МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

В.И. Рейзлин

ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

*Рекомендовано в качестве учебного пособия
Редакционно-издательским советом
Национального исследовательского
Томского политехнического университета*

Издательство
Томского политехнического университета
2011

УДК 519.6(075.8)
ББК 22.193
Р35

Рейзлин В.И.

Р35 Численные методы оптимизации: учебное пособие / В.И. Рейзлин; Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2011 – 105 с.

В пособии рассматриваются вопросы постановки задач оптимизации и численные методы их решения; одномерной и многомерной безусловной оптимизации; условной оптимизации; линейного программирования.

Предназначено для студентов, обучающихся по основной образовательной программе подготовки магистров направления 230100 «Информатика и вычислительная техника», а также может быть полезно студентам и аспирантам, применяющим в своей научной и учебной работе численные методы.

УДК 519.6(075.8)
ББК 22.193

Рецензенты

Доктор технических наук
начальник кафедры «Сети и системы связи»
ИКСИ Академии ФСБ РФ
И.А. Шалимов

Кандидат технических наук
заведующая лабораторией реологии нефти
Института химии нефти СО РАН
Н.В. Юдина

1. ВВЕДЕНИЕ

Оптимизация в широком смысле слова находит применение в науке, технике и в любой другой области человеческой деятельности.

Оптимизация – целенаправленная деятельность, заключающаяся в получении наилучших результатов при соответствующих условиях.

Поиски оптимальных решений привели к созданию специальных математических методов и уже в 18 веке были заложены математические основы оптимизации (вариационное исчисление, численные методы и др.). Однако до второй половины 20 века методы оптимизации во многих областях науки и техники применялись очень редко, поскольку практическое использование математических методов оптимизации требовало огромной вычислительной работы, которую без ЭВМ реализовать было крайне трудно, а в ряде случаев – невозможно. Особенно большие трудности возникали при решении задач оптимизации из-за большого числа параметров и их сложной взаимосвязи между собой. При наличии ЭВМ ряд задач оптимизации поддается решению.

1.1. Постановка задач оптимизации

При постановке задачи оптимизации необходимо:

1. Наличие объекта оптимизации и цели оптимизации. При этом формулировка каждой задачи оптимизации должна требовать экстремального значения лишь одной величины, то есть одновременно системе не должно приписываться два и более критерия оптимизации, так как практически всегда экстремум одного критерия не соответствует экстремуму другого.

Типичный пример неправильной постановки задачи оптимизации:

«Получить максимальную производительность при минимальной себестоимости». Ошибка заключается в том, что ставится задача поиска оптимума двух величин, противоречащих друг другу по своей сути.

Правильной постановкой задачи может быть:

а) получить максимальную производительность при заданной себестоимости;

б) получить минимальную себестоимость при заданной производительности.

В первом случае критерий оптимизации – производительность, а во втором – себестоимость.

2. Наличие ресурсов оптимизации, под которыми понимают возможность выбора значений некоторых параметров оптимизируемого объекта. Объект должен обладать определенными степенями свободы – управляющими воздействиями.

3. Возможность количественной оценки оптимизируемой величины, поскольку только в этом случае можно сравнивать эффекты от выбора тех или иных управляющих воздействий.

4. Учет ограничений.

Пример 1.

Задача о планировании выпуска продукции при ограниченных ресурсах

Нефтеперерабатывающий завод производит за месяц 1500000 л алкилата, 1200000 л крекинг-бензина и 1300000 л изопентола. В результате смешивания этих компонентов в пропорциях 1:1:1 и 3:1:2 получается бензин сорта А и Б соответственно. Стоимость 1000 л бензина сорта А и Б соответственно равна 16000 руб. и 20500 руб.

Определить месячный план производства бензина сорта А и Б, при котором стоимость выпущенной продукции будет максимальной.

Пример 2.

Задача о нахождении размеров нагруженной балки

Задана строительная конструкция, состоящая из балки А длиной L=35,6 см и жесткой опоры В. Балка А крепится на жесткой опоре В (рис. 1) с помощью сварного соединения. Балка изготавливается из стали марки 1010 и должна выдержать нагрузку F = 2721,5 кг. Размеры h, t, толщину b, ширину l сварного шва необходимо выбрать таким образом, чтобы полные затраты были минимальными.

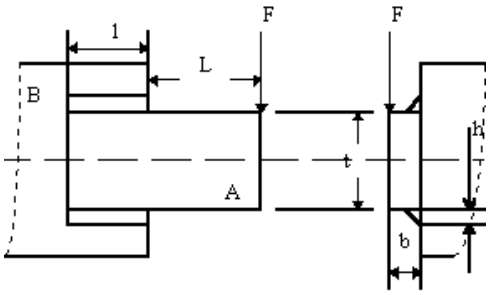


Рис. 1. *Нагруженная балка*

Пример 3.

Транспортная задача

В области имеются два цементных завода и три потребителя их продукции – домостроительные комбинаты. В табл. 1 указаны суточные объемы производства цемента, суточные потребности в нем комбинатов и стоимость перевозки 1 т цемента от каждого завода к каждому комбинату.

Таблица 1. *Числовые характеристики к транспортной задаче*

Заводы	Производство цемента (т/сут)	Стоимость перевозки 1 т цемента (ед.)		
		Комбинат 1	Комбинат 2	Комбинат 3
1	40	10	15	25
2	60	20	30	30
	Потребности в цементе (т/сут)	50	20	30

Требуется составить план суточных перевозок цемента таким образом, чтобы транспортные расходы были минимальны.

1.2. Математическая постановка задач оптимизации

1.2.1. Виды ограничений

Несмотря на то, что прикладные задачи относятся к совершенно разным областям, они имеют общую форму. Все эти задачи можно классифицировать как задачи минимизации вещественнозначной функции $f(x)$ на некотором множестве Ω N-мерного векторного аргумента $x=(x_1, x_2,..., x_n)$. Множество Ω задается ограничениями на компоненты вектора x, которые удовлетворяют системе уравнений $h_k(x)=0$, набору неравенств $g_i(x)\geq 0$, а также ограничены сверху и снизу, т.е. $x_i^{(u)} \geq x_i \geq x_i^{(l)}$.

Иногда множество Ω совпадает со всем N-мерным пространством. В этом случае задача отыскания максимума или минимума, которую также называют задачей оптимизации, называется безусловной.

Заметим, что если функция $f(x)$ имеет в точке x^* минимум, то функция $-f(x)$ в x^* имеем максимум. Поэтому для отыскания максимума применяются те же методы, что и для отыскания минимума. Далее мы, как правило, будем говорить только о задаче отыскания минимума.

Говорят, что функция $f(x)$ имеет локальный минимум x^*, если существует некоторая конечная ε-окрестность точки x^*, в которой выполняется

$$f(x^*) < f(x), \quad |x-x^*| \leq \varepsilon,\ x\in\Omega. \qquad (1.1)$$

У функции может быть много локальных минимумов. Если $f(x^*)$ – наименьший из всех минимумов, то говорят, что функция $f(x)$ достигает абсолютного минимума на множестве Ω. Этот минимум также называют глобальным.

Для нахождения абсолютного минимума надо найти все локальные минимумы, сравнить их и выбрать наименьшее значение. Поэтому задача отыскания глобального минимума сводиться к задаче (1.1), которую мы и будем рассматривать.

В последующем изложении функцию $f(x)$ будем называть ***целевой функцией***, уравнения $h_k(x)=0$ – ***ограничениями в виде равенств***, а неравенства $g_i(x)\geq 0$ – ***ограничениями в виде неравенств***. При этом предполагается, что все фигурирующие в задаче функции являются вещественнозначными, а число ограничений конечно.

Задача общего вида: минимизировать $f(x)$ (пишут $f(x)\to\min$) при ограничениях

$$h_k(x) = 0,\ k = 1,...,\ K,$$

$$g_j(x) \geq 0,\ j = 1,...,\ J,$$

$$x_i^{(u)} \geq x_i \geq x_i^{(l)},\ i = 1,...,\ N$$

называется задачей оптимизации с ***ограничениями*** или задачей ***условной*** оптимизации.

Задача, в которой нет ограничений, т.е.

$$J=K=0;$$

$$x_i^{(u)} = -x_i^{(l)} = \infty,\ i = 1,...,\ N,$$

называется оптимизационной задачей ***без ограничений*** или задачей ***безусловной*** оптимизации.

1.2.2. Критерии оптимальности

Обычно оптимизируемая величина связана с экономичностью работы рассматриваемого объекта (аппарат, цех, завод). Оптимизируемый вариант работы объекта должен оцениваться какой-то количественной мерой – ***критерием оптимальности***.

Критерием оптимальности называется количественная оценка оптимизируемого качества объекта.

На основании выбранного критерия оптимальности составляется ***целевая функция***, представляющая собой зависимость критерия оптимальности от параметров, влияющих на ее значение. Вид критерия оптимальности или целевой функции определяется конкретной задачей оптимизации. Таким образом, задача оптимизации сводится к нахождению экстремума целевой функции.

Наиболее общей постановкой оптимальной задачи является выражение критерия оптимальности в виде экономической оценки (производительность, себестоимость продукции, прибыль, рентабельность). Однако в частных задачах оптимизации, когда объект является частью технологического процесса, не всегда удается или не всегда целесообразно выделять прямой экономический показатель, который бы полностью характеризовал эффективность работы рассматриваемого объекта. В таких случаях критерием оптимальности может служить технологическая характеристика, косвенно оценивающая экономичность работы агрегата (время контакта, выход продукта, степень превращения, температура). Например, устанавливается оптимальный температурный профиль, длительность цикла «реакция–регенерация».

Рассмотрим более подробно требования, которые должны предъявляться к критерию оптимальности.

1. Критерий оптимальности должен выражаться количественно.

2. Критерий оптимальности должен быть единственным.

3. Критерий оптимальности должен отражать наиболее существенные стороны процесса.

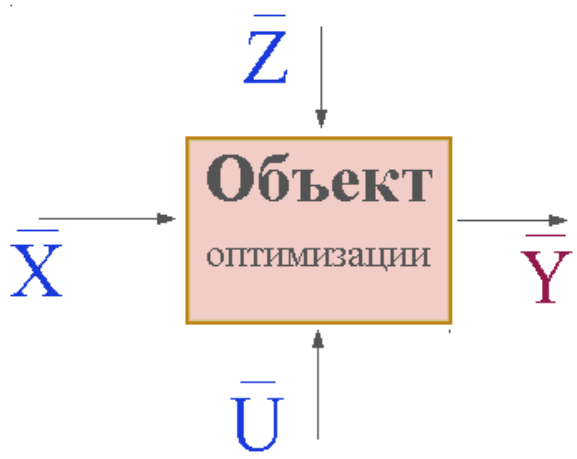
4. Желательно чтобы критерий оптимальности имел ясный физический смысл и легко рассчитывался.

Любой оптимизируемый объект схематично можно представить в соответствии с рис. 2.

При постановке конкретных задач оптимизации желательно критерий оптимальности записать в виде аналитического выражения.

В том случае, когда случайные возмущения невелики и их воздействие на объект можно не учитывать, критерий оптимальности может быть представлен как функция *входных*, *выходных* и *управляющих* параметров:

$$R = R(X_1, X_2, ..., X_N,\ Y_1,\ Y_2, ..., Y_N,\ U_1,\ U_2, ..., U_N).$$



$\bar{Y}$ – выходы объекта;
$\bar{X}$ – контролируемые входные параметры;
$\bar{U}$ – регулируемые входные параметры (управляющие параметры);
$\bar{Z}$ – неконтролируемые воздействия;

Рис. 2. Оптимизируемый объект

Так как $\bar{Y} = F(\bar{U})$, то при фиксированных $\bar{X}$ можно записать:

$$R = R(\bar{U}).$$

При этом всякое изменение значений управляющих параметров двояко сказывается на величине *R*:

- прямо, так как управляющие параметры непосредственно входят в выражение критерия оптимизации;
- косвенно – через изменение выходных параметров процесса, которые зависят от управляющих.

Если же случайные возмущения достаточно велики и их необходимо учитывать, то следует применять экспериментально-статистические методы, которые позволят получить модель объекта в виде функции, которая справедлива только для изученной локальной области и критерий оптимальности примет вид:

$$R = R(X, U).$$

В принципе, для оптимизации вместо математической модели можно использовать и сам объект, однако оптимизация опытным путем имеет ряд существенных недостатков:

а) необходим реальный объект;

б) необходимо изменять технологический режим в значительных пределах, что не всегда возможно;

в) длительность испытаний и сложность обработки данных. Наличие математической модели (при условии, что она достаточно надежно описывает процесс) позволяет значительно проще решить задачу оптимизации аналитическим либо численным методами.

В задачах оптимизации различают простые и сложные критерии оптимизации.

Критерий оптимальности называется ***простым***, если требуется определить экстремум целевой функции без задания условий на какие-либо другие величины. Такие критерии обычно используются при решении частных задач оптимизации (например, определение максимальной концентрации целевого продукта, оптимального времени пребывания реакционной смеси в аппарате и т.п.).

Критерий оптимальности называется ***сложным***, если необходимо установить экстремум целевой функции при некоторых условиях, которые накладываются на ряд других величин (например, определение максимальной производительности при заданной себестоимости, определение оптимальной температуры при ограничениях по термостойкости катализатора и др.).

Процедура решения задачи оптимизации обязательно включает, помимо выбора управляющих параметров, еще и установление ограничений на эти параметры (термостойкость, взрывобезопасность, мощность перекачивающих устройств). Ограничения могут накладываться как по технологическим, так и по экономическим соображениям.

Итак, для решения задачи оптимизации необходимо:

а) составить математическую модель объекта оптимизации

$$Y = F(X,\ U),$$

б) выбрать критерий оптимальности и составить целевую функцию

$$R = \varphi(Y) = F(X,\ U),$$

в) установить возможные ограничения, которые должны накладываться на переменные;

г) выбрать метод оптимизации, который позволит найти экстремальные значения искомых величин.

Принято различать задачи статической оптимизации для процессов, протекающих в установившихся режимах, и задачи динамической оптимизации.

В первом случае решаются вопросы создания и реализации оптимальной модели процесса, во втором – задачи создания и реализации системы оптимального управления процессом при неустановившихся режимах эксплуатации.

1.2.3. Классификация задач

Прежде всего, задачи оптимизации можно отнести по типу аргументов к дискретным (компоненты вектора x принимают дискретные или целочисленные значения) и к непрерывным (компоненты вектора x непрерывны). Для дискретных задач разработаны совершенно специфические методы оптимизации. В настоящем пособии они рассматриваться не будут.

Задачи оптимизации можно классифицировать в соответствии с видом функций f, h_k, g_i и размерностью вектора x. Задачи без ограничений, в которых x представляет собой одномерный вектор, называются ***с одной переменной*** и составляют простейший, но вместе с тем весьма важный подкласс оптимизационных задач. Задачи условной оптимизации, в которых функции h_k, и g_i являются линейными, носят название задач ***с линейными ограничениями***.

В таких задачах целевые функции могут быть либо линейными, либо нелинейными. Задачи, которые содержат только линейные функции вектора непрерывных переменных x, называются ***задачами линейного программирования***; ***в задачах целочисленного прграммирования*** компоненты вектора x должны принимать только целые значения.

Задачи с нелинейной целевой функцией и линейными ограничениями иногда называют ***задачами нелинейного программирования с линейными ограничениями***. Оптимизационные задачи такого рода можно классифицировать на основе структурных особенностей нелинейных целевых функций. Если $f(x)$ – квадратичная функция, то мы имеем дело ***с задачей квадратичного программирования***; если $f(x)$ есть отношение линейных функций, то соответствующая задача носит название задачи ***дробно-линейного программирования***, и т.д. Деление оптимизационных задач на эти классы представляет значительный интерес, поскольку специфические особенности тех или иных задач играют важную роль при разработке методов их решения.

Дерево классификации оптимизационных задач можно представить в следующей форме

Оптимизация:

Дискретная:

Целочисленное программирование;

Стохастическое программирование.

Непрерывная:

Безусловная:	Условная:
Глобальная оптимизация;	Линейное программирование;
Дифференцируемая оптимизация;	Нелинейные задачи;
Недифференцируемая оптимизация.	Стохастическое программирование.

Более подробная классификация приведена на рис. 3 (согласно источнику http://www-fp.mcs.anl.gov/otc/_vti_bin/shtml.dll/Guide/OptWeb/index.html/map).

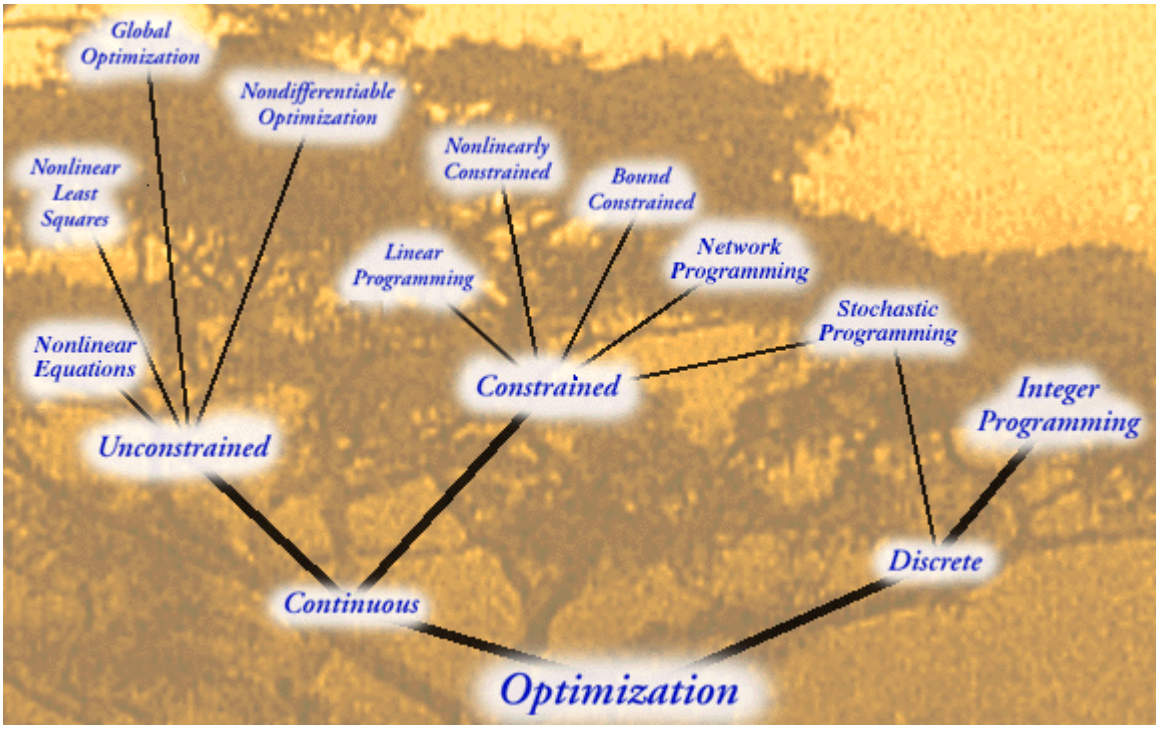


Рис. 3. *Дерево классификации оптимизационных задач*

2. ОДНОМЕРНАЯ ОПТИМИЗАЦИЯ

Оптимизация функции одной переменной – наиболее простой тип оптимизационных задач. Тем не менее, она занимает важное место в теории оптимизации. Это связано с тем, что задачи однопараметрической оптимизации достаточно часто встречаются в инженерной практике и, кроме того, находят свое применение при реализации более сложных итерактивных процедур многопараметрической оптимизации.

Разработано большое количество методов одномерной оптимизации и мы рассмотрим две группы таких методов:

- методы сужения интервала неопределенности;
- методы с использованием производных.

2.1. Методы сужения интервала неопределенности

Пусть требуется найти минимум функции $f(x)$ на некотором интервале $[a, b]$. Задача приближенного отыскания минимума в методах сужения интервала неопределенности состоит в том, чтобы найти множество абсцисс $x_1, x_2, ..., x_k$, в которых вычисляется функция, такое, что минимальное значение f^* лежит при некотором i в интервале $x_{i-1} \le x^* \le x_i$. Такой интервал называется интервалом неопределенности D. Очевидно, что сначала интервал неопределенности D совпадает с отрезком $[a, b]$.

Существуют несколько способов систематического сужения интервала неопределенности. Рассмотрим три из них.

2.1.1. Общий поиск

Пусть требуется найти минимум функции $f(x)$ на некотором интервале $[a, b]$. Если о функции $f(x)$ на этом интервале никакой дополнительной информации неизвестно, то для поиска минимума на $[a, b]$ можно применить простейший метод перебора, или, иначе, общего поиска.

В этом методе интервал $[a, b]$ делится на несколько равных частей с последующим вычислением значений функции в узлах полученной сетки. В качестве минимума принимается абсцисса с минимальным вычисленным значением функции (рис. 4).

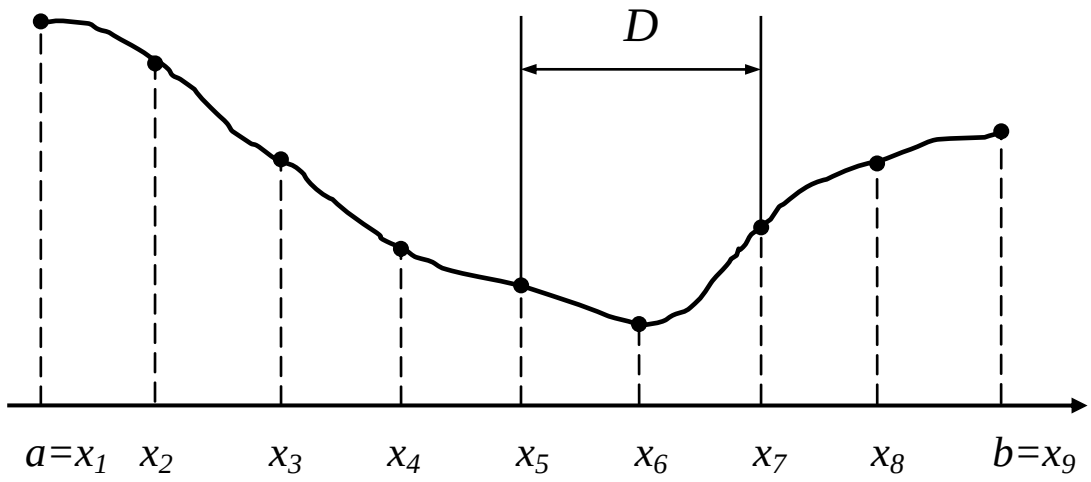


Рис. 4. *Иллюстрация к метлду общего поиска*

В результате интервал неопределенности сужается до двух шагов сетки.

Обычно говорят о дроблении интервала неопределенности, которое характеризуется коэффициентом α. Разделив интервал неопределенности на n равных частей, получим $n+1$ узел. Тогда $\alpha = \frac{2}{n}$. При этом необходимо вычислить функцию $N = n+1$ раза. Следовательно,

$$\alpha = \frac{2}{N-1},\ N = 3,4,5... \qquad (2.1)$$

Чтобы получить значение $\alpha = 0.01$ потребуется вычислить функцию в 201 точке, а при α=0.001 N=2001. Ясно, что эффективность этого метода с уменьшением интервала неопределенности быстро падает.

2.1.2. Унимодальные функции

Более эффективные методы можно построить, если предположить, что исследуемая функция имеет в рассматриваемом интервале только один минимум. Более точно: предположим, что в интервале $[a, b]$ имеется единственное значение x^* такое, что $f(x^*)$ – минимум $f(x)$ на $[a, b]$ и что $f(x)$ строго убывает для $x \le x^*$ и строго возрастает для $x \ge x^*$ (рис. 5). Такая функция называется унимодальной.

Для ее графика имеются три различные формы:

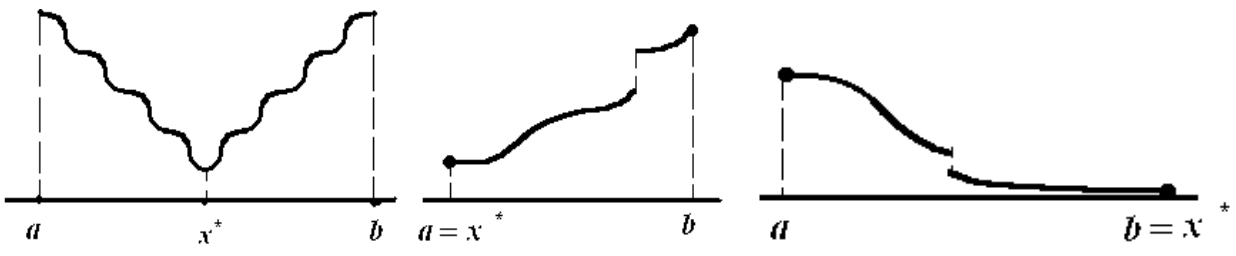


Рис. 5. *Унимодальные функции*

Заметим, что унимодальная функция не обязана быть гладкой или даже непрерывной.

Из предположения унимодальности следует, что для любых точек x_1, x_2 интервала $[a, b]$, таких, что $x_1 < x_2 \le x^*$ справедливо $f(x_2) < f(x_1)$. Аналогично, если $x^* \le x_1 < x_2$, то $f(x_2) > f(x_1)$. Обратно, если $x_1 < x_2$ и $f(x_1) > f(x_2)$, то $x_1 \le x^* \le b$, а если $f(x_1) < f(x_2)$, то $a \le x^* \le x_2$. Далее будем считать исследуемую функцию унимодальной.

2.1.3. Метод деления интервала пополам

Разделим интервал $[a, b]$ на две равные части (рис. 6), а затем каждую из частей еще на две равные части.

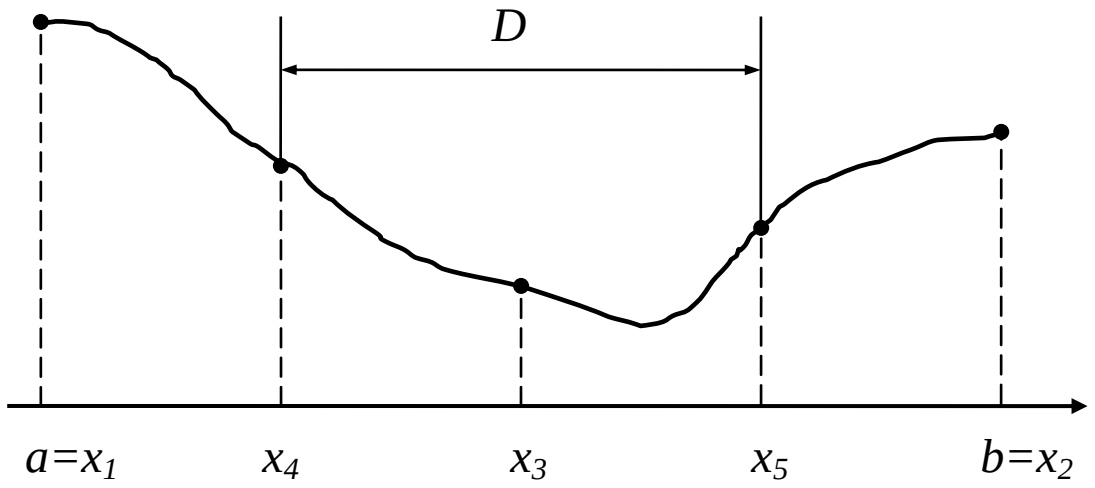


Рис. 6. *Иллюстрация к методу половинного деления*

Это первый этап поиска минимума. На нем после пяти вычислений функции (два – на краях и три – внутри интервала $[a, b]$) интервал неопределенности сужается вдвое, то есть на этом этапе $\alpha = 0{,}5$. Новый интервал неопределенности $[x_4, x_5]$ снова разделим пополам, а затем каждую половину снова пополам. Теперь значения функции по краям и в его середине уже известны. Поэтому для завершения поиска на этом этапе требуется вычислить только два значения функции, после чего интервал неопределенности снова уменьшится вдвое. Это преимущество рассмотренного метода сохранится и в дальнейшем.

После N вычислений функции коэффициент дробления интервала составляет

$$\alpha = (0{,}5)^{\frac{N-3}{2}},\ N = 5,7,9,... \qquad (2.2)$$

Здесь N=5,7,9,…, так как интервал неопределенности, начиная со второго этапа, уменьшается только после двух вычислений функции.

Из (2.1),(2.2) видно, что метод деления пополам эффективнее, чем общий поиск.

2.1.4. Метод золотого сечения

Деление интервала на неравные части позволяет найти еще более эффективный метод. Вычислим функцию на концах отрезка $[a, b]$ и положим $a = x_1,\ b = x_2$. Вычислим также функцию в двух внутренних точках $x_3,\ x_4$. Сравним все четыре значения функции и выберем среди них наименьшее (рис.°7). Пусть, например, наименьшим оказалось $f(x_3)$. Очевидно, минимум находиться в одном из прилегающих к нему отрезков. Поэтому отрезок $[x_4,\ b]$ можно отбросить и оставить отрезок $[a,\ x_4]$.

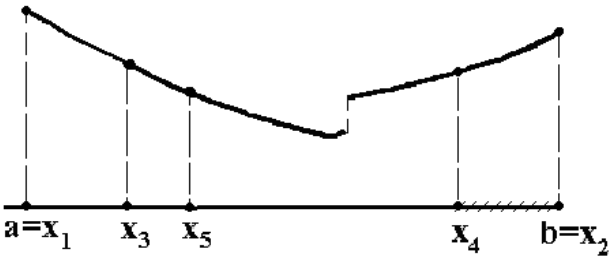


Рис. 7. *Иллюстрация к методу золотого сечения*

Первый шаг сделан. На отрезке $[a,\ x_4]$ снова надо выбрать две внутренние точки, вычислив в них и на концах значения функции и сделать следующий шаг. Но на предыдущем шаге вычислений мы уже нашли функцию на концах нового отрезка $[a,\ x_4]$ и в одной его внутренней точке x_3. Потому достаточно выбрать внутри $[a,\ x_4]$ еще одну точку x_5, определить в ней значение функции и провести необходимые сравнения. Это вчетверо уменьшает объем вычислений на одном шаге процесса. Как выгодно размещать точки? Каждый раз оставшийся отрезок делиться на три части и затем отбрасывается один из крайних отрезков.

Обозначим первоначальный интервал неопределенности через *D* (рис. 8).

