$$\sum_{i=1}^{n-1} 2(-\lambda_i + \lambda^*)\alpha_i^2 (h_i, h_i). \tag{2.13}$$

В силу произвольности вектора $s \in E_t$, для того, чтобы (2.13) было положительным, каждое $(-\lambda_i + \lambda^*)$ должно быть положительно. Таким образом, $\lambda^* > \lambda_i \ \forall i = 1, \ldots, n-1$, $-2\lambda^* < -2\lambda_i$.

Итак, точкой локального минимума будет та точка, множитель Лагранжа которой наибольший, то есть собственный вектор матрица A, с наименьшим собственным значением.

3. Выпуклая функция

Критерий выпуклости второго порядка выражается неравенством :

$$\nabla^2 f(x) = A \succ 0 \iff \forall s \in \mathbb{R}^n \to s^T A s \geq 0.$$

Факт положительной определенности можно установить с помощью критерий Сильвестра.

4. Методы

4.1. Градиентный метод

Рассмотрим градиентный метод для безусловной оптимизации. Общая схема метода выглядит так:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k p^k \\ \gamma_k &= const \\ p^k &= \nabla f(x^k) \end{aligned} \tag{4.1}$$

Данный метод сходится при следующих условиях:

- Градиент Липшицев с константой L
- $0 < \gamma < \frac{2}{L}$
- $f(x)$ ограниченая снизу

В таком случае градиент стремится к нулю, а $f(x)$ монотонно убывает на последовательности x^k.

Градиент функции $f(x) = \frac{1}{2}x^T Ax + b^T x + c$ Липшицев. Действительно:

$$\forall x, y \; \|\nabla f(x) - \nabla f(y)\| = \|Ax + b - (Ay + b)\| = \|A(x - y)\| \leq \|A\| \cdot \|x - y\|$$

То есть константа Липшица есть норма оператора A. Это число, так как в конечномерных пространствах все операторы непрерывны, что эквивалентно ограниченности.

$f(x)$ ограничена снизу на S, поскольку она непрерывна. В нашем случае стоит применить более общий метод - метод проекции градиента:

$$x^{k+1} = P_S(x^k - \gamma_k \nabla f(x^k)) = P_S((E - \gamma A)x^k - \gamma b), \tag{4.2}$$

где $P_S(x) = \underset{s \in S}{argmin} \|x - s\| = \frac{x}{\|x\|}$ - проекция x на S.

Поскольку оператор проектирования $P_S(x)$ обладает свойством Липшица с $L = 1$, то есть $\|P_S(x) - P_S(y)\| \leq \|x - y\|$, то условия сходимости те же, что и для безусловного метода.

4.2. Скорейший спуск

4.2.1. Немного о методе

Этот метод отличается от предыдущего тем, что шаг γ не постоянный, а выбирается в соответствии с определенным правилом, например, в соответсвии с правилом одномерной минимизации, которое будет описано ниже.

Сразу запишем метод для нашей задачи - с ограничением на принадлежность сфере:

$$\begin{aligned} x^{k+1} &= P_S(x^k - \gamma_k \nabla f(x^k)) \\ \gamma_k &= \underset{\gamma \geq 0}{argmin} f(P_S(x^k - \gamma \nabla f(x^k))) \end{aligned} \tag{4.3}$$

4.2.2. Выбор шага

$$\begin{aligned} x^{k+1} = P_S(x^k - \gamma_k \nabla f(x^k)) &= \frac{x^k - \gamma_k \nabla f(x^k)}{\|x^k - \gamma_k \nabla f(x^k)\|} \\ \gamma_k = \underset{\gamma \geq 0}{argmin} f(P_S(x^k - \gamma \nabla f(x^k))) &= \underset{\gamma \geq 0}{argmin} f\left(\frac{x^k - \gamma \nabla f(x^k))}{\|x^k - \gamma \nabla f(x^k))\|}\right). \end{aligned} \tag{4.4}$$

