НАО «Восточно-Казахстанский технический университет имени Д. Серикбаева»

Дисциплина «Методы оптимизации»

ЛЕКЦИЯ 5

Метод искусственного базиса

Лекция 5. Объём — 2 академических часа. Образовательная программа 6B06103 «Математическое и компьютерное моделирование».

1 Цель и связь с предыдущими темами

Симплексная таблица из лекции 4 позволяет улучшать уже известный допустимый базисный план. Однако после приведения ограничений к каноническому виду подходящего базиса может не оказаться. Метод искусственного базиса решает эту начальную задачу: временно вводит вспомогательные переменные и затем проверяет, можно ли сделать их равными нулю.

После занятия студент должен уметь отличать дополнительные переменные от искусственных; строить допустимый начальный базис расширенной задачи; выполнять две фазы симплекс-метода; распознавать несовместность ограничений; корректно удалять искусственные переменные; объяснять связь двухфазного подхода с методом большого штрафа.

В лекции используется прежнее соглашение: задача максимизации, оценки $\Delta_j = c_j - z_j$, положительная оценка указывает на возможное улучшение. Переменные для входа и выхода выбираются по правилу Бланда. Предполагается знание канонической формы, базиса и пересчёта таблицы из лекций 2–4. Искусственные переменные и M-задача рассматриваются в [1, § 11.1, с. 294], вспомогательная задача для поиска допустимого базиса — в [2, § 11.2.4].

2 Когда требуется искусственный базис

Рассмотрим равенства $Ax = b$, $x \geq 0$. Если в матрице имеются единичные столбцы, а соответствующие правые части неотрицательны, из этих столбцов можно составить допустимый базис. Например, ограничения типа «не более» при $b_i \geq 0$ после введения остатков часто сразу дают нужную единичную матрицу.

Иная ситуация возникает при равенствах и ограничениях типа «не менее». Для $x_1 + x_2 \geq 4$ вводится переменная избытка $s_1 \geq 0$:

$$x_1 + x_2 - s_1 = 4.$$

Если положить $x_1 = x_2 = 0$, получим $s_1 = -4$, что недопустимо. Само наличие отдельного столбца ещё не означает существования допустимого базиса. В таком случае временно добавляют искусственную переменную $a_1 \geq 0$:

$$x_1 + x_2 - s_1 + a_1 = 4.$$

Теперь план $x_1 = x_2 = s_1 = 0$, $a_1 = 4$ допустим для расширенного равенства. Он пока не является допустимым решением исходного ограничения. Искусственная переменная измеряет ту часть равенства, которую временно обеспечивает вспомогательный столбец.

Тип ограничения при неотрицательной правой части	Преобразованное равенство	Начальная базисная переменная
$a_i^T x \le b_i$	$a_i^T x + s_i = b_i$	Остаток s_i
$a_i^T x \ge b_i$	$a_i^T x - s_i + a_i^{\text{art}} = b_i$	Искусственная a_i^{art}
$a_i^T x = b_i$	$a_i^T x + a_i^{\text{art}} = b_i$	Искусственная a_i^{art}

В таблице a_i^T обозначает строку коэффициентов, а верхний индекс «art» отличает искусственную переменную от этих коэффициентов. Далее искусственные переменные обозначаются просто a_1, a_2. Все добавленные переменные неотрицательны.

Приведённое правило гарантирует простой начальный базис, но не требует вводить искусственную переменную там, где уже найден другой подходящий допустимый базис. Остатки и избытки сохраняют смысл исходных неравенств и могут быть ненулевыми в окончательном ответе. Искусственные переменные в решении исходной задачи должны быть равны нулю.

Если $b_i < 0$, сначала умножают всё ограничение на -1. Знак неравенства при этом меняется. Например, $-x_1 - x_2 \le -4$ переходит в $x_1 + x_2 \ge 4$. Добавление положительной искусственной переменной до нормировки правой части не устраняет отрицательность начального базисного значения.