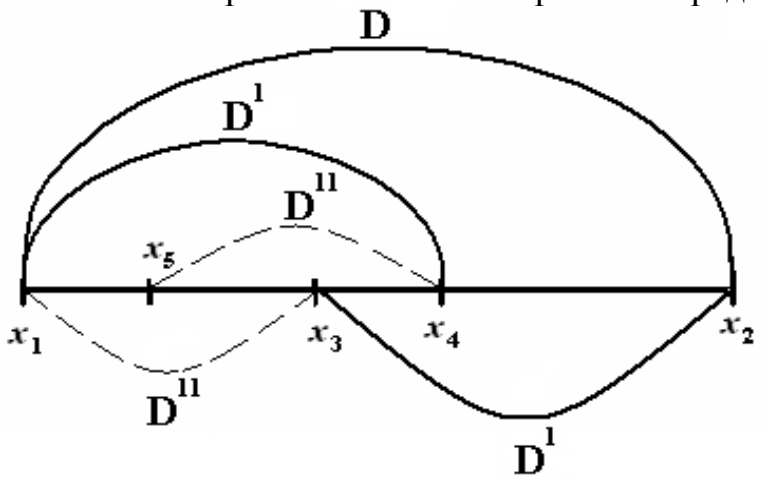


Рис. 8. *Деление интервала в методе золотого сечения*

Так как в общем случае может быть отброшен любой из отрезков x_1x_3 или x_4x_2, то выберем точки x_3 и x_4 так, чтобы длины этих отрезков были одинаковы: $x_3 - x_1 = x_2 - x_4$. После отбрасывания получится новый интервал неопределенности длины D'.

Обозначим отношение $\frac{D}{D'}$ буквой φ:

$$\varphi = \frac{D}{D'}.$$

Далее продолжим процесс аналогично. Для этого интервал D' разделим подобно интервалу D, то есть положим $\frac{D'}{D''} = \frac{D}{D'} = \varphi$, где D'' – следующий интервал неопределенности (иногда такое деление интервала называют делением в крайнем и среднем отношении: весь интервал относится к бóльшей части, как бóльшая часть к меньшей). Но D'' по длине равен отрезку, отброшенному на предыдущем этапе, то есть $D'' = D - D'$. Поэтому получим:

$$\frac{D}{D'} = \frac{D'}{D - D'} \Rightarrow \frac{D'}{D} = \frac{D}{D'} - 1.$$

Это приводит к уравнению $\frac{1}{\varphi} = \varphi - 1$ или, что то же

$$\varphi^2 - \varphi - 1 = 0.$$

Положительный корень этого уравнения дает

$$\varphi = \frac{\sqrt{5}+1}{2} \approx 1,6180.$$

Последнее число известно в математике как золотое отношение, а описанное деление отрезка, как золотое сечение. Потому рассматриваемый метод поиска минимума называют методом золотого сечения. Отношение $\frac{D}{D'} = \varphi \approx 1,618$ показывает, во сколько раз сокращается интервал неопределенности при одном добавочном вычислении функции. Учтем, что первые три вычисления еще не сокращают интервал неопределенности. Поэтому после N вычислений функции коэффициент дробления будет

$$\alpha = \left(\frac{1}{\varphi}\right)^{N-3} \approx (0,6180)^{N-3}. \qquad (2.3)$$

При $N \to \infty$ длина интервала неопределенности стремиться к нулю как геометрическая прогрессия со знаменателем $\frac{1}{\varphi}$, то есть метод золотого сечения всегда сходится. Очевидно, этот метод более эффективен, чем метод деления пополам, так как после N вычислений функции длина интервала неопределенности уменьшается при золотом сечении в $\varphi^{N-3} \approx (1,6180)^{N-3}$ раз, а в методе деления пополам в $2^{\frac{n-3}{2}} \approx (1,4142)^{N-3}$ раза.

Приведем теперь вычислительную схему метода. Имеем $\frac{D}{D'} = \varphi$, причем $D = x_2 - x_1$, $D' = x_4 - x_1$ или $x_2 - x_3$.

Поэтому $\frac{x_2 - x_1}{x_2 - x_3} = \varphi$, $\frac{x_2 - x_1}{x_4 - x_1} = \varphi$,

что дает

$$x_3 = x_2 - \frac{1}{4}(x_2 - x_1) = x_2 - \frac{\sqrt{5}-1}{2}(x_2 - x_1) \approx x_2 - 0,6180(x_2 - x_1), \qquad (2.4)$$

$$x_4 = x_1 + \frac{1}{4}(x_2 - x_1) = x_1 + \frac{\sqrt{5}-1}{2}(x_2 - x_1) \approx x_1 + 0,6180(x_2 - x_1). \qquad (2.5)$$

Так как длины отрезков x_1x_3 и x_4x_2 равны, последнее равенство можно переписать следующим образом:

$$x_4 = x_1 + x_2 - x_3. \qquad (2.6)$$

После сравнения может быть отброшена точка с любым номером, так что на следующих шагах оставшиеся точки будут перенумерованы беспорядочно. Пусть на данном отрезке есть четыре точки, x_i, x_j, x_k, x_l, из которых какие-то две являются концами отрезка.

Выберем ту точку, в которой функция принимает наименьшее значение; пусть это оказалась точка x_i:

$$f(x_i) < f(x_j),\ f(x_k),\ f(x_l) \qquad (2.7)$$

Затем отбрасываем ту точку, которая более удалена от x_i (это верно в методе золотого сечения). Пусть этой точкой оказалась x_l:

$$|x_l - x_i| > |x_j - x_i|,\ |x_k - x_i|.$$

Определим порядок распределения оставшихся трех точек на числовой оси; пусть, например, $x_k < x_i < x_j$. Пронумеруем эти точки, положив k=1, j=2, i=3. Тогда новую внутреннюю точку введем по формуле (2.6):

$$x_4 = x_1 + x_2 - x_3.$$

Ее номер теперь – 4.

Вычислим функцию $f(x_4)$ в этой точке. Выполним сравнение, отбросим одну точку, заново переименуем точки, введем новую точку по формуле (2.6) и т.д.

Минимум находится где-то внутри последнего отрезка: $x^* \in [x_1, x_2]$. Поэтому процесс прекращается, когда длина этого интервала неопределенности станет меньше заданной погрешности: $x_2 - x_1 < \varepsilon$.

Заметим, что если на [a, b] функция имеет несколько минимумов, то процесс сойдется к одному из них, но не обязательно к наименьшему.

Приведем таблицу сравнения методов поиска минимума по значениям коэффициента дробления интервала неопределенности после N вычислений функции:

N	Коэффициент дробления α		
	Общий поиск	Деление пополам	Золотое сечение
3	1	1	1
4	0,667	-	0,618
5	0,500	0,500	0,382
6	0,400	-	0,250
7	0,333	0,250	0,146
8	0,286	-	0,090
9	0,250	0,125	0,056
10	0,222	-	0,0345
19	0,111	0,00391	0,000453
20	0,105	-	0,000280
21	0,100	0,00195	0,000173

2.1.5. Установление первоначального интервала неопределенности

Рассмотренные выше методы поиска минимума, которые позволяют определить оптимум функции одной переменной путем уменьшения интервала поиска, носят название методов исключения интервалов.

Процесс применения методов поиска на основе исключения интервалов включает два этапа:

- этап установления границ интервала;
- этап уменьшения интервала.

Способы уменьшения интервала мы уже рассмотрели. Рассмотрим теперь этап установления границ интервала. Обычно используется эвристический метод, например, метод Свенна.

Итак, пусть требуется найти минимум функции $f(x)$ не на отрезке, а на всей оси x. Предположим снова, что функция $f(x)$ унимодальна. Выберем некоторое начальное приближение x_0, и сделаем из него шаг некоторой длины h: $x_1 = x_0 + h$ (рис. 9) Если $f(x_1)$ окажется большее, чем $f(x_0)$, то изменим направление шага и положим $x_1 = x_0 + h$. Пусть теперь $f(x_1) < f(x_0)$. Удвоим шаг

$h' = 2h$ и положим $x_2 = x_1 + h'$ и т.д., до тех пор, пока на некотором шаге не будет выполнено условие $f(x_n) > f(x_{n-1})$.

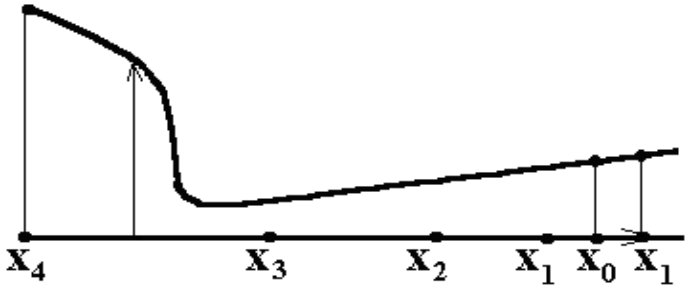


Рис. 9. *Иллюстрация к методу Свенна*

Теперь ясно, что минимум унимодальной функции лежит на отрезке $[x_4,\ x_3]$ и его можно найти одним из рассмотренных методов.

Главное достоинство поисковых методов состоит в том, что они основаны на вычислении только значений функции и, следовательно, не требуют выполнения условия дифференцируемости и записи целевой функции в аналитическом виде. Последнее свойство особенно ценно при имитационном моделировании.

Однако это достоинство поисковых методов является и их недостатком – скорость их сходимости невелика.

Заметим, что если функция $f(x)$ достаточно гладкая, например, имеет непрерывные первую и вторую производные, то скорость сходимости численных методов можно увеличить, применяя методы с использованием производных.

Рассмотренные методы оптимизации используют только значения функции $f(x)$. Такие методы называются методами 0-го порядка. Если предположить, что функция $f(x)$ дифференцируема, то можно предложить более быстрые методы, использующие производные. Методы, использующие ***первую*** производную, называются методами 1-го порядка и т. д.

2.2. Ньютоновские методы

Пусть функция $f(x)$ дважды дифференцируема. Как известно из математического анализа, условием минимума такой функции является равенство

$$f'(x^*) = 0. \qquad (2.8)$$

Это необходимое условие. Однако для того чтобы точка x^* была минимумом, должно также выполняться достаточное условие

$$f''(x^*) > 0. \qquad (2.9)$$

Итак, будем численно решать уравнение

$$f'(x) = 0. \qquad (2.10)$$

Зададим некоторое начальное приближение x_k и разложим в этой точке функцию в ряд Тейлора. Ограничимся лишь членами до второго порядка включительно, т.е. построим квадратичную модель функции:

$$\hat{f}(x) \approx f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2} f''(x_k)(x - x_k)^2. \qquad (2.11)$$

Если $f''(x_k) \neq 0$, $\hat{f}(x)$ будет иметь единственную стационарную точку. Найдем ее, для чего приравняем нулю производную $\hat{f}'(x)$:

$$\hat{f}'(x) = f'(x_k) + f''(x_k)(x - x_k) = 0.$$

Решим это уравнение относительно *x* и найденное решение примем за очередное, *k*+1-ое приближение к минимуму:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}. \qquad (2.12)$$

Формулу (2.12) можно получить иначе, если применить численный метод решения уравнения $g(x) = 0$, известный, как метод Ньютона:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}.$$

Надо только учесть, что теперь мы решаем уравнение $f'(x) = 0$, то есть положить $g(x) = f'(x)$.

У алгоритма (2.12) есть два недостатка. Во-первых, уравнение $f'(x) = 0$ может определять не только минимум, но и максимум. Во-вторых, модельная функция $\hat{f}(x)$ может сильно отличаться от оптимизируемой функции $f(x)$ и шаг $x_{k+1} - x_k$ может оказаться слишком большим (рис. 10):

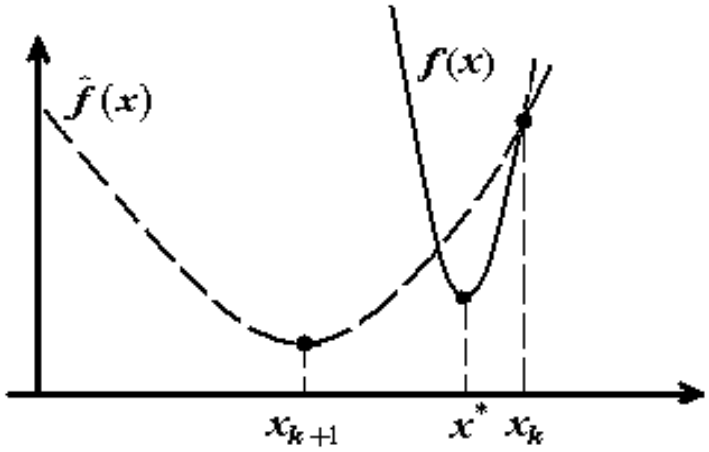


Рис. 10. *Иллюстрация к методу Ньютона*

Поэтому стратегию (2.12) следует уточнить.

Чтобы быть уверенными, что мы продвигаемся к минимуму, будем на каждом шаге проверять соотношение $f(x_{k+1}) < f(x_k)$. Если оно выполняется, то переходим к следующему шагу и т.д. Если же $f(x_{k+1}) > f(x_k)$, а $f'(x_k)(x_{k+1} - x_k) < 0$, то функция $f(x)$ должна первоначально уменьшаться в

направлении от x_k к x_{k+1}, поэтому следующую приемлемую точку можно найти, дробя шаг в обратном направлении, например, положив

$$x'_{k+1} = \frac{x_{k+1} + x_k}{2}.$$

Из формулы (2.12) видно, что выражение $f'(x_k)(x_{k+1} - x_k)$ отрицательно тогда и только тогда, когда $f''(x_k)$ положительна. Это означает, что если локальная модель, используемая для получения Ньютоновского шага, имеет минимум, а не максимум, то гарантируется существование подходящего направления шага. С другой стороны, если $f''(x_k) < 0$ и $f'(x_k)(x - x_k) > 0$, то при переходе от x_k к x_{k+1}, $f(x)$ первоначально увеличивается, поэтому шаг нужно сделать в противоположном направлении.

Критерий прекращения итераций при оптимизации можно выбрать в виде

$$\left| \frac{f'(x_{k+1})}{f(x_{k+1})} \right| < \varepsilon, \qquad (2.13)$$

где ε – заранее заданная точность.

Описанный метод с основным шагом (2.12) и приведенными уточнениями обычно называют методом Ньютона или ***Ньютона-Рафсона***.

В некоторых задачах производные функции $f(x)$ недоступны и метод Ньютона можно модифицировать.

Выберем начальное приближение x_k и малый шаг h. Рассмотрим три точки $x_k - h$, x_k, $x_k + h$. Тогда производные $f'(x_k)$ и $f''(x_k)$ можно аппроксимировать следующим образом:

$$f'(x_k) = \frac{f(x_k + h) - f(x_k - h)}{2h},$$

$$f''(x_k) = \frac{f'(x_k + h) - f'(x_k - h)}{2h} = \frac{\dfrac{f(x_k + h) - f(x_k)}{h} - \dfrac{f(x_k) - f(x_k - h)}{h}}{2h} =$$

$$= \frac{f(x_k + h) - 2f(x_k) + f(x_k - h)}{2h^2}.$$

Подставляя это в алгоритм (2.12), найдем:

$$x_{k+1} = x_k - h \frac{f(x_k + h) - f(x_k - h)}{f(x_k + h) - 2f(x_k) + f(x - h)}. \qquad (2.14)$$

Формула (2.14) дает основной шаг алгоритма, называемого квазиньютоновским методом или модифицированным методом Ньютона. Все соображения относительно шага $x_{k+1} - x_k$, приводимые при выводе метода ***Ньютона-Рафсона***, остаются в силе.

3. МИНИМУМ ФУНКЦИИ МНОГИХ ПЕРЕМЕННЫХ

3.1. Рельеф функции

Понятие «рельеф функции» удобно рассмотреть на примере функции двух переменных $z = F(x, y)$. Эта функция описывает некоторую поверхность в трехмерном пространстве с координатами *x*, *y*, *z*. Задача $F(x, y) \to \min$ означает поиск низшей точки этой поверхности.

Как в топографии, изобразим рельеф этой поверхности *линиями уровня*. Проведем равноотстоящие плоскости *z*=const и найдем линии их пересечения с поверхностью $F(x, y)$. Проекции этих линий на плоскость *x*, *y* называют ***линиями уровня***. Направление убывания функции будем указывать штрихами рядом с линиями уровня. По виду линий уровня условно выделим три типа рельефа: котловинный, овражный и неупорядоченный (рис. 11–15).

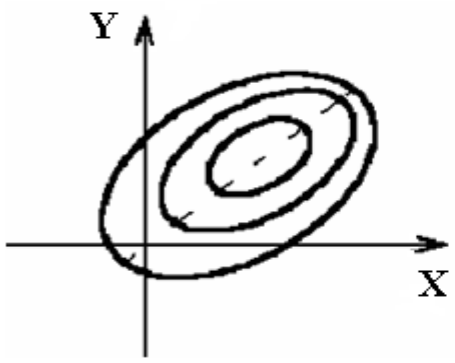


Рис. 11. *Котловинный рельеф*

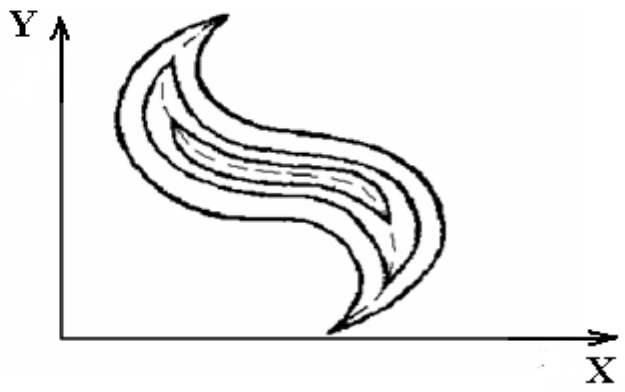


Рис. 12. *Овражный рельеф*

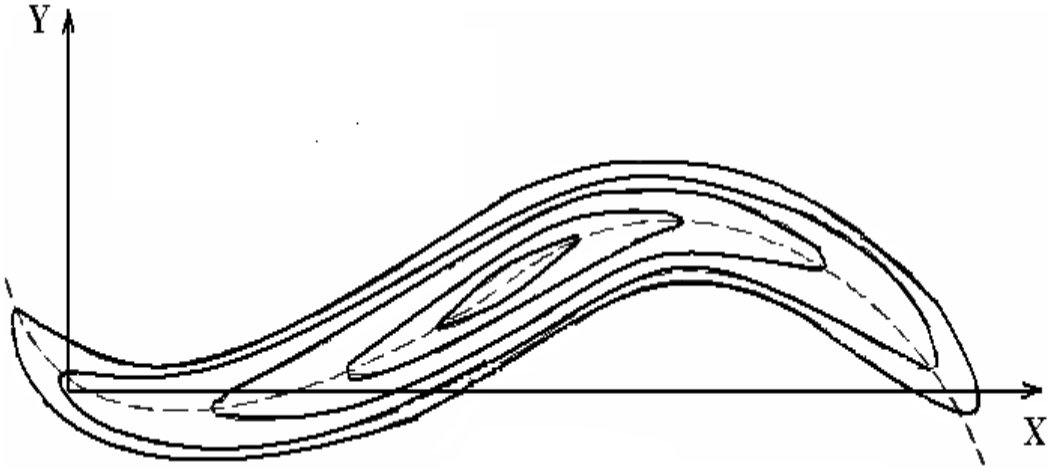


Рис. 13. *Разрешимый овраг*

При котловинном рельефе линии уровня похожи на эллипсы (рис. 11).

Рассмотрим овражный тип рельефа. Если линии уровня кусочно-гладкие (рис. 12), то выделим на каждой из них точку излома. Геометрическое место то-

чек излома назовем ***истинным оврагом***, если угол направлен в сторону возрастания функции, и ***гребнем***, – если в сторону убывания.

Чаще линии уровня всюду гладкие, но на них имеются участки с большой кривизной. Геометрические места точек с наибольшей кривизной назовем ***разрешимым оврагом*** или ***гребнем*** (рис. 13).

Например, рельеф функции $F(x,y)=10(y-\sin\ x)^2+0.1x^2$ (рис. 14) имеет ярко выраженный извилистый разрешимый овраг, «дно» которого – синусоида, а низшая точка – начало координат.

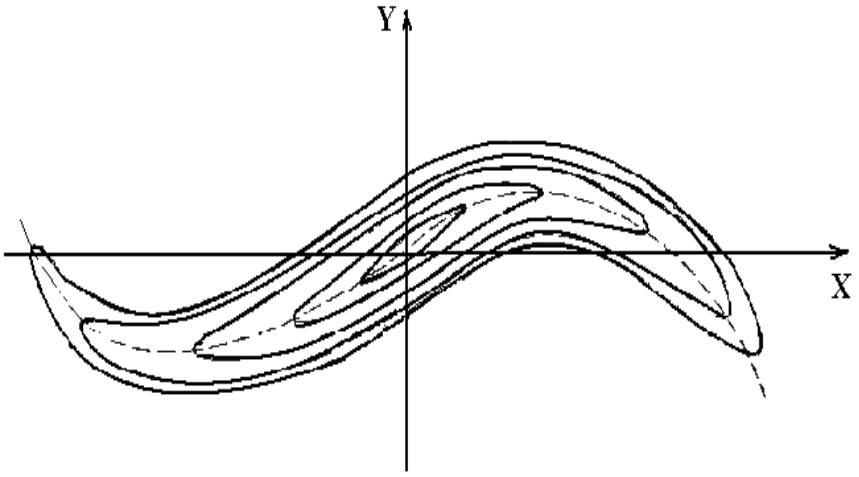


Рис. 14. *Рельеф функции* $F(x,y)=10(y-\sin\ x)^2+0.1x^2$

Неупорядоченный тип рельефа характеризуется наличием многих максимумов и минимумов. Так, функция $F(x,y)=(1+\sin^2\ x)(1+\sin^2\ y)$ (рис. 15) имеет минимумы в точках $x_k^*=\pi k,\ y_l^*=\pi l$ и максимумы в точках, сдвинутых относительно минимумов на $\frac{\pi}{2}$ по каждой координате.

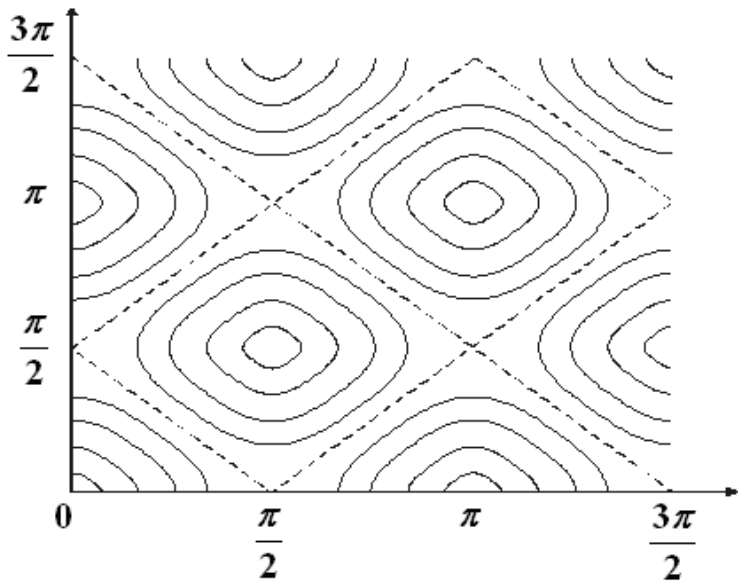


Рис. 15. *Рельеф функции* $F(x,y)=(1+\sin^2\ x)(1+\sin^2\ y)$

Все эффективные методы поиска минимума сводятся к построению траекторий, вдоль которых функция убывает; разные методы отличаются способами построения таких траекторий. Метод, приспособленный к одному типу рельефа, может оказаться плохим на рельефе другого типа.

3.2. Метод покоординатного спуска (Метод Гаусса)

Изложим этот метод на примере функции трех переменных $F(x, y, z)$.

Выберем нулевое приближение x_0, y_0, z_0. Фиксируем значение двух координат $y = y_0$, $z = z_0$. Тогда функция будет зависеть только от одной переменной x; обозначим ее через $f_1(x) = F(x, y_0, z_0)$. Используя какой-либо способ одномерной оптимизации, отыщем минимум функции $f_1(x)$ и обозначим его через x_1. Мы сделали шаг из точки (x_0, y_0, z_0) в точку (x_1, y_0, z_0) по направлению, параллельному оси x; на этом шаге значение функции уменьшилось.

Теперь из новой точки сделаем спуск по направлению, параллельному оси y, то есть рассмотрим функцию $f_2(y) = F(x_1, y, z_0)$, найдем ее минимум и обозначим его через y_1. Второй шаг приводит нас в точку (x_1, y_1, z_0). Из этой точки делаем третий шаг – спуск параллельно оси z и находим минимум функции $f_3(z) = F(x_1, y_1, z)$. Приход в точку (x_1, y_1, z_1) завершает цикл спусков или первую итерацию.

Далее будем повторять циклы. На каждом спуске функция не возрастает, и при этом значение функции ограничено снизу ее значением в минимуме $F^* = F(x^*, y^*, z^*)$. Следовательно, итерации сходятся к некоторому пределу $\tilde{F} \geq F^*$. Будет ли здесь иметь место равенство, то есть сойдутся ли спуски к минимуму и как быстро? Это зависит от функции и выбора нулевого приближения.

На примере функции двух переменных легко убедиться, что существуют случаи сходимости спуска по координатам к минимуму и случаи, когда этот спуск к минимуму не сходиться.

В самом деле, рассмотрим геометрическую трактовку этого метода (рис.°16 и 17):

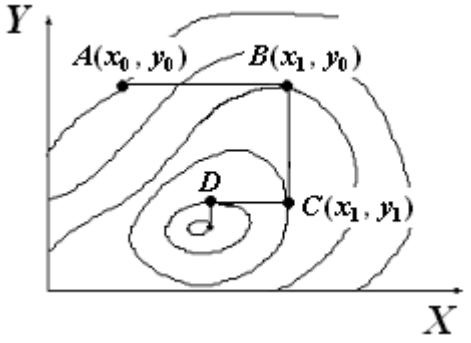


Рис. 16. *Метод Гаусса для котловинного рельефа*

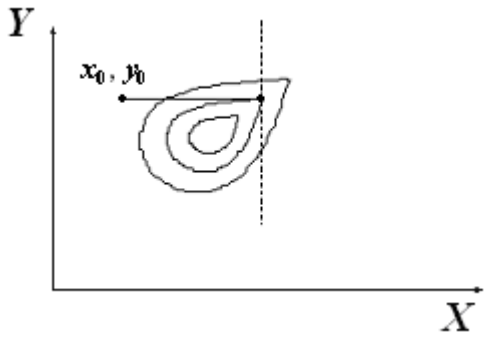


Рис. 17. *Метод Гаусса для истинного оврага*

Будем двигаться по выбранному направлению, то есть вдоль некоторой прямой в плоскости x, y.

В тех участках, где прямая пересекает линии уровня, мы при движении переходим от одной линии уровня к другой, так что при этом движении функция меняется (возрастает или убывает, в зависимости от направления движения). Только в той точке, где данная прямая касается линии уровня (рис. 16), функция имеет экстремум вдоль этого направления. Найдя такую точку, мы завершаем в ней спуск по первому направлению и должны начать спуск по второму.

Пусть линии уровня образуют истинный овраг. Тогда возможен случай (рис. 17), когда спуск по одной координате приводит нас на «дно оврага», а любое движение по следующей координате (пунктирная линия) ведет нас на подъем. Никакой дальнейший спуск по координатам невозможен, хотя минимум еще не достигнут. В данном случае процесс спуска по координатам не сходится к минимуму.

Наоборот, если функция достаточно гладкая, то в некоторой окрестности минимума процесс спуска по координатам сходится к этому минимуму. Однако скорость сходимости сильно зависит от формы линий уровня. Так, если рельеф функции имеет тип «разрешимый овраг», то при попадании траектории спуска в такой овраг сходимость становится настолько медленной, что расчет практически вести невозможно.

Обычно метод покоординатного спуска используют в качестве первой попытки при нахождении минимума.

3.3. Метод оврагов

Выберем произвольную точку ρ_0 и спустимся из нее (например, по координатам), делая не очень много шагов, то есть, не требуя высокой точности. Конечную точку спуска обозначим r_0. Если рельеф овражный, эта точка окажется вблизи дна оврага (рис. 18).

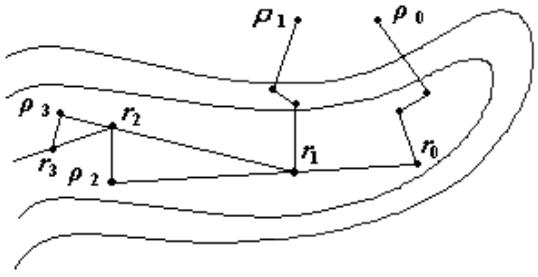


Рис. 18. *Иллюстрация к методу оврагов*

Теперь выберем другую точку ρ_1 не слишком далеко от первой. Из нее также сделаем спуск и попадем в некоторую точку r_1. Эта точка тоже лежит вблизи дна оврага. Проведем через точки r_0 и r_1 на дне оврага прямую – приблизительную линию дна оврага, передвинемся по этой линии в сторону убывания функции и выберем новую точку ρ_2 на этой прямой, на расстоянии h от точки r_1. Величина h называется овражным шагом и для каждой функции подбирается в ходе расчета.

Дно оврага не является отрезком прямой, поэтому точка ρ_2 на самом деле лежит не на дне оврага, а на его склоне. Из этой точки снова спустимся на дно и попадем в некоторую точку r_2. Затем соединим точки r_1 и r_2 прямой, наметим новую линию дна оврага и сделаем новый шаг по оврагу. Продолжим процесс до тех пор, пока значения функции на дне оврага, то есть в точках r_0, r_1,..., r_n убывают.

В случае, когда $F(r_{n+1}) > F(r_n)$, процесс надо прекратить и точку r_{n+1} не использовать. Метод оврагов рассчитан на то, чтобы пройти вдоль оврага и выйти в котловину около минимума. В этой котловине значение минимума лучше уточнять другими методами.

4. МЕТОДЫ С ИСПОЛЬЗОВАНИЕМ ПРОИЗВОДНЫХ

Методы спуска и их различные модификации, методы случайного поиска (п. 5), которые используют только значения функции, называются методами 0-го порядка. Они обычно имеют весьма малую скорость сходимости. Поэтому разработан ряд методов оптимизации, которые используют первые и вторые производные целевой функции (методы 1-го и 2-го порядка).

Прежде чем рассмотреть такие методы, введем ряд обозначений и напомним некоторые определения.

Вектор n-мерного пространства R^n будем обозначать столбцом:

$$u = \begin{bmatrix} x_1 \\ \dots \\ \dots \\ x_n \end{bmatrix}; \text{ тогда } u^{T=}\left[x_1,...,x_n\right].$$

Будем говорить, что функция $f: R^n \to R$ непрерывно дифференцируема в точке $x \in R^n$, если производные $\dfrac{\partial f(x)}{\partial x_i}$, $i = 1,...,n$ существуют и непрерывны. Тогда градиент функции f в точке x определяется как:

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1},...,\frac{\partial f(x)}{\partial x_n}\right]^T. \qquad (4.1)$$

Будем говорить, что функция $f(x)$ непрерывно дифференцируема в открытой области $D \subset R^n$, если она непрерывно дифференцируема в любой точке из D.

