МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

В. Г. Жадан

МЕТОДЫ ОПТИМИЗАЦИИ

ЧАСТЬ II

ЧИСЛЕННЫЕ АЛГОРИТМЫ

Допущено
Учебно-методическим объединением
высших учебных заведений Российской Федерации
по образованию в области прикладных математики и физики
в качестве учебного пособия для студентов вузов
по направлению подготовки «Прикладные математика и физика»

МОСКВА
МФТИ
2015

УДК 519.8(075)
ББК 22.18я73
Ж15

Рецензенты:
Кафедра исследования операций
факультета вычислительной математики и кибернетики
Московского государственного университета им. М. В. Ломоносова
(зав. кафедрой академик РАН *П. С. Краснощеков*)

Доктор физико-математических наук, профессор *А. В. Лотов*

Жадан, В. Г.

Ж15 Методы оптимизации. Ч. II. Численные алгоритмы : учебное пособие / В. Г. Жадан. – М. : МФТИ, 2015. – 320 с.
ISBN 978-5-7417-0571-1 (Ч. II)

Книга является учебным пособием по теории и численным методам решения оптимизационных задач. Написана на основе материалов курса «Методы оптимизации», читаемого автором в течение нескольких лет студентам 3 курса факультета управления и прикладной математики Московского физико-технического института (государственного университета). Первая часть книги посвящена основам выпуклого анализа и условиям оптимальности для различных классов оптимизационных задач.

Предназначено для студентов, специализирующихся в области прикладной математики. Может быть использовано для самостоятельного изучения основных разделов теории оптимизации.

УДК 519.8(075)
ББК 22.18я73

ISBN 978-5-7417-0571-1 (Ч. II)
ISBN 978-5-7417-0516-2

Оглавление

Предисловие

Данная вторая часть учебного пособия написана на основе лекций, читаемых во втором семестре двухсеместрового курса «Методы оптимизации» и является продолжением первой части [33]. Посвящена вторая часть в основном численным алгоритмам решения различных классов задач оптимизации в конечномерных пространствах. Рассматриваются численные методы решения задач одномерной и безусловной оптимизации. Приводятся также численные методы решения задач условной оптимизации, включая, в частности, задачи линейной, квадратичной и выпуклой оптимизации. Обсуждаются некоторые подходы к решению задач глобальной оптимизации и возможность обобщения методов нелинейной условной оптимизации для решения задач многокритериальной оптимизации.

Так как количество известных в настоящее время численных методов весьма огромно, то охватить их хотя бы в минимальном объеме нет никакой возможности. Поэтому при изложении материала автор старался рассмотреть главным образом те методы, которые исторически считаются важными или с теоретической, или с практической точек зрения. Делается это с целью продемонстрировать разные подходы и идеи к построению численных методов решения задач оптимизации.

Список методов, включенных в пособие, все-таки оказался весьма обширным. В силу недостатка времени не все эти методы удается затронуть на лекциях и семинарах с достаточной степенью подробности, часть из них предназначена для самостоятельного изучения.

Как и при написании первой части, автор считает своей приятной обязанностью выразить признательность коллегам по кафедре математических основ управления О.С. Федько, А.Г. Бирюкову, Р.В. Константинову, П.Е. Двуреченскому, А.А. Орлову, Ю.В. Дорну за совместную работу по преподаванию данного курса и за полезные замечания, а также благодарит В.У. Малкову за помощь в подготовке иллюстративного материала.

Введение

Численные методы являются основным средством решения оптимизационных задач, возникающих в практических приложениях. Поэтому их разработке постоянно уделялось повышенное внимание, начиная со второй половины XX века. К настоящему времени предложено много эффективных численных методов, предназначенных для решения разных классов задач. Одним из наиболее известных из них является симплекс-метод для задач линейного программирования. Разработанный в середине прошлого века независимо Л.В. Канторовичем [41] и Дж. Б. Данцигом [23], он оказался способным решать прикладные задачи большой размерности. Более того, его появление послужило толчком к развитию других численных методов, предназначенных для решения как линейных, так и нелинейных задач оптимизации.

В настоящем пособии рассматривается лишь незначительная часть предложенных к настоящему времени численных методов. Их выбор обусловлен, с одной стороны, желанием охватить основные классы задач конечномерной оптимизации, а именно: задачи одномерной и безусловной оптимизации, задачи линейного, квадратичного и нелинейного программирования. С другой стороны, хотелось привлечь внимание к тем методам, при построении которых использовались оригинальные подходы. Поэтому при выборе конкретных методов для включения в настоящее пособие предпочтение в первую очередь отдавалось методам, которые наиболее известны и которые стали по существу классическими. Помимо того, хотелось затронуть и некоторые современные методы, а также коснуться тех теоретических вопросов, которые интересовали внимание исследователей в последнее время.

Здесь несомненно следует упомянуть результаты, связанные со сложностью задач и эффективностью численных методов. На важность этих понятий при решении оптимизационных задач впервые указал выпускник ФУПМ Л.Г. Хачиян [79]. Ему удалось разработать метод решения задачи линейного программирования, который обладает по-

линомиальной трудоемкостью в отличие от симплекс-метода, теоретическая трудоемкость которого лишь экспоненциальная. Хотя метод эллипсоидов, предложенный им, не оказался работоспособным на практике, но он породил массу исследований, результатом которых стала разработка новых полиномиальных численных методов. Главным образом это так называемые методы внутренней точки, которые сопоставимы или превышают по эффективности симплекс-метод, особенно для задач большой и сверхбольшой размерности. Более того, удалось предложить полиномиальные методы для нелинейных выпуклых задач, перенести их на другие классы задач.

Многие методы оптимизации реализованы практически и собраны в виде пакетов программ для решения различных классов задач. Более того, были разработаны диалоговые системы оптимизации, позволяющие решать задачи в интерактивном режиме. Теоретические знания о численных методах, в том числе и приведенные в настоящем учебном пособии, могут помочь более осознанному применению подходящих численных методов из этих систем и пакетов.

Материал в учебном пособии излагается в следующем порядке.
В главе 1 даются основные понятия, связанные со сходимостью численных методов и скоростью сходимости. Глава 2 посвящена методам одномерной минимизации и методам безусловной оптимизации, в частности методу градиентного спуска, методу Ньютона, методу сопряженных градиентов. Методы решения задач линейного и квадратичного программирования рассматриваются в главе 3. Здесь главное внимание уделяется методам симплексного типа.

Изучение методов решения нелинейных задач условной оптимизации начинается с главы 4. В этой главе приводятся методы решения задач на допустимых множествах простого вида, в частности метод проекции градиента, методы условного и приведенного градиента. Главы 5 и 6 посвящены тем методам, которые основаны на решении последовательности более простых вспомогательных задач, для которых уже применимы методы из предыдущих глав. От ограничений теперь требуется, чтобы они описывались функционально через равенства и неравенства. В главе 7 рассматриваются методы, использующие функцию Лагранжа или ее модификации. Некоторые методы внутренней точки, предназначенные для решения задач линейного программирования, приводятся в главе 8. Глава 9 посвящена вопросам сложности задач и эффективности численных методов. Наконец, в главах 10 и 11 затрагиваются методы решения многоэкстремальных и многокритериальных задач оптимизации.

Для большинства основных утверждений, особенно касающихся сходимости методов, приводятся доказательства. Делается это с целью дать возможность интересующимся студентам самостоятельно ознакомиться с используемыми при этом подходами.

Список основных обозначений сохраняется полностью из первой части.

Список основных обозначений

$J^n = [1:n]$ — множество целых чисел от 1 до n;

$\mathbb{R}$ — множество вещественных чисел;

$\mathbb{R}_+$ — множество неотрицательных вещественных чисел;

$\mathbb{R}_{++}$ — множество положительных вещественных чисел;

$\mathbb{R}^n$ — n-мерное пространство вещественных векторов;

$\mathbb{R}^n_+ = \left\{x \in \mathbb{R}^n : \ x^i \geq 0, \ 1 \leq i \leq n\right\}$ — неотрицательный ортант $\mathbb{R}^n$;

$\mathbb{R}^n_{++} = \left\{x \in \mathbb{R}^n : \ x^i > 0, \ 1 \leq i \leq n\right\}$ — положительный ортант $\mathbb{R}^n$;

$\mathbb{R}^n_{-} = \left\{x \in \mathbb{R}^n : \ x^i \leq 0, \ 1 \leq i \leq n\right\}$ — неположительный ортант $\mathbb{R}^n$;

$\mathbb{R}^n_{--} = \left\{x \in \mathbb{R}^n : \ x^i < 0, \ 1 \leq i \leq n\right\}$ — отрицательный ортант $\mathbb{R}^n$;

x_+ — вектор с координатами $x^i_+ = \max\left[x^i, 0\right]$, $1 \leq i \leq n$;

x_- — вектор с координатами $x^i_- = \min\left[x^i, 0\right]$, $1 \leq i \leq n$;

$\langle x, y\rangle = \sum_{i=1}^n x^i y^i$ — евклидово скалярное произведение;

$0_n = [0, \ldots, 0]^T$ — нулевой n-мерный вектор;

0_{mn} — нулевая матрица размера $m \times n$;

I_n — единичная матрица порядка n;

$D(x)$ — диагональная матрица с вектором x на диагонали;

$\|x\|$ — норма вектора x (как правило, если не дано уточнения, имеется в виду евклидова норма);

$\|x\|_p = \left(\sum_{i=1}^n |x^i|^p\right)^{\frac{1}{p}}$ — p-я гельдеровская норма вектора x, $1 \leq p < \infty$ (при $p = 2$ совпадает с евклидовой нормой);

$\|x\|_\infty = \max_{1 \leq i \leq n} |x^i|$ — чебышовская норма вектора x (называемая также максимальной или кубической нормой);

$\|A\|$ — норма матрицы A (согласованная с нормой в пространствах векторов).

Глава 1

Основные определения

Нас будут интересовать алгоритмы решения оптимизационных задач вида

$$f_* = \min_{x \in X} f(x), \tag{1.1.1}$$

где как допустимое множество X, так и множество оптимальных решений $X_* \subseteq X$ предполагаются непустыми. Найти решение задачи (1.1.1) аналитически, применяя условия оптимальности или используя геометрический подход, удается лишь в самых простых случаях. Как правило, это возможно только тогда, когда целевая функция и допустимое множество выпуклы. Но в подавляющем большинстве случаев, особенно в практических приложениях, где зачастую размерность вектора x большая и функция $f(x)$ имеет сложный вид, решение задачи аналитически становится невозможным. Поэтому приходится обращаться к *численным методам*, т.е. к специальным алгоритмическим процедурам, которые находили бы решение задачи или хотя бы достаточно близкое к нему приближение.

При построении численных алгоритмов решения оптимизационных задач широко используются методы и приемы, наработанные в вычислительной математике. В частности, к ним можно отнести методы решения систем линейных и нелинейных уравнений, методы продолжения решения по параметру, методы погружения задачи в параметрическое семейство задач, масштабирование, монотонные преобразования функций и т.д. Практически все наиболее важные понятия и теоретические результаты, связанные со сходимостью итерационных методов, также заимствованы из вычислительной математики. Использование методов вычислительной математики при решении оптимизационных задач объясняется прежде всего желанием свести решение последних

к нахождению точек, удовлетворяющих условиям оптимальности. Как нам известно, многие условия оптимальности описываются через системы равенств и неравенств.

Наряду с этим при построении численных методов применяются подходы, которые существенным образом учитывают специфику оптимизационных задач. Прежде всего принимают во внимание, к какому классу относится целевая функция $f(x)$ и каким образом задается допустимое множество X, например выпуклые они или нет. Выделяют численные методы отыскания локальных и глобальных решений. Их так и называют *методы локальной и глобальной оптимизации*.

К другим наиважнейшим приемам, использующим природу условных оптимизационных задач, несомненно следует отнести *принцип Лагранжа*, заключающийся в сворачивании целевой функции с ограничениями. Этот принцип играет фундаментальную роль как при теоретическом исследовании задач, так и при построении численных методов. Из других приемов, оказывающихся полезными при построении численных методов, отметим идею *штрафования ограничений*, когда налагается «штраф» на целевую функцию в точках, выходящих за пределы допустимого множества. В другом варианте это использование *барьеров*, не позволяющих покидать допустимое множество.

Любой численный метод основан на вычислении значений целевой функции и функций, описывающих ограничения, а также, быть может, их производных. Метод называют *пассивным*, если точки, в которых проводятся вычисления, указываются заранее, причем независимо друг от друга. Метод называют *последовательным* или *итерационным*, если такие точки вычисления значений функций и их производных выбираются в процессе счета на основе информации, полученной в ходе вычислительного процесса. Подавляющее большинство численных методов, используемых на практике, оказываются итерационными процедурами.

Итерационный численный метод генерирует последовательность точек x_0, x_1, x_2, ..., которая, вообще говоря, является бесконечной. При этом начальная точка x_0 задается, а последующие точки вычисляются по предыдущим по определенным правилам. Точка x_k называется *k-м приближением* к решению, а вся совокупность точек $\{x_k\}$ — *траекторией*.

Переход от x_k к точке x_{k+1} называется *шагом или итерацией* метода. Способ этого перехода и составляет суть метода, поэтому часто метод записывается в виде итерационной схемы:

$$x_{k+1} = \Phi_k(x_k), \quad k = 0, 1, 2, \ldots \tag{1.1.2}$$

О рекуррентном соотношении (1.1.2) говорят как об *алгоритме* итерационного численного метода, а об отображении $\Phi_k(x)$ — как об его *алгоритмическом отображении*.

Приведенная схема (1.1.2) является *одношаговой*, так как в ней для определения последующего приближения x_{k+1} используется только предыдущее приближение x_k. Иногда рассматривают *многошаговые* схемы, в которых для нахождения x_{k+1} кроме x_k используются и другие приближения: x_{k-1}, x_{k-2} и т.д. Например, в *двухшаговых* схемах обычно используются x_k и x_{k-1}. Тогда перед началом вычислений надо помимо x_0 задать и x_1, а итерационная схема (1.1.2) заменяется на $x_{k+1} = \Phi_k(x_k, x_{k-1})$.

Порядком метода называют максимальный порядок производных целевой функции и ограничений, используемых при осуществлении итерации метода. Различают методы *нулевого, первого и второго порядков*. В методах нулевого порядка используются только значения функций, в методах первого порядка наряду со значениями функций задействованы также их первые производные. Если помимо значений функций и их первых производных применяются дополнительно вторые производные, то такие методы относятся к методам второго порядка. Известны также и методы более высокого порядка.

Метод называют *конечношаговым* или просто *конечным*, если любое начальное приближение корректно определяет траекторию, некоторая точка которой совпадает с решением задачи. Такие методы удается построить лишь для специальных классов задач. В противном случае метод называется *бесконечношаговым*, в нем траектория, вообще говоря, не попадает в точное решение задачи, а лишь аппроксимирует его. Разумеется, наибольший интерес представляют те бесконечношаговые численные методы, которые с увеличением числа итераций давали бы лучшую аппроксимацию решения задачи. Другими словами, желательно, чтобы метод обладал сходимостью к решению. Приведем некоторые общие определения, касающиеся основных понятий, связанных со сходимостью итерационных процессов типа (1.1.2).

Определение 1.1.1. *Метод называется сходящимся к решению* $x_* \in X_*$ *(из точки* x_0*), если* $\lim_{k\to\infty} x_k = x_*$.

Иногда метод генерирует последовательность $\{x_k\}$, которая не является сходящейся, но содержит сходящиеся подпоследовательности. Определение 1.1.1 в этом случае переформулируется следующим образом.

Определение 1.1.2. *Метод называется сходящимся к множеству решений X_* (из точки x_0), если* $\lim_{k\to\infty}\rho(x_k, X_*) = 0$, *где* $\rho(x, X)$ *— расстояние от точки x до множества X.*

В том случае, когда множество X_* есть компакт, из сходимости метода к X_* следует существование сходящихся подпоследовательностей у последовательности $\{x_k\}$.

Определение 1.1.3. *Последовательность $\{x_k\}$ называется минимизирующей, если*

$$\lim_{k\to\infty} f(x_k) = f_* = \inf_{x\in X} f(x). \tag{1.1.3}$$

При выполнении (1.1.3), говорят, что метод *сходится по функционалу*. Может оказаться так, что метод сходится по функционалу, но сходимости по x нет. В качестве примера приведем следующую задачу:

$$\min_{x\in X} f(x), \quad f(x) = \frac{x}{1+x^2}, \quad X = \mathbb{R}_+ = [0, +\infty),$$

в которой $f_* = 0$, $X_* = \{0\}$. Последовательность $\{x_k\}$ с $x_k = k + 1$, где $k \geq 0$, является минимизирующей для этой задачи, однако она не сходится к решению задачи — точке $x_* = 0$.

Одной из основных характеристик сходимости бесконечношаговых методов является *скорость сходимости*. Будем в основном рассматривать эту характеристику относительно сходимости по аргументу. Следует отметить, что используемая здесь терминология, касающаяся скорости сходимости, широко применяется при исследовании численных методов решения систем нелинейных уравнений и по существу полностью заимствована из этого раздела вычислительной математики.

Пусть последовательность $\{x_k\}$ сходится к точке x_*, являющейся решением задачи (1.1.1).

Определение 1.1.4. *Метод сходится к x_* с линейной скоростью, если можно указать такую константу $0 < C < 1$ и номер $K \geq 0$, что*

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|$$

для любых $k \geq K$.

Определение 1.1.5. *Метод сходится к x_* со сверхлинейной скоростью, если*

$$\|x_{k+1} - x_*\| \leq C_k\|x_k - x_*\|,$$

где $C_k > 0$ и $C_k \to 0$ при $k \to \infty$.

Определение 1.1.6. *Метод сходится к x_* с квадратичной скоростью, если можно указать такую константу $C > 0$ и номер $K \geq 0$, что*

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|^2$$

для любых $k \geq K$.

С линейной скоростью сходимости тесно связано такое понятие, как сходимость со *скоростью геометрической прогрессии* (по существу это одно и то же).

Определение 1.1.7. *Пусть можно указать константы $C > 0$, $0 < q < 1$ и номер $K \geq 0$ такие, что*

$$\|x_k - x_*\| \leq Cq^k$$

при $k \geq K$. Тогда говорят, что метод сходится со скоростью геометрической прогрессии, где q — знаменатель прогрессии.

Аналогичные определения скорости сходимости могут быть даны и касательно сходимости по функционалу. В них только x_k заменяется на $f(x_k)$, а x_* — на f_*. Например, линейная скорость сходимости по функционалу означает, что

$$|f(x_{k+1}) - f_*| \leq C|f(x_k) - f_*|$$

для некоторого $0 < C < 1$ и при всех $k \geq K$.

Вычислительный процесс не может длиться бесконечно долго, поэтому все бесконечношаговые методы должны иметь дополнительно *правила останова*, при выполнении которых расчеты прерываются. Эти правила носят эвристический характер и зависят от вида задачи. Они обычно различаются для задач с ограничениями и без них. Учитываются также гладкость функций, их вид (являются ли они линейными, квадратичными) и т.д. Правила останова зависят также от применяемого конкретного метода, а именно, какие величины вычисляются в нем в ходе итерационного процесса. В качестве примера приведем наиболее часто применяемые правила останова при решении задач безусловной минимизации.

Пусть ε_1, ε_2 и ε_3 — некоторые параметры требуемой точности. Тогда процесс прерывается, если выполняется какое-нибудь из неравенств:

1) $\|x_{k+1} - x_k\| \leq \varepsilon_1$,

2) $|f(x_{k+1}) - f(x_k)| \leq \varepsilon_2$.

Если функция $f(x)$ дифференцируемая, то можно проверять следующее условие:

3) $\|f_x(x_k)\| \le \varepsilon_3$,

указывающее на то, с какой точностью выполняется необходимое условие оптимальности в задаче.

Правила 1) и 2) основаны на использовании *абсолютных изменений* соответственно аргумента и значений целевой функции. Гораздо более обосновано использование их *относительных изменений*:

4) $\|x_{k+1} - x_k\| \le \varepsilon_1 (1 + \|x_{k+1}\|)$,

5) $|f(x_{k+1}) - f(x_k)| \le \varepsilon_2 (1 + |f(x_{k+1}|)$.

Приведенные правила весьма ненадежны и не гарантируют близости полученных в результате останова приближенных решений к точному решению задачи. Обычно эти правила дополняются заданием максимально возможного числа итераций, которые позволяется выполнить в ходе вычислительного процесса.

Правила останова задач математического программирования более сложные по сравнению с задачами безусловной оптимизации, так как в них дополнительно учитывается степень удовлетворения текущих точек x_k ограничениям задачи.

Глава 2

Методы одномерной и безусловной минимизации

2.1. Методы одномерной минимизации

Одной из наиболее простых с точки зрения постановки, но часто встречающейся в приложениях, является задача минимизации функции одного аргумента на отрезке:

$$f_* = \min_{x \in X} f(x), \tag{2.1.1}$$

где $x \in \mathbb{R}$, $X = [a, b]$. При этом считаем, что $a < b$ и функция $f(x)$ непрерывна на $[a, b]$.

Задача (2.1.1) нередко используется в качестве вспомогательной подзадачи во многих численных методах, предназначенных для решения более сложных оптимизационных задач. Поэтому эффективность решения (2.1.1) в сильной степени влияет на эффективность этих методов. Это вызывает повышенные требования к алгоритмам решения задачи (2.1.1).

Алгоритмы решения задачи (2.1.1) существенным образом зависят от свойств функции $f(x)$. Мы в данном разделе ограничимся рассмотрением функций, которые получили название *унимодальных*. Приведем определение унимодальности в общем случае для произвольных функций $f(x)$, не предполагая их непрерывности.

Определение 2.1.1. *Функция $f(x)$ называется унимодальной на $[a, b]$, если существует такая точка $x_* \in [a, b]$, что $f(x_1) > f(x_2)$ для любых $a \le x_1 < x_2 < x_*$ и $f(x_1) < f(x_2)$ для любых $x_* < x_1 < x_2 \le b$.*

Если на $[a, b]$ унимодальная функция $f(x)$ достигает своего минимума, то она достигает его именно в точке x_*. Для непрерывных функций свойство унимодальности означает обязательное наличие у функции $f(x)$ единственного локального минимума на $[a, b]$, который одновременно является и глобальным минимумом. Можно показать, что для непрерывной функции одного аргумента свойство унимодальности на отрезке $[a, b]$ равносильно их строгой квазивыпуклости на $[a, b]$. Графики разрывной и непрерывной унимодальных функций приведены соответственно на рис. 2.1 и рис. 2.2.

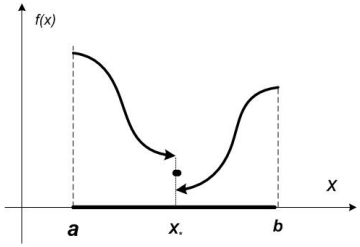

Рис. 2.1. Пример разрывной унимодальной функции

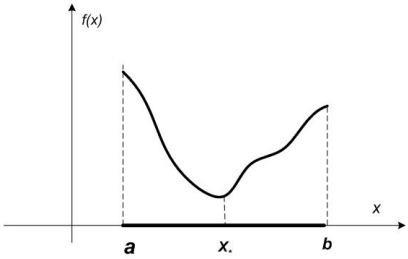

Рис. 2.2. Пример непрерывной унимодальной функции

Утверждение 2.1.1. *Пусть $f(x)$ — унимодальная на отрезке $[a, b]$ функция, достигающая своего минимума в точке $x_* \in [a, b]$. Пусть, кроме того, имеются две точки $x_1 \in [a, b]$ и $x_2 \in [a, b]$ такие, что $x_1 < x_2$. Тогда:*

1) если $f(x_1) \le f(x_2)$, то $x_* \in [a, x_2]$;

2) если $f(x_1) \geq f(x_2)$, *то* $x_* \in [x_1, b]$.

Доказательство. Докажем только первое утверждение. От противного, предположим, что $f(x_1) \leq f(x_2)$, но $x_* > x_2$. Тогда обязательно $x_1 < x_2 < x_*$ и в силу унимодальности функции $f(x)$ должно выполняться неравенство: $f(x_1) > f(x_2)$. Получено противоречие. ■

Для решения задач (2.1.1) с непрерывными унимодальными функциями достаточно эффективными являются **методы последовательной локализации решения**. В самом общем виде алгоритмы, принадлежащие к этому классу методов, можно описать следующим образом. Строится последовательность вложенных друг в друга *отрезков локализации* решения x_* задачи (2.1.1):

$$[a, b] \supseteq [a_0, b_0] \supset [a_1, b_1] \supset \cdots \supset [a_k, b_k] \supset \cdots . \tag{2.1.2}$$

Делается это таким образом, что каждый из отрезков $[a_k, b_k]$ содержит точку x_*. В качестве начального отрезка $[a_0, b_0]$, как правило, берется сам исходный отрезок $[a, b]$. Длины $\Delta_k = b_k - a_k$ отрезков $[a_k, b_k]$ образуют монотонно уменьшающуюся последовательности чисел

$$\Delta_0 > \Delta_1 > \Delta_2 > \cdots > \Delta_k > \cdots ,$$

стремящихся к нулю.

Если построен отрезок $[a_k, b_k]$, то в качестве *приближенного решения* задачи (2.1.1) берется произвольная точка x_k из отрезка $[a_k, b_k]$. Обычно в качестве такой точки целесообразно брать середину отрезка — точку

$$x_k = \frac{a_k + b_k}{2}. \tag{2.1.3}$$

Тогда гарантировано, что $|x_k - x_*| \leq \frac{\Delta_k}{2}$. В общем случае при выборе в качестве x_k произвольной точки из отрезка $[a_k, b_k]$ имеем только оценку: $|x_k - x_*| \leq \Delta_k$.

Для прерывания процесса построения отрезков (2.1.2) задаются *точностью решения задачи* (2.1.1) — положительным параметром ε. Под точностью в таких методах понимают оценку расстояния между точным решением x_* задачи (2.1.1) и найденным приближенным решением x_k, т.е. упоминавшуюся уже величину $|x_k - x_*|$. Процесс прерывают, когда выполняется неравенство: $|x_k - x_*| \leq \varepsilon$. Если в качестве точки x_k согласно (2.1.3) берется середина отрезка $[a_k, b_k]$, то для этого достаточно совершить K итераций, где K — наименьшее из целых неотрицательных чисел, для которого выполняется условие

$\Delta_K \le 2\varepsilon$. Следует, однако, подчеркнуть, что длина отрезка локализации решения в два раза превышает точность решения. Она совпадает с точностью решения только в том случае, когда нам из каких-то других соображений известно, с какой стороны от «срединной» точки находится решение задачи.

Разные алгоритмы, принадлежащие к классу методов последовательной локализации решения, отличаются между собой конкретными способами построения отрезков $[a_k, b_k]$ или, говоря более точно, конкретными способами перехода от отрезка $[a_k, b_k]$ к следующему отрезку $[a_{k+1}, b_{k+1}]$. При этом все эти методы существенным образом используют то обстоятельство, что минимизируемая функция $f(x)$ является унимодальной функцией на $[a, b]$.

Метод дихотомии (метод деления отрезка пополам) является одним из наиболее простых и интуитивно понятных методов решения задачи (2.1.1). Идея метода заключается в поиске на каждом шаге такого отрезка $[a_{k+1}, b_{k+1}]$ внутри отрезка $[a_k, b_k]$, который имел бы вдвое меньшую длину по сравнению с длиной отрезка $[a_k, b_k]$ и содержал бы решение задачи (2.1.1).

Начальная итерация. Полагаем $a_0 = a$, $b_0 = b$, $c_0 = \frac{a_0+b_0}{2}$ и вычисляем $f(c_0)$. Полагаем $k = 0$.

Общая k-я итерация

Шаг 1. Берем точку $y_k = \frac{a_k+c_k}{2}$ и вычисляем $f(y_k)$. Возможны два случая:

а) Если $f(y_k) \le f(c_k)$, то полагаем

$$a_{k+1} = a_k, \quad b_{k+1} = c_k, \quad c_{k+1} = y_k$$

и переходим на *Шаг* 3.

б) Если $f(y_k) > f(c_k)$, то берем новую точку $z_k = \frac{c_k+b_k}{2}$ и вычисляем $f(z_k)$.

Шаг 2. Если $f(c_k) \le f(z_k)$, то полагаем

$$a_{k+1} = y_k, \quad b_{k+1} = z_k, \quad c_{k+1} = c_k,$$

иначе

$$a_{k+1} = c_k, \quad b_{k+1} = b_k, \quad c_{k+1} = z_k.$$

Шаг 3. Увеличиваем номер итерации $k := k + 1$ и идем на шаг 1.

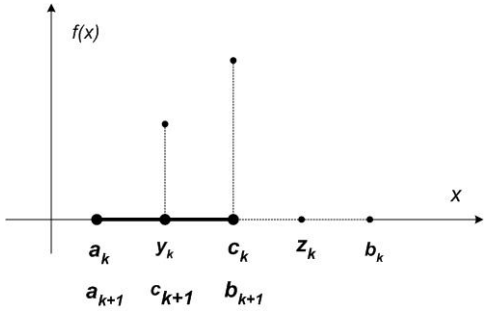

Рис. 2.3. Случай $f(y_k) \leq f(c_k)$

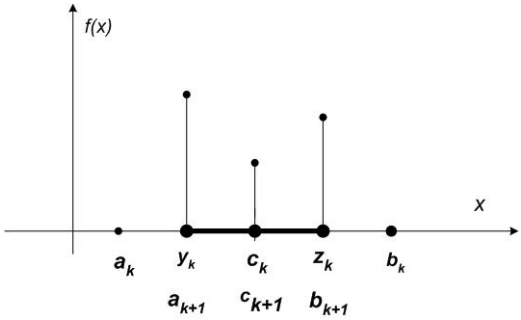

Рис. 2.4. Случай $f(c_k) \leq f(z_k)$

Возможные случаи перехода к новому отрезку в методе дихотомии приведены соответственно на рис. 2.3, 2.4 и 2.5.

Утверждение 2.1.1 гарантирует, что в любом случае после k итераций выполняется включение $x_* \in [a_{k+1}, b_{k+1}]$, а величина Δ_{k+1} оказывается равной

$$\Delta_{k+1} = b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k) = \cdots = \frac{1}{2^k}(b - a). \qquad (2.1.4)$$

В качестве очередного приближения к решению задачи — к точке x_* — берут любую точку из отрезка $[a_{k+1}, b_{k+1}]$, в частности «срединную точку» c_{k+1}, определяемую как

$$c_{k+1} = \frac{a_{k+1} + b_{k+1}}{2}.$$

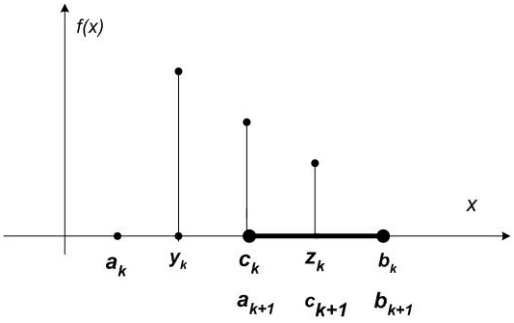

Рис. 2.5. Случай $f(c_k) > f(z_k)$

На каждой общей итерации работы алгоритма значение целевой функции $f(x)$ вычисляется самое большее в двух точках. На начальной итерации требуется вычислить только одно значение. Можно оценить ту точность, с которой будет определено решение задачи, в случае, когда допускается лишь ограниченное число вычислений значений функций $f(x)$. Это особенно актуально, если вычисление значения функции $f(x)$ весьма трудоемко, например требует решения задач Коши или краевых задач для систем дифференциальных уравнений.

Пусть позволяется сделать самое большое N вычислений значений функции $f(x)$, причем для простоты считаем, что N — нечетное число. Тогда методом дихотомии можно выполнить, по крайней мере, K итераций, где $K = \frac{N-1}{2}$ и получить решение — точку $x_{K+1} = c_{K+1}$ с погрешностью

$$|x_{K+1} - x_*| \le \tfrac{1}{2}\,(b_{K+1} - a_{K+1}) = \frac{1}{2^{\frac{N-1}{2}}}\,\frac{(b-a)}{2} \approx \approx 0.707^{N-1}\,\frac{(b-a)}{2}. \qquad (2.1.5)$$

Как видим, погрешность уменьшается не медленнее, чем со скоростью геометрической прогрессии со знаменателем $q \approx 0.707$.

Применяя формулу (2.1.5), можно оценить также количество итераций K, которое необходимо проделать методом, чтобы добиться заданной точности ε. Имеем

$$K \ge \lceil \log_2 \frac{b-a}{\varepsilon} - 1 \rceil,$$

где символом $\lceil a \rceil$ обозначено наименьшее целое число, аппроксимирующее величину a сверху.

Существуют также другие варианты метода дихотомии, например, разбивают текущий отрезок $[a_k, b_k]$ на два равных отрезка и в качестве следующего отрезка $[a_{k+1}, b_{k+1}]$ берут тот отрезок, который содержит решение задачи — точку x_*. Делается это за счет вычисления значения функции в точках, близких к общей границе обоих отрезков.

Метод золотого сечения. *Золотым сечением* называется такое деление отрезка на две неравные части, что отношение длины всего отрезка к большей части равно отношению большей части к меньшей. Эта замечательная пропорция встречается в природе, а также используется в человеческой деятельности.

Пусть имеется отрезок $[a, b]$. Предположим, что точка, которая делит отрезок в пропорции золотого сечения (обозначим ее y), находится ближе к левому концу отрезка — точке a. Тогда для нее выполняется соотношение

$$\frac{b-a}{b-y} = \frac{b-y}{y-a}, \tag{2.1.6}$$

из которого находим

$$y = y(a,b) = a + \frac{2}{3+\sqrt{5}}\,(b-a) \simeq a + 0.382\,(b-a)\,.$$

Если, напротив, данная точка находится ближе к правому концу отрезка, то для этой точки (обозначим ее z) вместо (2.1.6) имеет место соотношение

$$\frac{b-a}{z-a} = \frac{z-a}{b-z}. \tag{2.1.7}$$

Разрешая пропорцию (2.1.7), получаем

$$z = z(a,b) = a + \frac{\sqrt{5}-1}{2}\,(b-a) \simeq a + 0.618\,(b-a)\,.$$

Точка $y(a,b)$ носит название *меньшей золотой точки* отрезка $[a,b]$, точка $z(a,b)$ — *большей золотой точки* отрезка $[a,b]$ (см. рис. 2.6).

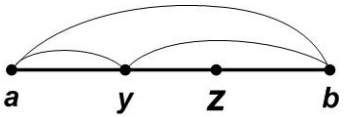

Рис. 2.6. Меньшая и большая точки золотого сечения

Если ввести величину

$$\tau = \frac{1+\sqrt{5}}{2},$$

то для меньшей и большей золотых точек оказываются справедливыми следующие выражения:

$$y(a,b) = a + \frac{1}{\tau^2}(b-a), \quad z(a,b) = a + \frac{1}{\tau}(b-a).$$

Меньшая и большая золотые точки обладают определенными свойствами, перечисленными в приведенном ниже утверждении, которое легко проверяется.

Утверждение 2.1.2. *Пусть y и z — меньшая и большая золотые точки отрезка $[a,b]$. Тогда:*

1) выполняются следующие равенства:

$$z - a = b - y = \frac{\sqrt{5}-1}{2}(b-a) = \frac{1}{\tau}(b-a);$$

2) y — большая золотая точка отрезка $[a,z]$, а z — меньшая золотая точка отрезка $[y,b]$.

Опишем теперь сам **алгоритм метода золотого сечения**.

Начальная итерация. Полагаем $a_1 = a$, $b_1 = b$, $y_1 = y(a,b)$, $z_1 = z(a,b)$ и вычисляем $f(y_1)$. Полагаем $k = 1$.

Общая k-я итерация

Шаг 1. Вычисляем то из значений $f(y_k)$ или $f(z_k)$, которое еще не вычислено. Если $f(y_k) \le f(z_k)$, то полагаем

$$a_{k+1} = a_k, \quad b_{k+1} = z_k, \quad z_{k+1} = y_k, \quad y_{k+1} = y(a_{k+1}, b_{k+1}).$$

В противном случае, т.е. когда $f(y_k) > f(z_k)$, полагаем

$$a_{k+1} = y_k, \quad b_{k+1} = b_k, \quad y_{k+1} = z_k, \quad z_{k+1} = z(a_{k+1}, b_{k+1}).$$

Шаг 2. Увеличиваем номер итерации $k := k+1$ и идем на Шаг 1.

Графически выбор нового отрезка локализации решения в первом и во втором случаях показан соответственно на рис. 2.7 и 2.8.

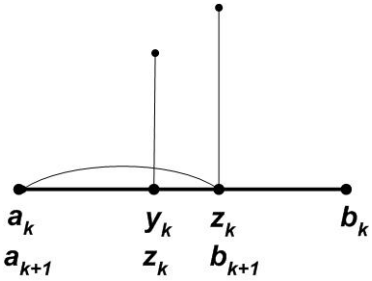

Рис. 2.7. Случай $f(y_k) \le f(z_k)$

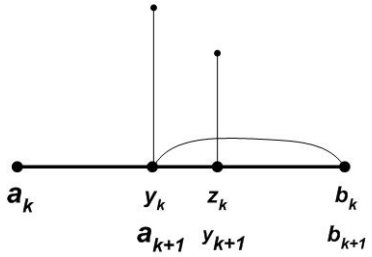

Рис. 2.8. Случай $f(y_k) > f(z_k)$

Согласно утверждению 2.1.2 на каждой k-й итерации

$$y_{k+1} = y(a_{k+1}, b_{k+1}), \quad z_{k+1} = z(a_{k+1}, b_{k+1}),$$

а согласно утверждению 2.1.1

$$x_* \in [a_{k+1}, b_{k+1}]\,.$$

Поэтому в качестве приближения x_{k+1} к решению задачи можно брать любую точку из отрезка $[a_{k+1}, b_{k+1}]$. Отметим также, что в методе золотого сечения на каждой итерации требуется вычислять значение целевой функции $f(x)$ лишь в одной точке.

Для длины отрезка локализации решения Δ_{k+1} получаем

$$\Delta_{k+1} = b_{k+1} - a_{k+1} = \frac{1}{\tau}\,(b_k - a_k) = \left(\frac{1}{\tau}\right)^k (b - a)\,. \tag{2.1.8}$$

Если позволяется вычислить значение функции $f(x)$ лишь в N точках, то можно сделать $K = N - 1$ шагов методом золотого сечения и

получить следующую оценку для близости точки x_N к x_*:

$$|x_N - x_*| \le \Delta_{K+1} = \left(\frac{1}{\tau}\right)^{N-1} (b-a) \approx 0.618^{N-1}\,(b-a)\,.$$

При N достаточно больших данная оценка лучше, чем получаемая в методе дихотомии. Для достижения точности ε, как видно из (2.1.8), следует провести количество итераций

$$K \ge \lfloor \frac{\ln \frac{b-a}{\varepsilon}}{\ln \tau} \rfloor + 1,$$

где символом $\lfloor a \rfloor$ обозначена целая часть числа a.

Метод Фибоначчи. Данный метод близок к методу золотого сечения, но в нем вместо пропорции золотого сечения используется последовательность чисел Фибоначчи, определяемая рекуррентным соотношением:

$$F_{i+1} = F_i + F_{i-1}, \quad i \ge 1. \tag{2.1.9}$$

При этом первые два числа полагаются равными: $F_0 = F_1 = 1$. Тогда для последующих чисел Фибоначчи согласно (2.1.9) получаем: $F_2 = 2$, $F_3 = 3$, $F_4 = 5$, $F_5 = 8$ и т.д.

Имеется также формула, выражающая числа Фибоначчи в явном виде посредством величины τ, используемой в золотом сечении. Данная точная формула получается, если решение рекуррентного уравнения (2.1.9) искать среди геометрических прогрессий с i-м членом, равным t^i. Тогда из соотношения

$$t^{i+1} = t^i + t^{i-1} \tag{2.1.10}$$

приходим к квадратному уравнению $t^2 - t - 1 = 0$, корнями которого являются числа

$$t_1 = \frac{1+\sqrt{5}}{2} = \tau, \qquad t_2 = \frac{1-\sqrt{5}}{2} = -\frac{1}{\tau}.$$

Последовательности $\{\tau^i\}$ и $\left\{\left(-\frac{1}{\tau}\right)^i\right\}$, как и любая их линейная комбинация с коэффициентами c_1 и c_2, удовлетворяют уравнению (2.1.10). Будем искать числа c_1 и c_2 из условия: $F_0 = F_1 = 1$. Тогда приходим к равенствам

$$c_1 + c_2 = 1, \qquad c_1\tau + c_2\left(-\frac{1}{\tau}\right) = 1.$$

Разрешая данную систему, получаем следующую формулу для чисел Фибоначчи (формулу Бине):

$$F_i = \frac{\tau^{i+1} - \left(-\frac{1}{\tau}\right)^{i+1}}{\sqrt{5}} = \frac{\left(\frac{1+\sqrt{5}}{2}\right)^{i+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{i+1}}{\sqrt{5}}.$$

Отсюда, в частности, вытекает приближенная формула:

$$F_i \approx \frac{\tau^{i+1}}{\sqrt{5}}, \tag{2.1.11}$$

справедливая при $i \to \infty$.

Опишем первые две итерации метода Фибоначчи для минимизации унимодальных функций. Для простоты будем считать, что отрезок $[a, b]$ совпадает с единичным отрезком $[0, 1]$. Это никоим образом не ограничивает общности рассуждений, так как всегда от произвольного отрезка $[a, b]$ можно перейти к единичному отрезку $[0, 1]$, сделав замену переменных: $x = a + y(b - a)$, где $y \in [0, 1]$.

Предположим, что задано целое число $N > 2$, связанное с тем количеством вычислений значений целевой функции $f(x)$, которое нам позволено сделать. На начальной итерации полагаем

$$a_1 = 0, \qquad b_1 = 1, \qquad \Delta_1 = 1$$

и определяем две точки:

$$y_1 = a_1 + \frac{F_{N-2}}{F_N} = b_1 - \frac{F_{N-1}}{F_N}, \quad z_1 = a_1 + \frac{F_{N-1}}{F_N} = b_1 - \frac{F_{N-2}}{F_N},$$

которые расположены симметрично относительно середины отрезка $[a_1, b_1]$ (единичного отрезка). Вычисляем значения целевой функции $f(y_1)$ и $f(z_1)$. Возможны два случая:

Случай 1. $f(y_1) \le f(z_1)$. Тогда берем: $a_2 = a_1$, $b_2 = z_1$, $z_2 = y_1$ и определяем точку

$$y_2 = a_2 + \frac{F_{N-3}}{F_N} = b_2 - \frac{F_{N-2}}{F_N}.$$

Случай 2. $f(y_1) > f(z_1)$. Теперь берем: $a_2 = y_1$, $b_2 = b_1$, $y_2 = z_1$ и определяем точку

$$z_2 = a_2 + \frac{F_{N-2}}{F_N} = b_2 - \frac{F_{N-3}}{F_N}.$$

В любом случае — в первом или во втором — получаем новый отрезок локализации решения $[a_2, b_2]$ с длиной

$$\Delta_2 = b_2 - a_2 = \frac{F_{N-1}}{F_N}\Delta_1 < \Delta_1.$$

Точки y_2 и z_2 снова расположены симметрично относительно середины отрезка $[a_2, b_2]$.

Далее переходим ко второй итерации, которая делается полностью аналогично первой. После выполнения второй итерации получаем отрезок локализации решения $[a_3, b_3]$ с длиной

$$\Delta_3 = \frac{F_{N-2}}{F_N}\Delta_1 < \Delta_2.$$

Последующие итерации также проводятся подобным образом. На k-м шаге мы получаем отрезок локализации с длиной

$$\Delta_{k+1} = \frac{F_{N-k}}{F_N}\Delta_1.$$

Всего мы можем выполнить подобным образом $N-2$ итерации. После этого имеем отрезок локализации решения

$$\Delta_{N-1} = \frac{F_2}{F_N}\Delta_1 = \frac{2}{F_N}$$

или, если вернуться к старым переменным,

$$\Delta_{N-1} = \frac{2}{F_N}(b-a).$$

Процесс на этом заканчивается. Меньшая и большая точки y_{N-1} и z_{N-1} при этом «схлопываются» в одну точку, совпадающую с серединой отрезка $[a_{N-2}, b_{N-2}]$. Эту точку и целесообразно брать в качестве приближенного решения задачи (2.1.1), т.е. точки x_{N-1}. Значение функции $f(x)$ в этой точке уже подсчитано.

Для расстояния от точки x_{N-1} до точного решения x_* справедлива оценка

$$|x_{N-1} - x_*| \le \frac{\Delta_{N-1}}{2} = \frac{1}{F_N}(b-a), \tag{2.1.12}$$

однако решение x_* может оказаться как слева от точки x_{N-1}, так и справа от нее. Поэтому отрезок локализации решения имеет удвоенную длину, т.е.

$$x_* \in [a_{N-2}, b_{N-2}], \quad b_{N-2} - a_{N-2} = \frac{2}{F_N}(b-a).$$

Чтобы определить, в какой из половин отрезка $[a_{N-2}, b_{N-2}]$ находится точка x_* и тем самым уменьшить длину отрезка локализации решения, следует провести дополнительное вычисление значения функции в какой-либо близлежащей к x_{N-1} точке $\bar{x}_{N-1}$ (справа или слева). Это приводит к длине отрезка локализации решения уже после N вычислений значения функции $f(x)$ либо $\frac{b-a}{F_N}$, либо $\frac{b-a}{F_N} + \delta$, где δ — расстояние между точками x_{N-1} и $\bar{x}_{N-1}$. В качестве приближенного решения x_N уже можно взять середину отрезка, содержащего решение задачи.

Неравенство (2.1.12) позволяет определить то число требуемых вычислений значений функций $f(x)$ и, стало быть, число итераций методом Фибоначчи, которое необходимо проделать, чтобы добиться заданной точности $\varepsilon > 0$. Используя (2.1.12) и приближенную формулу (2.1.11) для числа F_N, а также отбрасывая малую величину δ, приходим к следующей оценке на число N:

$$N \geq \lceil \frac{\ln \frac{\sqrt{5}(b-a)}{2\varepsilon}}{\ln \tau} \rceil - 1.$$

Оценка (2.1.12) близости точки x_N к решению задачи после $N-2$ итераций в методе Фибоначчи и вычисления дополнительного значения функции $f(x)$ в точке $\bar{x}_{N-1}$ меньше соответствующей длины в методе золотого сечения (последняя больше примерно на 17%). Но в методе золотого сечения можно в любой момент прервать вычисления и дальше их продолжить из найденных точек, если полученная точность нас по какой-то причине не устраивает. В методе Фибоначчи такое продолжение уже не допустимо. Требуется заново задать число N и заново проводить вычисления, правда, это можно делать, беря за основу уже найденный отрезок локализации решения.