3 Вспомогательная задача первой фазы

Пусть после приведения ограничений к нужному виду в расширенную систему введены k искусственных переменных. На первой фазе минимизируется их сумма:

$$w = a_1 + \cdots + a_k \to \min.$$

Чтобы использовать ту же табличную схему максимизации, решаем эквивалентную задачу

$$F_I = -a_1 - \cdots - a_k \to \max.$$

Коэффициенты всех обычных переменных в этой цели равны нулю, коэффициенты искусственных — -1. Исходная целевая функция на первой фазе не оптимизируется. Начальное значение w определяется искусственными базисными переменными; при стандартном построении оно равно сумме соответствующих правых частей.

3.1 Почему первая фаза проверяет совместность

Поскольку $a_i \ge 0$, всегда $w \ge 0$ и $F_I \le 0$. Если исходная система имеет допустимое решение, дополнение этого решения нулевыми искусственными переменными даёт допустимое решение расширенной системы с $w = 0$. Значит, оптимум вспомогательной задачи равен нулю.

Обратно, если найден допустимый план расширенной системы с $w = 0$, то каждое слагаемое неотрицательной суммы равно нулю. Искусственные столбцы перестают вносить вклад в равенства, а оставшиеся переменные удовлетворяют исходной системе. Поэтому

$$w^* = 0 \quad \Leftrightarrow \quad \text{исходные ограничения совместны.}$$

Если после завершения первой фазы $w^* > 0$, исходная задача не имеет допустимых решений. Здесь необходимо именно оптимальное значение. Положительные искусственные переменные в промежуточной таблице не являются доказательством несовместности.

При корректном построении первая фаза имеет допустимый начальный план и ограниченную сверху цель $F_I \le 0$. Обнаружение неограниченности на этой фазе означает ошибку в построении задачи, знаках или программе. Неограниченность исходной целевой функции может обнаружиться позднее, на второй фазе.

3.2 Почему нужно пересчитать строку оценок

Если искусственные переменные входят в базис, их коэффициенты $c_B = -1$ дают ненулевые значения $z_j = c_B^T B^{-1} A_j$. Поэтому строку оценок нельзя просто скопировать из вектора коэффициентов вспомогательной цели. Как и в лекции 4, сначала вычисляют

$$\Delta_j = c_j - c_B^T \overline{A}_j, \qquad F_{I,0} = c_B^T \overline{b}.$$

В корректной таблице оценки всех текущих базисных переменных равны нулю. Эта проверка особенно полезна при переходе от исходной цели к вспомогательной и обратно.

4 Полный пример с двумя фазами

Рассмотрим задачу

$$\max f = 3x_1 + 2x_2,$$

$$x_1 + x_2 \ge 4, \qquad x_1 + 2x_2 = 6, \qquad x_1 \le 5, \qquad x_1, x_2 \ge 0.$$

Введём избыток s_1, остаток s_2 и искусственные переменные a_1, a_2:

$$x_1 + x_2 - s_1 + a_1 = 4,$$

$$x_1 + 2x_2 + a_2 = 6,$$

$$x_1 + s_2 = 5.$$

Порядок столбцов: $(x_1, x_2, s_1, s_2, a_1, a_2)$. Начальный базис (a_1, a_2, s_2), начальные значения (4,6,5), а $F_I = -10$. Во всех таблицах последняя строка содержит оценки; в столбце b этой строки стоит текущее значение цели соответствующей фазы.

4.1 Начальная таблица первой фазы

Базис	c_B	b	x_1	x_2	s_1	s_2	a_1	a_2
a_1	-1	4	1	1	-1	0	1	0
a_2	-1	6	1	2	0	0	0	1
s_2	0	5	1	0	0	1	0	0
Δ	—	-10	2	3	-1	0	0	0

Для входа рассматриваются положительные оценки. По правилу Бланда входит x_1, хотя оценка x_2 больше. Отношения равны $4/1, 6/1, 5/1$. Из базиса выходит a_1, разрешающий элемент — единица в первой строке.