Пусть $f: R^n \to R$ непрерывно дифференцируема на некоторой открытой выпуклой области $D \subset R^n$. Тогда для $x \in D$ и произвольного ненулевого приращения $p \in R^n$ производная по направлению $p = \left[p_1,..., p_n\right]^T$ от функции $f(x)$ в точке x, определяемая как

$$\frac{\partial f(x)}{\partial p} \equiv \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon p) - f(x)}{\varepsilon},$$

существует и равна $\nabla f(x)^T \cdot p$, где символом '•' обозначено скалярное произведение.

Иначе можно записать

$$\frac{\partial f(x)}{\partial p} = \nabla f(x)^T \cdot p = \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} p_i. \qquad (4.2)$$

Будем говорить, что

функция $f: R^n \to R$ дважды непрерывно дифференцируема в $x \in R^n$, если производные $\dfrac{\partial^2 f(x)}{\partial x_i \partial x_j}$, $1 \le i, j \le n$, существуют и непрерывны.

Гессианом (матрицей Гессе) функции f в точке x называется матрица размера $n \times n$, и ее (i, j)-й элемент равен

$$H_{ij} = \nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j},\ 1 \le i, j \le n. \qquad (4.3)$$

Пусть $f: R^n \to R$ дважды непрерывно дифференцируема в открытой области $D \subset R^n$.

Тогда для любого $x \in D$ и произвольного ненулевого приращения $p \in R^n$ вторая производная по направлению p от функции f в точке x, определяемая как

$$\frac{\partial^2 f(x)}{\partial p^2} \equiv \lim_{\varepsilon \to \infty} \frac{\dfrac{\partial f}{\partial p}(x + \varepsilon p) - \dfrac{\partial f(x)}{\partial p}(x)}{\varepsilon},$$

существует и для нее выполняется равенство:

$$\frac{\partial^2 f(x)}{\partial p^2} = p^T \nabla^2 f(x)\, p. \qquad (4.4)$$

Пусть A – действительная симметричная матрица размером $n \times n$. Будем говорить, что A положительно определена, если для любого ненулевого вектора $u \in R^n$ выполняется неравенство

$u^T A u > 0$.

Матрица A положительно полуопределена, если $u^T A u \ge 0$ для всех $u \in R^n$.

Для того чтобы точка x^* была локальной точкой минимума $f(x)$ необходимо выполнение равенства $\nabla f(x^*) = 0$. Достаточное условие, кроме того, требует положительной определенности $\nabla^2 f(x^*)$, а необходимое – по крайней мере, положительной полуопределенности $\nabla^2 f(x^*)$.

Далее будем полагать, что $f(x)$, $\nabla f(x)$, $\nabla^2 f(x)$, существуют и непрерывны.

Все описываемые ниже методы основаны на итерационной процедуре, реализуемой в соответствии с формулой

$$x^{k+1} = x^k + \lambda^k s\left(x^k\right),$$

где x^k – текущее приближение к решению x^*, λ^k – параметр, характеризующий длину шага, s^k – направление поиска в n-мерном пространстве.

Рассмотрим метод первого порядка, использующий первые производные.

4.1. Градиентные методы

Градиент функции в любой точке *x* показывает направление наибольшего локального увеличения $f(x)$. Поэтому при поиске минимума можно попробовать двигаться в направлении, противоположном градиенту в данной точке, то есть в направлении наискорейшего спуска. Такой подход приведет к итерационной формуле, описывающей ***метод градиентного спуска***:

$$x^{k+1} = x^k - \lambda^k \nabla f\left(x^k\right) \text{ или}$$

$$x^{k+1} = x^k - \lambda^k \frac{\nabla f\left(x^k\right)}{\left\|\nabla f\left(x^k\right)\right\|} = x^k - \lambda^k s^k,$$

где $\left\|\nabla f\left(x^k\right)\right\|$ – норма градиента и, соответственно, s^k – единичный вектор.

(В качестве ***нормы*** вектора *u* можно выбрать так-называемую Гауссову норму $\| u \| = \sqrt{u_1^2 + ... + u_n^2}$.)

В зависимости от выбора параметра λ траектория спуска будет существенно различаться.

При большом значении λ траектория будет представлять собой колебательный процесс, а при слишком больших λ процесс может расходиться.

При малых λ траектория будет плавной, но и процесс будет сходиться медленно.

Параметр λ^k можно принимать постоянным или выбирать различным на каждой итерации. Иногда на каждом k-ом шаге параметр λ^k выбирают, производя одномерную минимизацию вдоль направления s^k с помощью какого-либо одномерного метода. Обычно такой процесс называют ***методом наискорейшего спуска***, или методом Коши.

Если λ^k определяется в результате одномерной минимизации, то есть $\lambda^k = arg \min_{\lambda} f\left(x^k + \lambda s^k\right)$, то градиент в точке очередного приближения будет ортогонален направлению предыдущего спуска (рис. 19).

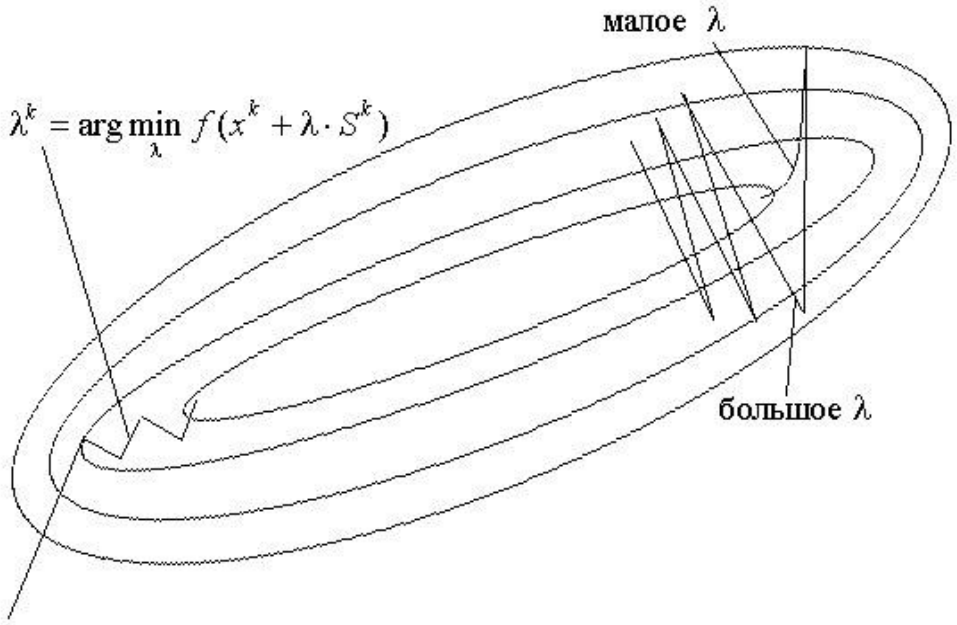


Рис. 19. *Иллюстрация к градиентным методам*

Одномерная оптимизация вдоль направления s^k улучшает сходимость метода, но одновременно возрастает число вычислений функции $f(x)$ на каждой итерации. Кроме того, вблизи экстремума норма $\left\|\nabla f\left(x\right)\right\|$ близка к нулю и сходимость здесь будет очень медленной.

Эффективность алгоритма зависит от вида минимизируемой функции. Так, для функции $f\left(x\right)=x_1^2+x_2^2$ метод Коши сойдется к минимуму за одну итерацию при любом начальном приближении, а для функции $f\left(x\right)=x_1^2+100x_2^2$ сходимость будет очень медленной.

Вообще, эффективность этого метода на овражных рельефах весьма плохая.

Этот метод используется очень редко.

4.2. Метод Ньютона

Построим квадратичную модель функции в окрестности точки x^k, разложив ее в ряд Тейлора до членов второго порядка:

$$\hat{f}\left(x^k+p\right)=f\left(x^k\right)+\nabla f^T\left(x^k\right)p+\frac{1}{2}p^T\nabla^2 f\left(x^k\right)p\,. \qquad (4.5)$$

Найдем точку $x^{k+1}=x^k+s^k$ из условия минимума квадратичной модели (4.5). Необходимым условием этого минимума будет $\nabla\hat{f}\left(x^{k+1}\right)=0$. Имеем:

$$\nabla\hat{f}\left(x^k+s^k\right)=0=\nabla f\left(x^k\right)+\nabla^2 f\left(x^k\right)s^k$$

и это приводит к следующему алгоритму:

на каждой итерации *k* решить систему уравнений

$$\nabla^2 f\left(x^k\right)s^k = -\nabla f\left(x^k\right) \qquad (4.6)$$

относительно s^k и положить

$$x^{k+1} = x^k + s^k. \qquad (4.7)$$

Алгоритм (4.6), (4.7) называется методом Ньютона. Положительной стороной метода Ньютона является то, что если x^0 достаточно близко к точке локального минимума x^* функции $f(x)$ с невырожденной матрицей Гессе $\nabla^2 f\left(x^*\right)$, то, как можно доказать, $\nabla^2 f\left(x^*\right)$ является положительно определенной и последовательность $\{x^k\}$, генерируемая алгоритмом (4.6), (4.7) будет сходиться к минимуму x^* (и скорость сходимости будет так называемой q-квадратичной).

***К недостаткам метода относятся*:**

1. Метод не сходится глобально, то есть требует достаточно хорошее начальное приближение x^0;

2. Требует аналитическое задание первых и вторых производных;

3. На каждом шаге необходимо решать систему уравнений (4.7);

4. В методе Ньютона нет ничего, что удерживало бы его от продвижения в сторону максимума или седловой точки, где ∇f тоже равен нулю.

Здесь каждый шаг направляется в сторону стационарной точки локальной квадратичной модели независимо от того, является ли стационарная точка минимумом, максимумом или седловой точкой. Этот шаг оправдан при минимизации, только когда Гессиан $\nabla^2 f\left(x^k\right)$ положительно определен.

Вообще говоря, метод Ньютона не обладает высокой надежностью.

4.3. Метод Марквардта

Этот метод является комбинацией методов градиентного спуска и Ньютона, в котором удачно сочетаются положительные свойства обоих методов. Движение в направлении антиградиента из точки x^0, расположенной на значительном расстоянии от точки минимума x^*, обычно приводит к существенному уменьшению целевой функции. С другой стороны, направление эффективного поиска в окрестности точки минимума определяются по методу Ньютона.

В соответствии с методом Марквардта, направление поиска s^k определяется равенством

$$\left(H^k + \lambda^k I\right)s^k = -\nabla f\left(x^k\right), \qquad (4.8)$$

а новая точка x^{k+1} задается формулой

$$x^{k+1} = x^k + s^k. \qquad (4.9)$$

В системе (1) I – единичная матрица, H^k – матрица Гессе, $\lambda^k \geq 0$.

В последней формуле коэффициент перед s^k взят равным 1, так как параметр λ^k в (4.8) позволяет менять и длину шага, и его направление.

На начальной стадии поиска параметру λ^0 приписывается некоторое большое значение, например 10^4, так что левая часть равенства Марквардта (4.8) при больших λ^0 примет вид

$$(H^0 + \lambda^0 I)s^k \approx (\lambda^0 I)s^k = \lambda^0 s^k. \qquad (4.10)$$

Таким образом, большим значением λ^0, как видно из (4.8) и (4.10), соответствует направление поиска

$$s^k = -\frac{1}{\lambda^0}\nabla f\left(x^k\right),$$

то есть направление наискорейшего спуска.

Из формулы (4.8) можно заключить, что при уменьшении λ^k до нуля вектор s^k изменяется от направления, противоположного градиенту, до направления, определяемому по Ньютону. Если после первого шага получена точка с меньшим значением целевой функции, то есть $f\left(x^1\right) < f\left(x^0\right)$, следует выбрать $\lambda^1 < \lambda^0$ и реализовать еще один шаг. В противном случае нужно положить $\lambda^0 = \beta\lambda^0$, где $\beta > 1$, и вновь реализовать предыдущий шаг.

Заметим, что недостатком метода Ньютона является то, что если матрица Гессе H^k не является положительно определенной, то Ньютоновский шаг s^k не приводит к убыванию функции. Поэтому «исправление» Гессиана в соответствии с формулой $H^k + \lambda^k I$ модифицирует матрицу и при соответствующем выборе λ^k делает ее положительно определенной, так как единичная матрица положительно определена.

Приведем теперь алгоритм метода:

1. Задать x^0 – начальное приближение к x^*, M – максимально допустимое количество итераций и ε – параметр сходимости.

2. Положить $k = 0$, $\lambda^0 = 10^4$.

3. Вычислить $\nabla f\left(x^k\right)$.

4. Проверить $\left\|\nabla f\left(x^k\right)\right\| < \varepsilon$?

(Можно взять $\left\|\nabla f\left(x^k\right)\right\| = \sqrt{\left(\frac{\partial f\left(x^k\right)}{\partial x_1}\right)^2 + \ldots\ldots + \left(\frac{\partial f\left(x^k\right)}{\partial x_n}\right)^2}$).

Если да, то перейти к п.11.

5. Проверить $k \geq M$? Если да, то перейти к п.11.

6. Вычислить шаг s^k, решив систему

$\left(H^k + \lambda^k I\right) \cdot s^k = -\nabla f\left(x^k\right)$.

7. Положить $x^{k+1} = x^k + s^k$.

8. Проверить: $f\left(x^{k+1}\right) < f\left(x^k\right)$?

Если да, то перейти к п.9, иначе к п.10.

9. Положить $\lambda^{k+1} = \frac{1}{2}\lambda^k$, $k = k+1$. Перейти к п.3.

10. Положить $\lambda^k = 2\lambda^k$. Перейти к п.7.

11. Вывод результатов:

x^k, $f\left(x^k\right)$, $\nabla f\left(x^k\right)$, $\left\|\nabla f\left(x^k\right)\right\|$, k.

Отметим, что в различных модификациях метода Ньютона требуется большое количество вычислений, так как на каждой итерации следует сначала вычислить элементы матрицы $n \times n$, а затем решать систему линейных уравнений. Применение конечно разностной аппроксимации первых и вторых производных только ухудшит ситуацию.

Поэтому в последнее время построено много так называемых квазиньютоновских методов, как с аналитическим вычислением градиента и матрицы Гессе, так и с их конечно разностной аппроксимацией. Эти методы опираются на возможность аппроксимации кривизны нелинейной целевой функции без явного формирования ее матрицы Гессе. Данные о кривизне накапливаются на основе наблюдения за изменением градиента во время спуска.

Различные формы таких методов, часто называемые методами секущих, показали неплохие результаты в научных и технических исследованиях.

5. УСЛОВНАЯ ОПТИМИЗАЦИЯ

Ряд инженерных задач связан с оптимизацией при наличии некоторого количества ограничений на управляемые переменные. Такие ограничения существенно уменьшают размеры области, в которой ищется оптимум. На первый взгляд может показаться, что уменьшение размеров допустимой области должно упростить процедуру поиска оптимума. Однако, напротив, процесс оптимизации становится более сложным, поскольку при наличии ограничений даже нельзя использовать применяемые нами выше условия оптимальности. При этом может нарушаться даже основное условие, в соответствии с которым оптимум должен достигаться в стационарной точке, характеризующейся нулевым градиентом. Например, безусловный минимум функции $f(x)=(x-2)^2$ имеет место в стационарной точке $x=2$. Но если задача минимизации решается с учетом ограничения $x\geq 4$, то будет найден ***условный минимум***, которому соответствует точка $x=4$. Эта точка не является стационарной точкой функции $f(x)$, так как $f'(4)=4$. Поэтому нужно изучить необходимые и достаточные условия оптимума в задачах с ограничениями, которые иначе называют задачами условной оптимизации.

5.1. Задачи с ограничениями в виде равенств

Рассмотрим задачу:

$$f(x)\to \min,\ x\in R^n$$

при ограничениях

$$h_k(x)=0,\ k=1,2,...,K.$$

Одним из методов ее решения является метод множителей Лагранжа.

5.1.1. Множители Лагранжа

С помощью метода множителей Лагранжа по существу устанавливаются необходимые условия, позволяющие идентифицировать точки оптимума в задачах оптимизации с ограничениями-равенствами. При этом задача с ограничениями преобразуется в эквивалентную задачу безусловной оптимизации, в которой фигурируют некоторые неизвестные параметры, называемые ***множителями Лагранжа***.

Рассмотрим задачу с одним ограничением-равенством:

$$f(x)\to \min,\ x\in R^n, \tag{5.1}$$

$$h_1(x)=0. \tag{5.2}$$

В соответствии с методом множителей Лагранжа эта задача преобразуется в следующую задачу безусловной минимизации:

$$L(x;\lambda) = f(x) - \lambda h_1(x) \to \min,\ x \in R^n. \qquad (5.3)$$

Функция $L(x;\lambda)$ называется функцией Лагранжа. Здесь λ – множитель Лагранжа.

Пусть при заданном значении $\lambda = \lambda^0$ безусловный минимум функции $L(x;\lambda)$ по переменной x достигается в точке $x = x^0$ и x^0 удовлетворяет уравнению

$$h_1(x^0) = 0.$$

Тогда, как не трудно видеть, x^0 минимизирует (5.1) с учетом (5.2), поскольку для всех значений x, удовлетворяющих (5.2), $h_1(x) = 0$ и $\min L(x;\lambda) = \min f(x)$.

Разумеется, нужно подобрать значение $\lambda = \lambda^0$ таким образом, чтобы координата точки безусловного минимума x^0 удовлетворяла равенству (5.2). Это можно сделать, если, рассматривая λ как переменную, найти безусловный минимум функции Лагранжа (5.3) в виде функции λ, а затем выбрать значение λ, при котором выполняется равенство (5.2).

Пример.

Решить задачу

$$f(x) = x_1^2 + x_2^2 \to \min$$

при ограничении

$$h_1(x) = 2x_1 + x_2 - 2 = 0.$$

Построим функцию Лагранжа:

$$L(x;\lambda) = x_1^2 + x_2^2 - \lambda(2x_1 + x_2 - 2)$$

и определим ее безусловный минимум. Найдем стационарную точку функции Лагранжа, приравняв нулю компоненты ее градиента:

$$\frac{\partial L}{\partial x_1} = 2x_1 - 2\lambda = 0 \Rightarrow\ x_1^0 = \lambda,$$

$$\frac{\partial L}{\partial x_2} = 2x_2 - \lambda = 0 \Rightarrow\ x_2^0 = \frac{\lambda}{2}.$$

Для того чтобы проверить, соответствует ли стационарная точка x^0 минимуму, вычислим матрицу Гессе функции Лагранжа, рассматриваемой как функция от x:

$$H_L(x;\lambda) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Эта матрица положительно определена, так как для любого ненулевого вектора $u^T = (a,b)$

$$u^T H_L u = (a,b)\cdot\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}\cdot\begin{pmatrix}a\\ b\end{pmatrix} = (2a,2b)\cdot\begin{pmatrix}a\\ b\end{pmatrix} = 2a^2 + 2b^2 > 0.$$

Это означает, что $L(x;\lambda)$ в точке $x_1^0 = \lambda$ и $x_2^0 = \dfrac{\lambda}{2}$ имеет точку глобального минимума. Оптимальное значение λ находится путем подстановки значений x_1^0 и x_2^0 в уравнение $2x_1 + x_2 = 2$, откуда

$$2\lambda + \frac{\lambda}{2} = 2 \text{ и } \lambda^0 = \frac{4}{5}.$$

Таким образом, условный минимум достигается при

$$x_1^0 = \frac{4}{5},\ x_2^0 = \frac{2}{5},$$

а минимальное значение $f(x^0;\lambda^0)$ есть $\dfrac{4}{5}$.

Очень часто оказывается, что решение системы

$$\frac{\partial L}{\partial x_j} = 0,\ \ j = 1,2,...,n$$

в виде явной функции переменной λ получить нельзя. Тогда значения x и λ находятся путем решения следующей системы, состоящей из $n+1$ уравнений с $n+1$ неизвестными:

$$\frac{\partial L}{\partial x_j} = 0,\ j = 1,2,...,n,$$

$$h_1(x) = 0.$$

Решить такую систему можно каким-либо численным методом.

Для каждого из решений $(x^0;\lambda^0)$ вычисляется матрица Гессе функции Лагранжа, рассматриваемой как функция от x. Если она положительно определена, то решение – точка минимума.

Метод множителей Лагранжа можно распространить на случай, когда задача имеет несколько ограничений в виде равенств:

$$f(x) \to \min,\ x \in R^n,$$

$$h_k(x) = 0,\ k = 1,2,...,K.$$

Функция Лагранжа принимает вид

$$L(x;\lambda) = f(x) - \sum_{k=1}^{K}\lambda_k h_k.$$

Здесь $\lambda_1,...,\lambda_K$ – множители Лагранжа, то есть неизвестные параметры, значения которых нужно определить. Приравнивая частные производные L по x нулю, получаем следующую систему

$$\frac{\partial L(x,\lambda)}{\partial x_1} = \frac{\partial L(x,\lambda)}{\partial x_2} = ... = \frac{\partial L(x,\lambda)}{\partial x_n} = 0.$$

Если найти решение этой системы в виде функций от вектора λ затруднительно, то можно расширить последнюю систему путем включения в неё ограничений-равенств:

$$h_1(x) = 0,\ h_2(x) = 0,...,h_K(x) = 0.$$

Решение расширенной системы, состоящей из $N+K$ уравнений с $N+K$ неизвестными, определяет стационарную точку функции L. Затем реализуется процедура проверки на минимум или максимум, которая проводится на основе вычисления элементов матрицы Гессе функции Лагранжа, рассматриваемой как функция от x.

5.2. Задачи с ограничениями в виде неравенств

Рассмотрим задачу

$$f(x) \to \min,\ x \in R^n \qquad (5.4)$$

с ограничениями-неравенствами

$$g_j(x) \le 0,\ j = 1,2,...,J. \qquad (5.5)$$

Пусть область (5.5) (обозначим ее D) – не пустое, ограниченное замкнутое множество. Функция Лагранжа для задачи (5.4) с ограничениями (5.5) определяется формулой

$$L(x;\lambda) = f(x) - \sum_{j=1}^{J} \lambda_j g_j(x) = f(x) - \lambda^T g(x), \qquad (5.6)$$

где λ – вектор множителей Лагранжа: $\lambda = (\lambda_1,...,\lambda_J)^T$, $g = (g_1,...,g_J)^T$.

В точке локального минимума x^* задачи (5.4), (5.5) каждое из ограничений (5.5) выполняется либо в виде равенства $g_j(x^*) = 0$, либо в виде неравенства $g_j(x^*) < 0$. Ограничения первого вида называются *активными ограничениями*. Остальные ограничения называются *неактивными ограничениями*.

Если точка $x^* \in D$ и ограничения $g_{j_k}(x^*) \le 0,\ j_k = 1,...,s,\ s \le J$ активны, то условие линейной независимости градиентов функций $g_{j_k}(x^*),\ j_k = 1,...,s,\ s \le J$ активных ограничений в точке x^* называется условием *регулярности ограничивающих функций* в точке x^*. Это условие означает, что, например, при n=2 количество ограничивающих функций, проходящих через точку x^*, не должно превышать 2 и в точке x^* векторы $\nabla g_1(x)$, $\nabla g_2(x)$ не должны быть коллинеарны. Например, на рис. 20 в ситуации (а) количество ограничивающих функций, про-

ходящих через точку x^*, превышает размерность вектора варьируемых параметров, в ситуации (б) в точке x^* градиенты $\nabla g_1(x)$, $\nabla g_2(x)$ ограничивающих функций коллинеарны.

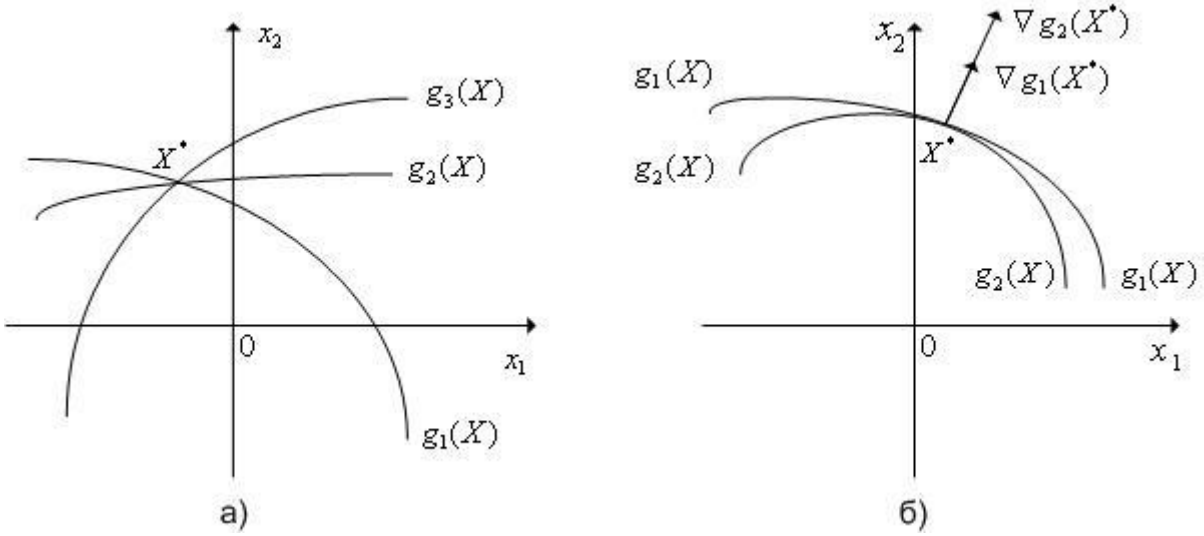


Рис. 20. *Ситуации, в которых не выполняется условие регулярности двумерной задачи*

Большое значение в теории и вычислительной практике имеет следующая теорема (*теорема Куна-Таккера для задачи условной оптимизации с ограничениями типа неравенств*).

Теорема. Пусть функция $f(x)$ и функции $g_j(x) \le 0,\ j = 1,2,...,J$ имеют непрерывные частные производные в некоторой окрестности точки x^* и пусть эта точка является точкой локального минимума функции $f(x)$ при ограничениях $g_j(x^*) \le 0$, удовлетворяющих в точке x^* условию регулярности ограничивающих функций. Тогда существуют такие неотрицательные множители Лагранжа $\lambda_1,...,\lambda_J$, что для функции Лагранжа $L(x;\lambda)$ точка x^* является стационарной точкой, т.е.

$$\nabla_x L(x^*;\lambda) = \nabla f(x^*) - \sum_{j=1}^{J} \lambda_j \nabla g_j(x^*) = 0.$$

Отметим, что теорема не запрещает того, чтобы все множители Лагранжа были равны нулю.

Смысл этой теоремы можно пояснить следующим примером.

Рассмотрим двумерную (n=2) задачу (5.4), (5.5), в которой область допустимых значений D задается тремя ограничивающими функциями. Положим, что множество D имеет вид, представленный на рис. 21.

Для всех граничных точек области D (рис. 21), очевидно, выполняются условия регулярности ограничивающих функций.

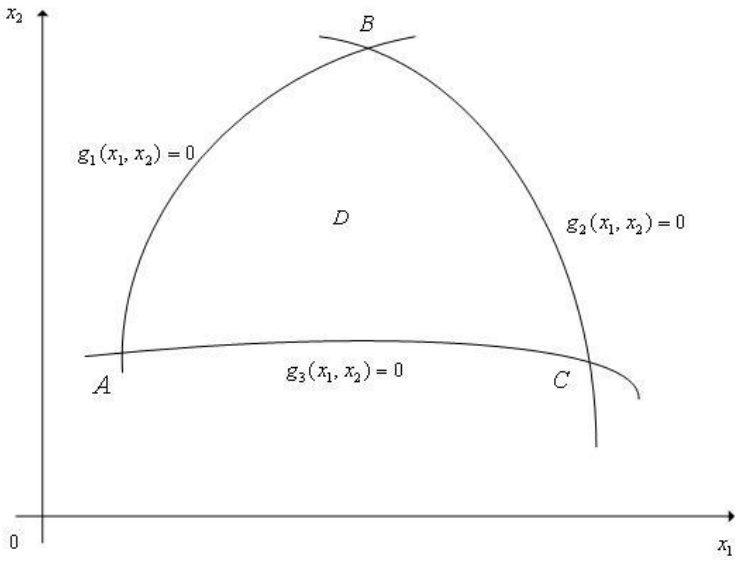


Рис. 21. *К теореме Куна-Таккера*

Если точка x^* находится внутри множества *D* (т.е. является стационарной точкой функции *f*(*x*)), то теорема будет справедлива, если положить все множители Лагранжа λ_i равными нулю.

Пусть теперь точка x^* находится на одной из дуг, например, на дуге *AB*, т.е. пусть ограничение $g_1(x) \le 0$ является активным ограничением, а остальные ограничения – неактивными ограничениями. Тогда в этой точке $g_1(x) = 0$ и справедливость теоремы вытекает из правила множителей Лагранжа для задачи с ограничениями типа равенств, если положить $\lambda_2 = \lambda_3 = 0$.

Пусть, наконец, точка x^* находится в одной из угловых точек множества *D*, например, в точке *B*, т.е. пусть ограничения $g_1(x) \le 0$, $g_2(x) \le 0$ являются активными ограничениями, а ограничение $g_3(x) \le 0$ – неактивным ограничением. Тогда можно положить $\lambda_3 = 0$ и справедливость теоремы вытекает из правила множителей Лагранжа для задачи с ограничениями типа равенств.

Теорема Куна-Таккера означает, что в ее условиях вместо задачи условной оптимизации (5.4), (5.5) можно решать задачу безусловной оптимизации функции Лагранжа (5.6).

Необходимым условием существования локального минимума этой задачи в некоторой точке x^* является условие

$$\nabla_x L\left(x^*;\lambda\right) = \nabla f(x^*) - \sum_{j=1}^{J} \lambda_j \nabla g_j(x^*) = 0.$$

5.2. Методы штрафных функций

Рассмотрим задачу условной оптимизации

$$f(x) \to \min,\ x \in R^n, \tag{5.7}$$

$$g_j(x) \le 0,\ j = 1,2,...,J, \tag{5.8}$$

$$h_k(x) = 0,\ k = 1,2,...,K, \tag{5.9}$$

$$x_i^{(l)} \le x_i \le x_i^{(u)},\ i = 1,2,...,n. \tag{5.10}$$

Такая задача также называется задачей *нелинейного программирования*.

Говорят, что точка x соответствует допустимому решению задачи нелинейного программирования, если для нее выполняются все ограничения, то есть соотношения (5.8–5.10).

Предполагается, что для вектора x^*, являющегося решением задачи нелинейного программирования, известно некоторое начальное приближение x^0, возможно недопустимое. В методах штрафных функций строится последовательность точек x^m, $m = 0,1,...,M,$ которая начинается с заданной точки x^0 и заканчивается точкой x^M, дающей наилучшее приближение к x^* среди всех точек построенной последовательности. В качестве x^m берутся точки решения вспомогательной задачи безусловной минимизации, полученной преобразованием исходной целевой функции с помощью так называемых *штрафных* функций. В этих методах исходная задача условной оптимизации преобразуется в последовательность задач безусловной оптимизации.