2.2. Методы спуска

2.2.1. Общая схема методов спуска

В данном разделе нас будут интересовать методы решения задач безусловной минимизации

$$\min_{x \in R^n} f(x), \tag{2.2.1}$$

где относительно $f(x)$ предполагается, что она является непрерывной достаточно гладкой функцией. Мы будем рассматривать в основном методы поиска локального минимума, которые, разумеется, если

функция оказывается выпуклой, позволяют найти и глобальный минимум.

Среди методов решения задачи (2.2.1) одними из основных являются *методы спуска*. В них строится последовательность точек $\{x_k\}$ такая, что $f(x_{k+1}) < f(x_k)$. При этом начальная точка x_0 задается, а для построения последующего приближения на k-й итерации берется *направление убывания* функции $f(x)$ в точке x_k и тем или иным способом подбирается некоторый *шаг*, двигаясь с которым вдоль направления убывания, получаем новую точку x_{k+1} с меньшим значением целевой функции. Разные методы отличаются друг от друга способом выбора направления убывания и способом выбора шага. Эти процедуры дополняются правилами остановки процесса, которые зависят от используемой информации и носят главным образом эвристический характер.

Дадим теперь более строгое описание методов спуска для решения задачи безусловной минимизации (2.2.1). Напомним, что согласно своему определению ненулевой вектор $s \in \mathbb{R}^n$ называется *направлением убывания* функции $f(x)$ в точке $x \in \mathbb{R}^n$, если $f(x + \alpha s) < f(x)$ для достаточно малых $\alpha > 0$. Множество всех направлений убывания функции $f(x)$ в точке x образуют конус. Обозначим его $\mathcal{K}_d(x)$. Имеет место следующий почти очевидный результат.

Утверждение 2.2.1. *Пусть функция $f(x)$ дифференцируема в точке $x \in \mathbb{R}^n$. Тогда:*

1) *для любого $s \in \mathcal{K}_d(x)$ выполнено $\langle f_x(x), s\rangle \le 0$;*

2) *если s удовлетворяет условию $\langle f_x(x), s\rangle < 0$, то $s \in \mathcal{K}_d(x)$.*

Доказательство. Докажем сначала второе утверждение. Имеем в силу дифференцируемости функции $f(x)$:

$$f(x + \alpha s) - f(x) = \alpha\langle f_x(x), s\rangle + o(\alpha) = \alpha\left[\langle f_x(x), s\rangle + \frac{o(\alpha)}{\alpha}\right] < 0,$$

если α достаточно мало. Таким образом, $s \in \mathcal{K}_d(x)$.

Перейдем теперь к доказательству первого утверждения. От противного, пусть $s \in \mathcal{K}_d(x)$ и пусть $\langle f_x(x), s\rangle > 0$. Тогда с помощью приведенного разложения убеждаемся, что s является направлением возрастания функции $f(x)$ в точке x. Следовательно, $s \notin \mathcal{K}_d(x)$. Мы пришли к противоречию. Поэтому выполняется неравенство $\langle f_x(x), s\rangle \le 0$. ■

Ниже рассматриваются *итерационные методы спуска*, описываемые следующим рекуррентным соотношением:

$$x_{k+1} = x_k + \alpha_k s_k, \quad s_k \in \mathcal{K}_d(x_k), \quad \alpha_k > 0, \tag{2.2.2}$$

где x_0 — задаваемое начальное приближение, α_k — *шаг спуска*, выбираемый так, чтобы $f(x_{k+1}) < f(x_k)$. Если $\mathcal{K}_d(x_k) \neq \emptyset$ для всех x_k, то получаем последовательность $\{f(x_k)\}$, которая является монотонно убывающей. Если же $\mathcal{K}_d(x_k) = \emptyset$ на некоторой итерации, то процесс (2.2.2) прерывается.

2.2.2. Правила выбора длины шага

Задача выбора шага спуска α_k является задачей *одномерной минимизации на луче*. Она играет важную роль в методах спуска, и от эффективности ее решения в сильной степени зависит эффективность всего метода, особенно в случае, когда вычисление значений функции $f(x)$ трудоемко. Довольно часто данная задача решается приближенно. Рассмотрим несколько наиболее известных правил, применяемых при выборе шага α_k.

Правило одномерной минимизации. Согласно этому правилу в качестве α_k берется решение задачи

$$f(x_k + \alpha_k s_k) = \min_{\alpha \geq 0} f(x_k + \alpha s_k). \tag{2.2.3}$$

Обозначим через $\phi(\alpha)$ функцию $\phi(\alpha) = f(x_k + \alpha s_k)$, зависящую от одной переменной. Тогда задачу (2.2.3) можно переписать как

$$\min_{\alpha \geq 0} \phi(\alpha).$$

Если $\alpha_k > 0$ и функция $f(x)$ в точке $x_{k+1} = x_k + \alpha_k s_k$ дифференцируема, то выполняется условие

$$0 = \phi'(\alpha_k) = \langle f_x(x_k + \alpha_k s_k), s_k \rangle = \langle f_x(x_{k+1}), s_k \rangle. \tag{2.2.4}$$

Таким образом, если $f_x(x_{k+1}) \neq 0_n$, то геометрически решение задачи (2.2.3) означает, что x_{k+1} является точкой касания луча, задаваемым направлением s_k, с поверхностью уровня функции $f(x)$, проходящей через точку x_{k+1}.

В определенном смысле выбор шага α_k согласно правилу одномерной минимизации наилучший, так как при этом способе получаем наибольшее убывание значения целевой функции вдоль используемого направления s_k. В некоторых частных случаях, когда целевая функция

$f(x)$ имеет простой вид, задачу (2.2.3) можно решить аналитически. Например, если $f(x)$ — сильно выпуклая квадратичная функция, т.е.

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle, \tag{2.2.5}$$

где A — симметричная положительно определенная матрица, то из равенства (2.2.4) получаем, что

$$\alpha_k = -\frac{\langle Ax_k + b, s_k\rangle}{\langle As_k, s_k\rangle}. \tag{2.2.6}$$

Напомним, что s_k должно быть направлением убывания функции (2.2.5) в точке x_k и, следовательно, $\langle f_x(x_k), s_k\rangle = \langle Ax_k + b, s_k\rangle < 0$.

В общем случае точное решение задачи (2.2.3) может оказаться весьма сложной проблемой и поэтому ее решают приближенно. В частности, заменяют задачу поиска шага α_k на луче на задачу поиска этого шага на отрезке:

$$\min_{\alpha\in\Delta}\phi(\alpha), \quad \Delta = [0, \bar{\alpha}],$$

где $\bar{\alpha}$ — фиксированная положительная величина. Если $\phi(\alpha)$ оказывается унимодальной функцией на отрезке Δ, то можно использовать процедуры одномерного поиска, рассмотренные в предыдущем параграфе. Однако чаще применяют специальные правила, которые позволяют найти некоторое приемлемое решение такой задачи за небольшое число итераций.

Правило Армихо. Это достаточно простой приближенный способ определения шага α_k.

Предположим, что функция $f(x)$ дифференцируема в точке x_k. Задаются два числа: $0 < \varepsilon < 1$ и $0 < \theta < 1$ и выбирают начальное значение длины шага $\bar{\alpha}$. Полагаем $\alpha = \bar{\alpha}$. Выбор α_k проводится согласно следующей двухэтапной процедуре.

Шаг 1. Проверяем выполнение условия

$$f(x_k + \alpha s_k) - f(x_k) \le \varepsilon\alpha\langle f_x(x_k), s_k\rangle, \tag{2.2.7}$$

которое называется *неравенством Армихо*.

Шаг 2. Если неравенство (2.2.7) не выполняется, то заменяем α на $\alpha := \theta\alpha$ и идем на шаг 1. В противном случае полагаем $\alpha_k = \alpha$ и заканчиваем процесс.

Геометрическая иллюстрация выбора шага по правилу Армихо приводится на рис. 2.9. Жирная полоса показывает отрезок значений α, для которых выполняется неравенство (2.2.7).

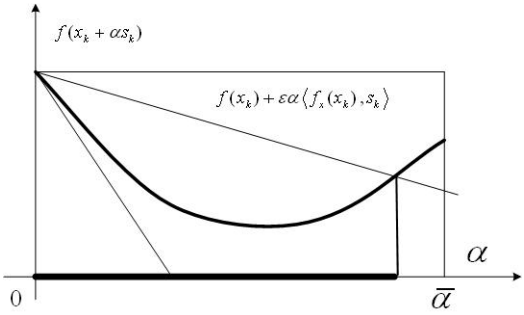

Рис. 2.9. Правило Армихо выбора шага

Покажем, что если для направления убывания s_k выполнено достаточное условие утверждения 2.2.1, т.е. неравенство $\langle f_x(x_k), s_k\rangle < 0$, то количество дроблений шага в описанной процедуре будет конечным.

Утверждение 2.2.2. *Пусть функция $f(x)$ дифференцируема в точке $x_k \in \mathbb{R}^n$ и $\langle f_x(x_k), s_k\rangle < 0$. Тогда неравенство Армихо (2.2.7) справедливо для α достаточно малых.*

Доказательство. Поскольку $\langle f_x(x_k), s_k\rangle < 0$, имеем

$$\begin{aligned} f(x_k + \alpha s_k) - f(x_k) &= \alpha\langle f_x(x_k), s_k\rangle + o(\alpha) = \\ &= \varepsilon\alpha\langle f_x(x_k), s_k\rangle + (1-\varepsilon)\alpha\langle f_x(x_k), s_k\rangle + o(\alpha) = \\ &= \varepsilon\alpha\langle f_x(x_k), s_k\rangle + \alpha\left[(1-\varepsilon)\langle f_x(x_k), s_k\rangle + \frac{o(\alpha)}{\alpha}\right] \le \varepsilon\alpha\langle f_x(x_k), s_k\rangle, \end{aligned}$$

если α достаточно мало. ■

Накладывая на функцию $f(x)$ дополнительные требования, можно добиться того, чтобы число дроблений α, после которых выполняется условие (2.2.7), было бы ограничено сверху равномерно по точкам x_k. Другими словами, существует оценка снизу $\alpha_{\min} > 0$ такая, что на всех итерациях выполняется $\alpha_k \ge \alpha_{\min}$.

Лемма 2.2.1. *Пусть функция $f(x)$ дифференцируема на $\mathbb{R}^n$, а ее производная непрерывна по Липшицу с константой L, т.е.*

$$\|f_x(x_1) - f_x(x_2)\| \le L\|x_1 - x_2\| \quad \forall x_1, x_2 \in \mathbb{R}^n. \tag{2.2.8}$$

Тогда, если $\langle f_x(x_k), s_k\rangle < 0$, то неравенство Армихо (2.2.7) выполняется для любого $\alpha \in (0, \hat{\alpha}_k]$, где

$$\hat{\alpha}_k = -\frac{2(1-\varepsilon)\langle f_x(x_k), s_k\rangle}{L\|s_k\|^2}. \tag{2.2.9}$$

Доказательство. Возьмем произвольные $x \in \mathbb{R}^n$ и $s \in \mathbb{R}^n$. Тогда по формуле Ньютона–Лейбница с учетом (2.2.8) и неравенства Коши–Буняковского получаем

$$|f(x+s) - f(x) - \langle f_x(x), s\rangle| = |\int_0^1 \langle f_x(x+\tau s) - f_x(x), s\rangle d\tau| \le$$
$$\le \int_0^1 \|f_x(x+\tau s) - f_x(x)\|\|s\| d\tau \le L\|s\|^2 \int_0^1 \tau d\tau = \tfrac{1}{2}L\|s\|^2. \tag{2.2.10}$$

Пусть теперь $x = x_k$, $s = s_k$ и $\alpha \in (0, \hat{\alpha}_k]$. Тогда на основании (2.2.10)

$$f(x_k + \alpha s_k) - f(x_k) \le \alpha\langle f_x(x_k), s_k\rangle + \tfrac{L}{2}\alpha^2\|s_k\|^2 =$$
$$= \alpha\left[\langle f_x(x_k), s_k\rangle + \tfrac{L}{2}\alpha\|s_k\|^2\right] \le$$
$$\le \alpha\left[\langle f_x(x_k), s_k\rangle - \tfrac{2L(1-\varepsilon)\langle f_x(x_k), s_k\rangle}{2L\|s_k\|^2}\|s_k\|^2\right] = \varepsilon\alpha\langle f_x(x_k), s_k\rangle.$$

Следовательно, для данных α неравенство (2.2.7) выполняется. ■

Из верхней оценки (2.2.9) видно, что если для всех $x_k \in \mathbb{R}^n$ выбирать s_k таким образом, чтобы для некоторого $\delta < 0$ выполнялось неравенство

$$\frac{\langle f_x(x_k), s_k\rangle}{\|s_k\|^2} \le \delta < 0, \tag{2.2.11}$$

то количество дроблений α будет конечным, ограниченным сверху одним и тем же числом для всех итераций.

Правило постоянного шага. Подставляя (2.2.11) в верхнюю оценку (2.2.9) для $\hat{\alpha}_k$, получаем

$$\alpha_k \le \hat{\alpha} = -\frac{2(1-\varepsilon)\delta}{L}.$$

Тогда можно заранее выбрать начальное $\bar{\alpha}_k = \alpha$ таким, чтобы $\alpha \le \hat{\alpha}$. В этом случае фактически реализуется простейшее правило выбора шага, когда α_k на всех итерациях берется одним и тем же, равным α.

Правило Голдстейна. В нем задаются два параметра: $0 < \varepsilon_1 < 1$ и $0 < \varepsilon_2 < 1$, причем $\varepsilon_1 < \varepsilon_2$. Шаг α на k-й итерации подбирается таким образом, чтобы он удовлетворял условиям

$$\varepsilon_1 \le \frac{f(x_k + \alpha s_k) - f(x_k)}{\alpha\langle f_x(x_k), s_k\rangle} \le \varepsilon_2. \tag{2.2.12}$$

Левое неравенство в (2.2.12) это по существу правило Армихо. Правое неравенство вводится для того, чтобы шаг не был достаточно малым, т.е. попадал в диапазон, указанный на рис. 2.10.

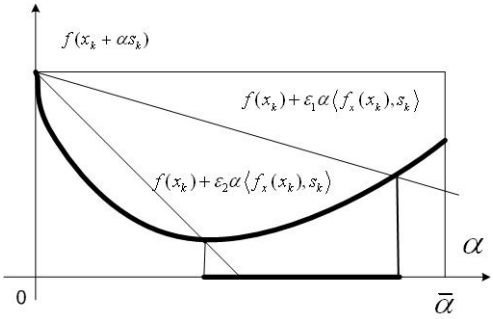

Рис. 2.10. Правило Годстейна выбора шага

Имеются также целый ряд других процедур для выбора шага. Укажем лишь одну из них.

Правило априорного выбора. Согласно этому правилу предварительно перед началом вычислительного процесса задается последовательность шагов $\{\alpha_k\}$ такая, что

$$\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{2.2.13}$$

Разумеется, что при данном выборе шага на некоторых итерациях может оказаться, что $f(x_{k+1}) \geq f(x_k)$, т.е. метод решения задачи (2.2.2) перестает быть методом спуска. Правило априорного выбора используется главным образом при минимизации негладких функций $f(x)$.

2.3. Метод градиентного спуска

Метод градиентного спуска применяется для минимизации дифференцируемых функций $f(x)$ на $\mathbb{R}^n$. В нем в качестве направления s_k на каждой итерации берется антиградиент функции $f(x)$ в текущей точке x_k, т.е. $s_k = -f_x(x_k)$. Сам метод описывается следующим рекуррентным соотношением:

$$x_{k+1} = x_k - \alpha_k f_x(x_k), \quad k = 0, 1, \ldots \tag{2.3.1}$$

Начальная точка x_0 задается. Если на некоторой k-й итерации оказывается $f_x(x_k) = 0_n$, то точка x_k является *стационарной*, в ней выпол-

няется необходимое условие локального минимума в задаче минимизации дифференцируемой функции на всем пространстве $\mathbb{R}^n$. Процесс (2.3.1) при этом прерывается.

Выбор в качестве направления спуска антиградиента является наилучшим с точки зрения использования *линейной аппроксимации* нелинейной целевой функции. В самом деле, если попытаться найти минимум линейного приближения $f(x)$ в точке x_k, т.е. минимум линейной функции

$$\phi(x) = f(x_k) + \langle f_x(x_k), x - x_k \rangle$$

на единочном шаре с центром в x_k (в евклидовой норме), то получаем, что точка x_*, доставляющая этот минимум, находится на границе шара и $x_* = x_k - \frac{f_x(x_k)}{\|f_x(x_k)\|}$. Таким образом, антиградиент $-f_x(x_k)$ является направлением *наискорейшего локального убывания* функции $f(x)$ в точке x_k.

Шаг α_k в итерационном процессе (2.3.1) может выбираться по одному из указанных выше способов. Рассмотрим вариант метода градиентного спуска (2.3.1), когда α_k выбирается по *правилу Армихо*. Тогда согласно (2.2.7) на каждой итерации выполняется неравенство

$$f\left(x_k - \alpha_k f_x(x_k)\right) - f(x_k) \leq -\varepsilon \alpha_k \|f_x(x_k)\|^2,$$

из которого следует, что $f(x_{k+1}) < f(x_k)$, если, конечно, точка x_k не стационарная. Более того, при $s_k = -f_x(x_k)$ получаем, что

$$\frac{\langle f_x(x_k), s_k \rangle}{\|s_k\|^2} = -1,$$

и эта величина не зависит от точки x_k. Поэтому, если градиент функции $f(x)$ непрерывен по Липшицу с константой L, то соответствующая константа $\hat{\alpha}_k$, которая входит в утверждение леммы 2.2.1, оказывается равной

$$\hat{\alpha}_k = \frac{2(1-\varepsilon)}{L}. \tag{2.3.2}$$

Таким образом, дробление начального шага $\bar{\alpha}_k$ производится всегда конечное число раз, причем не превышающее определенное пороговое значение. Данное пороговое значение не зависит от номера итерации. Поэтому существует оценка снизу на длину шага $\alpha_{\min} > 0$, и на всех итерациях выполняется: $\alpha_k \geq \alpha_{\min}$.

Теорема 2.3.1. *Пусть $f(x)$ — дифференцируемая и ограниченная снизу на $\mathbb{R}^n$ функция. Пусть, кроме того, ее градиент непрерывен по*

Липшицу на $\mathbb{R}^n$. *Тогда для метода градиентного спуска с правилом выбора шага по Армихо выполняется*

$$\lim_{k\to\infty} \|f_x(x_k)\| = 0. \tag{2.3.3}$$

Доказательство. Согласно (2.2.7) на каждой итерации имеет место неравенство

$$f(x_{k+1}) - f(x_k) \le -\varepsilon\alpha_k \|f_x(x_k)\|^2 \le 0, \tag{2.3.4}$$

т.е. последовательность значений функции $\{f(x_k)\}$ является невозрастающей. Так как по предположению функция $f(x)$ ограничена снизу на $\mathbb{R}^n$, то эта последовательность сходится и $f(x_{k+1}) - f(x_k) \to 0$ при $k \to \infty$. Но тогда в силу (2.3.4) и оценки снизу $\alpha_k \ge \alpha_{\min}$ получаем

$$0 \le \|f_x(x_k)\|^2 \le \frac{f(x_k) - f(x_{k+1})}{\varepsilon\alpha_k} \le \frac{f(x_k) - f(x_{k+1})}{\varepsilon\alpha_{\min}}.$$

Правая часть в этой цепочке неравенств стремится к нулю. Поэтому справедливо предельное равенство (2.3.3). ■

Теорема 2.3.1 утверждает, что любая предельная точка последовательности $\{x_k\}$ оказывается стационарной точкой функции $f(x)$, т.е. той точкой, в которой выполняется необходимое условие минимума. Если же функция $f(x)$ выпукла, то это будет решение задачи безусловной минимизации (2.2.1).

Можно получить и более сильное утверждение о методе градиентного спуска, гарантирующее уже сходимость самой последовательности $\{x_k\}$. Однако для этого следует предъявить к функции $f(x)$ более высокие по сравнению с условиями теоремы 2.3.1 требования. Предположим, что $f(x)$ — сильно выпуклая дважды дифференцируемая на $\mathbb{R}^n$ функция и ее матрица вторых производных $f_{xx}(x)$ удовлетворяет на $\mathbb{R}^n$ неравенствам

$$m\|s\|^2 \le \langle s, f_{xx}(x)s\rangle \le M\|s\|^2, \tag{2.3.5}$$

где $0 < m \le M < +\infty$. Тогда все множества подуровня этой функции выпуклы и ограничены. Поэтому для любого начального приближения x_0 последовательность $\{x_k\}$ оказывается ограниченной, следовательно, имеет предельные точки. Более того, из правого неравенства (2.3.5) вытекает, что градиент функции $f(x)$ непрерывен по Липшицу на $\mathbb{R}^n$. В этом случае в силу утверждения теоремы 2.3.1 последовательность

$\{x_k\}$ на самом деле оказывается сходящейся, причем сходится она к единственной точке минимума x_* функции $f(x)$ на $\mathbb{R}^n$. Мы пришли к следующему результату.

Теорема 2.3.2. *Пусть $f(x)$ — сильно выпуклая дважды дифференцируемая функция, для которой выполнено (2.3.5). Тогда метод градиентного спуска с правилом выбора шага по Армихо для любой начальной точки x_0 сходится к единственной точке минимума функции $f(x)$ на $\mathbb{R}^n$.*

Для выбора шага α_k может также применяться *правило постоянного шага* или *правило одномерной минимизации*. Вариант метода градиентного спуска, в котором шаг выбирается согласно правилу одномерной минимизации, называется *методом наискорейшего спуска* или *методом Коши*. Он характерен тем, что в нем на каждой итерации выполняется соотношение $\langle f_x(x_{k+1}), f_x(x_k)\rangle = 0$, т.е. используемые на соседних итерациях направления перемещения $s_k = -f_x(x_k)$ и $s_{k+1} = -f_x(x_{k+1})$ ортогональны друг другу (см. рис. 2.11).

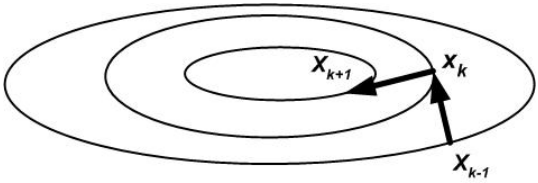

Рис. 2.11. Соседние итерации в наискорейшем спуске

При выполнении неравенств (2.3.5) в качестве константы Липшица L для градиентов можно взять константу M из правого неравенства (2.3.5). Тогда все верхние оценки $\hat{\alpha}_k$ в (2.3.2) при любом возможном выборе $\varepsilon \in (0,1)$ оказываются ограниченными сверху величиной $\frac{2}{M}$. Поэтому для постоянного шага $\alpha_k = \alpha$ такого, что $0 < \alpha < \frac{2}{M}$, получаем, что последовательность $\{f(x_k)\}$ оказывается монотонно убывающей.

Для обоих вариантов метода градиентного спуска как с правилом выбора шага из одномерной минимизации, так и по правилу постоянного шага сохраняется утверждение о сходимости теоремы 2.3.2, полученное ранее для варианта метода, в котором шаг выбирается по правилу Армихо.

Условия (2.3.5) позволяют также получить некоторые оценки на скорость сходимости метода. Предположим, что мы применяем пра-

вило постоянного шага, полагая $\alpha_k = \alpha$, где $0 < \alpha < \frac{2}{M}$. Если воспользоваться формулой Лагранжа и учесть, что $f_x(x_*) = 0_n$, то получаем

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \langle x_k - \alpha f_x(x_k) - x_*, x_{k+1} - x_* \rangle = \\ &= \langle x_k - x_* - \alpha \left(f_x(x_k) - f_x(x_*) \right), x_{k+1} - x_* \rangle = \\ &= \langle x_k - x_* - \alpha f_{xx}(\tilde{x}_k)(x_k - x_*), x_{k+1} - x_* \rangle \leq \\ &\leq \|I_n - \alpha f_{xx}(\tilde{x}_k)\| \, \|x_k - x_*\| \, \|x_{k+1} - x_*\|. \end{aligned} \tag{2.3.6}$$

Здесь $\tilde{x}_k$ — промежуточная точка между x_k и x_*, и в качестве матричной нормы используется спектральная норма, подчиненная евклидовой векторной норме.

Обозначим

$$C(\alpha) = \max \left\{ |1 - \alpha m|, \ |1 - \alpha M| \right\}.$$

Нетрудно видеть, что $C(\alpha) < 1$ при $0 < \alpha < \frac{2}{M}$, причем наименьшее свое значение константа $C(\alpha)$ достигает, когда $\alpha = \tilde{\alpha} = \frac{2}{m+M}$. Тогда

$$C(\tilde{\alpha}) = \frac{M - m}{M + m}.$$

На рис. 2.12 приводится график функции $C(\alpha)$.

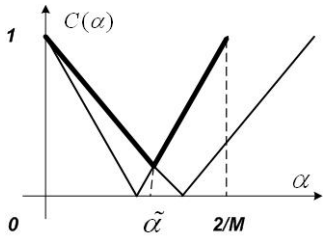

Рис. 2.12. Функция $C(\alpha)$

Из (2.3.5) следует, что $\|I_n - \alpha f_{xx}(\tilde{x}_k)\| \leq C(\alpha)$. Отсюда и из (2.3.6) приходим к оценке

$$\|x_{k+1} - x_*\| \leq C(\alpha) \|x_k - x_*\|, \tag{2.3.7}$$

которая позволяет сформулировать следующий результат.

Теорема 2.3.3. *Пусть $f(x)$ — дважды дифференцируемая сильно выпуклая на $\mathbb{R}^n$ функция, а ее матрица вторых производных $f_{xx}(x)$*

удовлетворяет условию (2.3.5). Тогда на каждой итерации в методе градиентного спуска с постоянным шагом выполняется неравенство (2.3.7).

Неравенство (2.3.7) показывает, что при сделанных предположениях метод градиентного спуска сходится с *линейной скоростью*, или, другими словами, со *скоростью геометрической прогрессии*. Однако при m, намного меньшим M, даже в наилучшем случае, когда $\alpha = \tilde{\alpha}$, соответствующая константа $C(\tilde{\alpha}) = \frac{M-m}{M+m}$ становится близкой к единице и скорость сходимости оказывается невысокой. Величины m и M фактически являются оценками соответственно снизу и сверху наименьших и наибольших собственных чисел матрицы $f_{xx}(x)$. Большой разброс собственных чисел говорит о том, что функция $f(x)$ имеет так называемый «овражный характер», т.е. ее линии уровня оказываются сильно вытянутыми. Функция $f(x)$ быстро меняется в одних направлениях и медленно меняется в других направлениях, что приводит к «зигзагообразному» движению траектории и, как следствие, к медленной скорости сходимости.

Для исправления этого недостатка были предложены специальные модификации метода градиентного спуска. Приведем одну из таких модификаций, которая предложена И.М. Гельфандом и получила название *овражного метода*. В этом методе наряду с точкой x_k берется близкая к ней точка y_k и из обеих точек делается спуск по обычному градиентному методу с малым шагом α_k по «склону оврага». В результате получаются две точки: $\bar{x}_{k+1}$ и $\bar{y}_{k+1}$. Далее делается большой шаг в направлении спуска, которое принадлежит прямой линии, соединяющей две найденные точки $\bar{x}_{k+1}$ и $\bar{y}_{k+1}$. Этот шаг приводит к перемещению вдоль «дна оврага» и к получению новой точки x_{k+1}. После этого итерация повторяется, берется новая точка y_{k+1}, близкая к x_{k+1}, и т.д. Данный овражный метод является эвристическим, но позволяет в некоторых случаях значительно ускорить сходимость.

Утверждение теоремы 2.3.3 о линейной скорости сходимости градиентного метода сохранится и для случая, когда шаг α_k выбирается либо по правилу Армихо, либо по правилу одномерной минимизации. Даже при минимизации квадратичной функции

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle \tag{2.3.8}$$

с симметричной положительно определенной матрицей A, применяя правило одномерной минимизации, приходим к следующей оценке по функционалу:

$$f(x_{k+1}) - f(x_*) \leq C\left(f(x_k) - f(x_*)\right). \tag{2.3.9}$$

Здесь $x_* = -A^{-1}b$ — точка минимума функции (2.3.8) на $\mathbb{R}^n$. Константа C определяется через наибольшее и наименьшее собственные значения $\lambda_{\max}$ и $\lambda_{\min}$ матрицы A и равняется

$$C = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 < 1. \tag{2.3.10}$$

Действительно, в этом случае шаг α_k вычисляется аналитически. Применяя формулу (2.2.6), получаем

$$\alpha_k = \frac{\|f_x(x_k)\|^2}{\langle f_x(x_k), Af_x(x_k)\rangle}, \quad f_x(x) = Ax + b.$$

Подставим данный шаг α_k в рекуррентную формулу (2.3.1) и найдем $f(x_{k+1})$. Имеем

$$\begin{array}{rcl} f(x_{k+1}) & = & f(x_k) - \alpha_k\|f_x(x_k)\|^2 + 2^{-1}\alpha_k^2\langle f_x(x_k), Af_x(x_k)\rangle = \\ & = & f(x_k) - \frac{\|f_x(x_k)\|^4}{2\langle f_x(x_k), Af_x(x_k))\rangle}. \end{array}$$

Далее, так как $f(x_*) = f(-A^{-1}b) = 2^{-1}\langle b, x_*\rangle$, то

$$\begin{aligned} f(x_k) - f(x_*) &= 2^{-1}\left[\langle x_k, Ax_k\rangle + \langle b, x_k\rangle + \langle b, x_k - x_*\rangle\right] = \\ &= 2^{-1}\left[\langle x_k, f_x(x_k)\rangle + \langle b, x_k - x_*\rangle\right] = \\ &= 2^{-1}\left[\langle x_k, f_x(x_k)\rangle + \langle b, x_k + A^{-1}b\rangle\right]. \end{aligned}$$

Отсюда, с учетом симметричности матрицы A^{-1} и очевидного равенства $A^{-1}f_x(x_k) = x_k + A^{-1}b$, приходим к

$$f(x_k) - f(x_*) = 2^{-1}\langle A^{-1}f_x(x_k), f_x(x_k)\rangle.$$

Таким образом,

$$f(x_{k+1}) - f(x_*) = C_k\left(f(x_k) - f(x_*)\right),$$

где

$$C_k = 1 - \frac{\|f_x(x_k)\|^4}{\langle f_x(x_k), Af_x(x_k)\rangle\ \langle f_x(x_k), A^{-1}f_x(x_k)\rangle}.$$

Воспользуемся теперь *неравенством Канторовича*, согласно которому

$$\langle y, Ay\rangle\ \langle y, A^{-1}y\rangle \le \frac{(\lambda_{\max} + \lambda_{\min})^2\|y\|^4}{4\lambda_{\max}\lambda_{\min}} \quad \forall y \in \mathbb{R}^n.$$

Тогда на каждой итерации $C_k \le C$, где константа C определяется согласно (2.3.10), и мы убеждаемся в справедливости оценки (2.3.9).

Обобщением градиентного метода, предназначенным для решения задач безусловной минимизации выпуклых негладких функций $f(x)$, является *субградиентный метод*. В нем вместо градиентов используются произвольные субградиенты из субдифференциала функции $f(x)$, а сам итерационный процесс может быть записан в виде

$$x_{k+1} = x_k - \alpha_k \frac{s_k}{\|s_k\|}, \quad s_k \in \partial f(x_k). \tag{2.3.11}$$

Шаг α_k берется исходя из правила априорного выбора, т.е. удовлетворяет соотношениям (2.2.13). Метод не является, вообще говоря, методом спуска. Однако если множество минимумов функции $f(x)$ на $\mathbb{R}^n$ не пусто и ограничено, то метод сходится к этому множеству.

2.4. Метод Ньютона

Метод Ньютона (другое его название *метод Ньютона–Рафсона*) хорошо известен как один из основных методов решения систем нелинейных уравнений. Его применение для решения системы $f_x(x) = 0_n$, являющейся необходимым условием для задачи минимизации дифференцируемой функции $f(x)$ на всем пространстве, приводит к варианту метода Ньютона для решения оптимизационных задач. Следует сразу сказать, что в своем классическом варианте оптимизационный метод Ньютона является локальным. Он одинаково хорошо находит как точки минимума функции $f(x)$, так и точки ее максимума. Но, с другой стороны, учет типа экстремальности задачи (ищется ли минимум функции $f(x)$ или ее максимум) позволяет проводить регулировку шага и делать его глобально сходящимся.

Метод Ньютона является *быстросходящимся*. Более высокая скорость сходимости по сравнению, например с методом градиентного спуска, достигается за счет того, что он относится к классу методов второго порядка. В нем используются вторые производные минимизируемой функции, поэтому каждая его итерация существенно более трудоемкая по сравнению с градиентным методом. Уменьшить вычислительные затраты при сохранении высокой скорости сходимости позволяют так называемые *квазиньютоновские методы*, в которых не вычисляются матрицы вторых производных, а строятся их аппроксимации с использованием лишь первых производных.

Пусть требуется решить задачу (2.2.1), в которой $f(x)$ является *выпуклой дважды непрерывно дифференцируемой функцией* на $\mathbb{R}^n$. Предполагаем также, что матрица вторых производных $f_{xx}(x)$ всюду

положительно определена, т.е. $f(x)$ по меньшей мере строго выпукла. Тогда если задача (2.2.1) имеет решение x_*, то это решение единственное. В этой точке выполняется необходимое условие

$$f_x(x_*) = 0_n. \tag{2.4.1}$$

Так как по предположению функция $f(x)$ выпуклая, то равенство (2.4.1) является одновременно и достаточным условием. Существование решения у системы нелинейных уравнений (2.4.1) гарантирует существование решения в задаче (2.2.1).

2.4.1. Метод Ньютона с постоянным шагом

Рассмотрим простейший вариант метода Ньютона для решения задачи (2.2.1). В нем строим последовательность точек $\{x_k\}$, при этом начальную точку x_0 задаем, а остальные точки определяем, исходя из следующих соображений. Пусть известно приближение x_k. Чтобы найти последующую точку x_{k+1}, разлагаем функцию $f(x)$ в ряд Тейлора в окрестности точки x_k вплоть до членов второго порядка малости:

$$\begin{aligned} f(x) = f(x_k) + \langle f_x(x_k), x - x_k\rangle + \\ + \tfrac{1}{2}\langle x - x_k, f_{xx}(x_k)(x - x_k)\rangle + \mathrm{o}(\|x - x_k\|^2). \end{aligned}$$

Возьмем квадратичную часть этой функции (квадратичное приближение $f(x)$ в точке x_k):

$$\phi(x) = f(x_k) + \langle f_x(x_k), x - x_k\rangle + \frac{1}{2}\langle x - x_k, f_{xx}(x_k)(x - x_k)\rangle$$

и найдем ее точку минимума, т.е. решим задачу

$$\min_{x \in R^n} \phi(x). \tag{2.4.2}$$

Так как $f_{xx}(x_k) > 0$, то квадратичная функция $\phi(x)$ является сильно выпуклой и, следовательно, решение задачи (2.4.2) существует и единственно. Необходимое и достаточное условие минимума для задачи (2.4.2) имеет вид

$$\phi_x(x) = f_x(x_k) + f_{xx}(x_k)(x - x_k) = 0_n.$$

Отсюда, решая относительно x эту линейную систему уравнений, получаем

$$x = \bar{x}_k = x_k - f_{xx}^{-1}(x_k) f_x(x_k).$$

Точку $\bar{x}_k$ и берем в качестве последующего приближения x_{k+1}.

Метод Ньютона формально можно записать в виде рекуррентной схемы (2.2.2), если положить $s_k = \bar{x}_k - x_k = -f_{xx}^{-1}(x_k) f_x(x_k)$ и $\alpha_k = 1$. Тогда приходим к рекуррентному соотношению: $x_{k+1} = x_k + s_k$, или

$$x_{k+1} = x_k - f_{xx}^{-1}(x_k) f_x(x_k). \tag{2.4.3}$$

Данный итерационный процесс есть не что иное, как *классический метод Ньютона* для решения системы уравнений (2.4.1). Одновременно он является и методом спуска для решения задачи безусловной минимизации (2.2.1), в которой выпуклая целевая функция обладает положительно определенными матрицами вторых производных. Действительно, так как матрица $f_{xx}(x_k)$ положительно определена, то и обратная матрица $f_{xx}^{-1}(x_k)$ также положительно определена. Поэтому

$$\langle f_x(x_k), s_k \rangle = -\langle f_x(x_k), f_{xx}^{-1}(x_k) f_x(x_k) \rangle < 0,$$

если только $f_x(x_k) \neq 0_n$. Величину $\langle f_x(x), f_{xx}^{-1}(x) f_x(x) \rangle^{\frac{1}{2}}$ называют *ньютоновским убыванием* в точке x.

Покажем, что при определенных дополнительных условиях метод Ньютона с постоянным шагом обладает *сверхлинейной скоростью* сходимости. Наложим на функцию $f(x)$ дополнительное требование, а именно: считаем, что она является сильно выпуклой функцией, причем для любых $s \in \mathbb{R}^n$ выполняются неравенства (2.3.5). Тогда для обратной матрицы $f_{xx}^{-1}(x)$ получаем

$$\frac{1}{M}\|s\|^2 \leq \langle s, f_{xx}^{-1}(x) s \rangle \leq \frac{1}{m}\|s\|^2, \tag{2.4.4}$$

где по-прежнему точка x и вектор s — произвольные из $\mathbb{R}^n$.

Пусть x_* — решение задачи (2.2.1). При сделанных предположениях относительно функции $f(x)$ данная точка x_* единственна. Имеем согласно (2.4.3)

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* - f_{xx}^{-1}(x_k) f_x(x_k) = \\ &= f_{xx}^{-1}(x_k) \left[-f_x(x_k) + f_{xx}(x_k)(x_k - x_*) \right]. \end{aligned} \tag{2.4.5}$$

Для любых x и s из $\mathbb{R}^n$ справедлива формула

$$\begin{aligned} f_x(x+s) &= f_x(x) + \textstyle\int_0^1 \frac{d}{d\tau} f_x(x + \tau s) d\tau = \\ &= f_x(x) + \textstyle\int_0^1 f_{xx}(x + \tau s) s d\tau = \\ &= f_x(x) + f_{xx}(x) s + \textstyle\int_0^1 \left[f_{xx}(x + \tau s) - f_{xx}(x) \right] s d\tau. \end{aligned} \tag{2.4.6}$$

Если теперь взять $x = x_k$, $s = x_* - x_k$, то, поскольку

$$f_x(x_k + s) = f_x(x_*) = 0_n,$$

получаем на основании (2.4.6)

$$\begin{aligned} &-f_x(x_k) + f_{xx}(x_k)(x_k - x_*) = \\ &\quad = \int_0^1 \left[f_{xx}\left(x_k + \tau(x_* - x_k)\right) - f_{xx}(x_k)\right](x_* - x_k)d\tau. \end{aligned}$$

После подстановки данного равенства в равенство (2.4.5) последнее переписывается в виде

$$\begin{aligned} &x_{k+1} - x_* = \\ &\quad = f_{xx}^{-1}(x_k) \int_0^1 \left[f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\right](x_* - x_k)d\tau. \end{aligned}$$

Отсюда и из (2.4.4), используя спектральную матричную норму, согласованную с евклидовой векторной нормой, получаем следующую оценку:

$$\begin{aligned} &\|x_{k+1} - x_*\| = \\ &\quad = \|f_{xx}^{-1}(x_k) \int_0^1 \left[f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\right](x_* - x_k)d\tau\| \le \\ &\quad \le \|f_{xx}^{-1}(x_k)\| \| \int_0^1 \left[f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\right](x_* - x_k)d\tau\| \le \\ &\quad \le \|f_{xx}^{-1}(x_k)\| \int_0^1 \| \left[f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\right](x_* - x_k)\|d\tau \le \\ &\quad \le \|f_{xx}^{-1}(x_k)\| \int_0^1 \|f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\|\|x_* - x_k\|d\tau \le \\ &\quad \le m^{-1}\|x_k - x_*\| \int_0^1 \|f_{xx}\left(x_k + \tau(x_* - x_k)\right) - f_{xx}(x_k)\|d\tau. \end{aligned} \tag{2.4.7}$$

Обозначим

$$C_k = m^{-1} \max_{0 \le \tau \le 1} \|f_{xx}\left(x_k + \tau(x_* - x_k)\right) - f_{xx}(x_k)\|.$$

Тогда согласно (2.4.7)

$$\|x_{k+1} - x_*\| \le C_k \|x_k - x_*\|. \tag{2.4.8}$$

Покажем, что если точка x_k достаточно близка к точке x_*, то константа C_k в оценке (2.4.8) меньше единицы. Более того, она стремится к нулю при $x_k \to x_*$. Действительно, все точки

$$\tilde{x}_k(\tau) = x_k + \tau(x_* - x_k) = \tau x_* + (1 - \tau)x_k$$

принадлежат отрезку, соединяющему точку x_k с x_*. Поэтому, устремляя $x_k \to x_*$, получаем, что $\|\tilde{x}_k(\tau) - x_k\| \to 0$. Следовательно,

$$\|f_{xx}(x_k + \tau(x_* - x_k)) - f_{xx}(x_k)\| = \|f_{xx}(\tilde{x}_k(\tau)) - f_{xx}(x_k)\| \to 0.$$

Это означает, что $C_k \to 0$ при $x_k \to x_*$.

Таким образом, существует окрестность $\Delta(x_*)$ точки x_* такая, что если $x_k \in \Delta(x_*)$, то $C_k < 1$ и согласно (2.4.8) выполняется включение $x_{k+1} \in \Delta(x_*)$, причем $\|x_{k+1} - x_*\| < \|x_k - x_*\|$, т.е. новая точка x_{k+1} оказывается в еще меньшей окрестности точки x_*. Другими словами, если на некоторой k-й итерации последовательность $\{x_k\}$ попадает в окрестность $\Delta(x_*)$ решения задачи x_*, то и все последующие точки этой последовательности не только остаются в этой окрестности, но и приближаются к x_*. Соответствующие константы C_k при этом стремятся к нулю. Мы приходим к выводу, что если начальное приближение x_0 взято достаточно близко к x_*, то траектория метода Ньютона (2.4.3) полностью определена и сходится к x_*, причем *скорость сходимости сверхлинейная.*

Если наложить на функцию $f(x)$ дополнительное требование, что ее вторая производная удовлетворяет на $\mathbb{R}^n$ условию Липшица с константой L, т.е.

$$\|f_{xx}(x) - f_{xx}(y)\| \le L\|x - y\| \quad \forall x, y \in \mathbb{R}^n, \tag{2.4.9}$$

то получаем, что в этом случае метод Ньютона обладает более сильной *квадратичной скоростью сходимости.* Действительно, тогда на основании (2.4.7) и (2.4.9) оценка (2.4.8) может быть уточнена:

$$\begin{aligned}\|x_{k+1} - x_*\| \le \\ \le m^{-1}\|x_k - x_*\| \int_0^1 \|f_{xx}\left(x_k + \tau(x_* - x_k)\right) - f_{xx}(x_k)\| d\tau \le \\ \le Lm^{-1}\|x_k - x_*\|^2 \int_0^1 \tau d\tau = L(2m)^{-1}\|x_k - x_*\|^2.\end{aligned}$$

Таким образом,

$$\|x_{k+1} - x_*\| \le C\|x_k - x_*\|^2,$$

где $C = \frac{L}{2m}$. Отсюда также следует, что если выделить область

$$\mathcal{D} = \left\{ x \in \mathbb{R}^n : \ \|x - x_*\| < \frac{2m}{L} \right\},$$

то в этой области метод оказывается сходящимся, так как при $x_k \in \mathcal{D}$ получаем, что $\|x_{k+1} - x_*\| < \|x_k - x_*\|$.

Суммируя все вышесказанное, приходим к следующему утверждению относительно классического метода Ньютона с постоянным единичным шагом для решения задачи (2.2.1).

Теорема 2.4.1. *Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая на $\mathbb{R}^n$ функция, для второй производной которой выполняются неравенства (2.3.5). Тогда метод Ньютона с постоянным шагом локально сходится к решению этой задачи — точке x_* — со сверхлинейной скоростью. Если, кроме того, для вторых*

производных выполнено условие Липшица (2.4.9), то данный метод локально сходится к x_* с квадратичной скоростью.

2.4.2. Метод Ньютона с переменным шагом

Одним из недостатков метода Ньютона с постоянным шагом является его локальная сходимость. В самом деле, применяемый даже для минимизации выпуклых функций, он не всегда может найти решение задачи. Приведем пример, поясняющий эту расходимость метода. Рассмотрим дважды непрерывно дифференцируемую функцию

$$f(x) = x \operatorname{arctg} x - \frac{1}{2}\ln(1+x^2), \quad x \in \mathbb{R}.$$

Так как $f'(x) = \operatorname{arctg} x$ и $f''(x) = (1+x^2)^{-1} > 0$, то данная функция является строго выпуклой и достигает своего минимума на $\mathbb{R}$ в нуле. Если воспользоваться методом Ньютона с постоянным шагом (равным единице) для минимизации $f(x)$, то он сходится лишь в том случае, когда начальное приближение взято достаточно близко к решению, а именно $|x_0| < \bar{x}$, где $\bar{x} \simeq 1.392$.

Чтобы расширить область сходимости метода Ньютона, применяют его вариант с переменным шагом (часто называемый *демпфированным методом Ньютона*). В этом методе итерационный процесс вместо (2.4.3) описывается следующим рекуррентным соотношением:

$$x_{k+1} = x_k - \alpha_k f_{xx}^{-1}(x_k) f_x(x_k), \tag{2.4.10}$$

т.е. вводится переменный шаг α_k, который регулируется тем или иным способом. Предположим, что α_k выбирается по правилу Армихо (2.2.7), причем начальный шаг $\bar{\alpha}$ полагается равным единице, а параметр ε удовлетворяет условию: $0 < \varepsilon < \frac{1}{2}$. Тогда на каждой k-й итерации выполняется неравенство

$$f(x_{k+1}) - f(x_k) \le -\varepsilon\alpha_k \langle f_x(x_k), f_{xx}^{-1}(x_k) f_x(x_k)\rangle. \tag{2.4.11}$$

Представим итерационный процесс (2.4.10) как метод спуска (2.2.2), в котором $s_k = -f_{xx}^{-1}(x_k) f_x(x_k)$. Используя формулу Тейлора, получаем

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_k + \alpha_k s_k) - f(x_k) = \\ &= \alpha_k \langle f_x(x_k), s_k\rangle + \tfrac{1}{2}\alpha_k^2 \langle s_k, f_{xx}(\tilde{x}_k) s_k\rangle = \\ &= \alpha_k \langle f_x(x_k), s_k\rangle + \tfrac{1}{2}\alpha_k^2 \langle s_k, f_{xx}(x_k) s_k\rangle + \\ &\qquad + \tfrac{1}{2}\alpha_k^2 \langle s_k, (f_{xx}(\tilde{x}_k) - f_{xx}(x_k))\, s_k\rangle, \end{aligned}$$

где $\tilde{x}_k \in [x_k, x_{k+1}]$.