Учтем вид функции: $f(x) = \frac{1}{2}x^T Ax$ при расчете шага:

$$\gamma_k = \underset{\gamma \geq 0}{argmin}\ \frac{1}{2}\left(\frac{x^k - \gamma Ax^k}{\|x^k - \gamma Ax^k\|}\right)^T A \frac{x^k - \gamma Ax^k}{\|x^k - \gamma Ax^k\|} = \tag{4.5}$$

$$= \underset{\gamma \geq 0}{argmin}\ \frac{1}{2}\ \frac{(x^k)^T Ax^k - 2\gamma (x^k)^T A^2 x^k + \gamma^2 (x^k)^T A^3 x^k}{< x^k - \gamma Ax^k, x^k - \gamma Ax^k >} = \tag{4.6}$$

$$= \underset{\gamma \geq 0}{argmin}\ \frac{1}{2}\ \frac{(x^k)^T Ax^k - 2\gamma (x^k)^T A^2 x^k + \gamma^2 (x^k)^T A^3 x^k}{(x^k - \gamma Ax^k)^T E (x^k - \gamma Ax^k)} = \tag{4.7}$$

$$= \underset{\gamma \geq 0}{argmin}\ \frac{1}{2}\ \frac{(x^k)^T Ax^k - 2\gamma (x^k)^T A^2 x^k + \gamma^2 (x^k)^T A^3 x^k}{(x^k)^T x^k - 2\gamma (x^k)^T Ax^k + \gamma^2 (x^k)^T A^2 x^k} = \tag{4.8}$$

$$= \underset{\gamma \geq 0}{argmin}\ \frac{1}{2}\ F(x^k, \gamma). \tag{4.9}$$

Поскольку в знаменателе стоит квадрат нормы вектора $x^k - \gamma Ax^k$, то, вводя обозначения $a_i(x^k) = (x^k)^T A^i x^k,\ i = 0, \dots 3$, знаменатель примет следующий вид:

$$a_2(x^k)\gamma^2 - 2\gamma a_1(x^k) + a_0(x^k) > 0. \tag{4.10}$$

Этот трехчлен должен быть строго больше нуля, чтобы сохранить смысл выражения.

Выясним, что нужно для того, чтобы (4.10) выполнялось. Для этого действительных корней не должно быть, а коэффициент при γ^2 должен быть больше нуля:

$$\begin{aligned} &d = (2a_1(x^k))^2 - a_2(x^k)a_0(x^k) < 0, \\ &a_2(x^k) > 0. \end{aligned} \tag{4.11}$$

Заметим, что последнее неравенство означает, что A^2 положительно определена.

Согласно необходимому условию экстремума, γ_k должно быть решением уравнения $\nabla_\gamma F(x^k, \gamma) = 0$:

$$\left(\frac{\gamma^2 a_3 - 2\gamma a_2 + a_1}{a_2\gamma^2 - 2\gamma a_1 + a_0}\right)' = 2\frac{\gamma^2(a_2^2 - a_3a_1) + \gamma(a_3a_0 - a_2a_1) + \overset{<0,(4.11)}{(a_1^2 - a_0a_2)}}{(a_2\gamma^2 - 2\gamma a_1 + a_0)^2} = 0 \tag{4.12}$$

$$\gamma_k = \frac{(a_2a_1 - a_3a_0) \pm \sqrt{(a_2a_1 - a_3a_0)^2 - 4(a_2^2 - a_3a_1)(a_1^2 - a_0a_2)}}{2(a_2^2 - a_3a_1)}. \tag{4.13}$$

