ИЗБРАННЫЕ ЛЕКЦИИ ПО ЭКСТРЕМАЛЬНЫМ ЗАДАЧАМ

Часть первая

Под редакцией проф. В. Н. Малозёмова

Санкт-Петербург
2017

УДК 519.85+517.988.38

И?? Избранные лекции по экстремальным задачам. Часть 1. Под ред. проф. В. Н. Малозёмова. — СПб.: Изд-во ВВМ, 2017. — 470 с.

ISBN 978-5-9651-1053-7

Основу данной книги составили общий и специальные курсы лекций по экстремальным задачам, которые читаются на математико-механическом факультете Санкт-Петербургского государственного университета для студентов отделения прикладной математики и информатики.

Книга состоит из двух частей. В первой части (главы 1–5) рассматриваются классические экстремальные задачи — линейные, квадратичные, нелинейные и вариационные. Вторая часть (главы 6, 7) посвящена негладким экстремальным задачам и чебышёвским приближениям.

Книга оформлена в виде отдельных лекций, которые можно читать практически независимо. Такой стиль поможет читателям, интересующимся конкретными вопросами, и студентам, готовящимся к экзаменам.

ISBN 978-5-9651-1053-7

Посвящается памяти
Владимира Фёдоровича Демьянова
(1938-2014)

Содержание

ГЛАВА 2. КВАДРАТИЧНЫЕ ЗАДАЧИ

ГЛАВА 4. ВАРИАЦИОННЫЕ ЗАДАЧИ

ГЛАВА 5. РАЗНОЕ

ПРЕДИСЛОВИЕ РЕДАКТОРА

Основу данной книги составили общий и специальные курсы лекций по экстремальным задачам, которые читались мной в разные годы (в течение пятидесяти лет) на математико-механическом факультете Санкт-Петербургского (Ленинградского) государственного университета. Общий курс «Экстремальные задачи» слушали студенты 3-го курса отделения прикладной математики и информатики, специальные курсы читались студентам старших курсов кафедры исследования операций. Чаще всего специальные курсы имели такие названия:

- «Численные методы нелинейного программирования»,
- «Чебышёвские приближения»,
- «Численные методы нелинейных чебышёвских приближений»,
- «Элементарные методы в экстремальных задачах».

К указанным источникам добавлены избранные доклады семинара по конструктивному негладкому анализу и недифференцируемой оптимизации («CNSA & NDO») и некоторые доклады семинара по дискретному гармоническому анализу и геометрическому моделированию («DHA & CAGD». В целом затронут широкий круг экстремальных задач как классического, так и неклассического типа.

Книга состоит из двух частей. Первая часть (главы 1–5) связана с общим курсом «Экстремальные задачи». Рассматриваются классические экстремальные задачи — линейные, квадратичные, нелинейные и вариационные. Вторая часть (главы 6, 7) посвящена негладким экстремальным задачам и чебышёвским приближениям.

Книга оформлена в виде отдельных лекций (докладов), которые можно читать практически независимо. Такой стиль поможет читателям, интересующимся конкретными вопросами, и студентам, готовящимся к экзаменам.

Список использованной литературы имеется в конце каждой лекции. Часто цитируемая книга

- Гавурин М. К., Малозёмов В. Н. Экстремальные задачи с линейными ограничениями. Л.: Изд-во ЛГУ, 1984. 176 с.

содержит основы теории экстремальных задач. В качестве общего источника информации по экстремальным задачам можно рекомендовать замечательный двухтомник:

- Васильев Ф. П. Методы оптимизации. Часть первая: Конечномерные задачи оптимизации. Принцип максимума. Динамическое программирование. М.: Изд-во МЦНМО, 2011. 620 с.

- Васильев Ф. П. Методы оптимизации. Часть вторая: Оптимизация в функциональных пространствах. Регуляризация. Аппроксимация. М.: Изд-во МЦНМО, 2011. 433 с.

Большую работу по подготовке данной книги к печати выполнил доц. Г. Ш. Тамасян. Приношу ему свою искреннюю благодарность. Благодарю также всех соавторов книги, которые помогли мне реализовать давно задуманную идею.

Январь 2017 г. *В. Н. Малозёмов*

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

- $\mathbb{Z}$, $\mathbb{R}$ — множества целых и вещественных чисел соответственно;
- M, N, $P, \ldots$ — конечные индексные множества;
- $k:j = \{k, k+1, \ldots, j\}$ — множество целых чисел от k до j включительно;
- $\mathbb{R}^N$ — линейное пространство векторов $x = x[N]$ с компонентами $x[j]$, $j \in N$; в случае $N = 1:n$ вместо $\mathbb{R}^N$ будем писать $\mathbb{R}^n$;
- $\mathbb{O} = \mathbb{O}[N]$ — вектор с компонентами $\mathbb{O}[j] = 0$, $j \in N$;
- $\langle x, y\rangle = x[N]\times y[N] = \sum_{j\in N} x[j]\times y[j]$ — скалярное произведение векторов x и y;
- $A = A[M, N]$ — матрица с элементами $A[k, j]$, $k \in M$, $j \in N$;
- $A[k, N]$, где $k \in M$, — k-я строка матрицы A;
- $A[M, j]$, где $j \in N$, — j-й столбец матрицы A;
- $A^T = A^T[N, M]$ — транспонированная матрица с элементами

$$A^T[j, k] = A[k, j], \quad j \in N, \ \ k \in M;$$

- $A[M_1, N_1]$, где $M_1 \subset M$, $N_1 \subset N$, — подматрица матрицы A;
- $y = Ax = A[M, N] \times x[N]$ — вектор с компонентами

$$y[k] = A[k, N] \times x[N], \quad k \in M;$$

- $v = uA = u[M] \times A[M, N]$ — вектор с компонентами

$$v[j] = u[M] \times A[M, j], \quad j \in N;$$

- $E = E[M, M]$ — единичная матрица, у которой $E[j, j] = 1$ при $j \in M$ и $E[j, k] = 0$ при $j \neq k$;
- $C = AB = A[M, N] \times B[N, P]$ — матрица с элементами

$$C[k, j] = A[k, N] \times B[N, j], \quad k \in M, \ \ j \in P;$$

- $:=$, $=:$ — равно по определению;
- $x[N] \geqslant y[N]$ означает, что $x[j] \geqslant y[j]$ при всех $j \in N$.

Отметим некоторые свойства подвекторов и подматриц.

1°. Пусть $N_1 \subset N$, $N_2 = N \setminus N_1$. Тогда

$$c[N] \times x[N] = c[N_1] \times x[N_1] + c[N_2] \times x[N_2].$$

Действительно,

$$c[N] \times x[N] = \sum_{j \in N} c[j] \times x[j] = \Big(\sum_{j \in N_1} + \sum_{j \in N_2} \Big) c[j] \times x[j] = \\ = c[N_1] \times x[N_1] + c[N_2] \times x[N_2].$$

2°. Пусть $N_1 \subset N$, $N_2 = N \setminus N_1$. Тогда

$$A[M, N] \times x[N] = A[M, N_1] \times x[N_1] + A[M, N_2] \times x[N_2]. \tag{1}$$

Аналогично, если $M_1 \subset M$, $M_2 = M \setminus M_1$, то

$$u[M] \times A[M, N] = u[M_1] \times A[M_1, N] + u[M_2] \times A[M_2, N]. \tag{2}$$

Проверим, например, равенство (1). При всех $k \in M$ имеем

$$A[k, N] \times x[N] = \sum_{j \in N} A[k, j] \times x[j] = \Big(\sum_{j \in N_1} + \sum_{j \in N_2} \Big) A[k, j] \times x[j] = \\ = A[k, N_1] \times x[N_1] + A[k, N_2] \times x[N_2].$$

Это равносильно (1).

Если учесть, что индексное множество есть объединение всех своих элементов, то в качестве следствия из (1) и (2) получаем важные формулы

$$Ax = \sum_{j \in N} A[M, j] \times x[j] = \sum_{j \in N} x[j] A[M, j],$$
$$uA = \sum_{k \in M} u[k] \times A[k, N].$$

Они означают, что вектор Ax равен линейной комбинации столбцов $A[M, j]$ матрицы A с коэффициентами $x[j]$, а вектор uA — линейной комбинации строк $A[k, N]$ матрицы A с коэффициентами $u[k]$.

3°. Справедливо равенство

$$A[M, N] \times B[N, P] = \sum_{j \in N} A[M, j] \times B[j, P]. \tag{3}$$

Оно проверяется непосредственным сравнением элементов с индексами (k, i) матриц, стоящих в левой и правой частях равенства (3).

При всей своей простоте формула (3) весьма содержательна. Она указывает на то, что произведение AB можно представить в виде суммы матриц, каждая из которых является произведением столбца матрицы A на соответствующую строку матрицы B.

4°. Пусть $N_1 \subset N$, $M_1 \subset M$. Тогда

$$E[N_1, N] \times x[N] = x[N_1], \tag{4}$$
$$u[M] \times E[M, M_1] = u[M_1]. \tag{5}$$

Таким образом, выделение подвектора — это линейная операция.

Проверим, например, равенство (4). Обозначим $N_2 = N \setminus N_1$. Согласно (1) имеем

$$\begin{aligned} E[N_1, N] \times x[N] &= E[N_1, N_1] \times x[N_1] + E[N_1, N_2] \times x[N_2] = \\ &= E[N_1, N_1] \times x[N_1] = x[N_1]. \end{aligned}$$

Аналогично, со ссылкой на формулу (2), доказывается равенство (5).

5°. Справедливо равенство

$$u[M] \times \big(A[M, N] \times x[N]\big) = \big(u[M] \times A[M, N]\big) \times x[N],$$

которое коротко можно переписать так:

$$\langle u, Ax \rangle = \langle uA, x \rangle. \tag{6}$$

Действительно,

$$\begin{aligned} \langle u, Ax \rangle &= \sum_{k \in M} u[k] \times \big(A[k, N] \times x[N]\big) = \\ = \sum_{k \in M} u[k] \times \Big(\sum_{j \in N} A[k, j] \times x[j]\Big) &= \sum_{k \in M} \sum_{j \in N} u[k] \times A[k, j] \times x[j], \\ \langle uA, x \rangle &= \sum_{j \in N} \big(u[M] \times A[M, j]\big) \times x[j] = \\ = \sum_{j \in N} \Big(\sum_{k \in M} u[k] \times A[k, j]\Big) \times x[j] &= \sum_{j \in N} \sum_{k \in M} u[k] \times A[k, j] \times x[j]. \end{aligned}$$

В правых частях двух последних равенств стоят повторные суммы, различающиеся лишь порядком суммирования. Они равны. Значит, равны и левые части.

Отметим, что векторы uA и $A^T u$ имеют одинаковые компоненты. Поэтому $\langle uA, x \rangle = \langle A^T u, x \rangle$. Равенство (6) можно переписать в виде

$$\langle u, Ax \rangle = \langle A^T u, x \rangle.$$

ГЛАВА 1. ЛИНЕЙНЫЕ ЗАДАЧИ

ДВОЙСТВЕННОСТЬ В ЛИНЕЙНОМ ПРОГРАММИРОВАНИИ

В. Н. Малозёмов

1°. Рассмотрим задачу линейного программирования

$$\begin{aligned} f(x) := c[N] \times x[N] &\to \inf, \\ A[M_1, N] \times x[N] &\geqslant b[M_1], \\ A[M_2, N] \times x[N] &= b[M_2], \\ x[N_1] &\geqslant \mathbb{O}[N_1], \end{aligned} \tag{1}$$

где $N_1 \subset N$. Вектор x, удовлетворяющий ограничениям задачи (1), называется *планом*. Множество планов обозначим Ω. Требуется найти *оптимальный план* — вектор $x^* \in \Omega$, на котором целевая функция $f(x)$ принимает наименьшее на Ω значение.

ТЕОРЕМА 1. *Оптимальный план существует тогда и только тогда, когда множество планов Ω непусто и целевая функция $f(x)$ ограничена снизу на Ω.*

2°. Обозначим $M = M_1 \cup M_2$, $N_2 = N \setminus N_1$ и запишем двойственную задачу линейного программирования

$$\begin{aligned} g(u) := b[M] \times u[M] &\to \sup, \\ u[M] \times A[M, N_1] &\leqslant c[N_1], \\ u[M] \times A[M, N_2] &= c[N_2], \\ u[M_1] &\geqslant \mathbb{O}[M_1]. \end{aligned} \tag{2}$$

Множество планов задачи (2) обозначим Λ.

ПЕРВАЯ ТЕОРЕМА ДВОЙСТВЕННОСТИ. *Из существования оптимального плана у одной из двойственных задач* (1), (2) *следует существование оптимального плана и у другой задачи. При этом справедливо соотношение двойственности*

$$\min_{x \in \Omega} f(x) = \max_{u \in \Lambda} g(u).$$

СЛЕДСТВИЕ. *Для того чтобы планы $x_0 \in \Omega$, $u_0 \in \Lambda$ двойственных задач были оптимальными, необходимо и достаточно, чтобы выполнялось равенство $f(x_0) = g(u_0)$.*

3°. Обычно исследуется пара двойственных задач линейного программирования и результаты формулируются одновременно для прямой и двойственной задач.

ТЕОРЕМА 2. *Для того чтобы обе задачи* (1) *и* (2) *имели оптимальные планы, необходимо и достаточно, чтобы множества их планов Ω и Λ были непусты.*

ВТОРАЯ ТЕОРЕМА ДВОЙСТВЕННОСТИ. *Планы x_0, u_0 двойственных задач* (1) *и* (2) *являются оптимальными тогда и только тогда, когда выполняются условия дополнительности*

$$\begin{aligned} &u_0[i] \times (A[i,N] \times x_0[N] - b[i]) = 0 \quad \forall i \in M_1, \\ &(c[j] - u[M] \times A[M,j]) \times x_0[j] = 0 \quad \forall j \in N_1. \end{aligned} \tag{3}$$

Условия дополнительности (3) можно переписать в любой из двух эквивалентных форм (они называются рабочими формами):

$$\begin{aligned} &u_0[M] \times A[M,j] = c[j], \text{ если } x_0[j] > 0,\ j \in N_1, \\ &u_0[i] = 0, \text{ если } A[i,N] \times x_0[N] > b[i],\ i \in M_1; \end{aligned}$$

или

$$\begin{aligned} &A[i,N] \times x_0[N] = b[i], \text{ если } u_0[i] > 0,\ i \in M_1, \\ &x_0[j] = 0, \text{ если } u_0[M] \times A[M,j] < c[j],\ j \in N_1. \end{aligned}$$

4°. Доказательства всех приведённых утверждений имеются в книге [1, с. 10–34].

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

МОДИФИЦИРОВАННЫЙ СИМПЛЕКС–МЕТОД*

В. Н. Малозёмов

Симплекс-метод решения задач линейного программирования является одним из выдающихся математических достижений 20-го столетия. В докладе излагается вариант симплекс-метода с обратной матрицей (модифицированный симплекс-метод) в том виде, в каком он читается мною в течение многих лет в курсе «Экстремальные задачи». Этим докладом я хотел бы обратить внимание читателей на замечательную, но уже забытую, книгу [1].

1°. Рассмотрим задачу линейного программирования в канонической форме

$$\begin{aligned} f(x) := c[N] \times x[N] &\to \inf, \\ A[M,N] \times x[N] &= b[M], \\ x[N] &\geqslant \mathbb{O}[N]. \end{aligned} \tag{1}$$

Вектор $x = x[N]$, удовлетворяющий ограничениям задачи (1), называется *планом*. Требуется найти план, доставляющий минимум целевой функции $f(x)$.

С планом x связан его *носитель*

$$N_+(x) = \big\{ j \in N \mid x[j] > 0 \big\}.$$

План x называется *базисным*, если столбцы $A[M,j]$ матрицы $A[M,N]$ при $j \in \in N_+(x)$ линейно независимы. Базисный план x называется *невырожденным*, если $|N_+(x)| = |M|$.

Невырожденному базисному плану x соответствует квадратная *базисная матрица* $A[M, N_+(x)]$ с линейно независимыми столбцами. Она обратима. Матрица

$$B[N_+(x), M] = \big(A[M, N_+(x)]\big)^{-1}$$

называется *обратной базисной матрицей*. По определению

$$\begin{aligned} B[N_+(x), M] \times A[M, N_+(x)] &= E[N_+(x), N_+(x)], \\ A[M, N_+(x) \times B[N_+(x), M] &= E[M, M]. \end{aligned}$$

*Семинар «DHA & CAGD». Избранные доклады. 20 ноября 2010 г.

Задачу (1) будем решать с помощью симплекс-метода, который позволяет от базисного плана x_0 перейти к «соседнему» базисному плану x_1 с меньшим значением целевой функции, $f(x_1) < f(x_0)$.

Предполагается, что выполнено *условие невырожденности*: все базисные планы задачи (1) невырождены. Роль этого условия существенна для доказательства конечной сходимости симплекс-метода.

2°. Перейдём к описанию метода. Возьмём начальный базисный план x_0 с носителем $N_+(x_0) =: N_+^{(0)}$ (вопрос о построении начального базисного плана рассмотрим позже). Проверим план x_0 на оптимальность.

Запишем двойственную задачу

$$\begin{gathered} b[M] \times u[M] \to \sup, \\ u[M] \times A[M,N] \leqslant c[N]. \end{gathered} \tag{2}$$

Условия дополнительности для плана x_0 имеют вид

$$u[M] \times A[M, N_+^{(0)}] = c[N_+^{(0)}].$$

Отсюда находим двойственный вектор

$$u_0[M] = c[N_+^{(0)}] \times B_0[N_+^{(0)}, M]. \tag{3}$$

Проверим, является ли вектор u_0 планом двойственной задачи. Для этого вычислим *оценки*

$$\Delta_0[j] = u_0[M] \times A[M,j] - c[j], \qquad j \in N \setminus N_+^{(0)}.$$

Если при всех $j \in N \setminus N_+^{(0)}$ выполняется неравенство $\Delta_0[j] \leqslant 0$, то u_0 — план двойственной задачи. По второй теореме двойственности x_0 — решение задачи (1) (и u_0 — решение двойственной задачи (2)).

3°. Предположим, что $\Delta_0[j_0] > 0$ при некотором $j_0 \in N \setminus N_+^{(0)}$. С учётом (3) перепишем это условие в виде

$$\Delta_0[j_0] = c[N_+^{(0)}] \times B_0[N_+^{(0)}, M] \times A[M, j_0] - c[j_0] > 0. \tag{4}$$

Обозначим

$$z_0[N_+^{(0)}] = B_0[N_+^{(0)}, M] \times A[M, j_0]. \tag{5}$$

Получим

$$\Delta_0[j_0] = c[N_+^{(0)}] \times z[N_+^{(0)}] - c[j_0] > 0. \tag{6}$$

Отметим, что в силу определения (5)

$$A[M, N_+^{(0)}] \times z_0[N_+^{(0)}] = A[M, j_0], \tag{7}$$

так что $z_0[N_+^{(0)}]$ есть вектор коэффициентов разложения столбца $A[M, j_0]$ по столбцам базисной матрицы $A[M, N_+^{(0)}]$. Доопределим вектор $z_0[N_+^{(0)}]$, положив

$$z_0[j] = \begin{cases} -1 & \text{при } j = j_0, \\ 0 & \text{при остальных } j \in N \setminus N_+^{(0)}. \end{cases}$$

Тогда соотношение (6) примет вид

$$\Delta_0[j_0] = c[N] \times z_0[N] > 0. \tag{8}$$

Соотношение (7) перепишется так:

$$A[M, N] \times z_0[N] = \mathbb{O}[M]. \tag{9}$$

4°. Введём луч

$$x(t) = x_0 - t\, z_0\,, \quad t > 0,$$

с направляющим вектором $-z_0$. Имеем

$$f\big(x(t)\big) = f(x_0) - t\, \Delta_0[j_0]. \tag{10}$$

Согласно (8) целевая функция при увеличении t убывает.

Отметим также, что согласно (9) при всех $t > 0$

$$A\, x(t) = b. \tag{11}$$

Предположим, что

$$z_0[N_+^{(0)}] \leqslant \mathbb{O}[N_+^{(0)}]. \tag{12}$$

Тогда и $z_0[N] \leqslant \mathbb{O}[N]$. Как следствие, $x(t) \geqslant \mathbb{O}$ при всех $t > 0$. Получаем, что вектор $x(t)$ является планом задачи (1) при всех $t > 0$. При этом в силу (10) $f\big(x(t)\big) \to -\infty$ при $t \to +\infty$.

Приходим к следующему заключению: при выполнении условия (12) задача (1) не имеет решения (целевая функция не ограничена снизу на множестве планов).

5°. Допустим, что найдётся индекс $s \in N_+^{(0)}$, на котором $z_0[s] > 0$. Обозначим через Γ_0 множество всех таких индексов:

$$\Gamma_0 = \big\{ s \in N_+^{(0)} \mid z_0[s] > 0 \big\}.$$

Очевидно, что $z_0[s] \leqslant 0$ при $s \in N_+^{(0)} \setminus \Gamma_0$.

При $t > 0$ имеем

$$x_0[s] - t\, z_0[s] > 0, \qquad s \in N_+^{(0)} \setminus \Gamma_0\,.$$

Кроме того,

$$x_0[j] - t\,z_0[j] = \begin{cases} 0 & \text{при } j \in \left(N \setminus N_+^{(0)}\right) \setminus \{j_0\}, \\ t & \text{при } j = j_0. \end{cases}$$

При $s \in \Gamma_0$ неравенство $x_0[s] - t\,z_0[s] \geqslant 0$ эквивалентно следующему

$$t \leqslant \frac{x_0[s]}{z_0[s]}\,.$$

Положим

$$t_0 = \min\left\{ \frac{x_0[s]}{z_0[s]} \;\middle|\; s \in \Gamma_0 \right\}.$$

Обозначим s_0 индекс, на котором достигается этот минимум (ниже будет показано, что при выполнении условия невырожденности такой индекс единствен). Очевидно, что $t_0 > 0$.

Введём вектор

$$x_1 = x_0 - t_0\,z_0\,. \tag{13}$$

В силу выбора t_0 имеем $x_1 \geqslant \mathbb{O}$, причём $x_1[s_0] = 0$. К этому нужно добавить, что согласно (11) $Ax_1 = b$. Значит, вектор x_1 является планом задачи (1).

Перепишем (13) в координатной форме:

$$\begin{aligned} &x_1[s] = x_0[s] - t_0\,z_0[s], \quad s \in N_+^{(0)} \setminus \{s_0\}; \\ &x_1[j_0] = t_0; \\ &x_1[j] = 0 \quad \text{при остальных } j \in N \text{ (включая } j = s_0). \end{aligned} \tag{14}$$

Покажем, что x_1 — *базисный* план*.

Обозначим $N_+^{(1)} = \left(N_+^{(0)} \setminus \{s_0\}\right) \cup \{j_0\}$. Согласно (14), $N_+(x_1) \subset N_+^{(1)}$. Базисность плана x_1 будет установлена, если выяснится, что столбцы матрицы $A[M, N_+^{(1)}]$ линейно независимы.

Запишем

$$\sum_{j \in N_+^{(0)} \setminus \{s_0\}} \alpha_j\,A[M, j] \;+\; \beta\,A[M, j_0] = \mathbb{O}[M] \tag{15}$$

и покажем, что это равенство возможно только тогда, когда все коэффициенты α_j, β равны нулю. Умножим обе части (15) слева на матрицу $B_0[N_+^{(0)}, M]$. С учётом (5) получим

$$\sum_{j \in N_+^{(0)} \setminus \{s_0\}} \alpha_j\,E[N_+^{(0)}, j] \;+\; \beta\,z_0[N_+^{(0)}] = \mathbb{O}[N_+^{(0)}]. \tag{16}$$

*Идея приводимого доказательства предложена И. В. Агафоновой.

В частности,

$$\sum_{j\in N_+^{(0)}\setminus\{s_0\}} \alpha_j\, E[s_0,j] \;+\; \beta\, z_0[s_0] = 0.$$

Сумма по $j \in N_+^{(0)} \setminus \{s_0\}$ равна нулю, а величина $z_0[s_0]$ по определению s_0 положительна. Значит, $\beta = 0$. Теперь из (16) следует, что и все коэффициенты α_j равны нулю. Линейная независимость столбцов матрицы $A[M, N_+^{(1)}]$, а с нею и базисность плана x_1, установлены.

В силу условия невырожденности $N_+(x_1) = N_+^{(1)}$. Действительно, нужно принять во внимание, что $N_+(x_1) \subset N_+^{(1)}$ и

$$\big|N_+(x_1)\big| = \big|M\big| = \big|N_+^{(0)}\big| = \big|N_+^{(1)}\big|.$$

Теперь понятно, почему минимум в определении t_0 достигается на единственном индексе. Иначе у плана x_1 появились бы лишние нулевые компоненты.

Отметим также, что согласно (10)

$$f(x_1) = f(x_0) - t_0\,\Delta_0[j_0].$$

Как следствие, $f(x_1) < f(x_0)$.

6°. Обратная базисная матрица

$$B_1[N_+^{(1)}, M] = \big(A[M, N_+^{(1)}]\big)^{-1}$$

порождена базисной матрицей $A[M, N_+^{(1)}]$, которая отличается от базисной матрицы $A[M, N_+^{(0)}]$ только одним столбцом. Естественно, что обратные базисные матрицы $B_1[N_+^{(1)}, M]$ и $B_0[N_+^{(0)}, M]$ должны быть связаны между собой. Чтобы разобраться в этом, потребуется некоторая подготовка.

Пусть $C = C[1:m, 1:m]$ — обратимая матрица со столбцами $C_1, \ldots, C_m$. Заменим в ней s-й столбец C_s столбцом P, разложение которого по базису $C_1, \ldots, C_m$ имеет вид

$$P = \sum_{j=1}^{m} z_j\, C_j\,.$$

Полученную матрицу обозначим D. Нетрудно понять, что

$$D = C\,U, \tag{17}$$

где матрица U отличается от единичной только s-м столбцом, который у U равен $(z_1, \ldots, z_m)^T$. Таким образом,

$$U = \begin{pmatrix} 1 & & z_1 & & \\ & \ddots & \vdots & & \\ & & z_s & & \\ & & \vdots & \ddots & \\ & & z_m & & 1 \end{pmatrix}.$$

ЛЕММА 1. *Если $z_s \neq 0$, то матрица D обратима и*

$$D^{-1} = V\,C^{-1}, \tag{18}$$

где матрица V отличается от единичной только s-м столбцом, который у V равен

$$\Big(-\frac{z_1}{z_s}, \ldots, -\frac{z_{s-1}}{z_s}, \frac{1}{z_s}, -\frac{z_{s+1}}{z_s}, \ldots, -\frac{z_m}{z_s}\Big)^T.$$

Доказательство. Достаточно проверить, что $U^{-1} = V$, то есть что

$$\begin{pmatrix} 1 & & z_1 & & \\ & \ddots & \vdots & & \\ & & z_s & & \\ & & \vdots & \ddots & \\ & & z_m & & 1 \end{pmatrix} \begin{pmatrix} 1 & & -z_1/z_s & & \\ & \ddots & \vdots & & \\ & & 1/z_s & & \\ & & \vdots & \ddots & \\ & & -z_m/z_s & & 1 \end{pmatrix} = E. \tag{19}$$

После этого справедливость леммы будет следовать из (17).

У произведения матриц UV из левой части (19) j-й столбец при $j \neq s$ равен единичному орту e_j. Запишем представление для s-го столбца:

$$\big(UV\big)_s = \sum_{j\neq s}\Big(-\frac{z_j}{z_s}\Big)\,e_j + \frac{1}{z_s}\sum_{j=1}^{m} z_j\,e_j\,.$$

Очевидно, что $(UV)_s = e_s$. Лемма доказана. □

Матрица V называется *мультипликатором*.

Распишем равенство (18) по строкам:

$$\begin{aligned} &D^{-1}[s,\cdot] = \frac{1}{z_s}\,C^{-1}[s,\cdot];\\ &D^{-1}[j,\cdot] = C^{-1}[j,\cdot] - z_j\,D^{-1}[s,\cdot] \quad \text{при } j \neq s. \end{aligned} \tag{20}$$

Строка $D^{-1}[s,\cdot]$ называется *рабочей строкой*.

Видим, что матрица D^{-1} легко пересчитывается по матрице C^{-1}.

7°. Обратимся к матрице $A[M, N_+^{(1)}]$. Она получается из обратимой базисной матрицы $A[M, N_+^{(0)}]$ заменой столбца с индексом s_0 на столбец с индексом j_0. При этом известно разложение вводимого столбца $A[M, j_0]$ по столбцам матрицы $A[M, N_+^{(0)}]$ (см. (7)). В указанном разложении коэффициент $z_0[s_0]$ у выводимого столбца по определению s_0 положителен. Мы находимся в условиях леммы, согласно которой матрица $A[M, N_+^{(1)}]$ обратима. Более того, на основании (20) для строк обратной матрицы $B_1[N_+^{(1)}, M]$ справедливы формулы пересчёта:

$$\begin{aligned} B_1[j_0, M] &= \frac{1}{z_0[s_0]}\, B_0[s_0, M] \quad \text{(рабочая строка)}; \\ B_1[j, M] &= B_0[j, M] - z_0[j]\, B_1[j_0, M] \quad \text{при } \ j \in N_+^{(0)} \setminus \{s_0\}. \end{aligned} \tag{21}$$

Строка $B_1[j_0, M]$ обратной матрицы соответствует столбцу $A[M, j_0]$, заменившему в матрице $A[M, N_+^{(0)}]$ столбец $A[M, s_0]$.

8°. Выведем формулу пересчёта для двойственного вектора. Аналогично (3) имеем

$$u_1[M] = c[N_+^{(1)}] \times B_1[N_+^{(1)}, M].$$

ЛЕММА 2. *Справедлива формула*

$$u_1[M] = u_0[M] - \Delta_0[j_0]\, B_1[j_0, M].$$

Доказательство. В силу (21)

$$\begin{aligned} u_1[M] \ &= \sum_{j \in N_+^{(0)} \setminus \{s_0\}} c[j]\, B_1[j, M] + c[j_0]\, B_1[j_0, M] = \\ &= \sum_{j \in N_+^{(0)} \setminus \{s_0\}} c[j]\, \big(B_0[j, M] - z_0[j]\, B_1[j_0, M]\big) + c[j_0]\, B_1[j_0, M] = \\ &= \sum_{j \in N_+^{(0)}} c[j]\, B_0[j, M] - c[s_0]\, B_0[s_0, M] - \\ &\qquad - \Bigg(\sum_{j \in N_+^{(0)} \setminus \{s_0\}} c[j]\, z_0[j] \ - \ c[j_0] \Bigg) B_1[j_0, M]. \end{aligned}$$

Воспользуемся равенством $B_0[s_0, M] = z_0[s_0]\, B_1[j_0, M]$ и формулой (6). Получим

$$\begin{aligned} u_1[M] &= u_0[M] - \Bigg(\sum_{j \in N_+^{(0)}} c[j]\, z_0[j] - c[j_0] \Bigg) B_1[j_0, M] = \\ &= u_0[M] - \Delta_0[j_0]\, B_1[j_0, M]. \end{aligned}$$

Лемма доказана. □

9°. Опишем общий шаг модифицированного симплекс-метода — переход от базисного плана x_k к базисному плану x_{k+1} с меньшим значением целевой функции. Считаем, что известны

$$x_k\,,\ N_+^{(k)},\ f(x_k),\ B_k[N_+^{(k)}, M],\ u_k[M]. \tag{22}$$

1) Последовательно вычисляем оценки

$$\Delta_k[j] = u_k[M] \times A[M, j] - c[j], \qquad j \in N \setminus N_+^{(k)}.$$

Если все $\Delta_k[j]$ неположительны, то x_k — оптимальный план. Для проверки правильности вычислений можно использовать соотношение двойственности

$$b[M] \times u_k[M] = f(x_k).$$

Процесс завершён. Иначе переходим к следующему пункту.

2) Берём индекс $j_k \in N \setminus N_+^{(k)}$, на котором $\Delta_k[j_k] > 0$. Вычисляем

$$z_k[N_+^{(k)}] = B_k[N_+^{(k)}, M] \times A[M, j_k].$$

Если $z_k[s] \leqslant 0$ при всех $s \in N_+^{(k)}$, то задача (1) не имеет решения (целевая функция не ограничена снизу на множестве планов). Процесс закончен. Иначе переходим к следующему пункту.

3) Вычисляем t_k по формуле

$$t_k = \min\left\{ \frac{x_k[s]}{z_k[s]} \;\middle|\; s \in N_+^{(k)},\ z_k[s] > 0 \right\}.$$

Обозначим s_k индекс, на котором достигается минимум.

4) Находим очередной базисный план x_{k+1}:

$$\begin{aligned} x_{k+1}[s] &= x_k[s] - t_k\, z_k[s], \quad s \in N_+^{(k)} \setminus \{s_k\}; \\ x_{k+1}[j_k] &= t_k\,, \\ x_{k+1}[j] &= 0 \quad \text{при остальных } j \in N. \end{aligned}$$

Обозначим $N_+^{(k+1)} = \big(N_+^{(k)} \setminus \{s_k\}\big) \cup \{j_k\}$.

5) Пересчитываем обратную базисную матрицу

$$\begin{aligned} B_{k+1}[j_k, M] &= \frac{1}{z_k[s_k]}\, B_k[s_k, M], \\ B_{k+1}[j, M] &= B_k[j, M] - z_k[j]\, B_{k+1}[j_k, M], \quad j \in N_+^{(k)} \setminus \{s_k\}. \end{aligned}$$

6) Пересчитываем значение целевой функции и двойственный вектор

$$f(x_{k+1}) = f(x_k) - t_k\,\Delta_k[j_k],$$
$$u_{k+1}[M] = u_k[M] - \Delta_k[j_k]\,B_{k+1}[j_k, M].$$

Проделав указанные действия, получим информацию вида (22):

$$x_{k+1}\,,\ N_+^{(k+1)},\ f(x_{k+1}),\ B_k[N_+^{(k+1)}, M],\ u_{k+1}[M].$$

Теперь можно переходить к очередной итерации.

Описанный метод сходится за конечное число шагов. Действительно, метод прекращает работу в двух случаях: либо когда очередной базисный план оптимален, либо когда выясняется, что задача не имеет решения. Иначе производится переход к следующему базисному плану с меньшим значением целевой функции. Конечность множества базисных планов и строгое уменьшение целевой функции гарантируют сходимость симплекс-метода за конечное число шагов.

10°. Остаётся разобраться с начальным базисным планом. Для его нахождения существует универсальный приём.

Будем считать, что в ограничениях задачи (1) все компоненты вектора b положительны. Рассмотрим вспомогательную задачу линейного программирования

$$\begin{gathered}\sum_{i\in M} y[i] \to \inf,\\ A[M,N]\times x[N] + E[M,M]\times y[M] = b[M],\\ x[N] \geqslant \mathbb{O}[N],\quad y[M] \geqslant \mathbb{O}[M].\end{gathered} \tag{23}$$

Множество её планов непусто (содержит $x = \mathbb{O}$, $y = b$) и целевая функция ограничена снизу (неотрицательна) на множестве планов. Значит, задача (23) имеет оптимальный базисный план. Его можно найти с помощью модифицированного симплекс-метода, взяв в качестве начального базисного плана $x = \mathbb{O}$, $y = b$. Базисной матрицей, равно как и обратной базисной матрицей, будет $E[M, M]$.

Значение целевой функции задачи (23) на оптимальном базисном плане будет либо положительным, либо равным нулю. Первый случай возможен только тогда, когда множество планов исходной задачи (1) пусто, то есть когда задача (1) не имеет решения. Во втором случае оптимальный базисный план имеет вид $(x_0, \mathbb{O})$, где x_0 — план задачи (1). Для носителя $N_+(x_0) =: N_+^{(0)}$ этого плана, вообще говоря, выполняется условие $|N_+^{(0)}| = |M|$. При этом по ходу реализации симплекс-метода найдена обратная базисная матрица $B_0[N_+^{(0)}, M]$.

План x_0 можно взять в качестве начального базисного плана для решения задачи (1). Вычислив

$$f(x_0) = c[N_+^{(0)}] \times x_0[N_+^{(0)}],$$
$$u_0[M] = c[N_+^{(0)}] \times B_0[N_+^{(0)}, M],$$

получим информацию вида (22) при $k = 0$. После этого можно приступить к решению задачи (1) с помощью модифицированного симплекс-метода.

11°. Конечная сходимость симплекс-метода доказана в предположении, что выполняется условие невырожденности. Однако симплекс-метод прекрасно работает и в вырожденном случае. Зацикливание не исключено, но оно возникает в редких случаях. Один из первых примеров такого рода описан в [2, с. 146–151]. Впрочем, симплекс-метод можно немного усовершенствовать, чтобы избежать зацикливания ([1, с. 71–81]). Усовершенствование состоит в том, что неоднозначный выбор индекса с положительной оценкой заменяется специальным образом организованным однозначным выбором.

ЛИТЕРАТУРА

1. Булавский В. А., Звягина Р. А., Яковлева М. А. *Численные методы линейного программирования.* М.: Наука, 1977. 368 с.

2. Гасс С. *Линейное программирование.* М.: Физматгиз, 1961. 303 с.

ВЫРОЖДЕННОСТЬ В ЗАДАЧАХ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ*

И. В. Агафонова, В А. Даугавет

Рассматривается задача линейного программирования в канонической форме:

$$f(x) := c[N] \times x[N] \to \min_{x \in \Omega}, \tag{1}$$

где

$$\Omega = \{x[N] \mid A[M,N] \times x[N] = b[M],\ x[N] \geqslant \mathbb{O}\},$$
$$M = \{1, 2, \ldots, m\}, \quad N = \{1, 2, \ldots, n\}.$$

Элементы векторов $c[N]$ и $b[M]$ и матрицы $A[M,N]$ — произвольные вещественные числа.

В докладе представлен конечный алгоритм для решения задачи (1), основанный на модифицированном симплекс-методе, называемом также симплекс-методом с обратной матрицей [1, 2]. Алгоритм не требует ни предварительных расчётов, ни наложения каких-либо дополнительных условий на A, b, c, в том числе условия невырожденности задачи, обеспечивающего конечную сходимость симплекс-метода (см. [1]). Конечную сходимость описываемого алгоритма гарантирует применение правила Блэнда для предотвращения зацикливания (см. Приложения А, В).

Алгоритм состоит из двух последовательных этапов, на которых симплекс-методом решаются приведённые ниже задачи линейного программирования: вспомогательная задача — на первом этапе (см. [3]) и некоторая задача, равносильная исходной, — на втором.

Первый этап алгоритма

Введём индексное множество $M' = \{n+1, n+2, \ldots, n+m\}$, новые переменные $x[i]$, $i \in M'$, называемые *искусственными*, и диагональную матрицу $\mathcal{E}[M, M']$ с компонентами

$$\mathcal{E}[i, i+n] = \begin{cases} \operatorname{sign} b[i], & \text{если } b[i] \neq 0, \\ 1, & \text{если } b[i] = 0. \end{cases}$$

*Семинар «DHA & CAGD». Избранные доклады. 11 декабря 2010 г.

Рассмотрим вспомогательную задачу линейного программирования

$$\begin{gathered}\mu(x) := \sum_{i \in M'} x[i] \to \min, \\ A[M, N] \times x[N] + \mathcal{E}[M, M'] \times x[M'] = b[M], \\ x[N] \geqslant \mathbb{O}, \quad x[M'] \geqslant \mathbb{O}.\end{gathered} \tag{2}$$

Обозначим $N' = N \cup M'$. В задаче (2) матрица ограничений имеет ранг m. Любой набор из m линейно независимых столбцов этой матрицы является базисом в пространстве $\mathbb{R}^m$ и может быть задан множеством индексов столбцов $\Gamma \subset N'$, $|\Gamma| = m$, которое, как и множество самих столбцов, будем кратко называть *базисом задачи* (2).

По каждому базису Γ строится базисное решение $x[N']$ системы ограничений-равенств задачи (2) — то единственное, в котором $x[N' \setminus \Gamma] = \mathbb{O}$. Если в этом решении оказывается $x[\Gamma] \geqslant \mathbb{O}$, то $x[N']$ — *базисный план* задачи (2).

Очевидным базисом задачи (2) является M', базисный план $x[N']$ имеет компоненты

$$x[i] = \begin{cases} 0, & i \in N, \\ \big|b[i-n]\big|, & i \in M'. \end{cases}$$

Обратная базисная матрица равна $\mathcal{E}$.

Так как целевая функция задачи ограничена снизу нулём, то задача (2) имеет решение.

Начиная с приведённого базисного плана, задачу (2) решают модифицированным симплекс-методом с применением правила Блэнда. В Приложении А помещено сжатое описание этого метода для задачи, записанной в виде (1), и, конечно, для (2) следует в этом описании под N понимать N' и очевидным образом изменить A и c.

При решении задачи (2) симплекс-методом искусственную переменную, как только она выйдет из базиса, рекомендуется сразу убирать из рассмотрения. Строго говоря, мы на первом этапе не решаем неизменную задачу (2) от начала и до конца, а переходим после каждого исключения искусственной переменной из базиса к решению новой, «укороченной» задачи, которую начинаем решать с имеющегося базисного плана.

В результате этого процесса будет получена оптимальная пара последней задачи такой цепочки:

- базис $\Gamma^0 = N^0 \cup M^0$, $N^0 \subset N$, $M^0 \subset M'$,
- план $x^0[N \cup M^0]$.

Оптимальное значение целевой функции равно $\mu^0 = \sum\limits_{i \in M^0} x^0[i]$.

Эта информация передаётся на следующий этап.

Второй этап алгоритма

Перед началом второго этапа возможны три ситуации.

1) $\mu^0 > 0$. Это означает, что множество планов Ω исходной задачи (1) пусто, то есть задача (1) не имеет решения.

2) $\mu^0 = 0$ и $M^0 = \varnothing$. Это означает, что в базисе не осталось искусственных переменных и вектор $x^0[N]$ является начальным базисным планом задачи (1) с базисом $N^0 \subseteq N$.

3) $\mu^0 = 0$ и $M^0 \neq \varnothing$. Это означает, что

$$x^0[M^0] = \mathbb{0}, \qquad x^0[N] \in \Omega,$$

но базисом является $\mathit{\Gamma}^0 = N^0 \cup M^0$.

В последнем случае сохраним искусственные базисные столбцы $\mathcal{E}[M, j]$, $j \in M^0$, но их знаки изменим на противоположные. Введём новое множество планов

$$\Omega_* = \Big\{ x[N \cup M^0] \;\Big|\; \begin{matrix} A[M,N] \times x[N] - \mathcal{E}[M, M^0] \times x[M^0] = b[M], \\ x[N] \geqslant \mathbb{0}, \qquad x[M^0] \geqslant \mathbb{0} \end{matrix} \Big\}. \tag{3}$$

Задача второго этапа

$$\varphi(x) := c[N] \times x[N] + \sum_{i \in M^0} x[i] \to \min_{x \in \Omega_*} \tag{4}$$

решается, как и задача (2), симплекс-методом с применением правила Блэнда. Для второго случая, когда $M^0 = \varnothing$, задача (4) совпадает с (1).

В Приложении С доказывается, что решение задачи (4) после отбрасывания равных нулю искусственных переменных даёт решение исходной задачи (1).

Приступая к решению задачи (4), мы уже имеем её базис $\mathit{\Gamma}^0$ и соответствующий ему базисный план x^0. Обратная базисная матрица получается из $B^0[\mathit{\Gamma}^0, M]$ заменой знаков строк её подматрицы $B^0[M^0, M]$ на противоположные.

Заметим, что если на итерации симплекс-метода будет выведена из базиса какая-либо искусственная переменная $x[i]$, $i \in M^0$, то её, как это делалось и на первом этапе, следует сразу исключить из рассмотрения.

Приложение А
Краткое описание одной итерации модифицированного симплекс-алгоритма с применением правила Блэнда

Решается задача (1). В начале итерации имеются:

- базис Γ;
- соответствующий этому базису базисный план $x[N]$;
- соответствующая этому базису обратная базисная матрица

$$B[\Gamma, M] = (A[M, \Gamma])^{-1}.$$

Проводятся действия:

1) **Проверка базисного плана x на оптимальность.** Найдём вектор двойственных переменных

$$u[M] = c[\Gamma] \times B[\Gamma, M]^1, \tag{5}$$

после чего вычислим *оценки* всех столбцов матрицы $A[M, N]$:

$$\Delta[N] = u[M] \times A[M, N] - c[N].$$

На самом деле $\Delta[\Gamma] = u[M] \times A[M, \Gamma] - c[\Gamma] = \mathbb{O}$, поэтому вычислять оценки $\Delta[j]$ нужно только при $j \in N \setminus \Gamma$.

Если $\Delta[N] \leqslant \mathbb{O}$, то вектор u — план задачи, двойственной к задаче (1), и

$$\begin{aligned} c[N] \times x[N] = c[\Gamma] \times x[\Gamma] = \big(u[M] \times A[M, \Gamma]\big) \times x[\Gamma] = \\ = u[M] \times \big(A[M, N] \times x[N]\big) = u[M] \times b[M]. \end{aligned}$$

Отсюда следует, что x — оптимальный план задачи (1) (и u — оптимальный план двойственной задачи). Алгоритм завершает свою работу.

Допустим, что $\Delta[j] > 0$ при некотором $j \in N \setminus \Gamma$. Обозначаем

$$J = \{j \in N \setminus \Gamma \mid \Delta[j] > 0\}.$$

Выберем *наибольшее* $j = j_0$ из J и перейдём к следующему пункту.

[1]Начиная со второй итерации, вектор u не вычисляется по этой формуле, а пересчитывается (см. [1]).

2) **Определение направления убывания целевой функции.** Вычисляем коэффициенты разложения столбца $A[M, j_0]$ по базису Γ:

$$z[\Gamma] = B[\Gamma, M] \times A[M, j_0].$$

Если $z[\Gamma] \leqslant \mathbb{O}$, то решения нет (целевая функция не ограничена снизу на множестве планов). Алгоритм завершает свою работу. Иначе переходим к следующему пункту.

3) **Определение длины шага.** Вычисляем

$$t = \min\left\{ \frac{x[p]}{z[p]} \;\middle|\; p \in \Gamma,\ \ z[p] > 0 \right\}.$$

Через S обозначаем множество индексов, на которых достигается минимум. Выбираем *наибольшее* $s_0 \in S$.

4) **Пересчёт плана:**

$$\begin{aligned} x[p] &:= x[p] - t\, z[p],\ p \in \Gamma \setminus \{s_0\}, \\ x[j_0] &:= t, \\ x[q] &:= 0 \text{ для остальных } q \in N. \end{aligned}$$

З а м е ч а н и е. (Будет использовано в Приложении В.)

> Значение целевой функции $f(x)$ на новом плане не больше, чем на прежнем: $f(x)$ станет меньше на величину $t\Delta[j_0] \geqslant 0$. Если $t = 0$, то из приведённых формул пересчёта плана следует, что *базисный план не изменится, хотя изменится базис*: равная нулю базисная компонента $x[s_0]$ заменяется компонентой $x[j_0]$, тоже нулевой.

5) **Смена базиса:** $\Gamma := \big(\Gamma \setminus \{s_0\}\big) \cup \{j_0\}$

Выбор *наибольших* индексов из J и S — это и есть правило Блэнда (см. Приложение В).

Приложение В
Зацикливание и правило Блэнда для его предотвращения

Симплекс-метод, решая задачу (1), строит последовательность базисов

$$\{\Gamma^{(k)}\},\ k = 1, 2, \ldots. \tag{6}$$

Каждый базис $\Gamma^{(k)}$ однозначно определяет базисный план $x^{(k)}$. При этом $f(x^{(k+1)}) \leqslant f(x^{(k)})$. Различных базисов у матрицы $A[M,N]$ конечное число. Если гарантировать, что ни один базис не войдёт в (6) больше одного раза, то симплекс-метод завершится за конечное число итераций[2].

Если в последовательности (6) есть совпадающие базисы $\Gamma^{(k)} = \Gamma^{(l)} = \Gamma$, то последовательность базисов между двумя вхождениями Γ будет циклически повторяться снова и снова. Это явление называется *зацикливанием*[3].

Существует несколько разных способов устранения зацикливания. Остановимся на правиле Блэнда [5]:

> При определении индекса j_0, подлежащего включению в базис[4], и индекса s_0, подлежащего исключению из базиса[5], следует брать максимальные индексы
>
> $$j_0 = \max_{j \in J} j,$$
> $$s_0 = \max_{s \in S} s.$$

Сразу заметим, что в этой формулировке можно заменить слово «максимальные» словом «минимальные» и тоже получить верное правило[6].

ТЕОРЕМА. *Применение правила Блэнда на итерациях симплекс-метода предотвращает зацикливание.*

Доказательство. Допустим, что в (6) базисы $\Gamma^{(k)}$ и $\Gamma^{(l)}$ совпали при некоторых $k < l$, и назовём *циклом* подпоследовательность

$$C = \{\Gamma^{(i)}\},\ i = k, k+1, \ldots, l.$$

На итерации симплекс-метода базис меняется, так что цикл C содержит не меньше двух различных базисов.

---

[2]Такую гарантию можно дать лишь тогда, когда задача (1) невырожденная и целевая функция строго убывает от итерации к итерации. Между тем большинство практических задач линейного программирования являются вырожденными [4].

[3]Видимо, следует уточнить, что повторение не обязательно будет бесконечным, если в переход от $\Gamma^{(k)}$ к $\Gamma^{(k+1)}$ включён элемент случайности, так что базис, следующий за Γ, не каждый раз один и тот же. Внесение такой случайности — один из способов борьбы с зацикливанием, который здесь не обсуждается.

[4]п. 1) Приложения А.

[5]п. 3) Приложения А.

[6]Минимальные индексы берут чаще. Здесь предпочтение отдано максимальным, чтобы из базиса быстрее уходили искусственные переменные (у них как раз большие индексы).

Отметим, что при всяком переходе от одного базиса цикла C к другому, согласно замечанию к п. 4) Приложения А, базисный план x не меняется, меняется лишь роль двух его нулевых компонент: одна из базисной становится небазисной, другая — наоборот.

Индексы, вошедшие в базисы $\Gamma^{(i)}$ из C, можно разделить на *неподвижные*, входящие в каждый базис цикла, и *подвижные* — все остальные. Каждый подвижный индекс должен хотя бы по одному разу пройти и процедуру исключения из некоторого базиса цикла, и процедуру включения в некоторый базис цикла. Точнее, если подвижный индекс принадлежит начальному базису цикла $\Gamma^{(k)}$, то он, как подвижный, когда-то выйдет из базиса, но снова вернётся, по крайней мере, в $\Gamma^{(l)}$. Если подвижный индекс не был в начальном базисе, то когда-то он войдёт в текущий базис и когда-то выйдет, поскольку в $\Gamma^{(l)}$ его нет. Множества неподвижных и подвижных индексов обозначим соответственно K и F.

Так как каждый индекс $j \in F$ в цикле хоть раз оказывался небазисным, то и соответствующие компоненты $x[j]$ базисного плана x, повторяющегося в цикле C, все побывали небазисными, то есть они равны нулю:

$$x[F] = \mathbb{O}. \tag{7}$$

Найдём минимальный индекс из подвижных:

$$r = \min_{j \in F} j.$$

После этого из всех базисов цикла C выберем такие Γ' и Γ'', что $r \in \Gamma'$, $r \notin \Gamma''$, причём при переходе от Γ' к следующему базису индекс r должен покинуть базис, замещаясь на некоторый подвижный индекс $j_0 \notin \Gamma'$ с положительной оценкой $\Delta'[j_0]$, а при переходе от Γ'' к следующему базису индекс r, имея положительную оценку $\Delta''[r]$, должен войти в базис вместо какого-то индекса, обозначение которого нам не понадобится.

Положим $N' = \Gamma' \cap F$. Множество N' состоит из подвижных индексов базиса Γ'. Очевидно, что $r \in N'$ и что

$$\Gamma' \setminus N' = K. \tag{8}$$

Обратимся к п. 1) описания симплекс-метода (Приложение А). По условию индекс r включается в базис, следующий за Γ'', поэтому $\Delta''[r] > 0$. Покажем, что

$$\Delta''[j] \leqslant 0 \quad \text{при всех} \quad j \in F \setminus \{r\}. \tag{9}$$

Действительно, если допустить, что $\Delta''[j] > 0$ при некотором $j \in F \setminus \{r\}$, то по правилу Блэнда $j < r$. Вместе с тем $r < j$, поскольку $j \in F \setminus \{r\}$. Приходим к противоречию.

Из (9), в частности, следует, что $\Delta''[j_0] \leqslant 0$ и

$$\Delta''[j] \leqslant 0 \quad \text{при} \quad j \in N' \setminus \{r\}. \tag{10}$$

Неподвижные индексы входят в базис Γ'', поэтому $\Delta''[K] = \mathbb{O}$. Согласно (8),

$$\Delta''[\Gamma' \setminus N'] = \mathbb{O}. \tag{11}$$

Теперь обратимся к п. 3) описания симплекс-алгоритма. По условию индекс $r \in \Gamma'$ выходит из базиса, поэтому $z[r] > 0$, где

$$z[\Gamma'] = B[\Gamma', M] \times A[M, j_0].$$

Покажем, что

$$z[j] \leqslant 0 \quad \text{при} \quad j \in N' \setminus \{r\}. \tag{12}$$

Действительно, если допустить, что $z[j] > 0$ при некотором $j \in N' \setminus \{r\}$, то по правилу Блэнда $j < r$ (учесть, что, согласно (7), $x[j] = x[r] = 0$). Вместе с тем $r < j$, поскольку $j \in F \setminus \{r\}$. Приходим к противоречию.

В силу положительности величин $\Delta''[r]$, $z[r]$ и неравенств (10), (12) получаем

$$\Delta''[N'] \times z[N'] > 0. \tag{13}$$

Установим противоположное неравенство. Этим завершится доказательство теоремы.

Напомним, что

$$u'[M] = c[\Gamma'] \times B[\Gamma', M].$$

Согласно (11), имеем

$$\begin{aligned}\Delta''[N'] \times z[N'] &= \Delta''[\Gamma'] \times z[\Gamma'] = \Delta''[\Gamma'] \times \big(B[\Gamma', M] \times A[M, j_0]\big) = \\ &= \big(u''[M] \times A[M, \Gamma'] - c[\Gamma']\big) \times B[\Gamma', M] \times A[M, j_0] + c[j_0] - c[j_0] = \\ &= u''[M] \times A[M, j_0] - c[j_0] - \big(u'[M] \times A[M, j_0] - c[j_0]\big) = \Delta''[j_0] - \Delta'[j_0].\end{aligned}$$

Как отмечалось выше, $\Delta''[j_0] \leqslant 0$, $\Delta'[j_0] > 0$, поэтому

$$\Delta''[N'] \times z[N'] < 0.$$

Получили противоречие с (13). Теорема доказана. □

Приложение C
Анализ задачи второго этапа

Рассмотрим множество (3), задающее ограничения задачи (4):

$$\Omega_* = \left\{ x[N \cup M^0] \;\middle|\; \begin{array}{c} A[M,N] \times x[N] - \mathcal{E}[M,M^0] \times x[M^0] = b[M], \\ x[N] \geqslant \mathbb{O}, \qquad x[M^0] \geqslant \mathbb{O} \end{array} \right\}.$$

ТЕОРЕМА. *Для любого вектора из Ω_* искусственная составляющая $x[M^0]$ равна нулю.*

Доказательство. Запишем задачу, двойственную задаче (2):

$$\begin{array}{c} b[M] \times u[M] \to \max, \\ u[M] \times A[M,N] \leqslant \mathbb{O}[N], \\ u[M] \times \mathcal{E}[M,M'] \leqslant \mathbb{1}[M'], \end{array} \tag{14}$$

где $\mathbb{1}$ — вектор из одних единиц.

Возьмём произвольный вектор $x[N \cup M^0] \in \Omega_*$. Умножим обе части равенства в определении Ω_* слева на $u^0[M]$ — оптимальный план задачи (14). Получим

$$u^0[M] \times A[M,N] \times x[N] - u^0[M] \times \mathcal{E}[M,M^0] \times x[M^0] = u^0[M] \times b[M]. \tag{15}$$

Отметим, что:

- Из ограничений (14) следует, что $u^0[M] \times A[M,N] \times x[N] \leqslant 0$.
- Первые n коэффициентов целевой функции задачи (2) равны нулю, а все следующие — единице. Согласно (5), имеем $u^0[M] = \sum\limits_{j \in M^0} B^0[j,M]$, где $B^0[\Gamma^0,M]$ — обратная базисная матрица. Поскольку все столбцы матрицы $\mathcal{E}[M,M^0]$ базисные, то

$$u^0[M] \times \mathcal{E}[M,M^0] = \sum_{j \in M^0} B^0[j,M] \times \mathcal{E}[M,M^0] = \mathbb{1}[M^0].$$

- Ко второму этапу алгоритм переходит только в том случае, когда оптимальное значение целевой функции задачи (2) равно нулю. По теореме двойственности правая часть (15) также будет равняться нулю.

С учётом отмеченного, на основании (15) получаем $\sum\limits_{i \in M^0} x[i] \leqslant 0$. Но $x[M^0] \geqslant \mathbb{O}$, значит $x[M^0] = \mathbb{O}$. □

Из доказанного утверждения следует, что искусственные составляющие всех планов задачи (4) нулевые, так что, решая её, мы получаем и решение задачи (1).

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Модифицированный симплекс-метод* // Семинар «DHA & CAGD». Избранные доклады. 20 ноября 2010 г. (`http://dha.spb.ru/reps10.shtml#1120`) [Данная книга, с. 15]

2. Карманов В. Г. *Математическое программирование.* 5-е изд. М.: ФИЗМАТЛИТ, 2004.

3. Булавский В. А. *Замечание о начале счёта в линейном программировании* / Сб. «Оптимальное планирование», № 15. Новосибирск, 1970. С. 76–78.

4. Dantzig G. B., Thapa M. N. *Linear Programming 2: Theory and Extensions.* Springer-Verlag, 2003.

5. Bland R. G. *New Finite Pivoting Methods for the Simplex Method* // Mathematics of Operations Research. 1977. Vol. 2. No. 2. P. 103–107.

ЕДИНСТВЕННОСТЬ РЕШЕНИЯ ЗАДАЧИ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ*

В. Н. Малозёмов

Данный доклад является естественным дополнением к докладу [1].

Рассмотрим задачу линейного программирования в канонической форме

$$\begin{aligned} c[N] \times x[N] &\to \inf, \\ A[M,N] \times x[N] &= b[M], \\ x[N] &\geqslant \mathbb{O}[N]. \end{aligned} \tag{1}$$

Пусть $x_0 = x_0[N]$ — базисный план (возможно, вырожденный) с носителем $N_+(x_0)$; $A[M,N_0]$, где $N_0 \supset N_+(x_0)$, $|N_0| = |M|$, — невырожденная базисная матрица, $B_0[N_0,M]$ — обратная базисная матрица. Такая ситуация возникает на каждом шаге модифицированного симплекс-метода. Вычислим двойственный вектор

$$u_0[M] = c[N_0] \times B_0[N_0,M] \tag{2}$$

и вектор оценок

$$\Delta_0[N \setminus N_0] = u_0[M] \times A[M, N \setminus N_0] - c[N \setminus N_0]. \tag{3}$$

Цель доклада — дать простое доказательство следующего утверждения.

ТЕОРЕМА. *Если* $\Delta_0[j] \leqslant 0$ *при всех* $j \in N \setminus N_0$, *то* x_0 — *оптимальный план. Если* $\Delta_0[j] < 0$ *при всех* $j \in N \setminus N_0$, *то* x_0 — *единственное решение задачи* (1).

Доказательство. Согласно (2), (3)

$$\Delta_0[N \setminus N_0] = c[N_0] \times B_0[N_0,M] \times A[M, N \setminus N_0] - c[N \setminus N_0].$$

Введём матрицу $Z_0 = Z_0[N, N \setminus N_0]$ по формулам

$$\begin{aligned} Z_0[N_0, N \setminus N_0] &= B_0[N_0,M] \times A[M, N \setminus N_0], \\ Z_0[N \setminus N_0, N \setminus N_0] &= -E[N \setminus N_0, N \setminus N_0]. \end{aligned} \tag{4}$$

*Семинар «DHA & CAGD». Избранные доклады. 17 декабря 2011 г.

В этом случае

$$\Delta_0[N \setminus N_0] = c[N] \times Z_0[N, N \setminus N_0]. \tag{5}$$

Действительно, в силу (4)

$$\begin{gathered} c[N] \times Z_0[N, N \setminus N_0] = c[N_0] \times Z_0[N_0, N \setminus N_0] + \\ + c[N \setminus N_0] \times Z_0[N \setminus N_0, N \setminus N_0] = c[N_0] \times B_0[N_0, M] \times A[M, N \setminus N_0] - \\ - c[N \setminus N_0] \times E[N \setminus N_0, N \setminus N_0] = \Delta_0[N \setminus N_0]. \end{gathered}$$

Далее, возьмём произвольный план $x = x[N]$ задачи (1) и покажем, что

$$x[N] = x_0[N] - Z_0[N, N \setminus N_0] \times x[N \setminus N_0]. \tag{6}$$

Будем проверять это равенство справа налево. Имеем

$$x_0[N \setminus N_0] - Z_0[N \setminus N_0, N \setminus N_0] \times x[N \setminus N_0] = E[N \setminus N_0, N \setminus N_0] \times x[N \setminus N_0] = x[N \setminus N_0].$$

Учитывая, что $A[M, N] \times x[N] = b[M]$ и $x_0[N_0] = B_0[N_0, M] \times b[M]$, получаем

$$\begin{gathered} x_0[N_0] - Z_0[N_0, N \setminus N_0] \times x[N \setminus N_0] = x_0[N_0] - B_0[N_0, M] \times \\ \times \big(A[M, N \setminus N_0] \times x[N \setminus N_0] + A[M, N_0] \times x[N_0] - A[M, N_0] \times x[N_0]\big) = \\ = x_0[N_0] - B_0[N_0, M] \times \big(A[M, N] \times x[N]\big) + B_0[N_0, M] \times A[M, N_0] \times x[N_0] = \\ = x_0[N_0] - B_0[N_0, M] \times b[M] + x[N_0] = x[N_0]. \end{gathered}$$

Формула (6) установлена.

Умножим (6) скалярно на $c[N]$. Согласно (5) придём к равенству

$$\begin{gathered} c[N] \times x[N] = c[N] \times x_0[N] - c[N] \times Z_0[N, N \setminus N_0] \times x[N \setminus N_0] = \\ = c[N] \times x_0[N] - \Delta_0[N \setminus N_0] \times x[N \setminus N_0]. \end{gathered} \tag{7}$$

По условию $x[N \setminus N_0] \geqslant \mathbb{O}[N \setminus N_0]$. Если $\Delta_0[j] \leqslant 0$ при всех $j \in N \setminus N_0$, то $\langle c, x\rangle \geqslant \langle c, x_0\rangle$ для всех планов x, то есть x_0 — оптимальный план.

Пусть $\Delta_0[j] < 0$ при всех $j \in N \setminus N_0$. Возьмём план x задачи (1), отличный от x_0. В силу (6) среди неотрицательных компонент вектора $x[N \setminus N_0]$ имеется хотя бы одна положительная. На основании (7) получаем $\langle c, x\rangle > \langle c, x_0\rangle$. Это значит, что x_0 — единственное решение задачи (1).

Теорема доказана. □

ЛИТЕРАТУРА

1. Агафонова И. В., Даугавет В. А. *Вырожденность в линейном программировании* // Семинар «DHA & CAGD». Избранные доклады. 11 декабря 2010 г. (`http://dha.spb.ru/reps10.shtml#1211`) [Данная книга, с. 25]

МУЛЬТИПЛИКАТИВНОЕ ПРЕДСТАВЛЕНИЕ ОБРАТНОЙ МАТРИЦЫ В МОДИФИЦИРОВАННОМ СИМПЛЕКС-МЕТОДЕ*

И. В. Романовский

Рассматривается задача линейного программирования в канонической форме:

$$f(x) = c[N] \times x[N] \to \min_{x \in \Omega}, \tag{1}$$

где

$$\Omega = \{x[N] \mid A[M,N] \times x[N] = b[M],\ x[N] \geqslant \mathbb{O}[N]\},$$
$$M = \{1, 2, \ldots, m\}, \quad N = \{1, 2, \ldots, n\}.$$

В докладе показано, как мультипликативное представление обратной матрицы влияет на алгоритмическую сторону модифицированного симплекс-метода. Обсуждаются также вопросы компьютерной реализации метода.

Введение

Как известно, в модифицированном симплекс-методе большое внимание уделяется способу представления обратной базисной матрицы $B[N', M]$ и среди таких представлений важное место занимает мультипликативное представление, в котором эта обратная матрица записывается в факторизованном виде — как произведение матриц-мультипликаторов[1]

$$B[N'_k, M] = D_k[N'_k, N'_{k-1}] \times D_{k-1}[N'_{k-1}, N'_{k-2}] \times \cdots \times D_1[N'_1, M],$$

а каждый мультипликатор отличается от единичной матрицы всего одним столбцом. Это представление было использовано ещё Г. Зойтендейком [1], а затем подробно описано Л. Лэсдоном [2], который сам ссылается на статью

*Семинар «DHA & CAGD». Избранные доклады. 5 февраля 2011 г.

[1]Мы выписываем здесь индексные множества мультипликаторов, так как с ними формула понятнее. Вычисления ведутся так, как будто все эти множества заменены множеством M.

Л. Ларсена [3]. Уже в 1962 появился краткий отчет Д. Смита и У. Орчард-Хейса об экспериментах с программной реализацией метода [4]. В моей книге 1977 г. [5] это представление, конечно, тоже излагается, хотя и вкратце[2].

Меня самого перейти на мультипликативное представление матрицы вынудили обстоятельства: я передал Н. Я. Краснеру в Воронежский университет свою программу модифицированного симплекс-метода, написанную на Алголе-60, и Краснер пожаловался, что программа медленно работает на задачах с двумя-тремя десятками ограничений. Тогда я переделал представление обратной матрицы на мультипликативное, и проблема снялась. С тех пор в основе наших программ лежит именно это представление.

Модуль обратной матрицы

Сейчас уже принято разрабатывать программы в объектно-ориентированном стиле и представлять все действия, связанные с обратной матрицей, как функции некоторого класса (по-старому говоря, модуля).

Таких действий (экспортируемых, т. е. вызываемых из программного окружения) совсем немного

Создание начальной матрицы. Имеется в виду обнуление списка мультипликаторов.

Изменение столбца. Замена одного из столбцов базисной матрицы требует добавления мультипликатора, изменяющего обратную матрицу. Этот мультипликатор добавляется в конец списка.

Умножение матрицы на вектор. Действие требуется для решения прямой системы $A[M, N'] \times x[N'] = b[M]$, выполняемого умножением вектора $b[M]$ слева на обратную матрицу $B[N', M]$.

Умножение вектора на матрицу. Действие требуется для решения двойственной системы $u[N'] \times A[M, N'] = c[N']$, выполняемого умножением вектора $c[N']$ справа на обратную матрицу $B[N', M]$.

Иногда набор операций над данными называют **кластером операций**.

В случае мультипликативного представления обратной матрицы обычно предусматривается ещё недоступное снаружи действие

Повторное обращение базисной матрицы. Это действие предназначено для сокращения информации о текущей базисной матрице.

[2]Популярность самого названия была так велика, что в стандартном учебнике для нашего экономического факультета им назван обычный метод обратной матрицы.

Умножение матрицы на вектор

Умножение вектора на обратную матрицу, составленную из r мультипликаторов, составляется из последовательных умножений на отдельные мультипликаторы $D_1, D_2, \ldots, D_r$. При этом удается выполнять «умножение на месте»: результат записывается на то место, где находился исходный вектор.

Мы можем сейчас считать, что все эти матрицы имеют одни и те же множества индексов и строк и столбцов $M = 1 : m$.

Напомним, что каждый мультипликатор D_k — это матрица, в которой один столбец, скажем, j_k — это произвольный столбец, обозначим его через $d_k[M]$, а остальные — столбцы единичной матрицы. Поэтому в произведении $D_k[M, M] \times x[M] = x'[M]$ получаем

$$\begin{aligned} x'[j_k] &:= d_k[j_k] \cdot x[j_k], \\ x'[j] &:= x'[j] + d_k[j] \cdot x[j_k], \quad j \neq j_k. \end{aligned}$$

Видно, что элементы вектора $x[M]$, которым соответствуют нулевые элементы вектора $d_k[M]$, не изменяются, а изменяемые элементы будут вычисляться правильно, если значение $x[j_k]$ появится при просмотре элементов столбца единообразно — в начале или в конце просмотра. Если оно появится в начале, то его нужно выписать отдельно для дальнейшего использования.

Возможна, таким образом, такая схема умножения на мультипликатор с сохранением результата на том же месте:

1. Прочесть пару $(d_k[j_k], j_k)$. Положить[3] $x_{\text{pivot}} := x[j_k]$. Положить $x[j_k] := := d_k[j_k] \cdot x[j_k]$.

2. Для каждой следующей пары $(d_k[j], j)$ положить $x[j] := x[j] + d_k[j] \cdot x_{\text{pivot}}$.

Здесь очень красиво это действие можно записать с помощью операции $+:=$ (читается «плюс-присвоить»; такая операция есть в ряде языков программирования и в наборе машинных команд).

Легко видеть, что если нужно выполнять умножение на последовательность матриц, то можно организовать «поточный» алгоритм»: будем записывать главный элемент в начале каждого мультипликатора, причём значение j_k будем записывать с минусом. Предположим, что у нас есть операция `GetNextPair(a,k)`, вырабатывающая логическое значение «следующая пара существует» и при благоприятном исходе записывающая в a и k вещественную и целочисленную компоненты пары. Алгоритм умножения будет выглядеть примерно так:

```
<записать в x[1:m] правые части системы>
<приготовиться к прямому просмотру мультипликаторов>
```

[3] Отмечу, что это не определение, а знак присваивания. Нужное нам значение помещается в безопасное место.

```
while GetNextPair(a,k) do
  if k < 0 then begin
    k := -k; xPivot := x[k]; x[k] := a
  end else
    x[k] +:= a*xPivot;
<записать из x[1:m] решение системы>
```

Умножение вектора на матрицу

Вычисление вектора $u[M] = c[N'] \times B[N', M]$, являющегося решением двойственной системы, также состоит из последовательных умножений на мультипликаторы, но эти умножения идут в обратном порядке — от D_k к D_1. В каждом таком умножении вычисление идёт по формулам

$$\begin{aligned} u'[k] &= \sum_i u[i] \cdot d[i], \\ u'[i] &= u[i] \quad \text{при } i \neq k. \end{aligned} \tag{$*$}$$

Стало быть, удобно, чтобы при просмотре элементов столбца $d_k[M]$ элемент d_k просматривался последним. Тогда можно просто вычислять сумму из $(*)$ и при появлении k-го элемента мультипликатора записать полученную сумму в нужную позицию вектора.

При описанном выше обратном просмотре последовательности пар обеспечивает нам правильный порядок и в перечислении мультипликаторов и при просмотре пар конкретного мультипликатора. Нужно только потребовать, чтобы такой обратный порядок просмотра был обеспечен. Будем считать, что у нас есть операция обратного просмотра `GetPrevPair(a,k)`, вырабатывающая логическое значение «предыдущая пара существует» и при благоприятном исходе записывающая в a и k вещественную и целочисленную компоненты пары. Алгоритм умножения будет выглядеть примерно так:

```
<записать в x[1:m] правые части системы>
<приготовиться к обратному просмотру мультипликаторов>
sum := 0;
while GetPrevPair(a,k) do
  if k < 0 then begin
    k := -k; x[k] := a*x[k] + sum;
    sum := 0
  end else
    sum +:= a*x[k];
<записать из x[1:m] решение системы>
```

Запись нового мультипликатора

Для записи нового мультипликатора нужно использовать операцию приписывания пары в конец набора. Назовём эту операцию `SavePair(a,r)`. После формирования массива d при известном индексе столбца r сначала записывается пара $(d[r], -r)$, а затем, в любом порядке, все ненулевые элементы столбца. Точнее, все элементы, абсолютная величина которых превосходит некоторую заранее выбранную величину `epsMultiplEntry`.

Повторное обращение базисной матрицы

При каждой итерации (модифицированного) симплекс-метода к последовательности мультипликаторов добавляется еще один, и умножения, трудоёмкость которых линейно зависит от длины последовательности, растёт. Считается целесообразным выполнять время от времени перевычисление (reinversion — переобращение) обратной матрицы, получая её из единичной по кратчайшему пути (длина его, если считать число мультипликаторов, равна числу неортов в текущем базисном множестве).

При обращении можно существенно сэкономить место за счет рационального выбора последовательности включения в формируемое заново базисное множество тех элементов, которые в нём должны быть. Я не буду рассказывать про это (очень интересное направление), отмечу, что у нас в лаборатории исследования операций этими вопросами издавна занимался С. С. Сурин (см., например, [6]).

Хранение информации о мультипликаторах

Набор мультипликаторов, как мы видели, представляет собой последовательность структур, каждая из которых составлена из двух элементов, и можно ожидать, что размер этой последовательности неудобно велик, но точно неизвестен. Он заслуживает серьёзного обсуждения.

Хранение информации о мультипликаторах должно удовлетворять следующим условиям.

- Информация состоит из последовательности пар (d, k), где поле d предназначено для хранения одного числа в формате с плавающей точкой (оно выравнивается на 8), а поле k предназначено для целочисленного индекса, и здесь достаточно короткого целого, выровненного на 2.
- Хранение должно обеспечивать эффективное выполнение следующих операций:
  - создание пустой последовательности или опустошение имеющейся,
  - добавление очередного элемента в конец последовательности,
  - просмотр последовательности от начала к концу,
  - просмотр последовательности от конца к началу.

Отметим некоторое неудобство хранения структуры, состоящей из пары полей неодинаковой длины, имеющих разную «степень выравнивания» — нам придется мириться с потерями либо памяти, либо эффективности.

Мы предлагаем следующее.

В современных компьютерах с огромными ресурсами оперативной памяти появилась конструкция `MemoryStream` — поток данных, размещаемый в оперативной памяти (среди языков, поддерживающих этот тип, назову `Delphi` и `C#`). Этот тип данных не требует особых действий по захвату памяти, она добавляется по мере надобности. Поток может иметь блочную структуру — состоять из элементов предписанного размера. Чтение потока может осуществляться от начала к концу и от конца к началу. Важно, что в отличие от массива изменение размера в потоке не требует от программы никаких специальных действий.

В соответствии со сказанным можно создать объект типа `MemoryStream`, фактически состоящий из блоков фиксированного размера. Например, блок может состоять из 1024 чисел формата с плавающей точкой, 1024 коротких целых чисел и возможно учетной информации — номера блока, его текущего заполнения и диапазона номеров пар. Чтение очередной пары будет состоять из возможной загрузки требуемого блока в пользовательский буфер, независимого чтения компонент пары и сдвига текущего индекса.

ЛИТЕРАТУРА

1. Zoutendijk G. *Methods of feasible directions*. Elsevier Publishing Co., 1960 (русский пер.: Г. Зойтендейк, Методы возможных направлений. М.: ИЛ, 1963).

2. Lasdon L. S. *Optimization theory for large systems*. MacMillan Co., 1970 (русский пер.: Л. Лэсдон, Оптимизация больших систем. М.: Наука, 1975).

3. Larsen L. J. *A modified inversion procedure for product form of the inverse linear programming codes* // Comm. of ACM. V. 5. 1962. P. 382–383.

4. Smith D. E., Orchard-Hays W. *Computational efficiency in product form LP codes* // In: R. L. Graves, Ph. Wolfe, eds. Recent Advances in Mathematical Programming. McGraw-Hill Book Co., 1963. P. 211–218.

5. Романовский И. В. *Алгоритмы решения экстремальных задач*. М.: Наука, 1977.

6. Брэгман Л. М., Прыгичев А. Н., Сурин С. С. *Повышение эффективности мультипликативного алгоритма метода последовательного улучшения плана* // В сб.: И. В. Романовский, ред. «Исследование операций и статистическое моделирование». Вып. 4. Изд-во ЛГУ, 1977. С. 3–49.

МАТРИЧНЫЕ ИГРЫ И ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ*

В. Н. Малозёмов

Аннотация. В докладе матричные игры анализируются с точки зрения линейного программирования. Приведены два нестандартных примера.

1°. Пусть задана квадратная или прямоугольная матрица $A = A[M, N]$ с вещественными элементами. Назовём её *матрицей платежей* (или *матрицей выигрышей*). Имеются два игрока, первый — *строчный* и второй — *столбцовый*.

Партия матричной игры заключается в следующем. Первый игрок произвольным образом выбирает индекс строки $i \in M$, второй игрок независимо выбирает индекс столбца $j \in N$. Число $A[i, j]$ есть величина выигрыша — такую сумму второй игрок выплачивает первому[1]. Матричная игра состоит из бесконечного числа таких партий.

В этой бесконечной серии каждый игрок должен выбрать свою *стратегию*. Стратегией первого игрока является вектор $p = p[M]$, компоненты которого удовлетворяют условиям

$$p[i] \geqslant 0 \quad \text{при всех} \quad i \in M; \quad \sum_{i \in M} p[i] = 1. \tag{1}$$

Здесь $p[i]$ — вероятность (частота) выбора i-й строки. Орты $e_i = e_i[M]$, $i \in M$, называются *чистыми стратегиями*. Остальные стратегии называются *смешанными*. Всё множество стратегий первого игрока обозначим $\mathcal{P}$.

Аналогично стратегией второго игрока является вектор $q = q[N]$, компоненты которого удовлетворяют условиям

$$q[j] \geqslant 0 \quad \text{при всех} \quad j \in N; \quad \sum_{j \in N} q[j] = 1. \tag{2}$$

*Семинар «CNSA & NDO». Избранные доклады. 14 апреля 2016 г.

[1]Строго говоря, величина выигрыша равна $\big|A[i, j]\big|$. Если $A[i, j] > 0$, то второй игрок платит первому сумму $A[i, j]$, если $A[i, j] < 0$, то первый игрок платит втором сумму $-A[i, j]$. При $A[i, j] = 0$ партия является ничейной. Во всех случаях первый игрок заинтересован в том, чтобы величина $A[i, j]$ была возможно большей, а второй игрок заинтересован в том, чтобы эта величина была возможно меньшей.

Орты $\hat{e}_j = \hat{e}_j[N]$, $j \in N$, называются *чистыми стратегиями*. Остальные стратегии называются *смешанными*. Всё множество стратегий второго игрока обозначим $\mathcal{Q}$.

В зависимости от стратегий игроков определяется средняя величина выигрыша в каждой партии:

$$a(p, q) = p[M] \times A[M, N] \times q[N]. \tag{3}$$

Допустим, что первый игрок выбрал стратегию p. Тогда его гарантированный выигрыш равен величине

$$\varphi(p) = \min_{q \in \mathcal{Q}} a(p, q). \tag{4}$$

Оптимальной естественно назвать ту стратегию p_*, на которой

$$\varphi(p_*) = \max_{p \in \mathcal{P}} \varphi(p). \tag{5}$$

Аналогично, если второй игрок выбрал стратегию q, то максимально возможный его «проигрыш» равен величине

$$\psi(q) = \max_{p \in \mathcal{P}} a(p, q). \tag{6}$$

Оптимальной естественно назвать ту стратегию q_*, на которой

$$\psi(q_*) = \min_{q \in \mathcal{Q}} \psi(q). \tag{7}$$

Справедливо следующее утверждение.

ТЕОРЕМА 1 (фон Нейман)**.** *Оптимальные стратегии игроков существуют. Для того чтобы пара стратегий p_*, q_* была оптимальной необходимо и достаточно, чтобы выполнялось равенство*

$$\varphi(p_*) = \psi(q_*). \tag{8}$$

2°. Для доказательства теоремы фон Неймана потребуется некоторая подготовка.

ЛЕММА. *Справедливы формулы*

$$\min_{q \in \mathcal{Q}} \langle c, q \rangle = \min_{j \in N} c[j], \tag{9}$$

$$\max_{p \in \mathcal{P}} \langle d, p \rangle = \max_{i \in M} d[i]. \tag{10}$$

Доказательство. Проверим, например, первое равенство. Обозначим

$$\alpha = \min_{j\in N} c[j], \quad J = \big\{j \in N \mid c[j] = \alpha\big\}.$$

При всех $q \in \mathcal{Q}$ имеем

$$\langle c, q\rangle - \alpha = \sum_{j\in N}\big(c[j] - \alpha\big) \times q[j] = \sum_{j\in N\setminus J}\big(c[j] - \alpha\big) \times q[j] \geqslant 0,$$

то есть $\langle c, q\rangle \geqslant \alpha$. Равенство достигается, когда $q[j] = 0$ при всех $j \in N \setminus J$. Формула (9) установлена.

Аналогично проверяется формула (10). □

На основании соотношений (4), (3) и (9) функция $\varphi(p)$ допускает представление

$$\begin{aligned}\varphi(p) &= \min_{q\in\mathcal{Q}}\big(p[M] \times A[M, N]\big) \times q[N] = \\ &= \min_{j\in N} p[M] \times A[M, j].\end{aligned} \tag{11}$$

На основании соотношений (6), (3) и (10) функция $\psi(q)$ допускает представление

$$\begin{aligned}\psi(q) &= \max_{p\in\mathcal{P}} p[M] \times \big(A[M, N] \times q[N]\big) = \\ &= \max_{i\in M} A[i, N] \times q[N].\end{aligned} \tag{12}$$

3°. Согласно (5) и (11), (7) и (12) оптимальные стратегии первого и второго игроков определяются как решения следующих экстремальных задач:

$$\varphi(p) := \min_{j\in N} p[M] \times A[M, j] \to \max_{p\in\mathcal{P}}, \tag{13}$$

$$\psi(q) := \max_{i\in M} A[i, N] \times q[N] \to \min_{q\in\mathcal{Q}}. \tag{14}$$

Задача (13) эквивалентна задаче линейного программирования (см. [1, с. 11–13])

$$\begin{aligned} s &\to \max, \\ -p[M] \times A[M, j] + s &\leqslant 0, \quad j \in N; \\ \sum_{i\in M} p[i] &= 1; \\ p[i] &\geqslant 0, \quad i \in M. \end{aligned} \tag{15}$$

При доказательстве эквивалентности плану p задачи (13) сопоставляется план (p, s) задачи (15), где

$$s = \min_{j\in N} p[M] \times A[M, j] = \varphi(p). \tag{16}$$

Далее, задача (14) также эквивалентна задаче линейного программирования

$$\begin{gathered} t \to \min, \\ -A[i,N] \times q[N] + t \geqslant 0, \quad i \in M; \\ \sum_{j \in N} q[j] \qquad = 1; \\ q[j] \geqslant 0, \quad j \in N. \end{gathered} \tag{17}$$

При доказательстве эквивалентности плану q задачи (14) сопоставляется план (q,t) задачи (17), где

$$t = \max_{i \in M} A[i,N] \times q[N] = \psi(q). \tag{18}$$

Матрица ограничений задачи (17) имеет вид

$$\begin{bmatrix} & & & 1 \\ & -A[M,N] & & \vdots \\ & & & 1 \\ 1 & \ldots\ldots & 1 & 0 \end{bmatrix}$$

Принципиальный факт заключается в том, что задачи линейного программирования (17) и (15) — двойственные! Это проверяется непосредственно.

Множества планов задач (17) и (15) непусты (по любому $q \in \mathcal{Q}$ легко подбирается подходящее t и по любому $p \in \mathcal{P}$ легко подбирается подходящее s). Значит, обе задачи имеют оптимальные планы (q_*, t_*) и (p_*, s_*). По эквивалентности, q_* и p_* — оптимальные планы задач (14) и (13) соответственно. Тем самым, установлено существование оптимальных стратегий у обоих игроков. Чтобы найти эти стратегии, нужно решить пару двойственных задач линейного программирования.

Покажем, что критерием оптимальности является равенство (8). Если p_*, q_* — оптимальные стратегии, то по эквивалентности (p_*, s_*) и (q_*, t_*) — оптимальные планы двойственных задач (15) и (17). Здесь, согласно (16) и (18), $s_* = \varphi(p_*)$, $t_* = \psi(q_*)$. По первой теореме двойственности $s_* = t_*$, что равносильно (8).

Наоборот, пусть для некоторых стратегий $p_* \in \mathcal{P}$, $q_* \in \mathcal{Q}$ выполнено равенство (8). Для планов (p_*, s_*), (q_*, t_*) задач (15), (17) при $s_* = \varphi(p_*)$, $t_* = \psi(q_*)$ имеет место равенство $s_* = t_*$. Значит, эти планы — оптимальные. По эквивалентности p_* и q_* — оптимальные стратегии первого и второго игроков.

Теорема доказана. □

Замечание 1. В силу (11) и (12) критерий оптимальности (8) можно переписать в виде

$$\min_{j \in N} p_*[M] \times A[M,j] = \max_{i \in M} A[i,N] \times q_*[N]. \tag{19}$$

Замечание 2. Критерий оптимальности (8) допускает ещё одну эквивалентную формулировку (в теории матричных игр она считается основной): для того чтобы пара стратегий p_*, q_* была оптимальной, необходимо и достаточно, чтобы при всех $p \in \mathcal{P}$ и всех $q \in \mathcal{Q}$ выполнялись неравенства

$$a(p, q_*) \leqslant a(p_*, q_*) \leqslant a(p_*, q). \tag{20}$$

Докажем это утверждение. Если p_*, q_* — оптимальные стратегии, то согласно (8) имеем

$$a(p_*, q_*) \leqslant \max_{p \in \mathcal{P}} a(p, q_*) = \psi(q_*) = \varphi(p_*) = \\ = \min_{q \in \mathcal{Q}} a(p_*, q) \leqslant a(p_*, q_*).$$

Отсюда следует, что

$$\max_{p \in \mathcal{P}} a(p, q_*) = a(p_*, q_*) = \min_{q \in \mathcal{Q}} a(p_*, q). \tag{21}$$

Это равносильно неравенствам (20).

Наоборот, пусть выполнены неравенства (20). Тогда справедливы соотношения (21). Их можно переписать в виде

$$\psi(q_*) = a(p_*, q_*) = \varphi(p_*).$$

В частности, $\varphi(p_*) = \psi(q_*)$. Утверждение доказано.

Величина $a(p_*, q_*)$ называется *ценой игры*.

Неравенства (20) характеризуют пару оптимальных стратегий p_*, q_* как *ситуацию равновесия*.

Замечание 3. Согласно (5) и (4)

$$\varphi(p_*) = \max_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} a(p, q).$$

Согласно (7) и (6)

$$\psi(q_*) = \min_{q \in \mathcal{Q}} \max_{p \in \mathcal{P}} a(p, q).$$

Пусть $A = A[M, N]$ — произвольная матрица. По теореме 1 существуют векторы p_* и q_*, такие что $\varphi(p_*) = \psi(q_*)$. Это значит, что для любой матрицы A выполняется равенство

$$\max_{p \in \mathcal{P}} \min_{q \in \mathcal{Q}} a(p, q) = \min_{q \in \mathcal{Q}} \max_{p \in \mathcal{P}} a(p, q).$$

В этой связи теорему фон Неймана часто называют *теоремой о минимаксе*.

4°. Выясним, когда ситуацию равновесия образует пара чистых стратегий.

ТЕОРЕМА 2. *Для того чтобы матричная игра с матрицей выигрышей $A = A[M, N]$ имела ситуацию равновесия в чистых стратегиях, необходимо и достаточно, чтобы выполнялось равенство*

$$\max_{i\in M}\min_{j\in N} A[i,j] = \min_{j\in N}\max_{i\in M} A[i,j]. \tag{22}$$

Доказательство. Необходимость. Пусть $p_* = e_{i_0}[M]$ и $q_* = \hat{e}_{j_0}[N]$ — оптимальные чистые стратегии. Согласно критерию оптимальности в форме (19) имеем

$$\min_{j\in N} A[i_0, j] = \max_{i\in M} A[i, j_0]. \tag{23}$$

С учётом равенства (23) при всех $i \in M$ получаем

$$\min_{j\in N} A[i,j] \leqslant A[i,j_0] \leqslant \max_{i'\in M} A[i',j_0] = \min_{j\in N} A[i_0,j].$$

Это значит, что

$$\max_{i\in M}\min_{j\in N} A[i,j] = \min_{j\in N} A[i_0,j]. \tag{24}$$

Аналогично при всех $j \in N$

$$\max_{i\in M} A[i,j] \geqslant A[i_0,j] \geqslant \min_{j'\in N} A[i_0,j'] = \max_{i\in M} A[i,j_0].$$

Это значит, что

$$\min_{j\in N}\max_{i\in M} A[i,j] = \max_{i\in M} A[i,j_0]. \tag{25}$$

На основании (24), (25) и (23) приходим к (22).

Достаточность. Обозначим через i_0 и j_0 внешние индексы, на которых достигается максимум и минимум в равенстве (22). Тогда

$$\min_{j\in N} A[i_0,j] = \max_{i\in M} A[i,j_0].$$

Перепишем это равенство в виде

$$\min_{j\in N} e_{i_0}[M] \times A[M,j] = \max_{i\in M} A[i,N] \times \hat{e}_{j_0}[N].$$

По критерию оптимальности в форме (19) чистые стратегии $p_* = e_{i_0}[M]$ и $q_* = \hat{e}_{j_0}[N]$ образуют ситуацию равновесия.

Теорема доказана. □

5°. Обратимся к примерам.

ПРИМЕР 1. В качестве матрицы выигрышей рассмотрим квадратную матрицу второго порядка с параметром c:

$$A = \begin{pmatrix} 1 & -1 \\ 0 & c \end{pmatrix}.$$

Построим график цены игры как функции параметра c.

Параметр c имеет два критических значения $c = -1$ и $c = 0$. Возможны три случая.

1) $c \leqslant -1$. Имеем

$$\max_i \min_j A[i,j] = \max\{-1, c\} = -1,$$
$$\min_j \max_i A[i,j] = \min\{1, -1\} = -1.$$

Если цену игры обозначить через $f(c)$, то по теореме 2 получаем

$$f(c) = -1 \quad \text{при} \quad c \leqslant -1. \tag{26}$$

2) $c \in (-1, 0]$. Имеем

$$\max_i \min_j A[i,j] = \max\{-1, c\} = c,$$
$$\min_j \max_i A[i,j] = \min\{1, c\} = c.$$

По теореме 2

$$f(c) = c \quad \text{при} \quad c \in (-1, 0]. \tag{27}$$

3) $c > 0$. Имеем

$$\max_i \min_j A[i,j] = \max\{-1, 0\} = 0,$$
$$\min_j \max_i A[i,j] = \min\{1, c\} > 0.$$

Равенство (22) нарушается. Будем искать решение в смешанных стратегиях.

Запишем задачи линейного программирования для второго и первого игроков:

$$\begin{array}{rr} t \to \min & \\ -q_1 + q_2 + t \geqslant 0 & \\ -c\,q_2 + t \geqslant 0 & \\ q_1 + q_2 = 1 & \\ q_1 \geqslant 0,\ q_2 \geqslant 0, & \end{array} \qquad \begin{array}{rr} s \to \max & \\ -p_1 + s \leqslant 0 & \\ p_1 - c\,p_2 + s \leqslant 0 & \\ p_1 + p_2 = 1 & \\ p_1 \geqslant 0,\ p_2 \geqslant 0. & \end{array} \tag{28}$$

Поскольку мы ищем решение в смешанных стратегиях, то можно воспользоваться условиями дополнительности

$$\begin{aligned} -q_1 + q_2 + t &= 0, & -p_1 + s &= 0,\\ -c\,q_2 + t &= 0, & p_1 - c\,p_2 + s &= 0. \end{aligned}$$

К этому следует добавить равенства

$$q_1 + q_2 = 1, \qquad p_1 + p_2 = 1.$$

Полученные системы линейных уравнений третьего порядка имеют единственные решения

$$q_* = \Big(\frac{1+c}{2+c},\ \frac{1}{2+c}\Big), \quad t_* = \frac{c}{2+c};$$
$$p_* = \Big(\frac{c}{2+c},\ \frac{2}{2+c}\Big), \quad s_* = \frac{c}{2+c}.$$

Векторы (q_*, t_*) и (p_*, s_*) являются планами двойственных задач линейного программирования (28), причём значения целевых функций на этих планах равны между собой,

$$t_* = s_* = \frac{c}{2+c}.$$

Указанные свойства гарантируют оптимальность векторов (q_*, t_*) и (p_*, s_*) и, как следствие, оптимальность стратегий q_* и p_*. При этом

$$f(c) = \frac{c}{2+c} \quad \text{при} \quad c > 0. \tag{29}$$

Объединяя формулы (26), (27), (29), приходим к окончательному результату (см. рис.)

$$f(c) = \begin{cases} -1 & \text{при} \quad c \leqslant -1,\\ c & \text{при} \quad c \in (-1, 0),\\ \frac{c}{2+c} & \text{при} \quad c > 0. \end{cases}$$

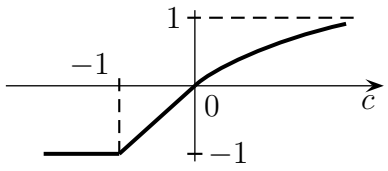


Рис. График цены игры как функции параметра c

Отметим, что

$$f'(-0) = 1, \quad f'(+0) = \frac{1}{2}.$$

ПРИМЕР 2. Найдём все квадратные матрицы второго порядка, которые порождают матричную игру со следующими свойствами:

1) цена игры равна нулю;

2) оптимальными стратегиями игроков являются векторы

$$p_* = \Big(\frac{2}{3},\ \frac{1}{3}\Big), \quad q_* = \Big(\frac{1}{4},\ \frac{3}{4}\Big).$$

Запишем матрицу выигрышей в общем виде

$$A = \begin{pmatrix} c & d \\ g & h \end{pmatrix}.$$

Ей соответствуют две задачи линейного программирования

$$\begin{array}{rr} t \to \inf & s \to \sup \\ -c\,q_1 - d\,q_2 + t \geqslant 0 & -c\,p_1 - g\,p_2 + s \leqslant 0 \\ -g\,q_1 - h\,q_2 + t \geqslant 0 & -d\,p_1 - h\,p_2 + s \leqslant 0 \\ q_1 + \quad q_2 \quad = 1 & p_1 + \quad p_2 \quad = 1 \\ q_1 \geqslant 0,\ q_2 \geqslant 0, & p_1 \geqslant 0,\ p_2 \geqslant 0. \end{array}$$

Оптимальные планы q_*, p_* этих задач известны. Известны и экстремальные значения целевых функций $t_* = s_* = 0$. В силу условий дополнительности имеем

$$\begin{array}{ll} -\dfrac{1}{4}c - \dfrac{3}{4}d = 0, & -\dfrac{2}{3}c - \dfrac{1}{3}g = 0, \\ -\dfrac{1}{4}g - \dfrac{3}{4}h = 0, & -\dfrac{2}{3}d - \dfrac{1}{3}h = 0. \end{array}$$

Отсюда следует, что

$$d = -\frac{1}{3}c, \quad g = -2c, \quad h = -2d = \frac{2}{3}c.$$

Равенство $g = -3h$ выполняется автоматически. Таким образом,

$$A = \begin{pmatrix} c & -\frac{1}{3}c \\ -2c & \frac{2}{3}c \end{pmatrix} = \frac{1}{3}c \begin{pmatrix} 3 & -1 \\ -6 & 2 \end{pmatrix}.$$

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

ТЕОРЕМА БОНДАРЕВОЙ–ШЕПЛИ*

Н. И. Наумова, Н. А. Соловьёва

Аннотация. В теореме Бондаревой–Шепли устанавливается критерий непустоты C-ядра кооперативной игры. В докладе приводится усовершенствованный вариант доказательства этой теоремы, в котором наряду с первой теоремой двойственности из линейного программирования используется лемма о базисном плане.

1°. Кооперативные игры были введены в книге [1]. *Кооперативной игрой* n лиц называется пара (N, v), где $N = \{1, \ldots, n\}$ — множество игроков, а v — отображение, которое каждому подмножеству S множества N ставит в соответствие вещественное число $v(S)$. При этом требуется только, чтобы $v(\varnothing) = 0$. Отображение v называется *характеристической функцией*.

Любое непустое подмножество S множества N называется *коалицией*. Число $v(S)$ интерпретируется как сумма, которую могут заработать вместе игроки из коалиции S, если будут действовать скоординированно. При этом $v(N)$ — это сумма, которую заработают все n игроков в случае согласованных действий.

C-ядром игры (N, v) называется множество

$$C(N,v) = \Big\{x \in R^n : \sum_{i=1}^{n} x[i] = v(N) \ \text{ и } \sum_{i\in S} x[i] \geqslant v(S) \ \text{ для любого } S \subset N,\ S \neq \varnothing,\ S \neq N\Big\}. \tag{1}$$

Предполагается, что все игроки вместе создали коалицию («большую фирму»), доход которой определяется величиной $v(N)$. Общий доход распределяется между участниками коалиции в соответствии с вектором x. Если C-ядро непусто, то суммарный доход $v(S)$ любой коалиции не превосходит суммарного дохода участников этой коалиции в рамках большой фирмы. В таком случае игрокам невыгодно уходить из большой коалиции N и создавать какую-либо частную коалицию S.

*Семинар «CNSA & NDO». Избранные доклады. 18 февраля 2016 г.

Непустота C-ядра является важной характеристикой кооперативной игры. При решении задачи о непустоте C-ядра используются результаты теории линейного программирования.

2°. Рассмотрим следующую задачу линейного программирования:

$$\begin{gathered}\sum_{i=1}^{n} x[i] \to \inf, \\ \sum_{i\in S} x[i] \geqslant v(S) \quad \text{при всех } S \subset N,\ S \neq \varnothing,\ S \neq N.\end{gathered} \tag{2}$$

Множество планов задачи (2) непусто. Например, вектор x с компонентами $x[i] = \max\{0, \max\limits_{S:i\in S} v(S)\}$ является планом задачи (2). Значение целевой функции задачи (2) ограничено снизу на множестве планов числом $\sum\limits_{i=1}^{n} v(\{i\})$. Следовательно (см., например, [2, с. 14]), задача (2) имеет решение. Обозначим через M_* минимальное значение целевой функции.

ПРЕДЛОЖЕНИЕ 1. *C-ядро $C(N, v)$ кооперативной игры (N, v) непусто тогда и только тогда, когда $v(N) \geqslant M_*$.*

Доказательство. Необходимость. Если $C(N, v) \neq \varnothing$, то любой вектор $x \in C(N, v)$ является планом задачи (2). Значит, $\sum\limits_{i=1}^{n} x[i] \geqslant M_*$. Непосредственно из определения C-ядра следует, что $v(N) \geqslant M_*$.
Достаточность. Предположим, что $v(N) \geqslant M_*$. Возьмём оптимальный план x_* задачи (2). Тогда вектор $x \in \mathbb{R}^n$ с компонентами

$$\begin{aligned} x[1] &= x_*[1] + v(N) - M_*, \\ x[j] &= x_*[j], \quad j = 2, \ldots, n, \end{aligned}$$

принадлежит C-ядру кооперативной игры. □

3°. Перенумеруем все непустые собственные подмножества S множества N в произвольном порядке. Получим последовательность

$$S_{\hat{1}},\ S_{\hat{2}},\ \ldots,\ S_{\hat{m}},$$

где $m = 2^n - 2$. Обозначим $\widehat{M} = \{\hat{1}, \ldots, \hat{m}\}$. Введём матрицу $\chi[N, \widehat{M}]$ с элементами

$$\chi[i, \hat{k}] = \begin{cases} 1, & i \in S_{\hat{k}}, \\ 0, & i \notin S_{\hat{k}}. \end{cases}$$

Для примера рассмотрим случай трёх игроков ($n = 3$). Зафиксируем порядок перебора всех непустых собственных подмножеств множества $N = \{1, 2, 3\}$:

$$S_{\hat{1}} = \{1\}, \quad S_{\hat{2}} = \{2\}, \quad S_{\hat{3}} = \{3\}, \quad S_{\hat{4}} = \{1, 2\}, \quad S_{\hat{5}} = \{1, 3\}, \quad S_{\hat{6}} = \{2, 3\}.$$

Тогда матрица $\chi[N, \widehat{M}]$ будет выглядеть так:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} \hat{1} & \hat{2} & \hat{3} & \hat{4} & \hat{5} & \hat{6} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}.$$

Введём вектор $v[\widehat{M}]$ с компонентами $v[\hat{k}] = v(S_{\hat{k}})$. Тогда задачу (2) можно переписать в следующем виде:

$$\begin{gathered} \langle \mathbb{1}, x \rangle \to \inf, \\ \chi^T[\widehat{M}, N] \times x[N] \geqslant v[\widehat{M}]. \end{gathered} \tag{3}$$

(Здесь $\mathbb{1}$ — вектор из единиц длины n.) Запишем задачу линейного программирования, двойственную к задаче (3):

$$\begin{gathered} \langle v, \lambda \rangle \to \sup, \\ \chi[N, \widehat{M}] \times \lambda[\widehat{M}] = \mathbb{1}[N], \\ \lambda[\widehat{M}] \geqslant \mathbb{0}[\widehat{M}]. \end{gathered} \tag{4}$$

По первой теореме двойственности [2, с. 32] задача (4) имеет решение и её экстремальное значение равно M_*. Заметим, что множество планов задачи (4) не зависит от характеристической функции v, а определяется только числом игроков.

4°. План задачи (4) называется *сбалансированным покрытием* (термин введён Л. Шепли [3]). *Минимальным сбалансированным покрытием* называется сбалансированное покрытие, носитель которого не содержит строго носителей других сбалансированных покрытий.

ПРЕДЛОЖЕНИЕ 2. *Вектор λ образует минимальное сбалансированное покрытие тогда и только тогда, когда он является базисным планом задачи* (4).

Доказательство. Необходимость. Приводимое доказательство необходимости составляет часть доказательства леммы о базисном плане [2, с. 14].

Рассмотрим вектор λ^1, который образует минимальное сбалансированное покрытие, но не является базисным планом. Носитель $\widehat{M}^1_+ \subset \widehat{M}$ плана λ^1 не пуст. Столбцы матрицы ограничений с индексами из $\widehat{M}^1_+$ являются линейно зависимыми, значит, система

$$\chi[N, \widehat{M}^1_+] \times z[\widehat{M}^1_+] = \mathbb{O}[N] \tag{5}$$

имеет ненулевое решение $z_0[\widehat{M}^1_+]$. Положим $z_0[\widehat{M} \setminus \widehat{M}^1_+] = \mathbb{O}[\widehat{M} \setminus \widehat{M}^1_+]$. С учётом однородности системы (5) можно считать, что у вектора z_0 есть положительные компоненты. Зададим луч $\lambda(t) = \lambda^1 - tz_0\,,\ t > 0$. При любом вещественном t верно равенство

$$\chi\,\lambda(t) = \chi\,\lambda^1 - t\,\chi z_0 = \mathbb{1}.$$

Вектор $\lambda(t)$ будет планом задачи (4), если $\lambda^1 - tz_0 \geqslant \mathbb{O}$. Заметим, что $\lambda^1[\hat{k}] - tz_0[\hat{k}] = 0$ при $\hat{k} \in \widehat{M} \setminus \widehat{M}^1_+$ и всех вещественных t. Кроме того, при $\hat{k} \in \widehat{M}^1_+$ таких, что $z_0[\hat{k}] \leqslant 0$, неравенство $\lambda^1[\hat{k}] - tz_0[\hat{k}] > 0$ верно при всех положительных t. Теперь рассмотрим $\hat{k} \in \widehat{M}^1_+$ такие, что $z_0[\hat{k}] > 0$. Обозначим

$$t_0 = \min\Big\{ \tfrac{\lambda^1[\hat{k}]}{z_0[\hat{k}]} \mid \hat{k} \in \widehat{M}^1_+ :\ z_0[\hat{k}] > 0 \Big\}.$$

Пусть $\hat{l}$ — индекс, на котором достигается минимум. Вектор $\lambda^2 = \lambda(t_0)$ — план задачи (4), при этом $\lambda^2[\hat{l}] = 0$. Таким образом, носитель плана λ^2 строго содержится в носителе плана λ^1, так что λ^1 не является минимальным сбалансированным покрытием.

Достаточность. Пусть λ^1 — базисный план задачи (4) с носителем $\widehat{M}^1_+$, который не будет минимальным сбалансированным покрытием. Тогда существует сбалансированное покрытие λ^2 с носителем $\widehat{M}^2_+$, такое, что $\widehat{M}^2_+ \subsetneqq \widehat{M}^1_+$. Так как для планов λ^1 и λ^2 верны равенства $\chi[N, \widehat{M}] \times \lambda^1[\widehat{M}] = \mathbb{1}[N]$, $\chi[N, \widehat{M}] \times \lambda^2[\widehat{M}] = \mathbb{1}[N]$, то

$$\chi[N, \widehat{M}] \times \lambda^1[\widehat{M}] - \chi[N, \widehat{M}] \times \lambda^2[\widehat{M}] = \mathbb{O}[N].$$

Это равенство можно переписать так:

$$\sum_{\hat{k} \in \widehat{M}^1_+} \chi[N, \hat{k}] \times \lambda^1[\hat{k}] - \sum_{\hat{k} \in \widehat{M}^2_+} \chi[N, \hat{k}] \times \lambda^2[\hat{k}] = \mathbb{O}[N]$$

или

$$\sum_{\hat{k} \in \widehat{M}^1_+} \chi[N, \hat{k}]\big(\lambda^1[\hat{k}] - \lambda^2[\hat{k}]\big) = \mathbb{O}[N].$$

При всех $\hat{k} \in \widehat{M}^1_+ \setminus \widehat{M}^2_+$ верно $\lambda^1[\hat{k}] - \lambda^2[\hat{k}] \neq 0$. Значит, столбцы $\chi[\,\cdot\,, \hat{k}]$ при $\hat{k} \in \widehat{M}^1_+$ линейно зависимы и план λ^1 не является базисным. $\square$

5°. Напомним формулировку леммы о базисном плане [2, с. 14]. Рассмотрим задачу линейного программирования в канонической форме:

$$\begin{gathered} \langle c, x\rangle \to \inf, \\ Ax = b, \\ x \geqslant \mathbb{O}. \end{gathered} \tag{6}$$

Предположим, что множество планов этой задачи непусто и целевая функция ограничена снизу на нём. Справедлива

ЛЕММА (о базисном плане)**.** *Пусть $b \neq \mathbb{O}$. Тогда любому плану задачи* (6) *можно сопоставить базисный план с меньшим либо равным значением целевой функции.*

Базисных планов конечное число, так как различным базисным планам соответствуют различные носители [2, с. 15]. Таким образом, для решения задачи достаточно перебрать все базисные планы и выбрать те, на которых достигается минимум целевой функции.

Из леммы о базисном плане, предложений 1 и 2 непосредственно следует

ТЕОРЕМА БОНДАРЕВОЙ–ШЕПЛИ. *C-ядро $C(N, v)$ кооперативной игры (N, v) непусто тогда и только тогда, когда*

$$v(N) \geqslant \max\Big\{ \sum_{\hat{k}\in\widehat{M}} \lambda[\hat{k}]\, v[\hat{k}] \;\Big|\; \lambda[\widehat{M}] \text{ — } \textit{минимальное сбалансированное покрытие} \Big\}.$$

Напомним, что $v[\hat{k}] = v(S_{\hat{k}})$, где $S_{\hat{k}}$ — непустое собственное подмножество множества N, имеющее номер $\hat{k}$ в произвольном заранее зафиксированном порядке.

Данный результат был опубликован О. Н. Бондаревой в 1963 году [4] и в 1967 году получен независимо Л. Шепли [3], который не использовал аппарат линейного программирования. Позднее эта теорема стала называться теоремой Бондаревой–Шепли.

6°. В качестве примера рассмотрим кооперативную игру трёх лиц. Зафиксируем тот же порядок непустых собственных подмножеств множества $N = \{1, 2, 3\}$, что и в п. 3°. Задача (2) принимает вид

$$x[1] + x[2] + x[3] \to \inf,$$

$$\begin{aligned}
x[1] &\geqslant v[\hat{1}],\\
x[2] &\geqslant v[\hat{2}],\\
x[3] &\geqslant v[\hat{3}],\\
x[1]+x[2] &\geqslant v[\hat{4}],\\
x[1]+x[3] &\geqslant v[\hat{5}],\\
x[2]+x[3] &\geqslant v[\hat{6}].
\end{aligned}$$

Перейдём к двойственной задаче:

$$\begin{gathered}
v[\hat{1}]\,\lambda[\hat{1}]+v[\hat{2}]\,\lambda[\hat{2}]+v[\hat{3}]\,\lambda[\hat{3}]+v[\hat{4}]\,\lambda[\hat{4}]+v[\hat{5}]\,\lambda[\hat{5}]+v[\hat{6}]\,\lambda[\hat{6}]\to\sup,\\
\begin{aligned}
\lambda[\hat{1}]+\lambda[\hat{4}]+\lambda[\hat{5}] &= 1,\\
\lambda[\hat{2}]+\lambda[\hat{4}]+\lambda[\hat{6}] &= 1,\\
\lambda[\hat{3}]+\lambda[\hat{5}]+\lambda[\hat{6}] &= 1,\\
\lambda[\hat{k}]\geqslant 0,\quad \hat{k}=\hat{1},\dots,\hat{6}.&
\end{aligned}
\end{gathered}\tag{7}$$

Нетрудно проверить, что базисными планами задачи (7) являются следующие векторы:

1) λ^1, где $\lambda^1[\hat{k}]=\begin{cases}1 & \text{при } \hat{k}=\hat{1},\hat{2},\hat{3},\\ 0 & \text{при } \hat{k}=\hat{4},\hat{5},\hat{6};\end{cases}$

2) λ^2, где $\lambda^2[\hat{k}]=\begin{cases}1 & \text{при } \hat{k}=\hat{1},\hat{6},\\ 0 & \text{при } \hat{k}=\hat{2},\hat{3},\hat{4},\hat{5};\end{cases}$

3) λ^3, где $\lambda^3[\hat{k}]=\begin{cases}1 & \text{при } \hat{k}=\hat{2},\hat{5},\\ 0 & \text{при } \hat{k}=\hat{1},\hat{3},\hat{4},\hat{6};\end{cases}$

4) λ^4, где $\lambda^4[\hat{k}]=\begin{cases}1 & \text{при } \hat{k}=\hat{3},\hat{4},\\ 0 & \text{при } \hat{k}=\hat{1},\hat{2},\hat{5},\hat{6};\end{cases}$

5) λ^5, где $\lambda^5[\hat{k}]=\begin{cases}\frac{1}{2} & \text{при } \hat{k}=\hat{4},\hat{5},\hat{6},\\ 0 & \text{при } \hat{k}=\hat{1},\hat{2},\hat{3}.\end{cases}$

Полный перебор носителей показывает, что других базисных планов у задачи (7) нет.

Теорема Бондаревой–Шепли утверждает, что для игры трёх лиц C-ядро непусто тогда и только тогда, когда

$$v(N)\geqslant\max\Big\{v[\hat{1}]+v[\hat{2}]+v[\hat{3}];\ v[\hat{1}]+v[\hat{6}];\ v[\hat{2}]+v[\hat{5}];\ v[\hat{3}]+v[\hat{4}];\ \tfrac{1}{2}\big(v[\hat{4}]+v[\hat{5}]+v[\hat{6}]\big)\Big\}.$$

7°. Алгоритм перечисления всех минимальных сбалансированных покрытий для кооперативных игр с произвольным числом игроков был предложен Б. Пелегом в работе [5].

ЛИТЕРАТУРА

1. Нейман Дж., Моргенштерн О. *Теория игр и экономическое поведение*. М.: Наука, 1970. 707 с.

2. Гавурин М. К., Малозёмов В. Н. Экстремальные задачи с линейными ограничениями. Л.: Изд-во Ленинградского ун-та, 1984. 176 с.

3. Shapley L. S. *On balanced sets and cores* // Naval Research Logistics Quarterly. 1967. V. 14. P. 453–460.

4. Бондарева О. Н. *Некоторые применения методов линейного программирования к теории кооперативных игр* // Проблемы кибернетики. 1963. Выпуск 10. С. 119–139.

5. Peleg B. *An inductive method for constructing minimal balanced collections of finite sets* // Naval Research Logistics Quarterly. 1965. V. 12. P. 155–162.

КОНЕЧНОМЕРНАЯ ПРОБЛЕМА МОМЕНТОВ*

В. Н. Малозёмов

Текст доклада соответствует лекции, которую автор много лет читал в курсе “Экстремальные задачи”.

1°. В линейном пространстве $\mathbb{R}^N$ векторов $x = x[N]$ рассмотрим линейный (однородный и аддитивный) функционал f. По определению

$$f(\alpha_1\, x_1 + \alpha_2\, x_2) = \alpha_1\, f(x_1) + \alpha_2\, f(x_2) \tag{1}$$

при любых x_1, x_2 из $\mathbb{R}^N$ и всех вещественных α_1, α_2.

Единичные орты в $\mathbb{R}^N$ обозначим $e_j = e_j[N]$. Очевидно, что

$$x = \sum_{j\in N} x[j]\, e_j.$$

Согласно (1)

$$f(x) = \sum_{j\in N} x[j]\, f(e_j).$$

Введём вектор $\xi = \xi[N]$ с компонентами $\xi[j] = f(e_j)$. Тогда

$$f(x) = \sum_{j\in N} x[j] \times \xi[j] = \langle \xi, x\rangle \qquad \forall\, x \in \mathbb{R}^N. \tag{2}$$

Получили общий вид линейного функционала в $\mathbb{R}^N$. Нетрудно понять, что вектор ξ, сопоставляемый функционалу f, определяется единственным образом.

2°. В пространстве $\mathbb{R}^N$ введём чебышёвскую норму

$$\|x\|_\infty = \max_{j\in N} \big|x[j]\big|.$$

Согласованная с ней норма линейного функционала f определяется так:

$$\|f\| = \max_{\|x\|_\infty=1} \big|f(x)\big|.$$

*Семинар «DHA & CAGD». Избранные доклады. 11 сентября 2010 г.

Покажем, что

$$\|f\| = \sum_{j\in N} \bigl|\xi[j]\bigr|. \tag{3}$$

Для вектора x с $\|x\|_\infty = 1$ согласно (2) имеем

$$\bigl|f(x)\bigr| \leqslant \sum_{j\in N} \bigl|x[j]\bigr| \times \bigl|\xi[j]\bigr| \leqslant \sum_{j\in N} \bigl|\xi[j]\bigr|. \tag{4}$$

При $\xi = \mathbb{O}$ равенство (3) справедливо, поскольку обе его части равны нулю. Пусть $\xi \neq \mathbb{O}$. Положим $x_0[j] = \operatorname{sign}\xi[j]$. Тогда $\|x_0\|_\infty = 1$ и

$$f(x_0) = \sum_{j\in N} \bigl|\xi[j]\bigr|. \tag{5}$$

Формула (3) следует из (4) и (5).

3°. *Моментом* функционала f называется его значение на некотором элементе $x \in \mathbb{R}^N$, то есть величина $f(x)$. Рассмотрим задачу: *среди линейных функционалов, имеющих заданные моменты*

$$f(a_i) = b[i], \qquad i \in M,$$

найти функционал с наименьшей нормой. Здесь a_i — фиксированные векторы из $\mathbb{R}^N$ и $b[i]$ — фиксированные вещественные числа. С учётом (2) и (3) эту задачу можно формализовать так:

$$\begin{gathered} \|\xi\|_1 := \sum_{j\in N} \bigl|\xi[j]\bigr| \to \inf, \\ \langle a_i, \xi\rangle = b[i], \quad i \in M. \end{gathered} \tag{6}$$

Обозначим через $A = A[M, N]$ матрицу со строками a_i, $i \in M$. Тогда ограничения задачи (6) примут вид

$$A\xi = b. \tag{7}$$

Будем изучать задачу (6) при следующих предположениях: вектор b отличен от нуля (иначе единственным решением задачи (6) является $\xi = \mathbb{O}$); система линейных уравнений (7) совместна. Последнее означает, что множество планов задачи (6) непусто.

Перейдём от (6) к эквивалентной задаче линейного программирования

$$\begin{aligned}\sum_{j\in N}\bigl(y[j]+z[j]\bigr)&\to\inf,\\ \xi[j]=y[j]-z[j],&\qquad j\in N;\\ A\xi&=b;\\ y[j]\geqslant 0,\quad z[j]\geqslant 0,&\quad j\in N.\end{aligned}$$

Сделаем ещё один эквивалентный переход, исключив вектор ξ:

$$\begin{aligned}\sum_{j\in N}\bigl(y[j]+z[j]\bigr)&\to\inf,\\ Ay-Az&=b,\\ y\geqslant\mathbb{O},\quad z&\geqslant\mathbb{O}.\end{aligned}\tag{8}$$

Теперь запишем двойственную задачу линейного программирования

$$\begin{aligned}\langle b,u\rangle&\to\sup,\\ u[M]\times A[M,j]&\leqslant 1,\quad j\in N;\\ -u[M]\times A[M,j]&\leqslant 1,\quad j\in N.\end{aligned}\tag{9}$$

В силу совместности системы (7) множество планов задачи (8) непусто, при этом целевая функция ограничена снизу нулём. Значит, задача (8) имеет решение. Как следствие получаем, что задачи (6), (9) также имеют решения и экстремальные значения целевых функций в задачах (6), (8) и (9) равны между собой. Это общее значение обозначим ν. Условие $b\neq\mathbb{O}$ в задаче (6) гарантирует, что $\nu>0$.

4°. Введём функцию

$$\varphi(u):=\max_{j\in N}\bigl|u[M]\times A[M,j]\bigr|=\|uA\|_\infty=\left\|\sum_{i\in M}u[i]\,a_i\right\|_\infty.\tag{10}$$

С её помощью ограничения задачи (9) можно переписать в виде

$$\varphi(u)\leqslant 1.$$

Перейдём к взаимной по Эйлеру задаче (целевая функция и функция, входящая в ограничение, меняются местами):

$$\begin{aligned}\varphi(u)&\to\inf,\\ \langle b,u\rangle&=1.\end{aligned}\tag{11}$$

Сначала покажем, что $\varphi(u) > 0$ на всех планах задачи (11). Допустим противное. Тогда найдётся вектор $u_0 \in \mathbb{R}^M$ со свойствами

$$\varphi(u_0) = 0, \qquad \langle b, u_0 \rangle = 1.$$

Согласно (10), $u_0 A = \mathbb{O}$. Возьмём вектор $\xi_0 \in \mathbb{R}^N$, удовлетворяющий системе (7) (по условию такой вектор существует). Получим

$$1 = \langle b, u_0 \rangle = \langle u_0, A\xi_0 \rangle = \langle u_0 A, \xi_0 \rangle = 0.$$

Противоречие убеждает нас в положительности $\varphi(u)$ на планах задачи (11).

ТЕОРЕМА 1. *Задача* (11) *имеем решение. Произведение экстремальных значений целевых функций в задачах* (11) *и* (6) *равно единице.*

Доказательство. Отметим, что функция $\varphi(u)$ положительно однородна, то есть

$$\varphi(\lambda u) = \lambda\, \varphi(u) \qquad \text{при } \lambda > 0 \text{ и всех } u \in \mathbb{R}^M.$$

Возьмём решение u^* задачи (9) и положим $u_* = \frac{1}{\nu}\, u^*$. Получим

$$\langle b, u_* \rangle = 1, \qquad \varphi(u_*) \leqslant \tfrac{1}{\nu}. \tag{12}$$

В частности, u_* — план задачи (11).

Пусть u — произвольный план задачи (11). Сопоставим ему вектор $\hat{u} = u/\varphi(u)$. Имеем $\varphi(\hat{u}) = 1$, так что $\hat{u}$ — план задачи (9). При этом

$$1 = \langle b, u \rangle = \varphi(u)\, \langle b, \hat{u} \rangle \leqslant \nu\, \varphi(u).$$

Отсюда в силу (12) следует, что

$$\varphi(u) \geqslant \tfrac{1}{\nu} \geqslant \varphi(u_*). \tag{13}$$

Установлено, что u_* — решение задачи (11).

Обозначим $\mu = \varphi(u_*)$. Подставив в (13) $u = u_*$, получим $\varphi(u_*) = \frac{1}{\nu}$, или $\mu\, \nu = 1$. Теорема доказана. □

На рис. иллюстрируется связь между решениями задач (9) и (11).

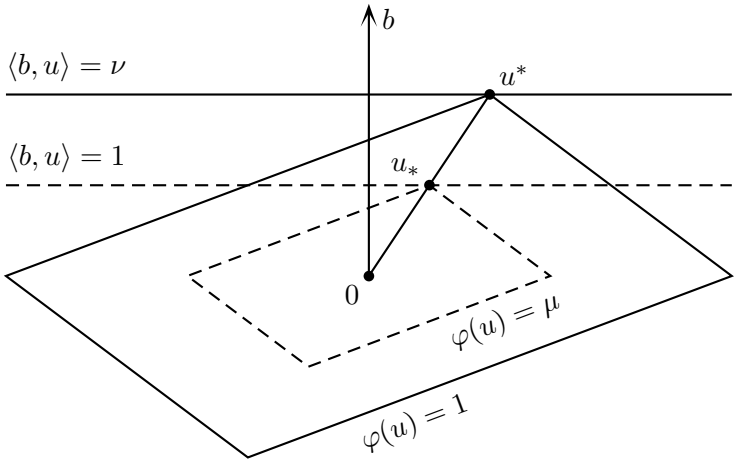


Рис.

5°. Пусть $a_0, a_1, \ldots, a_m$ — векторы из $\mathbb{R}^n$, причём a_0 не принадлежит линейной оболочке, натянутой на векторы $a_1, \ldots, a_m$:

$$a_0 \notin \operatorname{lin}(a_1, \ldots, a_m). \tag{14}$$

Рассмотрим задачу наилучшего приближения

$$\left\| a_0 - \sum_{i=1}^{m} x[i]\, a_i \right\|_\infty \to \inf_{x \in \mathbb{R}^m}. \tag{15}$$

Введём вектор $b = (1, 0, \ldots, 0) \in \mathbb{R}^{m+1}$ и обозначим через A матрицу со строками $a_0, a_1, \ldots, a_m$. Задача (15) эквивалентна задаче вида (11)

$$\begin{gathered} \|uA\|_\infty \to \inf, \\ \langle b, u \rangle = 1. \end{gathered} \tag{16}$$

Сопоставим последней задаче задачу вида (6)

$$\begin{aligned} &\|\xi\|_1 := \sum_{j=0}^{m} \big|\xi[j]\big| \to \inf, \\ &\langle a_0, \xi \rangle = 1, \\ &\langle a_i, \xi \rangle = 0, \quad i \in 1 : m. \end{aligned} \tag{17}$$

Мы хотим воспользоваться теоремой 1, но для этого нужно проверить, что множество планов задачи (17) непусто.

Допустим противное, то есть что система линейных уравнений

$$\begin{aligned} \langle a_0, \xi \rangle &= 1, \\ \langle a_i, \xi \rangle &= 0, \quad i \in 1 : m. \end{aligned}$$

не имеет решения. Тогда найдётся решение u_0 однородной сопряжённой системы

$$\sum_{i=0}^{m} u[i]\, a_i = \mathbb{O},$$

такое, что $u_0[0] \neq 0$. Положив $x_0[i] = -u_0[i]/u_0[0]$, $i \in 1 : m$, получим

$$a_0 = \sum_{i=1}^{m} x_0[i]\, a_i.$$

Но это противоречит условию (14).

На основании теоремы 1 приходим к такому заключению.

ТЕОРЕМА 2. *Задачи* (15) *и* (17) *имеют решения. Произведение экстремальных значений целевых функций в задачах* (15), (17) *равно единице.*

Обозначим соответствующие экстремальные значения через μ и ν. Тогда для любого плана ξ задачи (17) выполняется неравенство

$$\mu = \frac{1}{\nu} \geqslant \frac{1}{\|\xi\|_1}. \tag{18}$$

Равенство достигается на решении задачи (17). Формула (18) даёт оценку снизу для величины наилучшего приближения.

6°. Проблема моментов во всей её полноте рассматривается в монографии [1].

ЛИТЕРАТУРА

1. Крейн М. Г., Нудельман А. А. *Проблема моментов Маркова и экстремальные задачи.* М.: Наука, 1973. 552 с.

ПРИНЦИП МАКСИМУМА ДЛЯ ЛИНЕЙНЫХ ДИСКРЕТНЫХ СИСТЕМ*

В. Н. Малозёмов

1°. Рассмотрим линейную дискретную задачу оптимального управления [1, с. 152—157]:

$$\sum_{k=1}^{s+1}\langle c_k, x_k\rangle + \sum_{k=0}^{s}\langle b_k, u_k\rangle \to \sup, \tag{1}$$

$$x_{k+1} = A_k x_k + B_k u_k, \quad k = 0, 1, \ldots, s; \tag{2}$$

$$D_k u_k \leqslant d_k, \quad k = 0, 1, \ldots, s; \tag{3}$$

$$x_0 = \hat{x}. \tag{4}$$

Здесь A_k, B_k, D_k — матрицы с описанием

$$A_k = A_k[N, N], \quad B_k = B_k[N, M], \quad D_k = D_k[L, M].$$

Вектор $x_k = x_k[N]$ характеризует положение некоторой системы в момент времени k. С помощью вектора $u_k = u_k[M]$ осуществляется управляющее воздействие на систему. Этот вектор должен удовлетворять ограничениям (3). Траектория системы $\{x_0, x_1, \ldots, x_{s+1}\}$ однозначно определяется рекуррентным соотношением (2) и начальным условием (4) — после выбора последовательности управлений $\{u_k\}_{k=0}^{s}$.

Целевую функцию задачи (1) можно интерпретировать как оценку «качества» траектории (с учетом «цены» управления). Требуется выбрать последовательность управлений так, чтобы соответствующая траектория имела наивысшую оценку «качества».

2°. Задача (1)–(4) является задачей линейного программирования относительно блочного вектора неизвестных

$$(u_0, x_1, u_1, x_2, u_2, \ldots, x_s, u_s, x_{s+1}).$$

Представим атрибуты этой задачи в развёрнутом виде.

*Семинар «DHA & CAGD». Избранные доклады. 19 июня 2014 г.

	b_0	c_1	b_1	c_2	b_2	c_3	$\dots$	c_s	b_s	c_{s+1}	
p_0	$-B_0$	E	0	0	0	0	$\dots$	0	0	0	$= A_0\hat{x}$
q_0	D_0	0	0	0	0	0	$\dots$	0	0	0	$\leqslant d_0$
p_1	0	$-A_1$	$-B_1$	E	0	0	$\dots$	0	0	0	$= \mathbb{O}$
q_1	0	0	D_1	0	0	0	$\dots$	0	0	0	$\leqslant d_1$
p_2	0	0	0	$-A_2$	$-B_2$	E	$\dots$	0	0	0	$= \mathbb{O}$
q_2	0	0	0	0	D_2	0	$\dots$	0	0	0	$\leqslant d_2$
$\dots$	$\dots$										$\dots$
p_s	0	0	0	0	0	0	$\dots$	$-A_s$	$-B_s$	E	$= \mathbb{O}$
q_s	0	0	0	0	0	0	$\dots$	0	D_s	0	$\leqslant d_s$

Здесь $E = E[N, N]$ — единичная матрица. В крайнем левом столбце указаны блоки двойственных переменных.

Запишем двойственную задачу линейного программирования:

$$\begin{gathered} \langle A_0\hat{x}, p_0\rangle + \sum_{k=0}^{s}\langle d_k, q_k\rangle \to \inf, \\ -p_kB_k + q_kD_k = b_k, \quad k = 0, 1, \dots, s; \\ p_{k-1} - p_kA_k = c_k, \quad k = 1, \dots, s; \quad p_s = c_{s+1}; \\ q_k \geqslant \mathbb{O}, \quad k = 0, 1, \dots, s. \end{gathered} \tag{5}$$

Из ограничений задачи (5) выделим соотношения

$$\begin{gathered} p_{k-1} = p_kA_k + c_k, \quad k = s, s-1, \dots, 1; \\ p_s = c_{s+1}. \end{gathered} \tag{6}$$

Они позволяют однозначно определить половину блоков двойственных переменных p_s, p_{s-1}, $\dots$, p_0. Задача (5) (после отбрасывания постоянной величины $\langle A_0\hat{x}, p_0\rangle$) принимает вид

$$\begin{gathered} \sum_{k=0}^{s}\langle d_k, q_k\rangle \to \inf, \\ D_k^\top q_k = B_k^\top p_k + b_k, \quad k = 0, 1, \dots, s; \\ q_k \geqslant \mathbb{O}, \quad k = 0, 1, \dots, s. \end{gathered} \tag{7}$$

Слагаемое с индексом k в целевой функции задачи (7) зависит от блока q_k.

На q_k наложены ограничения

$$D_k^\top q_k = B_k^\top p_k + b_k, \quad q_k \geqslant \mathbb{O},$$

которые не зависят от других ограничений. В этом случае значение целевой функции задачи (7) будет наименьшим, если каждое слагаемое независимо примет наименьшее значение. Задача (7) распадается на $s+1$ независимых подзадач (при $k = 0, 1, \ldots, s$)

$$\begin{gathered} \langle d_k, q_k \rangle \to \inf, \\ D_k^\top q_k = B_k^\top p_k + b_k, \\ q_k \geqslant \mathbb{O}. \end{gathered} \tag{8}$$

Каждая задача вида (8) является задачей линейного программирования. Обозначим через u_k блок двойственных переменных и запишем двойственную задачу:

$$\begin{gathered} H_k(u_k) := \langle B_k^\top p_k + b_k, u_k \rangle \to \sup, \\ D_k u_k \leqslant d_k. \end{gathered} \tag{9}$$

Вычислив по формулам (6) блоки двойственных переменных p_s, p_{s-1}, …, p_0 задачи (1) и решив при $k = 0, 1, \ldots, s$ задачи (9), двойственные к задачам (8), найдём последовательность допустимых управлений u_0^*, u_1^*, …, u_s^*. Поскольку мы дважды переходили к двойственным задачам, то можно предположить, что построенная последовательность управлений будет оптимальной для задачи (1). Покажем, что это действительно так.

3°. Обозначим через U_k множество планов задачи (9).

ТЕОРЕМА (принцип максимума)**.** *Для того чтобы последовательность допустимых управлений* $\{u_0^*, u_1^*, \ldots, u_s^*\}$ *была оптимальной для задачи* (1)*, необходимо и достаточно, чтобы при каждом* $k = 0, 1, \ldots, s$ *выполнялось условие*

$$H_k(u_k^*) = \max_{u_k \in U_k} H_k(u_k), \tag{10}$$

то есть чтобы u_k^* *было решением задачи* (9)*.*

Доказательство. Необходимость. Пусть $\{u_k^*\}_{k=0}^s$ — оптимальная последовательность управлений для задачи (1). Это значит, что блочный вектор

$$(u_0^*, x_1^*, u_1^*, x_2^*, u_2^*, \ldots, x_s^*, u_s^*, x_{s+1}^*), \tag{11}$$

где блоки x_k^* последовательно вычисляются на основе рекуррентного соотношения (2) и начального условия (4), является решением задачи линейного

программирования (1). По первой теореме двойственности двойственная задача (5) также имеет решение

$$(p_0,\, q_0^*,\, p_1,\, q_1^*,\, \ldots,\, p_s,\, q_s^*), \tag{12}$$

в котором блоки p_k последовательно вычисляются (в обратном порядке) по формулам (6). По второй теореме двойственности выполняется условие дополнительности

$$\sum_{k=0}^{s} \langle d_k - D_k u_k^*, q_k^* \rangle = 0. \tag{13}$$

Поскольку $d_k - D_k u_k^* \geqslant \mathbb{O}$ и $q_k \geqslant \mathbb{O}$ при всех $k \in 0:s$, от из (13) следует, что

$$\langle d_k - D_k u_k^*, q_k^* \rangle = 0 \quad \text{при каждом } k \in 0:s. \tag{14}$$

Теперь при фиксированном $k \in 0:s$ рассмотрим пару двойственных задач линейного программирования (8) и (9). Векторы q_k^* и u_k^* являются планами этих задач и выполняется условие дополнительности (14). По второй теореме двойственности u_k^* — решение задачи (9), что и требовалось установить.

Достаточность. Допустим, что мы вычислили последовательность векторов $p_s,\, p_{s-1},\, \ldots,\, p_0$ по формулам (6) и при каждом $k \in 0:s$ нашли решение u_k^* задачи (9). Нужно проверить, что последовательность управлений $\{u_0^*,\, u_1^*,\, \ldots,\, u_s^*\}$ является оптимальной для задачи (1).

По первой теореме двойственности при каждом $k \in 0:s$ существует решение q_k^* задачи (8). По второй теореме двойственности выполняется условие дополнительности (14).

Отметим, что векторы (11) и (12) удовлетворяют ограничениям двойственных задач (1) и (5) соответственно, причем как следствие соотношений (14) выполняется условие дополнительности (13). По второй теореме двойственности вектор (11) является решением задачи (1), так что последовательность управлений $\{u_0^*,\, u_1^*,\, \ldots,\, u_s^*\}$ будет оптимальной.

Теорема доказана. □

4°. Принцип максимума позволяет свести большую задачу (1) к последовательности небольших задач вида (9). Имеется ещё одна особенность принципа максимума. Решение «сопряжённой» системы (6) зависит только от матриц A_k и векторов c_k. Это значит, что сведéние задачи (1) к последовательности задач (9) остаётся в силе при изменении матриц B_k, D_k и векторов b_k, d_k — атрибутов, связанных с управлением.

ЛИТЕРАТУРА

1. Ашманов С. А. *Линейное программирование.* М.: Наука, 1981. 340 с.

НАИЛУЧШЕЕ ЛИНЕЙНОЕ ОТДЕЛЕНИЕ ДВУХ МНОЖЕСТВ*

В. Н. Малозёмов, Е. К. Чернэуцану

На линейном уровне исследуется задача наилучшего приближённого отделения двух конечных множеств. Эта задача сводится к задаче негладкой оптимизации, при анализе которой используется вся мощь теории линейного программирования.

В идейном плане мы следуем работе [1].

1°. Пусть в пространстве $\mathbb{R}^n$ заданы два конечных множества

$$A = \{a_i\}_{i=1}^m \quad \text{и} \quad B = \{b_j\}_{j=1}^k.$$

Множества A и B называются *строго отделимыми*, если существуют ненулевой вектор $w \in \mathbb{R}^n$ и вещественное число γ, такие, что

$$\langle w, a_i\rangle < \gamma \quad \text{при всех } i \in 1:m, \tag{1}$$
$$\langle w, b_j\rangle > \gamma \quad \text{при всех } j \in 1:k. \tag{2}$$

При выполнении условий (1) и (2) говорят также, что гиперплоскость H, определяемая уравнением $\langle w, x\rangle = \gamma$, *строго отделяет* множество A от множества B.