Учтем теперь, что $f_x(x_k) = -f_{xx}(x_k)s_k$, и обозначим для сокращения записи

$$d_k = \langle f_x(x_k), s_k\rangle = -\langle s_k, f_{xx}(x_k)s_k\rangle = \\ = -\langle f_x(x_k), f_{xx}^{-1}(x_k)f_x(x_k)\rangle. \tag{2.4.12}$$

Имеем $d_k < 0$, когда $f_x(x_k) \neq 0_n$. Если функция $f(x)$ такова, что ее матрица вторых производных $f_{xx}(x)$ удовлетворяет условию (2.3.5), то для d_k выполняется неравенство

$$d_k = -\langle s_k, f_{xx}(x_k)s_k\rangle \leq -m\|s_k\|^2. \tag{2.4.13}$$

Тогда на основании этого неравенства и неравенства Коши–Буняковского получаем

$$\begin{aligned} &f(x_{k+1}) - f(x_k) = \\ &\quad = \alpha_k d_k \left[1 - \tfrac{\alpha_k}{2} + \tfrac{\alpha_k}{2d_k}\langle s_k, (f_{xx}(\tilde{x}_k) - f_{xx}(x_k))\, s_k\rangle\right] \leq \\ &\quad \leq \alpha_k d_k \left[1 - \tfrac{\alpha_k}{2} - \tfrac{\alpha_k}{2m\|s_k\|^2}\|f_{xx}(\tilde{x}_k) - f_{xx}(x_k)\|\|s_k\|^2\right] = \\ &\quad = \alpha_k d_k \left[1 - \tfrac{\alpha_k}{2} - \tfrac{\alpha_k}{2m}\|f_{xx}(\tilde{x}_k) - f_{xx}(x_k)\|\right]. \end{aligned}$$

Отсюда и из (2.4.12) видно, что неравенство (2.4.11) будет выполняться, если

$$1 - \frac{\alpha_k}{2} - \frac{\alpha_k}{2m}\|f_{xx}(\tilde{x}_k) - f_{xx}(x_k)\| \geq \varepsilon.$$

В частности, оно будет выполняться даже при $\alpha_k = 1$, когда

$$\frac{1}{2} - \frac{1}{2m}\|f_{xx}(\tilde{x}_k) - f_{xx}(x_k)\| \geq \varepsilon$$

или

$$\frac{1}{m}\|f_{xx}(\tilde{x}_k) - f_{xx}(x_k)\| \leq 1 - 2\varepsilon. \tag{2.4.14}$$

Поскольку по предположению $\varepsilon \in (0, \frac{1}{2})$, константа в правой части (2.4.14) положительна.

Неравенство (2.4.14) оказывается справедливым для точек x_k, достаточно близких к решению задачи (2.2.1) — точке x_*. Это вытекает из рассуждений, которые были проведены при обосновании локальной сходимости метода Ньютона с постоянным шагом $\alpha_k = 1$. Действительно, тогда можно указать такую достаточно малую окрестность $\Delta(x_*)$ точки x_*, что при $x_k \in \Delta(x_*)$ все последующие точки также остаются в этой окрестности. Но в этом случае согласно (2.4.8) обязательно выполняются неравенства: $\|x_{k+1} - x_*\| \leq C_k\|x_k - x_*\|$, где

$C_k < 1$, $C_k \to 0$. Поэтому $\tilde{x}_k \to x_k$ и, значит, левая часть в (2.4.14) стремится к нулю.

Убедимся теперь, что если шаг α_k выбирается по правилу Армихо, то какую бы начальную точку $x_0 \in \mathbb{R}^n$ мы ни взяли итерационный процесс (2.4.10) обязательно попадет в указанную окрестность $\Delta(x_*)$. Действительно, при выполнении условия (2.3.5) функция $f(x)$ является сильно выпуклой, а ее градиент непрерывен по Липшицу. Для d_k согласно (2.3.5) справедлива оценка (2.4.13), влекущая выполнение неравенства (2.2.11) при $\delta = -m$. Но тогда шаг α_k по правилу Армихо определяется за конечное количество дроблений начального шага, причем это количество не превосходит одно и то же число, которое не зависит от точек x_k. Все шаги α_k оказываются ограниченными снизу некоторым шагом $\alpha_{\min} > 0$. Последовательность значений функции $\{f(x_k)\}$ является *монотонно убывающей*. Последовательность точек $\{x_k\}$ оказывается принадлежащей компактному множеству

$$X = \{x \in \mathbb{R}^n : \ f(x) \le f(x_0)\}.$$

Так как функция $f(x)$ ограничена снизу на X и $\alpha_k \ge \alpha_{\min}$, то согласно (2.4.11)

$$\lim_{k\to\infty} \langle f_x(x_k), f_{xx}^{-1}(x_k) f_x(x_k)\rangle = 0.$$

В силу левого неравенства (2.4.4) это возможно тогда и только тогда, когда $\|f_x(x_k)\| \to 0$ при $k \to \infty$. Для сильно выпуклой непрерывно дифференцируемой функции $f(x)$ выполнение данного предельного равенства означает, что $x_k \to x_*$, где x_* — единственное решение задачи (2.2.1). Следовательно, точка x_k на некоторой итерации обязательно попадет в нужную окрестность $\Delta(x_*)$. После этого метод ведет себя как классический метод Ньютона с постоянным шагом $\alpha_k = 1$ и для него справедливы все полученные ранее оценки, касающиеся скорости сходимости.

Таким образом, нами получен следующий результат, который сформулируем в виде теоремы.

Теорема 2.4.2. *Пусть $f(x)$ — сильно выпуклая дважды непрерывно дифференцируемая функция, удовлетворяющая условию (2.3.5). Тогда метод Ньютона (2.4.10) с регулировкой шага по правилу Армихо сходится из любой начальной точки к единственной точке минимума функции $f(x)$ на $\mathbb{R}^n$ со сверхлинейной скоростью. Если, кроме того, выполнено условие Липшица (2.4.9) для вторых производных, то скорость сходимости квадратичная.*

Утверждение теоремы полностью сохранится, если в методе (2.4.10) выбирать шаг не по правилу Армихо, а по правилу одномерной минимизации, т.е. на каждой итерации решать вспомогательную задачу минимизации:

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x_k + \alpha s_k).$$

В общем случае, когда функция $f(x)$ не является сильно выпуклой, регулировка шага позволяет значительно расширить область сходимости метода Ньютона. Особенно это важно при отыскании локальных решений задачи (2.2.1).

Как уже не раз отмечалось, метод Ньютона обладает высокой скоростью сходимости. Минимум квадратичной выпуклой функции

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

где A — симметричная положительно определенная матрица, находится этим методом из любой начальной точки $x_0 \in \mathbb{R}^n$ по определению за одну итерацию. К недостаткам метода Ньютона следует отнести необходимость знать матрицу вторых производных $f_{xx}(x_k)$ на каждой итерации, хотя для нахождения направления s_k нет необходимости ее обращать. Обычно s_k определяют путем решения системы линейных алгебраических уравнений $f_{xx}(x_k)s_k = -f_x(x_k)$, в которой матрица системы симметричная.

Общепринятые эвристические рекомендации по использованию метода Ньютона и метода градиентного спуска для решения задач безусловной минимизации заключаются в том, чтобы применять градиентный спуск лишь на первом этапе вычислений для попадания в достаточно близкую окрестность решения, а затем продолжать расчеты методом Ньютона, если, разумеется, вычисление ньютоновских направлений не очень трудоемко.

2.4.3. Квазиньютоновские методы

Рассмотрим теперь класс методов, которые по своим свойствам близки к методу Ньютона, но в то же время не требуют вычисления матрицы вторых производных $f_{xx}(x)$. Основная идея, заложенная в их построение, заключается в использовании матриц, которые аппроксимируют матрицы $f_{xx}(x_k)$, постепенно приближаясь к матрице вторых производных в решении задачи.

С этой целью обратимся к итерационному процессу достаточно общего вида, предназначенного для решения задачи безусловной мини-

мизации (2.2.1) с непрерывно дифференцируемой функцией $f(x)$:

$$x_{k+1} = x_k + \alpha_k s_k, \quad s_k = -H_k f_x(x_k). \tag{2.4.15}$$

В нем на каждой итерации в качестве H_k будем брать симметричную положительно определенную матрицу.

Задаваемый рекуррентным соотношением (2.4.15) итерационный процесс относится к классу методов спуска, так как выполняется неравенство $\langle f_x(x_k), H_k f_x(x_k)\rangle < 0$, если $f_x(x_k) \neq 0_n$, т.е. если точка x_k не является стационарной. В случае, когда в качестве H_k на каждой итерации берется единичная матрица I_n, процесс (2.4.15) совпадает с методом градиентного спуска. Предполагая, что выпуклая функция $f(x)$ дважды дифференцируема и ее матрица вторых производных $f_{xx}(x)$ положительно определена на $\mathbb{R}^n$, можно прийти и к методу Ньютона с переменным шагом. В (2.4.15) в этом случае достаточно брать в качестве H_k матрицы $f_{xx}^{-1}(x_k)$.

Однако особый интерес представляют методы типа (2.4.15), в которых матрицы H_k, будучи отличными от $f_{xx}^{-1}(x_k)$, стремились к $f_{xx}^{-1}(x_*)$ в пределе, где x_* — решение задачи (2.2.1).

Пусть H_0 — произвольная симметричная положительно определенная матрица, и пусть на k-й итерации переход от матрицы H_k к матрице H_{k+1} осуществляется по формуле

$$H_{k+1} = H_k + \Delta H_k, \tag{2.4.16}$$

где ΔH_k — некоторая симметричная матрица. Если воспользоваться разложением

$$f_x(x_k) - f_x(x_{k+1}) = f_{xx}(x_{k+1})\,(x_k - x_{k+1}) + o(\|x_k - x_{k+1}\|),$$

то из его линейной части следует приближенное равенство

$$\Delta x_k = x_{k+1} - x_k \approx f_{xx}^{-1}(x_{k+1})\,(f_x(x_{k+1}) - f_x(x_k))\,.$$

Заменим его на обычное равенство, но при этом учтем наше желание, чтобы матрица H_{k+1} приближала $f_{xx}^{-1}(x_{k+1})$. Тогда приходим к условию

$$\Delta x_k = H_{k+1}\Delta y_k, \quad \Delta y_k = f_x(x_{k+1}) - f_x(x_k). \tag{2.4.17}$$

Равенство (2.4.17) носит название *квазиньютоновского условия*. Если пересчет матриц H_k проводится по рекуррентной формуле (2.4.16), то из него получаем:

$$\Delta H_k \Delta y_k = \Delta x_k - H_k \Delta y_k. \tag{2.4.18}$$

При этом дополнительно следует потребовать, чтобы матрица ΔH_k была симметричной и такой, чтобы преобразованная матрица H_{k+1} оставалась положительно определенной.

Система (2.4.18) является недоопределенной и существуют много способов удовлетворить равенству (2.4.18). Методы вида (2.4.15), в которых выполняется (2.4.16) и (2.4.18), получили название *квазиньютоновских*. Приведем некоторые наиболее популярные из них.

1. Рассмотрим сначала простейший вариант выбора матрицы ΔH_k, в котором матрица ΔH_k имеет ранг, равный единице. Тогда ΔH_k представима в виде: $\Delta H_k = \mu_k q_k q_k^T$, где q_k — некоторый ненулевой вектор, μ_k — константа. Обозначим

$$\Delta z_k = \Delta x_k - H_k \Delta y_k. \tag{2.4.19}$$

Равенство (2.4.18) при сделанном выборе ΔH_k сведется к следующему: $\mu_k q_k q_k^T \Delta y_k = \Delta z_k$. Отсюда следует, что оно будет выполняться, если $q_k^T \Delta y_k \neq 0$ и если $q_k = \Delta z_k$. Тогда $\mu_k = (q_k^T \Delta y_k)^{-1}$. Таким образом, получаем следующее выражение для матрицы-приращения:

$$\Delta H_k = \frac{(\Delta x_k - H_k \Delta y_k)\,(\Delta x_k - H_k \Delta y_k)^T}{\langle \Delta x_k - H_k \Delta y_k, \Delta y_k \rangle}, \tag{2.4.20}$$

справедливое, когда $\langle \Delta x_k - H_k \Delta y_k, \Delta y_k \rangle \neq 0$. Метод (2.4.15), (2.4.16), в котором матрица ΔH_k вычисляется по формуле (2.4.20), называется *методом Бройдена*.

2. Другой способ определения матрицы-приращения ΔH_k, но уже ранга два, также следует из (2.4.18). Так как в правой части присутствуют векторы Δx_k и $H_k \Delta y_k$, то будем строить ΔH_k в виде

$$\Delta H_k = \mu_1\ \Delta x_k\,(\Delta x_k)^T + \mu_2\ H_k \Delta y_k\,(H_k \Delta y_k)^T. \tag{2.4.21}$$

Если положить

$$\mu_1 = \frac{1}{\langle \Delta x_k, \Delta y_k \rangle}, \quad \mu_2 = -\frac{1}{\langle H_k \Delta y_k, \Delta y_k \rangle},$$

то получаем, что такая матрица (2.4.21) также удовлетворяет равенству (2.4.18). Метод (2.4.15), (2.4.16) с матрицей

$$\Delta H_k = \frac{\Delta x_k\,(\Delta x_k)^T}{\langle \Delta x_k, \Delta y_k \rangle} - \frac{H_k \Delta y_k\,(H_k \Delta y_k)^T}{\langle H_k \Delta y_k, \Delta y_k \rangle} \tag{2.4.22}$$

называется *методом Дэвидона–Флетчера–Пауэла*.

3. Приведем еще один пример матрицы-приращения ΔH_k, которая имеет ранг два и такой, что матрица H_{k+1} оказывается положительно определенной. Будем искать матрицу ΔH_k в виде

$$\Delta H_k = QUQ^T, \quad Q = [q_1, q_2], \quad q_1, q_2 \in \mathbb{R}^n, \quad U = \begin{bmatrix} a & c \\ c & b \end{bmatrix}, \quad (2.4.23)$$

причем в качестве вектора q_1 возьмем правую часть в квазиньютоновском равенстве (2.4.18) — вектор Δz_k. Условие $\langle \Delta z_k, \Delta y_k \rangle \neq 0$ при этом может и не выполняться. Кроме того, необязательно, чтобы ненулевые собственные значения симметричной матрицы ΔH_k (их всего два) имели одинаковые знаки.

Подставим разложенную согласно (2.4.23) матрицу ΔH_k в квазиньютоновское условие (2.4.18). Из-за того, что $q_1 = \Delta z_k$, следует равенство $UQ^T \Delta y_k = [1, 0]^T$ или, в подробной записи,

$$\begin{aligned} a\,\langle \Delta z_k, \Delta y_k \rangle + c\,\langle q_2, \Delta y_k \rangle &= 1, \\ c\,\langle \Delta z_k, \Delta y_k \rangle + b\,\langle q_2, \Delta y_k \rangle &= 0. \end{aligned}$$

Это система двух уравнений относительно трех переменных.

Пусть вектор q_2 выбран таким образом, что $\langle q_2, \Delta y_k \rangle \neq 0$. В этом случае можно взять $a = 0$. Тогда остальные элементы матрицы U принимают значения

$$c = \frac{1}{\langle q_2, \Delta y_k \rangle}, \quad b = -\frac{\langle \Delta z_k, \Delta y_k \rangle}{\langle q_2, \Delta y_k \rangle^2},$$

а для матрицы ΔH_k получаем

$$\Delta H_k = \frac{1}{\langle q_2, \Delta y_k \rangle} \left[\Delta z_k q_2^T + q_2 (\Delta z_k)^T - \frac{\langle \Delta z_k, \Delta y_k \rangle}{\langle q_2, \Delta y_k \rangle} q_2 q_2^T \right]. \quad (2.4.24)$$

Если положить $q_2 = \Delta x_k$, то после подстановки этого вектора в (2.4.24) приходим к симметричной матрице-приращению:

$$\Delta H_k = \frac{1}{\langle \Delta x_k, \Delta y_k \rangle} \left[\beta \Delta x_k (\Delta x_k)^T - H_k \Delta y_k (\Delta x_k)^T - \Delta x_k (H_k \Delta y_k)^T \right], \quad (2.4.25)$$

где

$$\beta = 1 + \frac{\langle H_k \Delta y_k, \Delta y_k \rangle}{\langle \Delta x_k, \Delta y_k \rangle}.$$

Если ввести вектор

$$p_k = \langle \Delta y_k, H_k \Delta y_k \rangle^{\frac{1}{2}} \left[\frac{\Delta x_k}{\langle \Delta x_k, \Delta y_k \rangle} - \frac{H_k \Delta y_k}{\langle \Delta y_k, H_k \Delta y_k \rangle} \right],$$

то матрицу (2.4.25) можно записать также (сравни с (2.4.22)) в виде

$$\Delta H_k = \frac{\Delta x_k \left(\Delta x_k\right)^T}{\langle \Delta x_k, \Delta y_k \rangle} - \frac{H_k \Delta y_k \left(H_k \Delta y_k\right)^T}{\langle H_k \Delta y_k, \Delta y_k \rangle} + p_k p_k^T.$$

Квазиньютоновский метод с матрицей-приращением (2.4.25) носит название *метода Бройдена–Флетчера–Гольдфарба–Шанно*.

Обратимся к матрице $\Delta H_k = QUQ^T$. Согласно известному результату из матричного анализа собственные значения матрицы AB совпадают с собственными значениями матрицы BA за исключением дополнительных нулевых значений. Поэтому матрица ΔH_k имеет те же самые ненулевые собственные значения, что и матрица $F = UQ^TQ$. Но матрица F квадратная порядка два, ее детерминант равен произведению собственных значений F. Поскольку $\det F = \det U \det Q^TQ$ и $\det U = ab - c^2 = -c^2 < 0$, то из положительной определенности матрицы Q^TQ и следующего отсюда неравенства $\det Q^TQ > 0$ получаем, что при $a = 0$ у матрицы ΔH_k ненулевые собственные значения имеют разные знаки.

Выбор значения $a = 0$ оправдан с точки зрения положительной определенности матрицы H_{k+1}. В самом деле, в общем случае, беря поправку ΔH_k ранга два вида (2.4.23), мы должны обеспечить положительную определенность матрицы $H_{k+1} = H_k + QUQ^T$, где H_k — положительно определенная матрица. Из положительной определенности матрицы H_k следует, что справедливо разложение $H_k = G^TG$ для некоторой неособой матрицы G. Поэтому

$$H_{k+1} = G^T \left(I_n + Q_1 U Q_1^T\right) G, \quad Q_1 = (G^T)^{-1} Q.$$

Матрица Q_1 размером $n \times 2$, как и матрица Q, также имеет ранг, равный двум.

Матрица H_{k+1} будет положительно определенной в том и только в том случае, когда положительно определенной будет матрица $I_n + Q_1 U Q_1^T$. Но у этой матрицы собственные значения лишь на единицу отличаются от собственных значений матрицы $Q_1 U Q_1$. Поэтому среди собственных значений $I_n + Q_1 U Q_1^T$ обязательно $n - 2$ собственные значения равны единицы. Остается рассмотреть только случай с ненулевыми собственными значениями матрицы $Q_1 U Q_1^T$, они совпадают с ненулевыми собственными значениями (2×2)-матрицы $UQ_1^TQ_1 = UQ^TH_k^{-1}Q$.

Возьмем (2×2)-матрицу $I_2 + UQ^TH_k^{-1}Q$. С учетом (2.4.23) она представима в виде

$$I_2 + UQ^TH_k^{-1}Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} q_1^T \\ q_2^T \end{bmatrix} H_k^{-1} \left[q_1 \; q_2\right].$$

Выпишем след и определитель этой матрицы и потребуем, чтобы они были положительными. Обозначая через $\langle h_1, h_2\rangle_{H_k^{-1}} = \langle h_1, H_k^{-1} h_2\rangle$ скалярное произведение между векторами h_1 и h_2 из $\mathbb{R}^n$, определяемое с помощью положительно определенной матрицы H_k^{-1}, получаем

$$2 + t_1 > 0, \quad 1 + t_1 + t_2 > 0, \tag{2.4.26}$$

где

$$t_1 = a\langle q_1, q_1\rangle_{H_k^{-1}} + 2c\langle q_1, q_2\rangle_{H_k^{-1}} + b\langle q_2, q_2\rangle_{H_k^{-1}},$$

$$t_2 = \left(ab - c^2\right)\left(\langle q_1, q_1\rangle_{H_k^{-1}}\langle q_2, q_2\rangle_{H_k^{-1}} - \langle q_1, q_2\rangle^2_{H_k^{-1}}\right).$$

Так как согласно неравенству Коши–Буняковского для линейно независимых векторов q_1 и q_2

$$\langle q_1, q_1\rangle_{H_k^{-1}}\langle q_2, q_2\rangle_{H_k^{-1}} - \langle q_1, q_2\rangle^2_{H_k^{-1}} > 0,$$

то при $ab = 0$ выполнение второго неравенства (2.4.26) влечет выполнение и первого неравенства (2.4.26). В самом деле, тогда $t_2 < 0$ и $2 + t_1 > 1 + t_1 > 1 + t_1 + t_2 > 0$. Поэтому далее будем рассматривать только второе из неравенств (2.4.26). Если $a = 0$, то оно сводится к следующему:

$$\begin{aligned} 1 + b\langle q_2, q_2\rangle_{H_k^{-1}} + 2c\langle q_1, q_2\rangle_{H_k^{-1}} - \\ - c^2\left(\langle q_1, q_1\rangle_{H_k^{-1}}\langle q_2, q_2\rangle_{H_k^{-1}} - \langle q_1, q_2\rangle^2_{H_k^{-1}}\right) > 0. \end{aligned} \tag{2.4.27}$$

В нашем случае

$$q_1 = \Delta z_k, \quad q_2 = \Delta x_k, \quad c = \langle \Delta x_k, \Delta y_k\rangle^{-1}, \quad b = -c^2\langle \Delta y_k, \Delta z_k\rangle.$$

Подставим эти величины в (2.4.27) и перегруппируем слагаемые. Тогда становится видно, что для выполнения неравенства (2.4.27) достаточно потребовать, чтобы $d_1 + d_2 > 0$, где введены обозначения

$$\begin{aligned} d_1 = \langle \Delta x_k, \Delta y_k\rangle^2 - \langle \Delta y_k, \Delta z_k\rangle\langle \Delta x_k, H_k^{-1}\Delta x_k\rangle + \\ + 2\langle \Delta x_k, \Delta y_k\rangle\langle \Delta x_k, H_k^{-1}\Delta z_k\rangle, \end{aligned}$$

$$d_2 = \langle \Delta x_k, H_k^{-1}\Delta z_k\rangle^2 - \langle \Delta x_k, H_k^{-1}\Delta x_k\rangle\langle \Delta z_k, H_k^{-1}\Delta z_k\rangle.$$

Учтем теперь, что согласно (2.4.19)

$$\langle \Delta x_k, H_k^{-1}\Delta z_k\rangle = \langle \Delta x_k, H_k^{-1}\Delta x_k\rangle - \langle \Delta x_k, \Delta y_k\rangle,$$

$$\langle \Delta y_k, \Delta z_k\rangle = \langle \Delta x_k, \Delta y_k\rangle - \langle \Delta y_k, H_k\Delta y_k\rangle.$$

Тогда

$$\begin{aligned} d_1 = \langle \Delta x_k, \Delta y_k \rangle^2 - \langle \Delta x_k, \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle + \\ + \langle \Delta y_k, H_k \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle - \\ - 2\langle \Delta x_k, \Delta y_k \rangle^2 + 2\langle \Delta x_k, \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle = \\ = -\langle \Delta x_k, \Delta y_k \rangle^2 + \langle \Delta x_k, \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle + \\ + \langle \Delta y_k, H_k \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle. \end{aligned}$$

Распишем далее в более подробном виде слагаемые в правой части величины d_2. Имеем

$$\begin{aligned} \langle \Delta x_k, H_k^{-1} \Delta z_k \rangle^2 = \big(\langle \Delta x_k, H_k^{-1} \Delta x_k \rangle - \langle \Delta x_k, \Delta y_k \rangle \big)^2 = \\ = \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle^2 + \langle \Delta x_k, \Delta y_k \rangle^2 - 2\langle \Delta x_k, \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle, \end{aligned}$$

$$\begin{aligned} \langle \Delta z_k, H_k^{-1} \Delta z_k \rangle = \langle \Delta x_k - H_k \Delta y_k, H_k^{-1} \Delta x_k - \Delta y_k \rangle = \\ = \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle - 2\langle \Delta x_k, \Delta y_k \rangle + \langle \Delta y_k, H_k \Delta y_k \rangle. \end{aligned}$$

После подстановки соответствующих выражений получаем

$$\begin{aligned} d_2 = \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle^2 + \langle \Delta x_k, \Delta y_k \rangle^2 - \\ - 2\langle \Delta x_k, \Delta y_k \rangle \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle - \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle^2 + \\ + 2\langle \Delta x_k, H_k^{-1} \Delta x_k \rangle \langle \Delta x_k, \Delta y_k \rangle - \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle \langle \Delta y_k, H_k \Delta y_k \rangle = \\ = \langle \Delta x_k, \Delta y_k \rangle^2 - \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle \langle \Delta y_k, H_k \Delta y_k \rangle. \end{aligned}$$

Поэтому

$$d_1 + d_2 = \langle \Delta x_k, H_k^{-1} \Delta x_k \rangle \langle \Delta x_k, \Delta y_k \rangle.$$

Мы приходим к выводу, что матрица H_{k+1} будет положительно определенной, если $\langle \Delta x_k, \Delta y_k \rangle > 0$. Данное неравенство всегда выполняется в случае строго монотонного градиентного отображения $f_x(x)$, т.е. когда

$$\langle f_x(x_1) - f_x(x_2), x_1 - x_2 \rangle > 0 \quad \forall x_1, x_2 \in \mathbb{R}^n, \quad x_1 \neq x_2.$$

У строго выпуклой непрерывно дифференцируемой функции $f(x)$ градиентное отображение строго монотонно.

Квазиньютоновские методы хотя и являются методами первого порядка, но обладают достаточно хорошей скоростью сходимости — сверхлинейной или даже квадратичной. Можно показать, что если шаг α_k выбирается из условия одномерной минимизации на всей действительной прямой, то любой из рассмотренных вариантов квазиньютоновских методов найдет минимум выпуклой квадратичной функции за конечное число шагов, не превышающее размерность задачи. Квазиньютоновские методы весьма популярны и широко применяются на практике.

Квазиньютоновские методы можно интерпретировать как различные методы секущих для решения системы уравнений $f_x(x) = 0_n$, в которых учитывается то обстоятельство, что матрица Якоби этого градиентного отображения (матрица Гессе $f_{xx}(x)$ функции $f(x)$) является симметричной.

Обратим внимание на еще одну достаточно интересную трактовку квазиньютоновских методов, связывающую их с так называемыми *методами переменной метрики*. Дело в том, что если имеется симметричная положительно определенная матрица A, то с ее помощью может быть задано скалярное произведение $\langle x, y\rangle_A = \langle x, Ay\rangle$ и соответствующая метрика $\|x - y\|^2 = \langle x - y, (x - y)\rangle_A$. В этом случае линейную часть приращения функции $f(x)$, а именно $\langle f_x(x), \Delta x\rangle$, можно представить в виде

$$\langle f_x(x), \Delta x\rangle = \langle AA^{-1} f_x(x), \Delta x\rangle = \langle A^{-1} f_x(x), \Delta x\rangle_A.$$

Таким образом, вектор $A^{-1} f_x(x)$ оказывается градиентом функции $f(x)$ в пространстве со скалярным произведением $\langle \cdot, \cdot\rangle_A$. С этой точки зрения итерационный процесс (2.4.15) есть не что иное, как градиентный метод, но в пространстве с меняющимся скалярным произведением, задаваемым матрицами H_k^{-1}. К выбору матриц H_k, являющихся приближениями матрицы, обратной к гессиану функции $f(x)$ в решении, можно подойти как к установлению некой метрики, более подходящей, чем евклидова, особенно, если гессиан близок к вырожденному. Задача состоит в том, чтобы заменить обратный гессиан на некую симметричную матрицу, которая всегда положительно определена, имеет, по возможности, меньшую норму и на «доступных векторах» определяет оператор, действующий так же, как и «точная» матрица.

2.5. Метод сопряженных градиентов

Метод сопряженных градиентов является одним из вариантов метода сопряженных направлений. Эти методы в своей первооснове тесно связаны с методами решения систем уравнений, а именно систем линейных уравнений.

2.5.1. Метод сопряженных направлений для квадратичных функций

Возьмем квадратичную функцию

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle, \tag{2.5.1}$$

где A — симметричная положительно определенная матрица. Данная функция является сильно выпуклой на $\mathbb{R}^n$, поэтому всегда существует ее точка минимума x_* на $\mathbb{R}^n$, причем единственная, и в этой точке x_* выполняется условие оптимальности: $f_x(x_*) = Ax_* + b = 0_n$. Отсюда получаем, что $x_* = -A^{-1}b$.

Пусть имеется произвольная точка $x_0 \in \mathbb{R}^n$ и пусть, кроме того, в $\mathbb{R}^n$ выделен набор из n линейно независимых векторов $s_1, \ldots, s_n$. Тогда точку x_* можно представить в виде

$$x_* = x_0 + \sum_{i=1}^{n} a^i s_i, \tag{2.5.2}$$

где a^i, $1 \le i \le n$, — некоторые коэффициенты.

Обозначим через a вектор из $\mathbb{R}^n$ с компонентами $a^1, \ldots, a^n$, через S — квадратную матрицу порядка n, столбцами которой являются векторы $s_1, \ldots, s_n$. Тогда разложение (2.5.2) перепишется в виде

$$x_* = x_0 + Sa. \tag{2.5.3}$$

Если (2.5.3) подставить в условие $f_x(x_*) = 0_n$, то приходим к равенству $A(x_0 + Sa) = -b$ или

$$ASa = -Ax_0 - b = -f_x(x_0).$$

Умножим обе части этого равенства слева на матрицу S^T. Тогда оно примет вид: $S^TASa = -S^Tf_x(x_0)$. Так как матрицы S и A неособые, то и симметричная матрица S^TAS также является неособой, поэтому существует ее обратная матрица и $a = -\left(S^TAS\right)^{-1} S^Tf_x(x_0)$. Подставляя данное a в (2.5.2), получаем

$$x_* = x_0 - S\left(S^TAS\right)^{-1} S^Tf_x(x_0).$$

Здесь система линейно независимых векторов $s_1, \ldots, s_n$ бралась произвольной без учета вида матрицы A. Если принять его во внимание, то можно получить более простое выражение для точки минимума x_*.

В дальнейшем нам понадобится следующее важное понятие.

Определение 2.5.1. *Векторы $s_1, \ldots, s_k$, где $k \leq n$, называются сопряженными относительно матрицы A или просто A-сопряженными, если они ненулевые и*

$$\langle s_i, As_j \rangle = 0, \quad 1 \leq i, j \leq k, \quad i \neq j.$$

A-сопряженные векторы называют также A-ортогональными, так как они действительно ортогональны между собой в случае, когда скалярное произведение в $\mathbb{R}^n$ задается с помощью симметричной положительно определенной матрицы A.

Утверждение 2.5.1. *Пусть векторы $s_1, \ldots, s_k$ сопряжены относительно положительно определенной матрицы A. Тогда они линейно независимы.*

Доказательство. Предположим противное, что векторы $s_1, \ldots, s_k$ линейно зависимы. Тогда найдется такой вектор s_i , что $s_i = \sum_{j \neq i} \alpha_j s_j$, где не все α_j равны нулю. В этом случае

$$\langle s_i, As_i \rangle = \langle s_i, \sum_{j \neq i} \alpha_j As_j \rangle = \sum_{j \neq i} \alpha_j \langle s_i, As_j \rangle = 0.$$

Отсюда, поскольку матрица A положительно определена, получаем, что $s_i = 0_n$. Мы пришли к противоречию. ■

Пусть теперь взятый нами набор $s_1, \ldots, s_n$ является набором из n сопряженных относительно матрицы A векторов. Такую совокупность векторов принято называть *A-сопряженной или A-ортогональной системой.* Для нее соответствующая матрица $S^T AS$ имеет диагональный вид:

$$S^T AS = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix},$$

где $d_i = \langle s_i, As_i \rangle > 0$, $1 \leq i \leq n$. Обратная матрица $(S^T AS)^{-1}$ также будет диагональной с элементами d_i^{-1} на диагонали, и мы получаем, что коэффициенты a^i равняются следующим величинам:

$$a^i = -\frac{\langle f_x(x_0), s_i \rangle}{\langle s_i, As_i \rangle}, \quad 1 \leq i \leq n.$$

Поясним *смысл коэффициентов* a_i в случае A-сопряженной системы векторов $s_1, \ldots, s_n$. Введем в рассмотрение точки

$$x_i = x_{i-1} + a^i s_i = x_0 + \sum_{j=1}^{i} a^j s_j, \quad 1 \le i \le n,$$

определяемые рекуррентным образом. Если обратиться к задаче

$$\min_{\alpha \in R} f(x_{i-1} + \alpha s_i), \tag{2.5.4}$$

то в силу необходимых условий минимума должно выполняться равенство

$$\frac{d}{d\alpha} f(x_{i-1} + \alpha s_i) = \langle f_x(x_{i-1} + \alpha s_i), s_i \rangle = 0.$$

Отсюда, так как $f_x(x) = Ax + b$, приходим к уравнению относительно коэффициента α:

$$\langle Ax_{i-1} + \alpha A s_i + b, s_i \rangle = 0$$

или

$$\langle s_i, As_i \rangle \alpha + \langle f_x(x_{i-1}), s_i \rangle = 0.$$

Таким образом, для решения α_* задачи (2.5.4) получаем

$$\alpha_* = -\frac{\langle f_x(x_{i-1}), s_i \rangle}{\langle s_i, As_i \rangle}.$$

Но, в силу A-сопряженности системы векторов $s_1, \ldots, s_n$,

$$\begin{aligned} \langle f_x(x_{i-1}), s_i, \rangle &= \langle A\left(x_0 + \textstyle\sum_{j=1}^{i-1} \alpha_j s_j\right) + b, s_i \rangle = \\ &= \langle Ax_0 + b, s_i \rangle = \langle f_x(x_0), s_i \rangle. \end{aligned}$$

Следовательно, $\alpha_* = a^i$.

Пусть l_i есть прямая в $\mathbb{R}^n$, задаваемая направлением s_i, и пусть $\tilde{l}_i$ есть её сдвиг на вектор x_{i-1}. Мы получили важный результат, а именно: решением задачи минимизации функции $f(x)$ на прямой $\tilde{l}_i$, проходящей через x_{i-1}, является точка $x_{i-1} + a_i s_i$, т.е. x_i.

Приведенные выше рассуждения позволяют проинтерпретировать процесс построения точек $x_1, \ldots, x_n$ с несколько иной точки зрения. Предположим, что задана начальная точка x_0. Берем произвольные направления $s_1, \ldots, s_n$, которые образуют A-сопряженную систему. Последовательно вычисляем точки $x_1, \ldots, x_n$, полагая

$$x_k = x_{k-1} + \alpha_k s_k, \quad f(x_{k-1} + \alpha_k s_k) = \min_{\alpha \in R} f(x_{k-1} + \alpha s_k), \tag{2.5.5}$$

где $k = 1, 2, \ldots, n$. Так как коэффициент α_k получается из условия минимизации функции $f(x_{k-1} + \alpha s_k)$ по α (он может быть как положительным, так и отрицательным), то на каждом из n шагов процесса (2.5.5) выполняется равенство

$$\langle f_x(x_k), s_k \rangle = 0, \quad k = 1, 2, \ldots, n. \tag{2.5.6}$$

Из вышесказанного следует, что процесс (2.5.5) для любого $x_0 \in \mathbb{R}^n$ позволяет найти минимум функции $f(x)$, причем за количество шагов, не превышающее n. Если на некотором k-м шаге оказывается: $f_x(x_k) = 0_n$, то расчеты заканчиваются, так как найдено решение задачи минимизации сильно выпуклой квадратичной функции (2.5.1).

Обратим внимание на еще одно свойство процесса (2.5.5). Обозначим через L_k k-мерное линейное подпространство, порожденное векторами $s_1, \ldots, s_k$, а через X_k — сдвиг этого подпространства на вектор x_0, т.е. $X_k = x_0 + L_k$.

Упражнение 1. *Покажите, что для любого $1 \leq k \leq n$ точка x_k доставляет минимум функции $f(x)$ на линейном многообразии X_k.*

Описанная процедура является общей схемой так называемых *методов сопряженных направлений*. Существует целое семейство таких методов, отличающихся друг от друга конкретным способом выбора направлений $s_1, \ldots, s_n$.

2.5.2. Метод сопряженных градиентов для квадратичных функций

Рассмотрим теперь *метод сопряженных градиентов*, который является одним из наиболее эффективных среди методов сопряженных направлений. Данный вариант метода сопряженных градиентов носит название *метода Хестенса–Штифеля*. Он был предложен в 1952 г. и изначально предназначался для решения систем линейных алгебраических уравнений $Ax = b$ с симметричной матрицей A.

В методе Хестенса–Штифеля вычисления проводятся согласно общей схеме (2.5.5), т.е. начальная точка x_0 задается, а последующие точки определяются по формуле

$$x_k = x_{k-1} + \alpha_k s_k, \tag{2.5.7}$$

где шаг α_k находится из задачи одномерной минимизации (2.5.4), причем в явном виде:

$$\alpha_k = \arg\min_{\alpha \in R^n} f(x_{k-1} + \alpha s_k) = -\frac{\langle f_x(x_{k-1}), s_k \rangle}{\langle s_k, A s_k \rangle}. \tag{2.5.8}$$

В качестве направления s_1 берется $s_1 = -f_x(x_0)$. Другие направления s_2, ..., s_n насчитываются постепенно от итерации к итерации, полагаясь равными

$$s_k = -f_x(x_{k-1}) + \beta_{k-1} s_{k-1}, \quad k = 2, 3, \ldots, n. \tag{2.5.9}$$

Число β_{k-1} выбирается из условия A-сопряженности двух «соседних» направлений s_k и s_{k-1}. Если расписать это условие $\langle s_k, As_{k-1}\rangle = 0$, подставив s_k из (2.5.9), то получаем

$$\langle -f_x(x_{k-1}) + \beta_{k-1} s_{k-1}, As_{k-1}\rangle = 0.$$

Отсюда находим

$$\beta_k = \frac{\langle f_x(x_k), As_k\rangle}{\langle s_k, As_k\rangle}, \quad k = 1, 2, \ldots, n-1. \tag{2.5.10}$$

Сформулируем теперь алгоритм метода сопряженных градиентов для минимизации выпуклых квадратичных функций.

Алгоритм метода Хестенса–Штифеля. Берем произвольную точку $x_0 \in \mathbb{R}^n$, вычисляем $f_x(x_0)$ и полагаем $k = 1$.

Шаг 1. Если $f_x(x_{k-1}) = 0_n$, то останавливаемся.

Шаг 2. Если $k > 1$, то вычисляем β_{k-1} по формуле (2.5.10).

Шаг 3. Полагаем $s_k = -f_x(x_{k-1})$, если $k = 1$. Иначе вычисляем s_k по формуле (2.5.9).

Шаг 4. Определяем α_k согласно формуле (2.5.8) и, используя формулу пересчета (2.5.7), находим x_k.

Шаг 5. Если $k < n$, то увеличиваем k, полагая $k := k + 1$, и идем на шаг 1.

Укажем простейшие свойства приведенного алгоритма.

Свойство 1. При вычислении коэффициента α_k по формуле (2.5.8) требуется, чтобы вектор s_k был ненулевым. Покажем, что случай, когда $s_k = 0_n$ на некотором k-м шаге, невозможен. Действительно, пусть $s_1 \neq 0_n$, ..., $s_{k-1} \neq 0_n$, а $s_k = 0_n$. Тогда обязательно $\langle f_x(x_{k-1}), s_k\rangle = 0$. Кроме того, согласно (2.5.9) и равенству (2.5.6), следующему из необходимого условия минимума при определении коэффициента α_{k-1},

$$\begin{aligned} 0 = \langle f_x(x_{k-1}), s_k\rangle = \\ = -\|f_x(x_{k-1})\|^2 + \beta_{k-1}\langle f_x(x_{k-1}), s_{k-1}\rangle = -\|f_x(x_{k-1})\|^2. \end{aligned}$$

Отсюда заключаем, что $f_x(x_{k-1}) = 0_n$, т.е. еще на предыдущей итерации найдена точка x_{k-1}, в которой выполняется необходимое и достаточное условие минимума выпуклой квадратичной функции (2.5.1) на

$\mathbb{R}^n$. Поэтому, как следует из описания алгоритма, он должен был бы остановиться ранее, не переходя на k-й шаг.

Свойство 2. По построению два «соседних» направления s_k и s_{k-1} сопряжены относительно матрицы A. Покажем, что для двух «соседних» градиентов $f_x(x_k)$ и $f_x(x_{k-1})$ выполняется равенство

$$\langle f_x(x_k), f_x(x_{k-1})\rangle = 0, \tag{2.5.11}$$

т.е. они ортогональны друг другу.

Данное равенство (2.5.11) легко проверяется, когда $k = 1$. Действительно, по построению два первых градиента $f_x(x_0)$ и $f_x(x_1)$ ортогональны друг другу, так как $s_1 = -f_x(x_0)$ и на основании (2.5.6) $\langle f_x(x_1), s_1\rangle = 0$. Поэтому $\langle f_x(x_1), f_x(x_0)\rangle = 0$.

Покажем, что и для $k > 1$ равенство $\langle f_x(x_k), f_x(x_{k-1})\rangle = 0$ сохраняется. Имеем: $f_x(x_k) = Ax_k + b$ и $x_k = x_{k-1} + \alpha_k s_k$. Следовательно, в силу (2.5.8)

$$f_x(x_k) = f_x(x_{k-1}) + \alpha_k A s_k = f_x(x_{k-1}) - \frac{\langle f_x(x_{k-1}), s_k\rangle}{\langle s_k, As_k\rangle} As_k.$$

Подставляя данное выражение для $f_x(x_k)$, получаем

$$\langle f_x(x_k), f_x(x_{k-1})\rangle = \|f_x(x_{k-1})\|^2 - \frac{\langle f_x(x_{k-1}), s_k\rangle}{\langle s_k, As_k\rangle}\langle f_x(x_{k-1}), As_k\rangle. \tag{2.5.12}$$

Но на основании (2.5.6) и (2.5.9)

$$f_x(x_{k-1}), s_{k-1}\rangle = 0, \quad s_k = -f_x(x_{k-1}) + \beta_{k-1}s_{k-1}. \tag{2.5.13}$$

Поэтому

$$\langle f_x(x_{k-1}), s_k\rangle = \langle f_x(x_{k-1}), -f_x(x_{k-1}) + \beta_{k-1}s_{k-1}\rangle = -\|f_x(x_{k-1})\|^2.$$

Таким образом, равенство (2.5.12) можно переписать как

$$\langle f_x(x_k), f_x(x_{k-1})\rangle = \|f_x(x_{k-1})\|^2 \left[1 + \frac{\langle f_x(x_{k-1}), As_k\rangle}{\langle s_k, As_k\rangle}\right]. \tag{2.5.14}$$

В силу A-сопряженности направлений s_{k-1} и s_k имеем с учетом второго равенства (2.5.13):

$$\langle s_k, As_k\rangle = \langle -f_x(x_{k-1}) + \beta_{k-1}s_{k-1}, As_k\rangle = -\langle f_x(x_{k-1}), As_k\rangle.$$

Следовательно,

$$\frac{\langle f_x(x_{k-1}), As_k\rangle}{\langle s_k, As_k\rangle} = -1.$$

Отсюда на основании (2.5.14) делаем вывод, что $\langle f_x(x_k), f_x(x_{k-1})\rangle = 0$, т.е. градиенты $f_x(x_k)$ и $f_x(x_{k-1})$ ортогональны друг другу для всех $k = 1, \ldots, n$.

Оказывается, приведенные свойства, касающиеся A-сопряженности двух «соседних» направлений s_k и s_{k-1} и ортогональности двух «соседних» градиентов $f_x(x_k)$ и $f_x(x_{k-1})$, распространяются на все направления s_1, ..., s_k и на все градиенты $f_x(x_0)$, ..., $f_x(x_{k-1})$.

Лемма 2.5.1. *Пусть A — положительно определенная симметричная матрица. Тогда какое бы ни было начальное приближение x_0, если для произвольного $1 \le k \le n$ алгоритм метода Хестенса–Штифеля сгенерировал направления $s_1, \ldots, s_k$, то они образуют систему векторов, сопряженных относительно матрицы A, а градиенты $f_x(x_0), \ldots, f_x(x_{k-1})$ — ортогональную систему в $\mathbb{R}^n$.*

Доказательство проведем индукцией по k. При $k = 2$ утверждение теоремы очевидно, так как s_1 и s_2 A-сопряжены по построению, а ортогональность градиентов $f_x(x_1)$ и $f_x(x_0)$ вытекает из условия выбора коэффициента α_1.

Предположим далее, что утверждение теоремы доказано для всех $k = 2, 3, \ldots, l-1$ и докажем его для $k = l$. Фактически надо показать, что

$$\langle As_l, s_i\rangle = 0, \quad \langle f_x(x_{l-1}), f_x(x_{i-1})\rangle = 0. \tag{2.5.15}$$

при $1 \le i < l$. Так как в силу *свойства* 2 оба равенства (2.5.15) выполняются для двух «соседних» направлений и двух «соседних» градиентов, то на самом деле их надо проверить при $i < l - 1$.