В случае, когда $(a_2^2 - a_3a_1) = 0$, получим одно решение

$$\gamma = \frac{a_1^2 - a_0a_2}{a_3a_0 - a_2a_1}$$

Заметим, что первый множитель в (4.13) во втором слагаемом под корнем есть произведение детерминантов квадратных уравнений в (4.9). Итак, получены две стационарные точки. Теперь нужно выяснить какая из них даёт минимум. Для этого нужно определить каков знак второй производной в каждой из стационарных точек. Точка с положительной второй производной будет точкой минимума:

$$F'' = \frac{\gamma^3(-2a_2\widehat{a_2}) + \gamma^2(-3a_2\widehat{a_1}) + \gamma(2\widehat{a_2}a_0 - 4a_2\widehat{a_0} + 2a_1\widehat{a_1}) + \widehat{a_1}a_0 + 4a_1\widehat{a_0}}{(\gamma^2 a_2 - 2\gamma a_1 + a_0)^3}, \tag{4.14}$$

где $\widehat{a_2} = a_2^2 - a_3a_1$, $\widehat{a_1} = a_3a_0 - a_2a_1$, $\widehat{a_0} = a_1^2 - a_0a_2$ Можно указать промежутки положительности F'' используя, например, формулу Кардано (напомним, что знаменатель положителен, см. (4.10), поэтому рассматриваем только числитель):

- Сперва приведем числитель к каноническому виду $y^3 + py + q$:
  Сделаем замену $\gamma = y - \frac{b}{3a} = y - \frac{3a_2\widehat{a_1}}{6a_2\widehat{a_2}}$.
  Тогда

$$p = \frac{6a_2\widehat{a_2}(2\widehat{a_2}a_0 - 4a_2\widehat{a_0} + 2a_1\widehat{a_1}) - (3a_2\widehat{a_1})}{(6a_2\widehat{a_2})}$$

$$q = \frac{2(-3a_2\widehat{a_1})^3 - 9(-2a_2\widehat{a_2})(-3a_2\widehat{a_1})(2\widehat{a_2}a_0 - 4a_2\widehat{a_0} + 2a_1\widehat{a_1}) + 27(-2a_2\widehat{a_2})^2(\widehat{a_1}a_0 + 4a_1\widehat{a_0})}{27(-2a_2\widehat{a_2})^3}$$

- Вычислим $Q = \left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2$ и определим его знак:

  $Q > 0$: один вещественный корень и два сопряженных комплексных корня,

  $Q = 0$: один однократный вещественный корень и один двукратный, или, если p = q = 0, то один трёхкратный вещественный корень,

  $Q < 0$ три вещественных корня.

- Введем обозначения:

$$\alpha = \sqrt[3]{-\frac{q}{2} + \sqrt{Q}}$$

$$\beta = \sqrt[3]{-\frac{q}{2} - \sqrt{Q}}$$

Тогда корни уравнения выражаются следующим образом:

$$\begin{aligned} y_1 &= \alpha + \beta, \\ y_{2,3} &= -\frac{\alpha+\beta}{2} \pm i\frac{\alpha-\beta}{2}\sqrt{3}. \end{aligned} \tag{4.15}$$

Переходя к исходной переменной:

$$\begin{aligned} \gamma_1 &= \alpha + \beta - \frac{3a_2\widehat{a_1}}{6a_2\widehat{a_2}}, \\ \gamma_{2,3} &= -\frac{\alpha+\beta}{2} \pm i\frac{\alpha-\beta}{2}\sqrt{3} - \frac{3a_2\widehat{a_1}}{6a_2\widehat{a_2}}. \end{aligned} \tag{4.16}$$

Считая, что имеет место случай $Q < 0$, а так же считая, что корни упорядочены по возрастанию, промежутки выпуклости будут следующими: $(-\inf,\ \gamma_1)\ \cup (\gamma_2,\ \gamma_3)$

Итак, найдя стационарные точки, нужно будет проверить их на принадлежноть вышеуказанным интервалам.

Возможно, дабы не погрязть в комплексных корнях, проще будет просто подставить найденные стационарные точки в (4.14) и посмотреть знак выражения.

4.2.3. Сходимость

Покажем, что метод сходится со скоростью геометрической прогрессии со знаменателем

$$q = \sup_{0 \le \gamma \le \gamma_3} \|E + \gamma A\|, \tag{4.17}$$

Где γ_3 - максимальное значение γ из (4.16).