Методы штрафных функций классифицируются в соответствии со способами учета ограничений-неравенств. В зависимости от того, являются ли элементы последовательности x^m допустимыми или недопустимыми точками, говорят о *методах внутренней и внешней точки* соответственно. Если последовательность x^m содержит точки обоих типов, метод называют *смешанным*.

Пусть необходимо решить задачу (5.7–5.10). Основная идея метода штрафных функций заключается в следующем. Строят вспомогательную функцию

$$Q(x,r,l) = f(x) + \sum_{j=1}^{J} r_j G_j\left(g_j(x)\right) + \sum_{k=1}^{K} l_k H_k\left(h_k(x)\right), \tag{5.11}$$

такую, что приближенное решение задачи (5.7–5.10) получается в результате решения *последовательности* задач безусловной минимизации функции

$$Q(x,r,l) \to \min,\ x \in R^n. \tag{5.12}$$

В методе ***внешних штрафных функций*** функции *H, G* выбираются таким образом, чтобы они становились отличными от нуля (положительными) при нарушении соответствующего ограничения (рис. 22). А так как мы минимизиру-

ем (5.11), то движение в сторону нарушения становится невыгодным. Внутри допустимой области в данном методе функции *H* и *G* должны быть равны нулю. Например, для ограничений-неравенств $G_j\left(g_j(x)\right)\to 0$ при $g_j(x)\to 0^+$.

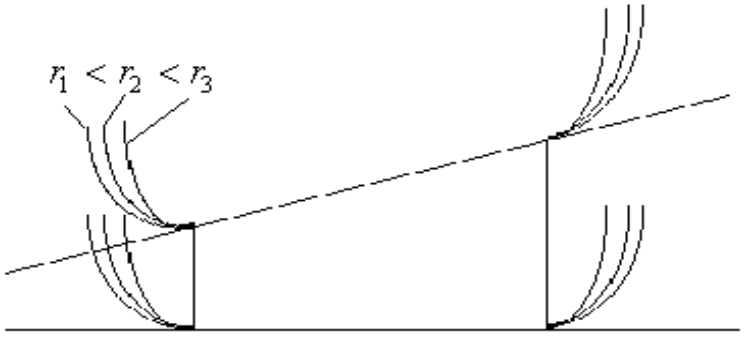


Рис. 22. *Поведение штрафных функций* $r_jG_j\left(g_j(x)\right)$ *в методах внешней точки*, j=1,2,3 (*снизу – функции* $r_jG_j(x)$)

Приближенное решение задачи (5.7–5.10) получается в результате решения последовательности задач (5.12) при $r_j,\ l_k\to\infty,\ j=1,...,J,\ k=1,...,K$.

В методе ***барьерных функций*** функции *H*, *G* выбираются отличными от нуля в допустимой области и такими, чтобы при приближении к границе допустимой области (изнутри) они возрастали, препятствуя выходу при поиске за границу области (рис. 23). В этом случае эти функции должны быть малыми (*положительными или отрицательными*) внутри допустимой области и большими *положительными* вблизи границы (внутри области). Например, для ограничений неравенств $G_j\left(g_j(x)\right)\to\infty$ при $g_j(x)\to 0^-$.

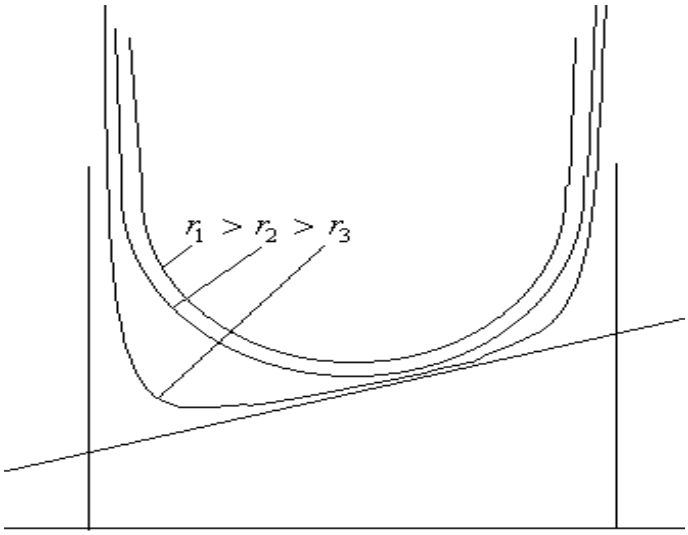


Рис. 23. *Поведение штрафных функций* $G_j\left(g_j(x)\right)\to\infty$ *в методах барьерных функций*, j=1,2,3

Такие методы называют еще *методами внутренней точки*. В алгоритмах, использующих функции штрафа данного типа, требуют, чтобы в процессе поиска точка x всегда оставалась внутренней точкой допустимой области. Прибли-

женное решение задачи (5.7–5.10) получается в результате решения последовательности задач (5.12) при $r_j,\ l_k \to 0,\ j=1,...,J,\ k=1,...,K$.

Для ограничений-равенств при выборе функций штрафов обычно требуют, чтобы $H_k\left(h_k(x)\right)\to 0$ при $h_k(x)\to 0$.

Это могут быть, например, функции вида:

$H_k\left(h_k(x)\right)=\left|h_k(x)\right|$, или

$H_k\left(h_k(x)\right)=\left(h_k(x)\right)^{\alpha}$, где α – четное (например, α =2).

Для ограничений-неравенств функции штрафа подбирают таким образом, чтобы

$G_j\left(g_j(x)\right)=0$, при $g_j(x)\le 0$,

$G_j\left(g_j(x)\right)>0$, при $g_j(x)>0$.

Этому требованию отвечают функции вида:

$$G_j\left(g_j(x)\right)=\frac{1}{2}\left\{g_j(x)+\left|g_j(x)\right|\right\}, \qquad (5.13)$$

$$G_j\left(g_j(x)\right)=\left[\frac{1}{2}\left\{g_j(x)+\left|g_j(x)\right|\right\}\right]^{\alpha}, \qquad (5.14)$$

при четном α. При α =2 штраф называют квадратичным.

В качестве барьерных функций для ограничений-неравенств могут служить функции вида:

$$G_j\left(g_j(x)\right)=\frac{1}{-g_j(x)}, \qquad (5.15)$$

$$G_j\left(g_j(x)\right)=-\ln\left(-g_j(x)\right). \qquad (5.16)$$

Логарифмический штраф (5.16) – это барьерная функция, не определенная в недопустимых точках (то есть для таких x, что $g(x)>0$). Поэтому в тех случаях, когда приходится иметь дело с недопустимыми точками (например, когда заданное начальное приближение x^0 не является допустимым), требуется специальная процедура, обеспечивающая попадание в допустимую область.

Штраф, заданный функцией (5.15), не имеет отрицательных значений в допустимой области. Этот штраф, как и предыдущий, является барьером; в этом случае также возникают трудности, связанные с возможным появлением недопустимых точек.

Часто функцию $Q(x,r,l)$ выбирают в виде

$$Q(x,r,l)=f(x)-R\sum_{j=1}^{J}\frac{1}{g_j(x)}+\frac{1}{R}\sum_{k=1}^{K}\left(h_k(x)\right)^2, \text{ или}$$

$$Q(x,r,l)=f(x)-R\sum_{j=1}^{J}\ln\left(-g_j(x)\right)+\frac{1}{R}\sum_{k=1}^{K}\left(h_k(x)\right)^2,$$

где положительный штрафной параметр $R \to 0$, монотонно убывая от итерации к итерации.

Последовательность действий при реализации методов штрафных функций или барьерных функций выглядит следующим образом:

1. На основании задачи (5.7–5.10) строим функцию (5.11). Выбираем начальной приближение x и начальные значения коэффициентов штрафа r_j, l_k, число итераций, точность безусловной оптимизации, точность соблюдения ограничений и т.д.

2. Решаем задачу (5.12).

3. Если полученное решение не удовлетворяет системе ограничений в случае использования метода штрафных функций, то увеличиваем значения коэффициентов штрафа и снова решаем задачу (5.12). В случае метода барьерных функций значения коэффициентов уменьшаются, чтобы можно было получить решение на границе.

4. Процесс прекращается, если найденное решение удовлетворяет системе ограничений с определенной точностью.

5.3. Метод факторов

Своеобразным и очень эффективным методом штрафов является метод факторов (или множителей), который основан на штрафе типа “квадрат срезки” для ограничений-неравенств.

Такой штраф определяется следующим образом:

$$S = R \cdot \langle g(x) \rangle^2, \tag{5.17}$$

где ***срезка*** t определяется так:

$$\langle t \rangle = \begin{cases} t, & \text{если } t \geq 0, \\ 0, & \text{если } t < 0. \end{cases} \tag{5.18}$$

Этот штраф внешний и стационарные точки функции $Q(x, R)$ могут оказаться недопустимыми. С другой стороны, недопустимые точки не создают в данном случае дополнительных сложностей по сравнению с допустимыми. Различие между ними состоит лишь в том, что в допустимых точках штраф равен нулю.

В методе факторов на каждой итерации производится безусловная минимизация функции

$$Q(x, \sigma, \tau) = f(x) + R\sum_{j=1}^{J}\left\{\langle g_j(x) + \sigma_j \rangle^2 - \sigma_j^2\right\} + R\sum_{k=1}^{K}\left\{\left[h_k(x) + \tau_k\right]^2 - \tau_k^2\right\}, \tag{5.19}$$

где R – постоянный весовой коэффициент, а угловые скобки обозначают операцию срезки. Параметры (факторы) σ_j и τ_k осуществляют сдвиг штрафных слагаемых. Компоненты векторов σ и τ меняются по ходу вычислений, однако в процессе решения каждой вспомогательной безусловной задачи оба эти вектора

остаются постоянными. Начальные значения факторов σ и τ можно выбрать нулевыми. Обозначим через x^m точку минимума функции $Q\left(x,\sigma^m,\tau^m\right)$, используемой на *m*-ой итерации.

При переходе к (*m*+1)-й итерации факторы пересчитываются по формулам

$$\sigma_j^{m+1}=\left\langle g_j\left(x^m\right)+\sigma_j^m\right\rangle,\ \ j=1,...,J\,, \tag{5.20}$$

$$\tau_k^{m+1}=h_k\left(x^m\right)+\tau_k^m,\qquad k=1,...,K\,. \tag{5.21}$$

Формулы пересчета таковы, что в результате сдвига при переходе к новой подзадаче штраф за нарушение ограничений возрастает, и вследствие этого точки x^m приближаются к допустимой области.

Для контроля сходимости метода используют последовательности x^m, σ^m, τ^m, $f\left(x^m\right)$, $g\left(x^m\right)$, $h\left(x^m\right)$. Прекращение основного процесса происходит, когда члены, по крайней мере, одной из этих последовательностей, перестают изменяться при пересчете факторов и последующей безусловной минимизации. Заметим, что величина положительного параметра *R* влияет на свойства метода, но конструктивного алгоритма его выбора не существует.

6. Случайный поиск

Регулярные, или детерминированные методы спуска, рассмотренные нами выше, не полноценны на неупорядоченном рельефе. Если экстремумов много, то спуск из одного начального приближения может сойтись только к одному из локальных минимумов, не обязательно абсолютному. Кроме того, с ростом размерности задач резко снижается эффективность регулярных методов поиска, которые требуют очень больших вычислительных ресурсов. Поэтому иногда для исследования таких сложных задач применяют случайный поиск.

6.1. Простой случайный поиск

Предполагают, что искомый минимум лежит в некотором n-мерном параллелепипеде. В этом параллелепипеде по равномерному закону выбирают случайным образом N пробных точек и вычисляют в них целевую функцию. Точку, в которой функция имеет минимальное значение, берут в качестве решения задачи. Однако даже при очень большом числе пробных точек вероятность того, что хотя бы одна точка попадает в небольшую окрестность локального минимума, ничтожно мала. Действительно, пусть $N=10^6$ и диаметр котловины около минимума составляет 10% от пределов изменения каждой координаты. Тогда объем этой котловины составляет 0.1^n часть объема n-мерного параллелепипеда. Уже при числе переменных $n>6$ практически ни одна точка в котловину не попадет.

Поэтому берут небольшое число точек $N=(5\div 20)\cdot n$ и каждую точку рассматривают как нулевое приближение. Из каждой точки совершают спуск, быстро попадая в ближайший овраг или котловину; когда шаги спуска быстро укорачиваются, его прекращают, не добиваясь высокой точности. Этого уже достаточно, чтобы судить о величине функции в ближайшем локальном минимуме с удовлетворительной точностью. Сравнивая окончательные значения функции на всех спусках между собой, можно изучить расположение локальных минимумов и сопоставить их величины. После этого можно отобрать нужные по смыслу задачи минимумы и провести в них дополнительные спуски для получения координат точек минимума с более высокой точностью.

При решении экстремальных задач на областях со сложной геометрией обычно эту область вписывают в n-мерный гиперпараллелепипед, в котором генерируют случайные точки по равномерному закону, оставляя только те, которые попадают в допустимую область (рис. 24).

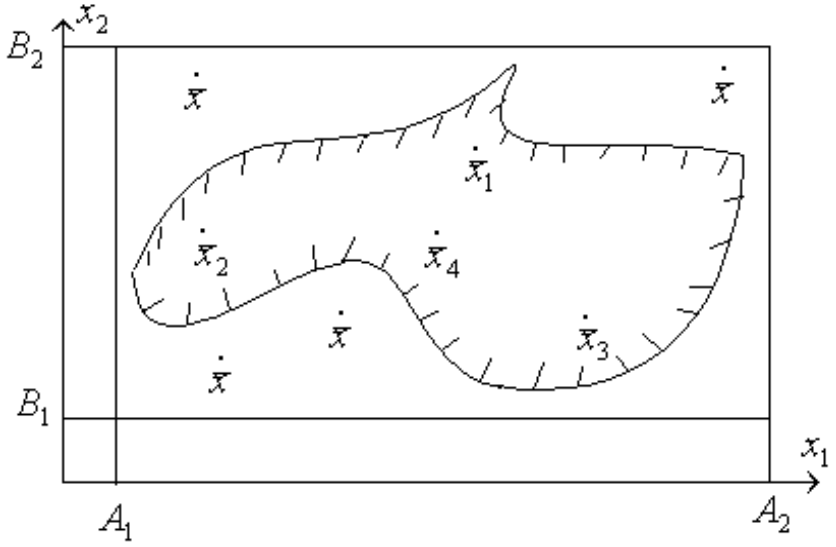


Рис. 24. *Построение включающег гиперпараллелепипеда* (*A, B – границы параллелепипеда*)

Различают направленный и ненаправленный случайный поиск.

6.2. Ненаправленный случайный поиск

Все случайные испытания строят совершенно не зависимо от результатов предыдущих. Сходимость такого поиска очень мала, но имеется важное преимущество: возможность решать многоэкстремальные задачи (искать глобальный экстремум). Примером является рассмотренный выше простой случайный поиск.

6.3. Направленный случайный поиск

В этом случае отдельные испытания связаны между собой. Результаты проведенных испытаний используются для формирования последующих. Скорость ходимости таких методов, как правило, выше, но сами методы обычно приводят к локальным экстремумам.

Приведем простейшие алгоритмы направленного случайного поиска.

6.3.1. Алгоритм парной пробы

В данном алгоритме четко разделены пробный и рабочий шаги. Пусть x^k – найденное на k-м шаге наименьшее значение минимизируемой функции $f(x)$. По равномерному закону генерируется случайный единичный вектор ξ и по обе стороны от исходной точки x^k делаются две пробы: проводят вычисление функции в точках $x^k_{1,2} = x^k \pm g \cdot \xi$, где g-величина пробного шага (рис. 25). Рабочий шаг делается в направлении наименьшего значения целевой функция. Очередное приближение определяется соотношением

$$x^{k+1} = x^k + \Delta x^k = x^k + a \cdot \xi \cdot \text{sign}\left(f(x^k - g\xi) - f(x^k + g\xi) \right),$$

где $\operatorname{sign}(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$

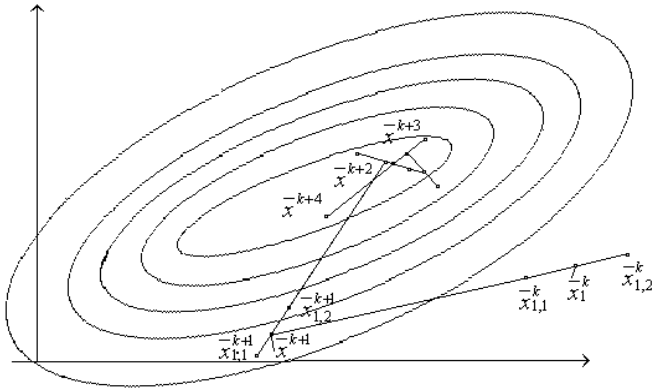


Рис. 25. *К алгоритму парной пробы*

Особенностью данного алгоритма является его повышенная тенденция к «блужданию». Даже найдя экстремум, алгоритм уводит систему в сторону.

6.3.2. Алгоритм наилучшей пробы

На k-м шаге мы имеем точку x^k. Генерируется m случайных единичных векторов $\xi_1, ..., \xi_m$. Делаются пробные шаги в направлениях $g \cdot \xi_1, ..., g \cdot \xi_m$ и в точках $x^k + g \cdot \xi_1, ..., x^k + g \cdot \xi_m$ вычисляются значения функции (рис. 26). Выбирается тот шаг, который приводит к наибольшему уменьшению функции: $\xi^* = \arg\min_{i=1,m} f\left(x^k + g \cdot \xi_i\right)$. В данном направлении делается шаг $\Delta x^k = \lambda \cdot \xi^*$. Параметр λ может определяться как результат минимизации по определенному направлению или выбирается по определенному закону.

С увеличением числа проб выбранное направление приближается к направлению антиградиента $-\nabla f\left(x\right)$.

Если функция $f\left(x\right)$ близка к линейной, то есть возможность ускорить поиск, выбирая вместе с наилучшей и наихудшую пробу. Тогда рабочий шаг можно делать или в направлении наилучшей пробы, или в направлении, противоположном наихудшей пробе.

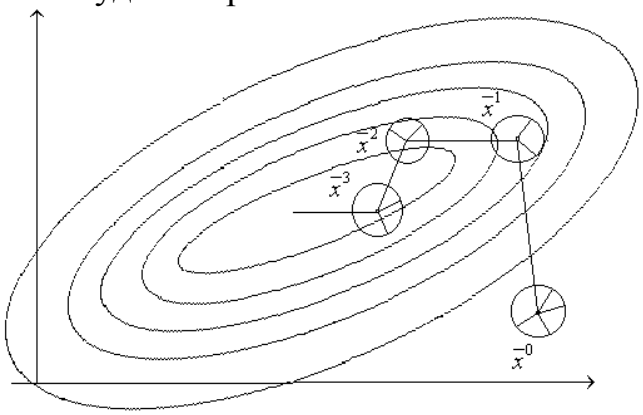


Рис. 26. *К алгоритму наилучшей пробы*

6.3.3. Метод статистического градиента

Из исходного состояния x^k делается m независимых проб $g \cdot \xi_1, ..., g \cdot \xi_m$ и вычисляются соответствующие значения минимизируемой функции в этих точках (рис. 27). Для каждой пробы запоминются приращения функции

$$\Delta f_j = f\left(x^k + g \cdot \xi_j\right) - f\left(x^k\right).$$

После этого формируют векторную сумму $\Delta f = \sum_{j=1}^{m} \xi_j \cdot \Delta f_j$. В пределе при $m \to \infty$ она совпадает с направлением градиента целевой функции. При конечном m вектор Δf представляет собой статистическую оценку направления градиента. Рабочий шаг делается в направлении Δf. Очередное приближение определяется соотношением

$$x^{k+1} = x^k - \lambda \cdot \frac{\Delta f}{\|\Delta f\|}.$$

При выборе оптимального значения λ, которое минимизирует функцию в заданном направлении, получают случайный вариант метода наискорейшего спуска. Существенным преимуществом перед детерминированными алгоритмами является возможность принятия решения о направлении рабочего шага при $m < n$. При $m = n$ и неслучайных ортогональных рабочих шагах, направленных вдоль осей координат, алгоритм вырождается в градиентный метод.

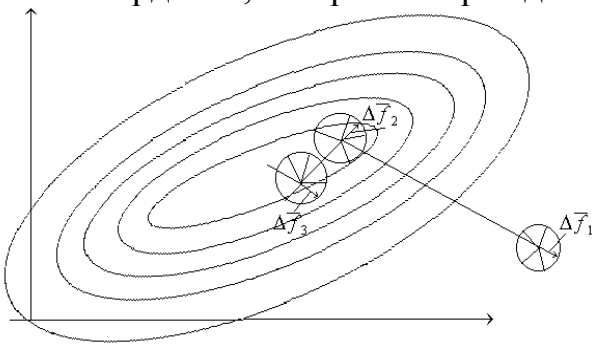


Рис. 27. *К алгоритму статистического градиента*

6.3.4. Алгоритм наилучшей пробы с направляющим гиперквадратом

Внутри допустимой области строится гиперквадрат. В этом гиперквадрате случайным образом разбрасывается m точек $x_1, ..., x_m$, в которых вычисляются значения функции (рис. 28). Среди построенных точек выбирают наилучшую. Опираясь на эту точку, строят новый гиперквадрат. Точка, в которой достигается минимум функции на k-м этапе, берется в качестве центра нового гиперквадрата на $(k+1)$-м этапе. Координаты вершин гиперквадрата на $(k+1)$-м этапе определяются соотношениями

$$a_i^{k+1} = x_i^{k+1} - \frac{b_i^k - a_i^k}{2},\ b_i^{k+1} = x_i^{k+1} + \frac{b_i^k - a_i^k}{2},$$

где x^k – наилучшая точка в гиперквадрате на k-м этапе.

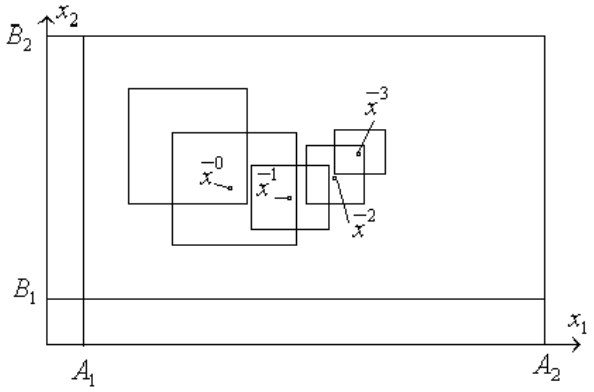


Рис. 28. *К алгоритму с направляющим гиперквадратом*

В новом гиперквадрате выполняем ту же последовательность действий, случайным образом разбрасывая m точек, и т.д.

Таким образом на 1-м этапе координаты случайных точек удовлетворяют неравенствам $a_i^1 \le x_i \le b_i^1$, $i = 1,...,n$, и $x^1 = \arg\min\limits_{j=\overline{1,m}} \{ f(x_j) \}$ – точка с минимальным значением целевой функции.

В алгоритме с *обучением* стороны гиперквадрата могут регулироваться в соответствии с изменением параметра α по некоторому правилу. В этом случае координаты вершин гиперквадрата на $(k+1)$-м этапе будут определяться соотношениями

$$a_i^{k+1} = x_i^{k+1} - \frac{b_i^k - a_i^k}{2\alpha},\ b_i^{k+1} = x_i^{k+1} + \frac{b_i^k - a_i^k}{2\alpha}.$$

Хорошо выбранное правило регулировки стороны гиперквадрата приводит к достаточно эффективному алгоритму поиска.

В алгоритмах случайного поиска вместо направляющего гиперквадрата могут использоваться направляющие гиперсферы, направляющие гиперконусы.

6.4. Алгоритмы глобального поиска

Случайный поиск приобретает решающее значение при решении многоэкстремальных задач. В общем случае решение многоэкстремальных задач без элемента случайности практически невозможно.

Алгоритм 1

В допустимой области D случайным образом выбирают точку $x_1 \in D$. Приняв ее за исходную и используя некоторый детерминированный метод или алгоритм направленного случайного поиска, осуществляется спуск в точку локального минимума $x_1^* \in D$ (рис. 29).

Затем выбирается новая случайная точка $x_2 \in D$ и по той же схеме осуществляется спуск в точку локального минимума $x_2^* \in D$ и т.д.

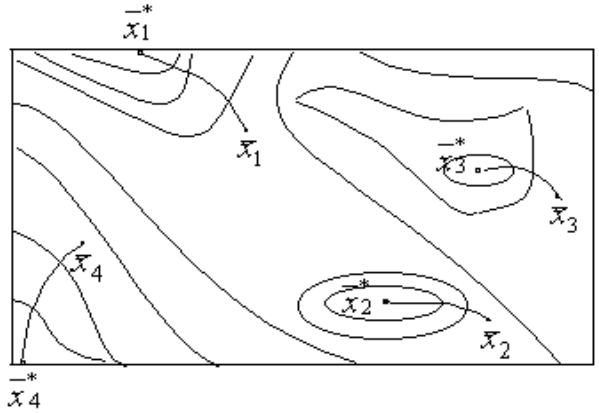


Рис. 29. *Алгоритм* 1 *глобального поиска*

Поиск прекращается, как только заданное число раз не удается найти точку локального экстремума со значением функции, меньшим предыдущих.

Алгоритм 2

Пусть получена некоторая точка локального экстремума $x_1^* \in D$. После этого переходим к ненаправленному случайному поиску до получения точки x_2 такой, что $f(x_2) < f(x_1^*)$.

Из точки x_2 с помощью детерминированного алгоритма или направленного случайного поиска получаем точку локального экстремума x_2^*, в которой заведомо выполняется неравенство $f(x_2^*) < f(x_1^*)$.

Далее с помощью случайного поиска определяем новую точку x_3, для которой справедливо неравенство $f(x_3) < f(x_2^*)$, и снова спуск в точку локального экстремума x_3^* и т.д.

Поиск прекращается, если при генерации некоторого предельного числа новых случайных точек не удается найти лучшей, чем предыдущий локальный экстремум, который и принимается в качестве решения.

Алгоритм 3

Пусть x_1^0 – некоторая исходная точка поиска в области D, из которой осуществляется спуск в точку локального экстремума x_1^* со значением $f(x_1^*)$. Далее из точки x_1^* двигаемся либо в случайном направлении, либо в направлении $x_1^* - x_1^0$ до тех пор, пока функция снова не станет убывать (выходим из «области притяжения» x_1^*). Полученная точка x_2^0 принимается за начало следующего спуска. В результате находим новый локальный экстремум x_2^* и значением функции $f(x_2^*)$ (рис. 30).

Если $f(x_2^*) < f(x_1^*)$, точка x_1^* забывается и ее место занимает точка x_2^*. Если $f(x_2^*) \geq f(x_1^*)$, то возвращаемся в точку x_1^* и движемся из нее в новом случайном направлении.

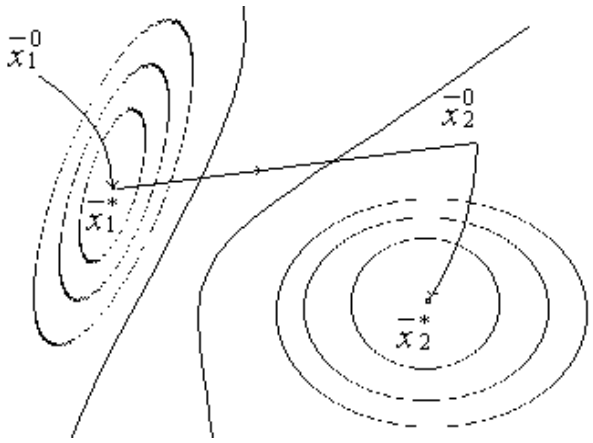


Рис. 30. *Алгоритм* 3 *глобального поиска*

Процесс прекращается, если не удается найти лучший локальный минимум после заданного числа попыток или «случайного» направления, в котором функция снова начинает убывать.

Этот метод позволяет найти глобальный экстремум в случае многосвязных допустимых областей.

Алгоритм 4

В допустимой области D разбрасывают m случайных точек и выбирают из них наилучшую, то есть ту, в которой значение функции минимально (рис. 31). Из выбранной точки осуществляют локальный спуск. Далее вокруг траектории спуска образуют запретную область. В оставшейся области случайным образом разбрасывают новую совокупность случайных точек, и из лучшей точки осуществляют спуск в точку локального экстремума. Вокруг новой траектории также строят запретную область и т.д.

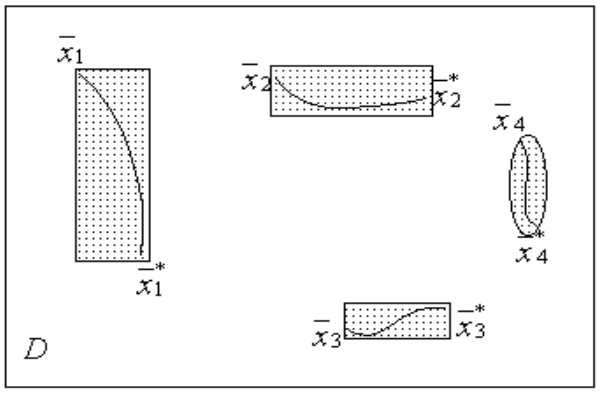


Рис. 31. *Алгоритм* 4 *глобального поиска*

Поиск прекращается, если в течение заданного числа попыток не удается найти лучшего локального экстремума.

Замечание. Комбинация случайного поиска с детерминированными методами применяется не только для решения многоэкстремальных задач. Часто к такой комбинации прибегают в ситуациях, когда детерминированные методы сталкиваются с теми или иными трудностями (застревают на дне узкого оврага, в седловой точке и т.д.). Шаг в случайном направлении порой позволяет преодолеть такую тупиковую ситуацию для детерминированного алгоритма.

7. ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ

7.1. Примеры задач линейного программирования

7.1.1. Задача об использовании сырья

Предположим, что изготовление продукции двух видов $П_1$ и $П_2$ требует использования четырех видов сырья S_1, S_2, S_3, S_4. Запасы сырья каждого вида ограничены и составляют соответственно b_1, b_2, b_3, b_4 условных единиц. Количество единиц сырья, необходимое для изготовления единицы каждого из видов продукции, известно и задается табл. 1.

Таблица 1

Виды сырья	Запасы сырья	Виды продукции	
		$П_1$	$П_2$
S_1	b_1	a_{11}	a_{12}
S_2	b_2	a_{21}	a_{22}
S_3	b_3	a_{31}	a_{32}
S_4	b_4	a_{41}	a_{42}
Доход		c_1	c_2

Таблица 2

Виды сырья	Запасы сырья	Виды продукции	
		$П_1$	$П_2$
S_1	19	2	3
S_2	13	2	1
S_3	15	0	3
S_4	18	3	0
Доход		7	5

Здесь $a_{ij}\,(i=1,...,4;\ j=1,2)$ означает количество единиц сырья вида S_i, необходимое для изготовления продукции вида $П_j$. В последней строке таблицы указан доход, получаемый предприятием от реализации одной единицы каждого вида продукции.