Воспользуемся следующими из формул пересчета (2.5.7) и (2.5.9) равенствами: $x_{l-1} = x_{l-2} + \alpha_{l-1}s_{l-1}$, $s_i = -f_x(x_{i-1}) + \beta_{i-1}s_{i-1}$. Тогда по предположению индукции

$$\begin{aligned}
&\langle f_x(x_{l-1}), f_x(x_{i-1})\rangle = \\
&\quad = \langle f_x(x_{l-2}), f_x(x_{i-1})\rangle + \alpha_{l-1}\langle As_{l-1}, f_x(x_{i-1})\rangle = \\
&\quad = \alpha_{l-1}\langle As_{l-1}, f_x(x_{i-1})\rangle = \alpha_{l-1}\langle As_{l-1}, \beta_{i-1}s_{i-1} - s_i\rangle = 0.
\end{aligned}$$

Итак, выполняется второе равенство (2.5.15).

По индукции имеем также

$$\langle As_l, s_i\rangle = -\langle Af_x(x_{l-1}), s_i\rangle + \beta_{l-1}\langle As_{l-1}, s_i\rangle = -\langle f_x(x_{l-1}), As_i\rangle. \tag{2.5.16}$$

Но поскольку $x_i = x_{i-1} + \alpha_i s_i$ и $f_x(x) = Ax + b$, то

$$f_x(x_i) - f_x(x_{i-1}) = Ax_{i-1} + \alpha_i As_i - Ax_{i-1} = \alpha_i As_i.$$

Отсюда

$$As_i = \frac{1}{\alpha_i} \left[f_x(x_i) - f_x(x_{i-1}) \right]. \tag{2.5.17}$$

Учтем равенство (2.5.17). После его подстановки в (2.5.16) приходим к

$$\langle As_l, s_i \rangle = \frac{1}{\alpha_i} \left[\langle f_x(x_{l-1}), f_x(x_{i-1}) \rangle - \langle f_x(x_{l-1}), f_x(x_i) \rangle \right] = 0.$$

Таким образом, первое равенство (2.5.15) также имеет место. ■

Как следует из всего вышесказанного, метод сопряженных градиентов находит точку минимума сильно выпуклой квадратичной функции $f(x)$ не более чем за n шагов. Можно показать, что на самом деле имеет место более сильный результат, в котором требования к функции $f(x)$ несколько ослаблены.

Теорема 2.5.1. *Если выпуклая квадратичная функция (2.5.1) достигает своего минимального значения на $\mathbb{R}^n$, то метод сопряженных градиентов находит ее точку минимума не более чем за n шагов.*

2.5.3. Метод сопряженных градиентов для произвольных функций

Рассмотрим теперь общую задачу безусловной минимизации

$$\min_{x \in R^n} f(x),$$

где $f(x)$ — произвольная непрерывно дифференцируемая функция. Нам хотелось бы построить обобщение метода сопряженных градиентов для решения этой задачи. Однако непосредственное перенесение формул, справедливых для квадратичных функций, на случай произвольных функций $f(x)$ встречает определенные трудности, которые, как оказывается, несложно обойти.

Дело в том, что для квадратичной функции $f(x)$ как сопряженные направления, так и длина шага α_k вычислялись по формулам, в которых присутствует матрица A, являющаяся матрицей Гессе для

функции (2.5.1). Так для коэффициента β_k и α_k были получены выражения

$$\beta_k = \frac{\langle f_x(x_k), As_k\rangle}{\langle s_k, As_k\rangle}, \quad \alpha_k = -\frac{\langle f_x(x_{k-1}), s_k\rangle}{\langle s_k, As_k\rangle}. \tag{2.5.18}$$

Поскольку нам не хотелось бы вводить в метод вычисление матрицы вторых производных $f_{xx}(x)$, то постараемся преобразовать эти выражения таким образом, чтобы исключить из них матрицу A. На самом деле, это следует проделать только для коэффициента β_k, а формулу для определения шага α_k из алгоритма вообще можно исключить, заменив ее непосредственным определением шага из задачи минимизации целевой функции $f(x)$ на прямой $\tilde{l}_k$, задаваемой направлением s_k. Равенства (2.5.6) при этом сохранятся.

Рассмотрим теперь один из основных подходов к замене выражения для коэффициента β_k, в котором уже не присутствует матрица A. Имеем согласно (2.5.7):

$$\begin{aligned} Ax_k &= A(x_{k-1} + \alpha_k s_k) = \\ &= Ax_{k-1} + b + \alpha_k As_k - b = f_x(x_{k-1}) + \alpha_k As_k - b, \end{aligned}$$

откуда для квадратичной функции $f(x)$ получаем

$$f_x(x_k) = Ax_k + b = f_x(x_{k-1}) + \alpha_k As_k.$$

Таким образом,

$$As_k = \alpha_k^{-1}\left[f_x(x_k) - f_x(x_{k-1})\right].$$

Подставляя данное выражение для As_k в формулу для β_k в (2.5.18), приходим с учетом утверждения леммы 2.5.1 и (2.5.6) к следующему выражению:

$$\beta_k = \frac{\langle f_x(x_k), f_x(x_k) - f_x(x_{k-1})\rangle}{\langle s_k, f_x(x_k) - f_x(x_{k-1})\rangle} = -\frac{\langle f_x(x_k), f_x(x_k)\rangle}{\langle s_k, f_x(x_{k-1})\rangle}. \tag{2.5.19}$$

Воспользуемся теперь тем, что на основании (2.5.9)

$$s_k = -f_x(x_{k-1}) + \beta_{k-1}s_{k-1}. \tag{2.5.20}$$

Тогда выражение (2.5.19) для β_k преобразуется к виду

$$\beta_k = \frac{\langle f_x(x_k), f_x(x_k)\rangle}{\langle f_x(x_{k-1}), f_x(x_{k-1})\rangle} = \frac{\|f_x(x_k)\|^2}{\|f_x(x_{k-1})\|^2}. \tag{2.5.21}$$

Данная формула для коэффициента β_k носит название *формулы Флетчера–Ривса*, на ее основании может быть построен численный метод минимизации произвольной дифференцируемой функции $f(x)$.

Метод Флетчера–Ривса. В этом методе, как и в случае квадратичной функции, точки x_k насчитываются согласно рекуррентной схеме (2.5.7), т.е. $x_k = x_{k-1} + \alpha_k s_k$. Начальная точка x_0 задается. В качестве первого направления s_1, как и в методе Хестенса–Штифеля, берется антиградиент: $s_1 = -f_x(x_0)$. Последующие направления s_k определяются по формуле (2.5.20), в которой коэффициент β_{k-1} вычисляется согласно (2.5.21). Шаг α_k находится из решения задачи одномерной минимизации:

$$f(x_{k-1} + \alpha_k s_k) = \min_{\alpha \in R} f(x_{k-1} + \alpha s_k). \tag{2.5.22}$$

Покажем, что s_k есть направление убывания целевой функции $f(x)$ в точке x_{k-1}. Действительно, на основании (2.5.20) и равенства (2.5.6), которое остается справедливым и для задачи выбора шага при минимизации произвольной не обязательно квадратичной функции,

$$\begin{aligned}\langle f_x(x_{k-1}), s_k\rangle &= -\langle f_x(x_{k-1}), f_x(x_{k-1}) - \beta_{k-1} s_{k-1}\rangle = \\ &= -\|f_x(x_{k-1})\|^2 + \beta_{k-1}\langle f_x(x_{k-1}), s_{k-1}\rangle = -\|f_x(x_{k-1})\|^2 < 0,\end{aligned}$$

если $f_x(x_{k-1}) \neq 0_n$. Таким образом, от задачи минимизации (2.5.22) на всей прямой можно перейти к задаче минимизации на луче:

$$f(x_{k-1} + \alpha_k s_k) = \min_{\alpha \geq 0} f(x_{k-1} + \alpha s_k). \tag{2.5.23}$$

В этом случае метод является методом спуска.

В методе Флетчера–Ривса направления s_1, s_2, ..., s_n уже не образуют сопряженную систему относительно какой-либо матрицы. Обычно в него вводится так называемая *процедура «обновления»*. Согласно этой процедуре после n шагов (напомним, что n — размерность вектора x) очередной коэффициент β_k не вычисляется по формуле (2.5.21), а полагается равным нулю, т.е. в качестве направления s_{k+1} берется теперь $s_{k+1} = -f_x(x_k)$, алгоритм начинает работать из текущей точки x_k как из начальной точки. При этом уменьшается влияние погрешностей решения одномерных задач минимизации (2.5.23) на вычислительный процесс.

Приведем утверждение о сходимости метода, предполагая, что $f(x)$ является дважды непрерывно дифференцируемой функцией и ее матрица вторых производных $f_{xx}(x)$ удовлетворяет условию (2.3.5). Так

как в этом случае функция $f(x)$ сильно выпукла, то задача безусловной минимизации этой функции всегда имеет решение x_*, причем единственное. Кроме того, решения одномерных задач минимизации (2.5.23) также всегда существуют.

Теорема 2.5.2. *Пусть для дважды непрерывно дифференцируемой функции $f(x)$ выполняется условие (2.3.5). Тогда метод сопряженных градиентов сходится к точке минимума x_* функции $f(x)$ со сверхлинейной скоростью в том смысле, что*

$$\|x_{(k+1)n} - x_*\| \le C_k \|x_{kn} - x_*\|,$$

где $C_k \to 0$ при $k \to \infty$.

Более того, если матрица вторых производных $f_{xx}(x)$ удовлетворяет на $\mathbb{R}^n$ условию Липшица (2.4.9), то

$$\|x_{(k+1)n} - x_*\| \le C \|x_{kn} - x_*\|^2$$

для k достаточно больших.

Согласно утверждению теоремы 2.5.2 рассмотренный метод сопряженных градиентов при минимизации сильно выпуклых функций обладает сверхлинейной и даже квадратичной скоростью сходимости, но в определенном смысле, когда n последовательных итераций объединяются как-бы в одну итерацию. Таким образом, метод, не требуя вычисления матриц вторых производных, оказывается достаточно эффективным. Метод сравнительно простой и является одним из основных методов безусловной минимизации, широко применяемым на практике. Итерационный процесс (2.5.7) в методе относится к двухшаговым схемам, а именно, из-за того, что $s_{k-1} = \alpha_{k-1}^{-1}(x_{k-1} - x_{k-2})$, на всех итерациях, кроме первой и кроме тех, на которых происходит обновление, его можно представить в виде

$$x_k = x_{k-1} - \alpha_k \left[f_x(x_{k-1}) - \frac{\beta_{k-1}}{\alpha_{k-1}} (x_{k-1} - x_{k-2}) \right],$$

т.е. выбор точки x_k зависит не только от предыдущей точки x_{k-1}, но и от точки x_{k-2}. Двухшаговые схемы, как правило, гораздо лучше учитывают поведение целевой функции, что приводит к ускорению сходимости. Отметим также, что имеются другие варианты метода сопряженных направлений, отличные от приведенного алгоритма Флетчера–Ривса. По сравнению с квазиньютоновскими методами в методе сопряженных градиентов нет необходимости хранить и пересчитывать матрицы, аппроксимирующие обратные матрицы Гессе минимизируемой функции.

Глава 3

Методы линейной и квадратичной оптимизации

3.1. Симплекс-метод для задач линейного программирования

В данном разделе нас будут интересовать численные методы решения задач линейного программирования, точнее, один из наиболее известных и популярных численных методов решения таких задач — симплекс-метод. Данный метод существенным образом использует специфику задач линейного программирования.

Подчеркнем также, что поскольку задачи линейного программирования являются частным случаем задач условной оптимизации, то методы решения последних, например методы решения задач выпуклого программирования, также могут применяться для решения задач линейного программирования. Некоторые из них будут рассмотрены в дальнейших разделах нашего курса.

Как уже говорилось [33], существуют различные формы постановок задач линейного программирования (например, каноническая, основная, стандартная и.д.), причем всегда можно от одной постановки перейти к другой. При этом используются следующие несложные приемы:

1. Равенство $Ax = b$ можно превратить в неравенство, записав его как: $Ax \leq b$, $-Ax \leq -b$.

2. Неравенство $Ax \leq b$ переходит в равенство, если ввести дополнительные неотрицательные переменные: $Ax + z = b$, где $z \geq 0_m$.

3. Любая свободная переменная x^j, т.е. переменная, на которую не наложено никакое ограничение, может быть представлена в виде разности двух неотрицательных переменных, а именно: $x^j = u^j - v^j$, где $u^j \geq 0$ и $v^j \geq 0$.

Задачи линейного программирования являются одними из наиболее простых, но в то же время и наиболее важных задач условной оптимизации. Известные в настоящее время численные методы позволяют находить на современных компьютерах решения задач линейного программирования весьма больших размеров, когда число переменных достигает несколько миллионов, а число ограничений — несколько десятков или даже сотен тысяч.

3.1.1. Оптимальные решения в задачах линейного программирования

Прежде чем перейти к построению численных методов решения задач линейного программирования, выясним, какие точки допустимого множества могут потенциально быть решениями задачи. В этом нам существенным образом помогут важные особенности задач линейного программирования, а именно, полиэдральность допустимого множества и линейность целевой функции.

Определение 3.1.1. *Точка $x \in X$ называется крайней точкой множества X, если она не принадлежит внутренности никакого отрезка, целиком лежащего в X.*

В случае полиэдрального множества X его крайние точки называют также *угловыми точками* или *вершинами*. На рис. 3.1 показаны угловые точки выпуклого многогранника на плоскости.

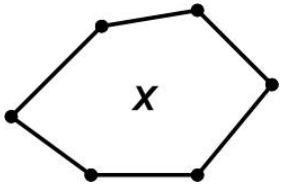

Рис. 3.1. Угловые точки допустимого множества

Обратимся теперь к полиэдральному множеству X вида

$$X = \left\{ x \in \mathbb{R}^n : \ \langle a_i, x \rangle \le b^i, \ 1 \le i \le l; \right. \\ \left. \langle a_i, x \rangle = b^i, \ \ l < i \le m \right\}, \tag{3.1.1}$$

где $a_i \in \mathbb{R}^n$, $1 \le i \le m$, $b \in \mathbb{R}^m$. Относительно l считаем, что это число может меняться от нуля до m, т.е. возможны случаи, когда множество X задается только ограничениями типа неравенства либо только ограничениями-равенствами.

Для точки $x \in X$ определим множество *активных ограничений* $J_0(x)$, положив

$$J_0(x) = \left\{ 1 \le i \le m : \ \langle a_i, x \rangle = b^i \right\}.$$

Понятно, что в данное множество активных ограничений всегда входят все ограничения типа равенства. В дальнейшем нам понадобится следующее понятие.

Определение 3.1.2. *Линейно независимые активные ограничения — это те активные ограничения, для которых соответствующие векторы a_i линейно независимы.*

Линейно независимые активные ограничения могут быть подмножеством множества всех активных ограничений в этой точке или совпадать с ним. Разумеется, число линейно независимых активных ограничений всегда не превышает число n. Имеет место следующий важный результат.

Утверждение 3.1.1. *Угловыми точками в множестве X вида (3.1.1) являются те и только те точки $x_v \in X$, которые определяются ровно n линейно независимыми активными ограничениями.*

Доказательство. *Необходимость.* Пусть x_v — угловая точка множества X и пусть r — число активных ограничений в этой точке, т.е. число индексов в множестве $J_0(x_v)$. Составим $r \times n$ матрицу $\tilde{A}$, строками которой являются соответствующие векторы a_i. Предположим, что максимальное число активных линейно независимых ограничений в точке x_v, равное k, меньше, чем n. Тогда ранг матрицы $\tilde{A}$ равен k. Поэтому линейная однородная система $\tilde{A}s = 0_r$ обязательно имеет ненулевое решение $s \in \mathbb{R}^n$.

Рассмотрим точку $x = x_v + \alpha s$, где $\alpha \in \mathbb{R}$. Для любого $\alpha \in \mathbb{R}$ и для всех активных ограничений (независимо от того, входят ли они в систему линейно независимых активных ограничений или нет) имеет место

$$\langle a_i, x \rangle = \langle a_i, x_v \rangle = b^i, \quad i \in J_0(x_v),$$

т.е. все они выполняются как равенства. Кроме того, так как $\langle a_i, x_v\rangle < b^i$ для $i \notin J_0(x_v)$, то всегда можно указать такое $\bar{\alpha} > 0$, что неравенство $\langle a_i, x\rangle \le b^i$ сохранится для любых $|\alpha| \le \bar{\alpha}$ и всех $i \notin J_0(x)$. Отсюда делаем вывод, что если взять отрезок $L = [x_v - \bar{\alpha}s, x_v + \bar{\alpha}s]$ с центром в точке x_v, то он целиком лежит во множестве X. Поэтому x_v не может быть угловой точкой X.

Достаточность. Пусть число линейно независимых активных ограничений в точке $x \in X$ равно n. Составим из соответствующих векторов a_i, $i \in J_0(x)$, матрицу $\tilde{A}$, взяв в качестве строк эти векторы. Данная матрица $\tilde{A}$ имеет полный ранг, равный n. Если x не угловая точка, то существуют не совпадающие между собой точки $x_1 \in X$ и $x_2 \in X$ такие, что $x = \frac{x_1+x_2}{2}$, т.е. x является серединой отрезка $[x_1, x_2]$, целиком лежащего в X.

Возьмем произвольный индекс $i \in J_0(x)$. Если $1 \le i \le l$, то в силу допустимости точек x_1 и x_2 для данного индекса выполняются неравенства: $\langle a_i, x_1\rangle \le b^i$, $\langle a_i, x_2\rangle \le b^i$. Поэтому

$$b^i = \langle a_i, x\rangle = \frac{1}{2}\langle a_i, x_1\rangle + \frac{1}{2}\langle a_i, x_2\rangle \le \frac{1}{2}b^i + \frac{1}{2}b^i = b^i,$$

откуда следует, что $\langle a_i, x_1\rangle = b^i$, $\langle a_i, x_2\rangle = b^i$. Эти равенства также имеют место и в случае, когда $l < i \le m$. Отсюда в силу произвольности индекса $i \in J_0(x)$ можно заключить, что линейная неоднородная система $\tilde{A}x = \tilde{b}$, где $\tilde{b}$ — подвектор вектора b, составленный из компонент b^i, $i \in J_0(x)$, имеет два отличных друг от друга решения. Но матрица $\tilde{A}$ полного ранга, равного n, поэтому данная система может иметь лишь единственное решение. Мы пришли к противоречию. Следовательно, x является угловой точкой множества X. ■

Согласно утверждению 3.1.1 каждая угловая точка x множества X определяется системой из n линейно независимых активных ограничений. Данная система может быть единственной, как показано на рис. 3.2, либо их может быть несколько (см. рис. 3.3, где таких систем три).

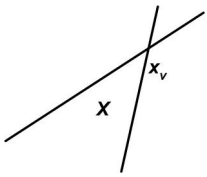

Рис. 3.2. Одна система линейно независимых активных ограничений

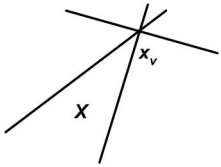

Рис. 3.3. Несколько систем линейно независимых активных ограничений

Обратимся теперь непосредственно к задаче линейного программирования. Допустимое множество в ней, независимо от того, в какой форме она задана, всегда есть полиэдральное множество типа (3.1.1). Если в задаче существует множество оптимальных решений, то обозначим его X_*. Следующий результат является одним из основных в теории задач линейного программирования. Приведем его без доказательства.

Теорема 3.1.1. *Пусть в задаче линейного программирования допустимое множество X содержит угловые точки и пусть множество оптимальных решений $X_* \subseteq X$ непусто. Тогда среди точек множества X_* найдется хотя бы одна угловая точка множества X.*

Из утверждения теоремы 3.1.1 следует важный вывод, а именно, что решение задачи линейного программирования может быть получено путем перебора только лишь угловых точек допустимого множества. Это и является основной идеей численного метода решения задач линейного программирования, называемого *симплекс-методом.*

3.1.2. Базис угловой точки

Суть симплекс-метода решения задачи линейного программирования заключается в *направленном переборе угловых точек* допустимого множества X с целью быстрейшего попадания в ту из них, которая принадлежит множеству оптимальных решений X_*. Поясним, каким образом это делается, на примере задачи линейного программирования в канонической форме:

$$\begin{array}{rcl} \langle c, x\rangle & \to & \min, \\ Ax & = & b, \\ x & \geq & 0_n, \end{array} \qquad (3.1.2)$$

где A — матрица размера $m \times n$, причем $m < n$. Ее столбцы в дальнейшем обозначаются a_i, $1 \le i \le n$. Предполагается, что A — матрица полного ранга, равного m, и что допустимое множество в задаче

$$X = \{x \in \mathbb{R}^n : \ Ax = b, \quad x \ge 0_n\}$$

не пусто. В этом случае, как можно показать, оно содержит, по крайней мере, одну угловую точку.

Применительно к задаче линейного программирования (3.1.2) в канонической форме можно дать другую *алгебраическую характеристику* угловой точки, учитывающую специальный вид допустимого множества X в этой задаче. Она связана с так называемым *базисным решением* системы $Ax = b$. Про решение x системы $Ax = b$ говорят как о ее базисном решении, если столбцы матрицы A, соответствующие ненулевым компонентам вектора x, линейно независимы. Базисное решение x системы $Ax = b$ называется *допустимым базисным решением*, если $x \ge 0_n$.

Покажем, что между угловыми точками множества X и допустимыми базисными решениями системы $Ax = b$ существует тесная связь, а более конкретно, что это, по существу, разные описания одних и тех же точек. Действительно, всего в задаче (3.1.2) $m + n$ ограничений. Из них m линейных равенств $Ax = b$, а также n линейных неравенств $x^j \ge 0$. Здесь $1 \le j \le n$. Как следует из утверждения 3.1.1, точка x из допустимого множества X может быть угловой только в том случае, когда в ней имеются активные ограничения в количестве не меньше, чем n штук, причем среди них найдется n линейно независимых. Ограничения типа равенства всегда активны в любой допустимой точке. Более того, все они линейно независимы, так как по предположению линейно независимы строки матрицы A. Но ограничений типа равенства всего m штук. Недостающие активные ограничения в количестве по меньшей мере $d = n - m$ штук должны добавляться из ограничений типа неравенства $x \ge 0_n$. Поэтому если точка $x \in X$ такова, что в ней меньше, чем d нулевых компонент, то она заведомо не может быть угловой точкой X. Чтобы x была угловой точкой X, необходимо, чтобы количество нулевых компонент в ней равнялось или превышало d.

Предположим теперь, что точка $x \in X$ такова, что в ней только $k \le m$ строго положительных компонент, а остальные компоненты — нулевые. Так как теперь количество нулевых компонент $n - k$ равно или превышает нужное число d, то такая точка в принципе может оказаться угловой. Все зависит от того, найдется ли в ней или не найдется n линейно независимых активных ограничений.

Пусть x^B — подвектор вектора $x \in X$, в котором собраны все положительные компоненты, а x^N — подвектор x, содержащий все нулевые компоненты. Не умаляя общности, считаем, что x^B состоит из первых k компонент x, а x^N — из последующих $n-k$ компонент x. Таким образом, имеет место разбиение: $x = [x^B, x^N]$, где $x^B > 0_k$, $x^N = 0_{n-k}$. В соответствии с разбиением вектора x разобьем также матрицу A на две подматрицы: $A = [A_B, A_N]$, где подматрица A_B состоит из первых k столбцов матрицы A, а A_N — из оставшихся $n-k$ столбцов A. Если среди всех ограничений задачи оставить только активные ограничения в точке x, то приходим к тому, что эта точка $x = [x^B, x^N]$ удовлетворяет следующей линейной системе из $n+m-k$ уравнений:

$$\begin{bmatrix} A_B & A_N \\ 0_{(n-k)k} & I_{n-k} \end{bmatrix} \begin{bmatrix} x^B \\ x^N \end{bmatrix} = \begin{bmatrix} b \\ 0_{n-k} \end{bmatrix}, \tag{3.1.3}$$

где I_{n-k} — единичная матрица порядка $n-k$. Так как $k \le m$, то количество уравнений в системе (3.1.3) равно или больше n.

Согласно утверждению 3.1.1 точка x будет угловой, если среди строк матрицы этой системы найдется n линейно независимых. Поскольку ранг по столбцам совпадает с рангом по строкам, то это действительно будет так, если данная матрица обладает полным рангом по столбцам, т.е. ее столбцы линейно независимы. Нетрудно видеть, что для этого достаточно, чтобы линейно независимыми были столбцы матрицы A_B. Понятно, что это может быть только в том случае, когда $k \le m$. Таким образом, условие, что *число положительных компонент не превышает* m, является необходимым для того, чтобы точка $x \in X$ была угловой. Мы убедились также, что угловым точкам допустимого множества в задаче (3.1.2) соответствуют допустимые базисные решения системы $Ax = b$.

Ниже нам потребуется еще одно важное понятие. Угловая точка $x \in X$ с k положительными компонентами называется *невырожденной*, если $k = m$ и *вырожденной*, если $k < m$.

Пусть x — невырожденная угловая точка допустимого множества X в задаче (3.1.2). Тогда линейно независимые столбцы a_{i_1}, …, a_{i_m} матрицы A, соответствующие положительным компонентам вектора x, называют *базисом угловой точки*, а матрицу, составленную из них, — *матрицей базиса*. В дальнейшем матрицу базиса будем обозначать B, т.е.

$$B = [a_{i_1}, \ldots, a_{i_m}].$$

В случае вырожденной угловой точки ее базис строится несколько иным образом. Берутся все столбцы матрицы A_B и к ним добавляются

произвольные другие $m-k$ столбцов матрицы A_N так, чтобы в совокупности они были бы линейно независимы. Другими словами, чтобы составленная из них матрица была бы неособой квадратной матрицей порядка m. По-прежнему эту матрицу будем обозначать B, а входящие в нее столбцы называть базисом точки x. У вырожденной угловой точки может быть несколько базисов, у невырожденной — только один.

Каноническая задача линейного программирования (3.1.2) называется *невырожденной*, если все угловые точки ее допустимого множества невырожденные. В дальнейшем будем предполагать, что задача (3.1.2) является невырожденной.

3.1.3. Симплекс-метод

Как уже говорилось, симплекс-метод — это целенаправленный переход от одной угловой точки допустимого множества к другой. Данный процесс удобно описывать через сопоставляемые этим угловым точкам базисы. По существу, это просто переход от одного базиса к другому, причем так, чтобы в новой угловой точке, соответствующей новому базису, значение целевой функции было меньше. Встает вопрос, каким образом это делать? Опишем данную процедуру подробно, условно ее можно разбить на три этапа.

Пусть x — невырожденная угловая точка в задаче (3.1.2) и пусть, для определенности, ее базис B состоит из первых m столбцов матрицы A. Подматрицу матрицы A, составленную из оставшихся $n-m$ столбцов, обозначим N. Тогда $A=[B,N]$ и вектор x разобьется на подвекторы:

$$x=\begin{bmatrix} x^B \\ x^N \end{bmatrix}, \quad x^B>0_m, \quad x^N=0_{n-m}. \tag{3.1.4}$$

Имеем также

$$b=Bx^B+Nx^N=Bx^B, \quad x^B=B^{-1}b. \tag{3.1.5}$$

Компоненты вектора x^B называют *базисными переменными*, а компоненты вектора x^N — *внебазисными переменными*. Разобьем также в соответствии с разбиением матрицы A вектор c, положив $c=[c^B,c^N]$. Через $J_B(x)$ будем обозначать множество индексов базисных переменных, через $J_N(x)$ — множество индексов внебазисных переменных. При сделанных допущениях $J_B(x)=[1:m]$, $J_N(x)=[m+1:n]$.

1. Выбор столбца матрицы N для ввода в базис. Чтобы перейти к новому базису, надо по крайней мере взять один столбец из

матрицы N и заменить им какой-то столбец матрицы B. Встает вопрос, какой именно столбец следует взять из N?

Чтобы выяснить это, возьмем произвольный столбец a_k из матрицы N и рассмотрим $n-m$ векторов:

$$x_k = x_k(\lambda) = \begin{bmatrix} x^B - \lambda B^{-1} a_k \\ 0 \\ \vdots \\ \lambda \\ \vdots \\ 0 \end{bmatrix}, \quad k \in J_N(x), \tag{3.1.6}$$

где $\lambda > 0$ и компонента λ является k-й компонентой вектора $x_k(\lambda)$. Имеем в силу (3.1.5) и (3.1.6)

$$Ax_k = Bx^B - \lambda BB^{-1}a_k + \lambda a_k = Bx^B = b,$$

поэтому ограничение типа равенства $Ax = b$ для всех векторов $x_k(\lambda)$ выполняется. Более того, так как $x^B > 0$ и $\lambda > 0$, то для λ положительных и достаточно малых получаем, что $x_k(\lambda) \geq 0_n$. Таким образом, $x_k(\lambda)$ — допустимая точка для $0 < \lambda \leq \tilde{\lambda}_k$, где $0 < \tilde{\lambda}_k \leq \infty$.

Рассмотрим, какие значения принимает целевая функция в точках $x_k(\lambda)$. Имеем

$$\begin{aligned} \langle c, x_k(\lambda)\rangle &= \langle c^B, x^B\rangle - \lambda\langle c^B, B^{-1}a_k\rangle + \lambda c^k = \\ &= \langle c^B, x^B\rangle - \lambda\left[\langle c^B, B^{-1}a_k\rangle - c^k\right]. \end{aligned}$$

Обозначим

$$\Delta_k = \langle c^B, B^{-1}a_k\rangle - c^k, \quad k \in J^n,$$

и составим из этих величин вектор $\Delta = [\Delta_1, \ldots, \Delta_n]$. Формально его компоненты определены при всех $k \in J^n$, а не только при $k \in J_N(x)$. Тогда

$$\langle c, x_k(\lambda)\rangle = \langle c, x\rangle - \lambda\Delta_k, \quad k \in J_N(x). \tag{3.1.7}$$

Вектор Δ иногда называют *вектором оценок замещения*. Как нетрудно проверить, если $k \in J_B(x)$, то $\Delta_k = 0$. Действительно, $B^{-1}a_k = e_k$ при этих k, где e_k — k-й единичный орт из $\mathbb{R}^m$. Поэтому $\langle c^B, e_k\rangle = c^k$, что и приводит к равенству $\Delta_k = 0$. При остальных k компонента Δ_k может принимать отличные от нуля значения. Понятно, что вектор a_k с номером $k \in J_N(x)$ целесообразно вводить в базис только в том случае, когда соответствующая оценка замещения Δ_k положительна.

В силу (3.1.7) это может привести к уменьшению значения целевой функции.

2. Проверка оптимальности решения или его отсутствия. Рассмотрим сначала случаи, которые позволяют сделать вывод, что текущая точка x либо оптимальна, либо в задаче (3.1.2) вообще не существует решения.

Случай оптимальности текущей точки. Смотрим оценки замещения Δ_k, $k \in J^n$. Если все $\Delta_k \leq 0$ (для этого фактически надо проверить оценки замещения только при $k \in J_N(x)$), то текущая точка x является оптимальной. Это видно из формулы (3.1.7), так как ввод в базис любого нового вектора a_k из N приводит к тому, что значение целевой функции увеличивается (по крайней мере, не убывает). Но можно обосновать это и более формальным образом. Обозначим $u = (B^{-1})^T c^B$. Тогда условие $\Delta_k \leq 0$, $k \in J^n$, может быть записано как

$$\Delta_k = \langle (B^{-1})^T c^B, a_k \rangle - c^k = \langle u, a_k \rangle - c^k \leq 0, \quad k \in J^n, \tag{3.1.8}$$

причем $\Delta_k = 0$, $k \in J_B(x)$. В матричной форме неравенства (3.1.8) принимают вид: $A^T u \leq c$. Отсюда следует, что если обратиться к задаче, двойственной к рассматриваемой задаче (3.1.2),

$$\begin{array}{rcl} \langle b, u \rangle & \to & \max, \\ A^T u & \leq & c, \end{array} \tag{3.1.9}$$

то точка $u = (B^{-1})^T c^B$ оказывается допустимой в ней. Кроме того,

$$\langle c, x \rangle = \langle c^B, x^B \rangle = \langle c^B, B^{-1} b \rangle = \langle (B^{-1})^T c^B, b \rangle = \langle u, b \rangle.$$

Мы получили, что точки x и u являются допустимыми и в них значения целевых функций совпадают. Поэтому по теореме двойственности для пары задач (3.1.2) и (3.1.9) обе точки являются оптимальными, каждая в своей задаче.

Случай отсутствия решения. Существует такой индекс $k \in J_N(x)$, что $\Delta_k > 0$ и $B^{-1} a_k \leq 0_m$. В этом случае можно сделать вывод, что допустимое множество X неограничено и существует луч, целиком принадлежащий допустимому множеству, вдоль которого целевая функция стремится к $-\infty$. Действительно, при этом предположении $x_k(\lambda) \geq 0_n$ при любом $\lambda \geq 0$. Кроме того, всегда по построению $A x_k(\lambda) = b$. Поэтому $x_k(\lambda) \in X$ для всех $\lambda \geq 0$. Имеем для этих $x_k(\lambda)$:

$$\langle c, x_k(\lambda) \rangle = \langle c, x \rangle - \lambda \Delta_k \to -\infty$$

при $\lambda \to +\infty$. В этом случае работа симплекс-метода также прерывается.

Рассмотрим, наконец, последнюю возможность, которая позволяет перейти в новую угловую точку множества X с новым базисом. Такой переход происходит, если для некоторого $k \in J_N(x)$ выполняется неравенство $\Delta_k > 0$, однако $B^{-1}a_k \not\leq 0_m$.

3. Выбор столбца матрицы B для вывода из базиса. Пусть найдутся такие $k \in J_N(x)$ и $i \in J_B(x)$, что $\Delta_k > 0$ и $(B^{-1}a_k)^i > 0$. В этом случае можно сделать итерационный шаг. Обозначим

$$J_B^+(x) = \left\{ i \in J_B(x) : \ \left(B^{-1}a_k\right)^i > 0 \right\}$$

и выберем из множества $J_B^+(x)$ такой индекс s, что

$$\lambda_* = \min_{i \in J_B^+(x)} \frac{(B^{-1}b)^i}{(B^{-1}a_k)^i} = \frac{(B^{-1}b)^s}{(B^{-1}a_k)^s}. \tag{3.1.10}$$

Заметим, что $\lambda_* > 0$, так как $(B^{-1}b)^i = (x^B)^i > 0$ для любого $i \in J_B(x)$.

Точка $\bar{x} = x_k(\lambda_*)$ — допустимая, а номер s, при котором достигается минимум в (3.1.10), — единственный. Это следует из предположения о невырожденности задачи. Более того, новая точка $\bar{x}$ также будет угловой точкой множества X. Базис $\bar{B}$ точки $\bar{x}$ получается из базиса $B = [a_1, \ldots, a_s, \ldots, a_m]$ старой точки x удалением столбца с номером s и помещением на его место столбца a_k, т.е.

$$\bar{B} = [a_1, \ldots, a_k, \ldots, a_m].$$

В новой точке $\bar{x}$ компонента $\bar{x}^s$ равняется нулю, а компонента $\bar{x}^k$, напротив, становится положительной. Кроме того,

$$\langle c, \bar{x} \rangle = \langle c, x \rangle - \lambda_* \Delta_k < \langle c, x \rangle,$$

т.е. при переходе из x в новую точку $\bar{x}$ значение целевой функции уменьшается. Итерация на этом заканчивается и метод идет на выполнение следующей итерации. Если решение существует, то метод всегда его найдет, хотя, быть может, только одну точку из возможного множества решений задачи, но при этом обязательно угловую точку допустимого множества X.

Утверждение 3.1.2. *Новая матрица $\bar{B}$ является неособой, т.е. определяет базис новой точки $\bar{x}$.*

Доказательство. Так как $b = Bx^B$, то вектор b принадлежит выпуклому конусу $\mathrm{pos}B$, натянутому на столбцы $a_1, \ldots, a_m$ матрицы B.

Более того, из-за линейной независимости этих столбцов следует, что данный конус симплициальный, т.е. имеет непустую внутренность. Поскольку $x^B > 0_m$, то вектор b принадлежит внутренности этого конуса и не может принадлежать ни одной его грани, порожденной столбцами матрицы B в количестве $m-1$ штук или меньше. Он также не может принадлежать ни одному $(m-1)$-мерному линейному подпространству, которые содержат эти грани.

Предположим, что новая матрица $\bar{B}$ оказалась особой. Это означает, что новый введенный столбец лежит в каком-то $(m-1)$-мерном линейном подпространстве, содержащем $(m-1)$-мерную грань конуса $\mathrm{pos}B$. В этом случае выпуклый конус $\mathrm{pos}\bar{B}$ также принадлежит этому подпространству. Так как $b = \bar{B}\bar{x}$, где $\bar{x} > 0_m$, то b принадлежит выпуклому конусу $\mathrm{pos}\bar{B}$, и, стало быть, оказывается лежащим в упомянутом линейном подпространстве размерности $m-1$. Мы пришли к противоречию. ■

Опишем теперь алгоритм решения задачи линейного программирования (3.1.2), основанный на приведенных рассуждениях.

Алгоритм симплекс-метода

Пусть дана начальная угловая точка x_0 и пусть найдена текущая угловая точка x_v.

Шаг 1. Ставим в соответствие точке x_v множества индексов базисных переменных $J_B = J_B(x_v)$ и внебазисных переменных $J_N = J_N(x_v)$, а также матрицу базиса B. Подсчитываем оценки замещения Δ_k для $k \in J_N$.

Шаг 2. Если $\Delta_k \le 0$ для любого $k \in J_N$, то текущая точка x_v является решением задачи. Процесс останавливается с найденным решением.

Шаг 3. Если существует индекс $k \in J_N$, для которого $\Delta_k > 0$ и $B^{-1}a_k \le 0_m$, то процесс прерывается с утверждением, что в задаче не существует решения.

Шаг 4. Если существует такое $k \in J_N$, что $\Delta_k > 0$ и $B^{-1}a_k \not\le 0$, то переходим в новую угловую точку $\bar{x}_v$ с меньшим значением целевой функции.

Шаг 5. Идем на шаг 1.

Данный алгоритм конечный, так как конечно множество угловых точек в задаче. Нетрудно видеть, что их число не может превышать количество возможных базисов $C_n^m = C_n^{n-m}$. На самом деле эта оценка на угловые точки несколько завышена, так как нас интересуют лишь те угловые точки, которым соответствуют допустимые базисные реше-

ния. Но даже для сравнительно скромных чисел m и n данная оценка на количество возможных угловых точек у допустимого множества X в задаче (3.1.2) может оказаться катастрофически большой. Тем не менее симплекс-метод является достаточно работоспособным и нашел широкое применение. Современные его реализации позволяют решать задачи больших и сверхбольших размеров, особенно с разреженными данными. Как показывает практика, в большинстве случаев количество требуемых итераций зависит главным образом от числа ограничений типа равенства m и равняется примерно нескольким m.

Симплекс-метод в достаточно законченной современной форме был предложен Дж.Б. Данцигом в 1947 году с целью отыскания оптимальных решений в практических задачах принятия решения. Но еще в 1939 году Л.В. Канторович опубликовал работу, в которой указал на важность линейного программирования как средства моделирования прикладных задач экономики. В данной работе был рассмотрен и метод решения таких задач, который является прообразом симплекс-метода.

Симплекс-метод в табличной форме

Одно из достоинств симплекс-метода заключается в возможности организовать его работу в табличной форме. Основой конструкции является симплекс-таблица, в которую помещается информация о задаче, связанная с текущей угловой точкой x множества X и с ее базисом — столбцами a_i, $i \in J_B(x)$.

Размер таблицы — $(m+1) \times (n+1)$. Будем считать, что строки и столбцы нумеруются от 0 до m и до n соответственно. Саму таблицу будем обозначать Z, а ее элементы z_{ij}, $0 \le i \le m$, $0 \le j \le n$. Конкретное заполнение таблицы, как уже говорилось, связано с текущей угловой точкой (точнее, с базисом данной угловой точки). По сути в ней содержатся разложения всех столбцов матрицы A и правой части b по векторам базиса этой точки, а также оценки замещения. Процесс преобразования таблицы проводится с помощью известного из линейной алгебры *метода Жордана-Гаусса*.

Пусть для определенности базис рассматриваемой угловой точки составляют первые m столбцов матрицы A, т.е. $J_B(x) = \{1, \ldots, m\}$, $J_N(x) = \{m+1, \ldots, n\}$. Тогда $A = [B, N]$, где $B = [a_1, \ldots, a_m]$ — базис, $N = [a_{m+1}, \ldots, a_n]$. Имеем $\det B \neq 0$ и из $x = [x^B, x^N]^T$, $x^N = 0_{n-m}$ следует, что $x^B = B^{-1}b$, причем $x^B > 0_m$.

Обозначим $a_0 = b$ и $\tilde{A} = [a_0, a_1, \ldots, a_n]$. По другому матрицу $\tilde{A}$ можно представить как $\tilde{A} = [b, B, N]$. Пусть $\tilde{Z}$ — это матрица Z без нулевой строки. Она имеет размер $m \times (n+1)$. Заполним $\tilde{Z}$, полагая

$\tilde{Z} = B^{-1}\tilde{A}$, или, в более подробной записи, $\tilde{Z} = [\tilde{z}_0, \tilde{Z}_B, \tilde{Z}_N]$, где

$$tilde z_0 = B^{-1}a_0 = B^{-1}b = x^B, \quad \tilde{Z}_B = B^{-1}B = I_m, \quad \tilde{Z}_N = B^{-1}N.$$

Таким образом,

$$z_{ij} = (B^{-1}a_j)^i, \quad 1 \le i \le m, \quad 0 \le j \le n.$$

Заполнение нулевой строки матрицы Z отличается от заполнения других строк, а именно, ее элементы с первого вплоть до последнего полагаются равными $z_{0j} = \Delta_j$, $1 \le j \le n$, где Δ — вектор оценок замещения в данной угловой точке x. Элемент z_{00} берется равным значению целевой функции в точке x, т.е. $z_{00} = \langle c, x\rangle$. С целью общности данное значение обозначим так же, как

$$\Delta_0 = \langle c, x\rangle = \langle c^B, x^B\rangle = \langle c^B, B^{-1}a_0\rangle = \langle c^B, B^{-1}a_0\rangle - \mathrm{c}^0,$$

где положено: $c^0 = 0$.

Удобно вверху над столбцами таблицы ставить номера соответствующих столбцов матрицы A, разложения которых даются в этих колонках. Также слева указываются номера столбцов базиса (в приведенной ниже таблице для наглядности указываются сами столбцы).

	$a_0 = b$	a_1	$\cdots$	a_m	a_{m+1}	$\cdots$	a_n
	$\Delta_0 = c^T x$	0	$\cdots$	0	Δ_{m+1}	$\cdots$	Δ_n
a_1 $\vdots$ a_m	$x^B = B^{-1}b$	$\tilde{Z}_B$	$=$	I_m	$\tilde{Z}_N$	$=$	$B^{-1}N$

После заполнения симплекс-таблицы проводится ее анализ. Возможны три ситуации:

1. Если $z_{0k} \le 0$ для всех $k \in J_N(x)$, то текущая точка x является решением задачи, дальнейшие расчеты прекращаются. Оптимальное решение и оптимальное значение целевой функции находятся в нулевом столбце матрицы Z, а именно:

$$x_*^B = [z_{10}, \ldots, z_{m0}]^T, \quad f_* = \langle c^B, x_*^B\rangle = z_{00}.$$

2. Если $z_{0k} > 0$ для некоторого $k \in J_N(x)$ и $z_{ik} \le 0$ для всех $1 \le i \le m$, то расчеты также прекращаются, при этом в задаче не существует решения и можно указать луч, принадлежащий допустимому множеству, вдоль которого значение целевой функции стремится к $-\infty$.

3. Если $z_{0k} > 0$ для некоторого $k \in J_N(x)$ и среди остальных компонент столбца с номером k имеются положительные компоненты, то определяются номера всех таких положительных компонент. Пусть $J_k^+(x) = \{i \in J_B(x) : \ z_{ik} > 0\}$. Подсчитывается величина

$$\lambda_* = \min_{i \in J_k^+(x)} \frac{z_{i0}}{z_{ik}} = \frac{z_{s0}}{z_{sk}}.$$

Элемент $z_{sk} > 0$ объявляется *ведущим элементом*. Столбец a_s выводится из базиса, а столбец a_k, напротив, вводится в базис. Все это соответствует переходу в новую угловую точку $\bar{x}$ с новым множеством базисных переменных $J_B(\bar{x}) = (J_B(x) \setminus \{s\}) \cup \{k\}$ и с новым базисом $\bar{B}$, в который уже входит столбец a_k. Слева вместо номера s ставится номер k. После этого производится перерасчет всех элементов таблицы.

Пусть $\bar{Z}$ — новая таблица, соответствующая новому базису $\bar{B}$. Перерасчет производится по следующим формулам:

$$\bar{z}_{ij} = z_{ij} - \frac{z_{ik}}{z_{sk}} z_{sj}, \quad 0 \le i \le m, \ i \ne s, \quad 0 \le j \le n, \tag{3.1.11}$$

$$\bar{z}_{sj} = \frac{1}{z_{sk}} z_{sj}, \quad 0 \le j \le n. \tag{3.1.12}$$

Таким образом, из всех строк, кроме s-й, вычитается s-я строка, умноженная предварительно на коэффициент $\frac{z_{ik}}{z_{sk}}$, где i — номер строки, из которой производится вычитание. Строка с номером s просто делится на ведущий элемент. При таком преобразовании в новой таблице $\bar{Z}$ в k-м столбце элемент $\bar{z}_{sk}$ становится равным единице, а остальные элементы $\bar{z}_{ik}$ оказываются равными нулю.

К формулам перерасчета (3.1.11), (3.1.12) можно прийти следующим образом. С учетом предположения о том, что базис в точке x состоит из первых m столбцов матрицы A и, значит, суммирование вместо $i \in J_B(x)$ можно ввести по i от единицы до m, имеем

$$a_k = B(B^{-1}a_k) = \sum_{i=1}^{m} z_{ik} a_i = \sum_{i=1, \ i \ne s}^{m} z_{ik} a_i + z_{sk} a_s,$$

и, поскольку ведущий элемент z_{sk} всегда положителен, получаем

$$a_s = \frac{1}{z_{sk}} a_k - \sum_{i=1, \ i \ne s}^{m} \frac{z_{ik}}{z_{sk}} a_i. \tag{3.1.13}$$

Далее, так как

$$a_j = \sum_{i=1}^{m} z_{ij} a_i, \quad 0 \le j \le n,$$

то, подставляя (3.1.13), приходим к следующему представлению векторов $a_0, \ldots, a_n$ в новом базисе.

$$\begin{array}{rcl} a_j & = & \sum_{i=1,\ i\ne s}^{m} z_{ij} a_i + z_{sj} a_s = \\ & = & \sum_{i=1,\ i\ne s}^{m} z_{ij} a_i + z_{sj} \left(\frac{1}{z_{sk}} a_k - \sum_{i=1,\ i\ne s}^{m} \frac{z_{ik}}{z_{sk}} a_i \right) = \\ & = & \sum_{i=1,\ i\ne s}^{m} \left(z_{ij} - \frac{z_{ik}}{z_{sk}} z_{sj} \right) a_i + \frac{z_{sj}}{z_{sk}} a_k, \end{array}$$

что полностью соответствует формулам пересчета (3.1.11), (3.1.12).