Напомним вид метода:

$$x^{k+1} = P_S(x^k - \gamma A x^k) = \frac{x^k - \gamma A x^k}{\|x^k - \gamma A x^k\|}, \tag{4.18}$$

S - единичная сфера,

$$\gamma_k = \underset{\gamma \ge 0}{argmin}\ \frac{1}{2}\left(\frac{x^k - \gamma A x^k}{\|x^k - \gamma A x^k\|}\right)^T A \frac{x^k - \gamma A x^k}{\|x^k - \gamma A x^k\|} \tag{4.19}$$

Пусть x^* - стационарная точка. Тогда для нее верно, что

$$(\nabla f(x^*), x - x^*) = (Ax^*, x - x^*)\ \forall x \in S. \tag{4.20}$$

Это условие эквивалентно следующему:

$$x^* = P_S(x^* - \gamma A x^*)\ \forall \gamma > 0. \tag{4.21}$$

Действительно, исходя из геометрического смысла: (4.20) эквивалентно тому, что

$S \cap Q = \{x \in \mathbb{R}^n : (\nabla f(x^*), x - x^*) < 0\} = \emptyset$. То есть множество направлений локального убывания не пересекается с множеством S. Допустим, что $x^* \in riS$ (относительная внутренность). В таком случае $\nabla f(x^*) = 0$ и $x^* = P_S(x^* - \gamma \nabla f(x^*)) = P_S(x^*) = x^*$. Пусть $x^* \in clS \setminus riS$. Предположим, что утверждение неверно. Тогда $\exists y \in S : y \ne x^*$ и $y = P_S(x^* - \gamma \nabla f(x^*)$. Тогда $(\nabla f(x^*), y - x^*) = (\nabla f(x^*), P_S(x^* - \gamma \nabla f(x^*) - x^*) \overset{(*)}{<} (\nabla f(x^*), x^* - \gamma \nabla f(x^*) - x^*) = -\gamma(\nabla f(x^*), \nabla f(x^*)) < 0$. Неравенство при использовании липшицевости $(*)$ строгое, в силу структуры множества Q. Это означает, что $y \in Q$, то есть $y \notin S$ - противоречие.

Итак, воспользуемся (4.21):

$$\|x^{k+1} - x^*\| = \|P_S(x^k - \gamma_k A x^k) - P_S(x^* - \gamma_k A x^k)\| \leq$$
$$\leq \|x^k - \gamma_k A x^k - x^* + \gamma_k A x^k\| = \|(E + \gamma_k A)(x^k - x^*)\| \leq$$
$$\leq \|x^k - x^*\| \cdot \|E + \gamma_k A\| \leq \|x^k - x^*\| \cdot q.$$

Таким образом $\|x^k - x^*\| \leq \|x^0 - x^*\| q^k$, $q = \sup\limits_{0 \leq \gamma \leq \gamma_3} \|E + \gamma A\|$.

5. Метод метод тяжелого шарика

Немного поясним смысл этих методов. В методах, описанных выше, никак не использовалась информация о том, что проделывал метод ранее. В следующих методах учитывается предыдущий шаг. Можно провести физическую аналогию с учётом информации с предыдущего шага: добавляется инерция, которая улучшает сходимость.

5.1. Метод тяжелого шарика

Рассмотрим метод тяжелого шарика для безусловной минимизации:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}), \quad \alpha > 0, \quad b \geq 0. \tag{5.1}$$

Слагаемое $\beta(x^k - x^{k-1})$ отражает ту самую "инерцию". Физически она проявляется в следующем: метод при больших β ("инерции") будет проскакивать незаметные, то есть неглубокие локальные (а может и не локальные) минимумы и идти дальше. Перейдем к вопросам о сходимости:

Если x^* - невырожденная точка минимума, то есть в ней выполнено достаточное условие точки минимума второго порядка (2.3), помимо этого выполнены следующие условия:

- $0 \leq \beta < 1$
- $0 < \alpha < 2\frac{1+\beta}{L}$
- $lE \leq \nabla^2 f(x^*) \leq LE$

тогда $\exists \varepsilon : \forall x^0, x^1 : \|x^0 - x^*\| < \varepsilon, \ \|x^1 - x^*\| < \varepsilon$ метод сходится к x^* со скоростью геометрической прогрессии. Здесь l и L - наименьшее и наибольшее собственные значения матрицы Гессе соответственно.