Требуется составить такой план выпуска продукции видов $П_1$ и $П_2$, при котором доход предприятия от реализации всей продукции оказался бы максимальным.

Математическую формулу поставленной задачи изучим на следующем числовом примере (см. табл. 2).

Допустим, что предприятие выпускает x_1 единиц продукции вида $П_1$ и x_2 единиц продукции вида $П_2$. Для этого потребуется $2x_1+3x_2$ единиц сырья S_1 (см. табл. 2). Так как в наличии имеется всего 19 единиц сырья S_1, то должно выполняться неравенство $2x_1+3x_2\le 19$. Неравенство (а не точное равенство) появляется в связи с тем, что максимальный доход может быть достигнут предприятием и в том случае, когда запасы сырья вида S_1 используются не полностью.

Аналогичные рассуждения, проведенные для остальных видов сырья, позволяют написать следующие неравенства:

$2x_1 + x_2 \le 13$ (сырье S_2),

$3x_2 \le 15$ (сырье S_3),

$3x_1 \le 18$ (сырье S_4).

При этих условиях доход F, получаемый предприятием, составит $F = 7x_1 + 5x_2$.

Таким образом, математически задачу можно сформулировать так. Дана система

$$\left.\begin{aligned} &2x_1 + 3x_2 \le 19, \\ &2x_1 + x_2 \le 13, \\ &3x_2 \le 15, \\ &3x_1 \le 18 \end{aligned}\right\} \qquad (7.1)$$

четырех линейных неравенств и линейная форма

$$F = 7x_1 + 5x_2. \qquad (7.2)$$

Требуется среди неотрицательных решений системы (7.1) выбрать такое, при котором форма F принимает наибольшее значение (максимизируется).

7.1.2. Задача об использовании мощностей оборудования

Предположим, что предприятию задан план производства по времени и номенклатуре: требуется за время T выпустить N_1 единиц продукции вида $П_1$ и N_2 вида $П_2$. Каждый из видов продукции может производиться двумя машинами A и B с различными мощностями. Эти мощности заданы в табл. 3. Здесь a_1 есть количество единиц продукции вида $П_1$, произведенной машиной A в единицу времени. Аналогичный смысл имеют и остальные данные из этой таблицы.

Таблица 3

	$П_1$	$П_2$
A	a_1	a_2
B	b_1	b_2

Таблица 4

	$П_1$	$П_2$
A	α_1	α_2
B	β_1	β_2

Таблица 5

	$П_1$	$П_2$
A	x_1	x_2
B	x_3	x_4

Расходы, вызванные изготовлением каждого из видов продукции на той или иной машине, различны и задаются табл. 4. В этой таблице α_1 выражает цену единицы рабочего времени машины A при изготовлении продукции вида $П_1$. Смысл остальных величин α_2, β_1, β_2 аналогичен.

Требуется составить оптимальный (наиболее рациональный) план работы машин, а именно найти, сколько времени каждая из машин A и B должна быть занята изготовлением каждого из видов продукции $П_1$ и $П_2$ с тем, чтобы стои-

мость всей продукции предприятия оказалась минимальной и в то же время был бы выполнен заданный план как по времени, так и по номенклатуре.

Найдем математическую формулировку поставленной задачи. Введем для неизвестных нам времен работы машин по изготовлению продукции следующие обозначения (табл. 5). Здесь, например, x_1 означает время работы машины A по изготовлению продукции $П_1$. Аналогичный смысл имеют величины x_2, x_3, x_4.

Поскольку машины A и B работают одновременно, то выполнение плана по времени будет обеспечиваться неравенствами

$$\left.\begin{aligned} x_1 + x_2 \le T, \\ x_3 + x_4 \le T. \end{aligned}\right\}$$

Изготовлением продукции $П_1$ машина A занята x_1 единиц времени. При этом за единицу времени она производит a_1 единиц продукции этого вида. Следовательно, всего машина A изготовляет $a_1 \cdot x_1$ единиц продукции $П_1$. Аналогично машина B изготовит $b_1 \cdot x_3$ единиц продукции вида $П_1$. Поэтому для обеспечения плана по номенклатуре должно выполняться равенство

$$a_1 x_1 + b_1 x_3 = N_1.$$

Аналогично для обеспечения плана по продукции $П_2$ необходимо выполнение равенства

$$a_2 x_2 + b_2 x_4 = N_2.$$

Далее, из условий задачи вытекает, что общая стоимость всей продукции составит

$$F = \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 x_3 + \beta_2 x_4 .$$

В итоге мы приходим к следующей математической задаче:

Задана система

$$\left.\begin{aligned} x_1 + x_2 \le T, \\ x_3 + x_4 \le T, \\ a_1 x_1 + b_1 x_3 = N_1, \\ a_2 x_2 + b_2 x_4 = N_2 \end{aligned}\right\} \qquad (7.3)$$

двух линейных неравенств и двух линейных уравнений и задана линейная форма

$$F = \alpha_1 x_1 + \alpha_2 x_2 + \beta_1 x_3 + \beta_2 x_4 . \qquad (7.4)$$

Требуется среди всех неотрицательных решений системы (7.3) найти такое, при котором форма F принимает наименьшее значение (минимизируется).

7.1.3. Транспортная задача

На двух станциях отправления A_1 и A_2 сосредоточено соответственно a_1 и a_2 единиц некоторого однородного груза. Этот груз следует доставить в три пункта назначения B_1, B_2, B_3. Причем в каждый из них должно быть завезено соответственно b_1, b_2, b_3 единиц этого груза. Стоимость перевозки единицы груза из пункта A_i в пункт B_j (обозначим c_{ij}) считаем заданной. Все данные полезно свести в табл. 7.

Таблица 6

Пункты назначения / Пункты отправления	Пункты назначения			Запасы груза
	B_1	B_2	B_3	
A_1	c_{11}	c_{12}	c_{13}	a_1
A_2	c_{21}	c_{22}	c_{23}	a_2
Потребность в грузе	b_1	b_2	b_3	$\sum a_i = \sum b_j$

Будем считать, что общий запас грузов на станциях отправления равен суммарной потребности в этом грузе всех станций назначения. Следовательно,

$$a_1 + a_2 = b_1 + b_2 + b_3. \qquad (7.5)$$

Требуется составить такой план перевозок, при котором их общая стоимость была бы наименьшей.

Обозначим через x_{ij} количество единиц груза, предназначенного к отправке из пункта A_i в пункт B_j. Тогда количество груза, который планируется к доставке в пункт B_1 из пунктов A_1 и A_2, составит

$$x_{11} + x_{21}.$$

Так как потребность в грузе B_1 равна b_1, то должно выполняться равенство:

$$x_{11} + x_{21} = b_1.$$

Аналогично получим равенства

$$x_{12} + x_{22} = b_2,$$

$$x_{13} + x_{23} = b_3.$$

С другой стороны, общее количество груза, отправленного со станции A_1, выражается суммой

$$x_{11} + x_{12} + x_{13},$$

которая, очевидно, обязана совпадать с запасом a_1 груза, сосредоточенным на этой станции, то есть

$$x_{11} + x_{12} + x_{13} = a_1.$$

Подобно этому

$$x_{21} + x_{22} + x_{23} = a_2.$$

Полученные соотношения легче запомнить, если все величины свести в так называемую *матрицу перевозок* (см. табл. 7). Тогда легко проверить, что сумма всех x_{ij}, расположенных на *i*-ой строке, равна запасу a_i в пункте назначения A_i. Сумма же всех x_{ij} из столбца *j* равна потребности b_j пункта назначения B_j.

Таблица 7

Пункты назначения / Пункты отправления	Пункты назначения			Запасы груза
	B_1	B_2	B_3	
A_1	x_{11}	x_{12}	x_{13}	a_1
A_2	x_{21}	x_{22}	x_{23}	a_2
Потребность в грузе	b_1	b_2	b_3	

Из условий задачи с очевидностью вытекает, что общая стоимость *F* всех перевозок равна

$$F = c_{11}x_{11} + c_{12}x_{12} + c_{13}x_{13} + c_{21}x_{21} + c_{22}x_{22} + c_{23}x_{23} = \sum_{i=1}^{2}\sum_{j=1}^{3} c_{ij}x_{ij} = \sum_{i,j} c_{ij}x_{ij}\ .$$

Таким образом, математическая формулировка транспортной задачи (по критерию стоимости перевозок) такова. Задана система

$$\left.\begin{aligned} &x_{11} + x_{21} = b_1,\\ &x_{12} + x_{22} = b_2,\\ &x_{13} + x_{23} = b_3,\\ &x_{11} + x_{12} + x_{13} = a_1,\\ &x_{21} + x_{22} + x_{23} = a_2. \end{aligned}\right\} \qquad (7.6)$$

пяти линейных алгебраических уравнений с шестью неизвестными и линейная форма

$$F = \sum_{i=1}^{2}\sum_{j=1}^{3} c_{ij}x_{ij} = \sum_{i,j} c_{ij}x_{ij}. \qquad (7.7)$$

Требуется среди всех неотрицательных решений x_{ij} системы (7.6) выбрать такое, при котором форма F минимизируется (достигает наименьшего значения). Отметим, что при решении транспортной задачи следует учитывать важное соотношение, вытекающее из самого условия задачи:

$$\sum_i a_i = \sum_j b_j . \qquad (7.5')$$

Впрочем, возможны и иные постановки транспортной задачи, когда условие (7.5') не выполнено. Однако мы их рассматривать не будем.

7.1.4. Задача о питании

Для сохранения здоровья и работоспособности человек должен потреблять в сутки некоторое количество питательных веществ, например белков, жиров, углеводов, воды и витаминов. Запасы этих ингредиентов в различных видах $\pi_i \left(i = 1,2,...\right)$ пищи различны. Ограничимся для простоты двумя видами пищи и зададим таблицу 8, в которой, например, число a_{11} указывает на запасы жиров в пище вида π_1. Смысл остальных чисел a_{ij} аналогичен.

Таблица 8

Питательные вещества	Норма	Виды пищи	
		π_1	π_2
B_1 – жиры	b_1	a_{11}	a_{12}
B_2 – белки	b_2	a_{21}	a_{22}
B_3 – углеводы	b_3	a_{31}	a_{32}
B_4 – вода	b_4	a_{41}	a_{42}
B_5 – витамины	b_5	a_{51}	a_{52}
Стоимость		c_1	c_2

Предположим далее, что стоимость некоторой единицы пищи вида π_i составляет c_i. Требуется так организовать питание, чтобы стоимость его была наименьшей, но организм получил бы не менее минимальной суточной нормы питательных веществ всех видов $B_i \left(i = 1,2,...,5\right)$.

Пусть x_1 и x_2 – количество пищи видов π_1 и π_2, приобретаемых человеком (подразумевается, что вся приобретаемая пища потребляется).

Математически задачу о питании можно сформулировать так.

Задана система из пяти линейных неравенств с двумя неизвестными x_1, x_2,

$$a_{i1}x_1 + a_{i2}x_2 \geq b_i, \ (i = 1,\ldots,5) \qquad (7.8)$$

и линейная форма относительно этих же неизвестных:

$$F = c_1x_1 + c_2x_2 . \qquad (7.9)$$

Требуется среди всех неотрицательных решений системы (7.8) найти такое, которое сообщает форме F наименьшее значение (минимизирует F).

Естественно, что рассмотренные выше задачи не исчерпывают всех встречающихся на практике типов задач линейного программирования. Тем не менее они дают общее представление об их характере. Несмотря на внешнюю непохожесть (в особенности по содержанию) задач 7.1 – 7.4, нельзя не заметить родственности их математической формы. Более того, в следующем параграфе мы покажем, что все эти различные примеры с помощью определенных приемов, в конце концов, сводятся к одной и той же задаче. Она называется основной задачей линейного программирования. К ее разбору мы и переходим.

7.2. Основная задача линейного программирования

Основная задача линейного программирования состоит в следующем. Задана система

$$\left.\begin{aligned} &a_{11}x_1 + \ldots + a_{1n}x_n = b_1, \\ &a_{21}x_1 + \ldots + a_{2n}x_n = b_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{m1}x_1 + \ldots + a_{mn}x_n = b_m \end{aligned}\right\} \qquad (7.10)$$

m линейных алгебраических уравнений с n неизвестными $x_1,...,x_n$ и линейная форма относительно этих же неизвестных:

$$F = c_1x_1 + ... + c_nx_n + c_0. \qquad (7.11)$$

Требуется среди всех неотрицательных решений системы (10) выбрать такое, при котором форма F принимает наименьшее значение (минимизируется).

Определение: Система (7.10) называется системой ограничений данной задачи.

Сами равенства (7.10) называются ограничениями-равенствами. Отметим, что кроме ограничений-равенств в основу задач входят также ограничения-неравенства $x_1 \geq 0,..., x_n \geq 0.$

Определение: Всякое неотрицательное решение $x_1^{(0)},...,x_n^{(0)}\left(x_i^{(0)} \geq 0;\ i = 1,...,n\right)$ системы (7.10) назовем ***допустимым***. Допустимое решение часто называют *планом задачи линейного программирования*.

Определение: Допустимое решение системы (7.10), минимизирующее форму F, назовем ***оптимальным***.

Рассмотрим следующие ограничения системы (7.10).

Задача имеет смысл лишь в том случае, когда система (7.10) совместна, то есть когда (согласно теореме Кронекера-Капелли) ранги основной и расширенной матриц системы совпадают. Этот общий ранг r не может превосходить числа n неизвестных. При $r = n$ решение $x_1^{(0)},...,x_n^{(0)}$ системы (7.10) единственно. Если это решение допустимо, то оно является оптимальным, так как никаких

других решений вообще нет. Если же среди чисел $x_i^{(0)}$ имеется хотя бы одно отрицательное, то задача не имеет решения.

Таким образом, интерес представляет лишь случай $r < n$, и только он будет нами рассматриваться в дальнейшем.

Каждую задача линейного программирования можно свести к форме основной задачи. Для этого нужно:

1. Уметь сводить задачу максимизации к задаче минимизации.

2. Уметь переходить от ограничений, заданных в виде неравенств, к некоторым им эквивалентным ограничениям-равенствам.

Покажем, как это сделать.

1. Очевидно, что форма F достигает наибольшей величины при тех же самых значениях неизвестных $x_1^{(0)},...,x_n^{(0)}$, при которых форма $F_1 = -F$ достигает наименьшей величины. Следовательно, максимизация формы F равносильна минимизации формы $F_1 = -F$. Тем самым задача максимизации сводится к задаче минимизации.

2. Допустим теперь, что среди ограничений задачи имеется некоторое неравенство. Его всегда можно записать в виде

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \beta \geq 0. \qquad (7.12)$$

Введем новую, так называемую добавочную, неизвестную, связанную с неизвестными $x_1,...,x_n$ уравнением

$$x_{n+1} = \alpha_1 x_1 + ... + \alpha_n x_n + \beta. \qquad (7.13)$$

Очевидно, что условие неотрицательности величины x_{n+1} эквивалентно выполнимости неравенства (7.12). Иными словами, если система $x_1^{(0)},...,x_n^{(0)},x_{n+1}^{(0)}$ неотрицательных значений $x_1,...,x_n,x_{n+1}$ удовлетворяет уравнению (7.13), то система $x_1^{(0)},...,x_n^{(0)}$ удовлетворяет неравенству (7.12). И обратно, если величины $x_1^{(0)},...,x_n^{(0)}$ неотрицательны и удовлетворяют неравенству (7.12), то величина $x_{n+1} = \alpha_1 x_1 + ... + \alpha_n x_n + \beta$, найденная из уравнения (7.13), окажется неотрицательной.

Итак, ограничение-наравенство (7.12) эквивалентно ограничению-равенству (7.13). Следовательно, ценою введения в задачу добавочных неизвестных удается все ограничения-неравенства заменить ограничениями-равенствами. При этом число добавочных неизвестных равно числу ограничений-неравенств в исходной задаче.

Посмотрим, какой вид примут рассмотренные выше задачи после сведения их к основной.

Задача 1: Если все члены неравенств системы ограничений (7.1) этой задачи перенести в правую часть, то они примут следующий вид:

$$\left.\begin{aligned} &0 \le 19 - 2x_1 - 3x_2, \\ &0 \le 13 - 2x_1 - x_2, \\ &0 \le 15 - 3x_2, \\ &0 \le 18 - 3x_1. \end{aligned}\right\} \qquad (7.14)$$

Введем четыре добавочные неизвестные и перейдем к новой системе ограничений-равенств:

$$\left.\begin{aligned} &x_3 = 19 - 2x_1 - 3x_2, \\ &x_4 = 13 - 2x_1 - x_2, \\ &x_5 = 15 - 3x_2, \\ &x_6 = 18 - 3x_1. \end{aligned}\right\} \qquad (7.15)$$

Вместо максимизации формы $F = 7x_1 + 5x_2$ будем минимизировать форму $F_1 = -7x_1 - 5x_2$. А это и есть основная задача линейного программирования.

Задача 2: Перепишем систему (7.3) ограничений этой задачи в виде:

$$\left.\begin{aligned} &0 \le T - x_1 - x_2, \\ &0 \le T - x_3 - x_4, \\ &a_1x_1 + b_1x_3 = N_1, \\ &a_2x_2 + b_2x_4 = N_2. \end{aligned}\right\} \qquad (7.16)$$

Введем добавочные неизвестные и перейдем к системе ограничений-равенств

$$\left.\begin{aligned} &x_5 = T_1 - x_1 - x_2, \\ &x_6 = T_2 - x_3 - x_4, \\ &a_1x_1 + b_1x_3 = N_1, \\ &a_2x_2 + b_2x_4 = N_2. \end{aligned}\right\} \qquad (7.17)$$

Среди неотрицательных решений системы (7.17) следует выбрать оптимальное для формы $F = \alpha_1x_1 + \alpha_2x_2 + \beta_1x_3 + \beta_2x_4$.

Задача 3: Транспортная задача уже имеет вид основной.

Задача 4: Систему ограничений-неравенств (7.8) заменяем системой

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + b_1 &= x_3, \\ a_{21}x_1 + a_{22}x_2 + b_2 &= x_4, \\ a_{31}x_1 + a_{32}x_2 + b_3 &= x_5, \\ a_{41}x_1 + a_{42}x_2 + b_4 &= x_6, \\ a_{51}x_1 + a_{52}x_2 + b_5 &= x_7. \end{aligned}\right\} \qquad (7.18)$$

ограничений-равенств. Минимизируемая форма $F = c_1x_1 + c_2x_2$ остается без изменения.

Часто оказывается полезным, а для некоторых целей (например, для геометрического толкования) необходимым свести основную задачу линейного программирования к другой, эквивалентной ей задаче. Все ограничения этой эквивалентной задачи состоят только из неравенств. К рассмотрению этой задачи мы и переходим.

7.3. Основная задача линейного программирования с ограничениями-неравенствами

Рассмотрим систему (7.10) ограничений-равенств основной задачи и рассмотрим случай $r < n$ (r – ранг системы, n – число неизвестных). Из линейной алгебры известно, что в этом случае r неизвестных линейно выражаются через остальные $r - n = k$ неизвестных (эти последние принято называть свободными). Всегда можно занумеровать неизвестные так, чтобы последние $x_{k+1},...,x_n$ выразились через первые $x_1,...,x_k$ (поскольку $n = k - r$, то неизвестных $x_{k+1},...,x_n$ в точности r штук).

$$\left.\begin{aligned} x_{k+1} &= \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1k}x_k + \beta_1, \\ x_{k+2} &= \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2k}x_k + \beta_2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= \alpha_{r1}x_1 + \alpha_{r2}x_2 + \ldots + \alpha_{rk}x_k + \beta_r. \end{aligned}\right\} \qquad (7.19)$$

Система (7.19) эквивалентна системе (7.10).

Если теперь вместо величин $x_{k+1},...,x_n$ в выражении для формы F подставить их значения из (7.19), то форма F примет другой вид:

$$F = \gamma_0 + \gamma_1x_1 + ... + \gamma_kx_k, \qquad (7.20)$$

то есть по-прежнему остается линейной, но выразится только через свободные неизвестные $x_1,...,x_k$.

Решая основную задачу, мы ограничиваемся лишь неотрицательными решениями системы (10) (или системы (7.19)). Поэтому с необходимостью должны выполняться неравенства

$$x_i \geq 0, (i = 1,2,...,k). \qquad (7.21)$$

Если при этом учесть равенства (7.19), то получаем для $i = k+1,\ k+2,...,n)$ неравенства

$$\left.\begin{array}{l} 0 \le \alpha_{11}x_1 + ... + \beta_1, \\ 0 \le \alpha_{21}x_1 + ... + \beta_2, \\ ..... \\ 0 \le \alpha_{r1}x_1 + ... + \beta_r. \end{array}\right\}$$

Для свободных неизвестных $x_i (i = 1,2,...,k)$ переписываем неравенства (7.21) и приходим к системе

$$\left.\begin{array}{l} x_1 \ge 0, \\ ..... \\ x_k \ge 0, \\ \alpha_{11}x_1 + \ldots + \alpha_{1k}x_k + \beta_1 \ge 0, \\ ..... \\ \alpha_{r1}x_1 + \ldots + \alpha_{rk}x_k + \beta_r \ge 0, \end{array}\right\} \qquad (7.22)$$

содержащей n линейных неравенств относительно k неизвестных. Ясно, что каждому допустимому решению системы (7.19) отвечает решение системы неравенств (7.22). И обратно, взяв произвольное решение $x_1^{(0)},...,x_k^{(0)}$ системы (7.22) и найдя по формулам (7.19) величин $x_{k+i}^{(0)} = \alpha_{i1}^{(0)}x_1 + ... + \alpha_{ik}^{(0)}x_k + \beta_i$, $(i = 1,...,r)$, получим, очевидно, неотрицательное решение системы (7.19), а значит, и системы (7.10). Тем самым вместо того, чтобы искать неотрицательные решения системы (7.10), можно искать все решения системы неравенств (7.22).

В результате мы пришли к следующей математической задаче.

Дана система (7.22), содержащая n линейных неравенств относительно k неизвестных $x_1,\ldots,x_k$ и линейная форма (7.20) относительно тех же неизвестных.

Требуется среди всех решений системы (7.22) выбрать такое, при котором форма F достигает наименьшего значения. Такое решение будем называть, как и прежде, оптимальным.

Из хода наших рассуждений очевидно, что основная задача линейного программирования эквивалентна только что сформулированной задаче. Эту задачу естественно называть основной задачей линейного программирования с ограничениями-неравенствами. Для раткости назовем ее основной задачей в форме (A) или просто задачей (A).

Проиллюстрируем наши рассуждения на задачах 1 – 4.

Задача 1: В результате сведения этой задачи к основной мы пришли к следующей системе ограничений:

$$\left.\begin{aligned} x_3 &= 19 - 2x_1 - 3x_2, \\ x_4 &= 13 - 2x_1 - x_2, \\ x_5 &= 15 - 3x_2, \\ x_6 &= 18 - 3x_1. \end{aligned}\right\} \qquad (7.15)$$

и форме $F = -7x_1 - 5x_2$.

Выпишем матрицу I системы (7.15):

$$I = \begin{pmatrix} 2 & 3 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Очевидно, что ранг ее $r = 4$, так как минор, составленный из четырех последних столбцов, не равен нулю. Ранг r рассмотренной матрицы не может быть, естественно, больше 4 (матрица I имеет лишь 4 строки). Таким образом, $r = 4,\ n = 6,$ а $k = n - r = 2$.

Согласно общей схеме теперь следует выразить четыре неизвестные через оставшиеся две. Из системы (7.15) видно, что четыре последние неизвестные $x_3,\ x_4,\ x_5,\ x_6$ уже выражены через первые неизвестные $x_1,\ x_2$. Форма F также выражается через $x_1,\ x_2$. Потребовав неотрицательности всех неизвестных, приходим к системе неравенств

$$\left.\begin{aligned} &x_1 \geq 0, \\ &x_2 \geq 0, \\ &19 - 2x_1 - 3x_2 \geq 0, \\ &13 - 2x_1 - x_2 \geq 0, \\ &15 - 3x_2 \geq 0, \\ &18 - 3x_1 \geq 0. \end{aligned}\right\} \qquad (7.23)$$

Тем самым эта задача приобрела форму задачи (А).

Задача 2: Конкретизируем задачу, придав коэффициентам следующие числовые значения:

$a_1 = 6,\ a_2 = 24,\ b_1 = 13,\ b_2 = 13,\ N_1 = 30,\ N_2 = 96,\ T = 6,\ \alpha_1 = 4,\ \alpha_2 = 47,$
$\beta_1 = 13,\ \beta_2 = 26.$

Тогда система ограничений и форма F примут вид:

$$\left.\begin{aligned} &x_1 + x_2 \le 6, \\ &x_3 + x_4 \le 6, \\ &6x_1 + 13x_3 = 30, \\ &24x_2 + 13x_4 = 96. \end{aligned}\right\} \qquad (7.24)$$

$$F = 4x_1 + 47x_2 + 13x_3 + 26x_4$$

Оставив пока в стороне первые два ограничения-неравенства, выразим из ограничений-равенств x_3 и x_4 через x_1 и x_2 (неизвестные x_1 и x_2 в нашей задаче мы выбираем в качестве свободных):

$$\left.\begin{aligned} x_3 &= \frac{1}{13}(30 - 6x_1), \\ x_4 &= \frac{1}{13}(96 - 24x_2). \end{aligned}\right\} \qquad (7.25)$$

Далее, подставим в левую часть исходных ограничений-неравенств и в форму F выражения несвободных неизвестных через свободные, в результате чего получим

$$x_1 + 4x_2 \ge 8,$$
$$F = 222 - 2x_1 - x_2.$$

Требование неотрицательности x_3 и x_4 эквивалентно неравенствам

$$\left.\begin{aligned} &\frac{1}{13}(30 - 6x_1) \ge 0, \\ &\frac{1}{13}(96 - 24x_2) \ge 0, \end{aligned}\right\}$$

которые после элементарных преобразований дают $x_1 \le 5$, $x_2 \le 4$. Учитывая теперь исходное неравенство $x_1 + x_2 \le 6$ и неотрицательность всех величин, получим следующую систему ограничений-неравенств:

$$\left.\begin{aligned} &x_1 \ge 0, \\ &x_2 \ge 0, \\ &x_1 \le 5, \\ &x_2 \le 4, \\ &x_1 + x_2 \le 6, \\ &x_1 + 4x_2 \ge 8. \end{aligned}\right\} \qquad (7.26)$$

При этом минимизируемая форма имеет вид

$$F = 222 - 2x_1 - x_2. \qquad (7.27)$$

Таким образом, мы пришли к задаче (А). Также можно свести к форме (А) задачу 3 и задачу 4.

7.4. Геометрическое толкование задач линейного программирования

Геометрический смысл основной задачи линейного программирования становится предельно прозрачным, если перейти от нее к эквивалентной ей задаче с ограничениями-неравенствами, то есть к задаче (А):

Дана система

$$\left.\begin{array}{l} x_1 \ge 0, \\ x_2 \ge 0, \\ \ldots\ldots\ldots \\ x_k \ge 0, \\ \alpha_{11}x_1 + \ldots + \alpha_{1k}x_k + \beta_1 \ge 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{r1}x_1 + \ldots + \alpha_{rk}x_k + \beta_r \ge 0, \end{array}\right\} \qquad (7.22)$$

содержащая $n = k = r$ линейных неравенств относительно k неизвестных $x_1, \ldots, x_k$, и линейная форма

$$F = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_k x_k \qquad (7.20)$$

относительно тех же неизвестных. Требуется среди всех решений системы (7.22) выбрать такое, которое минимизирует форму F.

Для большей наглядности геометрическому толкованию основной задачи в общем случае предпошлем разбор задач 1–4.

Рассмотрим задачу 1 в форме (А), она имеет вид:

$$\left.\begin{array}{l} x_1 \ge 0, \\ x_2 \ge 0, \\ 19 - 2x_1 - 3x_2 \ge 0, \\ 13 - 2x_1 - x_2 \ge 0, \\ 15 - 3x_2 \ge 0, \\ 18 - 3x_1 \ge 0. \end{array}\right\} \qquad (3.23)$$

$$F_1 = -7x_1 - 5x_2.$$

Введем на плоскости прямоугольную декартову систему координат x_1Ox_2. Известно, что геометрическое место точек на плоскости, координаты которых удовлетворяют системе линейных неравенств, образует выпуклый многоугольник. Этот многоугольник называется многоугольником решений данной системы неравенств. Стороны этого многоугольника располагаются на прямых, уравнения которых получаются, если в неравенствах системы знаки неравенств заменить точными равенствами. Сам же этот многоугольник есть пересечение полуплоскостей, на которые делит плоскость каждая из указанных прямых. В нашем случае такими прямыми являются

$$\left.\begin{aligned} &x_1 = 0, && \text{(I)}\\ &x_2 = 0, && \text{(II)}\\ &19 - 2x_1 - 3x_2 = 0, && \text{(III)}\\ &13 - 2x_1 - x_2 = 0, && \text{(IV)}\\ &15 - 3x_2 = 0, && \text{(V)}\\ &18 - 3x_1 = 0. && \text{(VI)} \end{aligned}\right\}$$

Вычертим эти прямые (см. рис. 32); стрелки указывают, какие полуплоскости в пересечении дают многоугольник решений.

Наряду с этим рассмотрим форму $F_1 = -7x_1 - 5x_2$. Она, очевидно, является линейной функцией координат (x_1, x_2) точки на плоскости. Поставим такой вопрос: как располагаются на плоскости те точки, в которых форма F_1 принимает одно и то же значение C? Для ответа на поставленный вопрос достаточно форму F_1 приравнять C и рассмотреть полученное уравнение

$$-7x_1 - 5x_2 = C. \qquad (7.28)$$

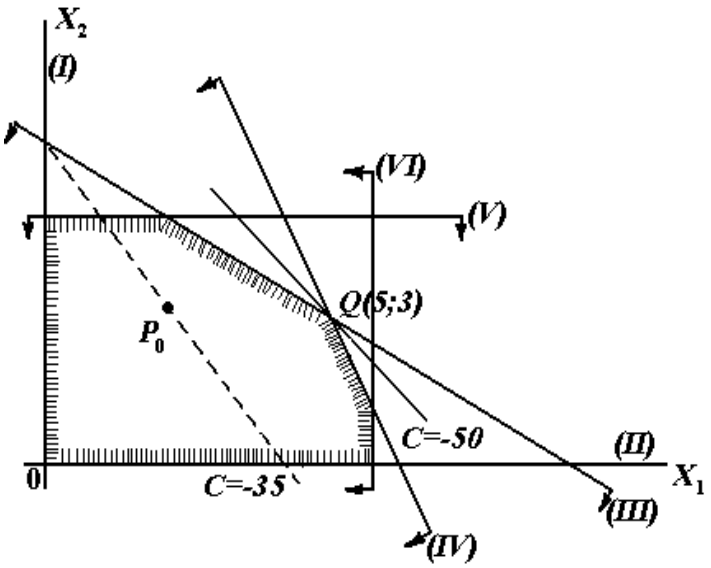


Рис. 32. *Иллюстрация к задаче 1*

Уравнение (7.28) определяет на плоскости некоторую прямую. Она и является искомым геометрическим местом точек, в которых F_1 принимает данное значение C (рис. 32).