Для элементов нулевой строки, учитывая, что $c^0 = 0$, имеем

$$z_{0j} = \langle c^B, B^{-1} a_j \rangle - c^j = \sum_{i=1}^{m} c^i z_{ij} - c^j, \quad 0 \le j \le n.$$

Поэтому, принимая во внимание, что в новой таблице коэффициенты разложения по k-му столбцу, введенному в базис, находятся в s-й строке, получаем

$$\begin{array}{rcl} \bar{z}_{0j} & = & \sum_{i=1,\ i\ne s}^{m} c^i \bar{z}_{ij} + c^k \bar{z}_{sj} - c^j = \\ & = & \sum_{i=1,\ i\ne s}^{m} c^i \left(z_{ij} - \frac{z_{ik}}{z_{sk}} z_{sj} \right) + \frac{z_{sj}}{z_{sk}} c^k - c^j = \\ & = & \sum_{i=1}^{m} c^i \left(z_{ij} - \frac{z_{ik}}{z_{sk}} z_{sj} \right) + \frac{z_{sj}}{z_{sk}} c^k - c^j = \\ & = & \sum_{i=1}^{m} c^i z_{ij} - c^j - \left(\sum_{i=1}^{m} c^i z_{ik} - c^k \right) \frac{z_{sj}}{z_{sk}} = z_{0j} - \frac{z_{0k}}{z_{sk}} z_{sj}. \end{array}$$

Так как разложения столбцов матрицы A, входящих в текущий базис, всегда дают единичные векторы, а соответствующие оценки замещения для этих столбцов всегда равны нулю, то нет особого смысла держать эти столбцы в таблице и поэтому их часто опускают. При этом, разумеется, следует указывать вверху таблицы номера тех столбцов матрицы A, которые входят в подматрицу N и которые сохраняются в таблице.

Выбор начальной угловой точки

Для нахождения начальной угловой точки применяются несколько способов.

Метод искусственного базиса. Считаем, что все компоненты вектора b неотрицательны, т.е. $b \ge 0_m$. Если же для какой-либо компоненты b^i вектора правой части b выполняется обратное неравенство

$b^i < 0$, то, умножая одновременно i-ю строку матрицы A и i-ю компоненту вектора b на -1, приходим к выполнению нужного требования. Строится *вспомогательная задача*:

$$\begin{array}{c} \langle \bar{e}, y \rangle \to \min, \\ Ax + y = b, \\ x \geq 0_n,\ y \geq 0_m, \end{array} \tag{3.1.14}$$

где $\bar{e}$ — m-мерный вектор, состоящий из единиц. Пространство переменных в этой задаче расширяется по сравнению с исходной задачей (3.1.2). В качестве начальной угловой точки берется точка

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0_n \\ b \end{bmatrix}. \tag{3.1.15}$$

Если $b > 0_m$, то данная угловая точка обязательно будет невырожденной.

Далее вспомогательная задача (3.1.14) решается симплекс-методом. Пусть найдено ее решение и пусть x_* и y_* — соответственно векторы x и y в решении. Обозначим $\mu_* = \langle \bar{e}, y_* \rangle$. Возможны две ситуации.

1) $\mu_* = 0$, т.е. $y_* = 0_m$. В этом случае x_* — угловая точка в исходной задаче.

2) $\mu_* > 0$, т.е. для хотя бы одной компоненты вектора y_* выполняется $y_*^i > 0$. В этом случае допустимое множество X в исходной задаче (3.1.2) должно быть пустым. Действительно, если $X \neq \emptyset$, то найдется $x \in X$. Но тогда точка

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0_m \end{bmatrix}.$$

будет решением вспомогательной задачи и в ней значение целевой функции $\langle \bar{e}, y \rangle$ равно нулю.

С использованием метода искусственного базиса реализуется *двухфазный симплекс-метод* решения задачи линейного программирования. Первая фаза состоит из решения вспомогательной задачи (3.1.14). Вторая фаза заключается в решении исходной задачи линейного программирования (3.1.2) из найденной угловой точки. Искусственная переменная при этом отбрасывается.

М-метод. В данном методе объединяются оба этапа нахождения начальной угловой точки и последующего решения исходной задачи.

Пусть M — достаточно большая положительная константа и пусть по-прежнему $b \geq 0_m$. Составляется так называемая *M-задача*:

$$\begin{gathered} \langle c, x\rangle + M\langle \bar{e}, y\rangle \to \min, \\ Ax + y = b, \\ x \geq 0_n,\ y \geq 0_m. \end{gathered} \tag{3.1.16}$$

В качестве начальной угловой точки может быть взята, например, точка (3.1.15).

Утверждение 3.1.3. *Пусть исходная задача линейного программирования (3.1.2) разрешима. Тогда найдется такое число $M_* \geq 0$, что для всех $M > M_*$ в любом решении $[x_*, y_*]$ задачи (3.1.16) точка x_* будет оптимальной для исходной задачи (3.1.2). При этом $y_* = 0_m$.*

Встает вопрос, как выбирать константу M в (3.1.16)? Можно показать, что если взять величину $M_* = \max_{1\leq i\leq m} |u_*^i|$, где u_* — оптимальное решение задачи (3.1.9), двойственной к (3.1.2), то утверждение 3.1.3 оказывается справедливым.

Модифицированный симплекс-метод

В обычном симплекс-методе работают с симплекс-таблицей Z. Ее элементы преобразуются по формулам (3.1.11), (3.1.12). В модифицированном симплекс-методе работают с матрицей базиса B, точнее, с обратной к ней матрицей B^{-1}. Объясняется это тем, что для получения информации о всех величинах, связанных с текущей угловой точкой, достаточно только знать матрицу B^{-1}. Действительно, имея B^{-1}, легко находим разложения всех столбцов матрицы A по текущему базису, а также величины

$$u = \left(B^{-1}\right)^T c^B, \quad \Delta_k = \langle u, a_k\rangle - c^k, \quad x^B = B^{-1}b.$$

Поэтому хранится только матрица B^{-1}, а новая матрица $\bar{B}^{-1}$, соответствующая новой угловой точке $\bar{x}$, вычисляется по старой матрице B^{-1} с использованием рекуррентных соотношений, подобных (3.1.11), (3.1.12).

Пусть B и $\bar{B}$ — соответственно базисы старой и новой угловых точек:

$$B = [a_1, \ldots, a_s, a_{s+1}, \ldots a_m], \quad \bar{B} = [a_1, \ldots, a_k, a_{s+1}, \ldots a_m].$$

Пусть, кроме того, $B^{-1} = [\beta_{ij}]$, $\bar{B}^{-1} = \left[\bar{\beta}_{ij}\right]$, $1 \leq i, j \leq m$. Формулы пересчета имеют следующий вид:

$$\bar{\beta}_{ij} = \beta_{ij} - \frac{z_{ik}}{z_{sk}}\beta_{sj}, \quad i \neq s; \qquad \bar{\beta}_{sj} = \frac{1}{z_{sk}}\beta_{sj}.$$

Справедливость этих формул проверяется путем умножения матрицы $\bar{B}$ на матрицу $\bar{B}^{-1}$.

Одним из достоинств модифицированного симплекс-метода является возможность построения такой вычислительной схемы, в которой в случае разреженной матрицы A участвуют лишь ненулевые элементы. Рассмотренный вариант модифицированного симплекс-метода называют иногда *симплекс-методом с обратной матрицей.*

3.1.4. Двойственный симплекс-метод

Так как задача, двойственная к задаче линейного программирования, сама является задачей линейного программирования (хотя и в другой форме), то понятно, что для нее также может быть разработан свой симплекс-метод, где уже осуществляется переход по угловым точкам допустимого множества двойственной задачи. Более того, как и в случае симплекс-метода для прямой задачи (3.1.2), когда наряду с решением x задачи (3.1.2) мы находим в конце решение u двойственной задачи (3.1.9), двойственный симплекс метод также позволяет получить в конце решение исходной задачи (3.1.2).

Предположим, что обе задачи линейного программирования (3.1.2) и (3.1.9) имеют решения. Тогда согласно теореме об оптимальных решениях в паре задач линейного программирования (см. [33]) эти решения x и u удовлетворяют следующей системе необходимых и достаточных условий:

$$\begin{aligned} D(x)v &= 0_n, \\ Ax &= b, \\ v &= c - A^T u, \end{aligned} \tag{3.1.17}$$

где $x \geq 0_n$, $v \geq 0_n$ и $D(x)$ — диагональная матрица с вектором x на диагонали. Вектор v, входящий в эти условия, называют *слабой двойственной переменной* или *двойственной невязкой.* Обозначим $v(u) = c - A^T u$. Если предположить, что в двойственной задаче все угловые точки допустимого множества

$$U = \{u \in \mathbb{R}^m : \ A^T u \leq c\}$$

невырожденные, т.е. определяются равно m активными ограничениями, то в любой угловой точке $u \in U$ у вектора $v = v(u)$ ровно m компонент равны нулю, а остальные строго положительные.

В рассмотренном в предыдущем разделе симплекс-методе выбирался базис, соответствующий некоторой угловой точке x допустимого множества X в прямой задаче (3.1.2). Если, используя этот базис, определить двойственную точку $u = (B^{-1})^T c^B$ и вектор $v = v(u)$, то

выполняются все равенства из системы условий (3.1.17) и неравенство $x \geq 0_n$. Не выполняется только неравенство $v \geq 0_n$. Это означает, что точка u не является допустимой в двойственной задаче, т.е. не принадлежит множеству U. Однако из n неравенств $A^T u \leq c$ в этой точке выполняются как строгие равенства ровно m из них. Такие точки $u \in \mathbb{R}^m$ принято называть *псевдовершинами* U.

Предполагая, что все псевдовершины также невырожденны, т.е. из n неравенств $A^T u \leq c$ могут обратиться в равенство не более чем m из них, получаем, что у вектора $v = v(u)$, соответствующего псевдовершине, m компонент равны нулю, а остальные отличны от нуля (либо положительны, либо отрицательны). Разбивая в соответствии с разбиением вектора $x = [x^B, x^N]^T$ вектор $v = v(u) = [v^B, v^N]^T$, получаем $v^B = 0_m$. Данный вектор v является обратным по отношению к вектору оценок замещения Δ, т.е. $v = -\Delta$. Если для данного базиса B оказывается, что $v \geq 0_n$, то текущая угловая точка $x \in X$ является решением прямой задачи (3.1.2), а точка $u = (B^{-1})^T c^B$ — решением двойственной задачи. Это следует, например, из того, что выполняются все условия (3.1.17). Если нет, то переходят к новому базису, при этом вводят в базис тот k-й столбец матрицы A (точнее, столбец ее подматрицы N), для которого соответствующая компонента v^k отрицательна. Переходя от одной вершины множества X к другой вершине, мы одновременно переходим от одной псевдовершины U к другой псевдовершине. Процесс продолжается до тех пор, пока вместо псевдовершины мы окажемся в вершине множества U.

Можно пойти несколько иным путем и, наоборот, брать такие базисы, чтобы им соответствовали вершины множества U, а вычисляемые по этим базисам точки x удовлетворяли равенству $Ax = b$, но, быть может, имели отрицательные компоненты. Другими словами, были бы *псевдовершинами* множества X. Это приводит к понятию *двойственного симплекс-метода*. По существу двойственный симплекс-метод это обычный симплекс-метод, но примененный к решению двойственной задачи и строящий параллельно последовательность псевдовершин множества X. Значения целевой функции $\langle c, x \rangle$ в таких псевдовершинах совпадают со значениями функции $\langle b, u \rangle$ в соответствующих вершинах множества U и в отличие от обычного симплекс-метода уже увеличиваются от итерации к итерации, приближаясь к оптимальному значению снизу.

Рассмотрим теперь двойственный симплекс-метод, предполагая, что двойственная задача (3.1.9) *невырожденная*, т.е. не вырожденны все угловые точки допустимого множества U. Тогда, как следует из утверждения 3.1.1, в каждой такой вершине u имеется ровно m активных

ограничений, т.е. среди n неравенств $A^T u \le c$ найдется ровно m, которые обращаются в равенство: $\langle a_i, u\rangle = c^i$. При этом обязательно векторы a_i линейно независимы. Данные векторы a_i, являясь столбцами матрицы A, образуют базис точки u, называемый *двойственным базисом*. Будем по-прежнему неособую матрицу, составленную из данных столбцов, обозначать через B, а подматрицу матрицы A, в которую входят оставшиеся столбцы, — через N. Считаем также для определенности, что двойственный базис B состоит из первых m столбцов, а для матрицы A и векторов x и c имеет место разбиение (3.1.4). Аналогичным образом разобьем вектор v, положив $v = [v^B, v^N]^T$. Имеем с учетом невырожденности точки u: $v^B = 0_m$, $v^N > 0_{n-m}$. Полагаем дополнительно $x^N = 0_{n-m}$. Для $u \in U$ обозначим также через $J_B(u)$ и $J_N(u)$ множества индексов базисных и небазисных компонент вектора $v(u)$, т.е. $J_B(u) = \{i \in J^n :\ v^i(u) = 0\}$, $J_N(u) = \{i \in J^n :\ v^i(u) > 0\}$.

Пусть несколько условно a_s^{-1} обозначает s-ю строку матрицы B^{-1} (s-й столбец матрицы $(B^{-1})^T$) и пусть $\theta \ge 0$. Рассмотрим m-мерные векторы

$$u_s(\theta) = u - \theta a_s^{-1}, \qquad s \in J_B(u),$$

и вычислим $x^B = B^{-1}b$.

Основной шаг двойственного симплекс-метода также можно разбить на три этапа.

1. Проверка оптимальности. Если $x^B \ge 0$, то точка $u \in U$ является оптимальным решением двойственной задачи (3.1.9), а точка $x = [x^B, x^N]^T$ — оптимальным решением исходной задачи (3.1.2). Это следует, в частности, из того, что для этих x и u выполнены все условия (3.1.17).

2. Выбор столбца для вывода из базиса. Берем произвольный индекс $s \in J_B(u)$, для которого $x^s < 0$. Если при этом окажется, что $(B^{-1}a_j)^s \ge 0$ для всех $j \in J_N(u)$, то допустимое множество U является неограниченным. Действительно, условие $(B^{-1}a_j)^s \ge 0$ переписывается как $\langle a_s^{-1}, a_j\rangle \ge 0$. Отсюда следует выполнение неравенства $N^T a_s^{-1} \ge 0_{n-m}$, используя которое получаем

$$v^N(u_s(\theta)) = c^N - N^T u + \theta N^T a_s^{-1} = \\ = v^N(u) + \theta N^T a_s^{-1} > \theta N^T a_s^{-1} \ge 0_{n-m}.$$

Для вектора $v^B(u_s(\theta))$ имеем соответственно

$$v^B(u_s(\theta)) = c^B - B^T u_s(\theta) = c^B - B^T u + \theta B^T a_s^{-1} = \theta e_s \ge 0_m,$$

где e_s — s-й единичный орт. Кроме того,

$$\langle b, u_s(\theta)\rangle = \langle b, u\rangle - \theta\langle b, a_s^{-1}\rangle = \langle b, u\rangle - \theta x^s \to +\infty$$

при $\theta \to +\infty$. Отсюда делаем вывод, что в двойственной задаче оптимальное значение целевой функции равно $+\infty$, следовательно, исходная задача (3.1.2) не имеет решения (множество X пустое). Далее предполагаем, что данный случай не реализуется.

3. Выбор столбца для ввода в базис. Пусть взят некоторый индекс $s \in J_B(u)$, для которого $x^s < 0$. Обозначим через $J_N^-(u)$ индексное множество $J_N^-(u) = \{j \in J_N(u) : \ (B^{-1}a_j)^s < 0\}$ и выберем из множества $J_N^-(u)$ такой индекс k, что

$$\theta_* = \min_{j \in J_N^-(u)} \frac{-v^j}{(B^{-1}a_j)^s} = \frac{-v^k}{(B^{-1}a_k)^s}.$$

После этого переходим в новую точку: $\bar{u} = u_s(\theta_*)$. Покажем, что $\bar{u}$ будет допустимой угловой точкой множества U. Действительно, для компоненты v^B имеем: $v^B(\bar{u}) = c^B - B^T u + \theta_* e_s = \theta_* e_s$. Помимо того, для всех $j \in J_N(u)$ выполняется неравенство

$$v^j(\bar{u}) = c^j - \langle a_j, u\rangle + \theta_* \langle a_j, a_s^{-1}\rangle = v^j + \theta_*(B^{-1}a_j)^s \geq 0,$$

причем для индекса $k \in J_N(u)$ оно переходит в равенство $v^k(\bar{u}) = 0$. Из-за предположения невырожденности следует, что данный индекс k может быть только единственным.

Двойственный базис $\bar{B}$, соответствующий новой точке $\bar{u}$, получается путем замены столбца a_s столбцом a_k матрицы A. При переходе в точку $\bar{u}$ значение целевой функции в двойственной задаче увеличивается:

$$\langle b, \bar{u}\rangle = \langle b, u\rangle - \theta_* \langle b, a_s^{-1}\rangle = \langle b, u\rangle - \theta_* x^s > \langle b, u\rangle.$$

Поскольку при переходе из одной угловой точки в другую угловую точку значение целевой функции увеличивается, а число угловых точек у множества U конечно, то в результате своей работы метод найдет решение двойственной задачи и, следовательно, решение прямой задачи.

3.2. Методы квадратичного программирования

Задачи квадратичного программирования представляют собой более сложный класс задач по сравнению с задачами линейного программирования. Даже рассматривая задачи квадратичного программирования с линейными ограничениями, уже нельзя утверждать, что среди решений задачи имеется решение, являющееся угловой точкой

допустимого множества. Тем не менее специфика этих задач дает возможность построить ряд алгоритмов нахождения их решения, причем за конечное число итераций. Это очень важно, так как довольно часто задачи квадратичного программирования используются в качестве вспомогательных подзадач в других более общих методах решения задач условной оптимизации.

Рассмотрим задачу квадратичного программирования с линейными ограничениями типа неравенства

$$\begin{array}{c} \min \ \frac{1}{2}\langle x, Cx\rangle + \langle d, x\rangle, \\ Ax \le b, \end{array} \tag{3.2.1}$$

где $d \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ и C — симметричная положительно определенная матрица порядка n. Матрица A со строками a_i имеет размер $m \times n$. Так как целевая функция в задаче (3.2.1) сильно выпуклая, то решение этой задачи существует и единственно, если, разумеется, допустимое множество $X = \left\{x \in \mathbb{R}^n : \ Ax \le b\right\}$ в ней не пустое.

Метод особых точек. Данный метод является одним из наиболее простых, по крайней мере, с идейной точки зрения, конечных методов решения задачи (3.2.1). Существует много различных конкретных способов реализации этого метода. Приведем здесь описание лишь одного из них.

Введем в рассмотрение наряду с допустимым множеством X множества вида

$$X(J) = \left\{x \in \mathbb{R}^n : \ \langle a_i, x\rangle = b^i, \ i \in J\right\},$$

где J — произвольное множество индексов из $J^m = [1:m]$, которое, в частности, может быть пустым. Если J — пустое множество, то $X(J)$ совпадает со всем пространством $\mathbb{R}^n$.

Точка $\bar{x} \in X$ называется *особой точкой* задачи (3.2.1), если она одновременно есть решение задачи

$$\min_{x \in X(J)} f(x), \qquad f(x) = \frac{1}{2}\langle x, Cx\rangle + \langle d, x\rangle, \tag{3.2.2}$$

при некотором индексном множестве $J \subseteq J^m$. Так как целевая функция в любой задаче вида (3.2.2) является сильно выпуклой, а число всевозможных индексных подмножеств из J^m конечно, то число особых точек в задаче (3.2.1) конечно.

Важный результат состоит в том, что решение задачи квадратичного программирования (3.2.1) всегда является ее особой точкой. Действительно, пусть $J_0(x)$ — множество индексов активных ограничений

задачи (3.2.1) в точке x, т.е. $J_0(x) = \left\{i \in J^m : \ \langle a_i, x\rangle = b^i\right\}$. Если точка x_* есть решение задачи (3.2.1), то она является локальным решением задачи

$$\min_{x\in\tilde{X}} f(x), \qquad \tilde{X} = \left\{x \in \mathbb{R}^n : \ \langle a_i, x\rangle \le b^i, \ i \in J_0(x_*)\right\}. \tag{3.2.3}$$

Но (3.2.3) — задача выпуклого программирования, поэтому ее локальное решение есть и глобальное решение. Отсюда также следует, что x_* будет и глобальным решением задачи (3.2.2), в которой $J = J_0(x_*)$. Поэтому x_* — особая точка задачи квадратичного программирования (3.2.1).

Таким образом, чтобы решить задачу (3.2.1), достаточно перебрать все ее особые точки, число которых, как уже отмечалось, конечно. В методе особых точек осуществляется целенаправленный перебор особых точек, причем значение целевой функции $f(x)$ уменьшается от итерации к итерации. В этом методе строится последовательность допустимых точек $x_k \in X$, и каждой такой точке ставится в соответствие индексное множество $J_k \subseteq J_0(x_k)$, т.е. выделяется некоторое множество ограничений-равенств из числа активных ограничений в этой точке.

Предположим, что известна начальная точка $x_0 \in X$ и что пришлось проводить последующие итерации. Пусть у нас на k-й итерации имеется особая точка $x_k \in X$, не являющаяся решением задачи (3.2.1). Пусть, кроме того, имеется индексное множество $J_k \subseteq J_0(x_k)$. Опишем переход к новой точке x_{k+1}.

Решаем задачу с ограничениями-равенствами:

$$\min_{x\in X(J_k)} f(x). \tag{3.2.4}$$

Данную задачу можно решить за конечное число итераций, переходя, например, к двойственной задаче. Обозначим через A_k подматрицу матрицы A, составленную из строк A с номерами из индексного множества J_k. Обозначим также через b_k соответствующий подвектор вектора b. Пусть n_k — количество индексов в множестве J_k. Если составить функцию Лагранжа:

$$L(x, u) = \frac{1}{2}\langle x, Cx\rangle + \langle d, x\rangle + \langle u, A_k x - b_k\rangle,$$

где $u \in \mathbb{R}^{n_k}$, то точка минимума этой функции по x в силу необходимых условий удовлетворяет системе уравнений $Cx + d + A_k^T u = 0$. Разрешая данное уравнение относительно x, получаем

$$x = x(u) = -C^{-1}\left(A_k^T u + d\right). \tag{3.2.5}$$

После подстановки найденного $x(u)$ в функцию Лагранжа приходим к двойственной функции:

$$\tilde{\phi}(u) = \tfrac{1}{2}\langle \left(A_k^T u + d\right), C^{-1}\left(A_k^T u + d\right)\rangle - \langle d, C^{-1}\left(A_k^T u + d\right)\rangle - \\ -\langle u, A_k C^{-1}\left(A_k^T u + d\right) + b\rangle.$$

Проведем группировку соответствующих слагаемых и поменяем знак у двойственной функции. Тогда она перейдет в функцию

$$\phi(u) = \frac{1}{2}\langle u, G_k u\rangle + \langle b, u\rangle - c, \quad G_k = A_k C^{-1} A_k^T, \quad c = \frac{1}{2}\langle d, C^{-1} d\rangle.$$

Данная функция $\phi(u)$ является выпуклой и квадратичной. Следовательно, решение задачи

$$\min_{u \in R^{n_k}} \phi(u) \tag{3.2.6}$$

может быть найдено за конечное число шагов, например с помощью метода сопряженных градиентов.

Если строки матрицы A_k линейно независимы, то симметричная матрица G_k оказывается положительно определенной, а функция $\phi(u)$ сильно выпуклой, решение задачи (3.2.6) существует и единственно. Пусть $\bar{u}_k$ — решение (3.2.6) и пусть $\bar{x}_k = x(\bar{u}_k)$ — соответствующее решение задачи (3.2.4), вычисленное по формуле (3.2.5). Возможны три случая.

Случай 1. Выполняется $\bar{x}_k \in X$ и $\bar{u}_k \geq 0_{n_k}$. Тогда точка $\bar{x}_k$, будучи особой, является решением поставленной задачи (3.2.1). В самом деле, в этом случае точка u_* из $\mathbb{R}^m_+$, в которой компоненты u_*^i совпадают с $\bar{u}^i$, когда $i \in J_k$, а остальные компоненты нулевые, является допустимой в двойственной задаче

$$\max_{u \in R^m_+} \phi(u). \tag{3.2.7}$$

В (3.2.7) двойственная целевая функция $\phi(u)$ соответствует уже всей задаче (3.2.1). Кроме того, значения целевых функций в прямой задаче (3.2.1) и в двойственной задаче (3.2.7) совпадают. Поэтому допустимая точка $\bar{x}$ не только особая, но и является решением (3.2.1).

Случай 2. Точка $\bar{x}_k$ особая, т.е. $\bar{x}_k \in X$, однако $\bar{u}_k \ngeq 0_{n_k}$. Полагаем $x_{k+1} = \bar{x}_k$ и $J_{k+1} = \{i \in J_k : \ \bar{u}_k^i > 0\}$. Имеет место неравенство $f(x_{k+1}) < f(x_k)$.

Случай 3. Точка $\bar{x}_k$ не является особой, т.е. она не принадлежит множеству X. В этом случае перейдем к другой точке x_{k+1}, положив

$$x_{k+1} = x_k + \alpha_k\left(\bar{x}_k - x_k\right),$$

где

$$\alpha_k = \max\{\alpha \geq 0 : \ x_k + \alpha(\bar{x}_k - x_k) \in X\}.$$

В множество индексов J_{k+1} включим все индексы из J_k, а также дополнительно все индексы активных ограничений из $J_0(x_{k+1})$, т.е. положим $J_{k+1} = J_k \cup J_0(x_{k+1})$. Таким образом, J_k является собственным подмножеством J_{k+1}, количество индексов в J_{k+1} хотя бы на единицу больше, чем в J_k.

Из-за того, что $\bar{x}_k$ — точка минимума на множестве $X(J_k)$, а функция $f(x)$ — сильно выпуклая, получаем, что $f(x_{k+1}) < f(x_k)$.

При реализации метода встает вопрос, как находить начальную точку $x_0 \in X$ и соответствующее множество J_0. Один из наиболее очевидных подходов состоит в том, чтобы решить вспомогательную задачу линейного программирования, аналогичную той, которая решается в методе искусственного базиса в линейном программировании:

$$\begin{gathered} \min \langle \bar{e}, y \rangle, \\ Ax - y \leq b, \\ y \geq 0_m. \end{gathered}$$

Если в решении $[x_*, y_*]$ этой задачи оказывается, что $y_* = 0_m$, то $x_* \in X$. Тогда полагаем: $x_0 = x_*$, $J_0 = J_0(x_0)$. В случае, когда $y_* \geq 0_m$, причем $y_* \neq 0_m$, получаем, что допустимое множество в исходной задаче квадратичного программирования пусто.

Упростить реализацию метода особых точек для решения задачи квадратичного программирования (3.2.1) можно путем перехода к двойственной задаче. Если отбросить константу в двойственной целевой функции и перейти от задачи на максимум к задаче на минимум, то она записывается в виде

$$\min_{u \in R^m_+} \phi(u), \qquad \phi(u) = \frac{1}{2}\langle u, AC^{-1}A^T u \rangle + \langle b, u \rangle. \tag{3.2.8}$$

Для решения задачи (3.2.8) можно также применять метод особых точек, причем теперь допустимое множество в (3.2.8) задается простыми ограничениями, а именно: $u \geq 0_m$. Если будет найдено решение этой задачи — точка u_*, то решение исходной задачи квадратичного программирования легко определяется по формуле, аналогичной (3.2.5) и имеющей вид:

$$x_* = x(u_*) = -C^{-1}\left(Au_* + d\right).$$

Заметим, что матрица $AC^{-1}A^T$ в целевой функции двойственной задачи (3.2.8) будет положительно определенной только тогда, когда

$\mathrm{rank}A = m$. В общем случае она только положительно полуопределенная. Однако существуют модификации метода особых точек, которые решают задачи с выпуклыми квадратичными функциями, а не обязательно с сильно выпуклыми.

Метод Вулфа. Для решения задач квадратичного программирования можно применять и технику перехода от одних допустимых базисных решений к другим базисным решениям аналогично тому, как это делается в симплекс-методе решения задач линейного программирования. Основой для таких методов служат условия Куна–Таккера, которые представляют собой ряд линейных равенств и дополнительное билинейное условие комплементарности.

Рассмотрим задачу квадратичного программирования, в которой ограничения заданы в канонической форме:

$$\begin{array}{c} \min\ \frac{1}{2}\langle x, Cx\rangle - \langle d, x\rangle, \\ Ax = b, \quad x \geq 0_n, \end{array} \tag{3.2.9}$$

где C — симметричная положительно определенная матрица, а относительно вектора b, не умаляя общности, предполагаем, что $b \geq 0_m$.

Составим для для задачи (3.2.9) функцию Лагранжа:

$$L(x, u, v) = \frac{1}{2}\langle x, Cx\rangle - \langle d, x\rangle + \langle Ax - b, u\rangle - \langle x, v\rangle,$$

где $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ и $v \in \mathbb{R}^n_+$. Необходимые и достаточные условия оптимальности для задачи (3.2.9) имеют вид

$$\begin{array}{rcl} L_x(x, u, v) & = & Cx - d + A^T u - v = 0_n, \\ L_u(x, u, v) & = & Ax - b = 0_m, \\ L_v(x, u, v) & = & -x \leq 0_n, \quad v \geq 0_n, \\ \langle L_v(x, u, v), v\rangle & = & D(v)x = D(x)v = 0_n. \end{array}$$

Здесь $D(v)$ — диагональная матрица с вектором v на диагонали. Таким образом, у нас есть $2n+m$ переменных и $n+m$ линейных ограничений типа равенства. Выпишем отдельно равенства и неравенства:

$$\begin{array}{c} Cx + A^T u - v = d, \\ Ax = b, \quad x \geq 0_n, \quad v \geq 0_n, \end{array} \tag{3.2.10}$$

и условия комплементарности:

$$x^i v^i = 0, \quad 1 \leq i \leq n. \tag{3.2.11}$$

Метод Вулфа основан на простой идее: вводят дополнительные переменные, причем таким образом, чтобы начальное базисное решение

было известно и при этом для основных переменных x и v условие комплементарности (3.2.11) выполняется. Решаются последовательно две вспомогательные задачи с целью приведения невязок (определяемых дополнительными компонентами) к нулю. Решения осуществляются путем перехода от одного базисного решения к другому, но таким образом, чтобы обеспечивалось выполнение условия комплементарности (3.2.11).

Предположим для простоты, что $d \geq 0_n$, и введем $m+n$ дополнительные неотрицательные переменные: $y \geq 0_m$ и $z \geq 0_n$. Перепишем теперь систему (3.2.10) в расширенном виде:

$$\begin{array}{rcl} Cx + A^T u - v + z & = & d, \\ Ax + y & = & b, \\ x \geq 0_n,\ v \geq 0_n, & & y \geq 0_m,\ z \geq 0_n. \end{array} \tag{3.2.12}$$

В качестве начального допустимого базисного решения тогда может быть взят вектор $[x, u, v, y, z]^T$ со следующими компонентами:

$$x = 0_n, \quad u = 0_m, \quad v = 0_n, \quad y = b, \quad z = d. \tag{3.2.13}$$

Если $b > 0_m$ и $d > 0_n$, то такое решение будет невырожденным.

Далее стараемся сделать дополнительные переменные y и z нулевыми, или, другими словами, удалить их из системы условий (3.2.12). Делается это в два этапа путем решения симплекс-методом вспомогательных задач линейного программирования. На первом этапе с целью удаления переменной y решается задача линейного программирования:

$$\begin{array}{rcl} \langle \bar{e}, y \rangle & \to & \min, \\ Cx + z & = & d, \\ Ax + y & = & b, \\ x \geq 0_n, \quad y \geq 0_m, & & z \geq 0_n, \end{array} \tag{3.2.14}$$

где, как и ранее, $\bar{e}$ — вектор, состоящий из единиц. Таким образом, на первом этапе стараются не только удалить переменную y, но и одновременно держать в системе условий (3.2.12) переменные u и v нулевыми. При этом, разумеется, условия комплементарности (3.2.11) полностью выполняются. В качестве начального допустимого базисного решения берется точка $[x, y, z]$ с компонентами из (3.2.13).

Если ограничения в задаче (3.2.9) совместные, то минимум целевой функции $\langle \bar{e}, y \rangle$ во вспомогательной задаче (3.2.14) равен нулю, и ее решением является вектор $y = 0_m$. После этого приступаем ко второму этапу. На втором этапе переменная y исключается из условий (3.2.12)

и решается вспомогательная задача линейного программирования:

$$\begin{gathered} \langle \bar{e}, z \rangle \to \min, \\ Cx + A^T u + z - v = d, \qquad Ax = b, \\ x \geq 0_n, \quad v \geq 0_n, \quad z \geq 0_n. \end{gathered} \tag{3.2.15}$$

В качестве начального допустимого базисного решения берется то решение, которое было получено на первом этапе.

При этом, чтобы соблюсти условие комплементарности (3.2.11), вводят *дополнительное правило*. Согласно этому правилу, если компонента x^i является базисной, то при переходе к новому базисному решению компонента v^i не должна становиться базисной. И наоборот, если v^i — базисная компонента, то x^i не может стать базисной.

Если в результате решения вспомогательной задачи (3.2.15) будет получено решение, $[x_*, u_*, v_*, z_*]$, в котором $z_* = 0_n$, то тем самым будет получено решение исходной задачи квадратичного программирования (3.2.9). Можно показать, что в случае положительно определенной матрицы C, всегда можно добиться нулевого решения $z_* = 0_n$.

Отметим также, что предположение $d \geq 0_n$ не является обременительным и легко может быть обойденным с помощью использования вместо одного вектора z двух векторов $z_1 \geq 0_n$ и $z_2 \geq 0_n$. В этом случае при задании начального допустимого базисного решения следует полагать $z_1^i = d^i$, $z_2^i = 0$, если $d^i \geq 0$. И наоборот, $z_1^i = 0$, $z_2^i = -d^i$, когда $d^i < 0$. Перед вторым этапом у нас только n компонент из $2n$ компонент векторов z_1 и z_2 могут оказаться базисными. Именно эти компоненты и учитываются далее, а остальные исключаются из рассмотрения.

Глава 4

Методы минимизации на множествах простого вида

Перейдем к рассмотрению других методов условной оптимизации, которые предназначены для задач с нелинейными целевыми функциями. Начнем с так называемых *прямых методов*. Применение этого названия обусловлено тем обстоятельством, что вычисления в таких методах проводятся только в пространстве исходных (прямых) переменных, входящих в постановку задачи. Двойственные переменные (множители Лагранжа) в них явным образом не используются.

В данной главе мы в основном коснемся методов решения только тех задач, в которых допустимое множество имеет достаточно простой вид, по крайней мере оно выпукло и замкнуто. В ряде случаев это позволяет учитывать его как единое множество без конкретизации описания с помощью равенств или неравенств. Другой случай простого допустимого множества — это когда оно является аффинным и, следовательно, задается через решения системы линейных уравнений. Поэтому появляется возможность понизить число переменных в задаче, выразив одни переменные через другие.

4.1. Метод проекции градиента

Пусть требуется найти

$$f_* = \min_{x \in X} f(x), \tag{4.1.1}$$

где X — выпуклое замкнутое множество в $\mathbb{R}^n$, $f(x)$ — дифференцируемая функция, определенная на некоторой области, содержащей множество X. Про такую задачу иногда говорят, что она является задачей с *прямым ограничением*, подчеркивая тем самым, что множество X здесь рассматривается как единое целое без уточнения его функционального описания.

Обозначим через $X_* \subseteq X$ множество оптимальных решений в задаче (4.1.1). Если $x_* \in X_*$, то в силу необходимых условий минимума дифференцируемой функции на выпуклом множестве данная точка x_* должна быть *стационарной*, т.е. быть решением вариационного неравенства с градиентным отображением $f_x(x)$:

$$\langle f_x(x_*), x - x_* \rangle \geq 0 \quad \forall\, x \in X. \tag{4.1.2}$$

Если $f(x)$ — выпуклая функция, то эти условия являются и достаточными.

В задаче безусловной минимизации, когда $X = \mathbb{R}^n$, одним из основных методов решения был метод градиентного спуска, в котором строилась последовательность точек $\{x_k\}$ согласно следующей рекуррентной схеме:

$$x_{k+1} = x_k - \alpha_k f_x(x_k), \quad k = 0, 1, \ldots, \tag{4.1.3}$$

где α_k — шаг спуска, определяемый по какому-либо правилу точной или приближенной одномерной минимизации. Однако если X отлично от всего пространства $\mathbb{R}^n$, то непосредственное применение градиентного метода затруднено, например из-за возможного выхода точек x_k за пределы множества X.

Рассмотрим теперь метод решения задачи (4.1.1), в котором все точки итерационного процесса принадлежат множеству X. В нем сочетаются идеи метода градиентного спуска и проектирования, поэтому он и получил название *метода проекции градиента*.

В методе проекции градиента, как и в методе градиентного спуска, начальная точка x_0 задается, причем $x_0 \in X$, а последующие точки вместо (4.1.3) вычисляются по рекуррентной схеме:

$$x_{k+1} = \pi_X \left(x_k - \alpha_k f_x(x_k) \right), \qquad k = 0, 1, \ldots, \tag{4.1.4}$$

где $\pi_X(a)$ — проекции точки a на множество X. Таким образом, вся последовательность $\{x_k\}$ оказывается принадлежащей допустимому множеству X (см. рис. 4.1). Шаг α_k в (4.1.4) выбирается по одной из модификаций процедур одномерного поиска, обсуждаемых ниже.

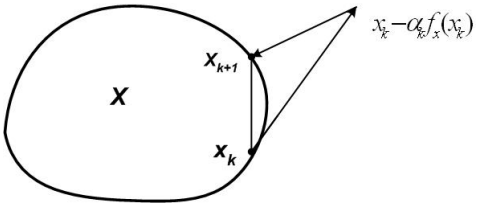

Рис. 4.1. Итерация в методе проекции градиента

Утверждение 4.1.1. *Пусть x_* — решение задачи (4.1.1), в которой X — выпуклое замкнутое множество. Тогда*

$$x_* = \pi_X \left(x_* - \alpha f_x(x_*) \right) \tag{4.1.5}$$

для любого $\alpha > 0$.

Доказательство. На основании необходимых условий минимума дифференцируемой функции на выпуклом замкнутом множестве X, если $x_* \in X$ — решение задачи, то эта точка должна быть стационарной, т.е. для нее выполняется неравенство (4.1.2).

Умножим неравенство (4.1.2) на $\alpha > 0$ и перепишем в виде

$$\langle x_* - (x_* - \alpha f_x(x_*)), x - x_* \rangle \geq 0 \quad \forall\, x \in X, \quad \forall\, \alpha > 0.$$

Отсюда на основании основного неравенства, справедливого для проекции точки на выпуклое множество (см. [33]), заключаем, что x_* — проекция точки $x_* - \alpha f_x(x_*)$ на множество X. Таким образом, имеет место равенство (4.1.5). ■

Заметим, что если $f(x)$ — выпуклая дифференцируемая функция, то выполнение равенства (4.1.5) является не только необходимым, но и достаточным для того, чтобы точка $x_* \in X$ была решением задачи (4.1.1). Итерационный процесс (4.1.4) есть не что иное, как метод простой итерации для решения уравнения (4.1.5).

Обсудим вопрос о выборе шага α_k в итерационном процессе (4.1.4). Возможно несколько подходов, упомянем только три из них.

1. *Правило постоянного шага.* Согласно этому правилу шаг α_k выбирается заранее, равным достаточно малой положительной величине.

2. *Правило одномерной минимизации.* Обозначим

$$\phi_k(\alpha) = f\left(x_k(\alpha)\right), \quad x_k(\alpha) = \pi_X \left(x_k - \alpha f_x(x_k) \right).$$

Согласно данному правилу шаг α_k является решением задачи

$$\phi_k(\alpha_k) = \min_{\alpha \geq 0} \phi_k(\alpha).$$

3. *Правило приближенной минимизации* на отрезке $[0, \hat{\alpha}]$, где $\hat{\alpha} > 0$ — некоторый заданный максимально возможный шаг. Для приближенной минимизации может применяться правило Армихо, в котором путем дробления начального шага $\alpha = \hat{\alpha}$ добиваются выполнения неравенства

$$f\left(x_k(\alpha)\right) - f(x_k) \leq \varepsilon \langle f_x(x_k), x_k(\alpha) - x_k \rangle, \tag{4.1.6}$$

где $0 < \varepsilon < 1$. Если неравенство (4.1.6) при данном α не выполняется, то его уменьшают, умножая на коэффициент $0 < \theta < 1$.

Предположим, что градиент функции $f(x)$ удовлетворяет условию Липшица:

$$\|f_x(x) - f_x(y)\| \leq L\|x - y\| \quad \forall\, x, y \in X, \tag{4.1.7}$$

и рассмотрим вопрос о сходимости итерационного процесса (4.1.4) в том случае, когда шаг α_k берется постоянным и ограниченным сверху:

$$\alpha_k = \alpha, \qquad 0 < \alpha < \frac{1}{L + c}, \tag{4.1.8}$$

где $c > 0$ — любое положительное число.

Теорема 4.1.1. *Пусть X — выпуклое замкнутое множество и $f(x)$ — дифференцируемая функция, градиент которой удовлетворяет условию Липшица (4.1.7). Пусть, кроме того, начальная точка $x_0 \in X$ такова, что множество $\tilde{X}(x_0) = \{x \in X : \ f(x) \leq f(x_0)\}$ компактное, а шаг α_k берется постоянным и таким, что выполняется неравенство (4.1.8). Тогда последовательность $\{x_k\}$, генерируемая методом проекции градиента, имеет предельные точки, которые являются стационарными точками в задаче (4.1.1). Если $f(x)$ — выпуклая функция, то каждая стационарная точка x_* есть решение задачи (4.1.1).*

Доказательство. Для проекции $\bar{a} \in X$ точки $a \in \mathbb{R}^n$ на множество X выполняется неравенство (см. [33])

$$\langle \bar{a} - a, x - \bar{a} \rangle \geq 0 \quad \forall x \in X. \tag{4.1.9}$$

Беря в нем $\bar{a} = x_{k+1}$, $a = x_k - \alpha f_x(x_k)$, $x = x_k$, приходим к

$$\langle x_{k+1} - (x_k - \alpha f_x(x_k)), x_k - x_{k+1} \rangle \geq 0. \tag{4.1.10}$$

Обозначим $s_k = x_{k+1} - x_k$. Неравенство (4.1.10) при этих обозначениях перепишется как $\langle s_k + \alpha f_x(x_k), s_k \rangle \le 0$ или

$$\|s_k\|^2 + \alpha \langle f_x(x_k), s_k \rangle \le 0. \tag{4.1.11}$$

Если представить $f(x_{k+1}) = f(x_k) + \langle f_x(\tilde{x}_k), s_k \rangle$, где $\tilde{x}_k \in [x_k, x_{k+1}]$, то на основании (4.1.7) и неравенства Коши–Буняковского

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \langle f_x(x_k), s_k \rangle + \langle f_x(\tilde{x}_k) - f_x(x_k), s_k \rangle \le \\ &\le f(x_k) + \langle f_x(x_k), s_k \rangle + L\|\tilde{x}_k - x_k\|\|s_k\| \le \\ &\le f(x_k) + \langle f_x(x_k), s_k \rangle + L\|s_k\|^2 = \\ &= f(x_k) + \alpha^{-1} \left[\|s_k\|^2 + \alpha \langle f_x(x_k), s_k \rangle \right] + \left(L - \alpha^{-1} \right) \|s_k\|^2. \end{aligned}$$

Отсюда с учетом (4.1.11) и (4.1.8) получаем

$$f(x_{k+1}) - f(x_k) \le \left(L - \alpha^{-1} \right) \|s_k\|^2 \le -c\|s_k\|^2. \tag{4.1.12}$$

Неравенство (4.1.12) и ограниченность непрерывной функции $f(x)$ на компактном множестве $\tilde{X}(x_0)$ позволяют заключить, что

$$\lim_{k \to \infty} \left(f(x_{k+1}) - f(x_k) \right) = 0,$$

а также

$$\lim_{k \to \infty} \|s_k\| = \lim_{k \to \infty} \|x_{k+1} - x_k\| = 0.$$

Так как последовательность $\{x_k\}$ принадлежит компактному множеству $\tilde{X}(x_0)$, существует сходящаяся подпоследовательность $\{x_{k_j}\}$. Пусть $\lim_{j \to \infty} x_{k_j} = x_*$. Имеем для этой подпоследовательности

$$\|s_{k_j}\| = \|\pi_X \left(x_{k_j} - \alpha_{k_j} f_x(x_{k_j}) \right) - x_{k_j}\| \to 0.$$

Поэтому в силу непрерывности оператора проектирования справедливо предельное равенство $\pi_X \left(x_* - \alpha f_x(x_*) \right) = x_*$. Отсюда, поскольку x_* является проекцией точки $x_* - \alpha f_x(x_*)$ на множество X, согласно (4.1.9) должно выполняться неравенство

$$\langle x_* - (x_* - \alpha f_x(x_*)), x - x_* \rangle \ge 0 \quad \forall\, x \in X.$$

Таким образом,

$$\langle f_x(x_*), x - x_* \rangle \ge 0 \quad \forall\, x \in X, \tag{4.1.13}$$

т.е. точка x_* является стационарной.

Если $f(x)$ — выпуклая функция, то условие (4.1.13) достаточно для того, чтобы точка x_* была бы решением задачи (4.1.1). ■

Рассмотренный метод проекции градиента является *релаксационным*, так как все точки итерационного процесса (4.1.4) принадлежат допустимому множеству и целевая функция убывает от итерации к итерации. Укажем на некоторые дополнительные свойства процесса (4.1.4).