2°. Введём функцию

$$f(g) = \frac{1}{m}\sum_{i=1}^{m}\big[\langle w, a_i\rangle - \gamma + c\big]_+ + \frac{1}{k}\sum_{j=1}^{k}\big[-\langle w, b_j\rangle + \gamma + c\big]_+, \tag{3}$$

где $g = (w, \gamma)$, $c > 0$ — параметр и $[u]_+ = \max\{0, u\}$. Ясно, что $f(g) \geqslant 0$ при всех g.

ТЕОРЕМА 1. *Множества A и B строго отделимы тогда и только тогда, когда существует вектор g_*, на котором $f(g_*) = 0$.*

*Семинар «DHA & CAGD». Избранные доклады. 6 октября 2012 г.

Доказательство. Пусть $f(g_*) = 0$ на некотором векторе $g_* = (w_*, \gamma_*)$. Покажем прежде всего, что $w_* \neq \mathbb{O}$. В противном случае

$$f(g_*) = (-\gamma_* + c)_+ + (\gamma_* + c)_+ = \begin{cases} -\gamma_* + c & \text{при } \gamma_* \leqslant -c, \\ 2c & \text{при } \gamma_* \in [-c, c], \\ \gamma_* + c & \text{при } \gamma_* \geqslant c. \end{cases}$$

Отсюда следует, что $f(g_*) \geqslant 2c$. Это противоречит условию $f(g_*) = 0$.

Далее, условие $f(g_*) = 0$ гарантирует, что все слагаемые

$$\big[\langle w_*, a_i\rangle - \gamma_* + c\big]_+ \quad \text{и} \quad \big[-\langle w_*, b_j\rangle + \gamma_* + c\big]_+$$

равны нулю. Это возможно лишь тогда, когда

$$\langle w_*, a_i\rangle - \gamma_* + c \leqslant 0 \quad \text{при всех } i \in 1:m, \tag{4}$$

$$-\langle w_*, b_j\rangle + \gamma_* + c \leqslant 0 \quad \text{при всех } j \in 1:k. \tag{5}$$

Остаётся отметить, что неравенства (4) и (5) обеспечивают выполнение условий строгой отделимости (1) и (2) с $w = w_*$, $\gamma = \gamma_*$.

Докажем обратное утверждение. Пусть выполнены условия (1) и (2). Обозначим

$$d := \min_{j\in 1:k}\langle w, b_j\rangle - \max_{i\in 1:m}\langle w, a_i\rangle > 0, \tag{6}$$

$$w_* = (\tfrac{2c}{d})w, \quad \gamma_* = \tfrac{1}{2}\big[\min_{j\in 1:k}\langle w_*, b_j\rangle + \max_{i\in 1:m}\langle w_*, a_i\rangle\big].$$

Согласно (6) и определению w_*

$$\min_{j\in 1:k}\langle w_*, b_j\rangle - \max_{i\in 1:m}\langle w_*, a_i\rangle = 2c.$$

Имеем

$$\max_{i\in 1:m}\langle w_*, a_i\rangle = 2\gamma_* - \min_{j\in 1:k}\langle w_*, b_j\rangle = 2\gamma_* - 2c - \max_{i\in 1:m}\langle w_*, a_i\rangle,$$

так что

$$\max_{i\in 1:m}\langle w_*, a_i\rangle = \gamma_* - c. \tag{7}$$

Аналогично

$$\min_{j\in 1:k}\langle w_*, b_j\rangle = 2\gamma_* - \max_{i\in 1:m}\langle w_*, a_i\rangle = 2\gamma_* + 2c - \min_{j\in 1:k}\langle w_*, b_j\rangle,$$

так что

$$\min_{j\in 1:k}\langle w_*, b_j\rangle = \gamma_* + c. \tag{8}$$

Положим $g_* = (w_*, \gamma_*)$. На основании (7) и (8) получим $f(g_*) = 0$.

Теорема доказана. □

3°. При доказательстве теоремы 1 описано преобразование вектора $g = (w, \gamma)$, порождающего строго отделяющую гиперплоскость $H = \{x \mid \langle w, x\rangle = \gamma\}$, в вектор $g_* = (w_*, \gamma_*)$, на котором $f(g_*) = 0$. Дело в том, что на само́м векторе g значение $f(g)$ может быть положительным (это зависит от параметра c).

ПРИМЕР 1. В качестве A и B возьмём одноточечные множества на плоскости $\mathbb{R}^2$: $A = \{a\}$, $B = \{b\}$, где $a = (0, 0)$ и $b = (0, 2)$. Вектор $g = (w, \gamma)$ с компонентами $w = (0, 1)$, $\gamma \in (0, 2)$ порождает прямую $x_2 = \gamma$, строго отделяющую точку a от точки b (см. рис. 1).

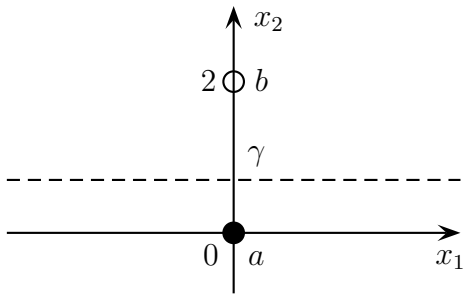


Рис. 1

Вместе с тем,

$$f(g) = [-\gamma + c]_+ + [-2 + \gamma + c]_+.$$

На рис. 2 представлен график $f(g)$ как функции от γ при $c \in (0, 1]$.

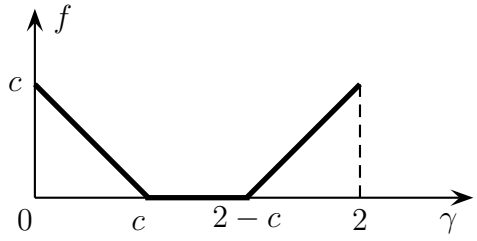


Рис. 2

Видим, что $f(g) = 0$ при $\gamma \in [c, 2 - c]$. При $\gamma \in (0, c) \cup (2 - c, 2)$ прямая $x_2 = \gamma$ по-прежнему строго отделяет точку a от точки b, но $f(g) > 0$.

4°. Рассмотрим экстремальную задачу

$$f(g) \to \min, \tag{9}$$

где $f(g)$ — функция вида (3). Эта задача эквивалентна задаче линейного программирования

$$\begin{gathered}\frac{1}{m}\sum_{i=1}^{m} y_i + \frac{1}{k}\sum_{j=1}^{k} z_j \to \min, \qquad (10)\\ -\langle w, a_i\rangle + \gamma + y_i \geqslant c, \quad i \in 1:m;\\ \langle w, b_j\rangle - \gamma + z_j \geqslant c, \quad j \in 1:k;\\ y_i \geqslant 0,\ i \in 1:m; \quad z_j \geqslant 0,\ j \in 1:k.\end{gathered}$$

Множество планов задачи (10) непусто (планом является вектор с компонентами $w = \mathbb{O}$, $\gamma = 0$, $y_i \equiv c$, $z_j \equiv c$) и целевая функция ограничена снизу нулём. Значит, задача (10) имеет решение. По эквивалентности существует решение и у задачи (9), причём минимальные значения целевых функций у этих задач равны между собой. Это общее значение обозначим через μ. Отметим также, что если $\bigl(w_*, \gamma_*, \{u_i^*\}, \{v_j^*\}\bigr)$ — решение задачи (10), то $g_* = \{w_*, \gamma_*\}$ — решение задачи (9).

5°. При $\mu = 0$ получим $f(g_*) = 0$. По теореме 1 вектор $g_* = (w_*, \gamma_*)$ порождает гиперплоскость $H = \bigl\{x \mid \langle w_*, x\rangle = \gamma_*\bigr\}$, строго отделяющую множество A от множества B.

Вектор g_* можно улучшить, пользуясь неединственностью решения задачи (9). Положим

$$\begin{gathered}w_0 = w_*/\|w_*\|,\\ \gamma_0 = \tfrac{1}{2}\bigl[\min_{j\in 1:k}\langle w_0, b_j\rangle + \max_{i\in 1:m}\langle w_0, a_i\rangle\bigr],\\ c_0 = \tfrac{1}{2}\bigl[\min_{j\in 1:k}\langle w_0, b_j\rangle - \max_{i\in 1:m}\langle w_0, a_i\rangle\bigr],\\ g_0 = (w_0, \gamma_0).\end{gathered}$$

Тогда при всех $i \in 1:m$

$$\langle w_0, a_i\rangle - \gamma_0 + c_0 = \langle w_0, a_i\rangle - \max_{i\in 1:m}\langle w_0, a_i\rangle \leqslant 0,$$

и при всех $j \in 1:k$

$$-\langle w_0, b_j\rangle + \gamma_0 + c_0 = -\langle w_0, b_j\rangle + \min_{j\in 1:k}\langle w_0, b_j\rangle \leqslant 0.$$

Значит, $f(g_0) = 0$ при $c = c_0$. Гиперплоскость $H_0 = \bigl\{x \mid \langle w_0, x\rangle = \gamma_0\bigr\}$ строго отделяет множество A от множества B, причём ширина отделяющей полосы равна $2c_0$.

6°. Как отмечалось в п. 4°, задача (9) всегда имеет решение. При $\mu > 0$ по теореме 1 множества A и B не допускают строгого линейного отделения. В этом случае будем говорить, что гиперплоскость $H_* = \{x \mid \langle w_*, x\rangle = \gamma_*\}$, порождаемая решением $g_* = (w_*, \gamma_*)$ задачи (9), является *наилучшей гиперплоскостью, приближённо отделяющей множество A от множества B* (при данном значении параметра c).

Здесь, однако, имеется тонкость: нет гарантии, что компонента w_* вектора g_* отлична от нулевой. Разберёмся в этой ситуации.

ТЕОРЕМА 2. *Для того чтобы задача* (9) *имела решение $g_* = (w_*, \gamma_*)$ с $w_* = \mathbb{O}$, необходимо и достаточно, чтобы выполнялось условие*

$$\frac{1}{m}\sum_{i=1}^{m} a_i = \frac{1}{k}\sum_{j=1}^{k} b_j. \tag{11}$$

Доказательство. Необходимость. При $w_* = \mathbb{O}$ легко вычисляется экстремальное значение целевой функции у задачи линейного программирования (10). Действительно,

$$\mu = f(g_*) = \min_{\gamma}\bigl\{[-\gamma + c]_+ + [\gamma + c]_+\bigr\} = 2c.$$

Такое же экстремальное значение имеет задача линейного программирования, двойственная к задаче (10). В силу разрешимости двойственной задачи совместна система

$$c\left(\sum_{i=1}^{m} u_i + \sum_{j=1}^{k} v_j\right) = 2c, \tag{12}$$

$$-\sum_{i=1}^{m} u_i a_i + \sum_{j=1}^{k} v_j b_j = \mathbb{O}, \tag{13}$$

$$\sum_{i=1}^{m} u_i - \sum_{j=1}^{k} v_j = 0, \tag{14}$$

$$0 \leqslant u_i \leqslant \tfrac{1}{m},\ i \in 1:m; \quad 0 \leqslant v_j \leqslant \tfrac{1}{k},\ j \in 1:k. \tag{15}$$

Из (12) и (14) следует, что

$$\sum_{i=1}^{m} u_i = 1, \quad \sum_{j=1}^{k} v_j = 1.$$

Принимая во внимание (15), заключаем, что все u_i равны $\frac{1}{m}$ и все v_j равны $\frac{1}{k}$. Теперь формула (13) становится эквивалентной равенству (11).

Достаточность. Запишем задачу, двойственную к (10):

$$c\Big(\sum_{i=1}^{m} u_i + \sum_{j=1}^{k} v_j\Big) \to \max$$

при ограничениях (13)–(15). В силу (11) набор $u_i \equiv \frac{1}{m}$, $v_j \equiv \frac{1}{k}$ является планом этой задачи. Значение целевой функции на нём равно $2c$.

Вместе с тем, на плане

$$w = \mathbb{O}, \quad \gamma = 0, \quad y_i \equiv c, \quad z_j \equiv c \tag{16}$$

задачи (10) значение целевой функции также равно $2c$. Отсюда следует, что план (16) задачи (10) с $w = \mathbb{O}$ является оптимальным.

Теорема доказана. □

ПРИМЕР 2. Рассмотрим на плоскости $\mathbb{R}^2$ два двухточечных множества

$$A = \big\{(0,0),(1,1)\big\}, \quad B = \big\{(1,0),(0,1)\big\}$$

(см. рис. 3). В данном случае выполняется условие (11). По теореме 2 задача (9) имеет решение $g_* = (w_*, \gamma_*)$ с $w_* = \mathbb{O}$. При этом $\mu = 2c$.

Покажем, что у задачи (9) существует другое решение $g_0 = (w_0, \gamma_0)$ с $w_0 \neq \neq \mathbb{O}$.

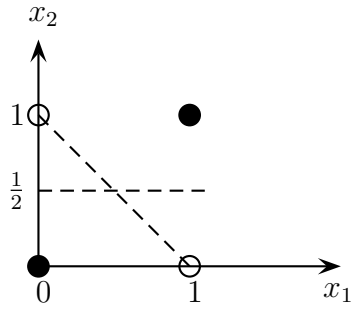


Рис. 3

Согласно (3)

$$f(g) = \tfrac{1}{2}\big\{[-\gamma + c]_+ + [w^1 + w^2 - \gamma + c]_+\big\} + \tfrac{1}{2}\big\{[-w^1 + \gamma + c]_+ + [-w^2 + \gamma + c]_+\big\}.$$

Здесь $w = (w^1, w^2)$. Положим

$$w_0 = (c, c), \quad \gamma_0 = c, \quad g_0 = (w_0, \gamma_0).$$

Тогда $f(g_0) = 2c$. Значит, на векторе g_0 достигается минимум функции $f(g)$. Гиперплоскость $H_0 = \{x \mid x_1 + x_2 = 1\}$ является наилучшей гиперплоскостью, приближённо отделяющей множество A от множества B.

Таким же свойством обладают вектор $g_1 = (w_1, \gamma_1)$ с $w_1 = (0, c)$, $\gamma_1 = \frac{c}{2}$ и гиперплоскость $H_1 = \{x \mid x_2 = \frac{1}{2}\}$ (см. рис. 3).

7°. Особенность, отмеченная в примере 2, имеет общий характер.

ТЕОРЕМА 3. *При $\mu > 0$ у задачи* (9) *существует решение $g_0 = (w_0, \gamma_0)$ с $w_0 \neq \mathbb{O}$.*

Доказательство. Допустим, что у решения $g_* = (w_*, \gamma_*)$ задачи (9) компонента w_* оказалась нулевой. Построим другое решение $g_0 = (w_0, \gamma_0)$ с $w_0 \neq \mathbb{O}$.

По теореме 2 выполняется соотношение (11) и $\mu = 2c$. Возьмём произвольный ненулевой вектор $h \in \mathbb{R}^n$ и рассмотрим задачу линейного программирования

$$\begin{gathered}
\langle h, w\rangle \to \min, \qquad (17)\\
-\frac{1}{m}\sum_{i=1}^{m} y_i - \frac{1}{k}\sum_{j=1}^{k} z_j = -2c;\\
-\langle w, a_i\rangle + \gamma + y_i \geqslant c, \quad i \in 1:m;\\
\langle w, b_j\rangle - \gamma + z_j \geqslant c, \quad j \in 1:k;\\
y_i \geqslant 0,\ i \in 1:m; \quad z_j \geqslant 0,\ j \in 1:k.
\end{gathered}$$

Вектор (16) удовлетворяет ограничениям задачи (17), то есть является её планом. Покажем, что этот план не может быть оптимальным.

В случае оптимальности плана (16) у задачи, двойственной к (17), должен существовать план с таким же (нулевым) значением целевой функции. Таким образом, должна быть совместной система

$$c\left(\sum_{i=1}^{m} u_i + \sum_{j=1}^{k} v_j - 2\zeta\right) = 0, \qquad (18)$$

$$-\sum_{i=1}^{m} u_i a_i + \sum_{j=1}^{k} v_j b_j = h, \qquad (19)$$

$$\sum_{i=1}^{m} u_i - \sum_{j=1}^{k} v_j = 0, \qquad (20)$$

$$0 \leqslant u_i \leqslant \tfrac{1}{m}\zeta,\ i \in 1:m; \quad 0 \leqslant v_j \leqslant \tfrac{1}{k}\zeta,\ j \in 1:k. \qquad (21)$$

Покажем, однако, что эта система несовместна.

Из (18) и (20) следует, что

$$\sum_{i=1}^{m} u_i = \zeta, \quad \sum_{j=1}^{k} v_j = \zeta.$$

В силу (21) получаем $u_i \equiv \frac{1}{m}\zeta$, $v_j \equiv \frac{1}{k}\zeta$. Равенство (19) принимает вид

$$\zeta\left(-\frac{1}{m}\sum_{i=1}^{m} a_i + \frac{1}{k}\sum_{j=1}^{k} b_j\right) = h.$$

Но это противоречит условию (11) (напомним, что $h \neq \mathbb{O}$).

Установлено, что план (16) задачи (17) с нулевым значением целевой функции не является оптимальным. Значит, существует план

$$\left(w_0,\ \gamma_0,\ \{u_i^0\},\ \{v_j^0\}\right) \tag{22}$$

с отрицательным значением целевой функции. У такого плана должно быть $w_0 \neq \mathbb{O}$.

Теперь отметим, что план (22) задачи (17) удовлетворяет ограничениям задачи (10) и на нём целевая функция задачи (10) принимает наименьшее возможное значение, равное $2c$ (напомним, что $\mu = 2c$). В силу эквивалентности задач (9) и (10) вектор $g_0 = (w_0, \gamma_0)$ с $w_0 \neq \mathbb{O}$ будет решением задачи (9).

Теорема доказана. □

Замечание. В качестве ненулевого вектора h можно взять, например, любую ненулевую разность $b_{j_0} - a_{i_0}$. В этом случае множество планов задачи, двойственной к (17), которое определяется условиями (19)– (21), будет непустым. Вместе с непустотой множества планов самóй задачи (17) это гарантирует наличие у задачи (17) оптимального плана.

8°. При $\mu > 0$ решение $g_0 = (w_0, \gamma_0)$ задачи (9) с $w_0 \neq \mathbb{O}$ можно привести к каноническому виду. Как и в п. 5° положим

$$\begin{gathered} w_1 = w_0/\|w_0\|, \\ \gamma_1 = \tfrac{1}{2}\big[\min_{j\in 1:k}\langle w_1, b_j\rangle + \max_{i\in 1:m}\langle w_1, a_i\rangle\big], \\ c_1 = \tfrac{1}{2}\big[\min_{j\in 1:k}\langle w_1, b_j\rangle - \max_{i\in 1:m}\langle w_1, a_i\rangle\big], \\ g_1 = (w_1, \gamma_1). \end{gathered}$$

В данном случае $c_1 \leqslant 0$. При $c_1 = 0$ гиперплоскость $H_1 = \{x \mid \langle w_1, x\rangle = \gamma_1\}$ нестрого отделяет множество A от множества B. При $c_1 < 0$ та же гиперплоскость H_1 является наилучшей, приближённо отделяющей множество A от множества B.

Согласно определению w_1, γ_1, c_1 имеем

$$\begin{aligned} \langle w_1, a_i\rangle - \gamma_1 + c_1 \leqslant 0, \quad i \in 1:m, \\ -\langle w_1, b_j\rangle + \gamma_1 + c_1 \leqslant 0, \quad j \in 1:k. \end{aligned}$$

При $c_1 < 0$ эти неравенства определяют “смешанную полосу”

$$c_1 \leqslant \langle w_1, x\rangle - \gamma_1 \leqslant -c_1,$$

которая содержит как точки множества A, так и точки множества B. Ширина смешанной полосы равна $2|c_1|$.

На рис. 4 приведён пример наилучшего приближённого отделения двух множеств.

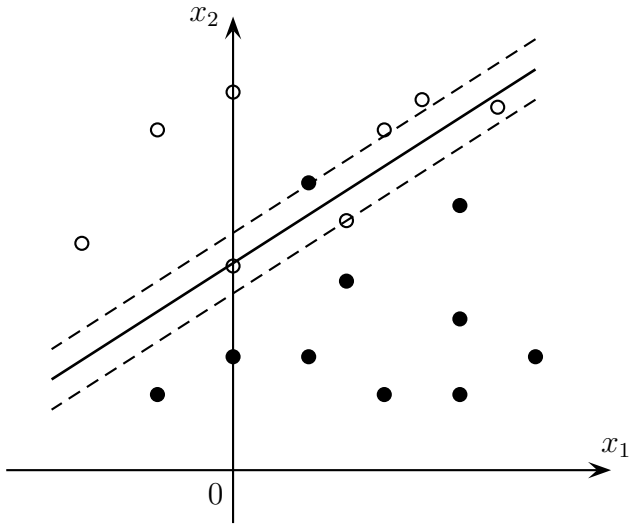


Рис. 4

9°. Чтобы подчеркнуть зависимость задачи (9) от параметра c, будем писать $f(g, c)$, $\mu(c)$ вместо $f(g)$ и μ. Очевидно, что при всех $c > 0$ справедливо формула

$$f(cg, c) = c\ f(g, 1).$$

Поэтому

$$\mu(c) = \min_g f(g, c) = \min_g f(cg, c) = c \min_g f(g, 1) = c\mu(1).$$

Более того, если g_1 — решение задачи (9) при $c = 1$, то вектор $g_c = c\ g_1$ будет решением задачи (9) при произвольном $c > 0$. Таким образом, аддитивный параметр $c > 0$ играет роль нормирующего множителя.

ЛИТЕРАТУРА

1. Bennett K. P., Mangassarian O. L. *Robust linear programming discrimination of two linearly inseparable sets* // Optimization Methods and Software. 1992. Vol. 1. P. 23–34.

СТРОГОЕ ПОЛИНОМИАЛЬНОЕ ОТДЕЛЕНИЕ ДВУХ МНОЖЕСТВ*

В. Н. Малозёмов, А. В. Плоткин

Аннотация. Как известно, задача *наилучшего линейного отделения* двух конечных множеств в евклидовом пространстве сводится к задаче линейного программирования. В этом докладе мы покажем, что задача *строгого полиномиального отделения* двух конечных множеств также сводится к задаче линейного программирования.

1°. Пусть в пространстве $\mathbb{R}^n$ заданы два конечных множества

$$A = \{a_i\}_{i=1}^m \quad \text{и} \quad B = \{b_j\}_{j=1}^k.$$

Рассмотрим обобщённый полином

$$P(x,t) = \sum_{s=1}^{r} x[s]u_s(t), \quad t \in \mathbb{R}^n,$$

где $u_s(t)$ — непрерывные функции от n переменных.

Будем говорить, что множества A и B *строго полиномиально отделимы*, если найдётся вектор коэффициентов $x_0 \in \mathbb{R}^r$, такой что

$$\begin{aligned} &P(x_0, a_i) \geqslant 1 \text{ при всех } i \in 1:m, \\ &P(x_0, b_j) \leqslant -1 \text{ при всех } j \in 1:k. \end{aligned} \tag{1}$$

При этом отделяющая «гиперповерхность» определяется уравнением

$$P(x_0, t) = 0.$$

Покажем, что построение отделяющего полинома $P(x_0, t)$ сводится к решению задачи линейного программирования.

*Семинар «CNSA & NDO». Избранные доклады. 20 октября 2016 г.

2°. Введём функцию

$$f(x) = \max\Big\{\max_{i\in 1:m}\big[1 - P(x,a_i)\big]_+,\ \max_{j\in 1:k}\big[1 + P(x,b_j)\big]_+\Big\},$$

где $[u]_+ = \max\{0,u\}$. Очевидно, что $f(x) \geqslant 0$ при всех $x \in \mathbb{R}^r$.

ЛЕММА. *Полином $P(x_0,t)$ строго отделяет множества A и B тогда и только тогда, когда $f(x_0) = 0$.*

Доказательство. Условия строгого отделения (1) можно переписать в эквивалентном виде

$$\max_{i\in 1:m}\big[1 - P(x_0,a_i)\big]_+ = 0 \quad \text{и} \quad \max_{j\in 1:k}\big[1 + P(x_0,b_j)\big]_+ = 0. \tag{2}$$

Теперь заключение леммы становится очевидным. □

Лемма показывает, что задача строгого полиномиального отделения множеств A и B сводится к минимизации функции $f(x)$ на $\mathbb{R}^r$.

3°. Рассмотрим экстремальную задачу

$$f(x) \to \min_{x\in\mathbb{R}^r}. \tag{3}$$

ТЕОРЕМА 1. *Задача* (3) *эквивалентна следующей задаче линейного программирования:*

$$\begin{gathered} w \to \min \\ P(x,a_i) + w \geqslant 1, \quad i \in 1:m; \\ -P(x,b_j) + w \geqslant 1, \quad j \in 1:k; \\ w \geqslant 0. \end{gathered} \tag{4}$$

Доказательство. Ограничения задачи (4) можно переписать так:

$$\begin{aligned} w &\geqslant \big[1 - P(x,a_i)\big]_+, \quad i \in 1:m; \\ w &\geqslant \big[1 + P(x,b_j)\big]_+, \quad j \in 1:k. \end{aligned} \tag{5}$$

Напомним [1, с. 11–13], что две задачи на минимум называются эквивалентными, если любому плану каждой из этих задач можно сопоставить план другой задачи с равным или меньшим значением целевой функции. В данном случае плану $x \in \mathbb{R}^r$ задачи (3) сопоставим вектор (x,w), где $w = f(x)$. Согласно (5), вектор (x,w) является планом задачи (4), при этом $w = f(x)$.

Наоборот, пусть (x,w) — план задачи (4). Тогда x — план задачи (3). При этом, в силу (5), $f(x) \leqslant w$.

Теорема доказана. □

4°. Задача линейного программирования (4) имеет решение, поскольку множество её планов непусто (за счёт w) и целевая функция ограничена снизу нулём. По лемме об эквивалентных экстремальных задачах задача (3) также имеет решение, причём минимальные значения целевых функций у задач (3) и (4) равны между собой. Обозначим это общее значение через w_*.

ТЕОРЕМА 2. *Величина w_* может принимать только два значения:* 0 *или* 1.

Доказательство. Пусть (x_*, w_*) — решение задачи (4). Ясно, что $w_* \geqslant 0$. Так как пара $x = \mathbb{O}$, $w = 1$ является планом задачи (4) со значением целевой функции, равным единице, то $w_* \leqslant 1$. Таким образом, $w_* \in [0, 1]$.

Предположим, что $w_* \in (0, 1)$. Согласно (4) для полинома $P(x_*, t)$ выполняются неравенства

$$P(x_*, a_i) \geqslant 1 - w_*, \quad i \in 1 : m;$$
$$P(x_*, b_j) \leqslant -(1 - w_*), \quad j \in 1 : k.$$

Пара $x_0 = \frac{1}{1-w_*} x_*$, $w_0 = 0$ удовлетворяет ограничениям задачи (4). Значит, необходимо $w_* \leqslant w_0$, что противоречит выбору w_*.

Для w_* остаются только две возможности: $w_* = 0$ или $w_* = 1$.

Теорема доказана. □

5°. Строгое полиномиальное отделение множеств A и B сводится к решению задачи линейного программирования (4). Если (x_*, w_*) — решение этой задачи, то при $w_* = 0$ полином $P(x_*, t)$ строго отделяет множества A и B. При $w_* = 1$ строгое полиномиальное отделение множеств A и B невозможно.

6°. Ниже мы приводим восемь примеров строгого отделения двух множеств на плоскости с помощью алгебраических полиномов 4-й степени от двух переменных (см. рисунки).

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

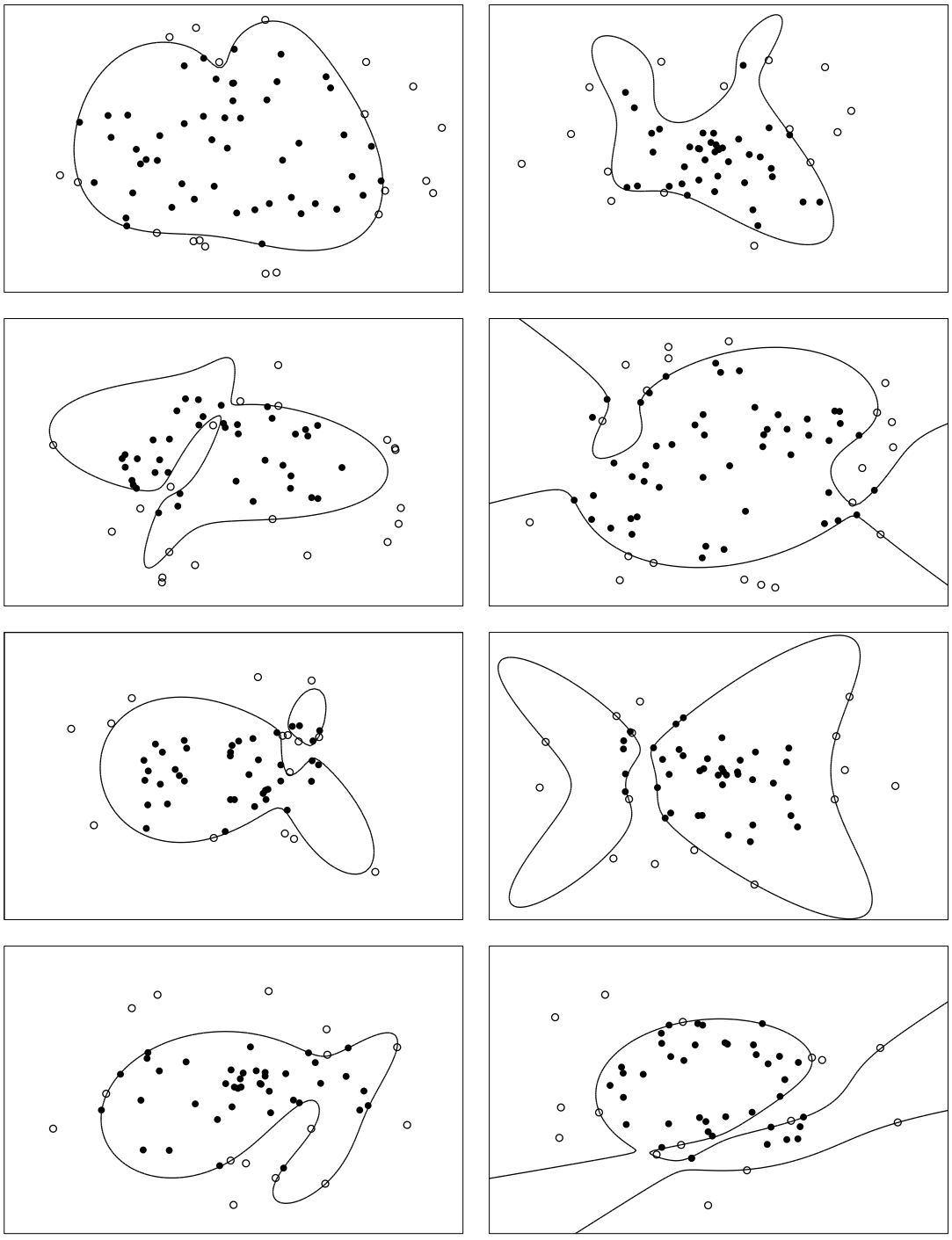


Рис. Строгое отделение двух множеств с помощью полиномов 4-й степени

РЕШЕНИЕ ЗАДАЧ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ В СРЕДЕ MATLAB*

А. Н. Сергеев, Н. А. Соловьёва, Е. К. Чернэуцану

В среде MATLAB задачи линейного программирования решаются с помощью функции `linprog`. Доклад посвящён описанию её возможностей.

1°. Функция `linprog` решает задачу линейного программирования в форме

$$\begin{aligned} \mathtt{f}^T \cdot \mathtt{x} &\to \inf, \\ \mathtt{A} \cdot \mathtt{x} &\leqslant \mathtt{b}, \\ \mathtt{Aeq} \cdot \mathtt{x} &= \mathtt{beq}, \\ \mathtt{lb} \leqslant \mathtt{x} &\leqslant \mathtt{ub}. \end{aligned} \tag{1}$$

Основными входными данными `linprog` являются: вектор коэффициентов целевой функции `f`, матрица ограничений-неравенств `A`, вектор правых частей ограничений-неравенств `b`, матрица ограничений-равенств `Aeq`, вектор правых частей ограничений-равенств `beq`, вектор `lb`, ограничивающий план `x` снизу, вектор `ub`, ограничивающий план `x` сверху. На выходе функция `linprog` даёт оптимальный план `x` задачи (1) и экстремальное значение целевой функции `fval`.

ПРИМЕР 1. Решим в MATLAB задачу линейного программирования

$$\begin{aligned} f(x) = 3x_1 + x_2 + 2x_3 &\to \inf, \\ x_1 + x_2 + x_3 &\geqslant 1, \\ 2x_1 + x_2 - x_3 &\geqslant -1, \\ x_1 - x_2 + x_3 &= 0, \\ 0 \leqslant x_1 &\leqslant 1, \\ 0 \leqslant x_2 &\leqslant 1, \\ 0 \leqslant x_3 &\leqslant 1. \end{aligned}$$

Соответствующая программа (m-файл)[1] выглядит так:

*Семинар «DHA & CAGD». Избранные доклады. 12 февраля 2011 г.

[1]Для отладки приведённых в докладе программ использовался MATLAB 7.11.0 (R2010b).

```
clear all
close all
clc % удаляются все текущие переменные из памяти MATLAB,
    закрываются все графические окна, очищается экран консоли
C = [3 1 2]; % задаётся вектор длины три
D = [1 1 1; 2 1 -1]; % строки матрицы разделяются точкой с
    запятой
B = [1 -1];
Aeq = [1 -1 1];
beq = [0];
lb = zeros(3,1); % задаётся нулевой вектор длины три
ub = [1 1 1];
f = C;
A = -D;
b = -B; % появляются знаки «-», так как ограничения-неравенства
    Dx ⩾ B приводятся к виду −Dx ⩽ −B
[x,fval] = linprog(f,A,b,Aeq,beq,lb,ub);
x
fval
```

Запустив программу, получим сообщение

```
Optimization terminated.
x =
  0
  0.5000
  0.5000
fval =
  1.5000
```

Дополнительно можно задать начальное приближение `x0`:

```
[x,fval] = linprog(f,A,b,Aeq,beq,lb,ub,x0).
```

Если какой-то из входных параметров отсутствует, на его место следует поставить квадратные скобки `[]`, за исключением случая, когда это последний параметр в списке. Например, если нужно решить задачу без ограничений-равенств, в которой не задано начальное приближение, то оператор вызова функции `linprog` будет выглядеть так:

```
[x,fval] = linprog(f,A,b,[],[],lb,ub).
```

(Квадратные скобки в конце списка, соответствующие начальному приближению, не ставятся.)

С помощью входного параметра `options` устанавливаются некоторые дополнительные настройки, в частности, выбирается алгоритм решения. MATLAB решает задачи линейного программирования двумя способами: алгоритмом внутренней точки (`Large-Scale Algorithm`) и вариантом симплекс-метода (`Medium-Scale Algorithm`). По умолчанию используется алгоритм внутренней точки. Чтобы выбрать симплекс-метод, нужно написать

```
options = optimset('LargeScale','off','Simplex','on');
[x,fval] = linprog(f,A,b,Aeq,beq,lb,ub,[],options).
```

Разберёмся с выходными данными. MATLAB позволяет выводить информацию о том, как завершилось решение задачи. За это отвечает параметр `exitflag`. Если значение `exitflag` равно 1, то найдено решение задачи, если равно 0, то превышено допустимое число итераций, если равно -2 — множество планов задачи пусто, если равно -3 — целевая функция не ограничена снизу на множестве планов. Интерпретация других значений параметра `exitflag` приведена в MATLAB Help. Для симплекс-метода допустимое число итераций (`MaxIter`) по умолчанию в 10 раз больше количества переменных. Значение `MaxIter` можно изменить. Чтобы установить допустимое число итераций равным, к примеру, 10, нужно написать

```
options =
optimset('LargeScale','off','Simplex','on','MaxIter',10);
[x,fval] = linprog(f,A,b,Aeq,beq,lb,ub,[],options).
```

Если после выполнения десятой итерации решение не будет найдено, параметр `exitflag` станет нулевым и на экране появится сообщение

```
Maximum number of iterations exceeded;
increase options.MaxIter.
```

Параметр `output` содержит информацию о процессе оптимизации, в частности, число итераций (`iterations`) и используемый алгоритм (`algorithm`). Другие поля параметра `output` описаны в MATLAB Help. Запустим с данными из примера 1 следующую программу:

```
options = optimset('LargeScale','off','Simplex','on');
[x,fval,exitflag,output] =
linprog(f,A,b,Aeq,beq,lb,ub,[],options);
exitflag
output.iterations
output.algorithm
```

На выходе получим:

```
Optimization terminated.
```

```
exitflag =
  1
ans =
  1
ans =
  medium scale: simplex
```

Это означает, что симплекс-метод успешно завершил работу, для нахождения решения потребовалось одна итерация.

Наконец, в выходном параметре `lambda` содержится решение двойственной задачи линейного программирования. Параметр `lambda` состоит из четырёх массивов: `lambda.ineqlin`, `lambda.eqlin`, `lambda.upper`, `lambda.lower`. В этих массивах находятся двойственные переменные, приписанные ограничениям-неравенствам, ограничениям-равенствам, ограничениям на план сверху и снизу соответственно. Подробное обсуждение этого вопроса отложим до п. 3°.

2°. При решении задачи линейного программирования возможны три выхода из процесса: найдено решение задачи, множество планов пусто, целевая функция не ограничена снизу на множестве планов. Продемонстрируем эти варианты на примерах.

ПРИМЕР 2. Решим в MATLAB задачу линейного программирования

$$\begin{aligned} f(x) = 4x_1 + x_2 \to \inf, \\ x_1 + x_2 \geqslant 2, \\ x_1 - x_2 \geqslant 1, \\ x_1 \geqslant 0,\ x_2 \geqslant 0. \end{aligned} \tag{2}$$

Соответствующая программа будет выглядеть так:

```
clear all
close all
clc
C = [4 1];
D = [1 1; 1 -1];
B = [2 1];
lb = zeros(2,1);
f = C;
A = -D;
b = -B;
options = optimset('LargeScale','off','Simplex','on')};
[x,fval,exitflag] = linprog(f,A,b,[],[],lb,[],[],options);
```

```
x
fval
exitflag
```

В результате работы программы получим:

```
Optimization terminated.
x =
  1.5000
  0.5000
fval =
  6.5000
exitflag =
  1
```

Найдено решение задачи (2).

ПРИМЕР 3. Решим в MATLAB задачу линейного программирования

$$\begin{aligned} f(x) = x_1 + x_2 &\to \inf, \\ -x_1 - x_2 &\geqslant -1, \\ x_1 + 4x_2 &\geqslant 8, \\ x_1 \geqslant 0,\ x_2 &\geqslant 0. \end{aligned} \tag{3}$$

Приведём результат работы соответствующей программы:

```
Exiting: The constraints are overly stringent; no feasible
starting point found.
x =
  0
  1
fval =
  1
exitflag =
  -2
```

Множество планов задачи (3) пусто.

ПРИМЕР 4. Решим в MATLAB задачу линейного программирования

$$\begin{aligned} f(x) = -x_1 - 3x_2 &\to \inf, \\ 2x_1 - x_2 &\geqslant 0, \\ -x_1 + x_2 &\geqslant -1, \\ x_1 \geqslant 0,\ x_2 &\geqslant 0. \end{aligned} \tag{4}$$

Запустив программу, решающую задачу (4), получим:

```
Exiting: The solution is unbounded and at infinity;
the constraints are not restrictive enough.
x =
  1.0e+016*
  1.0000
  2.0000
fval =
  -7.0000e+016
exitflag =
  -3
```

Целевая функция задачи (4) не ограничена снизу на множестве планов.

3°. Рассмотрим задачу линейного программирования в общем виде

$$\begin{aligned} f(x) := c[N]\times x[N] &\to \inf,\\ D[M_1, N]\times x[N] &\geqslant B[M_1],\\ D[M_2, N]\times x[N] &= B[M_2],\\ -x[N] &\geqslant -v[N],\\ x[N] &\geqslant w[N]. \end{aligned}$$

Переходя к обозначениям MATLAB, имеем:

$$\begin{gathered} \mathtt{f} = c[N],\ \ \mathtt{A} = -D[M_1, N],\ \ \mathtt{b} = -B[M_1],\\ \mathtt{Aeq} = D[M_2, N],\ \ \mathtt{beq} = B[M_2],\\ \mathtt{ub} = v[N],\ \ \mathtt{lb} = w[N]. \end{gathered}$$

Пусть $N = 1 : n$, $M_1 = 1 : s$, $M_2 = s + 1 : t$.

Вначале рассмотрим случай, когда $w[N] = \mathbb{O}$ или $w[N]$ не задан. Переменных в двойственной задаче будет $t+n$. Двойственные переменные содержатся в параметре `lambda`. Их можно найти по формулам

$$\begin{aligned} u_i^* &= \mathtt{lambda.ineqlin}(i), && i = 1, \ldots, s,\\ u_i^* &= \mathtt{lambda.eqlin}(i - s), && i = s + 1, \ldots, t,\\ u_i^* &= \mathtt{lambda.upper}(i - t), && i = t + 1, \ldots, t + n. \end{aligned} \tag{5}$$

Заметим, что

1) если $M_1 = \varnothing$, то считаем $s = 0$ и не используем первое соотношение из (5) (массив `lambda.ineqlin` пустой);

2) если $M_2 = \varnothing$, то считаем $t = s$ и не используем второе соотношение из (5) (массив `lambda.eqlin` пустой);

3) если не задан вектор верхних границ $v[N]$, то третье соотношение (5) не используется (массив `lambda.upper` нулевой).

Теперь предположим, что $w[N] \neq \mathbb{0}$. Пусть

$$w[k_i] \neq 0, \quad i = 1, \ldots, l, \quad l \leqslant n.$$

Тем самым выделяются индексы знаковых ограничений $N \setminus \{k_1, \ldots, k_l\}$. Переменных в двойственной задаче будет $t + n + l$. Их можно найти по формулам

$$\begin{aligned} u_i^* &= \texttt{lambda.ineqlin}(i), && i = 1, \ldots, s, \\ u_i^* &= \texttt{lambda.eqlin}(i - s), && i = s + 1, \ldots, t, \\ u_i^* &= \texttt{lambda.upper}(i - t), && i = t + 1, \ldots, t + n, \\ u_i^* &= \texttt{lambda.lower}(k_{i-t-n}), && i = t + n + 1, \ldots, t + n + l. \end{aligned} \tag{6}$$

Если не задан вектор верхних границ $v[N]$, то третье соотношение (6) не используется (массив `lambda.upper` нулевой), а четвёртое принимает вид

$$u_i^* = \texttt{lambda.lower}(k_{i-t}), \quad i = t + 1, \ldots, t + l. \tag{7}$$

ПРИМЕР 5. Рассмотрим задачу линейного программирования

$$\begin{aligned} f(x) = x_1 + x_2 &\to \inf, \\ x_1 - x_2 &\geqslant 2, \\ x_1 - 2x_2 &\geqslant 1, \\ -x_1 &\geqslant -4, \\ -x_2 &\geqslant -4, \\ x_1 \geqslant 0,\ x_2 &\geqslant 0. \end{aligned} \tag{8}$$

Составим двойственную задачу:

$$\begin{aligned} g(u) = 2u_1 + u_2 - 4u_3 - 4u_4 &\to \sup, \\ u_1 + u_2 - u_3 &\leqslant 1, \\ -u_1 - 2u_2 - u_4 &\leqslant 1, \\ u_1, u_2, u_3, u_4 &\geqslant 0. \end{aligned} \tag{9}$$

Решим задачу (8) в среде MATLAB. Программа будет выглядеть следующим образом:

```
clear all
close all
clc
```

```
C = [1 1];
D = [1 -1; 1 -2];
B = [2 1];
lb = zeros(2,1);
ub = [4 4];
f = C;
A = -D;
b = -B;
options = optimset('LargeScale','off','Simplex','on');
[x,fval,exitflag,output,lambda] =
linprog(f,A,b,[],[],lb,ub,[],options);
x
fval
lambda = structfun(@(t)(t.'),lambda,'UniformOutput',false)
% транспонируется содержимое lambda для лучшего отображения на
    экране консоли
```

На выходе получим оптимальный план $x_1^* = 2$, $x_2^* = 0$, минимальное значение целевой функции $f(x^*) = 2$ и `lambda`. Параметр `lambda` будет состоять из массивов

```
lambda =
ineqlin: [1 0]
eqlin: [1x0 double]
upper: [0 0]
lower: [0 2]
```

(Запись `eqlin: [1x0 double]` означает, что `eqlin` — пустой массив.)

Заметим, что в задаче (8) $t = s$ и $w[N] = \mathbb{O}$. По формулам (5)

$$u_1^* = 1,\ u_2^* = 0,\ u_3^* = 0,\ u_4^* = 0.$$

Легко проверить, что это план двойственной задачи (9), удовлетворяющий условиям дополнительности и соотношению двойственности $f(x^*) = 2 = g(u^*)$.

ПРИМЕР 6. Рассмотрим задачу линейного программирования

$$\begin{aligned} f(x) = {} & x_1 + x_2 \to \inf, \\ & x_1 - x_2 \geqslant 2, \\ & x_1 - 2x_2 \geqslant 1, \\ & x_1 \geqslant 0, \\ & x_2 \geqslant -1. \end{aligned} \tag{10}$$

Составим двойственную задачу:

$$\begin{gathered} g(u) = 2u_1 + u_2 - u_3 \to \sup, \\ u_1 + u_2 + u_3 \leqslant 1, \\ -u_1 - 2u_2 - u_3 = 1, \\ u_1, u_2, u_3 \geqslant 0. \end{gathered} \tag{11}$$

Решим задачу (10) в среде MATLAB. Для этого в программе из предыдущего примера достаточно заменить `lb = zeros(2,1)` на `lb = [0 -1]` и в списке аргументов функции `linprog` вместо `ub` поставить `[]`. На выходе получим оптимальный план $x_1^* = 1$, $x_2^* = -1$, минимальное значение целевой функции $f(x^*) = 0$ и `lambda`. Параметр `lambda` будет состоять из массивов

```
lambda =
ineqlin: [1 0]
eqlin: [1x0 double]
upper: [0 0]
lower: [0 2]
```

Заметим, что в задаче (10) отсутствует $v[N]$ и $w[1] = 0$, $w[2] \neq 0$ (то есть $l = 1$, $k_1 = 2$). По формулам (6) и (7)

$$u_1^* = 1,\ u_2^* = 0,\ u_3^* = 2.$$

Легко проверить, что это план двойственной задачи (11), удовлетворяющий условиям дополнительности и соотношению двойственности $f(x^*) = 0 = g(u^*)$.

ГЛАВА 2. КВАДРАТИЧНЫЕ ЗАДАЧИ

ТЕОРЕМА СУЩЕСТВОВАНИЯ РЕШЕНИЯ ДЛЯ ЗАДАЧИ КВАДРАТИЧНОГО ПРОГРАММИРОВАНИЯ*

В. Н. Малозёмов

1°. Рассмотрим задачу квадратичного программирования

$$\begin{gathered} Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf, \\ A[M,N] \times x[N] \geqslant b[M], \end{gathered} \tag{1}$$

где $D = D[N,N]$ — симметричная матрица. Множество планов (векторов $x = x[N]$, удовлетворяющих ограничениям задачи (1)) обозначим Ω. План x_* называется *оптимальным*, если

$$Q(x_*) = \inf_{x\in\Omega} Q(x).$$

ТЕОРЕМА. *Предположим, что матрица D неотрицательно определена на $\mathbb{R}^N$, множество Ω непусто и целевая функция $Q(x)$ ограничена снизу на Ω. Тогда у задачи* (1) *существует оптимальный план. Он может быть получен путём решения конечного числа систем линейных уравнений.*

В таком виде теорема существования решения для задачи квадратичного программирования была сформулирована М. К. Гавуриным. В статье [1] и книге [2, с. 111–112] приведено краткое доказательство этой теоремы. Здесь будет дано развёрнутое доказательство.

2°. Отметим некоторые свойства квадратичной функции.

При всех x, h из $\mathbb{R}^N$ и $t \in \mathbb{R}$ справедливо разложение

$$Q(x+th) = Q(x) + t\langle Dx + c, h\rangle + \tfrac{1}{2}\,t^2\langle Dh, h\rangle. \tag{2}$$

В частности,

$$Q(x+h) - Q(x) = \langle Dx + c, h\rangle + \tfrac{1}{2}\langle Dh, h\rangle. \tag{3}$$

Если из (3) найти выражение для $\langle Dx + c, h\rangle$ и подставить его в (2), то придём к формуле

$$Q(x+th) = Q(x) + t\big[Q(x+h) - Q(x)\big] - \tfrac{1}{2}\,t(1-t)\langle Dh, h\rangle. \tag{4}$$

*Семинар «DHA & CAGD». Избранные доклады. 22 января 2011 г.

3°. Доказательству теоремы предпошлём два вспомогательных утверждения.

ЛЕММА 1. *Если квадратичная функция $Q(x)$ ограничена снизу на $\mathbb{R}^N$, то существует точка, в которой $Q(x)$ достигает наименьшего на $\mathbb{R}^N$ значения.*

Доказательство. Обозначим $\mu_0 = \inf_{x\in\mathbb{R}^N} Q(x)$. Согласно (2) при фиксированном x имеем

$$Q(x) + t\langle Dx + c, h\rangle + \tfrac{1}{2}\, t^2\langle Dh, h\rangle \geqslant \mu_0 \quad \forall\, t \in \mathbb{R}.$$

Отсюда следует, что $\langle Dh, h\rangle \geqslant 0$ при всех $h \in \mathbb{R}^N$. В силу (3)

$$Q(x + h) - Q(x) \geqslant \langle Dx + c, h\rangle \quad \forall\, x,\, h \in \mathbb{R}^N.$$

Очевидно, что точка x_*, в которой $Dx_* + c = \mathbb{O}$, является точкой минимума функции $Q(x)$ на $\mathbb{R}^N$.

Покажем, что в условиях леммы система линейных уравнений

$$Dx = -c \tag{5}$$

имеет решение. Допустим противное. Тогда найдется вектор $h_0 \in \mathbb{R}^N$ со свойствами $Dh_0 = \mathbb{O}$, $\langle c, h_0\rangle \neq 0$. При фиксированном $x = x_0$ и всех $t \in \mathbb{R}$ согласно (2) имеем

$$Q(x_0 + th_0) - Q(x_0) = t\langle Dx_0 + c, h_0\rangle = t\langle c, h_0\rangle.$$

Это противоречит ограниченности снизу функции $Q(x)$ на $\mathbb{R}^N$.

Решение системы (5) и будет точкой минимума $Q(x)$ на $\mathbb{R}^N$. □

Перейдём к более сложной задаче

$$\begin{gathered} Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf, \\ A[M, N] \times x[N] = b[M]. \end{gathered} \tag{6}$$

Множество планов этой задачи обозначим ω.

ЛЕММА 2. *Если $\omega \neq \varnothing$ и квадратичная функция $Q(x)$ ограничена снизу на ω, то задача* (6) *имеет решение.*

Доказательство. Обозначим

$$\omega_0 = \{h \in \mathbb{R}^N \mid Ah = \mathbb{O}\}.$$

Возьмём $x_0 \in \omega$. Тогда $x_0 + th \in \omega$ при всех $h \in \omega_0$ и $t \in \mathbb{R}$.

Функция $Q(x_0+th)$ как функция от t ограничена снизу. Отсюда и из (2) следует, что $\langle Dh,h\rangle \geqslant 0$ при всех $h\in\omega_0$. Согласно (3)

$$Q(x+h)-Q(x)\geqslant\langle Dx+c,h\rangle \quad \text{для всех } x\in\omega \text{ и } h\in\omega_0.$$

Покажем, что существует пара $\{x_*,u_*\}$, где $x_*\in\omega$, такая, что

$$Dx_*+c=A^Tu_*.$$

В этом случае x_* будет точкой минимума функции $Q(x)$ на ω.

Собственно, нужно установить, что система линейных уравнений

$$\begin{aligned} Dx-A^Tu &= -c\\ -Ax &= -b \end{aligned} \tag{7}$$

имеет решение. Допустим противное. Тогда найдутся векторы $h_0\in\mathbb{R}^N$ и $v_0\in\in\mathbb{R}^M$ со свойствами

$$Dh_0-A^Tv_0=\mathbb{O},\qquad Ah_0=\mathbb{O},\qquad \langle c,h_0\rangle+\langle b,v_0\rangle\neq 0.$$

Согласно (2) при всех $t\in\mathbb{R}$ имеем

$$\begin{aligned} Q(x_0+th_0)-Q(x_0) &= t\big[\langle Dx_0,h_0\rangle+\langle c,h_0\rangle\big]+\tfrac12\,t^2\langle A^Tv_0,h_0\rangle=\\ &= t\big[\langle c,h_0\rangle+\langle x_0,Dh_0\rangle\big]=t\big[\langle c,h_0\rangle+\langle b,v_0\rangle\big]. \end{aligned}$$

Это противоречит ограниченности снизу функции $Q(x)$ на ω.

Таким образом, задача минимизации $Q(x)$ на ω в условиях леммы сводится к решению системы (7). □

4°. Обратимся к выпуклому многогранному множеству Ω, определяемому векторным неравенством

$$A[M,N]\times x[N]\geqslant b[M].$$

Нам потребуются некоторые структурные свойства Ω.

Обозначим $\Delta(x)=Ax-b$. Каждому индексному множеству $I\subset M$ сопоставим подмножество множества Ω вида

$$\Omega(I)=\big\{x\in\mathbb{R}^N\;\big|\;\Delta(x)[i]=0 \text{ при } i\in I;\quad \Delta(x)[i]>0 \text{ при } i\in M\setminus I\big\},$$

называемое *гранью* Ω (случай $I=\varnothing$ не исключается). Пусть Γ — множество тех I, которые порождают *непустые* грани $\Omega(I)$. Ясно, что $\Omega(I)\cap\Omega(I')=\varnothing$ при $I\neq I'$ и

$$\Omega=\bigcup_{I\subset\Gamma}\Omega(I). \tag{8}$$

Зафиксируем $I_0 \in \Gamma$. Множества

$$\partial\Omega(I_0) = \bigcup_{I \supset I_0,\ I \neq I_0} \Omega(I),$$
$$\omega(I_0) = \big\{x \in \mathbb{R}^N \mid \Delta(x)[i] = 0 \text{ при } i \in I_0\big\}$$

называются соответственно *относительной границей* и *аффинной оболочкой* грани $\Omega(I_0)$. Если $\varnothing \in \Gamma$, то по определению $\omega(\varnothing) = \mathbb{R}^N$.

ПРИМЕР. Рассмотрим на плоскости множество Ω, определяемое неравенствами $x_1 \geqslant 0,\ -x_1 \geqslant -1,\ x_2 \geqslant 0$ (см. рис.).

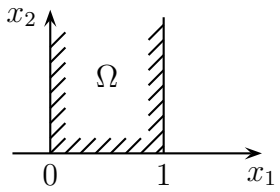


Рис.

Это множество имеет восемь граней

$$\Omega_0 = \{x \mid 0 < x_1 < 1,\ x_2 > 0\}, \qquad \Omega_{1,2} = \{x \mid x_1 = 0,\ x_1 = 1,\ x_2 > 0\},$$
$$\Omega_1 = \{x \mid x_1 = 0,\ x_2 > 0\}, \qquad \Omega_{1,3} = \{x \mid x_1 = 0,\ x_2 = 0\},$$
$$\Omega_2 = \{x \mid x_1 = 1,\ x_2 > 0\}, \qquad \Omega_{2,3} = \{x \mid x_1 = 1,\ x_2 = 0\},$$
$$\Omega_3 = \{x \mid 0 < x_1 < 1,\ x_2 = 0\}, \qquad \Omega_{1,2,3} = \{x \mid x_1 = 0,\ x_1 = 1,\ x_2 = 0\},$$

две из которых ($\Omega_{1,2}$ и $\Omega_{1,2,3}$) пустые.

Относительная граница грани Ω_3 состоит из двух точек $(0,0)$ и $(1,0)$ (грани $\Omega_{1,3}$ и $\Omega_{2,3}$), а аффинной оболочкой является ось $x_2 = 0$.

ЛЕММА 3. *Справедливо равенство*

$$\omega(I_0) = \Omega(I_0) \cup \partial\Omega(I_0) \cup \big(\omega(I_0) \setminus \Omega\big). \tag{9}$$

Доказательство. Обозначим множество из правой части равенства (9) через G. Проверим включение $\omega(I_0) \subset G$.

Пусть $x_0 \in \omega(I_0)$, так что $\Delta(x_0)[i] = 0$ при $i \in I_0$. Если при этом $\Delta(x_0)[i] > 0$ при $i \in M \setminus I_0$, то $x_0 \in \Omega(I_0)$. Если $\Delta(x_0)[i] \geqslant 0$ при $i \in M \setminus I_0$, причём хотя бы один раз неравенство выполняется как равенство, то $x_0 \in \partial\Omega(I_0)$. Наконец, если $\Delta(x_0)[i] < 0$ при некотором $i \in M \setminus I_0$, то $x_0 \in \omega(I_0) \setminus \Omega$. Получили, что $x_0 \in G$. Включение $\omega(I_0) \subset G$ установлено.

Обратное включение $G \subset \omega(I_0)$ очевидно. □

ЛЕММА 4 (об относительной границе)**.** *Пусть* $x_0 \in \Omega(I_0)$, $x_1 \in \omega(I_0) \setminus \Omega$. *Тогда на интервале* (x_0, x_1) *найдётся точка* $\hat{x}$, *принадлежащая* $\partial\Omega(I_0)$.

Доказательство. Обозначим $x(t) = tx_1 + (1-t)x_0$. Согласно определению $\Delta(x)$,

$$\Delta\big(x(t)\big) = t\Delta(x_1) + (1-t)\Delta(x_0). \tag{10}$$

В силу выбора x_0 и x_1 при всех $t \in \mathbb{R}$ имеем

$$\Delta\big(x(t)\big)[i] = 0, \quad i \in I_0.$$

Покажем, что существует $\hat{t} \in (0,1)$, на котором

$$\Delta\big(x(\hat{t})\big)[i] \geqslant 0, \quad i \in M \setminus I_0,$$

причём хотя бы один раз неравенство выполняется как равенство. Соответствующая точка $\hat{x} = x(\hat{t})$ будет требуемой.

На множестве индексов $\big\{i \in M \setminus I_0 \mid \Delta(x_1)[i] \geqslant 0\big\}$ согласно (10) при всех $t \in (0,1)$ выполняется строгое неравенство $\Delta\big(x(t)\big)[i] > 0$. Вместе с тем, условие $x_1 \in \omega(I_0) \setminus \Omega$ гарантирует, что множество $J_1 = \big\{i \in M \setminus I_0 \mid \Delta(x_1)[i] < 0\big\}$ непусто. Для $i \in J_1$ неравенство

$$\Delta\big(x(t)\big)[i] := \Delta(x_0)[i] - t\big(\Delta(x_0)[i] - \Delta(x_1)[i]\big) \geqslant 0$$

равносильно следующему

$$t \leqslant \frac{\Delta(x_0)[i]}{\Delta(x_0)[i] - \Delta(x_1)[i]}.$$

Положим

$$\hat{t} = \min_{i \in J_1} \frac{\Delta(x_0)[i]}{\Delta(x_0)[i] - \Delta(x_1)[i]}. \tag{11}$$

Ясно, что $\hat{t} \in (0,1)$. Точка $\hat{x} = x(\hat{t})$ — требуемая. Для неё $\Delta(\hat{x})[i] \geqslant 0$ при всех $i \in M \setminus I_0$ и $\Delta(\hat{x})[i] = 0$ на индексе $i \in J_1$, на котором достигается минимум в (11).

Лемма доказана. □

СЛЕДСТВИЕ. *Если* $\Omega(I_0) \neq \varnothing$, *но* $\partial\Omega(I_0) = \varnothing$, *то и* $\omega(I_0) \setminus \Omega = \varnothing$.

В этом случае, согласно (9), $\Omega(I_0) = \omega(I_0)$.

Замечание. Если $\Omega(I_0) \neq \varnothing$ и $\partial\Omega(I_0) \neq \varnothing$, то и $\omega(I_0) \setminus \Omega \neq \varnothing$.

Действительно, возьмём $x_0 \in \Omega(I_0)$ и $y_0 \in \partial\Omega(I_0)$, так что $\Delta(y_0)[i_1] = 0$ при некотором $i_1 \in M \setminus I_0$. Рассмотрим прямую $y(t) = ty_0 + (1-t)x_0$, проходящую через точки x_0 и y_0. Ясно, что $y(t) \in \omega(I_0)$ при всех $t \in \mathbb{R}$. Вместе с тем,

$$\Delta\big(y(t)\big)[i_1] = (1-t)\Delta(x_0)[i_1] < 0 \quad \text{при } t > 1.$$

Следовательно $y(t) \in \omega(I_0) \setminus \Omega$ при $t > 1$.

5°. Переходим к доказательству теоремы. Обозначим

$$\mu = \inf_{x\in\Omega} Q(x).$$

Согласно (8)

$$\mu = \min_{I\in\Gamma} \inf_{x\in\Omega(I)} Q(x). \tag{12}$$

Среди тех I, на которых достигается минимум в (12), выберем I с *наибольшим* количеством элементов. Обозначим его I_0. Имеем

$$\inf_{x\in\Omega(I_0)} Q(x) = \mu. \tag{13}$$

Предположим вначале, что $\partial\Omega(I_0) = \varnothing$. В этом случае, по следствию из леммы 4, $\Omega(I_0) = \omega(I_0)$, так что

$$\inf_{x\in\omega(I_0)} Q(x) = \mu.$$

На основании леммы 2 (или леммы 1 при $I_0 = \varnothing$) заключаем, что существует точка $x_* \in \omega(I_0)$, в которой $Q(x_*) = \mu$. По определению μ точка x_* — оптимальный план задачи (1).

Пусть $\partial\Omega(I_0) \neq \varnothing$. Обозначим

$$\mu' = \inf_{x\in\partial\Omega(I_0)} Q(x). \tag{14}$$

По определению относительной границы

$$\mu' = \min_{I\supset I_0,\, I\neq I_0} \inf_{x\in\Omega(I)} Q(x).$$

Очевидно, что $\mu' \geqslant \mu$. На самом деле, $\mu' > \mu$, ибо иначе (при $\mu' = \mu$) нашлась бы грань $\Omega(I_1)$ с $|I_1| > |I_0|$, на которой

$$\inf_{x\in\Omega(I_1)} Q(x) = \mu' = \mu.$$

Но это противоречит выбору I_0.

Итак, $\mu' > \mu$. Отсюда и из (13) следует, в частности, что существует точка $x_0 \in \Omega(I_0)$, в которой

$$Q(x_0) < \mu'. \tag{15}$$

Рассмотрим аффинную оболочку $\omega_0 = \omega(I_0)$ грани $\Omega(I_0)$. Покажем, что

$$\inf_{x\in\omega_0} Q(x) = \mu. \tag{16}$$

Предварительно установим неравенство

$$Q(x) > \mu' \quad \forall\, x \in \omega_0 \setminus \Omega. \tag{17}$$

Зафиксируем $x_1 \in \omega_0 \setminus \Omega$ (по замечанию к лемме 4 множество $\omega_0 \setminus \Omega$ непусто). Согласно лемме 4 на интервале (x_0, x_1) существует точка $\hat{x} = \hat{t}x_1 + (1-\hat{t})x_0$, $\hat{t} \in (0,1)$, принадлежащая $\partial\Omega(I_0)$. Воспользуемся формулой (4) и неотрицательной определённостью матрицы D на $\mathbb{R}^N$. Запишем

$$Q(\hat{x}) = Q\big(x_0 + \hat{t}(x_1 - x_0)\big) \leqslant Q(x_0) + \hat{t}[Q(x_1) - Q(x_0)] = \hat{t}Q(x_1) + (1-\hat{t})Q(x_0).$$

На основании определения μ' и (15) получим

$$\mu' \leqslant Q(\hat{x}) < \hat{t}Q(x_1) + (1-\hat{t})\mu'.$$

Отсюда следует, что $Q(x_1) > \mu'$. Неравенство (17) установлено. Оно гарантирует, что

$$\inf_{x \in \omega_0 \setminus \Omega} Q(x) \geqslant \mu'. \tag{18}$$

В силу леммы 3

$$\inf_{x \in \omega_0} Q(x) = \min\Big\{ \inf_{x \in \Omega(I_0)} Q(x),\ \inf_{x \in \partial\Omega(I_0)} Q(x),\ \inf_{x \in \omega_0 \setminus \Omega} Q(x) \Big\}.$$

Соотношения (13), (14), (18) и неравенство $\mu' > \mu$ приводят к (16).

По лемме 2 (или лемме 1 при $I_0 = \varnothing$) существует точка $x_* \in \omega_0$, в которой $Q(x_*) = \mu$. Соотношения (14) и (18) указывают, что необходимо $x_* \in \Omega(I_0)$. Значит, x_* — оптимальный план задачи (1).

Теорема доказана. □

6°. Теорема имеет конструктивный характер. Она определяет путь решения задачи (1). Нужно минимизировать квадратичную функцию $Q(x)$ на аффинных множествах вида

$$\omega(I) = \big\{ x \in \mathbb{R}^N \mid A[I,N] \times x[N] = b[I] \big\}$$

при всех $I \subset M$ (не забывая $I = \varnothing$). Это сводится к решению систем линейных уравнений вида (7) (или (5) при $I = \varnothing$). Нас интересуют решения, принадлежащие Ω. То из них, на котором $Q(x)$ принимает наименьшее значение, и будет оптимальным планом задачи (1).

7°. Вопрос о существовании решения у задачи квадратичного программирования рассматривался также в работах [3, 4]. Численным методам квадратичного программирования посвящена книга [5].

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Основы теории квадратичного программирования* // Вестник ЛГУ. 1980. № 1. С. 9–16.

2. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

3. Frank M., Wolfe P. *An algorithm for quadratic programming* // Naval Res. Logist. Quart. 1956. V. 3. No. 1-2. P. 95–110.

4. Eaves B. C. *On quadratic programming* // Manag. Sci. 1971. V. 17. No. 11. P. 698–711.

5. Даугавет В. А. *Численные методы квадратичного программирования.* СПб.: Изд-во СПбГУ, 2004. 128 с.

О МЕТОДЕ ПЕРЕБОРА ГРАНЕЙ В КВАДРАТИЧНОМ ПРОГРАММИРОВАНИИ*

В. Н. Малозёмов, Е. К. Чернэуцану

1°. Рассмотрим задачу квадратичного программирования

$$Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf_{x\in\Omega}, \tag{1}$$

где $D = D[N, N]$ — симметричная неотрицательно определённая матрица и Ω — выпуклое многогранное множество, определяемое системой линейных неравенств

$$A[M, N] \times x[N] \geqslant b[M].$$

В докладе [1] приведено подробное доказательство следующего утверждения: *для того чтобы задача* (1) *имела решение, необходимо и достаточно, чтобы множество планов* Ω *было непусто и чтобы целевая функция* $Q(x)$ *была ограничена снизу на* Ω.

Упомянутое доказательство конструктивно. В нём предлагается при различных $I \subset M$ (не исключая $I = \varnothing$) исследовать системы линейных уравнений

$$\begin{aligned} D[N, N] \times x[N] - A^T[N, I] \times u[I] &= -c[N], \\ -A[I, N] \times x[N] \qquad\qquad\quad &= -b[I]. \end{aligned} \tag{2}$$

Назовём множество $I \subset M$ *подходящим*, если система (2) имеет решение (x', u') и $x' \in \Omega$. Соответствующее решение (x', u') назовём *подходящей парой*. Среди подходящих множеств I следует выбрать то, которому соответствует подходящая пара (x', u') с наименьшим значением $Q(x')$. Найденное x' и будет оптимальным планом задачи (1).

Остаётся организовать перебор конечного числа подходящих множеств I так, чтобы соответствующие значения $Q(x')$ строго убывали. Получим конечный метод решения задачи (1), который и называется *методом перебора граней*.

*Семинар «DHA & CAGD». Избранные доклады. 10 сентября 2011 г.

2°. Описание метода перебора граней будем проводить в предположении, что выполнены *условия невырожденности*. Они включают два пункта:

(a) матрица D положительно определена на $\mathbb{R}^N$;

(b) при каждом $x \in \Omega$ строки матрицы $A\big[I(x), N\big]$, где

$$I(x) = \big\{i \in M \mid A[i, N] \times x[N] = b[i]\big\},$$

линейно независимы.

Выясним сначала, что дают условия невырожденности. Условие (a) гарантирует ограниченность снизу функции $Q(x)$ на $\mathbb{R}^N$. Точнее, выполняется неравенство

$$Q(x) \geqslant -\tfrac{1}{2}\langle D^{-1}c, c\rangle \quad \forall\, x \in \mathbb{R}^N. \tag{3}$$

Действительно, в силу положительной определённости матрицы D существует обратная матрица D^{-1}, которая также является положительно определённой. Точка минимума x_* строго выпуклой на $\mathbb{R}^N$ функции $Q(x)$ определяется из уравнения $Q'(x) = \mathbb{O}$ или $Dx + c = \mathbb{O}$. Отсюда следует, что $x_* = -D^{-1}c$ и

$$Q(x_*) = \tfrac{1}{2}\langle c, D^{-1}c\rangle - \langle c, D^{-1}c\rangle = -\tfrac{1}{2}\langle D^{-1}c, c\rangle.$$

Неравенство $Q(x) \geqslant Q(x_*)$ равносильно (3).

Таким образом, при выполнении условия (a) задача (1) имеет решение, только если множество её планов Ω непусто. Более того, в силу строгой выпуклости функции $Q(x)$ на $\mathbb{R}^N$ решение задачи (1) единственно.

Теперь выясним роль условия (b). Возьмём точку $x_0 \in \Omega$ с множеством индексов активных ограничений $I = I(x_0)$. По условию (b) строки матрицы $A[I, N]$ линейно независимы. Нетрудно проверить, что матрица

$$F[I, I] = A[I, N] \times D[N, N] \times A^T[N, I]$$

симметрична и положительно определена. Симметричность очевидна. Положительная определённость проверяется так: при $u \neq \mathbb{O}$ имеем

$$\langle Fu, u\rangle = \langle ADA^T u, u\rangle = \langle DA^T u, A^T u\rangle > 0,$$

поскольку $A^T u \neq \mathbb{O}$ в силу линейной независимости строк матрицы $A[I, N]$. Как симметричная и положительно определённая, матрица F имеет обратную матрицу F^{-1}, которая также является симметричной и положительно определённой.

Рассмотрим симметричную матрицу G системы (2):

$$G = \begin{pmatrix} D[N, N] & -A^T[N, I] \\ -A[I, N] & \mathbb{O}[I, I] \end{pmatrix}.$$

При выполнении условий невырожденности матрица G имеет обратную матрицу G^{-1}, при этом [2]

$$G^{-1} = \begin{pmatrix} E[N,N] & D^{-1}[N,N] \times A^T[N,I] \\ \mathbb{O}[I,N] & E[I,I] \end{pmatrix} \times \begin{pmatrix} D^{-1}[N,N] & \mathbb{O}[N,I] \\ \mathbb{O}[I,N] & -F^{-1}[I,I] \end{pmatrix} \times$$
$$\times \begin{pmatrix} E[N,N] & \mathbb{O}[I,N] \\ A[I,N] \times D^{-1}[N,N] & E[I,I] \end{pmatrix}.$$

То, что $G^{-1}G = E$, проверяется непосредственно.

Подчеркнём, что матрица G связана с планом x_0.

3°. Зададимся таким вопросом: когда подходящее индексное множество $I \subset M$ порождает решение задачи (1)?

ПРЕДЛОЖЕНИЕ 1. *Пусть $I \subset M$ — подходящее индексное множество и (x', u') — соответствующая подходящая пара. Тогда x' — оптимальный план задачи* (1)*, если $u'[I] \geqslant \mathbb{O}$.*

Доказательство. Пара (x', u') удовлетворяет соотношениям

$$D[N,N] \times x'[N] - A^T[N,I] \times u'[I] = -c[N], \tag{4}$$
$$-A[I,N] \times x'[N] = -b[I], \tag{5}$$
$$A[M \setminus I, N] \times x'[N] \geqslant b[M \setminus I], \tag{6}$$
$$u'[I] \geqslant \mathbb{O}. \tag{7}$$

Расширим множество I, включив в него индексы из $M \setminus I$, на которых неравенство (6) выполняется как равенство (если такие индексы найдутся). На вновь введённых индексах i положим $u'[i] = 0$. Пара (x', u') с преобразованным u' удовлетворяет соотношениям (4), (5), (7) и (6), в котором знак "$\geqslant$" следует заменить на знак строгого неравенства "$>$". Это значит, что $x' \in \Omega$, $I = I(x')$ и

$$Q'(x') = A^T\big[N, I(x')\big] \times u'\big[I(x')\big],$$
$$u'\big[I(x')\big] \geqslant \mathbb{O}.$$

По теореме Куна-Таккера [3] вектор x' является решением задачи (1). (Для этого достаточно, чтобы матрица D была неотрицательно определённой.)

Предложение доказано. □

4°. Нам потребуется одно вспомогательное утверждение.

ПРЕДЛОЖЕНИЕ 2. *Пусть* (x', u') *— решение системы* (2) *и* x *— произвольный вектор, удовлетворяющий уравнению*

$$A[I, N] \times x[N] = b[I].$$

Тогда при всех вещественных t *справедливо равенство*

$$Q(x + ts) = Q(x) - t(1 - \tfrac{t}{2})\langle Ds, s\rangle, \tag{8}$$

где $s = x' - x$.

Доказательство. В (4) и (5) подставим $x' = x + s$. Получим

$$D[N, N] \times s[N] - A^T[N, I] \times u'[I] = -Q'(x), \tag{9}$$
$$A[I, N] \times s[N] = \mathbb{O}. \tag{10}$$

Умножим (9) слева на $s[N]$. С учётом (10) придём к равенству

$$\langle Ds, s\rangle = -\langle Q'(x), s\rangle. \tag{11}$$

Теперь запишем разложение квадратичной функции

$$Q(x + ts) = Q(x) + t\langle Q'(x), s\rangle + \tfrac{1}{2}t^2\langle Ds, s\rangle. \tag{12}$$

Формула (8) очевидным образом следует из (12) и (11). □

При $t = 1$ равенство (8) принимает вид

$$Q(x') = Q(x) - \tfrac{1}{2}\langle Ds, s\rangle.$$

5°. Построим начальное подходящее множество $I_0 \subset M$. С этой целью возьмём произвольный план $x \in \Omega$ с множеством индексов активных ограничений $I = I(x)$ и решим систему линейных уравнений (2). В силу условия невырожденности система (2) имеет решение (x', u'). Если $x' \in \Omega$, то по определению I — подходящее множество. Положим $I_0 = I$. Соответствующей подходящей парой будет $(x_0, u_0) = (x', u')$.

Пусть $x' \notin \Omega$, то есть при некотором $i \in M \setminus I$ выполняется неравенство

$$A[i, N] \times x'[N] < b[i]. \tag{13}$$

Множество всех таких индексов обозначим J'. Докажем следующее утверждение (оно соответствует лемме об относительной границе [1]): *на интервале* (x, x') *найдётся точка* $\hat{x} \in \Omega$, *такая, что* $\big|I(\hat{x})\big| > \big|I(x)\big|$ *и*

$$Q(\hat{x}) < Q(x). \tag{14}$$

Для доказательства введём обозначения

$$\Delta(x) = Ax - b, \quad s = x' - x.$$

Ясно, что $s \neq \mathbb{O}$. Рассмотрим интервал (x, x'). Его параметрическое представление имеет вид

$$x(t) = tx' + (1-t)x = x + ts, \quad t \in (0,1).$$

При этом

$$\Delta\big(x(t)\big) = tAx' + (1-t)Ax - b = t\Delta(x') + (1-t)\Delta(x).$$

В силу выбора x и x' при всех $t \in (0,1)$ справедливо равенство

$$\Big(\Delta\big(x(t)\big)\Big)[i] = 0, \quad i \in I.$$

Далее, имеем

$$\big(\Delta(x)\big)[i] > 0,\ i \in M \setminus I; \quad \big(\Delta(x')\big)[i] \geqslant 0,\ i \in (M \setminus I) \setminus J',$$

поэтому при всех $t \in (0,1)$

$$\Big(\Delta\big(x(t)\big)\Big)[i] > 0, \quad i \in (M \setminus I) \setminus J'.$$

При $i \in J'$ неравенство

$$\Big(\Delta\big(x(t)\big)\Big)[i] := \big(\Delta(x)\big)[i] - t\Big(\big(\Delta(x)\big)[i] - \big(\Delta(x')\big)[i]\Big) \geqslant 0$$

равносильно следующему

$$t \leqslant \frac{\big(\Delta(x)\big)[i]}{\big(\Delta(x)\big)[i] - \big(\Delta(x')\big)[i]}.$$

У дроби из правой части этого неравенства числитель и знаменатель положительны (числитель — на основании того, что $J' \subset M \setminus I$, знаменатель — в силу (13)). Значение самой дроби лежит в интервале $(0,1)$. Положим

$$\hat{t} = \min_{i \in J'} \Big\{ \frac{\big(\Delta(x)\big)[i]}{\big(\Delta(x)\big)[i] - \big(\Delta(x')\big)[i]} \Big\}. \tag{15}$$

Ясно, что $\hat{t} \in (0,1)$. По построению точка $\hat{x} = x(\hat{t})$ принадлежит Ω. Соответствующее множество индексов активных ограничений $\hat{I} = I(\hat{x})$ содержит I и

те индексы из J', на которых достигается минимум в правой части (15). По крайней мере, $|\hat{I}| > |I|$.

Справедливость строгого неравенства (14) следует из (8) при $t = \hat{t}$.

Таким образом, от плана x при $x' \notin \Omega$ мы перешли к новому плану $\hat{x}$, такому, что $|I(\hat{x})| > |I(x)|$ и $Q(\hat{x}) < Q(x)$.

Теперь в качестве x возьмём $\hat{x}$ и повторим рассуждения. Тогда либо $x' \in \Omega$, что позволяет в качестве I_0 взять $I(x)$, либо $x' \notin \Omega$. Во втором случае от x мы перейдём к новому плану $\hat{x}$, у которого $|I(\hat{x})| > |I(x)|$ и $Q(\hat{x}) < Q(x)$. Далее процесс повторяется. В силу постоянного расширения множеств $I(\hat{x})$ и строгого убывания целевой функции $Q(x)$ описанная процедура конечна. За конечное число шагов мы придём к начальному подходящему множеству I_0 и соответствующей подходящей паре (x_0, u_0) с $x_0 \in \Omega$.

6°. Если $u_0[I_0] \geqslant \mathbb{O}$, то согласно предложению 1 вектор x_0 является решением задачи (1). Процесс закончен.

Предположим, что

$$u_0[i'] < 0 \quad \text{при некотором} \quad i' \in I_0. \tag{16}$$

Расширим при необходимости множество I_0, включив в него те индексы i из $M \setminus I_0$, на которых

$$A[i, N] \times x_0[N] = b[i].$$

Это действие обеспечивает равенство $I_0 = I(x_0)$. Положив $u_0[i] = 0$ на вновь введённых индексах i, сохраним соотношения

$$\begin{aligned} D[N, N] \times x_0[N] - A^T[N, I_0] \times u_0[I_0] &= -c[N], \\ -A[I_0, N] \times x_0[N] &= -b[I_0]. \end{aligned} \tag{17}$$

Введём индексное множество $I_0' = I_0 \setminus \{i'\}$. В силу условия невырожденности строки матрицы $A[I_0', N]$ линейно независимы. Решим задачу (2) при $I = I_0'$. Решение обозначим (x_0', u_0'). Таким образом,

$$\begin{aligned} D[N, N] \times x_0'[N] - A^T[N, I_0'] \times u_0'[I_0'] &= -c[N], \\ -A[I_0', N] \times x_0'[N] &= -b[I_0']. \end{aligned} \tag{18}$$

ПРЕДЛОЖЕНИЕ 3. *Вектор x_0' отличен от x_0. При этом $Q(x_0') < Q(x_0)$ и*

$$A[i', N] \times x_0'[N] > b[i']. \tag{19}$$

Доказательство. Допустив, что $x_0' = x_0$, на основании (17), (18) получим

$$A^T[N, I_0] \times u_0[I_0] = A^T[N, I_0'] \times u_0'[I_0']$$

или

$$A^T[N, I'_0] \times \big(u_0[I'_0] - u'_0[I'_0]\big) + A^T[N, i'] \times u_0[i'] = \mathbb{O}.$$

В силу (16) это противоречит линейной независимости строк матрицы $A[I_0,N]$. Значит $x'_0 \neq x_0$.

Обозначим $s_0 = x'_0 - x_0$. Формулы (17), (18) приводят к равенствам

$$D[N, N] \times s_0[N] - A^T[N, I'_0] \times \big(u_0[I'_0] - u'_0[I'_0]\big) + A^T[N, i'] \times u_0[i'] = \mathbb{O}, \quad (20)$$

$$A[I'_0, N] \times s_0[N] = \mathbb{O}. \quad (21)$$

Умножим (20) слева на s_0. На основании (21) получим

$$\langle Ds_0, s_0\rangle = -u_0[i'] \times \big(A[i', N] \times s_0[N]\big).$$

Отсюда, с учётом (16) и условия $s_0 \neq \mathbb{O}$, следует неравенство

$$A[i', N] \times s_0[N] > 0,$$

равносильное (19).

Как отмечалось после доказательства предложения 2,

$$Q(x'_0) = Q(x_0) - \tfrac{1}{2}\langle Ds_0, s_0\rangle.$$

В частности, $Q(x'_0) < Q(x_0)$.

Справедливость всех утверждений предложения 3 установлена. □

7°. Вернёмся к множеству I'_0 и соответствующему решению (x'_0, u'_0) системы (2) при $I_0 = I'_0$. Если $x'_0 \in \Omega$, то I'_0 — очередное подходящее множество. Обозначим $I_1 = I'_0$, $x_1 = x'_0$, $u_1 = u'_0$. Как доказано в предложении 3, $Q(x_1) < Q(x_0)$.

Предположим, что $x'_0 \notin \Omega$. В этом случае существует индекс $i \in M \setminus I'_0$, на котором

$$A[i, N] \times x'_0[N] < b[i].$$

Множество всех таких индексов обозначим J'_0. Отметим, что $i' \in M \setminus I'_0$. Более того, согласно (19)

$$i' \in (M \setminus I'_0) \setminus J'_0.$$

На интервале (x_0, x'_0) найдём точку $\hat{x}_0 \in \Omega$, такую, что $Q(\hat{x}_0) < Q(x_0)$. Для этого так же, как в п. 5°, запишем параметрическое представление интервала (x_0, x'_0)

$$x_0(t) = tx'_0 + (1-t)x_0, \quad t \in (0, 1),$$

и воспользуемся формулой

$$\Delta\big(x_0(t)\big) = t\Delta(x'_0) + (1-t)\Delta(x_0).$$

При всех $t \in (0,1)$ имеем

$$\Big(\Delta\big(x_0(t)\big)\Big)[i] = 0, \quad i \in I_0'. \tag{22}$$

Далее, возьмём $i \in (M \setminus I_0') \setminus J_0'$. Если $i = i'$, то в силу (19) при всех $t \in (0,1)$ получим

$$\Big(\Delta\big(x_0(t)\big)\Big)[i'] = t\big(\Delta(x_0')\big)[i'] > 0.$$

Пусть $i \neq i'$. Тогда $i \in M \setminus I_0$ и $i \notin J_0'$. Эти условия обеспечивают неравенства

$$\big(\Delta(x_0)\big)[i] > 0, \quad \big(\Delta(x_0')\big)[i] \geqslant 0.$$

Отсюда следует, что $\Big(\Delta\big(x_0(t)\big)\Big)[i] > 0$ при всех $t \in (0,1)$. Таким образом, при всех $t \in (0,1)$

$$\Big(\Delta\big(x_0(t)\big)\Big)[i] > 0, \quad i \in (M \setminus I_0') \setminus J_0'. \tag{23}$$

Остаётся разобраться с индексами из J_0'. Для них неравенство

$$\Big(\Delta\big(x_0(t)\big)\Big)[i] := \big(\Delta(x_0)\big)[i] - t\Big(\big(\Delta(x_0)\big)[i] - \big(\Delta(x_0')\big)[i]\Big) \geqslant 0$$

равносильно следующему

$$t \leqslant \frac{\big(\Delta(x_0)\big)[i]}{\big(\Delta(x_0)\big)[i] - \big(\Delta(x_0')\big)[i]}. \tag{24}$$

Индекс i из J_0' принадлежит $M \setminus I_0'$, но отличен от i', так что $i \in M \setminus I_0$. Значит, числитель дроби из правой части (24) положителен. По определению J_0' положителен и знаменатель. При этом значение самой дроби лежит в интервале $(0,1)$.

Положим

$$\hat{t}_0 = \min_{i \in J_0'} \Big\{ \frac{\big(\Delta(x_0)\big)[i]}{\big(\Delta(x_0)\big)[i] - \big(\Delta(x_0')\big)[i]} \Big\}. \tag{25}$$

Очевидно, что $\hat{t}_0 \in (0,1)$. Возьмём на интервале (x_0, x_0') точку $\hat{x}_0 = x_0(\hat{t}_0)$. По построению она принадлежит Ω. Согласно (22) и (23) множество индексов активных ограничений $\hat{I}_0 = I(\hat{x}_0)$ состоит из I_0' и тех индексов из J_0', на которых достигается минимум в правой части (25). По крайней мере, $|\hat{I}_0| > |I_0'|$. На основании предложения 2 (при $x = x_0$ и $I = I_0'$) выполняется строгое неравенство $Q(\hat{x}_0) < Q(x_0)$.

Теперь положим $x = \hat{x}_0$ и запустим процедуру построения подходящего множества из п. 5°. За конечное число шагов получим подходящее множество I_1 с подходящей парой (x_1, u_1), причем гарантируется, что $Q(x_1) < Q(x_0)$.

Аналогичным образом описывается переход от подходящего множества I_k к подходящему множеству I_{k+1} со строгим уменьшением целевой функции. В силу конечности системы подходящих множеств и строгого убывания целевой функции такой процесс конечен. На выходе получим подходящее множество I_* с подходящей парой (x_*, u_*), такой, что $u_*[I_*] \geqslant 0$. Согласно предложению 1, x_* — решение задачи (1).

8°. Решение системы (2) можно заменить решением экстремальной задачи

$$Q(x) \to \min,$$
$$A[I, N] \times x[N] = b[I].$$

Это позволяет использовать эффективные методы оптимизации (например, метод сопряжённых градиентов [4, с. 173–191]).

9°. Метод перебора граней без условия невырожденности анализируется в книге [5, гл. 3].

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Теорема существования решения для задачи квадратичного программирования* // Семинар «DHA & CAGD». Избранные доклады. 22 января 2011 г.
(`http://dha.spb.ru/reps11.shtml#0122`) [Данная книга, с. 91]

2. Малозёмов В. Н., Монако М. Ф., Петров А. В. *Формулы Фробениуса, Шермана-Моррисона и близкие вопросы.* Журн. вычисл. мат. и матем. физ. 2002. Т. 42. № 10. С. 1459–1465.

3. Малозёмов В. Н. *Теорема Куна-Таккера в дифференциальной форме* // Семинар «DHA & CAGD». Избранные доклады. 27 февраля 2010 г.
(`http://dha.spb.ru/reps10.shtml#0227`) [Данная книга, с. 210]

4. Пшеничный Б. Н., Данилин Ю. М. *Численные методы в экстремальных задачах.* М.: Наука, 1975. 320 с.

5. Даугавет В. А. *Численные методы квадратичного программирования.* СПб.: Изд-во СПбГУ, 2004. 128 с.

О МЕТОДЕ СОПРЯЖЁННЫХ ГРАДИЕНТОВ*

В. Н. Малозёмов

В докладе речь пойдёт об одном эффектном методе минимизации квадратичной функции, предложенном в начале 1950-х годов Хестинсом и Штифелем [1]. Будут использоваться некоторые сведения из книги [2, с. 433–454].

1°. Сопряжённые направления. Пусть D — симметричная положительно определённая матрица порядка n. Два ненулевых вектора s_1, s_2 из $\mathbb{R}^n$ называются D-ортогональными или *сопряжёнными*, если

$$\langle s_2, Ds_1\rangle = \langle Ds_2, s_1\rangle = 0.$$

Рис. 1 поясняет это определение.

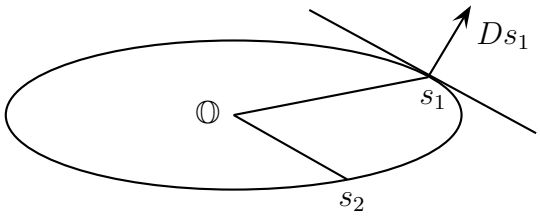


Рис. 1

Опишем динамический метод построения последовательности D-ортогональных векторов. Возьмём произвольный ненулевой вектор $y_1 \in \mathbb{R}^n$ и положим $s_1 = y_1$.

Пусть $y_2 \in \mathbb{R}^n$ — произвольный ненулевой вектор, линейно независимый с y_1. Сопряжённый к s_1 вектор s_2 будем искать в виде

$$s_2 = y_2 + \gamma_{21} s_1.$$

Отметим, что $s_2 = y_2 + \gamma_{21} y_1$. В силу линейной независимости y_1 и y_2 вектор s_2 отличен от нулевого при любом γ_{21}. Найдём γ_{21} из условия D-ортогональности $\langle s_2, Ds_1\rangle = 0$. Получим

$$\gamma_{21} = -\frac{\langle y_2, Ds_1\rangle}{\langle s_1, Ds_1\rangle}.$$

*Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г.

Продолжим процесс. Возьмём произвольный ненулевой вектор $y_3 \in \mathbb{R}^n$, линейно независимый с y_1 и y_2. Очередное сопряжённое направление s_3 будем искать в виде

$$s_3 = y_3 + \gamma_{31} s_1 + \gamma_{32} s_2.$$

Учитывая, что $s_3 = y_3 + \alpha_{31} y_1 + \alpha_{32} y_2$, в силу линейной независимости векторов y_1, y_2, y_3 заключаем, что $s_3 \neq \mathbb{O}$ при любых γ_{31}, γ_{32}. Коэффициенты γ_{31} и γ_{32} найдём из условий D-ортогональности

$$\langle s_3, Ds_1 \rangle = 0, \quad \langle s_3, Ds_2 \rangle = 0.$$

Получим

$$\gamma_{31} = -\frac{\langle y_3, Ds_1 \rangle}{\langle s_1, Ds_1 \rangle}, \quad \gamma_{32} = -\frac{\langle y_3, Ds_2 \rangle}{\langle s_2, Ds_2 \rangle}.$$

Пусть уже построены D-ортогональные векторы $s_1, s_2, \ldots, s_{k-1}$ с помощью последовательно привлекаемых линейно независимых векторов $y_1, y_2, \ldots, y_{k-1}$. Возьмём произвольный ненулевой вектор $y_k \in \mathbb{R}^n$, линейно независимый с $y_1, y_2, \ldots, y_{k-1}$. Сопряжённое направление s_k будем искать в виде

$$s_k = y_k + \sum_{j=1}^{k-1} \gamma_{kj} s_j. \tag{1}$$

В силу линейной независимости векторов y_1, y_2, $\ldots$, y_k вектор s_k отличен от нулевого при любых γ_{kj}. Найдём коэффициенты γ_{kj} из условий D-ортогональности: $\langle s_k, Ds_i \rangle = 0$ при всех $i \in 1 : k-1$. Получим

$$\gamma_{kj} = -\frac{\langle y_k, Ds_j \rangle}{\langle s_j, Ds_j \rangle}, \quad i \in 1 : k-1. \tag{2}$$

При $k = n$ будет построена полная в $\mathbb{R}^n$ система D-ортогональных векторов $s_1, s_2, \ldots, s_n$.

Эти векторы линейно независимы. Действительно, если

$$\sum_{j=1}^{n} c_j s_j = \mathbb{O},$$

то после умножения обеих частей этого равенства скалярно на Ds_i получим $c_i = 0$ при всех $i \in 1 : n$.

Обозначим через S матрицу со столбцами $s_1, s_2, \ldots, s_n$. В силу D-ортогональности

$$S^T(DS) = \Lambda, \tag{3}$$

где Λ — диагональная матрица с диагональными элементами $\Lambda[i,i] = \langle Ds_i, s_i\rangle$, $i \in 1:n$. Из (3) следует, что $D = (S^T)^{-1}\Lambda S^{-1}$ и $D^{-1} = S(\Lambda^{-1}S^T)$. Последнее равенство можно записать в виде

$$D^{-1} = \sum_{i=1}^{n} \frac{s_i s_i^T}{\langle Ds_i, s_i\rangle}. \tag{4}$$

2°. Обратимся к задаче минимизации квадратичной функции:

$$Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \min_{x\in\mathbb{R}^n}. \tag{5}$$

Будем считать, что матрица D симметрична и положительно определена.

Напомним, что для градиента квадратичной функции справедлива формула

$$Q'(x) = Dx + c.$$

Условие $Q'(x) = \mathbb{O}$ служит критерием оптимальности для задачи (5). Значит, решение экстремальной задачи (5) равносильно решению системы линейных уравнений

$$Dx = -c. \tag{6}$$

Это принципиальный факт.

Система (6) имеет единственное решение

$$x_* = -D^{-1}c. \tag{7}$$

Этот же вектор x_* является единственным решением задачи (5).

Предположим, что построена какая-нибудь система D-ортогональных векторов $s_1, s_2, \ldots, s_n$. Возьмём произвольный вектор $x_0 \in \mathbb{R}^n$ и вычислим $g_0 := Q'(x_0) = Dx_0 + c$. На основании (7) и (4) получим

$$\begin{aligned} x_* &= x_0 - D^{-1}(Dx_0 + c) = x_0 - D^{-1}g_0 = \\ &= x_0 - \sum_{i=1}^{n} \frac{\langle s_i, g_0\rangle}{\langle Ds_i, s_i\rangle} s_i = x_0 + \sum_{i=1}^{n} t_i s_i, \end{aligned}$$

где

$$t_i = -\frac{\langle g_0, s_i\rangle}{\langle Ds_i, s_i\rangle}.$$

Обозначим

$$x_k = x_0 + \sum_{i=1}^{k} t_i s_i.$$

Очевидно, что

$$x_k = x_{k-1} + t_k s_k, \quad k = 1, \dots, n. \tag{8}$$

При этом $x_n = x_*$.

Приходим к следующему выводу: *при любом выборе D-ортогональной системы векторов $s_1, s_2, \dots, s_n$ (точнее, при любом выборе линейно независимой системы векторов $y_1, y_2, \dots, y_n$) и произвольном начальном приближении x_0 вычисления по рекуррентной формуле* (8) *приводят к единственному решению задачи* (5).

Положим $g_k := Q'(x_k) = Dx_k + c$. Согласно (8)

$$g_k - g_{k-1} = t_k D s_k, \quad k \in 1 : n. \tag{9}$$

Отметим одну особенность метода сопряжённых направлений.

ПРЕДЛОЖЕНИЕ 1. *Справедливо равенство*

$$\langle g_k, s_i \rangle = 0, \quad i \in 1 : k. \tag{10}$$

Доказательство. Воспользуемся формулой (9). Получим

$$\begin{aligned} \langle g_k, s_i \rangle &= \big\langle (g_k - g_{k-1}) + (g_{k-1} - g_{k-2}) + \dots + (g_1 - g_0) + g_0, s_i \big\rangle = \\ &= \langle \sum_{j=1}^{k} t_j D s_j + g_0, s_i \rangle = t_i \langle D s_i, s_i \rangle + \langle g_0, s_i \rangle = 0. \end{aligned}$$ □

Умножим обе части равенства (9) скалярно на s_k. Приняв во внимание, что $\langle g_k, s_k \rangle = 0$, придём к основному представлению для коэффициентов t_k:

$$t_k = -\frac{\langle g_{k-1}, s_k \rangle}{\langle D s_k, s_k \rangle}, \quad k \in 1 : n. \tag{11}$$

Коэффициент t_k обладает экстремальным свойством, а именно, минимум функции $Q(x_{k-1} + t s_k)$ достигается при $t = t_k$. Это следует из (11) и разложения квадратичной функции

$$Q(x_{k-1} + t s_k) = Q(x_{k-1}) + t \langle g_{k-1}, s_k \rangle + \tfrac{1}{2} t^2 \langle D s_k, s_k \rangle. \tag{12}$$

Отметим, что разложение (12) при $t = t_k$ преобразуется к виду

$$Q(x_{k-1}) - Q(x_k) = \frac{1}{2} \frac{\langle g_{k-1}, s_k \rangle^2}{\langle D s_k, s_k \rangle}.$$

Указанное экстремальное свойство коэффициентов t_k допускает усиление.

ПРЕДЛОЖЕНИЕ 2. *Минимум квадратичной функции $Q(x)$ на аффинном множестве*

$$\mathcal{L} = \{x = x_0 + \sum_{i=1}^{k} z_i s_i\}$$

достигается только при $x = x_k$ (то есть при $z_i = t_i$, $i \in 1:k$).

Доказательство. Введём функцию

$$q_k(z) := Q(x_0 + \sum_{i=1}^{k} z_i s_i) = Q(x_0) + \sum_{i=1}^{k} \langle g_0, s_i \rangle z_i + \tfrac{1}{2} \sum_{i=1}^{k} \langle Ds_i, s_i \rangle z_i^2.$$

Видим, что $q_k(z)$ — квадратичная функция с диагональной матрицей квадратичной формы D_k, у которой $D_k[i,i] = \langle Ds_i, s_i \rangle$, $i \in 1:n$. Матрица D_k является симметричной и положительно определённой. Критерием оптимальности для экстремальной задачи

$$q_k(z) \to \min_{z \in \mathbb{R}^k}$$

служит условие $\langle Ds_i, s_i \rangle z_i = -\langle g_0, s_i \rangle$, $i \in 1:k$. Для компонент оптимального вектора z получаем формулу

$$z_i = -\frac{\langle g_0, s_i \rangle}{\langle Ds_i, s_i \rangle} = t_i, \quad i \in 1:k.$$

Это и требовалось доказать. □

3°. Метод сопряжённых градиентов. Рассмотрим конкретную последовательность привлекаемых к D-ортогонализации векторов и, тем самым, конкретный вариант метода сопряжённых направлений.

Возьмём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычислим градиент $g_0 = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то x_0 — решение задачи (5). Процесс закончен.

Пусть $g_0 \neq \mathbb{O}$. Положим $s_1 = -g_0$ (то есть $y_1 = -g_0$). По общей схеме вычисляем

$$x_1 = x_0 + t_1 s_1,$$

где

$$t_1 = -\frac{\langle g_0, s_1 \rangle}{\langle Ds_1, s_1 \rangle} = \frac{\langle g_0, g_0 \rangle}{\langle Ds_1, s_1 \rangle}.$$

Пересчитаем градиент $g_1 = g_0 + t_1 Ds_1$. Если $g_1 = \mathbb{O}$, то x_1 — решение задачи (5). Процесс закончен.

Пусть $g_1 \neq \mathbb{O}$. Отметим, что

$$\langle g_1, g_0 \rangle = \langle g_0, g_0 \rangle - t_1 \langle Ds_1, s_1 \rangle = 0.$$

Таким образом, вектор $y_2 = -g_1$ ортогонален y_1. Его можно привлечь к процессу D-ортогонализации. Полагаем

$$s_2 = -g_1 + \gamma_{21} s_1,$$

где

$$\gamma_{21} = -\frac{\langle y_2, Ds_1\rangle}{\langle s_1, Ds_1\rangle} = \frac{\langle g_1, g_1 - g_0\rangle}{t_1\langle Ds_1, s_1\rangle} = \frac{\langle g_1, g_1\rangle}{\langle g_0, g_0\rangle} =: b_1.$$

Приходим к формуле

$$s_2 = -g_1 + b_1 s_1.$$

Предположим, что уже построены x_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. При этом градиенты $g_0, g_1, \ldots, g_{k-1}$ попарно ортогональны, при $i \in 1 : k-1$

$$s_{i+1} = -g_i + b_i s_i, \qquad b_i = \frac{\langle g_i, g_i\rangle}{\langle g_{i-1}, g_{i-1}\rangle}, \tag{13}$$

и, по общему свойству метода сопряжённых напрвлений,

$$\langle g_{k-1}, s_i\rangle = 0, \quad i \in 1 : k-1. \tag{14}$$

Находим очередное приближение

$$x_k = x_{k-1} + t_k s_k,$$

где

$$t_k = -\frac{\langle g_{k-1}, s_k\rangle}{\langle Ds_k, s_k\rangle} = -\frac{\langle g_{k-1}, -g_{k-1} + b_{k-1}s_{k-1}\rangle}{\langle Ds_k, s_k\rangle} = \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle}.$$

Пересчитываем градиент $g_k = g_{k-1} + t_k Ds_k$. Если $g_k = \mathbb{O}$, то x_k — решение задачи (5). Процесс закончен.

Пусть $g_k \neq \mathbb{O}$. Покажем, что

$$\langle g_k, g_i\rangle = 0, \quad i \in 0 : k-1. \tag{15}$$

В силу D-ортогональности и формул (13), (14) при $i \in 0 : k-2$ имеем (считаем, что $s_0 = \mathbb{O}$)

$$\langle g_k, g_i\rangle = \langle g_{k-1} + t_k Ds_k, -s_{i+1} + b_i s_i\rangle = 0.$$

К этому нужно добавить, что

$$\begin{aligned}\langle g_k, g_{k-1}\rangle &= \langle g_{k-1} + t_k Ds_k, -s_k + b_{k-1}s_{k-1}\rangle = \\ &= -\langle g_{k-1}, s_k\rangle - t_k\langle Ds_k, s_k\rangle = 0.\end{aligned}$$

Соотношение (15) установлено.

Вектор $y_{k+1} = -g_k$ привлекаем к процессу D-ортогонализации. Запишем

$$s_{k+1} = -g_k + \sum_{i=1}^{k} \gamma_{k+1,i} s_i.$$

Здесь

$$\gamma_{k+1,i} = -\frac{\langle y_{k+1}, Ds_i\rangle}{\langle s_i, Ds_i\rangle} = \frac{\langle g_k, g_i - g_{i-1}\rangle}{t_i\langle Ds_i, s_i\rangle}.$$

Согласно (15) имеем $\gamma_{k+1,i} = 0$ при $i \in 1 : k-1$ и

$$\gamma_{k+1,k} = \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle} =: b_k.$$

Таким образом,

$$s_{k+1} = -g_k + b_k s_k.$$

Описание метода завершено. Он называется *методом сопряжённых градиентов*, поскольку именно текущие градиенты привлекаются к процессу D-ортогонализации.

4°. Вычислительная схема метода сопряжённых градиентов.

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычисляем градиент $g_0 = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то x_0 — решение задачи (5). Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

k-й шаг. Пусть уже имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$ и s_k. Последовательно вычисляем

$$\begin{aligned} t_k &= \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle}, \\ x_k &= x_{k-1} + t_k s_k, \\ g_k &= g_{k-1} + t_k Ds_k. \end{aligned}$$

Если $g_k = \mathbb{O}$, то x_k — решение задачи (5). Вычисления прекращаются. В противном случае находим

$$\begin{aligned} b_k &= \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle}, \\ s_{k+1} &= -g_k + b_k s_k. \end{aligned}$$

По крайней мере, при $k = n$ (а возможно и раньше) получим $x_k = x_*$.

На рис. 2 схематично представлен шаг метода сопряжённых градиентов.

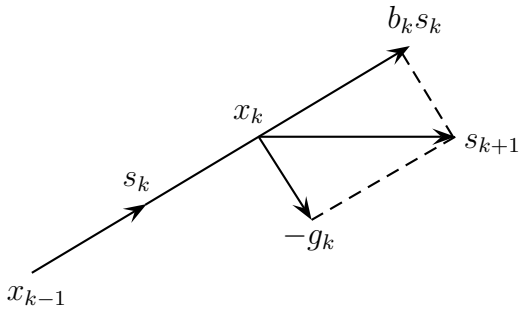


Рис. 2

5°. До сих пор мы предполагали, что матрица D в задаче (5) симметрична и положительно определена. Ослабим это предположение. Будем считать, что матрица D симметрична и неотрицательно определена. В этом случае квадратичная функция $Q(x)$ остаётся выпуклой на $\mathbb{R}^n$ и критерий оптимальности для задачи (5) сохраняет вид $Q'(x) = \mathbb{O}$.

ПРЕДЛОЖЕНИЕ 3. *Если квадратичная функция $Q(x)$ ограничена снизу на $\mathbb{R}^n$, то задача* (5) *имеет решение.*

Доказательство. Допустим противное. Тогда условие $Q'(x) = \mathbb{O}$ не выполняется ни при каком $x \in \mathbb{R}^n$, то есть система линейных уравнений $Dx = -c$ несовместна. Это в свою очередь означает, что найдется вектор $u_0 \in \mathbb{R}^n$, такой, что

$$Du_0 = \mathbb{O}, \quad \langle c, u_0\rangle \neq 0.$$

Возьмём произвольный вектор $x_0 \in \mathbb{R}^n$ и запишем разложение

$$\begin{aligned} Q(x_0 + tu_0) &= Q(x_0) + t\langle Dx_0 + c, u_0\rangle + \tfrac{1}{2}\langle Du_0, u_0\rangle = \\ &= Q(x_0) + t\big[\langle x_0, Du_0\rangle + \langle c, u_0\rangle\big] = Q(x_0) + t\langle c, u_0\rangle. \end{aligned} \tag{16}$$

Отсюда следует, вопреки условию предложения, что квадратичная функция $Q(x)$ неограничена снизу на прямой $x = x_0 + tu_0$, $t \in \mathbb{R}$.

Предложение доказано. □

Введём обозначение

$$\mathcal{P} = \{p \in \mathbb{R}^n \mid Dp = \mathbb{O}\}.$$

По-прежнему считаем, что квадратичная функция $Q(x)$ ограничена снизу на $\mathbb{R}^n$.

ПРЕДЛОЖЕНИЕ 4. *При всех $x \in \mathbb{R}^n$ и всех $p \in \mathcal{P}$ справедливо равенство*

$$\langle Q'(x), p\rangle = 0. \tag{17}$$

Доказательство. Согласно (16) при $p \in \mathcal{P}$

$$Q(x_0 + tp) = Q(x_0) + t\langle c, p\rangle.$$

Учитывая ограниченность снизу функции $Q(x)$ на $\mathbb{R}^n$, заключаем, что

$$\langle c, p\rangle = 0 \quad \forall p \in \mathcal{P}.$$

Теперь имеем

$$\langle Q'(x), p\rangle = \langle Dx + c, p\rangle = \langle Dx, p\rangle = \langle x, Dp\rangle = 0.$$

Предложение доказано □

Ортогональное дополнение к линейному множеству $\mathcal{P}$ обозначим $\mathcal{P}^\perp$. Формула (17) равносильна включению

$$Q'(x) \in \mathcal{P}^\perp \quad \forall x \in \mathbb{R}^n. \tag{18}$$

ПРЕДЛОЖЕНИЕ 5. *При всех $x \in \mathcal{P}^\perp$, $x \neq \mathbb{O}$, выполняется неравенство*

$$\langle Dx, x\rangle > 0.$$

Доказательство. Допустим, вопреки утверждению, что существует точка $x_0 \in \mathcal{P}^\perp$, $x_0 \neq \mathbb{O}$, в которой $\langle Dx_0, x_0\rangle = 0$. Этот факт можно проинтерпретировать следующим образом: точка x_0 доставляет минимум квадратичной функции $\varphi(x) = \frac{1}{2}\langle Dx, x\rangle$ на $\mathbb{R}^n$. Но тогда $\varphi'(x_0) = \mathbb{O}$, так что $Dx_0 = \mathbb{O}$. Получили, что $x_0 \in \mathcal{P}$. Вместе с включением $x_0 \in \mathcal{P}^\perp$ это гарантирует равенство $x_0 = \mathbb{O}$, противоречащее условию $x_0 \neq \mathbb{O}$.

Предложение доказано. □

6°. Метод сопряжённых градиентов в вырожденном случае.

Вернёмся к экстремальной задаче (5) при ослабленном условии на матрицу D. Будем считать, что матрица D симметрична и неотрицательно определена, причём $\operatorname{rank} D = r < n$. Вначале рассмотрим случай, когда квадратичная функция $Q(x)$ ограничена снизу на $\mathbb{R}^n$. Для решения задачи (5) формально воспользуемся методом сопряжённых градиентов. Вычисления будут продолжаться, пока градиенты $g_k = Q'(x_k)$ отличны от нуля. Согласно (18), $g_k \in \mathcal{P}^\perp$.

Так как сопряжённые направления s_k являются линейными комбинациями попарно ортогональных градиентов $g_0, g_1, \ldots, g_{k-1}$ и коэффициент при g_{k-1} в

такой линейной комбинации равен -1, то $s_k \in \mathcal{P}^\perp$ и $s_k \neq \mathbb{O}$. В силу предложения 5, $\langle Ds_k, s_k\rangle > 0$. Это гарантирует беспрепятственную реализацию вычислений, пока $g_k \neq \mathbb{O}$.

По условию $\operatorname{rank} D = r$, так что $\dim \mathcal{P} = n - r$ и $\dim \mathcal{P}^\perp = r$. Текущие градиенты g_k попарно ортогональны и принадлежат $\mathcal{P}^\perp$. Значит, по крайней мере, при $k = r$ (а возможно и раньше) получим $g_k = \mathbb{O}$. Согласно критерию оптимальности, соответствующее x_k является решением задачи (5).

Установлено замечательное свойство метода сопряжённых градиентов:
в случае, когда $\operatorname{rank} D = r < n$ *и квадратичная функция* $Q(x)$ *ограничена снизу на* $\mathbb{R}^n$, *метод сопряжённых градиентов решает задачу* (5) *не более, чем за* r *итераций.*

Теперь предположим, что квадратичная функция $Q(x)$ неограничена снизу на $\mathbb{R}^n$. Тогда задача (5) не имеет решения и $Q'(x) \neq \mathbb{O}$ при всех $x \in \mathbb{R}^n$. В процессе вычислений по методу сопряжённых градиентов при некотором k выполнится равенство $\langle Ds_k, s_k\rangle = 0$ (иначе процесс построения попарно ортогональных градиентов будет бесконечным). В этом случае квадратичная функция $Q(x)$ неограничена снизу на луче $x = x_{k-1} + ts_k$, $t > 0$, что следует из разложения

$$\begin{aligned} Q(x_{k-1} + ts_k) &= Q(x_{k-1}) + t\langle g_{k-1}, s_k\rangle = \\ &= Q(x_{k-1}) + t\langle g_{k-1}, -g_{k-1} + b_{k-1}s_{k-1}\rangle = Q(x_{k-1}) - t\|g_{k-1}\|^2. \end{aligned}$$

ЛИТЕРАТУРА

1. Hestenes M. R., Stiefel E. *Methods of conjudate gradients for solving linear systems* // J. Res. Nat. Bur. Standarts. 1952. Vol. 49. No. 6. P. 409–436.

2. Фаддеев Д. К., Фаддеева В. Н. *Вычислительные методы линейной алгебры.* М.: Физматгиз, 1960. 656 с.

ВАРИАНТЫ МЕТОДА СОПРЯЖЁННЫХ ГРАДИЕНТОВ*

В. Н. Малозёмов

1°. Напомним описание основного варианта метода сопряжённых градиентов для минимизации на $\mathbb{R}^n$ квадратичной функции

$$Q(x) = \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle$$

с симметричной положительно определённой матрицей D [1].

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычисляем градиент $g_0 = Q'(x_0) = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то x_0 — точка минимума. Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

k-й шаг. Пусть имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. Последовательно вычисляем

$$\begin{aligned} t_k &= \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle}, \qquad (1)\\ x_k &= x_{k-1} + t_k s_k,\\ g_k &= Q'(x_k) = g_{k-1} + t_k Ds_k. \end{aligned}$$

Если $g_k = \mathbb{O}$, то x_k — точка минимума. Вычисления прекращаются. В противном случае находим

$$b_k = \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle}, \qquad (2)$$

$$s_{k+1} = -g_k + b_k s_k. \qquad (3)$$

Описание метода завершено.

Метод сопряжённых градиентов обладает следующими свойствами:

$$\begin{aligned} \langle Ds_k, s_j\rangle &= 0 \quad \text{при} \quad k \neq j;\\ \langle g_k, s_j\rangle &= 0 \quad \text{при} \quad j \in 1:k; \qquad (4)\\ \langle g_k, g_j\rangle &= 0 \quad \text{при} \quad k \neq j. \qquad (5) \end{aligned}$$

Из последнего соотношения, в частности, следует, что $g_n = \mathbb{O}$ (если равенство $g_k = \mathbb{O}$ не встретится при $k < n$), то есть, по крайней мере, x_n будет точкой минимума.

*Семинар «CNSA & NDO». Избранные доклады. 29 октября 2015 г.

2°. Для последовательности $x_0, x_1, \ldots$, построенной методом сопряжённых градиентов, можно получить другое представление. На рис. 1 введены векторы p_k и p_{k+1}, исходя из условий

$$p_{k+1} = -g_k + \lambda(p_k + g_k), \tag{6}$$

$$\langle p_{k+1}, p_k + g_k \rangle = 0. \tag{7}$$

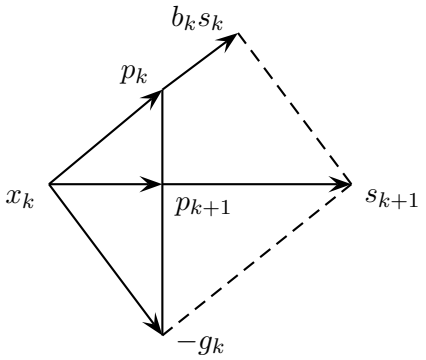


Рис. 1

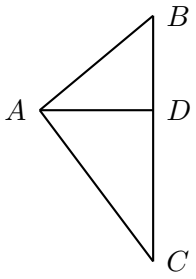


Рис. 2

На рис. 2 выделен треугольник ABC с $\overrightarrow{AB} = p_k$, $\overrightarrow{AC} = -g_k$ и $\overrightarrow{AD} = p_{k+1}$. Угол CAB прямой, поскольку, согласно (4), $\langle s_k, g_k \rangle = 0$. Вектор $p_{k+1} = \overrightarrow{AD}$ строится из условия, что отрезок AD перпендикулярен CB, то есть в прямоугольном треугольнике CAB из вершины A опускается перпендикуляр на сторону CB. При этом $\overrightarrow{AC} + \overrightarrow{CB} = \overrightarrow{AB}$, так что $\overrightarrow{CB} = p_k + g_k$.

Умножим равенство (6) скалярно на $p_k + g_k$. Учитывая (7) и ортогональность векторов p_k и g_k, получаем

$$0 = -\langle g_k, g_k \rangle + \lambda\big(\|p_k\|^2 + \|g_k\|^2\big).$$

Отсюда следует, что $\lambda = \frac{\|g_k\|^2}{\|p_k\|^2 + \|g_k\|^2}$ и

$$p_{k+1} = -g_k + \frac{\|g_k\|^2(p_k + g_k)}{\|p_k\|^2 + \|g_k\|^2} = \frac{\|p_k\|^2(-g_k) + \|g_k\|^2 p_k}{\|p_k\|^2 + \|g_k\|^2}. \tag{8}$$

К этому добавим условие $p_1 = -g_0$.

ТЕОРЕМА. *Справедливо равенство*

$$p_k = \frac{\|p_k\|^2}{\|g_{k-1}\|^2} s_k, \quad k = 1, 2, \ldots \tag{9}$$

Доказательство. Обозначим $\beta_k = \frac{\|p_k\|^2}{\|g_{k-1}\|^2}$, так что формулу (9) можно переписать в виде $p_k = \beta_k s_k$. При $k = 1$ равенство (9) выполняется. Сделаем индукционный переход от k к $k+1$.

Во-первых, отметим, что векторы p_k и g_k ортогональны, поскольку $\langle s_k, g_k\rangle = 0$. Далее, в силу (8), индукционного предположения (9) и (3) имеем

$$p_{k+1} = \frac{\|p_k\|^2}{\|p_k\|^2 + \|g_k\|^2}\Big(-g_k + \frac{\|g_k\|^2}{\|p_k\|^2}p_k\Big) =$$
$$= \frac{\|p_k\|^2}{\|p_k\|^2 + \|g_k\|^2}\Big(-g_k + \frac{\|g_k\|^2}{\|g_{k-1}\|^2}s_k\Big) = \frac{\|p_k\|^2}{\|p_k\|^2 + \|g_k\|^2}s_{k+1}.$$

Остаётся учесть, что

$$\|p_{k+1}\|^2 = \frac{\|p_k\|^4}{\big(\|p_k\|^2 + \|g_k\|^2\big)^2}\Big(\|g_k\|^2 + \frac{\|g_k\|^4}{\|p_k\|^2}\Big) = \frac{\|p_k\|^2\|g_k\|^2}{\|p_k\|^2 + \|g_k\|^2}.$$

Для коэффициента перед s_{k+1} получаем представление

$$\frac{\|p_k\|^2}{\|p_k\|^2 + \|g_k\|^2} = \frac{\|p_{k+1}\|^2}{\|g_k\|^2} = \beta_{k+1}.$$

Теорема доказана. □

Нетрудно проверить, что справедливо равенство

$$t_k s_k = \alpha_k p_k, \tag{10}$$

где

$$\alpha_k = \frac{\langle p_k, p_k\rangle}{\langle Dp_k, p_k\rangle}.$$

Действительно, согласно (9), (1) и определению β_k

$$\alpha_k = \frac{t_k}{\beta_k} = \frac{\|g_{k-1}\|^2\beta_k}{\langle Ds_k, s_k\rangle\beta_k^2} = \frac{\|p_k\|^2}{\langle Dp_k, p_k\rangle}.$$

3°. Воспользуемся векторами p_k для описания метода сопряжённых градиентов, имея в виду, что векторы p_k отличаются от s_k лишь положительным коэффициентом.

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычисляем градиент $g_0 = Q'(x_0) = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то x_0 — точка минимума. Вычисления прекращаются. Иначе полагаем $p_1 = -g_0$.

k-й шаг. Пусть имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$ и p_k. Последовательно вычисляем

$$\alpha_k = \frac{\langle p_k, p_k\rangle}{\langle Dp_k, p_k\rangle}, \tag{11}$$

$$x_k = x_{k-1} + \alpha_k p_k,$$
$$g_k = Q'(x_k) = g_{k-1} + \alpha_k D p_k.$$

Если $g_k = \mathbb{O}$, то x_k — точка минимума. Вычисления прекращаются. В противном случае находим

$$p_{k+1} = \frac{\|p_k\|^2(-g_k) + \|g_k\|^2 p_k}{\|p_k\|^2 + \|g_k\|^2}.$$

Описание метода завершено.

Данный вариант метода сопряжённых градиентов назовём *геометрическим вариантом*.

4°. Для шага α_k в геометрическом варианте метода сопряжённых градиентов наряду с формулой (11) справедливо представление

$$\alpha_k = \operatorname*{argmin}_{\alpha>0} Q(x_{k-1} + \alpha p_k). \tag{12}$$

Проверим это.

Функция $\varphi(\alpha) = Q(x_{k-1} + \alpha p_k)$ является выпуклой и

$$\varphi'(\alpha) = \big\langle Q'(x_{k-1} + \alpha p_k), p_k \big\rangle = \langle g_{k-1} + \alpha D p_k, p_k \rangle = \\ = \langle g_{k-1}, p_k \rangle + \alpha \langle D p_k, p_k \rangle.$$

При этом согласно (9), (3), (4) и определению β_k

$$\langle g_{k-1}, p_k \rangle = \langle g_{k-1}, \beta_k s_k \rangle = \big\langle g_{k-1}, \beta_k(-g_{k-1} + \beta_{k-1} s_{k-1}) \big\rangle = \\ = -\beta_k \langle g_{k-1}, g_{k-1} \rangle = -\langle p_k, p_k \rangle,$$

так что

$$\varphi'(\alpha) = -\langle p_k, p_k \rangle + \alpha \langle D p_k, p_k \rangle.$$

Отсюда следует, что $\varphi'(0) < 0$ и $\varphi'(\alpha) = 0$ только при $\alpha = \alpha_k$. Это гарантирует справедливость формулы (12).

В силу (10) шаг t_k вида (1) в основном варианте метода сопряжённых градиентов допускает аналогичное представление

$$t_k = \operatorname*{argmin}_{t>0} Q(x_{k-1} + t s_k).$$

Таким образом, в обоих вариантах неявно присутствует одномерная минимизация (линейный поиск). В следующих пунктах будет описан вариант метода сопряжённых градиентов без точного линейного поиска.

5°. Параллельно с последовательностью $\{x_k\}$, построенной с помощью основного варианта метода сопряжённых градиентов, рассмотрим ещё одну последовательность $\{\hat{x}_k\}$ вида

$$\hat{x}_k = \hat{x}_{k-1} + \hat{t}_k \hat{s}_k, \quad k = 1, 2, \ldots,$$

где $\hat{x}_0 = x_0$ и $\hat{s}_k = s_k$ при всех $k = 1, 2, \ldots$ В качестве шага $\hat{t}_k$ возьмём произвольное положительное число (например, обеспечивающее неравенство $Q(\hat{x}_{k-1} + \hat{t}_k \hat{s}_k) < Q(\hat{x}_{k-1})$). Выясним, как связаны последовательности $\{x_k\}$ и $\{\hat{x}_k\}$.

Введём обозначения:

$$\begin{gathered}
\theta_k = \hat{t}_k / t_k \quad (\text{так что } \hat{t}_k = \theta_k t_k), \\
\hat{z}_k = \hat{x}_k - \hat{x}_{k-1} = \hat{t}_k \hat{s}_k, \quad z_k = x_k - x_{k-1} = t_k s_k, \\
\hat{r}_k = \hat{g}_k - \hat{g}_{k-1} = D\hat{z}_k, \quad r_k = g_k - g_{k-1} = Dz_k, \\
u_k = \hat{x}_k - x_k, \qquad v_k = \hat{g}_k - g_k.
\end{gathered}$$

В силу равенства $\hat{x}_0 = x_0$ имеем $u_0 = \mathbb{O}$, $v_0 = \mathbb{O}$. Далее

$$\begin{gathered}
\hat{z}_k = \hat{t}_k \hat{s}_k = \theta_k t_k s_k = \theta_k z_k, \\
\hat{r}_k = D\hat{z}_k = \theta_k D z_k = \theta_k r_k, \\
u_k = (\hat{x}_{k-1} + \hat{z}_k) - (x_{k-1} + z_k) = u_{k-1} + (1 - \tfrac{1}{\theta_k})\hat{z}_k, \qquad (13) \\
v_k = \hat{g}_k - g_k = (\hat{g}_k - \hat{g}_{k-1}) + (\hat{g}_{k-1} - g_{k-1}) - (g_k - g_{k-1}) = \\
= \hat{r}_k + v_{k-1} - r_k = v_{k-1} + (1 - \tfrac{1}{\theta_k})\hat{r}_k. \qquad (14)
\end{gathered}$$

Из последней формулы и равенства $v_0 = \mathbb{O}$ следует, что

$$v_k = \sum_{j=1}^{k} (1 - \tfrac{1}{\theta_j})\hat{r}_j = \sum_{j=1}^{k} (\theta_j - 1) r_j. \qquad (15)$$

Умножим обе части соотношения (15) скалярно на $\hat{z}_k = \theta_k z_k$. Получим

$$\langle \hat{g}_k - g_k, \hat{z}_k \rangle = \sum_{j=1}^{k} (\theta_j - 1) \langle r_j, \hat{z}_k \rangle. \qquad (16)$$

Согласно (4)

$$\langle g_k, \hat{z}_k \rangle = \theta_k \langle g_k, z_k \rangle = \theta_k t_k \langle g_k, s_k \rangle = 0$$

и при $j < k$

$$\langle r_j, \hat{z}_k \rangle = \theta_k \langle Dz_j, z_k \rangle = \theta_k t_j t_k \langle Ds_j, s_k \rangle = 0.$$

Формула (16) принимает вид

$$\langle \hat{g}_k, \hat{z}_k \rangle = (\theta_k - 1)\langle r_k, \hat{z}_k \rangle = \tfrac{\theta_k - 1}{\theta_k}\langle \hat{r}_k, \hat{z}_k \rangle.$$

Учитывая, что $\langle \hat{r}_k, \hat{z}_k \rangle = \langle D\hat{z}_k, \hat{z}_k \rangle > 0$, получаем

$$\frac{\theta_k - 1}{\theta_k} = \frac{\langle \hat{g}_k, \hat{z}_k \rangle}{\langle \hat{r}_k, \hat{z}_k \rangle}.$$

Теперь рекуррентные соотношения (13) и (14) можно переписать так:

$$u_k = u_{k-1} + \frac{\langle \hat{g}_k, \hat{z}_k \rangle}{\langle \hat{r}_k, \hat{z}_k \rangle}\hat{z}_k,\ k = 1, 2, \ldots; \quad u_0 = \mathbb{O};$$
$$v_k = v_{k-1} + \frac{\langle \hat{g}_k, \hat{z}_k \rangle}{\langle \hat{r}_k, \hat{z}_k \rangle}\hat{r}_k,\ k = 1, 2, \ldots; \quad v_0 = \mathbb{O}.$$

При этом

$$x_k = \hat{x}_k - u_k, \quad g_k = \hat{g}_k - v_k.$$

6°. Переходим к описанию варианта метода сопряжённых градиентов без точного линейного поиска.

Нулевой шаг. Берём произвольное начальное приближение $\hat{x}_0 \in \mathbb{R}^n$. Вычисляем градиент $\hat{g}_0 = Q'(\hat{x}_0) = D\hat{x}_0 + c$. Если $\hat{g}_0 = \mathbb{O}$, то $\hat{x}_0$ — точка минимума. Вычисления прекращаются. Иначе полагаем

$$x_0 = \hat{x}_0,\ g_0 = \hat{g}_0,\ s_1 = -g_0,\ u_0 = \mathbb{O},\ v_0 = \mathbb{O}.$$

***k*-й шаг.** Пусть имеются

$$\hat{x}_{k-1},\ x_{k-1}, \hat{g}_{k-1} \neq \mathbb{O},\ g_{k-1} \neq \mathbb{O},\ s_k,\ u_{k-1},\ v_{k-1}.$$

Произвольно выбираем шаг $\hat{t}_k > 0$. Находим $\hat{z}_k = \hat{t}_k s_k$,

$$\hat{x}_k = \hat{x}_{k-1} + \hat{z}_k,$$
$$\hat{g}_k = Q'(\hat{x}_k) = \hat{g}_{k-1} + D\hat{z}_k.$$

Если $\hat{g}_k = \mathbb{O}$, то $\hat{x}_k$ — точка минимума. Вычисления прекращаются.

Пусть $\hat{g}_k \neq \mathbb{O}$. Тогда вычисляем

$$\hat{r}_k = \hat{g}_k - \hat{g}_{k-1}, \quad \hat{c}_k = \frac{\langle \hat{g}_k, \hat{z}_k \rangle}{\langle \hat{r}_k, \hat{z}_k \rangle},$$
$$u_k = u_{k-1} + \hat{c}_k\hat{z}_k, \quad v_k = v_{k-1} + \hat{c}_k\hat{r}_k,$$
$$x_k = \hat{x}_k - u_k, \quad g_k = Q'(x_k) = \hat{g}_k - v_k.$$

Если $g_k = \mathbb{O}$, то x_k — точка минимума. Вычисления прекращаются. Иначе полагаем $s_{k+1} = -g_k + b_k s_k$, где b_k имеет вид (2).

Описание метода завершено.

Из описания видно, что в методе сопряжённых градиентов без точного линейного поиска наряду с последовательностью $\{\hat{x}_k\}$ строится стандартная последовательность $\{x_k\}$ метода сопряжённых градиентов. Это гарантирует сходимость метода не более чем за n шагов.

7°. Следуя Соренсену, отметим, что коэффициент b_k допускает другое представление:

$$b_k = \frac{\langle g_k, \hat{r}_k \rangle}{\langle s_k, \hat{r}_k \rangle}. \tag{17}$$

Действительно, согласно (4) и (5)

$$\begin{aligned}
\langle g_k, \hat{r}_k \rangle &= \theta_k \langle g_k, g_k - g_{k-1} \rangle = \theta_k \langle g_k, g_k \rangle, \\
\langle s_k, \hat{r}_k \rangle &= \theta_k \langle s_k, g_k - g_{k-1} \rangle = -\theta_k \langle s_k, g_{k-1} \rangle = \\
&= -\theta_k \langle -g_{k-1} + b_{k-1} s_{k-1}, g_{k-1} \rangle = \theta_k \langle g_{k-1}, g_{k-1} \rangle.
\end{aligned}$$

Отсюда и из (2) следует (17).

8°. История метода сопряжённых градиентов за период с 1948 по 1976 годы (с обширной аннотированной библиографией) представлена в обзорной статье [2]. Современному состоянию в этой области посвящена книга [3].

ЛИТЕРАТУРА

1. Малозёмов В. Н. *О методе сопряжённых градиентов* // Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г. (`http://dha.spb.ru/reps12.shtml#0428`) [Данная книга, с. 108]

2. Golub G. H., O'Leary D. P. *Some History of the Conjugate Gradient and Lanczos Algorithms: 1948–1976* // SIAM Rev. 1989. Vol. 31, No. 1, pp. 50–102.

3. Pytlak R. *Conjugate Gradient Algorithms in Nonconvex Optimization.* Berlin: Springer, 2009.

ПРЕДОБУСЛАВЛИВАНИЕ В МЕТОДЕ СОПРЯЖЁННЫХ ГРАДИЕНТОВ*

В. Н. Малозёмов

1°. Пусть D — симметричная положительно определённая матрица порядка n и c — n-мерный вектор. Рассмотрим экстремальную задачу

$$Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \min_{x\in\mathbb{R}^n}. \tag{1}$$

Её единственное решение x_* определяется из уравнения

$$Dx = -c.$$

Возьмём произвольную симметричную положительно определённую матрицу B порядка n. Как известно, существует единственная симметричная положительно определённая матрица $B^{1/2}$ со свойством $B^{1/2}B^{1/2} = B$. Обозначим

$$A = B^{1/2}DB^{1/2}, \quad b = -B^{1/2}c.$$

Очевидно, что матрица A является симметричной и положительно определённой.

Наряду с задачей (1) рассмотрим ещё одну экстремальную задачу

$$F(y) := \tfrac{1}{2}\langle Ay, y\rangle - \langle b, y\rangle \to \min_{y\in\mathbb{R}^n}. \tag{2}$$

Её единственное решение y_* определяется из уравнения

$$Ay = b$$

или, в развёрнутой записи,

$$B^{1/2}DB^{1/2}y = -B^{1/2}c.$$

Ясно, что решения x_*, y_* задач (1) и (2) связаны соотношением

$$x_* = B^{1/2}y_*.$$

*Семинар «CNSA & NDO». Избранные доклады. 3 декабря 2015 г.

2°. Опишем общий шаг метода сопряжённых градиентов для решение задачи (2) (см., например, [1]).

***k*-й шаг.** К этому моменту уже имеются

$$y_{k-1},\ \tilde{g}_{k-1} := Ay_{k-1} - b \neq \mathbb{O},\ \tilde{s}_k.$$

Вычисляем

$$\tilde{t}_k = \frac{\langle \tilde{g}_{k-1}, \tilde{g}_{k-1} \rangle}{\langle A\tilde{s}_k, \tilde{s}_k \rangle}, \tag{3}$$

$$y_k = y_{k-1} + \tilde{t}_k \tilde{s}_k, \tag{4}$$

$$\tilde{g}_k = \tilde{g}_{k-1} + \tilde{t}_k A\tilde{s}_k. \tag{5}$$

Если $\tilde{g}_k = \mathbb{O}$, то y_k — точка минимума функции $F(y)$ на $\mathbb{R}^n$. Процесс завершён. Иначе вычисляем

$$\tilde{b}_k = \frac{\langle \tilde{g}_k, \tilde{g}_k \rangle}{\langle \tilde{g}_{k-1}, \tilde{g}_{k-1} \rangle}, \tag{6}$$

$$\tilde{s}_{k+1} = -\tilde{g}_k + \tilde{b}_k \tilde{s}_k. \tag{7}$$

Методом сопряжённых градиентов строится последовательность $y_0, y_1, \ldots, y_k, \ldots$, которая сходится к решению задачи (2) не более, чем за n шагов.

3°. Имея в виду решение задачи (1), перейдём от последовательности $\{y_k\}$ к последовательности $\{x_k\}$, где $x_k = B^{1/2} y_k$.

Если x_0 — произвольное начальное приближение для решения задачи (1), то в качестве согласованного начального приближения для решения задачи (2) следует взять $y_0 = B^{-1/2} x_0$, где $B^{-1/2} = (B^{1/2})^{-1}$. При этом

$$\tilde{g}_0 = Ay_0 - b = B^{1/2} D x_0 + B^{1/2} c = B^{1/2}(Dx_0 + c) = B^{1/2} g_0.$$

Отметим, что и в общем случае

$$\tilde{g}_k = Ay_k - b = B^{1/2}(Dx_k + c) = B^{1/2} g_k. \tag{8}$$

Умножим равенство (4) слева на матрицу $B^{1/2}$. Получим

$$x_k = x_{k-1} + \tilde{t}_k s_k,$$

где $s_k = B^{1/2} \tilde{s}_k$. Формула (3) согласно (8) принимает вид

$$\tilde{t}_k = \frac{\langle g_{k-1}, B g_{k-1} \rangle}{\langle D s_k, s_k \rangle}.$$

Из (5) и (8) следует, что

$$g_k = g_{k-1} + \tilde{t}_k D s_k.$$

В силу (8) условия $\tilde{g}_k = \mathbb{O}$ и $g_k = \mathbb{O}$ равносильны. Если $g_k = \mathbb{O}$, то и $\tilde{g}_k = \mathbb{O}$. В этом случае $y_k = y_*$ и

$$x_k = B^{1/2} y_k = B^{1/2} y_* = x_*.$$

Значит, x_k является решением задачи (1).

Пусть $g_k \neq \mathbb{O}$. На основании (8) формулу (6) можно переписать так:

$$\tilde{b}_k = \frac{\langle g_k, B g_k \rangle}{\langle g_{k-1}, B g_{k-1} \rangle}.$$

Умножим равенство (7) слева на матрицу $B^{1/2}$. Придём к формуле

$$s_{k+1} = -B g_k + \tilde{b}_k s_k.$$

Объединив полученные результаты, получим параметрический вариант метода сопряжённых градиентов для решения задачи (1). Параметром является симметричная положительно определённая матрица B.