Меняя значение C, получаем различные прямые, однако все они параллельны между собой, то есть образуют семейство параллельных прямых. Очевидно, что через каждую точку плоскости проходит одна прямая этого семейства. Каждую из прямых семейства (7.28) принято называть линией уровня (линией равных значений) формы F_1. При переходе от одной прямой к другой значение формы F_1 изменяется. На рис. 33 показаны прямые, отвечающие значениям C = –17.5; –35; –52.5. Вектор g указывает направление, двигаясь в котором мы переходим от больших значений формы к меньшим.

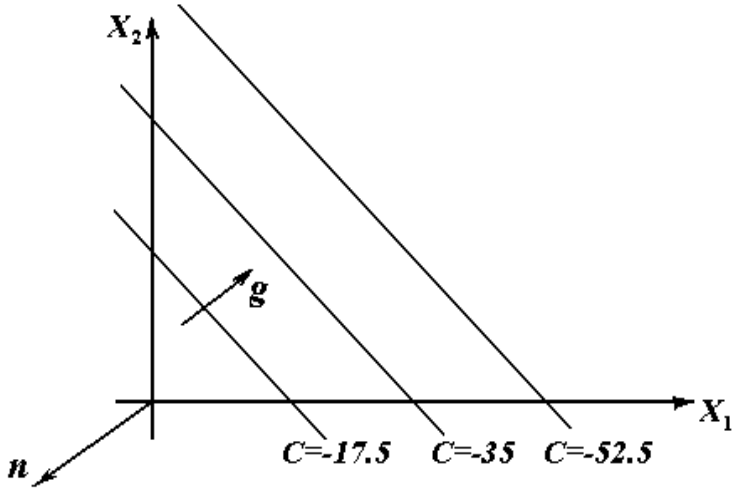


Рис. 33. *Линии уровня функции формы*

Из аналитической геометрии известно, что коэффициенты при переменных в уравнении прямой – это проекции вектора нормали *n*, перпендикулярного прямой. В нашем случае *n*={–7; –5}. Мы наблюдаем, что направление убывания формы F_1 противоположно направлению вектора *n*.

Обратимся вновь к рис. 32. Рассмотрим любую точку $P_0\left(x_1^{(0)}, x_2^{(0)}\right)$ многоугольника решений. Через эту точку проходит прямая семейства (7.28). Вдоль всей этой прямой форма F_1 принимает такое же значение, как и в точке P_0, т. е. $C_0 = -7x_1^{(0)} - 5x_2^{(0)}$.

На рис. 32 пунктиром показана прямая, отвечающая значению *C*=–35. Очевидно, что точка P_0 не соответствует оптимальному решению задачи. Действительно, внутри многоугольника решений можно найти точки, отвечающие значениям формы меньшим, чем C_0. Для этого достаточно перейти в направлении вектора *g* от прямой C_0 к другой параллельной ей прямой семейства (7.28), все еще пересекающей многоугольник решений. Теперь должно быть ясным, что оптимальное решение определится точкой *Q*(5, 3), а наименьшее значение формы F_1 равно

$$F_{1\min} = -7 \cdot 5 - 5 \cdot 3 = -50.$$

Итак, оптимальное решение задачи 1 найдено: $x_1 = 5,\ x_2 = 3$.

Если вспомнить условие этой задачи, то мы видим, что для наиболее рационального плана использования сырья, гарантирующего предприятию наибольший доход, следует выпускать 5 единиц продукции вида $П_1$ и 3 единицы вида $П_2$. При этом максимальный доход составит $F_{\max}$ = 50. Отметим, что при этом сырье видов S_1 и S_2 используется полностью, а S_3 и S_4 не полностью.

Замечание. Как известно, задача нахождения экстремальных точек функции рассматривается в курсе математического анализа. Там она решается методами дифференциального исчисления. Почему же, спрашивается, нельзя использовать эти методы для решения задач линейного программирования? Дело в том, что методы дифференциального исчисления позволяют определить только такие

экстремальные точки, которые находятся строго внутри рассматриваемой области, а не на ее границе. В задачах же линейного программирования оптимальное значение формы достигается всегда на границе многоугольника решений, как, например, в задаче 1. Вот почему методы дифференциального исчисления неприменимы для решения таких задач.

Задача 2: Дадим геометрическое толкование и найдем решение задачи 2. После приведения ее к задаче в форме (А) мы получили следующие ограничения и форму:

$$\left.\begin{aligned} &x_1 \ge 0, && \text{(I)}\\ &x_2 \ge 0, && \text{(II)}\\ &x_1 \le 5, && \text{(III)}\\ &x_2 \le 4, && \text{(IV)}\\ &x_1 + x_2 \le 6, && \text{(V)}\\ &x_1 + 4x_2 \ge 8. && \text{(VI)} \end{aligned}\right\} \qquad (7.29)$$

$$F = 222 - 2x_1 - x_2. \qquad (7.30)$$

Введем на плоскости систему координат x_1Ox_2 и вычерти многоугольник решений системы (7.29) подобно тому, как это сделано в задаче 1. Для этого заменим все неравенства из (7.29) равенствами и построим соответствующие прямые (см. рис. 34). На этом же рисунке построена одна из линий уровня формы (7.30), отвечающая значению F = 219. Вектор g указывает направление убывания формы F.

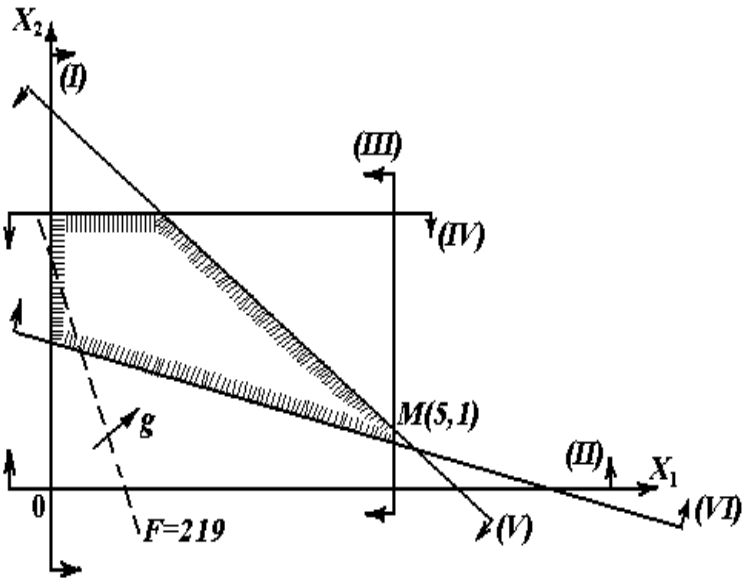


Рис. 34. *Иллюстрация к задаче 2*

Из рисунка видно, что наименьшее значение F на многоугольнике решений достигается в точке M(5, 1) пересечения прямых (III) и (V). Минимальное значение формы при $x_1 = 5,\ x_2 = 1$ находится из (7.30):

$$F = 222 - 2 \cdot 5 - 1 = 211.$$

Каков же экономический смысл полученного решения? Напомним, что величины $x_i\,(i = 1,2,3,4)$ означают времена работы машин A и B по изготовлению продукции $П_1$ и $П_2$. Величины $x_1 = 5$ и $x_2 = 1$ нами уже найдены, а величины x_3 и x_4 находим из уравнений (7.25):

$$\left.\begin{aligned} x_3 &= \frac{1}{13}(30 - 6x_1) = \frac{1}{13}(30 - 30) = 0, \\ x_4 &= \frac{1}{13}(96 - 24x_2) = \frac{1}{13}(96 - 24) = \frac{72}{13}. \end{aligned}\right\}$$

Таким образом, задача решена полностью.

Задача 3: Перейдем к геометрическому толкованию и заодно решим задачу 3. Ограничения и минимизируемая форма этой задачи (после приведения ее к виду (A)) таковы:

$$\left.\begin{aligned} &20 - x_{11} - x_{12} \geq 0, && \text{(I)} \\ &10 - x_{11} \geq 0, && \text{(II)} \\ &30 - x_{12} \geq 0, && \text{(III)} \\ &-10 + x_{11} + x_{12} \geq 0, && \text{(IV)} \\ &x_{11} \geq 0, && \text{(V)} \\ &x_{12} \geq 0, && \text{(VI)} \end{aligned}\right\} \qquad (7.31)$$

$$F = 330 - 2x_{11} - x_{12}. \qquad (7.32)$$

Введем систему координат на плоскости. На этот раз оси обозначим x_{11} и x_{12}. Вычертим многоугольник решений и одну из линий уровня формы F (см. рис. 35).

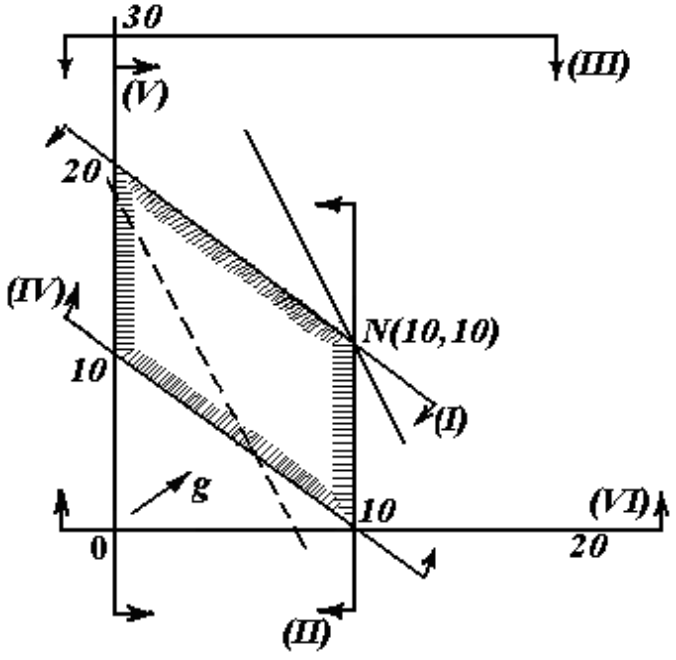


Рис. 35. *Иллюстрация к задаче 3*

Оптимальное решение дается точкой $N(10,10)$. Итак, $x_{11}=10$, $x_{12}=10$, $F_{\min}=300$.

Остальные значения x_{ij} находятся из уравнений (7.31):

$$\left.\begin{aligned} x_{13} &= 20 - x_{11} - x_{12} = 0, \\ x_{21} &= 10 - x_{11} = 0, \\ x_{22} &= 30 - x_{12} = 20, \\ x_{23} &= -10 + x_{11} + x_{12} = 10. \end{aligned}\right\}$$

Задача 4: Решим задачу о питании, принимая для известных по условию задачи величин следующие числовые значения (они носят иллюстрированный характер и не соответствуют действительности):

Таблица 9

Питательные вещества	Минимальная норма	Виды пищи	
		π_1	π_2
B_1-жиры	10	1	5
B_2-белки	12	3	2
B_3-углеводы	16	2	4
B_4-вода	10	2	2
B_5-витамины	1	1	0
Стоимость		2	3

Ограничения (7.8) и минимизируемая форма (7.9) задачи 4, при выбранных выше данных примут вид:

$$\left.\begin{aligned} x_1 + 5x_2 &\geq 10, && \text{(I)} \\ 3x_1 + 2x_2 &\geq 12, && \text{(II)} \\ 2x_1 + 4x_2 &\geq 16, && \text{(III)} \\ 2x_1 + 2x_2 &\geq 10, && \text{(IV)} \\ x_1 &\geq 1, && \text{(V)} \\ x_2 &\geq 0, && \text{(VI)} \end{aligned}\right\} \qquad (7.33)$$

$$F = 2x_1 + 3x_2. \qquad (7.34)$$

(Неравенство $x_1 \geq 0$ не включено в систему (7.33) потому, что оно является следствием неравенства (V): $x_1 \geq 1$.)

Вычертим многоугольник решений и одну из линий уровня формы F (рис. 36).

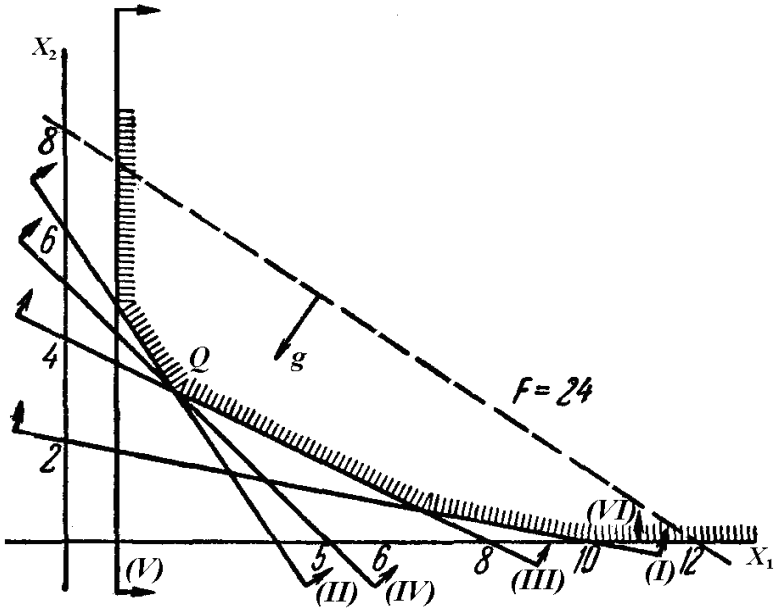


Рис. 36. *Иллюстрация к задаче о питании*

Оптимальное решение достигается в точке $Q(2;3)$, следовательно, $x_1 = 2,\ x_2 = 3,\ F = 13$.

Отметим, что в рассматриваемой задаче «многоугольник» решений не-ограничен сверху и потому не существует на многоугольнике наибольшего значения F. Это означает, очевидно, что питание можно организовать сколь угодно дорого.

Замечание: Изменим теперь стоимости видов пищи π_1 и π_2. Зададим их равными соответственно 3 и 2 единицам. В этом случае минимизировать следует форму

$$F = 3x_1 + 2x_2.$$

Линии уровня этой формы будут параллельны стороне $PQ\{\text{II}\}$ многоугольника решений. Оптимум формы достигается в любой точке стороны PQ (эта сторона располагается на линии уровня $F' = 12$).

Вывод: решенные задачи свидетельствуют о справедливости следующего положения:

если оптимальное решение задачи существует и единственно, то оно достигается в некоторой вершине многоугольника решений. Если же оптимальное решение не единственное, то таких решений бесчисленное множество, и они достигаются во всех точках некоторой стороны (и, в частности, в ограничивающих эту сторону вершинах) многоугольника.

Таким образом, всегда найдется вершина многоугольника решений, в которой достигается оптимальное решение (если оно, конечно, существует).

В задачах 1 – 4, приведенных к форме (А), число переменных, входящих в систему ограничений-неравенств, равнялось двум. Это обстоятельство давало возможность изобразить область решений системы неравенств в виде многоугольника на плоскости.

Рассмотрим простой пример задачи (А), когда число переменных равно трем.

Зададим систему неравенств

$$\left.\begin{aligned} &x_1 \geq 0,\\ &x_2 \geq 0,\\ &x_3 \geq 0,\\ &x_1 \leq 2,\\ &x_1 + x_2 + x_3 \leq 4. \end{aligned}\right\} \qquad (7.35)$$

Требуется в области решений этой системы минимизировать форму

$F = -x_1 - 2x_2 - 3x_3.$

Введем в пространстве прямоугольную систему координат x_1, x_2, x_3. Известно, что геометрическое место точек, координаты которых удовлетворяют системе линейных неравенств, образует выпуклый многогранник, называемый многогранником решений данной системы. Грани этого многогранника расположены на плоскостях, уравнения которых получаются, если в неравенствах системы знак неравенства заменить точным равенством. Сам многогранник решений является пересечением полупространств, на которые делит пространство каждая из указанных плоскостей. Вычертим эти плоскости (рис. 37).

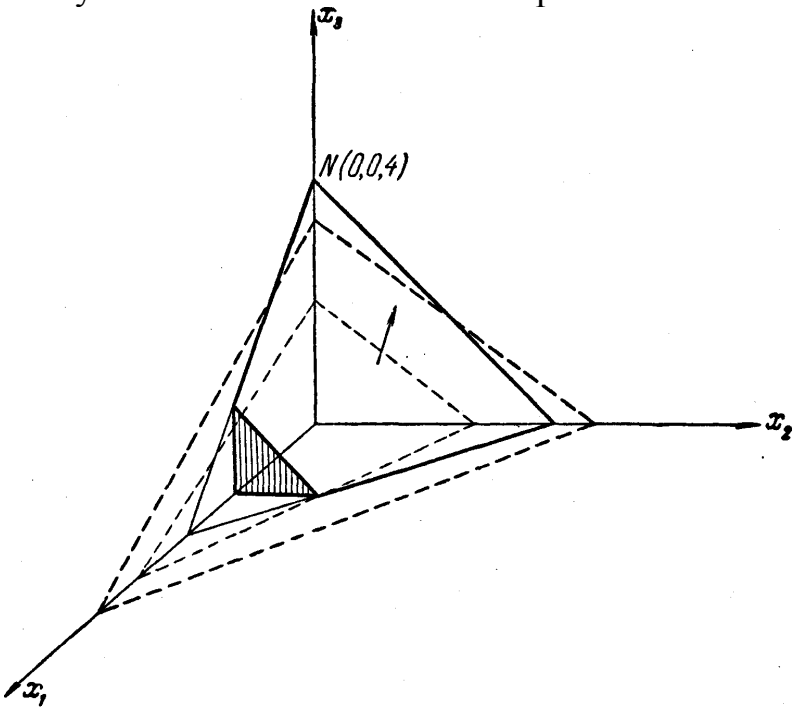


Рис. 37. *Иллюстрация к трехмерной задаче*

Форма $F = -x_1 - 2x_2 - 3x_3$ является линейной функцией координат (x_1, x_2, x_3) точек пространства. Координаты всех точек, в которых форма принимает одно и то же фиксированное значение C, удовлетворяют уравнению $F = C$, или, подробнее,

$$F = -x_1 - 2x_2 - 3x_3 = C. \qquad (7.36)$$

Это уравнение определяет в пространстве плоскость, называемую поверхностью уровня (поверхностью равных значений) формы *F*. Придавая *C* различные значения, получим семейство (7.36) параллельных между собой плоскостей. Подчеркнем снова, что перемещение по любой плоскости семейства (7.36) от одной ее точки к другой не вызывает изменений формы. В то же время переход от одной плоскости к другой сопровождается изменением значений формы *F*.

На рис. 37 изображены две плоскости

$$-x_1 - 2x_2 - 3x_3 = -6,$$

$$-x_1 - 2x_2 - 3x_3 = -10,$$

отвечающие значениям *F*, равным соответственно –6 и –10. Вектор на рис. 37 указывает направление, двигаясь в котором мы переходим от больших значений формы к меньшим. Известно, что коэффициенты при переменных в уравнении плоскости – это проекции вектора n, перпендикулярного к плоскости. Мы видим, что в нашем случае вектор $n = \{-1, -2, -3\}$ и его направление противоположно направлению, в котором убывает форма *F*. Из рис. 37 видно, что наименьшее значение формы достигается в вершине $N(0,0,4)$.

Следовательно, оптимальным будет решение

$$x_1 = 0,\ x_2 = 0,\ x_3 = 4,$$

$$F_{\min} = -12.$$

Замечание: Отметим, что в пространстве, как и на плоскости, оптимальное решение (если оно существует) достигается в некоторой вершине многоугольника решений.

Вернемся к геометрическому толкованию общей задачи (А). Напомним вид ее ограничений и форму:

$$\left.\begin{array}{l} x_1 \ge 0, \\ x_2 \ge 0, \\ \ldots\ldots\ldots \\ x_k \ge 0, \\ \alpha_{11}x_1 + \ldots + \alpha_{1k}x_k + \beta_1 \ge 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \alpha_{r1}x_1 + \ldots + \alpha_{rk}x_k + \beta_r \ge 0, \end{array}\right\} \qquad (7.22)$$

$$F = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_k x_k . \qquad (7.20)$$

Рассмотрим *k*-мерное пространство, в котором введем систему координат $x_1, x_2, \ldots, x_k$.

Известно, что областью решений системы (7.22) является некоторый выпуклый многогранник в рассматриваемом пространстве. Грани этого многогран-

ника лежат в гиперплоскостях, чьи уравнения можно получить, если всюду в (7.22) знаки неравенства заменить знаками точных равенств. Приравнивая форму F всевозможным постоянным C, получим уравнения

$$\gamma_0 + \gamma_1 x_1 + ... + \gamma_k x_k = C$$

семейства гиперплоскостей равных значений формы F.

При переходе от одной гиперплоскости к другой меняется значение F. По аналогии с плоским и трехмерным случаями имеет место общая теорема:

Имеет место общая

Теорема: Если оптимальное решение задачи (А) существует, то оно достигается на некоторой вершине многогранника решений.

(В принципе, возможен случай, когда оптимальных решений бесконечное множество, и все они лежат на некоторой гиперграни многогранника решений.)

Несмотря на кажущуюся наглядность этого факта, доказательство его выходит за рамки нашего курса и далеко не просто.

Решение таких многомерных задач весьма сложно и поэтому разработаны специальные вычислительные методы, один из которых называется симплекс метод.

7. СИМПЛЕКС МЕТОД ИЛИ МЕТОД ПОСЛЕДОВАТЕЛЬНОГО УТОЧНЕНИЯ ОЦЕНОК

Метод предназначен для решения общей задачи линейного программирования.

Пусть имеем следующую задачу:

$$Q\left(\bar{x}\right) = c_1 \cdot x_1 + c_2 \cdot x_2 + ... + c_n \cdot x_n \to \min,$$

с системой ограничений следующего вида:

$$\begin{cases} a_{11} \cdot x_1 + a_{12} \cdot x_2 + ... + a_{1n} \cdot x_n = b_1, \\ ..... \\ a_{m1} \cdot x_1 + a_{m2} \cdot x_2 + ... + a_{mn} \cdot x_n = b_m. \end{cases}$$

Разрешим эту систему относительно переменных $x_1, ..., x_m$:

$$\begin{cases} x_1 = a'_{1,m+1} \cdot x_{m+1} + ... + a'_{1,n} \cdot x_n + b'_1, \\ ..... \\ x_m = a'_{m,m+1} \cdot x_{m+1} + ... + a'_{m,n} \cdot x_n + b'_m. \end{cases} \qquad (7.3)$$

Векторы условий, соответствующие $x_1, ..., x_m$, образуют базис. Переменные $x_1, ..., x_m$ назовем базисными переменными. Остальные переменные задачи – небазисные.

Целевую функцию можно выразить через небазисные переменные:

$$Q\left(\bar{x}\right) = c'_{m+1} \cdot x_{m+1} + c'_{m+2} \cdot x_{m+2} + ... + c'_n \cdot x_n + c'_0 \to \min.$$

Если приравнять небазисные переменные нулю

$$x_{m+1} = 0,\ x_{m+2} = 0,\ ...;\ x_n = 0,$$

то соответствующие базисные переменные примут значения

$$x_1 = b'_1;\ x_2 = b'_2;\ ...;\ x_m = b'_m.$$

Вектор $\bar{x}$ с такими компонентами представляет собой угловую точку многогранника решений (допустимую) при условии, что $b'_i \geq 0$ (опорный план).

Теперь необходимо перейти к другой угловой точке с меньшим значением целевой функции. Для этого следует выбрать некоторую небазисную переменную и некоторую базисную так, чтобы после того, как мы «поменяем их местами», значение целевой функции уменьшилось. Такой направленный перебор в конце концов приведет нас к решению задачи.

Пример 1. Пусть

$$Q\left(\bar{x}\right) = x_4 - x_5 \to \min,$$

$$\left.\begin{aligned} x_1 + x_4 - 2x_5 &= 1, \\ x_2 - 2x_4 + x_5 &= 2, \\ x_3 + 3x_4 + x_5 &= 3. \end{aligned}\right\}$$

Выберем в качестве базисных следующие переменные $\{x_1, x_2, x_3\}$ и разрешим систему относительно этих переменных. Система ограничений примет следующий вид:

$$\left.\begin{aligned} x_1 &= 1 - x_4 + 2x_5, \\ x_2 &= 2 + 2x_4 - x_5, \\ x_3 &= 3 - 3x_4 - x_5. \end{aligned}\right\}$$

Переменные $\{x_4, x_5\}$ являются небазисными. Если взять $x_4 = 0$ и $x_5 = 0$, то получим угловую точку (опорный план)

$$\overset{-1}{x} = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 \end{bmatrix}^T,$$

которому соответствует $Q\left(\overset{-1}{x}\right) = 0$.

Значение целевой функции можно уменьшить за счет увеличения x_5. При увеличении x_5 величина x_1 также увеличивается, а x_2 и x_3 – уменьшаются. Причем величина x_2 раньше может стать отрицательной. Поэтому, вводя в базис переменную x_5, одновременно x_2 исключаем из базиса. В результате после очевидных преобразований получим следующие выражения для новой системы базисных переменных и целевой функции:

$$\left.\begin{aligned} x_5 &= 2 - x_2 + 2x_4, \\ x_1 &= 5 - 2x_2 + 3x_4, \\ x_3 &= 1 + x_2 - 5x_4, \end{aligned}\right\}$$

$$Q\left(\bar{x}\right) = -2 - x_4 + x_2 \to \min.$$

Соответствующий опорный план $\overset{-2}{x} = \begin{bmatrix} 5 & 0 & 1 & 0 & 2 \end{bmatrix}^T$ и $Q\left(\overset{-2}{x}\right) = -2$.

Целевую функцию можно уменьшить за счет увеличения x_4. Увеличение x_4 приводит к уменьшению только x_3. Поэтому вводим в базис переменную x_4, а x_3 исключаем из базиса. В результате получим следующие выражения для новой системы базисных переменных и целевой функции:

$$\left.\begin{aligned} x_4 &= \frac{1}{5} + \frac{1}{2}x_2 - \frac{1}{5}x_3, \\ x_1 &= \frac{28}{5} - \frac{7}{5}x_2 - \frac{3}{5}x_3, \\ x_5 &= \frac{12}{5} - \frac{3}{5}x_2 - \frac{2}{3}x_3, \end{aligned}\right\}$$

$$Q\left(\bar{x}\right) = -\frac{11}{5} + \frac{4}{5}x_2 + \frac{1}{5}x_3 \to \min.$$

Соответствующий опорный план $\overset{-3}{x} = \left[\frac{28}{5} \quad 0 \quad 0 \quad \frac{1}{5} \quad \frac{12}{5}\right]^T$ и значение целевой функции $Q\left(\overset{-3}{x}\right) = -\frac{11}{5}$. Так как все коэффициенты при небазисных переменных в целевой функции неотрицательны, то нельзя уменьшить целевую функцию за счет увеличения x_2 или x_3, следовательно, полученный план $\overset{-3}{x}$ является оптимальным.

Пример 2. Пусть имеем задачу

$$Q\left(\bar{x}\right) = -x_1 - x_2 \to \min,$$

$$\left.\begin{aligned} x_3 &= 1 + x_1 - x_2, \\ x_4 &= 2 - x_1 + 2x_2, \\ \bar{x} &\geq 0. \end{aligned}\right\}$$

Переменные $\{x_3, x_4\}$ – базисные, а $\{x_1, x_2\}$ – небазисные переменные. Опорный план $\overset{-0}{x} = [0 \quad 0 \quad 1 \quad 2]^T$, $Q\left(\overset{-0}{x}\right) = 0$.

Теперь вводим в базис переменную x_1, а x_4 исключаем из базиса. В результате получим следующие выражения для базисных переменных и целевой функции:

$$\left.\begin{aligned} x_1 &= 2 + 2x_2 - x_4, \\ x_3 &= 3 + x_2 - x_4, \end{aligned}\right\}$$

$$Q\left(\bar{x}\right) = -2 - 3x_2 + x_4.$$

Опорный план $\overset{-1}{x} = [2 \quad 0 \quad 3 \quad 0]^T$, значение целевой функции $Q\left(\overset{-1}{x}\right) = -2$.

Теперь можно заметить, что при увеличении x_2 значения переменных x_1 и x_3 также возрастают, то есть при $x_2 \to \infty$ в допустимой области $Q\left(\bar{x}\right) \to -\infty$ (задача не имеет решения).

Замечание. В процессе поиска допустимого плана может быть выявлена противоречивость системы ограничений.

7.1. Алгоритм симплекс метода

Формализованный алгоритм симплекс метода состоит из двух основных этапов:

1) построение опорного плана;

2) построение оптимального плана.

Проиллюстрируем алгоритм на рассмотренном ранее примере:

$$Q\left(\bar{x}\right) = x_4 - x_5 \to \min,$$

$$\left.\begin{aligned} x_1 + x_4 - 2x_5 &= 1, \\ x_2 - 2x_4 + x_5 &= 2, \\ x_3 + 3x_4 + x_5 &= 3, \end{aligned}\right\}$$

$$\bar{x} \geq \bar{0}.$$

В случае базисных переменных $\{x_1, x_2, x_3\}$ начальная симплексная таблица для данного примера будет выглядеть следующим образом:

	$-x_4$	$-x_5$	1
$x_1 =$	1	–2	1
$x_2 =$	–2	1	2
$x_3 =$	3	1	3
$Q\left(\bar{x}\right) =$	–1	1	0

Она уже соответствует опорному плану $\bar{x}^{-1} = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 \end{bmatrix}^T$ (столбец свободных членов).

Построение оптимального плана. Для того чтобы опорный план был оптимален, при минимизации целевой функции необходимо, чтобы коэффициенты в строке целевой функции были неположительными (в случае максимизации – неотрицательными). Т.е. при поиске минимума мы должны освободиться от положительных коэффициентов в строке $Q\left(\bar{x}\right)$.

Выбор разрешающего элемента. Если при поиске минимума в строке целевой функции есть коэффициенты больше нуля, то выбираем столбец с поло-

жительным коэффициентом в строке целевой функции в качестве разрешающего. Пусть это столбец с номером l.

Для выбора разрешающей строки (разрешающего элемента) среди положительных коэффициентов разрешающего столбца выбираем тот (строку), для которого отношение коэффициента в столбце свободных членов к коэффициенту в разрешающем столбце минимально:

$$\frac{b_r}{a_{rl}} = \min\left\{\frac{b_i}{a_{il}} \middle| \, a_{il} \geq 0\right\},$$

a_{rl} – разрешающий (направляющий) элемент, строка r – разрешающая.