В случае, когда мы работаем на сфере, метод стоит переписать следующим образом:

$$x^{k+1} = P_S(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})), \quad \alpha > 0, \quad b \geq 0. \tag{5.2}$$

6. Метод Ньютона

$$\begin{gathered} x^{k+1} = \underset{x \in S}{argmin} f_k(x) \\ f_k(x) = f(x^k) + (\nabla f(x^k), x - x^k) + \frac{1}{2}(\nabla^2 f(x^k)(x - x^k), x - x^k) \end{gathered} \tag{6.1}$$

Метод сходится при следующих условиях:

- Функция $f(x)$ достигаем минимум на S в точке x^*
- В окрестности x^* функция $f(x)$ - дважды дифференцируема
- Матрица Гессе $\nabla^2 f(x)$ - удовлетворяет условиям Липшица и положительно определена.

Перепишем второе уравнение для квадратичной функции:

$$f_k(x) = \frac{1}{2} {x^k}^T A x^k + b^T x^k + c + (Ax + b, x - x^k) + \frac{1}{2} A(x - x^k), x - x^k) =$$
$$= \frac{1}{2} {x^k}^T A x^k + b^k + c + x^T A x - x^T A x^k + b^T x -$$
$$- b^T x^k + \frac{1}{2} x^T A x - \frac{1}{2} {x^k}^T A x - \frac{1}{2} x^T A x^k + \frac{1}{2} x^k A x^k =$$
$$= {x^k}^T A x^k - \frac{3}{2} x^T A x^k - \frac{1}{2} {x^k}^T A x + \frac{3}{2} x^T A x + b^T x + c.$$

Решить задачу (6.1) можно с помощью метода множителей Лагранжа. Запишем функцию Лагранжа для (6.1):

$$L = {x^k}^T Ax^k - \frac{3}{2}x^T Ax^k - \frac{1}{2}{x^k}^T Ax + \frac{3}{2}x^T Ax + b^T x + c + \lambda(\|x\|^2 - 1) \tag{6.2}$$

Условия оптимальности на функцию Лагранжа будут выглядеть так:

$$\begin{aligned} 3Ax &= 2Ax^k - b + 2\lambda x, \\ \|x\|^2 - 1 &= 0. \end{aligned}$$

Помимо этого стоит потребовать положительной определенности матрицы, задающей вторую производную по x функции Лагранжа:

$$\forall s \in \mathbb{S}\ s^T L_{xx} s > 0.$$

Отсюда можно найти требуемый x.

7. Квадратичная функция с линейной частью

7.1. Постановка задачи

Минимизируем квадратичную функцию с линейной частью на единичной сфере.

$$\begin{aligned} \min_{x \in S} f(x) &= \frac{1}{2}x^T Ax + b^T x, \\ S &= \{x \in \mathbb{R}^n : \|x\|^2 = 1\}. \end{aligned}$$

7.2. Стационарные точки

Выясним какие точки являются стационарные. Запишем функцию Лагранжа:

$$L = \frac{1}{2}x^T Ax + b^T x + \lambda(\|x\|^2 - 1). \tag{7.1}$$

Используем необходимое условие первого порядка:

$$\nabla L = \begin{pmatrix} Ax + b + 2\lambda x \\ \|x\|^2 - 1 \end{pmatrix} = \vec{0}. \tag{7.2}$$

Отсюда следуют, что стационарные точки удовлетворяют:

$$\begin{aligned} (A + 2\lambda E)x &= -b, \\ x &\in S. \end{aligned} \tag{7.3}$$

Теперь займемся вопросом о том, какая стационарная точка дает минимум. Для этого зафиксируем x^*, удовлетворяющий (7.3). Определим множество $Et = Et(x^*) = \{s \in \mathbb{R}^n : (s, 2x^*) = 0\}$. Тогда x^* - точка минимума, если

$$(L_{xx}s, s) = ((A + 2\lambda E)s, s) > 0\ \forall s \in Et. \tag{7.4}$$

Вспомним, что каждое решение СЛАУ представимо в следующем виде:

$$x = y_0 + y, \tag{7.5}$$

где y_0 - решение однородной системы, а y - какое-либо решение неоднородной.