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычисляем градиент $g_0 = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то $x_0 = x_*$. Вычисления прекращаются. Иначе полагаем

$$s_1 = B^{1/2} \tilde{s}_1 = -B^{1/2} \tilde{g}_0 = -B g_0.$$

***k*-й шаг.** Имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. Вычисляем

$$\begin{aligned} \tilde{t}_k &= \frac{\langle g_{k-1}, B g_{k-1} \rangle}{\langle D s_k, s_k \rangle}, \\ x_k &= x_{k-1} + \tilde{t}_k s_k, \\ g_k &= g_{k-1} + \tilde{t}_k D s_k. \end{aligned} \tag{9}$$

Если $g_k = \mathbb{O}$, то $x_k = x_*$. Вычисления прекращаются. Иначе находим

$$\begin{aligned} \tilde{b}_k &= \frac{\langle g_k, B g_k \rangle}{\langle g_{k-1}, B g_{k-1} \rangle}, \\ s_{k+1} &= -B g_k + \tilde{b}_k s_k. \end{aligned}$$

Описание метода завершено.

Отметим, что последовательность $x_0, x_1, \ldots, x_k, \ldots$ сходится к x_* не более чем за n шагов.

4°. Коэффициент $\tilde{t}_k$ вида (9) обладает экстремальным свойством.

ПРЕДЛОЖЕНИЕ 1. *Минимум функции*

$$\varphi_k(t) = Q(x_{k-1} + ts_k)$$

на полуоси $(0, +\infty)$ *достигается при* $t = \tilde{t}_k$.

Доказательство. При $k = 1$ имеем

$$Q(x_0 + ts_1) = Q(x_0) + \langle g_0, s_1\rangle t + \tfrac{1}{2}t^2\langle Ds_1, s_1\rangle,$$

при этом $\langle g_0, s_1\rangle = -\langle g_0, Bg_0\rangle$. Ясно, что минимум квадратного трёхчлена $q_1(t)$ с положительным старшим коэффициентом достигается в единственной точке $t = \tilde{t}_1 > 0$.

Пусть $k \geqslant 2$. Запишем

$$Q(x_{k-1} + ts_k) = Q(x_{k-1}) + \langle g_{k-1}, s_k\rangle t + \tfrac{1}{2}t^2\langle Ds_k, s_k\rangle.$$

По свойству метода сопряжённых градиентов $\langle \tilde{g}_{k-1}, \tilde{s}_{k-1}\rangle = 0$, так что $\langle g_{k-1}, s_{k-1}\rangle = 0$. Как следствие,

$$\langle g_{k-1}, s_k\rangle = \langle g_{k-1}, -Bg_{k-1} + \tilde{b}_{k-1}s_{k-1}\rangle = -\langle g_{k-1}, Bg_{k-1}\rangle.$$

Теперь так же, как и в случае $k = 1$, заключаем, что минимум квадратного трёхчлена $q_k(t)$ с положительным старшим коэффициентом достигается в единственной точке $t = \tilde{t}_k > 0$. □

5°. Матрицу B обычно используют, когда в задаче (1) матрица D плохо обусловлена. Напомним, что числом обусловленности симметричной положительно определённой матрицы D называется отношение её наибольшего собственного числа к наименьшему, то есть величина

$$\varkappa(D) = \frac{\lambda_{\max}(D)}{\lambda_{\min}(D)}.$$

О плохой обусловленности матрицы D говорят, когда значение $\varkappa(D)$ велико.

Найдём число обусловленности матрицы $A = B^{1/2}DB^{1/2}$. По определению

$$\varkappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Отметим, что матрица $B^{1/2}AB^{-1/2}$ имеет те же собственные числа, что и матрица A. Вместе с тем, $B^{1/2}AB^{-1/2} = BD$. Значит,

$$\varkappa(A) = \frac{\lambda_{\max}(BD)}{\lambda_{\min}(BD)}.$$

Число обусловленности матрицы A будет близко к единице, когда матрица B является хорошим приближением к D^{-1}. В этом случае имеет смысл перейти от решения задачи (1) к решению задачи (2).

Матрица B называется *предобуславливателем*.

6°. Для получения качественных предобуславливателей можно использовать метод Хотеллинга (см., например, [2, с. 193–194]).

Пусть B_0 — произвольная симметричная положительно определённая матрица порядка n и $R_0 = E - DB_0$. Построим рекуррентную последовательность матриц

$$B_k = B_{k-1}(E - R_{k-1}), \quad R_k = E - DB_k, \quad k = 1, 2, \ldots$$

ПРЕДЛОЖЕНИЕ 2. *Если* $\|R_0\| \leqslant q < 1$, *то*

$$\|B_s - D^{-1}\| \leqslant \|B_0\| \frac{q^{2^s}}{1-q}, \quad s = 1, 2, \ldots$$

Здесь $\|\cdot\|$ — любая матричная норма.

7°. Простейшим предобуславливателем является предобуславливатель Якоби:

$$B = \operatorname{diag}(d_{11}^{-1}, \ldots, d_{nn}^{-1}).$$

ЛИТЕРАТУРА

1. Малозёмов В. Н. *О методе сопряжённых градиентов* // Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г. (`http://dha.spb.ru/reps12.shtml#0428`) [Данная книга, с. 108]

2. Фаддеев Д. К., Фаддеева В. Н. *Вычислительные методы линейной алгебры.* 4-е изд. стер. СПб.: «ЛАНЬ», 2009. 736 с.

КВАЗИНЬЮТОНОВСКИЕ МЕТОДЫ БЕЗУСЛОВНОЙ МИНИМИЗАЦИИ*

В. Н. Малозёмов, Е. К. Чернэуцану

Аннотация. В докладе анализируется общая схема построения квазиньютоновских методов безусловной минимизации, предложенная Ю. М. Данилиным [1]. Особенность квазиньютоновских методов состоит в том, что они позволяют найти точку минимума выпуклой квадратичной функции от n переменных не более чем за n шагов.

1°. Начнём с минимизации квадратичной функции:

$$f(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \min_{x\in\mathbb{R}^n}.$$

Здесь D — симметричная положительно определённая матрица.

Точка минимума x_* функции $f(x)$ на $\mathbb{R}^n$ существует, единственна и допускает представление

$$x_* = -D^{-1}c.$$

Вычисление x_* по этой формуле можно заменить итерационной схемой, в которой вместо атрибутов квадратичной функции (матрицы D и вектора c) используются значения функции f и её градиента.

2°. Возьмём некоторую линейно независимую систему векторов $y_0, y_1, \ldots, y_{n-1}$ и произвольные векторы $x_0, x_1, \ldots, x_{n-1}$. Положим

$$r_i = f'(x_i + y_i) - f'(x_i) = Dy_i, \quad i \in 0 : n-1. \tag{1}$$

Очевидно, что векторы $r_0, r_1, \ldots, r_{n-1}$ линейно независимы. Запишем

$$f'(x_i) - f'(x_*) = D(x_i - x_*).$$

Умножим последнее равенство скалярно на y_i. Учитывая, что $f'(x_*) = \mathbb{O}$, получаем

$$\big\langle D(x_i - x_*), y_i\big\rangle = \big\langle f'(x_i), y_i\big\rangle$$

*Семинар «CNSA & NDO». Избранные доклады. 16 апреля 2015 г.

или (в силу симметричности матрицы D и (1))

$$\langle r_i, x_* \rangle = \langle r_i, x_i \rangle - \big\langle f'(x_i), y_i \big\rangle.$$

Обозначив $d_i = \langle r_i, x_i \rangle - \big\langle f'(x_i), y_i \big\rangle$, придём к системе линейных уравнений относительно x_*:

$$\langle r_i, x_* \rangle = d_i, \quad i \in 0 : n-1. \tag{2}$$

Таким образом, задача минимизации квадратичной функции сводится к решению системы линейных уравнений (2), матрица и правая часть которой зависят только от градиента целевой функции.

3°. Выбор точек x_i находится в нашем распоряжении. В качестве x_0 возьмём произвольную точку из $\mathbb{R}^n$, а выбор остальных точек x_k подчиним условию

$$\langle r_i, x_k \rangle = d_i, \quad i \in 0 : k-1;\ k = 1, \ldots, n. \tag{3}$$

Подробнее: вычислив r_0 и d_0, точку x_1 строим так, чтобы

$$\langle r_0, x_1 \rangle = d_0; \tag{4}$$

точка x_2 должна удовлетворять двум уравнениям

$$\langle r_0, x_2 \rangle = d_0, \quad \langle r_1, x_2 \rangle = d_1$$

и так далее. Очевидно, что $x_n = x_*$.

Положим для определённости

$$x_{k+1} = x_k + \alpha_k p_k, \quad k \in 0 : n-1. \tag{5}$$

Уравнение (4) примет вид

$$\langle r_0, x_0 + \alpha_0 p_0 \rangle = d_0$$

или (с учётом определения d_0)

$$\alpha_0 \langle r_0, p_0 \rangle = -\langle f'(x_0), y_0 \rangle.$$

В качестве p_0 можно взять произвольный ненулевой вектор.

Допустим, что вектор x_k удовлетворяет системе уравнений (3) и потребуем, чтобы $\langle r_i, x_{k+1} \rangle = d_i$ при $i \in 0 : k$. Согласно (5) это условие можно переписать в виде

$$\langle r_i, x_k + \alpha_k p_k \rangle = d_i, \quad i \in 0 : k. \tag{6}$$

При $i = k$ получаем

$$\alpha_k \langle r_k, p_k \rangle = -\langle f'(x_k), y_k \rangle. \tag{7}$$

Для выполнения условия (6) при $i \in 0 : k-1$ в силу (3) достаточно, чтобы

$$\langle r_i, p_k \rangle = 0, \quad i \in 0 : k-1;\ k \in 1 : n-1. \tag{8}$$

Условие (7) используется для определения α_k.

4°. Обратимся к соотношению (8). Будем искать p_k в виде

$$p_k = H_k w_k,$$

где H_k — квадратная матрица порядка n. В качестве w_k обычно берут антиградиент $-f'(x_k)$. Имеем

$$\langle H_k^T r_i, w_k \rangle = 0, \quad i \in 0 : k-1. \tag{9}$$

Чтобы получить условие, аналогичное (3), потребуем, чтобы вектор $H_k^T r_i$ не зависел от k:

$$H_k^T r_i = z_i, \quad i \in 0 : k-1. \tag{10}$$

Тогда условие (9) примет вид

$$\langle z_i, w_k \rangle = 0, \quad i \in 0 : k-1;\ k \in 1 : n-1. \tag{11}$$

В частности, при $z_i = y_i$ матрица H_k должна удовлетворять условию

$$H_k^T D y_i = y_i, \quad i \in 0 : k-1. \tag{12}$$

Обозначим через Y матрицу со столбцами $y_0, y_1, \ldots, y_{n-1}$. Для H_n получим матричное уравнение

$$H_n^T D Y = Y.$$

Отсюда следует, что

$$H_n^T = Y(DY)^{-1} = D^{-1}.$$

В связи с этим условие (12) называют *квазиньютоновским*, а условие (10) — *обобщённым квазиньютоновским*. Отметим, что в вычислениях матрица H_n не участвует (последней используется матрица H_{n-1}).

5°. Матрицы H_k будем вводить последовательно:

$$H_{k+1} = H_k + \Delta H_k, \quad k \in 0, 1, \ldots, n-2.$$

В качестве H_0 берётся произвольная матрица (обычно $H_0 = I$, где I — единичная матрица порядка n).

Соотношение (10) при $k = 1$ примет вид

$$(H_0 + \Delta H_0)^T r_0 = z_0,$$

откуда следует, что

$$(\Delta H_0)^T r_0 = z_0 - H_0^T r_0.$$

Допустим, что H_k удовлетворяет условию (10). Потребуем, чтобы $H_{k+1}^T r_i = z_i$ при $i \in 0:k$, то есть чтобы

$$(H_k + \Delta H_k)^T r_i = z_i, \quad i \in 0:k. \tag{13}$$

При $i = k$ получаем

$$(\Delta H_k)^T r_k = z_k - H_k^T r_k. \tag{14}$$

Для выполнения условия (13) при $i \in 0:k-1$ достаточно, чтобы

$$(\Delta H_k)^T r_i = \mathbb{O}, \quad i \in 0:k-1. \tag{15}$$

Таким образом, обобщённое квазиньютоновское условие (10) выполняется для последовательности матриц $H_{k+1} = H_k + \Delta H_k$, где H_0 — произвольная матрица, а ΔH_k удовлетворяет условию (14) при $k \in 0:n-2$ и условию (15) при $k \in 1:n-2$.

6°. Соотношениям (14) и (15) можно удовлетворить, например, следующим образом (*схема Хуанга*):

$$(\Delta H_k)^T = \frac{z_k u_k^T}{\langle u_k, r_k\rangle} - \frac{H_k^T r_k v_k^T}{\langle v_k, r_k\rangle}, \tag{16}$$

если

$$\begin{gathered} \langle u_k, r_k\rangle \neq 0,\ \langle v_k, r_k\rangle \neq 0, \quad k \in 0:n-2; \\ \langle u_k, r_i\rangle = 0,\ \langle v_k, r_i\rangle = 0, \quad i \in 0:k-1,\ k \in 1:n-2. \end{gathered} \tag{17}$$

Для того чтобы схема была полной, к (16) нужно добавить условия (7) и (11).

Обычно полагают

$$\begin{aligned} u_k &= t_{k1} z_k + t_{k2} H_k^T r_k, \\ v_k &= t_{k3} z_k + t_{k4} H_k^T r_k, \end{aligned}$$

где $t_{k1}, \ldots, t_{k4}$ — произвольные числа, обеспечивающие выполнение условий (17). Тем самым, конкретная реализация схемы Хуанга определяется матрицей второго порядка

$$T_k = \begin{pmatrix} t_{k1} & t_{k2} \\ t_{k3} & t_{k4} \end{pmatrix}.$$

7°. Положим в схеме Хуанга

$$T_k = \begin{pmatrix} \rho_k & -1 \\ 1 & 0 \end{pmatrix},$$

так что

$$\begin{aligned} u_k &= \rho_k z_k - H_k^T r_k, \\ v_k &= z_k. \end{aligned}$$

Параметр ρ_k выберем из условия равенства знаменателей в формуле (16), то есть из условия $\langle u_k, r_k\rangle = \langle v_k, r_k\rangle$. Имеем

$$\langle \rho_k z_k - H_k^T r_k, r_k\rangle = \langle z_k, r_k\rangle,$$

откуда следует, что

$$\rho_k = 1 + \frac{\langle H_k^T r_k, r_k\rangle}{\langle z_k, r_k\rangle}.$$

Формула (16) принимает вид

$$(\Delta H_k)^T = \frac{z_k(\rho_k z_k - H_k^T r_k)^T}{\langle z_k, r_k\rangle} - \frac{H_k^T r_k z_k^T}{\langle z_k, r_k\rangle} =$$
$$= \left(1 + \frac{\langle H_k^T r_k, r_k\rangle}{\langle z_k, r_k\rangle}\right)\frac{z_k z_k^T}{\langle z_k, r_k\rangle} - \frac{H_k^T r_k z_k^T + z_k r_k^T H_k}{\langle z_k, r_k\rangle}.$$

Отметим, что в правой части этого равенства стоит симметричная матрица.

Положим $H_0 = I$. Тогда матрицы H_k будут симметричными. При этом

$$H_{k+1} = H_k + \left(1 + \frac{\langle H_k^T r_k, r_k\rangle}{\langle z_k, r_k\rangle}\right)\frac{z_k z_k^T}{\langle z_k, r_k\rangle} - \frac{H_k^T r_k z_k^T + z_k r_k^T H_k}{\langle z_k, r_k\rangle} =$$
$$= \left(I - \frac{z_k r_k^T}{\langle z_k, r_k\rangle}\right) H_k \left(I - \frac{r_k z_k^T}{\langle z_k, r_k\rangle}\right) + \frac{z_k z_k^T}{\langle z_k, r_k\rangle} =$$
$$= Q_k^T H_k Q_k + \frac{z_k z_k^T}{\langle z_k, r_k\rangle}. \tag{18}$$

Здесь

$$Q_k = I - \frac{r_k z_k^T}{\langle z_k, r_k\rangle}.$$

Отметим, что $Q_k r_k = \mathbb{O}$.

Дальнейший выбор параметров осуществим так:

$$w_k = -f'(x_k), \quad p_k = H_k w_k, \quad z_k = y_k = \alpha_k p_k,$$

где α_k удовлетворяет условию (7). Учитывая, что

$$r_k = f'(x_k + \alpha_k p_k) - f'(x_k) = f'(x_{k+1}) - f'(x_k) = \alpha_k D p_k = D z_k,$$

перепишем условие (7) в виде

$$\alpha_k^2 \langle D p_k, p_k\rangle = -\alpha_k \big\langle f'(x_k), p_k\big\rangle.$$

В качестве α_k возьмём величину

$$\alpha_k = -\frac{\langle f'(x_k), p_k\rangle}{\langle Dp_k, p_k\rangle}. \tag{19}$$

Отметим, что α_k является точкой минимума функции

$$\varphi_k(\alpha) = f(x_k + \alpha p_k).$$

Действительно,

$$\begin{aligned}\varphi_k'(\alpha) &= \langle f'(x_k + \alpha p_k), p_k\rangle = \langle D(x_k + \alpha p_k) + c, p_k\rangle = \\ &= \langle f'(x_k), p_k\rangle + \alpha\langle Dp_k, p_k\rangle,\end{aligned}$$

поэтому уравнение $\varphi_k'(\alpha) = 0$ имеет единственное решение $\alpha = \alpha_k$.

Условие $\varphi_k'(\alpha_k) = 0$ можно записать в виде

$$\langle f'(x_{k+1}), p_k\rangle = 0. \tag{20}$$

8°. В дальнейшем будем использовать обозначение $f_k' = f'(x_k)$.

ТЕОРЕМА. *Если градиенты $f_0', f_1', \ldots, f_{k-1}'$ отличны от нуля, то*

1) $z_{k-1} \neq \mathbb{O}$;

2) $\langle f_k', z_i\rangle = 0$, $i \in 0 : k-1$;

3) H_k — *положительно определённая матрица;*

4) $H_k D z_i = z_i$, $i \in 0 : k-1$;

5) $\langle Dz_i, z_j\rangle = 0$ *при* $i \neq j$; $i, j \in 0 : k-1$.

Доказательство проведём индукцией по k. При $k = 1$ имеем $p_0 = -H_0 f_0' = -f_0' \neq \mathbb{O}$,

$$\alpha_0 = -\frac{\langle f_0', -f_0'\rangle}{\langle Dp_0, p_0\rangle} > 0,$$

так что $z_0 = \alpha_0 p_0 \neq \mathbb{O}$. Согласно (20), $\langle f_1', z_0\rangle = \alpha_0\langle f_1', p_0\rangle = 0$.

Покажем, что матрица H_1 положительно определена. Приняв во внимание формулу (18) и тот факт, что $\langle z_0, r_0\rangle = \langle z_0, Dz_0\rangle > 0$, запишем

$$\langle H_1 x, x\rangle = \|Q_0 x\|^2 + \frac{\langle z_0, x\rangle^2}{\langle z_0, r_0\rangle} \geqslant 0.$$

Допустим, что $\langle H_1 x, x\rangle = 0$. Тогда $\langle z_0, x\rangle = 0$ и $Q_0 x = \mathbb{O}$. Из последнего равенства следует, что

$$\mathbb{O} = x - \frac{r_0\langle z_0, x\rangle}{\langle z_0, r_0\rangle} = x,$$

то есть $x = \mathbb{O}$. Положительная определённость матрицы H_1 установлена.

Далее, $H_1 D z_0 = H_1 r_0 = z_0$, поскольку $Q_0 r_0 = \mathbb{O}$. Условие 5) при $k = 1$ бессодержательно. Чтобы и по нему создать базу индукции, рассмотрим случай $k = 2$.

По условию теоремы $f_1' \neq \mathbb{O}$ и по доказанному матрица H_1 положительно определена. Значит, $p_1 = -H_1 f_1' \neq \mathbb{O}$. Согласно (19),

$$\alpha_1 = -\frac{\langle f_1', -H_1 f_1'\rangle}{\langle D p_1, p_1\rangle} > 0,$$

так что $z_1 = \alpha_1 p_1 \neq \mathbb{O}$. Теперь имеем

$$\begin{aligned}\langle z_1, D z_0\rangle &= -\alpha_1\langle H_1 f_1', D z_0\rangle = -\alpha_1\langle f_1', H_1 D z_0\rangle = \\ &= -\alpha_1\langle f_1', z_0\rangle = 0\end{aligned}$$

и

$$\langle f_2', z_0\rangle = \langle f_1' + D z_1, z_0\rangle = \langle f_1', z_0\rangle = 0.$$

Равенство $\langle f_2', z_1\rangle = 0$ справедливо в силу (20).

Проверим, что матрица H_2 положительно определена. Запишем

$$\langle H_2 x, x\rangle = \langle H_1 Q_1 x, Q_1 x\rangle + \frac{\langle z_1, x\rangle^2}{\langle z_1, r_1\rangle} \geqslant 0.$$

Мы приняли во внимание, что $\langle z_1, r_1\rangle = \langle z_1, D z_1\rangle > 0$. Если $\langle H_2 x, x\rangle = 0$, то $Q_1 x = \mathbb{O}$ и $\langle z_1, x\rangle = 0$. Отсюда следует, что

$$\mathbb{O} = Q_1 x = x - \frac{r_1\langle z_1, x\rangle}{\langle z_1, r_1\rangle} = x,$$

то есть $x = \mathbb{O}$. Положительная определённость матрицы H_2 установлена.

На основании равенства $Q_1 r_1 = \mathbb{O}$ получаем

$$H_2 D z_1 = H_2 r_1 = z_1.$$

Далее

$$\begin{aligned}Q_1 D z_0 &= \left(I - \frac{r_1 z_1^T}{\langle z_1, r_1\rangle}\right) D z_0 = D z_0,\\ Q_1^T D z_0 &= \left(I - \frac{z_1 r_1^T}{\langle z_1, r_1\rangle}\right) z_0 = z_0 - \frac{\langle D z_1, z_0\rangle}{\langle z_1, r_1\rangle} = z_0,\end{aligned}$$

поэтому

$$H_2 Dz_0 = Q_1^T H_1 Q_1 Dz_0 = Q_1^T H_1 Dz_0 = Q_1^T z_0 = z_0.$$

При $k = 2$ теорема доказана.

Сделаем индукционный переход от k к $k+1$. Так как $f'_k \neq \mathbb{O}$ и H_k — положительно определённая матрица, то $p_k = -H_k f'_k \neq \mathbb{O}$, $\alpha_k > 0$, $z_k \neq \mathbb{O}$. Матрица H_{k+1} положительно определена, что следует из положительной определённости матрицы H_k и условия $z_k \neq \mathbb{O}$.

При $i \in 0 : k-1$ в силу условий 4) и 2) имеем

$$\langle Dz_i, z_k\rangle = -\alpha_k \langle H_k f'_k, Dz_i\rangle = -\alpha_k \langle f'_k, z_i\rangle = 0.$$

Согласно (20), $\langle f'_{k+1}, z_k\rangle = 0$. При $i \in 0 : k-1$

$$\langle f'_{k+1}, z_i\rangle = \langle f'_k + Dz_k, z_i\rangle = \langle f'_k, z_i\rangle = 0.$$

Наконец, так же, как при $k = 2$, проверяется, что $H_{k+1} Dz_k = z_k$ и $H_{k+1} Dz_i = Q_k^T z_i = z_i$ при $i \in 0 : k-1$.

Теорема доказана. □

9°. Теперь можно описать метод минимизации квадратичной функции $f(x)$ с положительно определённой матрицей D, в котором используются только значения функции $f(x)$ и её градиента $f'(x)$.

Возьмём произвольное начальное приближение x_0. Если $f'(x_0) = \mathbb{O}$, то x_0 — точка минимума. Процесс заканчивается. Иначе полагаем $H_0 = I$.

Общая $(k+1)$-я итерация, перед началом которой имеются x_k, $f'(x_k) \neq \mathbb{O}$ и H_k, состоит из следующих шагов:

- находим направление спуска $p_k = -H_k f'(x_k)$;
- вычисляем α_k как точку минимума функции $\varphi_k(\alpha) = f(x_k + \alpha p_k)$;
- определяем $z_k = \alpha_k p_k$ и $x_{k+1} = x_k + z_k$;
- если $f'(x_{k+1}) = \mathbb{O}$, то процесс заканчивается. Иначе находим $r_k = f'(x_{k+1}) - f'(x_k)$;
- осуществляем подготовку к следующему шагу

$$H_{k+1} = \left(I - \frac{z_k r_k^T}{\langle z_k, r_k\rangle}\right) H_k \left(I - \frac{r_k z_k^T}{\langle z_k, r_k\rangle}\right) + \frac{z_k z_k^T}{\langle z_k, r_k\rangle}.$$

Таким образом, построены x_{k+1}, $f'(x_{k+1}) \neq \mathbb{O}$ и H_{k+1}. Описание метода завершено.

Подчеркнём, что в формировании матрицы H_{k+1} наряду с матрицей H_k участвуют векторы $z_k = x_{k+1} - x_k$ и $r_k = f'(x_{k+1}) - f'(x_k)$.

Последовательность $z_0, z_1, \ldots$ по свойству 5) из теоремы состоит из D-ортогональных векторов, которые, в частности, являются линейно независимыми. Их количество не может превысить n. Значит, по крайней мере, $f'(x_n) = \mathbb{O}$. Другими словами, точка минимума будет найдена не более, чем за n итераций.

10°. Описанный метод называется BFGS-методом по именам его авторов Бройдена [2], Флетчера [3], Гольдфарба [4] и Шанно [5]. Он может применяться для минимизации не только квадратичных функций, но и для минимизации произвольных гладких и даже негладких функций. По этому поводу см. книгу [6].

ЛИТЕРАТУРА

1. Пшеничный Б. Н., Данилин Ю. М. *Численные методы в экстремальных задачах* М.: Наука, 1975.

2. Broyden C. G. *The convergence of a class of double-rank minimization algorithms* // Jornal of the Institute of Mathematics and Its Applications **6** (1970): 76–90.

3. Fletcher R. *A New Approach to Variable Metric Algorithms* // Computer Journal **13** (1970): 317–322.

4. Goldfarb D. *A Family of Variable Metric Updates Derived by Variational Means* // Mathematics of Computation **24** (1970): 23–26.

5. Shanno D. F. *Conditioning of quasi-Newton methods for function minimization* // Math. Comput. **24** (1970): 647–656.

6. Nocedal J., Wright S. J. *Numerical Optimization.* 2nd edition. USA: Springer, 2006.

МЕТОД СОПРЯЖЁННЫХ ГРАДИЕНТОВ В КВАДРАТИЧНОМ ПРОГРАММИРОВАНИИ*

В. Н. Малозёмов, Е. К. Чернэуцану

Памяти Б. Н. Пшеничного
(1937–2000)

Данный доклад является продолжением доклада [1]. Здесь анализируется метод сопряжённых градиентов минимизации выпуклой квадратичной функции при наличии линейных ограничений. В идейном плане мы следуем Б. Н. Пшеничному [2, с. 173–191].

1°. Напомним описание метода сопряжённых градиентов решения следующей экстремальной задачи

$$Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf_{x\in\mathbb{R}^n}, \tag{1}$$

где D — симметричная неотрицательно определённая матрица (см. [1]).

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \mathbb{R}^n$ и вычисляем градиент $g_0 = Dx_0 + c$. Если $g_0 = \mathbb{O}$, то x_0 — решение задачи (1). Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

k-й шаг. Пусть уже имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. Последовательно вычисляем

$$t_k = \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle},$$
$$x_k = x_{k-1} + t_k s_k,$$
$$g_k = g_{k-1} + t_k D s_k.$$

Если $g_k = \mathbb{O}$, то x_k — решение задачи (1). Вычисления прекращаются. В противном случае находим

$$b_k = \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle},$$
$$s_{k+1} = -g_k + b_k s_k.$$

*Семинар «DHA & CAGD». Избранные доклады. 26 мая 2012 г.

Описание метода закончено.

Если квадратичная функция $Q(x)$ ограничена снизу на $\mathbb{R}^n$, то по крайней мере на r-м шаге, где $r = \operatorname{rank} D$, получим решение задачи (1). Если же $Q(x)$ неограничена снизу, то при некотором k выполнится равенство $\langle Ds_k, s_k\rangle = 0$. В этом случае $Q(x)$ на луче $x = x_{k-1} + ts_k$, $t > 0$, стремится к $-\infty$ при $t \to +\infty$.

2°. Усложним задачу (1), введя линейные ограничения-равенства:

$$\begin{gathered} Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf, \\ Ax = b. \end{gathered} \tag{2}$$

Считаем, что D — симметричная неотрицательно определённая матрица, а A — произвольная $(m \times n)$-матрица. Множество планов задачи (2) обозначим ω.

Напомним критерий оптимальности [3, с. 110]: *вектор $x^* \in \omega$ является решением задачи* (2) *тогда и только тогда, когда найдётся вектор $u^* \in \mathbb{R}^m$, такой, что*

$$Q'(x^*) = A^T u^*.$$

Для дальнейшего нам потребуется дополнительное предположение. Будем считать, что строки матрицы A линейно независимы. В этом случае $\operatorname{rank} A = m$.

Введём подпространство $\mathcal{L} = \{x \in \mathbb{R}^n \mid Ax = \mathbb{O}\}$ и обозначим через P матрицу ортогонального проектирования на $\mathcal{L}$. Как известно [4, с. 53],

$$P = E - A^T(AA^T)^{-1}A,$$

где E — единичная матрица порядка n. Матрица P обладает следующими свойствами:

1) $AP = 0, \quad PA^T = 0$;

2) $P^T = P, \quad PP = P$;

3) $\operatorname{rank} P = n - m$.

Сведём задачу (2) к задаче без ограничений. Возьмём произвольный план $x_0 \in \omega$ и сделаем замену переменных $x = x_0 + Py$. Рассмотрим функцию

$$\begin{gathered} q(y) := Q(x_0 + Py) = Q(x_0) + \langle Dx_0 + c, Py\rangle + \tfrac{1}{2}\langle DPy, Py\rangle = \\ = \tfrac{1}{2}\langle PDPy, y\rangle + \big\langle P(Dx_0 + c), y\big\rangle + Q(x_0). \end{gathered}$$

Видим, что $q(y)$ — квадратичная функция с симметричной неотрицательно определённой матрицей PDP. Отметим (см. [4, с. 15]), что

$$\operatorname{rank} PDP \leqslant \min\{\operatorname{rank} P, \operatorname{rank} D\} \leqslant n - m. \tag{3}$$

Кроме того,

$$q'(y) = PDPy + P(Dx_0 + c) = P\big[D(x_0 + Py) + c\big] = P[Dx + c] = PQ'(x). \quad (4)$$

Допустим, что y^* — точка минимума $q(y)$ на $\mathbb{R}^n$. Покажем, что точка $x^* = x_0 + Py^*$ является решением задачи (2). Имеем

$$Ax^* = Ax_0 + (AP)y^* = Ax_0 = b,$$

то есть $x^* \in \omega$. Далее, из условия $q'(y^*) = \mathbb{O}$ в силу (4) следует, что $PQ'(x^*) = \mathbb{O}$ или

$$Q'(x^*) - A^T(AA^T)^{-1}AQ'(x^*) = \mathbb{O}.$$

Положив

$$u^* = (AA^T)^{-1}AQ'(x^*),$$

получим

$$Q'(x^*) = A^T u^*.$$

По критерию оптимальности, x^* — решение задачи (2).

Таким образом, задача (2) сводится к минимизации выпуклой квадратичной функции $q(y)$ на $\mathbb{R}^n$.

3°. Запишем метод сопряжённых градиентов для минимизации $q(y)$ на $\mathbb{R}^n$. Предварительно найдём матрицу ортогонального проектирования P. Напомним, что в определение $q(y)$ входит некоторый план x_0 задачи (2).

Нулевой шаг. В качестве начального приближения возьмём вектор $y_0 = \mathbb{O}$. Вычислим градиент $g_0 = P(Dx_0 + c)$. Если $g_0 = \mathbb{O}$, то y_0 — точка минимума. Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

k-й шаг. Пусть уже имеются y_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. Вычисляем

$$t_k = \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle PDPs_k, s_k\rangle}, \quad (5)$$

$$y_k = y_{k-1} + t_k s_k,$$

$$g_k = g_{k-1} + t_k PDPs_k. \quad (6)$$

Если $g_k = \mathbb{O}$, то y_k — точка минимума. Вычисления прекращаются. В противном случае находим

$$b_k = \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle},$$

$$s_{k+1} = -g_k + b_k s_k.$$

Описание метода закончено.

На самом деле, формулы (5) и (6) допускают упрощение. Покажем, что

$$Ps_k = s_k, \quad k = 1, 2, \ldots \tag{7}$$

Имеем

$$Ps_k = s_k - A^T(AA^T)^{-1}(As_k). \tag{8}$$

По построению s_k есть линейная комбинация градиентов g_0, $g_1, \ldots, g_{k-1}$. Вместе с тем, согласно (4)

$$Ag_i = Aq'(y_i) = (AP)\big[D(x_0 + Py_i) + c\big] = 0.$$

Отсюда и из (8) следует (7).

Теперь формулы (5) и (6) принимают более простой вид:

$$t_k = \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle},$$
$$g_k = g_{k-1} + t_k PDs_k.$$

4°. Обозначим $x_k = x_0 + Py_k$. Очевидно, что $x_k \in \omega$ и согласно (7)

$$x_k - x_{k-1} = P(y_k - y_{k-1}) = t_k s_k.$$

Перепишем алгоритм предыдущего пункта в терминах x_k.

Предварительно находим матрицу ортогонального проектирования $P = E - A^T(AA^T)^{-1}A$.

Нулевой шаг. Берём произвольное начальное приближение $x_0 \in \omega$ и вычисляем $g_0 = P(Dx_0+c)$. Если $g_0 = \mathbb{O}$, то x_0 — решение задачи (2). Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

***k*-й шаг.** Пусть уже имеются x_{k-1}, $g_{k-1} \neq \mathbb{O}$, s_k. Последовательно вычисляем

$$t_k = \frac{\langle g_{k-1}, g_{k-1}\rangle}{\langle Ds_k, s_k\rangle},$$
$$x_k = x_{k-1} + t_k s_k,$$
$$g_k = g_{k-1} + t_k PDs_k.$$

Если $g_k = \mathbb{O}$, то x_k — решение задачи (2). Вычисления прекращаются. В противном случае находим

$$b_k = \frac{\langle g_k, g_k\rangle}{\langle g_{k-1}, g_{k-1}\rangle},$$
$$s_{k+1} = -g_k + b_k s_k.$$

Описание алгоритма решения задачи (2) закончено.

Если квадратичная функция $Q(x)$ ограничена снизу на ω, то квадратичная функция $q(y)$ ограничена снизу на $\mathbb{R}^n$. В силу (3), по крайней мере на $(n-m)$-м шаге данного алгоритма получим решение задачи (2). Из неограниченности $Q(x)$ снизу на ω следует неограниченность снизу на $\mathbb{R}^n$ функции $q(y)$ (иначе задача (2) имела бы решение). В этом случае при некотором k выполнится равенство $\langle Ds_k, s_k\rangle = 0$, которое гарантирует убывание $Q(x)$ на луче $x = x_{k-1} + ts_k$, $t > 0$, до $-\infty$ при $t \to +\infty$.

5°. В предыдущем пункте мы описали метод сопряжённых градиентов решения задачи (2). Он требует выбора начального приближения $x_0 \in \omega$, то есть решения системы линейных уравнений $Ax = b$. Покажем, что решение такой системы при любой $(m \times n)$-матрице A можно получить с помощью метода сопряжённых градиентов.

Рассмотрим задачу квадратичного программирования

$$\varphi(x) := \tfrac{1}{2}\|Ax - b\|^2 \to \min_{x\in\mathbb{R}^n}. \tag{9}$$

Запишем квадратичную функцию $\varphi(x)$ в стандартной форме:

$$\varphi(x) = \tfrac{1}{2}\langle Ax - b, Ax - b\rangle = \tfrac{1}{2}\langle A^TAx, x\rangle - \langle A^Tb, x\rangle + \tfrac{1}{2}\|b\|^2.$$

Матрица A^TA симметрична и неотрицательно определена. Кроме того, $\varphi(x) \geqslant$ $\geqslant 0$ при всех $x \in \mathbb{R}^n$. Значит, задача (9) имеет решение. Оно может быть получено методом сопряжённых градиентов. Для этого потребуется не более m шагов, поскольку $\operatorname{rank} A^TA \leqslant m$.

Пусть x_0 — решение задачи (9). Если $\varphi(x_0) = 0$, то $x_0 \in \omega$. В противном случае система $Ax = b$ решений не имеет ($\omega = \varnothing$). ‘

6°. Переходим к общей задаче квадратичного программирования

$$\begin{gathered} Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf, \\ \langle a_j, x\rangle \geqslant b_j, \quad j \in M_1; \\ \langle a_j, x\rangle = b_j, \quad j \in M_2. \end{gathered} \tag{10}$$

Считаем, что матрица D симметрична и неотрицательно определена. Множество планов задачи (10) обозначим Ω.

При фиксированном $x \in \Omega$ введём индексные множества активных ограничений

$$M_1(x) = \big\{j \in M_1 \mid \langle a_j, x\rangle = b_j\big\},$$
$$J(x) = M_1(x) \cup M_2.$$

Напомним критерий оптимальности [3, с. 88]: *вектор $x^* \in \Omega$ является решением задачи* (10) *тогда и только тогда, когда найдутся множители Лагранжа u_j^*, $j \in J(x^*)$, со свойствами*

$$Q'(x^*) = \sum_{j \in J(x^*)} u_j^* a_j,$$
$$u_j^* \geqslant 0, \quad j \in M_1(x^*).$$

В дальнейшем будем предполагать, что выполнено условие регулярности ограничений: *при всех $x \in \Omega$ векторы a_j, $j \in J(x)$, линейно независимы.*

7°. Опишем метод решения задачи (10), основанный на переборе активных ограничений.

В качестве начального приближения возьмём произвольный вектор $x_0 \in \Omega$. Обозначим $J_0 = J(x_0)$, так что

$$\langle a_j, x_0 \rangle > b_j \quad \text{при} \quad j \notin J_0. \tag{11}$$

Из строк a_j, $j \in J_0$, составим матрицу A_{J_0}. Найдём матрицу ортогонального проектирования

$$P_{J_0} = E - A_{J_0}^T (A_{J_0} A_{J_0}^T)^{-1} A_{J_0}$$

и вектор

$$u^0 = (A_{J_0} A_{J_0}^T)^{-1} A_{J_0} Q'(x_0).$$

В этом случае

$$P_{J_0} Q'(x_0) = Q'(x_0) - A_{J_0}^T u^0.$$

Имеются две возможности.

I) $P_{J_0} Q'(x_0) = \mathbb{O}$. По критерию оптимальности из п. 2° получается, что x_0 доставляет минимум квадратичной функции $Q(x)$ при ограничениях-равенствах

$$\langle a_j, x \rangle = b_j \quad j \in J_0.$$

Если к тому же $u_j^0 \geqslant 0$ при $j \in M_1(x_0)$, то, по критерию оптимальности из п. 6°, x_0 — решение задачи (10). Вычисления прекращаются.

Допустим, что нашёлся индекс $j_0 \in M_1(x_0)$, на котором $u_{j_0}^0 < 0$. Это значит, что ограничение $\langle a_{j_0}, x \rangle = b_{j_0}$ препятствует уменьшению функции $Q(x)$ на Ω. Уберём его. Введём индексное множество $J_0' = J_0 \setminus \{j_0\}$ и применим метод сопряжённых градиентов для минимизации $Q(x)$ при ограничениях

$$\langle a_j, x \rangle = b_j, \quad j \in J_0'. \tag{12}$$

В качестве начального приближения возьмём известную точку x_0. По алгоритму

$$x_1 = x_0 + t_1 s_1, \tag{13}$$

где $s_1 = -g_0$, $g_0 = P_{J_0'} Q'(x_0)$.

ПРЕДЛОЖЕНИЕ 1. *Вектор g_0 отличен от нулевого и*

$$\langle a_{j_0}, s_1\rangle = -(u^0_{j_0})^{-1}\|g_0\|^2 > 0, \tag{14}$$

$$\langle a_j, s_1\rangle = 0 \quad \textit{при} \quad j \in J'_0. \tag{15}$$

Доказательство. Имеем

$$g_0 = Q'(x_0) - A^T_{J'_0}v^0, \quad \text{где} \quad v^0 = (A_{J'_0}A^T_{J'_0})^{-1}A_{J'_0}Q'(x_0).$$

Вместе с тем, по условию I) $Q'(x_0) = A^T_{J_0}u^0$. Получаем

$$g_0 = A^T_{J_0}u^0 - A^T_{J'_0}v^0.$$

В правой части этого равенства стоит линейная комбинация столбцов a_j, $j \in J_0$, причём коэффициент при a_{j_0}, равный $u^0_{j_0}$, отличен от нуля. В силу условия регулярности ограничений векторы a_j, $j \in J_0$, линейно независимы. Значит, $g_0 \neq \mathbb{O}$.

Далее, перепишем условие I) в виде

$$Q'(x_0) - \sum_{j\in J'_0} u^0_j a_j - u^0_{j_0}a_{j_0} = \mathbb{O}.$$

Умножим это равенство скалярно на $s_1 = -g_0$. Получим

$$u^0_{j_0}\langle a_{j_0}, s_1\rangle = \big\langle Q'(x_0), s_1\big\rangle - \sum_{j\in J'_0} u^0_j\langle a_j, s_1\rangle. \tag{16}$$

Запишем

$$\langle a_j, s_1\rangle = -\big\langle a_j, P_{J'_0}Q'(x_0)\big\rangle = -\big\langle P_{J'_0}a_j, Q'(x_0)\big\rangle.$$

По свойству матрицы ортогонального проектирования $P_{J'_0}A^T_{J'_0} = 0$, поэтому $P_{J'_0}a_j = \mathbb{O}$ при $j \in J'_0$. При тех же j выполнится равенство $\langle a_j, s_1\rangle = 0$, что соответствует (15).

Теперь формула (16) принимает вид

$$u^0_{j_0}\langle a_{j_0}, s_1\rangle = \big\langle Q'(x_0), s_1\big\rangle. \tag{17}$$

Отметим, что

$$\begin{aligned}\big\langle Q'(x_0), s_1\big\rangle &= -\big\langle Q'(x_0), P_{J'_0}Q'(x_0)\big\rangle = -\big\langle Q'(x_0), P_{J'_0}P_{J'_0}Q'(x_0)\big\rangle = \\ &= -\big\langle P_{J'_0}Q'(x_0), P_{J'_0}Q'(x_0)\big\rangle = -\|g_0\|^2.\end{aligned} \tag{18}$$

Отсюда и из (17) следует (14).

Предложение доказано. □

8°. Вернёмся к формуле (13). Условие $g_0 \neq \mathbb{O}$ требуется при описании метода сопряжённых градиентов для перехода к точке x_1. В то же время неравенство (14) гарантирует, что в точках $x(t) = x_0 + ts_1$ при $t > 0$ ограничение с индексом j_0, активным в точке x_0, становится неактивным. Действительно,

$$\langle a_{j_0}, x(t)\rangle = \langle a_{j_0}, x_0\rangle + t\langle a_{j_0}, s_1\rangle = b_{j_0} + t\langle a_{j_0}, s_1\rangle > b_{j_0}. \tag{19}$$

Отметим, что шаг t_1 может выводить точку x_1 из Ω. Мы же хотим, чтобы все точки минимизирующей последовательности содержались в Ω. В этой связи введём ограничитель шага

$$\hat{t}_1 = \min_{j\in\hat{J}_0} \frac{\langle a_j, x_0\rangle - b_j}{-\langle a_j, s_1\rangle}, \tag{20}$$

где $\hat{J}_0 = \{j \in M_1 \cup M_2 \mid j \notin J_0',\ \langle a_j, s_1\rangle < 0\}$. (Если $\hat{J}_0 = \varnothing$, то по определению $\hat{t}_1 = +\infty$.)

ПРЕДЛОЖЕНИЕ 2. *Справедливо неравенство*

$$\hat{t}_1 > 0.$$

Доказательство. Достаточно проверить, что при всех $j \in \hat{J}_0$ дробь из правой части формулы (20) положительна.

Возьмём $j \in \hat{J}_0$. Тогда $j \notin J_0'$, так что либо $j = j_0$, либо $j \notin J_0$ (см. рис.).

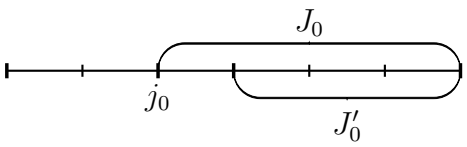


Рис.

Кроме того, $\langle a_j, s_1\rangle < 0$. Равенство $j = j_0$ невозможно в силу (14). Значит, $j \notin J_0$. Остаётся сослаться на неравенство (11) □

ПРЕДЛОЖЕНИЕ 3. *Если $t_1 < \hat{t}_1$, то точка x_1 вида* (13) *принадлежит* Ω *и*

$$J(x_1) = J_0'.$$

Доказательство. При $j \in J_0'$ согласно (15) имеем

$$\langle a_j, x_1\rangle = \langle a_j, x_0\rangle + t_1\langle a_j, s_1\rangle = \langle a_j, x_0\rangle = b_j,$$

то есть ограничения задачи (10) в точке x_1 при $j \in J_0'$ активны. Проверим, что при $j \notin J_0'$ ограничения в точке x_1 неактивны.

Возьмём $j \notin J_0'$ и предположим сначала, что $\langle a_j, s_1\rangle \geqslant 0$. При $j = j_0$ неравенство $\langle a_j, x_1\rangle > b_j$ следует из (19), а при $j \notin J_0$ — из (11).

Пусть теперь $j \notin J_0'$ и $\langle a_j, s_1\rangle < 0$, то есть $j \in \hat{J}_0$. На основании определения $\hat{t}_1$ и условия $t_1 < \hat{t}_1$ заключаем, что

$$\langle a_j, x_0\rangle - b_j \geqslant \hat{t}_1\big(-\langle a_j, s_1\rangle\big) > t_1\big(-\langle a_j, s_1\rangle\big).$$

Это неравенство приводится к виду $\langle a_j, x_1\rangle > b_j$.

Установлено, что $x_1 \in \Omega$ и $J(x_1) = J_0'$. Предложение доказано. □

9°. Снова вернёмся к формуле (13). Если $t_1 < \hat{t}_1$, то согласно предложению 3 точка x_1 принадлежит Ω и $J(x_1) = J_0'$. Применение метода сопряжённых градиентов для минимизации квадратичной функции $Q(x)$ при ограничениях (12) можно продолжить. Если же $t_1 \geqslant \hat{t}_1$, то полагаем $x_1 = x_0 + \hat{t}_1 s_1$, и точку x_1 принимаем за новое начальное приближение. В последнем случае множество $J(x_1)$ состоит из J_0' и тех индексов из $\hat{J}_0$, на которых в правой части (20) достигается минимум (множество активных ограничений расширяется). При этом $Q(x_1) < Q(x_0)$, что проверяется следующим образом. Выпуклая квадратичная функция одной переменной

$$\varphi(t) := Q(x_0 + ts_1) = Q(x_0) - t\|g_0\|^2 + \tfrac{1}{2}\langle Ds_1, s_1\rangle$$

достигает глобального минимума при $t = t_1$ (мы считаем, что $\langle Ds_1, s_1\rangle > 0$) и на отрезке $[0,\, t_1]$ строго убывает. В частности, $\varphi(0) > \varphi(\hat{t}_1)$, что равносильно требуемому неравенству.

В результате описанных действий мы либо найдём точку минимума $Q(x)$ при ограничениях (12) с множеством индексов активных ограничений J_0', либо на некоторой итерации метода сопряжённых градиентов выйдем на ограничитель шага

$$\hat{t}_k = \min_{j \in \hat{J}_{k-1}} \frac{\langle a_j, x_{k-1}\rangle - b_j}{-\langle a_j, s_k\rangle}, \tag{21}$$

где $\hat{J}_{k-1} = \big\{j \in M_1 \cup M_2 \mid j \notin J_0',\, \langle a_j, s_k\rangle < 0\big\}$. При $t_k \geqslant \hat{t}_k$ положим $x_k = x_{k-1} + \hat{t}_k s_k$. В обоих случаях полученную точку принимаем за очередное начальное приближение и работаем с ней как с x_0.

10°. До сих пор предполагалось, что выполнено условие I): $P_{J_0}Q'(x_0) = \mathbb{O}$. Рассмотрим вторую возможность.

II) $P_{J_0}Q'(x_0) \neq \mathbb{O}$. В этом случае методом сопряжённых градиентов решаем задачу минимизации $Q(x)$ при ограничениях

$$\langle a_j, x\rangle = b_j, \quad j \in J_0, \tag{22}$$

начиная с x_0. На каждой итерации вычисляем ограничитель шага $\hat{t}_k$ по формуле (21), заменив в определении $\hat{J}_{k-1}$ индексное множество J_0' на J_0. Если $t_k \geqslant \hat{t}_k$, то полагаем $x_k = x_{k-1} + \hat{t}_k s_k$ и принимаем эту точку за новое начальное приближение. Если же мы ни разу не выйдем на ограничитель шага, то методом сопряжённых градиентов получим точку минимума $Q(x)$ при ограничениях (22), которую примем за новое начальное приближение и будем работать с ней как с x_0.

Замечание 1. По ходу процесса может оказаться, что $\langle Ds_k, s_k\rangle = 0$. Тогда функция $Q(x)$ на луче $x = x_{k-1} + ts_k$, $t > 0$, стремится к $-\infty$ при $t_k \to +\infty$. Формально положим $t_k = +\infty$. Если при этом и $\hat{t}_k = +\infty$, то функция $Q(x)$ неограничена снизу на Ω.

11°. Описанным методом будет построена последовательность планов задачи (10), на которой целевая функция $Q(x)$ строго убывает. Покажем, что эта последовательность конечна.

Если минимизация на аффинном множестве идёт до конца, то соответствующее x_k будет точкой минимума $Q(x)$ при ограничениях

$$\langle a_j, x\rangle = b_j, \quad j \in J(x_k). \tag{23}$$

Выходить на ограничитель шага последовательно бесконечное число раз мы не можем, поскольку при этом расширяется множество активных ограничений. Значит, по ходу процесса мы систематически будем получать точки, в которых достигается минимум $Q(x)$ при ограничениях вида (23). Ясно, что таких точек конечное число и они не могут повторяться в силу строгой монотонности процесса.

Последний элемент построенной последовательности будет решением задачи (10). По описанию это единственное условие выхода из процесса (если попутно не выяснится, что задача (10) не имеет решения).

12°. Мы рассмотрели лишь принципиальную схему решения задачи квадратичного программирования (10). Вопроса организации вычислений мы не касались.

13°. Сделаем два заключительных замечания.

Замечание 2. В методе сопряжённых градиентов приходится вычислять произведение матрицы ортогонального проектирования P_{J_0} на вектор. Покажем, что вектор вида $P_{J_0}d$ можно найти без вычисления матрицы P_{J_0}. Методом сопряжённых градиентов решим вспомогательную задачу

$$\tfrac{1}{2}\|d - A_{J_0}^T u\| \to \min_{u\in\mathbb{R}^m}. \tag{24}$$

В данном случае критерий оптимальности имеет вид

$$A_{J_0}A_{J_0}^T u = A_{J_0}d,$$

так что решением задачи (24) является вектор

$$u_0 = (A_{J_0} A_{J_0}^T)^{-1} A_{J_0} d.$$

Получаем

$$P_{J_0} d = d - A_{J_0}^T (A_{J_0} A_{J_0}^T)^{-1} A_{J_0} d = d - A_{J_0}^T u_0.$$

Замечание 3. Начальное приближение $x_0 \in \Omega$ при решении задачи (10) можно искать, используя идеи из линейного программирования. Например, привести ограничения задачи (10) к каноническому виду и найти начальный базисный план [5].

ЛИТЕРАТУРА

1. Малозёмов В. Н. *О методе сопряжённых градиентов* // Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г. (`http://dha.spb.ru/reps12.shtml#0428`) [Данная книга, с. 108]

2. Пшеничный Б. Н., Данилин Ю. М. *Численные методы в экстремальных задачах.* М.: Наука, 1975. 320 с.

3. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

4. Малозёмов В. Н. *Линейная алгебра без определителей. Квадратичная функция.* СПб.: Изд-во СПбГУ, 1997. 80 с.

5. Малозёмов В. Н. *Модифицированный симплекс-метод* // Семинар «DHA & CAGD». Избранные доклады. 20 ноября 2010 г. (`http://dha.spb.ru/reps10.shtml#1120`) [Данная книга, с. 15]

ПРОЕКТИРОВАНИЕ ТОЧКИ НА ПОДПРОСТРАНСТВО И НА СТАНДАРТНЫЙ СИМПЛЕКС*

В. Н. Малозёмов

В докладе на двух примерах показывается, чем различаются классические и неклассические задачи оптимизации.

1°. Под классической задачей оптимизации понимается задача минимизации гладкой функции при наличии ограничений-равенств. Простейшим примером такой задачи является задача ортогонального проектирования точки на подпространство. Она формализуется следующим образом [1, с. 52–55]:

$$\begin{gathered} Q(x) := \tfrac{1}{2}\|x-c\|^2 \to \min, \\ A[M,N]\times x[N] = \mathbb{O}[M]. \end{gathered} \tag{1}$$

Здесь $c \in \mathbb{R}^N$ — фиксированная точка и $A = A[M,N]$ — матрица с линейно независимыми строками. Ограничения задачи (1) определяют подпространство L пространства $\mathbb{R}^N$. Речь идёт о нахождении точки $x_* \in L$, ближайшей к точке c, или, другими словами, об ортогональном проектировании точки c на подпространство L.

Целевая функция задачи (1) представляет собой выпуклую квадратичную функцию, что следует из представления

$$Q(x) = \tfrac{1}{2}\langle x-c, x-c\rangle = \tfrac{1}{2}\langle Ex, x\rangle - \langle c, x\rangle + \tfrac{1}{2}\|c\|^2.$$

Отметим также, что $Q'(x) = x - c$. Запишем критерий оптимальности для задачи (1) [2, с. 90–91]: для того чтобы план $x_* \in L$ был оптимальным, необходимо и достаточно, чтобы нашёлся вектор $u_* = u_*[M]$ со свойством

$$x_* - c = A^T u_*.$$

Таким образом, решение задачи (1) сводится к решению системы линейных уравнений

$$A^T u = x - c, \tag{2}$$

$$Ax = \mathbb{O}. \tag{3}$$

*Семинар «DHA & CAGD». Избранные доклады. 28 февраля 2013 г.

Эта система (относительно x и u) имеет единственное решение. Чтобы найти его, умножим уравнение (2) слева на матрицу A. Учитывая (3), получаем

$$(AA^T)u = -Ac. \tag{4}$$

Покажем, что квадратная матрица AA^T обратима. Для этого достаточно проверить, что однородная система $(AA^T)u = \mathbb{O}$ имеет только нулевое решение. Возьмём произвольное решение u_0 этой системы. Запишем

$$\mathbb{O} = \langle AA^T u_0, u_0\rangle = \langle A^T u_0, A^T u_0\rangle = \|A^T u_0\|^2.$$

Значит, $A^T u_0 = \mathbb{O}$. По условию строки матрицы A линейно независимы. Это гарантирует линейную независимость столбцов матрицы A^T. Из линейной независимости и условия $A^T u_0 = \mathbb{O}$ следует, что $u_0 = \mathbb{O}$. Тем самым, установлено, что матрица AA^T имеет обратную.

Умножим уравнение (4) слева на матрицу $(AA^T)^{-1}$. Получим

$$u_* = -(AA^T)^{-1}Ac.$$

Подставив это в (2), придём к решению задачи (1):

$$x_* = c + A^T u_* = c - A^T(AA^T)^{-1}Ac. \tag{5}$$

Обозначим

$$P = E - A^T(AA^T)^{-1}A. \tag{6}$$

Тогда формулу (5) можно переписать так:

$$x_* = Pc. \tag{7}$$

Матрица P называется *матрицей ортогонального проектирования на подпространство* L.

Подведём итог.

ТЕОРЕМА 1. *Единственное решение задачи ортогонального проектирования произвольной точки* $c \in \mathbb{R}^N$ *на подпространство* L *определяется формулой* (7)*, в которой* P *— матрица ортогонального проектирования вида* (6)*.*

В частном случае, когда подпространство L задаётся одним уравнением $\langle a, x\rangle = 0$, где a — единичная вектор-строка, $\|a\| = 1$, матрица P принимает вид

$$P = E - a^T a.$$

2°. Остановимся на основных свойствах матрицы ортогонального проектирования.

ТЕОРЕМА 2. *Матрица P вида* (6) *обладает следующими свойствами:*

1) $PA^T = 0$, $PP = P$;

2) *матрица P симметрична и неотрицательно определена;*

3) *ранг матрицы P равен $|N| - |M|$.*

Доказательство. Согласно (6) имеем

$$PA^T = A^T - A^T(AA^T)^{-1}AA^T = 0.$$

Учитывая это равенство, получаем также

$$PP = P(E - A^T(AA^T)^{-1}A) = P - (PA^T)(AA^T)^{-1}A = P.$$

Симметричность матрицы P следует из её определения. Неотрицательная определённость проверяется так:

$$\langle Px, x\rangle = \langle PPx, x\rangle = \langle Px, Px\rangle = \|Px\|^2 \geqslant 0.$$

Чтобы найти ранг матрицы P, введём подпространство

$$\mathcal{P} = \{x \in \mathbb{R}^N \mid Px = \mathbb{O}\}.$$

Как известно (см., например, [1, с. 22]), для размерности этого подпространства справедлива формула.

$$\dim \mathcal{P} = |N| - \operatorname{rank} P. \tag{8}$$

Согласно свойству 1) столбцы матрицы A^T принадлежат $\mathcal{P}$ и по условию задачи (1) они линейно независимы. Покажем, что любой вектор из $\mathcal{P}$ можно представить в виде линейной комбинации столбцов матрицы A^T.

Пусть $x_0 \in \mathcal{P}$. Согласно (6) имеем

$$\mathbb{O} = Px_0 = x_0 - A^T(AA^T)^{-1}Ax_0.$$

Обозначим $\lambda_0 = (AA^T)^{-1}Ax_0$. Тогда $x_0 = A^T\lambda_0$. Это и означает, что x_0 есть линейная комбинация столбцов матрицы A^T.

Установлено, что $|M|$ линейно независимых столбцов матрицы A^T являются базисом подпространства $\mathcal{P}$. По определению размерности линейного множества получаем $\dim \mathcal{P} = |M|$.

Теперь свойство 3) следует из формулы (8). □

3°. К неклассическим относятся экстремальные задачи, у которых в ограничениях присутствуют неравенства. В качестве примера рассмотрим задачу проектирования точки на стандартный симплекс. Задача формализуется так (см. [3]):

$$\begin{gathered} Q(x) := \tfrac{1}{2}\|x-c\|^2 \to \min, \\ \sum_{j=1}^{n} x_j = 1; \\ x_j \geqslant 0,\ j \in 1:n. \end{gathered} \tag{9}$$

В данном случае $N = 1:n$. Матрица ограничений задачи (9) имеет вид

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Ограничению-равенству сопоставим двойственную переменную λ, а ограничению-неравенству $x_j \geqslant 0$ — двойственную переменную u_j, $j \in 1:n$. Запишем критерий оптимальности [2, с. 91]: для того чтобы план x^* задачи (9) был оптимальным, необходимо и достаточно, чтобы нашлись числа $\lambda^*, u_1^*, \dots, u_n^*$, такие, что при всех $j \in 1:n$ выполняются соотношения (условия Куна-Таккера):

$$\begin{gathered} x_j^* - c_j = \lambda^* + u_j^*; \\ u_j^* x_j^* = 0,\ u_j^* \geqslant 0. \end{gathered}$$

Таким образом, решение задачи (9) сводится к решению следующей системы равенств и неравенств

$$\lambda + c_j = x_j - u_j,\ j \in 1:n; \tag{10}$$

$$u_j x_j = 0,\ u_j \geqslant 0,\ x_j \geqslant 0,\ j \in 1:n; \tag{11}$$

$$\sum_{j=1}^{n} x_j = 1. \tag{12}$$

Согласно (10) и (11) имеем

$$|\lambda + c_j| = |x_j - u_j| = x_j + u_j,\ j \in 1:n. \tag{13}$$

Последнее равенство очевидно при $u_j = 0$. В силу условия дополнительности $u_j x_j = 0$ оно выполняется и при $u_j > 0$.

Сложим равенства (10) и (13). Получим

$$x_j = \tfrac{1}{2}\big(\lambda + c_j + |\lambda + c_j|\big). \tag{14}$$

С помощью функции $t_+ = \frac{1}{2}\big(t + |t|\big)$ формулу (14) можно переписать в виде

$$x_j = (\lambda + c_j)_+, \quad j \in 1:n. \tag{15}$$

По определению

$$t_+ = \begin{cases} 0 & \text{при } t \leqslant 0, \\ t & \text{при } t \geqslant 0. \end{cases} \tag{16}$$

Значит,

$$x_j = \begin{cases} 0, & \text{если } \lambda + c_j < 0; \\ \lambda + c_j, & \text{если } \lambda + c_j \geqslant 0. \end{cases}$$

Ясно также, что

$$u_j = \tfrac{1}{2}\big(|\lambda + c_j| - (\lambda + c_j)\big), \quad j \in 1:n.$$

Найденные x_j и u_j удовлетворяют условиям (10) и (11) при всех λ. Остаётся уравнение (12), которое в силу (15) принимает вид

$$\varphi(\lambda) := \sum_{j=1}^{n} (\lambda + c_j)_+ = 1. \tag{17}$$

Покажем, что это уравнение имеет единственное решение.

Поменяем знаки у компонент c_j проектируемой точки c и числа $\{-c_j\}$ упорядочим по неубыванию. Получим последовательность $a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n$. Очевидно, что

$$\varphi(\lambda) = \sum_{j=1}^{n} \big(\lambda - (-c_j)\big)_+ = \sum_{j=1}^{n} (\lambda - a_j)_+. \tag{18}$$

Согласно (16) при всех $j \in 1:n$ имеем

$$(\lambda - a_j)_+ = 0, \quad \text{если } \lambda \leqslant a_1,$$
$$(\lambda - a_j)_+ = \lambda - a_j, \quad \text{если } \lambda \geqslant a_n,$$

поэтому

$$\begin{gathered} \varphi(\lambda) = 0 \quad \text{при } \lambda \leqslant a_1, \\ \varphi(\lambda) = \sum_{j=1}^{n} (\lambda - a_j) = n\lambda - \sum_{j=1}^{n} a_j \quad \text{при } \lambda \geqslant a_n. \end{gathered} \tag{19}$$

По тем же соображениям при $\lambda \in [a_k, a_{k+1}]$, $k \in 1 : n-1$,

$$\varphi(\lambda) = \sum_{j=1}^{k} (\lambda - a_j) = k\lambda - \sum_{j=1}^{k} a_j. \tag{20}$$

Отметим, что

$$\varphi(\lambda) = k(\lambda - a_k) + \varphi(a_k) \tag{21}$$

при $\lambda \in [a_k, a_{k+1}]$, $k \in 1 : n-1$, и при $\lambda \geqslant a_n$, $k = n$. В частности,

$$\varphi(a_{k+1}) = k(a_{k+1} - a_k) + \varphi(a_k), \quad k \in 1 : n-1. \tag{22}$$

На основании (18)–(20) заключаем, что функция $\varphi(\lambda)$ при $\lambda \geqslant a_1$ является непрерывной ломаной, строго возрастающей от 0 до $+\infty$ (см. рис.).

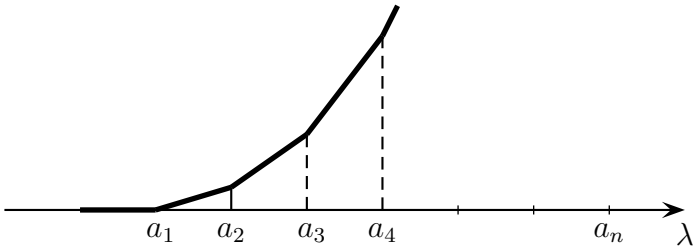


Рис. График ломаной $\varphi(\lambda)$

Это гарантирует существование и единственность точки λ^*, в которой $\varphi(\lambda^*) = 1$.

Установлено, что уравнение (17) имеет единственное решение λ^*. Отсюда следует, что система соотношений (10)–(12) имеет единственное решение и, наконец, что задача (9) имеет единственное решение — вектор x^* с компонентами

$$x_j^* = (\lambda^* + c_j)_+, \quad j \in 1 : n. \tag{23}$$

4°. Опишем простой алгоритм вычисления λ^*.

Ломаная $\varphi(\lambda)$ на отрезке $[a_1, a_n]$ полностью определяется своими значениями $\varphi_k = \varphi(a_k)$ в узлах $\lambda = a_k$, $k \in 1 : n$. Согласно (22) последовательность $\{\varphi_k\}$ можно вычислить рекуррентно:

$$\begin{aligned} \varphi_1 &= 0; \\ \varphi_{k+1} &= \varphi_k + k(a_{k+1} - a_k), \ k = 1, \ldots, n-1. \end{aligned} \tag{24}$$

Будем последовательно вычислять значения $\varphi_1, \varphi_2, \ldots$ по формуле (24), пока не встретим индекс k_0, на котором

$$\varphi_{k_0} < 1 \leqslant \varphi_{k_0+1}.$$

Если и $\varphi_n < 1$, то полагаем $k_0 = n$. На отрезке $[a_{k_0}, a_{k_0+1}]$ в случае $k_0 < n$ и полуоси $[a_{k_0}, +\infty)$ при $k_0 = n$ функция $\varphi(\lambda)$ имеет вид (21) (с заменой k на k_0). Она линейна. Решение уравнения $\varphi(\lambda) = 1$ находим элементарно:

$$\lambda^* = a_{k_0} + \tfrac{1}{k_0}(1 - \varphi_{k_0}). \tag{25}$$

5°. Подведём итог.

АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ (9)**.**

1) Меняем знаки у компонент c_j проектируемой точки c и числа $\{-c_j\}$ упорядочиваем по неубыванию. Получаем последовательность $a_1 \leqslant a_2 \leqslant \leqslant \ldots \leqslant a_n$.

2) Проводим последовательные вычисления по рекуррентной формуле (24), пока не встретим индекс k_0, на котором $\varphi_{k_0} < 1 \leqslant \varphi_{k_0+1}$. Если и $\varphi_n < 1$, то полагаем $k_0 = n$.

3) Вычисляем λ^* по формуле (25) и компоненты x_j^* проекции точки c на стандартный симплекс по формуле (23).

ПРИМЕР. Найдём проекцию точки

$$c = (1,\ -1,\ 0,\ 1,\ 0,\ \tfrac{2}{3})$$

на стандартный симплекс.

Упорядочим компоненты точки $-c = (-1,\ 1,\ 0,\ -1,\ 0,\ -\frac{2}{3})$ по неубыванию. Получим последовательность $a = (-1,\ -1,\ -\frac{2}{3},\ 0,\ 0,\ 1)$. Составим таблицу

k	1	2	3	4	5	6
a_k	-1	-1	$-\frac{2}{3}$	0	0	1

Проведём вычисления по формуле (24):

$$\varphi_1 = 0; \quad \varphi_2 = 0 + 1\big(-1 - (-1)\big) = 0;$$
$$\varphi_3 = 0 + 2\big(-\tfrac{2}{3} - (-1)\big) = \tfrac{2}{3};$$
$$\varphi_4 = \tfrac{2}{3} + 3\big(0 - (-\tfrac{2}{3})\big) = \tfrac{8}{3} > 1.$$

Получаем $\varphi_3 < 1 \leqslant \varphi_4$, так что $k_0 = 3$.

Согласно (25)

$$\lambda^* = -\tfrac{2}{3} + \tfrac{1}{3}(1 - \tfrac{2}{3}) = -\tfrac{5}{9}.$$

По формуле (23) находим компоненты проекции x^* точки c на стандартный симплекс:

$$x_1^* = (-\tfrac{5}{9} + 1)_+ = \tfrac{4}{9}; \quad x_2^* = (-\tfrac{5}{9} - 1)_+ = 0; \quad x_3^* = (-\tfrac{5}{9} + 0)_+ = 0;$$
$$x_4^* = (-\tfrac{5}{9} + 1)_+ = \tfrac{4}{9}; \quad x_5^* = (-\tfrac{5}{9} + 0)_+ = 0; \quad x_6^* = (-\tfrac{5}{9} + \tfrac{2}{3})_+ = \tfrac{1}{9}.$$

Таким образом,

$$x^* = (\tfrac{4}{9},\ 0,\ 0,\ \tfrac{4}{9},\ 0,\ \tfrac{1}{9}).$$

6°. В заключение отметим, что решение классической задачи минимизации (1) находится с помощью *формулы* (7), а решение неклассической задачи минимизации (9) — с помощью *алгоритма*.

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Линейная алгебра без определителей. Квадратичная функция.* СПб.: Изд-во СПбГУ, 1997. 80 с.

2. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

3. Малозёмов В. Н., Певный А. Б. *Быстрый алгоритм проектирования точки на симплекс* // Вестник СПбГУ. Сер. 1. 1992. Вып. 1 (№ 1). С. 112–113.

ЕЩЕ ОДИН БЫСТРЫЙ АЛГОРИТМ ПРОЕКТИРОВАНИЯ ТОЧКИ НА СТАНДАРТНЫЙ СИМПЛЕКС*

В. Н. Малозёмов, Г. Ш. Тамасян

1°. Задача ортогонального проектирования точки $c = (c_1, \ldots, c_n)$ на стандартный симплекс $\Lambda \subset \mathbb{R}^n$, определяемый условиями

$$\sum_{i=1}^{n} x_i = 1; \quad x_i \geqslant 0,\ i \in 1:n,$$

ставится следующим образом:

$$Q(x) := \tfrac{1}{2} \sum_{i=1}^{n} (x_i - c_i)^2 \to \min_{x \in \Lambda}. \tag{1}$$

Решение этой задачи существует и единственно. Обозначим его x^*.

В докладе [1] был описан быстрый алгоритм нахождения x^*. Идея алгоритма основана на чисто алгебраическом анализе условий оптимальности в форме Куна-Таккера для задачи (1). Этот анализ был выполнен в работе [2], опубликованной в 1992 г.

Ранее, в 1986 г., появилась работа [3], в которой также предлагался конечный алгоритм решения задачи (1). Этот алгоритм имеет геометрический характер, что подчеркивается в недавней работе [4].

В данном докладе мы даем усовершенствованный вариант описания и обоснования алгоритма из [3] и приводим результаты численных экспериментов по сравнению двух быстрых алгоритмов решения задачи (1).

2°. Начнем с описания алгоритма, идея которого предложена в [3].

Напомним, что $c = (c_1, \ldots, c_n)$. Обозначим $N = 1 : n$.

Предварительный шаг. В качестве начального приближения возьмем вектор $x^{(0)}$ с компонентами

$$x_i^{(0)} = c_i + \lambda, \quad i \in N, \tag{2}$$

*Семинар «DHA & CAGD». Избранные доклады. 5 сентября 2013 г.

где

$$\lambda = \frac{1}{n}\left(1 - \sum_{i\in N} c_i\right). \tag{3}$$

Общий шаг. Пусть имеется k-е приближение $x^{(k)}$. Если все компоненты вектора $x^{(k)}$ неотрицательны, то есть $x^{(k)} \geqslant \mathbb{O}$, то $x^{(k)}$ — искомая проекция точки c на стандартный симплекс Λ. Процесс заканчивается.

В противном случае, когда у вектора $x^{(k)}$ существует хотя бы одна отрицательная компонента, формируем индексное множество

$$I_k = \{i \in N \mid x_i^{(k)} \leqslant 0\}$$

и вычисляем

$$\lambda^{(k)} = \frac{1}{n_k}\left(1 - \sum_{i\in N\setminus I_k} x_i^{(k)}\right), \tag{4}$$

где $n_k = |N \setminus I_k|$. В качестве очередного приближения берем вектор $x^{(k+1)}$ с компонентами

$$x_i^{(k+1)} = \begin{cases} 0 & \text{при } i \in I_k; \\ x_i^{(k)} + \lambda^{(k)} & \text{при } i \in N \setminus I_k. \end{cases} \tag{5}$$

После этого возвращаемся к общему шагу.

Если у вектора $x^{(k+1)}$ имеется отрицательная компонента, то будет формироваться индексное множество $I_{k+1} = \{i \in N \mid x_i^{(k+1)} \leqslant 0\}$. Из определения $x^{(k+1)}$ следует, что множество I_{k+1} содержит I_k и те индексы $i \in N\setminus I_k$, на которых $x_i^{(k+1)} < 0$. Возможно, найдутся индексы $i \in N \setminus I_k$, на которых $x_i^{(k+1)} = 0$. Они тоже войдут в I_{k+1}.

Понятно, что I_k является собственным подмножеством множества I_{k+1}. Значит, по ходу процесса индексные множества I_k строго расширяются. Это гарантирует конечность процесса.

Остается проверить оптимальность точки $x^{(k)}$ при выполнении условия $x^{(k)} \geqslant \mathbb{O}$. Это мы сделаем позже, а пока приведем пример.

ПРИМЕР 1. Возьмем точку $c = (-1,\ 1,\ 0,\ -1,\ 0,\ \frac{2}{3})$ в $\mathbb{R}^6$ и найдем её проекцию на стандартный симплекс $\Lambda \subset \mathbb{R}^6$.

Предварительный шаг. Вычисляем $\lambda = \frac{1}{6}\big(1 + \frac{1}{3}\big) = \frac{2}{9}$. Имеем

$$x^{(0)} = \big(-\tfrac{7}{9},\ \tfrac{11}{9},\ \tfrac{2}{9},\ -\tfrac{7}{9},\ \tfrac{2}{9},\ \tfrac{8}{9}\big).$$

Первый шаг. Формируем множество $I_0 = \{1,\ 4\}$ и вычисляем

$$\lambda^{(0)} = \tfrac{1}{4}\big(1 - \tfrac{23}{9}\big) = -\tfrac{7}{18}.$$

Имеем

$$x^{(1)} = \left(0,\ \tfrac{15}{18},\ -\tfrac{3}{18},\ 0,\ -\tfrac{3}{18},\ \tfrac{9}{18}\right).$$

Второй шаг. Формируем множество $I_1 = \{1,\ 3,\ 4,\ 5\}$ и вычисляем

$$\lambda^{(1)} = \tfrac{1}{2}\left(1 - \tfrac{24}{18}\right) = -\tfrac{1}{6}.$$

Имеем

$$x^{(2)} = \left(0,\ \tfrac{2}{3},\ 0,\ 0,\ 0,\ \tfrac{1}{3}\right).$$

Выполняется условие $x^{(2)} \geqslant \mathbb{O}$, поэтому $x^{(2)}$ — искомая проекция.

3°. В общем случае строится конечная последовательность $x^{(0)}$, $x^{(1)}, \ldots$ точек из $\mathbb{R}^n$, удовлетворяющих условию

$$\sum_{i \in N} x_i^{(k)} = 1, \quad k = 0, 1, \ldots \tag{6}$$

Равенство (6) следует из (2) и (5). Вычисления заканчиваются, когда у очередной точки $x^{(k)}$ все компоненты будут неотрицательными.

ТЕОРЕМА. *Если $x^{(k)} \geqslant \mathbb{O}$, то $x^{(k)} = x^*$.*

Отдельно рассмотрим случай $k = 0$.

ЛЕММА 1. *Если $x^{(0)} \geqslant \mathbb{O}$, то $x^{(0)} = x^*$.*

Доказательство. Запишем критерий оптимальности для задачи (1) (см. [1]):

$$\begin{gathered} x_i - c_i = \lambda + u_i, \quad i \in N; \\ x_i u_i = 0, \quad u_i \geqslant 0, \quad x_i \geqslant 0, \quad i \in N; \\ \sum_{i \in N} x_i = 1. \end{gathered} \tag{7}$$

Согласно (6) и условию леммы имеем $x^{(0)} \in \Lambda$. Нетрудно проверить, что условия (7) выполняются при $x = x^{(0)}$, $u_i \equiv 0$ и λ вида (3). Значит, $x^{(0)} = x^*$.

Лемма доказана. □

Нам потребуется еще одно предварительное утверждение. Обозначим проекцию точки c на симплекс Λ через $\mathrm{Pr}_\Lambda(c)$.

ЛЕММА 2. *Справедливо равенство*

$$\mathrm{Pr}_\Lambda(x^{(0)}) = \mathrm{Pr}_\Lambda(c).$$

Доказательство. Пусть $y^{(0)} = \mathrm{Pr}_\Lambda(x^{(0)})$. По критерию оптимальности выполняются соотношения

$$y_i^{(0)} - x_i^{(0)} = \lambda^{(0)} + u_i^{(0)}, \quad i \in N;$$
$$y_i^{(0)} u_i^{(0)} = 0, \quad u_i^{(0)} \geqslant 0, \quad y_i^{(0)} \geqslant 0, \quad i \in N;$$
$$\sum_{i \in N} y_i^{(0)} = 1.$$

Учитывая (2), получаем

$$y_i^{(0)} - c_i = \left(\lambda + \lambda^{(0)}\right) + u_i^{(0)}, \quad i \in N;$$
$$y_i^{(0)} u_i^{(0)} = 0, \quad u_i^{(0)} \geqslant 0, \quad y_i^{(0)} \geqslant 0, \quad i \in N;$$
$$\sum_{i \in N} y_i^{(0)} = 1.$$

Значит, $y^{(0)} = \mathrm{Pr}_\Lambda(c)$.

Лемма доказана. □

4°. Переходим к доказательству теоремы. При $k = 0$ справедливость теоремы установлена в лемме 1.

Пусть $x^{(k+1)} \geqslant \mathbb{O}$ при некотором целом неотрицательном k. Согласно (6), $x^{(k+1)} \in \Lambda$. Перепишем формулу (5) в виде

$$x_i^{(k+1)} = x_i^{(k)} + \lambda^{(k)} + u_i^{(k)}, \quad i \in N, \tag{8}$$

где

$$u_i^{(k)} = \begin{cases} -x_i^{(k)} - \lambda^{(k)} & \text{при } i \in I_k; \\ 0 & \text{при } i \in N \setminus I_k. \end{cases} \tag{9}$$

По определению I_k имеем $x_i^{(k)} \leqslant 0$ при $i \in I_k$, причем одна из этих компонент строго отрицательна. Учитывая определение $\lambda^{(k)}$ и формулы (6) и (9), получаем

$$\lambda^{(k)} = \frac{1}{n_k} \sum_{i \in I_k} x_i^{(k)} < 0;$$
$$u_i^{(k)} > 0 \text{ при } i \in I_k, \qquad u_i^{(k)} = 0 \text{ при } i \in N \setminus I_k. \tag{10}$$

Формулы, аналогичные (8)–(10), справедливы и при меньших k, точнее

$$x_i^{(k-\nu+1)} = x_i^{(k-\nu)} + \lambda^{(k-\nu)} + u_i^{(k-\nu)}, \quad i \in N; \tag{11}$$

$$u_i^{(k-\nu)} = \begin{cases} -x_i^{(k-\nu)} - \lambda^{(k-\nu)} & \text{при } i \in I_{k-\nu}; \\ 0 & \text{при } i \in N \setminus I_{k-\nu}; \end{cases}$$

$$u_i^{(k-\nu)} > 0 \text{ при } i \in I_{k-\nu}, \qquad u_i^{(k-\nu)} = 0 \text{ при } i \in N \setminus I_{k-\nu}. \tag{12}$$

Здесь $\nu = 0, 1, \ldots, k$. Как отмечалось, $I_{k-\nu} \subset I_k$, поэтому $N \setminus I_k \subset N \setminus I_{k-\nu}$. Как следствие,

$$u_i^{(k-\nu)} = 0 \text{ при } i \in N \setminus I_k \text{ и всех } \nu = 0, 1, \ldots, k. \tag{13}$$

На основании (5) и (13) приходим к соотношению

$$x_i^{(k+1)} u_i^{(k-\nu)} = 0 \text{ при } i \in N \text{ и всех } \nu = 0, 1, \ldots, k. \tag{14}$$

Вернемся к формуле (8). Входящий в нее вектор $x^{(k)}$ согласно (11) можно выразить через $x^{(k-1)}$. В свою очередь $x^{(k-1)}$ можно выразить через $x^{(k-2)}$ и т. д. Наконец, $x^{(1)}$ можно выразить через $x^{(0)}$. В результате получим

$$x_i^{(k+1)} = x_i^{(0)} + \sum_{\nu=0}^{k} \lambda^{(k-\nu)} + \sum_{\nu=0}^{k} u_i^{(k-\nu)}, \quad i \in N.$$

Обозначим

$$\lambda^* = \sum_{\nu=0}^{k} \lambda^{(k-\nu)}, \quad u_i^* = \sum_{\nu=0}^{k} u_i^{(k-\nu)}$$

и перепишем последнюю формулу в виде

$$x_i^{(k+1)} - x_i^{(0)} = \lambda^* + u_i^*, \quad i \in N.$$

Отметим, что в силу (14)

$$x_i^{(k+1)} u_i^* = 0 \quad \forall\, i \in N.$$

В силу (12), $u_i^* \geqslant 0$ при всех $i \in N$ и по построению $x^{(k+1)} \in \Lambda$. Таким образом, выполнены все условия критерия оптимальности (7) при $x = x^{(k+1)}$, $c = x^{(0)}$, $\lambda = \lambda^*$ и $u = u^*$. Это гарантирует, что $x^{(k+1)} = \mathrm{Pr}_\Lambda(x^{(0)})$.

По лемме 2, $\mathrm{Pr}_\Lambda(x^{(0)}) = \mathrm{Pr}_\Lambda(c)$. Значит, $x^{(k+1)} = \mathrm{Pr}_\Lambda(c)$.

Теорема доказана. □

5°. В докладе [1] рассматривался другой метод проектирования точки $c = (c_1, \ldots, c_n)$ на стандартный симплекс. Напомним его описание.

1) Меняем знаки у компонент c_j точки c и числа $\{-c_j\}$ упорядочиваем по неубыванию. Получаем последовательность $a_1 \leqslant \ldots \leqslant a_n$.

2) Проводим последовательные вычисления по рекуррентной формуле

$$\begin{gathered}\varphi_1 = 0,\\ \varphi_{k+1} = \varphi_k + k(a_{k+1} - a_k), \quad k = 1, \dots, n-1,\end{gathered} \tag{15}$$

пока не встретим индекс k_0, на котором

$$\varphi_{k_0} < 1 \leqslant \varphi_{k_0+1}.$$

Если и $\varphi_n < 1$, то полагаем $k_0 = n$.

3) Вычисляем λ^* по формуле

$$\lambda^* = a_{k_0} + \tfrac{1}{k_0}(1 - \varphi_{k_0}). \tag{16}$$

Компоненты x_i^* проекции точки c на стандартный симплекс имеют вид

$$x_i^* = (\lambda^* + c_i)_+, \quad i \in 1:n, \tag{17}$$

где $(u)_+ = \max\{0, u\}$.

Найдем с помощью этого метода проекцию на стандартный симплекс точки c из разобранного в п. 2° примера 1. Напомним, что

$$c = (-1,\ 1,\ 0,\ -1,\ 0,\ \tfrac{2}{3}).$$

Имеем

$$-c = (1,\ -1,\ 0,\ 1,\ 0,\ -\tfrac{2}{3}), \quad a = (-1,\ -\tfrac{2}{3},\ 0,\ 0,\ 1,\ 1).$$

Составим таблицу

k	1	2	3	4	5	6
a_k	-1	$-\frac{2}{3}$	0	0	1	1

Проведем вычисления по формуле (15):

$$\begin{gathered}\varphi_1 = 0, \quad \varphi_2 = a_2 - a_1 = \tfrac{1}{3},\\ \varphi_3 = \varphi_2 + 2(a_3 - a_2) = \tfrac{5}{3} > 1.\end{gathered}$$

Значит, $k_0 = 2$. На основании (16) и (17) получаем

$$\begin{gathered}\lambda^* = a_2 + \tfrac{1}{2}(1 - \varphi_2) = -\tfrac{1}{3},\\ x^* = (0,\ \tfrac{2}{3},\ 0,\ 0,\ 0,\ \tfrac{1}{3}).\end{gathered}$$

Пришли к тому же результату, что и в п. 2°.

6°. С точки зрения трудоемкости второй алгоритм выглядит предпочтительней, поскольку в его основе лежит рекуррентное соотношение (15) для скалярных величин, в то время как в основе первого алгоритма лежит рекуррентное соотношение (5) для векторных величин.

Максимальная трудоемкость второго алгоритма достигается тогда, когда $\varphi_n < 1$. Можно описать все такие ситуации. Возьмем произвольную последовательность

$$0 = \varphi_1 \leqslant \varphi_2 \leqslant \ldots \leqslant \varphi_n < 1$$

и любое число a_n. Рекуррентное соотношение (15) обратимо. Положим

$$a_k = a_{k+1} - \tfrac{1}{k}(\varphi_{k+1} - \varphi_k), \quad k = n-1, n-2, \ldots, 1. \tag{18}$$

В качестве координат проектируемой точки c можно взять числа $\{-a_k\}$ в любом порядке.

ПРИМЕР 2. При $n = 4$ рассмотрим последовательность

$$\varphi_1 = 0, \;\; \varphi_2 = 0, \;\; \varphi_3 = \tfrac{2}{9}, \;\; \varphi_4 = \tfrac{5}{9},$$

и пусть $a_4 = \frac{2}{9}$. Согласно (18) имеем

$$\begin{aligned} a_3 &= a_4 - \tfrac{1}{3}(\varphi_4 - \varphi_3) = \tfrac{1}{9}, \\ a_2 &= a_3 - \tfrac{1}{2}(\varphi_3 - \varphi_2) = 0, \\ a_1 &= a_2 - \varphi_2 = 0. \end{aligned}$$

В качестве проектируемой возьмем следующую точку:

$$c = (-\tfrac{2}{9},\, 0,\, 0,\, -\tfrac{1}{9}). \tag{19}$$

С помощью второго алгоритма найдем ее проекцию на стандартный симплекс. Получим

$$x^* = (\tfrac{1}{9},\, \tfrac{1}{3},\, \tfrac{1}{3},\, \tfrac{2}{9}).$$

В общем случае второй алгоритм требует одну перестановку элементов массива длиной n и не более $4n - 2$ арифметических операций, где n – размерность пространства.

7°. Максимальная трудоемкость первого алгоритма достигается тогда, когда у всех членов последовательности $x^{(0)}, x^{(1)}, \ldots$ имеется только по одной отрицательной компоненте. В этом случае $x^{(n-1)}$ совпадает с одним из ортов пространства $\mathbb{R}^n$.

Приведем пример построения такой последовательности.

ПРИМЕР 3. Пусть $n = 4$ и $x^{(3)} = (0,\, 0,\, 0,\, 1)$. Это возможно, когда $x^{(2)} = (0,\, 0,\, x_3^{(2)},\, x_4^{(2)})$, где

$$x_3^{(2)} < 0, \quad x_4^{(2)} > 0, \quad x_3^{(2)} + x_4^{(2)} = 1. \tag{20}$$

Действительно, по алгоритму

$$\lambda^{(2)} = 1 - x_4^{(2)}, \quad x_4^{(3)} = x_4^{(2)} + \lambda^{(2)} = 1.$$

В качестве $x_4^{(2)}$ можно взять любое вещественное число, большее единицы, например, $x_4^{(2)} = 2$. Тогда $x_3^{(2)} = -1$. Вектор

$$x^{(2)} = (0,\, 0,\, -1,\, 2)$$

удовлетворяет всем условиям (20).

Запишем условия для компонент вектора $x^{(1)} = (0,\, x_2^{(1)},\, x_3^{(1)},\, x_4^{(1)})$:

$$\begin{gathered} x_2^{(1)} < 0, \quad x_3^{(1)} > 0, \quad x_4^{(1)} > 0, \quad x_2^{(1)} + x_3^{(1)} + x_4^{(1)} = 1, \\ x_3^{(1)} + \lambda^{(1)} = -1, \quad x_4^{(1)} + \lambda^{(1)} = 2, \end{gathered} \tag{21}$$

где $\lambda^{(1)} = \frac{1}{2}\bigl(1 - x_3^{(1)} - x_4^{(1)}\bigr) = \frac{1}{2}x_2^{(1)}$. Система линейных уравнений

$$\begin{aligned} x_2^{(1)} + x_3^{(1)} + x_4^{(1)} &= 1, \\ \tfrac{1}{2}x_2^{(1)} + x_3^{(1)} &= -1, \\ \tfrac{1}{2}x_2^{(1)} + x_4^{(1)} &= 2 \end{aligned}$$

эквивалентна следующей системе:

$$\begin{aligned} \tfrac{1}{2}x_2^{(1)} + x_3^{(1)} &= -1, \\ x_3^{(1)} - x_4^{(1)} &= -3. \end{aligned}$$

В качестве $x_3^{(1)}$ можно взять любое положительное число, например, $x_3^{(1)} = 1$. Тогда $x_4^{(1)} = 4$, $x_2^{(1)} = -4$. Вектор

$$x^{(1)} = \bigl(0,\, -4,\, 1,\, 4\bigr)$$

удовлетворяет всем условиям (21).

Запишем условия для компонент вектора $x^{(0)} = \bigl(x_1^{(0)},\, x_2^{(0)},\, x_3^{(0)},\, x_4^{(0)}\bigr)$:

$$\begin{gathered} x_1^{(0)} < 0, \quad x_2^{(0)} > 0, \quad x_3^{(0)} > 0, \quad x_4^{(0)} > 0, \\ x_1^{(0)} + x_2^{(0)} + x_3^{(0)} + x_4^{(0)} = 1, \\ x_2^{(0)} + \lambda^{(0)} = -4, \quad x_3^{(0)} + \lambda^{(0)} = 1, \quad x_4^{(0)} + \lambda^{(0)} = 4, \end{gathered} \tag{22}$$

где $\lambda^{(0)} = \frac{1}{3}\bigl(1 - x_2^{(0)} - x_3^{(0)} - x_4^{(0)}\bigr) = \frac{1}{3}x_1^{(0)}$. Система линейных уравнений

$$\begin{aligned} x_1^{(0)} + x_2^{(0)} + x_3^{(0)} + x_4^{(0)} &= 1,\\ \tfrac{1}{3}x_1^{(0)} + x_2^{(0)} &= -4,\\ \tfrac{1}{3}x_1^{(0)} + x_3^{(0)} &= 1,\\ \tfrac{1}{3}x_1^{(0)} + x_4^{(0)} &= 4 \end{aligned}$$

эквивалентна следующей системе

$$\begin{aligned} \tfrac{1}{3}x_1^{(0)} + x_2^{(0)} &= -4,\\ x_2^{(0)} - x_3^{(0)} &= -5,\\ x_3^{(0)} - x_4^{(0)} &= -3. \end{aligned}$$

В качестве $x_2^{(0)}$ можно взять любое положительное число, например, $x_2^{(0)} = 1$. Тогда $x_3^{(0)} = 6$, $x_4^{(0)} = 9$, $x_1^{(0)} = -15$. Вектор

$$x^{(0)} = \bigl(-15,\, 1,\, 6,\, 9\bigr).$$

удовлетворяет всем условиям (22).

Компоненты проектируемой точки c связаны с компонентами точки $x^{(0)}$ соотношением $c_i = x_i^{(0)} - \lambda$, $i \in 1:4$. При $\lambda = -16$ получим

$$c = \bigl(1,\, 17,\, 22,\, 25\bigr). \tag{23}$$

Нетрудно проверить, что при проектировании данной точки на стандартный симплекс первый алгоритм будет генерировать точки $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, $x^{(3)}$. У точки $x^{(3)} = (0,\, 0,\, 0,\, 1)$ все компоненты неотрицательные. Это гарантирует, что $x^{(3)} = \mathrm{Pr}_\Lambda(c)$.

В общем случае первый алгоритм требует не более $n^2 + 2n - 1$ арифметических операций.

8°. В примере 2 при проектировании точки c вида (19) на стандартный симплекс с помощью второго алгоритма потребовался максимально возможный объем вычислений. Воспользуемся для той же цели первым алгоритмом. Получим

$$\lambda = \frac{1}{4}\left(1 - \sum_{i=1}^{n} c_i\right) = \frac{1}{3},$$
$$x^{(0)} = \bigl(\tfrac{1}{9},\, \tfrac{1}{3},\, \tfrac{1}{3},\, \tfrac{2}{9}\bigr) = x^*,$$

то есть уже на предварительном шаге найдена искомая проекция.

В примере 3 при проектировании точки c вида (23) на стандартный симплекс с помощью первого алгоритма также потребовался максимально возможны объем вычислений. Воспользуемся вторым алгоритмом. Получим

$$a = (-25,\ -22,\ -17,\ -1),$$
$$\varphi_1 = 0, \quad \varphi_2 = a_2 - a_1 = 3 > 1,$$

так что $k_0 = 1$. Далее,

$$\lambda^* = a_1 + (1 - \varphi_1) = -24,$$
$$x^* = (0,\ 0,\ 0,\ 1).$$

Второй алгоритм приводит к требуемой проекции уже при $k_0 = 1$.

Таким образом, примеры показывают, что в случае, когда один из двух алгоритмов проектирования имеет максимальную трудоемкость, у второго алгоритма трудоемкость минимальна.

Мы провели массовые вычисления по сравнению эффективности двух алгоритмов проектирования точки на стандартный симплекс. Брались $m = 10000$ точек в n-мерном евклидовом пространстве при n, равном 100, 500, 1000 и 5000. Координаты точек формировались с помощью функции генерирования чисел по непрерывному равномерному распределению на интервале $(-m, m)$. Вычисления проводились на персональном компьютере с четырехъядерным процессором Intel Core 2 Quad с тактовой частотой 2.50 ГГц и оперативной памятью объемом 4 Гб.

В приводимых ниже таблицах указано время проектирования T (в сек.) всех $m = 10000$ точек в пространствах различных размерностей n.

Таблица 1 (первый алгоритм):

n	100	500	1000	5000
T	0.47	1.32	2.57	12.35

Таблица 2 (второй алгоритм):

n	100	500	1000	5000
T	0.34	0.96	1.75	8.93

ЛИТЕРАТУРА

1. Малоземов В. Н. *Проектирование точки на подпространство и на стандартный симплекс* // Семинар «DHA & CAGD». Избранные доклады. 28 февраля 2013 г.
(`http://dha.spb.ru/reps13.shtml#0228`) [Данная книга, с. 150]

2. Малоземов В. Н., Певный А. Б. *Быстрый алгоритм проектирования точки на симплекс* // Вестник СПбГУ. Сер. 1. 1992. Вып. 1 (№ 1). С. 112—113.

3. Michelot C. *A finite algorithm for finding the projection of a point onto the canonical simplex of* $\mathbb{R}^n$ // JOTA. 1986. Vol. 50. No 1. P. 195—200.

4. Causa A., Raciti F. *A purely geometric approach to the problem of computing the projection of a point on a simplex* // JOTA. 2013. Vol. 156. No 2. P. 524—528.

ПРОЕКТИРОВАНИЕ ТОЧКИ НА ТЕЛЕСНЫЙ СИМПЛЕКС*

В. Н. Малозёмов, Г. Ш. Тамасян

В докладе [1] рассматривались два быстрых алгоритма проектирования точки на стандартный симплекс $\Lambda \subset \mathbb{R}^n$. Теперь мы обратимся к задаче проектирования точки на телесный симплекс $\Omega \subset \mathbb{R}^n$, определяемый условиями

$$\sum_{i=1}^{n} x_i \leqslant 1; \quad x_i \geqslant 0,\ i \in 1:n.$$

Задача ставится так:

$$Q(x) := \tfrac{1}{2} \sum_{i=1}^{n} (x_i - c_i)^2 \to \min_{x \in \Omega}, \tag{1}$$

где $c_1, \ldots, c_n$ — координаты проектируемой точки c. Эта задача имеет единственное решение x^*.

Введем величину

$$h = \sum_{i=1}^{n} (c_i)_+,$$

где $(c_i)_+ = \max\{0, c_i\}$. Обозначим $c_+ = \big((c_1)_+, \ldots, (c_n)_+\big)$.

ТЕОРЕМА. *Если $h \leqslant 1$, то*

$$x^* = c_+. \tag{2}$$

При $h > 1$ справедливо равенство

$$x^* = \mathrm{Pr}_\Lambda(c_+). \tag{3}$$

Таким образом, при $h > 1$ задача (1) сводится к задаче проектирования точки c_+ на стандартный симплекс Λ.

*Семинар «DHA & CAGD». Избранные доклады. 11 октября 2013 г.

Доказательство. Запишем критерий оптимальности для задачи (1) (см. [2, с. 91]):

$$\begin{gathered} x_i - c_i = u_i - \lambda, \quad i \in 1:n; \\ \lambda(1 - x_1 - \ldots - x_n) = 0, \quad \lambda \geqslant 0; \\ u_i\, x_i = 0, \quad u_i \geqslant 0, \quad x_i \geqslant 0, \quad i \in 1:n; \\ x_1 + \ldots + x_n \leqslant 1. \end{gathered} \tag{4}$$

Нетрудно проверить, что в случае $h \leqslant 1$ все условия (4) выполняются при

$$x_i = (c_i)_+, \quad u_i = (c_i)_+ - c_i, \quad i \in 1:n; \qquad \lambda = 0.$$

В частности, равенство $\big((c_i)_+ - c_i\big)(c_i)_+ = 0$, справедливое при всех $i \in 1:n$, обеспечивает выполнение условия дополнительсти $u_i\, x_i = 0$, $i \in 1:n$. Значит, верна формула (2).

Допустим, что $h > 1$. Обозначим $\widehat{y} = \mathrm{Pr}_\Lambda(c_+)$. По критерию оптимальности (см. [1])

$$\begin{gathered} \widehat{y}_i - (c_i)_+ = \widehat{\lambda} + \widehat{u}_i, \quad i \in 1:n; \\ \widehat{u}_i\, \widehat{y}_i = 0, \quad \widehat{u}_i \geqslant 0, \quad \widehat{y}_i \geqslant 0, \quad i \in 1:n; \\ \widehat{y}_1 + \ldots + \widehat{y}_n = 1. \end{gathered} \tag{5}$$

Сложив равенства из первой строки условий (5), получим

$$1 - h = n\widehat{\lambda} + \sum_{i=1}^{n} \widehat{u}_i.$$

Отсюда следуют, что

$$\widehat{\lambda} < -\frac{1}{n} \sum_{i=1}^{n} \widehat{u}_i.$$

В частности, $\widehat{\lambda} < 0$.

Если все c_i неотрицательны, то положим

$$x_i = \widehat{y}_i, \quad u_i = \widehat{u}_i, \quad i \in 1:n; \qquad \lambda = -\widehat{\lambda}.$$

На основании (5) заключаем, что при этих данных выполняются все условия (4). Значит,

$$x^* = \widehat{y} = \mathrm{Pr}_\Lambda(c_+).$$

Допустим, что среди координат c_i имеется хотя бы одна отрицательная. Обозначим

$$I = \{i \in 1:n \mid c_i \leqslant 0\}.$$

При $i \in I$ равенство из первой строки условий (5) принимает вид

$$\widehat{y}_i = \widehat{\lambda} + \widehat{u}_i, \quad i \in I. \tag{6}$$

Умножим обе части последнего равенства на $\widehat{y}_i$. Учитывая условие дополнительности, получаем

$$\widehat{y}_i^2 = \widehat{\lambda}\,\widehat{y}_i, \quad i \in I. \tag{7}$$

Напомним, что $\widehat{\lambda} < 0$ и $\widehat{y}_i \geqslant 0$, поэтому из (7) и (6) следует, что

$$\widehat{y}_i = 0, \quad \widehat{u}_i = -\widehat{\lambda}, \quad i \in I. \tag{8}$$

Перепишем первое условие из (5) в виде

$$\begin{aligned} &\widehat{y}_i - c_i = (\widehat{u}_i - c_i) - (-\widehat{\lambda}), \quad i \in I; \\ &\widehat{y}_i - c_i = \widehat{u}_i - (-\widehat{\lambda}), \quad i \notin I. \end{aligned} \tag{9}$$

Положив $\lambda^* = -\widehat{\lambda}$,

$$u_i^* = \begin{cases} \widehat{u}_i - c_i, & i \in I; \\ \widehat{u}_i, & i \notin I, \end{cases}$$

придем к другому представлению формулы (9):

$$\widehat{y}_i - c_i = u_i^* - \lambda^*, \quad i \in 1:n.$$

При этом согласно (8) и определению множества I имеем $u_i^* > 0$ при $i \in I$. По определению $u_i^* = \widehat{u}_i \geqslant 0$ при $i \notin I$. Условие неотрицательности двойственных переменных u_i^* выполняется для всех $i \in 1:n$.

Осталось проверить условие дополнительности $u_i^*\,\widehat{y}_i = 0$, $i \in 1:n$. При $i \in I$ оно выполняется в силу (8), а при $i \notin I$ — в силу (5).

Таким образом, при $x_i = \widehat{y}_i$, $u_i = u_i^*$, $i \in 1:n$; $\lambda = \lambda^*$ выполняются все условия (4). Это гарантирует, что

$$x^* = \widehat{y} = \mathrm{Pr}_\Lambda(c_+).$$

Теорема доказана. □

ЛИТЕРАТУРА

1. Малоземов В. Н., Тамасян Г. Ш. *Ещё один быстрый алгоритм проектирование точки на стандартный симплекс* // Семинар «DHA & CAGD». Избранные доклады. 5 сентября 2013 г. (`http://dha.spb.ru/reps13.shtml#0905`) [Данная книга, с. 158]

2. Гавурин М. К., Малоземов В. Н. Экстремальные задачи с линейными ограничениями. Л.: Изд-во ЛГУ, 1984. 176 с.

МДМ-МЕТОДУ — 40 ЛЕТ*

В. Н. Малозёмов

1°. Пусть в пространстве $\mathbb{R}^n$ заданы m точек,

$$H = \{a_i\}_{i=1}^m.$$

Обозначим через G выпуклую оболочку множества H.

Ставится задача: *найти точку из G, ближайшую (в евклидовой норме) к началу координат.* Задачу можно записать так:

$$\|v\|^2 \to \min_{v \in G}. \tag{1}$$

Задача (1) имеет решение и оно единственно. Обозначим его v_*.

Вопрос о нахождении v_* возник как вспомогательная задача в оптимальном управлении [1] и негладкой оптимизации [2]. В 1971 году в работе [3] был предложен простой итерационный метод для решения задачи (1), который в дальнейшем получил название "МДМ-метод" (см., например [4, 5]). В данном докладе я возвращаюсь к анализу МДМ-метода с современных позиций.

2°. Отметим прежде всего, что при всех $v \in G$ выполняется неравенство

$$\langle v, v_* \rangle \geqslant \langle v_*, v_* \rangle. \tag{2}$$

Действительно, зафиксируем $v \in G$. В силу выпуклости множества G точка $v_* + t(v - v_*)$ при всех $t \in (0, 1)$ принадлежит G, поэтому

$$\langle v_*, v_* \rangle \leqslant \langle v_* + t(v - v_*), v_* + t(v - v_*) \rangle = \langle v_*, v_* \rangle + 2t\langle v_*, v - v_* \rangle + t^2\|v - v_*\|^2.$$

Отсюда следует, что

$$\langle v_*, v - v_* \rangle + \tfrac{1}{2}t\|v - v_*\|^2 \geqslant 0.$$

В пределе при $t \to +0$ получим неравенство, равносильное (2).

Неравенство (2) равносильно также следующему неравенству

$$\|v - v_*\|^2 \leqslant \|v\|^2 - \|v_*\|^2 \quad \forall v \in G. \tag{3}$$

*Семинар «DHA & CAGD». Избранные доклады. 10 декабря 2011 г.

3°. Обозначим через A матрицу со столбцами $a_1, \ldots, a_m$. Тогда любой вектор v из выпуклой оболочки G множества H допускает представление

$$v = Ap, \quad p \geqslant \mathbb{O}, \quad \sum_{i=1}^{m} p[i] = 1. \tag{4}$$

Первичным в этой формуле является вектор коэффициентов p. Носитель вектора p обозначим $M_+(p)$, так что

$$M_+(p) = \{i \in 1 : m \mid p[i] > 0\}.$$

Введём величину

$$\Delta(p) = \max_{i \in M_+(p)} \langle a_i, v\rangle - \min_{i \in 1:m} \langle a_i, v\rangle,$$

где $v = Ap$. Вектор p удовлетворяет условиям, указанным в формуле (4). Множество таких векторов обозначим P.

ЛЕММА 1. *При любом $v = Ap$, $p \in P$, справедливо неравенство*

$$\|v - v_*\|^2 \leqslant \Delta(p). \tag{5}$$

Доказательство. Согласно (2) имеем

$$\begin{aligned}\|v - v_*\|^2 = \|v\|^2 - 2\langle v, v_*\rangle + \|v_*\|^2 \leqslant \|v\|^2 - \langle v, v_*\rangle = \\ = \sum_{i \in M_+(p)} p[i]\langle a_i, v\rangle - \sum_{i=1}^{m} p_*[i]\langle a_i, v\rangle \leqslant \\ \leqslant \max_{i \in M_+(p)} \langle a_i, v\rangle - \min_{i \in 1:m} \langle a_i, v\rangle = \Delta(p).\end{aligned}$$

Лемма доказана. □

Из (5) следует, в частности, что

$$\Delta(p) \geqslant 0 \quad \forall p \in P. \tag{6}$$

ЛЕММА 2. *Равенство в* (6) *достигается тогда и только тогда, когда вектор $v = Ap$ является решением задачи* (1).

Доказательство. Неравенство (5) гарантирует оптимальность вектора $v = Ap$ в случае $\Delta(p) = 0$.

Наоборот, возьмём решение $v_* = Ap_*$ задачи (1) и покажем, что $\Delta(p_*) = 0$. Пусть

$$\Delta(p_*) = \langle a_{i'} - a_{i''}, v_*\rangle,$$

где $i' \in M_+(p_*)$, $i'' \in 1:m$. Введём вектор

$$\hat{v}_* = v_* - p_*[i'](a_{i'} - a_{i''})$$

(коэффициент при $a_{i'}$ передали вектору $a_{i''}$). Очевидно, что $\hat{v}_* \in G$. Имеем

$$\langle \hat{v}_*, v_* \rangle = \langle v_*, v_* \rangle - p_*[i']\langle a_{i'} - a_{i''}, v_* \rangle = \langle v_*, v_* \rangle - p_*[i']\Delta(p_*).$$

В силу (2) и положительности $p_*[i']$ получаем $\Delta(p_*) \leqslant 0$. Вместе с неравенством (6) это приводит к равенству $\Delta(p_*) = 0$.

Лемма доказана. □

4°. Обратимся к МДМ-методу. Возьмём начальное приближение $v_0 \in G$.

Пусть уже имеется k-е приближение $v_k = Ap_k$, $p_k \in P$. Опишем построение v_{k+1}.

Найдём индексы $i'_k \in M_+(p_k)$ и $i''_k \in 1:m$, такие, что

$$\max_{i \in M_+(p_k)} \langle a_i, v_k \rangle = \langle a_{i'_k}, v_k \rangle,$$
$$\min_{i \in 1:m} \langle a_i, v_k \rangle = \langle a_{i''_k}, v_k \rangle.$$

Для простоты будем использовать обозначения

$$a_{i'_k} = a'_k, \quad a_{i''_k} = a''_k.$$

В этом случае

$$\Delta_k := \Delta(p_k) = \langle a'_k - a''_k, v_k \rangle.$$

Если $\Delta_k = 0$, то, согласно лемме 2, v_k — решение задачи (1). Процесс закончен.

Пусть $\Delta_k > 0$. Введём вектор

$$\hat{v}_k = v_k - p'_k(a'_k - a''_k),$$

где $p'_k = p_k[i'_k]$. Очевидно, что $\hat{v}_k \in G$. Рассмотрим отрезок

$$v_k(t) = v_k + t(\hat{v}_k - v_k) = v_k - t\,p'_k(a'_k - a''_k), \quad t \in [0,1].$$

В силу выпуклости множества G все точки $v_k(t)$ этого отрезка принадлежат G. Выберем $t_k \in [0,1]$ из условия

$$\|v_k(t_k)\|^2 = \min_{t \in [0,1]} \|v_k(t)\|^2.$$

Положим $v_{k+1} = v_k(t_k)$ (см. рис.).

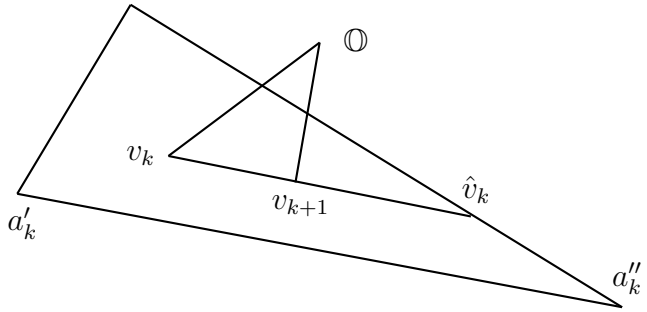


Рис.

Нетрудно понять, что

$$p_{k+1}[i] = \begin{cases} p_k[i], & \text{при } i \neq i'_k,\ i \neq i''_k, \\ (1-t_k)p_k[i'_k], & \text{при } i = i'_k, \\ p_k[i''_k] + t_k\, p_k[i'_k], & \text{при } i = i''_k. \end{cases}$$

Укажем явную формулу для t_k. Имеем

$$\begin{aligned} \|v_k(t)\|^2 &= \|v_k\|^2 + 2t\langle v_k, \hat{v}_k - v_k\rangle + t^2\|\hat{v}_k - v_k\|^2 = \\ &= \|v_k\|^2 - 2tp'_k\,\Delta_k + t^2\|\hat{v}_k - v_k\|^2. \end{aligned} \tag{7}$$

Абсолютный минимум $v_k(t)$ на $\mathbb{R}$ достигается в точке

$$\hat{t}_k = \frac{p'_k\,\Delta_k}{\|\hat{v}_k - v_k\|^2} = \frac{p'_k\,\Delta_k}{(p'_k)^2\,\|a'_k - a''_k\|^2} = \frac{\Delta_k}{p'_k\|a'_k - a''_k\|^2}. \tag{8}$$

Ясно, что $\hat{t}_k > 0$, поэтому

$$t_k = \begin{cases} \hat{t}_k, & \text{если } \hat{t}_k < 1, \\ 1, & \text{если } \hat{t}_k \geqslant 1. \end{cases}$$

Описание МДМ-метода завершено.

Построена последовательность $v_0, v_1, \ldots$ точек из G. Если она конечна, то последний её элемент является решением задачи (1). Вообще говоря, последовательность $\{v_k\}$ бесконечна. Такая ситуация возникает, когда

$$\Delta_k > 0 \quad \text{и, как следствие,} \quad \|v_{k+1}\| < \|v_k\| \quad \text{при всех} \quad k = 0, 1, \ldots. \tag{9}$$

Покажем, что в этом случае последовательность $\{v_k\}$ сходится к v_* — решению задачи (1).

5°. Начнём со вспомогательных утверждений.

ЛЕММА 3. *Справедливо предельное соотношение*

$$\lim_{k\to\infty} p'_k\,\Delta_k = 0. \tag{10}$$

Доказательство. Допустим противное, то есть что существует бесконечная подпоследовательность $\{p'_{k_s}\Delta_{k_s}\}$, такая, что

$$p'_{k_s}\Delta_{k_s} \geqslant \varepsilon > 0.$$

Обозначим через d диаметр множества G,

$$d = \max_{u\in G,\, v\in G} \|u - v\|.$$

Согласно (7) имеем

$$\|v_{k_s}(t)\|^2 \leqslant \|v_{k_s}\|^2 - 2t\varepsilon + t^2d^2 = \|v_{k_s}\|^2 - t\varepsilon - t(\varepsilon - td^2),$$

поэтому при $t \in [0, \varepsilon/d^2]$ будет

$$\|v_{k_s}(t)\|^2 \leqslant \|v_{k_s}\|^2 - t\varepsilon.$$

Положим $t_* = \min\{1, \varepsilon/d^2\}$. Тогда

$$\|v_{k_s+1}\|^2 = \min_{t\in[0,1]} \|v_{k_s}(t)\|^2 \leqslant \|v_{k_s}(t_*)\|^2 \leqslant \|v_{k_s}\|^2 - t_*\varepsilon.$$

Неограниченное число указанных понижений в монотонно убывающей последовательности $\left\{\|v_k\|^2\right\}$ противоречит неотрицательности её элементов. Лемма доказана. □

ЛЕММА 4. *Справедливо предельное соотношение*

$$\varliminf_{k\to\infty} \Delta_k = 0.$$

Доказательство. Допустим противное:

$$\varliminf_{k\to\infty} \Delta_k = \Delta' > 0.$$

В этом случае для достаточно больших $k \geqslant k_0$ будет выполнятся неравенство

$$\Delta_k \geqslant \Delta'/2. \tag{11}$$

Отсюда и из (10) следует, в частности, что

$$p'_k \to 0 \quad \text{при} \quad k \to \infty.$$

Далее, согласно (8) и (11)

$$\hat{t}_k \geqslant \frac{\Delta'}{2p'_k\, d^2},$$

так что $\hat{t}_k \to +\infty$ при $k \to \infty$. На основании определения t_k заключаем, что $t_k = 1$ при $k \geqslant k_1 \geqslant k_0$. Это приводит к соотношению

$$v_{k+1} = \hat{v}_k, \quad k \geqslant k_1. \tag{12}$$

Рассмотрим последовательность

$$v_{k_1},\, v_{k_1+1},\, v_{k_1+2}, \ldots. \tag{13}$$

Соотношение (12) показывает, что компоненты $p_k[i]$ вектора p_k, определяющего v_k, получаются путём перераспределения значений $p_{k_1}[i]$, $i \in 1 : m$, поэтому в последовательности (13) может быть лишь конечное число попарно различных элементов. Но это противоречит неравенству $\|v_{k+1}\| < \|v_k\|$, справедливому при всех $k = 0, 1, 2, \ldots$.

Лемма доказана. □

Теперь легко доказать утверждение о сходимости МДМ-метода.

ТЕОРЕМА 1. *При выполнении условия* (9) *последовательность* $\{v_k\}$*, построенная МДМ-методом, сходится к* v_*.

Доказательство. Согласно лемме 4 существует подпоследовательность $\{\Delta_{k_s}\}$, сходящаяся к нулю. По лемме 1 соответствующая подпоследовательность $\{v_{k_s}\}$ сходится к v_*. В частности, $\|v_{k_s}\| \to \|v_*\|$ при $s \to \infty$. В силу (9) вся последовательность $\big\{\|v_k\|\big\}$ строго убывает, поэтому $\|v_k\| \to \|v_*\|$ при $k \to \infty$. Остаётся сослаться на неравенство (3). Теорема доказана. □

6°. Отметим, что $v_* = \mathbb{O}$ тогда и только тогда, когда $\mathbb{O} \in G$. В этом пункте считаем, что $v_* \neq \mathbb{O}$.

Введём гиперплоскость

$$L = \big\{x \mid \langle v_*, x\rangle = \langle v_*, v_*\rangle\big\}.$$

ТЕОРЕМА 2. *Если* $v_* \neq \mathbb{O}$*, то, начиная с некоторого номера, все точки последовательности* $\{v_k\}$ *принадлежат* L.

Доказательство. Обозначим $M = 1 : m$,

$$M_0 = \{i \in M \mid a_i \in L\}, \quad M_1 = M \setminus M_0.$$

Согласно (2)

$$\langle a_i, v_* \rangle \geqslant \langle v_*, v_* \rangle, \quad i \in 1 : m,$$

поэтому при $i \in M_1$

$$\langle a_i, v_* \rangle > \langle v_*, v_* \rangle.$$

Положим

$$\tau = \min_{i \in M_1} \langle a_i, v_* \rangle - \langle v_*, v_* \rangle > 0.$$

Очевидно, что

$$\langle a_i, v_* \rangle \geqslant \langle v_*, v_* \rangle + \tau, \quad i \in M_1. \tag{14}$$

Воспользуемся теоремой 1, в силу которой найдётся индекс k_0, такой, что при $k \geqslant k_0$

$$\max_{i \in 1:m} \big| \langle a_i, v_k \rangle - \langle a_i, v_* \rangle \big| \leqslant \tfrac{\tau}{4}.$$

При тех же k и $i \in M_0$

$$\langle a_i, v_k \rangle \leqslant \langle a_i, v_* \rangle + \tfrac{\tau}{4} = \langle v_*, v_* \rangle + \tfrac{\tau}{4}, \tag{15}$$

в то время как при $i \in M_1$ согласно (14)

$$\langle a_i, v_k \rangle \geqslant \langle a_i, v_* \rangle - \tfrac{\tau}{4} \geqslant \langle v_*, v_* \rangle + \tfrac{3\tau}{4}. \tag{16}$$

Значит, при $k \geqslant k_0$

$$\min_{i \in 1:m} \langle a_i, v_k \rangle = \min_{i \in M_0} \langle a_i, v_k \rangle. \tag{17}$$

Запишем представление

$$v_k = \sum_{i \in M_0} p_k[i] a_i + \sum_{i \in M_1} p_k[i] a_i.$$

Покажем, что в последовательности

$$v_{k_0},\ v_{k_0+1},\ v_{k_0+2}, \ldots \tag{18}$$

встретится элемент v_k, у которого $p_k[i] = 0$ при всех $i \in M_1$.

Допустим противное. На основании (15) и (16) получим

$$\min_{i \in 1:m} \langle a_i, v_k \rangle \leqslant \langle v_*, v_* \rangle + \tfrac{\tau}{4},$$
$$\max_{i \in M_+(p_k)} \langle a_i, v_k \rangle \geqslant \langle v_*, v_* \rangle + \tfrac{3\tau}{4}.$$

Отсюда следует, что

$$\Delta_k \geqslant \tau/2, \quad k \geqslant k_0. \tag{19}$$

Далее, в силу определения M_0 и τ имеем

$$\langle v_k - v_*, v_* \rangle = \Big\langle \sum_{i=1}^{m} p_k[i](a_i - v_*), v_* \Big\rangle =$$
$$= \sum_{i \in M_1} p_k[i] \langle a_i - v_*, v_* \rangle \geqslant \tau \sum_{i \in M_1} p_k[i].$$

Левая часть этого неравенства стремится к нулю при $k \to \infty$, поэтому

$$\lim_{k \to \infty} \sum_{i \in M_1} p_k[i] = 0.$$

Выберем столь большое $k_1 \geqslant k_0$, чтобы при $k \geqslant k_1$ выполнялось неравенство

$$\sum_{i \in M_1} p_k[i] \leqslant \frac{\tau}{2d^2}.$$

На основании (8) и (19) получим

$$\hat{t}_k = \frac{\Delta_k}{p'_k \|a'_k - a''_k\|^2} \geqslant \frac{\tau}{2d^2 \sum_{i \in M_1} p_k[i]} \geqslant 1.$$

Значит, $t_k = 1$ и $v_{k+1} = \hat{v}_k$ при $k \geqslant k_1$. Как установлено при доказательстве леммы 4, отсюда следует, что в последовательности v_{k_1}, v_{k_1+1}, $v_{k_1+2}, \ldots$ может быть лишь конечное число попарно различных элементов. Это противоречит строгому убыванию $\|v_k\|$.

Итак, в последовательности (18) встретится элемент v_k, имеющий представление

$$v_k = \sum_{i \in M_0} p_k[i] a_i, \quad p_k \geqslant \mathbb{O}, \quad \sum_{i \in M_0} p_k[i] = 1. \tag{20}$$

В частности, v_k принадлежит гиперплоскости L. Согласно описанию МДМ-метода и соотношению (17) последующие элементы последовательности эту плоскость не покинут.

Теорема доказана. □

7°. Лемма 4 допускает усиление.

ТЕОРЕМА 3. *Справедливо предельное соотношение*

$$\lim_{k \to \infty} \Delta_k = 0.$$

Доказательство. В случае $v_* = \mathbb{O}$ (когда $v_k \to \mathbb{O}$) утверждение очевидно. Действительно,

$$\Delta_k \leqslant \max_{i\in 1:m}\langle a_i, v_k\rangle - \min_{i\in 1:m}\langle a_i, v_k\rangle.$$

Правая часть этого неравенства стремится к нулю при $k \to \infty$. Остаётся учесть неотрицательность Δ_k.

Пусть $v_* \neq \mathbb{O}$. Согласно (20) при больших k

$$\max_{i\in M_+(p_k)}\langle a_i, v_k\rangle \leqslant \max_{i\in M_0}\langle a_i, v_k\rangle.$$

Принимая во внимание (17), приходим к неравенству

$$\Delta_k \leqslant \max_{i\in M_0}\langle a_i, v_k\rangle - \min_{i\in M_0}\langle a_i, v_k\rangle.$$

В силу теоремы 1 и определения M_0 имеем

$$\lim_{k\to\infty}\big[\max_{i\in M_0}\langle a_i, v_k\rangle - \min_{i\in M_0}\langle a_i, v_k\rangle\big] = \max_{i\in M_0}\langle a_i, v_*\rangle - \min_{i\in M_0}\langle a_i, v_*\rangle = 0.$$

Остаётся учесть неотрицательность Δ_k.

Теорема доказана. □

8°. МДМ-метод позволяет за конечное число шагов строго отделить начало координат от G в случае, когда $\mathbb{O} \notin G$.

ТЕОРЕМА 4. *Условие $\mathbb{O} \notin G$ выполняется тогда и только тогда, когда при некотором k*

$$\Delta_k < \|v_k\|^2. \tag{21}$$

Неравенство (21) *гарантирует, что гиперплоскость*

$$\langle v_k, x\rangle - \langle v_k, a_k''\rangle = 0 \tag{22}$$

строго отделяет начало координат от множества G.

Доказательство. Пусть $\mathbb{O} \notin G$. Справедливость неравенства (21) при некотором k следует из того, что $\Delta_k \to 0$ и $v_k \to v_*$, $v_* \neq \mathbb{O}$.

Наоборот, на основании (21) и (5) получаем

$$\|v_k - v_*\|^2 < \|v_k\|^2.$$

Такое неравенство возможно только при $v_* \neq \mathbb{O}$.

Перепишем неравенство (21) в развёрнутом виде

$$\langle a_k' - a_k'', v_k\rangle < \langle v_k, v_k\rangle.$$

Отсюда следует, что

$$-\langle a''_k, v_k\rangle < \langle v_k, v_k - a'_k\rangle.$$

Так как

$$\langle v_k, v_k\rangle = \sum_{i\in M_+(p_k)} p_k[i]\langle a_i, v_k\rangle \leqslant \langle a'_k, v_k\rangle,$$

то

$$-\langle a''_k, v_k\rangle < 0. \tag{23}$$

Вместе с тем, при всех $v \in G$

$$\langle v_k, v\rangle = \sum_{i=1}^{m} p[i]\langle v_k, a_i\rangle \geqslant \langle a''_k, v_k\rangle,$$

так что

$$\langle v_k, v\rangle - \langle v_k, a''_k\rangle \geqslant 0. \tag{24}$$

На основании (23) и (24) заключаем, что гиперплоскость (22) строго отделяет начало координат от множества G.

Теорема доказана. □

ЛИТЕРАТУРА

1. Gilbert E. G. *An iterative procedure for computing the minimum of quadratic form on a convex set* // J. SIAM Control. 1966. Vol. 4. No. 1. P. 61–80.

2. Demyanov V. F. *Algorithms for some minimax problems* // J. Computer and System Sciences. 1968. Vol. 2. No. 4. P. 342–380.

3. Митчелл Б. Ф., Демьянов В. Ф., Малоземов В. Н. *Нахождение ближайшей к началу координат точки многогранника* // Вестник ЛГУ. 1971. № 19. С. 38–45.

4. Barbero A., Lopez J., Dorronsoro J. R. *An accelerated MDM algorithm for SVM Training* // European Symposium on Artificial Neural Networks — Advances in Computational Intelligence and Learning. Bruges (Belgium), 23–25 April 2008. P. 421–426.

5. Lazaro J. L. *On the relationship among the MDM, SMO and SVM-Light algorithms for Training Support Vector Machines* // Master's thesis. Universidad Autonoma de Madrid. Madrid, 2008.

О ЗАДАЧЕ ПРОЕКТИРОВАНИЯ НУЛЯ НА МНОГОГРАННИК*

В. Н. Малозёмов

В докладе представлены три варианта постановки задачи о проектировании начала координат на выпуклый многогранник.

1°. Пусть $a_1, \dots, a_m$ — точки из $\mathbb{R}^n$ и L — их выпуклая оболочка. Рассмотрим экстремальную задачу

$$\|x\|^2 \to \min_{x \in L}. \tag{1}$$

Решение этой задачи существует и единственно. Обозначим его x_*. Точка x_* имеет наименьшую евклидову норму среди всех точек из L. Она называется проекцией нуля на многогранник L.

Отметим, что

$$\langle x, x_* \rangle \geqslant \langle x_*, x_* \rangle \qquad \forall\, x \in L. \tag{2}$$

Действительно, возьмём произвольную точку x из L. Множество L выпуклое, поэтому при любом $\alpha \in (0,1)$ точка $x_* + \alpha(x - x_*)$ принадлежит L. По определению x_* имеем

$$\big\|x_* + \alpha(x - x_*)\big\|^2 \geqslant \|x_*\|^2,$$

что равносильно неравенству

$$2\alpha \langle x_*, x - x_* \rangle + \alpha^2 \|x - x_*\|^2 \geqslant 0 \qquad \forall\, \alpha \in (0,1).$$

Поделив на $\alpha > 0$ и перейдя к пределу при $\alpha \to +0$, получим неравенство, эквивалентное (2).

Из (2), в частности, следует, что

$$\langle a_i, x_* \rangle \geqslant \langle x_*, x_* \rangle, \qquad i \in 1 : m. \tag{3}$$

2°. Обозначим через Ω множество точек $y \in \mathbb{R}^n$, удовлетворяющих ограничениям

$$\langle a_i, y \rangle \geqslant \langle y, y \rangle, \qquad i \in 1 : m. \tag{4}$$

*Семинар «DHA & CAGD». Избранные доклады. 10 июня 2009 г.

Очевидно, что $\mathbb{O} \in \Omega$. Кроме того, согласно (3), $x_* \in \Omega$. Учитывая, что (4) эквивалентно неравенствам

$$\|y - \tfrac{1}{2}\, a_i\|^2 \leqslant \|\tfrac{1}{2}\, a_i\|^2, \qquad i \in 1:m,$$

заключаем, что Ω есть пересечение шаров с центрами $\frac{1}{2}\, a_i$ и радиусами $\|\frac{1}{2}\, a_i\|$ при $i \in 1:m$.

Рассмотрим экстремальную задачу

$$\|y\|^2 \to \max_{y \in \Omega}. \tag{5}$$

Решение этой задачи существует.

ТЕОРЕМА 1. *Единственным решением задачи* (5) *является* x_*.

Доказательство. Зафиксируем $y \in \Omega$ и введём вспомогательную задачу

$$\begin{aligned} \|v\|^2 &\to \min, \\ \langle v, y\rangle &\geqslant \langle y, y\rangle. \end{aligned} \tag{6}$$

Её единственным решением будет $v_* = y$ (см. рис. 1).

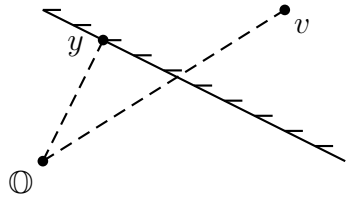


Рис. 1

Действительно, для любого плана v задачи (6) имеем

$$\|v\|^2 = \|(v - y) + y\|^2 \geqslant \|v - y\|^2 + \|y\|^2 \geqslant \|y\|^2.$$

Равенство $\|v\|^2 = \|y\|^2$ достигается только при $v = y$.

Множество планов задачи (6) обозначим $\Gamma(y)$. По определению Ω все точки a_i принадлежат $\Gamma(y)$, поэтому $\Gamma(y)$ содержит и их выпуклую оболочку, т. е. $L \subset \Gamma(y)$. Имеем

$$\|x_*\|^2 = \min_{x \in L} \|x\|^2 \geqslant \min_{v \in \Gamma(y)} \|v\|^2 = \|y\|^2.$$

Значит,

$$\|y\|^2 \leqslant \|x_*\|^2 \qquad \forall\, y \in \Omega.$$

Принимая во внимание, что $x_* \in \Omega$, заключаем, что x_* — решение задачи (5).

Проверим единственность решения. При $x_* = \mathbb{O}$ единственность очевидна. Пусть $x_* \neq \mathbb{O}$. Возьмём вектор $y_* \in \Omega$, такой, что $\|y_*\| = \|x_*\|$. Из условия $y_* \in \Omega$ следует, что

$$\langle x_*, y_* \rangle \geqslant \langle y_*, y_* \rangle = \|y_*\|^2.$$

Вместе с тем, в силу неравенства Коши-Буняковского

$$\langle x_*, y_* \rangle \leqslant \|x_*\| \cdot \|y_*\| = \|y_*\|^2.$$

Приходим к равенству

$$\langle x_*, y_* \rangle = \|x_*\| \cdot \|y_*\|,$$

которое в случае ненулевых x_*, y_* возможно лишь тогда, когда $y_* = \lambda\, x_*$ при некотором $\lambda > 0$. На самом деле, $\lambda = 1$, поскольку $\|y_*\| = \|x_*\|$. Это означает, что $y_* = x_*$.

Теорема доказана. □

Рис. 2 иллюстрирует содержание теоремы 1.

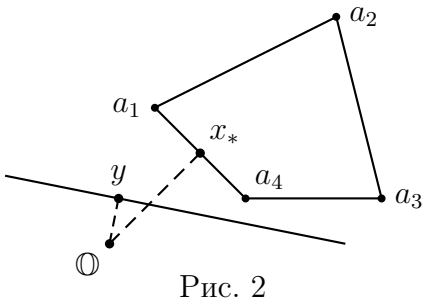


Рис. 2

3°. З. Р. Габидуллина в работе [1] ввела в рассмотрение ещё одну экстремальную задачу

$$\begin{gathered} \|z\|^2 \to \min, \\ \langle a_i, z \rangle \geqslant 1, \quad i \in 1:m. \end{gathered} \tag{7}$$

Множество планов этой задачи обозначим G.

ТЕОРЕМА 2. *Множество G пусто тогда и только тогда, когда $\mathbb{O} \in L$.*

Доказательство. Условие $\mathbb{O} \in L$ означает, что существуют неотрицательные числа $u[i]$, $i \in 1:m$, в сумме равные единице, такие, что

$$\sum_{i=1}^{m} u[i]\, a_i = \mathbb{O}.$$

Если допустить при этом, что $G \neq \varnothing$, то придём к ложному неравенству $0 \geqslant 1$. Таким образом, из условия $\mathbb{O} \in L$ следует, что $G = \varnothing$.

Наоборот, пусть $G = \varnothing$. Это значит, что система линейных неравенств

$$\langle a_i, z\rangle \geqslant 1, \qquad i \in 1:m,$$

несовместна. По теореме Фань-цзы [2, с. 25] найдутся числа $\lambda[i]$, $i \in 1:m$, удовлетворяющие условиям

$$\sum_{i=1}^{m} \lambda[i]\, a_i = \mathbb{O},$$

$$\lambda[i] \geqslant 0, \quad i \in 1:m; \quad \sum_{i=1}^{m} \lambda[i] > 0.$$

Положив $u[i] = \lambda[i] / \sum_{i=1}^{m} \lambda[i]$, получим

$$\sum_{i=1}^{m} u[i]\, a_i = \mathbb{O},$$

$$u[i] \geqslant 0, \quad i \in 1:m; \quad \sum_{i=1}^{m} u[i] = 1,$$

так что $\mathbb{O} \in L$. Теорема доказана. □

Теорема 2 позволяет сделать следующий вывод: *множество планов задачи* (7) *пусто тогда и только тогда, когда решением задачи* (1) *является вектор* $x_* = \mathbb{O}$.

Предположим, что $G \neq \varnothing$. В этом случае задача (7) имеем единственное решение. Обозначим его z_*. Очевидно, что $z_* \neq \mathbb{O}$.

ТЕОРЕМА 3. *Справедливы равенства*

$$x_* = \frac{z_*}{\|z_*\|^2}\,, \qquad z_* = \frac{x_*}{\|x_*\|^2}\,. \tag{8}$$

Доказательство. Проверим первое из равенств (8). Обозначим $\hat{y} = z_*/\|z_*\|^2$. При всех $i \in 1:m$ имеем

$$\langle a_i, \hat{y}\rangle = \frac{\langle a_i, z_*\rangle}{\|z_*\|^2} \geqslant \frac{1}{\|z_*\|^2} = \langle \hat{y}, \hat{y}\rangle.$$

Значит, $\hat{y}$ — план задачи (5). По теореме 1

$$\|\hat{y}\|^2 \leqslant \|x_*\|^2. \tag{9}$$

Далее, вектор $\hat{z} = x_*/\|x_*\|^2$ является планом задачи (7), поэтому $\|\hat{z}\|^2 \geqslant \geqslant \|z_*\|^2$, или $1/\|x_*\|^2 \geqslant \|z_*\|^2$, или

$$\|x_*\|^2 \leqslant \frac{1}{\|z_*\|^2} = \|\hat{y}\|^2. \tag{10}$$

Из (9) и (10) следует, что $\|\hat{y}\|^2 = \|x_*\|^2$. В силу единственности решения задачи (5) получаем $\hat{y} = x_*$, что соответствует первому равенству из (8).

Далее отметим, что $\|x_*\|^2 = 1/\|z_*\|^2$, так что $x_* = \|x_*\|^2 z_*$. Поделив на $\|x_*\|^2$, придём ко второму равенству из (8).

Теорема доказана. □

ЛИТЕРАТУРА

1. Габидуллина З. Р. *Теорема отделимости выпуклого многогранника от нуля пространства и её приложения в оптимизации* // Известия вузов. 2006. № 12. С. 21–26.

2. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

ПРОЕКТИРОВАНИЕ СИММЕТРИЧНОЙ МАТРИЦЫ НА КОНУС НЕОТРИЦАТЕЛЬНО ОПРЕДЕЛЁННЫХ МАТРИЦ И БЛИЗКИЕ ВОПРОСЫ*

В. Н. Малозёмов, С. Е. Михеев

1°. В линейном пространстве $\mathbb{R}^{n\times n}$ квадратных вещественных матриц порядка n введём скалярное произведение

$$\langle A, B\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} A[i,j]\times B[i,j]$$

и норму $\|A\| = \sqrt{\langle A, A\rangle}$. Напомним, что матрица $D \in \mathbb{R}^{n\times n}$ называется *симметричной*, если $D[i,j] = D[j,i]$ при всех $i, j \in 1:n$, и *неотрицательно определённой*, если $\langle Dx, x\rangle \geqslant 0$ при всех $x \in \mathbb{R}^n$. Множество симметричных неотрицательно определённых матриц порядка n обозначим $\mathcal{K}^n$. Очевидно, что $\mathcal{K}^n$ — выпуклый конус.

Возьмём симметричную матрицу $A \in \mathbb{R}^{n\times n}$ и рассмотрим задачу ортогонального проектирования матрицы A на конус $\mathcal{K}^n$ в следующей постановке:

$$F(X) := \|A - X\|^2 \to \inf_{X\in\mathcal{K}^n}. \tag{1}$$

В докладе приводится решение задачи (1), а также ближайших её обобщений.

2°. Нам потребуются некоторые свойства скалярного произведения и нормы матриц.

Из определений непосредственно следует, что

$$\|A + B\|^2 = \|A\|^2 + \|B\|^2 + 2\langle A, B\rangle. \tag{2}$$

Напомним, что *следом* квадратной матрицы $A \in \mathbb{R}^{n\times n}$ называется величина

$$\mathrm{Sp}(A) = \sum_{i=1}^{n} A[i,i].$$

*Семинар «CNSA & NDO». Избранные доклады. 18 декабря 2014 г.

ЛЕММА 1. *Справедлива формула*

$$\langle A, B\rangle = \mathrm{Sp}(AB^T). \tag{3}$$

Доказательство. При действиях с матрицами мы будем использовать индексную технику, описанную в [1].

Обозначим $N = 1 : n$. Имеем

$$\mathrm{Sp}(AB^T) = \sum_{i=1}^{n} A[i, N] \times B^T[N, i] =$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} A[i, j] \times B[i, j] = \langle A, B\rangle. \qquad \square$$

Очевидно, что $\mathrm{Sp}(C^T) = \mathrm{Sp}(C)$. Часто используется ещё одно свойство.

ЛЕММА 2. *Справедливо равенство*

$$\mathrm{Sp}(AB) = \mathrm{Sp}(BA). \tag{4}$$

Доказательство. Имеем

$$\mathrm{Sp}(AB) = \sum_{i=1}^{n} A[i, N] \times B[N, i] = \sum_{i=1}^{n}\sum_{j=1}^{n} A[i, j] \times B[j, i];$$
$$\mathrm{Sp}(BA) = \sum_{j=1}^{n} B[j, N] \times A[N, j] = \sum_{j=1}^{n}\sum_{i=1}^{n} B[j, i] \times A[i, j].$$

Отсюда очевидным образом следует (4). $\square$

Матрица $P \in \mathbb{R}^{n\times n}$ называется *ортогональной*, если

$$P^T P = PP^T = E.$$

ЛЕММА 3. *Если P и Q — ортогональные матрицы, то*

$$\|PCQ\| = \|C\|. \tag{5}$$

Доказательство. Согласно (3) и (4) имеем

$$\|PCQ\|^2 = \langle PCQ, PCQ\rangle = \mathrm{Sp}\big(PC(QQ^T)C^T P^T\big) =$$
$$= \mathrm{Sp}\big(P(CC^T P^T)\big) = \mathrm{Sp}\big((CC^T P^T)P\big) = \mathrm{Sp}(CC^T) = \|C\|^2.$$

Остаётся извлечь квадратный корень. $\square$

3°. Переходим к решению задачи (1). Как известно, для симметричной матрицы $A \in \mathbb{R}^{n\times n}$ существует ортогональная матрица P, такая, что

$$AP = P\Lambda, \tag{6}$$

где Λ — диагональная матрица, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, на диагонали которой стоят вещественные собственные числа матрицы A. Введём диагональную матрицу

$$\Lambda^+ = \operatorname{diag}(\lambda_1^+, \ldots, \lambda_n^+),$$

где $\lambda_i^+ = \max\{0, \lambda_i\}$.

ТЕОРЕМА 1. *Единственным решением задачи* (1) *является матрица*

$$X_* = P\Lambda^+ P^T. \tag{7}$$

При этом

$$F(X_*) = \sum_{\{i|\lambda_i<0\}} \lambda_i^2. \tag{8}$$

Для доказательства нам потребуется следующее вспомогательное утверждение.

ЛЕММА 4. *Пусть D и X — симметричные неотрицательно определённые матрицы порядка n. Тогда*

$$\langle D, X\rangle \geqslant 0.$$

Доказательство. Воспользуемся разложением $D = QVQ^T$, где V — диагональная матрица с неотрицательными диагональными элементами и Q — — ортогональная матрица. Согласно леммам 1 и 2 имеем

$$\langle D, X\rangle = \operatorname{Sp}\big(Q(VQ^TX)\big) = \operatorname{Sp}\big(V(Q^TXQ)\big) =$$
$$= \sum_{i=1}^{n} V[i,N] \times (Q^TXQ)[N,i] = \sum_{i=1}^{n} V[i,i] \times (Q^TXQ)[i,i].$$

У матрицы $Y = Q^TXQ$ диагональные элементы неотрицательны. Действительно,

$$Y[i,i] = \langle Q^TXQe_i, e_i\rangle = \langle XQe_i, Qe_i\rangle \geqslant 0.$$

Значит,

$$\langle D, X\rangle = \sum_{i=1}^{n} V[i,i] \times Y[i,i] \geqslant 0. \qquad \square$$

Доказательство теоремы 1. Возьмём произвольную матрицу $X \in \mathcal{K}^n$. В силу (2) имеем

$$F(X) = \big\|(A - X_*) + (X_* - X)\big\|^2 = F(X_*) + \|X_* - X\|^2 + \\ +2\langle A - X_*, X_* - X\rangle.$$

Согласно (6), $A = P\Lambda P^T$. Учитывая формулу (7), получаем

$$\langle A - X_*, X_*\rangle = \mathrm{Sp}\big(P(\Lambda - \Lambda^+)(P^TP)\Lambda^+P^T\big) = \\ = \mathrm{Sp}\big((\Lambda - \Lambda^+)\Lambda^+\big) = \sum_{i=1}^{n}(\lambda_i - \lambda_i^+)\lambda_i^+ = 0.$$

Значит,

$$F(X) = F(X_*) + \|X_* - X\|^2 + 2\langle X_* - A, X\rangle.$$

Матрица $X_* - A = P(\Lambda^+ - \Lambda)P^T$ симметрична и неотрицательно определена. По лемме 4, $\langle X_* - A, X\rangle \geqslant 0$. Приходим к неравенству

$$F(X) \geqslant F(X_*) + \|X_* - X\|^2. \tag{9}$$

Отсюда следует как оптимальность матрицы X_*, так и её единственность.

Формула (8) проверяется непосредственно:

$$F(X_*) = \|A - X_*\|^2 = \big\|P(\Lambda - \Lambda^+)P^T\big\|^2 = \\ = \|\Lambda - \Lambda^+\|^2 = \sum_{i=1}^{n}(\lambda_i - \lambda_i^+)^2 = \sum_{\{i|\lambda_i<0\}}\lambda_i^2.$$

Теорема доказана. □

Замечание. Неравенство (9) характеризует решение X_* задачи (1) как *сильно единственное.*

4°. В задаче (1) добавим ограничение $\mathrm{Sp}(X) \leqslant T$, где $T > 0$. Получим новую экстремальную задачу:

$$\begin{gathered} F(X) := \|A - X\|^2 \to \inf, \\ \mathrm{Sp}(X) \leqslant T, \quad X \in \mathcal{K}^n. \end{gathered} \tag{10}$$

Найдём её решение.

Обозначим через $\mathcal{G}^n$ множество планов задачи (10). Возьмём $X \in \mathcal{G}^n$. На основании формулы (6) и леммы 3 запишем

$$F(X) = \big\|P^T(A - X)P\big\|^2 = \|\Lambda - P^TXP\|^2.$$

Матрица $Y = P^T X P$ принадлежит конусу $\mathcal{K}^n$. Кроме того,

$$\mathrm{Sp}(Y) = \mathrm{Sp}\big(P^T(XP)\big) = \mathrm{Sp}\big(X(PP^T)\big) = \mathrm{Sp}(X) \leqslant T.$$

Обозначим $y_i = Y[i,i]$. Тогда

$$F(X) = \|\Lambda - Y\|^2 \geqslant \sum_{i=1}^{n} (\lambda_i - y_i)^2, \tag{11}$$

причём

$$\sum_{i=1}^{n} y_i \leqslant T; \quad y_i \geqslant 0,\ i \in 1:n. \tag{12}$$

Рассмотрим вспомогательную задачу: минимизировать функцию

$$f(y) = \sum_{i=1}^{n} (\lambda_i - y_i)^2$$

при ограничениях (12). Решение этой задачи существует и единственно. Обозначим его $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$. На основании (11) получаем

$$F(X) \geqslant f(\hat{y}) \quad \forall\ X \in \mathcal{G}^n. \tag{13}$$

Введём матрицу

$$\hat{X} = P\hat{Y}P^T, \tag{14}$$

где $\hat{Y} = \mathrm{diag}(\hat{y}_1, \ldots, \hat{y}_n)$.

ТЕОРЕМА 2. *Матрица $\hat{X}$ вида* (14) *является единственным решением задачи* (10).

Доказательство. Очевидно, что матрица $\hat{X}$ принадлежит множеству $\mathcal{G}^n$. При этом

$$F(\hat{X}) = \big\|P^T(A - \hat{X})P\big\|^2 = \|\Lambda - \hat{Y}\|^2 = \sum_{i=1}^{n} (\lambda_i - \hat{y}_i)^2 = f(\hat{y}).$$

Согласно (13)

$$F(X) \geqslant F(\hat{X}) \quad \forall\ X \in \mathcal{G}^n.$$

Оптимальность матрицы $\hat{X}$ установлена.

Равенство $F(X) = F(\hat{X})$ выполняется только тогда, когда $P^T X P = \hat{Y}$, то есть, когда $X = P\hat{Y}P^T = \hat{X}$. Это доказывает единственность решения задачи (10). □

Замечание. Для решения вспомогательной задачи минимизации функции $f(y)$ при ограничениях (12) разработан быстрый алгоритм [2].

5°. Обозначим через $\mathcal{M}^n_{\alpha,\beta}$ множество симметричных матриц порядка n, все собственные числа которых принадлежат отрезку $[\alpha, \beta]$. Рассмотрим ещё один вариант задачи (1):

$$F(X) := \|A - X\|^2 \to \inf_{X \in \mathcal{M}^n_{\alpha,\beta}}. \tag{15}$$

Укажем явное решение и этой задачи.

Возьмём спектральное разложение симметричной матрицы $A : A = P\Lambda P^T$, где $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ и P — ортогональная матрица. Введём диагональную матрицу

$$\check{\Lambda} = \operatorname{diag}(\check{\lambda}_1, \ldots, \check{\lambda}_n)$$

с диагональными элементами

$$\check{\lambda}_i = \begin{cases} \lambda_i, & \text{если } \lambda_i \in [\alpha, \beta]; \\ \alpha, & \text{если } \lambda_i < \alpha; \\ \beta, & \text{если } \lambda_i > \beta. \end{cases}$$

ТЕОРЕМА 3. *Матрица*

$$\check{X} = P\check{\Lambda}P^T$$

является единственным решением задачи (15).

Доказательство теоремы основано на следующем вспомогательном утверждении.

ЛЕММА 5. *Для диагональных элементов матрицы $D \in \mathcal{M}^n_{\alpha,\beta}$ выполняются неравенства*

$$\alpha \leqslant D[i,i] \leqslant \beta, \quad i \in 1 : n. \tag{16}$$

Доказательство. Воспользуемся спектральным разложением симметричной матрицы D: $D = QVQ^T$, где Q — ортогональная матрица и $V = \operatorname{diag}(v_1, \ldots, v_n)$, причём по условию все v_i принадлежат отрезку $[\alpha, \beta]$. Запишем

$$D[i,i] = (QV)[i,N] \times Q^T[N,i] =$$
$$= \sum_{j=1}^{n} \Big(\sum_{k=1}^{n} Q[i,k] \times V[k,j] \Big) \times Q[i,j] = \sum_{j=1}^{n} v_j \big(Q[i,j]\big)^2. \tag{17}$$

Вместе с тем, из условия $QQ^T = E$ следует, что

$$1 = (QQ^T)[i,i] = Q[i,N] \times Q^T[N,i] = \sum_{j=1}^{n} \big(Q[i,j]\big)^2,$$

то есть

$$\sum_{j=1}^{n}\big(Q[i,j]\big)^2 = 1 \quad \text{при всех} \quad i \in 1:n. \tag{18}$$

На основании (17), (18) и неравенств $\alpha \leqslant v_i \leqslant \beta$, $i \in 1:n$, приходим к (16). $\square$

Доказательство теоремы 3. Возьмём план X задачи (15). Как и раньше, имеем

$$F(X) = \|\Lambda - Y\|^2, \tag{19}$$

где $Y = P^T X P$. Покажем, что матрицы Y и X имеют одни и те же собственные числа. Действительно, пусть $XQ = QV$. Тогда

$$P^T X Q = P^T Q V. \tag{20}$$

Обозначим $U = P^T Q$. Очевидно, что U — ортогональная матрица. При этом $Q = PU$. Перепишем равенство (20) в виде

$$(P^T X P)U = UV.$$

Это и означает, что спектры матриц X и $Y = P^T X P$ совпадают. По лемме 5 для величин $y_i = Y[i,i]$ выполняются неравенства

$$\alpha \leqslant y_i \leqslant \beta, \quad i \in 1:n.$$

В силу (19)

$$F(X) \geqslant \sum_{i=1}^{n}(\lambda_i - y_i)^2. \tag{21}$$

Рассмотрим вспомогательную задачу

$$f(y) := \sum_{i=1}^{n}(\lambda_i - y_i)^2 \to \inf$$
$$\alpha \leqslant y_i \leqslant \beta, \quad i \in 1:n.$$

Эта задача имеет единственное решение $\check{y} = (\check{y}_1, \ldots, \check{y}_n)$, где $\check{y}_i = \check{\lambda}_i$. Учитывая (21), получаем

$$F(X) \geqslant f(\check{y}) \quad \forall\ X \in \mathcal{M}^n_{\alpha,\beta}. \tag{22}$$

Матрица $\check{X} = P\check{\Lambda}P^T$, указанная в формулировке теоремы 3, принадлежит множеству $\mathcal{M}^n_{\alpha,\beta}$. При этом

$$F(\check{X}) = \big\|P^T(A - \check{X})P\big\|^2 = \|\Lambda - \check{\Lambda}\| = \sum_{i=1}^{n}(\lambda_i - \check{\lambda}_i)^2 = f(\check{y}).$$

Согласно (22),

$$F(X) \geqslant F(\check{X}) \quad \forall\, X \in \mathcal{M}^n_{\alpha,\beta}.$$

Оптимальность матрицы $\check{X}$ установлена.

Равенство $F(X) = F(\check{X})$ выполняется только тогда, когда $P^T X P = \check{\Lambda}$, то есть, когда $X = P\check{\Lambda}P^T = \check{X}$. Это доказывает единственность решения задачи (15). □

6°. Близкие экстремальные задачи на множестве матриц рассматривались в работе [3].

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Линейная алгебра без определителей. Квадратичная функция.* СПб.: Изд-во СПбГУ, 1997. 80 с.

2. Малозёмов В. Н., Тамасян Г. Ш. *Проектирование точки на телесный симплекс* // Семинар «DHA & CAGD». Избранные доклады. 11 октября 2013 г. (`http://dha.spb.ru/reps13.shtml#1011`) [Данная книга, с. 169]

3. Coope I. D., Renaud P. F. *Trace inequalities with applications to orthogonal regression and matrix nearness problems* // J. Inequalities in Pure and Applied Math., 2009. Vol. 10, No. 4 (Article 92. 7 pp.).
(`http://jipam.vu.edu.au`)

РЕШЕНИЕ ЗАДАЧИ СИЛЬВЕСТРА В MATLAB*

М. А. Кольцов

1°. Постановка задачи. Будем рассматривать *задачу Сильвестра* — по заданному набору точек $a_i \in \mathbb{R}^n$, $i \in M = 1:m$, найти наименьший по объёму шар, их содержащий. Чтобы формально поставить задачу, введем переменную $x \in \mathbb{R}^n$ — центр искомого шара. Очевидно, что для того, чтобы шар $B_{\mathbb{R}^n}(x, r)$ содержал все заданные точки, необходимо, чтобы его радиус r удовлетворял следующему условию:

$$\forall\, i \in M \ \|a_i - x\| \leqslant r$$

То есть, в задаче Сильвестра требуется найти точку минимума функции $\max\limits_{i\in M}\{\|a_i - x\|\}$ на $\mathbb{R}^n$, при этом радиус искомого минимального шара будет равен значению функции в этой точке. Заметим, что выражение под знаком максимума можно заменить на $\frac{1}{2}\|a_i - x\|^2$, что не меняет точку минимума. Теперь можно сказать так: решение задачи Сильвестра — это точка x_*, являющаяся решением экстремальной задачи

$$\varphi(x) := \max_{i\in M}\{\tfrac{1}{2}\|a_i - x\|^2\} \to \min_{x\in\mathbb{R}^n}, \tag{1}$$

а радиус соответствующего минимального шара равен $\sqrt{2\varphi(x_*)}$. Доказательство существования и единственности такого решения имеется в книге [1].

Задача (1) сводится к задаче квадратичного программирования. Раскроем по определению нормы:

$$\varphi(x) = \tfrac{1}{2}\|x\|^2 + \max_{i\in M}\{-\langle a_i, x\rangle + \tfrac{1}{2}\|a_i\|^2\}$$

Теперь выражение под максимумом линейно по x, а значит сам максимум можно обозначить за t и ввести набор ограничений. То есть, исходная задача эквивалентна задаче квадратичного программирования

$$\frac{1}{2}\langle Ex, x\rangle + t \to \min$$
$$\langle a_i, x\rangle + t \geqslant \frac{1}{2}\|a_i\|^2, \quad i \in M$$

*Семинар «CNSA & NDO». Избранные доклады. 26 февраля 2015 г.

2°. Общий вид задачи квадратичного программирования и двойственной к ней. Чтобы двигаться дальше, необходимо сначала рассмотреть понятия задачи квадратичного программирование и двойственной задачи. Задача квадратичного программирования — это экстремальная задача вида

$$\begin{gathered} Q(x) = \frac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \min \\ A[M_1, N] \times x[N] \geqslant b[M_1] \\ A[M_2, N] \times x[N] = b[M_2] \\ x[N_1] \geqslant \mathbb{O}[N_1] \end{gathered}$$

где $D = D[N, N]$ — симметричная неотрицательно-определённая матрица. Как и в случае линейного программирования, можно ввести двойственную задачу и получить теорему, связывающую её с исходной задачей — первую теорему двойственности. При этом двойственная задача выглядит так:

$$\begin{gathered} -\frac{1}{2}\langle Dv, v\rangle + \langle b, u\rangle \to \max \\ -v[N] \times D[N, N_1] + u[M] \times A[M, N_1] \leqslant c[N_1] \\ -v[N] \times D[N, N_2] + u[M] \times A[M, N_2] = c[N_2] \\ u[M_1] \geqslant \mathbb{O}[M_1] \end{gathered}$$

Для удобства записи двойственной задачи можно использовать таблицу

	c	
v	$-D$	$-\frac{1}{2}Dv$
u	A	b

Тогда целевая функция получается скалярным умножением первого и последнего столбца таблицы, ограничения — первого и второго, а правые части ограничений стоят в первой строке, причём неравенствами являются ограничения, соответствующие знаковым ограничениям исходной задачи. Знаковые ограничения накладываются на двойственные переменные (компоненты вектора u), соответствующие исходным ограничениям-неравенствам.

3°. Двойственная задача для задача Сильвестра и её решение. Вернёмся к задаче Сильвестра. Из векторов-строк a_i составим матрицу A_0 и введём стандартные для квадратичной задачи обозначения: матрицу квадратичной формы D, вектор коэффициентов линейной формы c, матрицу ограничений A и столбец правых частей b. Получаем следующее:

$$D = \begin{pmatrix} & & 0 \\ E[N,N] & & \vdots \\ & & 0 \\ 0 & \cdots & 0 \quad 0 \end{pmatrix} \quad c = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix} \quad A = \begin{pmatrix} & & 1 \\ A_0 & & \vdots \\ & & 1 \end{pmatrix} \quad b[i] = \frac{1}{2}\|a_i\|^2$$

Переменной в такой задаче является пара $y = (x, t)$, а в двойственной вектор $z = ((v, s), u)$. Чтобы записать двойственную задачу, воспользуемся таблицей

	$0 \quad \cdots \quad 0$	1	
$v[N]$	$-E[N,N]$	0 $\vdots$ 0	$-\frac{1}{2}v[N]$
s	$0 \quad \cdots \quad 0$	0	0
u	A_0	1 $\vdots$ 1	b

Тогда по описанному общему алгоритму имеем

$$\begin{gathered} g(z) := -\frac{1}{2}\langle v, v\rangle + \langle b, u\rangle \to \max \\ -v + uA_0 = 0 \\ \sum_{i\in M} u[i] = 1 \\ u[M] \geqslant \mathbb{O}[M] \end{gathered}$$

Видно, что здесь можно исключить вектор v и получить задачу минимизации на стандартном симплексе

$$\begin{gathered} h(u) := \frac{1}{2}\|uA_0\|^2 - \langle b, u\rangle \to \min \\ \sum_{i\in M} u[i] = 1 \\ u[M] \geqslant \mathbb{O}[M] \end{gathered}$$

Для последней задачи существует оптимальный план u_*. В [1] доказано, что в таком случае вектор $x_* = u_*A_0$ является решением задачи Сильвестра. Так как два оптимальных плана пары двойственных задач имеют одно и то же значение целевой функции, то $\varphi(x_*) = -h(u_*)$ (на последнем шаге целевая функция была умножена на -1 чтобы перейти от задачи максимизации к задаче минимизации). Значит, радиус искомого минимального шара равен $\sqrt{-2h(u_*)}$.

4°. Решение задачи в среде MATLAB. Для удобства решения задачи Сильвестра была написана функция `sylvester`, принимающая единственный аргумент — матрицу A, строки которой образованы исходным множеством точек. Алгоритм, реализованный в функции, следующий:

1) Вычислить матрицу $D = AA^T$ и вектор b, соответствующие двойственной задаче.

2) Вызвать встроенную функцию `quadprog` для решения двойственной задачи и получить её решение — вектор u_*, и значение целевой функции f на нём.

3) Вычислить радиус минимального шара по формуле $r = \sqrt{-2f}$.

Написанная функция работает с пространствами любых размерностей, но для наглядности она была проверена на размерности 2. Для этого генерировались случайные множества точек, лежащих внутри и на границе круга $(x-4)^2 + (y-3)^2 \leqslant 4$ и для них решалась задача Сильвестра. Одно из таких решений с количеством точек 100 представлено на рис. 1.

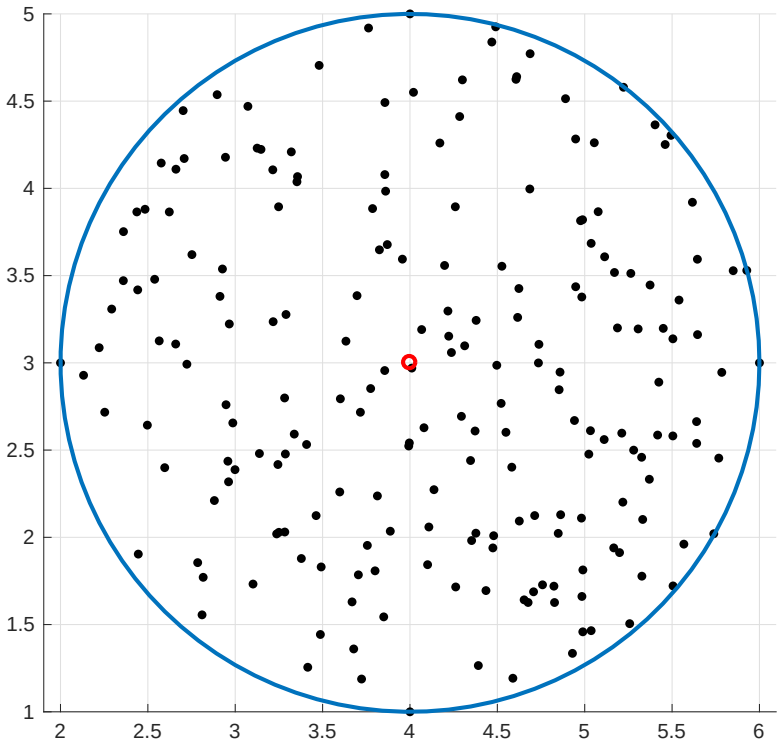


Рис. 1. Решение задачи Сильвестра для $m = 100$

Стоит отметить, что уже при попытке решить задачу для 200 точек, мы получаем ошибку:

```
Maximum number of iterations exceeded; increase options.MaxIter.
To continue solving the problem with the current solution as the
starting point, set x0 = x before calling quadprog.
```

Это связано с тем, что по умолчанию количество итераций, которые делает функция `quadprog`, ограничено числом 200. Чтобы решить эту проблему, можно поднять ограничение. При этом оказывается, что количество итераций зависит от количества точек примерно линейно (см. рис. 2).

Такой результат можно объяснить тем, что по умолчанию MATLAB выбирает для решения этой задачи алгоритм «active-set». Если же задать использование алгоритма «interior-point-convex», пригодного для решения выпуклых задач, то для 100 точек требуется 6 итераций, а для 500 — всего 9.

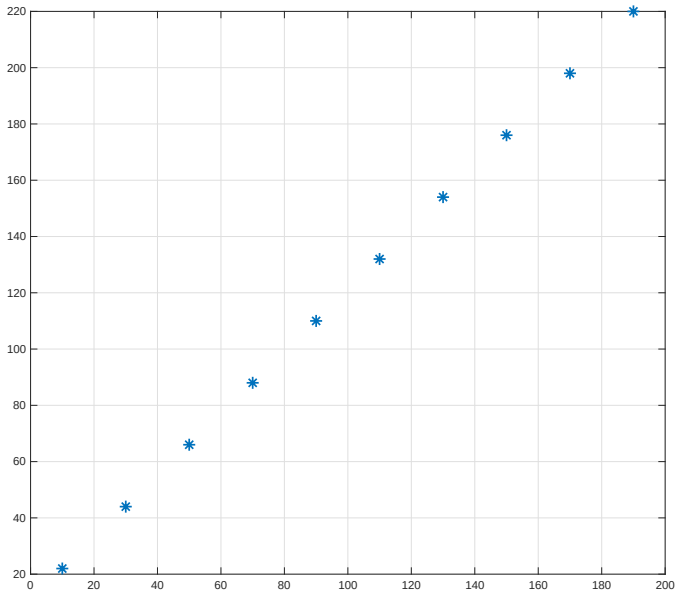


Рис. 2. Зависимость числа итераций от числа точек

Также можно протестировать программу в размерности 3. Для этого было сгенерировано 100 случайных точек внутри и на границе стандартного куба $[-1, 1]^3$. На этом наборе точек программа выдает ожидаемый ответ — центр шара близок к 0, а его радиус — к $\sqrt{3}$.

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

РЕШЕНИЕ ЗАДАЧ КВАДРАТИЧНОГО ПРОГРАММИРОВАНИЯ В СРЕДЕ MATLAB*

Н. А. Соловьёва, Е. К. Чернэуцану

В среде MATLAB задачи квадратичного программирования решаются с помощью функции `quadprog`. Доклад посвящён краткому описанию её возможностей.

1°. Функция `quadprog` решает задачу квадратичного программирования в форме

$$\begin{aligned} \tfrac{1}{2}\,\mathtt{x}^T\cdot\,\mathtt{H}\cdot\mathtt{x}+\mathtt{f}^T\cdot\,\mathtt{x}\to\inf,\\ \mathtt{A}\cdot\mathtt{x}\leqslant\mathtt{b},\\ \mathtt{Aeq}\cdot\mathtt{x}=\mathtt{beq},\\ \mathtt{lb}\leqslant\mathtt{x}\leqslant\mathtt{ub}. \end{aligned} \tag{1}$$

Основными входными параметрами `quadprog` являются: матрица `H` и вектор `f` из целевой функции, матрица ограничений-неравенств `A`, вектор правых частей ограничений-неравенств `b`, матрица ограничений-равенств `Aeq`, вектор правых частей ограничений-равенств `beq`, вектор `lb`, ограничивающий план `x` снизу, вектор `ub`, ограничивающий план `x` сверху. На выходе функция `quadprog` выдаёт оптимальный план `x` задачи (1) и экстремальное значение целевой функции `fval`.

Если матрица `H` несимметрична, то MATLAB заменяет её на $(\mathtt{H}+\mathtt{H}^T)/2$. (При этом значение целевой функции не меняется.)

ПРИМЕР 1. Решим в MATLAB задачу квадратичного программирования

$$\begin{aligned} f(x)=x_1^2+x_2^2+x_3^2-2x_1x_2-x_1-x_2+x_3\to\inf,\\ x_1+x_2+x_3\geqslant 1,\\ 2x_1+x_2-x_3\geqslant -1,\\ x_1-x_2+x_3=0,\\ 0\leqslant x_1\leqslant 1,\\ 0\leqslant x_2\leqslant 1,\\ 0\leqslant x_3\leqslant 1. \end{aligned}$$

*Семинар «DHA & CAGD». Избранные доклады. 12 февраля 2011 г.

Соответствующая программа (m-файл)* выглядит так:

```
clear all
close all
clc % удаляются все текущие переменные из памяти MATLAB,
    закрываются все графические окна, очищается экран консоли
D = [2 -2 0; -2 2 0; 0 0 2]; % строки матрицы разделяются
    точкой с запятой
C = [-1 -1 1]; % задаётся вектор длины три
G = [1 1 1; 2 1 -1];
B = [1 -1];
Aeq = [1 -1 1];
beq = [0];
lb = zeros(3,1); % задаётся нулевой вектор длины три
ub = [1 1 1];
H = D; % коэффициенты квадратичной части целевой функции
f = C; % коэффициенты линейной части целевой функции
A = -G;
b = -B; % появляются знаки «-», так как ограничения-неравенства
    Gx ⩾ B приводятся к виду −Gx ⩽ −B
[x,fval] = quadprog(H,f,A,b,Aeq,beq,lb,ub);
x
fval
```

Запустив программу, получим сообщение

```
Warning: Large-scale algorithm does not currently solve this
   problem formulation, using medium-scale algorithm instead.
Optimization terminated.

x =
  1.0000
  1.0000
  0.0000
fval =
  -2.0000
```

Предупреждение, содержащееся в первых двух строках, прокомментируем позже.

Дополнительно можно задать начальное приближение `x0`:

```
[x,fval] = quadprog(H,f,A,b,Aeq,beq,lb,ub,x0).
```

*Для отладки приведённых в докладе программ использовался MATLAB 7.11.0 (R2010b).

Если какой-то из входных параметров отсутствует, на его место следует поставить квадратные скобки `[]`, за исключением случая, когда это последний параметр в списке. Например, если нужно решить задачу без ограничений-равенств, в которой не задано начальное приближение, то оператор вызова функции `quadprog` будет выглядеть так:

```
[x,fval] = quadprog(H,f,A,b,[],[],lb,ub).
```

(Квадратные скобки в конце списка, соответствующие начальному приближению, не ставятся.)

С помощью входного параметра `options` устанавливаются некоторые дополнительные настройки, в частности, выбирается алгоритм решения. MATLAB решает задачи квадратичного программирования двумя способами: алгоритмом внутренней точки (`Large-Scale Algorithm`) и методом перебора граней (`Medium-Scale Algorithm`). По умолчанию используется алгоритм внутренней точки. Чтобы выбрать метод перебора граней, нужно написать

```
options = optimset('LargeScale','off');
[x,fval] = quadprog(H,f,A,b,Aeq,beq,lb,ub,[],options).
```

Заметим, что MATLAB использует алгоритм внутренней точки для задач двух типов: если есть только ограничения-равенства и матрица ограничений-равенств имеет полный ранг или если есть только нижние `lb` и верхние `ub` границы. Задачу, которая не относится к этим двум типам, MATLAB решает методом перебора граней, выдавая предупреждение (см. пример 1)

```
Warning: Large-scale algorithm does not currently solve this
   problem formulation, using medium-scale algorithm instead.
```

Разберёмся с выходными данными. MATLAB позволяет выводить информацию о том, как завершилось решение задачи. За это отвечает параметр `exitflag`. Если значение `exitflag` равно 1, то найдено решение задачи, если равно 0, то превышено допустимое число итераций, если равно -2 — множество планов задачи пусто, если равно -3 — целевая функция не ограничена снизу на множестве планов. Интерпретация других значений параметра `exitflag` приведена в MATLAB Help. Для метода перебора граней допустимое число итераций (`MaxIter`) по умолчанию равно 200. Значение `MaxIter` можно изменить. Чтобы установить допустимое число итераций равным, к примеру, 10, нужно написать

```
options =
optimset('LargeScale','off','MaxIter',10);
[x,fval] = quadprog(H,f,A,b,Aeq,beq,lb,ub,[],options).
```

Если после выполнения десятой итерации решение не будет найдено, параметр `exitflag` станет нулевым и на экране появится сообщение

```
Maximum number of iterations exceeded;
increase options.MaxIter.
```

Параметр `output` содержит информацию о процессе оптимизации, в частности, число итераций (`iterations`) и используемый алгоритм (`algorithm`). Другие поля параметра `output` описаны в MATLAB Help. Запустим с данными из примера 1 следующую программу:

```
options = optimset('LargeScale','off');
[x,fval,exitflag,output] =
quadprog(H,f,A,b,Aeq,beq,lb,ub,[],options);
exitflag
output.iterations
output.algorithm
```

На выходе получим:

```
Optimization terminated.
exitflag =
  1
ans =
  3
ans =
  medium scale: active-set
```

Это означает, что метод перебора граней успешно завершил работу. Для нахождения решения потребовалось три итерации.

2°. При решении задачи квадратичного программирования возможны три выхода из процесса: найдено решение задачи, множество планов пусто, целевая функция не ограничена снизу на множестве планов. Продемонстрируем эти варианты на примерах.

ПРИМЕР 2. Решим в MATLAB задачу квадратичного программирования

$$\begin{aligned} Q(x) = x_1^2 + x_1 + x_2 \to \inf, \\ x_1 - x_2 = 2, \\ x_1 \geqslant 0,\ x_2 \geqslant 0. \end{aligned} \tag{2}$$

Соответствующая программа будет выглядеть так:

```
clear all
close all
```

```
clc
D = [2 0; 0 0];
C = [1 1];
Aeq = [1 -1];
beq = [2];
lb = zeros(2,1);
H = D;
f = C;
options = optimset('LargeScale','off');
[x,fval,exitflag] =
quadprog(H,f,[],[],Aeq,beq,lb,[],[],options);
x
fval
exitflag
```

В результате работы программы получим:

```
Optimization terminated.
x =
  2
  0
fval =
  6
exitflag =
  1
```

Найдено решение задачи (2).

ПРИМЕР 3. Решим в MATLAB задачу квадратичного программирования

$$\begin{aligned} f(x) = x_1^2 + x_2^2 \to \inf, \\ x_1 + x_2 = 4, \\ x_1 + 2x_2 \geqslant 10, \\ x_1 \geqslant 0,\ x_2 \geqslant 0. \end{aligned} \tag{3}$$

Приведём результат работы соответствующей программы:

```
Exiting: The constraints are overly stringent; no feasible
starting point found.
x =
  -0.6180
  4.6180
fval =
  21.7082
```

```
exitflag =
  -2
```

Множество планов задачи (3) пусто.

ПРИМЕР 4. Решим в MATLAB задачу квадратичного программирования

$$\begin{aligned} f(x) = x_1^2 - x_1 - 4x_2 \to \inf, \\ x_1 + x_2 \geqslant 2, \\ x_1 \geqslant 0,\ x_2 \geqslant 0. \end{aligned} \tag{4}$$

Запустив программу, решающую задачу (4), получим:

```
Exiting: The solution is unbounded and at infinity;
the constraints are not restrictive enough.
x =
  1.0e+016*
  0
  4.0000
fval =
  -1.6000e+017
exitflag =
  -3
```

Целевая функция задачи (4) не ограничена снизу на множестве планов.

3°. Полное описание функции `quadprog` приведено в MATLAB Help.

ГЛАВА 3. НЕЛИНЕЙНЫЕ ЗАДАЧИ

ОСНОВНАЯ ЛЕММА НЕЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ*

В. Н. Малозёмов

1°. Термин «основная лемма» впервые появился в вариационном исчислении.

ОСНОВНАЯ ЛЕММА ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ. *Пусть $\alpha(t)$ и $\beta(t)$ — непрерывные на отрезке $[a,b]$ функции. Если для любой функции $h(t)$, непрерывно дифференцируемой на $[a,b]$ и удовлетворяющей граничным условиям $h(a)=0$, $h(b)=0$, выполняется равенство*

$$\int_a^b \big[\alpha(t)h'(t)+\beta(t)h(t)\big]=0,$$

то необходимо $\alpha(t)$ непрерывно дифференцируема и

$$\alpha'(t)\equiv\beta(t)\quad на\quad [a,b].$$

Эта лемма используется при выводе уравнения Эйлера.

По аналогии, в [1, с. 26–27] введена

ОСНОВНАЯ ЛЕММА ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ. *Допустим, что совместна система линейных соотношений*

$$\begin{gathered}\langle c,x\rangle=\mu,\\ Ax=b,\\ x\geqslant\mathbb{O},\end{gathered}$$

где $A=A[M,N]$ — некоторая матрица, однако при уменьшении μ она становится несовместной. Тогда совместна такая система:

$$\begin{gathered}\langle b,u\rangle=\mu,\\ uA\leqslant c.\end{gathered}$$

*Семинар «DHA & CAGD». Избранные доклады. 25 ноября 2008 г.

С помощью этой леммы выводятся необходимые условия оптимальности в задаче линейного программирования.

Следующее утверждение я называю основной леммой нелинейного программирования.

ОСНОВНАЯ ЛЕММА НЕЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ. *Рассмотрим систему нелинейных уравнений*

$$a_i(x) = 0, \quad i \in I, \tag{1}$$

где $x = x[N]$ *и* N, I *— конечные индексные множества,* $|I| < |N|$. *Пусть точка* x_0 *удовлетворяет системе* (1), *функции* $a_i'(x)$ *при всех* $i \in I$ *непрерывно дифференцируемы в окрестности точки* x_0 *и выполнено условие регулярности: градиенты* $a_i'(x_0)$, $i \in I$, *линейно независимы. Тогда для любого ненулевого вектора* $g_0 \in \mathbb{R}^N$, *ортогонального всем градиентам* $a_i'(x_0)$, *можно построить параметрическую кривую* $x = x(t)$, *непрерывно дифференцируемую в окрестности точки* $t = 0$, *и такую, что*

$$x(0) = x_0, \quad x'(0) = g_0, \tag{2}$$

$$a_i\big(x(t)\big) \equiv 0 \text{ при } t \in (-\delta, \delta) \text{ и всех } i \in I, \tag{3}$$

где δ *— некоторое положительное число.*

На рисунке иллюстрируется содержание этой леммы в случае, когда система (1) состоит из одного уравнения $a(x) = 0$ и $x = (x^1,\ x^2,\ x^3)$.

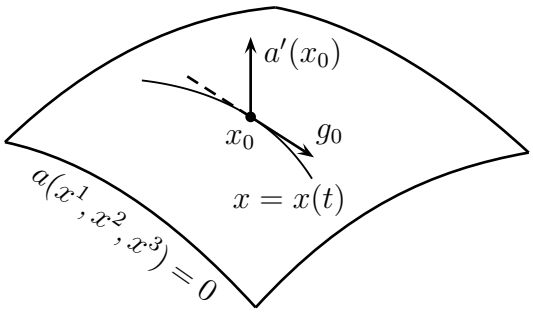


Рис.

В данном докладе приводится доказательство основной леммы нелинейного программирования. Эта лемма используется при выводе необходимых условий оптимальности первого порядка [2] и необходимого условия оптимальности второго порядка [3] в задаче нелинейного программирования.

2°. Начнём с предварительного анализа. Для любой вектор-функции $x = x(t)$, непрерывно дифференцируемой в окрестности точки $t = 0$, справедлива формула

$$\frac{d}{dt}a_i\big(x(t)\big) = \Big\langle a_i'\big(x(t)\big), x'(t)\Big\rangle.$$

По теореме о среднем для функций одной переменной имеем

$$a_i\big(x(t)\big) = a_i\big(x(0)\big) + \Big\langle a_i'\big(x(\xi_i)\big), x'(\xi_i)\Big\rangle t, \tag{4}$$

где $\xi_i \in (0, t)$. Значит, если обеспечить выполнение условий

$$\begin{gathered}\Big\langle a_i'\big(x(t)\big), x'(t)\Big\rangle = 0, \quad t \in (-\delta, \delta),\ i \in I;\\ x(0) = x_0.\end{gathered} \tag{5}$$

то из (4) будет следовать справедливость тождеств (3) (напомним, что точка x_0 удовлетворяет системе (1), так что $a_i\big(x(0)\big) = a_i(x_0) = 0$).

Условия (5) — это система дифференциальных уравнений. Приведём её к нормальной форме. Обозначим через $A'(x)$ матрицу со строками $a_i'(x)$, $i \in I$. Тогда систему (5) можно переписать в виде

$$\begin{gathered}A'(x)x' = \mathbb{O},\\ x(0) = x_0.\end{gathered} \tag{6}$$

По условию леммы строки матрицы $A'(x_0)$ линейно независисмы. Учитывая определение ранга матрицы, заключаем, что существует индексное множество $J \subset N$, такое, что $|J| = |I|$ и подматрица $A'(x_0)[I, J]$ обратима. С помощью этой подматрицы последнее условие леммы (условие ортогональности $A'(x_0)g_0 = \mathbb{O}$) запишем так:

$$A'(x_0)[I, J] \times g_0[J] + A'(x_0)[I, N \setminus J] \times g_0[N \setminus J] = \mathbb{O}[I].$$

Отсюда следует, что

$$g_0[J] = -\big(A'(x_0)[I, J]\big)^{-1} \times A'(x_0)[I, N \setminus J] \times g_0[N \setminus J]. \tag{7}$$

По аналогии, равенство $A'(x)x' = \mathbb{O}$ запишем в виде

$$A'(x)[I, J] \times x'[J] + A'(x)[I, N \setminus J] \times x'[N \setminus J] = \mathbb{O}[I]. \tag{8}$$

Рассмотрим нормальную систему дифференциальных уравнений

$$\begin{cases} x'(t)[N \setminus J] = g_0[N \setminus J], & (9)\\ x'(t)[J] = -\big(A'(x)[I, J]\big)^{-1} \times A'(x)[I, N \setminus J] \times g_0[N \setminus J], & (10)\\ x(0) = x_0. \end{cases}$$

Покажем, что решение этой системы является требуемой параметрической кривой.

Мы хотим воспользоваться теоремой Пеано (см., например, [4, с. 49]). Для этого нужно проверить, что правая часть автономной системы (9), (10) непрерывна в окрестности точки x_0. Элементы матрицы $A'(x)$ непрерывны в окрестности точки x_0 по условию леммы. У невырожденной матрицы $A'(x_0)[I, J]$ определитель отличен от нуля. По теореме о стабилизации знака он остаётся ненулевым и в некоторой окрестности точки x_0. Элементы обратной матрицы $\big(A'(x)[I, J]\big)^{-1}$ можно представить в виде отношения двух определителей, причём в знаменателе стоит определитель матрицы $A'(x)[I, J]$. Ясно, что эти отношения непрерывны в некоторой окрестности точки x_0.

Установлено, что правая часть системы (9), (10) непрерывна в некоторой окрестности точки x_0. По теореме Пеано эта система имеет решение $x = x(t)$, непрерывно дифференцируемое в окрестности точки $t = 0$ и удовлетворяющее начальному условию $x(0) = x_0$. Подставив в (9), (10) $t = 0$ и приняв во внимание равенство (7), получим $x'(0) = g_0$. Значит, выполнены условия (2) леммы.

Теперь подставим $g_0[N \setminus J] = x'(t)[N \setminus J]$ в правую часть (10) и умножим возникающее равенство слева на $A'(x)[I, J]$. Придём к формуле (8), равносильной соотношению $A'(x)x' = \mathbb{O}$. Последнее вместе с равенством $x(0) = x_0$ гарантирует, как отмечалось, выполнение условия (3) леммы.

Утверждение доказано.

ЛИТЕРАТУРА

1. Гавурин М. К., Малоземов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

2. Малозёмов В. Н. *Теорема Куна-Таккера в дифференциальной форме* // Семинар «DHA & CAGD». Избранные доклады. 27 февраля 2010 г. (`http://dha.spb.ru/reps10.shtml#0227`) [Данная книга, с. 210]

3. Малозёмов В. Н. *Условия оптимальности второго порядка в нелинейном программировании* // Семинар «DHA & CAGD». 16 октября 2010 г. (`http://dha.spb.ru/reps10.shtml#1016`) [Данная книга, с. 226]

4. Бибиков Ю. Н. *Общий курс обыкновенных дифференциальных уравнений.* Л.: Изд-во ЛГУ, 1981. 232 с.

ТЕОРЕМА КУНА–ТАККЕРА В ДИФФЕРЕНЦИАЛЬНОЙ ФОРМЕ*

В. Н. Малозёмов

1°. Рассмотрим гладкую задачу нелинейного программирования:

$$\textit{минимизировать } f(x) \tag{1}$$

при ограничениях

$$a_i(x) \geqslant 0, \quad i \in M_1;$$
$$a_i(x) = 0, \quad i \in M_2;$$
$$x \in U.$$

Здесь $U \subset \mathbb{R}^N$ — открытое множество и функции $f(x)$, $a_i(x)$ при всех $i \in M_1 \cup M_2$ непрерывно дифференцируемы на U.

Вектор $x \in \mathbb{R}^N$, удовлетворяющий ограничениям задачи (1), называется ее *планом*. Множество планов обозначим Ω.

Пусть $x \in \Omega$. Обозначим $M = M_1 \cup M_2$,

$$M_1(x) = \{i \in M_1 \mid a_i(x) = 0\},$$

$$I(x) = M_1(x) \cup M_2.$$

На рис. 1 схематично изображены введенные индексные множества. Ясно, что

$$M \setminus I(x) = M_1 \setminus M_1(x). \tag{2}$$

В частности,

$$a_i(x) > 0 \quad \text{при} \quad i \in M \setminus I(x). \tag{3}$$

Множество $I(x)$ называется *множеством индексов активных ограничений* (ограничение с индексом $i \in I(x)$ активно в том смысле, что выполняется как равенство). Если при $i \in I(x)$ градиенты $a_i'(x)$ линейно независимы, то ограничения в точке x называются *регулярными*.

Множество планов Ω задачи (1) и целевая функция $f(x)$ могут быть довольно сложными. Нас интересуют точки локального минимума.

*Семинар «DHA & CAGD». Избранные доклады. 27 февраля 2010 г.

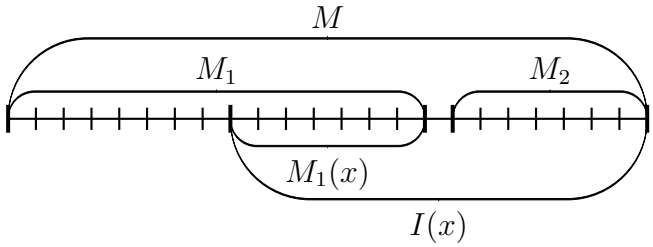


Рис. 1

План x_* называется *точкой локального минимума* в задаче (1), если при некотором $\delta > 0$

$$f(x) \geqslant f(x_*) \qquad \forall x \in \Omega \cap U_\delta(x_*), \tag{4}$$

где $U_\delta(x_*) = \{x \in \mathbb{R}^N \mid \|x - x_*\| < \delta\}$ — открытая δ-окрестность точки x_*.

Теперь мы можем сформулировать теорему Куна–Таккера в дифференциальной форме [1].

ТЕОРЕМА 1. *Пусть $x_* \in \Omega$ — точка локального минимума в задаче* (1) *и ограничения в этой точке регулярны. Тогда найдется вектор $u_* = u_*[I(x_*)]$, такой, что*

$$\begin{gathered} f'(x_*) = \sum_{i \in I(x_*)} u_*[i] a_i'(x_*), \\ u_*[i] \geqslant 0, \quad i \in M_1(x_*). \end{gathered} \tag{5}$$

Теорема Куна–Таккера входит в курс «Экстремальные задачи», который я читаю с 1986 г. на математико-механическом факультете СПбГУ для студентов третьего курса отделения прикладной математики и информатики. За основу было взято доказательство этой теоремы из книги [2], с. 33–42. Постепенно доказательство совершенствовалось. В докладе приведен современный вариант доказательства теоремы Куна–Таккера. Обсуждается вопрос о достаточности условий Куна–Таккера.

2°. Нам понадобятся два вспомогательных утверждения о линейных и нелинейных системах уравнений.

ЛЕММА 1 ([3], с. 25). *Система линейных уравнений*

$$A[N, M] \times u[M] = c[N]$$

имеет решение $u_*[M]$ со свойством $u_*[M_1] \geqslant \mathbb{O}[M_1]$, где $M_1 \subset M$, тогда и только тогда, когда для любого вектора $g \in \mathbb{R}^N$, удовлетворяющего условиям

$$g[N] \times A[N, M_1] \geqslant \mathbb{O}[M_1],$$

$$g[N] \times A[N, M \setminus M_1] = \mathbb{O}[M \setminus M_1],$$

выполняется неравенство $\langle c, g\rangle \geqslant 0$.

Теперь рассмотрим в $\mathbb{R}^N$ систему нелинейных уравнений

$$a_i(x) = 0, \quad i \in I. \tag{6}$$

ЛЕММА 2 ([2], с. 37–38). *Пусть* x_0 *удовлетворяет системе* (6). *Если при этом все функции* $a_i(x)$ *непрерывно дифференцируемы в окрестности* x_0 *и градиенты* $a_i'(x_0)$ *линейно независимы, то для любого ненулевого вектора* $g_0 \in \mathbb{R}^N$, *удовлетворяющего условию* $\langle a_i'(x_0), g_0\rangle = 0$ *при* $i \in I$, *можно построить в* $\mathbb{R}^N$ *параметрическую кривую* $x = x(t)$, *непрерывно дифференцируемую в окрестности точки* $t = 0$, *со свойствами*

$$x(0) = x_0, \qquad x'(0) = g_0,$$

$$a_i\big(x(t)\big) = 0 \ \textit{при} \ i \in I \ \textit{и малых} \ t.$$

На рис. 2 иллюстрируется содержание этой леммы в случае, когда система (6) состоит из одного уравнения $a(x) = 0$, и $x = (x^1, x^2, x^3)$.

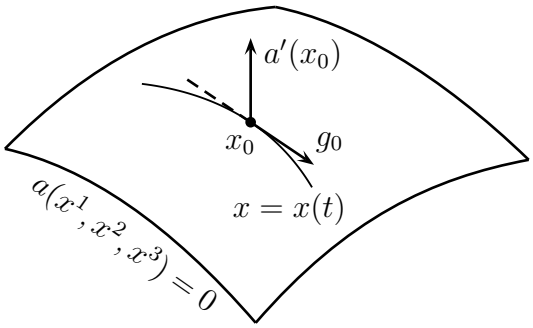


Рис. 2

Лемму 2 я называю *основной леммой нелинейного программирования.*

3°. Обратимся к доказательству теоремы Куна–Таккера. По существу, нужно проверить, что линейная относительно u_* система (5) совместна.

Воспользуемся леммой 1. Покажем, что для любого вектора $g_0 \in \mathbb{R}^N$, удовлетворяющего соотношениям

$$\begin{gathered} \langle a_i'(x_*), g_0\rangle \geqslant 0, \quad i \in M_1(x_*); \\ \langle a_i'(x_*), g_0\rangle = 0, \quad i \in M_2, \end{gathered} \tag{7}$$

выполняется неравенство $\langle f'(x_*), g_0\rangle \geqslant 0$. Отсюда будет следовать заключение теоремы 1.

Зафиксируем соответствующий вектор g_0. Можно считать, что $g_0 \neq \mathbb{O}$. Введем индексное множество

$$I_0 = \{i \in M_1(x_*) \mid \langle a_i'(x_*), g_0\rangle = 0\} \cup M_2.$$

Согласно (7),

$$\langle a_i'(x_*), g_0\rangle = 0, \quad i \in I_0, \tag{8}$$

$$\langle a_i'(x_*), g_0\rangle > 0, \quad i \in I(x_*) \setminus I_0. \tag{9}$$

Рассмотрим систему нелинейных уравнений

$$a_i(x) = 0, \quad i \in I_0.$$

Точка x_* ей удовлетворяет, поскольку $I_0 \subset I(x_*)$. На основании (8) и леммы 2 в $\mathbb{R}^N$ можно построить параметрическую кривую $x = x(t)$, непрерывно дифференцируемую в окрестности точки $t = 0$, со свойствами

$$x(0) = x_*, \qquad x'(0) = g_0, \tag{10}$$

$$a_i\big(x(t)\big) = 0 \ \text{ при } \ i \in I_0 \ \text{ и малых } \ t. \tag{11}$$

Покажем, что $x(t) \in \Omega$ при малых $t > 0$.

При $i \in I_0$ (в частности, при $i \in M_2$) выполняется равенство (11). Пусть $i \in I(x_*) \setminus I_0$. Согласно (10)

$$a_i\big(x(t)\big) = a_i\big(x(0)\big) + \big\langle a_i'\big(x(0)\big), x'(0)\big\rangle\, t + o(t) =$$

$$= t\Big[\langle a_i'(x_*), g_0\rangle + \frac{o(t)}{t}\Big].$$

Последнее выражение при малых $t > 0$ положительно в силу (9).

Осталось рассмотреть индексы $i \in M \setminus I(x_*)$. На них согласно (3) и теореме о стабилизации знака у непрерывной функции при малых t выполняется строгое неравенство $a_i\big(x(t)\big) > 0$. Таким образом, точка $x(t)$ при малых $t > 0$ удовлетворяет всем ограничениям задачи (1), т. е. принадлежит Ω.

По условию теоремы x_* — точка локального минимума в задаче (1), поэтому при малых $t > 0$

$$0 \leqslant f\big(x(t)\big) - f(x_*) = f\big(x(t)\big) - f\big(x(0)\big) = \big\langle f'\big(x(0)\big), x'(0)\big\rangle\, t + o(t) =$$

$$= t\Big[\langle f'(x_*), g_0\rangle + \frac{o(t)}{t}\Big].$$

Поделив это неравенство на $t > 0$ и перейдя к пределу при $t \to +0$, получим $\langle f'(x_*), g_0\rangle \geqslant 0$.

Теорема доказана. □

4°. Условие регулярности ограничений в точке локального минимума существенно для справедливости теоремы Куна–Таккера. Приведем соответствующий пример.

ПРИМЕР 1. На плоскости $x = (u, v)$ введем множество Ω с помощью неравенств

$$a_1(x) := u^3 - v \geqslant 0,$$
$$a_2(x) := -u^4 + v \geqslant 0.$$

Это множество представляет собой лунку (см. рис. 3).

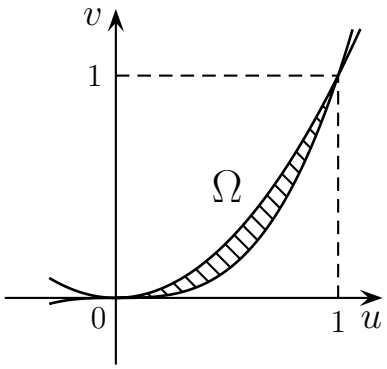


Рис. 3

Рассмотрим экстремальную задачу

$$f_1(x) := u \to \min_{x \in \Omega}. \tag{12}$$

Очевидно, что единственным решением задачи (12) является $x_* = (0, 0)$ (как единственная точка из Ω с наименьшей первой координатой). Выясним, как в этом случае выглядят соотношения (5).

Имеем

$$f_1'(x) = (1, 0), \qquad a_1'(x) = (3u^2, -1), \qquad a_2'(x) = (-4u^3, 1);$$

$$I(x_*) = M(x_*) = \{1, 2\}; \qquad f_1'(x_*) = (1, 0), \quad a_1'(x_*) = (0, -1), \quad a_2'(x_*) = (0, 1).$$

Соотношения (5) принимают вид

$$f_1'(x_*) = u_*[1]a_1'(x_*) + u_*[2]a_2'(x_*), \tag{13}$$

$$u_*[1] \geqslant 0, \quad u_*[2] \geqslant 0.$$

Однако равенство (13) не может выполняться ни при каких $u_*[1]$ и $u_*[2]$, поскольку по первой координате оно переписывается так: $1 = 0$.

Казалось бы, получено противоречие с теоремой 1. Но никакого противоречия нет. В точке x_* ограничения задачи (12) нерегулярны. Действительно, $a_1'(x_*)+a_2'(x_*)=0$, так что градиенты $a_1'(x_*)$ и $a_2'(x_*)$ линейно зависимы. А без условия регулярности ограничений справедливость теоремы 1 не гарантируется.

Впрочем, без гарантии условия (5) могут выполняться и при отсутствии регулярности ограничений. В качестве примера рассмотрим задачу минимизации функции $f_2(x) := v$ на той же лунке Ω. Ее единственным решением является $x_* = (0,0)$, причем, как отмечалось выше, ограничения в этой точке нерегулярны. Вместе с тем, $f_2'(x_*) = (0,1)$ и $f_2'(x_*) = a_2'(x_*)$. Видим, что условия (5) выполняются при $u_*[1]=0$, $u_*[2]=1$.

5°. В соотношения (5) входят первые производные функций, участвующих в постановке задачи (1), поэтому их называют необходимыми условиями оптимальности первого порядка. Выведем достаточные условия первого порядка для точки строгого локального минимума.

Напомним, что план x_* называется *точкой строгого локального минимума* в задаче (1), если при некотором $\delta > 0$

$$f(x) > f(x_*) \quad \forall x \in \Omega \cap \dot{U}_\delta(x_*),$$

где $\dot{U}_\delta(x_*) = U_\delta(x_*) \setminus \{x_*\}$ — проколотая δ-окрестность точки x_*.

Пусть в точке $x_* \in \Omega$ выполнено необходимое условие оптимальности (5). Дополнительно обозначим (см. рис. 4)

$$M_1^+(x_*) = \{i \in M_1(x_*) \mid u_*[i] > 0\},$$
$$I^+(x_*) = M_1^+(x_*) \cup M_2.$$

Отметим, что

$$u_*[i] = 0, \quad i \in I(x_*) \setminus I^+(x_*). \tag{14}$$

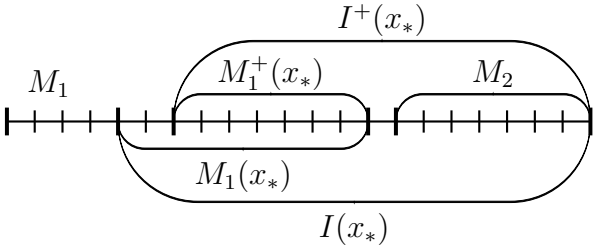


Рис. 4

Рассмотрим систему линейных соотношений

$$\begin{aligned} \langle a_i'(x_*), g\rangle &= 0, \quad i \in I^+(x_*); \\ \langle a_i'(x_*), g\rangle &\geqslant 0, \quad i \in I(x_*) \setminus I^+(x_*). \end{aligned} \tag{15}$$

Множество векторов g, удовлетворяющих (15), является конусом в $\mathbb{R}^N$. Обозначим его G_*.

ТЕОРЕМА 2. *Пусть в точке $x_* \in \Omega$ выполняются условия* (5). *Если при этом $G_* = \{\mathbb{O}\}$, то x_* — точка строгого локального минимума в задаче* (1).

Доказательство. Допустим противное. Тогда найдется последовательность планов $\{y_k\}$ со свойствами

$$y_k \neq x_* \text{ при всех } k; \quad y_k \to x_* \text{ при } k \to \infty;$$

$$f(y_k) \leqslant f(x_*) \text{ при всех } k.$$

Представим y_k в виде $y_k = x_* + \lambda_k g_k$, где $g_k = \dfrac{y_k - x_*}{\|y_k - x_*\|}$ и $\lambda_k = \|y_k - x_*\|$. Ясно, что $\lambda_k > 0$ и $\lambda_k \to +0$ при $k \to \infty$. По определению $\|g_k\| = 1$ при всех k, поэтому можно считать, что $g_k \to g_*$. В силу непрерывности нормы $\|g_*\| = 1$. Покажем, что $g_* \in G_*$.

При $i \in M_1(x_*)$ по теореме о среднем имеем

$$0 \leqslant a_i(y_k) - a_i(x_*) = \langle a_i'(\xi_{ik}), g_k\rangle \lambda_k, \tag{16}$$

где точка ξ_{ik} принадлежит отрезку $[x_*, y_k]$. Аналогично при $i \in M_2$

$$0 = a_i(y_k) - a_i(x_*) = \langle a_i'(\xi_{ik}), g_k\rangle \lambda_k, \tag{17}$$

где ξ_{ik} также принадлежит отрезку $[x_*, y_k]$. Поделив соотношения (16) и (17) на $\lambda_k > 0$ и перейдя к пределу при $k \to \infty$, получим

$$\begin{gathered} \langle a_i'(x_*), g_*\rangle \geqslant 0, \quad i \in M_1(x_*); \\ \langle a_i'(x_*), g_*\rangle = 0, \quad i \in M_2. \end{gathered} \tag{18}$$

Кроме того, из неравенства

$$0 \geqslant f(y_k) - f(x_*) = \langle f'(\eta_k), g_k\rangle \lambda_k,$$

где η_k принадлежит отрезку $[x_*, y_k]$, следует, что

$$\langle f'(x_*), g_*\rangle \leqslant 0. \tag{19}$$

Покажем, что на самом деле

$$\langle a_i'(x_*), g_*\rangle = 0, \quad i \in M_*^+(x_*). \tag{20}$$

Согласно (19), (5), (14) и (18) имеем

$$0 \geqslant \langle f'(x_*), g_*\rangle = \sum_{i \in I(x_*)} u_*[i] \langle a_i'(x_*), g_*\rangle =$$

$$= \sum_{i\in I^+(x_*)} u_*[i]\langle a_i'(x_*), g_*\rangle = \sum_{i\in M_1^+(x_*)} u_*[i]\langle a_i'(x_*), g_*\rangle.$$

В последней сумме все слагаемые неотрицательны, а сумма неположительна. Значит, все слагаемые равны нулю. Если учесть, что $u_*[i] > 0$ при $i \in M_1^+(x_*)$, то приходим к равенствам (20).

На основании (18) и (20) заключаем, что $g_* \in G_*$. Но это противоречит условию $G_* = \{\mathbb{O}\}$.

Теорема доказана. □

СЛЕДСТВИЕ. *Пусть в точке $x_* \in \Omega$ выполнено условие* (5). *Если при этом $|I^+(x_*)| = |N|$ и градиенты $a_i'(x_*)$ при $i \in I^+(x_*)$ линейно независимы, то x_* — точка строгого локального минимума в задаче* (1).

Действительно, в данном случае $G_* = \{\mathbb{O}\}$, что следует из условия

$$\langle a_i'(x_*), g\rangle = 0, \quad i \in I^+(x_*),$$

входящего в определение конуса G_*.

ПРИМЕР 2. Рассмотрим задачу минимизации функции $f_3(x) = -u$ на той же лунке Ω, что и в примере 1. Ее единственным решением является $x_* = (1, 1)$ (как единственная точка из Ω с наибольшей первой координатой). В данном случае

$$I(x_*) = \{1, 2\}; \qquad f_3'(x_*) = (-1, 0), \quad a_1'(x_*) = (3, -1), \quad a_2'(x_*) = (-4, 1),$$

так что

$$f_3'(x_*) = a_1'(x_*) + a_2'(x_*).$$

Условия (5) выполнены с $u_*[1] = u_*[2] = 1$. Более того, $I^+(x_*) = \{1, 2\}$ и градиенты $a_1'(x_*)$, $a_2'(x_*)$ линейно независимы. По следствию из теоремы 2 план x_* является точкой строгого локального минимума. Это соответствует отмеченному выше факту оптимальности x_*.

6°. Условия оптимальности (5) можно представить в другом виде.

Будем говорить, что в точке $x_* \in \Omega$ выполняются *условия Куна–Таккера*, если найдется вектор $u_* = u_*[M]$ со свойствами

$$f'(x_*) = \sum_{i\in M} u_*[i]a_i'(x_*); \tag{21}$$

$$u_*[i]a_i(x_*) = 0, \quad i \in M_1; \tag{22}$$

$$u_*[i] \geqslant 0, \quad i \in M_1. \tag{23}$$

ЛЕММА 3. *Условия оптимальности* (5) *равносильны условиям Куна–Таккера.*

Доказательство. Пусть выполнены условия (5). Дополним вектор $u_*[I(x_*)]$ компонентами $u_*[i] = 0$ при $i \in M \setminus I(x_*)$. Получим вектор $u_*[M]$. Очевидно, что он обладает свойствами (21) и (23). Что касается свойства (22), то следует рассмотреть два случая: $i \in M_1(x_*)$ и $i \in M_1 \setminus M_1(x_*) = M \setminus I(x_*)$ (см. (2)). В первом случае равенство выполняется, поскольку $a_i(x_*) = 0$. Во втором случае нужно учесть, что $u_*[i] = 0$.

Наоборот, пусть выполнены условия Куна–Таккера. В силу (22)

$$u_*[i] = 0 \ \text{ при } \ i \in M_1 \setminus M_1(x_*) = M \setminus I(x_*).$$

Очевидно, что вектор $u_*[I(x_*)]$ удовлетворяет условиям (5).

Лемма доказана. □

С учетом леммы 3 теорему 1 можно переформулировать так.

Пусть $x_* \in \Omega$ — точка локального минимума в задаче (1) *и ограничения в ней регулярны. Тогда в x_* выполняются условия Куна–Таккера.*

Отметим, что условия (21), (22), (23) называются соответственно *условием Лагранжа*, *условием дополнительности* и *условием неотрицательности.*

7°. Обратимся к задаче нелинейного программирования с линейными ограничениями:

$$\begin{gathered} f(x) \to \inf, \\ A[M_1, N] \times x[N] \geqslant b[M_1], \\ A[M_2, N] \times x[N] = b[M_2]. \end{gathered} \tag{24}$$

Множество планов, как и раньше, обозначим Ω. Предположим, что целевая функция $f(x)$ дифференцируема на некотором открытом множестве, содержащем Ω.

ТЕОРЕМА 3. *Пусть $x_* \in \Omega$ — точка локального минимума в задаче* (24). *Тогда найдется вектор $u_* = u_*[M]$, где $M = M_1 \cup M_2$, со свойствами*

$$\begin{gathered} f'(x_*)[N] = u_*[M] \times A[M, N]; \\ u_*[i] \times \big(A[i, N] \times x_*[N] - b[i]\big) = 0, \quad i \in M_1; \\ u_*[i] \geqslant 0, \quad i \in M_1. \end{gathered} \tag{25}$$

Условие регулярности ограничений в точке x_* здесь не требуется.

Простое доказательство теоремы 3, опирающееся на критерий оптимальности в линейном программировании, приведено в [3], с. 88.

ТЕОРЕМА 4. *Предположим, что целевая функция $f(x)$ задачи* (24) *выпукла и дифференцируема на некотором открытом выпуклом множестве, содержащем Ω. Если в точке $x_* \in \Omega$ выполняются условия* (25), *то x_* — точка глобального минимума $f(x)$ на Ω.*

По поводу доказательства см. [3], с. 91.

ЛИТЕРАТУРА

1. Kuhn H. W. *Nonlinear programming: a historical view* / In: SIAM-AMS Proceedings. Vol. IX. AMS, Providence, 1976. P. 1–26.

2. Фиакко А., Мак-Кормик Г. *Нелинейное программирование*. Пер. с англ. М.: Мир, 1972. 240 с.

3. Гавурин М. К., Малоземов В. Н. *Экстремальные задачи с линейными ограничениями*. Л.: Изд-во ЛГУ, 1984. 176 с.

ВОСПОЛЬЗУЕМСЯ ТЕОРЕМОЙ КУНА–ТАККЕРА*

В. Н. Малозёмов

Аннотация. На примере задачи квадратичного программирования показывается, как пользоваться критерием оптимальности в форме Куна–Таккера.

1°. Рассмотрим задачу квадратичного программирования:

$$\begin{gathered} Q(x) := \tfrac{1}{2}\langle Dx, x\rangle + \langle c, x\rangle \to \inf \\ A[M_1, N] \times x[N] \geqslant b[M_1] \\ A[M_2, N] \times x[N] = b[M_2]. \end{gathered} \tag{1}$$

Предполагается, что матрица D симметрична и неотрицательно определена.

Обозначим $M = M_1 \cup M_2$ и сформулируем критерий оптимальности (см., например, [1, с. 91]).

ТЕОРЕМА (Кун–Таккер)**.** *Для того, чтобы план x_* задачи* (1) *был оптимальным, необходимо и достаточно, чтобы нашёлся вектор $u_* = u_*[M]$, такой, что*

$$Q'(x_*) = \sum_{i \in M} u_*[i] \times A[i, N]; \tag{2}$$

$$u_*[i] \times \big(A[i, N] \times x_*[N] - b[i]\big) = 0, \quad i \in M_1; \tag{3}$$

$$u_*[i] \geqslant 0, \quad i \in M_1. \tag{4}$$

Условие (2) называется *условием Лагранжа*, условие (3) — *условием дополнительности*, а условие (4) — *условием неотрицательности*. Вместе условия (2)–(4) называются *условиями Куна–Таккера*.

Для практического применения более удобной является эквивалентная формулировка теоремы Куна–Таккера, в которой отсутствуют множители $u_*[i]$, гарантированно равные нулю.

ТЕОРЕМА. *Вектор $x_* \in \mathbb{R}^N$ будет решением задачи* (1) *тогда и только тогда, когда найдутся подмножество $I \subset M_1$ (не исключая $I = \varnothing$) и числа*

*Семинар «CNSA & NDO». Избранные доклады. 20 ноября 2014 г.

$u_*[i]$, $i \in I \cup M_2$, такие, что вектор $\big(x_*[N], u_*[I \cup M_2]\big)$ удовлетворяет системе линейных уравнений

$$Q'(x) = \sum_{i \in I \cup M_2} u[i] \times A[i, N], \tag{5}$$

$$A[i, N] \times x[N] = b[i], \quad i \in I \cup M_2, \tag{6}$$

причём

$$A[i, N] \times x_*[N] \geqslant b[i], \quad i \in M_1 \setminus I, \tag{7}$$

$$u_*[i] \geqslant 0, \quad i \in I. \tag{8}$$

Проверим эквивалентность условий Куна–Таккера (2)–(4) и условий (5)–(8).

Пусть в точке x_*, являющейся планом задачи (1), выполнены условия Куна–Таккера, то есть найдётся вектор u_* со свойствами (2)–(4). Положим

$$I = \big\{ i \in M_1 \mid A[i, N] \times x_*[N] = b[i] \big\}.$$

Отметим, что

$$A[i, N] \times x_*[N] - b[i] > 0 \quad \text{при} \quad i \in M_1 \setminus I. \tag{9}$$

В силу условия дополнительности (3)

$$u_*[i] = 0 \quad \text{при} \quad i \in M_1 \setminus I. \tag{10}$$

Покажем, что на векторе $\big(x_*[N], u_*[I \cup M_2]\big)$ выполняются условия (5)–(8). Действительно, (5) следует из (2) и (10), (7) — из (9), (8) — из (4), а (6) — из определения I и того факта, что x_* — план задачи (1).

Наоборот, пусть выполняются условия (5)–(8). Доопределим вектор u_*, положив

$$u_*[i] = 0 \quad \text{при} \quad i \in M_1 \setminus I. \tag{11}$$

В силу (6) и (7) вектор x_* является планом задачи (1). Условия (5) и (11) гарантируют выполнение равенства (2). Неравенство (4) следует из (8) и (11). Что касается условия дополнительности (3), то при $i \in I$ оно следует из (6), а при $i \in M_1 \setminus I$ — из (11).

2°. Вторая теорема позволяет описать конечный алгоритм решения задачи (1) при условии, что матрица D симметрична и неотрицательно определена. Нужно последовательно перебирать все подмножества I индексного множества M_1 (включая $I = \varnothing$) и при каждом I решать систему линейных

уравнений (5)–(6), пока не встретится решение (x_*, u_*), удовлетворяющее условиям (7), (8). Вектор x_* будет оптимальным планом задачи (1).

Для контроля правильности вычислений можно использовать равенство

$$\big\langle Q'(x_*), x_* \big\rangle = \sum_{i \in I \cup M_2} u_*[i] \times b[i],$$

которое выводится так: согласно (5) и (6)

$$\begin{aligned} \big\langle Q'(x_*), x_* \big\rangle &= \sum_{i \in I \cup M_2} u_*[i] \times \Big(\big(A[i, N] \times x_*[N] - b[i] \big) + b[i] \Big) = \\ &= \sum_{i \in I \cup M_2} u_*[i] \times b[i]. \end{aligned}$$

Если требуемого подмножества I не найдётся, то задача квадратичного программирования (1) не имеет решения.

3°. Рассмотрим несколько примеров. Во всех случаях квадратичная форма, входящая в целевую функцию, будет неотрицательно определённой. Это проверяется выделением полных квадратов. Например,

$$x_1^2 - x_1 x_2 + x_2^2 = (x_1 - \tfrac{1}{2} x_2)^2 + \tfrac{3}{4} x_2^2.$$

1)
$$\begin{gathered} Q(x) := x_1^2 - 2x_1 x_2 + 3x_2^2 \to \inf \\ 2x_1 - x_2 \geqslant 2 \\ -x_1 + x_2 = -1. \end{gathered}$$

Имеем $M_1 = \{1\}$, $M_2 = \{2\}$ (ограничения нумеруются в порядке их записи). Возьмём $I = \varnothing$. Система линейных уравнений (5)–(6) примет вид (уравнение (5) запишем в транспонированном виде — в виде равенства столбцов):

$$\begin{gathered} \begin{pmatrix} 2x_1 - 2x_2 \\ -2x_1 + 6x_2 \end{pmatrix} = u_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \\ -x_1 + x_2 = -1. \end{gathered}$$

Решение этой системы найти легко: $x_1^* = 1$, $x_2^* = 0$; $u_2^* = -2$. Условие (7) в данном случае выполняется:

$$2x_1^* - x_2^* = 2 \geqslant 2.$$

По теореме Куна–Таккера вектор $x_* = (1, 0)$ является оптимальным планом.

Контроль: $\big\langle Q'(x_*), x_* \big\rangle = 2 = u_2^* b_2$.

2) $$Q(x) := 3x_1^2 - 2x_1x_2 + x_2^2 - 4x_1 + 3x_2 \to \inf$$
$$x_1 - 3x_2 \geqslant -3$$
$$-x_1 + 2x_2 \geqslant 1.$$

Имеем $M_1 = \{1, 2\}$, $M_2 = \varnothing$.

Возьмём $I = \varnothing$. Система (5)–(6) примет вид $Q'(x) = \mathbb{O}$ или в подробной записи

$$6x_1 - 2x_2 - 4 = 0,$$
$$-2x_1 + 2x_2 + 3 = 0.$$

Её решение $x = (\frac{1}{4}, -\frac{5}{4})$ не удовлетворяет условию (7) (не является планом).

Возьмём $I = \{2\}$ и запишем систему линейных уравнений (5)–(6):

$$\begin{pmatrix} 6x_1 - 2x_2 - 4 \\ -2x_1 + 2x_2 + 3 \end{pmatrix} = u_2 \begin{pmatrix} -1 \\ 2 \end{pmatrix},$$
$$-x_1 + 2x_2 = 1.$$

Её решение $x_1^* = \frac{2}{3}$, $x_2^* = \frac{5}{6}$, $u_2^* = \frac{5}{3}$ удовлетворяет условиям (7), (8). Значит, $x_* = (\frac{2}{3}, \frac{5}{6})$ — оптимальный план.

Контроль: $\langle Q'(x_*), x_* \rangle = \frac{5}{3} = u_2^* b_2$.

3) $$Q(x) := x_1^2 - x_1x_2 + x_2^2 - 3x_1 + 5x_2 \to \inf$$
$$2x_1 + x_2 \geqslant 0$$
$$-x_1 + 3x_2 \geqslant -1$$
$$x_1 \geqslant 0$$
$$x_2 \geqslant 0.$$

Имеем $M_1 = \{1,\ 2,\ 3,\ 4\}$, $M_2 = \varnothing$.

Возьмём $I = \{2, 4\}$ и запишем систему (5)–(6):

$$\begin{pmatrix} 2x_1 - x_2 - 3 \\ -x_1 + 2x_2 + 5 \end{pmatrix} = u_2 \begin{pmatrix} -1 \\ 3 \end{pmatrix} + u_4 \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$
$$-x_1 + 3x_2 = -1,$$
$$x_2 = 0.$$

Её решение $x_1^* = 1$, $x_2^* = 0$, $u_2^* = 1$, $u_4^* = 1$ удовлетворяет условиям (7), (8). Значит, $x_* = (1, 0)$ — оптимальный план.

Контроль: $\langle Q'(x_*), x_* \rangle = -1 = u_2^* b_2 + u_4^* b_4$.

4) $$f(x) := x_1^2 + x_1x_2 + 2x_2^2 + 3|x_1 + x_2 - 1| \to \inf\limits_{x\in\mathbb{R}^2}.$$

Эта задача эквивалентна следующей задаче квадратичного программирования:

$$\begin{gathered} Q(x,t) := x_1^2 + x_1x_2 + 2x_2^2 + 3t \to \inf \\ -x_1 - x_2 + t \geqslant -1 \\ x_1 + x_2 + t \geqslant 1. \end{gathered}$$

(По поводу эквивалентности экстремальных задач см. [1, с. 11–12].)

Приступая к решению задачи квадратичного программирования, возьмём $I = \{1, 2\}$. Система (5)–(6) примет вид:

$$\begin{gathered} \begin{pmatrix} 2x_1 + x_2 \\ x_1 + 4x_2 \\ 3 \end{pmatrix} = u_1 \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} + u_2 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \\ -x_1 - x_2 + t = -1, \\ x_1 + x_2 + t = 1. \end{gathered}$$

Её решение $x_1^* = \frac{3}{4}$, $x_2^* = \frac{1}{4}$, $t_* = 0$, $u_1^* = \frac{5}{8}$, $u_2^* = \frac{19}{8}$ удовлетворяет условию (8). Значит, $(x_*, t_*) = (\frac{3}{4}, \frac{1}{4}, 0)$ — оптимальный план.

По эквивалентности, $x_* = (\frac{3}{4}, \frac{1}{4})$ — решение исходной задачи.

5) $$f(x) := x_1^2 + x_1x_2 + 2x_2^2 - 3|x_1 + x_2 - 1| \to \inf\limits_{x\in\mathbb{R}^2}. \tag{12}$$

Приём, который мы использовали при решении предыдущей задачи (у нее изменился только знак коэффициента перед модулем) приводит к задаче квадратичного программирования

$$\begin{gathered} Q(x,t) := x_1^2 + x_1x_2 + 2x_2^2 - 3t \to \inf \\ -x_1 - x_2 + t \geqslant -1 \\ x_1 + x_2 + t \geqslant 1. \end{gathered} \tag{13}$$

Однако задачи (12) и (13) не эквивалентны. Задача (13) не имеет решения. Это следует из того, что при фиксированных x_1, x_2 и положительных t, стремящихся к $+\infty$, вектор (x_1, x_2, t) удовлетворяет ограничениям задачи (13), а целевая функция на нём стремится к $-\infty$. В то же время, как мы покажем, задача (12) имеет решение.

Обозначим

$$\begin{aligned} P_1 &= \{x \in \mathbb{R}^2 \mid x_1 + x_2 - 1 \geqslant 0\}, \\ P_2 &= \{x \in \mathbb{R}^2 \mid x_1 + x_2 - 1 \leqslant 0\}. \end{aligned}$$

Ясно, что $P_1 \cup P_2 = \mathbb{R}^2$, поэтому

$$\inf_{x\in\mathbb{R}^2} f(x) = \min\Big\{ \inf_{x\in P_1} f(x), \inf_{x\in P_2} f(x)\Big\}. \tag{14}$$

Рассмотрим задачу минимизации функции $f(x)$ на P_1:

$$\begin{gathered} f(x) := x_1^2 + x_1x_2 + 2x_2^2 - 3x_1 - 3x_2 + 3 \to \inf \\ x_1 + x_2 \geqslant 1. \end{gathered} \tag{15}$$

Условие $f'(x) = \mathbb{O}$ приводит к системе уравнений

$$\begin{aligned} 2x_1 + x_2 &= 3, \\ x_1 + 4x_2 &= 3. \end{aligned}$$

Её решение $x_1 = \frac{9}{7}$, $x_2 = \frac{3}{7}$ удовлетворяет ограничению задачи (15). Значит, $\hat{x}_* = (\frac{9}{7}, \frac{3}{7})$ — оптимальный план. При этом $f(\hat{x}_*) = \frac{3}{7}$.

Рассмотрим задачу минимизации функции $f(x)$ на P_2:

$$\begin{gathered} f(x) := x_1^2 + x_1x_2 + 2x_2^2 + 3x_1 + 3x_2 - 3 \to \inf \\ -x_1 - x_2 \geqslant -1. \end{gathered} \tag{16}$$

Условие $f'(x) = \mathbb{O}$ приводит к системе уравнений

$$\begin{aligned} 2x_1 + x_2 &= -3, \\ x_1 + 4x_2 &= -3. \end{aligned}$$

Её решение $x_1 = -\frac{9}{7}$, $x_2 = -\frac{3}{7}$ удовлетворяет ограничению задачи (16). Значит, $\check{x}_* = (-\frac{9}{7}, -\frac{3}{7})$ — оптимальный план. При этом $f(\check{x}_*) = -\frac{39}{7}$.

На основании формулы (14) заключаем, что $\check{x}_* = (-\frac{9}{7}, -\frac{3}{7})$ — решение задачи (12).

ЛИТЕРАТУРА

1. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями.* Л.: Изд-во ЛГУ, 1984. 176 с.

УСЛОВИЯ ОПТИМАЛЬНОСТИ ВТОРОГО ПОРЯДКА В НЕЛИНЕЙНОМ ПРОГРАММИРОВАНИИ*

В. Н. Малозёмов

На современном уровне излагается теория условий оптимальности второго порядка в нелинейном программировании (ср. с [1], с. 42–52).

Данный доклад примыкает к докладу [2], посвящённому условиям оптимальности первого порядка.

1°. Рассмотрим задачу нелинейного программирования

$$\begin{gathered} f(x) \to \inf, \\ a_i(x) \geqslant 0, \quad i \in M_1 \\ a_i(x) = 0, \quad i \in M_2; \\ x \in U, \end{gathered} \tag{1}$$

где $U \subset \mathbb{R}^N$ — открытое множество и f, a_i при $i \in M_1 \cup M_2$ — дважды непрерывно дифференцируемые на U функции. Множество планов задачи (1) обозначим Ω.

Зафиксируем $x_* \in \Omega$ и введём индексные множества

$$\begin{gathered} M_1(x_*) = \{i \in M_1 \mid a_i(x_*) = 0\}, \\ I(x_*) = M_1(x_*) \cup M_2. \end{gathered}$$

Говорят, что в точке x_* ограничения задачи (1) *регулярны*, если градиенты $a_i'(x_*)$ при $i \in I(x_*)$ линейно независимы.

Обозначим $M = M_1 \cup M_2$. Говорят, что в точке x_* выполняются *условия Куна–Таккера*, если существует вектор $u_* = u_*[M]$ со свойствами

$$f'(x_*) = \sum_{i \in M} u_*[i]\, a_i'(x_*); \tag{2}$$

$$u_*[i]\, a_i(x_*) = 0, \qquad i \in M_1; \tag{3}$$

$$u_*[i] \geqslant 0, \qquad i \in M_1. \tag{4}$$

*Семинар «DHA & CAGD». Избранные доклады. 16 октября 2010 г.

С помощью функции Лагранжа

$$\mathcal{L}(x,u) = f(x) - \sum_{i\in M} u[i]\, a_i(x)$$

условие (2) можно переписать в виде

$$\mathcal{L}'_x(x_*,u_*) = \mathbb{O}. \tag{5}$$

С двойственным вектором u_* свяжем два индексных множества (см. рис. 1)

$$M_1^+(x_*) = \{i \in M_1(x_*) \mid u_*[i] > 0\},$$
$$I^+(x_*) = M_1^+(x_*) \cup M_2.$$

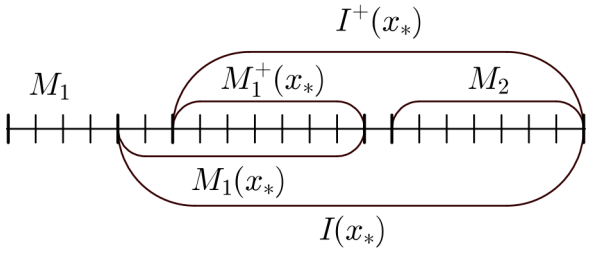


Рис. 1

Отметим, что в силу условия дополнительности (3) и условия неотрицательности (4)

$$u_*[i] = 0, \qquad i \in M \setminus I^+(x_*). \tag{6}$$

Рассмотрим систему линейных соотношений

$$\begin{aligned} &\big\langle a'_i(x_*), g\big\rangle = 0, && i \in I^+(x_*);\\ &\big\langle a'_i(x_*), g\big\rangle \geqslant 0, && i \in I(x_*) \setminus I^+(x_*). \end{aligned} \tag{7}$$

Множество векторов g, удовлетворяющих (7), является конусом в $\mathbb{R}^N$. Обозначим его G_*.

ТЕОРЕМА 1 (необходимое условие оптимальности второго порядка)**.** *Пусть $x_* \in \Omega$ — точка локального минимума в задаче* (1) *и ограничения в ней регулярны. Тогда в x_* выполняются условия Куна–Таккера, то есть найдётся вектор u_* со свойствами* (2)–(4)*. Более того, для всех $g \in G_*$ выполняется неравенство*

$$\big\langle \mathcal{L}''_{xx}(x_*,u_*)g, g\big\rangle \geqslant 0. \tag{8}$$

Доказательство. Выполнение условий Куна–Таккера считаем известным фактом [2]. Проверим выполнение условий второго порядка (8).

Зафиксируем ненулевой вектор $g \in G_*$ и введём индексное множество

$$I_g(x_*) = \{i \in I(x_*) \mid \langle a_i'(x_*), g\rangle\} = 0.$$

Согласно определениям

$$M_2 \subset I^+(x_*) \subset I_g(x_*) \subset I(x_*) \tag{9}$$

и

$$\langle a_i'(x_*), g\rangle > 0, \quad i \in I(x_*) \setminus I_g(x_*). \tag{10}$$

Рассмотрим систему нелинейных уравнений

$$a_i(x) = 0, \quad i \in I_g(x_*). \tag{11}$$

Согласно (9) точка x_* удовлетворяет (11), градиенты $a_i'(x_*)$ при $i \in I_g(x_*)$ линейно независимы и ненулевой вектор g ортогонален $a_i'(x_*)$ при всех $i \in I_g(x_*)$. По основной лемме нелинейного программирования [2] существует параметрическая кривая $x = x(t)$, непрерывно дифференцируемая в окрестности точки $t = 0$, такая, что

$$x(0) = x_*, \quad x'(0) = g, \tag{12}$$

$$a_i\big(x(t)\big) = 0 \text{ при } i \in I_g(x_*) \text{ и малых } t. \tag{13}$$

Покажем, что $x(t) \in \Omega$ при малых $t > 0$.

При $i \in I_g(x_*)$ выполняется равенство (13). В частности, оно выполняется при $i \in M_2$. При $i \in I(x_*) \setminus I_g(x_*)$ согласно (12) имеем

$$a_i\big(x(t)\big) = a_i\big(x(0)\big) + \Big\langle a_i'\big(x(0)\big), x'(0)\Big\rangle\, t + o(t) =$$

$$= a_i(x_*) + \langle a_i'(x_*), g\rangle\, t + o(t) = t\Big[\langle a_i'(x_*), g\rangle + \frac{o(t)}{t}\Big].$$

На основании (10) заключаем, что $a_i\big(x(t)\big) > 0$ при малых $t > 0$. Остаётся рассмотреть индексы $i \in M \setminus I(x_*)$, на которых $a_i(x_*) > 0$. Очевидно, что и на таких индексах будет выполняться неравенство $a_i\big(x(t)\big) > 0$ при малых t. Таким образом, действительно $x(t)$ является планом задачи (1) при малых $t > 0$.

Обратимся к функции Лагранжа. В силу (6), (13) и включения $I^+(x_*) \subset \subset I_g(x_*)$ имеем

$$\mathcal{L}\,(x(t), u_*) = f\,(x(t)) - \sum_{i \in I^+(x_*)} u_*[i]\, a_i\big(x(t)\big) = f\big(x(t)\big).$$

К этому нужно добавить, что $\mathcal{L}(x_*, u_*) = f(x_*)$.

Воспользуемся тем, что x_* — точка локального минимума. При малых $t > 0$ получим

$$0 \leqslant f\big(x(t)\big) - f(x_*) = \mathcal{L}\big(x(t), u_*\big) - \mathcal{L}(x_*, u_*) =$$
$$= \Big\langle \mathcal{L}'_x(x_*, u_*), x(t) - x_* \Big\rangle + \frac{1}{2}\Big\langle \mathcal{L}''_{xx}\big(\xi(t), u_*\big)\big(x(t) - x_*\big), x(t) - x \Big\rangle,$$

где $\xi(t) \in [x_*, x(t)]$. Отметим, что точка $\xi(t)$ может не принадлежать Ω. Нам важно, что $\xi(t) \to x_*$ при $t \to +0$. С учётом (5) перепишем последнее неравенство в эквивалентном виде

$$\Big\langle \mathcal{L}''_{xx}\left(\xi(t), u_*\right) \frac{x(t) - x(0)}{t}, \frac{x(t) - x(0)}{t} \Big\rangle \geqslant 0. \tag{14}$$

Согласно (12)

$$\lim_{t \to +0} \frac{x(t) - x(0)}{t} = g.$$

Кроме того, $\mathcal{L}''_{xx}\left(\xi(t), u_*\right) \to \mathcal{L}''_{xx}\left(x_*, u_*\right)$ при $t \to +0$ в силу дважды непрерывной дифференцируемости функции Лагранжа по x. Переходя к пределу в (14) при $t \to +0$, получаем (8).

Теорема доказана. □

2°. Обратимся к достаточным условиям строгого локального минимума. В [2] установлены достаточные условия первого порядка:

Если в точке $x_* \in \Omega$ выполнены условия Куна–Таккера и $G_* = \{\mathbb{O}\}$, то x_* — точка строгого локального минимума в задаче (1).

Рассмотрим случай, когда G_* состоит не только из нулевого вектора.

ТЕОРЕМА 2 (достаточные условия оптимальности второго порядка)**.** *Пусть в точке $x_* \in \Omega$ выполнены условия Куна–Таккера и для любого ненулевого вектора g из G_**

$$\big\langle \mathcal{L}''_{xx}(x_*, u_*)g, g \big\rangle > 0. \tag{15}$$

Тогда x_* — точка строгого локального минимума в задаче (1).

Доказательство проведём от противного. Предположим, что x_* не является точкой строгого локального минимума. В этом случае найдётся последовательность $\{y_k\}$ точек из Ω, отличных от x_*, со свойствами

$$y_k \to x_* \text{ при } k \to \infty, \tag{16}$$

$$f(y_k) \leqslant f(x_*) \text{ при всех } k. \tag{17}$$

Представим y_k в виде $y_k = x_* + \lambda_k g_k$, где $\lambda_k = \|y_k - x_*\|$ и $g_k = (y_k - x_*)/\lambda_k$. Очевидно, что $\|g_k\| = 1$. Из ограниченной последовательности $\{g_k\}$ можно выделить сходящуюся подпоследовательность. Сходящуюся подпоследовательность снова обозначим $\{g_k\}$. Имеем $g_k \to g_*$ при $k \to \infty$. По непрерывности нормы $\|g_*\| = 1$. Как показано в [2], $g_* \in G_*$. Согласно (15)

$$\big\langle \mathcal{L}''_{xx}(x_*, u_*)g_*, g_* \big\rangle > 0. \tag{18}$$

Вместе с тем, в силу (17), (3) и (4)

$$\mathcal{L}(y_k, u_*) - \mathcal{L}(x_*, u_*) = f(y_k) - f(x_*) - \sum_{i \in M_1} u_*[i]\big(a_i(y_k) - a_i(x_*)\big) \leqslant$$
$$\leqslant -\sum_{i \in M_1} u_*[i]\, a_i(y_k) \leqslant 0.$$

Воспользуемся формулой Тейлора и тем, что $y_k - x_* = \lambda_k g_k$. Получим

$$0 \geqslant \mathcal{L}(y_k, u_*) - \mathcal{L}(x_*, u_*) = \big\langle \mathcal{L}'_x(x_*, u_*), g_k \big\rangle \lambda_k +$$
$$+ \frac{1}{2}\big\langle \mathcal{L}''_{xx}(x_* + \theta_k\, \lambda_k\, g_k, u_*)g_k, g_k \big\rangle \lambda_k^2,$$

где $\theta_k \in (0, 1)$. Отсюда и из (5) следует, что

$$\big\langle \mathcal{L}''_{xx}(x_* + \theta_k\, \lambda_k\, g_k\, u_*)g_k, g_k\, \big\rangle \leqslant 0.$$

Учитывая, что согласно (16) $\lambda_k \to +0$ при $k \to \infty$ и что функция Лагранжа дважды непрерывно дифференцируема по x, в пределе при $k \to \infty$ приходим к неравенству

$$\big\langle \mathcal{L}''_{xx}(x_*, u_*)g_*, g_*\, \big\rangle \leqslant 0.$$

Но это противоречит (18).

Теорема доказана. □

3°. Приведём пример на использование условий оптимальности второго порядка.

ПРИМЕР. Рассмотрим задачу нелинейного программирования

$$f(x) := (x_1 + 1)^2 + x_2^2 \to \inf,$$
$$a(x) := x_1 - \alpha\, x_2^2 \geqslant 0,$$

где $\alpha \in \mathbb{R}$ — параметр. Рис. 2 поясняет, что единственным решением этой задачи (при $\alpha > 0$) является точка $x_* = (0, 0)$. Проверим, как в точке x_* выполняются условия оптимальности.

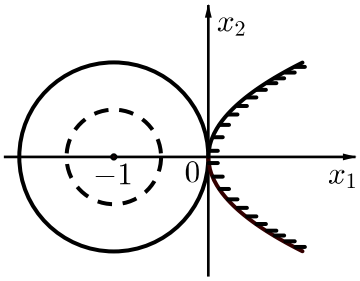


Рис. 2

Имеем

$$f'(x) = (2x_1 + 2, 2x_2), \quad a'(x) = (1, -2\alpha\, x_2);$$
$$f'(x_*) = (2, 0), \qquad\qquad a'(x) = (1, 0);$$
$$f'(x_*) = 2a'(x_*).$$

Положим $u_* = 2$. Видим, что условия Куна–Таккера (2)–(4) выполняются при всех $\alpha \in \mathbb{R}$.

Конус G_* в данном случае определяется условием $\langle a'(x_*), g\rangle = 0$ или $g_1 = 0$. Таким образом, G_* состоит из векторов вида $(0, g_2)$, где $g_2 \in \mathbb{R}$.

Далее,

$$f''(x_*) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad a''(x_*) = \begin{pmatrix} 0 & 0 \\ 0 & -2\alpha \end{pmatrix},$$

так что

$$\mathcal{L}''_{xx}(x_*, u_*) = f''(x_*) - u_* a''(x_*) = \begin{pmatrix} 2 & 0 \\ 0 & 2 + 4\alpha \end{pmatrix}.$$

При $g \in G_*$

$$\langle \mathcal{L}''_{xx}(x_*, u_*) g, g\rangle = (2 + 4\alpha) g_2^2.$$

Условия $g \in G_*$, $g \neq \mathbb{O}$ означают, что $g_2 \neq 0$. Ясно, что при $\alpha > -\frac{1}{2}$ и всех ненулевых g из G_* выполняется неравенство (15). По теореме 2, x_* — точка строгого локального минимума.

Вместе с тем, при $\alpha < -\frac{1}{2}$ неравенство (8) нарушается на любом ненулевом векторе g из G_*. По теореме 1 x_* не является даже точкой локального минимума. Рис. 3 поясняет эту ситуацию.

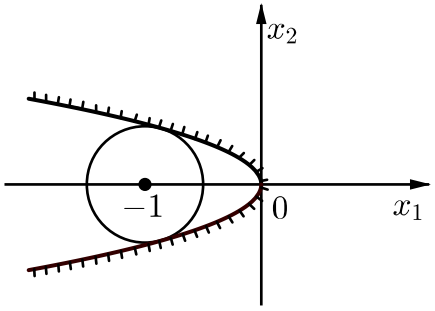


Рис. 3

Остаётся неопределённым случай $\alpha = -\frac{1}{2}$. Рассмотрим его. Перепишем ограничение задачи в эквивалентном виде

$$2x_1 + x_2^2 \geqslant 0.$$

Для любого плана x имеем

$$f(x) = x_1^2 + 2x_1 + 1 + x_2^2 \geqslant x_1^2 + 1 \geqslant 1.$$

Равенство $f(x) = 1$ достигается при выполнении условий $2x_1 + x_2^2 = 0$ и $x_1 = 0$, т. е. в точке $x_* = (0, 0)$. Таким образом, при $\alpha = -\frac{1}{2}$ точка x_* является единственным решением исходной задачи.

ЛИТЕРАТУРА

1. Фиакко А., Мак-Кормик Г. *Нелинейное программирование.* Пер. с англ. М.: Мир, 1972. 240 с.

2. Малоземов В. Н. *Теорема Куна–Таккера в дифференциальной форме* // Семинар «DHA & CAGD». Избранные доклады. 27 февраля 2010 г. (`http://dha.spb.ru/reps10.shtml#0227`) [Данная книга, с. 210]

О СООТНОШЕНИИ ДВОЙСТВЕННОСТИ В МАТЕМАТИЧЕСКОМ ПРОГРАММИРОВАНИИ*

А. В. Лазарев

1°. Рассмотрим в $\mathbb{R}^n$ задачу математического программирования

$$\begin{gathered} f(x) \to \inf, \\ g_i(x) \leqslant 0, \quad i \in 1:s; \\ x \in P, \end{gathered} \tag{1}$$

и двойственную к ней задачу

$$\varphi(y) := \inf\big\{L(x,y) \mid x \in P\big\} \to \sup_{y \in \mathbb{R}^s_+}, \tag{2}$$

где $\mathbb{R}^s_+ = \big\{y = (y_1, \ldots, y_s) \mid y_i \geqslant 0,\ i \in 1:s\big\}$ и

$$L(x,y) = f(x) + \sum_{i=1}^{s} y_i\, g_i(x)$$

— функция Лагранжа. Множество планов задачи (1) обозначим через X. Положим

$$f^* = \inf\big\{f(x) \mid x \in X\big\}, \qquad \varphi^* = \sup\big\{\varphi(y) \mid y \in \mathbb{R}^s_+\big\}.$$

Нас интересуют условия, при выполнении которых справедливо соотношение двойственности $f^* = \varphi^*$.

Этот фундаментальный вопрос изучался во многих работах (см., например, книги [1, 2, 3, 4] и библиографию в них). В данном докладе критерий, обеспечивающий справедливость соотношения двойственности, сформулирован в терминах ε-субдифференциала функции чувствительности задачи (1).

2°. Начнём с простого утверждения.

ЛЕММА 1. *При любых $x \in X$ и любых $y \in \mathbb{R}^s_+$ выполняется неравенство*

$$f(x) \geqslant \varphi(y). \tag{3}$$

*Семинар «DHA & CAGD». Избранные доклады. 17 мая 2008 г.

Доказательство. Возьмём $x \in X$ и $y \in \mathbb{R}^s_+$. Поскольку

$$\sum_{i=1}^{s} y_i\, g_i(x) \leqslant 0\,,$$

то

$$f(x) \geqslant \inf\Big\{ f(x) + \sum_{i=1}^{s} y_i\, g_i(x) \;\Big|\; x \in X \Big\} \geqslant \inf\big\{ L(x,y) \mid x \in P \big\} = \varphi(y)\,,$$

что и требовалось установить. □

Как следствие получаем неравенство

$$f^* \geqslant \varphi^*. \tag{4}$$

При $X \neq \varnothing$ оно следует из (3). Если же $X = \varnothing$, то $f^* = +\infty$ (по определению $\inf\{\varnothing\} = +\infty$, $\sup\{\varnothing\} = -\infty$). В этом случае неравенство (4) тривиально.

При $X = \varnothing$ соотношение двойственности $f^* = \varphi^*$ может как выполняться, так и не выполняться.

ПРИМЕР 1. Рассмотрим задачу минимизации функции одной переменной $f(x)$, тождественно равной нулю, при ограничениях

$$g(x) := x^2 + 1 \leqslant 0, \quad x \in \mathbb{R}\,.$$

Имеем $X = \varnothing$, $f^* = +\infty$,

$$\varphi(y) = \inf\big\{ y\,(x^2+1) \mid x \in \mathbb{R} \big\} = y \quad \text{при } y \geqslant 0\,,$$
$$\varphi^* = \sup\big\{ \varphi(y) \mid y \geqslant 0 \big\} = +\infty\,.$$

Получили $f^* = \varphi^*$.

ПРИМЕР 2. Рассмотрим задачу

$$f(x) := -x^4 \to \inf\,,$$
$$g(x) := x^2 + 1 \leqslant 0\,, \quad x \in \mathbb{R}\,.$$

Имеем

$$\varphi(y) = \inf\big\{ -x^4 + y\,(x^2+1) \mid x \in \mathbb{R} \big\} = -\infty \quad \text{при всех } y \geqslant 0\,,$$

так что $\varphi^* = -\infty$. Поскольку $f^* = +\infty$, то $f^* \neq \varphi^*$.

Предположим, что $X \neq \varnothing$. Тогда $\varphi^* < +\infty$. В противном случае из (4) следовало бы, что $f^* = +\infty$. Но это возможно лишь тогда, когда $X = \varnothing$.

При $f^* = -\infty$ неравенство (4) обращается в равенство. Но это неинтересный случай. В дальнейшем считаем, что

$$X \neq \varnothing \quad \text{и} \quad f^* > -\infty\,. \tag{5}$$

3°. Рассмотрим «возмущённую» задачу

$$f(x) \to \inf\,,$$
$$g_i(x) \leqslant v_i\,,\ i \in 1:s\,;$$
$$x \in P\,,$$

где $v = (v_1, \ldots, v_s)$ — произвольный вектор параметров. Множество планов этой задачи обозначим через $X(v)$.

Функция

$$F(v) = \inf\big\{f(x) \mid x \in X(v)\big\}$$

называется *функцией чувствительности* задачи (1). Очевидно, что $F(\mathbb{O}) = f^*$ и

$$F(v) \geqslant F(u) \quad \text{при } v \leqslant u\,.$$

В частности, при всех $i \in 1:s$ и $t > 0$

$$f^* \geqslant F(t\,e_i)\,.$$

Допустим, что $F(v) = -\infty$ при некотором $v \in \mathbb{R}^s$. В этом случае существует последовательность точек $x^k \in X(v)$, $k = 1, 2, \ldots$, удовлетворяющая условию $f(x^k) \to -\infty$. Для любого $y \in \mathbb{R}^s_+$ имеем

$$f(x^k) + \sum_{i=1}^{s} y_i\,g_i(x^k) \leqslant f(x^k) + \sum_{i=1}^{s} y_i\,v_i\,.$$

Отсюда следует, что

$$\varphi(y) := \inf\left\{f(x) + \sum_{i=1}^{s} y_i\,g_i(x)\,\middle|\,x \in P\right\} = -\infty \quad \text{при всех } y \in \mathbb{R}^s_+\,.$$

Значит, $\varphi^* = -\infty$. Мы предположили, что $f^* > -\infty$, поэтому соотношение двойственности в данном случае выполняться не может. В дальнейшем будем считать, что наряду с (5) выполняется условие

$$F(v) > -\infty \quad \text{при всех } v \in \mathbb{R}^s\,. \tag{6}$$

ЛЕММА 2. *При всех $y \in \mathbb{R}^s_+$ справедлива формула*

$$\varphi(y) = \inf\big\{F(v) + \langle y, v\rangle \mid v \in \mathbb{R}^s\big\}. \tag{7}$$

Доказательство. Возьмём $y \in \mathbb{R}^s_+$ и произвольный вектор $x \in P$. Обозначим $v_i = g_i(x)$, $i \in 1:s$; $v = (v_1, \ldots, v_s)$. Очевидно, что $x \in X(v)$, поэтому $f(x) \geqslant F(v)$. Далее

$$f(x) + \sum_{i=1}^{s} y_i\, g_i(x) \geqslant F(v) + \sum_{i=1}^{s} y_i\, v_i\,.$$

Отсюда следует, что

$$\varphi(y) \geqslant \inf\big\{F(v) + \langle y, v\rangle \mid v \in \mathbb{R}^s\big\}. \tag{8}$$

Допустим, что неравенство (8) выполняется как строгое. По определению инфимума найдётся вектор $w \in \mathbb{R}^s$, такой, что

$$\varphi(y) > F(w) + \langle y, w\rangle.$$

Обозначим через ε разность между левой и правой частями последнего неравенства. По определению функции чувствительности найдётся точка $x^* \in P$ со свойствами: $g_i(x^*) \leqslant w_i$, $i \in 1:s$; $f(x^*) \leqslant F(w) + \frac{\varepsilon}{2}$. Принимая во внимание, что $y \in \mathbb{R}^s_+$, получаем

$$F(w) + \sum_{i=1}^{s} y_i\, w_i \geqslant f(x^*) - \frac{\varepsilon}{2} + \sum_{i=1}^{s} y_i\, g_i(x^*)\,.$$

В таком случае

$$\varepsilon \leqslant \Big[f(x^*) + \sum_{i=1}^{s} y_i\, g_i(x^*)\Big] - \Big[F(w) + \sum_{i=1}^{s} y_i\, w_i\Big] \leqslant \frac{\varepsilon}{2}\,,$$

что невозможно. Лемма доказана. □

4°. Множества

$$\partial F(\mathbb{O}) = \big\{y \in \mathbb{R}^s \mid F(v) - F(\mathbb{O}) \geqslant \langle y, v\rangle \text{ при всех } v \in \mathbb{R}^s\big\},$$

$$\partial_\varepsilon F(\mathbb{O}) = \big\{y \in \mathbb{R}^s \mid F(v) - F(\mathbb{O}) \geqslant \langle y, v\rangle - \varepsilon \text{ при всех } v \in \mathbb{R}^s\big\}, \quad \varepsilon > 0\,,$$

называются соответственно *субдифференциалом* и *ε-субдифференциалом* функции чувствительности F в нуле.

Сформулируем основной результат доклада.

ТЕОРЕМА 1. *Пусть выполнены условия* (5), (6). *Для того чтобы имело место соотношение двойственности* $f^* = \varphi^*$, *необходимо и достаточно, чтобы множество* $\partial_\varepsilon F(\mathbb{O})$ *было непустым при всех* $\varepsilon > 0$.

Доказательство. Необходимость. Пусть $f^* = \varphi^*$. Поскольку $f^* = F(\mathbb{O})$, то

$$F(\mathbb{O}) = \varphi^* = \sup\left\{\varphi(y) \mid y \in \mathbb{R}^s_+\right\}.$$

Согласно (7)

$$F(\mathbb{O}) = \sup\left\{ \inf\left\{F(v) + \langle y, v\rangle \mid v \in \mathbb{R}^s\right\} \,\middle|\, y \in \mathbb{R}^s_+\right\}.$$

Возьмём последовательность положительных чисел $\{\varepsilon_k\}$, стремящуюся к нулю. По определению супремума при любом k найдётся вектор $y^k \in \mathbb{R}^s_+$, такой, что

$$F(\mathbb{O}) - \varepsilon_k \leqslant \inf\left\{F(v) + \langle y^k, v\rangle \mid v \in \mathbb{R}^s\right\} \leqslant$$
$$\leqslant F(v) + \langle y^k, v\rangle \quad \text{при всех } v \in \mathbb{R}^s.$$

Значит,

$$F(v) - F(\mathbb{O}) \geqslant \langle -y^k, v\rangle - \varepsilon_k \qquad \forall\, v \in \mathbb{R}^s.$$

Зафиксируем $\varepsilon > 0$ и выберем k из условия $\varepsilon_k < \varepsilon$. Получим

$$F(v) - F(\mathbb{O}) \geqslant \langle -y^k, v\rangle - \varepsilon \qquad \forall\, v \in \mathbb{R}^s,$$

то есть $-y^k \in \partial_\varepsilon F(\mathbb{O})$. Непустота множества $\partial_\varepsilon F(\mathbb{O})$ установлена.

Достаточность. Пусть $\partial_\varepsilon F(\mathbb{O}) \neq \varnothing$ при всех $\varepsilon > 0$. Возьмём последовательность положительных чисел $\{\varepsilon_k\}$, стремящуюся к нулю. При каждом k найдётся вектор $y^k \in \mathbb{R}^s$, такой, что

$$F(v) - F(\mathbb{O}) \geqslant \langle y^k, v\rangle - \varepsilon_k \qquad \forall\, v \in \mathbb{R}^s. \tag{9}$$

В частности, при всех $i \in 1:s$ и $t > 0$

$$f^* \geqslant F(t\,e_i) \geqslant F(\mathbb{O}) + t\,y^k_i - \varepsilon_k\,.$$

Отсюда следует, что $y^k \leqslant \mathbb{O}$ или $-y^k \in \mathbb{R}^s_+$. На основании (9) получаем

$$\inf\left\{F(v) + \langle -y^k, v\rangle \mid v \in \mathbb{R}^s\right\} \geqslant F(\mathbb{O}) - \varepsilon_k\,,$$

так что

$$\varphi^* \geqslant \varphi(-y^k) \geqslant F(\mathbb{O}) - \varepsilon_k = f^* - \varepsilon_k\,.$$

В пределе при $k \to +\infty$ приходим к неравенству $\varphi^* \geqslant f^*$, которое вместе с (4) гарантирует справедливость равенства $\varphi^* = f^*$.

Теорема доказана. □

5°. Вернёмся к двойственной задаче (2), в которой максимизируется вогнутая функция $\varphi(y)$ на множестве $\mathbb{R}^s_+$.

ТЕОРЕМА 2. *Предположим, что $f^* = \varphi^*$. Тогда множество решений задачи* (2) *совпадает с $-\partial F(\mathbb{O})$.*

Доказательство. Возьмём вектор $y^* \in -\partial F(\mathbb{O})$. По определению субдифференциала

$$F(v) - F(\mathbb{O}) \geqslant \langle -y^*, v\rangle \qquad \forall\, v \in \mathbb{R}^s.$$

Так же, как в теореме 1, проверяется, что $y^* \in \mathbb{R}^s_+$. Имеем

$$f^* = F(\mathbb{O}) \leqslant F(v) + \langle y^*, v\rangle \qquad \forall\, v \in \mathbb{R}^s,$$

так что согласно лемме 2

$$f^* \leqslant \inf\big\{F(v) + \langle y^*, v\rangle \mid v \in \mathbb{R}^s\big\} = \varphi(y^*) \leqslant \varphi^*.$$

Поскольку $f^* = \varphi^*$, то y^* — решение двойственной задачи.

Наоборот, пусть y^* — решение двойственной задачи. Тогда

$$f^* = \varphi^* = \inf\big\{F(v) + \langle y^*, v\rangle \mid v \in \mathbb{R}^s\big\}.$$

Значит, при всех $v \in \mathbb{R}^s$ справедливо неравенство

$$f^* = F(\mathbb{O}) \leqslant F(v) + \langle y^*, v\rangle.$$

Отсюда следует, что $-y^* \in \partial F(\mathbb{O})$. Теорема доказана. □

6°. Наиболее интересен случай, когда $f^* = \varphi^*$ и двойственная задача (2) имеет решение y^*.

Обозначим

$$X^* = \big\{x \in X \mid f(x) = f^*\big\}, \quad X^*(y^*) = \big\{x \in P \mid L(x, y^*) = \varphi^*\big\}.$$

ТЕОРЕМА 3. *Справедливо включение $X^* \subset X^*(y^*)$. При этом для любого $x^* \in X^*$ выполняется условие дополнительности*

$$y_i^*\, g_i(x^*) = 0, \qquad i \in 1:s\,. \tag{10}$$

Доказательство. Возьмём $x^* \in X^*$. Учитывая, что $x^* \in P$, $g_i(x^*) \leqslant 0$ при $i \in 1:s$ и $y^* \in \mathbb{R}^s_+$, получаем

$$\begin{aligned} L(x^*, y^*) &= f(x^*) + \sum_{i=1}^{s} y_i^*\, g_i(x^*) \leqslant f(x^*) = f^* = \\ &= \varphi^* = \varphi(y^*) = \inf\big\{L(x, y^*) \mid x \in P\big\} \leqslant L(x^*, y^*)\,. \end{aligned}$$

Значит,

$$L(x^*, y^*) = \varphi^* \quad \text{и} \quad \sum_{i=1}^{s} y_i^*\, g_i(x^*) = 0\,.$$

Первое равенство свидетельствует о том, что $x^* \in X^*(y^*)$, а из второго в силу неположительности слагаемых следует условие дополнительности (10). Теорема доказана. □

Отметим, что в общем случае равенство $X^* = X^*(y^*)$ гарантировать нельзя.

ПРИМЕР 3 ([1, с. 159]). Рассмотрим экстремальную задачу

$$f(x) := \max\left\{0, x_1 + \tfrac{1}{x_2}\right\} \to \inf\,,$$
$$g(x) := -x_1 \leqslant 0\,, \quad x \in P := \{x \in \mathbb{R}^2 \mid x_2 \geqslant 1\}.$$

В данном случае $X = \{x \in \mathbb{R}^2 \mid x_1 \geqslant 0, x_2 \geqslant 1\}$. Функция $f(x)$ положительна на X и $f(x^k) \to 0$ на последовательности $x^k = (0, k)$, $k = 1, 2, \ldots$, точек из X. Значит, $f^* := \inf\left\{f(x) \mid x \in X\right\} = 0$, но инфимум не достигается, то есть $X^* = \varnothing$.

Обратимся к двойственной задаче. Имеем

$$L(x, y) = \max\left\{0, x_1 + \tfrac{1}{x_2}\right\} - y\, x_1\,, \quad x \in P,\ y \in \mathbb{R}_+\,,$$

и

$$\varphi(0) = \inf\left\{ \max\left\{0, x_1 + \tfrac{1}{x_2}\right\} \,\middle|\, x \in P\right\} = 0\,.$$

Инфимум достигается на точках $x \in P$, у которых $x_1 + \frac{1}{x_2} \leqslant 0$.

Далее

$$\varphi^* = \sup\left\{\varphi(y) \mid y \in \mathbb{R}_+\right\} \geqslant \varphi(0) = 0 = f^*.$$

Учитывая (4), получаем $\varphi^* = f^*$ и $\varphi(0) = \varphi^*$. Таким образом, точка $y^* = 0$ является решением двойственной задачи. Множество $X^*(y^*)$ состоит из точек $x = (x_1, x_2)$, у которых $x_2 \geqslant 1$ и $x_1 + \frac{1}{x_2} \leqslant 0$.

Видим, что $X^* \neq X^*(y^*)$.

Вместе с тем, легко доказать такое утверждение:

Пусть $f^* = \varphi^*$ и y^* — решение двойственной задачи. Если $x^* \in X$, $x^* \in X^*(y^*)$ и выполнено условие дополнительности (10)*, то $x^* \in X^*$.*

Действительно,

$$f^* = \varphi^* = \varphi(y^*) = L(x^*, y^*) = f(x^*) + \sum_{i=1}^{s} y_i^*\, g_i(x^*) = f(x^*)\,,$$

так что $x^* \in X^*$.

7°. Отметим ещё одно свойство двойственной задачи.

ТЕОРЕМА 4. *Пусть $f^* = \varphi^*$ и двойственная задача* (2) *не имеет решения. Тогда любая максимизирующая последовательность $\{y^k\}$ точек из $\mathbb{R}^s_+$, такая, что $\varphi(y^k) \to \varphi^*$ при $k \to +\infty$, является неограниченной.*

Доказательство. Допустим противное, что существует ограниченная максимизирующая последовательность $\{y^k\}$ точек из $\mathbb{R}^s_+$. Для определённости будем считать, что $y^k \to y^*$ при $k \to +\infty$, $y^* \in \mathbb{R}^s_+$. Обозначим

$$\varepsilon_k = \varphi^* - \varphi(y^k)\,, \quad k = 1, 2, \ldots$$

По условию последовательность $\{\varepsilon_k\}$ стремится к нулю.

Имеем

$$\varphi(y^k) = \varphi^* - \varepsilon_k = f^* - \varepsilon_k = F(\mathbb{O}) - \varepsilon_k\,.$$

Согласно (7)

$$\inf\left\{F(v) + \langle y^k, v\rangle \mid v \in \mathbb{R}^s\right\} \geqslant F(\mathbb{O}) - \varepsilon_k\,,$$

так что

$$F(v) - F(\mathbb{O}) \geqslant \langle -y^k, v\rangle - \varepsilon_k \qquad \forall\, v \in \mathbb{R}^s.$$

В пределе получаем

$$F(v) - F(\mathbb{O}) \geqslant \langle -y^*, v\rangle \qquad \forall\, v \in \mathbb{R}^s.$$

Это значит, что $-y^* \in \partial F(\mathbb{O})$. По теореме 2 вектор y^* является решением двойственной задачи. Но по условию двойственная задача не имеет решения. Получили противоречие. Теорема доказана. □

ЛИТЕРАТУРА

1. Сухарев А. Г., Тимохов А. В., Фёдоров В. В. *Курс методов оптимизации.* М.: Наука, 1986.

2. Эльстер К.-Х., Рейнгардт Р., Шойбле М., Донат Г. *Введение в нелинейное программирование.* М.: Мир, 1985.

3. Лоран П.-Ж. *Аппроксимация и оптимизация.* М.: Мир, 1975.

4. Рокафеллар Р. *Выпуклый анализ.* М.: Мир, 1973.

НЕОБХОДИМЫЕ УСЛОВИЯ ГЛОБАЛЬНОЙ ОПТИМАЛЬНОСТИ*

А. В. Лазарев

Данный доклад примыкает к докладу [1].

1°. Рассмотрим в $\mathbb{R}^n$ задачу математического программирования

$$\begin{gathered} f(x) \to \inf, \\ g_i(x) \leqslant 0, \quad i \in 1:s; \\ x \in P. \end{gathered} \tag{1}$$

Предполагается, что $P \subset \mathbb{R}^n$ — выпуклое множество, и что функции f, $g_1, \ldots, g_s$ дифференцируемы на P (если P не открытое множество, то дифференцируемы на некотором открытом множестве, содержащем P). Кроме того, считаем, что множество планов X задачи (1) непусто и величина $f^* = \inf\{f(x) \mid x \in X\}$ конечна.

Введём функцию Лагранжа

$$L(x, y) = f(x) + \sum_{i=1}^{s} y_i\, g_i(x)$$

и наряду с (1) рассмотрим двойственную задачу

$$\varphi(y) := \inf\{L(x, y) \mid x \in P\} \to \sup_{y \in \mathbb{R}^s_+}, \tag{2}$$

где $\mathbb{R}^s_+ = \{y = (y_1, \ldots, y_s) \mid y_i \geqslant 0 \text{ при всех } i \in 1:s\}$. Обозначим $\varphi^* = \sup\{\varphi(y) \mid y \in \mathbb{R}^s_+\}$ и в дальнейшем будем предполагать, что выполнено соотношение двойственности

$$f^* = \varphi^*. \tag{3}$$

Зафиксируем $v = (v_1, \ldots, v_s)$. Множество векторов x, удовлетворяющих возмущённым ограничениям

$$g_i(x) \leqslant v_i,\ i \in 1:s; \quad x \in P,$$

*Семинар «DHA & CAGD». Избранные доклады. 9 сентября 2008 г.

обозначим $X(v)$. Функция $F(v) = \inf\{f(x) \mid x \in X(v)\}$ называется функцией чувствительности задачи (1). Будем говорить, что выполнено условие *глобальной регулярности* задачи (1), если субдифференциал функции чувствительности в нуле непуст, то есть $\partial F(\mathbb{O}) \neq \varnothing$.

ТЕОРЕМА 1. *Допустим, что выполнены соотношение двойственности* (3) *и условие глобальной регулярности. Тогда если* x^* *— глобальное решение задачи* (1)*, то при любом* $y^* \in -\partial F(\mathbb{O})$ *точка* x^* *является решением задачи*

$$L(x, y^*) \to \inf_{x \in P}. \tag{4}$$

При этом

$$\langle L'_x(x^*, y^*),\, x - x^* \rangle \geqslant 0\,, \quad x \in P\,; \tag{5}$$

$$y_i^*\, g_i(x^*) = 0\,, \quad i \in 1:s\,. \tag{6}$$

Доказательство. Напомним [1, теорема 2], что в случае $f^* = \varphi^*$ множество $-\partial F(\mathbb{O})$ совпадает с множеством решений двойственной задачи (2). То, что x^* является решением задачи (4) и выполнены условия дополнительности (6), установлено в [1, теорема 3].

Осталось проверить справедливость неравенства (5). Оно становится очевидным, если учесть, что x^* — оптимальный план задачи (4), функция Лагранжа $L(x, y^*)$ дифференцируема на P и P — выпуклое множество (см., например, [2, с. 87]). Теорема доказана. □

Замечание. Если x^* — внутренняя точка множества P, то условие (5) эквивалентно условию Лагранжа

$$f'(x^*) + \sum_{i=1}^{s} y_i^*\, g_i'(x^*) \;=\; \mathbb{O}\,.$$

2°. Приведём простое достаточное условие глобальной регулярности.

ТЕОРЕМА 2. *Пусть для задачи* (1) *выполнено соотношение двойственности* (3) *и существует точка* $z \in P$*, такая, что* $g_i(z) < 0$ *при всех* $i \in 1:s$*. Тогда* $\partial F(\mathbb{O}) \neq \varnothing$*.*

Доказательство. Предположим, что $\partial F(\mathbb{O}) = \varnothing$. Это значит, что двойственная задача (2) не имеет решения [1, теорема 2]. Возьмём максимизирующую последовательность планов $\{y^k\}$ задачи (2). Имеем $y^k \in \mathbb{R}^s_+$ при всех k и $\varphi(y^k) \to \varphi^*$. Как установлено в [1, теорема 4], последовательность $\{y^k\}$ неограниченна. Можно считать, что $\|y^k\| \to \infty$.

Обозначим $\varepsilon_k = \varphi^* - \varphi(y^k)$. Ясно, что $\varepsilon_k > 0$ и $\varepsilon_k \to 0$ при $k \to \infty$. В силу равенств $\varphi^* = f^*$ и $f^* = F(\mathbb{O})$ получаем

$$\varphi(y^k) = F(\mathbb{O}) - \varepsilon_k\,, \quad k = 1, 2, \ldots \tag{7}$$

Напомним, что

$$\varphi(y) = \inf\left\{F(v) + \langle y, v\rangle \mid v \in \mathbb{R}^s\right\}$$

[1, лемма 2]. Отсюда и из (7) следует, что

$$F(v) - F(\mathbb{O}) \geqslant \langle y^k, -v\rangle - \varepsilon_k \quad \text{при всех } v \in \mathbb{R}^s. \tag{8}$$

Введём вектор $v^* = (v_1^*, \ldots, v_s^*)$ с компонентами $v_i^* = g_i(z) < 0$. Подставив в (8) $v = v^*$, придём к неравенству

$$F(v^*) - F(\mathbb{O}) \geqslant \langle y^k, -v^*\rangle - \varepsilon_k\,, \quad k = 1, 2, \ldots$$

Положим $c = \min\left\{-v_i^* \mid i \in 1:s\right\} > 0$. Поскольку $y^k \in \mathbb{R}^s_+$, то

$$\langle y^k, -v^*\rangle \geqslant c\sum_{i=1}^{s} |y^k| \geqslant c\,\|y^k\|\,,$$

так что

$$F(v^*) - F(\mathbb{O}) \geqslant c\,\|y^k\| - \varepsilon_k\,, \quad k = 1, 2, \ldots \tag{9}$$

Величина $F(v^*)$ конечна, так как $X(v^*) \neq \varnothing$ и $X(v^*) \subset X$. В то же время правая часть неравенства (9) стремится к $+\infty$ при $k \to \infty$. Полученное противоречие завершает доказательство теоремы. □

3°. Исследуем случай $\partial F(\mathbb{O}) = \varnothing$ при сохранении соотношения двойственности.

ТЕОРЕМА 3. *Существует вектор $u^* \in \mathbb{R}^s_+$, $u^* \neq \mathbb{O}$, такой, что оптимальный план x^* задачи* (1) *является решением следующей задачи*

$$\sum_{i=1}^{s} u_i^*\, g_i(x) \to \inf_{x \in P}. \tag{10}$$

При этом выполняются соотношения

$$\left\langle \sum_{i=1}^{s} u_i^*\, g_i'(x),\, x - x^* \right\rangle \geqslant 0\,, \quad x \in P\,; \tag{11}$$

$$u_i^*\, g_i(x^*) = 0\,, \quad i \in 1:s\,. \tag{12}$$

Доказательство. По условию

$$f(x^*) = f^* = \varphi^* = \sup_{y \in \mathbb{R}^s_+} \varphi(y)\,.$$

Возьмём последовательность положительных чисел $\{\varepsilon_k\}$, стремящуюся к нулю. Согласно определению супремума найдётся последовательность векторов $\{y^k\}$, такая, что $y^k \in \mathbb{R}^s_+$ и

$$f(x^*) - \varepsilon_k \leqslant \varphi(y^k)\,, \quad k = 1, 2, \ldots \tag{13}$$

В частности, $\varphi(y^k) \to \varphi^*$. Отсюда, как отмечалось, следует, что последовательность $\{y^k\}$ неограниченна. Можно считать, что $\|y^k\| \to \infty$.

Положим $u^k = y^k/\|y^k\|$. Из последовательности единичных векторов $\{u^k\}$ можно выделить сходящуюся подпоследовательность. Не умаляя общности, будем считать, что $u^k \to u^*$, при этом $u^* \in \mathbb{R}^s_+$, $\|u^*\| = 1$. Покажем, что u^* — требуемый вектор.

Согласно (13)

$$f(x^*) - \varepsilon_k \leqslant \inf\left\{L(x, y^k) \mid x \in P\right\} \leqslant f(x^*) + \sum_{i=1}^{s} y_i^k\, g_i(x^*) \leqslant f(x^*)\,.$$

Значит,

$$\sum_{i=1}^{s} y_i^k\, g_i(x^*) \to 0 \quad \text{при} \quad k \to \infty$$

и тем более

$$\sum_{i=1}^{s} u_i^k\, g_i(x^*) \to 0 \quad \text{при} \quad k \to \infty.$$

В пределе получаем

$$\sum_{i=1}^{s} u_i^*\, g_i(x^*) = 0\,. \tag{14}$$

Отсюда следуют условия дополнительности (12).

Покажем, что

$$\sum_{i=1}^{s} u_i^*\, g_i(x) \geqslant 0 \quad \text{при всех} \quad x \in P.$$

Вместе с (14) это будет гарантировать оптимальность x^* применительно к задаче (10). Допустим вопреки утверждению, что существует вектор $z \in P$, такой, что $\sum\limits_{i=1}^{s} u_i^*\, g_i(z) < 0$. Имеем

$$f(x^*) - \varepsilon_k \leqslant f(z) + \sum_{i=1}^{s} y_i^k\, g_i(z) = \left\|y^k\right\| \Big[f(z)/\|y^k\| + \sum_{i=1}^{s} u_i^k\, g_i(z)\Big]\,.$$

Выражение в квадратных скобках обозначим β_k. Ясно, что

$$\beta_k \to \beta^* = \sum_{i=1}^{s} u_i^* \, g_i(z) < 0 \,.$$

При больших k будет $\beta_k \leqslant \frac{1}{2}\,\beta^*$. Следовательно,

$$f(x^*) \leqslant \varepsilon_k + \left\|y^k\right\| \beta_k \leqslant \varepsilon_k + \tfrac{1}{2}\,\beta^* \left\|y^k\right\| . \tag{15}$$

По условию величина $f(x^*)$ конечна. В то же время правая часть неравенства (15) стремится к $-\infty$ при $k \to \infty$. Получили противоречие.

Установлено, что x^* — решение задачи (10). Неравенство (11) является простым следствием этого факта, выпуклости множества P и дифференцируемости целевой функции на P.

Теорема доказана. □

Отметим, что условия (11) и (12) не зависят от функции f.

4°. Приведём два характерных примера.

ПРИМЕР 1. Рассмотрим экстремальную задачу

$$\begin{gathered} f(x) := x_1^3 \to \min \,, \\ g_1(x) := -x_1^3 + x_2 \leqslant 0 \,, \\ g_2(x) := -x_1^3 - x_2 \leqslant 0 \,, \\ x \in \mathbb{R}^2 \,. \end{gathered} \tag{16}$$

В данном случае $P = \mathbb{R}^2$. Множество планов X этой задачи представлено на рисунке.

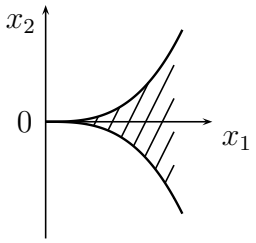


Рис.

Очевидно, что задача имеет единственное решение $x^* = \mathbb{O}$, и $f^* = 0$.

Запишем функцию Лагранжа

$$L(x, y) = x_1^3 + y_1 \,(-x_1^3 + x_2) + y_2 \,(-x_1^3 - x_2)$$

и целевую функцию двойственной задачи

$$\varphi(y) = \max \left\{ L(x, y) \mid x \in \mathbb{R}^2 \right\} .$$

При $y^* = \left(\frac{1}{2}, \frac{1}{2}\right)$ имеем $L(x, y^*) \equiv 0$, так что $\varphi(y^*) = 0$. Получаем

$$f^* = \varphi(y^*) \leqslant \varphi^* \leqslant f^* \,.$$

Отсюда следует, что y^* — решение двойственной задачи, и $\varphi^* = 0$. В частности, выполняется соотношение двойственности $f^* = \varphi^*$. Поскольку двойственная задача имеет решение, то выполняется также условие глобальной регулярности.

Решение $x^* = \mathbb{O}$ задачи (16) является внутренней точкой множества P. Теорема гарантирует справедливость равенств

$$f'(x^*) + y_1^*\, g_1'(x^*) + y_2^*\, g_2'(x^*) = \mathbb{O}\,,$$
$$y_1^*\, g_1(x^*) = 0\,, \quad y_2^*\, g_2(x^*) = 0\,.$$

Это можно проверить и непосредственно, если учесть, что $f'(x^*) = (0,0)$, $g_1'(x^*) = (0,1)$, $g_2'(x^*) = (0,-1)$, $y^* = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Отметим, что градиенты $g_1'(x^*)$ и $g_2'(x^*)$ линейно зависимы. Значит, условие локальной регулярности ограничений в точке x^* не выполняется. В таком случае классическая теорема Куна-Таккера не работает.

Соотношения (5), (6) и (11), (12) могут выполняться и тогда, когда $f^* > \varphi^*$.

ПРИМЕР 2. Рассмотрим экстремальную задачу

$$f(x) := x - 1 \to \min\,,$$
$$g(x) := x^2 - 1 = 0\,,$$
$$x \in P = R_+\,.$$

Ограничения можно записать в эквивалентном виде

$$g_1(x) := x^2 - 1 \leqslant 0\,,$$
$$g_2(x) := -x^2 + 1 \leqslant 0\,,$$
$$x \in P = R_+\,.$$

Задача имеет единственное решение $x^* = 1$, и $f^* = 0$.

Обратимся к функции Лагранжа

$$L(x,y) = x - 1 + y_1\,(x^2 - 1) + y_2\,(-x^2 + 1) =$$
$$= x^2\,(y_1 - y_2) + x + (y_2 - y_1 - 1).$$

Для целевой функции двойственной задачи получаем формулу

$$\varphi(y) = \min\left\{L(x,y) \mid x \geqslant 0\right\} = \begin{cases} -1 & \text{при} \quad y_1 - y_2 = 0, \\ -\infty & \text{при} \quad y_1 - y_2 < 0, \\ y_2 - y_1 - 1 & \text{при} \quad y_1 - y_2 > 0. \end{cases}$$

Здесь $y \in \mathbb{R}^2_+$. Очевидно, что $\varphi^* = -1$. Значит, $f^* > \varphi^*$. Вместе с тем,

$$f'(x^*) = 1\,, \quad g_1'(x^*) = 2\,, \quad g_2'(x^*) = -2$$

и

$$f'(x^*) + \tfrac{1}{2}\, g_2'(x^*) = 0\,, \quad g_1'(x^*) + g_2'(x^*) = 0\,.$$

Приходим к следующим выводам: при $y^* = \left(0, \frac{1}{2}\right)$ выполняются соотношения (5), (6), а при $u^* = (1, 1)$ — соотношения (11), (12).

ЛИТЕРАТУРА

1. Лазарев А. В. *О соотношении двойственности в математическом программировании* // Семинар «DHA & CAGD». Избранные доклады. 17 мая 2008 г. (`http://dha.spb.ru/reps08.shtml#0517`) [Данная книга, с. 233]

2. Гавурин М. К., Малозёмов В. Н. *Экстремальные задачи с линейными ограничениями*. Л.: Изд-во ЛГУ, 1984. 176 с.

ГЛОБАЛЬНАЯ РЕГУЛЯРНОСТЬ В МАТЕМАТИЧЕСКОМ ПРОГРАММИРОВАНИИ*

Манлио Гаудиозо, В. Н. Малозёмов

Делается попытка переосмыслить и дополнить содержание докладов [1, 2].