Для перехода к следующей симплексной таблице (следующему опорному плану с меньшим значением целевой функции) делается шаг модифицированного жорданова исключения с разрешающим элементом a_{rl}.

Если в разрешающем столбце нет положительных коэффициентов, то целевая функция не ограничена снизу (при максимизации – не ограничена сверху).

Шаг модифицированного жорданова исключения над симплексной таблицей.

1. На месте разрешающего элемента ставится 1 и делится на разрешающий элемент.

2. Остальные элементы разрешающего столбца меняют знак на противоположный и делятся на разрешающий элемент.

3. Остальные элементы разрешающей строки делятся на разрешающий элемент.

4. Все остальные элементы симплексной таблицы вычисляются по следующей формуле:

$$a_{ij} = \frac{a_{ij} \cdot a_{rl} - a_{rj} \cdot a_{il}}{a_{rl}} = a_{ij} - \frac{a_{rj} \cdot a_{il}}{a_{rl}}.$$

	$-x_4$	$-x_5$	1
x_1	1	–2	1
x_2	–2	1	2
x_3	3	1	3
$Q(\bar{x})$	–1	1	0

Разрешающий элемент, который соответствует замене базисной переменной x_2 на небазисную переменную x_5.

	$-x_4$	$-x_2$	1
x_1	–3	2	5
x_5	–2	1	2
x_3	5	–1	1
$Q(\bar{x})$	1	–1	–2

Разрешающий элемент, который соответствует замене базисной переменной x_3 на небазисную переменную x_4.

	$-x_3$	$-x_2$	1
x_1	3/5	7/5	28/5
x_5	2/5	3/5	12/5
x_4	1/5	-1/5	1/5
$Q(\bar{x})$	-1/5	-4/5	-11/5

Все коэффициенты в строке целевой функции отрицательны, т.е. мы нашли оптимальное решение.

Построение опорного плана.

Пусть необходимо решить задачу:

$$Q(\bar{x}) = c_1 \cdot x_1 + c_2 \cdot x_2 + ... + c_n \cdot x_n \to \min(\max),$$

$$\begin{cases} a_{1,1} \cdot x_1 + ... + a_{1,n} \cdot x_n = b_1, \\ ..... \\ a_{m,1} \cdot x_1 + ... + a_{m,n} \cdot x_n = b_m, \\ a_{m+1,1} \cdot x_1 + ... + a_{m+1,n} \cdot x_n \le b_{m+1}, \\ ..... \\ a_{m+p,1} \cdot x_1 + ... + a_{m+p,n} \cdot x_n \le b_{m+p}. \end{cases}$$

Введем дополнительные переменные, чтобы преобразовать ограничения-неравенства к равенствам. В ограничениях-равенствах дополнительные переменные должны быть нулевыми. Тогда система ограничений принимает вид:

$$\begin{cases} 0 = b_1 - a_{1,1} \cdot x_1 - ..... - a_{1,n} \cdot x_n, \\ ..... \\ 0 = b_m - a_{m,1} \cdot x_1 - ..... - a_{m,n} \cdot x_n, \\ x_{n+1} = b_{m+1} - a_{m+1,1} \cdot x_1 - ... - a_{m+1,n} \cdot x_n, \\ ..... \\ x_{n+p} = b_{m+p} - a_{m+p,1} \cdot x_1 - ... - a_{m+p,n} \cdot x_n, \end{cases}$$

где $x_{n+i} \ge 0$, $i = 1,..., p$.

В качестве базисных переменных будем брать систему дополнительно введенных переменных. Тогда симплексная таблица для преобразованной задачи будет иметь следующий вид:

	$-x_1$	$-x_2$	….	$-x_S$	….	$-x_n$	1
0	$a_{1,1}$	$a_{1,2}$	….	$a_{1,S}$	….	$a_{1,n}$	b_1
….	….	….	….	….	….	….	….
0	$a_{m,1}$	$a_{m,2}$	….	$a_{m,S}$	….	$a_{m,n}$	b_m
x_{m+1}	$a_{m+1,1}$	$a_{m+1,2}$	….	$a_{m+1,s}$	….	$a_{m+1,n}$	b_{m+1}
….	….	….	….	….	….	….	…
x_{m+p}	$a_{m+p,1}$	$a_{m+p,2}$	….	$a_{m+p,S}$	….	$a_{m+p,n}$	b_{m+p}
$Q(\bar{x})$	$-c_1$	$-c_2$	….	$-c_S$	….	$-c_n$	0

Правила выбора разрешающего элемента при поиске опорного плана

1. При условии отсутствия “0-строк” (ограничений-равенств) и “свободных” переменных (т.е. переменных, на которые не наложено требование неотрицательности).

- Если в столбце свободных членов симплексной таблицы нет отрицательных элементов, то опорный план найден.
- Есть отрицательные элементы в столбце свободных членов, например $b_i < 0$. В такой строке ищем отрицательный коэффициент a_{il}, и этим самым определяем разрешающий столбец l. Если не найдем отрицательный a_{il}, *то система ограничений несовместна* (противоречива).
- В качестве разрешающей выбираем строку, которой соответствует минимальное отношение: $\frac{b_r}{a_{rl}} = \min_i \left\{ \frac{b_i}{a_{il}} \middle| \frac{b_i}{a_{il}} > 0 \right\}$, где r – номер разрешающей строки. Таким образом, a_{rl} – разрешающий элемент.
- После того, как разрешающий элемент найден, делаем шаг модифицированного жорданова исключения с направляющим элементом a_{rl} и переходим к следующей симплексной таблице.

2. В случае присутствия ограничений-равенств и “свободных” переменных поступают следующим образом.

- Выбирают разрешающий элемент в “0-строке” и делают шаг модифицированного жорданова исключения, после чего вычеркивают этот разрешающий столбец. Данную последовательность действий продолжают до тех пор, пока в симплексной таблице остается хотя бы одна “0-строка” (при этом таблица сокращается).

- Если же присутствуют и свободные переменные, то необходимо данные переменные сделать базисными. И после того, как свободная переменная станет базисной, в процессе определения разрешающего элемента при поиске опорного и оптимального планов данная строка не учитывается (но преобразуется).

7.2. Вырожденность в задачах линейного программирования

Рассматривая симплекс-метод, мы предполагали, что задача линейного программирования является невырожденной, т.е. каждый опорный план содержит ровно m положительных компонент, где m – число ограничений в задаче. В вырожденном опорном плане число положительных компонент оказывается меньше числа ограничений: некоторые базисные переменные, соответствующие данному опорному плану, принимают нулевые значения. Используя геометрическую интерпретацию для простейшего случая (рис. 38), когда $n-m=2$ (число небазисных переменных равно 2), легко отличить вырожденную задачу от невырожденной. В вырожденной задаче в одной вершине многогранника условий пересекается более двух прямых, описываемых уравнениями вида $x_i=0$. Это значит, что одна или несколько сторон многоугольника условий стягиваются в точку.

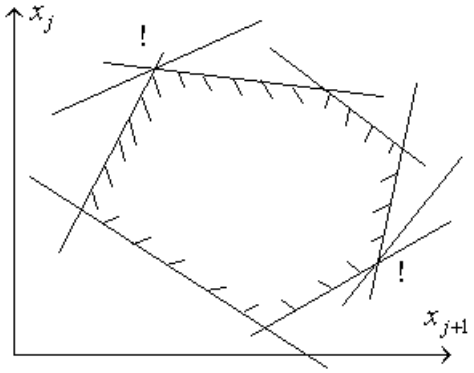


Рис. 38. *К понятию вырожденности*

Аналогично при $n-m=3$ в вырожденной задаче в одной вершине пересекается более 3-х плоскостей $x_i=0$.

В предположении о невырожденности задачи находилось только одно значение $\theta=\min\limits_i\left\{\dfrac{b_i}{a_{il}}\middle|\dfrac{b_i}{a_{il}}>0\right\}$, по которому определялся индекс выводимого из базиса вектора условий (выводимой из числа базисных переменной).

В вырожденной задаче $\min\limits_i\left\{\dfrac{b_i}{a_{il}}\middle|\dfrac{b_i}{a_{il}}>0\right\}$ может достигаться на нескольких индексах сразу (для нескольких строк). В этом случае в находимом опорном плане несколько базисных переменных будут нулевыми.

Если задача линейного программирования оказывается вырожденной, то при плохом выборе вектора условий, выводимого из базиса, может возникнуть бесконечное движение по базисам одного и того же опорного плана. Это – так называемое явление зацикливания. Хотя в практических задачах линейного про-

граммирования зацикливание является довольно редким, возможность его не исключена.

Один из приемов борьбы с вырожденностью состоит в преобразовании задачи путем «незначительного» изменения вектора правых частей системы ограничений на величины ε_i таким образом, чтобы задача стала невырожденной, и, в то же время, чтобы это изменение не повлияло реально на оптимальный план задачи.

Чаще реализуемые алгоритмы включают в себя некоторые простые правила, снижающие вероятность возникновения зацикливания или его преодоления.

Пусть переменную x_j необходимо сделать базисной. Рассмотрим множество индексов E_0, состоящее из тех i, для которых достигается $\theta_0 = \min\limits_i \left\{ \dfrac{b_i}{a_{il}} \middle| \dfrac{b_i}{a_{il}} > 0 \right\}$. Множество индексов i, для которых выполняется данное условие обозначим через E_0. Если E_0 состоит из одного элемента, то из базиса исключается вектор условий A_i (переменная x_i делается небазисной).

Если E_0 состоит более чем из одного элемента, то составляется множество E_1, которое состоит из $i \in E_0$, на которых достигается $\theta_1 = \min\limits_{i \in E_0} \left\{ \dfrac{a_{i1}}{a_{il}} \right\}$. Если E_1 состоит из одного индекса k, то из базиса выводится переменная x_k. В противном случае составляется множество E_2 и т.д.

Практически правилом надо пользоваться, если зацикливание уже обнаружено.

7.3. Двойственность задачи линейного программирования

Рассмотрим задачу (7.1) максимизации линейной формы и, одновременно, задачу (7.2) минимизации:

$$\left.\begin{aligned} Q\left(\bar{x}\right) &= \bar{p}^T \cdot \bar{x} \to \max, \\ A \cdot \bar{x} &\le \bar{b}, \\ \bar{x} &\ge 0, \end{aligned}\right\} \tag{7.1}$$

$$\left.\begin{aligned} W\left(\bar{u}\right) &= \bar{b}^T \cdot \bar{u} \to \min, \\ A^T \cdot \bar{x} &\ge \bar{p}, \\ \bar{u} &\ge 0. \end{aligned}\right\} \quad x \in \mathbb{R} \tag{7.2}$$

Задача (7.2) называется двойственной по отношению к прямой (7.1) (и наоборот).

Пример: Предприятие выпускает три вида продукции. Каждая продукция требует обработки на трех различных типах установок. Ресурс времени каждого типа установок ограничен. Известна прибыль от единицы каждого вида продукции: p_1, p_2, p_3. Если количество выпускаемой продукции каждого вида есть x_1, x_2, x_3, то прибыль определяется по формуле

$$Q\left(\bar{x}\right) = p_1 \cdot x_1 + p_2 \cdot x_2 + p_3 \cdot x_3 \to \max$$

при ограничениях следующего вида:

$$a_{11} \cdot x_1 + a_{12} \cdot x_2 + a_{13} \cdot x_3 \le b_1,$$

$$a_{21} \cdot x_1 + a_{22} \cdot x_2 + a_{23} \cdot x_3 \le b_2,$$

$$a_{31} \cdot x_1 + a_{32} \cdot x_2 + a_{33} \cdot x_3 \le b_3,$$

$$\bar{x} \ge \bar{0},$$

где b_1, b_2, b_3 – ресурсы времени установок первого, второго и третьего типов. Величины a_{ij} определяют количество ресурса времени установки i-го типа, которое необходимо для выпуска одной единицы продукции j-го вида.

Двойственная к ней задача будет иметь вид:

$$W\left(\bar{u}\right) = b_1 \cdot u_1 + b_2 \cdot u_2 + b_3 \cdot u_3 \to \min$$

с ограничениями:

$$\begin{cases} a_{11} \cdot u_1 + a_{21} \cdot u_2 + a_{31} \cdot u_3 \ge p_1, \\ a_{12} \cdot u_1 + a_{22} \cdot u_2 + a_{32} \cdot u_3 \ge p_2, \\ a_{13} \cdot u_1 + a_{23} \cdot u_2 + a_{33} \cdot u_3 \ge p_3, \\ \bar{u} \ge \bar{0}. \end{cases}.$$

Здесь u_1 – это оценка (цена), соответствующая одной единице ограниченного ресурса, соответствующего первой установке. И она равна величине, на которую могла бы увеличиться суммарная прибыль, если бы количество этого ограниченного ресурса увеличилось на единицу, и если это увеличение было бы использовано оптимально. Иными словами, u_1 – это количество прибыли, недополученной из-за нехватки единицы ограниченного ресурса b_1. Аналогичным образом можно интерпретировать смысл величин u_2 и u_3.

Преобразования при решении прямой и двойственной задач

Пусть имеются прямая и двойственная задачи следующего вида:

Прямая задача:

$$Q\left(\bar{x}\right) = \bar{p}^T \cdot \bar{x} \to \max,$$

$$A \cdot \bar{x} \le \bar{b},$$

$$\bar{x} \ge \bar{0}.$$

Представим ограничения в виде:

$$\bar{y} = -A \cdot \bar{x} + \bar{b} \ge \bar{0},$$

$$\bar{x} \ge \bar{0}.$$

Двойственная к ней задача: $W(\bar{u}) = \bar{b}^T \cdot \bar{u} \to \min,$ $A^T \cdot \bar{u} \geq \bar{p},$ $\bar{u} \geq \bar{0}.$	$\bar{v} = A^T \cdot \bar{u} - \bar{p} \geq 0,$ $\bar{u} \geq \bar{0}.$

Для ограничений прямой задачи симплексная таблица имеет вид:

	$-x_1$	…	$-x_S$	…	$-x_n$	1
$y_1 =$	a_{11}	…	a_{1S}	…	a_{1n}	b_1
…	…	…	…	…	…	…
$y_r =$	a_{r1}	…	a_{rS}	…	a_{rn}	b_r
…	…	…	…	…	…	…
$y_m =$	a_{m1}	…	a_{mS}	…	a_{mn}	b_m
$Q(\bar{x}) =$	$-p_1$	…	$-p_S$	…	$-p_n$	0

Пусть a_{rS} – разрешающий элемент, сделаем шаг модифицированного жорданова исключения:

	$-x_1$	…	$-y_r$	…	$-x_n$	1
$y_1 =$	b_{11}	…	$-a_{1S}$	…	b_{1n}	$b_{1,n+1}$
…	…	…	…	…	…	…
$x_S =$	a_{r1}	…	1	…	a_{rn}	b_r
…	…	…	…	…	…	…
$y_m =$	b_{m1}	…	$-a_{mS}$	…	b_{mn}	$b_{m,n+1}$
$Q(\bar{x}) =$	$b_{m+1,1}$	…	p_S	…	$b_{m+1,n}$	$b_{m+1,n+1}$

Здесь $b_{ij} = a_{ij} \cdot a_{rS} - a_{rj} \cdot a_{iS}$ и всю данную таблицу следует разделить еще на a_{rS}.

Симплексную таблицу для двойственной задачи запишем, развернув ее на 90°. Получаем:

	$v_1 =$	…	$v_s =$	…	$v_n =$	$W =$
u_1	a_{11}	…	a_{1S}	…	a_{1n}	b_1
…	…	…	…	…	…	…
u_r	a_{r1}	…	a_{rS}	…	a_{rn}	b_r
…	…	…	…	…	…	…
u_m	a_{m1}	…	a_{mS}	…	a_{mn}	b_m
1	$-p_1$	…	$-p_S$	…	$-p_n$	0

Пусть a_{rS} – направляющий элемент. Сделаем шаг обыкновенного жорданова исключения (отличие от модифицированного состоит в том, что элементы в разрешающей строке меняют знаки, а в столбце знаки сохраняются; в остальном преобразование остается тем же):

	$v_1 =$	…	$u_r =$	…	$v_n =$	$W =$
u_1	b_{11}	…	$-a_{1S}$	…	b_{1n}	$b_{1,n+1}$
…	…	…	…	…	…	…
v_s	a_{r1}	…	1	…	a_{rn}	b_r
…	…	…	…	…	…	…
u_m	b_{m1}	…	$-a_{mS}$	…	b_{mn}	$b_{m,n+1}$
1	$b_{m+1,1}$	…	p_S	…	$b_{m+1,n}$	$b_{m+1,n+1}$

Здесь $b_{ij} = a_{ij} \cdot a_{rS} - a_{rj} \cdot a_{iS}$ и всю данную таблицу также следует разделить еще на a_{rS}.

Замечание: Не следует забывать при преобразованиях, что в данном случае у нас таблица развернута.

Таким образом, нетрудно заметить, что шаг модифицированного жорданова исключения над симплексной таблицей прямой задачи соответствует шагу обыкновенного жорданова исключения над симплексной таблицей двойственной задачи. Эти взаимно двойственные задачи можно совместить в одной симплексной таблице:

	$v_1 =$ $-x_1$	…	$v_s =$ $-x_S$	…	$v_n =$ $-x_n$	$W =$ 1
u_1 $y_1 =$	a_{11}	…	a_{1S}	…	a_{1n}	b_1
…	…	…	…	…	…	…
u_r $y_r =$	a_{r1}	…	a_{rS}	…	a_{rn}	b_r
…	…	…	…	…	…	…
u_m $y_m =$	a_{m1}	…	a_{mS}	…	a_{mn}	b_m
1 $Q(\bar{x}) =$	$-p_1$	…	$-p_S$	…	$-p_n$	0

Можно показать, что, решая основную задачу линейного программирования, решаем и двойственную к ней. И наоборот. Причем, $\max Q = \min W$.

Теоремы двойственности

Основная теорема двойственности линейного программирования

Пусть рассматривается пара двойственных задач:

$$\left.\begin{aligned} Q(\bar{x}) = \bar{p}^T \cdot \bar{x} \to \max, \\ A \cdot \bar{x} \le \bar{b}, \\ \bar{x} \ge 0, \end{aligned}\right\} \qquad (7.1)$$

$$\left.\begin{array}{l} W\left(\bar{u}\right)=\bar{b}^{T}\cdot\bar{u}\to\min,\\ A^{T}\cdot\bar{x}\geq\bar{p},\\ \bar{u}\geq 0. \end{array}\right\}\quad x\in\mathbb{R} \tag{7.2}$$

Если одна из этих задач обладает оптимальным решением, то и двойственная к ней задача также имеет оптимальное решение. Причем экстремальные значения соответствующих линейных форм равны: $\max Q=\min W$.

Если же у одной из этих задач линейная форма не ограничена, то двойственная к ней задача противоречива.

Доказательство: Пусть основная задача (7.1) имеет конечное решение и получена окончательная симплексная таблица:

		$u_1=$	....	$u_s=$	$v_{s+1}=$	....	$v_n=$	$W=$
		$-y_1$	....	$-y_s$	$-x_{s+1}$	....	$-x_n$	1
v_1	$x_1=$	$b_{1,1}$	....	$b_{1,s}$	$b_{1,s+1}$	....	$b_{1,n}$	$b_{1,n+1}$
....	....	....	....	....	....	....	....	....
v_s	$x_s=$	$b_{s,1}$	....	$b_{s,s}$	$b_{s,s+1}$	....	$b_{s,n}$	$b_{s,n+1}$
u_{s+1}	$y_{s+1}=$	$b_{s+1,1}$	....	$b_{s+1,s}$	$b_{s+1,s+1}$	....	$b_{s+1,n}$	$b_{s+1,n+1}$
....	....	....	....	....	....	....	....	....
u_m	$y_m=$	$b_{m,1}$	....	$b_{m,s}$	$b_{m,s+1}$	....	$b_{m,n}$	$b_{m,n+1}$
1	$Q=$	q_1	....	q_s	q_{s+1}	....	q_n	q_0

Так как данная таблица, по предположению, соответствует оптимальному решению задачи (7.1), то $b_{1,n+1}\geq 0$, ..., $b_{m,n+1}\geq 0$ и $q_1\geq 0$, ..., $q_n\geq 0$. При этом $\max Q=q_0$ достигается при $y_1=...=y_s=x_{s+1}=...=x_n=0$.

Рассмотрим полученную таблицу двойственной задачи. Полагая значения переменных слева (небазисных) равными нулю:

$$v_1=...=v_s=u_{s+1}=...=u_m=0,$$

найдем

$$u_1=q_1\geq 0,\ \ldots,\ u_s=q_s\geq 0,\ v_{s+1}=q_{s+1}\geq 0,\ \ldots,\ v_n=q_n\geq 0.$$

Следовательно, получено опорное решение

$$u_1=q_1,\ \ldots,\ u_s=q_s,\ u_{s+1}=0,\ \ldots,\ u_m=0.$$

Из последнего столбца находим, что

$$W=b_{1,n+1}\cdot v_1+...+b_{s,n+1}\cdot v_s+b_{s+1,n+1}\cdot u_{s+1}+...+b_{m,n+1}\cdot u_m+q_0$$

и в точке

$$v_1=...=v_s=u_{s+1}=...=u_m=0$$

значение W будет минимальным в силу того, что $b_{i,n+1} \geq 0$, $i = 1,...,m$. Следовательно, $\max Q = \min W$.

Пусть теперь линейная форма прямой задачи неограничена, т.е. для некоторой верхней переменной, например, y_s, соответствующий коэффициент $q_s < 0$, а все коэффициенты этого столбца симплексной таблицы неположительны: $b_{1,s} \leq 0$, $b_{2,s} \leq 0$, …, $b_{m,s} \leq 0$. Тогда из таблицы для двойственной задачи:

$$u_s = b_{1,s} \cdot v_1 + ... + b_{s,s} \cdot v_s + b_{s+1,s} \cdot u_{s+1} + ... + b_{m,s} \cdot u_m + q_s \leq q_s < 0,$$

то есть система ограничений двойственной задачи противоречива. Так как из неотрицательности $v_1, ..., v_s, u_{s+1}, ..., u_m$ следует неположительность u_s (нельзя сделать ее положительной), то есть система несовместна.

Теорема доказана.

Вторая теорема двойственности:

Если хотя бы одно оптимальное решение одной из двойственных задач обращает i-е ограничение этой задачи в строгое неравенство, то i-я компонента (т.е. x_i или u_i) каждого оптимального решения второй двойственной задачи равна нулю.

Если же i-я компонента хотя бы одного оптимального решения одной из двойственных задач положительна, то каждое оптимальное решение другой двойственной задачи обращает i-е ограничение в строгое равенство.

Иначе говоря, оптимальные решения $\bar{x}^*$ и $\bar{u}^*$ пары двойственных задач удовлетворяют условиям

$$x_j^* \cdot \left[\sum_{i=1}^{m} a_{ij} \cdot u_i^* - p_j \right] = 0, \quad j = \overline{1,n}, \tag{7.3}$$

$$u_i^* \cdot \left[\sum_{j=1}^{n} a_{ij} \cdot x_j^* - b_i \right] = 0, \quad i = \overline{1,m}. \tag{7.4}$$

Доказательство:

Пусть $\bar{x}^*$ и $\bar{u}^*$ – оптимальные решения пары двойственных задач. Тогда для

$$Q(\bar{x}) = \sum_{j=1}^{n} p_j x_j \to \max,$$

$$W(\bar{u}) = \sum_{i=1}^{m} b_i u_i \to \min$$

они удовлетворяют следующим ограничениям:

$$\left.\begin{array}{l} a_{i1}\cdot x_1^* + a_{i2}\cdot x_2^* + ... + a_{in}\cdot x_n^* \le b_i, \quad i=\overline{1,m}, \\ x_j^* \ge 0, \quad j=\overline{1,n}, \\ a_{1j}\cdot u_1^* + a_{2j}u_2^* + ... + a_{mj}u_m^* \ge p_j, \quad j=\overline{1,n}, \\ u_i^* \ge 0, \quad i=\overline{1,m}. \end{array}\right\}. \qquad (7.5)$$

Умножим (7.5) соответственно на u_i^* и x_j^* и просуммируем полученные выражения:

$$\sum_{j=1}^{n} p_j \cdot x_j^* \le \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\cdot u_i^* \cdot x_j^* \le \sum_{i=1}^{m} b_i \cdot u_i^* . \qquad (7.6)$$

Из основной теоремы двойственности следует

$$\sum_{j=1}^{n} p_j \cdot x_j^* = \sum_{i=1}^{m} b_i \cdot u_i^* . \qquad (7.7)$$

И с учетом (7.6) получаем:

$$\sum_{j=1}^{n} p_j \cdot x_j^* = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}\cdot u_i^* \cdot x_j^* ,$$

$$\sum_{i=1}^{m} b_i \cdot u_i^* = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}\cdot x_j^* \cdot u_i^* .$$

Первое из этих выражений можем переписать в виде

$$\sum_{j=1}^{n} x_j^* \left(\sum_{i=1}^{m} a_{ij}\cdot u_i^* - p_j \right) = 0 ,$$

и так как все x_j^* и выражения в скобках неотрицательны, то, опуская суммирование, получим:

$$x_j^* \cdot \left(\sum_{i=1}^{m} a_{ij}\cdot u_i^* - p_j \right) = 0, \quad j = 1,...,n .$$

Аналогично найдём:

$$u_i^* \cdot \left(\sum_{j=1}^{n} a_{ij}\cdot x_j^* - b_i \right) = 0, \quad i = 1,...,m .$$

Что и требовалось доказать.

Справедлива и обратная теорема.

7.4. Метод последовательного уточнения оценок

Этот метод иногда называют еще двойственным симплекс-методом. Ранее говорилось, что одновременно с решением прямой задачи решается и двойственная задача. Если проследить за получающимися преобразованиями двойственной таблицы и переменных u_i и x_j и записать таблицу для двойственной задачи в обычном виде, то получим описание нового метода – метода последовательного уточнения оценок.

Пусть дана задача:

$$W\left(\bar{u}\right) = b_1 \cdot u_1 + ... + b_r \cdot u_r + ... + b_m \cdot u_m \to \min,$$

$$v_1 = a_{11} \cdot u_1 + ... + a_{r1} \cdot u_r + ... + a_{m1} \cdot u_m - p_1 \geq 0,$$

.....

$$v_s = a_{1s} \cdot u_1 + ... + a_{rs} \cdot u_r + ... + a_{ms} \cdot u_m - p_r \geq 0,$$

.....

$$v_n = a_{1n} \cdot u_1 + ... + a_{rn} \cdot u_2 + ... + a_{m,n} \cdot u_m - p_n \geq 0,$$

$$\bar{v} \geq \bar{0}.$$

Симплексая таблица, построенная для данной задачи, будет иметь вид:

	u_1	….	u_r	….	u_m	1
$v_1 =$	a_{11}	….	a_{1r}	….	a_{1m}	$-p_1$
….	….	….	….	….	….	….
$v_s =$	a_{s1}	….	a_{sr}	….	a_{sm}	$-p_s$
….	….	….	….	….	….	….
$v_n =$	a_{n1}	….	a_{nr}	….	a_{nm}	$-p_n$
$W =$	b_1	….	b_r	….	b_m	0

В методе последовательного уточнения оценок сначала избавляются от отрицательности в W-строке (получают псевдоплан), а затем, перебирая псевдопланы, ищут оптимальный план (первый найденный опорный).

Правило выбора разрешающего элемента для избавления от отрицательности в W-строке:

1) Если все коэффициенты W-строки неотрицательны, то 0 является оценкой снизу для целевой функции W и можно переходить к отысканию оптимального решения. Иначе выбираем некоторый $b_r < 0$ и рассматриваем r-й столбец.

2) Находим в r-м столбце какой-нибудь из отрицательных элементов, например, $a_{rs} < 0$. Тогда строку с номером s, содержащую a_{rs}, выбираем в качестве разрешающей строки. Если все коэффициенты r-го столбца неотрицательны, то либо W *неограничена снизу* ($W \to -\infty$), либо *система ограничений противоречива*. (Из противоречивости двойственной не следует неограниченность прямой задачи).

3) Находим неотрицательное отношение коэффициента *W*-строки к коэффициентам разрешающей (*s*-й) строки. В качестве разрешающего берем тот элемент разрешающей *s*-й строки, для которого это отношение положительно и минимально, т.е. выбираем некоторый коэффициент a_{sk}, для которого

$$\frac{b_k}{a_{sk}} = \min_j \left\{ \frac{b_j}{a_{sj}} \middle| \frac{b_j}{a_{sj}} > 0 \right\}.$$

Выбрав разрешающий элемент, делаем шаг обыкновенного жорданова исключения. Указанная последовательность действий выполняется до тех пор, пока все коэффициенты *W*-строки не станут неотрицательными. Например, будет получена следующая таблица

	v_1	....	v_s	u_{s+1}	....	u_m	1
$u_1 =$	b_{11}	....	b_{s1}	$b_{s+1,1}$	....	b_{m1}	q_1
....	....	....	....	....	....	....	....
$u_s =$	b_{1s}	....	b_{ss}	$b_{s+1,s}$	....	b_{ms}	q_s
$v_{s+1} =$	$b_{1,s+1}$	....	$b_{s,s+1}$	$b_{s+1,s+1}$	....	$b_{m,s+1}$	q_{s+1}
....	....	....	....	....	....	....	....
$v_n =$	$b_{1,n}$	....	$b_{s,n}$	$b_{s+1,n}$	....	b_{mn}	q_n
$W =$	b_1	....	b_s	b_{s+1}	....	b_m	q_0

Если все $q_1, ..., q_n$ неотрицательны, то таблица соответствует оптимальному решению и $q_0 = \min W$, иначе, q_0 – оценка снизу для *W*.

Правило выбора разрешающего элемента при поиске оптимального решения:

1) В качестве разрешающей строки берем строку, содержащую отрицательный коэффициент, например, $q_s < 0$, и строка с номером s будет разрешающей.

2) В качестве разрешающего выбираем тот положительный коэффициент b_{ks} строки s, для которого

$$\frac{b_k}{b_{sk}} = \min_j \left\{ \frac{b_j}{b_{sj}} \middle| \frac{b_j}{b_{sj}} > 0 \right\}.$$

Если в строке с номером s нет положительных коэффициентов, то *ограничения задачи противоречивы*.

После выбора разрешающего элемента делаем шаг обыкновенного жорданова исключения.

После конечного числа шагов либо найдем оптимальное решение, либо можно убедимся в противоречивости ограничений задачи.

Замечание: Если в симплекс-методе мы приближаемся к оптимальному решению при поиске минимума *сверху*, передвигаясь по опорным планам, то в методе последовательного уточнения оценок при поиске минимума к оптимальному решению *снизу*, причем промежуточные планы (*псевдопланы*) не являются

опорными (лежат вне многогранника решений). Первое же допустимое решение (опорный план) будет оптимальным.