Первая строка сохраняется, из второй и третьей вычитается первая. Новый базис — (x_1, a_2, s_2), значения базисных переменных (4,2,1). Значение вспомогательной цели увеличивается до $F_I = -2$, то есть сумма искусственных переменных уменьшается до $w = 2$.

4.2 Таблица после первого перехода

Базис	c_B	b	x_1	x_2	s_1	s_2	a_1	a_2
x_1	0	4	1	1	-1	0	1	0
a_2	-1	2	0	1	1	0	-1	1
s_2	0	1	0	-1	1	1	-1	0
Δ	—	-2	0	1	1	0	-2	0

Теперь входит x_2. В его столбце положительные коэффициенты имеются в первой и второй строках: отношения равны 4 и 2. Коэффициент третьей строки отрицателен и в проверке отношений не участвует. Выходит a_2, разрешающая строка вторая.

После нормировки этой строки из первой вычитают вторую, а к третьей прибавляют вторую. Получается базис (x_1, x_2, s_2) со значениями (2,2,3). Искусственные переменные небазисные и равны нулю.

4.3 Завершение первой фазы

Базис	c_B	b	x_1	x_2	s_1	s_2	a_1	a_2
x_1	0	2	1	0	-2	0	2	-1
x_2	0	2	0	1	1	0	-1	1
s_2	0	3	0	0	2	1	-2	1
Δ	—	0	0	0	0	0	-1	-1

Все оценки неположительны, $F_I = 0$, следовательно, $w^* = 0$. Исходные ограничения совместны. После удаления искусственных столбцов остаётся допустимый план $(x_1, x_2, s_1, s_2) = (2,2,0,3)$.

Этот план ещё не обязательно оптимален для f. Первая фаза отвечала на вопрос о допустимости. При найденном плане исходная целевая функция равна $3 \cdot 2 + 2 \cdot 2 = 10$.

4.4 Возврат к исходной целевой функции

Удаляем столбцы a_1, a_2, возвращаем $c = (3,2,0,0)^T$ и пересчитываем строку оценок по текущему базису (x_1, x_2, s_2). Строки ограничений сохраняются. Коэффициенты цели базисных переменных теперь равны $(3,2,0)^T$.

Базис	c_B	b	x_1	x_2	s_1	s_2
x_1	3	2	1	0	-2	0
x_2	2	2	0	1	1	0
s_2	0	3	0	0	2	1

Базис	c_B	b	x_1	x_2	s_1	s_2
Δ	—	10	0	0	4	0

Положительная оценка $\Delta_{s_1} = 4$ показывает, что избыток первого ограничения следует увеличить. В столбце s_1 положительны коэффициенты второй и третьей строк. Отношения равны $2/1 = 2$ и $3/2$. Выходит s_2, входит s_1, разрешающий элемент равен 2.

Новая третья строка получается делением на 2. В первой строке коэффициент ведущего столбца равен -2, поэтому к первой строке прибавляется удвоенная новая третья. Из второй строки вычитается новая третья. Получаем заключительную таблицу.

Базис	c_B	b	x_1	x_2	s_1	s_2
x_1	3	5	1	0	0	1
x_2	2	$\frac{1}{2}$	0	1	0	$-\frac{1}{2}$
s_1	0	$\frac{3}{2}$	0	0	1	$\frac{1}{2}$
Δ	—	16	0	0	0	-2

Все оценки неположительны. Итоговый план

$$x_1^* = 5, \qquad x_2^* = \frac{1}{2}, \qquad s_1^* = \frac{3}{2}, \qquad s_2^* = 0,$$

$$f^* = 3 \cdot 5 + 2 \cdot \frac{1}{2} = 16.$$

В данном примере последовательность целей первой фазы равна $-10, -2, 0$, второй фазы — $10, 16$. Эти последовательности относятся к разным функциям; объединять их в одну монотонную последовательность значений исходной цели нельзя.