Рассмотрим y_0 подробнее:

$$\begin{aligned} (A + 2\lambda E)y_0 &= 0 \\ Ay_0 &= -2\lambda y_0. \end{aligned} \tag{7.6}$$

Таким образом решение однородной системы - собственный вектор оператора A, соответствующий собственному значению -2λ. Без ограничения общности, будем считать, что это собственный вектор с номером n. Обратимся к (7.4). Поскольку A - симметричный, то существует *ортонормированный базис из*

собственных векторов A: $\{h_i\}_{i=1}^n$. Разложим s по собственным векторам: $s = \sum_{i=1}^n \alpha_i h_i$.

$$\begin{aligned}
\Big((A+2\lambda_n)s, s\Big) &= \left((A+2\lambda_n)\sum_{i=1}^n \alpha_i h_i, \sum_{i=1}^n \alpha_i h_i\right) = \\
&= \left(A\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) + 2\lambda_n\left(\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) + \alpha_n^2((A+\underbrace{\lambda_n E}_{=0})h_n, h_n) = \\
&= \left(\sum_{i=1}^{n-1} \alpha_i(-2\lambda_i) h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) + 2\lambda_n\left(\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) = \\
&= \left(\sum_{i=1}^{n-1} \alpha_i(-2\lambda_i) h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) + 2\lambda_n\left(\sum_{i=1}^{n-1} \alpha_i h_i, \sum_{i=1}^{n-1} \alpha_i h_i\right) = \\
&= -2\sum_{i=1}^{n-1} \lambda_i \alpha_i^2 (h_i, h_i) + 2\sum_{i=1}^{n-1} \lambda_n \alpha_i^2 (h_i, h_i) = \\
&= 2\sum_{i=1}^{n-1} \alpha_i^2 (h_i, h_i)(\lambda_n - \lambda_i) > 0.
\end{aligned} \tag{7.7}$$

Для того, чтобы неравенство было верным, $\lambda_n > \lambda_i \ \forall i = 1, \ldots, n-1$.

Таким образом минимумом является решение (7.3), где λ - наименьшее собственное значение A, помимо этого решение складывается из собственного вектора, отвечающего наименьшему собственному значению A и частного решения системы. И, безусловно, вектор должен быть единичным.

Рассуждения, приведенные выше, дают знание о том, из чего складывается решение, но на деле толку от этого мало. На практике же стоит сразу решать задачу численно, ибо (7.3). Аналитического решения не имеет.

7.3. Метод сопряженных градиентов

Заметим, что ограничение $(x, x) = 1$ является частным случаем ограничения

$$X^T X = I, \ X \in \mathbb{M}^{n \times p}, \tag{7.8}$$

которое наызвается *Stiefel Manifold*.

Метод, который далее будет описан, взят из статьи [1] и предназначен для работы с матрицами. Ниже будет рассмотрен случай $p = 1$, который отражает единичную сферу.

Посвятим немного внимания вопросу о геометрии (7.8), пробежавшись по основым вещам. Сначала обсудим как эти вещи описаны в статье, а потом "спроецируемся"на сферу.