1. Пусть $f(x)$ — выпуклая дифференцируемая функция, градиент которой удовлетворяет условию Липшица (4.1.7). Пусть, кроме того, множество X ограничено. Тогда для любого x_k справедливо неравенство

$$f(x_k) - f_* \le \frac{C}{k}, \tag{4.1.14}$$

где $C > 0$.

Действительно, предположим, что шаг α_k на каждой итерации ограничен снизу: $\alpha_k > \alpha_*$. Такая ограниченность, конечно, имеет место при постоянном шаге. Обозначим $\Delta_k = f(x_k) - f_* \ge 0$. В силу выпуклости функции $f(x)$ по критерию первого порядка выполняется неравенство

$$\Delta_k = f(x_k) - f(x_*) \le \langle f_x(x_k), x_k - x_* \rangle,$$

которое перепишем в виде

$$\begin{aligned}\Delta_k \le \langle f_x(x_k), x_k - x_{k+1} \rangle - \alpha_k^{-1} \langle x_{k+1} - x_k + \alpha_k f_x(x_k), x_* - x_{k+1} \rangle + \\ + \alpha_k^{-1} \langle x_{k+1} - x_k, x_* - x_{k+1} \rangle.\end{aligned} \tag{4.1.15}$$

Но $x_{k+1} = \pi_X(x_k - \alpha_k f_x(x_k))$, $x_* \in X$. Следовательно, согласно основному неравенству (4.1.9), имеющему место для проекции на выпуклое замкнутое множество,

$$\langle x_{k+1} - x_k + \alpha_k f_x(x_k), x_* - x_{k+1} \rangle \ge 0.$$

Отсюда и из (4.1.15), используя дополнительно неравенство Коши–Буняковского, приходим к оценке

$$\Delta_k \le \beta_k \|x_{k+1} - x_k\|, \qquad \beta_k = \|f_x(x_k)\| + \alpha_*^{-1} \|x_* - x_{k+1}\|.$$

Так как по предположению множество X ограничено, то можно указать такую константу $c_1 > 0$, что $\|f_x(x)\| \le c_1$ для любых $x \in X$. Кроме того, $\|x_* - x\| \le d$ для этих x, где d — диаметр множества X. Поэтому все $\beta_k \le c_2$ для некоторого $c_2 > 0$. Следовательно, можно перейти к независимой от номера итерации оценке: $\Delta_k \le c_3 \|x_{k+1} - x_k\|$, в которой $c_3 > \mathrm{c}_1 + \alpha_*^{-1} \mathrm{c}_2$.

Далее, на каждой итерации согласно (4.1.12) выполняется неравенство

$$\Delta_{k+1} - \Delta_k \le -c \|x_{k+1} - x_k\|^2,$$

на основании которого получаем

$$\Delta_k - \Delta_{k+1} \geq c\|x_{k+1} - x_k\|^2 \geq \mathrm{c}_4 \Delta_k^2, \quad \mathrm{c}_4 = \frac{c}{c_3^2}. \tag{4.1.16}$$

Если все $\Delta_k > 0$, то после деления обеих частей неравенства (4.1.16) на $\Delta_k \Delta_{k+1} > 0$ приходим к неравенству

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \geq c_4 \frac{\Delta_k}{\Delta_{k+1}}.$$

Просуммировав эти неравенства от начального $k = 1$ до текущего $k-1$, получаем с учетом того, что $\Delta_{k+1} < \Delta_k$,

$$\frac{1}{\Delta_k} \geq \frac{1}{\Delta_1} + c_4 \sum_{i=1}^{k-1} \frac{\Delta_i}{\Delta_{i+1}} \geq \frac{1}{\Delta_1} + c_4(k-1).$$

Таким образом,

$$\Delta_k \leq \left[c_4(k-1) + \Delta_1^{-1}\right]^{-1} \leq c_4^{-1}(k-1)^{-1} \leq \frac{2}{c_4 k},$$

когда $k > 1$. Отсюда делаем вывод, что справедлива оценка (4.1.14), в которой $C = \frac{2}{c_4}$. Следует сразу отметить, что данная оценка не является хорошей — даже при довольно сильных предположениях о задаче скорость сходимости по функционалу оказывается низкой.

2. Предположим теперь, что $f(x)$ — дифференцируемая сильно выпуклая функция с константой θ и x_* — ее точка минимума на множестве X. У сильно выпуклых функций такая точка x_* обязательно существует, причем единственная. Предположим также, что градиент $f(x)$ удовлетворяет условию Липшица (4.1.7) и что шаг α_k берется постоянным из интервала $(0, \frac{4\theta}{L^2})$. Тогда

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|, \tag{4.1.17}$$

где $0 < C < 1$. В самом деле, пусть $\alpha_k = \alpha$, где $\alpha \in (0, \frac{4\theta}{L^2})$. Так как оператор проектирования является нерастягивающим, то, используя утверждение 4.1.1 и неравенство

$$\langle f_x(x) - f_x(y), x - y\rangle \geq 2\theta\|x - y\|^2 \quad \forall\, x, y \in \mathbb{R}^n,$$

справедливое для сильно выпуклых функций (см. [33]), получаем

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|\pi_X(x_k - \alpha f_x(x_k)) - \pi_X(x_* - \alpha f_x(x_*))\|^2 \leq \\ &\leq \|x_k - \alpha f_x(x_k) - x_* + \alpha f_x(x_*)\|^2 = \\ &= \|x_k - x_*\|^2 + \alpha^2\|f_x(x_k) - f_x(x_*)\|^2 - \\ &\qquad -2\alpha\langle f_x(x_k) - f_x(x_*), x_k - x_*\rangle \leq \\ &\leq (1 + \alpha^2 L^2 - 4\theta\alpha)\|x_k - x_*\|^2. \end{aligned} \tag{4.1.18}$$

Поэтому имеет место (4.1.17) при $C = \sqrt{1 + \alpha^2 L^2 - 4\theta\alpha}$. Неравенство (4.1.17) означает, что метод проекции градиента с постоянным шагом сходится к точке минимума сильно выпуклой функции на X с линейной скоростью (уже по аргументу).

3. Если дополнительно предположить, что x_* — точка острого минимума функции $f(x)$ на X, т.е. для некоторого $\gamma > 0$ выполняется

$$f(x) - f(x_*) \geq \gamma \|x - x_*\| \quad \forall\, x \in X, \tag{4.1.19}$$

то метод проекции градиента (4.1.4) с постоянным шагом находит решение задачи за конечное число шагов. Действительно, для выпуклой на выпуклом множестве X дифференцируемой функции $f(x)$ неравенство (4.1.19) эквивалентно

$$\langle f_x(x_*), x - x_* \rangle \geq \gamma \|x - x_*\| \qquad \forall\, x \in X.$$

Тогда, беря любое $x \in X$, имеем

$$\begin{aligned}
&\langle x_k - \alpha f_x(x_k) - x_*, x - x_* \rangle = \\
&\quad = \langle x_k - x_* - \alpha(f_x(x_k) - f_x(x_*)), x - x_* \rangle - \alpha \langle f_x(x_*), x - x_* \rangle \leq \\
&\quad \leq (1 + \alpha L)\|x_k - x_*\| \|x - x_*\| - \alpha\gamma \|x - x_*\| = \\
&\quad = [(1 + \alpha L)\|x_k - x_*\| - \alpha\gamma]\ \|x - x_*\| \leq 0,
\end{aligned}$$

когда $\|x_k - x_*\| \leq \frac{\alpha\gamma}{(1+\alpha L)}$. Так как согласно (4.1.18) $x_k \to x_*$, отсюда получаем

$$\langle x_* - (x_k - \alpha f_x(x_k)), x - x_* \rangle \geq 0 \quad \forall\, x \in X$$

для k достаточно больших. Но в силу (4.1.9) данное неравенство означает, что x_* есть проекция вектора $x_k - \alpha f_x(x_k)$ на множество X, т.е. точка x_{k+1} совпадает с решением задачи — точкой x_*.

Чтобы уменьшить количество проектирований, наряду с основной схемой метода проекции градиента рассматривают ее модификацию:

$$x_{k+1} = x_k + \lambda_k s_k, \quad s_k = \pi_X\left(x_k - \alpha_k f_x(x_k)\right) - x_k. \tag{4.1.20}$$

Здесь α_k то же самое, что и в основной схеме, а λ_k выбирается из условия

$$\lambda_k = \arg \min_{0 \leq \lambda \leq 1} f\left(x_k + \lambda s_k\right)$$

или из условия

$$\lambda_k = \arg \min_{0 \leq \lambda \leq \lambda_*} f\left(x_k + \lambda s_k\right),$$

где λ_* — максимальное значение параметра λ, для которого сдвинутая точка $x_k + \lambda s_k$ не выходит за пределы множества X. Схему (4.1.20) принято называть *схемой с ускорением*.

Метод проекции градиента применяют в основном для решения тех задач, в которых проектирование на множество X достаточно простое, например, это неотрицательный ортант пространства, параллелепипед, подпространство, полупространство и т.д. Если $X = \mathbb{R}^n_+$, то согласно (4.1.4) метод проекции градиента принимает совсем простой вид:

$$x_{k+1} = [x_k - \alpha_k f_x(x_k)]_+, \quad x_0 \in \mathbb{R}^n_+,$$

где a_+ — положительная срезка вектора $a \in \mathbb{R}^n$, т.е. вектор с компонентами $a^i_+ = \max[0, a^i]$, $1 \le i \le n$.

В случае задачи минимизации негладкой выпуклой функции $f(x)$ можно применять *метод проекции субградиента*, который строится аналогично методу проекции градиента, но с применением уже субградиентов:

$$x_{k+1} = \pi_X \left(x_k - \alpha_k \frac{s_k}{\|s_k\|} \right), \quad s_k \in \partial f(x_k).$$

Шаг α_k выбирают обычно по правилу, используемому в субградиентном методе.

Упражнение 2. *Пусть множество X есть линейное многообразие, задаваемое как множество решений неоднородной системы линейных уравнений: $Ax = b$, где A — $(m \times n)$-матрица, $b \in \mathbb{R}^m$. Считая, что $m < n$ и что матрица A полного ранга, опишите итерационный процесс метода проекции градиента для минимизации дифференцируемой функции $f(x)$ на этом множестве.*

4.2. Метод условного градиента и условный метод Ньютона

Рассмотрим другой прямой метод решения задачи (4.1.1), получивший название *метода условного градиента*. Он также относится к классу релаксационных методов и основан на идее линеаризации целевой функции.

Предположим, что в задаче (4.1.1) множество X является *выпуклым компактом*. Предполагаем также, что $f(x)$ — дифференцируемая функция. Итерационный процесс в методе описывается следующим рекуррентным соотношением:

$$x_{k+1} = x_k + \alpha_k s_k, \tag{4.2.1}$$

где $s_k = \bar{x}_k - x_k$ и точка $\bar{x}_k$ есть решение вспомогательной задачи

$$\min_{x \in X} \langle f_x(x_k), x - x_k \rangle. \tag{4.2.2}$$

Шаг α_k определяется из решения одномерной задачи минимизации

$$\alpha_k = \arg \min_{0 \le \alpha \le 1} f(x_k + \alpha s_k),$$

что соответствует наискорейшему спуску. Возможно также приближенное решение этой задачи, например с использованием правила Армихо. О направлении s_k говорят как об *условном антиградиенте* функции $f(x)$ в точке x_k (относительно множества X).

Обратим внимание, что вспомогательная задача (4.2.2) эквивалентна задаче минимизации линейного приближения целевой функции. В самом деле, если точка $\bar{x}_k$ есть решение (4.2.2), то эта же точка будет и решением задачи

$$\min_{x \in X} f_k(x), \quad f_k(x) = f(x_k) + \langle f_x(x_k), x - x_k \rangle, \tag{4.2.3}$$

как, впрочем, и точкой минимума функции $\langle f_x(x_k), x \rangle$ на X.

Определение 4.2.1. *Множество X называется строго выпуклым, если для любых отличных друг от друга точек $x_1 \in X$, $x_2 \in X$ и $0 < \lambda < 1$ выполняется включение $\lambda x_1 + (1 - \lambda) x_2 \in intX$.*

Упражнение 3. *Покажите, что в случае строго выпуклого множества X решение задачи (4.2.3) единственное, когда $f_x(x_k) \neq 0_n$.*

Обозначим через $\eta_k = \langle f_x(x_k), s_k \rangle$ минимальное значение целевой функции во вспомогательной задаче (4.2.2). Поскольку $x_k \in X$, обязательно $\eta_k \le 0$. Возможны два случая:

1) $\eta_k = 0$. Тогда $\langle f_x(x_k), x - x_k \rangle \ge 0 \quad x \in X$. Выполнение данного неравенства означает, что x_k является стационарной точкой в задаче (4.1.1).

2) $\eta_k < 0$. Тогда из $\langle f_x(x_k), s_k \rangle < 0$ следует, что возможно движение вдоль направления s_k с уменьшением значения целевой функции $f(x)$.

Предположим, что шаг α_k выбирается по правилу Армихо:

$$f(x_k + \alpha_k s_k) - f(x_k) \le \varepsilon \alpha_k \eta_k, \tag{4.2.4}$$

где $0 < \varepsilon < 1$, $0 < \alpha_k \le 1$.

Теорема 4.2.1. *Пусть X — выпуклый компакт и $f(x)$ — дифференцируемая функция, градиент которой удовлетворяет на X условию Липшица (4.1.7). Пусть, кроме того, шаг α_k выбирается путем дробления начального шага, равного единице, по правилу Армихо (4.2.4). Тогда для любого начального приближения $x_0 \in X$ метод условного градиента порождает последовательность точек $\{x_k\}$, которая имеет предельные точки, и любая предельная точка x_* является стационарной для задачи (4.1.1). Если $f(x)$ — выпуклая функция на X, то x_* — решение задачи (4.1.1).*

Доказательство. Считая, что $\eta_k < 0$, обозначим

$$\bar{\alpha}_k = \frac{2(\varepsilon - 1)\eta_k}{L\|s_k\|^2} > 0.$$

Аналогично тому, как это делается в методе градиентного спуска с правилом Армихо выбора шага α_k, можно показать, что неравенство (4.2.4) заведомо выполняется, когда $\alpha_k \leq \bar{\alpha}_k$. Поэтому, производя дробление шага, начиная с единичного шага, мы в принципе можем столкнуться с двумя случаями. В первом случае получаем $\alpha_k = 1$ (это соответствует тому, что неравенство (4.2.4) выполняется сразу без дробления). Во втором случае $\alpha_k < 1$, т.е. произведено хотя бы одно дробление. Но это возможно только при $\bar{\alpha}_k < 1$, и мы дроблением шага добиваемся выполнения условия (4.2.4). Но тогда обязательно $\alpha_k \geq \theta\bar{\alpha}_k$, где $0 < \theta < 1$ — параметр дробления шага в правиле Армихо. Таким образом, всегда

$$\min\{1, \theta\bar{\alpha}_k\} \leq \alpha_k \leq 1. \tag{4.2.5}$$

Пусть последовательность $\{x_k\}$ бесконечная. Имеем $\{x_k\} \subset X$ и $f(x_{k+1}) < f(x_k)$. Множество X — компактное, функция $f(x)$ — непрерывная. Поэтому существуют пределы:

$$\lim_{k\to\infty} [f(x_{k+1}) - f(x_k)] = 0, \quad \lim_{k\to\infty} (\alpha_k\eta_k) = 0. \tag{4.2.6}$$

Шаг α_k на каждой итерации удовлетворяет неравенству (4.2.5), в котором, подчеркнем, величина $\bar{\alpha}_k$ зависит от η_k и s_k.

Из компактности множества X следует, что можно указать константу $c > 0$, для которой $\|s_k\| \leq c$ на каждой итерации. Отсюда получаем, что $\bar{\alpha}_k \geq \frac{2(\varepsilon-1)\eta_k}{Lc^2}$. Обозначим

$$\delta = \frac{2\theta(1-\varepsilon)}{Lc^2} > 0.$$

Имеем либо $\alpha_k = 1$, либо в силу (4.2.5) $\alpha_k \geq -\delta\eta_k > 0$, поэтому из (4.2.6) следует предельное равенство

$$\lim_{k\to\infty} \eta_k = 0. \tag{4.2.7}$$

Так как η_k — минимальное значение функции $\langle f_x(x_k), x - x_k\rangle$ на X, то

$$\eta_k \leq \langle f_x(x_k), x - x_k\rangle \quad \forall\, x \in X. \tag{4.2.8}$$

Последовательность $\{x_k\}$ принадлежит компактному множеству X и, значит, имеет предельные точки. Пусть $x_{k_j} \to x_* \in X$. Для этих k_j согласно (4.2.8) выполняются неравенства

$$\eta_{k_j} \leq \langle f_x(x_{k_j}), x - x_{k_j}\rangle \quad \forall\, x \in X.$$

Отсюда, переходя к пределу, получаем с учетом (4.2.7)

$$\langle f_x(x_*), x - x_*\rangle \geq 0 \quad \forall\, x \in X, \tag{4.2.9}$$

т.е. x_* является стационарной точкой для задачи (4.1.1).

Если $f(x)$ — выпуклая функция, то условие (4.2.9) достаточно для того, чтобы точка x_* была решением задачи (4.1.1). ■

Сделаем несколько замечаний.

Замечание 1. *Утверждение теоремы сохранится, если вместо правила Армихо выбора шага α_k применять правило одномерной минимизации.*

Замечание 2. *Если $f(x)$ — выпуклая дифференцируемая на X функция, градиент которой удовлетворяет условию Липшица, то имеет место следующая оценка:*

$$f(x_k) - f_* \leq \frac{C}{k}, \qquad k = 1, 2, \ldots, \tag{4.2.10}$$

где $C > 0$.

К оценке (4.2.10) можно прийти, проводя примерно те же самые рассуждения, что при получении оценки (4.1.14) в методе проекции градиента. Пусть $x_* \in X_*$ и пусть по-прежнему $\Delta_k = f(x_k) - f_*$. Имеем в силу выпуклости функции $f(x)$:

$$0 \leq \Delta_k = f(x_k) - f(x_*) \leq \langle f_x(x_k), x_k - x_*\rangle. \tag{4.2.11}$$

Кроме того, поскольку точка $\bar{x}_k$ есть решение задачи (4.2.2), то

$$\eta_k = \langle f_x(x_k), s_k\rangle = \langle f_x(x_k), \bar{x}_k - x_k\rangle \leq \langle f_x(x_k), x_* - x_k\rangle.$$

Отсюда и из (4.2.11) следует неравенство

$$0 \le \Delta_k \le \langle f_x(x_k), x_k - \bar{x}_k \rangle = -\eta_k. \tag{4.2.12}$$

Обратимся теперь к неравенству (4.2.4), обращение которого дает

$$f(x_k) - f(x_{k+1}) \ge -\varepsilon \alpha_k \eta_k, \tag{4.2.13}$$

причем, как было получено, α_k либо равняется единице, если не производилось дробление начального шага, либо $\alpha_k \ge -\delta\eta_k$, когда шаг приходится дробить. Но $\eta_k \to 0$. Поэтому можно заключить, что найдется константа $c_1 > 0$ такая, что $\alpha_k \ge -c_1\eta_k$ на всех итерациях. Согласно (4.2.13) тогда опять на всех итерациях справедливо неравенство

$$f(x_k) - f(x_{k+1}) \ge c_2 \eta_k^2,$$

где $c_2 = \varepsilon c_1$. Данное неравенство с учетом (4.2.12) можно переписать в виде

$$\Delta_k - \Delta_{k+1} \ge c_2 \Delta_k^2. \tag{4.2.14}$$

Из (4.2.14), как было показано в предыдущем параграфе, может быть получено (4.2.10).

Таким образом, метод условного градиента обладает той же достаточно медленной скоростью сходимости по функционалу, что и метод проекции градиента.

Замечание 3. *Если x_* — точка острого минимума функции $f(x)$ на X, то метод условного градиента сходится за конечное число шагов.*

Понятно, что метод условного градиента целесообразно применять только для решения тех задач, в которых достаточно просто находить минимум линейной функции $\langle c, x \rangle$ на множестве X, например, если X есть некоторый шар в пространстве $\mathbb{R}^n$:

$$X = \left\{ x \in \mathbb{R}^n : \ \|x - a\| \le r \right\}.$$

Тогда, решая задачу минимизации функции $\langle c, x \rangle$ на множестве X, получаем, что ее решением является точка $\bar{x} = a - r\|c\|^{-1}c$.

Если X — полиэдр, то вспомогательная задача (4.2.2) для нахождения точки $\bar{x}$ есть задача линейного программирования.

Предположим теперь, что $f(x)$ — выпуклая дважды дифференцируемая на $\mathbb{R}^n$ функция со строго положительной матрицей вторых производных $f_{xx}(x)$ всюду на X, а множество X только выпукло и

замкнуто. Если задачу (4.2.3) минимизации линейного приближения функции $f(x)$ в точке x_k заменить на задачу минимизации ее квадратичного приближения в этой точке, то можно получить вариант метода Ньютона, который назовем *условным методом Ньютона* для решения задачи (4.1.1). В нем начальная точка $x_0 \in X$. Если найдена точка $x_k \in X$, то последующая точка x_{k+1} полагается равной $\bar{x}_k$, где $\bar{x}_k$ есть решение задачи

$$\min_{x \in X} \phi(x), \tag{4.2.15}$$

в которой

$$\phi(x) = f(x_k) + \langle f_x(x_k), x - x_k \rangle + \frac{1}{2} \langle x - x_k, f_{xx}(x_k)(x - x_k) \rangle.$$

Так как матрица $f_{xx}(x_k)$ положительно определена, решение задачи (4.2.15) всегда существует.

Накладывая на $f(x)$ дополнительные требования, можно показать, что данный вариант метода Ньютона сохраняет сверхлинейную или даже квадратичную скорость сходимости.

Теорема 4.2.2. *Пусть $X \subset \mathbb{R}^n$ — выпуклое замкнутое множество и $f(x)$ — дважды дифференцируемая функция, для второй производной которой выполняется левое неравенство (2.3.5) и условие Липшица (2.4.9). Тогда условный метод Ньютона локально сходится к решению задачи (4.1.1) с квадратичной скоростью.*

Доказательство. Возьмем точку $x_* \in X$, являющуюся решением задачи (4.1.1). Такая точка существует и единственна, поскольку согласно (2.3.5) $f(x)$ — сильно выпуклая функция. Оценим $\|x_{k+1} - x_*\|$. Из-за того, что точка x_{k+1} есть точка минимума функции $\phi(x)$ на X, из условий оптимальности следует неравенство

$$\langle f_x(x_k) + f_{xx}(x_k)(x_{k+1} - x_k), x - x_{k+1} \rangle \geq 0 \quad \forall\, x \in X.$$

Полагая в этом неравенстве $x = x_*$ и учитывая необходимые условия минимума функции $f(x)$ на X, получаем

$$\begin{aligned} 0 &\leq \langle f_x(x_k) + f_{xx}(x_k)(x_{k+1} - x_k), x_* - x_{k+1} \rangle = \\ &= \langle f_x(x_k) - f_x(x_*) + f_{xx}(x_k)(x_{k+1} - x_k), x_* - x_{k+1} \rangle + \\ &\qquad + \langle f_x(x_*), x_* - x_{k+1} \rangle \leq \\ &\leq \langle f_x(x_k) - f_x(x_*) + f_{xx}(x_k)(x_{k+1} - x_k), x_* - x_{k+1} \rangle = \Delta. \end{aligned} \tag{4.2.16}$$

Но

$$\begin{aligned} f_x(x_k) &= f_x(x_*) + \textstyle\int_0^1 f_{xx}(x_* + \tau(x_k - x_*))(x_k - x_*) d\tau = \\ &= f_x(x_*) + f_{xx}(x_k)(x_k - x_*) + \Delta_1, \end{aligned} \tag{4.2.17}$$

где

$$\Delta_1 = \int_0^1 [f_{xx}(x_* + \tau(x_k - x_*)) - (f_{xx}(x_k)](x_k - x_*)d\tau.$$

Имеем согласно неравенству (2.4.9)

$$\begin{array}{rcl} \|\Delta_1\| & = & \| \int_0^1 [f_{xx}(x_* + \tau(x_k - x_*)) - (f_{xx}(x_k)](x_k - x_*)d\tau\| \leq \\ & \leq & L\|x_k - x_*\|^2 \int_0^1 \tau d\tau = \frac{L}{2}\|x_k - x_*\|^2. \end{array}$$

Поэтому для правой части Δ неравенства (4.2.16) после подстановки (4.2.17) получаем

$$\begin{array}{rcll} \Delta & = & \langle f_{xx}(x_k)(x_{k+1} - x_*), x_* - x_{k+1}\rangle + \langle \Delta_1, x_* - x_{k+1}\rangle \leq & \\ & \leq & -m\|x_{k+1} - x_*\|^2 + \|\Delta_1\|\|x_{k+1} - x_*\| = & (4.2.18) \\ & = & \left[-m\|x_{k+1} - x_*\| + \frac{L}{2}\|x_k - x_*\|^2\right] \|x_{k+1} - x_*\|. & \end{array}$$

Так как в силу (4.2.16) $\Delta \geq 0$, то неравенство (4.2.18) возможно только, если $x_{k+1} = x_*$, либо когда

$$-m\|x_{k+1} - x_*\| + \frac{L}{2}\|x_k - x_*\|^2 \geq 0.$$

Последнее неравенство влечет оценку

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|^2, \qquad (4.2.19)$$

где $C = \frac{L}{2m}$. Если точка $x_0 \in X$ взята достаточно близко к x_*, а именно, когда выполняется неравенство $\|x_0 - x_*\| < C^{-1}$, то неравенство (4.2.19) показывает, что метод Ньютона будет сходиться к x_*, причем с квадратичной скоростью. ■

Применение метода Ньютона для решения задачи условной минимизации (4.1.1) оправдано лишь в том случае, когда решение вспомогательной задачи по нахождению точки $\bar{x}_k$ сравнительно просто. Например, если X — полиэдральное множество, то вспомогательная задача будет задачей квадратичного программирования.

4.3. Метод приведенного градиента

Идея метода приведенного градиента заключается в переходе к безусловной оптимизации в другом пространстве, причем желательно меньшей размерности. Особенно просто это сделать в том случае,

когда у нас допустимое множество задается ограничениями типа равенства. Опишем один из вариантов метода приведенного градиента на примере задачи с линейными ограничениями-равенствами:

$$\min_{x\in X} f(x), \qquad X = \{x \in \mathbb{R}^n : \ Ax = b\}, \tag{4.3.1}$$

где $f(x)$ — дифференцируемая функция, A — матрица размера $m \times n$, причем $m < n$. Предполагаем, что матрица A имеет полный ранг, равный m, т.е. строки A линейно независимы.

В задаче (4.3.1) в отличие от ранее рассмотренных случаев допустимое множество X уже определяется функционально с помощью равенств. Однако поскольку эти ограничения линейные, допустимое множество имеет сравнительно простой вид, а именно X есть линейное многообразие в $\mathbb{R}^n$ размерности $d = n - m$. Методы приведенного градиента для задачи (4.3.1) основаны на учете специфики данного множества. Здесь прежде всего мы рассмотрим вариант метода, в котором проводится исключение части переменных. Делать это можно разными способами.

Пусть B — невырожденная квадратная подматрица порядка m матрицы A. Не умаляя общности, считаем, что B состоит из первых m столбцов матрицы A. Другими словами, матрица A представима в виде: $A = [B, N]$. Разобьем в соответствии с разбиением матрицы A также вектор x на две части $x = [x^B, x^N]^T$. Тогда если $x \in X$, то

$$Ax = Bx^B + Nx^N = b. \tag{4.3.2}$$

Мы можем разрешить уравнение (4.3.2) относительно x^B и выразить x^B через переменную x^N, которую принято называть *независимой переменной* и размерность которой равна d. Имеем

$$x^B = x^B(x^N) = B^{-1}\left(b - Nx^N\right). \tag{4.3.3}$$

Рассмотрим теперь наряду с функцией $f(x) = f(x^B, x^N)$ функцию

$$\phi(x^N) = f(x^B(x^N), x^N), \tag{4.3.4}$$

зависящую только от независимой переменной x^N. От решения задачи (4.3.1) теперь можно перейти к задаче безусловной минимизации

$$\min_{x^N \in R^d} \phi(x^N). \tag{4.3.5}$$

Между градиентами функции $\phi(x^N)$ и градиентами исходной функции $f(x)$ существует связь. Используя формулу дифференцирования

сложной функции, получаем

$$\phi_{x^N}(x^N) = f_{x^N}(x^B(x^N), x^N) + \left(\frac{d\,x^B(x^N)}{d\,x^N}\right)^T f_{x^B}(x^B(x^N), x^N). \quad (4.3.6)$$

Но согласно (4.3.3): $\frac{d\,x^B(x^N)}{d\,x^N} = -B^{-1}N$. Таким образом,

$$\phi_{x^N}(x^N) = Rf_x(x), \quad R = \left[-N^T(B^T)^{-1} \mid I_d\right], \quad (4.3.7)$$

где $x = [x^B, x^N]^T \in X$ и $x^B = x^B(x^N)$.

Можно подойти к исключению переменных в задаче (4.3.1) с несколько иных позиций, а именно считать, что проводится замена переменных, и в новых переменных ограничения задачи принимают очень простой вид. Например, они сводятся к равенству нулю части новых переменных.

Пусть $y = [y^B, y^N] \in \mathbb{R}^n$ и пусть

$$y^B = Ax - b, \quad y^N = x^N. \quad (4.3.8)$$

В матричной записи такой переход можно записать как

$$y = y(x) = \begin{bmatrix} y^B(x) \\ y^N(x) \end{bmatrix} = \begin{bmatrix} B & N \\ 0_{dm} & I_d \end{bmatrix} \begin{bmatrix} x^B \\ x^N \end{bmatrix} - \begin{bmatrix} b \\ 0_d \end{bmatrix}.$$

У данного линейного преобразования существует обратное преобразование:

$$x = x(y) = \begin{bmatrix} x^B(y) \\ x^N(y) \end{bmatrix} = \begin{bmatrix} B^{-1} & -B^{-1}N \\ 0_{dm} & I_d \end{bmatrix} \begin{bmatrix} y^B \\ y^N \end{bmatrix} + \begin{bmatrix} B^{-1}b \\ 0_d \end{bmatrix}.$$

Чтобы выполнялись ограничения задачи (4.3.1), следует положить $y^B = 0_m$. Тогда приходим к целевой функции $\phi(y^N)$, зависящей только от второй компоненты y^N и совпадающей в нашем случае со второй компонентой x^N. Данная функция получается как $\phi(y^N) = \tilde{f}(0, y^N)$, где функция $\tilde{f}(y)$ есть $\tilde{f}(y) = f(x(y))$. Фактически $\phi(y^N)$ имеет вид (4.3.4).

Вместо преобразования (4.3.8) можно также обратиться к более общему линейному преобразованию $y(x)$, в котором

$$y^B(x) = G_B(Ax - b), \qquad y^N(x) = G_N x - p, \quad (4.3.9)$$

где G_B — неособая матрица порядка m, матрица G_N имеет размер $d \times n$, вектор $p \in \mathbb{R}^d$. Из невырожденности матрицы G_B следует, что

условие $y^B = 0_m$ выполняется тогда и только тогда, когда выполняется равенство $Ax = b$. Использованное нами ранее преобразование (4.3.8) является частным случаем более общего преобразования (4.3.9) и соответствует тому, что $G_B = I_m$, $G_N = [0_{d\,m} \mid I_d]$ и $p = 0_d$.

В матричном виде преобразование $y(x)$ запишется как

$$y(x) = Qx - q, \qquad Q = \begin{bmatrix} G_B A \\ G_N \end{bmatrix}, \qquad q = \begin{bmatrix} G_B b \\ p \end{bmatrix}. \tag{4.3.10}$$

Если строки матриц A и G_N линейно независимы, то матрица Q будет неособой. Тогда существует обратное преобразование:

$$x(y) = Q^{-1}(y + q).$$

Пусть, как и ранее, $\tilde{f}(y) = f(x(y))$. Имеем $f(x) = \tilde{f}(Qx - q)$. Дифференцируя это равенство, получаем

$$f_x(x) = Q^T \tilde{f}_y(y(x)) = A^T G_B^T \tilde{f}_{y^B}(y(x)) + G_N^T \tilde{f}_{y^N}(y(x)). \tag{4.3.11}$$

Откуда, в частности,

$$\tilde{f}_y(y) = (Q^T)^{-1} f_x(x(y)). \tag{4.3.12}$$

Если $y_* = [y_*^B, y_*^N]$ — точка минимума функции $\tilde{f}(y)$ при $y^B \equiv 0_m$ или, что, по существу, одно и то же, если y_*^N — решение задачи

$$\min_{y^N \in R^d} \phi(y^N), \qquad \phi(y^N) = \tilde{f}(0_m, y^N), \tag{4.3.13}$$

то $\tilde{f}_{y^N}(y_*) = 0_d$. Тогда из (4.3.12), используя выражение (4.3.11), приходим к равенству: $f_x(x_*) = A^T u_*$, где $x_* = x(y_*)$ и $u_* = G_B^T \tilde{f}_{y^B}(y_*)$. Отсюда делаем вывод, что соответствующая точка x_* является стационарной точкой в задаче (4.3.1), поскольку пара $[x_*, u_*]$ образует точку Каруша–Куна–Таккера.

Для решения задачи (4.3.13) можно применять различные методы безусловной минимизации и, в частности, метод градиентного спуска. В качестве примера приведем метод градиентного спуска для простейшего случая (4.3.13) — задачи (4.3.5). Тогда расчеты следует проводить в пространстве переменной y^N, размерность которой равна d. Так как для преобразования (4.3.3) $y^N = x^N$, то после перехода к переменной x^N метод запишется в виде рекуррентной схемы:

$$x_{k+1}^N = x_k^N - \alpha_k \phi_{x^N}(x_k^N), \tag{4.3.14}$$

в которой $\alpha_k > 0$ и

$$\phi_{x^N}(x_k^N) = -N^T(B^{-1})^T f_{x^B}(x_k) + f_{x^N}(x_k).$$

Здесь $x_k = [x_k^B, x_k^N]^T$, причем $x_k^B = B^{-1}(b - x_k^N)$. Кроме того, принята во внимание связь (4.3.6) между градиентами.

Можно поступать и по другому, а именно: расчеты проводить в пространстве исходных переменных x, но вместо градиентов целевой функции $f(x)$ в этом пространстве брать образы в x-пространстве градиентов функции $\tilde{f}(y)$ с нулевыми первыми компонентами $\tilde{f}_{y^B}(y)$, т.е. в которых положено $\tilde{f}_{y^B}(y) = 0_m$. Но сдвигу Δy в y-пространстве согласно (4.3.10) соответствует сдвиг $\Delta x = Q^{-1}\Delta y$ в x-пространстве. Отсюда, используя формулу (4.3.12), приходим к вектору

$$f_x^r(x) = Q^{-1}\begin{bmatrix} 0_m \\ \tilde{f}_{y^N}(y(x)) \end{bmatrix} = Q^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I_d \end{bmatrix}\tilde{f}_y(y(x)) = \\ = Q^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I_d \end{bmatrix}(Q^T)^{-1} f_x(x).$$

Вектор $f_x^r(x)$ носит название *приведенного* или *редуцированного градиента*. Так как $\Delta y^B = 0_m$, а вектор Δy^B в силу (4.3.10) связан с Δx зависимостью $y^B = G_B A \Delta x$, то $A\Delta x = 0_m$. Следовательно, для $f_x^r(x)$ выполняется равенство: $Af_x^r(x) = 0_m$, т.е. он лежит в линейном подпространстве параллельном допустимому множеству X.

С использованием приведенного градиента итерационный процесс запишется в виде

$$x_{k+1} = x_k - \alpha_k f_x^r(x_k), \quad x_0 \in X, \tag{4.3.15}$$

и проводится он в отличие, например, от (4.3.14) в n-мерном пространстве. Выбор шага α_k осуществляется на основе минимизации исходной целевой функции $f(x)$, а не видоизмененной по сравнению с $f(x)$ функции $\tilde{f}(y)$. Если $x_0 \in X$, то все последующие точки x_k также принадлежат X. Итерационные процессы типа (4.3.15) можно трактовать так же, как образы в x-пространстве итерационных процессов, проводимых в y-пространстве.

Рассмотрим теперь один важный частный случай процесса (4.3.15). Дело в том, что за счет выбора матриц G_B и G_N матрицу Q можно сделать ортогональной, т.е. такой, что $QQ^T = Q^TQ = I_n$. В самом деле, для этого достаточно взять, например, такие G_B и G_N, для которых выполняются соотношения:

$$G_B G_B^T = (AA^T)^{-1}, \qquad G_N G_N^T = I_d, \\ G_N^T G_N = I - A^T(AA^T)^{-1}A, \qquad AG_N^T = 0_m.$$

Тогда $Q^{-1} = Q^T = [A^T G_B^T \mid G_N^T]$ и мы имеем

$$f_x^r(x) = G_N^T G_N f_x(x) = P f_x(x), \quad P = I - A^T (AA^T)^{-1} A.$$

Таким образом, в этом случае приведенный градиент является ортогональной проекцией $f_x(x)$ на линейное подпространство, параллельное допустимому множеству X. Если подставить данный приведенный градиент в общую схему спуска (4.3.15), то получим следующий итерационный процесс:

$$x_{k+1} = x_k - \alpha_k P f_x(x_k), \quad x_0 \in X. \tag{4.3.16}$$

Метод (4.3.16) может быть обобщен на решение задачи (4.3.1), в которой допустимое множество X задается ограничениями-неравенствами:

$$X = \left\{ x \in \mathbb{R}^n : \ Ax \le b \right\}.$$

Используемый здесь подход заключается в учете *активных ограничений*, т.е. таких ограничений $\langle a_i, x \rangle \le b^i$, которые в точке x выполняются как равенства. Здесь a_i есть i-я строка матрицы A.

4.4. Метод приведенных ньютоновских направлений

Рассмотрим теперь метод второго порядка для решения задачи (4.3.1). Аналогично методу Ньютона для задачи безусловной минимизации в нем целевая функция заменяется ее квадратичным приближением.

Пусть в (4.3.1) целевая функция $f(x)$ является выпуклой дважды дифференцируемой функцией с положительно определенными матрицами вторых производных $f_{xx}(x)$. Предположим, что нам известна точка $x \in X$, т.е. точка, удовлетворяющая равенству $Ax = b$. Будем искать новую точку $\bar{x} \in X$ в виде $\bar{x} = x + s$, где $s \in \mathbb{R}^n$. Тогда, используя квадратичное приближение целевой функции в точке x, приходим к следующей задаче условной минимизации:

$$\begin{gathered} f(x) + \langle f_x(x), s \rangle + \frac{1}{2} \langle s, f_{xx}(x) s \rangle \to \min_{s \in R^n}, \\ As = 0_m. \end{gathered} \tag{4.4.1}$$

Составим для задачи (4.4.1) функцию Лагранжа:

$$L(s, u) = f(x) + \langle f_x(x), s \rangle + \frac{1}{2} \langle s, f_{xx}(x) s \rangle + \langle u, As \rangle, \quad u \in \mathbb{R}^m.$$

В решении задачи (4.4.1) — точке s — должны выполняться необходимые условия минимума (условия Куна–Таккера):

$$L_s(s,u) = f_x(x) + f_{xx}(x)s + A^T u = 0_n, \quad L_u(s,u) = As = 0_m, \quad (4.4.2)$$

где u — соответствующий множитель Лагранжа. В матричном виде система уравнений (4.4.2) записывается как

$$\begin{bmatrix} f_{xx}(x) & A^T \\ A & 0_{mm} \end{bmatrix} \begin{bmatrix} s \\ u \end{bmatrix} = \begin{bmatrix} -f_x(x) \\ 0_m \end{bmatrix}. \quad (4.4.3)$$

Матрица системы (4.4.3) симметричная, но не положительно определенная. Если она неособая, то, разрешая систему (4.4.3), можно найти направление s. Один из наиболее простых подходов к решению системы (4.4.3), как, впрочем, и других подобных систем, получающихся из необходимых условий оптимальности, состоит в исключении одной переменной. Разрешая первое уравнение из (4.4.2) относительно s, находим: $s = -f_{xx}^{-1}(x)[f_x(x) + A^T u]$. После подстановки данного s во второе равенство из (4.4.2) приходим к уравнению относительно u:

$$Af_{xx}^{-1}(x)A^T u + Af_{xx}^{-1}(x)f_x(x) = 0_m.$$

При сделанных предположениях относительно целевой функции $f(x)$ и матрицы A (напомним, что матрица A имеет полных ранг, равный m) матрица $Af_{xx}^{-1}(x)A^T$ является положительно определенной и мы получаем

$$u = -\left(Af_{xx}^{-1}(x)A^T\right)^{-1} Af_{xx}^{-1}(x)f_x(x).$$

Таким образом,

$$s = -f_{xx}^{-1}(x)\left[I_n - A^T\left(Af_{xx}^{-1}(x)A^T\right)^{-1} Af_{xx}^{-1}(x)\right] f_x(x). \quad (4.4.4)$$

В случае, когда $s \neq 0_n$, данное направление является допустимым, т.е. оно не выводит за пределы допустимого множества X. Кроме того, s является направлением, вдоль которого целевая функция не возрастает. В самом деле

$$\langle f_x(x), s\rangle = -\langle Q(x)f_x(x), P(x)\ Q(x)f_x(x)\rangle \leq 0.$$

Здесь $Q(x) = f_{xx}^{-\frac{1}{2}}(x)$ и

$$P(x) = I_n - Q(x)A^T\left(AQ^2(x)A^T\right)^{-1} AQ(x).$$

Матрица $P(x)$ есть матрица ортогонального проектирования на подпространство, ортогональное пространству столбцов матрицы $Q(x)A^T$.

Найденному направлению s можно дать и несколько иную интерпретацию. А именно, допустимое множество в задаче (4.3.1) является аффинным и задается с помощью системы линейных алгебраических уравнений. Но, как нам известно, существует другой способ представления аффинного множества в виде сдвинутого линейного подпространства L, параллельного X. Подпространство L есть нульпространство матрицы A, его размерность равна $d = n - m$. Пусть B — матрица размером $n \times d$, столбцы которой задают базис в подпространстве L. Пусть, кроме того, $\tilde{x}$ — произвольная точка из X. Тогда

$$X = \left\{ x \in \mathbb{R}^n : \; x = x(y) = By + \tilde{x}, \; y \in \mathbb{R}^d \right\}. \tag{4.4.5}$$

Представление допустимого множества в виде (4.4.5) дает еще одну возможность сведения задачи условной оптимизации (4.3.1) к задаче безусловной минимизации функции меньшего числа переменных:

$$\min_{y \in R^d} \tilde{f}(y), \quad \tilde{f}(y) = f(By + \tilde{x}).$$

Градиент функции $\tilde{f}(y)$ и ее матрица Гессе равняются соответственно

$$\tilde{f}_y(y) = B^T f_x(x(y)), \quad \tilde{f}_{yy}(y) = B^T f_{xx}(x(y))B.$$

Поэтому для ньютоновского направления $r \in \mathbb{R}^d$ при минимизации функции $\tilde{f}(y)$ получаем выражение

$$r = -\left(B^T f_{xx}(x(y))B\right)^{-1} B^T f_x(x(y)).$$

Данному направлению в y-пространстве соответствует направление в исходном x-пространстве:

$$s = Br = -B\left(B^T f_{xx}(x)B\right)^{-1} B^T f_x(x). \tag{4.4.6}$$

Покажем, что направление s из (4.4.6) есть на самом деле найденное нами ранее направление (4.4.4). С этой целью возьмем множитель Лагранжа:

$$u = -(AA^T)^{-1} A \left[f_x(x) + f_{xx}(x)s\right] \tag{4.4.7}$$

и покажем, что пара $[s, u]$ удовлетворяет системе (4.4.2). Поскольку

$$AB = 0_{md}, \tag{4.4.8}$$

то второе уравнение из (4.4.2) обязательно удовлетворяется.

Проверим выполнение первого уравнения из (4.4.2). Прежде всего заметим, что для u вида (4.4.7) выполняется

$$A\left(f_{xx}(x)s + A^T u + f_x(x)\right) = 0_m. \tag{4.4.9}$$

Кроме того, подставляя s из (4.4.6) и повторно учитывая равенство (4.4.8), получаем

$$B^T\left(f_{xx}(x)s + A^T u + f_x(x)\right) = 0_d. \tag{4.4.10}$$

Так как строки матриц A и B^T линейно независимы и в совокупности образуют базис в пространстве $\mathbb{R}^n$, то из (4.4.9) и (4.4.10) следует, что $f_{xx}(x)s + A^T u + f_x(x) = 0_n$. Таким образом, указанные s и u удовлетворяют системе (4.4.2). Поскольку матрица этой системы неособая, то у системы может быть только единственное решение. Поэтому полученное направление (4.4.6) можно интерпретировать как ньютоновское направление в некоторой редуцированной задаче.

Глава 5

Методы линеаризации и последовательного квадратичного программирования

Идея использования линейных или квадратичных приближений для решения нелинейных задач является одной из основных в вычислительной математике. Широко она используется и в оптимизации, когда решение задачи нелинейного программирования ищется с помощью решений последовательности задач линейного или квадратичного программирования.

5.1. Метод возможных направлений

Метод возможных направлений применяется для решения задачи

$$f_* = \min_{x \in X} f(x), \quad X = \left\{ x \in \mathbb{R}^n : \; g^i(x) \leq 0, \; 1 \leq i \leq m \right\}, \tag{5.1.1}$$

в которой $f(x)$ и $g^i(x)$ — дифференцируемые функции, $1 \leq i \leq m$. Таким образом, в отличие от задачи (4.1.1) допустимое множество X теперь задается функционально с помощью ограничений типа неравенства. Множество решений (5.1.1) обозначаем по-прежнему X_*.

Точка $x_* \in X$ является *стационарной* для задачи (5.1.1), если существует такое $u_* \in \mathbb{R}^m_+$, что $L_x(x_*, u_*) = 0$ и $\langle g(x_*), u_* \rangle = 0$. Здесь $L(x, u) = f(x) + \langle g(x), u \rangle$ — функция Лагранжа, составленная для задачи (5.1.1).

Напомним встречавшиеся ранее определения.