4.5 Независимая проверка ответа

Подставим решение в исходные ограничения: $5 + 1/2 \geq 4$, $5 + 2 \cdot (1/2) = 6$, $5 \leq 5$, обе переменные неотрицательны. Для проверки оптимальности исключим x_1 из равенства: $x_1 = 6 - 2x_2$. Остальные ограничения дают $1/2 \leq x_2 \leq 2$, а функция принимает вид $f = 18 - 4x_2$. Максимум достигается при $x_2 = 1/2$ и равен 16.

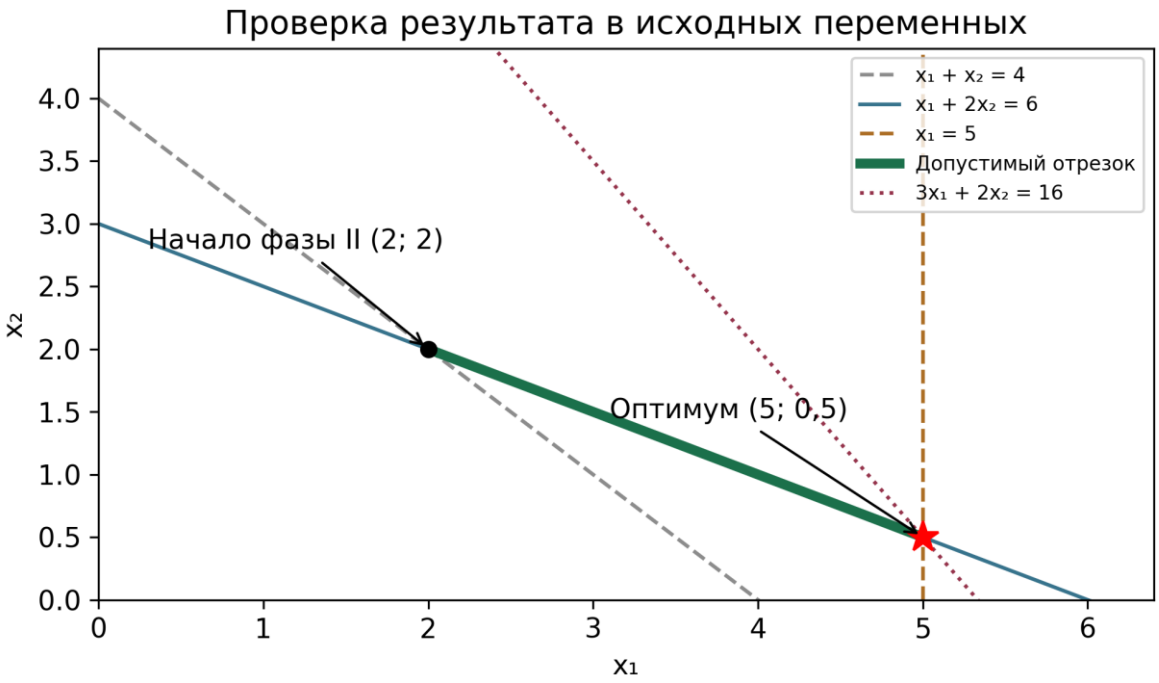


Рисунок 1 — Геометрическая проверка двухфазного решения. Точка начала второй фазы допустима, но ещё не оптимальна.

5 Если искусственная переменная осталась в базисе

Условие $w^* = 0$ означает, что значения всех искусственных переменных равны нулю. Однако некоторые из них могут остаться базисными при нулевой правой части. Это допустимая вырожденная ситуация. Она требует очистки базиса перед удалением искусственных столбцов.

Если в строке нулевой искусственной базисной переменной есть ненулевой коэффициент при обычной небазисной переменной, выполняют обмен по этому коэффициенту. Он может быть отрицательным: правая часть разрешающей строки равна нулю, поэтому правые части остальных строк не изменяются и допустимость сохраняется. Это технический обмен базиса, а не обычный улучшающий шаг с положительной оценкой.