7.3.1. Касательное и нормальное подпространства

Рассуждая интуитивно, касательным подпростанством к точке на многообразии будет касательная плоскость. Попробуем получить что-то более формальное для матриц. Продифференцируем соотношение из (7.8):

$$\begin{aligned}
(X^T X)' &= (I)' \\
X^T\Delta + \Delta^T X &= 0 \\
X^T\Delta &= -X\Delta^T.
\end{aligned} \tag{7.9}$$

То есть $X^T\Delta$ - skew-symmetric (антисимметрична). Антисимметричной матрицей A называется матрица A, удовлетворяющая соотношению $A^T = -A$. Приведем пример антисимметричной матрицы:

$$\begin{pmatrix} 0 & 2 & -1 \\ -2 & 0 & -4 \\ 1 & 4 & 0 \end{pmatrix}. \tag{7.10}$$

Последнее соотношение из (7.9), при размере матрицы X $n \times p$, дает $\frac{p(p-1)}{2}$ ограничений, из чего следует, что касательное подпространство имеет размерность $np - \frac{p(p-1)}{2}$.

На сфере касательным подпространсвом к точке будет плоскость, касательная к точке сферы. А размерностью этой плоскости будет $n - 1$.

Перейдем к вопросу о нормальном подпространстве. Чтобы ввести это понятие, нам понадобится скалярное произведение. Возьмем стандартное скалярное произведение:

$$\langle \Delta_1, \Delta_2 \rangle = tr\Delta_1^T \Delta_2. \tag{7.11}$$

Ясно видно, что в случае, когда $\Delta_{1,2}$ - векторы, это есть всеми любимая и знакомая сумма покомпонентных произведений. Итак, сохраняя тот же X, нормальным подпространством в точке X будут все матрицы N: $tr\Delta^T N = 0\ \forall\Delta$ из касательного подпространства в точке X. Очевидно, что любая матрица из нормального подпространства в точке X представима в виде $N = YS$, где S - $p\times p$ симметричная. Поскольку очевидные вещи легко доказываются, докажем, что это так:

$$\begin{aligned} \Delta \in Tg(X) &\iff X^T\Delta = -\Delta^T X \\ N \in Norm(X) &\iff tr(\Delta^T N) = 0\ \forall\Delta \in Tg(X) \\ tr(\underbrace{\{\Delta^T Y\}}_{=Y^T\Delta} S) &= 0. \end{aligned} \tag{7.12}$$

Поскольку $Y^T\Delta$ - антисимметрична, значит на диагонали у этой матрицы стоят нули, а S - симметрична, то последнее равенство верно.

Таким образом нормальное пространство это в точности $\{YS\}$, $S \in O_p$.

7.3.2. Точки на многообразии

Пусть $\mathbb{M}^{n\times p} \ni Q : Q^T Q = I$. Точкой на многообразии является класс эквивалентности

$$[Q] = \left\{ Q \begin{pmatrix} I_p & 0 \\ 0 & Q_{n-p} \end{pmatrix} : Q_{n-p} \in O_{n-p} \right\}, \tag{7.13}$$

где O_{n-p} - ортогональные квадратные матрицы размера $n-p$. По смыслу этот класс эквивалентности состоит из матриц, у которых первые p столбцов совпадают. Не стоит пугаться такого страшного крокодила, в случае $p = 1$ нас будут интересовать лишь точки на сфере.

Введем понятия горизонтального и вертикального подпространств. Вертикальным подпространством в точке $[Q]$ будем называть множество векторов, касательное к $[Q]$. Горизонтальным - множество касательных векторов к Q, ортогональных к вертикальному.

В точке Q вертикальное подпространство выражается следующим образом:

$$\Phi = Q \begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix},$$

где $C \in \mathbb{M}^{n-p\times n-p}$ - антисимметрична.

Горизонтальное подпространство в точке Q выражается так:

$$\Delta = Q \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix},$$

где $A \in \mathbb{M}^{p\times p}$ - антисимметрична.

7.3.3. Геодезические линии

Геодезические линии на (7.8) задаются следующим соотношением:

$$X(t) = Q \exp\left\{ t \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \right\} I_{n,p} \tag{7.14}$$

Сформулируем важную теорему, которая дает удобное выражение, дающее геодезические линии, которое выражено в терминах начальной позиции и направления движения.