1°. Рассмотрим задачу математического программирования

$$\begin{gathered} f(x) \to \inf, \\ g_i(x) \leqslant 0, \quad i \in 1:s; \\ x \in P. \end{gathered} \tag{1}$$

Предполагается, что $P \subset \mathbb{R}^n$ — произвольное непустое множество (возможно, дискретное), $f, g_1, \ldots, g_s$ — произвольные конечные функции, заданные на P. Обозначим

$$X = \big\{x \in P \mid g_i(x) \leqslant 0,\ i \in 1:s\big\},$$

$$f^* = \inf\big\{f(x) \mid x \in X\big\}.$$

Введём функцию Лагранжа

$$L(x,y) = f(x) + \sum_{i=1}^{s} y_i\, g_i(x)$$

и запишем двойственную задачу

$$\varphi(y) := \inf\big\{L(x,y) \mid x \in P\big\} \to \sup_{y \in \mathbb{R}^s_+}. \tag{2}$$

Здесь $\mathbb{R}^s_+$ — множество векторов $y = (y_1, \ldots, y_s)$ с неотрицательными компонентами. Отметим, что функция Лагранжа $L(x,y)$ при фиксированном x аффинна по y, поэтому целевая функция двойственной задачи $\varphi(y)$ *вогнута* на $\mathbb{R}^s_+$. Обозначим

$$\varphi^* = \sup\big\{\varphi(y) \mid y \in \mathbb{R}^s_+\big\}.$$

*Семинар «DHA & CAGD». Избранные доклады. 28 октября 2008 г.

Наряду с (1), (2) рассмотрим вспомогательную параметрическую задачу

$$\begin{gathered} f(x) \to \inf, \\ g_i(x) \leqslant v_i, \quad i \in 1:s; \\ x \in P. \end{gathered} \tag{3}$$

Обозначим

$$X(v) = \bigl\{x \in P \mid g_i(x) \leqslant v_i,\ i \in 1:s\bigr\}.$$

Функция

$$F(v) = \inf\bigl\{f(x) \mid x \in X(v)\bigr\}, \quad v \in \mathbb{R}^s,$$

называется *функцией чувствительности* для задачи (1).

Исходная задача (1) вкладывается в задачу (3) (при $v = \mathbb{O}$). Функция $F(v)$ отражает характер этого вложения. Вместе с тем, функция чувствительности связана и с целевой функцией двойственной задачи (2). Справедливо следующее утверждение.

ЛЕММА. *При всех $y \in \mathbb{R}^s_+$ функция $\varphi(y)$ допускает представление*

$$\varphi(y) = \inf\bigl\{F(v) + \langle y, v\rangle \mid v \in \mathbb{R}^s\bigr\}.$$

При исследовании пары двойственных задач (1), (2) функция чувствительности играет важную роль.

2°. Относительно задачи (1) сделаем естественные предположения

$$X \neq \varnothing, \qquad f^* > -\infty, \tag{4}$$

которые гарантируют, что f^* — конечная величина.

Из определений следует, что $f^* \geqslant \varphi^*$. Первый вопрос, на который нужно ответить: когда выполняется *соотношение двойственности* $f^* = \varphi^*$?

ТЕОРЕМА 1. *Для того чтобы имело место соотношение двойственности, необходимо и достаточно, чтобы ε-субдифференциал функции чувствительности в нуле, $\partial_\varepsilon F(\mathbb{O})$, был непустым при всех $\varepsilon > 0$.*

3°. Условие

$$\partial F(\mathbb{O}) \neq \varnothing \tag{5}$$

назовём условием *глобальной регулярности* задачи (1). По существу, оно характеризует «регулярность» вложения задачи (1) в параметрическое семейство задач (3).

ТЕОРЕМА 2. *Условие глобальной регулярности выполняется тогда и только тогда, когда $f^* = \varphi^*$ и двойственная задача* (2) *имеет решение.*

По ходу доказательства этой теоремы выясняется, что

- любая точка из $-\partial F(\mathbb{O})$ является решением задачи (2);
- при выполнении соотношения двойственности $f^* = \varphi^*$ любое решение двойственной задачи принадлежит множеству $-\partial F(\mathbb{O})$.

4°. Приведём одно достаточное условие, гарантирующее глобальную регулярность задачи (1).

ТЕОРЕМА 3. *Пусть $f^* = \varphi^*$ и выполнено условие Слейтера: существует точка $z \in P$, такая, что $g_i(z) < 0$ при всех $i \in 1:s$. Тогда $\partial F(\mathbb{O}) \neq \varnothing$.*

5°. Предположим, что выполнено условие глобальной регулярности (5) задачи (1). Возьмём решение двойственной задачи y^* и рассмотрим задачу Лагранжа

$$L(x, y^*) \to \inf_{x \in P}. \tag{6}$$

Её экстремальное значение равно $\varphi(y^*) = \varphi^*$.

ТЕОРЕМА 4 (принцип Лагранжа)**.** *Любое решение x^* глобально регулярной задачи* (1) *является решением задачи* (6)*. При этом выполняется условие дополнительности*

$$y_i^*\, g_i(x^*) = 0\,, \quad i \in 1:s\,. \tag{7}$$

Приведём пример, когда задачи (1) и (6) имеют единственные, но разные решения.

ПРИМЕР. Рассмотрим экстремальную задачу

$$\begin{gathered} f(x) := x - 1 \to \inf\,, \\ g_1(x) := x^2 - 1 \leqslant 0\,, \\ g_2(x) := -x^2 + 1 \leqslant 0\,, \\ P = \mathbb{R}_+\,. \end{gathered}$$

Она имеет единственное решение $x^* = 1$; при этом $f^* = 0$.

Запишем функцию Лагранжа

$$\begin{aligned} L(x, y) &= x - 1 + y_1\,(x^2 - 1) + y_2\,(-x^2 + 1) = \\ &= x^2\,(y_1 - y_2) + x - (y_1 - y_2 + 1)\,. \end{aligned}$$

Для целевой функции двойственной задачи получаем формулу

$$\varphi(y) := \inf\big\{L(x, y) \mid x \geqslant 0\big\} = \begin{cases} -(y_1 - y_2 + 1) & \text{при } y_1 - y_2 \geqslant 0\,, \\ -\infty & \text{при } y_1 - y_2 < 0\,. \end{cases}$$

Здесь $y \in \mathbb{R}^2_+$. Ясно, что $\varphi^* = -1$.

Решением двойственной задачи является любой вектор $y^* \geqslant \mathbb{O}$ с $y_1^* = y_2^*$. Для таких векторов задача Лагранжа принимает вид

$$L(x, y^*) := x - 1 \to \inf_{x \geqslant 0}.$$

Её единственное решение $\hat{x} = 0$ отлично от решения $x^* = 1$ исходной задачи.

В данном случае не выполняется соотношение двойственности, поскольку $f^* = 0$, $\varphi^* = -1$.

6°. Обратимся к достаточным условиям глобальной оптимальности. Говорят [3, с. 144], что пара $\{x^*, y^*\}$, где $x^* \in P$, $y^* \in \mathbb{R}^s_+$, удовлетворяет условию *глобальной оптимальности*, если

$$(\alpha) \qquad L(x^*, y^*) = \min_{x \in P} L(x, y^*)\,;$$

$$(\beta) \qquad y_i^*\, g_i(x^*) = 0\,, \qquad i \in 1 : s\,;$$

$$(\gamma) \qquad g_i(x^*) \leqslant 0\,, \qquad i \in 1 : s\,.$$

ТЕОРЕМА 5. *Если $\{x^*, y^*\}$ — глобально оптимальная пара, то x^* — решение задачи* (1)*, y^* — решение задачи* (2) *и $f(x^*) = \varphi(y^*)$.*

Доказательство. Согласно (γ), $x^* \in X$. Для любого $x \in X$ в силу (β) и (α) имеем

$$f(x^*) = f(x^*) + \sum_{i=1}^{s} y_i^*\, g_i(x^*) = L(x^*, y^*) \leqslant L(x, y^*) =$$
$$= f(x) + \sum_{i=1}^{s} y_i^*\, g_i(x) \leqslant f(x)\,.$$

Это значит, что x^* — решение задачи (1), $f(x^*) = f^*$. Далее

$$\varphi^* \geqslant \varphi(y^*) = \inf\big\{L(x, y^*) \mid x \in P\big\} = L(x^*, y^*) =$$
$$= f(x^*) + \sum_{i=1}^{s} y_i^*\, g_i(x^*) = f(x^*) = f^* \geqslant \varphi^*.$$

Отсюда следует, что $\varphi(y^*) = \varphi^*$ и $f(x^*) = \varphi(y^*)$.

Теорема доказана. □

ТЕОРЕМА 6. *Глобально оптимальная пара $\{x^*, y^*\}$ является седловой точкой функции Лагранжа, то есть при всех $x \in P$ и $y \in \mathbb{R}^s_+$*

$$L(x^*, y) \leqslant L(x^*, y^*) \leqslant L(x, y^*)\,. \tag{8}$$

Доказательство. При всех $y \in \mathbb{R}^s_+$ имеем

$$L(x^*, y) = f(x^*) + \sum_{i=1}^{s} y_i\, g_i(x^*) \leqslant f(x^*)\,,$$

поэтому

$$\sup\left\{L(x^*, y) \mid y \in \mathbb{R}^s_+\right\} \leqslant f(x^*)\,.$$

Учитывая равенство $f(x^*) = \varphi(y^*)$, получаем

$$L(x^*, y^*) \leqslant \sup\left\{L(x^*, y) \mid y \in \mathbb{R}^s_+\right\} \leqslant f(x^*) =$$
$$= \varphi(y^*) = \inf\left\{L(x, y^*) \mid x \in P\right\} \leqslant L(x^*, y^*)\,.$$

Приходим к соотношениям

$$\sup\left\{L(x^*, y) \mid y \in \mathbb{R}^s_+\right\} = L(x^*, y^*) = \inf\left\{L(x, y^*) \mid x \in P\right\},$$

равносильным (8). Теорема доказана. □

7°. В качестве простейшего объекта применения общих результатов рассмотрим задачу линейного программирования

$$\begin{aligned} f(x) :=& \langle c, x\rangle \to \inf\,,\\ & A\,x \leqslant b\,,\\ & P = \mathbb{R}^n_+\,. \end{aligned} \tag{9}$$

Как задача линейного программирования она имеет двойственную

$$\begin{aligned} \psi(u) :=& \langle b, u\rangle \to \sup\,,\\ & u\,A \leqslant c\,, \quad u \leqslant \mathbb{O}\,. \end{aligned} \tag{10}$$

Множество планов задачи (10) обозначим U. В условиях (4) обе задачи (9) и (10) имеют решения и выполняется соотношение двойственности

$$f^* := \min_{x\in X} f(x) = \max_{u\in U} \psi(u) \;=:\; \psi^*.$$

Выясним, как связаны задачи (10) и (2). Имеем

$$\varphi(y) := \inf\left\{\langle c, x\rangle + \langle y, A\,x - b\rangle \mid x \in \mathbb{R}^n_+\right\} =$$
$$= -\langle b, y\rangle + \inf\left\{\langle c + y\,A, x\rangle \mid x \in \mathbb{R}^n_+\right\}.$$

Ясно, что

$$\inf\left\{\langle c + y\,A, x\rangle \mid x \in \mathbb{R}^n_+\right\} = \begin{cases} 0\,, & \text{если } c + y\,A \geqslant \mathbb{O},\\ -\infty\,, & \text{в противном случае.}\end{cases}$$

Поэтому задача (2) принимает вид

$$\begin{aligned} &\varphi(y) := -\langle b, y\rangle \to \sup\,, \\ &c + y\,A \geqslant \mathbb{O}\,, \quad y \geqslant \mathbb{O}\,. \end{aligned} \tag{11}$$

Задачи (10) и (11) эквивалентны. Множества их решений U^* и Y^* связаны соотношением $Y^* = -U^*$, и $\varphi^* = \psi^*$.

Получили, что $f^* = \psi^* = \varphi^*$ и двойственная задача (11) имеет решение. По теореме 2 в условиях (4) задача линейного программирования (9) глобально регулярна.

ТЕОРЕМА 7. *Множество решений задачи* (10) *совпадает с субдифференциалом функции чувствительности задачи* (9) *в нуле, то есть* $U^* = \partial F(\mathbb{O})$.

Доказательство. Запишем вспомогательную параметрическую задачу

$$\begin{gathered} f(x) := \langle c, x\rangle \to \inf\,, \\ A\,x \leqslant b + v\,, \\ P = \mathbb{R}^n_+\,. \end{gathered}$$

Возьмём $u^* \in U^*$ и покажем, что

$$F(v) - F(\mathbb{O}) \geqslant \langle u^*, v\rangle \quad \forall\, v \in \mathbb{R}^s. \tag{12}$$

При любом $x \in X(v)$ имеем

$$\langle c, x\rangle \geqslant \langle c, x\rangle + \langle u^*, b + v - A\,x\rangle = \langle b, u^*\rangle + \langle u^*, v\rangle +$$

$$+ \langle c - u^*A, x\rangle \geqslant \langle b, u^*\rangle + \langle u^*, v\rangle = f^* + \langle u^*, v\rangle\,.$$

Отсюда следует, что при всех $v \in \mathbb{R}^s$

$$F(v) \geqslant f^* + \langle u^*, v\rangle\,.$$

Учитывая равенство $f^* = F(\mathbb{O})$, приходим к (12).

Наоборот, пусть $u^* \in \partial F(\mathbb{O})$, так что выполняется неравенство (12). Подставляя в (12) на место v единичные орты e_i и принимая во внимание, что $F(e_i) \leqslant F(\mathbb{O})$, получаем $u^*_i \leqslant 0$. Значит, $u^* \leqslant \mathbb{O}$.

Далее, в силу (12)

$$F(v) + \langle -u^*, v\rangle \geqslant F(\mathbb{O}) \qquad \forall\, v \in \mathbb{R}^s.$$

По лемме из п. 1°, $\varphi(-u^*) \geqslant F(\mathbb{O})$. Соотношения

$$\varphi^* \geqslant \varphi(-u^*) \geqslant F(\mathbb{O}) = f^* \geqslant \varphi^*$$

убеждают нас в том, что вектор $-u^*$ является решением задачи (11). В этом случае, как отмечалось, u^* — решение задачи (10).

Теорема доказана. □

К данной теореме примыкает такой результат (см. [4]).

ТЕОРЕМА 8. *Если u — план двойственной задачи линейного программирования* (10)*, то $u \in \partial_\varepsilon F(\mathbb{O})$, где $\varepsilon = f^* - \langle b, u\rangle$.*

Доказательство. Нужно проверить, что

$$F(v) - F(\mathbb{O}) \geqslant \langle u, v\rangle - \varepsilon \qquad \forall\, v \in \mathbb{R}^s.$$

При $X(v) = \varnothing$ это очевидно. Пусть $X(v) \neq \varnothing$. Для любого $x \in X(v)$ имеем

$$\langle c, x\rangle \geqslant \langle c, x\rangle + \langle u, b + v - A\,x\rangle = \langle b, u\rangle + \langle u, v\rangle +$$
$$+ \langle c - uA, x\rangle \geqslant f^* - \varepsilon + \langle u, v\rangle = F(\mathbb{O}) + \langle u, v\rangle - \varepsilon\,.$$

Остаётся взять инфимум по $x \in X(v)$.

Теорема доказана. □

8°. Современное состояние теории чувствительности в оптимизации представлено в книге [5].

ЛИТЕРАТУРА

1. Лазарев А. В. *О соотношении двойственности в математическом программировании* // Семинар «DHA & CAGD». Избранные доклады. 17 мая 2008 г. (`http://dha.spb.ru/reps08.shtml#0517`) [Данная книга, с. 233]

2. Лазарев А. В. *Необходимые условия глобальной оптимальности* // Семинар «DHA & CAGD». Избранные доклады. 9 сентября 2008 г. (`http://dha.spb.ru/reps08.shtml#0909`) [Данная книга, с. 241]

3. Shapiro J. *Mathematical Programming: Structures and Algorithms.* J. Wiley & Sons, 1979.

4. De Leone R., Gaudioso M., Monaco M. F. *Nonsmooth optimization methods for parallel decomposition of multicommodity flow problems* // Annals of Operation Research. 1993. Vol. 44. P. 299–311.

5. Измаилов А. Ф. *Чувствительность в оптимизации.* М.: Физматлит, 2006.

О СЕДЛОВЫХ ТОЧКАХ ФУНКЦИИ ЛАГРАНЖА*

Н. И. Наумова

Данный доклад примыкает к докладу [1].

1°. Рассмотрим задачу математического программирования

$$\begin{gathered} f(x) \to \inf, \\ g_j(x) \leqslant 0, \quad i \in 1:s; \\ x \in P. \end{gathered} \tag{1}$$

Предположим, что $P \subset \mathbb{R}^n$ — произвольное непустое множество и f, $g_1, \ldots,$ g_s — произвольные конечные функции, заданные на P.

Введём функцию Лагранжа

$$L(x,u) = f(x) + \sum_{i=1}^{s} u_i\, g_i(x).$$

Говорят [2, с. 144], что пара $\{x^*, u^*\}$, где $x^* \in P$, $u^* \in \mathbb{R}^s_+$, удовлетворяет условию глобальной оптимальности, если

$$(\alpha) \qquad L(x^*,u^*) = \min_{x\in P} L(x,u^*);$$

$$(\beta) \qquad u_i^*\, g_i(x^*) = 0, \quad i \in 1:s;$$

$$(\gamma) \qquad g_i(x^*) \leqslant 0, \quad i \in 1:s.$$

ТЕОРЕМА 1. *Пара $\{x^*, u^*\}$ удовлетворяет условию глобальной оптимальности тогда и только тогда, когда она является седловой точкой функции Лагранжа, то есть*

$$L(x^*,u) \leqslant L(x^*,u^*) \leqslant L(x,u^*) \tag{2}$$

при всех $x \in P$ и $u \in \mathbb{R}^s_+$.

*Семинар «DHA & CAGD». Избранные доклады. 25 ноября 2008 г.

Доказательство. То, что глобально оптимальная пара является седловой точкой функции Лагранжа, установлено в [1]. Проверим обратное утверждение.

Пусть выполнены соотношения (2). Из первого неравенства следует (α). Проверим, что выполняется условие (γ).

Если $g_{i_0}(x^*) > 0$ при некотором $i_0 \in 1:s$, то, положив при $t > 0$

$$u_i(t) = \begin{cases} t & \text{при } i = i_0, \\ 0 & \text{при остальных } i \in 1:s, \end{cases}$$

получим

$$L\big(x^*, u(t)\big) = f(x^*) + t\, g_{i_0}(x^*).$$

Очевидно, что $L\big(x^*, u(t)\big) \to +\infty$ при $t \to +\infty$. Но это противоречит левому неравенству в (2). Итак, $g_i(x^*) \leqslant 0$ при всех $i \in 1:s$.

Далее, согласно (2),

$$f(x^*) = L(x^*, \mathbb{O}) \leqslant L(x^*, u^*),$$

поэтому

$$L(x^*, u^*) = f(x^*) + \sum_{i=1}^{s} u_i^*\, g_i(x^*) \leqslant f(x^*) \leqslant L(x^*, u^*).$$

Приходим к равенству

$$\sum_{i=1}^{s} u_i^*\, g_i(x^*) = 0,$$

равносильному условию (β) в силу неположительности всех слагаемых.

Теорема доказана. □

2°. Вернёмся к задаче (1) и обозначим

$$X = \big\{x \in P \mid g_i(x) \leqslant 0,\ i \in 1:s\big\},$$

$$f^* = \inf\big\{f(x) \mid x \in X\big\}.$$

Будем предполагать, что $X \neq \varnothing$ и $f^* > -\infty$. Это, в частности, гарантирует конечность f^*.

Запишем двойственную задачу

$$\varphi(u) := \inf\big\{L(x,u) \mid x \in P\big\} \to \sup_{u \in \mathbb{R}^s_+}.$$

Обозначим $\varphi^* = \sup\big\{\varphi(u) \mid u \in \mathbb{R}^s_+\big\}$.

В [3] получен критерий выполнения соотношения двойственности $f^* = \varphi^*$ в терминах ε-субдифференциала функции чувствительности. Другой вариант критерия, известный в теории игр, связан с ε-седловыми точками функции Лагранжа [4, с. 94–96].

Напомним, что при $\varepsilon > 0$ пара $\{x^\varepsilon, u^\varepsilon\}$, где $x^\varepsilon \in P$, $u^\varepsilon \in \mathbb{R}^s_+$, называется ε-седловой точкой функции Лагранжа, если

$$-\varepsilon + L(x^\varepsilon, u) \leqslant L(x^\varepsilon, u^\varepsilon) \leqslant L(x, u^\varepsilon) + \varepsilon \tag{3}$$

при всех $x \in P$ и $u \in \mathbb{R}^s_+$.

ТЕОРЕМА 2. *Для того, чтобы выполнялось соотношение двойственности $f^* = \varphi^*$, необходимо и достаточно, чтобы при всех $\varepsilon > 0$ у функции Лагранжа существовала ε-седловая точка.*

Доказательство. Необходимость. Покажем прежде всего, что

$$f^* = \inf_{x \in P} \sup_{u \in \mathbb{R}^s_+} L(x, u). \tag{4}$$

При $x \in X$ имеем

$$\sup_{u \in \mathbb{R}^s_+} L(x, u) = \sup_{u \in \mathbb{R}^s_+} \Big\{ f(x) + \sum_{i=1}^{s} u_i \, g_i(x) \Big\} = f(x). \tag{5}$$

При $x \in P \setminus X$

$$\sup_{u \in \mathbb{R}^s_+} L(x, u) = +\infty. \tag{6}$$

Действительно, если $x \in P \setminus X$, то $g_{i_0}(x) > 0$ при некотором $i_0 \in 1 : s$. Положив при $t > 0$

$$u_i(t) = \begin{cases} t & \text{при } i = i_0, \\ 0 & \text{при остальных } i \in 1 : s, \end{cases}$$

получим

$$\sup_{u \in \mathbb{R}^s_+} L(x, u) \geqslant \sup_{t > 0} L\big(x, u(t)\big) = \sup_{t > 0} \big\{ f(x) + t \, g_{i_0}(x) \big\} = +\infty.$$

Соотношение (6) установлено.

На основании (5) и (6) запишем

$$\inf_{x \in P} \sup_{u \in \mathbb{R}^s_+} L(x, u) = \min \Big\{ \inf_{x \in X} \sup_{u \in \mathbb{R}^s_+} L(x, u), \inf_{x \in P \setminus X} \sup_{u \in \mathbb{R}^s_+} L(x, u) \Big\} =$$
$$= \inf_{x \in X} \sup_{u \in \mathbb{R}^s_+} L(x, u) = \inf_{x \in X} f(x) = f^*.$$

Это соответствует (4).

Теперь соотношение двойственности принимает вид

$$\inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) = \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u). \tag{7}$$

Доказывая необходимость условий теоремы, мы считаем, что равенство (7) выполняется и что величина, стоящая в его левой части, конечна. Нужно проверить, что при всех $\varepsilon > 0$ у функции Лагранжа существует ε-седловая точка.

Зафиксируем $\varepsilon > 0$. По определению точной нижней и точной верхней границ по ε найдутся точки $x^\varepsilon \in P$ и $u^\varepsilon \in \mathbb{R}^s_+$, такие, что

$$\inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) + \frac{\varepsilon}{2} \geqslant \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u),$$
$$\sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u) - \frac{\varepsilon}{2} \leqslant \inf_{x\in P} L(x,u^\varepsilon).$$

С учётом (7) получаем

$$-\frac{\varepsilon}{2} + \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u) \leqslant \inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) =$$
$$= \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u) \leqslant \inf_{x\in P} L(x,u^\varepsilon) + \frac{\varepsilon}{2},$$

так что

$$-\varepsilon + \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u) \leqslant \inf_{x\in P} L(x,u^\varepsilon) \leqslant L(x^\varepsilon,u^\varepsilon). \tag{8}$$

Аналогично,

$$\frac{\varepsilon}{2} + \inf_{x\in P} L(x,u^\varepsilon) \geqslant \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u) =$$
$$= \inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) \geqslant \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u) - \frac{\varepsilon}{2},$$

так что

$$L(x^\varepsilon,u^\varepsilon) \leqslant \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u) \leqslant \inf_{x\in P} L(x,u^\varepsilon) + \varepsilon. \tag{9}$$

Из (8) и (9) очевидным образом следует (3). Значит, $(x^\varepsilon,u^\varepsilon)$ есть ε-седловая точка функции Лагранжа.

Достаточность. Согласно (3)

$$-\varepsilon + \inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) \leqslant -\varepsilon + \sup_{u\in\mathbb{R}^s_+} L(x^\varepsilon,u) \leqslant L(x^\varepsilon,u^\varepsilon) \leqslant$$

$$\leqslant \inf_{x\in P} L(x,u^\varepsilon)+\varepsilon \leqslant \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u)+\varepsilon.$$

Поэтому

$$\inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) \leqslant \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u)+2\varepsilon.$$

Ввиду произвольности $\varepsilon>0$ получаем неравенство

$$\inf_{x\in P}\sup_{u\in\mathbb{R}^s_+} L(x,u) \leqslant \sup_{u\in\mathbb{R}^s_+}\inf_{x\in P} L(x,u).$$

Обратное неравенство верно всегда. Значит, выполняется равенство (7), эквивалентное соотношению двойственности.

Теорема доказана. □

ЛИТЕРАТУРА

1. Гаудиозо М., Малозёмов В. Н. *Глобальная регулярность в математическом программировании* // Семинар «DHA & CAGD». Избранные доклады. 28 октября 2008 г. (`http://dha.spb.ru/reps08.shtml#1028`) [Данная книга, с. 248]

2. Shapiro J. *Mathematical Programming: Structures and Algorithms.* J. Wiley & Sons, 1979.

3. Лазарев А. В. *О соотношении двойственности в математическом программировании* // Семинар «DHA & CAGD». Избранные доклады. 17 мая 2008 г. (`http://dha.spb.ru/reps08.shtml#0517`) [Данная книга, с. 233]

4. Воробьёв Н. Н. *Теория игр для экономистов-кибернетиков.* М.: Наука, 1985.

СХОДИМОСТЬ МЕТОДА СОПРЯЖЁННЫХ ГРАДИЕНТОВ ДЛЯ ОБЩЕЙ ЗАДАЧИ БЕЗУСЛОВНОЙ МИНИМИЗАЦИИ*

А. В. Плоткин

Рассмотрим задачу безусловной оптимизации вида

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

где функция $f(x)$ непрерывно дифференцируема на $\mathbb{R}^n$. Градиент целевой функции будем обозначать через $g(x)$. Нашей целью является нахождение стационарной точки x_* функции $f(x)$, в которой $g(x_*) = \mathbb{O}$.

Одним из методов нахождения стационарной точки является метод сопряжённых градиентов [1, 2, 3]. В данной работе исследуется сходимость геометрического варианта метода, который был предложен в докладе [4]. Перейдём к описанию вычислительной схемы данного варианта метода сопряжённых градиентов.

Предварительный шаг. Возьмем произвольное начальное приближение $x_1 \in \mathbb{R}^n$ и вычислим значение градиента $g_1 = g(x_1)$. Если $g_1 = \mathbb{O}$, то x_1 — стационарная точка. В противном случае задаем первое направление спуска $d_1 = -g_1$.

Общий шаг. Пусть имеются x_k, d_k. Вычислим следующее приближение x_{k+1} по формуле

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

где шаг спуска α_k — приближенное решение задачи одномерной минимизации

$$f(x_k + \alpha d_k) \to \min_{\alpha > 0}. \tag{2}$$

Далее найдем значение градиента $g_{k+1} = g(x_{k+1})$. Если $g_{k+1} = \mathbb{O}$, то x_{k+1} — стационарная точка. В противном случае вычисляем следующее направление спуска d_{k+1} по правилу

$$\begin{aligned} d_{k+1} &= -g_{k+1} + \lambda_k (g_{k+1} + d_k), \\ \lambda_k &= \frac{||g_{k+1}||^2}{||g_{k+1}||^2 + ||d_k||^2}. \end{aligned} \tag{3}$$

*Семинар «CNSA & NDO». Избранные доклады. 28 апреля 2016 г.

На рис. 1 продемонстрировано применение правила (3) для получения нового направления.

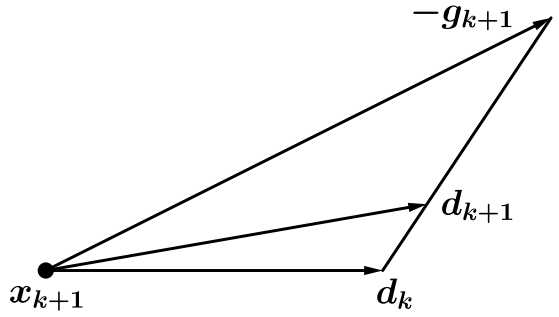


Рис. 1. Вычисление нового направления

Отметим, что, если $g_{k+1} \perp d_k$, то новое направление d_{k+1} будет являться перпендикуляром к гипотенузе треугольника, образованного векторами $-g_{k+1}$ и d_k.

Под сходимостью метода сопряжённых градиентов подразумевается выполнение условия

$$\lim_{k\to\infty} ||g_k|| = 0.$$

Исследовать сходимость геометрического варианта метода сопряжённых градиентов будем при следующих условиях на целевую функцию:

УСЛОВИЕ 1. *Множество уровня* $\mathcal{L} = \{x \in \mathbb{R}^n \mid f(x) \leqslant f(x_1)\}$, *где* x_1 — *начальное приближение, ограничено.*

УСЛОВИЕ 2. *В некоторой окрестности* $\mathcal{U}$ *множества* $\mathcal{L}$ *градиент* $g(x)$ *функции* $f(x)$ *удовлетворяет условию Липшица, то есть существует константа* $L > 0$, *такая что*

$$||g(x) - g(y)|| \leqslant L||x - y|| \qquad \forall x, y \in \mathcal{U}.$$

Теперь определим метод приближённого решения задачи (2). Введём обозначение $\varphi_k(\alpha) = f(x_k + \alpha d_k)$ и перепишем задачу (2) в упрощённом виде

$$\varphi_k(\alpha) \to \min_{\alpha>0}. \tag{4}$$

ОПРЕДЕЛЕНИЕ 1. Будем называть метод линейного поиска *точным*, если в качестве решения задачи (4) выбирается первая стационарная точка функции $\varphi_k(\alpha)$:

$$\alpha_k^* = \inf\{\alpha > 0 \mid \varphi_k'(\alpha) = 0\} \tag{5}$$

(см., например, рис. 2). Отметим, что

$$\varphi_k'(\alpha) = \langle g(x_k + \alpha d_k), d_k\rangle.$$

Данный метод труднореализуем на практике, однако часто используется при анализе сходимости различных вариантов метода сопряжённых градиентов.

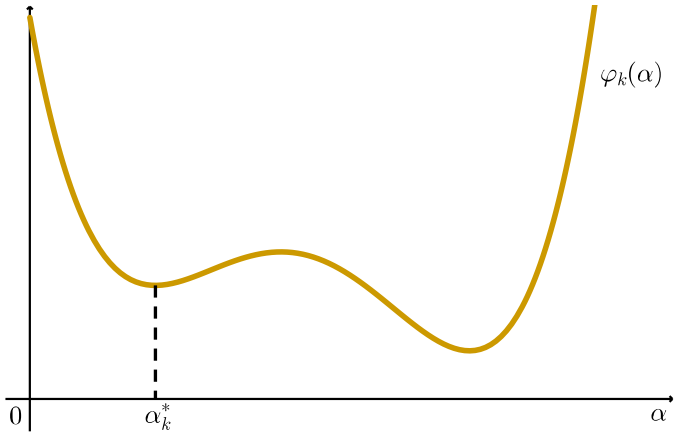


Рис. 2. Точный линейный поиск

Из определения точного линейного поиска имеем

$$g_{k+1} \perp d_k, \tag{6}$$

из чего следует, что условие

$$\varphi_k'(0) < 0 \tag{7}$$

выполняется на каждой итерации. Условие (7) называется *условием убывания* и совместно с условием 1 гарантирует как существование стационарной точки функции $\varphi_k(\alpha)$, так и положительность α_k^*.

Сформулируем теорему о сходиомости геометрического варианта метода сопряжённых градиентов с точным линейным поиском.

ТЕОРЕМА 1. *Пусть функция $f(x)$ удовлетворяет условиям 1, 2. Если в геометрическом варианте метода сопряженных градиентов шаг спуска выбирается с помощью точного линейного поиска* (5)*, то*

$$\varliminf_{k\to\infty} ||g_k|| = 0. \tag{8}$$

Для доказательства факта сходимости нам потребуется теорема Зойтендейка. Рассмотрим другой способ выбора шага спуска, который используется в теореме Зойтендейка. В качестве решения задачи (4) выбирается значение α_k, удовлетворяющее условиям Вулфа:

$$\varphi_k(\alpha_k) \leqslant \varphi_k(0) + \mu\alpha_k\varphi_k'(0), \tag{9}$$

$$\varphi_k'(\alpha_k) \geqslant \nu\varphi_k'(0), \tag{10}$$

где $0 < \mu < \nu < 1$ — фиксированные параметры (см., рис. 3).

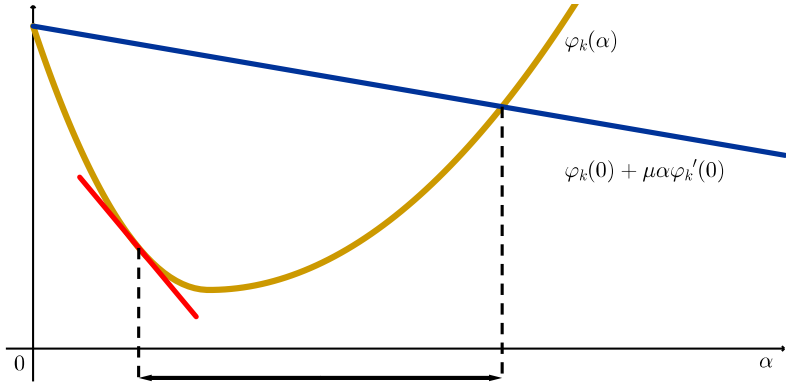


Рис. 3. Условия Вулфа

При выполнении условия убывания (7) и условия 1 такое α_k существует [5].

Введём обозначение Θ_k для угла между векторами $-g_k$ и d_k, так что

$$\cos\Theta_k = \frac{\langle -g_k, d_k\rangle}{||g_k||\cdot||d_k||}, \tag{11}$$

и сформулируем теорему Зойтендейка.

ТЕОРЕМА 2 (Зойтендейк)**.** *Пусть функция $f(x)$ удовлетворяет условиям 1, 2. Рассмотрим любой итеративный метод вида* (1)*. Если условие убывания* (7) *выполняется на каждой итерации, а значение α_k удовлетворяет условиям Вулфа* (9), (10)*, то существует константа $\lambda > 0$, такая что*

$$\varphi_k(0) - \varphi_k(\alpha_k) \geqslant \lambda\cos^2\Theta_k||g_k||^2 \qquad \forall k \geqslant 1. \tag{12}$$

Доказательство. Согласно (10)

$$\varphi_k'(\alpha_k) - \varphi_k'(0) \geqslant (\nu - 1)\varphi_k'(0).$$

В силу условия Липшица имеем

$$\varphi_k'(\alpha_k) - \varphi_k'(0) \leqslant L\alpha_k||d_k||^2.$$

Комбинируя эти неравенства, получаем

$$\alpha_k \geqslant \frac{\nu - 1}{L||d_k||^2}\varphi_k'(0) = \frac{1 - \nu}{L||d_k||^2}\langle -g_k, d_k\rangle,$$

что совместно с условием (9) приводит к неравенству

$$\varphi_k(0) - \varphi_k(\alpha_k) \geqslant \frac{\mu(1 - \nu)}{L||d_k||^2}\big(\langle -g_k, d_k\rangle\big)^2.$$

Используя определение угла Θ_k и функции $\varphi_k(\alpha)$, перепишем полученный результат в окончательном виде

$$\varphi_k(0) - \varphi_k(\alpha_k) \geqslant \lambda \cos^2 \Theta_k ||g_k||^2, \tag{13}$$

где $\lambda = \mu(1-\nu)/L > 0$. □

ЛЕММА 1. *Заключение теоремы Зойтендейка остаётся справедливым, если шаг спуска выбирается с помощью точного линейного поиска* (5).

Доказательство. Покажем, что неравенство (13) сохраняется для α_k^*. Рассмотрим два случая. Пусть α_k^* удовлетворяет условиям Вулфа, тогда неравенство (13) выполняется по теореме Зойтендейка. Если же это не так, то нарушаться может только первое условие Вулфа (9). Положим

$$h_k(\alpha) = \varphi_k(\alpha) - \big(\varphi_k(0) + \nu\alpha\varphi_k'(\alpha)\big).$$

Ясно, что $h_k(0) = 0$ и найдется точка $\alpha_k' > 0$, в которой $h_k(\alpha_k') = 0$, при этом $\alpha_k' < \alpha_k^*$. Обозначим через α_k'' точку минимума функции $h_k(\alpha)$ на отрезке $[0, \alpha_k']$. В ней $h_k'(\alpha_k'') = 0$ (рис. 4).

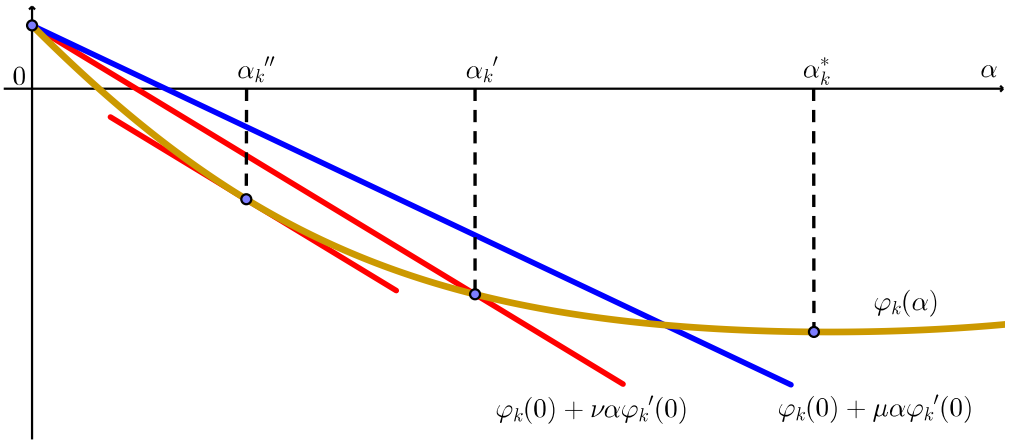


Рис. 4. Введение точек α_k' и α_k''

Имеем $\varphi_k'(\alpha_k'') = \nu\varphi_k'(0)$ и $\varphi_k(\alpha_k'') < \varphi_k(0) + \mu\alpha_k''\varphi_k'(0)$, то есть для α_k'' выполняются условия Вулфа (9), (10). Так как $\alpha_k'' < \alpha_k^*$, то $\varphi_k(\alpha_k'') > \varphi_k(\alpha_k^*)$, а значит,

$$\varphi_k(0) - \varphi_k(\alpha_k^*) > \varphi_k(0) - \varphi_k(\alpha_k'') \geqslant \lambda \cos^2 \Theta_k ||g_k||^2.$$

□

Вернёмся к доказательству теоремы сходимости.

Доказательство теоремы 1. Пусть соотношение (8) неверно. Тогда существует константа $\varepsilon > 0$, такая что

$$||g_k||^2 \geqslant \varepsilon \qquad \forall k \geqslant 1. \tag{14}$$

По лемме 1

$$\varphi_k(0) - \varphi_k(\alpha_k^*) \geqslant \lambda \cos^2 \Theta_k ||g_k||^2, \tag{15}$$

где $\lambda > 0$. Из формулы (3) и условия (6) следует, что (см. рис. 5)

$$\cos^2 \Theta_k ||g_k||^2 = ||d_k||^2.$$

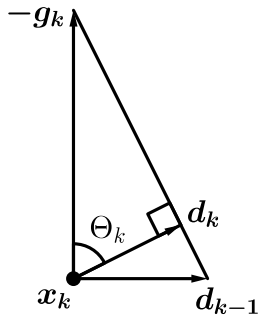


Рис. 5. Случай $-g_k \perp d_{k-1}$

Неравенство (15) принимает вид

$$\varphi_k(0) - \varphi_k(\alpha_k^*) \geqslant \lambda ||d_k||^2. \tag{16}$$

Докажем по индукции, что

$$||d_k||^2 \geqslant \frac{\varepsilon}{k} \qquad \forall k \geqslant 1. \tag{17}$$

При $k = 1$ неравенство верно. Сделаем индукционный переход от k к $k+1$. На основании формулы (3) и условия (6) имеем

$$||d_{k+1}||^2 = \frac{||g_{k+1}||^2 ||d_k||^2}{||g_{k+1}||^2 + ||d_k||^2} = \frac{1}{\frac{1}{||d_k||^2} + \frac{1}{||g_{k+1}||^2}}.$$

В силу индукционного предположения и неравенства (14)

$$\frac{1}{\frac{1}{||d_k||^2} + \frac{1}{||g_{k+1}||^2}} \geqslant \frac{1}{\frac{k}{\varepsilon} + \frac{1}{\varepsilon}} = \frac{\varepsilon}{k+1},$$

так что

$$||d_{k+1}||^2 \geqslant \frac{\varepsilon}{k+1}.$$

Неравенство (17) установлено. Объединив (16) и (17), получим

$$\varphi_k(0) - \varphi_k(\alpha_k^*) \geqslant \lambda\frac{\varepsilon}{k} \qquad \forall k \geqslant 1.$$

Теперь запишем

$$f(x_N) = f(x_1) - \sum_{k=1}^{N-1}[f(x_k) - f(x_{k+1})] =$$
$$= f(x_1) - \sum_{k=1}^{N-1}[\varphi_k(0) - \varphi_k(\alpha_k^*)] \leqslant f(x_1) - \lambda\varepsilon\sum_{k=1}^{N-1}\frac{1}{k}.$$

Отсюда следует, что

$$\lim_{N\to\infty} f(x_N) = -\infty,$$

а это противоречит условию 1. □

В дальнейшем планируется провести анализ сходимости геометрического варианта метода споряжённых градиентов при неточном линейном поиске.

ЛИТЕРАТУРА

1. W. W. Hager, H. Zhang. *A Survey of Nonlinear Conjugate Gradient Methods* // Pacific Journal of Optimization. 2006. Vol. 2. P. 335–358.

2. Y. H. Dai. *Nonlinear Conjugate Gradient Methods* // Wiley Encyclopedia of Operations Research and Management Science. 15.02.2011.

3. Малозёмов В. Н. *О методе сопряжённых градиентов* // Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г. (`http://dha.spb.ru/reps12.shtml#0428`) [Данная книга, с. 108]

4. Малозёмов В. Н. *Варианты метода сопряжённых градиентов* // Семинар «CNSA & NDO». Избранные доклады. 29 октября 2015 г. (`http://apmath.spbu.ru/cnsa/reps15.shtml#1029`) [Данная книга, с. 118]

5. Pytlak R. Conjugate Gradient Algorithms in Nonconvex Optimization. Berlin: Springer, 2009. P. 478.

ОПТИМАЛЬНЫЙ ГРАДИЕНТНЫЙ МЕТОД МИНИМИЗАЦИИ ВЫПУКЛЫХ ФУНКЦИЙ*

М. В. Долгополик

Аннотация. В докладе обсуждается в некотором смысле оптимальный градиентный метод минимизации гладких выпуклых функций, предложенный Ю.Е. Нестеровым [1, 2, 3]. В изложении данного метода мы следуем разделу 2.2.1 книги [3].

1°. Оценивающие последовательности. Рассмотрим задачу глобальной минимизации выпуклой функции f, определённой и непрерывно дифференцируемой на всём пространстве $\mathbb{R}^n$. Для простоты изложения мы будем предполагать, что градиент функции f удовлетворяет условию Липшица на $\mathbb{R}^n$ с константой $L > 0$, т.е.

$$\|f'(x) - f'(y)\| \leqslant L\|x - y\| \quad \forall x, y \in \mathbb{R}^n,$$

где $\|\cdot\|$ — евклидова норма. Также будем предполагать, что функция f ограничена снизу и достигает глобального минимума.

Для построения и анализа метода Нестерова мы воспользуемся техникой оценивающих последовательностей ([3], раздел 2.2.1).

ОПРЕДЕЛЕНИЕ 1. Пара последовательностей $(\{\varphi_k(x)\}, \{\lambda_k\})$, где $\varphi_k(x)$ — функция определенная на $\mathbb{R}^n$, а λ_k — неотрицательное число, называется *оценивающей последовательностью* для функции $f(x)$, если для любого вектора $x \in \mathbb{R}^n$ справедливо неравенство

$$\varphi_k(x) \leqslant (1 - \lambda_k)f(x) + \lambda_k\varphi_0(x) \quad \forall k \in \{0\} \cup \mathbb{N}.$$

Замечание 1. Здесь и далее мы предполагаем, что все последовательности нумеруются начиная с нуля, т.е. что индекс последовательности k принадлежит множеству $\{0\} \cup \mathbb{N}$.

В общем случае, оценивающие последовательности не несут почти никакой информации о функции f, т.к. в качестве функции $\varphi_k(x)$ можно выбрать

*Семинар «CNSA & NDO». Избранные доклады. 10 ноября 2016 г.

любую миноранту функции $(1-\lambda_k)f(x)+\lambda_k\varphi_0(x)$. Оценивающие последовательности оказываются полезным инструментом лишь в том случае, когда они определённым образом связаны с некоторой последовательностью точек $\{x_k\}$, построенной согласно какому-нибудь итерационному методу минимизации функции f.

Обозначим через x^* точку глобального минимума функции f, и положим $f^*=f(x^*)$.

ЛЕММА 1. *Пусть для некоторой последовательности $\{x_k\}$ справедливо неравенство*

$$f(x_k)\leqslant\varphi_k^*:=\min_{x\in\mathbb{R}^n}\varphi_k(x)\quad\forall k\in\{0\}\cup\mathbb{N}. \tag{1}$$

Тогда

$$f(x_k)-f^*\leqslant\lambda_k\big(\varphi_0(x^*)-f^*\big)\quad\forall k\in\{0\}\cup\mathbb{N}.$$

Доказательство. По определению последовательности $\{x_k\}$ будет

$$f(x_k)\leqslant\min_{x\in\mathbb{R}^n}\varphi_k(x)\quad\forall k\in\{0\}\cup\mathbb{N}.$$

Воспользовавшись определением оценивающей последовательности, получим, что

$$f(x_k)\leqslant\min_{x\in\mathbb{R}^n}\Big((1-\lambda_k)f(x)+\lambda_k\varphi_0(x)\Big)\quad\forall k\in\{0\}\cup\mathbb{N}$$

и, следовательно,

$$f(x_k)\leqslant(1-\lambda_k)f(x^*)+\lambda_k\varphi_0(x^*)\quad\forall k\in\{0\}\cup\mathbb{N}.$$

Перенеся $f(x^*)=f^*$ в левую часть, придём к неравенству

$$f(x_k)-f^*\leqslant\lambda_k\big(\varphi_0(x^*)-f^*\big),$$

что и требовалось доказать. □

Таким образом, если имеется некоторая последовательность точек $\{x_k\}$, и функции $\varphi_k(x)$ из оценивающей последовательности выбраны так, чтобы их глобальный минимум φ_k^* не превосходил значения функции f в очередной точке x_k, т.е. $\varphi_k^*\geqslant f(x_k)$, то можно сразу получить оценку

$$f(x_k)-f^*\leqslant\lambda_k\big(\varphi_0(x^*)-f^*\big)\quad\forall k\in\{0\}\cup\mathbb{N}.$$

Если, вдобавок, коэффициенты $\lambda_k\geqslant 0$ выбраны таким образом, что $\lambda_k\to 0$ при $k\to\infty$, то можно заключить, что последовательность $\{x_k\}$ является минимизирующей последовательностью для функции f. Кроме того, в этом случае можно легко получить оценку скорости сходимости последовательности $\{f(x_k)\}$ к f^* исходя из скорости сходимости последовательности $\{\lambda_k\}$ к нулю.

Для того чтобы воспользоваться описанным выше результатом, необходимо построить оценивающую последовательность для функции f и последовательность $\{x_k\}$, удовлетворяющую неравенству (1). Первым делом мы рассмотрим простую итеративную процедуру построения оценивающих последовательностей.

ЛЕММА 2. *Пусть $\varphi_0\colon \mathbb{R}^n \to \mathbb{R}$ — произвольная функция и $\lambda_0 = 1$. Пусть также $\{y_k\} \subset \mathbb{R}^n$ и $\{\alpha_k\} \subset (0,1)$ — произвольные последовательности. Тогда пара последовательностей $(\{\varphi_k(x)\}, \{\lambda_k\})$, определяемая рекуррентными соотношениями*

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \tag{2}$$

$$\varphi_{k+1}(x) = (1-\alpha_k)\varphi_k(x) + \alpha_k\Big(f(y_k) + \langle f'(y_k), x - y_k\rangle\Big), \tag{3}$$

является оценивающей последовательностью для функции f.

Доказательство. Для того чтобы доказать данную лемму, необходимо проверить, что для любого $x \in \mathbb{R}^n$ справедливо неравенство

$$\varphi_k(x) \leqslant (1-\lambda_k)f(x) + \lambda_k\varphi_0(x) \quad \forall k \in \{0\} \cup \mathbb{N}. \tag{4}$$

Воспользуемся методом математической индукции. Пусть сначала $k = 0$. Поскольку $\lambda_0 = 1$, то

$$\varphi_0(x) \leqslant (1-\lambda_0)f(x) + \lambda_0\varphi_0(x) \equiv \varphi_0(x) \quad \forall x \in \mathbb{R}^n.$$

Предположим теперь, что неравенство (4) выполняется для некоторого $k \in \in \{0\} \cup \mathbb{N}$. Покажем, что тогда это неравенство выполнено и для $k+1$. Действительно, по определению

$$\varphi_{k+1}(x) = (1-\alpha_k)\varphi_k(x) + \alpha_k\Big(f(y_k) + \langle f'(y_k), x - y_k\rangle\Big).$$

Поскольку функция f выпукла, то $f(x) - f(y_k) \geqslant \langle f'(y_k), x - y_k\rangle$. Следовательно,

$$\varphi_{k+1}(x) \leqslant (1-\alpha_k)\varphi_k(x) + \alpha_k f(x).$$

Заметим, что

$$\alpha_k = \big(1 - (1-\alpha_k)\lambda_k\big) - (1-\alpha_k)(1-\lambda_k).$$

Поэтому

$$\varphi_{k+1}(x) \leqslant \big(1 - (1-\alpha_k)\lambda_k\big)f(x) + (1-\alpha_k)\big(\varphi_k(x) - (1-\lambda_k)f(x)\big).$$

Так как по нашему предположению неравенство (4) выполнено для k, то $\varphi_k(x) - (1-\lambda_k)f(x) \leqslant \lambda_k\varphi_0(x)$, откуда

$$\varphi_{k+1}(x) \leqslant \big(1-(1-\alpha_k)\lambda_k\big)f(x) + (1-\alpha_k)\lambda_k\varphi_0(x).$$

Напомним, что по определению $\lambda_{k+1} = (1-\alpha_k)\lambda_k$. Поэтому

$$\varphi_{k+1}(x) \leqslant (1-\lambda_{k+1})f(x) + \lambda_{k+1}\varphi_0(x),$$

т.е. неравенство (4) выполнено для $k+1$. Следовательно, согласно методу математической индукции данное неравенство выполнено для всех k. □

Замечание 2. Пусть оценивающая последовательность $(\{\varphi_k(x)\},\{\lambda_k\})$ построена по описанной выше итеративной процедуре. Отметим, что простым достаточным условием, гарантирующим сходимость λ_k к 0 является условие $\sum\limits_{k=0}^{\infty}\alpha_k = +\infty$. Действительно, по определению

$$\lambda_k = \prod_{s=1}^{k}(1-\alpha_s).$$

Оценим величину λ_k сверху. Для этого воспользуемся неравенством

$$1 - t \leqslant e^{-t} \quad \forall t \geqslant 0,$$

справедливость которого следует из неотрицательности производной функции $h(t) = e^{-t} + t - 1$ для всех $t \geqslant 0$ и того факта, что $h(0) = 0$.

Имеем

$$\lambda_k = \prod_{s=1}^{k}(1-\alpha_s) \leqslant e^{-\sum_{s=1}^{k}\alpha_s}.$$

Отсюда, учитывая, что $\sum\limits_{k=0}^{\infty}\alpha_k = +\infty$, получаем $\lim\limits_{k\to\infty}\lambda_k = 0$, что и требовалось доказать.

Предыдущая лемма описывает простой итеративный способ построения оценивающей последовательности для функции f. Необходимо выбрать «начальное приближение» $\varphi_0(x)$, а также две последовательности $\{y_k\} \subset \mathbb{R}^n$ и $\{\alpha_k\} \subset (0,1)$. После этого оценивающая последовательность строится согласно рекуррентным соотношениям (2), (3). Отметим, что последовательности $\{y_k\}$ и $\{\alpha_k\}$ выступают в качестве параметров, которые необходимо выбирать таким образом, чтобы гарантировать справедливость неравенства (1).

Покажем, что если в качестве «начального приближения» $\varphi_0(x)$ выбрать простую квадратичную функцию, то последовательность $\{\varphi_k(x)\}$ можно вычислить в явном виде.

ЛЕММА 3. *Пусть*

$$\varphi_0(x) = \varphi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2,$$

и предположим, что последовательности $\{\lambda_k\}$ *и* $\{\varphi_k(x)\}$ *определяются рекуррентными соотношениями* (2), (3). *Тогда*

$$\varphi_k(x) \equiv \varphi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2 \quad \forall k \in \mathbb{N},$$

где последовательности $\{\gamma_k\}$, $\{v_k\}$ *и* $\{\varphi_k^*\}$ *определяются следующим образом:*

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k,$$
$$v_{k+1} = v_k - \frac{\alpha_k}{\gamma_{k+1}} f'(y_k),$$
$$\varphi_{k+1}^* = (1 - \alpha_k)\varphi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 + \alpha_k\langle f'(y_k), v_k - y_k\rangle.$$

Доказательство. Согласно рекуррентному соотношению (3) функция $\varphi_k(x)$ имеет вид

$$\varphi_{k+1}(x) = (1 - \alpha_k)\varphi_k(x) + \alpha_k\Big(f(y_k) + \langle f'(y_k), x - y_k\rangle\Big).$$

Поскольку функция $\varphi_0(x)$ является квадратичной, а каждая последующая функция $\varphi_{k+1}(x)$ получается из предыдущей $\varphi_k(x)$ домножением на константу $(1 - \alpha_k)$ и прибавлением линейной функции, то все функции $\varphi_k(x)$ являются квадратичными и имеют вид

$$\varphi_k(x) = \varphi_k^* + \langle x - v_k, A_k(x - v_k)\rangle, \tag{5}$$

для некоторых матриц A_k, векторов v_k и чисел φ_k^*. Вычислим данные величины в явном виде.

Так как $\varphi_0''(x) \equiv \gamma_0 E_n$, где E_n — единичная матрица размерности n, и

$$\varphi_{k+1}''(x) \equiv (1 - \alpha_k)\varphi_k''(x),$$

то

$$A_{k+1} = \gamma_{k+1}E_n, \quad \gamma_{k+1} = (1 - \alpha_k)\gamma_k.$$

Поэтому

$$\varphi_k(x) = \varphi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2$$

и

$$\varphi_{k+1}(x) = (1 - \alpha_k)\left(\varphi_k^* + \frac{\gamma_k}{2}\|x - v_k\|^2\right) + \alpha_k\Big(f(y_k) + \langle f'(y_k), x - y_k\rangle\Big).$$

Заметим, что вектор v_k из (5) является точкой глобального минимума функции $\varphi_k(x)$. Поэтому справедливо равенство

$$0 = \varphi'_{k+1}(v_{k+1}) = (1-\alpha_k)\gamma_k(v_{k+1}-v_k) + \alpha_k f'(y_k) = 0,$$

откуда

$$v_{k+1} = v_k - \frac{\alpha_k}{\gamma_{k+1}} f'(y_k). \tag{6}$$

Теперь вычислим φ^*_{k+1}. Из рекуррентного соотношения (3) следует, что

$$\varphi^*_{k+1} + \frac{\gamma_{k+1}}{2}\|y_k - v_{k+1}\|^2 = \varphi_{k+1}(y_k) = (1-\alpha_k)\left(\varphi^*_k + \frac{\gamma_k}{2}\|y_k - v_k\|^2\right) + \alpha_k f(y_k). \tag{7}$$

В силу (6) имеем

$$v_{k+1} - y_k = v_k - y_k - \frac{\alpha_k}{\gamma_{k+1}} f'(y_k).$$

Поэтому

$$\frac{\gamma_{k+1}}{2}\|v_{k+1} - y_k\|^2 = \frac{\gamma_{k+1}}{2}\|v_k - y_k\|^2 - \alpha_k\langle f'(y_k), v_k - y_k\rangle + \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2.$$

Подставляя данное выражение в (7), получаем

$$\varphi^*_{k+1} = (1-\alpha_k)\varphi^*_k + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 + \alpha_k\langle f'(y_k), v_k - y_k\rangle,$$

что и требовалось доказать. □

2°. Метод Нестерова. Перейдём к построению численного метода минимизации функции f. Пусть задано начальное приближение x_0. Мы будем одновременно строить последующие приближения x_k и подбирать параметры y_k и α_k оценивающей последовательности $(\{\varphi_k\}, \{\lambda_k\})$, определяемой соотношениями (2), (3), с целью гарантировать справедливость неравенства

$$f(x_k) \leqslant \varphi^*_k := \min_{x\in\mathbb{R}^n} \varphi_k(x) \quad \forall k \in \{0\} \cup \mathbb{N}.$$

Если при этом удастся показать, что последовательность $\{\lambda_k\}$ стремится к нулю, то, воспользовавшись леммой 1, сразу получим, что последовательность $\{x_k\}$ является минимизирующей последовательностью для функции f. Кроме того, оценив скорость сходимости последовательности $\{\lambda_k\}$ к нулю, мы получим оценку скорости сходимости построенного метода.

Поскольку при построении численного метода нам потребуется оценивать снизу величину φ^*_k, что делать проще зная её явный вид, то, учитывая лемму 3, положим

$$\varphi_0(x) = \varphi^*_0 + \frac{\gamma_0}{2}\|x - v_0\|^2.$$

Пусть очередное приближение x_k, удовлетворяющее неравенству

$$\varphi_k^* \geqslant f(x_k) \tag{8}$$

уже построено. Покажем как определить следующее приближение x_{k+1}.

По лемме 3 имеем

$$\varphi_{k+1}^* = (1-\alpha_k)\varphi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 + \alpha_k\langle f'(y_k), v_k - y_k\rangle.$$

Учитывая (8), получаем

$$\varphi_{k+1}^* \geqslant (1-\alpha_k)f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 + \alpha_k\langle f'(y_k), v_k - y_k\rangle.$$

Так как функция f выпукла, то $f(x_k) - f(y_k) \geqslant \langle f'(y_k), x_k - y_k\rangle$. Поэтому

$$\varphi_{k+1}^* \geqslant f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2 + (1-\alpha_k)\Big\langle f'(y_k), \frac{\alpha_k}{1-\alpha_k}(v_k - y_k) + x_k - y_k\Big\rangle. \tag{9}$$

Разберёмся сначала с первыми двумя слагаемыми:

$$f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|f'(y_k)\|^2.$$

Напомним, что необходимо выбрать x_{k+1} так, чтобы $\varphi_{k+1}^* \geqslant f(x_{k+1})$. Выберем в качестве x_{k+1} любую точку удовлетворяющую неравенству

$$f(y_k) - \frac{1}{2L}\|f'(y_k)\|^2 \geqslant f(x_{k+1}). \tag{10}$$

В частности, можно положить $x_{k+1} = y_k - \dfrac{1}{L}f'(y_k)$.

Существование по крайней мере одной точки x_{k+1}, удовлетворяющей неравенству (10), следует из хорошо известного неравенства

$$f(y) - f(x) - \langle f'(x), y - x\rangle \leqslant \frac{L}{2}\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n \tag{11}$$

(см., например, [3], Теорема 2.1.5). Для того чтобы из данного неравенства получить неравенство (10) достаточно положить $x = y_k$ и $y = x_{k+1} = y_k - \dfrac{1}{L}f'(y_k)$.

Определим α_k из уравнения

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k.$$

Тогда

$$\frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2L}.$$

Подставляя данное выражение в (9) и используя определение точки x_{k+1} (см. (10)), приходим к неравенству

$$\varphi_{k+1}^* \geqslant f(x_{k+1}) + (1-\alpha_k)\Big\langle f'(y_k), \frac{\alpha_k}{1-\alpha_k}(v_k - y_k) + x_k - y_k\Big\rangle.$$

Теперь естественно определить y_k из уравнения

$$\frac{\alpha_k}{1-\alpha_k}(v_k - y_k) + x_k - y_k = 0,$$

т.е.

$$y_k = \alpha_k v_k + (1-\alpha_k)x_k.$$

Данный выбор y_k гарантирует выполнение неравенства $\varphi_{k+1}^* \geqslant f(x_{k+1})$ для всех $k \in \{0\} \cup \mathbb{N}$.

Таким образом, мы приходим к следующей теоретической схеме метода минимизации функции f, который принято называть оптимальным градиентным методом или методом Нестерова.

1) Выберем $x_0 \in \mathbb{R}^n$ и $\gamma_0 > 0$. Положим $v_0 = x_0$.

2) Переход от k-го приближения к $k+1$-ому осуществляется следующим образом:

   (a) Найдём $\alpha_k \in (0,1)$ из уравнения

   $$L\alpha_k^2 = (1-\alpha_k)\gamma_k.$$

   Положим $\gamma_{k+1} = (1-\alpha_k)\gamma_k$.

   (b) Выберем

   $$y_k = \alpha_k v_k + (1-\alpha_k)x_k$$

   и вычислим $f(y_k)$ и $f'(y_k)$.

   (c) Найдём x_{k+1} такое, что

   $$f(x_{k+1}) \leqslant f(y_k) - \frac{1}{2L}\|f'(y_k)\|^2.$$

   (d) Положим

   $$v_{k+1} = v_k - \frac{\alpha_k}{\gamma_{k+1}} f'(y_k).$$

Покажем сначала, что для любого $k \in \{0\} \cup \mathbb{N}$ действительно найдётся решение квадратного уравнения

$$L\alpha_k^2 = (1 - \alpha_k)\gamma_k,$$

принадлежащее интервалу $(0, 1)$. Для этого воспользуемся методом математической индукции. При $k = 0$ искомое решение существует, так как непрерывная функция $h_0(t) = Lt^2 - (1 - t)\gamma_0$ принимает на концах отрезка $[0, 1]$ значения разных знаков ($h_0(0) = -\gamma_0 < 0$ и $h_0(1) = L > 0$). Предположим теперь, что требуемое решение α_k существует для некоторого $k \in \{0\} \cup \mathbb{N}$. По определению

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k = L\alpha_k^2,$$

то есть $\gamma_{k+1} > 0$. Теперь, воспользовавшись тем, что функция $h_{k+1}(t) = Lt^2 - (1-t)\gamma_{k+1}$ принимает на концах отрезка $[0, 1]$ значения разных знаков, получим существование решения $\alpha_{k+1} \in (0, 1)$, что и требовалось показать. Таким образом, все параметры α_k корректно определены.

Перейдём теперь к анализу сходимости построенного метода.

ТЕОРЕМА 1. *Пусть последовательность $\{x_k\}$ построена по методу Нестерова. Тогда*

$$f(x_k) - f^* \leqslant \frac{2L(L + \gamma_0)}{(\sqrt{\gamma_0}k + 2\sqrt{L})^2}\|x_0 - x^*\|^2 \quad \forall k \in \{0\} \cup \mathbb{N}.$$

В частности, если $\gamma_0 = L$, то

$$f(x_k) - f^* \leqslant \frac{4L}{(k + 2)^2}\|x_0 - x^*\|^2 \quad \forall k \in \{0\} \cup \mathbb{N}.$$

Доказательство. Напомним, что функция $\varphi_0(x)$ была выбрана в виде

$$\varphi_0(x) = \varphi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2.$$

Положим $\varphi_0^* = f(x_0)$. Тогда по построению $\varphi_k^* \geqslant f(x_k)$ для всех $k \in \{0\} \cup \mathbb{N}$. Следовательно, по лемме 1 будет

$$f(x_k) - f^* \leqslant \lambda_k \left(f(x_0) - f^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2\right) \quad \forall k \in \{0\} \cup \mathbb{N}.$$

Воспользовавшись неравенством $f(x_0) - f^* \leqslant L\|x_0 - x^*\|^2/2$, которое следует из неравенства (11) при $y = x_0$ и $x = x^*$, получим

$$f(x_k) - f^* \leqslant \frac{\lambda_k}{2}(L + \gamma_0)\|x_0 - x^*\|^2. \tag{12}$$

Оценим коэффициенты λ_k. По определению $\gamma_{k+1} = (1-\alpha_k)\gamma_k$ и $\lambda_{k+1} = (1-\alpha_k)\lambda_k$ (см. (2)). Так как $\lambda_0 = 1$, то получаем, что $\gamma_k = \gamma_0\lambda_k$. Поэтому

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k = \gamma_{k+1} = \gamma_0\lambda_{k+1},$$

то есть

$$\lambda_{k+1} = \frac{L}{\gamma_0}\alpha_k^2. \tag{13}$$

Для любого $k \in \{0\} \cup \mathbb{N}$ имеем

$$\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} = \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k\lambda_{k+1}}} = \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k\lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})}.$$

Поскольку $\lambda_k = \prod_{s=1}^{k}(1-\alpha_k)$ и $\alpha_k \in (0,1)$, то $\{\lambda_k\}$ — невозрастающая последовательность. Следовательно,

$$\sqrt{\lambda_k\lambda_{k+1}}\big(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}}\big) = \lambda_k\frac{\sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k}}\big(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}}\big) =$$
$$= \lambda_k\left(\sqrt{\lambda_{k+1}} + \frac{\sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k}}\sqrt{\lambda_{k+1}}\right) \leqslant 2\lambda_k\sqrt{\lambda_{k+1}}.$$

Значит

$$\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} \geqslant \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k\sqrt{\lambda_{k+1}}} = \frac{\lambda_k - (1-\alpha_k)\lambda_k}{2\lambda_k\sqrt{\lambda_{k+1}}} = \frac{\alpha_k}{2\sqrt{\lambda_{k+1}}}.$$

Отсюда и из (13) получим

$$\frac{1}{\sqrt{\lambda_{k+1}}} - \frac{1}{\sqrt{\lambda_k}} \geqslant \frac{1}{2}\sqrt{\frac{\gamma_0}{L}}.$$

Поэтому, учитывая что $\lambda_0 = 1$, приходим к неравенству

$$\frac{1}{\sqrt{\lambda_k}} \geqslant 1 + \frac{k}{2}\sqrt{\frac{\gamma_0}{L}}$$

или, что эквивалентно,

$$\lambda_k \leqslant \frac{4L}{(\sqrt{\gamma_0}k + 2\sqrt{L})^2}.$$

Подставляя данное неравенство в (12), окончательно получаем

$$f(x_k) - f^* \leqslant \frac{2L(L+\gamma_0)}{(\sqrt{\gamma_0}k + 2\sqrt{L})^2}\|x_0 - x^*\|^2.$$

Теорема доказана. □

З а м е ч а н и е 3. Отметим, что для традиционных методов оптимизации, таких как метод градиентного спуска и различные варианты метода сопряжённых градиентов, справедлива следующая оценка скорости сходимости:

$$f(x_k) - f^* = O\left(\frac{1}{k}\right)$$

(см., например, [4, 5]). С другой стороны, согласно предыдущей теореме, для метода Нестерова справедлива лучшая оценка

$$f(x_k) - f^* \leqslant O\left(\frac{1}{k^2}\right).$$

Можно показать, что данная оценка в некотором смысле оптимальна (см. главу 2 книги [3]), что и объясняет называние «оптимальный градиентный метод». Подробнее по поводу оптимальности численных методов оптимизации см. книгу [6].

ЛИТЕРАТУРА

1. Нестеров Ю.Е. *Метод решения задачи выпуклого программирования со скоростью сходимости* $O(1/k^2)$ // Докл. АН СССР, 1983, т. 269, № 3, с. 543–548.

2. Нестров Ю.Е. *Об одном классе методов безусловной минимизации выпуклой функции, обладающих высокой скоростью сходимости* // Ж. вычисл. матем. и матем. физ, 1984, т. 24, № 7, с. 1090–1093.

3. Nesterov Y. *Introductory Lectures on Convex Optimization. A Basic Course.* Dordrecht: Kluwer Academic Publishers, 2004. 236 p.

4. Поляк Б.Т. *Градиентные методы минимизации функционалов* // Ж. вычисл. матем. и матем. физ., 1963, том 3, № 4, с. 643–653.

5. Любич Ю.И., Майстровский Г.Д. *Общая теория релаксационных процессов для выпуклых функционалов* // УМН, 1970, том 25, вып. 1(151), с. 57–112.

6. Немировский А.С., Юдин Д.Б. *Сложность задач и эффективность методов оптимизации.* М.: Наука, 1979. 384 с.

МЕТОД ЗАРЯЖЕННЫХ ШАРИКОВ*

М. Э. Аббасов

Аннотация. Механические аналогии в ряде случаев дают возможность строить эффективные алгоритмы для решения задач математического программирования. Хорошо известен метод тяжелого шарика, позволяющий решать задачи безусловной оптимизации. Бесспорным преимуществом таких методов является наглядность, идейная прозрачность, а также уверенность в их сходимости, проистекающая из законов механики. В настоящей работе рассматривается задача поиска минимального расстояния между точкой и гладким выпуклым замкнутым множеством, а также задача поиска минимального расстояния между двумя такими множествами. Для их решения предлагается новый алгоритм [1], базирующийся на механических принципах. Нужно отметить, что рассматриваемые в статье задачи имеют многочисленные практические приложения и возникают, в частности, во многих разделах математики. Например, в негладком анализе вектор, на котором достигается минимальное расстояние от начала координат до субдифференциала или экзостера, определяет направление наискорейшего спуска [2, 3, 4]. Поэтому не удивительно, что решение данной проблемы привлекает внимание многих исследователей [5, 6, 7, 8, 9, 10].

1°. Вспомогательные сведения и постановка задачи. Идея перехода от исходной оптимизационной задачи к некоторой механической системе, стремящейся с течением времени к равновесному положению, совпадающему с решением исходной задачи, позволяет строить новые эффективные итерационные алгоритмы. Для этого вначале составляют дифференциальные уравнения движения, а затем переходят к разностной схеме их решения. Такой подход не нов и подробно рассмотрен в [11], где класс получаемых таким образом методов называется методами установления. Очевидно, речь идет об установлении равновесия в нестационарной механической системе, которой заменяется исходная стационарная задача. Одним из наиболее известных представителей этого класса является метод тяжелого шарика. В нем для решения задачи поиска минимума функции $f(x)$ предлагается в присутствии сил тяжести и сопротивления поместить на поверхность $y = f(x)$ тяжелый шарик, который

*Семинар «CNSA & NDO». Избранные доклады. 21 мая 2015 г.

может двигаться только по данной поверхности (см. рис. 1). Ясно, что шарик в конечном итоге займет положение, соответствующее (локальному) минимуму потенциала $f(x)$, то есть остановится в решении исходной задачи.

Рис. 1. Метод тяжелого шарика

Уравнение движения

$$\frac{d^2x}{dt^2} + \mu\frac{dx}{dt} + \frac{\nabla f(x)}{1 + \|\nabla f(x)\|^2} = 0,$$

в окрестности равновесия (то есть точки x_*, в которой $\nabla f(x_*) = 0$) мало отличается от уравнения

$$\frac{d^2x}{dt^2} + \mu\frac{dx}{dt} + \nabla f(x) = 0.$$

Откуда, переходя к разностной схеме, приходим к алгоритму с двумя параметрами α, β

$$x_{n+1} = x_n + \alpha(x_n - x_{n-1}) + \beta\nabla f(x_n),$$

сходящемуся с линейной скоростью. За счет выбора параметров можно добиться получения наилучшего (в смысле количества итераций) линейного процесса.

Схема большинства известных методов установления описывается уравнениями вида

$$A_0\left(x, \frac{dx}{dt}\right)\frac{dx}{dt} + A_1\left(x, \frac{dx}{dt}, \nabla f(x)\right) = 0$$

или

$$B_0\left(x, \frac{dx}{dt}\right)\frac{d^2x}{dt^2} + B_1\left(x, \frac{dx}{dt}, \nabla f(x)\right) = 0,$$

где

$$A_0(x_*, 0) \neq 0,\ A_1(x_*, 0, 0) = 0,$$
$$B_0(x_*, 0) \neq 0,\ B_1(x_*, 0, 0) = 0$$

и выполнены условия диссипативности, обеспечивающие сходимость к точке экстремума x_*. Отметим, что в [11] подробно рассмотрен процесс установления, обеспечивающий линейную скорость сходимости, и показано, что за счет выбора параметров метода можно добиться оптимальности процесса.

Изложение будет вестись для n-мерного евклидова пространства. Пусть $\widetilde{x}$ произвольная точка из $\mathbb{R}^n$. Введем множество $X = \{x \in \mathbb{R}^n |\ f(x) \leq 0\}$, где $f : \mathbb{R}^n \to \mathbb{R}$ выпуклая, непрерывно дифференцируемая функция, и рассмотрим экстремальную задачу

$$\begin{cases} \|x - \widetilde{x}\| \longrightarrow \inf \\ x \in X \end{cases} \tag{1}$$

Очевидно, что X — замкнутое, выпуклое множество. Выбирая произвольное $x \in X$, можем гарантировать, что минимум достигается на множестве $X \bigcap B_{\|x\|}(\widetilde{x})$, где $B_{\|x\|}(\widetilde{x})$ замкнутый шар радиуса $\|x\|$ с центром в точке $\widetilde{x}$. Это множество замкнуто и выпукло как пересечение двух замкнутых выпуклых множеств. Кроме того, оно и ограничено, как пересечение двух множеств, одно из которых ограничено. Поэтому исходная задача имеет решение, т. е. существует $x_* \in X$, на котором искомый инфимум достигается. Кроме того, без уменьшения общности можно считать $\widetilde{x} = 0_n$. С учетом сказанного задачу (1) можно переписать в виде

$$\begin{cases} \|x\| \longrightarrow \min \\ x \in X \end{cases} \tag{2}$$

Ее мы и будем рассматривать в дальнейшем.

2°. Идейная основа алгоритма и его реализация. Для решения задачи мысленно поместим в начало координат отрицательный заряд q и зафиксируем его. Выберем в множестве X произвольную точку $\widehat{x}$, поместим в неё положительно заряженный шарик массой m и зарядом q. Силой тяжести будем пренебрегать. Шарик начнет двигаться в направлении начала координат по прямой, соединяющей шарики, до тех пор пока не столкнется с границей множества X. Координату точки столкновения можно найти, решив систему

$$\begin{cases} f(x) = 0 \\ x = \lambda\widehat{x},\ \lambda \in (-\infty, \infty), \end{cases} \tag{3}$$

и выбрав из двух решений то, которое находится ближе к началу координат. В результате получим $x_0 \in X$ (см. рис. 2).

Далее в отсутствии силы трения шарик двигается по внутренней стороне поверхности, ограничивающей X, достигая положения равновесия, которое в данном случае совпадает с решением задачи (2), и начинает колебаться около этого положения. Действительно, именно в этой точке сила нормальной реакции поверхности уравновешивается силой Кулона. В любой же другой точке касательная составляющая силы Кулона отлична от нуля и направлена к положению равновесия. Для того чтобы колебания стали затухающими и процесс сошелся к искомому решению, можно ввести силу сопротивления,

пропорциональную скорости и направленную противоположно ей. Выписав дифференциальные уравнения движения и перейдя к разностной схеме их решения, получим итерационный алгоритм решения задачи (2).

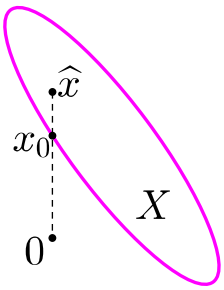


Рис. 2. Нахождение начальной точки x_0

2.1. Уравнения движения. Обозначим координаты шарика $x(t)$, где t время. Известно, что кулоновская сила равна

$$F(t) = -\frac{c_1 q^2}{\|x\|^3} x,$$

где c_1 — электрическая постоянная. Сила нормальной реакции перпендикулярна поверхности в точке x, направлена внутрь множества и по величине равна нормальной составляющей кулоновской силы,

$$N(t) = -\frac{\nabla f(x)}{\|\nabla f(x)\|^2} \langle F(t), \nabla f(x) \rangle.$$

Силу сопротивления определим формулой

$$R(t) = -c_2 \dot{x},$$

где c_2 — коэффициент сопротивления.

По второму закону Ньютона

$$m\ddot{x}(t) = F(t) + N(t) + R(t). \tag{4}$$

2.2. Переход к разностной схеме. Перейдём от (4) к системе первого порядка с помощью введения n-мерного вектора фиктивных переменных $z(t)$:

$$\begin{cases} \dot{x} = z \\ \dot{z} = \psi(x, z) \end{cases} \tag{5}$$

где

$$\psi(x,z) = -\frac{p_1}{\|x\|^3}x + \frac{p_1\langle x, \nabla f(x)\rangle}{\|x\|^3\|\nabla f(x)\|^2}\nabla f(x) - p_2 z$$

и p_1, p_2 — параметры, зависящие от m, q, c_1, c_2. Полученную систему (5) будем решать методом ломанных Эйлера. Выбираем $x(0) = x_0$, $z(0) = 0_n$ и некоторое малое положительное δ — длину шага. Пусть у нас есть x_{k-1}, z_{k-1}, тогда

$$\begin{cases} x_k = x_{k-1} + \delta z_{k-1} \\ z_k = z_{k-1} + \delta\psi(x_{k-1}, z_{k-1}) \end{cases} \tag{6}$$

Отметим, что схема Эйлера решения системы (5) эквивалентна применению градиентного спуска для функции, чей градиент совпадает с правой частью этой системы. Поэтому решение интересующей нас задачи (2) есть минимум такой функции. Но если градиент указанной функции удовлетворяет условию Липшица, то существует постоянный шаг, при котором градиентный спуск сходится. Правая часть (5), очевидно, удовлетворяет условию Липшица, поэтому можно утверждать, что найдется такой малый шаг δ, при котором алгоритм (6) будет сходящимся. В дальнейшем при расчетах будем использовать часто применяющееся на практике значение $\delta = 0.001$.

Проблема тут заключается в том, что на каждой итерации мы вместо движения по истинной траектории совершаем малый шаг вдоль касательной к этой траектории (см. рис. 3). С учетом выпуклости множества X это приводит к тому, что с ростом k точка x_k будет все дальше отдаляться от истинной траектории, пролегающей по границе множества X.

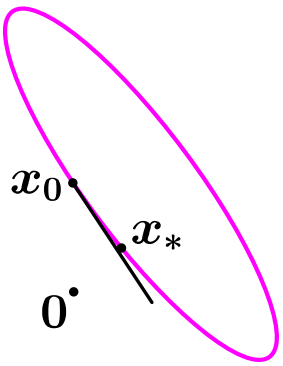


Рис. 3. Траектория решения по схеме (6) выводит за границы множества с ростом k. Здесь x_* — положение равновесия

Для преодоления этой проблемы введем специальную процедуру коррекции, заключающейся в проектировании (возвращении) точки x_k на границу

множества X. Пусть из x_{k-1}, совершив итерацию по (6), попадаем в точку $\widetilde{x}_{k-1}$. Так как $\widetilde{x}_{k-1}$ при малом δ находится в малой окрестности границы множества X, можем считать, что прямая, проходящая через $\widetilde{x}_{k-1}$ параллельно $\nabla f(\widetilde{x}_{k-1})$ пересекает поверхность множества X под прямым углом. Поэтому точку этого пересечения будем считать проекцией $\widetilde{x}_{k-1}$ на границу X и именно ее брать в качестве x_k. Таким образом

$$x_k = \widetilde{x}_{k-1} + \frac{\nabla f(\widetilde{x}_{k-1})}{\|\nabla f(\widetilde{x}_{k-1})\|}\xi,$$

где ξ некий скалярный параметр, определяемый из условия

$$f\left(\widetilde{x}_{k-1} + \frac{\nabla f(\widetilde{x}_{k-1})}{\|\nabla f(\widetilde{x}_{k-1})\|}\xi\right) = 0.$$

Разлагая эту функцию в окрестности $\xi = 0$ и учитывая близость $\widetilde{x}_{k-1}$ к границе X, отбрасываем все слагаемые выше первого порядка. Получаем уравнение

$$f(\widetilde{x}_{k-1}) + \left\langle \nabla f(\widetilde{x}_{k-1}), \frac{\nabla f(\widetilde{x}_{k-1})}{\|\nabla f(\widetilde{x}_{k-1})\|} \right\rangle \xi = 0.$$

Значит,

$$x_k = \widetilde{x}_{k-1} - \frac{\nabla f(\widetilde{x}_{k-1})}{\|\nabla f(\widetilde{x}_{k-1})\|^2} f(\widetilde{x}_{k-1}).$$

Окончательно алгоритм (6) перепишется в виде двухэтапной процедуры

$$\begin{cases} \widetilde{x}_{k-1} = x_{k-1} + \delta z_{k-1} \\ x_k = \widetilde{x}_{k-1} - \dfrac{\nabla f(\widetilde{x}_{k-1})}{\|\nabla f(\widetilde{x}_{k-1})\|^2} f(\widetilde{x}_{k-1}) \\ z_k = z_{k-1} + \delta\psi(x_{k-1}, z_{k-1}) \end{cases} \tag{7}$$

Отметим, что необходимость процедуры коррекции возникла только из-за замены непрерывного процесса дискретным аналогом. Полученная схема (7) есть ни что иное как корректная процедура численного решения системы (5). На рис. 4 приведена иллюстрация работы алгоритма (7).

2.3. Критерий остановки и скорость сходимости. В положении равновесия сила нормальной реакции уравновешивается силой Кулона, что означает коллинеарность векторов x_k и $\nabla f(x_k)$. Поэтому в качестве критерия остановки можно взять условие

$$\sqrt{\sum_{i=2}^{n} \left(\frac{x_k^i}{x_k^1} - \frac{f'_{x^i}(x_k)}{f'_{x^1}(x_k)}\right)^2} < \varepsilon, \tag{8}$$

где $\varepsilon > 0$ — произвольное малое число.

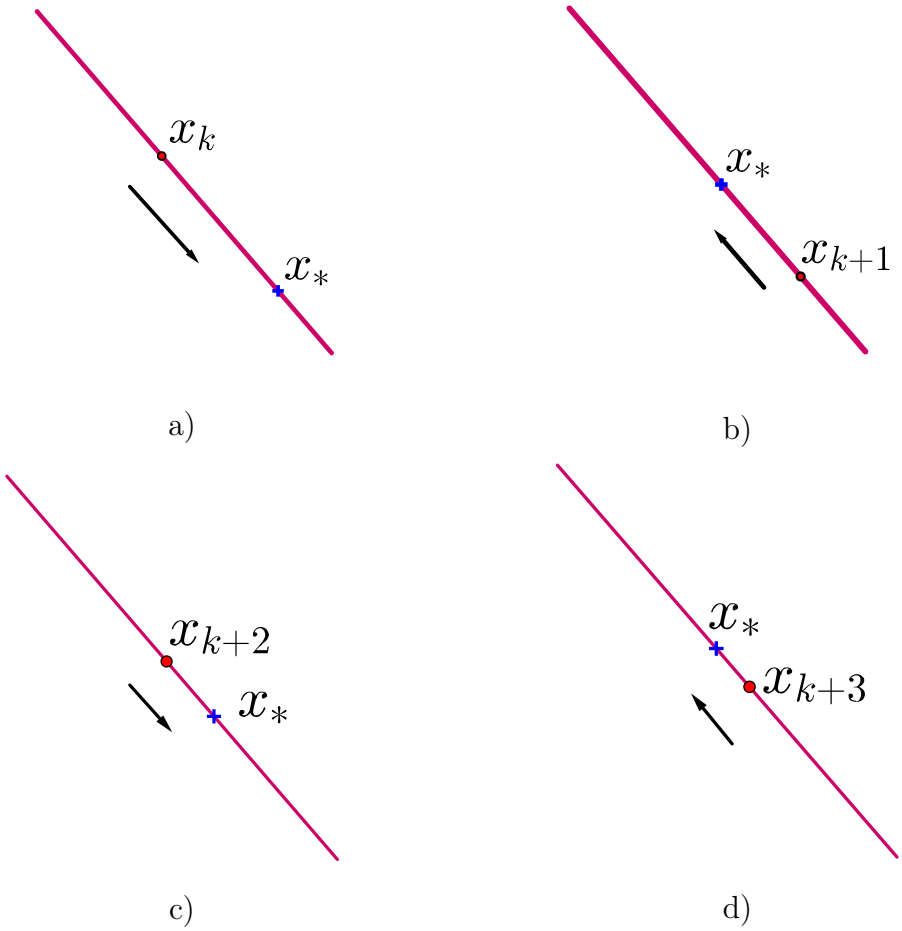


Рис. 4. Иллюстрация работы алгоритма (7): a) точка движется к положению равновесия x_*, b) точка "проскакивает" x_*, останавливается и меняет направление движения, c) точка опять "проскакивает" x_*, но останавливается ближе к x_*, d) точка останавливается еще ближе к x_*

Отметим, что коллинеарность векторов x_k и $\nabla f(x_k)$ можно получить и из необходимых условия экстремума для решаемой нами задачи условной оптимизации. Действительно, функция Лагранжа в данном случае может быть записана в виде

$$L(x, \lambda) = \|x\|^2 + \lambda f(x),$$

где λ — некоторый скаляр. Отсюда, решение должно удовлетворять условию

$$2x + \lambda \nabla f(x) = 0,$$

что опять же означает коллинеарность x_k и $\nabla f(x_k)$.

Из механики самого процесса очевидно, что решение системы (5) асимптотически устойчиво. Кроме того, для двумерного случая можно построить первое приближение системы в отклонениях

$$\begin{pmatrix}\dot\xi_1\\ \dot\xi_2\\ \dot\xi_3\\ \dot\xi_4\end{pmatrix} = \begin{pmatrix}0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ p_1a_1 & p_1b_1 & -p_2 & 0\\ p_1a_2 & p_1b_2 & 0 & -p_2\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\\ \xi_3\\ \xi_4\end{pmatrix},$$

где a_i, b_i — некоторые значения зависящие от x_*, $f(x_*)$ и $\nabla f(x_*)$. Корни характеристического уравнения для этой системы могут быть найдены с помощью формул Феррари:

$$-\frac{p_2}{2} - \frac{\sqrt{2p_1(A-B)+p_2^2}}{2}, \quad -\frac{p_2}{2} + \frac{\sqrt{2p_1(A-B)+p_2^2}}{2},$$

$$-\frac{p_2}{2} - \frac{\sqrt{2p_1(A+B)+p_2^2}}{2}, \quad -\frac{p_2}{2} + \frac{\sqrt{2p_1(A+B)+p_2^2}}{2},$$

где

$$A = a_1 + b_2, \quad B = \sqrt{a_1^2 - 2a_1b_2 + b_2^2 + 4a_2b_1}.$$

Отсюда следует, что по крайней мере в окрестности решения система экспоненциально устойчива, а значит алгоритм (7) сходится с линейной скоростью.

За счет выбора параметров p_1, p_2 можно существенно влиять на скорость сходимости полученного линейного процесса.

2.4. Численные эксперименты. В табл. 1 приведены результаты численного эксперимента для задачи поиска минимального расстояния от нуля до эллипсоида

$$f(x) = \langle Ax, x\rangle + \langle b, x\rangle + c,$$

где

$$A = \begin{pmatrix}4 & 5/2\\ 5/2 & 2\end{pmatrix}, \ b = \begin{pmatrix}-30\\ -21\end{pmatrix}, \ c = 47.$$

Необходимо отметить, что при $p_1 \gg p_2$ происходит очень медленное затухание колебаний, а при $p_1 \ll p_2$ сопротивление очень велико. И то и другое ведет к медленной сходимости процесса (см. табл. 1). Оптимальные значения p_1, p_2, очевидно, находятся между этими двумя предельными случаями.

p_1	p_2	$criteria$[a]	N[b]	$\|x - x_*\|$[c]	t[d]
100	0.05	9.3×10^{-9}	238411	0.87×10^{-6}	1.245
100	0.1	6.9×10^{-9}	71601	8.53×10^{-6}	0.379
100	0.5	3.5×10^{-9}	36100	0.15×10^{-6}	0.195
100	1	4.2×10^{-9}	19183	17.5×10^{-9}	0.104
100	5	3.1×10^{-9}	5009	80.8×10^{-9}	0.027
100	10	10×10^{-9}	8216	20.2×10^{-9}	0.044
100	20	10×10^{-9}	19588	20.1×10^{-9}	0.105
100	30	10×10^{-9}	30041	20.1×10^{-9}	0.163
100	50	10×10^{-9}	50608	20×10^{-9}	0.273
100	100	10×10^{-9}	101665	20×10^{-9}	0.555
100	200	10×10^{-9}	203556	20×10^{-9}	1.105
100	2000	10×10^{-9}	2036340	20×10^{-9}	10.528

[a] Значение выражение (8),

[b] Количество итерации,

[c] Расстояние от истинного решения,

[d] Время работы алгоритма в секундах

Таблица 1. Результаты применения алгоритма (7) при $\varepsilon = 10^{-8}$, $\delta = 10^{-3}$.

3°. Нахождение минимального расстояния между двумя множествами. Пусть $X = \{x \in \mathbb{R}^n |\ f_1(x) \leq 0\}$, $Y = \{y \in \mathbb{R}^n |\ f_2(y) \leq 0\}$, где $f_1,\ f_2 : \mathbb{R}^n \to \mathbb{R}$ — выпуклые, непрерывно дифференцируемые функции. Экстремальная задача

$$\begin{cases} \|x - y\| \longrightarrow \inf \\ x \in X \\ y \in Y \end{cases}$$

может быть решена с применением того же подхода. Можно поместить в указанные множества разнозаряженные шарики. Сначала они будут двигаться по прямой, их соединяющей, до столкновения с границами множеств в точках $x_0 \in X$ и $y_0 \in Y$ (см. рис. 5).

Кулоновские силы, действующие на первую и вторую точку, будут равны соответственно

$$F_1(t) = \frac{c_1 q^2}{\|x - y\|^3}(y - x) \text{ и } \ F_2(t) = \frac{c_1 q^2}{\|x - y\|^3}(x - y).$$

Для сил нормальной реакции имеем

$$N_1(t) = -\frac{\|F_1(t)\| \cdot \nabla f_1(x)}{\|x - y\| \cdot \|\nabla f_1(x)\|^2} \langle y - x, \nabla f_1(x) \rangle ,$$

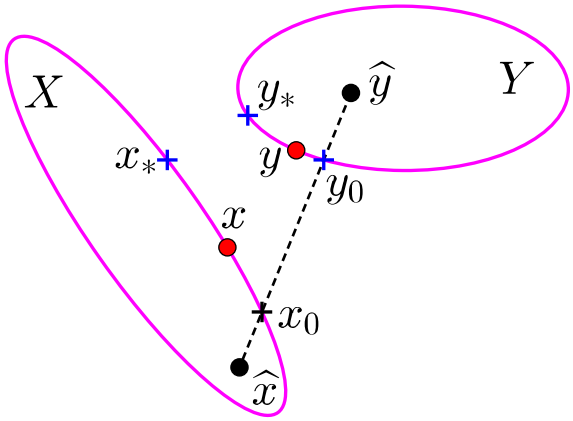


Рис. 5. Нахождение начальных точек x_0 и y_0

$$N_2(t) = -\frac{\|F_2(t)\| \cdot \nabla f_2(y)}{\|x-y\| \cdot \|\nabla f_2(y)\|^2} \langle x-y, \nabla f_2(y)\rangle .$$

Силы сопротивления определим формулами

$$R_1(t) = -c_2\dot{x},\ R_2(t) = -c_2\dot{y}.$$

Здесь по-прежнему c_1 — электрическая постоянная, c_2 — коэффициент сопротивления, q — заряд шариков. Окончательно, принимая массу шариков равной m, получаем дифференциальные уравнения движения:

$$\begin{cases} m\ddot{x}(t) = F_1(t) + N_1(t) + R_1(t) \\ m\ddot{y}(t) = F_2(t) + N_2(t) + R_2(t) \end{cases}$$

Понижая порядок системы за счет введения фиктивных переменных z_1, z_2, переходим к системе

$$\begin{cases} \dot{x} = z_1 \\ \dot{y} = z_2 \\ \dot{z_1} = \psi_1(x, y, z_1, z_2) \\ \dot{z_2} = \psi_2(x, y, z_1, z_2) \end{cases} \tag{9}$$

где

$$\psi_1(x,y,z_1,z_2) = \frac{p_1}{\|x-y\|^3}\left[y - x + \frac{\langle x-y, \nabla f_1(x)\rangle}{\|\nabla f_1(x)\|^2}\nabla f_1(x)\right] - p_2 z_1,$$

$$\psi_2(x,y,z_1,z_2) = \frac{p_1}{\|x-y\|^3}\left[x - y + \frac{\langle y-x, \nabla f_2(y)\rangle}{\|\nabla f_2(y)\|^2}\nabla f_2(y)\right] - p_2 z_2.$$

Полученную систему (9) будем опять же решать методом ломанных Эйлера, введя дополнительные корректирующие процедуры, аналогично тому, как это делалось выше.

Выбираем $x(0) = x_0$, $y(0) = y_0$, $z_1(0) = 0_n$, $z_2(0) = 0_n$ и некоторое малое положительное δ — длину шага. Пусть у нас есть x_{k-1}, y_{k-1}, $z_{1,k-1}$, $z_{2,k-1}$, тогда

$$\begin{cases} \widetilde{x}_{k-1} = x_{k-1} + \delta z_{1,k-1} \\ \widetilde{y}_{k-1} = y_{k-1} + \delta z_{2,k-1} \\ x_k = \widetilde{x}_{k-1} - \dfrac{\nabla f_1(\widetilde{x}_{k-1})}{\|\nabla f_1(\widetilde{x}_{k-1})\|^2} f_1(\widetilde{x}_{k-1}) \\ y_k = \widetilde{y}_{k-1} - \dfrac{\nabla f_2(\widetilde{y}_{k-1})}{\|\nabla f_2(\widetilde{y}_{k-1})\|^2} f_2(\widetilde{y}_{k-1}) \\ z_{1,k} = z_{1,k-1} + \delta\psi_1(x_{k-1}, y_{k-1}, z_{1,k-1}, z_{2,k-1}) \\ z_{2,k} = z_{2,k-1} + \delta\psi_2(x_{k-1}, y_{k-1}, z_{1,k-1}, z_{2,k-1}) \end{cases}$$

В положении равновесия силы нормальных реакций уравновешиваются соответствующими силами Кулона, что означает коллинеарность векторов $x_k - y_k$, $\nabla f_1(x_k)$ и $\nabla f_2(y_k)$. Поэтому в качестве критерия остановки можно взять условие

$$\sqrt{\sum_{i=2}^{n} \left(\frac{x_k^i - y_k^i}{f'_{1,x^i}(x_k)} - \frac{x_k^1 - y_k^1}{f'_{1,x^1}(x_k)} \right)^2} + \sqrt{\sum_{i=2}^{n} \left(\frac{x_k^i - y_k^i}{f'_{2,y^i}(y_k)} - \frac{x_k^1 - y_k^1}{f'_{2,y^1}(y_k)} \right)^2} < \varepsilon,$$

где $\varepsilon > 0$ — произвольное малое число.

4°. Благодарности. Автор выражает признательность д.ф.-м.н., профессору В.Н. Малозёмову за поддержку настоящей работы, ряд важных замечаний и полезные обсуждения.

ЛИТЕРАТУРА

1. Abbasov M. E. *Charged balls method for finding minimum distance between two smooth closed convex sets* // [to appear in] Optimization, 2015.

2. Demyanov V. F., Rubinov A. M. *Constructive Nonsmooth Analysis.* Approximation & Optimization, 1995. 7. Peter Lang, Frankfurt am Main, iv+416 pp.

3. Demyanov V. F., Malozemov V. N. *Introduction to minimax.* New York: Dover, 1990. 307 p.

4. Abbasov M. E., Demyanov V. F. *Proper and adjoint exhausters in nonsmooth analysis: optimality conditions* // J. Glob. Optim., 2013. Vol. 56, Issue 2, P. 569–585.

5. Косолап А. И. *Квадратичные оптимизационные задачи компьютерной геометрии* // Искусственный интеллект, 2010. №1. С. 70–75.

6. Lin A., Han S. P. *On the distance between two ellipsoids* // SIAM Journal on Optimization, 2002. Vol. 13. P. 298–308.

7. Hu S.-M., Wallner J. *A second order algorithm for orthogonal projection onto curves and surfaces* // Computer Aided Geometric Design, 2005. Vol. 22. P. 251–260.

8. Лебедев Д. М., Полякова Л. Н. *Задача проектирования нулевой точки на квадрику* // Вестник СПбГУ. Сер. 10, 2013. Вып. 1. С. 11–17.

9. Утешев А. Ю. *Вычисление расстояний между геометрическими объектами* // (`http://pmpu.ru/vf4/algebra2/optimiz/distance`)

10. Утешев А. Ю., Яшина М. В. *Нахождение расстояния от эллипсоида до плоскости и квадрики в* $\mathbb{R}^n$ // Доклады АН, 2008. Т. 419, № 4. С. 471–474.

11. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. *Численные методы.* М.: Наука, 1987, 600 с.

НАХОЖДЕНИЕ МИНИМАЛЬНОГО РАССТОЯНИЯ МЕЖДУ ДВУМЯ ГЛАДКИМИ КРИВЫМИ В ТРЁХМЕРНОМ ПРОСТРАНСТВЕ*

М. Э. Аббасов

Аннотация. В докладе [1] рассматривался метод заряженных шариков, который применялся для решения важных задач вычислительной геометрии:

• ортогонального проектирования точки на выпуклое, замкнутое множество, размерность которого совпадала с размерностью пространства,

• поиска минимального расстояния между двумя такими множествами.

В данной работе рассматривается задача поиска минимального расстояния между двумя гладкими кривыми в трехмерном пространстве, часто возникающая в астрономии. Показывается, что и она может быть эффективно решена с помощью той же идеи.

В заключении приводятся численные примеры, иллюстрирующие работу предложенного алгоритма.

1°. Постановка задачи и используемые обозначения. Изложение будет вестись для 3-х мерного евклидова пространства. Рассмотрим в $\mathbb{R}^3$ кривые

$$r_1(u) = (x_1(u), y_1(u), z_1(u)), \ r_2(v) = (x_2(v), y_2(v), z_2(v)),$$

где $r_1, r_2 \colon \mathbb{R} \to \mathbb{R}^3$ – непрерывно дифференцируемые вектор-функции, скалярных параметров u и v соответственно. Будем искать точки

$$r_1^* = (x_1(u_*), y_1(u_*), z_1(u_*)), \ r_2^* = (x_2(v_*), y_2(v_*), z_2(v_*)),$$

на которых достигается минимальное расстояние между данными кривыми.

2°. Идейная основа алгоритма и его реализация. Как и прежде, предлагается на одну из кривых поместить положительно заряженный шарик, а на другую – отрицательно заряженный, причем шарики могут двигаться только по соответствующим кривым. Считаем, что движение происходит только под действием силы Кулона и пропорциональной скорости движения силы сопротивления. Последняя обеспечивает диссипацию (рассеивание) энергии, благодаря которой с ростом времени шарики стремятся к точкам r_1^*, r_2^*. Описываем дифференциальные уравнения движения. С помощью перехода к разностной схеме их решения, получим итерационный алгоритм решения данной задачи.

*Семинар «CNSA & NDO». Избранные доклады. 8 сентября 2016 г.

2.1. Уравнения движения. Так как шарики могут двигаться только по кривым, их положение в каждый конкретный момент времени t определяется параметрами u и v наших кривых. То есть координаты шариков есть функции времени $r_1(u(t))$, $r_2(v(t))$. Очевидно, нормальная составляющая силы Кулона для каждого шарика компенсируется нормальной реакцией соответствующей кривой, а потому при составлении уравнений движения нужно учитывать лишь силу вязкого трения, а также касательную составляющую силы Кулона. Последнюю можно получить спроецировав саму силу на касательное направление, которое для первой кривой задается вектором $\tau_1(u) = \dfrac{\dot r_1(u)}{\|\dot r_1(u)\|}$, а для второй – вектором $\tau_2(u) = \dfrac{\dot r_2(u)}{\|\dot r_2(u)\|}$. Силы Кулона, действующие на первый и второй шарики, с точностью до некоторой константы[1] равны соответственно $\dfrac{r_2 - r_1}{\|r_2 - r_1\|^3}$, $-\dfrac{r_2 - r_1}{\|r_2 - r_1\|^3}$. С помощью второго закона Ньютона уравнения движения можно записать в виде

$$\begin{cases} \dfrac{d^2 r_1}{dt^2} = p_1 \left\langle \dfrac{r_2 - r_1}{\|r_2 - r_1\|^3}, \tau_1 \right\rangle \tau_1 - p_2 \dfrac{dr_1}{dt} \\ \dfrac{d^2 r_2}{dt^2} = p_1 \left\langle \dfrac{r_1 - r_2}{\|r_2 - r_1\|^3}, \tau_2 \right\rangle \tau_2 - p_2 \dfrac{dr_2}{dt} \end{cases} \tag{1}$$

Здесь p_1, p_2 константы. Учитывая, что

$$\frac{d^2 r_1}{dt^2} = \frac{d^2 r_1}{du^2} \left(\frac{du}{dt}\right)^2 + \frac{d^2 u}{dt^2} \frac{dr_1}{du},$$

$$\frac{d^2 r_2}{dt^2} = \frac{d^2 r_2}{dv^2} \left(\frac{dv}{dt}\right)^2 + \frac{d^2 v}{dt^2} \frac{dr_2}{dv},$$

можем переписать (1) в виде

$$\begin{cases} \ddot u \dfrac{dr_1}{du} + \dot u^2 \dfrac{d^2 r_1}{du^2} + p_2 \dot u \dfrac{dr_1}{du} = p_1 \left\langle \dfrac{r_2 - r_1}{\|r_2 - r_1\|^3}, \tau_1 \right\rangle \tau_1 \\ \ddot v \dfrac{dr_2}{dv} + \dot v^2 \dfrac{d^2 r_2}{dv^2} + p_2 \dot v \dfrac{dr_2}{dv} = p_1 \left\langle \dfrac{r_1 - r_2}{\|r_2 - r_1\|^3}, \tau_2 \right\rangle \tau_2 \end{cases} \tag{2}$$

Так как $\langle \tau_1, \tau_1 \rangle = 1$, то

$$2 \left\langle \frac{d\tau_1}{du}, \tau_1 \right\rangle = 0.$$

Таким образом, $n_1 = \frac{d\tau_1}{du}$ ортогонально τ_1. С другой стороны очевидно, что n_1 лежит в соприкасающейся плоскости кривой. Поэтому n_1 есть главная нормаль нашей кривой. Аналогично $n_2 = \frac{d\tau_2}{dv}$ есть главная нормаль второй кривой.

[1] зависящей от электрической постоянной и зарядов шаров

Трехгранники Френе для наших кривых определяются векторами:

- касательных $\tau_1(t)$, $\tau_2(t)$;
- главных нормалей $n_1(t)$, $n_2(t)$;
- бинормалей b_1, b_2.

Скалярно умножим теперь уравнения (2) слева и справа на соответствующие касательные направления, вдоль которых и происходит движение[2].

Получаем

$$\begin{cases} \ddot{u}\left\|\dfrac{dr_1}{du}\right\| + \dot{u}^2\left\langle \tau_1, \dfrac{d^2r_1}{du^2}\right\rangle + p_2\dot{u}\left\|\dfrac{dr_1}{du}\right\| = p_1\left\langle \dfrac{r_2-r_1}{\|r_2-r_1\|^3}, \tau_1\right\rangle \\ \ddot{v}\left\|\dfrac{dr_2}{dv}\right\| + \dot{v}^2\left\langle \tau_2, \dfrac{d^2r_2}{dv^2}\right\rangle + p_2\dot{v}\left\|\dfrac{dr_2}{du}\right\| = p_1\left\langle \dfrac{r_1-r_2}{\|r_2-r_1\|^3}, \tau_2\right\rangle \end{cases} \tag{3}$$

Вводя фиктивные переменные $\xi = \dot{u}$, $\eta = \dot{v}$, понижаем порядок системы (3):

$$\begin{cases} \dot{u} = \xi \\ \dot{\xi} = p_1\left\langle \dfrac{r_2-r_1}{\|r_2-r_1\|^3}, \tau_1\right\rangle \left\|\dfrac{dr_1}{du}\right\|^{-1} - p_2\xi - \left\langle \tau_1, \dfrac{d^2r_1}{du^2}\right\rangle \left\|\dfrac{dr_1}{du}\right\|^{-1}\xi^2 \\ \dot{v} = \eta \\ \dot{\eta} = p_1\left\langle \dfrac{r_1-r_2}{\|r_2-r_1\|^3}, \tau_2\right\rangle \left\|\dfrac{dr_2}{dv}\right\|^{-1} - p_2\eta - \left\langle \tau_2, \dfrac{d^2r_2}{dv^2}\right\rangle \left\|\dfrac{dr_2}{dv}\right\|^{-1}\eta^2 \end{cases} \tag{4}$$

Применим явный метод Эйлера для решения системы (4) с шагом δ. Предполагая, что заданы u_k, ξ_k, v_k, η_k, получаем алгоритм решения исходной задачи:

$$\begin{cases} u_{k+1} = u_k + \delta\xi_k \\ \xi_{k+1} = \xi_k + \delta\left(\varphi_1(u_k, v_k)\left\|\dfrac{dr_1}{du}\right\|^{-1} - p_2\xi_k - \psi_1(u_k)\left\|\dfrac{dr_1}{du}\right\|^{-1}\xi_k^2\right) \\ v_{k+1} = v_k + \delta\eta_k \\ \eta_{k+1} = \eta_k + \delta\left(\varphi_2(u_k, v_k)\left\|\dfrac{dr_2}{dv}\right\|^{-1} - p_2\eta_k - \psi_2(v_k)\left\|\dfrac{dr_2}{dv}\right\|^{-1}\eta_k^2\right) \end{cases}$$

где

$$\varphi_1(u,v) = p_1\left\langle \frac{r_2(u)-r_1(v)}{\|r_2(v)-r_1(u)\|^3}, \tau_1(u)\right\rangle, \quad \psi_1(u) = \left\langle \frac{d^2r_1(u)}{du^2}, \tau_1(u)\right\rangle,$$

[2]Домножение на вектор главной нормали приводит к тривиальному решению.

$$\varphi_2(v,v) = p_1 \left\langle \frac{r_1(v) - r_2(u)}{\|r_2(v) - r_1(u)\|^3}, \tau_2(v) \right\rangle, \quad \psi_2(v) = \left\langle \frac{d^2 r_2(v)}{dv^2}, \tau_2(v) \right\rangle.$$

В точках r_1^*, r_2^*, очевидно, касательная составляющая сил Кулона должна равняться нулю, поэтому в качестве критерия остановки можно взять условие

$$|\varphi_1(u_k, v_k)| + |\varphi_2(u_k, v_k)| \le \varepsilon, \tag{5}$$

где $\varepsilon > 0$ некоторое малое число.

З а м е ч а н и е 1. Отметим, что решение задачи, вообще говоря, неединственно, поэтому получаемое в ходе применения алгоритма решение зависит от выбора начальных точек.

3°. Численные эксперименты. Приведем численные примеры решения задач по описанному алгоритму. Вычисления везде велись в математическом пакете Matlab 2013a.

3.1. Пример 1. Найдем расстояние между кривыми (см. рис. 1)

$$r_1(u) = (3\sin u,\ 5\cos u,\ 0)\,; \quad r_2(v) = (3\sin v,\ 5\cos v,\ 7 - 3.75\cos v)\,.$$

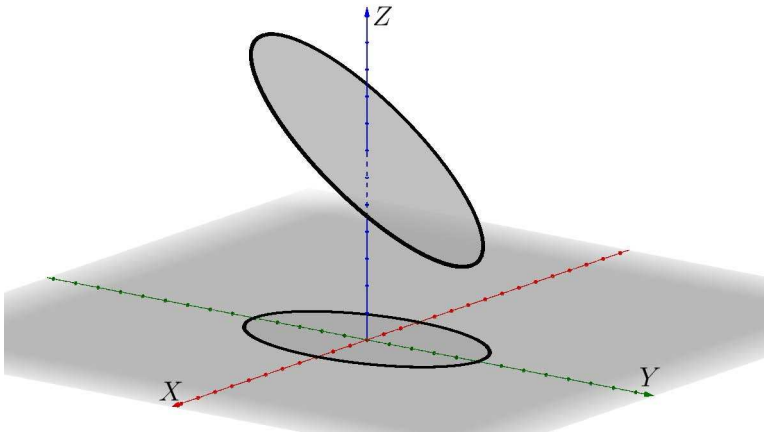


Рис. 1. Кривые из примера 1.

В Табл. 1 приведены результаты вычислений при различных значениях параметров метода p_1, p_2, δ.

3.2. Пример 2. Найдем расстояние между кривыми (см. рис. 2)

$$r_1(u) = (3\sin u,\ 5\cos u,\ 0)\,; \quad r_2(v) = (6\sin v - 6,\ 2\cos v,\ -1.5\cos v)\,.$$

В Табл. 2 приведены результаты вычислений при различных значениях параметров метода p_1, p_2, δ.

p_1	p_2	δ	$criteria^a$	N^b	$\|r_1 - r_1^*\|^c$	$\|r_2 - r_2^*\|^c$	t^d
100	2	0.1	$3.27 \cdot 10^{-7}$	212	$2.1 \cdot 10^{-8}$	$4.6 \cdot 10^{-8}$	0.0068
100	10	0.1	$10 \cdot 10^{-7}$	1268	$4.8 \cdot 10^{-7}$	$2.5 \cdot 10^{-7}$	0.0424
100	2	0.01	$10 \cdot 10^{-7}$	1815	$4.3 \cdot 10^{-7}$	$2.5 \cdot 10^{-7}$	0.0596
100	10	0.01	$10 \cdot 10^{-7}$	12732	$4.8 \cdot 10^{-7}$	$2.5 \cdot 10^{-7}$	0.3850
10	2	0.01	$10 \cdot 10^{-7}$	21793	$4.8 \cdot 10^{-6}$	$2.5 \cdot 10^{-6}$	0.6737
10	2	0.1	$10 \cdot 10^{-7}$	2175	$4.8 \cdot 10^{-6}$	$2.5 \cdot 10^{-6}$	0.0717

[a]Значение выражение (5), [b]Количество итерации, [c]Расстояние от истинного решения, [d]Время работы алгоритма, с.

Таблица 1. Решение задачи из примера 1 при $\varepsilon = 10^{-6}$, $u_0 = 1$, $v_0 = 1$, $\xi_0 = 0$, $\eta_0 = 0$ и различных значениях параметра метода.

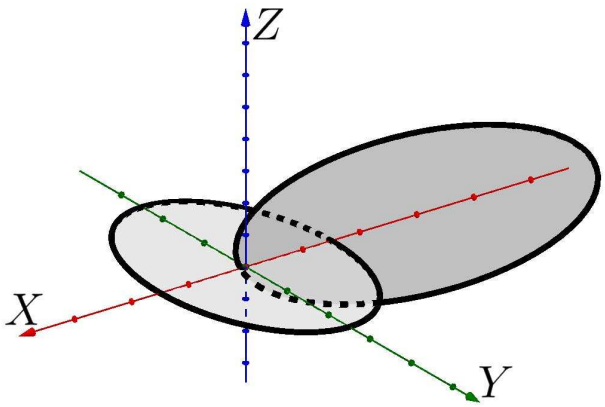


Рис. 2. Кривые из примера 2.

p_1	p_2	δ	$criteria^a$	N^b	$\|r_1 - r_1^*\|^c$	$\|r_2 - r_2^*\|^c$	t^d
100	10	0.1	$8.2 \cdot 10^{-5}$	114	$2.7 \cdot 10^{-7}$	$15.8 \cdot 10^{-7}$	0.0072
100	20	0.1	$6.2 \cdot 10^{-5}$	201	$4.7 \cdot 10^{-7}$	$7.7 \cdot 10^{-7}$	0.0099
100	20	0.01	$9.9 \cdot 10^{-5}$	2043	$8.5 \cdot 10^{-7}$	$11.7 \cdot 10^{-7}$	0.0628
200	20	0.1	$1.8 \cdot 10^{-5}$	93	$0.1 \cdot 10^{-7}$	$1.3 \cdot 10^{-7}$	0.0030

[a]Значение выражение (5), [b]Количество итерации, [c]Расстояние от истинного решения, [d]Время работы алгоритма, с.

Таблица 2. Решение задачи из примера 2 при $\varepsilon = 10^{-4}$, $u_0 = 0$, $v_0 = 0$, $\xi_0 = 0$, $\eta_0 = 0$ и различных значениях параметра метода.

3.3. Пример 3. Найдем расстояние между винтовой линией

$$r_1(u) = (\sin u,\ \cos u,\ u)$$

и окружностью (см. рис. 3)

$$r_2(v) = (3\sin v,\ 3\cos v,\ 0)\,.$$

В Табл. 3 приведены результаты вычислений при различных значениях параметров метода p_1, p_2, δ.

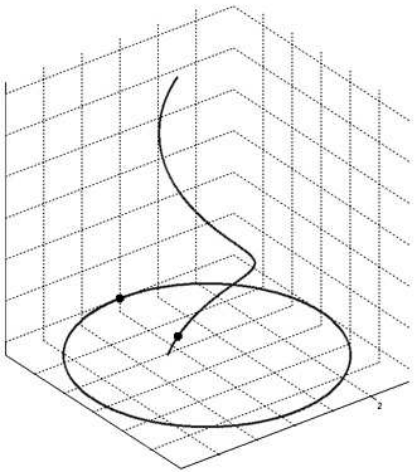


Рис. 3. Кривые из примера 3 с изображением точек, на которых достигается минимальное расстояние

p_1	p_2	δ	$criteria^a$	N^b	$\|r_1 - r_1^*\|^c$	$\|r_2 - r_2^*\|^c$	t^d
10	1	0.3	$7.4 \cdot 10^{-5}$	295	$17.25 \cdot 10^{-5}$	$43.74 \cdot 10^{-5}$	0.0422
100	10	0.1	$9.9 \cdot 10^{-5}$	1151	$2.89 \cdot 10^{-5}$	$7.88 \cdot 10^{-5}$	0.0401
100	20	0.1	$9.96 \cdot 10^{-5}$	2319	$2.9 \cdot 10^{-5}$	$7.91 \cdot 10^{-5}$	0.0748
200	20	0.1	$9.8 \cdot 10^{-5}$	1231	$1.43 \cdot 10^{-5}$	$3.89 \cdot 10^{-5}$	0.0422

[a]Значение выражение (5), [b]Количество итерации, [c]Расстояние от истинного решения, [d]Время работы алгоритма, с.

Таблица 3. Решение задачи из примера 3 при $\varepsilon = 10^{-4}$, $u_0 = 1$, $v_0 = 1$, $\xi_0 = 0$, $\eta_0 = 0$ и различных значениях параметра метода.

4°. Благодарности. Автор выражает признательность д.ф.-м.н., профессору В. Н. Малозёмову за помощь в постановке задачи и полезные замечания.

Работа выполнена при поддержке Санкт-Петербургского Государственного Университета в рамках гранта 9.38.205.2014, а также при финансовой поддержке РФФИ в рамках научного проекта № 16-31-00056 мол_а.

ЛИТЕРАТУРА

1. Аббасов М. Э. *Метод заряженных шариков* // Семинар «CNSA & NDO». Избранные доклады. 21 мая 2015 г.
(`http://apmath.spbu.ru/cnsa/reps15.shtml#0521`) [Данная книга, с. 278]

ПОСТРОЕНИЕ МИНИМАЛЬНОГО ЭЛЛИПСОИДА: АЛГОРИТМ ШОРА*

М. А. Кольцов

1°. Постановка задачи. В докладе [1] анализировалось решение *задачи Сильвестра* — задачи нахождения шара минимального объёма, который содержит заданное множество точек. Эту задачу можно обобщить: искать не шар, а *эллипсоид* минимального объёма. В отличие от задачи Сильвестра, которую можно решить точно с помощью квадратичного программирования, задачу нахождения минимального эллипсоида удаётся решить лишь приближённо. Данная статья посвящена одному алгоритму, строящему такое приближённое решение.

Теперь поставим задачу формально. Дано множество точек $c_j \in \mathbb{R}^N$, $j \in 1 : m$. Требуется построить эллипсоид $E \subset \mathbb{R}^N$ минимального объёма, который содержит все точки c_j. Эллипсоид E будем задавать таким способом:

$$E = \{x \in \mathbb{R}^n | \langle M(x-c), x-c\rangle \leqslant 1\} \ ,$$

где $c \in \mathbb{R}^N$ — центр эллипсоида, а M — симметричная положительно-определённая матрица порядка n.

Известно, что объём эллипсоида вычисляется по формуле

$$V = w_n \left(\det M\right)^{-1/2} \ ,$$

где w_n — объём единичного n-мерного шара. Таким образом, задачу нахождения минимального эллипсоида можно сформулировать как экстремальную задачу

$$\begin{aligned} &V := w_n \left(\det M\right)^{-1/2} \to \min \\ &\langle M(c_j - c), c_j - c\rangle \leqslant 1,\ j \in 1 : m \\ &M \text{ положительно определена} \end{aligned}$$

Отбросим множитель w_n, возведём целевую функцию в степень -2 и перейдём к задаче максимизации определителя матрицы M. В итоге получается

*Семинар «CNSA & NDO». Избранные доклады. 14 мая 2015 г.

экстремальная задача

$$\begin{aligned} &f := \det M \to \max \\ &\langle M(c_j - c), c_j - c\rangle \leqslant 1,\ j \in 1:m \\ &M \text{ положительно определена} \end{aligned} \tag{1}$$

Если среди c_j найдутся $n+1$ афинно независимых точек, то решение задачи (1) существует и единственно (см статью [2]). Нас же пока интересует только алгоритм построения.

2°. Оператор сжатия пространства. Алгоритм построения минимального эллипсоида использует оператор сжатия пространства, поэтому рассмотрим сначала его свойства.

ОПРЕДЕЛЕНИЕ. Пусть $\xi \in \mathbb{R}^n$, $\|\xi\| = 1$ и $\gamma \in (0,1)$. Тогда оператор сжатия R_γ определяется формулой

$$R_\gamma = E - (1-\gamma)\xi\xi^T$$

Здесь $\xi\xi^T$ — это матрица проектирования на прямую $x = \lambda\xi$, $\lambda \in \mathbb{R}$. Вычисление значения R_γ на конкретном векторе $y \in \mathbb{R}^n$ позволяет понять, как действует этот оператор:

$$R_\gamma y = (E - (1-\gamma)\xi\xi^T)y = y - (1-\gamma)\xi\xi^T y = y - (1-\gamma)\cdot\langle\xi, y\rangle\cdot\xi$$

Из этого равенства и определения R_γ очевидны следующие свойства:

1) R_γ — симметричная матрица

2) $R_\gamma\xi = \gamma\xi$

3) $R_\gamma p = p$, если $\langle\xi, p\rangle = 0$

Значит, R_γ имеет собственное число $\lambda_0 = \gamma$ кратности 1 с собственным вектором ξ и собственное число $\lambda_1 = 1$ кратности $n-1$ с собственным подпространством, ортогональным ξ. Если $y = \alpha\xi + p$, где $\langle\xi, p\rangle = 0$, то $R_\gamma y = \gamma\alpha\xi + p$.

Кроме этого, матрица R_γ положительно определена при всех $\gamma \in (0,1)$.

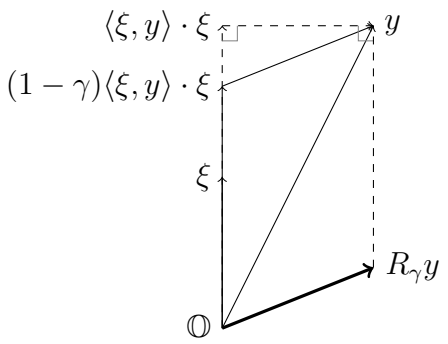


Рис. 1. Оператор сжатия с $\gamma = 0.2$

ЛЕММА 1. *Оператор R_γ обратим и обратный к нему оператор задаётся формулой*

$$R_\gamma^{-1} = E + \frac{1-\gamma}{\gamma}\xi\xi^T$$

Доказательство. Вычислим произведение $R_\gamma \cdot R_\gamma^{-1}$:

$$\begin{gathered}\left(E-(1-\gamma)\xi\xi^T\right)\left(E+\frac{1-\gamma}{\gamma}\xi\xi^T\right) = \\ = E + \frac{1-\gamma}{\gamma}\xi\xi^T - (1-\gamma)\xi\xi^T - \frac{(1-\gamma)^2}{\gamma}\xi\underbrace{\xi^T\xi}_{=1}\xi^T = \\ = E + \frac{1-\gamma}{\gamma}\left(\xi\xi^T - \gamma\xi\xi^T - (1-\gamma)\xi\xi^T\right) = E\end{gathered}$$

□

К обратному оператору применимы те же рассуждения о собственных числах и собственных векторах, в частности, вектор ξ собственный с собственным числом $\frac{1}{\gamma}$.

Теперь поймём, как оператор сжатия действует на шары и эллипсоиды в $\mathbb{R}^n$. Пусть, для начала, имеется шар с центром в точке c и радиусом r, который задан уравнением

$$\langle x-c, x-c\rangle \leqslant r^2$$

Подействуем на него оператором R_γ, получив новые точки $y = R_\gamma x$ и новый центр $a = R_\gamma c$. Так как R_γ обратим, то $x = R_\gamma^{-1}y$ и $c = R_\gamma^{-1}a$. Подставим эти выражения в уравнение шара

$$\begin{gathered}\langle x-c, x-c\rangle \leqslant r^2 \\ \langle R_\gamma^{-1}(y-a), R_\gamma^{-1}(y-a)\rangle \leqslant r^2 \\ \langle (R_\gamma^{-2})(y-a), y-a\rangle \leqslant r^2 \\ \langle \frac{R_\gamma^{-2}}{r^2}(y-a), y-a\rangle \leqslant 1\end{gathered}$$

Получилось уравнение эллипсоида с центром в точке $a = R_\gamma c$ и матрицей $M := \frac{R_\gamma^{-2}}{r^2}$. Как было установлено выше, матрица M имеет собственный вектор ξ, которому соответствует собственное число $\frac{1}{r^2\gamma^2}$, а остальные собственные векторы ортогональны ξ и имеют собственные числа $\frac{1}{r^2}$. Таким образом, оператор сжатия превращает шар радиуса r в эллипсоид, одной из полуосей которого является ξ с длиной $r\gamma$, а длины остальных полуосей равны r — шар «сплющивается» в направлении ξ.

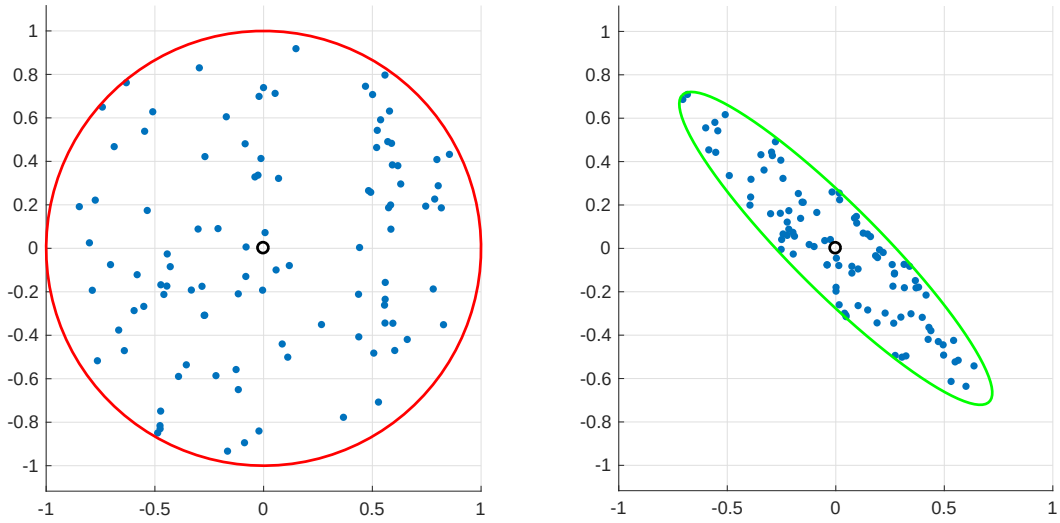
Разобравшись с шаром, подействуем теперь оператором R_γ на эллипсоид, заданный уравнением

$$\langle M(x-c), x-c\rangle \leqslant 1$$

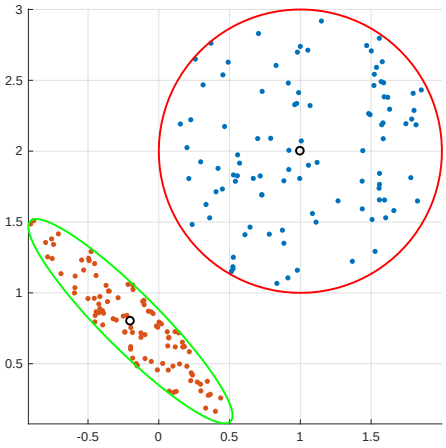
Положим аналогично $x = R_\gamma^{-1}y$, $c = R_\gamma^{-1}a$ и получим новое уравнение эллипсоида в виде

$$\langle R_\gamma^{-1} M R_\gamma^{-1}(y-a), y-a\rangle \leqslant 1$$

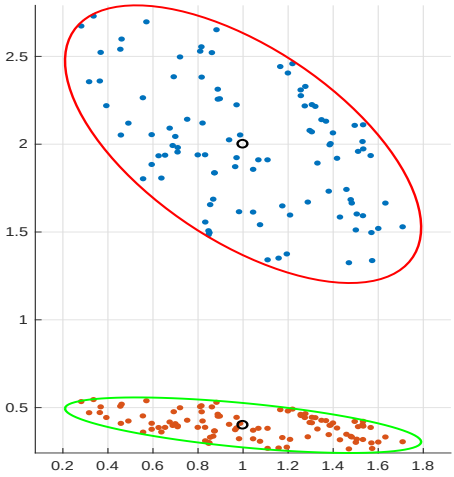
Наглядное представление о работе оператора сжатия даёт рис. 2.



(a) Простейший случай — сжатие единичного круга в направлении $(1,1)$ с $\gamma = 0.2$



(b) Сжатие круга в направлении $(1,1)$ с $\gamma = 0.2$



(c) Сжатие эллипса в направлении $(0,1)$ с $\gamma = 0.4$

Рис. 2. Действие оператора сжатия на множество, выделенное красной линией. Результат выделен зелёным цветом

3°. Итеративный алгоритм построения минимального эллипсоида. Для задачи о минимальном эллипсоиде известен итеративный алгоритм, строящий последовательные приближения к ответу. Этот алгоритм, описанный Н.З. Шором в статье [3], основан на достаточно простых идеях.

Введём вспомогательную задачу: вложим множество точек c_j в гиперплоскость $x_{n+1} = 1$ пространства $\mathbb{R}^{n+1}$. Обозначим $\tilde{c}_j = \begin{pmatrix} c_j \\ 1 \end{pmatrix}$. В новом пространстве будем искать минимальный эллипсоид с фиксированным центром в нуле. В статье [3] утверждается, что сечение такого эллипсоида плоскостью $x_{n+1} = 1$ и будет решением исходной задачи (см рис. 3).

Поставим формально вспомогательную задачу в терминах матрицы $\tilde{B}$ порядка $n + 1$:

$$\begin{aligned} &f := \det \tilde{B} \to \max \\ &\langle \tilde{B}\tilde{c}_j, \tilde{c}_j \rangle \leqslant 1, \ j \in 1 : m \\ &\tilde{B} \text{ положительно определена} \end{aligned} \tag{2}$$

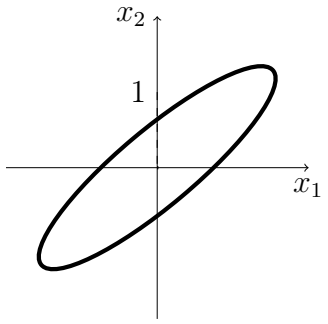


Рис. 3. Сечение эллипса с центром в нуле прямой $x_2 = 1$

Рассмотрим вопрос построения сечения подробнее. Итак, пусть

$$\tilde{E} = \left\{ \tilde{x} \in \mathbb{R}^{n+1} | \langle \tilde{B}\tilde{x}, \tilde{x} \rangle \leqslant 1 \right\}$$

— минимальный эллипсоид, содержащий точки $\tilde{c}_j$, $j \in 1 : m$. Возьмём две различные точки $\tilde{c}_{j_0}$ и $\tilde{c}_{j_1}$. Отметим, что

$$\langle \tilde{B}(\tilde{c}_{j_0} + \tilde{c}_{j_1}), \tilde{c}_{j_0} + \tilde{c}_{j_1} \rangle + \langle \tilde{B}(\tilde{c}_{j_0} - \tilde{c}_{j_1}), \tilde{c}_{j_0} - \tilde{c}_{j_1} \rangle = 2\langle \tilde{B}\tilde{c}_{j_0}, \tilde{c}_{j_0} \rangle + 2\langle \tilde{B}\tilde{c}_{j_1}, \tilde{c}_{j_1} \rangle \tag{3}$$

Обозначим

$$\tilde{x}_0 = \frac{1}{2}(\tilde{c}_{j_0} + \tilde{c}_{j_1}) = \begin{pmatrix} x_0 \\ 1 \end{pmatrix}, \quad \tilde{x}_1 = \frac{1}{2}(\tilde{c}_{j_0} - \tilde{c}_{j_1}) = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}$$

Поделив равенство (3) на 4, получим

$$\langle \tilde{B}\tilde{x}_0, \tilde{x}_0\rangle = \frac{1}{2}\langle \tilde{B}\tilde{c}_{j_0}, \tilde{c}_{j_0}\rangle + \frac{1}{2}\langle \tilde{B}\tilde{c}_{j_1}, \tilde{c}_{j_1}\rangle - \langle \tilde{B}\tilde{x}_1, \tilde{B}\tilde{x}_1\rangle$$

По определению точки $\tilde{c}_{j_0}$ и $\tilde{c}_{j_1}$ принадлежат $\tilde{E}$ и $\tilde{x}_1 \neq \mathbb{O}$, поэтому

$$\langle \tilde{B}\tilde{x}_0, \tilde{x}_0\rangle < 1 \tag{4}$$

Представим матрицу $\tilde{B}$ в виде

$$\tilde{B} = \begin{pmatrix} B & b \\ b^T & \hat{b} \end{pmatrix},$$

где B — главный минор порядка n. Очевидно, что B симметрична и положительно определена (по критерию Сильвестра). Аккуратно раскроем скобки в (4):

$$\begin{gathered} \langle \tilde{B}\tilde{x}_0, \tilde{x}_0\rangle = \langle \begin{pmatrix} Bx_0 + b \\ \langle b, x_0\rangle + \hat{b} \end{pmatrix}, \begin{pmatrix} x_0 \\ 1 \end{pmatrix}\rangle = \langle Bx_0 + b, x_0\rangle + \langle b, x_0\rangle + \hat{b} = \\ = \langle B(x_0 + B^{-1}b), x_0 + B^{-1}b\rangle - \langle B(x_0 + B^{-1}b), B^{-1}b\rangle + \langle b, x_0\rangle + \hat{b} = \\ = \langle B(x_0 + B^{-1}b), x_0 + B^{-1}b\rangle - \langle B^{-1}b, b\rangle + \hat{b} < 1 \end{gathered} \tag{5}$$

ЛЕММА 2. *Справедливы неравенства*

$$0 < -\langle B^{-1}b, b\rangle + \hat{b} < 1$$

Доказательство. Пусть, во-первых, $\tilde{x} = \begin{pmatrix} -B^{-1}b \\ 1 \end{pmatrix} \neq \mathbb{O}$. Матрица $\tilde{B}$ положительно определена, так что $\langle \tilde{B}\tilde{x}, \tilde{x}\rangle > 0$. Рассуждения, аналогичные (5), приводят к равенству $\langle \tilde{B}\tilde{x}, \tilde{x}\rangle = -\langle B^{-1}b, b\rangle + \hat{b}$, что доказывает левое неравенство.

Для доказательство правого неравенства заметим, что в неравенстве (5) слагаемое $\langle B(x + B^{-1}b), x + B^{-1}b\rangle$ неотрицательное в силу положительной определённости B. Тогда имеем

$$1 > \langle B(x + B^{-1}b), x + B^{-1}b\rangle - \langle B^{-1}b, b\rangle + \hat{b} \geqslant -\langle B^{-1}b, b\rangle + \hat{b}$$

Неравенство доказано. □

Теперь последнее неравенство из (5) поделим на положительное число $1 + \langle B^{-1}b, b\rangle - \hat{b}$, введём обозначения

$$c := -B^{-1}b, \quad M := B/(1 + \langle B^{-1}b, b\rangle - \hat{b}) \tag{6}$$

и окончательно получим

$$\langle M(x-c), x-c\rangle \leqslant 1 \text{ — уравнение минимального эллипсоида в } \mathbb{R}^n$$

Алгоритм решения задачи (2) устроен следующим образом. Текущее множество точек $\tilde{c}_j^{(k)}$ хранится в виде столбцов матрицы X_k, текущая матрица последовательности операторов сжатия обозначается A_k. Изначально X_0 состоит из исходных точек $\tilde{c}_j \in \mathbb{R}^{n+1}$, а $A_0 = E_{n+1}$. Один шаг работы алгоритма состоит из нескольких простых действий:

- Найти среди векторов $\tilde{c}_j^{(k)}$ вектор $\tilde{c}_{j_k}^{(k)}$ с максимальной нормой $\tau_{k+1} = \|\tilde{c}_{j_k}^{(k)}\|$
- Вычислить единичный вектор направления сжатия $\xi_{k+1} = \dfrac{\tilde{c}_{j_0}^{(k)}}{\tau_{k+1}}$
- Построить оператор сжатия R_{k+1} вдоль вектора ξ_{k+1} с коэффициентом α_{k+1}
- Вычислить новые точки $\tilde{c}_j^{(k+1)}$, умножив матрицу R_{k+1} на матрицу X_k, и сохранить результат в X_{k+1}
- Добавить матрицу R_{k+1} к последовательности операторов сжатия, умножив её на матрицу A_k, и сохранить результат в A_{k+1}

Коэффициенты α_k в соответствии со статьей [3] выбираются из условий

$$\alpha_k = 1 - \beta_k, \quad \beta_k > 0, \quad \beta_k \xrightarrow[k\to\infty]{} 0, \quad \sum_{k=1}^{\infty} \beta_k = \infty$$

Пусть вычисления остановлены после шага m. Тогда из описания алгоритма следует, что

$$X_m = R_m \cdot R_{m-1} \cdots R_1 \cdot X_0 = A_m \cdot X_0$$

Также известно, что точки $\tilde{c}_j^{(m)}$ лежат внутри сферы с радиусом τ_m и центром в начале координат. С учётом этих двух фактов можно получить матрицу искомого эллипсоида:

$$\begin{gathered}
\langle \tilde{c}_j^{(m)}, \tilde{c}_j^{(m)}\rangle \leqslant \tau_m^2 \\
\langle A_m\tilde{c}_j, A_m\tilde{c}_j\rangle \leqslant \tau_m^2 \\
\langle A_m^T A_m\tilde{c}_j, \tilde{c}_j\rangle \leqslant \tau_m^2 \\
\langle \frac{A_m^T A_m}{\tau_m^2}\tilde{c}_j, \tilde{c}_j\rangle \leqslant 1
\end{gathered}$$

Матрица $\tilde{B} := A_m^T A_m / \tau_m^2$ симметрична, положительно определена и является (приближённым) решением вспомогательной задачи о минимальном эллипсоиде. Окончательное решение основной задачи получается по формулам (6).