Например, равенство $-x + a = 0$ при $x, a \geq 0$ и цели первой фазы $-a$ уже имеет оптимум нуль с базисом a. Для исключения искусственной переменной разрешается обмен по коэффициенту -1, после чего базисной становится $x = 0$.

Если после мысленного удаления искусственных столбцов вся такая строка превращается в $0 = 0$, она избыточна и удаляется. Так происходит для зависимых ограничений $x = 1$ и $2x = 2$. После введения двух искусственных переменных и одного перехода одна из строк может содержать только искусственные столбцы и нулевую правую часть.

Нельзя удалить искусственный столбец, пока он является единственным базисным столбцом сохраняемой строки. Иначе исчезнет корректное представление базиса. Порядок действий: вывести нулевые искусственные базисные переменные, удалить избыточные строки, удалить искусственные столбцы, вернуть исходную цель и пересчитать оценки.

6 Несовместность и неограниченность

Для системы $x \geq 2, x \leq 1, x \geq 0$ введём переменные:

$$x - s + a = 2, \qquad x + t = 1, \qquad x, s, a, t \geq 0.$$

Из второго равенства следует $x \leq 1$. Первое даёт $a = 2 - x + s \geq 1$. Минимально возможное значение искусственной переменной равно 1, поэтому $w^* = 1 > 0$: исходные ограничения несовместны. Это заключение не зависит от коэффициентов исходной целевой функции.

Другой исход возникает у задачи $\max x$ при $x \geq 1$. Ограничение совместно; первая фаза может найти точку $x = 1$. Но на второй фазе допустимы значения $x = 1 + t$ при любом $t \geq 0$, и цель не ограничена сверху. Таким образом, отсутствие допустимых точек и отсутствие конечного максимума — разные свойства задачи.

7 Связь с методом большого штрафа

Вместо двух последовательно решаемых задач можно записать расширенную цель

$$F_M = c^T x - M \sum_{i=1}^{k} a_i \to \max, \qquad M > 0.$$

Для задачи минимизации знак штрафа противоположный: к исходной цели прибавляют $M\sum a_i$. Метод называют методом большого M. Его идея — сделать положительные искусственные переменные невыгодными [1, § 11.1].

Символическое «достаточно большое M» означает приоритет устранения искусственных переменных. Произвольно выбранная числовая константа не обеспечивает этот приоритет для любой задачи. Например, исходная задача $\max(-1000x)$ при $x = 1, x \geq 0$ имеет единственное допустимое решение. Если записать $x + a = 1$ и взять $M = 100$, расширенная цель равна $-1000x - 100a = -100 - 900x$. Она максимальна при $x = 0$, $a = 1$, хотя исходная система совместна.

Слишком большой числовой штраф может, в свою очередь, ухудшать точность вычислений с плавающей точкой. Двухфазный подход явно разделяет приоритеты: сначала минимизируется сумма искусственных переменных, затем оптимизируется исходная функция на допустимом множестве. Для учебной программы используется именно этот вариант с точной рациональной арифметикой.

8 Проверка примера на Python

В архиве two_phase.py выполняет обе фазы, используя табличный алгоритм из simplex_tableau.py. Матрица передаётся уже в расширенной канонической форме. Автоматическое преобразование произвольных неравенств не выполняется: составление этой формы является частью учебной работы.

Ключевые исходные данные основного примера:

```
from two_phase import two_phase
A = [[1, 1, -1, 0, 1, 0],
     [1, 2,  0, 0, 0, 1],
     [1, 0,  0, 1, 0, 0]]
b = [4, 6, 5]
c = [3, 2, 0, 0, 0, 0]
result = two_phase(A, b, c,
                   basis=[4, 5, 3], artificial=[4, 5])
print(result['status'], result['w'])
print(result['phase2']['x'], result['phase2']['value'])
```

Индексы Python начинаются с нуля. Список basis=[4,5,3] означает (a_1, a_2, s_2). Вектор с задаёт исходную цель, поэтому коэффициенты искусственных переменных в нём нулевые. Цель первой фазы формируется отдельно внутри функции.