Теорема 7.1. *Пусть* $X(t) = Q \exp\left\{ t \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \right\} I_{n,p}$, *с начальными условиями* $X(0) = X$ *и* $\dot{X}(0) = H$, *тогда*

$$X(t) = YM(t) + (I - XX^T)H \int_0^t M(t)dt, \tag{7.15}$$

где $M(t)$ *- решение следующего дифференциального уравнения:*

$$\ddot{M} - A\dot{M} + CM = 0 \quad M(0) = I_p,\ \dot{M}(0) = A, \tag{7.16}$$

$$M(t) = I_{n,p}^T \exp\left\{ t \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \right\} I_{n,p}.$$

Теперь сформулируем следствие, которым будет активно пользоваться в самом методе.

Следствие 7.1. *Пусть X и H - $n \times p$ матрицы, такие, что $X^T X = I_p$, $A = Y^T H$ - антисимметрична. Тогда геодезическая на Stiefel manifold* (7.8)*, исходящая из X в направлении H выражается кривой*

$$X(t) = YM(t) + QN(t), \tag{7.17}$$

где

$$QR := K = (I - XX^T)H \tag{7.18}$$

QR-разложение матрицы K, $Q \in \mathbb{M}^{n \times p}$, $R \in \mathbb{M}^{p \times p}$, $M(t)$ и $N(t)$ - $p \times p$ матрицы, заданные через матричную экспоненту

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp\left\{ t \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \right\} \begin{pmatrix} I_p \\ 0 \end{pmatrix}.$$

Итак, переведем вышесказанное на язык "$p = 1$".

Заметим, что QR разложение будет производиться не по матрице, а по вектору. Если раскладывается вектор $x : x \neq \vec{0}$, то $Q = \frac{x}{\|x\|}$, $R = \|x\|$. Обратим внимание, что выражения для $M(t), N(t)$ в нашем случае выглядит так:

$$\begin{pmatrix} M(t) \\ N(t) \end{pmatrix} = \exp\left\{ t \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \right\} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{7.19}$$

Легко видеть, что умножение на вектор $(1, 0)^T$ матрицы 2×2 даст вектор, который является первым столбцом матрицы, на которую он был умножен. Это поможет сократить вычисления.

7.3.4. Метод

Сразу переведем метод из статьи [1] на случай $p = 1$.

- Начальное приближение x_0: $(x_0, x_0) = 1$. Вычислим $G_0 = f_{x_0} - x_0 f_{x_0}^T x_0 = Ax_0 + b - x_0(Ax_0 + b)^T x_0$, $H_0 = -G_0$.
- Для $k = 0, 1, ...$
  - Минимизируем $f(x_k(t))$ по t, где

    $$x_k(t) = x_k M(t) + QN(t),$$

    QR - QR-разложение вектора $(I - x_k x_k^T)H_k$, $\widehat{A} = x_k^T H_k$, $M(t)$, $N(t)$ из (7.19).
  - $t_k = t_{min}$ $x_{k+1} = x_k(t_k)$.
  - Параллельный перенос вектора H_k в точку x_{k+1} вдоль геодезической:

    $$\tau H_k = H_k M(t_k) - x_k R^T N(t_k)$$
    $$\tau G_k = G_k$$

  - Вычисляем новое направление:

    $$H_{k+1} = -G_{k+1} + \gamma_k \tau H_k,$$
    $$\gamma_k = \frac{\langle G_{k+1} - \tau G_k, G_{k+1} \rangle}{\langle G_k, G_k \rangle},$$
    $$\langle \Delta_1, \Delta_2 \rangle = tr \Delta_1^T (I - \frac{1}{2} x x^T) \Delta_2$$

  - Если $k + 1 = 0 \; mod \; (n - 1)$, то $H_{k+1} = -G_{k+1}$.

Список литературы

[1] Alan Edelman, T.A. Arias, Steven T. Smith *LaTeX: THE GEOMETRY OF ALGORITHMS WITH ORTHOGONALITY CONSTRAINTS* 1998.