Определение 5.1.1. *Направление $s \in \mathbb{R}^n$ называется возможным относительно множества X в точке $x \in X$, если можно указать $\bar{\alpha} > 0$ такое, что $x + \alpha s \in X$ для $0 \le \alpha \le \bar{\alpha}$.*

Определение 5.1.2. *Направление $s \in \mathbb{R}^n$ называется направлением убывания функции $f(x)$ в точке $x \in \mathbb{R}^n$, если можно указать $\bar{\alpha} > 0$ такое, что $f(x + \alpha s) < f(x)$ для $0 < \alpha \le \bar{\alpha}$.*

Совокупность всех возможных относительно множества X в точке $x \in X$ направлений является конусом. Обозначим его $K_X(x)$. Если X — выпуклое множество и векторы $x \in X$, $y \in X$, то $s = y - x \in K_X(x)$. В случае, когда множество X имеет вид (5.1.1), также несложно указать направление из $K_X(x)$ в точке $x \in X$. Действительно, обозначим через $J_0(x) = \{1 \le i \le m : \ g^i(x) = 0\}$ *множество индексов активных ограничений* в точке x. Если $s \in K_X(x)$, то $\langle g^i_x(x), s \rangle \le 0$ для всех $i \in J_0(x)$. Более того, если $\langle g^i_x(x), s \rangle < 0$ для любого $i \in J_0(x)$, то обязательно $s \in K_X(x)$.

Что касается множества всех направлений убывания функции $f(x)$ в точке $x \in \mathbb{R}^n$, то оно также является конусом. Его обозначим $K_d(x)$. Если функция $f(x)$ дифференцируема в x, то любой вектор $s \in \mathbb{R}^n$, для которого $\langle f_x(x), s \rangle < 0$, принадлежит конусу $K_d(x)$.

Основная идея метода возможных направлений заключается в следующем: в текущей точке $x \in X$ найти направление $s \in \mathbb{R}^n$, которое является одновременно возможным относительно множества X и направлением убывания целевой функции $f(x)$ в этой точке, и осуществить вдоль этого направления спуск с некоторым шагом. При этом следующая точка должна остаться в множестве X.

Опишем теперь общую схему метода, которая реализует данную идею. Пусть $x_0 \in X$ задано и пусть найдена точка x_k. Новую точку x_{k+1} вычисляем согласно следующему рекуррентному соотношению:

$$x_{k+1} = x_k + \alpha_k s_k,$$

где $s_k \in K_X(x_k) \cap K_d(x_k)$, а шаг α_k выбирается таким образом, чтобы по крайней мере выполнялись условия: $x_{k+1} \in X$ и $f(x_{k+1}) < f(x_k)$.

В рамках этой достаточно общей эвристической схемы реализуются несколько конкретных вариантов метода возможных направлений.

Схема 1. В ней для определения вектора s_k на каждой итерации решается вспомогательная задача

$$\begin{array}{rcll} \min\ \sigma, & & & \\ \langle f_x(x_k), s\rangle & \leq & \sigma, & \\ \langle g_x^i(x_k), s\rangle & \leq & \sigma, & i \in J_0(x_k), \\ s & \in & S, & \end{array} \tag{5.1.2}$$

где S — некоторое компактное множество в $\mathbb{R}^n$, содержащее начало координат. Обычно в качестве множества S берется единичный шар в какой-нибудь норме, например:

$$B_\infty = \left\{s \in \mathbb{R}^n :\ \|s\|_\infty \leq 1\right\} \quad \text{или} \quad B_2 = \left\{s \in \mathbb{R}^n :\ \|s\|_2 \leq 1\right\},$$

причем выбор в качестве S множества B_∞ приводит к тому, что вспомогательная задача (5.1.2) оказывается задачей линейного программирования.

Пусть $[\sigma_k, s_k]$ — решение задачи (5.1.2). Если $\sigma_k < 0$, то $s_k \in K_X(x_k)$ и $s_k \in K_d(x_k)$, т.е. направление s_k оказывается одновременно и допустимым и направлением убывания. В случае, когда $J_0(x) = \emptyset$ и $S = B_2$, оно просто совпадает с направлением антиградиента функции $f(x)$ в точке x_k.

Шаг α_k выбирается посредством какой-либо процедуры одномерного поиска при дополнительном условии, что $\alpha_k \leq \bar{\alpha}_k$, где $\bar{\alpha}_k$ — максимальный шаг, для которого $x_k + \alpha_k s_k \in X$.

Если же в решении $[\sigma_k, s_k]$ оказывается, что $\sigma_k = 0$, то точка x_k при выполнении в ней *условия регулярности ограничений Мангасариана–Фромовица*, заключающемуся в существовании такого $\bar{s} \in \mathbb{R}^n$, что $\langle g_x^i(x_k), \bar{s}\rangle < 0$, $i \in J_0(x_k)$, является стационарной точкой для задачи (5.1.1).

Данный вариант метода возможных направлений может либо вообще не сходиться, либо «застревать» вблизи точек, не являющихся стационарными. Это происходит потому, что в выборе направления s_k участвуют лишь активные ограничения в точке x_k и совсем не принимаются во внимание близкие «почти активные» ограничения. В связи с этим рассмотрим другой вариант метода, в котором эти «почти активные» ограничения уже учитываются.

Схема 2 *(Зойтендейка).* Положим

$$J_\varepsilon(x) = \left\{1 \leq i \leq m :\ g^i(x) \geq -\varepsilon\right\},$$

где $\varepsilon > 0$. Об ограничениях $g^i(x)$ с индексами $i \in J_\varepsilon(x)$ говорят как об *ε-активных ограничениях* в точке $x \in X$.

Перед началом расчетов задается $x_0 \in X$ и $\varepsilon_0 > 0$, а также дополнительный параметр $0 < \theta < 1$. Пусть на k-м шаге получена точка $x_k \in X$ и известно $\varepsilon_k > 0$. Для нахождения s_k решается вспомогательная задача

$$\begin{array}{rcl} \min\ \sigma, & & \\ \langle f_x(x_k), s\rangle & \leq & \sigma, \\ \langle g^i_x(x_k), s\rangle & \leq & \sigma, \quad i \in J_{\varepsilon_k}(x_k), \\ s & \in & S. \end{array} \tag{5.1.3}$$

Если в ее решении $[\sigma_k, s_k]$ величина σ_k удовлетворяет неравенству $\sigma_k \leq -\varepsilon_k$, то s_k берется в качестве возможного направления и направления убывания. Если нет, то параметр ε_k уменьшают, полагая $\varepsilon_k := \theta\varepsilon_k$, и снова решают вспомогательную задачу (5.1.3). Этот вариант метода возможных направлений при определенных предположениях уже может обладать сходимостью.

В предложенном методе начальная точка x_0 бралась из допустимого множества X. Чтобы ее найти, в общем случае требуется решить некоторую вспомогательную задачу, которая сама по себе может оказаться достаточно сложной. Другой путь — это рассмотреть вариант метода, в котором этап поиска начальной точки объединен с основным этапом поиска решения задачи (5.1.1) на допустимом множестве. Заодно будет дана и несколько иная трактовка методов возможных направлений.

Предположим далее, что (5.1.1) есть задача выпуклого программирования, т.е. как целевая функция $f(x)$, так и все функции $g^i(x)$, где $1 \leq i \leq m$, выпуклы. Предположим также, что ограничения в (5.1.1) удовлетворяют условию регулярности Слейтера, согласно которому существует точка $\bar{x} \in X$, которая принадлежит множеству $X_0 = \left\{x \in \mathbb{R}^n :\ g(x) < 0_m\right\}$. Таким образом, выпуклое множество X_0 заведомо непустое. Более того, его замыкание совпадает с X.

Введем величину $\eta \in \mathbb{R}$ и обозначим

$$\psi(x) = \max_{1 \leq i \leq m} g^i(x).$$

Используя функцию $\psi(x)$, составим вспомогательную функцию, зависящую от $n+1$ переменной:

$$M(x, \eta) = \max\,[f(x) - \eta, \psi(x)]. \tag{5.1.4}$$

В более развернутом виде функцию $M(x, \eta)$ можно записать как

$$M(x, \eta) = \max\,[f(x) - \eta, g^1(x), \ldots, g^m(x)].$$

Функция $M(x,\eta)$ является негладкой и строго положительной, когда $x \notin X$. Для задачи выпуклого программирования она выпукла по x.

Обозначим через $X(\eta)$ множество решений задачи минимизации функции $M(x,\eta)$ на $\mathbb{R}^n$ при фиксированном η:

$$X(\eta) = \operatorname{Arg} \min_{x \in R^n} M(x,\eta).$$

Считаем, что решения таких задач существуют.

Теорема 5.1.1. *Точка $x_* \in \mathbb{R}^n$ является решением задачи (5.1.1) тогда и только тогда, когда можно указать $\eta_* \in \mathbb{R}$ такое, что*

$$M(x_*,\eta_*) = 0, \qquad x_* \in X(\eta_*). \tag{5.1.5}$$

Доказательство. *Необходимость.* Возьмем $x_* \in X_*$ и положим $\eta_* = f(x_*)$. Отсюда в силу допустимости x_* сразу приходим к выполнению равенства $M(x_*,\eta_*) = 0$.

Покажем теперь, что $x_* \in X(\eta_*)$. Поскольку $\psi(x) > 0$ для любого $x \notin X$, то для этих x имеет место неравенство

$$M(x,\eta_*) \geq \psi(x) > 0 = M(x_*,\eta_*). \tag{5.1.6}$$

Если $x \in X$, то, учитывая, что $f(x) \geq f(x_*) = \eta_*$ и $\psi(x) \leq 0$, получаем

$$M(x,\eta_*) = f(x) - \eta_* \geq f(x_*) - \eta_* = M(x_*,\eta_*), \tag{5.1.7}$$

причем неравенство будет строгое, если $x \notin X_*$. Из (5.1.6) и (5.1.7) следует, что $x_* \in X(\eta_*)$.

Достаточность. Пусть выполнено (5.1.5). Проверим сначала, что $x_* \in X$. В самом деле, так как $M(x_*,\eta_*) = 0$, то $\psi(x_*) \leq 0$ и, следовательно, $x_* \in X$.

Убедимся теперь, что $x_* \in X_*$. Для любых $x \in X$ в силу (5.1.5) имеет место неравенство

$$M(x,\eta_*) \geq M(x_*,\eta_*) = 0. \tag{5.1.8}$$

Предположим, что $\eta_* = f(x_*)$. Отсюда

$$M(x,\eta_*) = f(x) - \eta_* = f(x) - f(x_*) \geq 0,$$

если $x \in X_0$. В силу непрерывности неравенство $f(x) \geq f(x_*)$ остается справедливым и для всех $x \in X$. Таким образом, $x_* \in X_*$.

Покажем теперь, что случай, когда η_* было бы отлично от $f(x_*)$, невозможен. Действительно, допущение $\eta_* < f(x_*)$ сразу приводит к

нарушению равенства $M(x_*, \eta_*) = 0$. Если $\eta_* > f(x_*)$, то обязательно $\psi(x_*) = 0$ и, следовательно, x_* — граничная точка X. Так как ограничения в задаче (5.1.1) удовлетворяют условию Слейтера, то найдется $\bar{x} \in X_0$. Из выпуклости функции $\psi(x)$ для всех $x_\lambda = \lambda x_* + (1-\lambda)\bar{x}$, $0 < \lambda < 1$, следует неравенство

$$\psi(x_\lambda) \le \lambda\psi(x_*) + (1-\lambda)\psi(\bar{x}) < 0.$$

Поскольку по предположению $f(x_*) - \eta_* < 0$, в силу непрерывности для достаточно малых λ будем иметь: $M(x_\lambda, \eta_*) = \psi(x_\lambda) < 0$, что противоречит включению $x_* \in X(\eta_*)$. ■

Утверждение теоремы 5.1.1 показывает, что в принципе решение задачи (5.1.1) сводится к отысканию точек x_* и η_*, удовлетворяющих условиям (5.1.5). Находить такие точки можно, например, используя подход, основанный на следующей идее: будем минимизировать функцию $M(x, \eta)$ по x, применяя тот или иной метод спуска, но одновременно на каждой итерации этого метода будем уточнять значение параметра η, причем таким образом, чтобы выполнялось равенство $M(x, \eta) = 0$. Разумеется, точного выполнения данного равенства можно добиться только в том случае, когда текущая точка x оказывается в допустимом множестве X. В остальных точках $x \notin X$ будем выбирать η, исходя из пожелания, чтобы значение функции $M(x, \eta)$ как можно меньше отличалось от нулевого значения.

Рассмотрим в связи со сказанным следующий итерационный процесс:

$$\eta_k = f(x_k), \qquad x_{k+1} = x_k + \alpha_k s_k, \tag{5.1.9}$$

где $x_0 \in \mathbb{R}^n$, направление $s_k = s(x_k, \eta_k)$ является направлением убывания функции $M(x, \eta_k)$ по x в точке x_k. Шаг $\alpha_k > 0$ выбирается тем или иным способом из условия минимизации функции $M(x, \eta)$ по x вдоль направления убывания.

Вспомогательная функция $M(x, \eta)$ имеет специальный вид, а именно: при каждом фиксированном η, рассматриваемая как функция от переменной x, она является так называемой *функцией максимума*. Общий вид функции максимума следующий:

$$y(x) = \max_{1 \le i \le m} h^i(x), \tag{5.1.10}$$

где $h^i(x)$ — некоторые функции, $1 \le i \le m$. Если все функции $h^i(x)$ выпуклы на $\mathbb{R}^n$, то такая функция $y(x)$ также выпукла на $\mathbb{R}^n$. Поэтому в каждой точке x существует субдифференциал $\partial y(x)$.

Для нахождения точек минимума функции максимума разработаны специальные методы, основанные на использовании направлений убывания таких функций. Рассмотрим некоторые из них, предполагая при этом, что все функции $h^i(x)$, $1 \le i \le m$, по крайней мере непрерывно дифференцируемы.

Для произвольного $\varepsilon > 0$ и $x \in \mathbb{R}^n$ определим индексное множество

$$J_\varepsilon(x) = \left\{1 \le i \le m : \ h^i(x) \ge y(x) - \varepsilon\right\}.$$

Функции $h^i(x)$, индексы которых принадлежат множеству $J_\varepsilon(x)$, называются *ε-активными* в точке x. Введем, кроме того, множество

$$K(x) = \left\{ s \in \mathbb{R}^n : \ y'(x;s) = \max_{i \in J_0(x)} \langle h^i_x(x), s\rangle < 0 \right\}.$$

Здесь $y'(x;s)$ — производная функции $y(x)$ по направлению s. Нетрудно видеть, что $K(x)$ есть конус. Каждое направление s из $K(x)$ является направлением убывания функции $y(x)$ в точке x. Если x не является *стационарной точкой* функции $y(x)$, т.е. точкой, где выполняется необходимое условие локального минимума $y(x)$, то конус $K(x)$ не пустой.

Приведем два способа выбора направления убывания функции $y(x)$.

1. *Направление ε-наискорейшего спуска.* Данное направление находится из решения следующей минимаксной задачи:

$$\sigma = \min_{s \in S} \max_{i \in J_\varepsilon(x)} \langle h^i_x(x), s\rangle,$$

где S, как и в случае задачи (5.1.2), некоторое компактное множество, содержащее начало координат. Беря в качестве S множество B_∞ (единичный шар в норме $\|\cdot\|_\infty$), приходим к вспомогательной задаче линейного программирования:

$$\begin{array}{rl} \min \sigma, & \\ \langle h^i_x(x), s\rangle \le & \sigma, \quad i \in J_\varepsilon(x), \\ |s^j| \le 1, & 1 \le j \le n. \end{array} \tag{5.1.11}$$

Если $\sigma < 0$, то $s \in K(x)$.

2. Другой способ построения направления убывания функции $y(x)$ в точке x основан на идее линеаризации. В самом деле, если заменить все функции $h^i(x)$, определяющие $y(x)$, их линейными приближениями и далее найти минимум функции максимума от этих линейных приближений, то полученное решение будет направлением, вдоль которого функция $y(x)$ убывает. Таким образом, мы приходим к следующей

минимаксной задаче:

$$\sigma = \min_{s \in S} \max_{1 \le i \le m} p^i(s), \tag{5.1.12}$$

где

$$p^i(s) = h^i(x) + \langle h^i_x(x), s \rangle - y(x).$$

Компактное множество S введено с той же целью, что и в (5.1.11).

Если в качестве S взять множество B_∞, то решение задачи (5.1.12) эквивалентно решению задачи линейного программирования:

$$\begin{array}{ll} \min \sigma, & \\ h^i(x) + \langle h^i_x(x), s \rangle - y(x) \le \sigma, & 1 \le i \le m, \\ |s^j| \le 1, & 1 \le j \le n. \end{array} \tag{5.1.13}$$

Отметим, что решение задачи (5.1.13), как правило, более сложное, чем решение соответствующей задачи (5.1.11). В частности, размерность задачи линейного программирования (5.1.13) превышает размерность задачи (5.1.11).

Беря в итерационном процессе (5.1.9) в качестве векторов s_k направления ε-наискорейшего спуска, построенные по первому способу, мы приходим к классическому методу возможных направлений Зойтендейка, уже рассмотренному нами для случая, когда в качестве стартовой точки x_0 берется точка из X. В этом случае направления s_k находятся из решения вспомогательной задачи (5.1.3), которая может быть, в частности, задачей линейного программирования. Второй способ построения направления убывания функции максимума применяется в *методе возможных направлений Топкиса–Вейнотта.*

Упражнение 4. *Сформулируйте вспомогательную задачу, которую требуется решать в методе Топкиса–Вейнотта на каждой итерации. Считайте при этом, что итерации проводятся по рекуррентной схеме (5.1.9) и* $S = B_\infty$.

5.2. Метод линеаризации

Рассмотрим снова задачу (5.1.1), причем опять считаем, что все функции, определяющие эту задачу, по меньшей мере непрерывно дифференцируемы. По-прежнему множество оптимальных решений в задаче (5.1.1) обозначаем через X_*. Через $\tilde{X}$ обозначим множество *стационарных точек* в задаче (5.1.1), т.е. множество точек $x_* \in \mathbb{R}^n$, которые вместе с вектором $u_* \in \mathbb{R}^m$ образуют точку Каруша–Куна–Таккера.

Построим другой метод решения задачи (5.1.1), основанный на использовании линейных приближений как целевой функции, так и ограничений. В этом методе на каждом шаге для нахождения направления перемещения решается вспомогательная задача линейного программирования. Однако теперь найденное из решения вспомогательной задачи направление является не направлением убывания вспомогательной функции (5.1.4), а направлением убывания другой вспомогательной функции:

$$M(x,t) = f(x) + t \max_{0\le i\le m} g^i(x). \tag{5.2.1}$$

Здесь t — достаточно большое положительное число. Кроме того, для удобства дальнейших построений введено дополнительное ограничение $g^0(x) \le 0$ с функцией $g^0(x)$, тождественно равной нулю для всех $x \in \mathbb{R}^n$. Допустимое множество X от этого, разумеется, не изменится. Отметим также, что функция $M(x,t)$ является негладкой относительно переменной x. В случае, когда (5.1.1) — задача выпуклого программирования, $M(x,t)$ выпукла по x.

Если точка x_* есть решение задачи (5.1.1) и в ней выполнены достаточные условия второго порядка, то можно показать, что x_* при достаточно большом t будет также точкой минимума функции $M(x,t)$ по x (быть может, локальной). Таким образом, беря достаточно большое число t и минимизируя функцию $M(x,t)$ по x, мы можем в принципе найти решение исходной задачи (5.1.1). Эта идея и лежит в основе метода линеаризации, предложенного Б.Н. Пшеничным. Разные способы выбора направления убывания функции $M(x,t)$ приводят к разным реализациям этого метода. Здесь мы рассмотрим один из них, в котором направление убывания находится из решения вспомогательной задачи линейного программирования.

Вспомогательная функция (5.2.1) имеет вид функции максимума, так как

$$M(x,t) = \max_{0\le i\le m} h^i(x), \quad h^i(x) = f(x) + tg^i(x), \quad 0 \le i \le m.$$

При сделанных предположениях, касающихся функций $f(x)$ и $g^i(x)$, $0 \le i \le m$, функция $M(x,t)$ дифференцируема относительно x по любому ненулевому направлению $s \in \mathbb{R}^n$, и ее производная $M_x(x,t;\,s)$ по этому направлению равна

$$M_x(x,t;\,s) = \langle f_x(x), s\rangle + t \max_{i\in J_0^0(x)} \langle g_x^i(x), s\rangle, \tag{5.2.2}$$

где

$$J_0^0(x) = \left\{0 \le i \le m :\ g^i(x) = \phi(x)\right\}, \qquad \phi(x) = \max_{0\le i\le m} g^i(x).$$

Нетрудно видеть, что $\phi(x) \geq 0$ для любого $x \in \mathbb{R}^n$, причем $\phi(x) \equiv 0$, когда $x \in X$.

Точку $[x_*, t_*] \in \mathbb{R}^n \times \mathbb{R}_+$ назовем *особой точкой* функции $M(x,t)$, если

$$\inf_{\|s\|=1} M_x(x_*, t_*;\, s) \geq 0, \qquad M_t(x_*, t_*) = \phi(x_*) = 0. \tag{5.2.3}$$

В силу второго равенства любая особая точка $[x_*, t_*]$ функции $M(x,t)$ такова, что $x_* \in X$.

Первое неравенство в (5.2.3) указывает на то, что в точке x_* выполняется необходимое условие локального минимума функции $M(x, t_*)$ на всем пространстве. Оно справедливо тогда и только тогда, когда существуют такие множители $w^i \geq 0$, $i \in J_0^0(x_*)$, что $\sum_{i \in J_0^0(x_*)} w^i = 1$ и имеет место равенство

$$\sum_{i \in J_0^0(x_*)} w^i [f_x(x_*) + t g_x^i(x_*)] = 0. \tag{5.2.4}$$

Это следует из того, что при выполнении первого неравенства (5.2.3) $s = 0_n$ является точкой минимума на $\mathbb{R}^n$ положительно однородной и выпуклой относительно переменной s на $\mathbb{R}^n$ функции

$$R(s) = \langle f_x(x_*), s \rangle + t \max_{i \in J_0^0(x_*)} \langle g_x^i(x_*), s \rangle.$$

Поэтому субдифференциал этой функции в нуле $\partial R(0_n)$ должен содержать нулевую точку. Учитывая вид субдифференциала для функции максимума (см. [33]), приходим к (5.2.4).

Чтобы найти направление убывания функции $M(x,t)$ по x, можно было бы использовать подходы, приведенные, например, в предыдущем разделе. Однако здесь мы поступим несколько иначе, а именно: будем в качестве направления s в точке $x \in \mathbb{R}^n$ брать решение следующей вспомогательной задачи:

$$\begin{array}{rcl} \langle f_x(x), s \rangle + c\,\sigma & \to & \min_{s,\sigma}, \\ g^i(x) + \langle g_x^i(x), s \rangle & \leq & 0, \quad i \in J_\delta(x), \\ |s^j| - \sigma & \leq & 1, \quad 1 \leq j \leq n, \quad \sigma \geq 0. \end{array} \tag{5.2.5}$$

Здесь c — произвольная константа такая, что $c \geq c(x) = 1 + \sqrt{n}\|f_x(x)\|$, индексное множество $J_\delta(x)$, зависящее от параметра $\delta \geq 0$, определяется следующим образом:

$$J_\delta(x) = \left\{ 1 \leq i \leq m :\ g^i(x) \geq \phi(x) - \delta \right\}.$$

Утверждение 5.2.1. *Если допустимое множество X в задаче (5.2.5) непусто, то решение задачи (5.2.5) существует и конечно.*

Доказательство. Обозначим через $z \in \mathbb{R}^n \times \mathbb{R}_+$ пару $z = [s, \sigma]$, через $y(z)$ — целевую функцию в задаче (5.2.5), т.е.

$$y(z) = \langle f_x(x), s\rangle + c\sigma.$$

Если предположить противное, то найдется последовательность точек $z_k = [s_k, \sigma_k]$, удовлетворяющих ограничениям задачи (5.2.5), и таких, что $\lim_{k\to\infty} \|z_k\| = \infty$. При этом, естественно, $\lim_{k\to\infty} \sigma_k = \infty$. В то же время для последовательности значений целевой функции выполняется $y(z_{k+1}) \le y(z_k)$ для всех $k \ge 1$. Отсюда приходим к выводу, что $\limsup_{k\to} y(z_k) \le y(z_0)$.

С другой стороны, поскольку $\|s_k\| \le \sqrt{n}(1+\sigma_k)$, в силу неравенства Коши–Буняковского справедлива оценка

$$\begin{aligned} y(z_k) &\ge c(x)\sigma_k - \|f_x(x)\|\|s_k\| \ge \\ &\ge c(x)\sigma_k - \sqrt{n}(1+\sigma_k)\|f_x(x)\| \ge \sigma_k - \sqrt{n}\|f_x(x)\|, \end{aligned}$$

поэтому $\limsup_{k\to\infty} y(z_k) = \infty$, что невозможно. ■

Двойственной к (5.2.5) будет задача

$$\begin{array}{rcl} \sum_{i\in J_\delta(x)} u^i g^i(x) - \sum_{j=1}^n \left(v_1^j + v_2^j\right) & \to & \max_{u,v_1,v_2}, \\ \sum_{i\in J_\delta(x)} u^i g_x^i(x) + v_1 - v_2 & = & -f_x(x), \\ \sum_{j=1}^n \left(v_1^j + v_2^j\right) & \le & c, \\ v_1 \ge 0_n, \quad v_2 \ge 0_n, & u^i \ge 0, & i \in J_\delta(x). \end{array} \tag{5.2.6}$$

Здесь $v_1 = [v_1^1, \ldots, v_1^n]$ и $v_2 = [v_2^1, \ldots, v_2^n]$ — множители Лагранжа, соответствующие двусторонним ограничениям: $-(1+\sigma) \le s^j \le 1+\sigma$, $1 \le j \le n$.

Предположим, что u и v_1, v_2 есть решение задачи (5.2.6), s и σ — решение задачи (5.2.5). На основании теоремы двойственности оптимальные значения в прямой задаче (5.2.5) и в двойственной задаче (5.2.6) совпадают, поэтому имеет место равенство

$$\langle f_x(x), s\rangle + c\sigma = \sum_{i\in J_\delta(x)} u^i g^i(x) - \sum_{j=1}^n \left(v_1^j + v_2^j\right). \tag{5.2.7}$$

Лемма 5.2.1. *Пусть ненулевое направление s является решением задачи (5.2.5). Тогда для производной по направлению $M_x(x,t;\, s)$ функции $M(x,t)$ справедлива оценка*

$$M_x(x,t;\, s) \le \langle f_x(x), s\rangle - t\phi(x). \tag{5.2.8}$$

Более того, если $x \in X$, то данная производная не зависит от t и совпадает с производной функции $f(x)$ по s, т.е.

$$M_x(x,t;\, s) = \langle f_x(x), s\rangle. \tag{5.2.9}$$

Доказательство. Для производной функции $M(x,t)$ по направлению s справедливо выражение (5.2.2). Но когда $i \in J_0^0(x)$, $i \neq 0$, то обязательно $i \in J_\delta(x)$ и выполняется $\langle g_x^i(x), s\rangle \leq -g^i(x) = -\phi(x)$. Это неравенство имеет место и для $i = 0$, причем независимо от выбора точки x, следовательно:

$$\max_{i \in J_0^0(x)} \langle g_x^i(x), s\rangle \leq -\phi(x).$$

Отсюда приходим к оценке (5.2.8).

Пусть теперь $x \in X$. При этом предположении множество $J_0^0(x)$ либо состоит из единственного нулевого индекса, либо содержит помимо его такие индексы i, для которых $g^i(x) = 0$. Но тогда этот ненулевой индекс i обязательно входит в индексное множество $J_\delta(x)$, и поскольку направление s есть решение вспомогательной задачи (5.2.5), то выполняется неравенство $\langle g_x^i(x), s\rangle \leq -g^i(x) = 0$. Поэтому в любом случае получаем, что выражение (5.2.2) для производной $M_x(x,t;\, s)$ принимает вид (5.2.9). ■

Утверждение 5.2.2. *Пусть $x \notin \tilde{X}$. Пусть, кроме того, направление s вместе с σ являются решением задачи (5.2.5), а множители u_i, $i \in J_\delta(x)$, вместе с v_1 и v_2 являются решением двойственной задачи (5.2.6). Тогда для любого*

$$t > t_\delta(u) = \sum_{i \in J_\delta(x)} u^i$$

направление s есть направление убывания функции $M(x,t)$ по x.

Доказательство. Воспользуемся равенством (5.2.7) и оценкой (5.2.8). Принимая во внимание, что

$$t\phi(x) \geq \phi(x) \sum_{i \in J_\delta(x)} u^i(x) \geq \sum_{i \in J_\delta(x)} u^i g^i(x), \tag{5.2.10}$$

приходим к неравенству

$$M_x(x,t;\, s) \leq \langle f_x(x), s\rangle - t\phi(x) \leq -\left[c\sigma + \sum_{j=1}^{n} \left(v_1^j + v_2^j\right)\right]. \tag{5.2.11}$$

Более того, при $x \notin X$ неравенство (5.2.10), а следовательно, и (5.2.11) строгие, так как в этом случае $\phi(x) > 0$. Отсюда заключаем, что $M_x(x,t;\, s) < 0$ при данных x.

Допустим далее, что $x \in X$. Тогда для производной $M_x(x,t;\, s)$ справедливо выражение (5.2.9). Кроме того, так как $\phi(x) = 0$, то из (5.2.11) следует, что $\langle f_x(x), s\rangle \le 0$.

Покажем, что равенство нулю возможно только при $x \in \tilde{X}$. Действительно, если $\langle f_x(x), s\rangle = 0$, то согласно (5.2.11)

$$c\sigma + \sum_{j=1}^{n}(v_1^j + v_2^j) = 0,$$

т.е. $\sigma = 0$, $v_1 = v_2 = 0_n$. Поэтому из первого ограничения в (5.2.6) получаем

$$f_x(x) + \sum_{i \in J_\delta(x)} u^i g_x^i(x) = 0_n. \tag{5.2.12}$$

Более того, равенство (5.2.7) с учетом допущения $\langle f_x(x), s\rangle = 0$ сводится к следующему:

$$\sum_{i \in J_\delta(x)} u^i g^i(x) = 0. \tag{5.2.13}$$

Отсюда вытекает, что $u^i = 0$, когда $i \in J_\delta(x)$ и $g^i(x) < 0$. Полагая все остальные $u^i = 0$ при $i \notin J_\delta(x)$, из (5.2.12) и (5.2.13) приходим к равенствам

$$f_x(x) + \sum_{i=1}^{m} u^i g_x^i(x) = 0_n, \quad u^i g^i(x) = 0, \quad 1 \le i \le m. \tag{5.2.14}$$

Вместе с учетом неравенства $g(x) \le 0_m$ это означает, что пара $[x,u]$ является точкой Каруша–Куна–Таккера для задачи (5.1.1). Следовательно, $x \in \tilde{X}$. ■

Выясним теперь, как решения вспомогательной задачи (5.2.5) связаны с особыми точками функции $M(x,t)$.

Утверждение 5.2.3. *Пусть $[x,t]$ — особая точка функции $M(x,t)$. Тогда среди решений вспомогательной задачи (5.2.5) имеется нулевое решение $s = 0_n$, $\sigma = 0$. Обратно, если $s = 0_n$ и $\sigma = 0$ являются решением задачи (5.2.5), то $[x,t]$ — особая точка функции $M(x,t)$ для любого $t \ge t_\delta(u)$, где u — соответствующее решение двойственной задачи* (5.2.6).

Доказательство. Предположим сначала, что $[x,t]$ — особая точка функции $M(x,t)$. Тогда пара $s=0_n$, $\sigma=0$ является допустимой для задачи (5.2.5). Кроме того, x — стационарная точка в задаче (5.1.1). Действительно, в этом случае $x\in X$ и имеет место равенство (5.2.4), в котором $x_*=x$. Поэтому если обозначить $u^i=tw^i$, $i\in J_0(x)$, и положить $u^i=0$, $i\notin J_0(x)$, то приходим к равенствам (5.2.14), которые означают, что точка x является стационарной. Но тогда, беря эти самые u^i, $i\in J_\delta(x)$, и $v_1=v_2=0_n$, получаем из (5.2.14), что данная точка является допустимой в двойственной задаче (5.2.6), причем в силу (5.2.7) значение целевой функции в двойственной задаче совпадает со значением целевой функции в прямой задаче (5.2.5), когда берутся $s=0_n$ и $\sigma=0$. Это означает, что данная нулевая точка является оптимальной в задаче (5.2.5).

Допустим теперь, что $s=0_n$, $\sigma=0$ являются решением задачи (5.2.5) и $t\geq t_\delta(u)$. В силу допустимости для задачи (5.2.5) нулевого решения $s=0_n$ имеем: $g^i(x)\leq 0$, $i\in J_\delta(x)$, поэтому $\phi(x)=0$ и выполняется второе равенство из (5.2.3) для $x_*=x$. Кроме того, из условия дополняющей нежесткости для (5.2.5) следует, что $v_1=v_2=0_n$ и $u^i=0$, когда $g^i(x)<0$. В этом случае из ограничений задачи (5.2.6) вытекает равенство

$$f_x(x)+\sum_{i\in J_0(x)}u^ig_x^i(x)=0_n. \tag{5.2.15}$$

Если положить $w^i=\frac{u^i}{t}$, $i\in J_0(x)$, и $w^0=1-\frac{t_\delta(u)}{t}$, то из (5.2.15) следует (5.2.4), гарантирующее выполнение при $t\geq t_\delta(u)$ левого неравенства (5.2.3). ■

Утверждения 5.2.2 и 5.2.3 позволяют свести нахождение стационарных точек задачи (5.1.1) к нахождению особых точек функции $M(x,t)$. Последние могут быть найдены с помощью метода спуска вдоль направлений, определяемых из решения вспомогательной задачи линейного программирования (5.2.5). Наличие нулевого решения у вспомогательной задачи указывает на то, что текущая точка является точкой минимума по x функции $M(x,t)$. Все это, разумеется верно, когда параметр t взят достаточно большим.

Построим теперь метод решения задачи, основанный на решении вспомогательных задач (5.2.5).

Алгоритм метода линеаризации. Задаемся начальной точкой $x_0\in\mathbb{R}$. Пусть, кроме того, заданы достаточно большая величина $t>0$, начальный шаг $\alpha>0$ и параметры $\delta>0$ и $0<\varepsilon<1$.

Общая k-я итерация

Шаг 1. Пусть известна точка x_k. Решаем вспомогательную задачу (5.2.5) при $x = x_k$ и находим ее решение s_k и σ_k. Если $s_k = 0_n$ и $\sigma_k = 0$, то останавливаемся.

Шаг 2. Переходим в новую точку x_{k+1}, полагая

$$x_{k+1} = x_k + \alpha_k s_k, \tag{5.2.16}$$

где α_k — шаг спуска, выбираемый путем деления пополам начального шага α до тех пор, пока не будет выполнено условие

$$M(x_k + \alpha_k s_k, t) \le M(x_k, t) + \varepsilon \alpha_k P(x_k, t, s_k), \tag{5.2.17}$$

где

$$P(x, t, s) = \langle f_x(x), s\rangle - t\phi(x).$$

Шаг 3. Полагаем $k := k + 1$ и идем на шаг 1.

Сделаем два замечания относительно алгоритма метода линеаризации.

1. Значение функции $P(x, t, s)$ является оценкой сверху для производной по направлению $M_x(x, t; s)$.

2. Если предположить, что константа t выбрана настолько большой, что множество подуровня

$$\mathcal{L}(x_0; t) = \left\{x \in \mathbb{R}^n : \ M(x, t) \le M(x_0, t)\right\}$$

ограничено и градиенты функций $f(x)$ и $g^i(x)$, $1 \le i \le m$, удовлетворяют на $\mathcal{L}(x_0; t)$ условию Липшица с константой L:

$$\|f_x(x_1) - f_x(x_2)\| \le L\|x_1 - x_2\|, \quad \|g^i_x(x_1) - g^i_x(x_2)\| \le L\|x_1 - x_2\|, \tag{5.2.18}$$

то, используя разложения Тейлора для функций $f(x)$ и $g^i(x)$, можно получить оценку снизу для α_k:

$$\alpha_k \ge \frac{1}{2}\min\left\{\alpha, \ \frac{\delta}{\phi(x_k) + K\|s_k\|}, \ \frac{(\varepsilon - 1)P(x_k, t, s_k)}{(t + 1)L\|s_k\|^2}\right\}. \tag{5.2.19}$$

Здесь: K — величина, ограничивающая $\|f_x(x)\|$ на компактном множестве $\mathcal{L}(x_0; t)$. Из (5.2.19) видно, что если x_k не является стационарной точкой в задаче (5.1.1), то для того, чтобы подобрать α_k, надо делить начальное α пополам только конечное число раз.

Покажем, что метод линеаризации при определенных дополнительных предположениях обладает сходимостью.

Лемма 5.2.2. *Пусть начальная точка x_0 и величина t таковы, что множество $\mathcal{L}(x_0;t)$ ограничено. Пусть, кроме того, для любого $x \in \mathcal{L}(x_0;t)$ задача линейного программирования (5.2.5) разрешима, и для соответствующих множителей Лагранжа u^i, $i \in J_\delta(x)$, выполняется оценка сверху: $\sum_{i\in J_\delta(x)} u^i \le t-1$. Тогда можно указать достаточно большое $C > 0$ такое, что для любого $x \in \mathcal{L}(x_0;t)$ и для любого решения $[s,\sigma]$ задачи (5.2.5), построенной в этой точке, имеет место неравенство $\sigma < C$.*

Доказательство. Для любого $x \in \mathcal{L}(x_0;t)$ справедлива цепочка неравенств

$$\begin{array}{rcl} \langle f_x(x), s\rangle + c\sigma & \ge & c\sigma - \|f_x(x)\|\|s\| \ge \\ & \ge & c\sigma - \sqrt{n}\|f_x(x)\|(1+\sigma) \ge \\ & \ge & \sigma - \sqrt{n}\|f_x(x)\| \ge \sigma - \sqrt{n}K. \end{array} \tag{5.2.20}$$

С другой стороны, согласно (5.2.7) имеет место оценка

$$\begin{array}{rcl} \langle f_x(x), s\rangle + c\sigma & = & \sum_{i\in J_\delta(x)} u^i g^i(x) - \sum_{j=1}^n \left(v_1^j + v_2^j\right) \le \\ & \le & \sum_{i\in J_\delta(x)} u^i g^i(x) \le t\phi(x) \le tC_1, \end{array} \tag{5.2.21}$$

где $C_1 = \sup_{x\in\mathcal{L}(x_0;t)} \phi(x)$. В силу непрерывности функции $\phi(x)$ на $\mathcal{L}(x_0;t)$ константа C_1 существует и конечна. Полагая $C = tC_1 + \sqrt{n}K$, из (5.2.20) и (5.2.21) приходим к требуемому утверждению. ■

Ниже для простоты предполагается, что константа c во всех вспомогательных задачах берется равной $c(x)$.

Теорема 5.2.1. *Пусть выполнены предположения леммы 5.2.2. Пусть, кроме того, градиенты функций $f(x)$ и $g^i(x)$, $1 \le i \le m$, удовлетворяют на $\mathcal{L}(x_0;t)$ условию Липшица (5.2.18). Тогда любая предельная точка $\bar{x}$ последовательности $\{x_k\}$, получающейся в результате работы метода (5.2.16), является стационарной точкой в задаче (5.1.1).*

Доказательство. Выделим из последовательности $\{x_k\}$ сходящуюся подпоследовательность. Такие подпоследовательности обязательно существуют, так как все точки последовательности $\{x_k\}$ принадлежат ограниченному множеству $\mathcal{L}(x_0;t)$. Пусть $x_{k_l} \to \bar{x}$.

Так как непрерывная функция $M(x,t)$ ограничена снизу на компактном множестве $\mathcal{L}(x_0;t)$, то, в силу (5.2.17), $\alpha_{k_l} P(x_{k_l}, t, s_{k_l}) \to 0$. При этом обязательно

$$P(x_{k_l}, t, s_{k_l}) \to 0. \tag{5.2.22}$$

В самом деле, если (5.2.22) не имеет места, то $\alpha_{k_l} \to 0$. Но согласно оценке (5.2.19) это возможно только тогда, когда выполняется (5.2.22), ибо, согласно лемме 5.2.2, все s_k ограничены в совокупности на $\mathcal{L}(x_0; t)$.

Убедимся сначала, что $\phi(\bar{x}) = 0$, т.е. $\bar{x} \in X$. Действительно, не умаляя общности, можно считать, что последовательность $\{x_{k_l}\}$ взята такой, что множество $J_\delta(x_{k_l})$ одно и то же для всех x_{k_l}. Обозначим его J_δ. В силу непрерывности всех функций $g^i(x)$, $1 \le i \le m$, должно выполняться включение $J_\delta \subseteq J_\delta(\bar{x})$.

По последовательности $\{x_{k_l}\}$ определяются последовательности решений прямых и двойственных задач (5.2.5), (5.2.6), причем в силу сделанных предположений и утверждения леммы 5.2.2 все они ограничены в совокупности. Поэтому из них можно выделить сходящиеся подпоследовательности. Считаем для простоты, что $s_{k_l} \to \bar{s}$, $\sigma_{k_l} \to \bar{\sigma}$, $u^i_{k_l} \to \bar{u}^i$, $i \in J_\delta$, и $(v_1)_{k_l} \to \bar{v}_1$, $(v_2)_{k_l} \to \bar{v}_2$.

Из двойственных соотношений (5.2.7), переходя в них к пределу, с учетом непрерывности получаем

$$\langle f_x(\bar{x}), \bar{s}\rangle - \sum_{i \in J_\delta} \bar{u}^i g^i(\bar{x}) = -\bar{c}\bar{\sigma} - \sum_{j=1}^{n} \left(\bar{v}_1^j + \bar{v}_2^j\right), \tag{5.2.23}$$

где $\bar{c} = c(\bar{x})$. При этом правая часть в (5.2.23) неположительна. Но в силу (5.2.22)

$$\langle f_x(\bar{x}), \bar{s}\rangle - t\phi(\bar{x}) = 0. \tag{5.2.24}$$

Если теперь предположить, что $\bar{x} \notin X$, то из (5.2.23) и (5.2.24) приходим к противоречию. Следовательно, $\bar{x} \in X$ и $\langle f_x(\bar{x}), \bar{s}\rangle = 0$. Кроме того, равенство (5.2.23) возможно только в том случае, когда $\bar{\sigma} = 0$, $\bar{v}_1 = \bar{v}_2 = 0_n$ и $\bar{u}^i = 0$, если $g^i(\bar{x}) < 0$.

Рассмотрим теперь задачи (5.2.5) и (5.2.6) при $x = \bar{x}$. Если взять в качестве v_1, v_2 и u^i, $i \in J_\delta$, соответственно $\bar{v}_1$, $\bar{v}_2$ и $\bar{u}^i$, а также положить $u^i = 0$, $i \in J_\delta(\bar{x}) \setminus J_\delta$, то эта точка является допустимой для двойственной задачи (5.2.6), построенной в точке $x = \bar{x}$. С другой стороны, точка $[s, \sigma] = [0_n, 0]$ является допустимой для прямой задачи (5.2.5) для данного $\bar{x}$. Поскольку значения целевых функций совпадают, то, согласно теореме двойственности, $[s, \sigma] = [0_n, 0]$ есть решение прямой задачи (5.2.5). В силу утверждения 5.2.3 это может быть только тогда, когда $\bar{x}$ является особой точкой функции $M(x, t)$. Поэтому выполняется (5.2.4). Отсюда, обозначая $u^i_* = tw^i$, $i \in J_0(x_*)$, и полагая все остальные $u^i_* = 0$, приходим к выводу, что x_* — стационарная точка в задаче (5.1.1). ■

Если (5.1.1) — задача выпуклого программирования, то любая ее стационарная точка является одновременно решением (5.1.1). Кроме

того, в этом случае более слабые требования к допустимому множеству X в задаче (5.1.1), а именно условие Липшица и ограниченность X, гарантируют выполнение предположений леммы 5.2.2.

5.3. Методы последовательного квадратичного программирования

Методы последовательного квадратичного программирования довольно часто называют *SQP-методами* (SQP — сокращение английского термина *sequential quadratic programming*). В этих методах решение нелинейной задачи условной оптимизации заменяется решением последовательности задач квадратичного программирования. Последние же строятся с использованием линейных и квадратичных приближений функций, входящих в постановку задачи. Понятно из общих соображений, что квадратичные приближения гораздо точнее аппроксимируют исходную нелинейную задачу и, следовательно, можно надеяться, что они будут обладать большей скоростью сходимости по сравнению с методами, основанными только на линейных приближениях. Хотя методы последовательного квадратичного программирования можно строить для задач, содержащих как ограничения типа неравенства, так ограничения типа равенства, здесь для простоты изложения ограничимся задачей, в которой присутствуют только ограничения-равенства.

Пусть требуется решить задачу

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : \ g(x) = 0_m\right\}, \tag{5.3.1}$$

в которой $m < n$ и предполагается, что как целевая функция $f(x)$, так и вектор-функция $g(x) = [g^1(x), \ldots, g^m(x)]^T$, определяющая ограничения, по крайней мере дважды непрерывно дифференцируемы. Через $L(x, u)$ будем обозначать функцию Лагранжа:

$$L(x, u) = f(x) + \langle g(x), u\rangle, \quad u \in \mathbb{R}^m, \tag{5.3.2}$$

составленную для задачи (5.3.1).

Предположим, что задана точка $x \in \mathbb{R}^n$. Возьмем вместо функций $f(x)$ и $g^i(x)$ их квадратичные приближения в этой точке:

$$\phi(s) = f(x) + \langle f_x(x), s\rangle + \frac{1}{2}\langle s, f_{xx}(x)s\rangle,$$

$$\psi^i(s) = g^i(x) + \langle g^i_x(x), s\rangle + \frac{1}{2}\langle s, g^i_{xx}(x)s\rangle, \quad 1 \le i \le m,$$

и рассмотрим вспомогательную задачу:

$$\begin{array}{c} \phi(s) \ \to \ \min, \\ \psi^i(s) = 0, \quad 1 \le i \le m. \end{array} \tag{5.3.3}$$

Задача (5.3.3) является задачей квадратичного программирования с квадратичными ограничениями типа равенства. Присутствие в (5.3.3) квадратичных ограничений существенно усложняет ее решение. Однако имеется возможность упростить задачу (5.3.3) и заменить ее задачей с линейными ограничениями. Действительно, составим для задачи (5.3.3) функцию Лагранжа:

$$M(s,u) = \phi(s) + \langle \psi(s), u \rangle,$$

где $u \in \mathbb{R}^m$, $\psi(s) = [\psi^1(s), \ldots, \psi^m(s)]^T$. Если точка s_* есть решение задачи (5.3.3) и выполнено какое-нибудь условие регулярности ограничений, например линейная независимость градиентов ограничений, то в силу необходимых условий экстремума найдется $u_* \in \mathbb{R}^m$ такое, что

$$M_s(s_*, u_*) = 0_n. \tag{5.3.4}$$

В более подробном виде с использованием функции Лагранжа $L(x,u)$ для исходной задачи (5.3.1) равенство (5.3.4) записывается как

$$L_x(x, u_*) + L_{xx}(x, u_*)s_* = 0_n, \tag{5.3.5}$$

где

$$L_{xx}(x,u) = f_{xx}(x) + H(x,u), \quad H(x,u) = \sum_{i=1}^{m} u^i g^i_{xx}(x),$$

есть матрица вторых производных функции Лагранжа (5.3.2) по прямым переменным.