Команда python two_phase.py печатает все пять таблиц. При успешном решении status равен optimal, w равен нулю, а результат второй фазы содержит план (5,1/2,3/2,0) и значение 16. При infeasible вторая фаза не запускается. При unbounded конечного оптимального значения нет.

Используются объекты Fraction. Для дробных коэффициентов передавайте строки вида '1/3' либо точные дроби. Передача числа с плавающей точкой сохраняет его двоичное приближение. Процедура очистки отдельно допускает отрицательный ненулевой разрешающий элемент в строке с нулевой правой частью.

9 Самостоятельная работа

1. Для ограничений $-2x_1 - x_2 \le -6$ и $x_1 - x_2 = 1$ выполните нормировку и запишите расширенную систему с неотрицательными переменными. Укажите искусственный начальный базис.
2. В основном примере вычислите начальные оценки первой фазы по формуле $c_j - z_j$. Объясните, почему $\Delta_{x_1} = 2$, хотя коэффициент x_1 во вспомогательной цели равен нулю.
3. Повторите переход между второй и третьей таблицами первой фазы. Проверьте значения всех искусственных переменных.
4. Решите двухфазным методом задачу $\min(x_1 + 2x_2)$ при $x_1 + x_2 \ge 3$, $x_1 + 3x_2 \ge 5$, $x \ge 0$. Для табличной максимизации поменяйте знак исходной цели.
5. Докажите несовместность системы $x_1 + x_2 \ge 4$, $x_1 + x_2 \le 2$, $x \ge 0$ через положительное минимальное значение искусственной переменной.
6. Для ограничений $x_1 + x_2 = 2, 2x_1 + 2x_2 = 4, x \ge 0$ объясните, какая строка может оказаться избыточной при очистке базиса.
7. Для задачи $\max x_1$ при $x_1 - x_2 \ge 1, x \ge 0$ найдите допустимый начальный план и укажите луч, доказывающий неограниченность исходной цели.

Ориентиры для проверки. В задании 4 оптимум (2,1), минимальное значение 4. Для доказательства достаточно сложить исходные неравенства с весами 1/2 и 1/2: получим $x_1 + 2x_2 \ge 4$, и в указанной точке достигается равенство. В задании 5 минимальная сумма искусственных переменных равна 2 при стандартном добавлении одной искусственной переменной к первому ограничению. В задании 7 подходит луч $(1 + t, 0), t \ge 0$.

10 Контрольные вопросы

1. Чем искусственная переменная отличается от остатка и избытка?

2. Почему отрицательную правую часть нужно обработать до построения стандартного искусственного базиса?
3. Какая функция оптимизируется на первой фазе и почему она ограничена?
4. Почему нулевая сумма искусственных переменных доказывает допустимость исходного плана?
5. Можно ли сделать вывод о несовместности по промежуточной таблице?
6. Почему при возврате исходной цели нужно пересчитывать оценки?
7. Может ли нулевая искусственная переменная остаться в базисе?
8. Когда при очистке допустим отрицательный разрешающий элемент?
9. При каком условии строка может быть удалена как избыточная?
10. Чем неограниченность исходной цели отличается от несовместности ограничений?
11. Почему фиксированное значение M не подходит для всех задач?
12. Как проверить ответ независимо от последовательности таблиц?

Литература

1. Пантелеев А. В., Летова Т. А. Методы оптимизации в примерах и задачах: учебное пособие. 4-е изд., испр. — Санкт-Петербург: Лань, 2015. — 512 с. § 11.1, с. 292–298.
2. Kochenderfer M. J., Wheeler T. A. Algorithms for Optimization. — Cambridge, MA: MIT Press, 2019. § 11.2.4, с. 202, 205–206: вспомогательная задача и поиск начального допустимого базиса.
3. Материалы курса «Методы оптимизации». Лекции 2–4: каноническая форма, базис, оценки и симплексная таблица.