4°. Практическое испытание алгоритма. Алгоритм был реализован в среде MATLAB и испытан на тестовых данных. Коэффициенты β_k были равны $1/(k+1)$. В качестве тестовых данных было сгенерировано множество случайных точек внутри эллипсоида с центром в точке $(1, 2)$ с полуосями $2, 1$, который вытянут в направлении $(-1, 1)$.

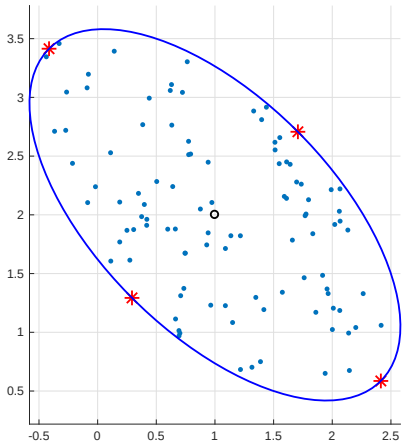
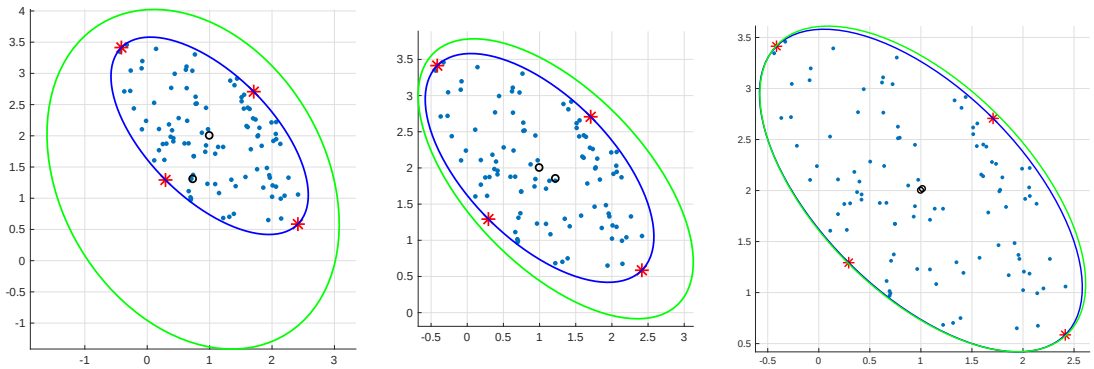


Рис. 4. Тестовое множество точек. Синим отмечена граница минимального эллипсоида

Алгоритм сходится к решению достаточно хорошо, уже после 50-ти шагов построенный эллипсоид близок к искомому.



(a) 10 шагов (b) 30 шагов (c) 50 шагов

Рис. 5. Результат работы алгоритма. Зелёным обозначен построенный эллипсоид

ЛИТЕРАТУРА

1. Кольцов М. А. *Решение задачи Сильвестра в MATLAB* // Семинар «CNSA & NDO». Избранные доклады. 26 февраля 2015 г. (`http://apmath.spbu.ru/cnsa/reps15.shtml#0226`) [Данная книга, с. 195]

2. Danzer L., Laugwitz D., Lenz H. *Uber das Lownersche Ellipsoid und sein Analogon unter den einem Eikorper einbeschriebenen Ellipsoiden* // Arch. Math, 1957. Vol. 8, No 3, pp. 214–219.

3. Шор Н. З., Стеценко С. И. *Алгоритм последовательного сжатия пространства для построения описанного эллипсоида минимального объёма* // Исследование методов решения экстремальных задач. Киев: Ин-т кибернетики им. В. М. Глушкова, 1990. С. 25–29.

ПОСТРОЕНИЕ МИНИМАЛЬНОГО ЭЛЛИПСОИДА: АЛГОРИТМ ХАЧИЯНА*

М. А. Кольцов

1°. Постановка задачи. Будем искать эллипсоид минимального объёма, который содержит все точки некоторого множества точек $a_i \in \mathbb{R}^n$, $i \in 1 : m$. Для существования решения будем предполагать, что аффинная оболочка множества $\{a_i\}$ совпадает с $\mathbb{R}^n$.

Эллипсоид будем задавать таким образом:

$$\mathcal{E} = \{x \in \mathbb{R}^n | \langle M(x-c), x-c\rangle \leqslant 1\},$$

где M — симметричная положительно-определённая матрица, c — центр эллипсоида. Объём вычисляется по формуле

$$\mathrm{vol}(\mathcal{E}) = \omega_n (\det M)^{-1/2},$$

где $\omega_n = \dfrac{\pi^{n/2}}{\Gamma(n/2+1)}$ — объём единичного шара в n-мерном пространстве.

Пренебрегая множителем ω_n и переходя к логарифму целевой функции, задачу о минимальном эллипсоиде можно формально поставить как

$$\begin{gathered} f(M,c) := -\ln\det M \to \min \\ \langle M(a_i - c), a_i - c\rangle \leqslant 1, \quad i \in 1:m, \\ M \succ 0, \end{gathered} \tag{1}$$

где запись $M \succ 0$ означает положительную определённость матрицы M.

Таким образом, неизвестными являются одновременно и центр эллипсоида c, и его матрица M. Попробуем избавиться от переменной c. Вложим точки a_i в пространство $\mathbb{R}^{n+1}$ по правилу $q_i := \begin{pmatrix} a_i \\ 1 \end{pmatrix}$. Заметим, что из предположения о существовании решения и определения аффинной независимости следует, что среди q_i найдётся набор из $n+1$ линейно независимых векторов. То есть, множество $\{q_i\}$ содержит базис $\mathbb{R}^{n+1}$.

*Семинар «CNSA & NDO». Избранные доклады. 21 апреля 2016 г.

В пространстве $\mathbb{R}^{n+1}$ будем искать минимальный эллипсоид с центром в начале, который содержит точки q_i. Обозначим матрицу этого эллипсоида через $\widetilde{B}$. Тогда новую задачу можно поставить как

$$\begin{aligned} f(\widetilde{B}) := -\ln\det\widetilde{B} &\to \min \\ \langle \widetilde{B}q_i, q_i\rangle &\leqslant 1, \quad i \in 1:m, \\ \widetilde{B} &\succ 0. \end{aligned} \tag{2}$$

Пусть найдена матрица $\widetilde{B}^*$ — решение задачи (2). Решение исходной задачи (1) будем искать как сечение эллипсоида $\left\{x|\langle\widetilde{B}^*\tilde{x},\tilde{x}\rangle \leqslant 1\right\}$ плоскостью $x_{n+1}=1$. Обозначим блоки этой матрицы через

$$\widetilde{B}^* = \begin{pmatrix} B & b \\ b^\intercal & \widehat{b} \end{pmatrix}.$$

В докладе [1] показано, что в таком случае матрица M и вектор c получаются по формулам

$$M := B/\big(1+\langle B^{-1}b,b\rangle - \widehat{b}\big), \quad c := -B^{-1}b.$$

На протяжении оставшейся части статьи будет рассматриваться только задача (2).

2°. Вспомогательные сведения. Пусть q — вектор, а M — симметричная матрица соответствующей размерности. Найдём другое представление для $q^\intercal Mq$, используя индексную технику:

$$\begin{aligned} q^\intercal Mq = \langle Mq,q\rangle = \langle\sum_{k=1}^{n} M[1:n,k]q[k],q\rangle = \sum_{k=1}^{n} q[k]\langle M[1:n,k],q\rangle = \\ = \sum_{k=1}^{n} q[k]\left(\sum_{j=1}^{n} M[j,k]q[j]\right) = \sum_{k=1}^{n}\sum_{j=1}^{n} M[j,k]\left(q[j]q[k]\right) = \\ = \sum_{k=1}^{n} M[j,k]\left(qq^\intercal\right)[j,k] = \operatorname{tr}(M\cdot qq^\intercal). \end{aligned} \tag{3}$$

Нам понадобится ещё одно равенство, называемое «лемма об определителе матрицы» в англоязычной литературе [2]. Пусть A — обратимая матрица, u, v – вектора. Тогда

$$\det(A+uv^\intercal) = (1+v^\intercal A^{-1}u)\cdot\det A. \tag{4}$$

Прежде чем проверить это равенство, заметим, что $\det(A+uv^\intercal) = \det A\cdot\det(E+(A^{-1}u)v^\intercal)$. Следовательно, достаточно проверить только случай $A=E$.

В этом случае необходимое равенство следует из свойств определителя и равенства

$$\begin{pmatrix} E & 0 \\ v^\intercal & 1 \end{pmatrix} \cdot \begin{pmatrix} E + uv^\intercal & u \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} E & 0 \\ -v^\intercal & 1 \end{pmatrix} = \begin{pmatrix} E & u \\ 0 & 1 + v^\intercal u \end{pmatrix}.$$

Действительно, определитель матрицы справа равен $1 + v^\intercal u$, а определитель средней матрицы в левой части равен $\det(E + uv^\intercal)$. Кроме того, определители остальных матриц равны 1. Значит, равенство доказано.

Рассмотрим теперь общую задачу выпуклого программирования с ограничениями:

$$\begin{aligned} f(x) &\to \min \\ h_i(x) &\leqslant 0, \quad i \in 1 : m, \\ x &\in D, \end{aligned}$$

где D – открытое выпуклое множество. Введём неотрицательные числа λ_i, соответствующие ограничениям задачи. Тогда *функцией Лагранжа* [3] этой задачи называется функция

$$\mathcal{L}(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x),$$

а числа λ_i называются *множителями Лагранжа*.

Используя функцию Лагранжа, можно записать двойственную по Лагранжу задачу:

$$\begin{aligned} g(\lambda) := \inf_{x \in D} \mathcal{L}(x, \lambda) &\to \max \\ \lambda_i &\geqslant 0, \quad i \in 1 : m. \end{aligned}$$

Для любых планов x и λ прямой и двойственной задач соответственно верно $f(x) \geqslant g(\lambda)$. Отсюда, если x^* и λ^* — решения пары двойственных задач, то

$$f(x^*) \geqslant g(\lambda^*).$$

3°. Вывод двойственной задачи для задачи о минимальном эллипсоиде. Попробуем применить двойственность Лагранжа к задаче (2). Введём множители Лагранжа $p_i \geqslant 0$, соответствующие ограничениям $\langle \widetilde{B} q_i, q_i \rangle - 1 \leqslant 0$, а ограничение $M \succ 0$ будем считать нефункциональным и включим в область определения $\mathcal{L}$. Тогда функция Лагранжа записывается как

$$\mathcal{L}(\widetilde{B}, p) := -\ln\det \widetilde{B} + \sum_{i=1}^{m} p_i \big(q_i^\intercal \widetilde{B} q_i - 1\big).$$

Или, с учётом формулы (3),

$$\mathcal{L}(\widetilde{B}, p) := -\ln\det \widetilde{B} + \sum_{i=1}^{m} p_i \operatorname{tr}(q_i q_i^\intercal \cdot \widetilde{B}) - \sum_{i=1}^{m} p_i.$$

Здесь и далее будем считать, что $\operatorname{dom}\mathcal{L} = \left\{(\widetilde{B}, p) | \widetilde{B} \succ 0, p \in \mathbb{R}^m\right\}$.

По определению целевая функция двойственной к (2) задачи равна

$$g(p) := \inf_{\widetilde{B}\succ 0} \mathcal{L}(\widetilde{B}, p) = \inf_{\widetilde{B}\succ 0} \left[-\ln\det\widetilde{B} + \sum_{i=1}^{m} p_i \operatorname{tr}(q_i q_i^\intercal \cdot \widetilde{B}) - \sum_{i=1}^{m} p_i \right].$$

Для того, чтобы упростить это выражение, найдём инфимум аналитически. Известно (см. [3]), что функции $-\ln\det\widetilde{B}$ и $\operatorname{tr}(q_i q_i^\intercal \cdot \widetilde{B})$ выпуклы по $\widetilde{B}$ на открытом множестве положительно определённых матриц. Значит, инфимум достигается в точке, где производная по $\widetilde{B}$ выражения под ним равна нулю. Возьмём эту производную:

$$\frac{\partial\mathcal{L}(\widetilde{B}, p)}{\partial\widetilde{B}} = -\widetilde{B}^{-1} + \sum_{i=1}^{m} p_i q_i q_i^\intercal = -\widetilde{B}^{-1} + QPQ^\intercal,$$

где Q — матрица, составленная из точек q_i, а $P = \operatorname{diag}(p)$. Для дальнейших рассуждений необходимо, чтобы матрица $QPQ^\intercal$ была положительно определена. Проверим, так ли это.

Зафиксируем произвольный ненулевой вектор $x \in \mathbb{R}^{n+1}$. По определению, для положительной определённости $QPQ^\intercal$ необходимо, чтобы $\langle QPQ^\intercal x, x\rangle > 0$. Распишем скалярное произведение:

$$\langle \sum_{i=0}^{m} p_i q_i q_i^\intercal \cdot x, x\rangle = \sum_{i=1}^{m} p_i \langle q_i q_i^\intercal \cdot x, x\rangle = \sum_{i=1}^{m} p_i \langle q_i, x\rangle^2.$$

Так как среди q_i содержится базис $\mathbb{R}^{n+1}$, то вектор x не может быть ортогонален сразу всем q_i. Значит, если для всех i $p_i > 0$, то матрица $QPQ^\intercal$ заведомо положительно определена. На самом деле, достаточно, чтобы числа p_i были положительны на индексах i, соответствующих точкам базиса. Будем теперь рассматривать только такие p, что $QPQ^\intercal \succ 0$. В следующей части будет описан алгоритм, на каждом шаге которого выполнено условие $p > 0$.

Таким образом, единственным решением уравнения $\dfrac{\partial\mathcal{L}(\widetilde{B}, p)}{\partial\widetilde{B}} = 0$ при фиксированном p является матрица $\widetilde{B} = (QPQ^\intercal)^{-1}$. Подставив это значение в $\mathcal{L}(\widetilde{B}, p)$, получим выражение для целевой функции двойственной задачи $g(p)$:

$$g(p) = \ln\det QPQ^\intercal + \sum_{i=1}^{m} p_i \operatorname{tr}\left[q_i q_i^\intercal \cdot (QPQ^\intercal)^{-1}\right] - \sum_{i=1}^{m} p_i.$$

Обозначим $A = QPQ^\intercal$ и распишем выражение под первой суммой:

$$\sum_{i=1}^{m} p_i \operatorname{tr}\left[q_i q_i^\intercal \cdot (QPQ^\intercal)^{-1}\right] = \operatorname{tr}\left(\sum_{i=1}^{m} p_i q_i q_i^\intercal\right) A^{-1} = \operatorname{tr} AA^{-1} = n + 1. \qquad (5)$$

В итоге имеем

$$g(p) = \ln\det QPQ^\intercal - \sum_{i=1}^{m} p_i + n + 1. \tag{6}$$

Значит, двойственная задача имеет вид

$$\begin{aligned} g(p) := \ln\det QPQ^\intercal - \sum_{i=1}^{m} p_i + n + 1 &\to \max \\ p_i \geqslant 0, \quad i \in 1:m. & \end{aligned} \tag{7}$$

Чтобы ещё упростить вид целевой функции, запишем для этой задачи условия Куна–Таккера. Чтобы сделать это, посчитаем сначала производную от $\ln\det QPQ^\intercal$, аккуратно применив правило цепочки:

$$\begin{aligned} \frac{\partial}{\partial p_i}\ln\det QPQ^\intercal = \operatorname{tr}\left(\frac{\partial \ln\det QPQ^\intercal}{\partial QPQ^\intercal}\cdot\frac{\partial QPQ^\intercal}{\partial p_i}\right) = \\ = \operatorname{tr}\left((QPQ^\intercal)^{-1}\cdot q_i q_i^\intercal\right) = q_i^\intercal (QPQ^\intercal)^{-1} q_i. \end{aligned}$$

Следовательно, условия Куна–Таккера для (7) с множителями λ_i, соответствующими ограничениям $p_i \geqslant 0$, выглядят как

$$\begin{aligned} q_i^\intercal (QPQ^\intercal)^{-1} q_i - 1 + \lambda_i &= 0, \\ \lambda_i p_i &= 0, \\ \lambda_i &\geqslant 0. \end{aligned}$$

Избавимся от переменных λ_i:

$$q_i^\intercal (QPQ^\intercal)^{-1} q_i \leqslant 1, \tag{8}$$

$$p_i\left(1 - q_i^\intercal (QPQ^\intercal)^{-1} q_i\right) = 0. \tag{9}$$

Просуммировав (9) по i и воспользовавшись ещё раз формулой (5), получаем

$$\sum_{i=1}^{m} p_i = n + 1.$$

Теперь двойственную задачу можно записать как

$$\begin{aligned} g(p) := \ln\det QPQ^\intercal &\to \max \\ 1^\intercal p &= n + 1, \\ p &\geqslant 0. \end{aligned}$$

С помощью замены переменных $u := \dfrac{p}{n+1}$, $U := \operatorname{diag}(u) = \dfrac{1}{n+1}P$ получаем итоговый вид двойственной задачи

$$\begin{aligned} \widehat{g}(u) := \ln\det QUQ^\intercal &\to \max \\ 1^\intercal u &= 1, \\ u &\geqslant 0. \end{aligned} \tag{10}$$

Таким образом, мы свели нахождение минимального эллипсоида к максимизации функции на стандартном симплексе в $\mathbb{R}^m$.

Пусть найдено решение двойственной задачи u^*. Из того, что в точке u^* достигается максимум функции $\widehat{g}$, следует, что $\widehat{g}(u^*) > -\infty$, а значит $QU^*Q^\intercal$ положительно определена. Положим $p^* = (n+1)u^*$, $\widetilde{B}^* = (n+1)\cdot(QU^*Q^\intercal)^{-1}$. В силу (8) эта матрица будет удовлетворять ограничениям задачи (2). Очевидно, что выполняется равенство $f(\widetilde{B}^*) = g(p^*)$. В силу неравенства между целевыми функциями пары двойственных задач, для любой матрицы $\widetilde{B}$, удовлетворяющей ограничениям, верно $f(\widetilde{B}) \geqslant g(p^*) = f(\widetilde{B}^*)$, следовательно, $\widetilde{B}^*$ — решение задачи (2).

4°. Алгоритм Хачияна. В статье [4] описан простой алгоритм, решающий двойственную задачу (10). Приведём его вывод.

Во-первых, заметим, что формула (8) для изменённой задачи (10) имеет вид

$$q_i^\intercal (QUQ^\intercal)^{-1} q_i \leqslant n+1.$$

Как было вычислено выше, $\widehat{g}_j(u) := \dfrac{\partial}{\partial u_j}\widehat{g}(u) = q_j^\intercal (QUQ^\intercal)^{-1} q_j$. Значит, если u^* — решение, то

$$\widehat{g}_j(u^*) \leqslant n+1, \quad \forall j \in 1:m. \tag{11}$$

Пусть на k-м шаге имеется вектор u_k. Найдём, с какой ошибкой u_k удовлетворяет условию (11). То есть, найдём такое минимальное число ε, что

$$\widehat{g}_j(u_k) \leqslant (1+\varepsilon)(n+1), \quad \forall j \in 1:m.$$

Пусть $R \subset 1:m$ — множество индексов, на которых неравенство обращается в равенство. Тогда для $j \notin R$

$$\widehat{g}_j(u_k) < (1+\varepsilon)(n+1) = \widehat{g}_r(u_k), \quad \forall r \in R.$$

Выберем какое-нибудь $r \in R$. Из формулы видно, что $\widehat{g}_r(u_k) = \max\bigl\{\widehat{g}_j(u_k) \mid j \in 1:m\bigr\}$ и

$$\varepsilon = \frac{\widehat{g}_r(u_k) - (n+1)}{n+1}.$$

Хачиян предлагает следующее: в качестве u_{k+1} выбрать точку из отрезка $[u_k, e_r]$, где e_r — r-й орт $\mathbb{R}^m$, максимизирующую $\widehat{g}(u_{k+1})$. То есть, выбрать число $\alpha \in (0,1)$ и положить

$$u_{k+1} = (1-\alpha)u_k + \alpha e_r.$$

Покажем, как найти такое α. Для этого распишем $\widehat{g}(u_{k+1})$:

$$\begin{aligned}\widehat{g}(u_{k+1}) &= \ln\det\left[Q\left((1-\alpha)U_k + \alpha E_r\right)Q^\intercal\right] = \ln\det\left[(1-\alpha)QU_kQ^\intercal + \alpha q_r q_r^\intercal\right] = \\ &= \ln\left[(1-\alpha)^{n+1}\det\left(QU_kQ^\intercal + \frac{\alpha}{1-\alpha}q_rq_r^\intercal\right)\right].\end{aligned}$$

Применим формулу (4):

$$\det\left(QU_kQ^\intercal + \frac{\alpha}{1-\alpha}q_rq_r^\intercal\right) = \left(1 + \frac{\alpha}{1-\alpha}q_r^\intercal(QU_kQ^\intercal)^{-1}q_r\right)\det QU_kQ^\intercal.$$

Таким образом,

$$\begin{aligned}\widehat{g}(u_{k+1}) &= \ln\left[(1-\alpha)^{n+1}\left(1 + \frac{\alpha}{1-\alpha}q_r^\intercal(QU_kQ^\intercal)^{-1}q_r\right)\det QU_kQ^\intercal\right] = \\ &= \ln\left[(1-\alpha)^n\big(1+\alpha(\widehat{g}_r(u_k)-1)\big)\det QU_kQ^\intercal\right] = \\ &= n\ln(1-\alpha) + \ln\big(1+\alpha(\widehat{g}_r(u_k)-1)\big) + \ln\det QU_kQ^\intercal.\end{aligned}$$

Возьмём производную по α:

$$\frac{\mathrm{d}\widehat{g}(u_{k+1})}{\mathrm{d}\alpha} = -\frac{n}{1-\alpha} + \frac{\widehat{g}_r(u_k)-1}{1+\alpha(\widehat{g}_r(u_k)-1)}.$$

Приравнивая нулю и решая относительно α, получаем единственное решение

$$\alpha = \frac{\widehat{g}_r(u_k)-(n+1)}{(n+1)(\widehat{g}_r(u_k)-1)}.$$

Так как $\widehat{g}_r(u_k) = (1+\varepsilon)(n+1) > n+1 > 1$, то $\alpha > 0$. Кроме того,

$$\alpha < 1 \Leftrightarrow \widehat{g}_r(u_k) - n - 1 < n\cdot\widehat{g}_r(u_k) - n + \widehat{g}_r(u_k) - 1 \Leftrightarrow \widehat{g}_r(u_k) > 0.$$

Значит, α действительно лежит в $(0,1)$. Так как $\widehat{g}(u_{k+1})$ выпукла по α, u_{k+1} — искомая точка максимума на отрезке $(0,1)$.

В качестве начального приближения u_0 можно взять вектор $(1,\ldots,1)/m$. Так как $u_0 > 0$, то $QU_0Q^\intercal \succ 0$, и целевая функция $\widehat{g}(u_0)$ корректно определена. Кроме того, на каждом шаге $\alpha_k < 1$, значит, если $u_k > 0$, то и $u_{k+1} > 0$. Таким образом, алгоритм корректный.

Опишем ещё раз последовательность действий на k-м шаге алгоритма:

- Вычислить $\nabla\widehat{g}(u_k)$.
- Найти максимальную компоненту градиента $\widehat{g}_r(u_k)$.
- Вычислить $\varepsilon = \dfrac{\widehat{g}_r(u_k) - (n+1)}{n+1}$.
- Если ε меньше требуемой погрешности ε_0, закончить вычисления. В противном случае, вычислить $\alpha := \dfrac{\widehat{g}_r(u_k) - (n+1)}{(n+1)(\widehat{g}_r(u_k) - 1)}$.
- Положить $u_{k+1} := (1-\alpha)u_k + \alpha e_r$ и перейти на следующий шаг.

Кроме того, заметим, что на каждом шаге матрица $QUQ^\intercal$ умножается на $1-\alpha$ и к ней прибавляется одноранговая матрица $\alpha q_r q_r^\intercal$. В таком случае нет необходимости каждый раз заново вычислять обратную матрицу, а можно применить формулу обновления (см. [4]):

$$(QU_{k+1}Q^\intercal)^{-1} = \left(1 + \frac{\varepsilon}{n\cdot(1+\varepsilon)}\right)(QU_kQ^\intercal)^{-1} - \frac{\varepsilon}{n\cdot(1+\varepsilon)^2} b_k b_k^\intercal,$$

где $b_k := (QU_kQ^\intercal)^{-1} q_r$.

Были проведены численные эксперименты при $\varepsilon_0 = 10^{-5}$ и различных размерностях основного пространства n и различном количестве двойственных переменных m (то есть количестве точек q_i). Результаты приведены в табл. 1. В столбце «точность» содержится относительная ошибка δ, то есть

$$\delta = \frac{\left|(\det M)^{-1/2} - (\det M^*)^{-1/2}\right|}{(\det M^*)^{-1/2}},$$

где M^* — известная матрица минимального эллипсоида.

Таблица 1. Численные результаты алгоритма Хачияна

n	m	число итераций	точность	время (с)
2	104	199993	$7\cdot10^{-6}$	17.2
2	504	199994	$7\cdot10^{-6}$	78.8
5	510	499990	$1.4\cdot10^{-5}$	261.3

Если в алгоритме критерий останова по ε (по точности выполнения ограничений) заменить на останов по внутренней сходимости, то число итераций и время работы сократится, но относительная точность не изменится. Отметим, что даже в случае с четырьмя точками в $\mathbb{R}^2$, одна из которых лежит внутри треугольника, образованного тремя другими, алгоритм делает порядка 80 тысяч шагов и работает около 4 секунд.

5°. Комбинированный алгоритм. Как было продемонстрировано в предыдущем разделе, применение одного только алгоритма Хачияна не приносит желаемых результатов из-за низкой точности и большого времени работы. Поэтому были предприняты попытки изменить алгоритм так, чтобы он работал быстрее и точнее.

В ходе численных экспериментов была обнаружена следующая закономерность — в начале своей работы алгоритм делает достаточно большие шаги α, быстро получая эллипсоид, близкий к искомому. Затем же алгоритм начинает делать очень короткие шаги, с $\alpha \approx 10^{-5}$, постепенно увеличивая эллипсоид и приближаясь к решению. Это связано с тем, что на каждом шаге алгоритм производит изменение в направлении только одной координаты, как бы «подтягивая» эллипсоид к точке, которая дальше всего от него находится. Таким образом, если текущий эллипсоид уже достаточно близок к оптимальному, то алгоритму приходится поочерёдно немного подтягивать его к точкам, лежащим вне.

На основе этого наблюдения и статьи [5] возникла идея улучшения алгоритма. Во-первых, в 1948 году Джон доказал, что минимальный эллипсоид определяется не более $(n^2+3n)/2$ точками, где n — размерность пространства. То есть, в нашем случае только $l = \big((n+1)^2 + 3(n+1)\big)/2$ точек из исходных m определяют искомый эллипсоид. Кроме того, из формулы (8) и её эквивалента для вектора u видно, что те точки a_i, для которых $\widehat{g}_i(u) \geqslant n+1$, лежат на границе или вне эллипсоида, задаваемого u. Значит, если их количество не превосходит l, можно предположить, что только эти точки и задают эллипсоид, который мы ищем. Будем дальше работать уже только с ними. Для корректности задания целевой функции на этих точках необходимо потребовать, чтобы они так же содержали аффинно независимое множество. Таким образом, в «активное» множество всегда будет входить не более l и не менее $n+1$ точек.

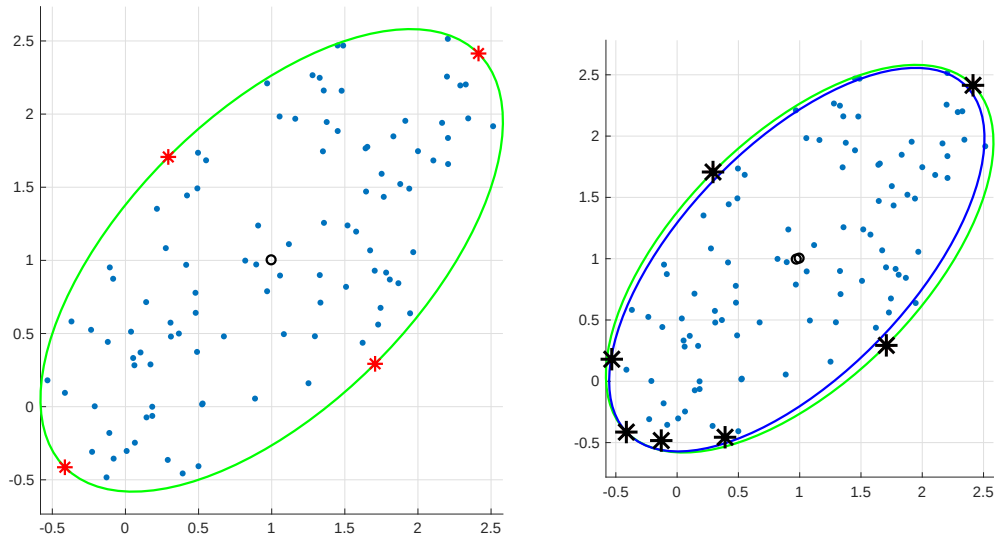
Во-вторых, раз увеличение эллипсоида в направлении только одной точки приводит к очень маленьким шагам, разумно попытаться увеличивать его сразу в нескольких направлениях. В первую очередь для этого был испытан обычный градиентный метод наискорейшего подъёма. В процедуре линейного поиска использовано условие Армихо. В качестве направления подъёма выбирается направление градиента, спроецированное на стандартный симплекс. Кроме того, линейный поиск останавливается при достижении какой-либо компонентой вектора нуля. Таким образом обеспечивается выполнение ограничений двойственной задачи на каждом шаге градиентного метода. Критерий останова используется тот же, что и в алгоритме Хачияна.

На рис. наглядно представлена работа такого комбинированного подхода, а подробные результаты его применения приведены в табл. 2.

Таблица 2. Численные результаты комбинированного алгоритма

n	m	алгоритм Хачияна	активные точки	градиентный метод	точность	время (с)
2	104	18	7	40	$2 \cdot 10^{-9}$	0.3
2	504	123	9	85	$8 \cdot 10^{-8}$	0.7
5	510	92	26	87	$2 \cdot 10^{-7}$	0.8
10	1020	80	74	220	$2 \cdot 10^{-6}$	2.4

Здесь столбцы «алгоритм Хачияна» и «градиентный метод» содержат количество шагов соответствующих алгоритмов, в столбце «активные точки» содержится число точек в активном множестве после остановки алгоритма Хачияна, а столбец «точность» имеет тот же смысл, что и в предыдущей таблице. Для вычислений использовалась погрешность $\varepsilon_0 = 10^{-5}$.



(a) Тестовое множество и известный минимальный эллипсоид

(b) Активное множество и построенный алгоритмом Хачияна эллипсоид

Рис. Иллюстрация работы комбинированного алгоритмапри $n = 2$, $m = 104$

К сожалению, в некоторых случаях алгоритм зацикливается или выдает неверный ответ. Такое происходит, например, при $n = 30$, $m = 560$ — в этом случае градиентный метод начиная с 1890-го шага начинает делать циклические шаги, не имея возможности никак остановиться. При этом достигается относительная точность только 10^{-4}. Это может быть связано с недостатками процедуры линейного поиска, а также с общей неприменимостью простых градиентных методов вблизи решения.

ЛИТЕРАТУРА

1. Кольцов М.А. *Построение минимальноˊго эллипсоида: Алгоритм Шора* // Семинар «CNSA & NDO». Избранные доклады. 14 мая 2015 г. (`http://apmath.spbu.ru/cnsa/reps15.shtml#0514`) [Данная книга, с. 297]

2. Brookes, M. *The Matrix Reference Manual*, 2011 (электронный ресурс), (`http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html`).

3. Boyd S., Vandenberghe L. *Convex optimization*. Cambridge University Press, 2004.

4. Khachiyan, L. G. *Rounding of Polytopes in the Real Number Model of Computation* // Mathematics of Operations Research, 1996, vol. 21, no. 2, pp. 307–320.

5. Sun P, Freund R. *Computation of Minimum-Volume Covering Ellipsoids* // Operations Research, 2004, vol. 52, no. 5, pp. 690–706.

НАХОЖДЕНИЕ СТАЦИОНАРНЫХ ТОЧЕК В ЗАДАЧАХ БЕЗУСЛОВНОЙ ОПТИМИЗАЦИИ В MATLAB*

М. А. Кольцов, А. В. Плоткин

1°. Постановка задачи. Рассмотрим *задачу безусловной оптимизации*

$$f(x) \to \operatorname{extr}_{x\in\mathbb{R}^n},$$

где $f \in C^1(\mathbb{R}^n)$. Запишем необходимое условие экстремума в точке x_*:

$$g(x_*) := f'(x_*) = \mathbb{O}. \tag{1}$$

Все точки из $\mathbb{R}^n$, удовлетворяющие условию (1), называются *стационарными*. Система уравнений $g(x) = \mathbb{O}$, в общем случае, является нелинейной. Решению таких систем с помощью `MATLAB` и посвящён данный доклад.

2°. Функция fsolve. В среде `MATLAB` имеется функция `fsolve`, предназначенная для решения систем нелинейных уравнений вида $F(x) = \mathbb{O}$, $x \in \mathbb{R}^n$. Решение системы происходит путём минимизации невязки $G(x) = ||F(x)||^2$.

Функция `fsolve` имеет следующий формат вызова:

```
[xval, fval, exitflag] = fsolve(fun, x0, options)
```

Поясним входные параметры функции `fsolve`:

- `fun` — указатель на функцию $F(x)$, задающую левую часть решаемой системы уравнений $\big($в нашем случае — градиент $g(x)\big)$,
- `x0` — начальное приближение к решению,
- `options` — параметр, устанавливающий некоторые дополнительные настройки, в частности, алгоритм решения.

Значение параметра `options` задаётся с помощью стандартной функции `optimoptions` следующим образом:

```
options =
optimoptions(@fsolve, Опция 1, Значение 1, Опция 2, Значение 2, ...)
```

*Семинар «CNSA & NDO». Избранные доклады. 24 ноября 2016 г.

Приведём основные доступные настройки:

- ▷ `Algorithm` — используемый алгоритм решения. Возможные значения: «trust-region-dogleg» (по умолчанию), «trust-region», «levenberg-marquardt» (подробнее о данных алгоритмах см. [1, с. 730–731]).
- ▷ `OptimalityTolerance` — мера оптимальности точки (по умолчанию 10^{-6}). Если $||G'(x_k)||_\infty <$ `OptimalityTolerance`, то вычисления прекращаются.
- ▷ `StepTolerance` — минимальный допустимый размер шага (по умолчанию 10^{-6}).
- ▷ `FunctionTolerance` — минимальное допустимое уменьшение невязки $G(x)$ за одну итерацию (по умолчанию 10^{-6}).
- ▷ `MaxIterations` — максимальное число итераций (по умолчанию 400).
- ▷ `MaxFunctionEvaluations` — допустимое количество вычислений функции F (по умолчанию $100 \cdot n$).
- ▷ `SpecifyObjectiveGradient` — использовать или нет (`true` или `false`) второй выходной параметр функции `fun` — матрицу Якоби функции F в точке x. Если `false` (по умолчанию), то матрица приближается конечными разностями. В нашем случае матрицей Якоби является матрица вторых производных функции $f(x)$.
- ▷ `OutputFcn` — указатель на дополнительную функцию, которая будет выполняться после каждой итерации алгоритма.

Полный список доступных настроек приведён в руководстве [1, с. 722–728].

Перейдём к описанию выходных параметров функции `fsolve`:

- `xval` — найденное решение,
- `fval` — значение $F($`xval`$)$,
- `exitflag` — значение, сигнализирующее об окончании вычислений.

Поясним возможные значения r параметра `exitflag`:
$r > 0$ — конечное значение невязки мало $\left(G(\texttt{xval}) < \sqrt{\texttt{FunctionTolerance}}\right)$,
$r = 0$ — превышено максимальное число итераций или количество вычислений функции F (см. `MaxIterations` и `MaxFunctionEvaluations`),
$r < 0$ — конечное значение невязки велико $\left(G(\texttt{xval}) \geqslant \sqrt{\texttt{FunctionTolerance}}\right)$.

3°. Информация о процессе вычислений. Из описания выходных параметров функции ясно, что `fsolve` предоставляет информацию только о конечном значении x и $F(x)$. Однако зачастую бывает интересна информация о всём процессе вычислений (к примеру, для построения графиков). Для получения данной информации может быть использована опция `OutputFcn`.

Функция, указатель на которую передаётся в опцию `OutputFcn`, должна принимать от `fsolve` три аргумента:

- `x` — текущее значение x,
- `optimValues` — структура, содержащая информацию о текущей итерации,
- `state` — стадия выполнения `fsolve`.

Функция может вернуть «`true`», если требуется прекратить вычисления, и «`false`» в противном случае.

Приведём пример простейшей реализации такой функции:

```
function stop = outfun(x, optimValues, state)
  stop = false;                             % не прерываем вычисления
  if isequal(state, 'iter')                 % если конец итерации
    xhist = [xhist; x];                     % сохраняем значение x
    ghist = [ghist; optimValues.fval'];     % сохраняем значение g(x)
  end
end
```

По окончании работы `fsolve` в переменных `xhist`, `ghist` будут содержаться значения x и $g(x)$ на каждой итерации.

Перечень доступной в структуре `optimValues` информации, а также примеры более сложных функций приведены в руководстве [1, с. 166–172, 568–577].

4°. Пример 1: обобщённая трёхдиагональная функция. Обратимся к банку тестовых функций [2] и опробуем `fsolve` на обобщённой трёхдиагональной функции (Generalized Tridiagonal function):

$$f(x) = \sum_{i=1}^{n-1} (x_i + x_{i+1} - 3)^2 + (x_i - x_{i+1} + 1)^4.$$

Запишем градиент данной функции:

$$\begin{aligned}
\frac{\partial f}{\partial x_1} &= 2\,(x_1 + x_2 - 3) + 4\,(x_1 - x_2 + 1)^3, \\
\frac{\partial f}{\partial x_i} &= 2\,(x_i + x_{i+1} - 3) + 4\,(x_i - x_{i+1} + 1)^3 + \\
&\quad + 2\,(x_{i-1} + x_i - 3) - 4\,(x_{i-1} - x_i + 1)^3, \quad i \in 2 : n-1, \\
\frac{\partial f}{\partial x_n} &= 2\,(x_{n-1} + x_n - 3) - 4\,(x_{n-1} - x_n + 1)^3.
\end{aligned}$$

Пусть $n = 5$. Банк тестовых функций предлагает в качестве начального приближения $x_0 = [2, \ldots, 2]$. Запустим `fsolve` с этим начальным приближением и после 6 итераций получим ответ: $x_* \approx [1.037, 1.370, 1.500, 1.630, 1.963]$, $f(x_*) \approx 2.27875$ (см. рис. 1).

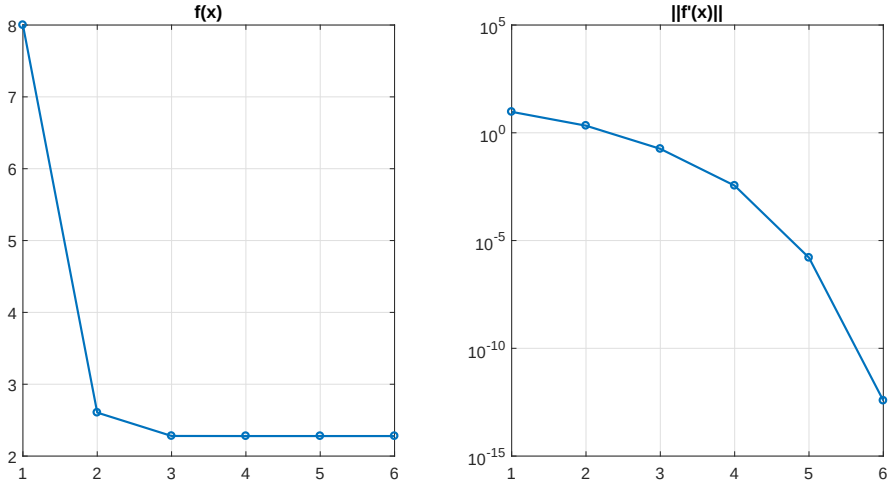


Рис. 1. Поведение обобщённой трёхдиагональной функции по итерациям

5°. Пример 2: функция EG2. Рассмотрим теперь функцию EG2 из того же банка тестовых функций:

$$f(x) = \sum_{i=1}^{n-1} \sin(x_1 + x_i^2 - 1) + \frac{1}{2} \sin x_n^2.$$

Запишем градиент данной функции:

$$\frac{\partial f}{\partial x_1} = (1+2x_1)\cos(x_1+x_1^2-1) + \sum_{i=2}^{n-1}\cos(x_1+x_i^2-1),$$
$$\frac{\partial f}{\partial x_i} = 2\,x_i\cos(x_1+x_i^2-1), \quad i \in 2:n-1,$$
$$\frac{\partial f}{\partial x_n} = x_n\cos x_n^2.$$

Также вычислим матрицу вторых производных:

$$\frac{\partial^2 f}{\partial x_1^2} = -(1+2x_1)^2\sin(x_1+x_1^2-1) + 2\,\cos(x_1+x_1^2-1) - \sum_{i=2}^{n-1}\sin(x_1+x_i^2-1),$$
$$\frac{\partial^2 f}{\partial x_i^2} = -4\,x_i^2\sin(x_1+x_i^2-1) + 2\,\cos(x_1+x_i^2-1), \quad i \in 2:n-1,$$
$$\frac{\partial^2 f}{\partial x_n^2} = -2\,x_n^2\sin x_n^2 + \cos x_n^2,$$
$$\frac{\partial^2 f}{\partial x_1 \partial x_i} = -2\,x_i\sin(x_1+x_i^2-1), \quad i \in 2:n-1.$$

Пусть $n = 5$. Банк тестовых функций предлагает в качестве начального приближения $x_0 = [1,\ldots,1]$. Запустим `fsolve` с этим начальным приближением и следующими настройками: `SpecifyObjectiveGradient = true`, `OptimalityTolerance` $= 10^{-12}$. После 7 итераций получим ответ: $x_* \approx$ $\approx [1.180, 1.180, 1.180, 1.180, 1.253]$, $f(x_*) \approx 4.5$ (см. рис. 2).

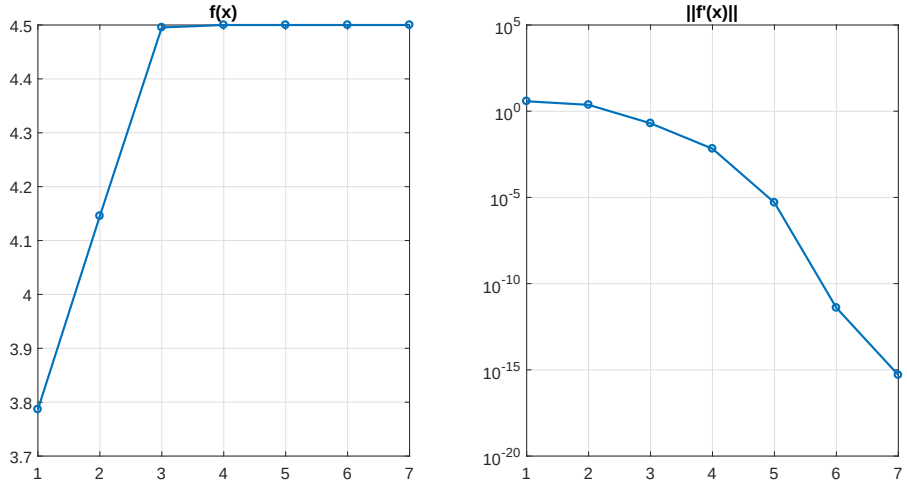


Рис. 2. Поведение функции `EG2` по итерациям

Найденное значение x_* является точкой *локального максимума* (матрица вторых производных отрицательно определена в этой точке).

Интересно заметить, что задача *минимизации* функции `EG2` может быть решена аналитически (решение представлено Даниилом Быковым). Действительно, функция является суммой синусов. Если бы в какой-то точке все синусы принимали значение -1, то эта точка была бы точкой глобального минимума. Попробуем этого добиться.

Во-первых, ясно, что надо положить $x_n = \sqrt{\dfrac{3\pi}{2}}$, так как x_n не выходит в другие слагаемые. Далее, минимизируем первое слагаемое:

$$\begin{aligned}
&\sin(x_1 + x_1^2 - 1) = -1,\\
&x_1 + x_1^2 - 1 = \frac{3\pi}{2},\\
&x_1 = \frac{-1 \pm \sqrt{5 + 6\pi}}{2}.
\end{aligned}$$

Теперь можно найти остальные переменные:

$$x_i = \pm\sqrt{1 + \frac{3\pi}{2} - x_1}, \quad i \in 2 : n-1.$$

Значение целевой функции в найденной точке равно $-n + \dfrac{1}{2}$.

6°. Пример 3: обобщённая функция Розенброка. Рассмотрим теперь обобщённую функцию Розенброка (Generalized Rosenbrock function):

$$f(x) = \sum_{i=1}^{n-1} \left[100\,(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right].$$

Запишем градиент данной функции:

$$\begin{aligned}
\frac{\partial f}{\partial x_1} &= -400\,x_1(x_2 - x_1^2) - 2\,(1 - x_1),\\
\frac{\partial f}{\partial x_i} &= -400\,x_i(x_{i+1} - x_i^2) - 2\,(1 - x_i) + 200\,(x_i - x_{i-1}^2), \quad i \in 2 : n-1,\\
\frac{\partial f}{\partial x_n} &= 200\,(x_n - x_{n-1}^2).
\end{aligned}$$

Частным случаем (при $n = 2$) является обычная функция Розенброка, которая подробно изучена. Опробуем `fsolve` сначала на этой двумерной функции. В качестве начального приближения возьмём $x_0 = [-1.2, 1]$. В ответ получим следующее сообщение от `MATLAB`:

```
fsolve stopped because it exceeded the function evaluation limit,
options.MaxFunctionEvaluations = 200 (the default value).
```

Поведение функции Розенброка по итерациям приведено на рис. 3 и рис. 4.

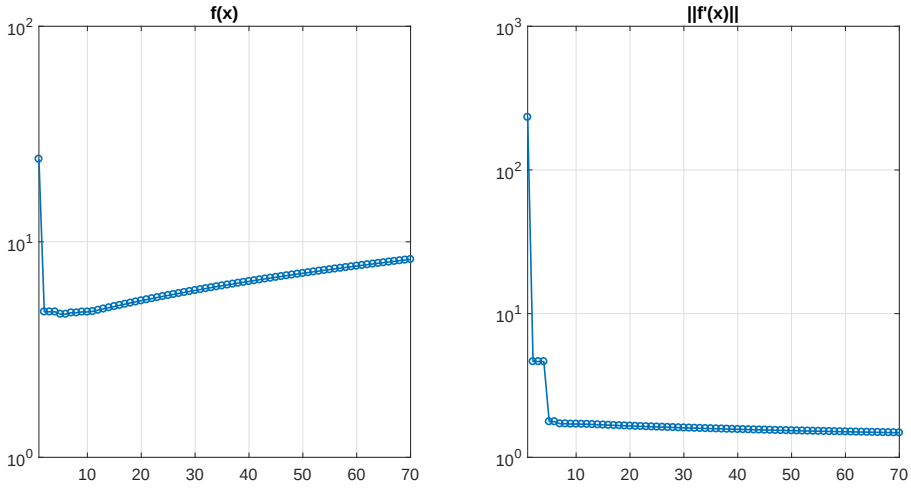


Рис. 3. Поведение функции Розенброка по итерациям при $x_0 = [-1.2, 1]$

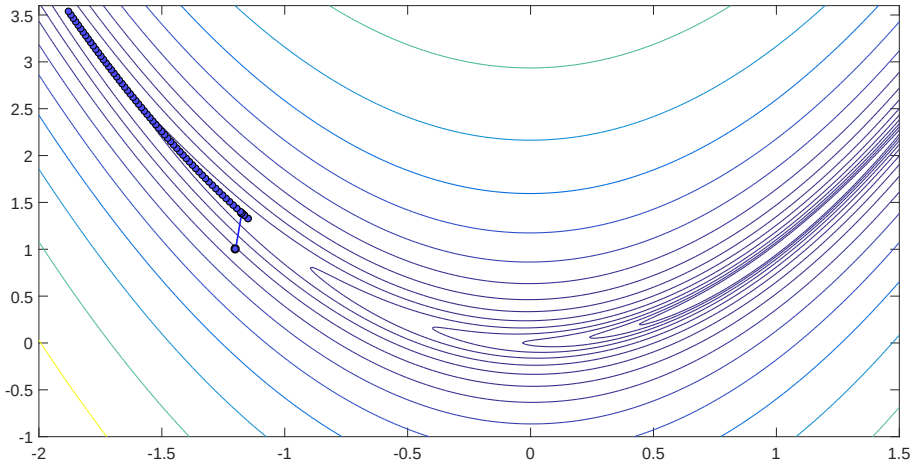


Рис. 4. Поведение аргумента по итерациям при $x_0 = [-1.2, 1]$

Увеличение параметра `MaxFunctionEvaluations` в данном случае не приводит к успеху.

Попробуем другое начальное приближение: $x_0 = [-0.5, 1.5]$. После 29 итераций получим ответ: $x_* \approx [1, 1]$, $f(x_*) \approx 0$ (см. рис. 5 и рис. 6).

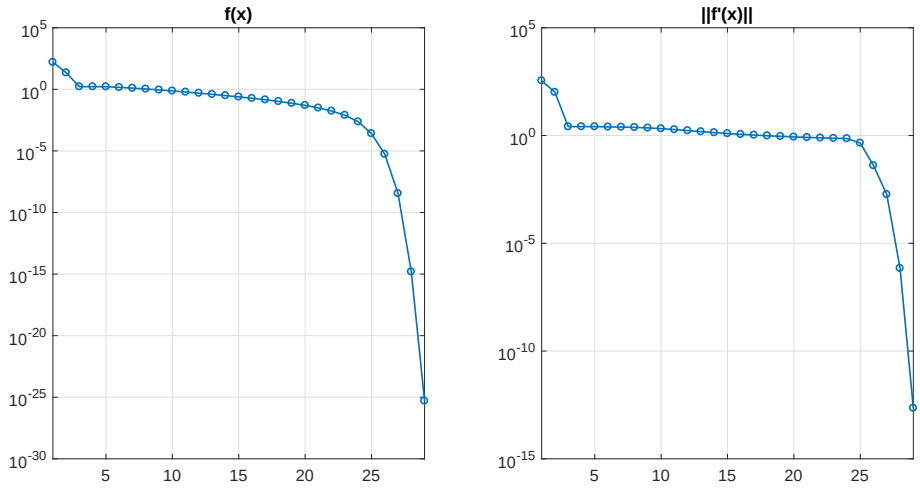


Рис. 5. Поведение функции Розенброка по итерациям при $x_0 = [-0.5, 1.5]$

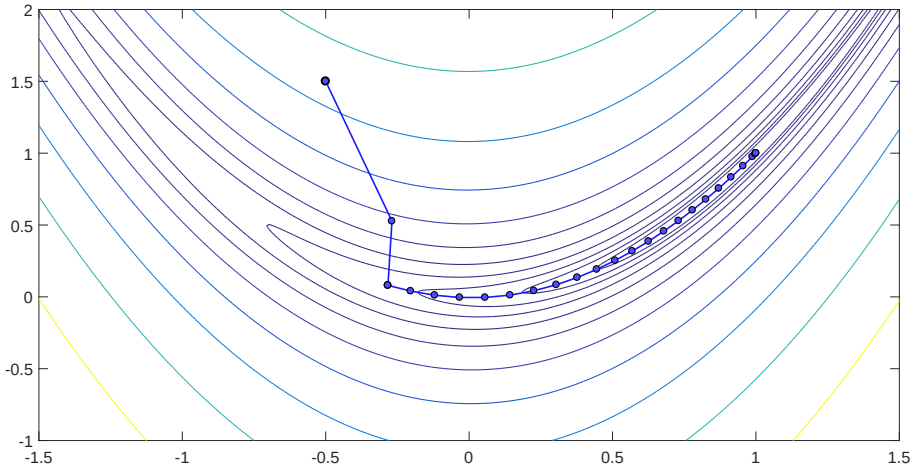


Рис. 6. Поведение аргумента по итерациям при $x_0 = [-0.5, 1.5]$

Найденное значение x_* является точкой глобального минимума функции Розенброка.

Вернёмся к обобщённой функции. Пусть $n = 5$. Банк тестовых функций предлагает в качестве начального приближения $x_0 = [-1.2, 1, -1.2, 1, -1.2]$.

Запустим `fsolve` с этим начальным приближением и после 43 итераций получим ответ: $x_* \approx [-0.962, 0.936, 0.881, 0.778, 0.605]$, $f(x_*) \approx 3.93084$ (см. рис. 7).

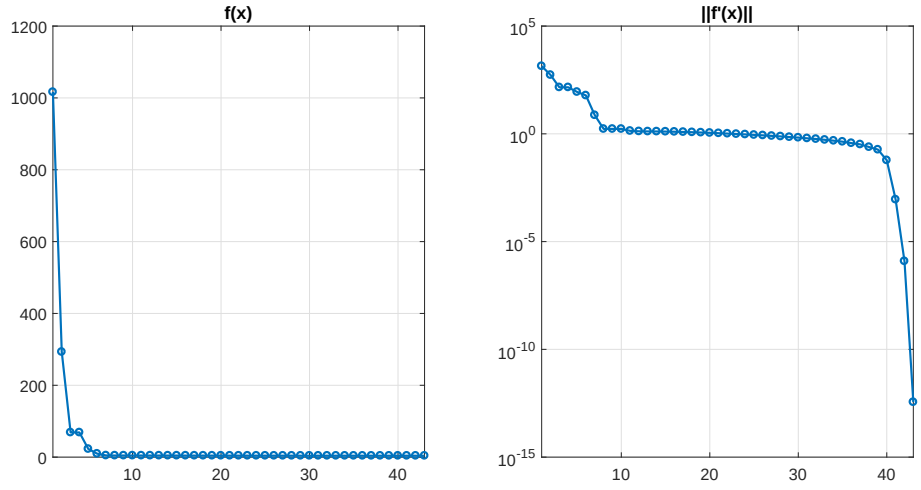


Рис. 7. Поведение обобщённой функции Розенброка по итерациям

Подробное исследование обобщённой функции Розенброка проведено в статье [3].

7°. Если `fsolve` не находит решение. Приведём несколько рекомендаций на случай, когда функция `fsolve` не смогла найти решение.

1) Попробуйте изменить начальное приближение. Задавая различные начальные приближения, вы увеличиваете шанс на успех.

2) Проверьте, является ли функция $F(x)$ гладкой — `fsolve` может плохо сходиться для функций, которые не являются гладкими.

3) Попробуйте изменить настройки точности, в особенности это касается параметров `OptimalityTolerance` и `StepTolerance`.

ЛИТЕРАТУРА

1. The MathWorks. *Optimization Toolbox User's Guide R2016b.* (`http://www.mathworks.com/help/pdf\_doc/optim/optim\_tb.pdf`)

2. Andrei N. *An unconstrained optimization test functions collection.* Advanced Modeling and Optimization. 2008. Vol. 10, No. 1, pp. 147–161.

3. Kok S., Sandrock C. *Locating and characterizing the stationary points of the extended rosenbrock function.* Evolutionary Computation. 2009. Vol. 17, No. 1, pp. 437–453.

ГЛАВА 4. ВАРИАЦИОННЫЕ ЗАДАЧИ

КВАДРАТИЧНЫЕ ВАРИАЦИОННЫЕ ЗАДАЧИ*

В. Н. Малозёмов

Аннотация. Эта статья написана на основе трёх лекций, которые автор прочитал в университете Калабрии (Италия) в мае 1999 г. по приглашению проф. Манлио Гаудиозо. Статья представляет собой нестандартное введение в вариационное исчисление, математически строгое и согласованное в идейном плане с теорией конечномерных экстремальных задач.

1°. Постановка задачи. Критерий оптимальности

1.1. Обозначим через $C[a,b]$ и $C^1[a,b]$ линейные пространства непрерывных и непрерывно дифференцируемых на отрезке $[a,b]$ функций соответственно. Зафиксируем функции p, q, f из $C[a,b]$ и определим на $C^1[a,b]$ *интегральный квадратичный функционал в канонической форме*:

$$Q(x) = \int_a^b \Big\{ p(t)\big[x'(t)\big]^2 + q(t)\big[x(t)\big]^2 - 2f(t)x(t) \Big\} dt.$$

Нас интересует квадратичная вариационная задача следующего вида:

$$\begin{gathered} Q(x) \to \inf, \\ x(a) = A, \quad x(b) = B, \quad x \in C^1[a,b]. \end{gathered} \tag{1}$$

Функцию x, удовлетворяющую ограничениям задачи (1), назовём её *планом*. Множество планов обозначим через Ω. Решение x_* задачи (1) характеризуется тем, что $x_* \in \Omega$ и $Q(x) \geqslant Q(x_*)$ для всех $x \in \Omega$. Решение будем называть также *оптимальным планом*.

1.2. Введём *множество допустимых вариаций*

$$C_0^1[a,b] = \big\{ h \in C^1[a,b] \mid h(a) = h(b) = 0 \big\}.$$

Очевидно, что из условий $x \in \Omega$, $h \in C_0^1[a,b]$ следует, что $x + \alpha h \in \Omega$ при всех вещественных α.

*Журнал «Вестник молодых учёных. Прикл. мат. и мех.». 2000. № 3. С. 12–22.

Запишем разложение

$$Q(x+\alpha h) = \int_a^b \big\{p(x'+\alpha h')^2 + q(x+\alpha h)^2 - 2f\cdot(x+\alpha h)\big\}dt =$$
$$= Q(x) + 2\alpha\int_a^b \big\{px'h' + (qx-f)h\big\}dt + \alpha^2\int_a^b \big\{p(h')^2 + qh^2\big\}dt.$$

Обозначим

$$l(x;h) = \int_a^b \big\{px'h' + (qx-f)h\big\}dt,$$
$$D(h) = \int_a^b \big\{p(h')^2 + qh^2\big\}dt.$$

Тогда

$$Q(x+\alpha h) = Q(x) + 2\alpha\, l(x;h) + \alpha^2 D(h). \tag{2}$$

ЛЕММА 1. *Если существует допустимая вариация h_0, для которой $D(h_0) < 0$, то*

$$\inf_{x\in\Omega} Q(x) = -\infty. \tag{3}$$

Доказательство. Возьмём произвольный план x_0. Согласно (2) выражение $Q(x_0+\alpha h_0)$ представляет собой квадратичный трёхчлен относительно α с отрицательным старшим коэффициентом. Понятно, что $Q(x_0+\alpha h_0)\to-\infty$ при $\alpha\to\infty$, откуда и следует (3). □

Таким образом, задача (1) содержательна только при условии

$$D(h) \geqslant 0 \quad \forall h\in C_0^1[a,b], \tag{4}$$

то есть когда интегральная квадратичная форма $D(h)$ *неотрицательно определена* на $C_0^1[a,b]$. В дальнейшем это условие будем считать выполненным.

1.3. Разложение (2) и условие (4) позволяют легко получить критерий оптимальности.

ТЕОРЕМА 1. *Для того чтобы план x_* задачи* (1) *был оптимальным, необходимо и достаточно, чтобы*

$$l(x_*;h) = 0 \quad \forall h\in C_0^1[a,b]. \tag{5}$$

Доказательство. Необходимость. Возьмём допустимую вариацию h. Поскольку x_* — точка минимума $Q(x)$ на Ω, то

$$0 \leqslant Q(x_*+\alpha h) - Q(x_*) = 2\alpha\, l(x_*;h) + \alpha^2 D(h).$$

Поделим это неравенство на $\alpha > 0$ и перейдём к пределу при $\alpha \to +0$. Получим $l(x_*; h) \geqslant 0$. Отметим, что вместе с h допустимой вариацией является и $-h$. По доказанному $l(x_*; -h) \geqslant 0$, или $-l(x_*; h) \geqslant 0$. Объединяя два неравенства $l(x_*; h) \geqslant 0$ и $-l(x_*; h) \geqslant 0$, приходим к равенству (5).

Достаточность. Возьмём любой план x и положим $h = x - x_*$. Очевидно, что $h \in C_0^1[a, b]$. Соласно (2), (5) и (4) имеем

$$Q(x) = Q(x_* + h) = Q(x_*) + D(h) \geqslant Q(x_*).$$

Теорема доказана. □

1.4. Теперь в критерии оптимальности (5) нужно избавиться от h (переформулировать его в терминах функций p, q и f). Для этого потребуется некоторая подготовка.

ЛЕММА 2. *Пусть* $u \in C[a, b]$. *Если*

$$\int_a^b u(t)h'(t)dt = 0 \quad \forall h \in C_0^1[a, b],$$

то $u(t) \equiv \text{const}$ *на* $[a, b]$.

Доказательство. Положим $u_1(t) = \int_a^t u(\tau)d\tau$ и введём линейную функцию $p_1(t) = c_0 t + c_1$, интерполирующую $u_1(t)$ в точках $t = a$ и $t = b$:

$$p_1(a) = u_1(a), \quad p_1(b) = u_1(b).$$

Разность $h_1(t) = u_1(t) - p_1(t)$ является допустимой вариацией, причём $h_1'(t) = u(t) - c_0$.

Имеем

$$\int_a^b \big[u(t) - c_0\big]^2 dt = \int_a^b \big[u(t) - c_0\big] h_1'(t)dt = \\ = \int_a^b u(t)h_1'(t)dt - c_0 \int_a^b h_1'(t)dt = \int_a^b u(t)h_1'(t)dt.$$

Последний интеграл в силу условия леммы равен нулю, так что

$$\int_a^b \big[u(t) - c_0\big]^2 dt = 0.$$

Отсюда очевидным образом следует требуемое. □

ОСНОВНАЯ ЛЕММА ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ. *Пусть функции u, v принадлежат $C[a,b]$. Для того чтобы выполнялось равенство*

$$\int_a^b \big[u(t)h'(t) + v(t)h(t)\big]dt = 0 \quad \forall h \in C_0^1[a,b], \tag{6}$$

необходимо и достаточно, чтобы

$$u \in C^1[a,b] \quad u \quad u'(t) \equiv v(t) \quad на \quad [a,b].$$

Доказательство. Необходимость. Положим $v_1(t) = \int_a^t v(\tau)d\tau$. При $h \in$ $\in C_0^1[a,b]$ имеем

$$\int_a^b v(t)h(t)dt = \int_a^b v_1'(t)h(t)dt = v_1(t)h(t)\Big|_a^b - \int_a^b v_1(t)h'(t)dt = \\ = -\int_a^b v_1(t)h'(t)dt.$$

Условие (6) принимает вид

$$\int_a^b \big[u(t) - v_1(t)\big]h'(t)dt = 0 \quad \forall h \in C_0^1[a,b].$$

По лемме 2, $u(t) - v_1(t) \equiv \text{const}$ или

$$u(t) \equiv v_1(t) + \text{const} \quad \text{на} \quad [a,b].$$

Отсюда следует и непрерывная дифференцируемость функции $u(t)$, и тождество $u'(t) \equiv v(t)$ на $[a,b]$.

Достаточность. При $h \in C_0^1[a,b]$ имеем

$$\int_a^b \big[u(t)h'(t) + v(t)h(t)\big]dt = \int_a^b \big[u(t)h'(t) + u'(t)h(t)\big]dt = \\ = \int_a^b \big[u(t)h(t)\big]'dt = u(t)h(t)\Big|_a^b = 0,$$

что и требовалось доказать. □

1.5. Вернёмся к критерию оптимальности (5) и перепишем его в развёрнутом виде:

$$\int_a^b \big\{px_*'h' + (qx_* - f)h\big\}dt = 0 \quad \forall h \in C_0^1[a,b]. \tag{7}$$

ТЕОРЕМА 2. *Для того чтобы план x_* задачи* (1) *был оптимальным, необходимо и достаточно, чтобы $px'_* \in C^1[a,b]$ и*

$$-(px'_*) + qx_* = f \quad \textit{на} \quad [a,b]. \tag{8}$$

Доказательство непосредственно следует из теоремы 1 и основной леммы вариационного исчисления при $u(t) = p(t)x'_*(t)$, $v(t) = q(t)x_*(t) - f(t)$. □

Установлено, что решение задачи (1) сводится к решению краевой задачи

$$\begin{gathered} \mathfrak{L}(x;t) := -\frac{d}{dt}\left(p\frac{dx}{dt}\right) + qx = f, \\ x(a) = A, \quad x(b) = B. \end{gathered} \tag{9}$$

Дифференциальный оператор $\mathfrak{L}(x;t)$ называется *оператором Штурма–Лиувилля.*

1.6. Напомним, что теорема 2 справедлива при выполнении условия (4), которое в развёрнутом виде выглядит так:

$$\int_a^b \left\{p(h')^2 + qh^2\right\}dt \geqslant 0 \quad \forall h \in C_0^1[a,b]. \tag{10}$$

Избавиться в этом условии от h (переформулировать его в терминах p и q) — непростая задача, решению которой посвящён следующий раздел. Пока же отметим, что неравенство (10) очевидно выполняется, если функции $p(t)$ и $q(t)$ неотрицательны на $[a,b]$.

Менее тривиальным является следующее утверждение.

ТЕОРЕМА 3. *Если $D(h) \geqslant 0$ на $C_0^1[a,b]$, то необходимо*

$$p(t) \geqslant 0 \quad \textit{на} \quad [a,b]. \tag{11}$$

Доказательство. Допустим противное. Тогда найдётся точка $t_0 \in (a,b)$, в которой $p(t_0) < 0$. Положим $\varepsilon_0 = -p(t_0)/2$ и выберем $\delta_0 > 0$ так, чтобы $[t_0 - \delta_0, t_0 + \delta_0] \subset (a,b)$ и $\bigl|p(t) - p(t_0)\bigr| \leqslant \varepsilon_0$ при $t \in [t_0 - \delta_0, t_0 + \delta_0]$. В этом случае $p(t) \leqslant p(t_0) + \varepsilon_0 = -\varepsilon_0$, то есть

$$p(t) \leqslant -\varepsilon_0 \quad \text{на} \quad [t_0 - \delta_0, t_0 + \delta_0]. \tag{12}$$

При $\delta \in (0, \delta_0]$ введём вариацию

$$h_\delta(t) = \begin{cases} \sqrt{\frac{\delta}{\pi}}\Bigl[1 + \cos\bigl(\frac{\pi}{\delta}(t - t_0)\bigr)\Bigr], & \text{если } t \in [t_0 - \delta, t_0 + \delta], \\ 0 & \text{при остальных } t \in [a,b]. \end{cases}$$

Поскольку $h'_\delta(t) = -\sqrt{\frac{\pi}{\delta}}\sin\big(\frac{\pi}{\delta}(t-t_0)\big)$ при $t \in [t_0-\delta, t_0+\delta]$, то $h_\delta \in C_0^1[a,b]$. Кроме того,

$$0 \leqslant h_\delta(t) \leqslant 2\sqrt{\tfrac{\delta}{\pi}} \quad \text{на} \quad [a,b]. \tag{13}$$

Согласно (12) и (13) имеем

$$D(h_\delta) = \int\limits_{t_0-\delta}^{t_0+\delta} \big\{p(h'_\delta)^2 + qh_\delta^2\big\}dt \leqslant -\varepsilon_0\frac{\pi}{\delta}\int\limits_{t_0-\delta}^{t_0+\delta}\sin^2\big(\tfrac{\pi}{\delta}(t-t_0)\big)dt + \\ + \frac{4\delta}{\pi}\int_a^b \big|q(t)\big|dt = -\varepsilon_0\int_{-\pi}^{\pi}\sin^2(\tau)d\tau + \frac{4\delta}{\pi}\int_a^b\big|q(t)\big|dt.$$

Очевидно, что $D(h_\delta) < 0$ при малых $\delta > 0$. Но это противоречит предположению о неотрицательности $D(h)$ на $C_0^1[a,b]$. Теорема доказана. □

1.7. Условие (11) называется *условием Лежандра*. Оно необходимо для неотрицательной определенности $D(h)$ на $C_0^1[a,b]$, но не достаточно. Более того, даже *усиленное условие Лежандра* $p(t) > 0$ на $[a,b]$ не гарантирует неотрицательной определенности $D(h)$ на $C_0^1[a,b]$.

ПРИМЕР. Возьмём

$$D_\lambda(h) = \int_0^\pi \big\{(h')^2 - \lambda h^2\big\}dt, \quad h \in C_0^1[0,\pi].$$

Здесь $p(t) \equiv 1$ — усиленное условие Лежандра выполнено. Вместе с тем, на допустимой вариации $h_0(t) = \sin(t)$ имеем

$$D_\lambda(h_0) = \int_0^\pi \big\{\cos^2(t) - \lambda\sin^2(t)\big\}dt = \frac{\pi}{2}(1-\lambda),$$

так что при $\lambda > 1$ интегральная квадратичная форма $D_\lambda(h)$ не является неотрицательно определенной на $C_0^1[0,\pi]$.

2°. Критерий неотрицательной определённости интегральной квадратичной формы

2.1. Будем исследовать интегральную квадратичную форму

$$D(h) = \int_a^b \big\{p(h')^2 + qh^2\big\}dt, \quad h \in C_0^1[a,b],$$

при следующих предположениях

$$q \in C[a,b], \quad p \in C^1[a,b], \quad p(t) > 0 \quad \text{на} \quad [a,b]. \tag{14}$$

Рассмотрим на $[a,b]$ дифференциальное уравнение

$$(ph')' = qh. \tag{15}$$

Учитывая (14), перепишем его в виде $ph''+p'h'-qh=0$, или, что равносильно,

$$h'' + \frac{p'}{p}h' - \frac{q}{p}h = 0. \tag{16}$$

Уравнение (16) называется *уравнением Якоби*. Это однородное дифференциальное уравнение второго порядка, разрешённое относительно старшей производной, с непрерывными на $[a,b]$ коэффициентами. Из теории обыкновенных дифференциальных уравнений известно, что такое уравнение имеет единственное на $[a,b]$ решение при любых начальных условиях вида $h(c)=A$, $h'(c)=A'$, где $c\in[a,b]$. Назовём *главным решением уравнения Якоби* то решение $h_0(t)$, у которого

$$h_0(a)=0, \quad h_0'(a)=1. \tag{17}$$

ТЕОРЕМА ЯКОБИ. *Пусть выполнены условия* (14). *Для того чтобы квадратичная форма $D(h)$ была неотрицательно определённой на $C_0^1[a,b]$, необходимо и достаточно, чтобы главное решение уравнения Якоби $h_0(t)$ было положительным на (a,b).*

2.2. Доказательство. Достаточность. Допустим, что $h_0(t)>0$ на (a,b). Покажем, что

$$D(h) = \int_a^b p\big(h' - \frac{h}{h_0}h_0'\big)^2 dt \qquad \forall h\in C_0^1[a,b]. \tag{18}$$

Отметим прежде всего, что функция h/h_0 непрерывна на $[a,b]$. Действительно, по правилу Лопиталя для $h\in C_0^1[a,b]$ имеем

$$\lim_{t\to a+0}\frac{h(t)}{h_0(t)} = \lim_{t\to a+0}\frac{h'(t)}{h_0'(t)} = h'(a).$$

Далее, предел дроби $h(t)/h_0(t)$ при $t\to b-0$ равен нулю, если $h_0(b)>0$. Если же $h_0(b)=0$, то $h_0'(b)\neq 0$ [иначе $h_0(t)\equiv 0$ в силу единственности решения уравнения Якоби с начальными условиями $h_0(b)=0$, $h_0'(b)=0$, что противоречит (17)]. В этом случае

$$\lim_{t\to b-0}\frac{h(t)}{h_0(t)} = \lim_{t\to b-0}\frac{h'(t)}{h_0'(t)} = \frac{h'(b)}{h_0'(b)}.$$

Установлено, что $h/h_0 \in C[a,b]$. Значит, под интегралом в правой части (18) стоит непрерывная функция.

Проверим справедливость равенства (18). При малых $\varepsilon > 0$ имеем

$$\int\limits_{a+\varepsilon}^{b-\varepsilon} p\Big(h' - \frac{h}{h_0}h_0'\Big)^2 dt = \int\limits_{a+\varepsilon}^{b-\varepsilon} p(h')^2 dt - 2\int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{ph_0'}{h_0} hh'dt + \int\limits_{a+\varepsilon}^{b-\varepsilon} p\Big(\frac{h_0'}{h_0}\Big)^2 h^2 dt. \tag{19}$$

Но

$$-2\int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{ph_0'}{h_0} hh'dt = -\int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{ph_0'}{h_0} dh^2 = -\frac{ph_0'}{h_0}h^2\Big|_{a+\varepsilon}^{b-\varepsilon} + \int\limits_{a+\varepsilon}^{b-\varepsilon} \Big(\frac{ph_0'}{h_0}\Big)' h^2 dt =$$

$$= -p\frac{h}{h_0}h_0' h\Big|_{a+\varepsilon}^{b-\varepsilon} + \int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{(ph_0')'}{h_0} h^2 dt - \int\limits_{a+\varepsilon}^{b-\varepsilon} \Big(\frac{h_0'}{h_0}\Big)^2 h^2 dt.$$

Поскольку h_0 удовлетворяет уравнению (15), то

$$\int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{(ph_0')'}{h_0} h^2 dt = \int\limits_{a+\varepsilon}^{b-\varepsilon} qh^2 dt.$$

Значит,

$$-2\int\limits_{a+\varepsilon}^{b-\varepsilon} \frac{ph_0'}{h_0} hh'dt = \int\limits_{a+\varepsilon}^{b-\varepsilon} qh^2 dt - \int\limits_{a+\varepsilon}^{b-\varepsilon} p\Big(\frac{h_0'}{h_0}\Big)^2 h^2 dt - p\frac{h}{h_0}h_0' h\Big|_{a+\varepsilon}^{b-\varepsilon}.$$

Подставляя это в (19), получаем

$$\int\limits_{a+\varepsilon}^{b-\varepsilon} p\Big(h' - \frac{h}{h_0}h_0'\Big)^2 dt = \int\limits_{a+\varepsilon}^{b-\varepsilon} \big\{p(h')^2 + qh^2\big\} dt - p\frac{h}{h_0}h_0' h\Big|_{a+\varepsilon}^{b-\varepsilon}.$$

В данном равенстве перейдём к пределу при $\varepsilon \to +0$. Функция $p(h/h_0)h_0'$ непрерывна на $[a,b]$ и $h \in C_0^1[a,b]$, так что двойная подстановка стремится к нулю при $\varepsilon \to +0$. Теперь ясно, что предельный переход приводит к формуле (18), из которой, в частности, следует, что $D(h) \geqslant 0$ при всех $h \in C_0^1[a,b]$. Достаточность установлена. □

2.3. Для доказательства необходимости потребуется некоторая подготовка.

ЛЕММА О СКРУГЛЕНИИ УГЛОВ. *Пусть* $D(h) \geqslant 0$ *на* $C_0^1[a,b]$. *Если функция* $\hat{h}$ *удовлетворяет условиям*

$$\hat{h} \in C[a,b], \quad \hat{h}(a) = \hat{h}(b) = 0;$$
$$\hat{h} \in C^1[a,\xi], \quad \hat{h} \in C^1[\xi,b] \quad \textit{при некотором} \quad \xi \in (a,b),$$

то $D(\hat{h}) \geqslant 0$.

Характерный вид функции $\hat{h}$ изображён на рис. 1.

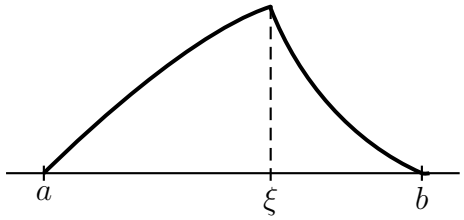


Рис. 1. Вид вариации $\hat{h}$

Доказательство. Введём вспомогательную функцию (см. рис. 2)

$$g(t) = \begin{cases} \left(\frac{1-|t|}{2}\right)^2 & \text{при } t \in [-1,1], \\ 0 & \text{при остальных } t \in \mathbb{R}. \end{cases}$$

На интервале $(0,1)$ имеем

$$g'(t) = \frac{d}{dt}\Big(\frac{1-t}{2}\Big)^2 = -\frac{1}{2}(1-t).$$

Учитывая чётность g, запишем

$$g'(+0) = -\frac{1}{2}, \quad g'(-0) = \frac{1}{2}, \quad \big|g'(t)\big| \leqslant \frac{1}{2} \quad \text{при} \quad t \neq 0. \tag{20}$$

К этому нужно добавить, что $0 \leqslant g(t) \leqslant 1/4$ на $\mathbb{R}$.

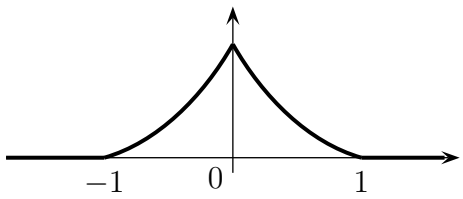


Рис. 2. Вид функции $g(t)$

Положим при $\delta > 0$

$$g_\delta(t) = \delta g\Big(\frac{t}{\delta}\Big).$$

Очевидно, что $g_\delta(t) = 0$ вне $[-\delta, \delta]$. Кроме того,

$$0 \leqslant g_\delta(t) \leqslant \delta/4 \quad \text{на} \quad \mathbb{R}. \tag{21}$$

Поскольку $g'_\delta(t) = g'(t/\delta)$, то согласно (20) при всех $\delta > 0$

$$g'_\delta(+0) = -\frac{1}{2}, \quad g'_\delta(-0) = \frac{1}{2}, \quad \big|g'_\delta(t)\big| \leqslant \frac{1}{2} \quad \text{при} \quad t \neq 0. \tag{22}$$

Введём обозначение $\alpha_1 = \hat{h}'(\xi+0) - \hat{h}'(\xi-0)$ и при малых $\delta > 0$ рассмотрим функцию

$$h_\delta(t) = \hat{h}(t) + \alpha_1 g_\delta(t - \xi).$$

Покажем, что $h_\delta \in C_0^1[a, b]$. Сомнение вызывает только дифференцируемость $h_\delta(t)$ при $t = \xi$. Имеем согласно (22)

$$h'_\delta(\xi + 0) = \hat{h}'(\xi + 0) - \frac{1}{2}\alpha_1, \quad h'_\delta(\xi - 0) = \hat{h}'(\xi - 0) + \frac{1}{2}\alpha_1.$$

Учитывая определение α_1, получаем $h'_\delta(\xi+0) - h'_\delta(\xi-0) = 0$, то есть $h'_\delta(\xi+0) = h'_\delta(\xi - 0)$. Установлено, что $h_\delta \in C_0^1[a, b]$ при малых $\delta > 0$.

Согласно условию леммы $D(h_\delta) \geqslant 0$. Распишем это неравенство подробно:

$$\begin{aligned}
0 &\leqslant D\big(\hat{h} + \alpha_1 g_\delta(\cdot - \xi)\big) = \\
&= \int_a^b \Big\{p\big(\hat{h}' + \alpha_1 g'_\delta(\cdot - \xi)\big)^2 + q\big(\hat{h} + \alpha_1 g_\delta(\cdot - \xi)\big)^2\Big\}dt = \\
&= D(\hat{h}) + \alpha_1 \int_a^b q\big[2\hat{h} g_\delta(\cdot - \xi) + \alpha_1 g_\delta^2(\cdot - \xi)\big]dt + \\
&\quad + \alpha_1 \int\limits_{\xi-\delta}^{\xi+\delta} p\Big[2\hat{h}' g'_\delta(\cdot - \xi) + \alpha_1\big(g'_\delta(\cdot - \xi)\big)^2\Big]dt =: \\
&=: D(\hat{h}) + \alpha_1 I_\delta^{(0)} + \alpha_1 I_\delta^{(1)}.
\end{aligned} \tag{23}$$

Оценим интегралы $I_\delta^{(0)}$ и $I_\delta^{(1)}$. На основании (21) и (22) имеем

$$|I_\delta^{(0)}| \leqslant \frac{\delta}{2}\int_a^b |q|\big(|\hat{h}| + |\alpha_1|\delta/8\big)dt,$$

$$|I_\delta^{(1)}| \leqslant \Big(\int\limits_{\xi-\delta}^{\xi} + \int\limits_{\xi}^{\xi+\delta}\Big)p\big(|\hat{h}'| + |\alpha_1|/4\big)dt.$$

Очевидно, что $I_\delta^{(0)} \to 0$ и $I_\delta^{(1)} \to 0$ при $\delta \to +0$. Переходя в (23) к пределу при $\delta \to +0$, получаем $D(\hat{h}) \geqslant 0$. Лемма доказана. □

2.4. Нам понадобится ещё одно вспомогательное утверждение.

ЛЕММА 3. *Пусть в некоторой точке $\xi \in (a, b]$ главное решение уравнения Якоби обращается в ноль, $h_0(\xi) = 0$. Тогда*

$$\int_a^\xi \left[p(h_0')^2 + qh_0^2\right] dt = 0. \tag{24}$$

Доказательство. Согласно (15) и определению h_0 имеем

$$\int_a^\xi p(h_0')^2 dt = \int_a^\xi ph_0'\, dh_0 = -\int_a^\xi (ph_0')' h_0\, dt = -\int_a^\xi qh_0^2\, dt,$$

откуда и следует (24). □

Замечание. Если $h_0(b) = 0$, то $h_0 \in C_0^1[a, b]$ и $D(h_0) = 0$.

2.5. Переходим к доказательству необходимости в теореме Якоби. Пусть $D(h) \geqslant 0$ на $C_0^1[a, b]$. Покажем, что главное решение $h_0(t)$ уравнения Якоби положительно на (a, b).

В противном случае найдётся точка $\xi \in (a, b)$, в которой $h_0(\xi) = 0$. При этом $h_0'(\xi) \neq 0$ (в силу единственности решения уравнения Якоби — см. пункт 2.2).

Введём функцию

$$\hat{h}_0(t) = \begin{cases} h_0(t) & \text{при } t \in [a, \xi], \\ 0 & \text{при } t \in [\xi, b]. \end{cases}$$

По лемме 3

$$D(\hat{h}_0) = \int_a^\xi \left\{p(h_0')^2 + qh_0^2\right\} dt = 0. \tag{25}$$

Возьмём произвольную вариацию $h \in C_0^1[a, b]$ и запишем разложение

$$D(\hat{h}_0 + \alpha h) = D(\hat{h}_0) + \alpha D(h) + 2\alpha \int_a^\xi [ph_0'h' + qh_0 h] dt.$$

Имеем

$$\int_a^\xi ph_0'h'\, dt = \int_a^\xi (ph_0')dh = ph_0'h\Big|_a^\xi - \int_a^\xi qh_0 h\, dt.$$

Учитывая (25), получаем

$$D(\hat{h}_0 + \alpha h) = \alpha^2 D(h) + 2\alpha\, p(\xi) h_0'(\xi) h(\xi).$$

Выберем допустимую вариацию h так, чтобы выполнялось неравенство

$$\lambda := 2p(\xi)h_0'(\xi)h(\xi) < 0.$$

Тогда

$$D(\hat{h}_0 + \alpha h) = \alpha\big[\lambda + \alpha D(h)\big]. \tag{26}$$

При малых $\alpha > 0$ правая часть (26) отрицательна. Вместе с тем, $D(\hat{h}_0 + \alpha h) \geqslant$ $\geqslant 0$, поскольку функция $\hat{h}_0 + \alpha h$ удовлетворяет условиям леммы о скруглении углов. Полученное противоречие завершает доказательство теоремы Якоби. □

2.6. Условие $h_0(t) > 0$ при $t \in (a, b)$ называется *условием Якоби*. При доказательстве достаточности в теореме Якоби было установлено, что выполнение условия Якоби гарантирует для интергального квадратичного функционала $D(h)$ представление (18). Этот факт имеет дополнительное следствие.

ЛЕММА 4. *Пусть выполнено условие Якоби. Если допустимая вариация h_* такова, что $D(h_*) = 0$, то $h_*(t) = \lambda h_0(t)$ при некотором вещественном λ, где h_0 — главное решение уравнения Якоби.*

Доказательство. Согласно (18)

$$D(h_*) = \int_a^b p\big(h_*' - \frac{h_*}{h_0}h_0'\big)^2 dt.$$

По условию леммы $D(h_*) = 0$. Принимая во внимание, что $p(t) > 0$ на $[a, b]$, получаем

$$h_*' - \frac{h_*}{h_0}h_0' = 0 \quad \text{на} \quad (a, b).$$

Перепишем это равенство в эквивалентном виде

$$\frac{h_*'h_0 - h_*h_0'}{h_0^2} = 0 \quad \text{на} \quad (a, b).$$

Это значит, что $\big(\frac{h_*}{h_0}\big)' = 0$ на (a, b). Теперь очевидно, что $\frac{h_*}{h_0} \equiv \text{const}$ или $h_*(t) =$ $= \lambda h_0(t)$ при $t \in (a, b)$. По непрерывности последнее равенство выполняется на всём отрезке $[a, b]$.

Лемма доказана. □

3°. Критерий положительной определённости интегральной квадратичной формы

3.1. Квадратичная форма $D(h)$ называется *положительно определённой* на $C_0^1[a,b]$, если $D(h) > 0$ при всех $h \in C_0^1[a,b]$, не равных тождественно нулю на $[a,b]$.

ТЕОРЕМА 4. *Пусть выполнены условия* (14). *Для того чтобы квадратичная форма $D(h)$ была положительно определённой на $C_0^1[a,b]$, необходимо и достаточно, чтобы главное решение $h_0(t)$ уравнения Якоби было положительным на $(a,b]$ (включая точку $t=b$).*

Доказательство. Необходимость. Очевидно, что положительно определённая форма $D(h)$ является и неотрицательно определённой. По теореме Якоби $h_0(t) > 0$ на (a,b). Остаётся проверить, что $h_0(b) > 0$. Допустим противное: $h_0(b) = 0$. Тогда $h_0 \in C_0^1[a,b]$. Более того, $D(h_0) = 0$. Это следует из леммы 3 при $\xi = b$. Получили противоречие с положительной определённостью $D(h)$, поскольку главное решение $h_0(t)$ уравнения Якоби не равно тождественно нулю на $[a,b]$ ($h_0'(a) = 1$).

Достаточность. Так как, в частности, $h_0(t) > 0$ на (a,b), то по теореме Якоби $D(h) \geqslant 0$ на $C_0^1[a,b]$. Допустим, что $D(h_*) = 0$ на некоторой допустимой вариации h_*. В силу леммы 4 имеем $h_*(t) = \lambda h_0(t)$ на $[a,b]$. В частности, $h_*(b) = \lambda h_0(b)$. В этом равенстве $h_*(b) = 0$ по определению допустимой вариации и $h_0(b) > 0$ по условию теоремы. Значит, $\lambda = 0$, так что $h_*(t) \equiv 0$ на $[a,b]$. Установлено, что $D(h) > 0$ на всех допустимых вариациях, не равных тождественно нулю на $[a,b]$.

Теорема доказана. □

3.2. Условие $h_0(t) > 0$ при $t \in (a,b]$ называется *усиленным условием Якоби.*

ТЕОРЕМА 5. *При выполнении усиленного условия Якоби существует константа $\mu > 0$, такая, что*

$$D(h) \geqslant \mu \int_a^b (h')^2 dt \qquad \forall h \in C_0^1[a,b]. \tag{27}$$

Доказательство. Теория Якоби развивается в предположениях (14). В частности, считается, что выполнено усиленное условие Лежандра $p(t) > 0$ при $t \in [a,b]$.

Обозначим

$$\mu_0 = \min_{t\in[a,b]} p(t).$$

Ясно, что $\mu_0 > 0$. При $\mu \in (0, \mu_0)$ рассмотрим семейство интегральных квадратичных функционалов

$$D_\mu(h) = \int_a^b \big[(p-\mu)(h')^2 + qh^2\big]\,dt.$$

Покажем, что при некотором $\mu \in (0, \mu_0)$ функционал $D_\mu(h)$ положительно определён на $C_0^1[a,b]$. Отсюда очевидным образом будет следовать неравенство (27).

Запишем уравнение Якоби для $D_\mu(h)$:

$$h'' + \frac{p'}{p-\mu}h' - \frac{q}{p-\mu}h = 0. \tag{28}$$

Обозначим через $h(t,\mu)$ главное решение уравнения Якоби. Оно удовлетворяет начальным условиям

$$h(a,\mu) = 0, \quad h'(a,\mu) = 1.$$

Коэффициенты уравнения (28) зависят от параметра μ. Они непрерывны на множестве $[a,b] \times (-\mu_0, \mu_0)$, причём $h(t,0) = h_0(t)$. По теореме о непрерывной зависимости решения линейного однородного дифференциального уравнения от параметра найдётся $\mu_1 \in (0, \mu_0)$, такое, что функции $h(t,\mu)$ и $h'(t,\mu)$ будут непрерывными по совокупности переменных на множестве $[a,b]\times[-\mu_1,\mu_1]$.

Возьмём $\varepsilon = \frac{1}{2}$. По нему найдётся $\delta > 0$ со свойством

$$\big|h'(t,\mu) - h'(a,0)\big| \leqslant \frac{1}{2} \quad \text{при} \quad 0 \leqslant t-a \leqslant \delta \quad \text{и} \quad 0 \leqslant \mu \leqslant \delta.$$

В частности, при тех же t и μ

$$h'(t,\mu) \geqslant h'(a,0) - \frac{1}{2} = \frac{1}{2}$$

и

$$h(t,\mu) = h(t,\mu) - h(a,\mu) = h'(\xi,\mu)(t-a) \geqslant \frac{1}{2}(t-a). \tag{29}$$

Пусть при $t \in [a+\delta, b]$ будет $h(t,0) \geqslant \varepsilon_1 > 0$. Тогда найдётся $\delta_1 \in (0,\delta]$ со свойством

$$\big|h(t,\mu) - h(t,0)\big| \leqslant \frac{1}{2}\varepsilon_1 \quad \text{при} \quad t \in [a+\delta, b] \quad \text{и} \quad 0 \leqslant \mu \leqslant \delta_1.$$

В частности, при тех же t и μ

$$h(t,\mu) \geqslant h(t,0) - \frac{1}{2}\varepsilon_1 \geqslant \frac{1}{2}\varepsilon_1. \tag{30}$$

На основании (29) и (30) заключаем, что при $\mu = \delta_1$ главное решение $h(t,\mu)$ уравнения (28) положительно на $(a,b]$. Это гарантирует положительную определённость функционала $D_\mu(h)$ на $C_0^1[a,b]$.

Теорема доказана. □

4°. Описание всего множества решений. Вернёмся к квадратичной вариационной задаче (1). Пусть x_* — некоторое её решение. Согласно (2) и (5) имеем

$$Q(x_* + h) = Q(x_*) + D(h) \quad \forall h \in C_0^1[a, b].$$

Отсюда следует, что в случае положительной определённости интегральной квадратичной формы $D(h)$ решение x_* единственно. Если к тому же выполнены условия (14), то в силу теоремы 5 при некотором $\mu > 0$ справедливо неравенство

$$Q(x_* + h) \geqslant Q(x_*) + \mu \int_0^1 (h')^2 dt \qquad \forall h \in C_0^1[a, b].$$

А что можно сказать, когда $h_0(t) > 0$ при $t \in (a, b)$, но $h_0(b) = 0$, то есть когда форма $D(h)$ только неотрицательно определена?

ТЕОРЕМА 6. *Пусть x_* — некоторое решение задачи* (1)*. Если выполнено сформулированное выше условие, то всё множество решений задачи* (1) *допускает представление*

$$x(t) = x_*(t) + \lambda h_0(t) \quad \forall \lambda \in \mathbb{R}. \tag{31}$$

Доказательство. В данном случае $h_0 \in C_0^1[a, b]$ и $D(h_0) = 0$ в силу леммы 3 при $\xi = b$.

Возьмём план x вида (31). Пользуясь разложением (2) и формулой (5), получаем

$$Q(x) = Q(x_* + \lambda h_0) = Q(x_*) + \lambda^2 D(h_0) = Q(x_*),$$

так что на плане x, как и на плане x_*, функционал $Q(x)$ достигает наименьшего значения.

Наоборот, пусть x — оптимальный план. Обозначим $h_* = x - x_*$. Ясно, что $h_* \in C_0^1[a, b]$. Имеем

$$Q(x) = Q(x_* + h_*) = Q(x_*) + D(h_*).$$

Так как $Q(x) = Q(x_*)$, то $D(h_*) = 0$. По лемме 4, $h_* = \lambda h_0$ при некотором вещественном λ. Значит, $x = x_* + h_* = x_* + \lambda h_0$.

Теорема доказана. □

5°. Неограниченность снизу функционала $Q(x)$. Если $h_0(\xi) = 0$ в некоторой точке $\xi \in (a, b)$, то по теореме Якоби квадратичная форма $D(h)$ не будет неотрицательно определённой на множестве $C_0^1[a, b]$. В этом случае согласно лемме 1 справедливо соотношение (3), то есть функционал $Q(x)$ не ограничен снизу на множестве планов Ω. Следующее утверждение дополняет этот результат.

ТЕОРЕМА 7. *Пусть* $h_0(t) > 0$ *на* (a,b), *но* $h_0(b) = 0$. *При выполнении условия*

$$\int_a^b f h_0\,dt \neq p(b)Bh_0'(b) - p(a)A \tag{32}$$

справедливо соотношение (3).

Доказательство. В данном случае $h_0 \in C_0^1[a,b]$ и $D(h_0) = 0$. Возьмём произвольный план x_0 и запишем разложение

$$Q(x_0 + \alpha h_0) = Q(x_0) + 2\alpha l(x_0; h_0), \tag{33}$$

где

$$l(x_0; h_0) = \int_a^b \bigl[px_0'h_0' + (qx_0 - f)h_0\bigr]dt.$$

Так как $(ph_0')' = qh_0$, то

$$\int_a^b px_0'h_0'dt = \int_a^b ph_0'dx_0 = ph_0'x_0\Big|_a^b - \int_a^b qh_0x_0dt.$$

Значит,

$$l(x_0; h_0) = px_0h_0'\Big|_a^b - \int_a^b f h_0\,dt.$$

При выполнении условия (32) коэффициент $2l(x_0; h_0)$ в разложении (33) отличен от нуля, откуда и следует заключение теоремы. □

При $f(t) \equiv 0$ на $[a,b]$ условие (32) принимает простой вид

$$p(b)Bh_0'(b) \neq p(a)A.$$

6°. Схема решения квадратичной вариационной задачи

6.1. Рассмотрим квадратичную вариационную задачу

$$\begin{gathered} Q(x) := \int_a^b \bigl\{p(x')^2 + qx^2 - 2fx\bigr\}dt \to \inf, \\ x(a) = A, \quad x(b) = B, \quad x \in C^1[a,b]. \end{gathered} \tag{32}$$

Предположим, что

$$q \in C[a,b], \quad p \in C^1[a,b], \quad p(t) > 0 \quad \text{на} \quad [a,b].$$

Проведённый в предыдущих разделах анализ позволяет предложить следующую схему решения задачи (32).

1) Находим главное решение уравнения Якоби

$$-(ph')' + qh = 0,$$
$$h(a) = 0, \quad h'(a) = 1.$$

Обозначим его $h_0(t)$.

2) Если $h_0(\xi) = 0$ при некотором $\xi \in (a, b)$, то

$$\inf_{x\in\Omega} Q(x) = -\infty.$$

3) Пусть $h_0(t) > 0$ при $t \in (a, b)$. Решаем краевую задачу Штурма–Лиувилля

$$-(px')' + qx = f,$$
$$x(a) = A, \quad x(b) = B.$$

Пусть x_* — некоторое её решение. (Отметим, что существование решения не гарантируется, равно как не исключается наличие бесконечного множества решений — см. примеры ниже.)

4) Если $h_0(b) > 0$, то x_* — единственное решение задачи (32).

Если $h_0(b) = 0$, то всё множество решений задачи (32) допускает представление

$$x(t) = x_*(t) + \lambda h_0(t), \quad \lambda \in \mathbb{R}.$$

Замечание 1. Дифференциальные операторы в уравнениях Якоби и Штурма-Лиувилля одинаковые. Только уравнение Якоби однородное, а уравнение Штурма–Лиувилля неоднородное, и для уравнения Якоби решается задача Коши, а для уравнения Штурма–Лиувилля — краевая задача.

Замечание 2. Выполнение усиленного условия Лежандра $p(t) > 0$ при $t \in \in [a, b]$ существенно для теории Якоби. Если $p(t) \geqslant 0$ и $q(t) \geqslant 0$ при $t \in [a, b]$, то квадратичная форма $D(h)$ неотрицательно определена на $C_0^1[a, b]$. В этом случае следует сразу переходить к решению краевой задачи для уравнения Штурма-Лиувилля.

6.2. Обратимся к примерам.

ПРИМЕР 1. Рассмотрим квадратичную вариационную задачу

$$Q(x) := \int_0^{\pi/2} \left\{(x')^2 - x^2\right\} dt \to \inf,$$
$$x(0) = 0, \quad x(\pi/2) = 1, \quad x \in C^1[0, \pi/2]. \tag{33}$$

В данном случае $p(t) \equiv 1$, $q(t) \equiv -1$, $f(t) \equiv 0$. Все необходимые условия выполнены. Найдём главное решение уравнения Якоби

$$-h'' - h = 0, \quad h(0) = 0, \quad h'(0) = 1.$$

Общее решение уравнения Якоби имеет вид

$$h(t) = c_1 \cos(t) + c_2 \sin(t).$$

Выделим главное решение: $h(0) = 0 \Rightarrow c_1 = 0$; $h'(0) = 1 \Rightarrow c_2 = 1$. Таким образом, $h_0(t) = \sin(t)$. Имеем $h_0(t) > 0$ на $(0, \pi/2]$.

Решим краевую задачу Штурма–Лиувилля

$$-x'' - x = 0, \quad x(0) = 0, \quad x(\pi/2) = 1.$$

Из общего решения $x(t) = c_1 \cos(t) + c_2 \sin(t)$ выделим то, которое удовлетворяет краевым условиям: $x(0) = 0 \Rightarrow c_1 = 0$; $x(\pi/2) = 1 \Rightarrow c_2 = 1$. Получаем $x_*(t) = \sin(t)$. Поскольку $h_0(\pi/2) > 0$, то x_* — единственное решение задачи (33).

ПРИМЕР 2. В задаче (33) заменим $\pi/2$ на π. Придём к такой задаче:

$$\begin{gathered} Q(x) := \int_0^\pi \left\{(x')^2 - x^2\right\} dt \to \inf, \\ x(0) = 0, \quad x(\pi) = 1, \quad x \in C^1[0, \pi]. \end{gathered} \tag{34}$$

Главное решение уравнения Якоби не изменится, $h_0(t) = \sin(t)$, причём $h_0(t)$ остаётся положительным на $(0, \pi)$.

Обратимся к краевой задаче Штурма–Лиувилля:

$$-x'' - x = 0, \quad x(0) = 0, \quad x(\pi) = 1. \tag{35}$$

Из общего решения $x(t) = c_1 \cos(t) + c_2 \sin(t)$ выделим то, которое удовлетворяет первому краевому условию: $x(0) = 0 \Rightarrow c_1 = 0$. Второе краевое условие $x(\pi) = 1$ принимает вид $c_2 \sin(\pi) = 1$. Оно не выполняется ни при каком c_2. Значит, задача (35) не имеет решения.

В данном случае $\inf\limits_{x \in \Omega} Q(x) = -\infty$. Это следует из теоремы 7.

ПРИМЕР 3. В задаче (34) заменим второе краевое условие $x(\pi) = 1$ на $x(\pi) = 0$. У новой задачи

$$\begin{gathered} Q(x) := \int_0^\pi \left\{(x')^2 - x^2\right\} dt \to \inf, \\ x(0) = 0, \quad x(\pi) = 0, \quad x \in C^1[0, \pi] \end{gathered} \tag{36}$$

главное решение уравнения Якоби остаётся тем же, $h_0(t) = \sin(t)$. Сохраняется и его положительность на $(0, \pi)$. Что касается краевой задачи Штурма–Лиувилля

$$-x'' - x = 0, \quad x(0) = 0, \quad x(\pi) = 0,$$

то её очевидным решением является функция $x_*(t) \equiv 0$. Учитывая равенство $h_0(\pi) = 0$, приходим к выводу о том, что всё множество решений задачи (36) допускает представление

$$x(t) = x_*(t) + \lambda h_0(t) = \lambda \sin(t), \quad \lambda \in \mathbb{R}.$$

6.3. В заключение отметим, что использованный в данной статье элементарный подход позволяет решать и более сложные вариационные задачи (см., например, [1]). Дальнейшие сведения о вариационном исчислении можно найти в книгах [2, 3, 4, 5].

ДОБАВЛЕНИЕ

ВЫПУКЛОСТЬ КВАДРАТИЧНОГО ФУНКЦИОНАЛА

На выпуклом множестве Ω,

$$\Omega = \big\{x \in C^1[a,b] \mid x(a) = A,\ x(b) = B\big\},$$

рассмотрим квадратичный функционал

$$Q(x) = \int_a^b \big[p(x')^2 + qx^2 - 2fx\big]\,dt,$$

где p, q и f — непрерывные функции.

ТЕОРЕМА. *Для того чтобы функционал $Q(x)$ был выпуклым на Ω, необходимо и достаточно, чтобы интегральная квадратичная форма*

$$D(h) = \int_a^b \big[p(h')^2 + qh^2\big]\,dt$$

была неотрицательно определённой на $C_0^1[a,b]$.

Доказательство. Достаточность. Нужно проверить, что при всех x_0, x_1 из Ω и всех $\lambda \in [0,1]$ выполняется неравенство

$$Q\big(\lambda x_1 + (1-\lambda)x_0\big) \leqslant \lambda Q(x_1) + (1-\lambda)Q(x_0). \tag{37}$$

Зафиксируем x_0, x_1 из Ω и положим $h = x_1 - x_0$. Ясно, что $h \in C_0^1[a,b]$. При всех $\lambda \in \mathbb{R}$ согласно (2) имеем

$$Q(x_0 + \lambda h) = Q(x_0) + 2\lambda l(x_0; h) + \lambda^2 D(h). \tag{38}$$

В частности, при $\lambda = 1$

$$Q(x_1) = Q(x_0) + 2l(x_0; h) + D(h).$$

Отсюда следует, что

$$2l(x_0; h) = Q(x_1) - Q(x_0) - D(h).$$

Подставив это в (38), получим

$$\begin{aligned} Q\big(\lambda x_1 + (1-\lambda)x_0\big) &= Q(x_0) + \lambda\big[Q(x_1) - Q(x_0) - D(h)\big] + \lambda^2 D(h) = \\ &= \lambda Q(x_1) + (1-\lambda)Q(x_0) - \lambda(1-\lambda)D(h). \end{aligned} \tag{39}$$

По условию теоремы $D(h) \geqslant 0$, что гарантирует справедливость неравенства (37) при $\lambda \in [0,1]$.

Необходимость. Зафиксируем $h \in C_0^1[a,b]$, возьмём произвольную функцию x_0 из Ω и положим $x_1 = x_0 + h$. Ясно, что $x_1 \in \Omega$. Воспользуемся равенством (39) при $\lambda = \frac{1}{2}$. Запишем

$$Q\big(\tfrac{1}{2}(x_1 + x_0)\big) - \tfrac{1}{2}\big[Q(x_1) + Q(x_0)\big] = -\tfrac{1}{4}D(h).$$

Левая часть этого равенства неположительна в силу выпуклости $Q(x)$ на Ω. Значит, $D(h) \geqslant 0$.

Теорема доказана. □

Замечание. Конкретный вид функционалов $l(x;h)$ и $D(h)$ не играл роли. Использовалось только разложение (38).

ЛИТЕРАТУРА

1. Кирушев В. А., Малозёмов В. Н., Певный А. Б. *Хвост дракона* // Журн. вычисл. мат. и матем. физ., 1997. Т. 37. № 11. С. 1362–1369.

2. Буслаев В. С. *Вариационное исчисление.* Л.: Изд-во ЛГУ, 1980. 287 с.

3. Алексеев В. М., Тихомиров В. М., Фомин С. В. *Оптимальное управление.* М.: Наука, 1979. 429 с.

4. Смирнов В. И. *Курс высшей математики.* Т. IV. Часть I. Изд. 6-е. М.: Наука, 1974. 336 с.

5. Михлин С. Г. *Курс математической физики.* М.: Физматгиз, 1968. 575 с.

ОБ ОДНОЙ КУБИЧЕСКОЙ ВАРИАЦИОННОЙ ЗАДАЧЕ*

В. Н. Малозёмов, Г. Ш. Тамасян

1°. Квадратичные вариационные задачи изучены детально [1]. При выполнении усиленного условия Лежандра и условия Якоби они сводятся к решению краевой задачи для линейного дифференциального уравнения Штурма–Лиувилля.

При переходе к кубическим вариационным задачам ситуация становится менее определенной.

2°. В качестве примера рассмотрим кубическую вариационную задачу вида

$$\begin{gathered} J(x) := \int_{-1}^{1} \left[(x')^3 + 24tx\right] dt \to \operatorname{extr}, \\ x(-1) = -1, \quad x(1) = 1, \quad x \in C^1[-1,1]. \end{gathered} \tag{1}$$

Покажем, что у этой задачи существует стационарная кривая, которая однако не принадлежит пространству $C^2[-1,1]$; при этом

$$\inf J(x) = -\infty, \qquad \sup J(x) = \infty, \tag{2}$$

где инфимум и супремум берутся по множеству всех планов задачи (1).

Подынтегральную функцию обозначим буквой F,

$$F(t,x,x') = (x')^3 + 24tx.$$

Стационарной кривой называется решение уравнение Эйлера

$$F'_x - \frac{d}{dt}F'_{x'} = 0,$$

удовлетворяющее краевым условиям. В данном случае приходим к нелинейной задаче:

$$\frac{d}{dt}(x')^2 = 8t, \tag{3}$$

*Семинар «CNSA & NDO». Избранные доклады. 11 февраля 2016 г.

$$x(-1) = -1, \quad x(1) = 1. \tag{4}$$

Решение уравнения (3) будем искать в виде алгебраического полинома второй степени, поскольку его производная есть полином первой степени, в квадрате — полином второй степени, дифференцирование которого снова приводит к полиному первой степени. Несложные вычисления показывают, что необходимо

$$x(t) = \varepsilon t^2 + c, \tag{5}$$

где $\varepsilon = \pm 1$ и c — произвольная вещественная константа.

К сожалению, ни одна кривая вида (5), которые называют экстремалями, не удовлетворяет краевым условиям (4). Стационарную кривую получим, склеивая две различные экстремали. Положим

$$x_*(t) = \begin{cases} t^2 & \text{при } t \in [0,1], \\ -t^2 & \text{при } t \in [-1,0]. \end{cases} \tag{6}$$

Имеем $x'_*(t) = 2|t|$. Ясно, что функция x_* принадлежит пространству $C^1[-1,1]$, но не принадлежит $C^2[-1,1]$. Последний факт соответствует теореме Гильберта, так как вторая производная $F''_{x'x'}$ на функции x_* в точке $t = 0$ обращается в ноль.

Функция x_* удовлетворяет уравнению (3) и краевым условиям (4). По определению она является стационарной кривой для вариационной задачи (1).

3°. Переходим к доказательству предельных соотношений (2).

Введём параметрическое семейство планов (см. рис. 1)

$$x_\alpha(t) = \begin{cases} \left(\dfrac{t-\alpha}{1+\alpha}\right)^2 - 2 & \text{при } t \in [-1,\alpha], \\ 3\left(\dfrac{t-\alpha}{1-\alpha}\right)^2 - 2 & \text{при } t \in [\alpha,1], \end{cases}$$

где $\alpha \in (-1,1)$.

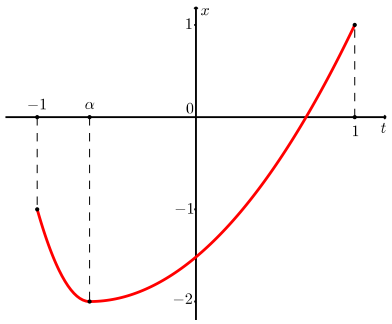


Рис. 1. График функции $x_\alpha(t)$

Вычислим значение $J(x_\alpha)$. Имеем

$$\int_{-1}^{\alpha}(x'_\alpha)^3dt = \frac{2}{(1+\alpha)^6}(t-\alpha)^4\Big|_{-1}^{\alpha} = -\frac{2}{(1+\alpha)^2},$$
$$\int_{\alpha}^{1}(x'_\alpha)^3dt = \frac{54}{(1-\alpha)^6}(t-\alpha)^4\Big|_{\alpha}^{1} = \frac{54}{(1-\alpha)^2},$$
$$24\int_{-1}^{1}tx_\alpha dt = -4(\alpha^2+4\alpha-3).$$

Значит,

$$J(x_\alpha) = -\frac{2}{(1+\alpha)^2} + \frac{54}{(1-\alpha)^2} - 4(\alpha^2+4\alpha-3).$$

Отсюда очевидным образом следуют предельные соотношения (2) (см. рис. 2).

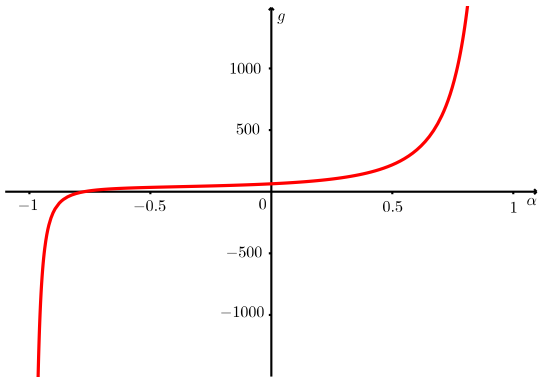


Рис. 2. График функции $g(\alpha) = J(x_\alpha)$ на интервале $(-1, 1)$

4°. Интересно, что получится, если применить метод наискорейшего спуска к решению задачи (1). Напомним описание этого метода [2, 3].

Рассмотрим простейшую вариационную задачу

$$\begin{aligned} &J(x) := \int_a^b F(t, x(t), x'(t))\,dt \to \inf, \\ &x(a) = A, \quad x(b) = B, \quad x \in C^1[a,b]. \end{aligned} \tag{7}$$

Здесь функция $F(t,u,v)$ непрерывна вместе с $\frac{\partial F}{\partial u}$ и $\frac{\partial F}{\partial v}$ на множестве $[a,b] \times \times \mathbb{R} \times \mathbb{R}$. Решение будем искать в виде

$$x(t) = A + \int_a^t z(\tau)\,d\tau, \tag{8}$$

где $z \in C[a,b]$. В этом случае $x'(t) = z(t)$ на $[a,b]$ и $x(a) = A$.

Перепишем задачу (7) в терминах функции z:

$$\begin{aligned} f(z) &= \int_a^b F(t,x,z)\,dt \to \inf, \\ \varphi(z) &:= A + \int_a^b z(t)\,dt - B = 0. \end{aligned} \tag{9}$$

Как известно [2, с. 182], функционал f дифференцируем по Гато, причём его производная Гато в точке z имеет вид

$$Q(t,z) = \int_t^b \frac{\partial F(\tau, x(\tau), z(\tau))}{\partial x} d\tau + \frac{\partial F(t, x(t), z(t))}{\partial z}.$$

Направление наискорейшего спуска функционала $f(z)$ при ограничении $\varphi(z) = 0$ вычисляется по формуле

$$G(t,z) = -\frac{q(t,z)}{\|q(t,z)\|}\,,$$

где

$$q(t,z) = Q(t,z) - \frac{1}{b-a}\int_a^b Q(t,z)\,dt \tag{10}$$

и $\|q(t,z)\| = \sqrt{\int_a^b q^2(t,z)\,dt}$.

Опишем алгоритм решения задачи (9), который мы будем использовать при расчетах. В качестве начального приближения возьмём произвольную функцию $z_0 \in C[a,b]$, удовлетворяющую ограничению $\varphi(z_0) = 0$. По формуле (10) вычислим $q(t,z_0)$. Если $q(t,z_0) \equiv 0$ на $[a,b]$, то z_0 — стационарная точка. Процесс заканчивается. Иначе переходим к построению z_1.

Общая $(k+1)$-я итерация, перед началом которой имеются z_k и $q(t,z_k) \not\equiv 0$, состоит из следующих шагов:

- находим направление спуска $G_k(t) = G(t,z_k)$;
- вычисляем $\gamma_k > 0$ как точку минимума функции $\psi_k(\gamma) = f(z_k + \gamma G_k)$ на полуоси $(0,+\infty)$;
- определяем $z_{k+1} = z_k + \gamma_k G_k$.

Если $q(t,z_{k+1}) \equiv 0$ на $[a,b]$, то z_{k+1} — стационарная точка. Вычисления прекращаются. Иначе переходим к очередной итерации.

Отметим, что все точки последовательности $z_0, z_1, \ldots$ удовлетворяют ограничению задачи (9). Функции вида (8)

$$x_k(t) = A + \int_a^t z_k(\tau)\,d\tau, \quad k = 0, 1, \ldots, \tag{11}$$

удовлетворяют ограничениям задачи (7) и образуют для неё минимизирующую последовательность.

5°. Воспользуемся описанным алгоритмом для решения задачи (1), заменив в ней операцию «extr» на «inf». Имеем $a = -1$, $b = 1$, $A = -1$, $B = 1$, $F(t, u, v) = v^3 + 24tu$ и

$$Q(t, z) = 12(1 - t^2) + 3z^2(t).$$

В качестве начального приближения возьмём функцию $x_0(t) = t$. Ей соответствует $z_0(t) \equiv 1$. Вычисляем

$$f(z_0) = 18, \qquad Q(t, z_0) = -12t^2 + 15,$$
$$G_0(t) = \sqrt{\tfrac{5}{128}}(12t^2 - 4).$$

Далее

$$\psi_0(\gamma) := f(z_0 + \gamma G_0) = \tfrac{\sqrt{10}}{7}\gamma^3 + 3\gamma^2 - \tfrac{8\sqrt{10}}{5}\gamma + 18.$$

График этой функции изображён на рис. 3. Единственным положительным корнем функции $\psi_0'(\gamma)$ является

$$\gamma_0 = \tfrac{\sqrt{10}}{30}\left(\sqrt{777} - 21\right) \approx 0.725.$$

Для первого приближения z_1 получаем представление

$$z_1(t) := z_0(t) + \gamma_0 G_0(t) = \tfrac{\sqrt{777}-21}{4}t^2 + \tfrac{33-\sqrt{777}}{12}.$$

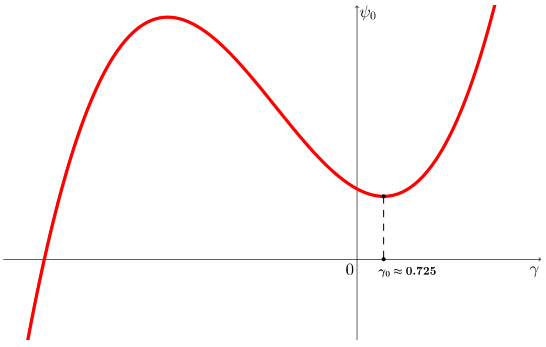


Рис. 3. График функции $\psi_0(\gamma)$

Переходим к построению z_2. Вычисляем

$$f(z_1) = 39 - \tfrac{37}{45}\sqrt{777} \approx 16.081,$$
$$Q(t, z_1) = \tfrac{29-\sqrt{777}}{64}(504t^4 - 432t^2 + 407 + 11\sqrt{777}),$$
$$G_1(t) = \frac{3(29-\sqrt{777})}{16\sqrt{809-29\sqrt{777}}}(35t^4 - 30t^2 + 3).$$

Далее,

$$\psi_1(\gamma) := f(z_1 + \gamma G_1) \approx -0.343\gamma^3 + 3.893\gamma^2 - 0.955\gamma + 16.081.$$

График этой функции изображён на рис. 4.

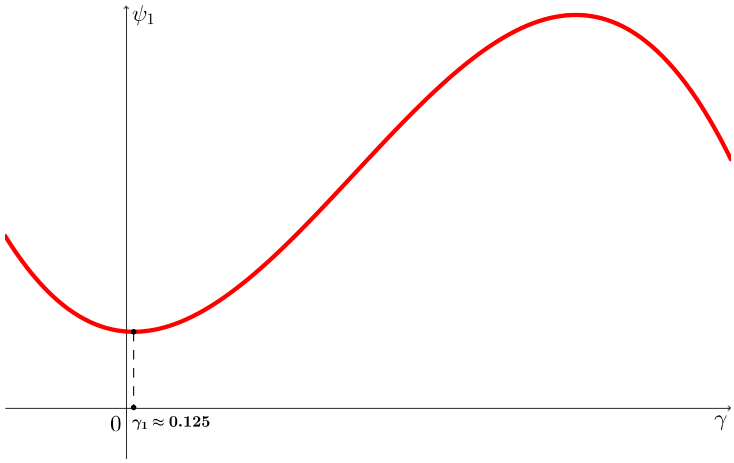


Рис. 4. График функции $\psi_1(\gamma)$

Старший коэффициент у полинома $\psi_1(\gamma)$ отрицательный, поэтому

$$\inf_{\gamma>0} \psi_1(\gamma) = -\infty.$$

Отсюда следует, что рассматриваемая кубическая вариационная задача (1) не имеет решения.

Можно ограничиться локальным минимумом функции ψ_1, который достигается в точке $\gamma_1 \approx 0.125$, и положить $z_2 = z_1 + \gamma_1 G_1$. Если и в дальнейшем на этапе поиска величины шага спуска ограничиваться информацией о локальном минимуме, то получим сходящуюся последовательность $\{z_k\}$.

В таблице приведены результаты расчетов для первых пяти приближений. Напомним, что стационарная кривая $x_*(t)$ определяется формулой (6). При этом $z_*(t) = 2|t|$. Функции x_k и z_k связаны соотношением (11).

Таблица. Результаты расчетов

k	$f(z_k)$	$\|q(t,z_k)\|$	$\|z_*-z_k\|$	$\max\limits_{-1\leqslant t\leqslant 1}\|x_*(t)-x_k(t)\|$
0	18	5.060	0.816	0.250
1	16.081	0.955	0.214	0.053
2	16.022	0.437	0.132	0.033
3	16.008	0.248	0.090	0.020
4	16.004	0.151	0.069	0.015
5	16.002	0.086	0.057	0.012

На рисунках 5 и 6 изображены графики функций z_k и x_k при k, равном 0, 1, 5, и графики предельных функций z_* и x_*.

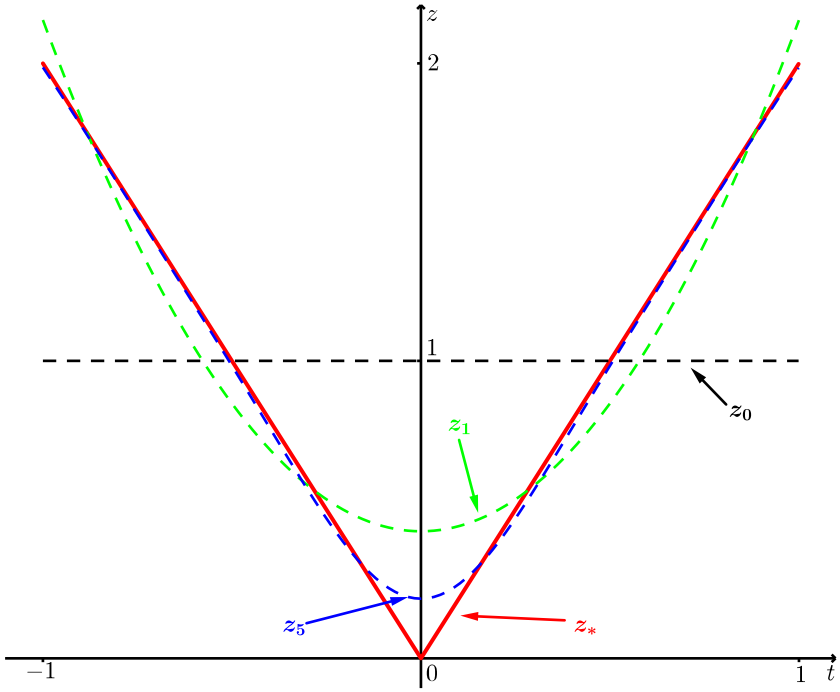


Рис. 5. Графики функций z_k и z_*

Функция x_5 имеет вид

$$x_5(t) \approx 0.192203\,t + 1.739081\,t^3 - 2.853252\,t^5 + 5.459985\,t^7 - 8.941935\,t^9 + \\ + 12.372540\,t^{11} - 14.610029\,t^{13} + 14.640463\,t^{15} - 12.266217\,t^{17} + \\ + 8.470679\,t^{19} - 4.748481\,t^{21} + 2.121426\,t^{23} - 0.736246\,t^{25} + \\ + 0.190414\,t^{27} - 0.034208\,t^{29} + 0.003763\,t^{31} - 0.000189\,t^{33}.$$

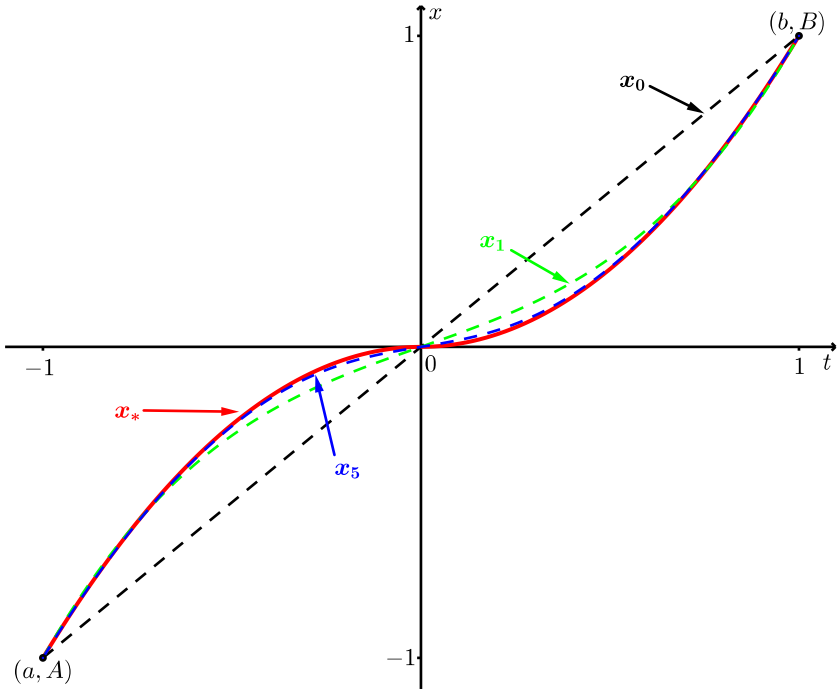


Рис. 6. Графики функций x_k и x_*

6°. В заключение отметим, что функция $q(t,z)$ является ортогональной проекцией производной Гато $Q(t,z)$ на подпространство пространства $C[a,b]$, определяемое условием

$$\int_a^b z(t)\,dt = 0. \tag{12}$$

Это следует из приведённой ниже леммы.

Пусть $z_0(t)$ — произвольная непрерывная на отрезке $[a,b]$ функция. Рассмотрим экстремальную задачу: *минимизировать функционал*

$$\int_a^b [z(t) - z_0(t)]^2\,dt$$

по всем функциям $z \in C[a,b]$, *удовлетворяющим ограничению* (12).

ЛЕММА. *Данная задача имеет единственное решение*

$$z_*(t) = z_0(t) - c_0, \tag{13}$$

где $c_0 = \frac{1}{b-a}\int_a^b z_0(t)\,dt$.

Доказательство. Для любого плана z в силу (12) и определения c_0 имеем

$$0 \leqslant \int_a^b [z-(z_0-c_0)]^2\,dt = \int_a^b [(z-z_0)+c_0]^2\,dt =$$
$$= \int_a^b (z-z_0)^2\,dt - 2\int_a^b z_0c_0\,dt + \int_a^b c_0^2\,dt = \int_a^b (z-z_0)^2\,dt - (b-a)c_0^2.$$

Значит,

$$\int_a^b (z-z_0)^2\,dt \geqslant (b-a)c_0^2.$$

Равенство достигается только при $z = z_0 - c_0$. Лемма доказана. □

ДОБАВЛЕНИЕ 1

Экстремальное свойство коэффициентов Фурье

Это добавление навеяно леммой из п. 6°.

Пусть в пространстве $L_2[a,b]$ заданы функция x_0 и ортонормированная система $\{\xi_1,\dots,\xi_n\}$. Рассмотрим экстремальную задачу:

$$\textit{минимизировать} \quad \int_a^b [x-x_0]^2\,dt$$

по всем функциям x из $L_2[a,b]$, удовлетворяющим условиям

$$\int_a^b x\,\xi_k\,dt = 0, \quad k \in 1:n. \tag{14}$$

ТЕОРЕМА 1. *Данная задача имеет единственное решение*

$$x_* = x_0 - \sum_{k=1}^n c_k\xi_k,$$

где $c_k = \int_a^b x_0\xi_k\,dt$ — коэффициенты Фурье функции x_0.

Доказательство. Обозначим

$$p_0 = \sum_{k=1}^n c_k\xi_k.$$

Имеем

$$\int_a^b x_0 p_0\,dt = \sum_{k=1}^n c_k \int_a^b x_0\xi_k\,dt = \sum_{k=1}^n c_k^2,$$
$$\int_a^b p_0^2\,dt = \sum_{k,j=1}^n c_k c_j \int_a^b \xi_k\xi_j\,dt = \sum_{k=1}^n c_k^2. \tag{15}$$

Возьмём произвольную функцию $x \in L_2[a,b]$, удовлетворяющую условиям (14). Для неё в силу (15)

$$0 \leqslant \int_a^b [x-(x_0-p_0)]^2 dt = \int_a^b [(x-x_0)+p_0)]^2 dt =$$
$$= \int_a^b [x-x_0]^2\,dt - 2\int_a^b x_0 p_0\,dt + \int_a^b p_0^2\,dt = \int_a^b [x-x_0]^2\,dt - \sum_{k=1}^n c_k^2.$$

Остаётся переписать полученное неравенство в виде

$$\int_a^b [x-x_0]^2\,dt \geqslant \sum_{k=1}^n c_k^2$$

и отметить, что неравенство выполняется как равенство только тогда, когда $x = x_0 - p_0$.

Теорема доказана. □

ДОБАВЛЕНИЕ 2

Эквивалентные определения стационарной кривой

Напомним, что стационарной кривой для простейшей вариационной задачи (7) называется план этой задачи, удовлетворяющий уравнению Эйлера. Можно дать эквивалентное определение стационарной кривой в терминах функции $q(t,z)$ вида (10).

ТЕОРЕМА 2. *Для того чтобы план x_* задачи* (7) *был стационарной кривой, необходимо и достаточно, чтобы для производной $z_* = x_*'$ выполнялось соотношение*

$$q(t,z_*) \equiv 0 \quad \textit{на} \quad [a,b]. \tag{16}$$

Доказательство. Необходимость. Пусть план x_* является стационарной кривой. Проинтегрировав уравнение Эйлера по отрезку $[t,b]$, получим

$$\int_t^b F_x'(\tau, x_*(\tau), z_*(\tau))\,d\tau - F_{x'}'(b, x_*(b), z_*(b)) + F_{x'}'(t, x_*(t), z_*(t)) = 0$$

или

$$Q(t, z_*) \equiv \text{const} \quad \text{на} \quad [a, b].$$

Теперь тождество (16) следует из определения (10) функции q.
Достаточность. Пусть x_* — план задачи (7) и $z_* = x'_*$. Тогда

$$x_*(t) = A + \int_a^t z_*(\tau)\, d\tau.$$

Продифференцируем тождество (16). Получим

$$-F'_x(t, x_*(t), x'_*(t)) + \frac{d}{dt} F'_{x'}(t, x_*(t), x'_*(t)) \equiv 0 \quad \text{на} \quad [a, b].$$

Это означает, что x_* удовлетворяет уравнению Эйлера. По условию x_* — план, так что кривая x_* является стационарной.

Теорема доказана. □

Замечание. В п. 4° основного текста мы неявно пользовались следующим утверждением: *если функция z принадлежит $C[a, b]$, удовлетворяет ограничению задачи* (9) *и $q(t, z) \equiv 0$ на $[a, b]$, то функция x вида* (8) *является стационарной кривой для задачи* (7). Справедливость этого утверждения следует из теоремы 2, если учесть, что x — план задачи (7).

ЛИТЕРАТУРА

1. Малоземов В. Н. *Квадратичные вариационные задачи* // Вестник молодых учёных. Прикл. мат. и мех. 2000. № 3. С. 12–22.

2. Демьянов В. Ф. *Условия экстремума и вариационное исчисление.* М.: Высшая школа, 2005. 335 с.

3. Долгополик М. В., Тамасян Г. Ш. *Об эквивалентности методов наискорейшего и гиподифференциального спусков в некоторых задачах условной оптимизации* // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2014. Т. 14, вып. 4, ч. 2, с. 532–542.

ПЕРВЫЙ И ВТОРОЙ ДИФФЕРЕНЦИАЛЫ ИНТЕГРАЛЬНОГО ФУНКЦИОНАЛА*

В. Н. Малозёмов

1°. Рассмотрим интегральный функционал вида

$$J(x) = \int_a^b F\big(t, x(t), x'(t)\big)dt. \tag{1}$$

Здесь $F(t, y, z)$ — функция трёх переменных, заданная и непрерывная на некотором открытом связном множестве $U \subset \mathbb{R}^3$. Функционал $J(x)$ определён на функциях $x = x(t)$, непрерывно дифференцируемых на отрезке $[a, b]$, и таких, что параметрическая кривая

$$\Gamma(x) = \Big\{\big(t, x(t), x'(t)\big) \mid t \in [a, b]\Big\}$$

содержится в U. Множество таких функций x обозначим $\Omega^{\rm o}$ и назовём *естественной областью определения* функционала $J(x)$.

В линейном пространстве $C^1[a, b]$ непрерывно дифференцируемых на отрезке $[a, b]$ функций введём норму

$$\|x\|_1 = \max_{t\in[a,b]}\big|x(t)\big| + \max_{t\in[a,b]}\big|x'(t)\big|.$$

В качестве подготовительного шага докажем, что естественная область определения $\Omega^{\rm o}$ функционала $J(x)$ открыта в $C^1[a, b]$.

2°. Зафиксируем функцию $x_0 \in \Omega^{\rm o}$. Наряду с кривой $\Gamma_0 = \Gamma(x_0)$ рассмотрим её окрестность ("трубку")

$$\Gamma_\delta = \Big\{(t, u, v) \mid t \in [a, b],\ \big|u - x_0(t)\big| + \big|v - x_0'(t)\big| \leqslant \delta\Big\}, \quad \delta > 0. \tag{2}$$

ЛЕММА 1. *При любом $\delta > 0$ множество Γ_δ ограничено и замкнуто в $\mathbb{R}^3$.*

*Семинар «CNSA & NDO». Избранные доклады. 5 декабря 2013 г.

Доказательство. Ограниченность Γ_δ следует из условия $t \in [a,b]$ и неравенства

$$|u| + |v| \leqslant \delta + \|x_0\|_1.$$

Проверим замкнутость Γ_δ.

Пусть последовательность точек (t_k, u_k, v_k) принадлежит Γ_δ и сходится к (t_*, u_*, v_*). В частности, $t_k \to t_*$ и $t_* \in [a,b]$. По определению Γ_δ имеем

$$\big|u_k - x_0(t_k)\big| + \big|v_k - x_0'(t_k)\big| \leqslant \delta.$$

В пределе при $k \to \infty$ получаем

$$\big|u_* - x_0(t_*)\big| + \big|v_* - x_0'(t_*)\big| \leqslant \delta.$$

Это значит, что предельная точка (t_*, u_*, v_*) принадлежит Γ_δ.

Лемма доказана. □

ЛЕММА 2. *Существует $\delta_0 > 0$, такое, что $\Gamma_{\delta_0} \subset U$.*

Доказательство. Допустим противное. В этом случае для любой убывающей и стремящейся к нулю последовательности положительных чисел $\{\delta_k\}$ найдутся точки (t_k, u_k, v_k) из Γ_{δ_k}, которые не принадлежат U,

$$(t_k, u_k, v_k) \notin U, \ \ k = 1, 2, \ldots \tag{3}$$

По определению Γ_{δ_k} имеем

$$\big|u_k - x_0(t_k)\big| + \big|v_k - x_0'(t_k)\big| \leqslant \delta_k. \tag{4}$$

Последовательность $\big\{(t_k, u_k, v_k)\big\}$ ограничена, поэтому из неё можно выделить сходящуюся подпоследовательность. Можно считать, что

$$(t_k, u_k, v_k) \to (t_*, u_*, v_*) \quad \text{при} \quad k \to \infty, \tag{5}$$

где $t_* \in [a,b]$. Переходя в (4) к пределу при $k \to \infty$, получаем

$$\big|u_* - x_0(t_*)\big| + \big|v_* - x_0'(t_*)\big| = 0,$$

то есть $u_* = x_0(t_*)$, $v_* = x_0'(t_*)$. Как следствие,

$$(t_*, u_*, v_*) = \big(t_*, x_0(t_*), x_0'(t_*)\big).$$

Значит, точка (t_*, u_*, v_*) принадлежит Γ_0.