Пример. Решить следующую задачу методом последовательного уточнения оценок:

$$L(\bar{x}) = -2x_1 - x_2 \to \min,$$
$$x_1 - 2x_2 + 3 \ge 0,$$
$$-3x_1 - 7x_2 + 21 \ge 0,$$
$$-x_1 + x_2 + 2 \ge 0,$$
$$-5x_1 - 4x_2 + 20 \ge 0,$$
$$x_i \ge 0, \ \ i = 1, 2.$$

	x_1	x_2	1
$y_1 =$	1	–2	3
$y_2 =$	–3	–7	21
$y_3 =$	–1	1	2
$y_4 =$	–5	–4	20
$L =$	–2	–1	0

	y_3	x_2	1
$y_1 =$	–1	–1	5
$y_2 =$	3	–10	15
$x_1 =$	–1	1	2
$y_4 =$	5	–9	10
$L =$	2	–3	–4

	y_3	y_1	1
$x_2 =$	–1	–1	5
$y_2 =$	13	10	–35
$x_1 =$	–2	–1	7
$y_4 =$	14	9	–35
$L =$	5	3	–19

	y_3	y_2	1
$x_2 =$	0,3	–0,1	1,5
$y_1 =$	–1,3	0,1	3,5
$x_1 =$	–0,7	–0,1	3,5
$y_4 =$	2,3	0,9	–3,5
$L =$	1,1	0,3	–8,5

	y_3	y_4	1
$x_2 =$	5/9	–1/9	10/9
$y_1 =$	14/9	1/9	35/9
$x_1 =$	4/9	–1/9	28/9
$y_2 =$	–23/9	10/9	35/9
$L =$	1/3	1/3	–22/3

Ответ:

$$L_{\min} = -7\frac{1}{3}; \quad \bar{x} = \left(3\frac{1}{9}, 1\frac{1}{9}\right).$$

7.5. Методы решения транспортной задачи

Транспортная задача линейного программирования формулируется следующим образом. Необходимо минимизировать транспортные расходы

$$Q(X) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} \cdot x_{ij} \to \min$$

при ограничениях

$$\left.\begin{array}{ll} \sum\limits_{i=1}^{m} x_{ij} = b_j, & j = \overline{1,n}, \\ \sum\limits_{j=1}^{n} x_{ij} = a_i, & i = \overline{1,m}, \\ x_{ij} \geq 0, \quad i = \overline{1,m}, & j = \overline{1,n}, \end{array}\right\},$$

где c_{ij} – стоимость перевозки единицы продукции из пункта i в пункт j; x_{ij} – планируемая величина перевозок из пункта i в пункт j (план перевозок X – матрица размерности $m \times n$);

b_j – потребности в продукте в пункте j;

a_i – запасы в пункте i.

Предполагается, что модель *закрытого* типа, то есть $\sum\limits_{j=1}^{n} b_j = \sum\limits_{i=1}^{m} a_i$.

Если модель открытого типа $\left(\sum\limits_{j=1}^{n} b_j \neq \sum\limits_{i=1}^{m} a_i \right)$, то ее всегда можно привести к закрытому типу введением фиктивного пункта производства или фиктивного пункта потребления:

Если $\sum\limits_{j=1}^{n} b_j < \sum\limits_{i=1}^{m} a_i$, то $b_{n+1} = \sum\limits_{i=1}^{m} a_i - \sum\limits_{j=1}^{n} b_j$,

тогда $\sum\limits_{j=1}^{n+1} b_j = \sum\limits_{i=1}^{m} a_i$, причем $c_{i,n+1} = 0 \quad \forall i$.

Если $\sum\limits_{j=1}^{n} b_j > \sum\limits_{i=1}^{m} a_i$, то $a_{m+1} = \sum\limits_{j=1}^{n} b_j - \sum\limits_{i=1}^{m} a_i$, $\sum\limits_{j=1}^{n} b_j = \sum\limits_{i=1}^{m+1} a_i$ и $c_{m+1,j} = 0 \quad \forall j$.

Транспортная задача представляет собой задачу линейного программирования и, естественно, ее можно решить с использованием метода последовательного улучшения плана или метода последовательного уточнения оценок. В этом случае основная трудность бывает связана с числом переменных задачи ($m \times n$) и числом ограничений ($m+n$). Поэтому специальные алгоритмы оказываются более эффективными. К таким алгоритмам относятся *метод потенциалов* и *венгерский метод*.

Алгоритм метода потенциалов, его называют еще модифицированным распределительным алгоритмом, начинает работу с некоторого опорного плана транспортной задачи (допустимого плана перевозок). Для построения опорного плана обычно используется один из двух методов: *метод северо-западного угла* или *метод минимального элемента*.

7.5.1. Метод северо-западного угла

Он позволяет найти некоторый допустимый план перевозок. Составим транспортную таблицу некоторой задачи.

a_i \ b_j	30	80	20	30	90
120	2 30	4 80	2 10	3	8
30	3	5	6 10	6 20	2
40	6	8	7	4 10	5 30
60	3	4	2	1	4 60

В данном случае имеем задачу закрытого типа, т.к.

$$\sum_{i=1}^{4} a_i = 250 = \sum_{j=1}^{5} b_j .$$

При построении плана должны учитывать, что сумма перевозок в столбце должна оказаться равной потребностям в данном пункте, а сумма перевозок в строке запасу в пункте, соответствующем данной строке.

Заполнение начинается с верхнего левого угла таблицы. Величина перевозки устанавливается равной минимальной из величин: величины остатка запасов в пункте *i* или величины еще неудовлетворенного спроса в пункте *j*.

Если ресурс в данной строке исчерпан, то переходим к перевозке в следующей строке текущего столбца (на одну строку вниз).

Если потребности для данного пункта (столбца) удовлетворены, то переходим к следующей перевозке текущей строки в следующем столбце.

Затраты на перевозку по построенному плану равны

$$Q = 30\times 2 + 4\times 80 + 2\times 10 + 6\times 10 + 6\times 20 + 4\times 10 + 5\times 30 + 4\times 60 = 1010 .$$

Естественно, что найденный план далек от оптимального.

7.5.2. Метод минимального элемента

В таблице отыскивается $\min\{c_{ij}\}$ и в первую очередь заполняется соответствующая клетка: $x_{ij} = \min\{a_i, b_j\}$. Затем вычеркивается остаток соответствующей строки, если $a_i < b_j$, или столбца, если $a_i > b_j$, и корректируем остатки запасов и неудовлетворенного спроса.

В оставшихся клетках таблицы снова отыскивается минимальная стоимость перевозки и заполняется соответствующая клетка и т.д.

a_i \ b_j	30	80	20	30	90
120	2 30	4 80	2	3	8 10
30	3	5	6	6	2 30
40	6	8	7	4	5 40
60	3	4	2 20	1 30	4 10

Затраты на перевозку по построенному плану равны

$$Q = 30\times 2 + 4\times 80 + 8\times 10 + 2\times 30 + 5\times 40 + 2\times 20 + 1\times 30 + 4\times 10 = 830.$$

Этот план лучше, но утверждать, что он оптимален, нельзя.

Определение 1. Набором называется произвольная совокупность перевозок транспортной таблицы.

Определение 2. Цепью называют такие наборы, когда каждая пара соседних клеток в цепи расположены либо в одном столбце, либо в одной строке.

Определение 3. Циклом называется цепь, крайние элементы которой находятся либо в одной строке, либо в одном столбце.

7.5.3. Метод потенциалов

Метод позволяет находить оптимальный план перевозок транспортной таблицы. В основе лежит следующая теорема.

Теорема. Для того, чтобы некоторый план $X = [x_{ij}]_{m\times n}$ транспортной задачи был оптимальным, необходимо и достаточно, чтобы ему соответствовала такая система $m+n$ чисел $u_1, u_2, ..., u_m$; $v_1, v_2, ..., v_n$, для которой выполняются условия

$$v_j - u_i \le c_{ij},\ i = \overline{1,m},\ j = \overline{1,n}, \qquad (7.8)$$

$$v_j - u_i = c_{ij},\ \forall x_{ij} > 0. \qquad (7.9)$$

Величины u_i и v_j называются потенциалами соответствующих пунктов отправления и пунктов назначения. Условия (7.8-7.9) называются условиями потенциальности.

План X будем называть потенциальным, если для него существует система u_i и v_j, удовлетворяющая (7.8–7.9). Тогда теорема коротко формулируется следующим образом.

Теорема. Для оптимальности транспортной задачи необходимо и достаточно, чтобы потенциальный план был оптимален.

Достаточность. Пусть план X потенциален, так что существует система u_i и v_j, удовлетворяющая (7.8–7.9). Тогда для любого допустимого плана $X' = [x'_{ij}]_{m\times n}$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x'_{ij} \ge \sum_{i=1}^{m}\sum_{j=1}^{n}\left(v_j - u_i\right)x'_{ij} = \sum_{j=1}^{n} v_j \sum_{i=1}^{m} x'_{ij} - \sum_{i=1}^{m} u_i \sum_{j=1}^{n} x'_{ij} =$$

$$= \sum_{j=1}^{n} v_j b_j - \sum_{i=1}^{m} u_i a_i = \sum_{j=1}^{n} v_j \sum_{i=1}^{m} x_{ij} - \sum_{i=1}^{m} u_i \sum_{j=1}^{n} x_{ij} =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{m} v_j x_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n} u_i x_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m}\left(v_j - u_i\right)x_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} c_{ij}x_{ij},$$

т.е. стоимость перевозок по любому плану X' не меньше стоимости перевозок по потенциальному плану X.

Следовательно, план X оптимален.

Необходимость. Будем рассматривать транспортную задачу, как задачу линейного программирования с минимизацией линейной формы

$$Q(X) = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} \cdot x_{ij} \to \min$$

при соответствующих ограничениях. Заполним симплексную таблицу и рассмотрим двойственную к ней задачу, что легко получить из таблицы. Прямую таблицу будем заполнять, повернув.

	$0=-u_1$	…	$0=-u_i$	…	$0=-u_m$	$0=-v_1$	…	$0=-v_j$	…	$0=-v_n$	$Q=1$
$x_{11}\ y_{11}=$	-1	…	0	…	0	1	…	0	…	0	c_{11}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{1n}\ y_{1n}=$	–1	…	0	…	0	0	…	0	…	1	c_{1n}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{i1}\ y_{i1}=$	0	…	–1	…	0	1	…	0	…	0	c_{i1}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{ij}\ y_{ij}=$	0	…	–1	…	0	0	…	1	…	0	c_{ij}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{in}\ y_{in}=$	0	…	–1	…	0	0	…	0	…	1	c_{in}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{m1}\ y_{m1}=$	0	…	0	…	–1	1	…	0	…	0	C_{m1}
…	…	…	…	…	…	…	…	…	…	…	…
$x_{mn}\ y_{mn}=$	0	…	0	…	–1	0	…	0	…	1	C_{mn}
1 $w=$	a_1	…	a_i	…	a_n	$-b_1$	…	$-b_j$	…	$-b_n$	0

Получаем, что двойственная задача имеет вид:

$$w = \sum_{j=1}^{n} b_j v_j - \sum_{i=1}^{m} a_i u_i \to \max$$

при ограничениях

$$y_{ij} = u_i - v_j + c_{ij} \geq 0,\ i = \overline{1,m},\ j = \overline{1,n},$$

т.е. $v_j - u_i \leq c_{ij},\ i = \overline{1,m},\ j = \overline{1,n}$.

Пусть $X = [x_{ij}]_{m \times n}$ – оптимальное решение транспортной задачи. Тогда на основании теоремы двойственности двойственная задача имеет оптимальное решение

$$u_1^*, ..., u_m^*; v_1^*, ..., v_n^*.$$

Убедимся, что эти числа являются потенциалами соответствующих пунктов транспортной задачи. Действительно, все u_i^*, v_j^* как опорное решение двойственной задачи удовлетворяют неравенствам (7,8).

Если $x_{ij} > 0$, то по второй теореме двойственности соответствующее ограничение

$$y_{ij}^* = u_i^* - v_j^* + c_{ij} \geq 0$$

двойственной задачи обращается в строгое равенство

$$v_j^* - u_i^* = c_{ij}.$$

Алгоритм метода потенциалов

Алгоритм метода потенциалов состоит из предварительного этапа и повторяющегося основного этапа.

Предварительный этап.

Каким-либо способом ищется допустимый план X (методом северо-западного угла или минимального элемента).

1. Для полученного плана строится система $m+n$ чисел $u_1,...,u_m$, $v_1, ..., v_n$, таких, что $v_j - u_i = c_{ij},\ \forall x_{ij} > 0$.

2. Построенная система u_i и v_j исследуется на потенциальность (то есть план X исследуется на оптимальность). Для этого проверяется $v_j - u_i \leq c_{ij}$, $\forall x_{ij} = 0$.

Если система непотенциальна, то переходят к основному этапу (т.к. план не оптимален), иначе оптимальный план найден.

Основной этап.

1. Улучшаем план, то есть от плана X переходим к плану X' такому, что $Q(X) \geq Q(X')$.

2. Для плана X' строим новую систему $u_i', v_j',\ i = \overline{1,m},\ j = \overline{1,n}$, такую, что $v_j' - u_j' = c_{ij},\ \forall x_{ij} > 0$.

3. Исследуем систему u_i', v_j' на потенциальность. Если система непотенциальна, то переходим на п.1. Иначе найден оптимальный план.

Найдем методом потенциалов оптимальное решение задачи, взяв в качестве опорного план, построенный методом северо-западного угла (1-й шаг предварительного этапа).

v_j / u_i	v_1	v_2	v_3	v_4	v_5
u_1	2 30	4 80	2 10	3	8
u_2	3	5	6 10	– 6 20	+ 2
u_3	6	8	7	+ 4 10	– 5 30
u_4	3	4	2	1	4 60

2. Строим систему потенциалов:

$v_1 - u_1 = 2,$ $v_2 - u_1 = 4,$ $v_3 - u_1 = 2,$
$v_3 - u_2 = 6,$ $v_4 - u_2 = 6,$ $v_4 - u_3 = 4,$
$v_5 - u_3 = 5,$ $v_5 - u_4 = 4.$

Число неизвестных больше числа уравнений, поэтому можем взять, например, $u_1 = 0$ и найти значения остальных потенциалов, $u_2 = -4$, $u_3 = -2$, $u_4 = -1$, $v_1 = 2$, $v_2 = 4$, $v_3 = 2$, $v_4 = 2$, $v_5 = 3$.

3. Проверяем систему на потенциальность:

$v_1 - u_2 = 6 \le 3,$ $v_1 - u_3 = 4 \le 6,$ $v_1 - u_4 = 3 \le 3,$
$v_2 - u_2 = 8 \le 5,$ $v_2 - u_3 = 6 \le 8,$ $v_2 - u_4 = 5 \le 4,$
$v_3 - u_3 = 4 \le 7,$ $v_3 - u_4 = 3 \le 2,$ $v_4 - u_1 = 2 \le 3,$
$v_4 - u_4 = 3 \le 1,$ $v_5 - u_1 = 3 \le 8,$ $v_5 - u_2 = 7 \le 2,$

Система непотенциальна.

Переходим к общему этапу.

1. Выбираем клетку, для которой неравенство вида $v_j - u_i \le c_{ij}$ нарушается в наибольшей степени, то есть, находится число

$$\alpha_{i_0 j_0} = \max_{i,j} \left\{ \alpha_{ij} = v_j - u_i - c_{ij} > 0 \right\}$$

среди тех клеток, для которых условие (1) не выполняется: $\alpha_{i_0 j_0} = \alpha_{25} = 5$.

Начиная с клетки $i_0 j_0$ в направлении против часовой стрелки строится цепь из заполненных клеток таблицы (цикл). Совершая обход по цепи, помечаем

клетки, начиная с $i_0 j_0$, попеременно знаками + и –. Клетки со знаками + образуют положительную полуцепь, а со знаками – отрицательную полуцепь. В клетках отрицательной полуцепи ищем минимальную перевозку

$$\theta = \min\left\{x_{ij}^{-}\right\}.$$

Теперь улучшаем план следующим образом: перевозки отрицательной полуцепи уменьшаем на величину θ, а перевозки положительной полуцепи увеличиваем на θ. Новые

$$x_{ij}^{'} = \begin{cases} x_{ij}^{-} - \theta, \\ x_{ij}^{+} + \theta, \\ x_{ij}. \end{cases}$$

В нашем примере $\theta = \min\left\{x_{ij}^{-}\right\}$=20.

1. Новому плану соответствует таблица.

v_j / u_i	v_1	v_2	v_3	v_4	v_5
u_1	2 30	4 80	2 10	3	8
u_2	3	5	– 6 10	6 0	+ 2 20
u_3	6	8	7	4 30	5 10
u_4	3	4	+ 2	1	– 4 60

Затраты на перевозку по построенному плану равны:

$$Q = 30\times 2 + 4\times 80 + 2\times 10 + 6\times 10 + 4\times 30 + 2\times 20 + 5\times 10 + 4\times 60 = 910.$$

2. Строим систему потенциалов

$v_1 - u_1 = 2$, $v_2 - u_1 = 4$, $v_3 - u_1 = 2$,
$v_3 - u_2 = 6$, $v_5 - u_2 = 2$, $v_4 - u_3 = 4$,
$v_5 - u_3 = 5$, $v_5 - u_4 = 4$.

Полагаем $u_1 = 0$ и находим значения остальных потенциалов:

$u_2 = -4$, $u_3 = -7$, $u_4 = -6$, $v_1 = 2$, $v_2 = 4$, $v_3 = 2$, $v_4 = -3$, $v_5 = -2$.

3. Проверяем систему на потенциальность:

$v_1 - u_2 = 6 \le 3$, $v_1 - u_3 = 9 \le 6$, $v_1 - u_4 = 8 \le 3$,
$v_2 - u_2 = 8 \le 5$, $v_2 - u_3 = 11 \le 8$, $v_2 - u_4 = 10 \le 4$,
$v_3 - u_3 = 9 \le 7$, $v_3 - u_4 = 8 \le 2$, $v_4 - u_1 = -3 \le 3$,
$v_4 - u_4 = 3 \le 1$, $v_5 - u_1 = -2 \le 8$, $v_4 - u_2 = 1 \le 6$,

Система непотенциальна.

1. Находим $\alpha_{i_0 j_0} = \alpha_{43} = 6$, строим цикл, $\theta = \min\left\{x_{ij}^-\right\}$=10. Улучшаем план. Новому плану соответствует таблица.

u_i \ v_j	v_1	v_2	v_3	v_4	v_5
u_1	2 30	4 80	2 10	3	8
u_2	3	5	6 0	6 0	2 30
u_3	6	8	7	– 4 30	+ 5 10
u_4	3	4	2 10	+ 1	4 50

Затраты на перевозку по построенному плану равны:

$Q = 30\times 2 + 4\times 80 + 2\times 10 + 2\times 10 + 4\times 30 + 2\times 30 + 5\times 10 + 4\times 50 = 850$.

2. Строим систему потенциалов

$v_1 - u_1 = 2$, $v_2 - u_1 = 4$, $v_3 - u_1 = 2$,

$v_3 - u_4 = 2$, $v_5 - u_2 = 2$, $v_4 - u_3 = 4$,

$v_5 - u_3 = 5$, $v_5 - u_4 = 4$.

Полагаем $u_1 = 0$ и находим значения остальных потенциалов: $u_2 = 2$, $u_3 = -1$, $u_4 = 0$, $v_1 = 2$, $v_2 = 4$, $v_3 = 2$, $v_4 = 3$, $v_5 = 4$.

3. Проверяем систему на потенциальность:

$v_1 - u_2 = 0 \le 3$, $v_1 - u_3 = 3 \le 6$, $v_1 - u_4 = 2 \le 3$,

$v_2 - u_2 = 2 \le 5$, $v_2 - u_3 = 5 \le 8$, $v_2 - u_4 = 4 \le 4$,

$v_3 - u_3 = 3 \le 7$, $v_3 - u_2 = 0 \le 6$, $v_4 - u_1 = 3 \le 3$,

$v_4 - u_4 = 3 \le 1$, $v_5 - u_1 = 4 \le 8$, $v_4 - u_2 = 1 \le 6$,

Система непотенциальна.

1. Находим $\alpha_{i_0 j_0} = \alpha_{44} = 2$, строим цикл, $\theta = \min\left\{x_{ij}^-\right\}$=30. Улучшаем план. Новому плану соответствует таблица.

v_j / u_i	v_1	v_2	v_3	v_4	v_5
u_1	2 30	4 80	2 10	3	8
u_2	3	5	6	6	2 30
u_3	6	8	7	4 0	5 40
u_4	3	4	2 10	1 30	4 20

Затраты на перевозку по построенному плану равны:

$$Q = 30\times 2 + 4\times 80 + 2\times 10 + 2\times 10 + 1\times 30 + 2\times 30 + 5\times 40 + 4\times 20 = 790.$$

2. Строим систему потенциалов

$v_1 - u_1 = 2,$ $v_2 - u_1 = 4,$ $v_3 - u_1 = 2,$

$v_3 - u_4 = 2,$ $v_5 - u_2 = 2,$ $v_4 - u_4 = 1,$

$v_5 - u_3 = 5,$ $v_5 - u_4 = 4.$

Полагаем $u_1 = 0$ и находим значения остальных потенциалов:

$u_2 = 2,\ u_3 = -1,\ u_4 = 0,\ v_1 = 2,\ v_2 = 4,\ v_3 = 2,\ v_4 = 1,\ v_5 = 4.$

3. Проверяем систему на потенциальность:

$v_1 - u_2 = 0 \le 3,$ $v_1 - u_3 = 3 \le 6,$ $v_1 - u_4 = 2 \le 3,$

$v_2 - u_2 = 2 \le 5,$ $v_2 - u_3 = 5 \le 8,$ $v_2 - u_4 = 4 \le 4,$

$v_3 - u_3 = 3 \le 7,$ $v_3 - u_2 = 0 \le 6,$ $v_4 - u_1 = 1 \le 3,$

$v_4 - u_4 = 1 \le 1,$ $v_5 - u_1 = 4 \le 8,$ $v_4 - u_2 = -1 \le 6,$

Система *потенциальна*, следовательно, план *оптимален* и окончательные затраты $Q_{\min} = 790.$

Определение 4. Допустимый опорный план транспортной задачи называется невырожденным, если число заполненных клеток транспортной таблицы, т.е. число положительных перевозок $x_{ij} > 0$, равно $m + n + 1$, где m – число пунктов отправления, n – число пунктов назначения.

Определение 5. Если допустимый опорный план содержит менее $m + n + 1$ элементов $x_{ij} > 0$, то он называется вырожденным, а транспортная задача называется вырожденной транспортной задачей.

Следующая теорема позволяет определить вырожденность задачи до ее решения.

Теорема. Для невырожденной транспортной задачи необходимо и достаточно отсутствие такой неполной группы пунктов производства, суммарный

объем производства которой точно совпадает с суммарными потребностями некоторой группы пунктов потребления.

Другими словами, это условие означает, что для любых двух систем индексов $i_1, i_2, ..., i_t$, $j_1, j_2, ..., j_S$, где $t + S < n + m$, имеет место неравенство $\sum_{k=1}^{t} a_{i_k} \neq \sum_{k=1}^{S} b_{j_k}$. (Доказательство не сложно, от противного.)

Для решения транспортной задачи методом потенциалов строится система потенциалов $v_j - u_i = c_{ij}$, $\forall x_{ij} > 0$. Если опорное решение невырожденно, то число неизвестных на 1 больше числа уравнений. При вырожденном опорном решении число этих уравнений еще меньше. По аналогии симплекс-методом, в невырожденном решении $x_{ij} > 0$ представляют собой базисные переменные, а $x_{ij} = 0$ – небазисные. Если опорное решение вырожденно, то часть базисных переменных принимает нулевые значения.

Пусть первое опорное решение, найденное методом северо-западного угла или методом минимального элемента, является вырожденным. Тогда, чтобы решать задачу методом потенциалов необходимо выбрать в качестве базисных переменных некоторые перевозки $x_{ij} = 0$ и для них также составить уравнения $v_j - u_i = c_{ij}$ по условию (2) теоремы. Какие перевозки вида $x_{ij} = 0$ включать в базисные? Выбираются такие клетки таблицы с $x_{ij} = 0$, чтобы из базисных переменных нельзя было организовать ни одного цикла!

При переходе к новому улучшенному плану задачи в небазисные переменные переводится перевозка в отрицательной полуцепи, которая находится следующим образом $\theta = \min\left\{x_{ij}^{-}\right\}$. В вырожденной задаче это значение может достигаться на нескольких перевозках x_{ij} отрицательной полуцепи. В этом случае на каждом шаге в небазисные переменные переводится та минимальная перевозка отрицательной полуцепи, которая связана с пунктом производства, имеющим меньший номер. Это правило уменьшает вероятность возникновения зацикливания, что само по себе достаточно редкое явление в практических задачах.

СПИСОК ЛИТЕРАТУРЫ

1. Березин И.С., Жидков Н.П. Методы вычислений. Том 1 и 2. – М.: “Наука”, 1994.

2. Бахвалов Н.С. Численные методы. – М.: “Наука”, 1993.

3. Калиткин Н.Н. Численные методы. – М.: “Наука”, 1991.

4. Зуховицкий С.И., Авдеева Л.И. Линейное и выпуклое программирование. – М.: “Наука”, 1994.

5. Реклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике: В 2-х кн. Кн.1. Пер. с англ.- М.: Мир, 1986.

6. Реклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике: В 2-х кн. Кн.2. Пер. с англ.- М.: Мир, 1986.

7. Атманов С.А. Линейное программирование. М.: “Наука”, 1981.

8. Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М., Методы оптимизации. М.: – Наука, 1978.

9. Лесин В.В., Лисовец Ю.П. Основы методов оптимизации. – М.: Изд-во МАИ, 1995. – 344 с.

10. Растригин Л.А. Случайный поиск в задачах оптимизации многопараметрических систем. Рига, Зинатне,1965. 212 с.

11. Измаилов А.Ф., Солодов М.В. Численные методы оптимизации: Учеб. пособие. – М.: ФИЗМАТЛИТ, 2005. – 304 с.

12. Растригин Л.А. Статистические методы поиска. М.: Наука, 1968. 376 с.

13. Растригин Л.А. Случайный поиск. М.: Знание, 1979. 64 с.

14. Электронный учебник: В.И. Рейзлин. Методы оптимизации. Томск, ТПУ: http://109.123.146.125/

15. NEOS Wiki – электронный ресурс: http://wiki.mcs.anl.gov/NEOS/index.php/NEOS_Wiki, метод доступа – свободный.

16. Optimization – From Wikipedia: http://en.wikipedia.org/wiki/Optimization_(mathematics)

СОДЕРЖАНИЕ

1. ВВЕДЕНИЕ 3
1.1. Постановка задач оптимизации 3
1.2. Математическая постановка задач оптимизации 5
1.2.1. Виды ограничений 5
1.2.2. Критерии оптимальности 6
1.2.3. Классификация задач 9
2. ОДНОМЕРНАЯ ОПТИМИЗАЦИЯ 11
2.1. Методы сужения интервала неопределенности 11
2.1.1. Общий поиск 11
2.1.2. Унимодальные функции 12
2.1.3. Метод деления интервала пополам 13
2.1.4. Метод золотого сечения 14
2.1.5. Установление первоначального интервала неопределенности 17
2.2. Ньютоновские методы 18
3. МИНИМУМ ФУНКЦИИ МНОГИХ ПЕРЕМЕННЫХ 21
3.1. Рельеф функции 21
3.2. Метод покоординатного спуска (Метод Гаусса) 23
3.3. Метод оврагов 24
4. МЕТОДЫ С ИСПОЛЬЗОВАНИЕМ ПРОИЗВОДНЫХ 26
4.1. Градиентные методы 28
4.2. Метод Ньютона 29
4.3. Метод Марквардта 30
5. УСЛОВНАЯ ОПТИМИЗАЦИЯ 33
5.1. Задачи с ограничениями в виде равенств 33
5.1.1. Множители Лагранжа 33
5.2. Задачи с ограничениями в виде неравенств 36
5.2. Методы штрафных функций 39
5.3. Метод факторов 42
6. Случайный поиск 44
6.1. Простой случайный поиск 44
6.2. Ненаправленный случайный поиск 45

6.3. Направленный случайный поиск ... 45
6.3.1. Алгоритм парной пробы ... 45
6.3.2. Алгоритм наилучшей пробы ... 46
6.3.3. Метод статистического градиента ... 47
6.3.4. Алгоритм наилучшей пробы с направляющим гиперквадратом ... 47
6.4. Алгоритмы глобального поиска ... 48
7. ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ ... 51
7.1. Примеры задач линейного программирования ... 51
7.1.1. Задача об использовании сырья ... 51
7.1.2. Задача об использовании мощностей оборудования ... 52
7.1.3. Транспортная задача ... 54
7.1.4. Задача о питании ... 56
7.2. Основная задача линейного программирования ... 57
7.3. Основная задача линейного программирования с ограничениями-неравенствами ... 60
7.4. Геометрическое толкование задач линейного программирования ... 64
7. СИМПЛЕКС МЕТОД ИЛИ МЕТОД ПОСЛЕДОВАТЕЛЬНОГО УТОЧНЕНИЯ ОЦЕНОК ... 74
7.1. Алгоритм симплекс метода ... 77
7.2. Вырожденность в задачах линейного программирования ... 81
7.3. Двойственность задачи линейного программирования ... 82
7.4. Метод последовательного уточнения оценок ... 89
7.5. Методы решения транспортной задачи ... 91
7.5.1. Метод северо-западного угла ... 93
7.5.2. Метод минимального элемента ... 93
7.5.3. Метод потенциалов ... 94
СПИСОК ЛИТЕРАТУРЫ ... 102
СОДЕРЖАНИЕ ... 103

Учебное издание

Рейзлин Валерий Израилевич

ЧИСЛЕННЫЕ МЕТОДЫ ОПТИМИЗАЦИИ

Учебное пособие

Издано в авторской редакции

Научный редактор
доктор технииических наук
профессор В.К. Погребной

Отпечатано в Издательстве ТПУ в полном соответствии с качеством предоставленного оригинал-макета

Подписано к печати 30.11.2011. Формат 60x84/17. Бумага «Снегурочка».
Печать XEROX. Усл. печ. л. 6.1. Уч.- изд. л. 5,53.
Заказ . Тираж 100 экз.

Национальный исследовательскийТомский политехнический университет
Система менеджмента качества
Издательства Томского политехнического университета сертифицирована
NATIONAL QUALITY ASSURANCE по стандарту BS EN ISO 9001:2008

ИЗДАТЕЛЬСТВО ТПУ. 634050, г. Томск, пр. Ленина, 30
Тел./факс: 8(3822)56-35-35, www.tpu.ru