По условию $x_0 \in \Omega^\circ$, так что $\Gamma_0 \subset U$. В частности, $(t_*, u_*, v_*) \in U$. В силу открытости множества U вместе с точкой (t_*, u_*, v_*) ему принадлежат и близкие точки. Но это противоречит совокупности условий (5) и (3).

Лемма доказана. □

В дальнейшем с функцией $x_0 \in \Omega^{\mathrm{o}}$ мы будем связывать ограниченное и замкнутое (компактное) множество Γ_{δ_0} вида (2), содержащееся в $U \subset \mathbb{R}^3$.

Множество Γ_{δ_0} можно представить в другом виде:

$$\Gamma_{\delta_0} = \Big\{ \big(t, x_0(t) + u, x_0'(t) + v\big) \mid t \in [a,b],\ |u| + |v| \leqslant \delta_0 \Big\}.$$

ТЕОРЕМА 1. *Естественная область определения Ω^o функционала $J(x)$ открыта в $C^1[a,b]$.*

Доказательство. Зафиксируем функцию $x_0 \in \Omega^{\mathrm{o}}$. По лемме 2 существует $\delta_0 > 0$, такое, что $\Gamma_{\delta_0} \subset U$. Покажем, что $x_0 + h \in \Omega^{\mathrm{o}}$ при условии $\|h\|_1 \leqslant \delta_0$, то есть, что $\Gamma(x_0 + h) \subset U$. Для этого достаточно проверить включение

$$\Gamma(x_0 + h) \subset \Gamma_{\delta_0} \quad \text{при} \quad \|h\|_1 \leqslant \delta_0. \tag{6}$$

Пусть точка $\big(t, x_0(t) + h(t), x_0'(t) + h'(t)\big)$ принадлежит $\Gamma(x_0 + h)$. Обозначим $u = h(t)$, $v = h'(t)$. По условию $|u| + |v| \leqslant \delta_0$, поэтому

$$\big(t, x_0(t) + u, x_0'(t) + v\big) \in \Gamma_{\delta_0}.$$

Включение (6), а вместе с ним и теорема, доказаны. □

Таким образом, для каждой функции $x_0 \in \Omega^{\mathrm{o}}$ существует $\delta_0 > 0$, такое, что

$$x_0 + h \in \Omega^{\mathrm{o}} \quad \text{при} \quad \|h\|_1 \leqslant \delta_0.$$

3°. Переходим к вопросу о первом дифференциале функционала $J(x)$. Будем предполагать, что подынтегральная функция $F(t,y,z)$ непрерывно дифференцируема на множестве U, $F \in C^1(U)$.

Дифференцируемость функционала $J(x)$ в точке $x = x_0$ связана с разложением

$$J(x_0 + h) = J(x_0) + \ell(x_0; h) + o\big(\|h\|_1\big), \tag{7}$$

в котором $\ell(x_0; h)$ — линейный непрерывный функционал на $C^1[a,b]$ и $o\big(\|h\|_1\big)/\|h\|_1 \to 0$ при $\|h\|_1 \to 0$. Покажем, что разложение (7) при $x_0 \in \Omega^{\mathrm{o}}$ и $\|h\|_1 \leqslant \delta_0$ возможно.

Запишем

$$J(x_0 + h) - J(x_0) = \int_a^b \Big[F\big(t, x_0(t) + h(t), x_0'(t) + h'(t)\big) - F\big(t, x_0(t), x_0'(t)\big) \Big] dt.$$

Подынтегральную функцию обозначим через $H(t)$. При фиксированном t отрезок, соединяющий точки $\big(t, x_0(t), x_0'(t)\big)$ и $\big(t, x_0(t) + h(t), x_0'(t) + h'(t)\big)$ содержится в Γ_{δ_0}, а значит, и в U. По теореме о среднем для функции двух переменных имеем

$$H(t) = \tilde{F}_x' h + \tilde{F}_{x'}' h' = F_x' h + F_{x'}' h' + \big[(\tilde{F}_x' - F_x') h + (\tilde{F}_{x'}' - F_{x'}') h' \big].$$

Здесь частные производные F'_x и $F'_{x'}$ функции F по второму и третьему аргументу вычисляются в точке $\big(t, x_0(t), x'_0(t)\big)$, а $\tilde F'_x$ и $\tilde F'_{x'}$ — в средней точке

$$\big(t, x_0(t) + \theta h(t), x'_0(t) + \theta h'(t)\big).$$

Величина $\theta \in (0,1)$ зависит от t. Приходим к представлению

$$J(x_0+h) - J(x_0) = \int_a^b H(t)dt = \int_a^b [F'_x h + F'_{x'} h']dt +$$
$$+ \int_a^b \big[(\tilde F'_x - F'_x)h + (\tilde F'_{x'} - F'_{x'})h'\big]dt =: \ell(x_0;h) + \omega_1(x_0;h).$$

Функционал

$$\ell(x_0;h) = \int_a^b \Big[F'_x\big(t, x_0(t), x'_0(t)\big)h(t) + F'_{x'}\big(t, x_0(t), x'_0(t)\big)h'(t)\Big]dt \tag{8}$$

является линейным и ограниченным на $C^1[a,b]$. Его линейность очевидна. Ограниченность проверяется так:

$$\big|\ell(x_0;h)\big| \leqslant \|h\|_1 \int_a^b \big[|F'_x| + |F'_{x'}|\big]dt =: K_1\|h\|_1.$$

Из ограниченности следует непрерывность $\ell(x_0;h)$ на $C^1[a,b]$. Действительно,

$$\big|\ell(x_0;h_1) - \ell(x_0;h_2)\big| = \big|\ell(x_0;h_1-h_2)\big| \leqslant K_1\|h_1-h_2\|_1.$$

Займёмся оценкой слагаемого $\omega_1(x_0;h)$. Напомним, что $\|h\|_1 \leqslant \delta_0$. Зафиксируем $\varepsilon > 0$. Функции F'_x и $F'_{x'}$, как функции трёх переменных непрерывны на компактном (по лемме 1) множестве Γ_{δ_0}, а значит, и равномерно непрерывны на нём. Поэтому по $\varepsilon > 0$ найдётся такое $\delta \in (0, \delta_0/2]$, что

$$\Big|F'_x\big(t, x_0(t)+u, x'_0(t)+v\big) - F'_x\big(t, x_0(t), x'_0(t)\big)\Big| < \frac{\varepsilon}{b-a},$$
$$\Big|F'_{x'}\big(t, x_0(t)+u, x'_0(t)+v\big) - F'_{x'}\big(t, x_0(t), x'_0(t)\big)\Big| < \frac{\varepsilon}{b-a}$$

при всех $t \in [a,b]$, как только $|u| < \delta$, $|v| < \delta$ (в этом случае $|u| + |v| < 2\delta \leqslant \delta_0$, так что аргументы у производных F'_x и $F'_{x'}$ содержатся в Γ_{δ_0}). В частности,

$$|\tilde F'_x - F'_x| < \frac{\varepsilon}{b-a}, \quad |\tilde F'_{x'} - F'_{x'}| < \frac{\varepsilon}{b-a}$$

при всех $t \in [a,b]$, как только $\|h\|_1 < \delta$. Теперь при $\|h\|_1 < \delta$ имеем

$$\big|\omega_1(x_0;h)\big| \leqslant \frac{\varepsilon}{b-a}\int_a^b \big[|h| + |h'|\big]dt \leqslant \varepsilon\|h\|_1,$$

а это и означает, что $\omega_1(x_0;h) = o\big(\|h\|_1\big)$. Справедливость разложения (7) установлена.

4°. Покажем, что в разложении (7) линейный функционал $\ell(x_0;h)$ определяется единственным образом. Пусть

$$J(x_0+h)-J(x_0)=\ell_1(x_0;h)+o\big(\|h\|_1\big),$$
$$J(x_0+h)-J(x_0)=\ell_2(x_0;h)+o\big(\|h\|_1\big).$$

Тогда $\ell_1(x_0;h)-\ell_2(x_0;h)=o\big(\|h\|_1\big)$ при $\|h\|_1\leqslant\delta_0$. Покажем, что

$$\ell_1(x_0;h)=\ell_2(x_0;h)\quad\forall h\in C^1[a,b].$$

Зафиксируем $h_0\in C^1[a,b]$, $h_0\not\equiv 0$. При малых $\lambda>0$ будет $\|\lambda h_0\|_1\leqslant\delta_0$ и

$$\ell_1(x_0;\lambda h_0)-\ell_2(x_0;\lambda h_0)=o\big(\|\lambda h_0\|_1\big).$$

В силу линейности функционалов ℓ_1 и ℓ_2 при положительных λ последнее равенство можно переписать в виде

$$\ell_1(x_0;h_0)-\ell_2(x_0;h_0)=\frac{o\big(\|\lambda h_0\|_1\big)}{\|\lambda h_0\|_1}\|h_0\|_1.$$

В пределе при $\lambda\to+0$ получим $\ell_1(x_0;h_0)=\ell_2(x_0;h_0)$. Равенство $\ell_1(x_0;0)=\ell_2(x_0;0)=0$ следует из линейности функционалов. Таким образом, $\ell_1(x_0;h)=\ell_2(x_0;h)$ при всех $h\in C^1[a,b]$.

Единственный линейный непрерывный функционал $\ell(x_0;h)$ в формуле (7) называется *первым дифференциалом (Фреше) функционала $J(x)$ вида* (1) *в точке $x=x_0$* и обозначается $dJ(x_0;h)$. С учётом формулы (8) приходим к следующему заключению.

ТЕОРЕМА 2. *При $F\in C^1(U)$ интегральный функционал $J(x)$ вида* (1) *дифференцируем в каждой точке x_0 своей естественной области определения Ω^o и*

$$dJ(x_0;h)=\int_a^b[F'_x h+F'_{x'}h']dt.$$

При этом справедливо разложение

$$J(x_0+h)=J(x_0)+dJ(x_0;h)+o\big(\|h\|_1\big),$$

где $o\big(\|h\|_1\big)/\|h\|_1\to 0$ при $\|h\|_1\to 0$.

5°. Теперь предположим, что подынтегральная функция $F(t,y,z)$ в определении функционала $J(x)$ дважды непрерывно дифференцируема на множестве U, $F\in C^2(U)$. В этом случае для разности $H(t)$ из п. 3° можно записать разложение

$$H(t)=F'_x h+F'_{x'}h'+\tfrac12\big[F''_{xx}h^2+2F''_{xx'}hh'+F''_{x'x'}(h')^2\big]+$$
$$+\frac12\Big\{\big[\tilde F''_{xx}-F''_{xx}\big]h^2+2\big[\tilde F''_{xx'}-F''_{xx'}\big]hh'+\big[\tilde F''_{x'x'}-F''_{x'x'}\big](h')^2\Big\}.$$

Интегрируя, получаем

$$J(x_0+h)-J(x_0)=dJ(x_0;h)+\tfrac12 D(x_0;h)+\omega_2(x_0;h), \tag{9}$$

где

$$D(x_0;h)=\int_a^b\big[F''_{xx}h^2+2F''_{xx'}hh'+F''_{x'x'}(h')^2\big]dt. \tag{10}$$

Оценку для $\omega_2(x_0;h)$ проведём по той же схеме, что и для $\omega_1(x_0;h)$. По $\varepsilon>0$ найдём $\delta\in(0,\delta_0/2]$, такое, что

$$|\tilde F''_{xx}-F''_{xx}|<\frac{2\varepsilon}{b-a},\quad |\tilde F''_{xx'}-F''_{xx'}|<\frac{2\varepsilon}{b-a},\quad |\tilde F''_{x'x'}-F''_{x'x'}|<\frac{2\varepsilon}{b-a}$$

при всех $t\in[a,b]$ и $\|h\|_1<\delta$. В этом случае

$$\big|\omega_2(x_0;h)\big|\leqslant\frac{\varepsilon}{b-a}\int_a^b\big[|h|^2+2|hh'|+|h'|^2\big]dt\leqslant\varepsilon\|h\|_1^2.$$

Значит, $\omega_2(x_0;h)=o\big(\|h\|_1^2\big)$. Разложение (9) принимает вид

$$J(x_0+h)=J(x_0)+dJ(x_0;h)+\tfrac12 D(x_0;h)+o\big(\|h\|_1^2\big). \tag{11}$$

Функционал $D(x_0;h)$ определён на $C^1[a,b]$ и обладает тем свойством, что

$$D(x_0;\lambda h)=\lambda^2 D(x_0;h)\quad\forall\lambda\in\mathbb{R}. \tag{12}$$

Это свойство гарантирует его единственность. Действительно, допустим, что наряду с (11) справедливо разложение

$$J(x_0+h)=J(x_0)+dJ(x_0;h)+\tfrac12 D_1(x_0;h)+o\big(\|h\|_1^2\big),$$

в котором функционал $D_1(x_0;h)$ обладает свойством

$$D_1(x_0;\lambda h)=\lambda^2 D_1(x_0;h). \tag{13}$$

Тогда

$$D(x_0;h) - D_1(x_0;h) = o\big(\|h\|_1^2\big) \quad \text{при} \quad \|h\|_1 \leqslant \delta_0.$$

Покажем, что $D(x_0;h) = D_1(x_0;h)$ при всех $h \in C^1[a,b]$.

Зафиксируем функцию $h_0 \in C^1[a,b]$, $h_0 \not\equiv 0$. При малых $\lambda > 0$ будет $\|\lambda h_0\| \leqslant \delta_0$ и

$$D(x_0;\lambda h_0) - D_1(x_0;\lambda h_0) = o\big(\|\lambda h_0\|_1^2\big).$$

В силу (12) и (13) при положительных λ последнее равенство можно переписать в виде

$$D(x_0;h_0) - D_1(x_0;h_0) = \frac{o\big(\|\lambda h_0\|_1^2\big)}{\|\lambda h_0\|_1^2}\|h_0\|_1^2.$$

В пределе при $\lambda \to +0$ получаем $D(x_0;h_0) = D_1(x_0;h_0)$.

Равенства $D(x_0;0) = D_1(x_0;0) = 0$ следуют из (12) и (13).

Функционал $D(x_0;h)$ вида (10) называется *вторым дифференциалом (Фреше) функционала $J(x)$ в точке $x = x_0$* и обозначается $d^2J(x_0;h)$.

Подведём итог.

ТЕОРЕМА 3. *При $F \in C^2(U)$ функционал $J(x)$ вида (1) дважды дифференцируем в каждой точке x_0 своей естественной области определения Ω^o и*

$$d^2J(x_0;h) = \int_a^b [F''_{xx}h^2 + 2F''_{xx'}hh' + F''_{x'x'}(h')^2]dt,$$

где

$$F''_{xx} = F''_{xx}\big(t,x_0(t),x_0'(t)\big), \quad F''_{xx'} = F''_{xx'}\big(t,x_0(t),x_0'(t)\big), \quad F''_{x'x'} = F''_{x'x'}\big(t,x_0(t),x_0'(t)\big).$$

При этом справедливо разложение

$$J(x_0+h) = J(x_0) + dJ(x_0;h) + \tfrac{1}{2}d^2J(x_0;h) + o\big(\|h\|_1^2\big),$$

где $o\big(\|h\|_1^2\big)/\|h\|_1^2 \to 0$ при $\|h\|_1 \to 0$.

НЕОБХОДИМЫЕ УСЛОВИЯ ОПТИМАЛЬНОСТИ ПЕРВОГО И ВТОРОГО ПОРЯДКОВ В ПРОСТЕЙШЕЙ НЕЛИНЕЙНОЙ ЗАДАЧЕ ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ*

В. Н. Малозёмов

Аннотация. Рассматривается простейшая нелинейная задача вариационного исчисления. С помощью первого и второго дифференциалов нелинейного интегрального функционала, основной леммы вариационного исчисления и теоремы Якоби о критерии неотрицательной определённости интегральной квадратичной формы выводятся необходимые условия локального минимума первого и второго порядков для этой задачи.

1°. Рассмотрим вариационную задачу

$$J(x) := \int_a^b F\big(t, x(t), x'(t)\big)dt \to \inf, \tag{1}$$

$$x(a) = A, \quad x(b) = B. \tag{2}$$

Здесь $F = F(t, x, y)$ — функция трёх переменных, заданная и непрерывно дифференцируемая на некотором открытом линейно связном множестве $U \subset \mathbb{R}^3$. *Естественной областью определения* Ω^o функционала $J(x)$ является множество функций $x \in C^1[a, b]$, таких, что параметрическая кривая

$$\Gamma(x) = \Big\{ \big(t, x(t), x'(t)\big) \mid t \in [a, b] \Big\}$$

содержится в U. Функция $x \in \Omega^o$, удовлетворяющая краевым условиям (2), называется *планом* вариационной задачи (1). Требуется найти план, доставляющий минимум функционалу $J(x)$.

2°. В пространстве $C^1[a, b]$ введём норму

$$\|x\|_1 = \max_{t\in[a,b]}\big|x(t)\big| + \max_{t\in[a,b]}\big|x'(t)\big|.$$

В докладе [1] показано, что естественная область определения Ω^o функционала $J(x)$ открыта в $C^1[a, b]$ в том смысле, что вместе с x_0 содержит при

*Семинар «CNSA & NDO». Избранные доклады. 8 декабря 2016 г.

некотором $\delta > 0$ все функции вида $x_0 + h$ с $\|h\|_1 \leqslant \delta$. Там же установлено разложение

$$J(x_0 + h) = J(x_0) + \ell(x_0; h) + o\big(\|h\|_1\big), \tag{3}$$

где

$$\ell(x_0; h) = \int_a^b \Big[F'_x\big(t, x_0(t), x'_0(t)\big)h(t) + F'_{x'}\big(t, x_0(t), x'_0(t)\big)h'(t)\Big]dt \tag{4}$$

и $o\big(\|h\|_1\big)/\|h\|_1 \to 0$ при $\|h\|_1 \to 0$. Линейный по h функционал $\ell(x_0; h)$ определён при всех $h \in C^1[a, b]$. Он называется *первым дифференциалом функционала $J(x)$ в точке x_0* и обозначается $dJ(x_0; h)$.

3°. Вернёмся к задаче (1), (2) и введём множество

$$C_0^1[a, b] = \big\{h \in C^1[a, b] \mid h(a) = 0,\ h(b) = 0\big\}.$$

Ясно, что вместе с планом $x_0 \in \Omega$ множество Ω содержит функции вида $x_0 + \alpha h$ при всех $h \in C_0^1[a, b]$ и малых α.

План x_0 называется *точкой локального минимума* функционала $J(x)$, если при некотором $\delta > 0$ выполняется неравенство $J(x_0 + h) \geqslant J(x_0)$ для всех $h \in C_0^1[a, b]$, таких, что $\|h\|_1 \leqslant \delta$. Отсюда, в частности, следует, что в точке локального минимума

$$J(x_0 + \alpha h) \geqslant J(x_0) \tag{5}$$

для любой фиксированной функции $h \in C_0^1[a, b]$ и малых α.

Выясним, какими свойствами обладает точка локального минимума.

ТЕОРЕМА 1. *Если $x_0 \in \Omega$ — точка локального минимума функционала $J(x)$, то*

$$dJ(x_0; h) = 0 \quad \forall h \in C_0^1[a, b]. \tag{6}$$

Доказательство. При $h(t) \equiv 0$ равенство (6) тривиально (см. формулу (4)). Возьмём $h_0 \in C_0^1[a, b]$, $h_0(t) \not\equiv 0$ на $[a, b]$. Согласно (5) и (3) при малых $\alpha > 0$ имеем

$$0 \leqslant J(x_0 + \alpha h_0) - J(x_0) = dJ(x_0; \alpha h_0) + o\big(\|\alpha h_0\|_1\big) =$$
$$= \alpha\Big[dJ(x_0; h_0) + \frac{o\big(\|\alpha h_0\|_1\big)}{\|\alpha h_0\|_1} \cdot \|h_0\|_1\Big].$$

Поделим это неравенство на $\alpha > 0$ и перейдём к пределу при $\alpha \to +0$. Получим $dJ(x_0; h_0) \geqslant 0$. Отметим, что вместе с h_0 множеству $C_0^1[a, b]$ принадлежит и $-h_0$. Значит, $dJ(x_0; -h_0) \geqslant 0$ или $-dJ(x_0; h_0) \geqslant 0$. Объединив два неравенства $dJ(x_0; h_0) \geqslant 0$ и $-dJ(x_0; h_0) \geqslant 0$, придём к равенству $dJ(x_0; h_0) = 0$.

Теорема доказана. □

4°. Формула (4) позволяет переписать условие (6) так:

$$\int_a^b \big[F'_x h + F'_{x'} h'\big]\,dt = 0 \quad \forall h \in C_0^1[a,b]. \tag{7}$$

Обозначим

$$u(t) = F'_{x'}\big(t, x_0(t), x_0'(t)\big), \quad v(t) = F'_x\big(t, x_0(t), x_0'(t)\big).$$

Тогда равенство (7) примет вид

$$\int_a^b \big[u(t)h'(t) + v(t)h(t)\big]\,dt = 0 \quad \forall h \in C_0^1[a,b]. \tag{8}$$

Воспользуемся основной леммой вариационного исчисления (см. например [2]), согласно которой условие (8) выполняется тогда и только тогда, когда $u \in C^1[a,b]$ и $u'(t) = v(t)$ на $[a,b]$. Придём к следующему заключению.

ТЕОРЕМА 2. *Если $x_0 \in \Omega$ — точка локального минимума функционала $J(x)$, то необходимо*

1) $u(t) := F'_{x'}\big(t, x_0(t), x_0'(t)\big) \in C^1[a,b]$;
2) $\frac{d}{dt} F'_{x'}\big(t, x_0(t), x_0'(t)\big) = F'_x\big(t, x_0(t), x_0'(t)\big)$ *при всех* $t \in [a,b]$.

Получили необходимые условия оптимальности первого порядка. Их смысл в том, что точка локального минимума функционала $J(x)$ является решением краевой задачи

$$\frac{d}{dt} F'_{x'}\big(t, x(t), x'(t)\big) = F'_x\big(t, x(t), x'(t)\big), \tag{9}$$

$$x(a) = A, \quad x(b) = B. \tag{10}$$

Дифференциальное уравнение (9) называется *уравнением Эйлера*.

Введём ещё два понятия:

- *экстремаль* — любая функция класса $C^1[a,b]$, удовлетворяющая уравнению Эйлера;
- *стационарная кривая* — экстремаль, удовлетворяющая краевым условиям (10).

Обычно ищут стационарную кривую, после чего пытаются доказать, что она является либо решением задачи (1), (2), либо, по крайней мере, точкой локального минимума функционала $J(x)$. Следует отметить также принципиальный факт: если решение задачи (1), (2) существует и стационарная кривая единственна, то эта стационарная кривая будет единственным решением задачи (1), (2).

5°. Возьмём произвольную экстремаль $x = x(t)$. В случае, когда $F \in C^2(U)$ и $x \in C^2[a,b]$, можно записать

$$\frac{d}{dt}F'_{x'} = F''_{x't} + F''_{x'x}\,x' + F''_{x'x'}\,x''.$$

Таким образом, при указанных условиях экстремаль удовлетворяет дифференциальному уравнению второго порядка

$$F''_{x'x'}\,x'' = F'_x - F''_{x't} - F''_{x'x}\,x'.$$

Однако экстремаль может не принадлежать классу $C^2[a,b]$. Соответствующий пример приведён в докладе [3]. Ясность в этом вопросе наводит следующее утверждение.

ТЕОРЕМА ГИЛЬБЕРТА О ДИФФЕРЕНЦИРУЕМОСТИ. *Пусть $F \in C^2(U)$. Тогда экстремаль $x_0(t)$ дважды непрерывно дифференцируема на множестве*

$$T = \Big\{t \in [a,b] \mid F''_{x'x'}\big(t, x_0(t), x'_0(t)\big) \neq 0\Big\}.$$

Доказательство. Зафиксируем точку $t_0 \in T$. Равенство (9), определяющее экстремаль, представляет собой формулу для производной по t функции $F'_{x'}\big(t, x_0(t), x'_0(t)\big)$. По определению производной разностное отношение

$$\frac{1}{\Delta t}\Big[F'_{x'}\big(t_0+\Delta t, x_0(t_0+\Delta t), x'_0(t_0+\Delta t)\big) - F'_{x'}\big(t_0, x_0(t_0), x'_0(t_0)\big)\Big] \tag{11}$$

при $\Delta t \to 0$ имеет предел, равный $F'_x\big(t_0, x_0(t_0), x'_0(t_0)\big)$. Воспользуемся теоремой о среднем, согласно которой выражение (11) допускает представление

$$\tilde F''_{x't} + \tilde F''_{x'x}\frac{x_0(t_0+\Delta t)-x_0(t_0)}{\Delta t} + \tilde F''_{x'x'}\frac{x'_0(t_0+\Delta t)-x'_0(t_0)}{\Delta t}, \tag{12}$$

где $\tilde F''_{x't}$, $\tilde F''_{x'x}$, $\tilde F''_{x'x'}$ суть частные производные $F''_{x't}$, $F''_{x'x}$, $F''_{x'x'}$, вычисленные в средней точке

$$\Big(t_0+\theta\Delta t, x_0(t_0)+\theta\big(x_0(t_0+\Delta t)-x_0(t_0)\big),\ x'_0(t_0)+\theta\big(x'_0(t_0+\Delta t)-x'_0(t_0)\big)\Big).$$

Так как $t_0 \in T$, то $\tilde F''_{x'x'} \neq 0$ при малых Δt.

Запишем очевидное равенство

$$\frac{x'_0(t_0+\Delta t)-x'_0(t_0)}{\Delta t} = \frac{1}{\tilde F''_{x'x'}}\Big[\tilde F''_{x'x'}\frac{x'_0(t_0+\Delta t)-x'_0(t_0)}{\Delta t} +$$
$$+\tilde F''_{x'x}\frac{x_0(t_0+\Delta t)-x_0(t_0)}{\Delta t} + \tilde F''_{x't}\Big] - \frac{1}{\tilde F''_{x'x'}}\Big[\tilde F''_{x'x}\frac{x_0(t_0+\Delta t)-x_0(t_0)}{\Delta t} + \tilde F''_{x't}\Big].$$

В первой квадратной скобке из первой части этого равенства стоит выражение (12), которое имеет предел при $\Delta t \to 0$. Имеет предел и выражение, стоящее во второй квадратной скобке. Отсюда следует, что экстремаль $x_0(t)$ имеет вторую производную в точке $t = t_0$, причём

$$x_0''(t_0) = \frac{1}{F''_{x'x'}}\Big[F'_x - F''_{x'x}\,x_0'(t_0) - F''_{x't}\Big]. \tag{13}$$

Частные производные функции F вычисляются в точке $\big(t_0, x_0(t_0), x_0'(t_0)\big)$.

Правая часть формулы (13) непрерывна на T. Значит, $x_0 \in C^2(T)$.

Теорема доказана. □

Следствие. Пусть на экстремали x_0 выполняется условие

$$F''_{x'x'}\big(t_0, x_0(t_0), x_0'(t_0)\big) \neq 0 \quad \forall t \in [a, b].$$

Тогда $x_0 \in C^2[a, b]$.

6°. В связи с теоремой Гильберта вводятся следующие определения.

Функционал $J(x)$ вида (1) с $F \in C^2(U)$ называется

- *регулярным*, если $F''_{x'x'}(t, x, y) \neq 0$ на U;
- *положительно регулярным*, если $F''_{x'x'}(t, x, y) > 0$ на U;
- *отрицательно регулярным*, если $F''_{x'x'}(t, x, y) < 0$ на U.

У регулярного функционала все экстремали принадлежат классу $C^2[a, b]$. Положительно регулярный функционал появляется, например, в задаче о минимальной поверхности вращения (см. доклад [4]).

7°. Сделаем замечание общего характера. Пусть $x \in C^2[a, b]$. Тогда

$$\frac{d}{dt}\big(F - x'F'_{x'}\big) = \big(F'_t + F'_x\,x' + F'_{x'}x''\big) - \big(x''F'_{x'} - x'\frac{d}{dt}F'_{x'}\big) =$$
$$= F'_t - x'\left(\frac{d}{dt}F'_{x'} - F'_x\right).$$

Отсюда следует, что решение краевой задачи

$$\begin{aligned} &\frac{d}{dt}\big(F - x'F'_{x'}\big) = F'_t, \\ &x(a) = A, \quad x(b) = B, \end{aligned} \tag{14}$$

принадлежащее $C^2[a, b]$, у которого производная обращается в ноль лишь в конечном числе точек отрезка $[a, b]$, является стационарной кривой.

В случае $F = F(x, x')$ задача (14) принимает вид

$$\begin{gathered} F - x' F'_{x'} \equiv \text{const}, \\ x(a) = A, \quad x(b) = B. \end{gathered} \tag{15}$$

Отметим, что нелинейное дифференциальное уравнение (15) имеет первый порядок.

8°. Переходим к необходимым условиям оптимальности второго порядка. Предположим, что $F \in C^2(U)$. Возьмём $x_0 \in \Omega^{\text{o}}$ и воспользуемся тем, что функционал $J(x)$ вида (1) допускает разложение (см. [1])

$$J(x_0 + h) = J(x_0) + dJ(x_0; h) + \tfrac{1}{2} D(x_0; h) + o\big(\|h\|_1^2\big), \tag{16}$$

где

$$D(x_0; h) = \int_a^b \big[F''_{xx} h^2 + 2F''_{xx'} hh' + F''_{x'x'} (h')^2\big]\, dt \tag{17}$$

(частные производные функции F вычисляются в точке $\big(t, x_0(t), x'_0(t)\big)$) и $o\big(\|h\|_1^2\big)/\|h\|_1^2 \to 0$ при $\|h\|_1 \to 0$. Интегральная квадратичная форма $D(x_0; h)$ определена при всех $h \in C^1[a, b]$. Она называется *вторым дифференциалом функционала $J(x)$ в точке x_0* и обозначается $d^2 J(x_0; h)$.

ТЕОРЕМА 3. *Если $x_0 \in \Omega$ — точка локального минимума функционала $J(x)$, то*

$$d^2 J(x_0; h) \geqslant 0 \quad \forall h \in C_0^1[a, b]. \tag{18}$$

Доказательство. При $h(t) \equiv 0$ равенство (18) очевидно (см. формулу (17)). Возьмём $h_0 \in C_0^1[a, b]$, $h_0(t) \not\equiv 0$ на $[a, b]$. По теореме 1, $dJ(x_0, h_0) = 0$. При малых ненулевых α функция $x_0 + \alpha h_0$ является планом задачи (1), (2). Согласно (16),

$$\begin{aligned} 0 \leqslant J(x_0 + \alpha h_0) - J(x_0) &= \tfrac{1}{2} D(x_0; \alpha h_0) + o\big(\|\alpha h_0\|_1^2\big) = \\ &= \alpha^2 \Big[\tfrac{1}{2} D(x_0; h_0) + \frac{o\big(\|\alpha h_0\|_1^2\big)}{\|\alpha h_0\|_1^2} \cdot \|h_0\|_1^2\Big]. \end{aligned}$$

Поделим это неравенство на α^2, после чего перейдём к пределу при $\alpha \to 0$. Получим $\frac{1}{2} D(x_0; h_0) \geqslant 0$, что равносильно (18). □

Получили необходимое условие оптимальности второго порядка. Оно заключается в том, что в точке локального минимума x_0 функционала $J(x)$ интегральная квадратичная форма $D(x_0; h)$ должна быть неотрицательно определённой на множестве $C_0^1[a, b]$.

9°. Вопрос о неотрицательной определённости интегральной квадратичной формы вида

$$D(h) = \int_a^b \big[p(h')^2 + qh^2\big]\,dt, \quad h \in C_0^1[a,b], \tag{19}$$

рассматривался в докладе [2]. Там приведено доказательство критерия неотрицательной определённости при следующих предположениях:

$$p \in C^1[a,b], \quad q \in C[a,b], \quad p(t) > 0 \text{ на } [a,b]. \tag{20}$$

В формулировке критерия участвует функция $h_0(t)$ — решение дифференциального уравнения

$$(ph')' = qh, \tag{21}$$

удовлетворяющее начальным условиям

$$h(a) = 0, \quad h'(a) = 1.$$

Уравнение (21) называется *уравнением Якоби*, а функция $h_0(t)$ — *главным решением уравнения Якоби*.

Напомним формулировку критерия неотрицательной определённости.

ТЕОРЕМА ЯКОБИ. *В предположениях* (20) *квадратичная форма* $D(h)$ *вида* (19) *неотрицательно определена на* $C_0^1[a,b]$ *тогда и только тогда, когда главное решение уравнения Якоби положительно на интервале* (a,b).

10°. Чтобы применить теорему Якоби к форме $D(x_0;h)$, нужно сначала привести эту форму к каноническому виду (19). Обозначим

$$p(t) = F''_{x'x'}\big(t, x_0(t), x_0'(t)\big),\ u(t) = F''_{xx'}\big(t, x_0(t), x_0'(t)\big),\ v(t) = F''_{xx}\big(t, x_0(t), x_0'(t)\big).$$

В силу (17)

$$D(x_0;h) = \int_a^b \big[p(h')^2 + 2uhh' + vh^2\big]\,dt. \tag{22}$$

Предположим, что $p(t) > 0$ на $[a,b]$, то есть

$$F''_{x'x'}\big(t, x_0(t), x_0'(t)\big) > 0 \quad \forall t \in [a,b].$$

В этом случае по теореме Гильберта о дифференцируемости стационарная кривая $x_0(t)$ принадлежит классу $C^2[a,b]$. В частности, при $F \in C^3(U)$ функции $p(t)$ и $u(t)$ непрерывно дифференцируемы на $[a,b]$.

Для любой функции $h \in C_0^1[a,b]$ имеем

$$2\int_a^b uhh'dt = \int_a^b udh^2 = -\int_a^b u'h^2dt.$$

Обозначим $q = v - u'$ и перепишем формулу (22) в терминах p и q:

$$D(x_0; h) = \int_a^b \bigl[p(h')^2 + qh^2\bigr]dt.$$

Квадратичная форма $D(x_0; h)$ приведена к каноническому виду, причём выполнены условия (20).

На основании теоремы 3 и теоремы Якоби приходим к следующему заключению.

ТЕОРЕМА 4. *Пусть $F \in C^3(U)$. Если $x_0 \in \Omega$ — точка локального минимума функционала $J(x)$ и*

$$F''_{x'x'}\bigl(t, x_0(t), x_0'(t)\bigr) > 0 \quad \forall t \in [a, b],$$

то функция $h_0(t)$, удовлетворяющая дифференциальному уравнению

$$(ph')' = (v - u')h \tag{23}$$

и начальным условиям $h(a) = 0$, $h'(a) = 1$, положительна на интервале (a, b).

Это другая форма необходимого условия оптимальности второго порядка.

11°. В некоторых случаях функцию $h_0(t)$ из теоремы 4 можно построить даже не выписывая уравнение Якоби (23), а используя только уравнение Эйлера.

Сохранив предположение $F \in C^3(U)$, рассмотрим параметрическую задачу Коши для уравнения Эйлера

$$\begin{gathered}\frac{d}{dt}F'_{x'}\bigl(t, x(t), x'(t)\bigr) = F'_x\bigl(t, x(t), x'(t)\bigr),\\ x(a) = A, \quad x'(a) = \alpha.\end{gathered} \tag{24}$$

Здесь α — параметр. Пусть при $\alpha = \alpha_0$ решением задачи (24) является стационарная кривая $x_0(t)$. В общем случае решение задачи (24) обозначим через $x(t, \alpha)$. В частности, $x(t, \alpha_0) = x_0(t)$.

Дальнейшее справедливо лишь тогда, когда функция двух переменных $x(t, \alpha)$ трижды непрерывно дифференцируема на множестве

$$[a, b] \times (\alpha_0 - \varepsilon, \alpha_0 + \varepsilon)$$

при некотором $\varepsilon > 0$.

Подставив в (24) $x(t,\alpha)$ вместо $x(t)$, получим тождества по α. Продифференцируем первое из них по α, после чего подставим $\alpha=\alpha_0$. Придём к соотношению

$$\frac{d}{dt}\big[F''_{x'x}\cdot x'_\alpha(t,\alpha_0)+F''_{x'x'}\cdot x''_{t\alpha}(t,\alpha_0)\big]=F''_{xx}\cdot x'_\alpha(t,\alpha_0)+F''_{xx'}\cdot x''_{t\alpha}(t,\alpha_0)$$

(частные производные функции F вычисляются в точке $\big(t,x_0(t),x'_0(t)\big)$). Воспользуемся принятыми ранее обозначениями и перепишем последнее равенство в виде

$$\frac{d}{dt}[ux'_\alpha+px''_{t\alpha}]=vx'_\alpha+ux''_{t\alpha}.$$

Так как

$$\frac{d}{dt}(ux'_\alpha)=u'x'_\alpha+ux''_{\alpha t}$$

и $x''_{\alpha t}(t,x_0)=x''_{t\alpha}(t,x_0)$, то

$$\frac{d}{dt}(px''_{\alpha t})=(v-u')x'_\alpha. \tag{25}$$

Обозначим

$$h_0(t)=x'_\alpha(t,\alpha_0). \tag{26}$$

Согласно (25)

$$(ph'_0)'=(v-u')h_0,$$

то есть функция h_0 удовлетворяет уравнению Якоби (23).

Теперь обратимся к тождествам

$$x(a,\alpha)=A,\quad x'_t(a,\alpha)=\alpha.$$

Продифференцируем их по α, после чего подставим $\alpha=\alpha_0$. Получим

$$x'_\alpha(a,\alpha_0)=0,\quad x''_{t\alpha}(a,\alpha_0)=1$$

или в других обозначениях $h_0(a)=0$, $h'_0(a)=1$.

Таким образом, функция $h_0(t)$ вида (26) удовлетворяет всем условиям из теоремы 4.

12°. В качестве примера рассмотрим вариационную задачу

$$\begin{gathered}J(x):=\int_0^1\frac{x}{(x')^2}dt\to\inf\\ x(0)=1,\quad x(1)=4.\end{gathered} \tag{27}$$

Построим стационарную кривую для этой задачи и проверим выполнение на ней необходимого условия оптимальности второго порядка.

В данном случае $F(t,x,y) = x/y^2$. Функция F определена при $y \neq 0$. Принимая во внимание краевые условия, открытое линейно связное множество U введём следующим образом:

$$U = \big\{(t,x,y) \mid x > 0,\ y > 0\big\}.$$

Очевидно, что $F \in C^3(U)$.

Вычислим частные производные:

$$F'_x = \frac{1}{y^2}, \quad F'_y = -\frac{2x}{y^3}, \quad F''_{yy} = \frac{6x}{y^4}.$$

Так как $F''_{yy} > 0$ на U, то функционал $J(x)$ положительно регулярный. По теореме Гильберта о дифференцируемости все экстремали задачи (27) принадлежат классу $C^2[0,1]$.

Функция F не зависит явно от t. Это позволяет записать уравнение Эйлера в виде (15):

$$\frac{x}{(x')^2} + x'\frac{2x}{(x')^3} = \frac{3}{4a^2} \qquad (a > 0)$$

или $4a^2x = (x')^2$. В соответствии с определением множества U мы ищем экстремали со свойствами $x > 0$, $x' > 0$. Приходим к дифференциальному уравнению

$$x' = 2a\sqrt{x}.$$

Его решением является двухпараметрическое семейство функций

$$x(t) = (at + b)^2 \tag{28}$$

при $a > 0$. Коэффициенты a и b найдём из граничных условий

$$b^2 = 1, \quad (a+b)^2 = 4.$$

При $b = 1$ получаем положительное $a = 1$. При $b = -1$ будет $a = 3$.

Таким образом, выделены две функции (см. рис.)

$$x_0(t) = (t+1)^2, \quad x_1(t) = (3t-1)^2.$$

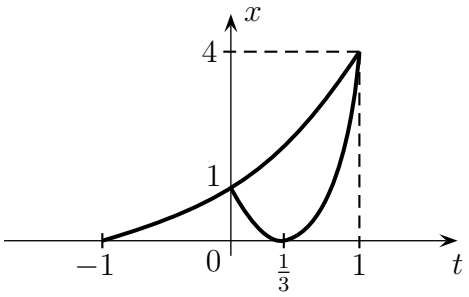


Рис. Графики функций $x_0(t)$ и $x_1(t)$.

Функция $x_1(t)$ — побочная. У неё производная обращается в ноль при $t = \frac{1}{3}$. Что касается функции $x_0(t)$, то она принадлежит классу $C^2[0,1]$ и её производная не обращается в ноль на отрезке $[0,1]$. Это гарантирует стационарность $x_0(t)$ (см. п. 7°). Покажем, что на x_0 выполняется необходимое условие оптимальности второго порядка.

В двухпараметрическом семействе экстремалей (28) выделим однопараметрическое семейство, удовлетворяющее начальным условиям

$$x(0) = 1, \quad x'(0) = \alpha.$$

Данные условия приводят к равенствам $b^2 = 1$, $2ab = \alpha$. Учитывая их, записываем

$$x(t,\alpha) = \Big(\frac{\alpha}{2b}t + b\Big)^2 = b^2\Big(\frac{\alpha}{2b^2}t + 1\Big)^2 = \Big(\frac{\alpha}{2}t + 1\Big)^2.$$

Стационарная кривая x_0 получается при $\alpha = 2$, то есть $\alpha_0 = 2$. Согласно (26)

$$h_0(t) = x'_\alpha(t,\alpha_0) = t(t+1).$$

Очевидно, что $h_0(t) > 0$ при всех $t \in (0,1)$. Это и требовалось установить.

Вычислим значение функционала $J(x)$ на стационарной кривой x_0:

$$J(x_0) = \int_0^1 \frac{(t+1)^2}{4(t+1)^2}dt = \frac{1}{4}.$$

Вместе с тем, на побочной кривой x_1, являющейся планом задачи (27), имеем

$$J(x_1) = \int_0^1 \frac{(3t-1)^2}{3t(3t-1)^2}dt = \frac{1}{36},$$

то есть $J(x_1)$ в девять раз меньше, чем $J(x_0)$! К сожалению, кривая x_1 находится вне теории. На ней подынтегральная функция F имеет устранимую особенность.

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Первый и второй дифференциалы интегрального функционала* // Семинар «DHA & CAGD». Избранные доклады. 5 декабря 2013 г. (`http://dha.spb.ru/reps13.shtml#1205`) [Данная книга, с. 357]

2. Малозёмов В. Н. *Квадратичные вариационные задачи* // Вестник молодых учёных. Прикл. мат. и мех. 2000. № 3. С. 12–22.

3. Малозёмов В. Н., Тамасян Г. Ш. *Об одной кубической вариационной задаче* // Семинар «CNSA & NDO». Избранные доклады. 11 февраля 2016 г. (`http://apmath.spbu.ru/cnsa/reps16.shtml#0211`) [Данная книга, с. 346]

4. Малозёмов В. Н. *О минимальной поверхности вращения* // Вестник СПбГУ. Сер. 10. 2006. Вып. 1. С. 52–56.

ДОСТАТОЧНЫЕ УСЛОВИЯ СТРОГОГО ЛОКАЛЬНОГО МИНИМУМА В КЛАССИЧЕСКОЙ ВАРИАЦИОННОЙ ЗАДАЧЕ*

В. Н. Малозёмов

Аннотация. Этот доклад примыкает к докладам [1] и [2]. При выводе достаточных условий строгого локального минимума используются первый и второй дифференциалы нелинейного интегрального функционала.

1°. Рассмотрим классическую вариационную задачу

$$J(x) := \int_a^b F\big(t, x(t), x'(t)\big)dt \to \inf \tag{1}$$

$$x(a) = A, \quad x(b) = B. \tag{2}$$

Считаем, что $F \in C^3(U)$, где $U \subset \mathbb{R}^3$ — открытое линейно связное множество. Напомним некоторые определения и результаты, связанные с задачей (1), (2).

Функция $x \in C^1[a,b]$ называется *допустимой*, если она удовлетворяет краевым условиям (2) и если параметрическая кривая

$$Z(x) = \Big\{\big(t, x(t), x'(t)\big) \mid t \in [a,b]\Big\}$$

содержится в U. Множество допустимых функций обозначим Ω.

Первым дифференциалом функционала $J(x)$ в точке $x_0 \in \Omega$ называется линейный интегральный функционал

$$\ell(x_0; h) = \int_a^b \Big[F'_x\big(t, x_0(t), x'_0(t)\big)h(t) + F'_{x'}\big(t, x_0(t), x'_0(t)\big)h'(t)\Big]dt,$$

определённый на множестве $C^1[a,b]$.

Обозначим $C^1_0[a,b] = \big\{h \in C^1[a,b] \mid h(a) = 0, h(b) = 0\big\}$.

*Семинар «CNSA & NDO». Избранные доклады. 2 марта 2017 г.

УТВЕРЖДЕНИЕ 1. *Равенство*

$$\ell(x_0; h) = 0 \quad \forall h \in C_0^1[a, b]$$

справедливо тогда и только тогда, когда выполняются два условия:

1) *функция* $F'_{x'}\big(t, x_0(t), x_0'(t)\big)$ *непрерывно дифференцируема на* $[a, b]$*;*

2) $\frac{d}{dt}F'_{x'}\big(t, x_0(t), x_0'(t)\big) = F'_x\big(t, x_0(t), x_0'(t)\big)$, $t \in [a, b]$.

Доказательство приводится в [3, Основная лемма вариационного исчисления].

Функция $x_0 \in \Omega$, удовлетворяющая условиям 1) и 2), называется *стационарной*. Как известно [2, п. 4°], допустимая функция x_0, в которой достигается локальный минимум функционала $J(x)$ вида (1), является стационарной функцией. Имея в виду достаточные условия строгого локального минимума, в дальнейшем будем считать, что x_0 — стационарная функция. Более того, предположим, что на ней выполняется *усиленное условие Лежандра*:

$$p(t) := F''_{x'x'}\big(t, x_0(t), x_0'(t)\big) > 0 \quad \text{на} \quad [a, b]. \tag{3}$$

По теореме Гильберта о дифференцируемости [2, п. 5°] из (3) следует, что $x_0 \in C^2[a, b]$.

2°. Вторым дифференциалом функционала $J(x)$ в точке x_0 называется интегральная квадратичная форма

$$D(x_0; h) = \int_a^b \big[F''_{xx}h^2 + 2F''_{xx'}hh' + F''_{x'x'}(h')^2\big]dt, \tag{4}$$

определённая на множестве $C^1[a, b]$. Аргумент у частных производных функции F имеет вид $\big(t, x_0(t), x_0'(t)\big)$. Нас интересует вопрос о положительной определённости формы $D(x_0; h)$ на множестве $C_0^1[a, b]$.

Наряду с обозначением $p(t)$ (см. (3)) будем использовать ещё два обозначения

$$u(t) = F''_{xx'}\big(t, x_0(t), x_0'(t)\big), \quad v(t) = F''_{xx}\big(t, x_0(t), x_0'(t)\big).$$

Перепишем формулу (4) в новых обозначениях:

$$D(x_0; h) = \int_a^b \big[p(h')^2 + 2uhh' + vh^2\big]dt.$$

Напомним, что $F \in C^3(U)$. В силу (3), $x_0 \in C^2[a, b]$. Отсюда, в частности, следует, что $u \in C^1[a, b]$. При $h \in C_0^1[a, b]$ имеем

$$2\int_a^b uhh'dt = \int_a^b udh^2 = -\int_a^b u'h^2dt.$$

Значит,

$$D(x_0;h)=\int_a^b\big[p(h')^2+(v-u')h^2\big]dt,\quad h\in C_0^1[a,b]. \tag{5}$$

Квадратичная форма $D(x_0;h)$ приведена к каноническому виду.

Теперь можно воспользоваться теоремой Якоби [3, п. 3°]. Напомним, что функция $h_0\in C^1[a,b]$, удовлетворяющая уравнению Якоби

$$(ph')'=(v-u')h \tag{6}$$

и начальным условиям $h(a)=0$, $h'(a)=1$, называется *главным решением уравнения Якоби.*

ТЕОРЕМА ЯКОБИ. *Пусть $F\in C^3(U)$, x_0 — стационарная кривая и выполнено усиленное условие Лежандра* (3). *В этом случае интегральная квадратичная форма $D(x_0;h)$ будет положительно определённой на $C_0^1[a,b]$ тогда и только тогда, когда главное решение уравнения Якоби $h_0(t)$ положительно на полуоткрытом интервале $(a,b]$.*

Условие $h_0(t)>0$ при $t\in(a,b]$ называется *усиленным условием Якоби.*

Нам потребуется ещё один результат, характеризующий положительно определённую квадратичную форму $D(x_0;h)$ [3, п. 3°].

УТВЕРЖДЕНИЕ 2. *Если на стационарной кривой x_0 выполняются усиленное условие Лежандра и усиленное условие Якоби, то найдётся положительное число μ, такое что*

$$D(x_0;h)\geqslant\mu\int_a^b(h')^2dt\quad\forall h\in C_0^1[a,b]. \tag{7}$$

3°. Допустимая функция x_0 называется точкой строгого локального минимума в задаче (1), (2), если найдётся $\delta>0$, такое что при всех $h\in C_0^1[a,b]$ со свойствами $h(t)\not\equiv 0$ на $[a,b]$, $\|h\|_1\leqslant\delta$ справедливо неравенство $J(x_0+h)>J(x_0)$. Здесь

$$\|h\|_1=\max_{t\in[a,b]}\big|h(t)\big|+\max_{t\in[a,b]}\big|h'(t)\big|.$$

ТЕОРЕМА (о достаточных условиях строгого локального минимума). *Пусть $F\in C^3(U)$. Если x_0 — стационарная функция, на которой выполнены усиленное условие Лежандра и усиленное условие Якоби, то найдутся $\delta>0$ и $\alpha>0$, такие, что*

$$J(x_0+h)\geqslant J(x_0)+\alpha\int_a^b(h')^2dt \tag{8}$$

при всех $h\in C_0^1[a,b]$ со свойством $\|h\|_1\leqslant\delta$.

Отметим, что $\int_a^b (h')^2 dt > 0$ при $h \in C_0^1[a,b]$, $h(t) \not\equiv 0$ на $[a,b]$.

Доказательство. Воспользуемся разложением [1, п. 5°]

$$J(x_0+h) = J(x_0) + \ell(x_0;h) + \tfrac{1}{2}D(x_0;h) + \omega_2(x_0;h). \tag{9}$$

Здесь $\ell(x_0;h)$, $D(x_0;h)$ — первый и второй дифференциалы функционала $J(x)$ в точке x_0 и $\omega_2(x_0;h)$ — остаточный член, обладающий следующим свойством: по $\varepsilon > 0$ найдётся $\delta > 0$, такое что

$$\big|\omega_2(x_0;h)\big| \leqslant \frac{\varepsilon}{b-a}\int_a^b \big[|h|^2 + 2|hh'| + |h'|^2\big]dt \tag{10}$$

при условии $\|h\|_1 \leqslant \delta$. Нас интересуют приращения h только из $C_0^1[a,b]$, ибо только в этом случае функция $x_0 + h$ удовлетворяет краевым условиям (2). То, что x_0 — стационарная кривая, гарантирует равенство $\ell(x_0;h) = 0$ при всех $h \in C_0^1[a,b]$. На основании (9) и (7) получаем

$$J(x_0+h) - J(x_0) \geqslant \tfrac{1}{2}D(x_0;h) - \big|\omega_2(x_0;h)\big| \geqslant \tfrac{1}{2}\mu\int_a^b (h')^2 dt - \big|\omega_2(x_0;h)\big|. \tag{11}$$

Покажем, что найдётся такое $\delta > 0$, что для $h \in C_0^1[a,b]$, $\|h\|_1 \leqslant \delta$ выполняется неравенство

$$\big|\omega_2(x_0;h)\big| \leqslant \tfrac{1}{4}\mu\int_a^b (h')^2 dt. \tag{12}$$

Тогда из (11) будет следовать требуемое неравенство (8) с $\alpha = \frac{1}{4}\mu$.

Обратимся к формуле (10) и проверим, что при всех $h \in C_0^1[a,b]$

$$\int_a^b \big[|h|^2 + 2|hh'| + |h'|^2\big]dt \leqslant (b-a+1)^2\int_a^b (h')^2 dt. \tag{13}$$

Действительно, имеем $h(t) = \int_a^t h'(\tau)d\tau$. По неравенству Коши – Буняковского

$$h^2(t) = \Big(\int_a^t 1\cdot h'(\tau)d\tau\Big)^2 \leqslant (b-a)\int_a^b (h')^2 dt.$$

Интегрируя, получаем

$$\int_a^b h^2 dt \leqslant (b-a)^2\int_a^b (h')^2 dt. \tag{14}$$

Далее

$$\int_a^b |hh'|dt \leqslant \Big(\int_a^b h^2 dt\Big)^{1/2}\Big(\int_a^b (h')^2 dt\Big)^{1/2} \leqslant (b-a)\int_a^b (h')^2 dt. \tag{15}$$

На основании (14) и (15) приходим к (13).

Положим

$$\varepsilon = \frac{\mu(b-a)}{4(b-a+1)^2}.$$

Как отмечалось, по этому ε найдётся $\delta \geqslant 0$, такое, что при $h \in C_0^1[a,b]$, $\|h\|_1 \leqslant \delta$ выполняется неравенство (10). Теперь из (10) и (13) следует неравенство (12).

Теорема доказана. □

4°. В качестве примера на использование условий оптимальности рассмотрим задачу Эйлера:

$$J(x) := \int_0^1 \left[(x')^2 + \sigma \cos x\right] dt \to \inf$$
$$x(0) = 0, \quad x(1) = 0,$$

где $\sigma > 0$ — параметр. Требуется найти точную верхнюю границу тех σ, при которых функция $x_0(t) \equiv 0$ на $[0,1]$ будет точкой строгого локального минимума функционала $J(x)$.

Имеем $F = (x')^2 + \sigma \cos x$, $F'_x = -\sigma \sin x$, $F'_{x'} = 2x'$,

$$F''_{xx} = -\sigma \cos x, \quad F''_{xx'} \equiv 0, \quad F''_{x'x'} \equiv 2.$$

Очевидно, что функция $x_0(t) \equiv 0$ на $[0,1]$ удовлетворяет условиям 1) и 2) утверждения 1 при всех $\sigma > 0$. Значит, при всех $\sigma > 0$ функция x_0 является стационарной. Выполнение усиленного условия Лежандра очевидно. Проверим, когда на x_0 выполняется усиленное условие Якоби.

В данном случае $p(t) \equiv 2$, $u(t) \equiv 0$, $v(t) = -\sigma$, так что уравнение Якоби имеет вид

$$2h'' + \sigma h = 0.$$

Начальным условиям $h(0) = 0$, $h'(0) = 1$ удовлетворяет функция

$$h_0(t) = \sqrt{\frac{2}{\sigma}} \sin \sqrt{\frac{\sigma}{2}} t.$$

По определению $h_0(t)$ — главное решение уравнения Якоби (см. рис.)

Обозначим через t_0 наименьший положительный корень фнукции $h_0(t)$. Ясно, что $t_0 = \pi\sqrt{\frac{2}{\sigma}}$. Если $t_0 > 1$ ($\sigma < 2\pi^2$), то выполнено усиленное условие Якоби. Значит, при $\sigma < 2\pi^2$ функция $x_0(t) \equiv 0$ на $[0,1]$ является точкой строгого локального минимума. При $\sigma > 2\pi^2$ ($t_0 < 1$) нарушается необходимое условие локального минимума второго порядка [2, п. 10°], то есть x_0 не будет

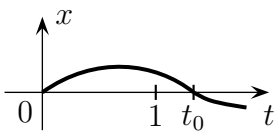


Рис. Графики функции $h_0(t)$.

даже точкой локального минимума. Из сказанного следует, что точная верхняя граница тех $\sigma > 0$, при которых x_0 является точкой строгого локального минимума функционала $J(x)$, равна $2\pi^2$.

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Первый и второй дифференциалы интегрального функционала* // Семинар «DHA & CAGD». Избранные доклады. 5 декабря 2013 г. (`http://dha.spb.ru/reps13.shtml#1205`) [Данная книга, с. 357]

2. Малозёмов В. Н. *Необходимые условия оптимальности первого и второго порядков в простейшей нелинейной задаче вариационного исчисления* // Семинар «CNSA & NDO». Избранные доклады. 8 декабря 2016 г. (`http://apmath.spbu.ru/cnsa/reps16.shtml#1208`) [Данная книга, с. 364]

3. Малозёмов В. Н. *Квадратичные вариационные задачи* // Вестник молодых учёных. Прикл. мат. и мех. 2000. № 3. С. 12–22.

ДИФФЕРЕНЦИРУЕМОСТЬ ПО ФРЕШЕ ОДНОГО НЕЛИНЕЙНОГО ФУНКЦИОНАЛА*

М. В. Долгополик

Аннотация. В докладе обсуждается дифференцируемость по Фреше одного нелинейного функционала, возникающего в задачах вариационного исчисления в связи с изучением точных штрафных функций и метода наискорейшего спуска.

1°. Первый и второй дифференциалы интегрального функционала. Рассмотрим интегральный функционал вида

$$J(x) = \int_a^b F\big(t, x(t), x'(t)\big)\, dt. \tag{1}$$

Здесь $F(t, y, z)$ — функция трёх переменных, заданная и непрерывная на некотором открытом связном множестве $U \subset \mathbb{R}^3$. Функционал $J(x)$ определён на функциях $x = x(t)$, непрерывно дифференцируемых на отрезке $[a, b]$, и таких, что параметрическая кривая

$$\Gamma(x) = \Big\{ \big(t, x(t), x'(t)\big) \;\Big|\; t \in [a, b] \Big\}$$

содержится в U. Множество таких функций x обозначим через Ω^o и назовём *естественной областью определения* функционала $J(x)$.

В линейном пространстве $C^1[a, b]$ непрерывно дифференцируемых на отрезке $[a, b]$ функций введём норму

$$\|x\|_{1,\infty} = \max_{t\in[a,b]} \big|x(t)\big| + \max_{t\in[a,b]} \big|x'(t)\big|.$$

В стандартном курсе вариационного исчисления доказывается справедливость следующих утверждений (см., например, доклад [1]).

ТЕОРЕМА 1. *Естественная область определения Ω^o функционала $J(x)$ открыта в $C^1[a, b]$.*

*Семинар «CNSA & NDO». Избранные доклады. 3 марта 2016 г.

ТЕОРЕМА 2. *При $F \in C^1(U)$ интегральный функционал $J(x)$ дифференцируем по Фреше в каждой точке x_0 своей естественной области определения Ω^o и*

$$dJ(x_0;h) = \int_a^b \Big[F'_x\big(t,x_0(t),x'_0(t)\big)h(t) + F'_{x'}\big(t,x_0(t),x'_0(t)\big)h'(t)\Big]\,dt.$$

При этом справедливо разложение

$$J(x_0+h) = J(x_0) + dJ(x_0;h) + o(\|h\|_{1,\infty}),$$

где $o(\|h\|_{1,\infty})/\|h\|_{1,\infty} \to 0$ при $\|h\|_{1,\infty} \to 0$.

ТЕОРЕМА 3. *При $F \in C^2(U)$ функционал $J(x)$ дважды дифференцируем по Фреше в каждой точке x_0 своей естественной области определения Ω^o и*

$$d^2J(x_0;h) = \int_a^b \left[F''_{xx}h^2 + 2F''_{xx'}hh' + F''_{x'x'}(h')^2\right]dt,$$

где

$$F''_{xx} = F''_{xx}\big(t,x_0(t),x'_0(t)\big), \quad F''_{xx'} = F''_{xx'}\big(t,x_0(t),x'_0(t)\big),$$
$$F''_{x'x'} = F''_{x'x'}\big(t,x_0(t),x'_0(t)\big).$$

При этом справедливо разложение

$$J(x_0+h) = J(x_0) + dJ(x_0;h) + \tfrac{1}{2}d^2J(x_0;h) + o(\|h\|^2_{1,\infty}),$$

где $o(\|h\|^2_{1,\infty})/\|h\|^2_{1,\infty} \to 0$ при $\|h\|_{1,\infty} \to 0$.

2°. Дифференцируемость по Фреше одного нелинейного функционала. В книге [2] был разработан альтернативный подход к исследованию задач вариационного исчисления, основанный на применении теории точных штрафных функций. В рамках данного подхода рассматривается интегральный функционал вида

$$I(z) = \int_a^b F\Big(t, A + \int_a^t z(\tau)\,d\tau, z(t)\Big)\,dt,$$

где A — константа, соответствующая граничному условию при $t = a$.

Функционал $I(z)$ получается из интегрального функционала $J(x)$ вида (1) с помощью замены $x' \to z$. Данный переход позволяет преобразовать стандартные граничные условия

$$x(a) = A, \quad x(b) = B$$

в линейное ограничение–равенство вида

$$\int_a^b z(t)\,dt = B - A,$$

которое проще учитывать при построении численных методов [3, 4]. Однако, следует отметить, что аналогичные численные методы можно построить и без перехода от функционала $J(x)$ к функционалу $I(z)$ (см., например, [5]).

В книге [2] и последующих работах по применению точных штрафных функций к задачам вариационного исчисления доказывалась лишь дифференцируемость по Гато функционала $I(z)$. В данном разделе мы покажем, что при естественных предположениях функционал $I(z)$ является дифференцируемым по Фреше, как и функционал $J(x)$.

Напомним, что $F(t, y, z)$ — функциях трёх переменных, заданная и непрерывная на некотором открытом связном множестве $U \subset \mathbb{R}^3$. Функционал $I(z)$ определён на функциях $z = z(t)$, непрерывных на отрезке $[a, b]$, и таких, что параметрическая кривая

$$\Gamma_0(z) = \left\{ \left(t, A + \int_a^t z(\tau)\,d\tau, z(t) \right) \;\middle|\; t \in [a, b] \right\}$$

содержится в U. Множество таких функций z обозначим через Z^o и назовём *естественной областью определения* функционала $I(z)$. Отметим, что если функции x и z связаны соотношением $x(t) = A + \int_a^t z(\tau)\,d\tau$, то $\Gamma(x) = \Gamma_0(z)$.

В линейном пространстве $C[a, b]$ непрерывных на отрезке $[a, b]$ функций введём норму

$$\|z\|_\infty = \max_{t\in[a,b]} \big|z(t)\big|.$$

Определим также линейный интегральный оператор

$$\big(Tz\big)(t) = \int_a^t z(\tau)\,d\tau.$$

Нам потребуется следующее вспомогательное утверждение.

ЛЕММА 1. *Оператор T является линейным непрерывным оператором из $C[a, b]$ в $C^1[a, b]$. При этом $\|T\| \leqslant (b - a + 1)$.*

Доказательство. Ясно, что для любого $z \in C[a, b]$ будет $Tz \in C^1[a, b]$. Поскольку

$$\|Tz\|_{1,\infty} = \max_{t\in[a,b]} \left| \int_a^t z(\tau)\,d\tau \right| + \max_{t\in[a,b]} \big|z(t)\big| \leqslant (b - a + 1)\|z\|_\infty,$$

то оператор T непрерывен и $\|T\| \leqslant (b - a + 1)$. □

Покажем, что естественная область определения функционала $I(z)$ открыта в $C[a,b]$. Для этого зафиксируем функцию $z_0 \in Z^o$ и докажем, что найдётся $\delta > 0$ такое, что $z_0 + h \in Z^o$ для любого $h \in C[a,b]$, удовлетворяющего неравенству $\|h\|_\infty < \delta$.

Определим $x_0(t) = A + \int_a^t z_0(\tau)\,d\tau$. Как было отмечено выше, $\Gamma(x_0) = \Gamma_0(z_0)$. Поэтому $x_0 \in \Omega^o$. Следовательно, по теореме 1 найдётся $r > 0$ такое, что для любого $\eta \in C^1[a,b]$, удовлетворяющего неравенству $\|\eta\|_{1,\infty} < r$, будет $x_0 + \eta \in \Omega^o$.

Положим $\delta = r/(b-a+1)$. Тогда для любой функции $h \in C[a,b]$ такой, что $\|h\|_\infty < \delta$ будет $\|Th\|_{1,\infty} < r$ и $x_0 + Th \in \Omega^o$, то есть $\Gamma(x_0 + Th) \subset U$. Заметим, что $\Gamma_0(z_0 + h) = \Gamma(x_0 + Th)$. Поэтому $z_0 + h \in Z^o$. Значит, множество Z^o открыто в $C[a,b]$.

Таким образом, справедливо следующее утверждение.

ТЕОРЕМА 4. *Естественная область определения Z^o функционала $I(z)$ открыта в $C[a,b]$.*

Докажем теперь дифференцируемость по Фреше функционала $I(z)$.

ТЕОРЕМА 5. *При $F \in C^1(U)$ интегральный функционал $I(z)$ дифференцируем по Фреше в каждой точке z_0 своей естественной области определения Z^o и*

$$dI(z_0;h) = \int_a^b Q(z_0,t)h(t)\,dt,$$

где

$$Q(z,t) = \int_t^b F'_x\Big(\tau, A + \int_a^\tau z(\xi)\,d\xi, z(\tau)\Big)\,d\tau + F'_{x'}\Big(t, A + \int_a^t z(\tau)\,d\tau, z(t)\Big).$$

При этом справедливо разложение

$$I(z_0 + h) = I(z_0) + dI(z_0;h) + o(\|h\|_\infty),$$

где $o(\|h\|_\infty)/\|h\|_\infty \to 0$ при $\|h\|_\infty \to 0$.

Доказательство. Зафиксируем функцию $z_0 \in Z^o$ и положим

$$x_0(t) = A + \int_a^t z_0(\tau)\,d\tau.$$

Выберем произвольное $\varepsilon > 0$. По теореме 2 существует $\delta > 0$ такое, что для любого $v \in C^1[a,b]$, удовлетворяющего неравенству $0 < \|v\|_{1,\infty} < \delta$, будет

$$\frac{1}{\|v\|_{1,\infty}}\Big|J(x_0 + v) - J(x_0) - dJ(x_0;v)\Big| < \varepsilon.$$

Поэтому для любого $h \in C[a,b]$ такого, что $0 < \|h\|_\infty < \delta/(b-a+1)$, справедливо неравенство

$$\frac{1}{\|Th\|_{1,\infty}}\Big|J(x_0+Th) - J(x_0) - dJ(x_0;Th)\Big| < \varepsilon.$$

Отсюда, учитывая, что $\|Th\|_{1,\infty} \geqslant \|h\|_\infty$ и $J(x_0+Th) = I(z_0+h)$, получаем, что

$$\frac{1}{\|h\|_\infty}\Big|I(z_0+h) - I(z_0) - dJ(x_0;Th)\Big| < \varepsilon,$$

для любого $h \in C[a,b]$, удовлетворяющего неравенству $0 < \|h\|_\infty < \delta/(b-a+1)$. Таким образом, справедливо разложение

$$I(z_0+h) = I(z_0) + dJ(x_0;Th) + o(\|h\|_\infty),$$

где $o(\|h\|_\infty)/\|h\|_\infty \to 0$ при $\|h\|_\infty \to 0$.

По теореме 2 имеем

$$dJ(x_0;Th) = \int_a^b \Big[F'_x\big(t,x_0(t),x'_0(t)\big)\cdot\int_a^t h(\tau)d\tau + F'_{x'}\big(t,x_0(t),x'_0(t)\big)h(t)\Big]\,dt. \quad (2)$$

Проинтегрируем первое слагаемое по частям. Обозначим

$$u(t) = \int_t^b F'_x\big(\tau,x_0(\tau),x'_0(\tau)\big)d\tau.$$

Тогда

$$\int_a^b \Big[F'_x\big(t,x_0(t),x'_0(t)\big)\cdot\int_a^t h(\tau)d\tau\Big]\,dt = \int_a^b \Big[(-u'(t))\cdot\int_a^t h(\tau)d\tau\Big]\,dt =$$
$$= \Big((-u(t))\cdot\int_a^t h(\tau)d\tau\Big)\Big|_a^b + \int_a^b u(t)h(t)\,dt = \int_a^b u(t)h(t)\,dt.$$

Теперь формула (2) принимает вид

$$dJ(x_0;Th) = \int_a^b \left[\int_t^b F'_x\big(\tau,x_0(\tau),x'_0(\tau)\big)d\tau + F'_{x'}\big(t,x_0(t),x'_0(t)\big)\right]h(t)\,dt =$$
$$= \int_a^b Q(z_0,t)h(t)\,dt =: dI(z_0;h),$$

что и требовалось доказать. □

Рассуждая аналогичным образом и используя теорему 3, нетрудно проверить, что справедливо следующее утверждение.

ТЕОРЕМА 6. *При $F \in C^2(U)$ функционал $I(z)$ дважды дифференцируем по Фреше в каждой точке z_0 своей естественной области определения Z^o и*

$$d^2I(z_0;h) = d^2J(x_0;Th) = \int_a^b \left[F''_{xx}(Th)^2 + 2F''_{xx'}(Th)h + F''_{x'x'}h^2\right] dt,$$

где

$$F''_{xx} = F''_{xx}\big(t, x_0(t), z_0(t)\big), \quad F''_{xx'} = F''_{xx'}\big(t, x_0(t), z_0(t)\big),$$
$$F''_{x'x'} = F''_{x'x'}\big(t, x_0(t), z_0(t)\big)$$

и $x_0(t) = A + Tz_0$. При этом справедливо разложение

$$I(z_0 + h) = I(z_0) + dI(z_0;h) + \tfrac{1}{2}d^2I(z_0;h) + o(\|h\|^2_\infty),$$

где $o(\|h\|^2_\infty)/\|h\|^2_\infty \to 0$ при $\|h\|_\infty \to 0$.

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Первый и второй дифференциалы интегрального функционала* // Семинар «DHA & CAGD». Избранные доклады. 5 декабря 2013 г. (`http://dha.spb.ru/reps13.shtml#1205`) [Данная книга, с. 357]

2. Демьянов В. Ф. *Условия экстремума и вариационное исчисление.* М.: Высш. шк., 2004. 335 с.

3. Тамасян Г. Ш. *Гиподифференциальный спуск в вариационных задачах* // Семинар «CNSA & NDO». Избранные доклады. 25 сентября 2014 г. (`http://apmath.spbu.ru/cnsa/reps14.shtml#0925`) [Данная книга, с. 393]

4. Малозёмов В. Н., Тамасян Г. Ш. *Об одной кубической вариационной задаче* // Семинар «CNSA & NDO». Избранные доклады. 11 февраля 2016 г. (`http://apmath.spbu.ru/cnsa/reps16.shtml#0211`) [Данная книга, с. 346]

5. Поляк Б. Т. *Градиентные методы минимизации функционалов* // Ж. вычисл. матем. и матем. физ., 1963, том 3, № 4, С. 643–653.

О МИНИМАЛЬНОЙ ПОВЕРХНОСТИ ВРАЩЕНИЯ*

В. Н. Малозёмов

Перу Николая Максимовича Гюнтера принадлежит один из лучших учебников по вариационному исчислению [1]. Его книга была подписана к печати 27 мая 1941 г. Как отметил в предисловии В. И. Смирнов, последняя корректура была сдана Н. М. Гюнтером 29 апреля 1941 г. за пять дней до его кончины, последовавшей 4 мая 1941 г. С волнением берем мы в руки это сочинение замечательного математика.

Учебник Н. М. Гюнтера отличают полнота и тщательность анализа изучаемых вопросов. Он и сегодня читается с большим интересом. Вместе с тем за прошедшие 65 лет вариационное исчисление получило дальнейшее развитие как вширь, так и вглубь [2, 3, 4, 5]. Даже некоторые конкретные вопросы, рассмотренные в [1], теперь мы понимаем лучше. В этой связи остановимся на классической задаче о минимальной поверхности вращения и дополним анализ Н. М. Гюнтера случая двух стационарных кривых.

1°. Напомним постановку задачи о минимальной поверхности вращения:

$$J(y) := 2\pi \int\limits_{-a}^{a} y(x)\sqrt{1+[y'(x)]^2}dx \to \min, \qquad y(-a) = y(a) = A. \tag{1}$$

Здесь a, A – положительные константы. Решение будем искать на множестве функций $y(x)$, положительных и непрерывно дифференцируемых на отрезке $[-a, a]$.

Как известно [1, с. 36–40], двухпараметрическое семейство лагранжевых кривых (экстремалей) для задачи (1) определяется формулой

$$y(x) = \lambda \operatorname{ch} \frac{x+c}{\lambda}. \tag{2}$$

В силу симметричности краевых условий $c = 0$. Константу $\lambda > 0$ найдём из уравнения

$$\lambda \operatorname{ch} \frac{a}{\lambda} = A. \tag{3}$$

*Вестник СПбГУ. Сер. 10. 2006. Вып. 1. С. 52–56.

Обозначив $t = a/\lambda$, $\varphi(t) = \operatorname{ch}(t)/t$, перепишем последнее уравнение в виде

$$\varphi(t) = \frac{A}{a}. \tag{4}$$

Исследуем уравнение (4).

Функция $\varphi(t)$ на полуоси $(0, +\infty)$ имеет производную

$$\varphi'(t) = \frac{t \operatorname{sh} t - \operatorname{ch} t}{t^2} = \frac{\operatorname{sh} t}{t^2}(t - \operatorname{cth} t).$$

Отметим, что

$$\operatorname{cth} t = \frac{e^t + e^{-t}}{e^t - e^{-t}} = 1 + \frac{2}{e^{2t} - 1}.$$

Ясно (рис. 1), что у разности $t - \operatorname{cth} t$ существует единственный положительный корень t_0, причем $t_0 > 1$. Эта разность отрицательна при $0 < t < t_0$ и положительна при $t > t_0$. Значит, t_0 является единственной точкой минимума функции $\varphi(t)$ на $(0, +\infty)$ (рис. 2). При этом

$$\varphi(t_0) = \frac{\operatorname{ch} t_0}{\operatorname{cth} t_0} = \operatorname{sh} t_0 = \frac{1}{\sqrt{\operatorname{cth}^2 t_0 - 1}} = \frac{1}{\sqrt{t_0^2 - 1}} =: k_0.$$

Нетрудно сосчитать, что $k_0 = 1.50888$.

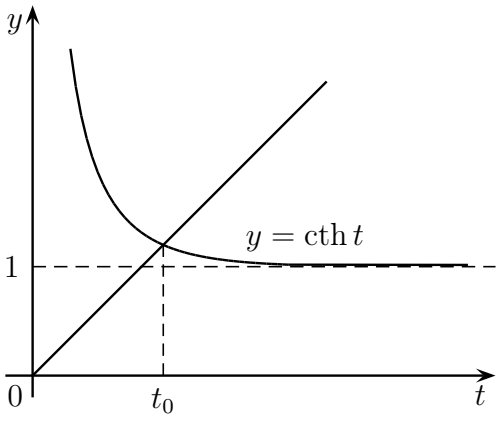


Рис. 1. Корень производной функции $\varphi(t)$

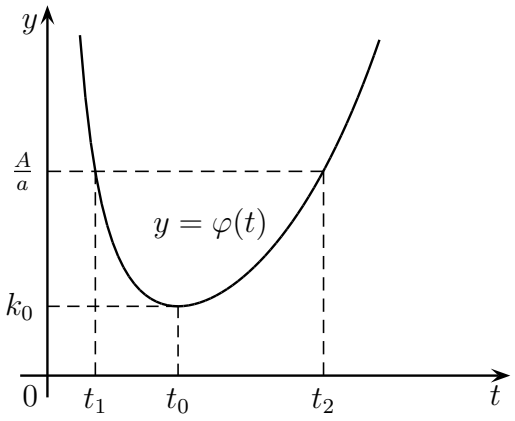


Рис. 2. Точка минимума функции $\varphi(t)$

Приходим к следующему выводу:

при $A/a > k_0$ уравнение (4) имеет два решения $0 < t_1 < t_2$,

при $A/a = k_0$ уравнение (4) имеет одно решение t_0,

при $A/a < k_0$ уравнение (4) не имеет решений.

Нас интересует первый случай, когда уравнение (4) имеет два решения. Соответственно уравнение (3) также имеет два решения $\lambda_1 = a/t_1$, $\lambda_2 = a/t_2$, причём $\lambda_1 > \lambda_2 > 0$. Это значит, что при $A/a > k_0$ существуют две стационарные кривые $y_k(x) = \lambda_k \operatorname{ch}(x/\lambda_k)$, $k = 1, 2$ (рис. 3). На обеих кривых выполняется усиленное условие Лежандра. Какой из них отдать предпочтение с точки зрения поставленной задачи — минимизации поверхности вращения?

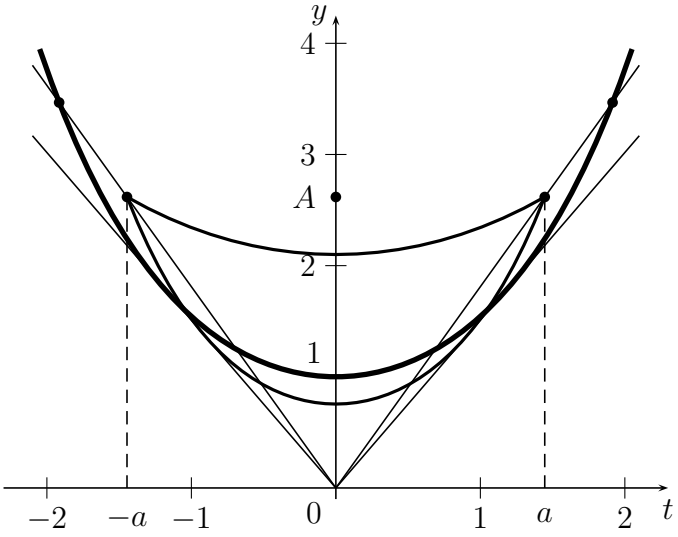


Рис. 3. Графики гиперболического косинуса (жирная линия) и двух стационарных кривых[1]

2°. Для ответа на этот вопрос нужно записать уравнение Якоби, соответствующее стационарной кривой $y_k(x)$, и найти его главное решение. Уравнение Якоби в данном случае имеет вид [1, с. 85]

$$u'' - \frac{2}{\lambda_k} \operatorname{th}\Big(\frac{x}{\lambda_k}\Big) u' + \frac{1}{\lambda_k^2} u = 0.$$

Главное решение $u_k(x)$ определяется начальными условиями $u(-a) = 0$, $u'(-a) = 1$. Как показано в [1, с. 85],

$$u_k(x) = \frac{\lambda_k \operatorname{ch} t_k - a \operatorname{sh} t_k}{\operatorname{ch}^2 t_k} \operatorname{sh} \frac{x}{\lambda_k} - \frac{\operatorname{sh} t_k}{\operatorname{ch}^2 t_k} \Big(x \operatorname{sh} \frac{x}{\lambda_k} - \lambda_k \operatorname{ch} \frac{x}{\lambda_k} \Big). \tag{5}$$

Справедливость формулы (5) легко проверить, однако вывести её совсем непросто.

3°. Существует более регулярный способ нахождения $u_k(x)$ [5, с. 124–127]. Для этого в двухпараметрическом семействе экстремалей (2) нужно выделить однопараметрическое семейство $y(x, \alpha)$, исходя из условий $y(-a) = A$, $y'(-a) = \alpha$. Здесь α – параметр. Распишем указанные условия подробно

$$\lambda \operatorname{ch} \frac{-a + c}{\lambda} = A, \quad \operatorname{sh} \frac{-a + c}{\lambda} = \alpha.$$

[1]Рисунок выполнен М.И. Григорьевым.

Отсюда следует, что $(-a+c)/\lambda = \operatorname{arsh}\alpha$ и

$$\lambda = \frac{A}{\sqrt{1+\alpha^2}}, \quad c = a + \lambda \operatorname{arsh}\alpha.$$

Таким образом,

$$y(x,\alpha) = \lambda(\alpha)\operatorname{ch}\frac{x+c\,(\alpha)}{\lambda(\alpha)}. \tag{6}$$

В этом семействе стационарная кривая $y_k(x)$ соответствует параметру $\alpha = \alpha_k$, такому, что $\lambda(\alpha_k) = \lambda_k$ и $c(\alpha_k) = 0$. Получаем

$$\operatorname{arsh}\alpha_k = -a/\lambda_k = -t_k, \quad \alpha_k = -\operatorname{sh} t_k.$$

Факт теории заключается в том, что

$$u_k(x) = y'_\alpha(x,\alpha_k). \tag{7}$$

Воспользуемся этой формулой. Поскольку

$$\lambda'(\alpha_k) = -\frac{A\alpha_k}{(1+\alpha_k^2)^{3/2}} = \frac{A\operatorname{sh} t_k}{\operatorname{ch}^3 t_k} = \frac{a}{t_k}\,\frac{\operatorname{sh} t_k}{\operatorname{ch}^2 t_k} = \frac{\lambda_k \operatorname{sh} t_k}{\operatorname{ch}^2 t_k},$$

$$c\,'(\alpha_k) = \lambda'(\alpha_k)\,\operatorname{arsh}\alpha_k + \frac{\lambda_k}{\sqrt{1+\alpha_k^2}} = -\frac{t_k\lambda_k \operatorname{sh} t_k}{\operatorname{ch}^2 t_k} + \frac{\lambda_k}{\operatorname{ch} t_k} =$$

$$= \frac{\lambda_k \operatorname{ch} t_k - a\operatorname{sh} t_k}{\operatorname{ch}^2 t_k},$$

то, согласно (6),

$$y'_\alpha(x,\alpha_k) = \lambda'(\alpha_k)\operatorname{ch}\frac{x}{\lambda_k} + \lambda_k \operatorname{sh}\Big(\frac{x}{\lambda_k}\Big)\Big[\frac{c'(\alpha_k)\lambda_k - x\lambda'(\alpha_k)}{\lambda_k^2}\Big] =$$

$$= c'(\alpha_k)\operatorname{sh}\frac{x}{\lambda_k} - \frac{\lambda'(\alpha_k)}{\lambda_k}\Big[x\operatorname{sh}\frac{x}{\lambda_k} - \lambda_k \operatorname{ch}\frac{x}{\lambda_k}\Big] =$$

$$= \frac{\lambda_k \operatorname{ch} t_k - a\operatorname{sh} t_k}{\operatorname{ch}^2 t_k}\operatorname{sh}\frac{x}{\lambda_k} - \frac{\operatorname{sh} t_k}{\operatorname{ch}^2 t_k}\Big[x\operatorname{sh}\frac{x}{\lambda_k} - \lambda_k \operatorname{ch}\frac{x}{\lambda_k}\Big].$$

Теперь (5) следует из (7).

Отметим, в частности, что

$$u_k(a) = \frac{2\lambda_k \operatorname{sh} t_k}{\operatorname{ch} t_k} - \frac{2a\operatorname{sh}^2 t_k}{\operatorname{ch}^2 t_k} = -\frac{2a\,t_k \operatorname{sh} t_k}{\operatorname{ch}^2 t_k}\;\frac{t_k \operatorname{sh} t_k - \operatorname{ch} t_k}{t_k^2} =$$

$$= -\frac{2a^2}{A}\;\operatorname{th}(t_k)\,\varphi'(t_k). \tag{8}$$

4°. Если стационарная кривая $y_k(x)$ является точкой (слабого) локального минимума функционала $J(y)$, то в ней выполняется необходимое условие оптимальности второго порядка: главное решение уравнения Якоби $u_k(x)$ положительно на $(-a,a)$. Поскольку $\varphi'(t_2) > 0$ (см. рис. **??**), то, согласно (8), $u_2(a) < 0$. Это значит, что $y_2(x)$ не является точкой локального минимума.

В силу той же формулы (8) $u_1(a) > 0$. Покажем, что $u_1(x) > 0$ на $(-a,a]$, то есть на $y_1(x)$ выполняется усиленное условие Якоби. В этом случае теория гарантирует, что $y_1(x)$ — точка строгого локального минимума.

Имеем

$$u_1'(x) = \frac{1}{\lambda_1^2}\operatorname{ch}\Big(\frac{x}{\lambda_1}\Big)\Big[c'(\alpha_1)\lambda_1 - x\lambda'(\alpha_1)\Big].$$

В квадратных скобках стоит линейная по x функция, поэтому $u_1'(x)$ обращается в нуль на $(-a,a)$ не более чем в одной точке. Так как $u_1(-a) = 0$, $u_1'(-a) = 1$ и $u_1(a) > 0$, то необходимо $u_1(x) > 0$ на $(-a,a]$. Действительно, если допустить, что существует точка $\xi \in (-a,a)$, в которой $u_1(\xi) = 0$, то в точке максимума η функции $u_1(x)$ на $[-a,\xi]$ будет $u_1(\eta) > 0$ и $u_1'(\eta) = 0$. Более того, в точке минимума $u_1(x)$ на $[\eta,a]$ производная $u_1'(x)$ ещё раз обратится в нуль, что, как отмечалось, невозможно.

5°. Предпочтительность стационарной кривой $y_1(x)$ по сравнению с $y_2(x)$ можно установить более непосредственным путем, если показать, что $J(y_1) < J(y_2)$. Именно этим мы теперь и займемся.

ТЕОРЕМА. *Справедливо неравенство* $J(y_1) < J(y_2)$.

Доказательство. Воспользуемся формулами

$$2\operatorname{ch}^2 x = \operatorname{ch} 2x + 1, \quad \operatorname{sh} 2x = 2\operatorname{sh} x \operatorname{ch} x.$$

Получим

$$J(y_k) = 2\pi\lambda_k \int\limits_{-a}^{a} \operatorname{ch}^2 \frac{x}{\lambda_k}\,dx = 2\pi\lambda_k \int\limits_{0}^{a} \Big(\operatorname{ch}\frac{2x}{\lambda_k} + 1\Big)dx =$$

$$= 2\pi\lambda_k\Big(\frac{\lambda_k}{2}\operatorname{sh}\frac{2a}{\lambda_k} + a\Big) = 2\pi\Big(\lambda_k^2 \operatorname{sh} t_k \operatorname{ch} t_k + a\lambda_k\Big). \tag{9}$$

Поскольку $\operatorname{ch}(t_k)/t_k = A/a$, то

$$(2\pi a^2)^{-1} J(y_k) = \frac{\operatorname{sh} t_k \operatorname{ch} t_k}{t_k^2} + \frac{1}{t_k} = \Big(\frac{A}{a}\Big)^2 \operatorname{th} t_k + \frac{1}{t_k}\,. \tag{10}$$

Введём функцию

$$\psi(t) = \Big(\frac{A}{a}\Big)^2 \operatorname{th} t + \frac{1}{t}\,, \quad t \in (0, +\infty),$$

и вычислим её производную:

$$\psi'(t) = \Big(\frac{A}{a}\Big)^2 \frac{1}{\operatorname{ch}^2 t} - \frac{1}{t^2} = \frac{1}{\operatorname{ch}^2 t}\Big[\Big(\frac{A}{a}\Big)^2 - \varphi^2(t)\Big].$$

Согласно определению функции $\varphi(t)$ и выбору точек t_1, t_2 имеем $\psi'(t_1) = 0$, $\psi'(t_2) = 0$, $\psi'(t) > 0$ при $t \in (t_1, t_2)$. В частности, $\psi(t_1) < \psi(t_2)$. Переписав формулу (10) в виде

$$(2\pi a^2)^{-1} J(y_k) = \psi(t_k),$$

придём к заключению теоремы. □

З а м е ч а н и е. Из (9) следует, что

$$J(y_k) = 2\pi\Big(A\sqrt{A^2 - \lambda_k^2} + a\lambda_k\Big), \quad k = 1, 2.$$

Действительно, нужно учесть, что $\operatorname{ch} t_k = A/\lambda_k$ и

$$\lambda_k^2 \operatorname{sh} t_k \operatorname{ch} t_k = A\lambda_k \operatorname{sh} t_k = A\lambda_k \sqrt{\operatorname{ch}^2 t_k - 1} = A\sqrt{A^2 - \lambda_k^2}\,.$$

6°. Более общие результаты, связанные с минимальными поверхностями, представлены в [6].

ЛИТЕРАТУРА

1. Гюнтер Н. М. *Курс вариационного исчисления.* Л.; М.: Гостехиздат, 1941. 308 с.

2. Янг Л. *Лекции по вариационному исчислению и теории оптимального управления* // Пер. с англ. М. Г. Элуашвили; под ред. В. М. Алексеева. М.: Мир, 1974. 488 с.

3. Алексеев В. М., Тихомиров В. М., Фомин С. В. *Оптимальное управление.* М.: Наука, 1979. 430 с.

4. Буслаев В. С. *Вариационное исчисление.* Л.: Изд-во Ленингр. ун-та, 1980. 287 с.

5. Коша А. *Вариационное исчисление* // Пер. с венг. Д. Валовича; под ред. Ш. А. Алимова. М.: Высшая школа, 1983. 280 с.

6. Тужилин А. А., Фоменко А. Т. *Элементы геометрии и топологии минимальных поверхностей.* М.: Наука, 1991. 176 с.

ГИПОДИФФЕРЕЦИАЛЬНЫЙ СПУСК В ВАРИАЦИОННЫХ ЗАДАЧАХ*

Г. Ш. Тамасян

Аннотация. В докладе сравниваются численные реализации двух прямых методов решения задач вариационного исчисления. А именно, хорошо известный метод Ритца и метод гиподифференциального спуска. Приводится пример, демонстрирующий преимущества второго метода.

1°. Постановка задачи. Пусть x_0, $x_1 \in \mathbb{R}$ и $T > 0$ фиксированы. Через $C^1[0,T]$ обозначим класс непрерывно дифференцируемых на $[0,T]$ функций.

Рассмотрим *простейшую* (или основную) задачу вариационного исчисления: требуется минимизировать функционал

$$I(x) = \int_0^T F(x(t), \dot{x}(t), t)\, dt \tag{1}$$

на множестве

$$\Omega = \{x \in C^1[0,T] \mid x(0) = x_0, \quad x(T) = x_1\}. \tag{2}$$

Здесь функция $F(x,z,t)$ непрерывна вместе с F'_x и F'_z по всем своим аргументам на $\mathbb{R} \times \mathbb{R} \times [0,T]$.

Положим

$$\varphi_1(z) = \int_0^T z(t)\, dt + x_0 - x_1, \quad \varphi(z) = |\varphi_1(z)|. \tag{3}$$

Введём множество

$$Z = \{z \in C[0,T] \mid \varphi(z) = 0\},$$

где $C[0,T]$ — множество непрерывных на $[0,T]$ функций.

Пусть v_1, $v_2 \in C[0,T]$. На множестве $C[0,T]$ введём скалярное произведение

$$\langle v_1, v_2 \rangle = \int_0^T v_1(t) v_2(t)\, dt.$$

*Семинар «CNSA & NDO». Избранные доклады. 25 сентября 2014 г.

Рассмотрим функционал

$$f(z) = \int_0^T F(x_0 + \int_0^t z(\tau)\,d\tau, z(t), t)\,dt.$$

Задача

$$I(x) \longrightarrow \inf_{x\in\Omega} \tag{4}$$

эквивалентна задаче

$$f(z) \longrightarrow \inf_{z\in Z} \tag{5}$$

в том смысле, что если $x^* \in \Omega$ — решение задачи (4), то функция $z^*(t) = \dot{x}^*(t)$ является решением задачи (5); и обратно, если $z^* \in Z$ доставляет минимум функции f на множестве Z, то функция $x^*(t) = x_0 + \int_0^t z^*(\tau)\,d\tau$ является решением задачи (4).

Мы будем решать задачу (5) с помощью теории точных штрафных функций [1, 5, 2, 4, 3, 6, 7]. При $\lambda \geqslant 0$ рассмотрим штрафную функцию

$$\Phi_\lambda(z) = f(z) + \lambda\varphi(z). \tag{6}$$

В работах [3, 5] показано, что для достаточно большого λ функция Φ_λ является точной штрафной функцией, т. е., что для достаточно большого λ любая точка глобального минимума штрафной функции Φ_λ является решением исходной задачи.

Таким образом, задача минимизации функционала f на множестве Z сведена к задаче минимизации функционала $\Phi_\lambda(z)$ на множестве $C[0,T]$.

2°. Метод гиподифференциального спуска (МГС). Функция Φ_λ является гиподифференцируемой в точке $z \in C[0,T]$ (см. [5, 8, 10]). Действительно, для любого $v \in C[0,T]$ справедливо разложение

$$\Phi_\lambda(z+\varepsilon v) = \Phi_\lambda(z) + \max_{[a,w]\in d\Phi_\lambda(z)} [a + \varepsilon\langle w, v\rangle] + o(\varepsilon, v),$$

где $o(\varepsilon, v)/\varepsilon \to 0$ при $\varepsilon \downarrow 0$. Множество $d\Phi_\lambda(z) \subset \mathbb{R}\times C[0,T]$ — гиподифференциал функции Φ_λ в точке z. Он имеет вид

$$d\Phi_\lambda(x) = \operatorname{co}\left\{ \begin{pmatrix} 0 \\ Q(t,z) + \lambda\ \operatorname{sign}\varphi_1(z) \end{pmatrix}, \begin{pmatrix} -2\lambda\varphi(z) \\ Q(t,z) - \lambda\ \operatorname{sign}\varphi_1(z) \end{pmatrix} \right\}. \tag{7}$$

Здесь $Q(t,z)$ — градиент функционала f в точке z

$$Q(t,z) = \int_t^T \frac{\partial F(x(\tau), z(\tau), \tau)}{\partial x} d\tau + \frac{\partial F(x(t), z(t), t)}{\partial z} \tag{8}$$

и

$$\varphi_1(z) = \int_0^T z(t)\,dt + x_0 - x_1.$$

В пространстве $\mathbb{R} \times C[0,T]$ введём норму

$$\big\|[a,w]\big\| = \sqrt{a^2 + \langle w, w\rangle}.$$

ЛЕММА 1 ([5, 3])**.** *Для того чтобы в точке $z^* \in C[0,T]$ функция Φ_λ достигала своего наименьшего значения, необходимо, а в случае, когда функция f выпукла и достаточно, чтобы*

$$\begin{pmatrix} 0 \\ \mathbb{O} \end{pmatrix} \in d\Phi_\lambda(z^*). \tag{9}$$

Точка z^*, которая удовлетворяет условию (9), называется *стационарной.*

ЛЕММА 2. *Пусть $z \in C[0,T]$ не является стационарной точкой. Направлением спуска функционала Φ_λ в точке z является функция*

$$g(t,z) = -\frac{q_2^*(t,z)}{||q_2^*(t,z)||}, \tag{10}$$

где

$$q_2^*(t,z) = \begin{cases} Q(t,z) + \lambda \operatorname{sign}\varphi_1(z), & \text{если } \gamma^* < 0; \\ Q(t,z) - \dfrac{\int_0^T Q(t,z)dt - \lambda\varphi^2(z)\ \operatorname{sign}\varphi_1(z)}{T + \varphi^2(z)}, & \text{если } \gamma^* \in [0,1]; \\ Q(t,z) - \lambda \operatorname{sign}\varphi_1(z), & \text{если } \gamma^* > 1; \end{cases} \tag{11}$$

и

$$\gamma^* = \frac{T + \dfrac{\operatorname{sign}\varphi_1(z)}{\lambda}\displaystyle\int_0^T Q(t,z)dt}{2[T + \varphi^2(z)]}. \tag{12}$$

В отличие от направления наискорейшего спуска [5, 10], направление $g(t,z)$ является непрерывным по z.

Доказательство. Найдем минимальный по норме гипоградиент

$$[q_1^*, q_2^*] \in d\Phi_\lambda(z).$$

Отметим, что гиподифференциал $d\Phi_\lambda(z)$ (см. (7)) является отрезком, поэтому каждый элемент $\widetilde{q} \in d\Phi_\lambda(z)$ можно описать следующим образом:

$$\widetilde{q}(\gamma) = \begin{pmatrix} -2\gamma\lambda\varphi(z) \\ Q(t,z) + (1-2\gamma)\lambda\ \operatorname{sign}\varphi_1(z) \end{pmatrix}, \qquad \gamma \in [0,1].$$

Поиск минимального по норме гипоградиента сводится к решению задачи

$$\min_{\gamma\in[0,1]} \left\|\widetilde{q}(\gamma)\right\|^2 = \min_{\gamma\in[0,1]} h(\gamma) = h(\gamma^*), \tag{13}$$

где

$$h(\gamma) = (2\gamma\lambda\varphi(z))^2 + \int_0^T \left(Q(t,z) + (1-2\gamma)\lambda \operatorname{sign}\varphi_1(z)\right)^2 dt.$$

Ясно, что решение γ^* данной задачи существует и единственно. Для минимального по норме гипоградиента получим формулу $[q_1^*, q_2^*] = \widetilde{q}(\gamma^*)$, где

$$\gamma^* = \frac{T + \dfrac{\operatorname{sign}\varphi_1(z)}{\lambda}\displaystyle\int_0^T Q(t,z)dt}{2[T+\varphi^2(z)]}, \qquad q_1^*(z) = \begin{cases} 0, & \gamma^* < 0; \\ -2\gamma^*\lambda\varphi(z), & \gamma^* \in [0,1]; \\ -2\lambda\varphi(z), & \gamma^* > 1; \end{cases}$$

$$q_2^*(t,z) = \begin{cases} Q(t,z) + \lambda \operatorname{sign}\varphi_1(z), & \gamma^* < 0; \\ Q(t,z) + (1-2\gamma^*)\lambda \operatorname{sign}\varphi_1(z), & \gamma^* \in [0,1]; \\ Q(t,z) - \lambda \operatorname{sign}\varphi_1(z), & \gamma^* > 1. \end{cases}$$

Лемма доказана. □

Опишем схему метода гиподифференциального спуска для простейшей задачи вариационного исчисления (1), (2).

Зафиксируем $\varepsilon > 0$.

1) Выберем $z_0 \in P[0,T]$.

2) Переход от k-го приближения к $(k+1)$-му осуществляется в следующем порядке ($k \geqslant 0$):

   (a) Вычислим по формулам (11), (12) функцию $q_2^*(t, z_k)$ — вторую компоненту минимального по норме гипоградиента функционала F_λ в точке z_k.

   (b) Проверим выполнение условия $\|q_2^*(t, z_k)\| < \varepsilon$. Если оно выполнено, то процесс прекращается.

   (c) Построим направления спуска $g_k := g(t, z_k)$ по формуле (10) и найдём $\beta_k \geqslant 0$ такое, что

$$\min_{\beta\geqslant 0} \Phi_\lambda(z_k - \beta g_k) = \Phi_\lambda(z_k - \beta_k g_k).$$

   (d) Положим $z_{k+1} = z_k - \beta_k g_k$.

3°. Метод Ритца. Напомним (см. монографию С.Г. Михлина [9]), что основные этапы численной реализации вариационных методов, в частности, метода Ритца, сводятся к следующему:

1) выбор системы координатных функций;

2) составление системы Ритца и её решение;

3) учёт влияния погрешностей, допущенных при составлении и решении системы Ритца, на точность приближенного решения.

Зафиксируем некоторое натуральное число m. Идея метода Ритца состоит в том, что m-е приближенное решение x_m задачи (1), (2) ищется в виде линейной комбинации координатных функций $\{\psi_k(t)\}$:

$$x_m(t) = \widetilde{x}(t) + \sum_{k=1}^{m} c_k^{(m)} \psi_k(t). \tag{14}$$

Здесь $\widetilde{x}(t)$ — произвольная функция, удовлетворяющая краевым условиям (2), $c_k^{(m)}$ — неизвестные пока вещественные постоянные («коэффициенты Ритца»), а координатные функции удовлетворяют условиям $\psi_k(0) = \psi_k(T) = 0$, $k = 1 : m$.

На линейных комбинациях $x_m(t)$ функционал (1) обращается в функцию аргументов $c_1^{(m)}, \ldots, c_m^{(m)}$:

$$\Phi(c_1^{(m)}, \ldots, c_m^{(m)}) := I(x_m). \tag{15}$$

Далее, остаётся найти коэффициенты $c_k^{(m)}$, $k = 1 : m$, так, чтобы функция (15) принимала минимальное значение.

Замечание. Отметим три основных недостатка метода Ритца при решении нелинейных задач. Во-первых, трудность построения координатных функций, удовлетворяющих заданным граничным условиям, при сколь-нибудь сложной форме области. Во-вторых, трудоёмкость составления системы Ритца, связанная с тем, что коэффициенты этой системы обычно выражаются через некоторые интегралы. В-третьих, численное решение системы Ритца, вообще говоря, нелинейной относительно $c_1^{(m)}, \ldots, c_m^{(m)}$.

4°. Численный эксперимент. Следующий пример иллюстрирует одну из «слабых» сторон метода Ритца.

ПРИМЕР 1. (см. [9, стр. 339].) Найдём функцию, удовлетворяющую краевым условиям

$$x(0) = 1, \quad x(1) = 0$$

и доставляющую минимум интегралу

$$I(x) = \int_0^1 \left[\frac{(x')^4}{48} + (x')^2 + x^2 - 6x \right] dt.$$

Известно точное решение $x^*(t) = 1 - t^2$, $I(x^*) = -\frac{31}{15} \approx -2{,}0666666$.

В [9] предлагается выбрать координатные функции вида

$$\psi_k(t) := \sin[(2k-1)\pi t], \quad k = 1, 2, \ldots.$$

Очевидно, что они удовлетворяют условию $\psi_k(0) = \psi_k(1) = 0$. В качестве функции $\widetilde{x}(t)$ (см. (14)), положим

$$\widetilde{x}(t) = 1 - t.$$

Покажем, как строится приближение пятого порядка:

$$x_5(t) = 1 - t + \sum_{k=1}^{5} c_k \sin[(2k-1)\pi t].$$

Подставив x_5 в функционал $I(x)$ (см. (15)), получим

$$P(c_1, c_2, c_3, c_4, c_5) := I(x_5) =$$

$$\begin{aligned}
= \quad & \frac{1}{8064\,\pi} \Big[63\,\pi^5 {c_1}^4 + 252\,\pi^5 {c_1}^3 c_2 + 2268\,\pi^5 {c_1}^2 {c_2}^2 + 3780\,\pi^5 {c_1}^2 c_2 c_3 + \\
+ \quad & 6300\,\pi^5 {c_1}^2 {c_3}^2 + 8820\,\pi^5 {c_1}^2 c_3 c_4 + 12348\,\pi^5 {c_1}^2 {c_4}^2 + 15876\,\pi^5 {c_1}^2 c_4 c_5 + \\
+ \quad & 20412\,\pi^5 {c_1}^2 {c_5}^2 + 11340\,\pi^5 c_1 {c_2}^2 c_3 + 15876\,\pi^5 c_1 {c_2}^2 c_4 + 52920\,\pi^5 c_1 c_2 c_3 c_4 + \\
+ \quad & 68040\,\pi^5 c_1 c_2 c_3 c_5 + 95256\,\pi^5 c_1 c_2 c_4 c_5 + 56700\,\pi^5 c_1 {c_3}^2 c_5 + 5103\,\pi^5 {c_2}^4 + \\
+ \quad & 20412\,\pi^5 {c_2}^3 c_5 + 56700\,\pi^5 {c_2}^2 {c_3}^2 + 111132\,\pi^5 {c_2}^2 {c_4}^2 + 183708\,\pi^5 {c_2}^2 {c_5}^2 + \\
+ \quad & 132300\,\pi^5 c_2 {c_3}^2 c_4 + 476280\,\pi^5 c_2 c_3 c_4 c_5 + 39375\,\pi^5 {c_3}^4 + 308700\,\pi^5 {c_3}^2 {c_4}^2 + \\
+ \quad & 510300\,\pi^5 {c_3}^2 {c_5}^2 + 555660\,\pi^5 c_3 {c_4}^2 c_5 + 151263\,\pi^5 {c_4}^4 + 1000188\,\pi^5 {c_4}^2 {c_5}^2 + \\
+ \quad & 413343\,{c_5}^4 \pi^5 + 4536\,\pi^3 {c_1}^2 + 40824\,\pi^3 {c_2}^2 + 113400\,\pi^3 {c_3}^2 + 222264\,\pi^3 {c_4}^2 + \\
+ \quad & 367416\,{c_5}^2 \pi^3 + 4032\,\pi\,{c_1}^2 + 4032\,\pi\,{c_2}^2 + 4032\,\pi\,{c_3}^2 + 4032\,\pi\,{c_4}^2 + \\
+ \quad & 4032\,{c_5}^2 \pi - 13272\,\pi - 80640\,c_1 - 26880\,c_2 - 16128\,c_3 - 11520\,c_4 - 8960\,c_5 \Big].
\end{aligned}$$

Составим систему Ритца. Приравняем нулю частные производные функции P по переменным c_j, $j = 1 : 5$:

$$\begin{aligned}\frac{\partial P}{\partial c_1} &= \frac{1}{8064\,\pi}\Big[252\,\pi^5 {c_1}^3 + 756\,\pi^5 {c_1}^2 c_2 + 4536\,\pi^5 c_1 {c_2}^2 + \\ &+ 7560\,\pi^5 c_1 c_2 c_3 + 12600\,\pi^5 c_1 {c_3}^2 + 17640\,\pi^5 c_1 c_3 c_4 + 24696\,\pi^5 c_1 {c_4}^2 + \\ &+ 31752\,\pi^5 c_1 c_4 c_5 + 40824\,\pi^5 c_1 {c_5}^2 + 11340\,\pi^5 {c_2}^2 c_3 + 15876\,\pi^5 {c_2}^2 c_4 + \\ &+ 52920\,\pi^5 c_2 c_3 c_4 + 68040\,\pi^5 c_2 c_3 c_5 + 95256\,\pi^5 c_2 c_4 c_5 + 56700\,\pi^5 {c_3}^2 c_5 + \\ &+ 9072\,\pi^3 c_1 + 8064\,\pi\,c_1 - 80640\Big] = 0,\end{aligned}$$

$$\begin{aligned}\frac{\partial P}{\partial c_2} &= \frac{1}{8064\,\pi}\Big[252\,\pi^5 {c_1}^3 + 4536\,\pi^5 {c_1}^2 c_2 + 3780\,\pi^5 {c_1}^2 c_3 + 22680\,\pi^5 c_1 c_2 c_3 + \\ &+ 31752\,\pi^5 c_1 c_2 c_4 + 52920\,\pi^5 c_1 c_3 c_4 + 68040\,\pi^5 c_1 c_3 c_5 + 95256\,\pi^5 c_1 c_4 c_5 + \\ &+ 20412\,\pi^5 {c_2}^3 + 61236\,\pi^5 {c_2}^2 c_5 + 113400\,\pi^5 c_2 {c_3}^2 + 222264\,\pi^5 c_2 {c_4}^2 + \\ &+ 367416\,\pi^5 c_2 {c_5}^2 + 132300\,\pi^5 {c_3}^2 c_4 + 476280\,\pi^5 c_3 c_4 c_5 + 81648\,\pi^3 c_2 + \\ &+ 8064\,\pi\,c_2 - 26880\Big] = 0,\end{aligned}$$

$$\begin{aligned}\frac{\partial P}{\partial c_3} &= \frac{1}{8064\,\pi}\Big[3780\,\pi^5 {c_1}^2 c_2 + 12600\,\pi^5 {c_1}^2 c_3 + 8820\,\pi^5 {c_1}^2 c_4 + 11340\,\pi^5 c_1 {c_2}^2 + \\ &+ 52920\,\pi^5 c_1 c_2 c_4 + 68040\,\pi^5 c_1 c_2 c_5 + 113400\,\pi^5 c_1 c_3 c_5 + 113400\,\pi^5 {c_2}^2 c_3 + \\ &+ 264600\,\pi^5 c_2 c_3 c_4 + 476280\,\pi^5 c_2 c_4 c_5 + 157500\,\pi^5 {c_3}^3 + 617400\,\pi^5 c_3 {c_4}^2 + \\ &+ 1020600\,\pi^5 c_3 {c_5}^2 + 555660\,\pi^5 {c_4}^2 c_5 + 226800\,\pi^3 c_3 + 8064\,c_3\pi - 16128\Big] = 0,\end{aligned}$$

$$\begin{aligned}\frac{\partial P}{\partial c_4} &= \frac{1}{8064\,\pi}\Big[8820\,\pi^5 {c_1}^2 c_3 + 24696\,\pi^5 {c_1}^2 c_4 + 15876\,\pi^5 {c_1}^2 c_5 + 15876\,\pi^5 c_1 {c_2}^2 + \\ &+ 52920\,\pi^5 c_1 c_2 c_3 + 95256\,\pi^5 c_1 c_2 c_5 + 222264\,\pi^5 {c_2}^2 c_4 + 132300\,\pi^5 c_2 {c_3}^2 + \\ &+ 476280\,\pi^5 c_2 c_3 c_5 + 617400\,\pi^5 {c_3}^2 c_4 + 1111320\,\pi^5 c_3 c_4 c_5 + 605052\,\pi^5 {c_4}^3 + \\ &+ 2000376\,\pi^5 c_4 {c_5}^2 + 444528\,\pi^3 c_4 + 8064\,\pi\,c_4 - 11520\Big] = 0,\end{aligned}$$

$$\begin{aligned}\frac{\partial P}{\partial c_5} &= \frac{1}{8064\,\pi}\Big[15876\,\pi^5 {c_1}^2 c_4 + 40824\,\pi^5 {c_1}^2 c_5 + 68040\,\pi^5 c_1 c_2 c_3 + \\ &+ 95256\,\pi^5 c_1 c_2 c_4 + 56700\,\pi^5 c_1 {c_3}^2 + 20412\,\pi^5 {c_2}^3 + 367416\,\pi^5 {c_2}^2 c_5 + \\ &+ 476280\,\pi^5 c_2 c_3 c_4 + 1020600\,\pi^5 {c_3}^2 c_5 + 555660\,\pi^5 c_3 {c_4}^2 + 2000376\,\pi^5 {c_4}^2 c_5 + \\ &+ 1653372\,\pi^5 {c_5}^3 + 734832\,\pi^3 c_5 + 8064\,\pi\,c_5 - 8960\Big] = 0.\end{aligned}$$

Далее, эта нелинейная система в [9] решалась методом Ньютона – Канторовича с начальным приближением

$$c_1 = c_2 = c_3 = c_4 = c_5 = 0.$$

Приведём результаты третьего приближения:

$$x_5(t) = 1 - t + \sum_{k=1}^{5} c_k \sin[(2k-1)\pi t],$$

где $c_1 = 0{,}25801628$; $c_2 = 0{,}00956007$; $c_3 = 0{,}00206907$; $c_4 = 0{,}00075838$; $c_5 = 0{,}00036394$. Отметим, что

$$\|x_5 - x^*\| = 1{,}78 \cdot 10^{-4}, \quad \|x_5' - z^*\| = 7{,}33 \cdot 10^{-2}, \quad I(x_5) = -2{,}066601195.$$

Решим этот же пример методом гиподифференциального спуска. Пусть штрафной параметр λ равен 100 и точность вычислений $\varepsilon = 10^{-4}$.

В качестве начального приближения выберем $z_0(t) \equiv -1$. Приведём результаты счёта. На первом шаге по формуле (11) имеем

$$q_2^*(t, z_0) = t^2 + 4t - \frac{7}{3}, \qquad \|q_2^*(t, z_0)\| = 1{,}4453 > \varepsilon.$$

Условие останова не выполнено, поэтому переходим к поиску β_0 — величины шага спуска в направлении g_0 (см. (10)). Так как z_0 — допустимая точка, т. е. удовлетворяет краевым условиям, то, согласно работе [10], для всех β функции (одной переменной) $\Phi_\lambda(z_0 - \beta g_0)$ и $f(z_0 - \beta g_0)$ равны, при этом

$$f(z_0 - \beta g_0) = \frac{3743}{22680}\beta^4 + \frac{197}{5670}\beta^3 + \frac{9671}{3780}\beta^2 - \frac{94}{45}\beta - \frac{79}{48}.$$

В связи с тем, что градиент данной функции равен нулю в единственной (вещественной) точке $\beta_0^* = 0{,}396958$, она и будет точкой минимума. Получили следующее приближение

$$z_1(t) = -0{,}396952\, t^2 - 1{,}587808\, t - 0{,}0737786.$$

Приближения, полученные с помощью МГС.

k	I	$\|g_k\|$	$\|z_k - z^*\|$	$\|x_k - x^*\|$
0	$-1{,}64583333$	1,4453	0,57735	0,18257
1	$-2{,}06561147$	0,0713	0,02991	$4{,}7 \cdot 10^{-3}$
2	$-2{,}06664366$	0,0107	$4{,}3 \cdot 10^{-3}$	$8{,}8 \cdot 10^{-4}$
3	$-2{,}06666606$	$1{,}6 \cdot 10^{-3}$	$7{,}2 \cdot 10^{-4}$	$9{,}6 \cdot 10^{-5}$

Из таблицы видно, что уже на третьем шаге результаты лучше, чем приближение, полученное методом Ритца.

ЛИТЕРАТУРА

1. Ерёмин И.И. *Метод «штрафов» в выпуклом программировании* // Доклады АН СССР, 1967. Т. 143, № 4. С. 748–751.

2. Di Pillo G., Facchinei F. *Exact penalty functions for nondifferentiable programming problems*, in Nonsmooth Optimization and Related Topics, F. Clarke, V. F. Demyanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 89–107.

3. Демьянов В. Ф. *Точные штрафные функции в задачах негладкой оптимизации* // Вестн. С.-Петерб. ун-та. Сер. 1, 1994. Вып. 4 (№ 22). С. 21–27.

4. Demyanov V. F., Di Pillo G., Facchinei F. *Exact penalization via Dini and Hadamard conditional derivatives* // Optim. Methods Softw, 1998. vol. 9, no. 1–3. pp. 19–36.

5. Демьянов В. Ф. *Условия экстремума и вариационное исчисление.* М.: Высшая школа, 2005. 335 с.

6. Долгополик М. В. *Точные штрафные функции в негладкой оптимизации* // Семинар «CNSA & NDO». Избранные доклады. 8 мая 2014 г. (`http://apmath.spbu.ru/cnsa/reps14.shtml#0508`)

7. Долгополик М. В. *Обзор по точным штрафным функциям* // Семинар «CNSA & NDO». Избранные доклады. 2 октября 2014 г. (`http://apmath.spbu.ru/cnsa/reps14.shtml#1002`)

8. Демьянов В. Ф., Рубинов А. М. *Основы негладкого анализа и квазидифференциальное исчисление.* М.: Наука. Гл. ред. физ.-мат. лит., 1990. 432 с.

9. Михлин С. Г. *Численная реализация вариационных методов.* М.: Наука, 1966. 432 с.

10. Долгополик М.В., Тамасян Г.Ш. *Об эквивалентности методов наискорейшего и гиподифференциального спусков в некоторых задачах условной оптимизации* // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2014. Т. 14, вып. 4, ч. 2, с. 532–542.

СХОДИМОСТЬ МЕТОДА ГИПОДИФФЕРЕНЦИАЛЬНОГО СПУСКА В КЛАССИЧЕСКИХ ЗАДАЧАХ ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ*

М. В. Долгополик

Аннотация. В докладе обсуждаются вопросы сходимости метода гиподифференциального спуска в простейшей задаче вариационного исчисления [1, 2, 3]. Показывается, что метод гиподифференциального спуска для данной задачи совпадает с методом проекции градиента, и приводятся общие теоремы о сходимости данного метода, содержащие оценки скорости сходимости.

1°. Постановка задачи. Рассмотрим классическую задачу вариационного исчисления

$$J(x) = \int_a^b F(x(t), x'(t), t)\, dt \to \inf, \tag{1}$$

$$x(a) = A, \quad x(b) = B, \quad x \in W_2^1[a, b]. \tag{2}$$

Здесь $F(x, z, t)$ — функция трёх переменных, заданная и непрерывно дифференцируема на $\mathbb{R}^2 \times [a, b]$, а $W_2^1[a, b]$ — пространство Соболева на отрезке $[a, b]$ (см., например, [4, 5]). Как обычно, мы будем считать, что пространство $W_2^1[a, b]$ состоит из всех абсолютно непрерывных функций $x\colon [a, b] \to \mathbb{R}$ таких, что $x' \in L_2[a, b]$.

Следует отметить, что в качестве основного пространства в задаче (1), (2) было выбрано пространство Соболева, поскольку в этом пространстве существенно упрощается анализ сходимости метода гиподифференциального спуска. Кроме того, использование пространства Соболева позволяет при естественных предположениях гарантировать существование оптимального плана задачи (1), (2). Однако, у данного подхода имеются свои ограничения. Для того чтобы гарантировать, что функционал $J(x)$ корректно определён и всюду конечен (т.е. не принимает значений $+\infty$ и $-\infty$), необходимо налагать *условия роста* на подынтегральную функцию $F(x, z, t)$. Действительно, если, например, $[a, b] = [0, 1]$ и $F(x, z, t) = z^4$, то для функции $x(t) = t^{3/4}$, принадлежащей

*Семинар «CNSA & NDO». Избранные доклады. 15 сентября 2016 г.

пространству $W_2^1[0,1]$, будет

$$J(x) := \int_0^1 x'(t)^4\,dt = \frac{81}{256}\int_0^1 \frac{1}{t}\,dt = +\infty.$$

Поэтому везде далее мы будем предполагать, что функция $F(x,z,t)$ удовлетворяет следующему условию: для любого $M>0$ существуют $c_1, c_2 > 0$ такие, что

$$|F(x,z,t)| \leqslant c_1|z|^2 + c_2 \quad \forall x \in [-M,M], \quad z \in \mathbb{R}, \quad t \in [a,b].$$

Нетрудно проверить, что при выполнении данного условия функционал $J(x)$ будет определён и конечен для всех $x \in W_2^1[a,b]$.

Поскольку при исследовании метода гиподифференциального спуска нам потребуется рассматривать производную Гато функционала $J(x)$, то помимо условия роста для функции $F(x,z,t)$, нам потребуются также условия роста и для производных этой функции, гарантирующие, что функционал $J(x)$ дифференцируем по Гато. А именно, далее мы будем предполагать, что для любого $M>0$ существуют $d_i > 0$, $i \in 1\colon 4$ такие, что для всех $x \in [-M,M]$, $z \in \mathbb{R}$ и $t \in [a,b]$ справедливы неравенства:

$$\Big|F'_x(x,z,t)\Big| \leqslant d_1|z|^2 + d_2, \quad \Big|F'_z(x,z,t)\Big| \leqslant d_3|z| + d_4.$$

Можно проверить, что при выполнении этого условия функционал $J(x)$ дифференцируем по Гато в каждой точке $x \in W_2^1[a,b]$ и его производная Гато имеет вид

$$J'(x;h) = \int_a^b \Big(F'_x(x(t),x'(t),t)h(t) + F'_z(x(t),x'(t),t)h'(t)\Big)\,dt \quad \forall h \in W_2^1[a,b].$$

(см., например, [6], Теорема 4.12).

2°. Метод гиподифференциального спуска для классической задачи вариационного исчисления. Опишем, как строится метод гиподифференциального спуска для решения задачи (1), (2). Для этого, следуя идеям В.Ф. Демьянова [1], «перейдём в пространство производных», то есть сделаем замену переменных $x' \to z$. Учитывая граничное условие $x(a) = A$, получим, что

$$x(t) = A + \int_a^t z(\tau)\,d\tau \quad \forall t \in [a,b].$$

Поэтому граничное условие $x(b) = B$ принимает вид

$$\int_a^b z(t)\,dt = B - A.$$

Следовательно, задача (1), (2) преобразуется к виду

$$I(z) = \int_a^b F\Big(A + \int_a^t z(\tau)\,d\tau, z(t), t\Big)\,dt \to \inf, \tag{3}$$

$$\int_a^b z(t)\,dt = B - A, \quad z \in L_2[a,b]. \tag{4}$$

Будем решать данную задачу с помощью теории точных штрафных функций [1]. Для этого введём штрафную функцию

$$\Phi_\lambda(z) = I(z) + \lambda\varphi(z),$$

где $\lambda \geqslant 0$ — штрафной параметр и

$$\varphi(z) = |\varphi_1(z)|, \quad \varphi_1(z) = \int_a^b z(t)\,dt + A - B.$$

Можно показать, что при некоторых дополнительных предположениях штрафная функция $\Phi_\lambda(z)$ является точной, то есть найдётся $\lambda^* \geqslant 0$ такое, что для любого $\lambda \geqslant \lambda^*$ задача (3), (4) эквивалента следующей задаче безусловной минимизации штрафной функции:

$$\Phi_\lambda(z) \to \inf, \quad z \in L_2[a,b]. \tag{5}$$

Поэтому далее мы будем строить численный метод решения задачи (5).

Нетрудно проверить, что штрафная функция $\Phi_\lambda(z)$ является *гиподифференцируемой* [1, 7] в каждой точке пространства $L_2[a,b]$, то есть для любых z, $h \in L_2[a,b]$ справедливо разложение:

$$\Phi_\lambda(z+\alpha h) - \Phi_\lambda(z) = \max_{[a,v]\in \underline{d}\Phi_\lambda(z)} \big(a + \alpha\langle v, h\rangle\big) + o(\alpha),$$

где $o(\alpha)/\alpha \to 0$ при $\alpha \to +0$, $\langle\cdot,\cdot\rangle$ — скалярное произведение в $L_2[a,b]$ и

$$\underline{d}\Phi_\lambda(z) = \operatorname{co}\Big\{\big(0, Q[z] + \lambda \operatorname{sign}\varphi_1(z)\big), \big(-2\lambda\varphi(z), Q[z] - \lambda\operatorname{sign}\varphi_1(z)\big)\Big\},$$

$$Q[z](t) = \int_t^b F_x'\Big(A + \int_a^\tau z(\xi)\,d\xi, z(\tau), \tau\Big)\,d\tau + F_z'\Big(A + \int_a^t z(\tau)\,d\tau, z(t), t\Big).$$

Заметим, что оператор $Q[z]$ является градиентом Гато функционала $I(z)$ в точке z (см. [1]).

Множество $\underline{d}\Phi_\lambda(z)$ называется *гиподифференциалом* функции $\Phi_\lambda(z)$ в точке z, а его элементы называются *гипоградиентами* данной функции в рассматриваемой точке. Необходимое условие минимума штрафной функции $\Phi_\lambda(z)$ можно записать в терминах гиподифференциала данной функции [1, 7]. А

именно, если $z^* \in L_2[a,b]$ является точкой локального минимума функционала $\Phi_\lambda(z)$, то

$$(0, \mathbb{O}) \in \underline{d}\Phi_\lambda(z^*). \tag{6}$$

Легко видеть, что данное условие эквивалентно уравнению Эйлера в интегральной форме.

Если в точке $z \in L_2[a,b]$ не выполнено необходимое условие минимума (6), то найдём наименьший по L_2-норме гипоградиент функционала Φ_λ в данной точке, то есть решим задачу

$$a^2 + \langle v, v\rangle \to \inf_{[a,v]\in\underline{d}\Phi_\lambda(z)}.$$

Оптимальный план данной задачи можно легко найти аналитически [2]. В случае когда $\varphi(z) = 0$, оптимальный план имеет вид $(0, G[z])$, где

$$G[z](t) = Q[z](t) - \frac{1}{b-a}\int_a^b Q[z](\tau)\,d\tau \quad \forall t \in [a,b]. \tag{7}$$

Можно показать [7], что направление $v = -G[z]$ является направление спуска функционала $\Phi_\lambda(z)$, то есть $\Phi_\lambda(z+\beta v) < \Phi_\lambda(z)$ для всех достаточно малых $\beta > 0$. Кроме того, заметим, что

$$\int_a^b \big(z(t) - \beta G[z](t)\big)\,dt = \int_a^b z(t)\,dt \quad \forall \beta \in \mathbb{R}.$$

Поэтому, в частности, направление $G[z]$ не выводит из множества планов задачи (3), (4).

Теперь мы можем записать теоретическую схему метода гиподифференциального спуска для минимизации функционала $\Phi_\lambda(z)$.

1) Выберем $z_0 \in L_2[a,b]$ такое, что $\int_a^b z_0(t)\,dt = B - A$.

2) Переход от k-го приближения к $(k+1)$-му осуществляется в следующем порядке ($k \geqslant 0$):

   (a) Вычислим направление спуска $G[z_k]$.

   (b) Найдём $\beta_k \geqslant 0$ такое, что

   $$\min_{\beta\geqslant 0} \Phi_\lambda(z_k - \beta G[z_k]) = \Phi_\lambda(z_k - \beta_k G[z_k]).$$

   (c) Положим $z_{k+1} = z_k - \beta_k G[z_k]$.

Заметим, что в качестве начального приближения в методе гиподифференциального спуска выбирается некоторый план z_0 задачи (3), (4). Поскольку направление $G[z_k]$ не выводит из множества планов данной задачи, то для любого $k \in \mathbb{N}$ точки z_k также будут планами задачи (3), (4) и $\Phi_\lambda(z_k - \beta G[z_k]) = I(z_k - \beta G[z_k])$ для всех $\beta \geqslant 0$.

3°. Сходимость метода гиподифференциального спуска. Перейдём теперь к вопросу сходимости метода гиподифференциального спуска, описанного выше. Данный метод был построен, как метод минимизации негладкой штрафной функции для задачи (3), (4), которая по построению эквивалентна исходной классической задаче вариационного исчисления. Для того чтобы исследовать сходимость метода гиподифференциального спуска необходимо посмотреть на этот метод с другой стороны.

Зафиксируем произвольный план z_0 задачи (3), (4), и введём линейное подпространство

$$Z = \Big\{ z \in L_2[a,b] \;\Big|\; \int_a^b z(t)\,dt = 0 \Big\}.$$

Ясно, что множество планов задачи (3), (4) совпадает с линейным многообразием $z_0 + Z$. Поэтому задачу (3), (4) можно переписать в следующем виде:

$$I(z) \to \inf, \quad z \in z_0 + Z.$$

Будем решать данную задачу методом проекции градиента. В качестве начального приближения возьмём точку z_0. Для того чтобы совершить один шаг по методу проекции градиента необходимо найти ортогональную проекцию градиента Гато функционала $I(z)$ на линейное подпространство Z. Напомним, что оператор

$$Q[z](t) = \int_t^b F'_x\Big(A + \int_a^\tau z(\xi)\,d\xi, z(\tau), \tau\Big)\,d\tau + F'_z\Big(A + \int_a^t z(\tau)\,d\tau, z(t), t\Big)$$

является градиентом Гато функционала $I(z)$. Кроме того, заметим, что оператор ортогонального проектирования на подпространство Z имеет вид

$$Pr_Z[v](t) = v(t) - \frac{1}{b-a}\int_a^b v(\tau)\,d\tau \quad \forall v \in L_2[a,b].$$

(см. [3], пункт 6). Следовательно, проекция градиента функционала $I[z]$ на подпространство Z совпадает с направлением $G[z]$, используемым на каждой итерации метода гиподифференциального спуска (см. (7)). Таким образом, метод гиподифференциального спуска для классической задачи вариационного исчисления совпадает с методом проекции градиента для функционала $I(z)$ при условии, что в качестве начального приближения в методе гиподифференциального спуска выбирается план задачи (3), (4).

З а м е ч а н и е. Следует отметить, что совпадение метода гиподифференциального спуска для минимизации негладкой штрафной функции и метода проекции градиента для решения исходной задачи не обусловлено какими-либо

особенностями конкретной задачи, а является общим фактом для всех задач с линейными ограничениями-равенствами [8]. Помимо этого отметим, что метод проекции градиента для решения классической задачи вариационного исчисления был рассмотрен Б.Т. Поляком ещё в начале 60-х годов [9].

Поскольку метод гиподифференциального спуска совпадает с методом проекции градиента, то для исследования сходимости данного метода можно использовать общие теоремы о сходимости градиентных методов для функционалов, определённых на банаховых пространствах [9, 10] (см. также [11], глава XV).

Пусть везде далее $\{z_k\}$, $k \geqslant 0$ — последовательность, построенная по методу гиподифференциального спуска (или, что тоже самое, по методу проекции градиента). Обозначим через $\|\cdot\|_2$ — норму пространства $L_2[a,b]$, а через $\|\cdot\|_\infty$ — равномерную норму на отрезке $[a,b]$ (т.е. стандартную норму пространства $C[a,b]$). Воспользовавшись общими теоремами из [9, 10, 11], нетрудно проверить, что справедливы следующие утверждения.

ТЕОРЕМА 1. *Пусть множество*

$$\left\{ z \in Z \;\middle|\; I(z+z_0) \leqslant I(z_0) \right\}$$

ограничено в $L_2[a,b]$, и оператор $z \to Q[z]$, действующий из $L_2[a,b]$ в $L_2[a,b]$, удовлетворяет условию Липшица на любом ограниченном множестве. Тогда $\|G[z_k]\|_2 \to 0$ при $k \to \infty$.

ТЕОРЕМА 2. *Пусть множество*

$$\left\{ z \in Z \;\middle|\; I(z+z_0) \leqslant I(z_0) \right\} \tag{8}$$

ограничено в $L_2[a,b]$, и оператор $z \to Q[z]$ удовлетворяет условию Липшица на любом ограниченном множестве. Предположим также, что функционал $I(z)$ является выпуклым. Тогда существует оптимальный план z^* задачи (3), (4) *и*

$$I(z_k) - I(z^*) = O\left(\frac{1}{k}\right). \tag{9}$$

Обозначим

$$x_k(t) = \int_a^t z_k(\tau)\, d\tau \quad \forall k \geqslant 0.$$

Заметим, что функции x_k являются планами задачи (1), (2).

ТЕОРЕМА 3. *Пусть множество*

$$\left\{ z \in Z \;\middle|\; I(z+z_0) \leqslant I(z_0) \right\}$$

ограничено в $L_2[a,b]$, *и оператор* $z \to Q[z]$ *удовлетворяет условию Липшица на любом ограниченном множестве. Предположим также, что функционал* $I(z)$ *является сильно выпуклым. Тогда существует единственный оптимальный план* z^* *задачи* (3), (4) *и найдутся* $C_1, C_2 > 0$ *и* $q \in (0,1)$ *такие, что*

$$\|z_k - z^*\|_2 \leqslant C_1 q^k, \quad \|x_k - x^*\|_\infty \leqslant C_2 q^k \quad \forall k \geqslant 0.$$

Для того чтобы применять приведённые выше теоремы к конкретным функционалам, необходимо уметь проверять ограниченность множества

$$\Big\{ z \in Z \;\Big|\; I(z + z_0) \leqslant I(z_0) \Big\},$$

липшицевость отображения $z \to Q[z]$ и выпуклость функционала $I(z)$. Ниже мы приводим простые достаточные условия, гарантирующие выполнение указанных предположений.

ЛЕММА 1 ([6], теорема 4.1)**.** *Предположим, что существуют* $0 \leqslant p < 2$, $\alpha_1 > 0$ *и* $\alpha_2, \alpha_3 \in \mathbb{R}$ *такие, что*

$$F(x,z,t) \geqslant \alpha_1 z^2 + \alpha_2 |x|^p + \alpha_3 \quad \forall (x,z,t) \in \mathbb{R}^2 \times [a,b]. \tag{10}$$

Тогда для любого плана z_0 *задачи* (3), (4) *множество*

$$\Big\{ z \in Z \;\Big|\; I(z + z_0) \leqslant I(z_0) \Big\}$$

ограничено в $L_2[a,b]$.

Доказательство. Зафиксируем произвольное $z \in L_2[a,b]$ и положим $x(t) = A + \int_a^t z(\tau)\,d\tau$. Воспользовавшись неравенством (10), получим

$$F(x(t), z(t), t) \geqslant \alpha_1 z(t)^2 + \alpha_2 |x(t)|^p + \alpha_3 \quad \text{для п.в. } t \in [a,b].$$

Интегрируя данное неравенство по t от a до b, имеем

$$I(z) \geqslant \alpha_1 \big(\|z\|_2\big)^2 - |\alpha_2| \int_a^b |x(t)|^p\,dt + \alpha_3 (b-a).$$

Из определения функции $x(t)$ следует, что

$$|x(t)| \leqslant \big|A\big| + \int_a^b |z(t)|\,dt \quad \forall t \in [a,b].$$

Отсюда, воспользовавшись неравенством Коши-Буняковского, получим

$$\|x\|_\infty \leqslant \big|A\big| + \sqrt{b-a}\|z\|_2. \tag{11}$$

Следовательно

$$I(z) \geqslant \alpha_1\big(\|z\|_2\big)^2 - |\alpha_2|(b-a)\Big(|A| + \sqrt{b-a}\|z\|_2\Big)^p + \alpha_3(b-a).$$

Поскольку $0 \leqslant p < 2$, то из предыдущего неравенства вытекает, что $I(z) \to +\infty$ при $\|z\|_2 \to +\infty$. Поэтому множество $\Big\{z \in Z \;\Big|\; I(z+z_0) \leqslant I(z_0)\Big\}$ ограничено в $L_2[a,b]$. □

ЛЕММА 2. *Пусть для любого $M > 0$ существует $L > 0$ такое, что для любых $t \in [a,b]$, $x_1, x_2 \in [-M, M]$ и $z_1, z_2 \in \mathbb{R}$ справедливы неравенства*

$$\Big|F'_x(x_1, z_1, t) - F'_x(x_2, z_2, t)\Big| \leqslant L|x_1 - x_2| + L|z_1 - z_2|, \tag{12}$$

$$\Big|F'_z(x_1, z_1, t) - F'_z(x_2, z_2, t)\Big| \leqslant L|x_1 - x_2| + L|z_1 - z_2|. \tag{13}$$

Тогда оператор $z \to Q[z]$ удовлетворяет условию Липшица на любом ограниченном подмножестве пространства $L_2[a,b]$.

Доказательство. Пусть $K \subset L_2[a,b]$ — непустое ограниченное множество. Из ограниченности множества K и неравенства (11) следует, что найдётся такое $M > 0$, что

$$\|x\|_\infty \leqslant M \quad \forall x \in \Big\{x \in W_2^1[a,b] \;\Big|\; x(t) = A + \int_a^t z(\tau)\,d\tau,\ z \in K\Big\}.$$

Для данного M найдётся $L > 0$, для которого выполнены неравенства (12) и (13).

Зафиксируем произвольные $z_1, z_2 \in K$ и обозначим $x_i(t) = A + \int_a^t z_i(\tau)\,d\tau$, $i \in \{1,2\}$. Воспользовавшись неравенствами (12) и (13), а также неравенствами Минковского и Коши-Буняковского, получим, что

$$\sqrt{\int_a^b \Big(F'_z(x_1(t), z_1(t), t) - F'_z(x_2(t), z_2(t), t)\Big)^2 dt} \leqslant L\|x_1 - x_2\|_2 + L\|z_1 - z_2\|_2 \tag{14}$$

и

$$\begin{aligned}\Big|\int_t^b \Big(F'_x(x_1(\tau), z_1(\tau), \tau) - F'_x(x_2(\tau), z_2(\tau), \tau)\Big)\,d\tau\Big| \leqslant \\ \leqslant \int_a^b \Big|F'_x(x_1(t), z_1(t), t) - F'_x(x_2(t), z_2(t), t)\Big|\,dt \leqslant \\ \leqslant L\int_a^b |x_1(t) - x_2(t)|\,dt + L\int_a^b |z_1(t) - z_2(t)|\,dt \leqslant \\ \leqslant L\sqrt{b-a}\|x_1 - x_2\|_2 + L\sqrt{b-a}\|z_1 - z_2\|_2. \end{aligned} \tag{15}$$

Напомним, что

$$Q[z](t) = \int_t^b F'_x\Big(A + \int_a^\tau z(\xi)\,d\xi, z(\tau), \tau\Big)\,d\tau + F'_z\Big(A + \int_a^t z(\tau)\,d\tau, z(t), t\Big).$$

Применяя неравенство Минковского и неравенства (14), (15), имеем

$$\begin{aligned}\|Q[z_1] - Q[z_2]\|_2 \leqslant \\ \leqslant \sqrt{\int_a^b \left(\int_t^b \Big[F'_x(x_1(\tau), z_1(\tau), \tau) - F'_x(x_2(\tau), z_2(\tau), \tau)\Big]\,d\tau\right)^2 dt} + \\ + \sqrt{\int_a^b \Big(F'_z(x_1(t), z_1(t), t) - F'_z(x_2(t), z_2(t), t)\Big)^2 dt} \leqslant \\ \leqslant L\big(b - a + \sqrt{b-a}\big)\Big(\|x_1 - x_2\|_2 + \|z_1 - z_2\|_2\Big).\end{aligned}$$

С помощью неравенства Коши–Буняковского нетрудно показать, что

$$\|x_1 - x_2\|_2 \leqslant (b-a)\|z_1 - z_2\|_2,$$

откуда

$$\big\|Q[z_1] - Q[z_2]\big\|_2 \leqslant L\big(b - a + \sqrt{b-a}\big)(b - a + 1)\|z_1 - z_2\|_2.$$

Следовательно, оператор $Q[z]$ удовлетворяет условию Липшица на множестве K. □

ЛЕММА 3. *Пусть функция $F(x, z, t)$ дважды непрерывно дифференцируема, и пусть существует $\mu_0 > 0$ такое, что для любых $(x, z, t) \in \mathbb{R}^2 \times [a, b]$ и $h_1, h_2 \in \mathbb{R}$ справедливо неравенство*

$$F''_{xx}(x, z, t)h_1^2 + 2F''_{xz}(x, z, t)h_1h_2 + F''_{zz}(x, z, t)h_2^2 \geqslant \mu_0 h_2^2 \tag{16}$$

(в частности, достаточно предполагать, что для любого $t \in [a, b]$ функция $(x, z) \to F(x, z, t)$ сильно выпукла с константой сильной выпуклости $\mu \geqslant \mu_0$). Предположим также, что для любого $M > 0$ найдутся $\alpha_i > 0$, $i \in 1\colon 5$ такие, что для любых $x \in [-M, M]$, $z \in \mathbb{R}$ и $t \in [a, b]$ справедливы неравенства

$$\big|F''_{xx}(x, z, t)\big| \leqslant \alpha_1 z^2 + \alpha_2, \quad \big|F''_{xz}(x, z, t)\big| \leqslant \alpha_3|z| + \alpha_4, \quad \big|F''_{zz}(x, z, t)\big| \leqslant \alpha_5. \tag{17}$$

Тогда функционал $I(z)$ является сильно выпуклым.

Доказательство. Для краткости мы приведём лишь общую идею доказательства. Воспользовавшись неравенствами (17) и теоремой Лебега о мажорируемой сходимости, можно показать, что функционал $I(z)$ является дважды дифференцируемым по Гато в каждой точке $z \in L_2[a,b]$ и его вторая производная Гато имеет вид

$$I''[z](h,h) = \int_a^b \Big[F''_{xx}(t)\Big(\int_a^t h(\tau)\,d\tau\Big)^2 + 2F''_{xz}(t)\Big(\int_a^t h(\tau)\,d\tau\Big)h(t) + F''_{zz}(t)h(t)^2\Big]\,dt,$$

где

$$F''_{xx}(t) = F''_{xx}\big(x(t),z(t),t\big), \quad F''_{xz}(t) = F''_{xz}\big(x(t),z(t),t\big),$$
$$F''_{zz}(t) = F''_{zz}\big(x(t),z(t),t\big).$$

и $x(t) = A + \int_a^t z(\tau)\,d\tau$. Теперь применяя неравенство (16), получим, что

$$I''[z](h,h) \geqslant \mu_0(\|h\|_2)^2 \quad \forall z,h \in L_2[a,b],$$

откуда заключаем, что функционал $I(z)$ является сильно выпуклым. □

Приведём простой пример функционала, удовлетворяющего всем указанным выше условиям. Пусть

$$J(x) = \int_a^b \Big(\gamma(t)x'(t)^2 + \omega(t)x'(t) + f(x(t),t)\Big)\,dt,$$

где функции $\gamma(t)$ и $\omega(t)$ непрерывны и $\gamma(t) > 0$ на $[a,b]$, а функция $f(x,t)$ дважды непрерывно дифференцируема и выпукла по x при каждом $t \in [a,b]$. Нетрудно проверить, что для этого функционала выполнены все предположения лемм 1–3. Поэтому с помощью теоремы 3 можно заключить, что в данном случае существует единственный оптимальный план x^* исходной задачи, и для любого начального приближения $z_0 \in L_2[a,b]$, удовлетворяющего ограничению $\int_a^b z_0(t)\,dt = B - A$, найдутся $C_1, C_2 > 0$ и $q \in (0,1)$ такие, что

$$\|z_k - z^*\|_2 \leqslant C_1 q^k, \quad \|x_k - x^*\|_\infty \leqslant C_2 q^k \quad \forall k \geqslant 0,$$

где $z^* = dx^*/dt$, то есть метод гиподифференциального спуска сходится со скоростью геометрической прогрессии.

ЛИТЕРАТУРА

1. Демьянов В. Ф. *Условия экстремума и вариационное исчисление.* М.: Высш. шк., 2004. 335 с.

2. Тамасян Г. Ш. *Гиподифференциальный спуск в вариационных задачах* // Семинар «CNSA & NDO». Избранные доклады. 25 сентября 2014 г. (`http://apmath.spbu.ru/cnsa/reps14.shtml#0925`) [Данная книга, с. 393]

3. Малозёмов В. Н., Тамасян Г. Ш. *Об одной кубической вариационной задаче* // Семинар «CNSA & NDO». Избранные доклады. 11 февраля 2016 г. (`http://apmath.spbu.ru/cnsa/reps16.shtml#0211`) [Данная книга, с. 346]

4. Осмоловский В. Г. *Нелинейная задача Штурма–Лиувилля: Учебное пособие.* СПб.: Изд–во СПбГУ, 2003. 257 с.

5. Leoni G. *A first course in Sobolev spaces.* Providence, RI: American Mathematical Society, 2009. 607 p.

6. Dacorogna B. *Direct methods in the calculus of variations.* New York: Springer Science+Business Media, 2008. 622 p.

7. Демьянов В. Ф., Долгополик М. В. *Кодифференцируемые функции в банаховых пространствах: методы и приложения к задачам вариационного исчисления* // Вестн. С.-Петербург. ун-та. Сер. 10. Прикл. матем. Информ. Проц. упр., 2013, вып. 3, с. 48–66.

8. Долгополик М. В., Тамасян Г. Ш. *Об эквивалентности методов наискорейшего и гиподифференциального спусков в некоторых задачах условной оптимизации* // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика, 2014, том 14, вып. 4(2), с. 532–542.

9. Поляк Б. Т. *Градиентные методы минимизации функционалов* // Ж. вычисл. матем. и матем. физ., 1963, том 3, № 4, с. 643–653.

10. Любич Ю. И., Майстровский Г. Д. *Общая теория релаксационных процессов для выпуклых функционалов* // УМН, 1970, Т. 25, вып. 1(151), с. 57–112.

11. Канторович Л. В., Акилов Г. П. *Функциональный анализ.* СПб.: Невский Диалект; БХВ-Петербург, 2004. 816 с.

12. Adams R. A. *Sobolev spaces.* New York: Academic Press, 1975. 286 p.

ГЛАВА 5. РАЗНОЕ

НЕРАВЕНСТВА И ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ*

В. Н. Малозёмов

Аннотация. Доклад посвящён элементарным методам в экстремальных задачах.

1°. Рассмотрим несколько конкретных экстремальных задач.

ЗАДАЧА 1. *Найти прямоугольный параллелепипед наибольшего объёма при заданной площади его поверхности.*

Эта задача легко формализуется. Пусть x_1, x_2, x_3 — длины рёбер параллелепипеда. Требуется максимизировать функцию

$$V(x) = x_1x_2x_3$$

при ограничениях

$$a(x) := x_1x_2 + x_2x_3 + x_3x_1 - p = 0,$$
$$x_1 > 0,\ x_2 > 0,\ x_3 > 0.$$

Здесь $p > 0$ — половина площади поверхности параллелепипеда.

Имеем задачу на условный экстремум. Согласно общим рекомендациям [1, с. 609–624] составим функцию Лагранжа

$$L(x, \lambda) = x_1x_2x_3 - \lambda(x_1x_2 + x_2x_3 + x_3x_1 - p)$$

и запишем необходимые условия экстремума

$$L'_{x_1}(x, \lambda) := x_2x_3 - \lambda(x_2 + x_3) = 0, \tag{1}$$
$$L'_{x_2}(x, \lambda) := x_1x_3 - \lambda(x_1 + x_3) = 0, \tag{2}$$
$$L'_{x_3}(x, \lambda) := x_1x_2 - \lambda(x_1 + x_2) = 0, \tag{3}$$
$$x_1x_2 + x_2x_3 + x_3x_1 = p. \tag{4}$$

Нужно найти решение этой системы с положительными x_1, x_2, x_3.

*Семинар «CNSA & NDO». Избранные доклады. 4 сентября 2014 г.

Ясно, что $\lambda \neq 0$. Согласно (1) и (2),

$$\frac{x_2 + x_3}{x_2 x_3} = \frac{x_1 + x_3}{x_1 x_3},$$

откуда следует, что $x_1 = x_2$. Согласно (2) и (3),

$$\frac{x_1 + x_3}{x_1 x_3} = \frac{x_1 + x_2}{x_1 x_2},$$

откуда следует, что $x_2 = x_3$. Значит, $x_1 = x_2 = x_3$. На основании (4) получаем

$$x_1^* = x_2^* = x_3^* = \sqrt{\tfrac{p}{3}}.$$

При этом

$$\lambda^* = \frac{x_2^* x_3^*}{x_2^* + x_3^*} = \tfrac{1}{2}\sqrt{\tfrac{p}{3}}.$$

Точка $x^* = (x_1^*,\ x_2^*,\ x_3^*)$ является стационарной. Покажем, что она удовлетворяет достаточному условию строгого локального максимума. Для этого найдём матрицу вторых производных функции Лагранжа

$$L''_{xx}(x^*, \lambda^*) = \begin{pmatrix} 0 & x_3^* - \lambda^* & x_2^* - \lambda^* \\ x_3^* - \lambda^* & 0 & x_1^* - \lambda^* \\ x_2^* - \lambda^* & x_1^* - \lambda^* & 0 \end{pmatrix} = \tfrac{1}{2}\sqrt{\tfrac{p}{3}} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

и запишем формулу для второго дифференциала

$$\left\langle L''_{xx}(x^*, \lambda^*)h,\ h \right\rangle = \sqrt{\tfrac{p}{3}}(h_1 h_2 + h_1 h_3 + h_2 h_3).$$

Нужно проверить, что второй дифференциал принимает отрицательные значения на всех ненулевых векторах h, удовлетворяющих ограничению

$$\left\langle a'(x^*), h \right\rangle = 0,$$

которое в данном случае принимает вид

$$2\ \sqrt{\tfrac{p}{3}}(h_1 + h_2 + h_3) = 0.$$

Имеем

$$0 = \tfrac{1}{2}\sqrt{\tfrac{p}{3}}(h_1 + h_2 + h_3)^2 = \tfrac{1}{2}\sqrt{\tfrac{p}{3}}(h_1^2 + h_2^2 + h_3^2) + \left\langle L''_{xx}(x^*, \lambda^*)h,\ h \right\rangle,$$

откуда следует, что

$$\left\langle L''_{xx}(x^*, \lambda^*)h,\ h \right\rangle = -\tfrac{1}{2}\sqrt{\tfrac{p}{3}}(h_1^2 + h_2^2 + h_3^2) < 0.$$

Таким образом, теория условного экстремума позволила установить, что в задаче о прямоугольном параллелепипеде наибольшего объёма при заданной площади его поверхности точка x^* с равными компонентами (что соответствует кубу) является точкой строгого локального максимума.

2°. Остаётся вопрос: будет ли x^* точкой *глобального* максимума? Теория условного экстремума не даёт ответа на этот вопрос («молчит наука»). Удивительно, но разобраться в данной ситуации помогает «совсем простая штука» — неравенство Коши между средним геометрическим и средним арифметическим положительных чисел $x_1, x_2, \ldots, x_n$, которое записывается так:

$$\sqrt[n]{x_1 x_2 \ldots x_n} \leqslant \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{5}$$

Неравенство (5) выполняется как равенство только тогда, когда $x_1 = x_2 = \cdots = x_n$.

В книге О. А. Иванова [2, с. 67–70] приводится пять доказательств неравенства Коши. На самом деле, их значительно больше. Одно из наиболее изящных доказательств будет представлено в Приложении 1.

Вернёмся к задаче о прямоугольном параллелепипеде. Согласно (5) при $n = 3$ и (4) имеем

$$V^2(x) = (x_1x_2)(x_2x_3)(x_3x_1) \leqslant \left(\frac{x_1x_2 + x_2x_3 + x_3x_1}{3}\right)^3 = \left(\frac{p}{3}\right)^3,$$

так что

$$V(x) \leqslant \left(\frac{p}{3}\right)^{3/2}.$$

Это неравенство выполняется как равенство только тогда, когда $x_1x_2 = x_2x_3 = x_3x_1$, то есть только при $x_1^* = x_2^* = x_3^* = \sqrt{\frac{p}{3}}$. В точке $x^* = (x_1^*,\ x_2^*,\ x_3^*)$ достигается максимальное значение функции $V(x)$, равное $(\frac{p}{3})^{3/2}$, и x^* — единственная точка, удовлетворяющая ограничениям задачи, с таким свойством. Значит, x^* — единственная точка глобального максимума.

3°. Рассмотрим ещё две экстремальные задачи, при решении которых можно эффектно использовать неравенство Коши.

ЗАДАЧА 2. *Среди треугольников, имеющих заданный периметр* $2p$*, найти треугольник с наибольшей площадью.*

Р е ш е н и е. Для площади треугольника справедлива формула Герона

$$S(x) = \sqrt{p(p - x_1)(p - x_2)(p - x_3)}.$$

Здесь x_1, x_2, x_3 — длины сторон треугольника и $x = (x_1,\ x_2,\ x_3)$. Требуется максимизировать $S(x)$ при ограничениях

$$x_1 + x_2 + x_3 = 2p,$$
$$0 < x_i < p, \quad i \in 1:3.$$

В силу неравенства Коши (5) имеем

$$S^2(x) = p\big[(p-x_1)(p-x_2)(p-x_3)\big] \leqslant p(\tfrac{p}{3})^3,$$

так что

$$S(x) \leqslant \frac{p^2}{3\sqrt{3}}.$$

Это неравенство выполняется как равенство только тогда, когда $p - x_1 = p - x_2 = p - x_3$, то есть только при $x_1^* = x_2^* = x_3^* = \frac{2}{3}p$. Значит, единственным решением задачи 2 является равносторонний треугольник.

В следующей задаче ответ не столь очевиден.

ЗАДАЧА 3. *Лист бумаги имеет форму круга. Из него вырежем сектор с углом φ радиан, а края оставшейся части склеим. Получим боковую поверхность прямого кругового конуса, длина образующей которого равна радиусу круга R. Требуется найти угол φ, порождающий конус наибольшего объёма.*

Р е ш е н и е. Длина дуги вырезанного сектора равна φR, так что длина окружности, лежащей в основании конуса, равна $2\pi R - \varphi R$. Обозначим через r радиус данной окружности. Тогда $2\pi r = 2\pi R - \varphi R$ и

$$\varphi = 2\pi(1 - \tfrac{r}{R}). \tag{6}$$

Высота конуса h равна $\sqrt{R^2 - r^2}$, а его объём выражается формулой

$$V(r) = \tfrac{1}{3}\pi r^2 h = \tfrac{1}{3}\pi r^2\sqrt{R^2 - r^2}.$$

В силу неравенства Коши (5) имеем

$$V^2(r) = \tfrac{4}{9}\pi^2\big[\tfrac{r^2}{2}\cdot\tfrac{r^2}{2}\cdot(R^2 - r^2)\big] \leqslant \tfrac{4}{9}\pi^2(\tfrac{R^2}{3})^3.$$

Значит,

$$V(r) \leqslant \frac{2\sqrt{3}}{27}\pi R^3.$$

Это неравенство выполняется как равенство только тогда, когда $\frac{r^2}{2} = R^2 - r^2$, то есть только при $r^* = \sqrt{\frac{2}{3}}R$. Принимая во внимание соотношение (6), приходим к заключению: задача 3 имеет единственное решение

$$\varphi^* = 2\pi(1 - \sqrt{\tfrac{2}{3}}).$$

4°. При решении экстремальных задач наряду с неравенством Коши используется неравенство Коши-Буняковского:

$$\sum_{i=1}^{n} x_i y_i \leqslant \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \cdot \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}. \tag{7}$$

Если не все x_i и не все y_i равны нулю, то неравенство (7) выполняется как равенство только тогда, когда $y_i = \lambda x_i$, $i \in 1:n$, при некотором $\lambda > 0$.

Следствием неравенства Коши-Буняковского является неравенство между средним арифметическим и средним квадратичным положительных чисел $x_1, x_2, \ldots, x_n$:

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \leqslant \left(\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}\right)^{1/2}. \tag{8}$$

Неравенство (8) выполняется как равенство только тогда, когда $x_1 = x_2 = \ldots = x_n$.

Оригинальное доказательство этих утверждений мы приведём в Приложении 2. А пока рассмотрим три конкретных задачи.

ЗАДАЧА 4. *Найти наибольшее и наименьшее значения линейной формы*

$$f(x) = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n,$$

в которой не все коэффициенты c_i равны нулю, при ограничении

$$x_1^2 + x_2^2 + \ldots + x_n^2 = 1. \tag{9}$$

Р е ш е н и е. Согласно (7) для любого вектора $x = (x_1, x_2, \ldots, x_n)$, удовлетворяющего ограничению (9), имеем

$$f(x) \leqslant \left(\sum_{i=1}^{n} c_i^2\right)^{1/2} =: \|c\|.$$

Равенство достигается только тогда, когда $x_i = \lambda c_i$, $i \in 1:n$, при некотором $\lambda > 0$. В силу (9), $\lambda = \frac{1}{\|c\|}$ и

$$x_i^* = \tfrac{c_i}{\|c\|}, \quad i \in 1:n.$$

Получили, что вектор $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ является единственной точкой максимума функции $f(x)$ при ограничении (9).

Далее запишем

$$-f(x) = (-c_1)x_1 + (-c_2)x_2 + \ldots + (-c_n)x_n.$$

Аналогично предыдущему для любого вектора x, удовлетворяющего ограничению (9), получим $-f(x) \leqslant A$ или $f(x) \geqslant -A$, причём равенство достигается только тогда, когда $x_i = \lambda(-c_i)$, $i \in 1:n$, при некотором $\lambda > 0$. Отсюда легко приходим к выводу о том, что вектор $-x^*$ является единственной точкой минимума функции $f(x)$ при ограничении (9).

ЗАДАЧА 5. *Найти наименьшее значения квадратичной функции*

$$f(x) = x_1^2 + x_2^2 + \ldots + x_n^2$$

при ограничении

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n \geqslant b, \tag{10}$$

где не все коэффициенты a_i равны нулю.

Р е ш е н и е. При $b \leqslant 0$ очевидной точкой минимума является нулевой вектор x. Поэтому будем считать, что $b > 0$.

Согласно неравенствам (10) и (7) имеем

$$0 < b \leqslant \sum_{i=1}^{n} a_ix_i \leqslant \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \cdot \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

Отсюда следует, что

$$f(x) \geqslant \left(\tfrac{b}{\|a\|}\right)^2.$$

Равенство достигается только тогда, когда

$$\sum_{i=1}^{n} a_ix_i = b$$

и

$$\sum_{i=1}^{n} a_ix_i = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2} \cdot \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

Это соответствует тому, что $x_i = \lambda a_i$, $i \in 1:n$, при некотором $\lambda > 0$ и

$$\lambda \sum_{i=1}^{n} a_i^2 = b.$$

Получаем

$$\lambda = \tfrac{b}{\|a\|^2} \quad \text{и} \quad x_i^* = \tfrac{ba_i}{\|a\|^2},\ i \in 1:n.$$

Вектор $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ при $b > 0$ является единственной точкой минимума квадратичной функции $f(x)$ на полупространстве, определяемом неравенством (10).

ЗАДАЧА 6. *Обозначим через x_1, x_2, x_3 длины сторон треугольника. Найти наименьшее значение величины*

$$x_1^2 + x_2^2 + x_3^2$$

при условии, что площадь треугольника равна S.

Р е ш е н и е. Положим $p = \frac{1}{2}(x_1 + x_2 + x_3)$. На основании формулы Герона для площади треугольника и неравенств (5), (8) для средних величин получаем

$$S^2 = p\big[(p-x_1)(p-x_2)(p-x_3)\big] \leqslant p(\tfrac{p}{3})^3 = 3(\tfrac{p}{3})^4 = \\ = \frac{3}{16}\left(\frac{x_1+x_2+x_3}{3}\right)^4 \leqslant \frac{3}{16}\left(\frac{x_1^2+x_2^2+x_3^2}{3}\right)^2. \qquad (11)$$

Отсюда следует, что

$$x_1^2 + x_2^2 + x_3^2 \geqslant (4\sqrt{3})S.$$

Равенство в этом неравенстве достигается только тогда, когда оба неравенства в (11) выполняются как равенства. А это возможно только при $x_1 = x_2 = x_3$.

Приходим к заключению: наименьшее значение величины $x_1^2+x_2^2+x_3^2$ равно $(4\sqrt{3})S$ и достигается на равностороннем треугольнике.

5°. Вернёмся к задаче 1 о прямоугольном параллелепипеде наибольшего объёма. Было установлено, что в этой задаче имеется единственная стационарная точка $x^* = (x_1^*,\ x_2^*,\ x_3^*)$ с равными компонентами $x_1^* = x_2^* = x_3^* = \sqrt{\frac{p}{3}}$. То, что x^* является точкой глобального максимума, можно выяснить с помощью соображений, отличных от приведённых в п. 2°. Достаточно доказать, что в задаче 1 максимизации функции $V(x)$ на множестве планов Ω точка максимума $\hat{x}$ существует и ограничение в ней регулярно. Тогда согласно теории условного экстремума точка $\hat{x}$ будет стационарной, а поскольку стационарная точка единственна, это x^*, то необходимо $x^* = \hat{x}$.

Покажем, что точка максимума $\hat{x}$ существует[1]. Отягчающим обстоятельством является неограниченность множества планов Ω. Этому множеству принадлежат, например, точки

$$x^{(n)} = \big(n,\ \tfrac{p}{2n},\ \tfrac{pn}{2n^2+p}\big), \quad n = 1, 2, \ldots$$

Вместе с тем, справедливо следующее утверждение: если для плана $x \in \Omega$ выполняется неравенство

$$\|x\|_\infty := \max\{x_1, x_2, x_3\} \geqslant N,$$

[1]Доказательство предложил М. В. Долгополик.

то

$$V(x) \leqslant \tfrac{p^2}{N}. \tag{12}$$

Действительно, пусть, например, $x_1 \geqslant N$. В силу ограничения (4) имеем $x_1x_2 \leqslant p$, $x_1x_3 \leqslant p$, так что

$$V(x) \leqslant (x_1x_2)(x_1x_3)\tfrac{1}{x_1} \leqslant \tfrac{p^2}{N}.$$

Введём обозначение

$$\mu = \sup_{x\in\Omega} V(x).$$

Ясно, что $\mu > 0$. Возьмём максимизирующую последовательность планов $x^{(k)} \in \Omega$, для которых

$$\lim_{k\to\infty} V(x^{(k)}) = \mu. \tag{13}$$

Последовательность $\{x^{(k)}\}$ ограничена. В противном случае найдётся подпоследовательность $\{x^{(k_j)}\}$, такая, что $\|x^{(k_j)}\|_\infty \to +\infty$ при $k_j \to \infty$. Согласно (12),

$$V(x^{(k_j)}) \to 0 \quad \text{при} \quad k_j \to \infty,$$

что противоречит (13) и условию $\mu > 0$.

Выделим из последовательности $\{x^{(k)}\}$ сходящуюся подпоследовательность. Можно считать, что вся последовательность $\{x^{(k)}\}$ сходится при $k \to \infty$ к некоторой точке $\hat{x}$. В пределе получаем

$$V(\hat{x}) = \lim_{k\to\infty} V(x^{(k)}) = \mu,$$

то есть

$$V(\hat{x}) = \mu. \tag{14}$$

К этому нужно добавить, что $\hat{x}$ принадлежит Ω, так как согласно (14) все компоненты вектора $\hat{x}$ положительны, а ограничение (4) при $x = \hat{x}$ выполняется из соображений непрерывности.

Таким образом, установлено, что у задачи 1 точка максимума $\hat{x}$ существует. Ограничение в ней регулярно в силу того, что градиент

$$a'(\hat{x}) = (\hat{x}_2 + \hat{x}_3, \hat{x}_1 + \hat{x}_3, \hat{x}_1 + \hat{x}_2)$$

отличен от нулевого вектора.

ПРИЛОЖЕНИЕ 1

Доказательство неравенства Коши

Начнём со вспомогательного утверждения.

ЛЕММА 1. *Пусть $x_1, x_2, \ldots, x_n$ — положительные числа, не равные между собой, и*

$$x_1 x_2 \ldots x_n = 1. \tag{15}$$

Тогда

$$x_1 + x_2 + \ldots + x_n > n.$$

Доказательство. Воспользуемся методом математической индукции.

При $n = 2$ имеем $x_1 > 0$, $x_2 > 0$, $x_1 \neq x_2$ и $x_1 x_2 = 1$, поэтому

$$0 < (x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1 x_2 = (x_1 + x_2)^2 - 4.$$

Отсюда следует, что $x_1 + x_2 > 2$.

Сделаем индукционный переход от n к $n+1$. Возьмём $n+1$ положительных чисел $x_1, \ldots, x_n, x_{n+1}$, которые не все равны между собой и удовлетворяют условию

$$x_1 \ldots x_n x_{n+1} = 1.$$

Ясно, что среди них существует число, меньшее единицы, и число, большее единицы. Пусть для определённости $x_n < 1$, $x_{n+1} > 1$. Тогда

$$(1 - x_n)(x_{n+1} - 1) > 0$$

и

$$x_n x_{n+1} < x_n + x_{n+1} - 1. \tag{16}$$

Рассмотрим n положительных чисел $x_1, \ldots, x_{n-1}, (x_n x_{n+1})$, произведение которых равно единице. Если не все они равны между собой, то по индукционному предположению

$$x_1 + \ldots + x_{n-1} + (x_n x_{n+1}) > n. \tag{17}$$

Если же $x_1 = \ldots = x_{n-1} = (x_n x_{n+1}) = 1$ (это возможно, например, при $x_1 = \ldots = x_{n-1} = 1$, $x_n = \frac{1}{2}$, $x_{n+1} = 2$), то строгое неравенство (17) нужно заменить равенством. В обоих случаях

$$x_1 + \ldots + x_{n-1} + (x_n x_{n+1}) \geqslant n.$$

С учётом (16) получаем

$$x_1 + \ldots + x_{n-1} + x_n + x_{n+1} - 1 > n,$$

что равносильно требуемому. □

Теперь легко доказать неравенство Коши. Пусть $x_1, x_2, \ldots, x_n$ — положительные числа, не все равные между собой. Обозначим

$$y_i = \frac{x_i}{\sqrt[n]{x_1 x_2 \ldots x_n}}, \quad i \in 1:n.$$

Числа y_i положительные, не все равные между собой и $y_1 y_2 \ldots y_n = 1$. По лемме $y_1 + y_2 + \ldots + y_n > n$, что равносильно неравенству

$$\frac{x_1 + x_2 + \ldots + x_n}{n} > \sqrt[n]{x_1 x_2 \ldots x_n}. \tag{18}$$

Если все x_i равны между собой (и только в этом случае) строгое неравенство (18) нужно заменить равенством.

ПРИЛОЖЕНИЕ 2

Доказательство неравенства Коши-Буняковского

Введём обозначения

$$x = (x_1, x_2, \ldots, x_n), \quad y = (y_1, y_2, \ldots, y_n),$$
$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i, \quad \|x\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} = \sqrt{\langle x, x\rangle}.$$

ЛЕММА 2. *Для ненулевых векторов x, y справедливо равенство Коши-Буняковского* [2]

$$\langle x, y\rangle = \left(1 - \tfrac{1}{2}\left\|\tfrac{x}{\|x\|} - \tfrac{y}{\|y\|}\right\|^2\right)\|x\| \cdot \|y\|. \tag{19}$$

Доказательство. Рассмотрим сначала случай, когда $\|x\| = \|y\| = 1$. Имеем

$$\|x - y\|^2 = \langle x - y, x - y\rangle = \|x\|^2 - 2\langle x, y\rangle + \|y\|^2 = 2 - 2\langle x, y\rangle.$$

Отсюда следует, что

$$\langle x, y\rangle = 1 - \tfrac{1}{2}\|x - y\|^2. \tag{20}$$

Если x, y — произвольные ненулевые векторы, то подставив в (20) $\frac{x}{\|x\|}$ и $\frac{y}{\|y\|}$ вместо x и y, придём к (19). □

[2] См. [3, с. 33].

Равенство (19) для ненулевых векторов x, y делает очевидными как неравенство Коши-Буняковского

$$\langle x, y\rangle \leqslant \|x\| \cdot \|y\|, \tag{21}$$

так и условие обращения его в равенство

$$\tfrac{x}{\|x\|} = \tfrac{y}{\|y\|}. \tag{22}$$

Условие (22) равносильно тому, что $y = \lambda x$ при некотором $\lambda > 0$.

Положив в (21) $y_i = \frac{1}{n}$, $i \in 1 : n$, придём к неравенству между средним арифметическим и средним квадратичным положительных чисел $x_1, x_2, \ldots, x_n$:

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \leqslant \left(\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}\right)^{1/2}. \tag{23}$$

Равенство в (23) достигается только тогда, когда $x_1 = x_2 = \ldots = x_n$.

ЛИТЕРАТУРА

1. Зорич В. А. *Математический анализ.* Часть 1. Изд-е 4-е. М.: МЦНМО, 2002. 664 с.

2. Иванов О. А. *Избранные главы элементарной математики.* СПб.: Изд-во СПбГУ, 1995. 224 с.

3. Малозёмов В. Н. *Линейная алгебра без определителей. Квадратичная функция.* СПб.: Изд-во СПбГУ, 1997. 80 с.

СТУДЕНТЫ РЕШАЮТ ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ...*

В. Н. Малозёмов

С 1986 г. я читаю на математико-механическом факультете Санкт-Петербургского государственного университета лекции по экстремальным задачам для студентов 3-го курса отделения прикладной математики и информатики. На лекциях студентам предлагаются для самостоятельного решения как стандартные задачи (проверка плана общей задачи линейного программирования на оптимальность, решение задачи квадратичного программирования с помощью теоремы Куна–Таккера, решение квадратичных вариационных задач), так и менее стандартные. Иногда студенты находят оригинальные, красивые решения нестандартных задач. Примеры таких решений вместе с комментариями к ним я приведу в этом докладе.

ЗАДАЧА 1. *Пусть D — симметричная положительно определённая матрица порядка n и c — ненулевой n-мерный вектор. Найдите глобальное решение экстремальной задачи*

$$\begin{gathered} f(x) := \langle c, x\rangle \to \max, \\ \langle Dx, x\rangle = 1 \end{gathered} \tag{1}$$

и докажите, что оно единственно.

Р е ш е н и е (М. Кольцов, Е. Ржевская, 2014 г.). Запишем формально необходимые условия локального максимума:

$$c = uDx, \tag{2}$$

$$\langle Dx, x\rangle = 1. \tag{3}$$

Так как c — ненулевой вектор, то и $u \neq 0$.

Умножим (2) скалярно на x. Учитывая (3), получаем

$$\langle c, x\rangle = u. \tag{4}$$

*Семинар «CNSA & NDO». Избранные доклады. 12 февраля 2015 г.

Далее, у симметричной положительно определённой матрицы D существует обратная матрица D^{-1}, которая также является симметричной и положительно определённой. Из (2) следует, что

$$x = \frac{1}{u} D^{-1} c. \tag{5}$$

Условие (3) принимает вид

$$\langle c, D^{-1} c \rangle = u^2,$$

так что

$$u = \pm \sqrt{\langle c, D^{-1} c \rangle}.$$

На основании (4) и того факта, что мы максимизируем функцию $\langle c, x \rangle$, выбираем

$$u = \sqrt{\langle c, D^{-1} c \rangle}.$$

Согласно (5) приходим к формуле

$$x_* = \frac{D^{-1} c}{\sqrt{\langle c, D^{-1} c \rangle}}. \tag{6}$$

Покажем, что $f(x) < f(x_*)$ для любого плана x, отличного от x_*. Этим и завершится решение задачи (1).

Возьмём план x, отличный от x_*, и обозначим $h = x - x_*$, $h \neq \mathbb{O}$. Имеем

$$1 = \langle Dx, x \rangle = \big\langle D(x_* + h), x_* + h \big\rangle = \langle Dx_*, x_* \rangle + 2\langle Dx_*, h \rangle + \langle Dh, h \rangle.$$

Отсюда следует, что

$$2\langle Dx_*, h \rangle = -\langle Dh, h \rangle < 0.$$

В силу (6) приходим к неравенству

$$\langle c, h \rangle < 0.$$

Окончательно получаем

$$f(x) = f(x_* + h) = \langle c, x_* \rangle + \langle c, h \rangle < \langle c, x_* \rangle = f(x_*),$$

то есть $f(x) < f(x_*)$. □

К о м м е н т а р и й. Я имел в виду другое решение, основанное на связи задачи (1) с взаимной по Эйлеру задачей

$$\begin{gathered} g(x) := \langle Dx, x \rangle \to \min, \\ \langle c, x \rangle = 1 \end{gathered} \tag{7}$$

(по сравнению с задачей (1) целевая функция и функция, входящая в ограничение, поменялись местами). У задачи (7) целевая функция выпуклая и

ограничение линейное. Её решение x_0 находится стандартным путём из критерия оптимальности. Оно единственно и имеет вид

$$x_0 = \frac{D^{-1}c}{\langle c, D^{-1}c\rangle}.$$

Обозначим

$$\mu = \langle Dx_0, x_0\rangle = \frac{1}{\langle c, D^{-1}c\rangle}.$$

Вектор $x_* = \frac{1}{\sqrt{\mu}}x_0$ является планом задачи (1). При этом

$$\langle c, x_*\rangle = \frac{1}{\sqrt{\mu}}. \tag{8}$$

Покажем, что для любого плана x_1 задачи (1) выполняется неравенство

$$\langle c, x_1\rangle \leqslant \frac{1}{\sqrt{\mu}}. \tag{9}$$

Если $\langle c, x_1\rangle \leqslant 0$, то это очевидно. Пусть $\langle c, x_1\rangle > 0$. Вектор

$$x_2 = \frac{x_1}{\langle c, x_1\rangle}$$

удовлетворяет ограничению задачи (7), поэтому

$$\langle Dx_2, x_2\rangle \geqslant \langle Dx_0, x_0\rangle = \mu.$$

Отсюда следует, что

$$\frac{1}{\big[\langle c, x_1\rangle\big]^2} \geqslant \mu \quad \text{и} \quad \langle c, x_1\rangle \leqslant \frac{1}{\sqrt{\mu}}.$$

На основании (8) и (9) заключаем, что x_* — решение задачи (1).

Единственность решения задачи (1) связана с единственностью решения задачи (7).

Отмечу также, что задача 1 имеет элементарное решение, основанное на обобщённом неравенстве Коши–Буняковского. Введём в $\mathbb{R}^n$ обобщённое скалярное произведение

$$\langle x, y\rangle_D = \langle Dx, y\rangle$$

и обобщённую норму $\|x\|_D = \sqrt{\langle x, x\rangle_D} = \sqrt{\langle Dx, x\rangle}$. В книге [1, с. 33] показано, что справедливо неравенство

$$\langle x, y\rangle_D \leqslant \|x\|_D \cdot \|y\|_D, \tag{10}$$

причём при ненулевых x, y неравенство обращается в равенство только тогда, когда

$$\frac{x}{\|x\|_D} = \frac{y}{\|y\|_D}.$$

Воспользуемся этим утверждением для решения задачи 1. Согласно (10) для любого плана x имеем

$$f(x) = \big\langle D(D^{-1}c), x\big\rangle \leqslant \|D^{-1}c\|_D \cdot \|x\|_D = \sqrt{\langle c, D^{-1}c\rangle}.$$

Неравенство выполняется как равенство только тогда, когда

$$\frac{D^{-1}c}{\|D^{-1}c\|_D} = \frac{x}{\|x\|_D}.$$

Отсюда следует, что

$$x = \frac{D^{-1}c}{\sqrt{\langle c, D^{-1}c\rangle}}.$$

Этот вектор является единственным решением задачи (1).

ЗАДАЧА 2. *Найдите глобальное решение вариационной задачи*

$$\begin{gathered} J(x) := \int_0^1 x^2 (x')^2 dt \to \min, \\ x(0) = 1,\ x(1) = \sqrt{2},\ x \in C^1[0,1], \end{gathered} \tag{11}$$

и докажите его единственность.

Р е ш е н и е (А. Петров, 1998 г.). Нам потребуется следующее вспомогательное утверждение.

ЛЕММА. *Для любой непрерывной на отрезке* $[0,1]$ *функции* $y(t)$ *справедливо неравенство*

$$\Big(\int_0^1 y(t)dt\Big)^2 \leqslant \int_0^1 y^2(t)dt. \tag{12}$$

Неравенство выполняется как равенство тогда и только тогда, когда $y(t) \equiv \mathrm{const}$.

Д о к а з а т е л ь с т в о. При любом вещественном α имеем

$$\begin{aligned} 0 \leqslant \int_0^1 (y-\alpha)^2 dt &= \int_0^1 y^2 dt - 2\alpha \int_0^1 y dt + \alpha^2 = \\ &= \Big(\alpha - \int_0^1 y dt\Big)^2 - \Big(\int_0^1 y dt\Big)^2 + \int_0^1 y^2 dt. \end{aligned}$$

Положив $\alpha = \int_0^1 y dt$, придём к требуемому неравенству.

Если неравенство выполняется как равенство, то

$$0 \leqslant \int_0^1 (y-\alpha)^2 dt = \Big(\alpha - \int_0^1 y dt\Big)^2.$$

Положив и здесь $\alpha = \int_0^1 y dt$, получим

$$\int_0^1 (y-\alpha)^2 dt = 0.$$

Отсюда следует, что $y(t) \equiv \alpha$.

То, что при $y(t) \equiv \text{const}$ неравенство (12) выполняется как равенство, очевидно. Лемма доказана. □

Переходим к решению задачи (11). Возьмём план x и обозначим $y = xx'$. В силу (12) имеем

$$J(x) = \int_0^1 y^2 dt \geqslant \Big(\int_0^1 y dt\Big)^2 = \Big(\int_0^1 xx' dt\Big)^2 = \frac{1}{4}x^2\Big|_0^1 = \frac{1}{4}.$$

По лемме неравенство выполняется как равенство только тогда, когда $xx' \equiv \equiv \frac{1}{2}\alpha$, то есть при $x^2 = \alpha t + \beta$. Учитывая краевые условия, получаем $\beta = 1$, $\alpha = 1$. Таким образом,

$$x^2(t) = t + 1, \quad t \in [0,1].$$

В частности, на отрезке $[0,1]$ функция $x^2(t)$ не обращается в нуль. Значит, и $x(t)$ на отрезке $[0,1]$ не обращается в нуль. В силу краевых условий она должна быть положительной. Получаем

$$x_*(t) = \sqrt{t+1}.$$

Эта функция является единственным решением задачи 2.

Комментарий. Неравенства играют важную роль при решении экстремальных задач. При этом должны быть известны все случаи, когда неравенство выполняется как равенство. Подробнее об этом см. в докладе [2].

ЗАДАЧА 3. *Пусть A — невырожденная матрица порядка n и $B = x_0 y_0^T$, где x_0, y_0 — ненулевые n-мерные векторы. Найдите формулу для обратной матрицы $(A-B)^{-1}$.*

Решение (А. Петров, 1998 г.; А. Проскурников, 2001 г.). Имеем

$$A - B = (E - BA^{-1})A. \tag{13}$$

Далее,

$$(E - BA^{-1})x_0 = x_0 - x_0 y_0^T A^{-1} x_0.$$

Обозначим $\eta = y_0^T A^{-1} x_0 = \langle y_0, A^{-1}x_0\rangle$. Тогда

$$(E - BA^{-1})x_0 = (1-\eta)x_0.$$

Если $\eta = 1$, то матрица $E - BA^{-1}$ необратима (вместе с $A - B$ — см. (13)).

Предположим, что $\eta \neq 1$. В этом случае

$$(1-\eta)^{-1}(E - BA^{-1})x_0 y_0^T A^{-1} = x_0 y_0^T A^{-1} = BA^{-1} = \\ = E - (E - BA^{-1}).$$

Получаем

$$(E - BA^{-1})\big(E + (1-\eta)^{-1}BA^{-1}\big) = E.$$

Отсюда следует, что

$$(E - BA^{-1})^{-1} = E + (1-\eta)^{-1}BA^{-1}.$$

В силу (13)

$$(A - B)^{-1} = A^{-1} + (1-\eta)^{-1}A^{-1}BA^{-1}. \tag{14}$$

К о м м е н т а р и й. Формула (14) хорошо известна. Она называется *формулой Шермана–Моррисона*. Её справедливость, когда она записана, легко проверить непосредственно. Однако мне было интересно, как формула (14) *выводится*. А. Петров и А. Проскурников блестяще с этим разобрались.

В дальнейшем мы с А. Петровым и Ф. Монако продолжали заниматься данной темой и опубликовали статью [3]. Рассмотрим частную ситуацию, которая возникает при описании симплекс-метода. У обратимой матрицы A k-й столбец A_k заменяется на столбец C. Как при этом изменится обратная матрица?

Новая матрица имеет вид

$$S = A - (A_k - C)e_k^T,$$

где e_k — k-й орт. В данном случае

$$1 - \eta = 1 - \big\langle e_k, A^{-1}(A_k - C)\big\rangle = \langle e_k, A^{-1}C\rangle.$$

Обозначим $z = A^{-1}C$. Тогда $1 - \eta = z[k]$. Критерием обратимости матрицы S является условие $z[k] \neq 0$. При выполнении этого условия в силу (14) получаем

$$S^{-1} = A^{-1} + \big(z[k]\big)^{-1}A^{-1}(A_k - C)e_k^T A^{-1} = \\ = A^{-1} + \big(z[k]\big)^{-1}(e_k - z)e_k^T A^{-1}. \tag{15}$$

Матричное равенство (15) принимает простой вид, если расписать его по строкам:

$$S^{-1}[k, N] = \big(z[k]\big)^{-1}A^{-1}[k, N]; \\ S^{-1}[i, N] = A^{-1}[i, N] - z[i]S^{-1}[k, N], \quad i \in N \setminus \{k\}.$$

Здесь $N = 1 : n$.

ЗАДАЧА 4. Пусть A — $(n \times m)$-матрица, $n < m$, b — ненулевой n-мерный вектор. Рассмотрим две экстремальные задачи

$$\begin{matrix} \|\xi\|_\infty \to \min, \\ A\xi = b, \end{matrix} \qquad (16) \qquad \begin{matrix} \|A^T\eta\|_1 \to \min, \\ \langle b, \eta\rangle = 1. \end{matrix} \qquad (17)$$

Здесь $\|\xi\|_\infty = \max\limits_{j\in 1:m}\big|\xi[j]\big|$ и $\|y\|_1 = \sum\limits_{j=1}^{m}\big|y[j]\big|$, где $y = A^T\eta$.

Докажите следующее утверждение: *если множество планов задачи* (16) *непусто, то обе задачи* (16) *и* (17) *имеют решения; при этом минимальные значения μ, ν их целевых функций связаны соотношением $\mu \cdot \nu = 1$.*

Р е ш е н и е (А. Демьянов, 2002 г.). Прежде всего покажем, что задача (16) имеет решение. Запишем эквивалентную ей задачу линейного программирования:

$$\begin{gathered} \rho \to \min, \\ A\xi = b, \\ -\rho \leqslant \xi[j] \leqslant \rho, \quad j \in 1:m. \end{gathered} \qquad (18)$$

Множество планов задачи (18) непусто $\big($оно содержит вектор (ξ_0, ρ_0), где ξ_0 — план задачи (16) и $\rho_0 = \max\limits_{j\in 1:m}\big|\xi_0[j]\big|\big)$ и целевая функция на множестве планов ограничена снизу нулём. Это гарантирует существование оптимального плана у задачи (18). В силу эквивалентности существует оптимальный план и у задачи (16). Обозначим его ξ_*. По определению

$$\mu = \|\xi_*\|_\infty.$$

Так как $b \neq \mathbb{O}$, то $\mu > 0$.

Теперь обратимся к задаче (17) и запишем эквивалентную ей задачу линейного программирования. Воспользуемся тем, что компоненты $y[j]$ вектора $y = A^T\eta$ допускают представление

$$y[j] = \alpha[j] - \beta[j], \quad \alpha[j] \geqslant 0,\ \beta[j] \geqslant 0,$$

при этом $\big|y[j]\big| = \alpha[j] + \beta[j]$. Получим

$$\begin{gathered} \sum_{j=1}^{m}\big(\alpha[j] + \beta[j]\big) \to \min, \\ -y[j] + \alpha[j] - \beta[j] = 0, \quad j \in 1:m; \\ \alpha[j] \geqslant 0,\ \beta[j] \geqslant 0, \quad j \in 1:m; \\ \langle b, \eta\rangle = 1. \end{gathered} \qquad (19)$$

Введём вектор $e = (1, 1, \ldots, 1)$ и перепишем задачу (19) в векторном виде

$$\begin{aligned} \langle e, \alpha + \beta\rangle &\to \min, \\ -A^T\eta + \alpha - \beta &= \mathbb{O}, \\ \langle b,\eta\rangle &= 1, \\ \alpha \geqslant \mathbb{O},\ \beta &\geqslant \mathbb{O}. \end{aligned} \tag{20}$$

Перейдём к двойственной задаче:

$$\begin{aligned} v &\to \max, \\ -Au + vb &= \mathbb{O}, \\ u \leqslant e,\ -u &\leqslant e. \end{aligned} \tag{21}$$

Нетрудно проверить, что вектор $(u_*, v_*) := (\xi_*/\mu, 1/\mu)$ является планом задачи (21) (учесть, что вторая строчка ограничений равносильна условию $\|u\|_\infty \leqslant 1$). Более того, этот план — оптимальный. Действительно, допустим противное. Тогда у задачи (21) существует план (u_0, v_0) с бóльшим значением целевой функции, то есть

$$v_0 > \frac{1}{\mu}.$$

Положим $\hat{\xi} = \frac{1}{v_0}u_0$. Вектор $\hat{\xi}$ удовлетворяет ограничению задачи (16) и

$$\|\hat{\xi}\|_\infty = \frac{1}{v_0}\|u_0\|_\infty \leqslant \frac{1}{v_0} < \mu.$$

Это противоречит определению μ.

Таким образом, задача (21) имеет решение и максимальное значение её целевой функции равно $1/\mu$. По первой теореме двойственности в линейном программировании задача (20) также имеет решение и минимальное значение её целевой функции равно $1/\mu$. Наконец, в силу эквивалентности задача (17) имеет решение и минимальное значение её целевой функции равно $1/\mu$. По определению

$$v = \frac{1}{\mu},$$

так что $\mu \cdot \nu = 1$.

К о м м е н т а р и й. На лекциях я рассматривал «симметричную» пару экстремальных задач

$$\begin{aligned} \|\xi\|_1 &\to \min, \\ A\xi &= b \end{aligned} \qquad \text{и} \qquad \begin{aligned} \|A^T\eta\|_\infty &\to \min, \\ \langle b, \eta\rangle &= 1. \end{aligned}$$

Соответствующие результаты представлены в [4].

ЗАДАЧА 5. *Пусть $u(t)$ — непрерывная на отрезке $[a,b]$ функция и*

$$C_0^n[a,b] = \big\{h \in C^n[a,b] \mid h^{(k)}(a) = h^{(k)}(b) = 0 \ \ \forall k \in 1:n-1\big\}.$$

Докажите, что при выполнении условия

$$\int_a^b u(t)h^{(n)}(t)dt = 0 \quad \forall h \in C_0^n[a,b] \tag{22}$$

функция $u(t)$ является алгебраическим полиномом степени не выше $n-1$.

Р е ш е н и е (Д. Петров, 1999 г.). Положим

$$u_0(t) = u(t); \quad u_k(t) = \int_a^t u_{k-1}(\tau)d\tau, \quad k = 1,\dots,n.$$

Введём интерполяционный полином Эрмита $p_{2n-1}(t)$, исходя из условий

$$p_{2n-1}^{(k)}(a) = u_n^{(k)}(a); \quad p_{2n-1}^{(k)}(b) = u_n^{(k)}(b), \quad k \in 0:n-1.$$

Рассмотрим функцию $\hat{h}(t) = u_n(t) - p_{2n-1}(t)$. Очевидно, что $\hat{h} \in C_0^n[a,b]$. При этом

$$\hat{h}^{(n)}(t) = u(t) - p_{2n-1}^{(n)}(t). \tag{23}$$

Полином $g_{n-1}(k) = p_{2n-1}^{(n)}(t)$ имеет степень не выше $n-1$ и

$$\int_a^b g_{n-1}(t)\hat{h}^{(n)}(t)dt = 0. \tag{24}$$

Формула (24) проверяется интегрированием по частям:

$$\int_a^b g_{n-1}\hat{h}^{(n)}dt = -\int_a^b g'_{n-1}\hat{h}^{(n-1)}dt = \ldots = (-1)^n\int_a^b g_{n-1}^{(n)}\hat{h}\ dt = 0.$$

На основании (22)–(24) получаем

$$\int_a^b \big[u(t) - g_{n-1}(t)\big]^2 dt = \int_a^b \big[u(t) - g_{n-1}(t)\big]\hat{h}^{(n)}(t)dt =$$
$$= \int_a^b u(t)\hat{h}^{(n)}(t)dt - \int_a^b g_{n-1}(t)\hat{h}^{(n)}(t)dt = 0.$$

Отсюда следует, что $u(t) \equiv g_{n-1}(t)$ на $[a,b]$, то есть функция $u(t)$ на отрезке $[a,b]$ совпадает с алгебраическим полиномом $g_{n-1}(t)$ степени не выше $n-1$.

К о м м е н т а р и й. Это замечательное решение побудило меня на лекциях при доказательстве основной леммы вариационного исчисления использовать линейную интерполяцию.

ЗАДАЧА 6. *Рассмотрим вариационную задачу*

$$J(x) := \int_{-1}^{1} x^2(1-x')^2 dt \to \inf,$$
$$x(-1) = 0, \quad x(1) = 1, \quad x \in C^1[-1,1].$$

Найдите инфимум функционала $J(x)$ на множестве планов и докажите, что этот инфимум не достигается.

Решение (А. Герасимов, 1999 г.; И. Анисимова, 2004 г.). Очевидно, что $J(x) \geqslant 0$ на всех планах x. На функции

$$x_*(t) = \begin{cases} 0 & \text{при } t \in [-1,0], \\ t & \text{при } t \in [0,1] \end{cases}$$

выполняется равенство $J(x_*) = 0$. Но x_* не принадлежит пространству $C^1[-1,1]$. Эта функция является решением исходной задачи в более широком пространстве кусочно гладких функций.

Покажем, что на $C^1[-1,1]$ задача не имеет решения. Введём последовательность функций

$$x_n(t) = \begin{cases} 0 & \text{при } t \in [-1, -\frac{1}{n}], \\ \frac{1}{4n}(nt+1)^2 & \text{при } t \in [-\frac{1}{n}, \frac{1}{n}], \\ t & \text{при } t \in [\frac{1}{n}, 1]. \end{cases}$$

Полином $p_2(t) = \frac{1}{4n}(nt+1)^2$ удовлетворяет условиям

$$p_2(-\tfrac{1}{n}) = 0, \quad p_2'(-\tfrac{1}{n}) = 0,$$
$$p_2(\tfrac{1}{n}) = \tfrac{1}{n}, \quad p_2'(\tfrac{1}{n}) = 1.$$

Значит, функции x_n принадлежат $C^1[-1,1]$ при всех натуральных n. Кроме того, все $x_n(t)$ удовлетворяют граничным условиям. Вычислим значения функционала J на планах x_n:

$$J(x_n) = \int_{-1}^{1} x_n^2(t)\big[1 - x_n'(t)\big]^2 dt =$$
$$= \frac{1}{64n^2} \int_{-1/n}^{1/n} (nt+1)^4 (nt-1)^2 dt = \frac{1}{64n^3} \int_{-1}^{1} (u+1)^4 (u-1)^2 du.$$

Очевидно, что $J(x_n) \to 0$ при $n \to \infty$. Приходим к следующему выводу: инфимум $J(x)$ на функциях $x \in C^1[-1,1]$, удовлетворяющих граничным условиям $x(-1) = 0$, $x(1) = 1$, равен нулю. Покажем, что этот инфимум не достигается.

Предположим, что $J(x_0) = 0$ на некотором плане x_0. Тогда

$$x_0(t)\big[1 - x_0'(t)\big] \equiv 0 \quad \text{на } [-1, 1]. \tag{25}$$

По теореме о среднем

$$1 = x_0(1) - x_0(-1) = 2x_0'(\xi),$$

где $\xi \in (-1, 1)$. Получаем $x_0'(\xi) = \frac{1}{2}$. В некоторой окрестности точки ξ будут выполняться неравенства $\frac{1}{4} < x_0'(t) < \frac{3}{4}$. Согласно (25) в той же окрестности $x_0(t) \equiv 0$. Отсюда, в частности, следует, что $x_0'(\xi) = 0$. Это противоречит равенству $x_0'(\xi) = \frac{1}{2}$.

Противоречие убеждает нас в том, что инфимум функционала $J(x)$ на множестве планов не достигается.

К о м м е н т а р и й. Если вариационная задача не имеет решения, то можно попытаться расширить основное пространство, на котором определён функционал, так, чтобы обеспечить существование решения.

ЛИТЕРАТУРА

1. Малозёмов В. Н. *Линейная алгебра без определителей. Квадратичная функция.* СПб.: Изд-во СПбГУ, 1997. 80 с.

2. Малозёмов В. Н. *Неравенства и экстремальные задачи* // Семинар «CNSA & NDO». Избранные доклады. 4 сентября 2014 г. (`http://apmath.spbu.ru/cnsa/reps14.shtml#0904`) [Данная книга, с. 413]

3. Малозёмов В. Н., Монако М. Ф., Петров А. В. *Формулы Фробениуса, Шермана-Моррисона и близкие вопросы.* // Журн. вычисл. мат. и матем. физ. 2002. Т. 42. № 10. С. 1459–1465.

4. Малозёмов В. Н. *Конечномерная проблема моментов* // Семинар «DHA & CAGD». Избранные доклады. 11 сентября 2010 г. (`http://dha.spb.ru/reps10.shtml#0911`) [Данная книга, с. 59]

ЦИКЛИЧЕСКИЕ ФУНКЦИИ И ЭКСТРЕМАЛЬНЫЕ ЗАДАЧИ*

В. Н. Малозёмов

1°. Начнём с простого примера. Возьмём циклическую функцию

$$G_n(x) = \frac{x_1^2}{x_1 + x_2} + \frac{x_2^2}{x_2 + x_3} + \cdots + \frac{x_{n-1}^2}{x_{n-1} + x_n} + \frac{x_n^2}{x_n + x_1}$$

и рассмотрим экстремальную задачу

$$\begin{gathered} \textit{минимизировать } G_n(x) \textit{ при ограничениях :} \\ \textit{все компоненты вектора } x = (x_1, \ldots, x_n) \textit{ положительны и} \\ \sum_{k=1}^{n} x_k = a, \quad a > 0. \end{gathered} \tag{1}$$

Покажем, что задача (1) при $n \geqslant 2$ имеет единственное решение

$$x^* = (\tfrac{a}{n}, \tfrac{a}{n}, \ldots, \tfrac{a}{n}).$$

Нам потребуется вспомогательное утверждение.

ЛЕММА 1. *Для вещественных чисел $c_1, \ldots, c_n$ и положительных $b_1, \ldots, b_n$ справедливо неравенство*

$$\frac{c_1^2}{b_1} + \cdots + \frac{c_n^2}{b_n} \geqslant \frac{(c_1 + \cdots + c_n)^2}{b_1 + \cdots + b_n}. \tag{2}$$

Неравенство выполняется как равенство только тогда, когда $c_k = \lambda b_k$, $k \in 1 : n$, при некотором вещественном λ.

Доказательство. По неравенству Коши-Буняковского

$$\begin{aligned} (c_1 + \cdots + c_n)^2 &= \Big(\sqrt{b_1}\frac{c_1}{\sqrt{b_1}} + \cdots + \sqrt{b_n}\frac{c_n}{\sqrt{b_n}}\Big)^2 \leqslant \\ &\leqslant (b_1 + \cdots + b_n)\Big(\frac{c_1^2}{b_1} + \cdots + \frac{c_n^2}{b_n}\Big). \end{aligned}$$

Отсюда очевидным образом следует (2).

Неравенство Коши-Буняковского выполняется как равенство только тогда, когда $\frac{c_k}{\sqrt{b_k}} = \lambda\sqrt{b_k}$, то есть когда $c_k = \lambda b_k$ при всех $k \in 1 : n$. □

*Семинар «CNSA & NDO». Избранные доклады. 27 августа 2015 г.

Теперь легко получить решение задачи (1). Согласно лемме для любого плана $x = (x_1, \dots, x_n)$ имеем

$$G_n(x) \geqslant \frac{(x_1 + \dots + x_n)^2}{2(x_1 + \dots + x_n)} = \frac{a}{2}.$$

Неравенство выполняется как равенство (а в этом случае достигается минимальное значение целевой функции) только тогда, когда $x_k = \lambda(x_k + x_{k+1})$, $k \in 1 : n$, где $x_{n+1} = x_1$. Сложив эти равенства, получим $a = 2\lambda a$, так что $\lambda = \frac{1}{2}$ и $x_1 = x_2 = \dots = x_n = \frac{a}{n}$.

2°. Возьмём более сложную циклическую функцию

$$F_n(x) = \frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_4} + \dots + \frac{x_{n-1}}{x_n + x_1} + \frac{x_n}{x_1 + x_2}.$$

У неё каждое слагаемое зависит от трёх соседних переменных.

Рассмотрим экстремальную задачу: *минимизировать* $F_n(x)$ *при тех же ограничениях, что и в задаче* (1). Множество планов обозначим через Ω_n.

Начнём с того, что укажем решение данной задачи при $n \in 3 : 6$.

3°. Пусть $n = 3$. Нам потребуется неравенство

$$\frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_1} + \frac{x_3}{x_1 + x_2} \geqslant \frac{3}{2}, \tag{3}$$

справедливое при всех положительных x_1, x_2, x_3. Проверим его.

У каждой дроби в числителе добавим и вычтем знаменатель. Получим

$$\frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_1} + \frac{x_3}{x_1 + x_2} = (x_1 + x_2 + x_3)\Big(\frac{1}{x_2 + x_3} + \frac{1}{x_3 + x_1} + \frac{1}{x_1 + x_2}\Big) - 3. \tag{4}$$

Теперь воспользуемся неравенством между средним гармоническим и средним арифметическим, которое запишем в виде

$$\frac{1}{b_1} + \frac{1}{b_2} + \frac{1}{b_3} \geqslant \frac{9}{b_1 + b_2 + b_3}.$$

Подставив это в (4), придём к (3).

Неравенство (3) выполняется как равенство только тогда, когда $x_2 + x_3 = x_3 + x_1 = x_1 + x_2$, то есть когда $x_1 = x_2 = x_3$. Отсюда следует, что минимум функции $F_3(x)$ на Ω_3 достигается в единственной точке $x^* = (\frac{a}{3}, \frac{a}{3}, \frac{a}{3})$ и равен $\frac{3}{2}$.

4°. Пусть $n = 4$. Докажем, что при всех положительных x_1, x_2, x_3, x_4 справедливо неравенство

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\frac{x_3}{x_4+x_1}+\frac{x_4}{x_1+x_2}\geqslant 2. \tag{5}$$

Оно выполняется как равенство только тогда, когда $x_1 = x_3$ и $x_2 = x_4$.

Действительно, по неравенству между средним геометрическим и средним арифметическим имеем

$$\begin{aligned}\frac{x_1}{x_2+x_3}+\frac{x_3}{x_4+x_1}&=\frac{x_1(x_4+x_1)+x_3(x_2+x_3)}{(x_2+x_3)(x_4+x_1)}\geqslant\\ &\geqslant\frac{4(x_1^2+x_3^2+x_1x_4+x_2x_3)}{(x_1+x_2+x_3+x_4)^2},\\ \frac{x_2}{x_3+x_4}+\frac{x_4}{x_1+x_2}&=\frac{x_2(x_1+x_2)+x_4(x_3+x_4)}{(x_3+x_4)(x_1+x_2)}\geqslant\\ &\geqslant\frac{4(x_2^2+x_4^2+x_1x_2+x_3x_4)}{(x_1+x_2+x_3+x_4)^2}.\end{aligned}$$

Сложив эти неравенства, получим

$$F_4(x)\geqslant\frac{2\big[(x_1+x_2)^2+(x_2+x_3)^2+(x_3+x_4)^2+(x_4+x_1)^2\big]}{(x_1+x_2+x_3+x_4)^2}.$$

Теперь оценим числитель с помощью неравенства между средним квадратичным и средним арифметическим, которое запишем в виде

$$b_1^2+b_2^2+b_3^2+b_4^2\geqslant\tfrac14(b_1+b_2+b_3+b_4)^2.$$

Придём к (5).

Неравенство (5) выполняется как равенство только тогда, когда все промежуточные неравенства выполняются как равенства, то есть когда

$$\begin{gathered}x_2+x_3=x_4+x_1,\quad x_3+x_4=x_1+x_2,\\ x_1+x_2=x_2+x_3=x_3+x_4=x_4+x_1.\end{gathered}$$

Это возможно лишь в случае $x_1 = x_3$, $x_2 = x_4$, то есть при $x = (x_1, x_2, x_1, x_2)$. Точки такого вида, принадлежащие Ω_4, должны удовлетворять условиям

$$x_1+x_2=\tfrac{a}{2},\quad x_1>0,\ x_2>0. \tag{6}$$

Таким образом, минимум функции $F_4(x)$ на Ω_4 равен 2 и достигается на векторах вида $x = (x_1, x_2, x_1, x_2)$, удовлетворяющих условиям (6).

Отметим, что функция $F_4(x)$ определена и в предельных точках $(0, \frac{a}{2}, 0, \frac{a}{2})$ и $(\frac{a}{2}, 0, \frac{a}{2}, 0)$ и принимает на них значение 2.

5°. Анализ случаев $n = 5$ и $n = 6$ основан на другом подходе. Будем минимизировать функцию $F_n(x)$ по всем векторам $x = (x_1, \dots, x_n)$ с положительными компонентами, удовлетворяющими условию

$$\sum_{k=1}^{n} x_k = n. \tag{7}$$

Так как $F_n(\lambda x) = F_n(x)$ при положительных λ, то решение $\hat{x}$ такой задачи связано с решением x^* задачи минимизации $F_n(x)$ на Ω_n соотношением $x^* = \frac{a}{n}\hat{x}$.

Положим $x_k = 1 + 2h_k$, $k \in 1 : n$, где $h_k > -\frac{1}{2}$. Согласно (7)

$$\sum_{k=1}^{n} h_k = 0. \tag{8}$$

Имеем

$$F_n(x) = \sum_{k=1}^{n} \frac{1 + 2h_k}{2 + 2(h_{k+1} + h_{k+2})}.$$

Здесь $h_{n+1} = h_1$, $h_{n+2} = h_2$.

Воспользуемся неравенством

$$\frac{1}{1+u} \geqslant 1 - u \quad \text{при} \quad u > -1,$$

которое выполняется как равенство только при $u = 0$. Согласно (8) получим

$$F_n(x) \geqslant \frac{1}{2} \sum_{k=1}^{n} (1 + 2h_k)\big(1 - (h_{k+1} + h_{k+2})\big) =$$
$$= \frac{n}{2} - \sum_{k=1}^{n} h_k(h_{k+1} + h_{k+2}).$$

Обозначим

$$Q_n(h) = \sum_{k=1}^{n} h_k(h_{k+1} + h_{k+2}).$$

Тогда

$$F_n(x) \geqslant \frac{n}{2} - Q_n(h). \tag{9}$$

При $n = 5$ имеем $h_6 = h_1$, $h_7 = h_2$ и

$$Q_5(h) = h_1h_2 + h_1h_3 + h_4h_1 + h_5h_1 +$$

$$
\begin{aligned}
&+ h_2h_3 + h_2h_4 + h_5h_2 + \\
&+ h_3h_4 + h_3h_5 + \\
&+ h_4h_5 = \\
= \sum_{1\leqslant i<j\leqslant 5} h_ih_j &= \frac{1}{2}\Big(\sum_{k=1}^{5} h_k\Big)^2 - \frac{1}{2}\sum_{k=1}^{5} h_k^2 = -\frac{1}{2}\sum_{k=1}^{5} h_k^2.
\end{aligned}
$$

Значит,

$$F_5(x) \geqslant \frac{5}{2} + \frac{1}{2}\sum_{k=1}^{5} h_k^2 \geqslant \frac{5}{2}.$$

Равенство $F_5(x) = \frac{5}{2}$ выполняется только тогда, когда все h_k равны нулю, то есть на векторе $\hat{x} = (1, 1, 1, 1, 1)$.

Приходим к следующему выводу: единственным решением задачи минимизации функции $F_5(x)$ на Ω_5 является вектор $x^* = \frac{a}{5}\hat{x}$.

При $n = 6$ имеем $h_7 = h_1$, $h_8 = h_2$ и

$$
\begin{aligned}
Q_6(h) = h_1h_2 + h_1h_3 + &\quad + h_5h_1 + h_6h_1 + \\
+ h_2h_3 + h_2h_4 + &\quad + h_6h_2 + \\
+ h_3h_4 + h_3h_5 + &\\
+ h_4h_5 + h_4h_6 + &\\
+ h_5h_6 = &\\
= \sum_{1\leqslant i<j\leqslant 6} h_ih_j - h_1h_4 - h_2h_5 - h_3h_6 = &\\
= \frac{1}{2}\Big(\sum_{k=1}^{6} h_k\Big)^2 - \frac{1}{2}\big[(h_1+h_4)^2 + (h_2+h_5)^2 + (h_3+h_6)^2\big]&.
\end{aligned}
$$

Согласно (8) и (9)

$$F_6(x) \geqslant 3 + \frac{1}{2}\big[(h_1+h_4)^2 + (h_2+h_5)^2 + (h_3+h_6)^2\big] \geqslant 3.$$

Равенство $F_6(x) = 3$ выполняется только тогда, когда

$$
\begin{gathered}
h_2 + h_3 = 0,\ h_3 + h_4 = 0,\ h_4 + h_5 = 0,\ h_5 + h_6 = 0,\ h_6 + h_1 = 0,\ h_1 + h_2 = 0; \\
h_1 + h_4 = 0,\ h_2 + h_5 = 0,\ h_3 + h_6 = 0,
\end{gathered}
$$

то есть когда $h_2 = h_4 = h_6$, $h_1 = h_3 = h_5$, $h_1 + h_2 = 0$. Это соответствует представлению

$$\hat{x}_1 = 1 + h,\ \hat{x}_2 = 1 - h,\ \hat{x}_3 = 1 + h,\ \hat{x}_4 = 1 - h,\ \hat{x}_5 = 1 + h,\ \hat{x}_6 = 1 - h, \quad (10)$$

где $|h| < 1$.

Приходим к следующему выводу: минимальное значение функции $F_6(x)$ на Ω_6 равно 3 и достигается на векторах $x^* = \frac{a}{6}\hat{x}$, где компоненты вектора $\hat{x}$ имеют вид (10).

Отметим, что функция $F_6(x)$ определена и в предельных точках $\hat{x}$, соответствующих $h = 1$ и $h = -1$, и принимает на них значение 3.

6°. Решение задачи о минимизации функции $F_n(x)$ на множестве Ω_n при $n \in 3 : 6$ было получено на основании того факта, что при указанных n и положительных $x_1, \ldots, x_n$ выполняется неравенство

$$F_n(x) \geqslant \frac{n}{2}.$$

В 1954 г. Г. Шапиро доказал это неравенство при $n = 3$ и $n = 4$ и выдвинул гипотезу, что неравенство справедливо и при всех $n > 4$ [1]. Гипотеза Шапиро вызвала широкий интерес. Коллективными усилиями было установлено, что неравенство Шапиро выполняется при $n \leqslant 13$ и нечётных n от 15 до 23. При остальных n неравенство нарушается (см., например [2]).

Возникает вопрос: в случае, когда неравенство Шапиро нарушается, насколько минимум $F_n(x)$ на Ω_n отличается от $\frac{n}{2}$. Ответ на этот вопрос содержится в замечательной теореме В. Г. Дринфельда [3].

ТЕОРЕМА. *При $n \geqslant 7$ для всех векторов x с положительными компонентами справедлива оценка*

$$F_n(x) > c\ \frac{n}{2},$$

где $c = 0.989$.

Для доказательства нам потребуется вспомогательное предложение.

ЛЕММА 2. *Пусть $u_1 \geqslant u_2 \geqslant \ldots \geqslant u_n$ — упорядоченная по невозрастанию последовательность $x_1, x_2, \ldots, x_n$ и $v_1 \leqslant v_2 \leqslant \ldots \leqslant v_n$ — упорядоченная по неубыванию последовательность $y_1, y_2, \ldots, y_n$. Тогда*

$$\sum_{k=1}^{n} x_k y_k \geqslant \sum_{k=1}^{n} u_k v_k. \tag{11}$$

Доказательство. За счёт перестановки слагаемых неравенство (11) можно переписать в виде

$$\sum_{k=1}^{n} u_k y_k^0 \geqslant \sum_{k=1}^{n} u_k v_k,$$

где $y_1^0, \ldots, y_n^0$ — некоторая перестановка чисел $y_1, \ldots, y_n$. Если $v_1 = y_1^0$, то

$$\sum_{k=1}^{n} u_k y_k^0 = u_1 v_1 + \sum_{k=2}^{n} u_k y_k^0. \tag{12}$$

Допустим, что $v_1 = y^0_{k_0}$ при $k_0 > 1$. Так как

$$(u_1 - u_{k_0})(y^0_1 - v_1) \geqslant 0$$

или

$$u_1 y^0_1 + u_{k_0} y^0_{k_0} \geqslant u_1 v_1 + u_{k_0} y^0_1,$$

то

$$\sum_{k=1}^{n} u_k y^0_k \geqslant u_1 v_1 + \sum_{k=2}^{n} u_k y^1_k, \tag{13}$$

где $y^1_{k_0} = y^0_1$ и $y^1_k = y^0_k$ при $k \neq k_0$. При этом

$$\min_{k \in 2:n} y^1_k = v_2.$$

Очевидно, что и в (12)

$$\min_{k \in 2:n} y^0_k = v_2.$$

С помощью таких же соображений в правых частях неравенств (12) и (13) можно выделить слагаемое $u_2 v_2$ и т. д. В результате придём к неравенству (11). □

Доказательство теоремы. Обозначим $\xi_k = x_{k+1}/x_k$, $k \in 1:n$. Тогда

$$F_n(x) = \sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} = \sum_{k=1}^{n} \frac{1}{\xi_k(1 + \xi_{k+1})}.$$

Здесь $\xi_{n+1} = \xi_1$, так как по определению $x_{n+1} = x_1$, $x_{n+2} = x_2$. Отметим также, что $\xi_1 \xi_2 \cdots \xi_n = 1$.

Упорядочим числа $\xi_1, \xi_2, \ldots, \xi_n$ по неубыванию. Получим последовательность $0 < y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$. По лемме 2

$$F_n(x) \geqslant \sum_{k=1}^{n} \frac{1}{y_k(1 + y_{n-k+1})}.$$

Положим

$$r_k = \frac{1}{y_k(1 + y_{n-k+1})} + \frac{1}{y_{n-k+1}(1 + y_k)}.$$

Очевидно, что

$$F_n(x) \geqslant \frac{1}{2} \sum_{k=1}^{n} r_k. \tag{14}$$

Обозначим $\eta_k = y_k y_{n-k+1}$. Имеем $\eta_1\eta_2\dots\eta_n = 1$ и

$$r_k = \frac{1}{\eta_k}\Big(1 + \frac{\eta_k - 1}{(1+y_k)(1+y_{n-k+1})}\Big).$$

Так как

$$(1+y_k)(1+y_{n-k+1}) = 1 + \eta_k + 2\frac{y_k + y_{n-k+1}}{2} \geqslant (1+\sqrt{\eta_k})^2,$$

то

$$r_k \geqslant \begin{cases} 1/\eta_k \quad \text{при} \quad \eta_k \geqslant 1, \\ \frac{1}{\eta_k}\Big(1 + \frac{\sqrt{\eta_k}-1}{\sqrt{\eta_k}+1}\Big) = \frac{2}{\eta_k + \sqrt{\eta_k}} \quad \text{при} \quad \eta_k < 1. \end{cases}$$

Пусть $z_k = \ln \eta_k$. Из последнего неравенства следует, что

$$r_k \geqslant \min\big\{e^{-z_k},\ 2(e^{z_k} + e^{z_k/2})^{-1}\big\}. \tag{15}$$

При этом

$$z_1 + z_2 + \dots + z_n = \ln(\eta_1\eta_2\dots\eta_n) = 0. \tag{16}$$

Введём функции $f(z) = e^{-z}$, $g(z) = 2(e^z + e^{z/2})^{-1}$,

$$\psi(z) = \min\big\{f(z), g(z)\big\}.$$

На основании (14) и (15) получаем

$$\frac{2}{n}\ F_n(x) \geqslant \frac{1}{n}\sum_{k=1}^{n} \psi(z_k). \tag{17}$$

Теперь мы хотим воспользоваться неравенством Йенсена. Для этого построим выпуклую огибающую функции $\psi(z)$ (см. рис.).

Функции $f(z)$ и $g(z)$ обладают следующими свойствами: они строго выпуклые, принимают одинаковое значение при $z = 0$, $f(0) = g(0) = 1$, и удовлетворяют неравенствам

$$f(z) > g(z) \text{ при } z < 0 \quad \text{и} \quad f(z) < g(z) \text{ при } z > 0.$$

Ясно, что построение выпуклой огибающей функции $\psi(z)$ (обозначим её $\tilde{\psi}(z)$) сводится к нахождению общей касательной функций $f(z)$ и $g(z)$ между точками касания.

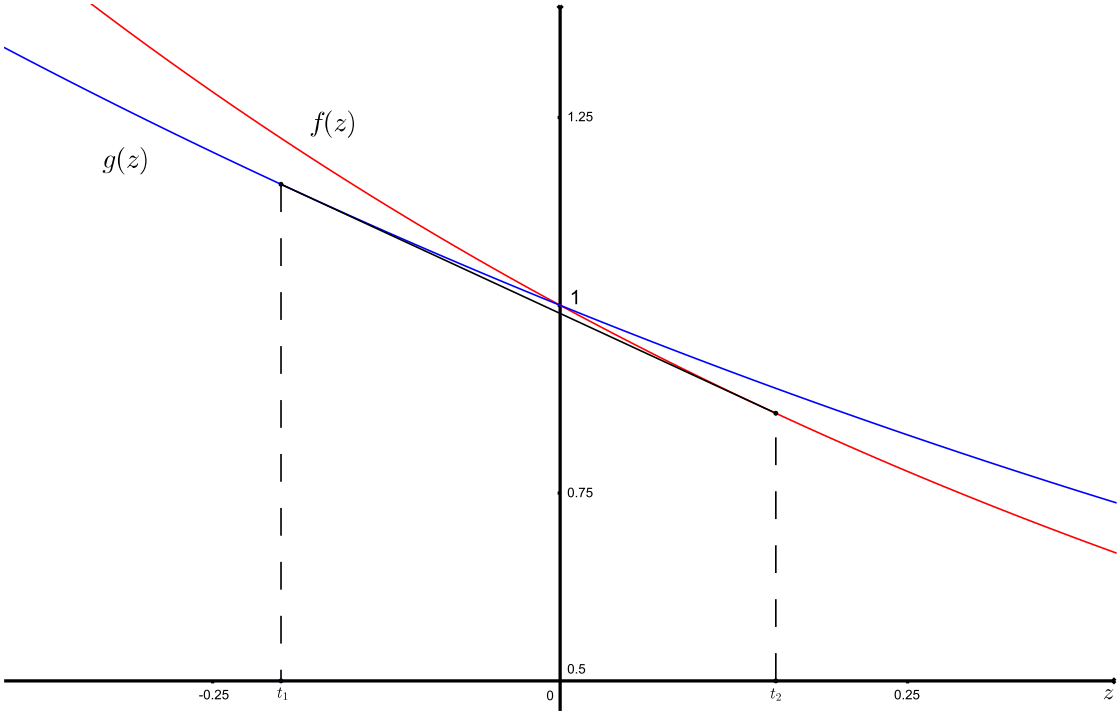


Рис. Выпуклая огибающая функции $\psi(z)$

Абсциссы точек касания обозначим $z = t_1$ и $z = t_2$, $t_1 < t_2$. Запишем систему уравнений относительно t_1 и t_2:

$$\begin{cases} g'(t_1) = f'(t_2), \\ f'(t_2)(t_1 - t_2) + f(t_2) = g(t_1). \end{cases}$$

Из первого уравнения следует, что

$$\frac{2e^{t_1} + e^{t_1/2}}{(e^{t_1} + e^{t_1/2})^2} = e^{-t_2},$$

так что

$$t_2 = -\ln\Big(\frac{2e^{t_1} + e^{t_1/2}}{(e^{t_1} + e^{t_1/2})^2}\Big).$$

Второе уравнение преобразуется к виду

$$\frac{2e^{t_1} + e^{t_1/2}}{e^{t_1} + e^{t_1/2}} \left(t_1 + \ln\Big(\frac{2e^{t_1} + e^{t_1/2}}{(e^{t_1} + e^{t_1/2})^2}\Big) - 1\right) + 2 = 0.$$

Решение последнего уравнения находим численно:

$$t_1 = -0.200811\ldots$$

При этом $t_2 = 0.155195\ldots$ Отметим, что

$$\tilde{\psi}(0) = -f'(t_2)t_2 + f(t_2) = 0.989134\ldots$$

Вернёмся к формуле (17). Принимая во внимание неравенство $\psi(z) \geqslant \tilde{\psi}(z)$, выпуклость $\tilde{\psi}(z)$, неравенство Йенсена и равенство (16), получаем

$$\frac{2}{n}\, F_n(x) \geqslant \frac{1}{n}\sum_{k=1}^{n} \tilde{\psi}(z_k) \geqslant \tilde{\psi}\big(\frac{1}{n}\sum_{k=1}^{n} z_k\big) = \\ = \tilde{\psi}(0) > 0.989,$$

что равносильно утверждению теоремы. □

7°. Приношу благодарность Г. Ш. Тамасяну и Е. К. Чернэуцану за помощь при подготовке данного доклада.

ЛИТЕРАТУРА

1. Shapiro H. S. *Problem 4603* // Amer. Math. Monthly. 1954. Vol. 61. P. 571.

2. Храбров А. *Неравенство Шапиро.* (`http://olympiads.mccme.ru/lktg/2010/5/5-1ru.pdf`)

3. Дринфельд В. Г. *Об одном циклическом неравенстве* // Матем. заметки. 1971. Том. 9. Вып. 2. С. 113–119.

МИНИМИЗАЦИЯ ЦИКЛИЧЕСКОЙ ФУНКЦИИ*

А. В. Плоткин

1°. Неравенство Шапиро. Рассмотрим циклическую функцию вида

$$F_n(x) = \frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_4} + \ldots + \frac{x_{n-1}}{x_n + x_1} + \frac{x_n}{x_1 + x_2},$$

где $x_i \geqslant 0$, $i \in 1 : n$, и все знаменатели отличны от нуля.

В 1954 г. Г. Шапиро выдвинул гипотезу о том, что при $n \geqslant 3$ справедливо следующее неравенство:

$$F_n(x) \geqslant \frac{n}{2}.$$

На тот момент у него имелось доказательство только для случаев $n = 3$ и $n = 4$.

Отметим важное свойство этого неравенства:

ЛЕММА. *Если неравенство Шапиро неверно при некотором $n = k$, то оно неверно и при $n = k + 2$.*

Доказательство. Пусть $F_n(x') < \frac{n}{2}$. Рассмотрим $x'' = (x'_1, \ldots, x'_n, x'_1, x'_2)$. Заметим, что $F_{n+2}(x'') = F_n(x') + 1$, следовательно, $F_{n+2}(x'') < \frac{n+2}{2}$. □

Гипотеза Шапиро вызвала широкий интерес математиков (см., например, [1, 2]), но только к 1989 г. коллективными усилиями удалось установить, что неравенство Шапиро верно дня четных $n \leqslant 12$ и нечетных $n \leqslant 23$. Для $n = 14$ и $n = 25$ неравенство нарушается, а значит, по доказанной лемме, нарушается и для всех четных $n > 14$ и нечетных $n > 25$.

2°. Постановка задачи и метод решения. Целью работы являлся поиск значений x, при которых нарушается неравенство Шапиро в случаях $n = 14$ и $n = 25$. Для этого рассмотрим экстремальную задачу следующего вида:

$$F_n(x) \to \min, \quad x_i \geqslant 0,\ i \in 1 : n.$$

Стоит отметить, что при $n \geqslant 3$ для всех векторов x с положительными компонентами справедлива оценка Дринфельда

$$F_n(x) > c\,\frac{n}{2},$$

где $c = 0.989$ [2].

*Семинар «CNSA & NDO». Избранные доклады. 17 декабря 2015 г.

Сделаем замену $x_i = y_i^2$ и перейдем к задаче безусловной оптимизации:

$$G_n(y) := \frac{y_1^2}{y_2^2 + y_3^2} + \frac{y_2^2}{y_3^2 + y_4^2} + \ldots + \frac{y_{n-1}^2}{y_n^2 + y_1^2} + \frac{y_n^2}{y_1^2 + y_2^2} \to \min_{y \in \mathbb{R}^n}.$$

Решать поставленную задачу будем методом сопряженных градиентов без точного линейного поиска [3, 4, 5]. Опишем вычислительную схему данного метода.

Нулевой шаг. Берем начальное приближение y_0 и вычисляем градиент $g_0 = G'(y_0)$. Если $g_0 = \mathbb{O}$, то возвращаем y_0 в качестве ответа. Вычисления прекращаются. Иначе полагаем $s_1 = -g_0$.

k-й шаг. Пусть уже имеются y_{k-1}, $g_{k-1} \neq \mathbb{O}$ и s_k. Если s_k не является направлением убывания $G(y)$, то возвращаемся на нулевой шаг с начальным приближением y_{k-1}. Иначе находим t_k из условия Армихо (см. ниже) и вычисляем

$$y_k = y_{k-1} + t_k s_k,$$
$$g_k = G'(y_k).$$

Если $g_k = \mathbb{O}$, то возвращаем y_k в качестве ответа. Вычисления прекращаются. В противном случае, если $k = n$, то возвращаемся на нулевой шаг с начальным приближением y_k. Если же $k < n$, то находим

$$b_k = \frac{\langle g_k, g_k - g_{k-1} \rangle}{\langle s_k, g_k - g_{k-1} \rangle},$$
$$s_{k+1} = -g_k + b_k s_k.$$

Условие Армихо. Опишем, как по правилу Армихо находить t_k. Пусть заданы параметры $\lambda > 0$, $\delta, c \in (0, 1)$. Величина t_k выбирается итеративно. Изначально $t_k := \lambda$. На каждой итерации проверяется выполнение условия

$$G(y_k + t_k s_k) < G(y_k) + c\, t_k \langle G'(y_k), s_k \rangle.$$

В случае его выполнения вычисления прекращаются. Иначе полагаем $t_k := \delta t_k$ и переходим к следующей итерации. Процедура выбора обязательно завершится, так как s_k — направление убывания $G(y)$. При решении использовались значения $\lambda = 1$, $\delta = 0.5$, $c = 0.1$.

3°. Начальное приближение. Запуск алгоритма из случайных начальных приближений редко приводил к тому, что в получившихся точках значение функции было меньше $\frac{n}{2}$. Анализ успешных случаев при $n = 14$ позволил сделать предположение о структуре решения. На этом основании начальное приближение выбиралось следующим:

$$y_0 = [\alpha, \beta_1, \ldots, \alpha, \beta_7],$$

где α — случайная величина, равномерно распределенная на отрезке $[0.7, 0.9]$, а β_i — независимые случайные величины, равномерно распределенные на отрезке $[0, 0.2]$. Алгоритм был запущен 10000 раз из точек такого вида и каждый раз получалась точка, значение функции в которой было меньше $\frac{n}{2}$.

При $n = 25$ успеха удалось добиться, когда в качестве начального приближения брались точки вида

$$y_0 = [\alpha, \beta_1, \ldots, \alpha, \beta_{12}, \alpha],$$

где α и β_i — такие же случайные величины, что и при $n = 14$. Алгоритм был снова 10000 раз запущен из точек указанного вида и вновь в каждом случае привел к решению. Стоит отметить, что значение функции в точках, взятых в качестве начального приближения, для $n = 14$ очень близко к $\frac{n}{2}$, а для $n = 25$ уже достаточно сильно отличается от $\frac{n}{2}$.

4°. Результаты вычислений. Было получено большое количество значений x, при которых $F_n(x) < \frac{n}{2}$ для $n = 14$ и $n = 25$. Многие из них были приведены к целочисленному виду.

1) $n = 14$

- $x = [78, 7, 75, 8, 71, 6, 70, 3, 71, 1, 74, 1, 77, 3]$, $F_{14}(x) = 6.99996$;
- $x = [73, 8, 70, 6, 68, 3, 69, 1, 72, 1, 75, 3, 76, 7]$, $F_{14}(x) = 6.99994$;
- $x = [74, 6, 72, 3, 73, 1, 76, 1, 79, 3, 80, 7, 77, 8]$, $F_{14}(x) = 6.99991$.

2) $n = 25$

- $x = [39, 39, 38, 28, 36, 19, 35, 11, 35, 4, 39, 0, 46, 0, 55, 0, 65, 0, 77, 12, 79, 29, 68, 38, 53]$, $F_{25}(x) = 12.49896$;
- $x = [49, 48, 47, 35, 44, 23, 43, 13, 44, 5, 48, 0, 57, 0, 68, 0, 80, 0, 95, 15, 97, 36, 83, 47, 65]$, $F_{25}(x) = 12.49885$;
- $x = [43, 43, 41, 30, 39, 20, 38, 12, 39, 5, 43, 0, 51, 0, 60, 0, 71, 0, 84, 13, 86, 32, 74, 41, 57]$, $F_{25}(x) = 12.49881$.

5°. Заключение. Отметим, что метод сопряженных градиентов без точного линейного поиска оказался эффективным при решении данной задачи. В дальнейшем планируется испытание этого метода на различных целевых функциях с другими способами выбора t_k.

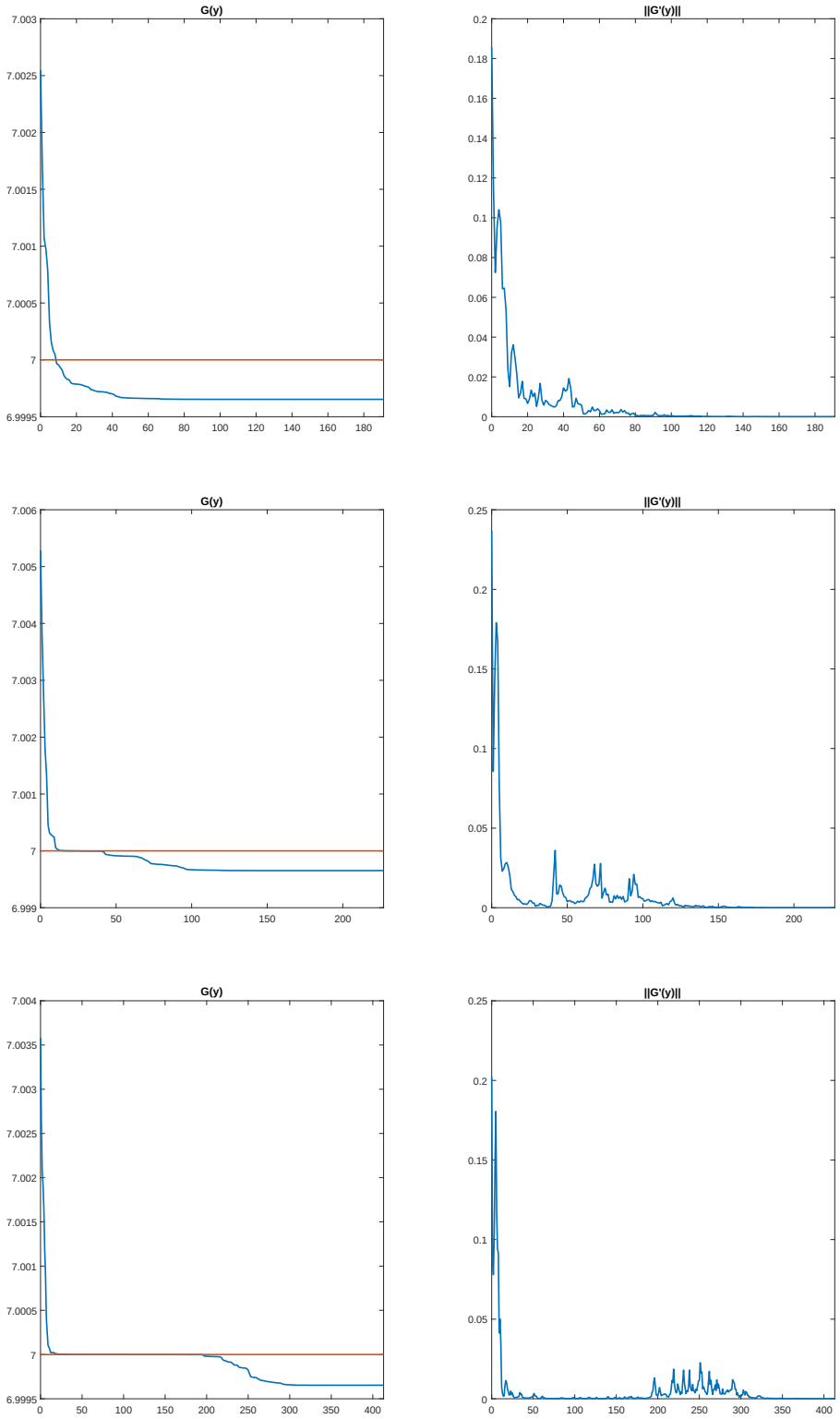


Рис. 1. Поведение функции $G(y)$ и её градиента по итерациям при $n = 14$ и различных начальных приближениях

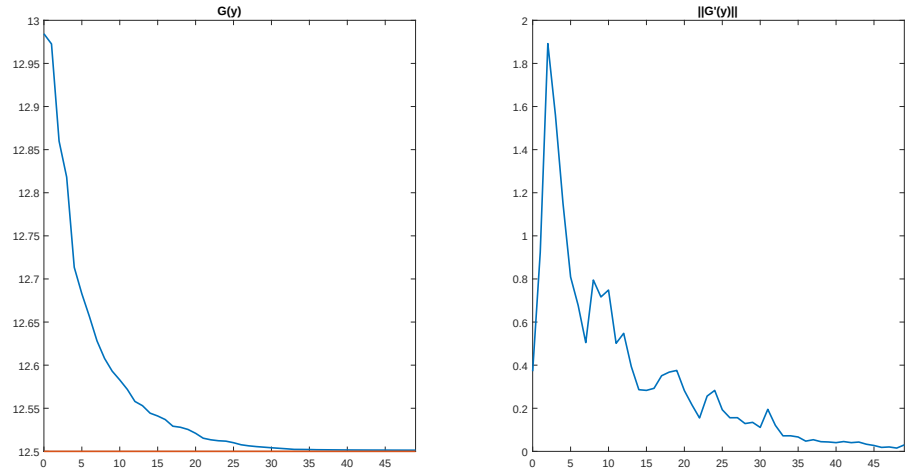


Рис. 2. Поведение функции $G(y)$ и её градиента по итерациям при $n = 25$ до 50-й итерации

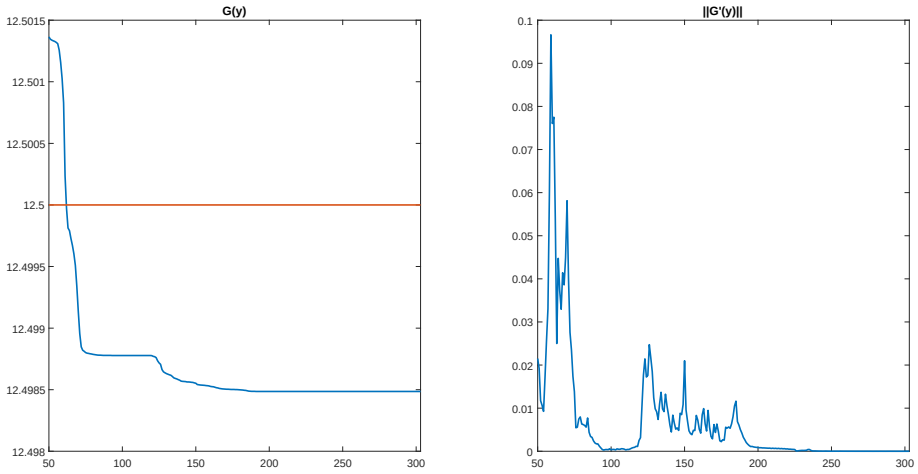


Рис. 3. Поведение функции $G(y)$ и её градиента по итерациям при $n = 25$ после 50-й итерации

ЛИТЕРАТУРА

1. А. Храбров. *Неравенство Шапиро* (`http://olympiads.mccme.ru/lktg/2010/5/5-1ru.pdf`)

2. В. Н. Малоземов. *Циклические функции и экстремальные задачи* // Семинар «CNSA & NDO». Избранные доклады. 27 августа 2015 г. (`http://apmath.spbu.ru/cnsa/reps15.shtml#0827`) [Данная книга, с. 435]

3. Малозёмов В. Н. *О методе сопряжённых градиентов* // Семинар «DHA & CAGD». Избранные доклады. 28 апреля 2012 г. (`http://dha.spb.ru/reps12.shtml#0428`) [Данная книга, с. 108]

4. Малозёмов В. Н. *Варианты метода сопряжённых градиентов* // Семинар «CNSA & NDO». Избранные доклады. 29 октября 2015 г. (`http://apmath.spbu.ru/cnsa/reps15.shtml#1029`) [Данная книга, с. 118]

5. Pytlak R. Conjugate Gradient Algorithms in Nonconvex Optimization. Berlin: Springer, 2009. P. 478.

НЕКОТОРЫЕ СВОЙСТВА ДИСКРЕТНОГО МАКСИМУМА*

В. Н. Малозёмов

1°. Пусть I — конечное индексное множество, α_i при $i \in I$ и c — произвольные вещественные числа. В докладе будут указаны нестандартные приложения элементарного равенства

$$\max_{i\in I}\{\alpha_i + c\} = \max_{i\in I}\{\alpha_i\} + c. \tag{1}$$

2°. Вначале проверим само равенство (1). Обозначим $A = \max\limits_{i\in I}\{\alpha_i\}$. При всех $i \in I$ имеем $\alpha_i + c \leqslant A + c$. Значит,

$$\max_{i\in I}\{\alpha_i + c\} \leqslant A + c. \tag{2}$$

Вместе с тем,

$$\alpha_i = (\alpha_i + c) - c \leqslant \max_{i\in I}\{\alpha_i + c\} - c,$$

так что

$$A \leqslant \max_{i\in I}\{\alpha_i + c\} - c$$

и

$$A + c \leqslant \max_{i\in I}\{\alpha_i + c\}. \tag{3}$$

Объединяя неравенства (2) и (3), приходим к равенству (1).

3°. В. Ф. Демьянов обратил внимание на то, что наряду с очевидным неравенством

$$\max_{i\in I}\{\alpha_i + \beta_i\} \leqslant \max_{i\in I}\{\alpha_i\} + \max_{i\in I}\{\beta_i\}$$

выполняется обратное неравенство

$$\max_{i\in I}\{\alpha_i + \beta_i\} \geqslant \max_{i\in I}\{\alpha_i\} + \max_{i\in R}\{\beta_i\}, \tag{4}$$

где $R = \{i \in I \mid \alpha_i = A\}$. Действительно, согласно (1),

$$\max_{i\in I}\{\alpha_i + \beta_i\} \geqslant \max_{i\in R}\{\alpha_i + \beta_i\} =$$

*Семинар «CNSA & NDO». Избранные доклады. 14 мая 2015 г.

$$= \max_{i\in R}\{A+\beta_i\} = A + \max_{i\in R}\{\beta_i\} = \max_{i\in I}\{\alpha_i\} + \max_{i\in R}\{\beta_i\}.$$

Неравенство (4) обобщается на бóльшее число слагаемых. Например

$$\max_{i\in I}\{\alpha_i+\beta_i+\gamma_i\} \geqslant \max_{i\in R}\{A+\beta_i+\gamma_i\} = A + \max_{i\in R}\{\beta_i+\gamma_i\}.$$

Обозначим $B = \max\limits_{i\in R}\{\beta_i\}$ и $R_1 = \{i \in R \mid \beta_i = B\}$. Согласно (1),

$$\max_{i\in R}\{\beta_i+\gamma_i\} \geqslant \max_{i\in R_1}\{B+\gamma_i\} = B + \max_{i\in R_1}\{\gamma_i\}.$$

Окончательно получаем

$$\max_{i\in I}\{\alpha_i+\beta_i+\gamma_i\} \geqslant \max_{i\in I}\{\alpha_i\} + \max_{i\in R}\{\beta_i\} + \max_{i\in R_1}\{\gamma_i\}.$$

С помощью указанных в этом пункте соображений в книге [1, с. 69–76] выводится разложение дискретной функции максимума по направлению.

4°. Обозначим $f_i = \alpha_i + \beta_i$, $i \in I$.

ЛЕММА 1. *Справедливо равенство*

$$\max_{i\in I}\{f_i\} = \max_{i\in I}\Big\{\alpha_i - \sum_{j\in I, j\neq i}\beta_j\Big\} + \sum_{i\in I}\beta_i. \tag{5}$$

Например, в случае

$$\alpha = (1,4,7,0), \quad \beta = (2,1,-3,1)$$

равенство (5) принимает вид

$$\max\{3,5,4,1\} = \max\{2,4,3,0\} + 1.$$

В книге [2, с. 116] доказательство этой леммы занимает полторы страницы. На самом деле, равенство (5) легко следует из (1). Действительно, запишем

$$\alpha_i + \beta_i = \alpha_i + \Big(\sum_{i\in I}\beta_i - \sum_{j\in I, j\neq i}\beta_j\Big) =$$
$$= \Big(\alpha_i - \sum_{j\in I, j\neq i}\beta_j\Big) + \sum_{i\in I}\beta_i.$$

Взяв максимум по $i \in I$ и воспользовавшись равенством (1) при $c = \sum\limits_{i\in I}\beta_i$, получим (5).

5°. Укажем на ещё одно приложение равенства (1).

ЛЕММА 2. *Справедливы равенства*

$$\max_{i\in I}\min_{j\in I}\{\alpha_i+\beta_j\}=\min_{j\in I}\max_{i\in I}\{\alpha_i+\beta_j\}= \\ =\max_{i\in I}\{\alpha_i\}+\min_{j\in I}\{\beta_j\}. \tag{6}$$

Доказательство. Отметим, что наряду с (1) выполняется равенство

$$\min_{j\in I}\{\beta_j+c\}=\min_{j\in I}\{\beta_j\}+c. \tag{7}$$

Имеем

$$\max_{i\in I}\min_{j\in I}\{\alpha_i+\beta_j\}=\max_{i\in I}\big\{\alpha_i+\min_{j\in I}\{\beta_j\}\big\}.$$

Воспользовавшись равенством (1) при $c=\min\limits_{j\in I}\{\beta_j\}$, получим

$$\max_{i\in I}\min_{j\in I}\{\alpha_i+\beta_j\}=\max_{i\in I}\{\alpha_i\}+\min_{j\in I}\{\beta_j\}.$$

Аналогично

$$\min_{j\in I}\max_{i\in I}\{\alpha_i+\beta_j\}=\min_{j\in I}\big\{\max_{i\in I}\{\alpha_i\}+\beta_j\big\}.$$

Воспользовавшись равенством (7) при $c=\max\limits_{i\in I}\{\alpha_i\}$, получим

$$\min_{j\in I}\max_{i\in I}\{\alpha_i+\beta_j\}=\max_{i\in I}\{\alpha_i\}+\min_{j\in I}\{\beta_j\}.$$

Лемма доказана. □

6°. Следующее свойство дискретного максимума не зависит от равенства (1).

Введём полиэдральную функцию

$$\varphi(x)=\max_{i\in I}\big\{\langle a_i,x\rangle+b_i\big\},\quad x\in\mathbb{R}^n.$$

Обозначим через F множество $(n+1)$-мерных векторов $f_i=\begin{pmatrix}a_i\\ b_i\end{pmatrix}$, $i\in I$, через G — выпуклую оболочку множества F, $G=\mathrm{co}(F)$, и через K — совокупность всех крайних точек выпуклого множества G. Ясно, что $K\subset F$.

Пусть $J = \{j \in I \mid f_j \in K\}$.

ЛЕММА 3. *Справедливо равенство*

$$\varphi(x) = \max_{j\in J}\big\{\langle a_j, x\rangle + b_j\big\}.$$

Таким образом, максимум по I сводится к максимуму по подмножеству J.

Доказательство. По теореме Крейна-Мильмана [3, с. 32] любой вектор f_i при $i \in I \setminus J$ допускает представление

$$f_i = \sum_{j\in J} \lambda_j f_j,$$

где $\lambda_j \geqslant 0$ и $\sum\limits_{j\in J} \lambda_j = 1$. В подробной записи:

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} = \sum_{j\in J} \lambda_j \begin{pmatrix} a_j \\ b_j \end{pmatrix}.$$

Отсюда следует, что

$$\begin{aligned}\langle a_i, x\rangle + b_i = \sum_{j\in J} \lambda_j\big[\langle a_j, x\rangle + b_j\big] \leqslant \\ \leqslant \max_{j\in J}\big\{\langle a_j, x\rangle + b_j\big\}.\end{aligned}$$

Теперь имеем

$$\begin{aligned}\varphi(x) = \max\Big\{\max_{j\in J}\big\{\langle a_j, x\rangle + b_j\big\},\ \max_{i\in I\setminus J}\big\{\langle a_i, x\rangle + b_i\big\}\Big\} = \\ = \max_{j\in J}\big\{\langle a_j, x\rangle + b_j\big\}.\end{aligned}$$

Лемма доказана. □

7°. В заключение приведём основные свойства часто встречающейся плюсиковой функции

$$[x]_+ = \max\{0, x\},\ \ x \in \mathbb{R}.$$

А именно:

(I) $x \leqslant [x]_+ \leqslant |x|$;

(II) $[tx]_+ = t[x]_+$ при $t \geqslant 0$;

(III) $[x+y]_+ \leqslant [x]_+ + [y]_+$;

(IV) $\bigl|[x]_+ - [y]_+\bigr| \leqslant |x-y|$;

(V) $\max_{s\in 1:h}[x_s]_+ = [\max_{s\in 1:h} x_s]_+$;

(VI) $\min_{s\in 1:h}[x_s]_+ = [\min_{s\in 1:h} x_s]_+$;

(VII) $\max_{s\in 1:h}[x_s+y_s]_+ \leqslant \max_{s\in 1:h}[x_s]_+ + \max_{s\in 1:h}[y_s]_+$;

(VIII) $\bigl|\max_{s\in 1:h}[x_s]_+ - \max_{s\in 1:h}[y_s]_+\bigr| \leqslant \max_{s\in 1:h}|x_s-y_s|$;

(IX) $\min_{s\in 1:h}[x_s]_+ \leqslant \min_{s\in 1:h}[y_s]_+ + \max_{s\in 1:h}|x_s-y_s|$;

(X) $\bigl|\min_{s\in 1:h}[x_s]_+ - \min_{s\in 1:h}[y_s]_+\bigr| \leqslant \max_{s\in 1:h}|x_s-y_s|$.

Докажем некоторые из этих свойств.

Свойство III. Согласно I имеем $x+y \leqslant [x]_+ + [y]_+$. Вместе с тем, $0 \leqslant [x]_+ + [y]_+$. Поэтому

$$[x+y]_+ = \max\{0, x+y\} \leqslant [x]_+ + [y]_+.$$

Свойство IV. Имеем $x = y + (x-y) \leqslant [y]_+ + |x-y|$, так что $[x]_+ \leqslant [y]_+ + |x-y|$. Аналогично $[y]_+ \leqslant [x]_+ + |y-x|$. Из двух последних неравенств следует требуемое.

Свойство V. Если $\max_{s\in 1:h} x_s \leqslant 0$, то утверждение очевидно $(0=0)$. В противном случае и правая, и левая части доказываемого соотношения равны $\max_{s\in 1:h} x_s$.

Свойство VI. Если $\min_{s\in 1:h} x_s \geqslant 0$, то обе части доказываемого соотношения равны $\min_{s\in 1:h} x_s$. В противном случае утверждение очевидно $(0=0)$.

Свойство IX. В случае $\min_{s\in 1:h} x_s \leqslant 0$ утверждение в силу свойства VI очевидно. Пусть все x_s положительны. Тогда

$$[x_s]_+ = x_s = y_s + (x_s - y_s) \leqslant [y_s]_+ + \max_{s\in 1:h}|x_s-y_s|.$$

Отсюда следует, что

$$\min_{s\in 1:h}[x_s]_+ \leqslant [y_s]_+ + \max_{s\in 1:h}|x_s-y_s|$$

и

$$[y_s]_+ \geqslant \min_{s\in 1:h}[x_s]_+ - \max_{s\in 1:h}|x_s-y_s|.$$

Взяв в левой части минимум по $s \in 1:h$, придём к неравенству, которое равносильно требуемому.

ЛИТЕРАТУРА

1. Демьянов В. Ф., Малозёмов В. Н. *Введение в минимакс.* М.: Наука, 1972. 368 с.

2. Демьянов В. Ф., Рубинов А. М. *Основы негладкого анализа и квазидифференциальное исчисление.* М.: Наука, 1990. 432 с.

3. Лейхтвейс К. Выпуклые множества. М.: Наука, 1985. 336 с.

ПАРАМЕТРИЧЕСКИЕ ВАРИАНТЫ НЕРАВЕНСТВА ТРЕУГОЛЬНИКА*

В. Г. Малинов, Н. А. Соловьёва

Аннотация. В докладе приводятся параметрические варианты неравенства треугольника в евклидовом пространстве.

1°. Через E^n будем обозначать n-мерное евклидово пространство.

ЛЕММА 1. *Для любых $u, v, w \in E^n$ и произвольного $\varepsilon > 0$ верно двойное неравенство*

$$(1-\varepsilon)\|u-v\|^2+(1-\tfrac{1}{\varepsilon})\|v-w\|^2 \leqslant \|u-w\|^2 \leqslant (1+\varepsilon)\|u-v\|^2+(1+\tfrac{1}{\varepsilon})\|v-w\|^2.$$

Доказательство. Запишем

$$\|u-w\|^2 = \|u-v\|^2 + 2\langle u-v, v-w\rangle + \|v-w\|^2. \tag{1}$$

Заметим, что для любых вещественных чисел a, b и любого положительного ε верно

$$\left(\sqrt{\varepsilon}\,|a| - \tfrac{1}{\sqrt{\varepsilon}}|b|\right)^2 \geqslant 0.$$

Отсюда следует, что

$$2|ab| \leqslant \varepsilon a^2 + \tfrac{1}{\varepsilon}b^2.$$

С учётом последнего неравенства получим

$$2\Big|\langle u-v, v-w\rangle\Big| \leqslant 2\|u-v\|\|v-w\| \leqslant \varepsilon\|u-v\|^2 + \tfrac{1}{\varepsilon}\|v-w\|^2. \tag{2}$$

На основании (1) и (2) приходим к требуемому результату. □

2°. Рассмотрим произвольные $u, v, w \in E^n$. Обозначим $\ell_1 = \|u-v\|^2$, $\ell_2 = \|u-w\|^2$, $\ell_3 = \|v-w\|^2$, $s = \ell_1 + \ell_2 + \ell_3$. Предположим, что $\ell_2,\ \ell_3 > 0$. Рассмотрим квадратное уравнение

$$\ell_2\varepsilon^2 - s\varepsilon + \ell_3 = 0. \tag{3}$$

*Семинар «CNSA & NDO». Избранные доклады. 2 марта 2017 г.

Заметим, что уравнение (3) имеет вещественные корни, так как дискриминант этого уравнения неотрицателен. В самом деле,

$$s^2 - 4\ell_2\ell_3 = (\ell_1 + \ell_2 + \ell_3)^2 - 4\ell_2\ell_3 = \ell_1^2 + 2\ell_1\ell_2 + 2\ell_1\ell_3 + (\ell_2 - \ell_3)^2 \geqslant 0.$$

Корни уравнения (3) имеют вид

$$\varepsilon_{1,2} = \frac{s \mp \sqrt{s^2 - 4\ell_2\ell_3}}{2\ell_2}.$$

Ясно, что $0 < \varepsilon_1 \leqslant \varepsilon_2$.

ЛЕММА 2. *Для любого ε из отрезка $[\varepsilon_1, \varepsilon_2]$ верно неравенство*

$$\|u - v\|^2 \geqslant (\varepsilon - 1)\|u - w\|^2 - (1 - \tfrac{1}{\varepsilon})\|v - w\|^2. \tag{4}$$

Доказательство. Перепишем (4) в терминах ℓ_1, ℓ_2, ℓ_3:

$$\ell_1 \geqslant (\varepsilon - 1)\ell_2 - (1 - \tfrac{1}{\varepsilon})\ell_3.$$

Приведя подобные, получим неравенство

$$s \geqslant \varepsilon\ell_2 + \tfrac{1}{\varepsilon}\ell_3.$$

Умножим обе его части на ε (любое ε из отрезка $[\varepsilon_1, \varepsilon_2]$ положительно). Имеем

$$\ell_2\varepsilon^2 - s\varepsilon + \ell_3 \leqslant 0. \tag{5}$$

Неравенство (5) выполняется, так как ε расположено между корнями ε_1 и ε_2 квадратного уравнения (3) и $\ell_2 > 0$. Значит, выполняется и неравенство (4). □

Замечание 1. Если $\ell_3 = 0$, а $\ell_2 > 0$, то $\varepsilon_1 = 0$. Тогда неравенство (4) выполняется при $\varepsilon \in (0, \varepsilon_2]$, где $\varepsilon_2 = \frac{s}{\ell_2}$.

Замечание 2. Если $\ell_1 = 0$, а $\ell_2 = \ell_3$, то $\varepsilon_1 = \varepsilon_2 = 1$. В этом случае неравенство (4) становится тривиальным.

3°. Посмотрим, как зависят границы ε_1, ε_2 «допустимого» отрезка в лемме 2 от взаимного отношения величин $\|u - v\|$, $\|u - w\|$, $\|v - w\|$. Будем предполагать, что для точек u, v, w выполняется неравенство треугольника.

Вначале ограничимся случаем, когда

$$\|u - w\| = \|u - v\| = \alpha\|v - w\|. \tag{6}$$

Для выполнения неравенства треугольника потребуем, чтобы $2\alpha \geqslant 1$, то есть $\alpha \geqslant \frac{1}{2}$. Имеем

$$\ell_1 = \ell_2 = \alpha^2 \ell_3, \quad s = (2\alpha^2 + 1)\ell_3.$$

Вычислим границы ε_1, ε_2 отрезка, на котором выполняется неравенство (4):

$$\varepsilon_{1,2}(\alpha) = \frac{(2\alpha^2 + 1) \mp \sqrt{4\alpha^4 + 1}}{2\alpha^2}. \tag{7}$$

Обозначим через $L(\alpha)$ длину отрезка $[\varepsilon_1, \varepsilon_2]$:

$$L(\alpha) = \sqrt{4 + \tfrac{1}{\alpha^4}}\,. \tag{8}$$

Запишем производную функции $L(\alpha)$:

$$L'(\alpha) = \frac{-2}{\alpha^3 \sqrt{4\alpha^4 + 1}}.$$

Ясно, что функция $L(\alpha)$ убывает на полуоси $[\frac{1}{2}, +\infty)$.

Приведём несколько примеров.

ПРИМЕР 1. Рассмотрим вырожденный случай, когда точки u, v, w расположены на одной прямой:

$$\|u - w\| = \|u - v\| = \tfrac{1}{2}\|v - w\|.$$

Вычислим границы «допустимого» отрезка (см. (7)):

$$\varepsilon_{1,2}(\tfrac{1}{2}) = 3 \mp \sqrt{5}.$$

Получим $\varepsilon_1 \approx 0.76$, $\varepsilon_2 \approx 5.24$. Неравенство (4) будет выполняться, например, при любом ε из отрезка $[0.8, 5.2]$. По формуле (8) найдём длину $L(\frac{1}{2})$ отрезка $[\varepsilon_1, \varepsilon_2]$:

$$L(\tfrac{1}{2}) = 2\sqrt{5} \approx 4.47.$$

ПРИМЕР 2. Пусть

$$\|u - w\| = \|u - v\| = \tfrac{3}{5}\|v - w\|,$$

то есть $\alpha = \frac{3}{5}$. Запишем выражения (7) для границ «допустимого» отрезка:

$$\varepsilon_{1,2}(\tfrac{3}{5}) = \frac{43 \mp \sqrt{949}}{18}.$$

Вычислим $\varepsilon_1 \approx 0.68$, $\varepsilon_2 \approx 4.10$. Найдём (см. (8)) длину отрезка $[\varepsilon_1, \varepsilon_2]$:

$$L(\tfrac{3}{5}) = \frac{\sqrt{949}}{9} \approx 3.42.$$

ПРИМЕР 3. Пусть

$$\|u-w\| = \|u-v\| = \tfrac{2}{3}\|v-w\|,$$

то есть $\alpha = \frac{2}{3}$. Выражения (7) для границ «допустимого» отрезка будут выглядеть так:

$$\varepsilon_{1,2}(\tfrac{2}{3}) = \frac{17 \mp \sqrt{145}}{8}.$$

Вычислим $\varepsilon_1 \approx 0.62$, $\varepsilon_2 \approx 3.63$. По формуле (8) найдём длину отрезка $[\varepsilon_1, \varepsilon_2]$:

$$L(\tfrac{2}{3}) = \frac{\sqrt{145}}{4} \approx 3.01.$$

Примеры 1, 2, 3 иллюстрируют случай, когда для сторон треугольника с вершинами u, v, w выполняется соотношение (6). Приведём примеры с другим соотношением сторон треугольника.

ПРИМЕР 4. Пусть

$$\|u-v\| = \tfrac{2}{5}\|v-w\|, \quad \|u-w\| = \tfrac{4}{5}\|v-w\|.$$

В этом случае $\ell_1 = \frac{16}{25}\ell_3$, $\ell_2 = \frac{4}{25}\ell_3$, $s = \frac{45}{25}\ell_3$. Границы «допустимого» отрезка будут иметь вид

$$\varepsilon_{1,2} = \frac{45 \mp \sqrt{1625}}{8}.$$

Вычислим $\varepsilon_1 \approx 0.59$, $\varepsilon_2 \approx 10.66$, $\varepsilon_2 - \varepsilon_1 \approx 10.08$.

ПРИМЕР 5. Если принять

$$\|u-v\| = \tfrac{3}{5}\|v-w\|, \quad \|u-w\| = \|v-w\|,$$

то $\ell_1 = \frac{16}{25}\ell_3$, $\ell_2 = \ell_3$, $s = \frac{66}{25}\ell_3$. Запишем выражения для границ «допустимого» отрезка:

$$\varepsilon_{1,2} = \frac{66 \mp \sqrt{1856}}{50}.$$

Вычислим $\varepsilon_1 \approx 0.46$, $\varepsilon_2 \approx 2.18$, $\varepsilon_2 - \varepsilon_1 \approx 1.72$.

ЛИПШИЦЕВА НЕПРЕРЫВНОСТЬ ВЫПУКЛОЙ ФУНКЦИИ*

В. Н. Малозёмов, А. В. Плоткин

Аннотация. Для функций, выпуклых на открытом выпуклом множестве, установлена липшицева непрерывность с неулучшаемой константой Липшица.

Пусть $U \subset \mathbb{R}^n$ — открытое выпуклое множество и $f(x)$ — выпуклая на U функция. Возьмём точку $x_0 \in U$. Выберем число $\beta > 0$ так, чтобы $x_0 \pm \beta e_k \in U$ при всех $k \in 1{:}n$. Здесь e_k — единичные орты.

ТЕОРЕМА. *Справедливо неравенство*

$$|f(x) - f(x_0)| \leqslant L \left\|x - x_0\right\|_1 \tag{1}$$

при условии, что $\left\|x - x_0\right\|_1 \leqslant \beta$. *Константа* L *определяется формулой*

$$L = \max_{k \in 1:n} \left| \frac{f(x_0 \pm \beta e_k) - f(x_0)}{\beta} \right|. \tag{2}$$

Доказательство. Обозначим $h_k = \beta e_k$. Любой вектор $x \in \mathbb{R}^n$ допускает представление

$$x = x_0 + \sum_{k=1}^{n} w_k h_k. \tag{3}$$

Рассмотрим систему уравнений

$$\begin{aligned} w_k \ &= u_k - v_k, \\ |w_k| &= u_k + v_k. \end{aligned}$$

Получим

$$u_k = \tfrac{1}{2}\left(|w_k| + w_k\right), \quad v_k = \tfrac{1}{2}(|w_k| - w_k).$$

Очевидно, что числа u_k, v_k неотрицательные. При этом

$$|(x - x_0)_k| = \beta |w_k| = \beta(u_k + v_k).$$

Как следствие,

$$\left\|x - x_0\right\|_1 = \beta \sum_{k=1}^{n} (u_k + v_k). \tag{4}$$

*Семинар «CNSA & NDO». Избранные доклады. 24 ноября 2016 г.

Перепишем формулу (3) в виде

$$x - x_0 = \sum_{k=1}^{n} u_k h_k + \sum_{k=1}^{n} v_k(-h_k).$$

Обозначим $u_{n+k} = v_k$, $h_{n+k} = -h_k$. Тогда

$$x - x_0 = \sum_{k=1}^{2n} u_k h_k, \tag{5}$$

где все коэффициенты u_k неотрицательные.

Зафиксируем точку x, удовлетворяющую условию $\|x - x_0\|_1 \leqslant \beta$. Согласно (4),

$$\sum_{k=1}^{2n} u_k = \sum_{k=1}^{n}(u_k + v_k) \leqslant 1. \tag{6}$$

Из (5) следует, что

$$x = \left(1 - \sum_{k=1}^{2n} u_k\right) x_0 + \sum_{k=1}^{2n} u_k(x_0 + h_k).$$

Все коэффициенты в этом представлении неотрицательны и в сумме равны единице, а точки $x_0, x_0+h_1, \ldots, x_0+h_{2n}$ принадлежат выпуклому множеству U. По неравенству Йенсена для выпуклых функций

$$f(x) - f(x_0) \leqslant \sum_{k=1}^{2n} u_k\big[f(x_0 + h_k) - f(x_0)\big] \leqslant L\beta \sum_{k=1}^{2n} u_k,$$

где константа L определяется формулой (2). В силу (6) и (4)

$$\beta \sum_{k=1}^{2n} u_k = \beta \sum_{k=1}^{n}(u_k + v_k) = \|x - x_0\|_1, \tag{7}$$

так что

$$f(x) - f(x_0) \leqslant L\,\|x - x_0\|_1. \tag{8}$$

Оценим разность $f(x_0) - f(x)$. Введём точку $y = 2x_0 - x$. Имеем

$$y - x_0 = x_0 - x = \sum_{k=1}^{n} u_k(-h_k) + \sum_{k=1}^{n} v_k h_k.$$

Положив $v_{n+k} = u_k$ при $k \in 1:n$, запишем

$$y - x_0 = \sum_{k=1}^{2n} v_k h_k.$$

При этом согласно (6)

$$\sum_{k=1}^{2n} v_k = \sum_{k=1}^{n} (v_k + u_k) \leqslant 1. \tag{9}$$

Как и раньше, воспользуемся представлением

$$y = \left(1 - \sum_{k=1}^{2n} v_k\right) x_0 + \sum_{k=1}^{2n} v_k (x_0 + h_k)$$

и неравенством Йенсена для выпуклых функций. Получим

$$f(y) - f(x_0) \leqslant \sum_{k=1}^{2n} v_k \big[f(x_0 + h_k) - f(x_0)\big] \leqslant L\beta \sum_{k=1}^{2n} v_k.$$

Из (6), (9) и (7), в частности, следует, что

$$\beta \sum_{k=1}^{2n} v_k = \beta \sum_{k=1}^{2n} u_k = \|x - x_0\|_1 .$$

Значит,

$$f(y) - f(x_0) \leqslant L \|x - x_0\|_1 . \tag{10}$$

Теперь отметим, что $x_0 = \frac{1}{2}(y + x)$. В силу выпуклости функции f

$$f(x_0) \leqslant \tfrac{1}{2}\big[f(y) + f(x)\big],$$

так что

$$f(y) \geqslant 2f(x_0) - f(x). \tag{11}$$

Объединив (10) и (11), придём к неравенству

$$f(x_0) - f(x) \leqslant L \|x - x_0\|_1 . \tag{12}$$

Из (8) и (12) следует (1). Теорема доказана. □

Эта теорема является усилением результата из книги [1, с. 61].

Отметим, что для выпуклой на $\mathbb{R}^n$ функции $f(x) = \|x\|_1$ при $x_0 = \mathbb{O}$ и произвольном $\beta > 0$ неравенство (1) выполняется как равенство (с $L = 1$).

ЛИТЕРАТУРА

1. Пшеничный Б. Н. *Выпуклый анализ и экстремальные задачи.* М.: Наука, 1980. 320 с.

НЕУЛУЧШАЕМАЯ ЛОКАЛЬНАЯ КОНСТАНТА ЛИПШИЦА ДЛЯ ВЫПУКЛОЙ ФУНКЦИИ*

М. Э. Аббасов, В. Н. Малозёмов

Аннотация. Для функций, выпуклых на открытом выпуклом множестве в евклидовом пространстве с чебышевской нормой, установлена липшицева непрерывность с неулучшаемой и легко вычислимой константой Липшица. Аналогичный результат в случае ℓ_1-нормы получен в [1].

1°. Пусть $U \subset \mathbb{R}^n$ — открытое выпуклое множество и $f(x)$ — функция, выпуклая на U. Обозначим $N = 2^n$ и введём систему векторов $\{h_i\}$, $i \in 1:N$, вида $(\pm 1, \ldots, \pm 1)^T$. Возьмём точку $x_0 \in U$. Выберем число $\beta > 0$ так, чтобы $x_0 + \beta h_i \in U$ при всех $i \in 1:N$.

ТЕОРЕМА. *Для любого вектора x, удовлетворяющего условию $\|x - x_0\|_\infty \leqslant \leqslant \beta$, выполняется неравенство*

$$|f(x) - f(x_0)| \leqslant L\|x - x_0\|_\infty, \tag{1}$$

где

$$L = \max_{i=1:N} \left| \frac{f(x_0 + \beta h_i) - f(x_0)}{\beta} \right|. \tag{2}$$

2°. Для доказательства теоремы нам потребуется одно вспомогательное утверждение.

С произвольным ненулевым вектором $x = \left(x^{(1)}, \ldots, x^{(n)}\right)^T$ свяжем подсистему $\left\{h_i^{(x)}\right\}$, $i \in 1:N/2$, системы $\left\{h_i^{(x)}\right\}$, $i \in 1:N$. Для этого выделим индекс $k \in 1:n$, на котором $\|x\|_\infty = \left|x^{(k)}\right|$. Подсистему $\left\{h_i^{(x)}\right\}$ составим из векторов

$$(\pm 1, \ldots, \pm 1, \underbrace{\operatorname{sign} x^{(k)}}_{k\text{-я позиция}}, \pm 1, \ldots, \pm 1)^T.$$

*Семинар «CNSA & NDO». Избранные доклады. 22 декабря 2016 г.

ЛЕММА. *Вектор x допускает представление*

$$x = \sum_{i=1}^{N/2} \lambda_i h_i^{(x)}, \tag{3}$$

где коэффициенты λ_i неотрицательны и

$$\sum_{i=1}^{N/2} \lambda_i = \|x\|_\infty. \tag{4}$$

Доказательство. Обозначим через A матрицу со столбцами $\left\{h_i^{(x)}\right\}$, $i \in \in 1 : N/2$. Покажем, что система $A\lambda = x$ имеет неотрицательное решение λ. По теореме Фаркаша достаточно проверить, что из условия $u^T A \geqslant \mathbb{O}$ следует неравенство $\langle x, u\rangle \geqslant 0$.

Распишем условие $u^T A \geqslant \mathbb{O}$ подробно:

$$u^{(k)} \operatorname{sign} x^{(k)} + \sum_{i\neq k} \left(\pm u^{(i)}\right) \geqslant 0.$$

В частности,

$$u^{(k)} \operatorname{sign} x^{(k)} + \sum_{i\neq k} \left|u^{(i)}\right| \geqslant 0,$$

$$u^{(k)} \operatorname{sign} x^{(k)} - \sum_{i\neq k} \left|u^{(i)}\right| \geqslant 0.$$

Отсюда следует, что $u^{(k)} \operatorname{sign} x^{(k)} \geqslant 0$ и

$$\sum_{i\neq k} \left|u^{(i)}\right| \leqslant u^{(k)} \operatorname{sign} x^{(k)} = \left|u^{(k)}\right|. \tag{5}$$

Обратимся к скалярному произведению $\langle x, u\rangle$. Запишем

$$\begin{aligned}\langle x, u\rangle =& x^{(k)}u^{(k)} + \sum_{i\neq k} x^{(i)}u^{(i)} = \left|x^{(k)}\right| \cdot |u^{(k)}| + \sum_{i\neq k} x^{(i)}u^{(i)} \geqslant \\ \geqslant& \left|x^{(k)}\right| \cdot |u^{(k)}| - \sum_{i\neq k} \left|x^{(k)}\right| \cdot |u^{(i)}| = \left|x^{(k)}\right| \left(|u^{(k)}| - \sum_{i\neq k} |u^i|\right).\end{aligned}$$

Согласно (5), $\langle x, u\rangle \geqslant 0$.

Установлено, что система $A\lambda = x$ имеет неотрицательное решение λ. Это значит, что справедливо представление (3) с неотрицательными коэффициентами λ_i. Равенство (4) тоже выполняется, поскольку

$$\|x\|_\infty = \left|x^{(k)}\right| = x^{(k)} \operatorname{sign} x^{(k)} = \sum_{i=1}^{N/2} \lambda_i \left(\operatorname{sign} x^{(k)}\right)^2 = \sum_{i=1}^{N/2} \lambda_i.$$

Лемма доказана. □

3°. Пусть $f(x)$ — функция, выпуклая на открытом выпуклом множестве $U \subset \mathbb{R}^n$. Возьмём точку $x_0 \in U$ и число $\beta > 0$ такое, что $x_0 + \beta h_i \in U$ при всех $i \in 1 : N$.

Зафиксируем точку x, отличную от x_0 и удовлетворяющую неравенству $\|x - x_0\| \leqslant \beta$. Обозначим $u = x - x_0$. Согласно лемме ненулевой вектор u допускает представление

$$u = \sum_{i=1}^{N/2} \lambda_i h_i^{(u)}, \tag{6}$$

где коэффициенты λ_i неотрицательны и

$$\sum_{i=1}^{N/2} \lambda_i = \|u\|_\infty.$$

Введём новые коэффициенты $\alpha_i = \lambda_i/\beta$ и перепишем (6) в виде

$$u = \sum_{i=1}^{N/2} \alpha_i \left(\beta h_i^{(u)}\right). \tag{7}$$

Здесь α_i неотрицательны и

$$\sum_{i=1}^{N/2} \alpha_i = \frac{\|x - x_0\|_\infty}{\beta} \leqslant 1.$$

Из (7) следует, что

$$x = x_0 + \sum_{i=1}^{N/2} \alpha_i \left(\beta h_i^{(u)}\right) = \left(1 - \sum_{i=1}^{N/2} \alpha_i\right) x_0 + \sum_{i=1}^{N/2} \alpha_i \left(x_0 + \beta h_i^{(u)}\right).$$

В этом разложении вектора x по векторам x_0, $x_0 + \beta h_1^{(u)}$, …, $x_0 + \beta h_{N/2}^{(u)}$, принадлежащим множеству U, коэффициенты неотрицательны и в сумме равны

единице. Значит, $x \in U$. По неравенству Йенсена получаем

$$f(x) - f(x_0) \leqslant \sum_{i=1}^{N/2} \lambda_i \frac{f\big(x_0 + \beta h_i^{(u)}\big) - f(x_0)}{\beta} \leqslant L \sum_{i=1}^{N/2} \lambda_i = L\|x - x_0\|_\infty. \tag{8}$$

Теперь оценим $f(x_0) - f(x)$. Введём вектор $y = 2x_0 - x$. В силу (7) имеем

$$y - x_0 = x_0 - x = \sum_{i=1}^{N/2} \alpha_i \left(-\beta h_i^{(u)}\right).$$

Перепишем это равенство в эквивалентном виде

$$y = \left(1 - \sum_{i=1}^{N/2} \alpha_i\right) x_0 + \sum_{i=1}^{N/2} \alpha_i \left(x_0 - \beta h_i^{(u)}\right).$$

Отсюда следует, что $y \in U$. По неравенству Йенсена

$$f(y) - f(x_0) \leqslant \sum_{i=1}^{N/2} \lambda_i \frac{f\big(x_0 - \beta h_i^{(u)}\big) - f(x_0)}{\beta} \leqslant L \sum_{i=1}^{N/2} \lambda_i = L\|x - x_0\|_\infty. \tag{9}$$

Отметим, что $x_0 = \frac{1}{2}(x + y)$. В силу выпуклости функции f

$$f(x_0) \leqslant \frac{1}{2}\left[f(y) + f(x)\right],$$

так что

$$f(y) \geqslant 2f(x_0) - f(x). \tag{10}$$

Объединив (9) и (10), придём к неравенству

$$f(x_0) - f(x) \leqslant L\|x - x_0\|_\infty. \tag{11}$$

На основании (8) и (11) заключаем, что справедливо требуемое неравенство (1) с константой L вида (2).

Теорема доказана. □

Замечание 1. Константа Липшица L в неравенстве (1) неулучшаема. Это следует из того, что при $f(x) = \|x\|_\infty$, $x_0 = \mathbb{O}$ и произвольном $\beta > 0$ неравенство (1) выполняется как равенство.

Замечание 2. Если $f(x_0 + \beta h_i) = f(x_0)$ при $i \in 1 : N$, то $f(x) = f(x_0)$ для всех x, удовлетворяющих условию $\|x - x_0\|_\infty \leqslant \beta$.

ЛИТЕРАТУРА

1. Малоземов В. Н., Плоткин А. В. *Липшицева непрерывность выпуклой функции* // Семинар «CNSA & NDO». Избранные доклады. 24 ноября 2016 г. (`http://apmath.spbu/cnsa/reps16.shtml#1124b`) [Данная книга, с. 461]

СПИСОК АВТОРОВ КНИГИ

– Аббасов Меджид Эльхан оглы (кандидат физ.-мат. наук, доцент)
abbasov.majid@gmail.com

– Агафонова Ирина Витальевна (кандидат физ.-мат. наук, доцент)
i.agafonova@spbu.ru

– Ангелов Тодор Ангелов (аспирант)
angelov.t@gmail.com

– Гаудиозо Манлио (профессор Калабрийского университета, Италия)
gaudioso@deis.unical.it

– Гхашим Мохамад (кандидат физ.-мат. наук, доцент Севастопольского университета)
mohgh110@yahoo.com

– Даугавет Валентина Александровна (кандидат физ.-мат. наук, доцент)
vadaug@yandex.ru

– Долгополик Максим Владимирович (кандидат физ.-мат. наук)
maxim.dolgopolik@gmail.com

– Кольцов Максим Алексеевич (бакалавр)
kolmax94@gmail.com

– Лазарев Алексей Викторович (кандидат физ.-мат. наук, доцент Петрозаводского университета)
lazarev_av@sampo.ru

– Малинов Валерий Григорьевич (кандидат физ.-мат. наук, доцент)
vgmalinov@mail.ru

– Малозёмов Василий Николаевич (доктор физ.-мат. наук, профессор)
v.malozemov@spbu.ru

– Михеев Сергей Евгеньевич (доктор физ.-мат. наук, профессор)
him2@mail.ru

– Наумова Наталия Ивановна (кандидат физ.-мат. наук, доцент)
nataliai.naumova@mail.ru

– Плоткин Артём Владимирович (бакалавр)
avplotkin@gmail.com

– Полякова Людмила Николаевна (доктор физ.-мат. наук, профессор)
lnpol07@mail.ru

– Романовский Иосиф Владимирович (доктор физ.-мат. наук, профессор)
josephromanovsky@gmail.com

– Соловьёва Наталья Анатольевна (кандидат физ.-мат. наук)
vinyo@mail.ru

– Сукач Михаил Петрович (аспирант)
nskmike@yandex.ru

– Тамасян Григорий Шаликович (кандидат физ.-мат. наук, доцент)
g.tamasyan@spbu.ru

– Чернэуцану Екатерина Константиновна (кандидат физ.-мат. наук)
katerinache@yandex.ru

– Чумаков Андрей Александрович (аспирант)
andrew1991.spb@gmail.com

Научное издание

ИЗБРАННЫЕ ЛЕКЦИИ ПО ЭКСТРЕМАЛЬНЫМ ЗАДАЧАМ

Часть первая

Под редакцией проф. В. Н. Малозёмова