Возьмем теперь вместо $\psi^i(s)$ линейные приближения функций $g^i(x)$:

$$\tilde{\psi}^i(s) = g^i(x) + \langle g^i_x(x), s \rangle, \quad 1 \le i \le m.$$

В совокупности они составляют линейную m-мерную вектор-функцию $\tilde{\psi}(s) = [\tilde{\psi}^1(s), \ldots, \tilde{\psi}^m(s)]^T$. Изменим также целевую функцию $\phi(s)$, положив

$$\tilde{\phi}(s) = \phi(s) + \frac{1}{2}\langle s, H(x,u_*)s \rangle = f(x) + \langle f_x(x), s \rangle + \frac{1}{2}\langle s, L_{xx}(x,u_*)s \rangle.$$

Данная функция $\tilde{\phi}(s)$ отличается от $\phi(s)$ только тем, что теперь в квадратичном слагаемом вместо матрицы вторых производных целевой функции $f_{xx}(x)$ используется матрица вторых производных функции Лагранжа $L_{xx}(x,u_*)$.

Если заменить $\phi(s)$ и $\psi(s)$ соответственно на функции $\tilde{\phi}(s)$ и $\tilde{\psi}(s)$ и рассмотреть вместо (5.3.3) задачу квадратичного программирования с линейными ограничениями

$$\begin{array}{c} f(x) + \langle f_x(x), s\rangle + \frac{1}{2}\langle s, L_{xx}(x, u_*)s\rangle \;\to\; \min, \\ g^i(x) + \langle g^i_x(x), s\rangle = 0, \quad 1 \le i \le m, \end{array} \tag{5.3.6}$$

то, составляя для нее функцию Лагранжа

$$\tilde{M}(s, u) = \tilde{\phi}(s) + \langle \tilde{\psi}(s), u\rangle, \quad u \in \mathbb{R}^m,$$

и выписывая условия оптимальности для решения $\tilde{s}_*$ уже задачи (5.3.6), получаем

$$\tilde{M}_s(\tilde{s}_*, \tilde{u}_*) = L_x(x, \tilde{u}_*) + L_{xx}(x, u_*)\tilde{s}_* = 0_n, \tag{5.3.7}$$

где $\tilde{u}_*$ — соответствующий вектор множителей Лагранжа для решения $\tilde{s}_*$ задачи (5.3.6).

Вычтем равенство (5.3.7) из (5.3.5). Тогда приходим к следующему соотношению:

$$L_{xx}(x, u_*)\,(\tilde{s}_* - s_*) = -g^T_x(x)\,(\tilde{u}_* - u_*)\,,$$

из которого видно, что в принципе решение задачи с линейными ограничениями (5.3.7) может быть близким к решению задачи с квадратичными ограничениями, если не очень сильно отличаются друг от друга соответствующие множители Лагранжа $\tilde{u}_*$ и u_*.

Данное свойство вспомогательных задач (5.3.3) и (5.3.6) наводит на мысль, что можно отказаться от решения задач (5.3.3) и заменить их решением гораздо более простых задач (5.3.6) с линейными ограничениями. Метод последовательного квадратичного программирования основан именно на этой идее. В нем с помощью решений вспомогательных задач вида (5.3.6) строятся последовательности точек $\{x_k\}$ и $\{u_k\}$ такие, что последовательность $\{x_k\}$ сходится к решению x_* задачи (5.3.1). Конечно, заключение о близости решения $\tilde{s}_*$ задачи (5.3.6) к решению s_* задачи (5.3.3) основывается на том, что в (5.3.6) добавочная матрица $H(x, u_*)$ берется с множителем Лагранжа u_*, который, разумеется, заранее нам не известен. Это препятствие обходится с помощью использования вместо неизвестных множителей u_* множителей Лагранжа u_k, найденных на предыдущих итерациях.

Опишем теперь итерационный алгоритм метода последовательного квадратичного программирования для решения задачи (5.3.1). Пусть заданы x_0 и u_0. Предположим, что на некотором k-м шаге получено

приближение x_k, u_k. Находим новые точки x_{k+1}, u_{k+1}, решая вспомогательную задачу квадратичного программирования:

$$\begin{gathered} f(x_k) + \langle f_x(x_k), x - x_k\rangle + \tfrac{1}{2}\langle x - x_k, L_{xx}(x_k, u_k)(x - x_k)\rangle \;\to\; \min, \\ g(x_k) + g_x(x_k)\,(x - x_k) = 0_m. \end{gathered} \tag{5.3.8}$$

В качестве x_{k+1} берется решение этой задачи, а в качестве u_{k+1} — соответствующий вектор множителей Лагранжа, вычисляемый из условий оптимальности для задачи (5.3.8).

Задача (5.3.8) имеет сравнительно простой вид, и условия оптимальности сводятся к системе линейных уравнений, которым удовлетворяют $\Delta x_k = x_{k+1} - x_k$, а также u_{k+1}. Эти условия оптимальности записываются следующим образом:

$$\begin{gathered} L_x(x_k, u_{k+1}) + L_{xx}(x_k, u_k)\Delta x_k = 0_n, \\ g(x_k) + g_x(x_k)\Delta x_k = 0_m. \end{gathered} \tag{5.3.9}$$

Обозначим дополнительно $\Delta u_k = u_{k+1} - u_k$. Если учесть, что

$$L_x(x_k, u_{k+1}) = L_x(x_k, u_k) + g_x^T(x_k)\Delta u_k,$$

то система (5.3.9) принимает вид

$$\begin{gathered} L_{xx}(x_k, u_k)\Delta x_k + g_x^T(x_k)\Delta u_k = -L_x(x_k, u_k), \\ g_x(x_k)\Delta x_k = -g(x_k), \end{gathered}$$

или, используя матричную запись,

$$\begin{bmatrix} L_{xx}(x_k, u_k) & g_x^T(x_k) \\ g_x(x_k) & 0_{mm} \end{bmatrix} \begin{bmatrix} \Delta x_k \\ \Delta u_k \end{bmatrix} = - \begin{bmatrix} L_x(x_k, u_k) \\ g(x_k) \end{bmatrix}.$$

Отсюда сразу следует еще одна интерпретация метода последовательного квадратичного программирования. Фактически получается, что итерация метода по определению Δx_k и Δu_k есть не что иное, как ньютоновская итерация, примененная к решению системы $n+m$ нелинейных уравнений

$$L_x(x, u) = 0_n, \quad g(x) = 0_m.$$

Эти условия являются необходимыми условиями первого порядка для задачи (5.3.1).

Указанная связь метода последовательного квадратичного программирования с методом Ньютона позволяет сравнительно просто установить локальную сходимость метода, причем со сверхлинейной скоростью. Для этого нам потребуются достаточные условия второго порядка для задачи (5.3.1). Напомним, что точка $x_* \in X$ будет решением

задачи (5.3.1), если можно указать такой вектор множителей Лагранжа $u_* \in \mathbb{R}^m$, что $L_x(x_*, u_*) = 0_n$ и

$$\langle s, L_{xx}(x_*, u_*)s\rangle > 0 \tag{5.3.10}$$

для любого ненулевого вектора $s \in K(x_*) = \{s \in \mathbb{R}^n : \ g_x(x_*)s = 0_m\}$.

Теорема 5.3.1. *Пусть в точке $x_* \in X$, являющейся решением задачи (5.3.1), выполнены достаточные условия второго порядка (5.3.10). Пусть, кроме того, матрица Якоби $g_x(x_*)$ имеет полный ранг, равный m. Тогда метод последовательного квадратичного программирования локально сходится к точке $[x_*, u_*]$ со сверхлинейной скоростью.*

Доказательство. Покажем, что при сделанных предположениях матрица

$$\begin{bmatrix} L_{xx}(x_*, u_*) & g_x^T(x_*) \\ g_x(x_*) & 0_{mm} \end{bmatrix} \tag{5.3.11}$$

неособая. Для этого достаточно убедиться, что линейная система уравнений

$$\begin{bmatrix} L_{xx}(x_*, u_*) & g_x^T(x_*) \\ g_x(x_*) & 0_{mm} \end{bmatrix} \begin{bmatrix} s \\ u \end{bmatrix} = 0_{n+m} \tag{5.3.12}$$

имеет только тривиальное (нулевое) решение.

В самом деле, предположим от противного, что существует нетривиальное решение $z = [s, u]^T$ системы (5.3.12). Тогда согласно второму уравнению этой системы $g_x(x_*)s = 0_m$, т.е. $s \in K(x_*)$. Из первого уравнения (5.3.12) имеем

$$L_{xx}(x_*, u_*)s + g_x(x_*)^T u = 0_n. \tag{5.3.13}$$

Если $s \neq 0_n$, то, умножая это равенство слева на s^T, получаем

$$\langle s, L_{xx}(x_*, u_*)s\rangle + \langle s, g_x(x_*)^T u\rangle = \langle s, L_{xx}(x_*, u_*)s\rangle = 0,$$

что невозможно из-за нарушения условия (5.3.10). Если $s = 0_n$, обязательно $u \neq 0_m$. Равенство (5.3.13) в этом случае переходит в $g_x(x_*)^T u = 0_n$, что противоречит условиям теоремы, поскольку по предположению матрица $g_x(x_*)$ имеет полный ранг, равный m. Таким образом, система (5.3.12) обладает только тривиальным решением и, следовательно, матрица (5.3.11) неособая.

Утверждение о локальной сверхлинейной сходимости метода последовательного квадратичного программирования вытекает из общих

условий сходимости метода Ньютона для систем нелинейных уравнений. ■

В добавление к утверждению теоремы 5.3.1 скажем, что метод последовательного квадратичного программирования будет локально сходиться с квадратичной скоростью, если дополнительно потребовать, чтобы матрицы вторых производных функций $f(x)$ и $g^i(x)$, $1 \le i \le m$, были непрерывны по Липшицу. Это опять же вытекает из свойств метода Ньютона.

Как уже отмечалось, метод последовательного квадратичного программирования может применяется также и для задач, содержащих ограничения-неравенства. Например, если требуется решить задачу (5.1.1), то теперь вместо вспомогательной задачи (5.3.8) решается задача квадратичного программирования:

$$\begin{gathered} f(x_k) + \langle f_x(x_k), x - x_k\rangle + \tfrac{1}{2}\langle x - x_k, L_{xx}(x_k, u_k)(x - x_k)\rangle \;\to\; \min, \\ g(x_k) + g_x(x_k)\,(x - x_k) \le 0_m. \end{gathered} \tag{5.3.14}$$

В отличие от задачи с равенствами эту задачу уже нельзя свести к решению системы линейных уравнений вида (5.3.9), поэтому приходится решать задачу (5.3.14), привлекая численные методы квадратичного программирования. Если матрица $L_{xx}(x_k, u_k)$ положительно определенная, то задача (5.3.14) имеет единственное решение, разумеется, при непустом допустимом множестве. В случае, когда (5.1.1) — задача выпуклого программирования с сильно выпуклой целевой функцией $f(x)$, матрица $L_{xx}(x_k, u_k)$ на каждой итерации оказывается положительно определенной.

Упражнение 5. *Покажите, что решение задачи (5.3.8) не изменится, если в ней вместо $L_{xx}(x_k, u_k)$ взять матрицу*

$$L_{xx}(x_k, u_k) + c g_x^T(x_k) g_x(x_k),$$

где $c > 0$. Будет ли такая матрица положительно определенной для точек $[x_k, u_k]$, близких к точке Каруша–Куна–Таккера $[x_*, u_*]$, когда в ней выполняются условия теоремы 5.3.1 для задачи (5.3.1)?

Глава 6

Методы последовательной безусловной минимизации

Среди методов решения задач условной оптимизации широкое применение нашли *методы последовательной безусловной минимизации*. Основная идея этих методов состоит в том, чтобы свести решение задачи с ограничениями к однократной или многократной минимизации некоторой вспомогательной функции на всем пространстве или на множестве простой структуры. Существует много способов построения вспомогательных функций, все они приводят к различным численным методам. Среди них наиболее популярными являются *методы штрафных функций (внутренних или внешних)*, а также различные *методы центров (методы параметризации целевой функции)*.

6.1. Методы штрафных функций

Рассмотрим задачу условной оптимизации, в которой допустимое множество описывается функциональным образом с помощью ограничений типа неравенства

$$f_* = \min_{x \in X} f(x), \quad X = \left\{ x \in \mathbb{R}^n : \; g(x) \leq 0_m \right\}, \tag{6.1.1}$$

где функция $f(x)$ и вектор-функция $g(x) = [g^1(x), \ldots, g^m(x)]^T$ непрерывны на $\mathbb{R}^n$. Считаем, что *допустимое множество* X не пусто и что задача имеет решение. Множество всех точек из X, являющихся решением (6.1.1), обозначим X_*.

Идея «штрафования» ограничений — одна из основных в оптимизации. Она проста и заключается в переходе от исходной задачи условной

оптимизации (6.1.1) к минимизации вспомогательной функции, определенной на всем пространстве.

Пусть $\delta(x|X)$ — индикаторная функция множества X:

$$\delta(x|X) = \begin{cases} 0, & x \in X, \\ +\infty, & x \notin X. \end{cases}$$

Составим с ее помощью вспомогательную функцию

$$P(x) = f(x) + \delta(x|X)$$

и рассмотрим вместо (6.1.1) задачу безусловной минимизации:

$$\min_{x \in R^n} P(x). \tag{6.1.2}$$

Задачи (6.1.1) и (6.1.2) формально эквивалентны и можно в принципе было бы решать (6.1.2) вместо (6.1.1). Однако то, что теперь функция $P(x)$ задана на всем пространстве $\mathbb{R}^n$ не очень помогает, так как она разрывна и принимает бесконечные значения. Избавиться от этого недостатка можно, аппроксимируя функцию $\delta(x|X)$ последовательностью функций $\delta_k(x|X)$ с более хорошими с вычислительной точки зрения свойствами. Важно только, чтобы эта последовательность функций стремились к функции $\delta(x|X)$ в пределе. Тогда вместо (6.1.2) решается последовательность задач

$$\min_{x \in R^n} P_k(x), \quad P_k(x) = f(x) + \delta_k(x|X),$$

где

$$\lim_{k \to \infty} \delta_k(x|X) = \delta(x|X). \tag{6.1.3}$$

Разные подходы к выбору последовательности функций $\delta_k(x|X)$ приводят к разным классам методов, получивших название *методов штрафных функций.*

6.1.1. Методы внешних штрафных функций

Обратимся сначала к методам, построенным на основе внешней аппроксимации функции $\delta(x|X)$. В этих методах от последовательности функций $\{\delta_k(x|X)\}$ требуется, чтобы каждая из функций $\delta_k(x|X)$ была определена и непрерывна на всем пространстве $\mathbb{R}^n$. Кроме того, желательно, чтобы она обладала следующими свойствами:

1. $\delta_k(x|X) = 0, \quad x \in X;$

2. $\delta_k(x|X) > 0, \quad x \notin X;$

3. $\delta_{k+1}(x|X) > \delta_k(x|X), \quad x \notin X;$

4. $\lim_{k\to\infty} \delta_k(x|X) = +\infty, \quad x \notin X.$

Если эти свойства выполняются, то данная последовательность функций $\delta_k(x|X)$ удовлетворяет предельному равенству (6.1.3) (см. рис. 6.1).

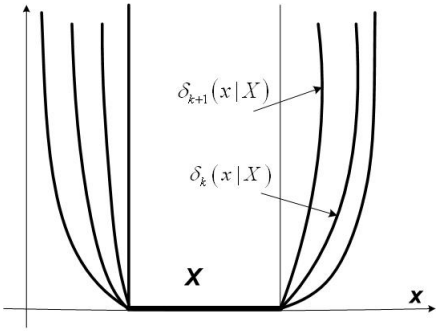

Рис. 6.1. Последовательность внешних аппроксимирующих функций

Встает вопрос, как наиболее просто строить эти последовательности функций $\{\delta_k(x|X)\}$. Рассмотрим один из возможных подходов.

Определение 6.1.1. *Штрафом (внешним) для произвольного множества $X \subseteq \mathbb{R}^n$ называется любая функция $\psi(x)$, удовлетворяющая условиям: $\psi(x) = 0$ при $x \in X$ и $\psi(x) > 0$ при $x \in \mathbb{R}^n \setminus X$.*

С помощью штрафа $\psi(x)$ строим семейство *внешних штрафных функций* достаточно общего вида:

$$P(x,t) = f(x) + t\psi(x), \tag{6.1.4}$$

где $t > 0$ — *коэффициент штрафа.* Если t увеличивается, то увеличивается и вклад добавки к целевой функции $f(x)$ в недопустимых точках в (6.1.4).

Функция $P(x,t)$ обладает следующими свойствами: $P(x,t) = f(x)$ при $x \in X$ и

$$\lim_{t\to\infty} P(x,t) = +\infty \quad \forall x \notin X.$$

Таким образом, в качестве $\delta_k(x|X)$ можно брать

$$\delta_k(x|X) = t_k\psi(x),$$

где $\{t_k\}$ — монотонно возрастающая последовательность положительных чисел, стремящихся к бесконечности, т.е. $t_{k+1} > t_k$ и $t_k \to +\infty$.

Рассмотрим теперь способ, с помощью которого достаточно просто строить внешние штрафы.

Определение 6.1.2. *Непрерывную функцию $B(g)$, определенную на $\mathbb{R}^m$, назовем внешней свертывающей функцией, если $B(g) = 0$ при $g \in \mathbb{R}^m_-$ и $B(g) > 0$ для всех $g \notin \mathbb{R}^m_-$.*

В качестве внешних свертывающих функций могут использоваться, например, следующие функции:

$$B(g) = \|g_+\|_p, \quad B(g) = \frac{1}{p}\|g_+\|_p^p, \tag{6.1.5}$$

где $\|\cdot\|_p$ — p-я гельдеровская норма, $1 \le p < \infty$, и

$$g_+ = [g_+^1, \ldots, g_+^m]^T, \quad g_+^i = \max[0, g^i], \quad 1 \le i \le m.$$

Обе функции (6.1.5) являются неубывающими на $\mathbb{R}^m$, т.е. из $g_2 \ge g_1$ следует, что $B(g_2) \ge B(g_1)$. Вторая из них при $p > 1$ непрерывно дифференцируема на $\mathbb{R}^m$. Параметр p у первой функции можно полагать также равным $+\infty$. Тогда $B(g) = \max_{1 \le i \le m} |g_+^i|$.

Беря внешнюю свертывающую функцию $B(g)$, получаем внешний штраф $\psi(x) = B(g(x))$. Подставляя этот внешний штраф в (6.1.4), приходим к *внешней штрафной функции*:

$$P(x,t) = f(x) + tB(g(x)). \tag{6.1.6}$$

На рис. 6.2 показан вид внешней штрафной функции при разных значениях коэффициента штрафа.

Обозначим через $X(t)$ множество точек, доставляющих минимум функции $P(x,t)$ на $\mathbb{R}^n$ при фиксированном значении коэффициента штрафа t, т.е. множество

$$X(t) = \operatorname{Arg}\min_{x \in R^n} P(x,t).$$

Определение 6.1.3. *Функция $P(x,t)$ называется точной внешней штрафной функцией для задачи (6.1.1), если существует такое множество T значений коэффициента штрафа t, что $X(t) = X_*$ при $t \in T$.*

Составим для задачи (6.1.1) функцию Лагранжа:

$$L(x,u) = f(x) + \langle g(x), u\rangle, \quad u \in \mathbb{R}^m_+.$$

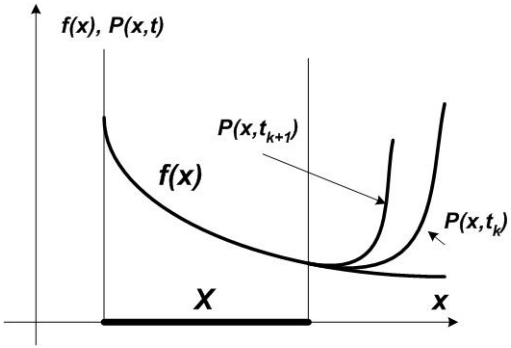

Рис. 6.2. Последовательность внешних штрафных функций

Точка $[x_*, u_*]$ является седловой точкой функции Лагранжа, если

$$L(x_*, u) \le L(x_*, u_*) \le L(x, u_*) \quad \forall x \in \mathbb{R}^n, \quad \forall u \in \mathbb{R}^m_+. \tag{6.1.7}$$

Как следует из достаточных условий оптимальности, если у функции Лагранжа существует седловая точка $[x_*, u_*]$, то точка x_* является решением задачи (6.1.1). При этом $f_* = L(x_*, u_*)$.

Теорема 6.1.1. *Пусть в задаче (6.1.1) существует седловая точка $[x_*, u_*]$ функции Лагранжа. Пусть, кроме того, внешняя сворачивающая функция $B(g)$ такова, что*

$$0 < B^0(u_*) < +\infty,$$

где $B^0(u)$ — полярная функция к функции $B(g)$. Тогда $P(x,t)$ — точная штрафная функция для задачи (6.1.1) и

$$T = \{t \in \mathbb{R} : \ t > B^0(u_*)\}.$$

Доказательство. Имеем из правого неравенства в определении (6.1.7) седловой точки

$$f_* = f(x_*) = L(x_*, u_*) \le L(x, u_*) = f(x) + \langle g(x), u_* \rangle. \tag{6.1.8}$$

Воспользуемся неравенством Минковского–Малера, чтобы оценить скалярное произведение в правой части (6.1.8). Тогда

$$f_* \le f(x) + \langle g(x), u_* \rangle \le f(x) + B(g(x)) B^0(u_*).$$

Так как функция $B(g(x))$ неотрицательна на $\mathbb{R}^n$, то, беря $t \geq t_*$, где $t_* = B^0(u_*)$, получаем

$$f_* \leq f(x) + B(g(x))B^0(u_*) \leq f(x) + tB(g(x)) = P(x,t).$$

Таким образом, $P(x,t) \geq f_*$ для всех $x \in \mathbb{R}^n$ и любого $t \geq t_*$.

Но $B(g(x_*)) = 0$, поскольку $x_* \in X$. В этом случае справедливо равенство $P(x_*,t) = f(x_*) = f_*$. Отсюда заключаем, что $x_* \in X(t)$ и, следовательно, $X_* \subseteq X(t)$ для любого $t \geq t_*$. Это, в частности, означает:

$$P(x,t) = f_* \quad \forall\, x \in X(t), \quad t \geq t_*. \tag{6.1.9}$$

Покажем, что при $t > t_*$ множества X_* и $X(t)$ совпадают. Доказательство проведем от противного: предположим, что существуют $t_1 > t_*$ и $\bar{x} \in X(t_1)$ такие, что $\bar{x} \notin X_*$. Точка $\bar{x}$ не может принадлежать допустимому множеству X, так как иначе из (6.1.9) следовало бы, что $f(\bar{x}) = f_*$. Поэтому обязательно $\bar{x} \notin X$. Но тогда

$$f_* = f(\bar{x}) + t_1 B(g(\bar{x})), \tag{6.1.10}$$

причем $B(g(\bar{x})) > 0$.

Возьмем теперь $t_* < t_2 < t_1$. Так как $t_1 > t_*$, то такое t_2 всегда найдется. Тогда на основании (6.1.10)

$$P(\bar{x},t_2) = f(\bar{x}) + t_2 B(g(\bar{x})) < f(\bar{x}) + t_1 B(g(\bar{x})) = f_*,$$

что противоречит сказанному выше. Таким образом, действительно имеет место совпадение множеств $X(t)$ и X_* при $t > t_*$. ■

Если внешняя штрафная функция $P(x,t)$ имеет вид

$$P(x,t) = f(x) + t\|g_+(x)\|_p, \quad 1 < p < \infty,$$

то $B(g) = \|g_+\|_p$. Поскольку $u_* \geq 0_m$, то, как было показано в [33], $B^0(u_*) = \|u_*\|_{p_*}$, где числа p и p_* связаны соотношением $p^{-1} + p_*^{-1} = 1$. Таким образом, $P(x,t)$ является точной внешней штрафной функцией при $t > \|u_*\|_{p_*}$.

Отметим также, что для того, чтобы внешняя штрафная функция $P(x,t)$ вида (6.1.6) была бы точной для задачи (6.1.1), в общем случае требуется, чтобы она была негладкой.

Опишем теперь основной вариант метода внешней штрафной функции, не предполагая, что $P(x,t)$ является точной внешней штрафной функцией.

Алгоритм метода внешних штрафных функций. Задаем монотонно возрастающую последовательность коэффициентов штрафа $t_k \to +\infty$ и полагаем $k = 1$.

Общая k-я итерация

Шаг 1. Решая задачу безусловной минимизации

$$\min_{x \in R^n} P(x, t_k), \tag{6.1.11}$$

находим точку $x_k \in X(t_k)$.

Шаг 2. Полагаем $k := k + 1$ и идем на шаг 1.

Если все задачи безусловной минимизации (6.1.11) имеют решения, то в результате работы алгоритма получается последовательность точек $\{x_k\}$. Желательно, чтобы эта последовательность имела предельные точки и чтобы каждая предельная точка x_* была бы решением задачи (6.1.1).

Рассмотрим вопрос о сходимости метода внешних штрафных функций. Предположим, что все множества $X(t)$ при $t \geq 0$ ограничены в совокупности, т.е. принадлежат ограниченному множеству $\tilde{X}$. Данное условие заведомо выполняется, если существует такая точка $\bar{x} \in X$, что множество $\mathcal{L}(\bar{x}) = \left\{x \in \mathbb{R}^n : \ f(x) \leq f(\bar{x})\right\}$ ограничено. Тогда последовательность $\{x_k\}$ принадлежит $\tilde{X}$ и, следовательно, имеет предельные точки.

Теорема 6.1.2. *Пусть x_* — предельная точка последовательности $\{x_k\}$. Тогда $x_* \in X_*$ и*

$$\lim_{k \to \infty} P(x_k, t_k) = f(x_*) = f_*. \tag{6.1.12}$$

Доказательство. Не умаляя общности, считаем, что сама последовательность $\{x_k\}$ является сходящейся, т.е. $x_* = \lim_{k \to \infty} x_k$.

Учтем, что $P(x, t) = f(x)$, когда $x \in X$. Примем также во внимание, что $P(x, t_k) \to \infty$, когда $x \notin X$ и $t_k \to \infty$. Тогда

$$\begin{aligned}
&\min_{x \in R^n} \lim_{t_k \to \infty} P(x, t_k) = \min_{x \in X} \lim_{t_k \to \infty} P(x, t_k) = \\
&\quad = \min_{x \in X} f(x) = \lim_{t_k \to \infty} \min_{x \in X} f(x) = \\
&\quad = \lim_{t_k \to \infty} \min_{x \in X} P(x, t_k) \geq \\
&\quad \geq \lim_{t_k \to \infty} \min_{x \in R^n} P(x, t_k) = \lim_{k \to \infty} P(x_k, t_k).
\end{aligned}$$

Отсюда, в частности, следует

$$\lim_{k \to \infty} P(x_k, t_k) \leq \min_{x \in X} f(x) = f_*. \tag{6.1.13}$$

Покажем, что $x_* \in X$. Действительно, если допустить противное, то $\psi(x_*) > 0$, и, в силу непрерывности, найдется такое $c > 0$, что $\psi(x_k) \geq c$ для достаточно больших номеров k. Но последовательность $\{|f(x_k)|\}$ ограничена, поскольку ограничена последовательность $\{x_k\}$ и функция $f(x)$ непрерывна. Поэтому

$$\begin{aligned}\lim_{k\to\infty} P(x_k, t_k) = \lim_{k\to\infty} [f(x_k) + t_k\psi(x_k)] \geq \\ \geq \lim_{k\to\infty} [f(x_k) + ct_k] = +\infty.\end{aligned}$$

Это противоречит неравенству (6.1.13).

Наконец, из неотрицательности функции $\psi(x)$ и непрерывности $f(x)$ следует

$$\lim_{k\to\infty} P(x_k, t_k) = \lim_{k\to\infty} [f(x_k) + t_k\psi(x_k)] \geq \lim_{k\to\infty} f(x_k) = f(x_*).$$

Таким образом,

$$f_* \geq \lim_{k\to\infty} P(x_k, t_k) \geq f(x_*).$$

Но $x_* \in X$. Отсюда, поскольку $f(x_*) \geq f_*$, заключаем, что $f(x_*) = f_*$ и имеет место предельное равенство (6.1.12). ■

Предположим, что в задаче (6.1.1) существует седловая точка $[x_*, u_*]$ функции Лагранжа $L(x, u)$. Тогда, используя функцию, сопряженную к функции $B(g)$, получаем с помощью неравенства Юнга–Фенхеля:

$$\begin{aligned} f_* &= f(x_*) = L(x_*, u_*) \leq L(x, u_*) = f(x) + \langle g(x), u_*\rangle \leq \\ &\leq f(x) + tB(g(x)) + (tB)^*(u_*) = P(x, t) + (B^*t)(u_*). \end{aligned} \tag{6.1.14}$$

Здесь $(B^*t)(u)$ — функция $B^*(u)$, умноженная справа на константу $t > 0$. По определению такой операции $(B^*t)(u) = tB^*(\frac{u}{t})$.

Неравенство (6.1.14) справедливо для любого $x \in \mathbb{R}^n$. Перепишем его в виде

$$P(x, t) \geq f_* - (B^*t)(u_*). \tag{6.1.15}$$

Обозначим правую часть этого неравенства через $q(t)$. Так как она не зависит от x, то из (6.1.15) получаем

$$\min_{x\in R^n} P(x, t) \geq f_* - (B^*t)(u_*) = q(t).$$

Возьмем в качестве $B(g)$ функцию

$$B(g) = \frac{1}{p}\|g_+\|_p^p, \quad 1 < p < \infty. \tag{6.1.16}$$

Тогда при $u \geq 0_m$:

$$B^*(u) = \frac{1}{p_*}\|u\|_{p_*}^{p_*}, \quad p^{-1} + p_*^{-1} = 1,$$

и поскольку $1 < p_* < \infty$, то для оценки снизу $q(t)$ имеем: $q(t) < f_*$ и

$$q(t) = f_* - \frac{\|u_*\|_{p_*}^{p_*}}{p_* t^{(p_*-1)}} \to f_*$$

при $t \to \infty$.

В частном случае, когда $p = 2$, функция $P(x,t)$ принимает вид

$$P(x,t) = f(x) + \frac{t}{2}\sum_{i=1}^{m}\left(g_+^i(x)\right)^2, \tag{6.1.17}$$

а оценка $q(t)$ с учетом того, что $p_* = 2$ при $p = 2$, становится равной

$$q(t) = f_* - \frac{1}{2t}\|u_*\|_2^2.$$

Функция (6.1.17) носит название *квадратичной штрафной функции*. Она является одной из основных вспомогательных функций, применяемой в методах внешних штрафных функции.

К достоинствам метода внешних штрафных функций следует отнести также то, что он применим для решения задач, содержащих помимо ограничений типа неравенства также ограничения типа равенства

$$\min_{x\in X} f(x), \quad X = \{x \in \mathbb{R}^n : \ g(x) \leq 0_m, \ h(x) = 0_l\}, \tag{6.1.18}$$

где $h : \mathbb{R}^n \to \mathbb{R}^l$. Для этой задачи квадратичная штрафная функция принимает вид

$$P(x,t) = f(x) + \frac{t}{2}\left[\sum_{i=1}^{m}\left(g_+^i(x)\right)^2 + \sum_{j=1}^{l}|h^j(x)|^2\right].$$

Заметим также, что если в задаче дополнительно присутствует ограничение простого вида $x \in \Pi$, например, допустимое множество X задается следующим образом:

$$X = \{x \in \Pi : \ g(x) \leq 0_m, \ h(x) = 0_l\},$$

то целесообразно не включать это ограничение в число общих функциональных ограничений, а вспомогательную задачу минимизации функции $P(x, t_k)$ на всем пространстве $\mathbb{R}^n$ заменить на задачу ее минимизации на множестве Π, т.е. вместо (6.1.11) решать задачу

$$\min_{x \in \Pi} P(x, t_k).$$

Все основные свойства алгоритма метода внешних штрафных функций при этом сохраняются.

6.1.2. Методы внутренних штрафных функций

Методы внутренних штрафных функций (называемые также *методами барьерных функций*) применяются только для решения задач с ограничениями типа неравенства, т.е. для задач вида (6.1.1). Пусть

$$X_0 = \left\{ x \in \mathbb{R}^n : \ g(x) < 0_m \right\}.$$

Считаем, что множество X *регулярно*, т.е. множество X_0 не пусто и $X_0 = \mathrm{int} X$. Из этого предположения следует, что $\mathrm{cl} X_0 = X$. Обозначим через ∂X границу множества X.

Относительно последовательности функций $\{\delta_k(x|X)\}$ теперь будем предполагать, что каждая функция $\delta_k(x|X)$ является непрерывной и определена на $\mathbb{R}^n$, но конечные значения принимает только на X_0. Наделим ее следующими свойствами (см. рис. 6.3):

1. $\delta_k(x|X) < +\infty$ для всех $x \in X_0$.

2. $\delta_k(x|X) = +\infty$ для всех $x \notin X_0$.

Кроме того, наложим на $\{\delta_k(x|X)\}$ следующие требования:

3. $\lim_{k \to \infty} |\delta_k(x|X)| = 0$ для всех $x \in X_0$.

4. если $\{\delta_k(x)\}$ — последовательность положительных на X_0 функций, то $\delta_k(x|X) > \delta_{k+1}(x|X)$ для любого $x \in X_0$.

Согласно этим условиям $\delta_k(\bar{x}|X) = +\infty$ для всех $\bar{x} \in \partial X$, причем выполняется предельное равенство

$$\lim_{x \to \bar{x}, \ x \in X_0} \delta_k(x|X) = +\infty.$$

Поэтому аппроксимация функции $\delta(x|X)$ последовательностью функций $\{\delta_k(x|X)\}$ осуществляется только с точностью до границы, так как

$$\lim_{k \to \infty} \delta_k(x|X) \neq \delta(x|X), \quad x \in X \setminus X_0.$$

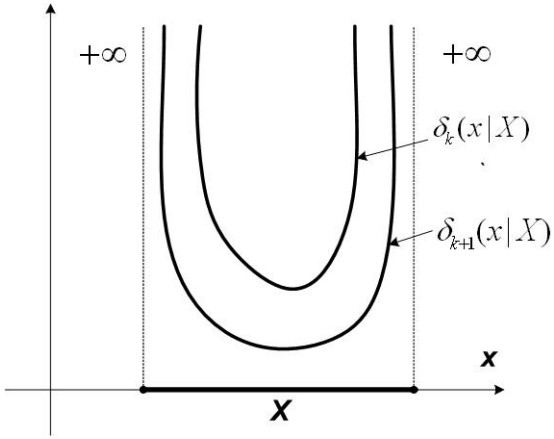

Рис. 6.3. Последовательность внутренних аппроксимирующих функций

Определение 6.1.4. *Внутренним штрафом или барьером будем называть непрерывную функцию* $\psi(x)$ *такую, что* $\psi(x) < +\infty$ *тогда и только тогда, когда* $x \in X_0$, *причем* $\lim_{x\to\bar{x},\ x\in X_0} \psi(x) = +\infty$, *если* $\bar{x} \in \partial X$.

Используя барьер, можно построить *внутреннюю штрафную функцию (барьерную функцию)* в виде

$$P(x,t) = f(x) + t\psi(x),$$

где $t > 0$ — коэффициент внутреннего штрафа.

Обозначим теперь

$$X(t) = \operatorname{Arg}\min_{x\in X_0} P(x,t).$$

Алгоритм метода внутренних штрафных функций (барьерных функций). Задаем монотонно убывающую последовательность коэффициентов $t_k \downarrow 0$ и полагаем $k = 1$.

Общая k*-я итерация*

Шаг 1. Решаем вспомогательную задачу

$$\min_{x\in X_0} P(x,t_k) \tag{6.1.19}$$

и находим точку $x_k \in X(t_k)$.

Шаг 2. Полагаем $k := k + 1$ и идем на шаг 1.

В результате работы алгоритма получаем последовательность точек $\{x_k\} \subset X_0$. Если она ограничена, то имеет предельные точки $x_* \in X$.

Далее будем предполагать, что X — компактное множество. Тогда точки $x_k \in X(t_k)$ существуют и $x_k \in X_0$. Полагаем также для простоты, что $\psi(x) \geq 0$ для всех $x \in X_0$.

Теорема 6.1.3. *Любая предельная точка x_* последовательности $\{x_k\}$ есть решение задачи (6.1.1) и*

$$\lim_{k\to\infty} P(x_k, t_k) = f(x_*) = f_*.$$

Доказательство. Не ограничивая общности, считаем, что последовательность $\{x_k\}$ является сходящейся. Тогда $x_k \to x_*$. Так как X — замкнутое множество, то $x_* \in X$.

Если $x_* \notin X_*$, то из предположения $\mathrm{cl}X_0 = X$ следует, что найдется $y \in X_0$ такое, что $f(y) < f(x_*)$ и, значит, $f(y) < \lim_{k\to\infty} f(x_k)$. Поэтому можно указать такое $\delta > 0$, что $f(x_k) > f(y) + \delta$ для k достаточно больших. Но тогда, поскольку $\psi(x) \geq 0$, для этих k справедлива оценка:

$$f(y) + \delta < f(x_k) \leq f(x_k) + t_k\psi(x_k) = P(x_k, t_k). \tag{6.1.20}$$

Кроме того, для $y \in X_0$ из-за того, что

$$\lim_{k\to\infty} t_k\psi(y) = 0, \tag{6.1.21}$$

имеет место равенство

$$\lim_{k\to\infty} P(y, t_k) = f(y). \tag{6.1.22}$$

Отсюда и из (6.1.20) получаем, что для достаточно больших k выполняются неравенства: $P(y, t_k) < P(x_k, t_k)$, что противоречит определению точки $x_k \in X(t_k)$, поскольку $y \in X_0$. Таким образом, $x_* \in X_*$.

Далее, опять в силу определения точек $x_k \in X(t_k)$,

$$\begin{aligned} P(x_{k+1}, t_{k+1}) &= f(x_{k+1}) + t_{k+1}\psi(x_{k+1}) \leq \\ &\leq f(x_k) + t_{k+1}\psi(x_k) \leq \\ &\leq f(x_k) + t_k\psi(x_k) = P(x_k, t_k), \end{aligned}$$

т.е. последовательность $\{P(x_k, t_k)\}$ является невозрастающей. Но она ограничена снизу, поэтому существует предел

$$\lim_{k\to\infty} P(x_k, t_k) = P_* \geq f_*.$$

Покажем, что строгое неравенство $P_* > f_*$ невозможно и на самом деле имеет место равенство $P_* = f_*$. Действительно, иначе нашлись бы ε-окрестность $\Delta_\varepsilon(X_*)$ множества X_* и $\delta > 0$ такие, что для некоторого $y \in \Delta_\varepsilon(X_*) \cap X_0$ оказываются справедливыми неравенства

$$P(x_k, t_k) > f(y) + \delta \tag{6.1.23}$$

для k достаточно больших. Но согласно (6.1.21)

$$P(y, t_k) = f(y) + t_k \psi(y) \leq f(y) + \frac{\delta}{2},$$

если k достаточно большое. Отсюда и из (6.1.23)

$$P(x_k, t_k) > P(y, t_k) + \frac{\delta}{2}$$

для этих k, что невозможно в силу определения точек x_k. ■

Встает вопрос, как строить внутренние штрафы $\psi(x)$? Будем поступать аналогично, тому, как это делалось в методе внешних штрафных функций, используя аппарат свертывающих функций.

Пусть $\bar{\mathbb{R}}$ — расширенная прямая, т.е. прямая $\mathbb{R}$, пополненная элементами $\{+\infty\}$ и $\{-\infty\}$. Пусть, кроме того, $\mathbb{R}^m_{--}$ — внутренность ортанта $\mathbb{R}^m_-$. Функцию $Q(z)$, отображающую $\mathbb{R}^m$ в $\bar{\mathbb{R}}$, назовем непрерывной на $\mathbb{R}^m$, если она непрерывна в обычном смысле во всех точках, в которых она конечна, а в остальных точках $z \in \mathbb{R}^m$, в которых она принимает бесконечные значения, $\lim\limits_{z_k \to z} Q(z_k) = Q(z)$ для любой последовательности $z_k \to z$.

Определение 6.1.5. *Непрерывная функция $B(g) : \mathbb{R}^m \to \bar{\mathbb{R}}$ называется внутренней свертывающей функцией, если $B(g) < +\infty$ при $g \in \mathbb{R}^m_{--}$ и $B(g) = +\infty$ при $g \notin \mathbb{R}^m_{--}$.*

В качестве примера внутренних свертывающих функций приведем следующую функцию:

$$B(g) = -\frac{1}{p}\|g_-\|_p^p, \quad g \in \mathbb{R}^m_{--}, \tag{6.1.24}$$

где $-\infty < p < 0$ и g_- — отрицательная срезка вектора $g \in \mathbb{R}^m$, т.е. $g^i_- = \min[g^i, 0]$ для всех $1 \leq i \leq m$. Кроме того, полагаем $B(g) = +\infty$, когда $g \notin \mathbb{R}^m_{--}$. Разумеется, функция $\|g\|_p$ при $p < 0$ уже нормой не является, однако для однообразия обозначений знак нормы здесь сохранен.

Если воспользоваться функцией $\|g\|_0$, определенной как

$$\|g\|_0 = \sqrt{m}\left(\prod_{i=1}^{m}|g^i|\right)^{\frac{1}{m}}, \quad g \in \mathbb{R}^m, \tag{6.1.25}$$

то приходим к внутренней свертывающей функции, дополняющей семейство функции (6.1.24) при значении параметра $p = 0$:

$$B(g) = -\ln\left(\|g_-\|_0\right) - \frac{1}{2}, \quad g \in \mathbb{R}^m_{--}. \tag{6.1.26}$$

Здесь опять считается, что $B(g) = +\infty$ при $g \notin \mathbb{R}^m_{--}$.

Используя внутренние свертывающие функции, строим внутренние штрафные (барьерные) функции в виде

$$P(x,t) = f(x) + tB(g(x)), \quad t > 0. \tag{6.1.27}$$

В частном случае, беря внутреннюю свертывающую функцию (6.1.24) при значении параметра $p = -1$, приходим к внутренней штрафной функции

$$P(x,t) = f(x) + t\sum_{i=1}^{m}\frac{1}{-g^i(x)}, \quad x \in X_0,$$

которая носит название *обратной барьерной функции.*

Аналогичным образом, используя внутреннюю свертывающую функцию (6.1.25), получаем другую внутреннюю штрафную функцию

$$P(x,t) = f(x) - t\left[\frac{1}{m}\sum_{i=1}^{m}\ln\left(-g^i(x)\right) + \frac{1}{2}\left(1 + \ln m\right)\right], \quad x \in X_0. \tag{6.1.28}$$

Если опустить константы, которые не влияют на решение вспомогательных задач минимизации (6.1.19), то функцию (6.1.28) можно переписать как

$$P(x,t) = f(x) - t\sum_{i=1}^{m}\ln\left(-g^i(x)\right), \quad x \in X_0.$$

Данную функцию принято называть *логарифмической барьерной функцией.*

Получим оценки оптимального значения вспомогательной функции $M(x,y)$ в задаче минимизации (6.1.19) в предположении, что в задаче нелинейного программирования (6.1.1) существует седловая точ-

ка $[x_*, u_*]$ функции Лагранжа. Вновь применяя неравенство Юнга–Фенхеля, как и в (6.1.14), приходим к оценке

$$\min_{x \in X_0} P(x,t) \geq f_* - (B^*_{R^m_-} t)(u_*) = q(t). \tag{6.1.29}$$

Пусть внутренняя свертывающая функция $B(g)$ имеет вид (6.1.24), где $-\infty < p < 0$. В [33] было показано, что сопряженной к такой функции $B(g)$ является функция

$$B^*_{R^m_-}(u) = -\frac{1}{p_*} \|u\|^{p_*}_{p_*},$$

которая определена на $\mathbb{R}^m_+$. При этом p и p_* связаны тем же двойственным соотношением $p^{-1} + p_*^{-1} = 1$, что и для $p > 1$. Тогда имеем

$$q(t) = f_* + \frac{t}{p_*} \left\| \frac{u_*}{t} \right\|^{p_*}_{p_*} = f_* + \frac{t^{1-p_*}}{p_*} \|u_*\|^{p_*}_{p_*}. \tag{6.1.30}$$

Так как $0 < p_* < 1$ при $p < 0$, то в силу (6.1.30) нижняя оценка $q(t)$ превышает f_* и стремится к f_* при $t \downarrow 0$. В частном случае при $p = -1$ для обратной барьерной функции $P(x,t)$ получаем, что

$$q(t) = f_* + 2\sqrt{t} \sum_{i=1}^{m} \sqrt{u^i_*}. \tag{6.1.31}$$

Если в качестве внутренней свертывающей функции $B(g)$ используется функция (6.1.26), то для нее, как было показано в [33],

$$B^*_{R^m_-}(u) = -\ln(\|u\|_0) - \frac{1}{2}, \quad u \in \mathbb{R}^m_+,$$

т.е. она в некотором смысле самосопряженная, только u принадлежит другому множеству, а именно $u \in \mathbb{R}^m_+$. Поэтому в случае, когда $p = 0$, для общей логарифмической барьерной функции (6.1.28) имеем соответственно

$$q(t) = f_* + t \left[\frac{1}{m} \sum_{i=1}^{m} \ln u^i_* - \ln t + \frac{1}{2}(1 + \ln m) \right]. \tag{6.1.32}$$

Следовательно, нижняя оценка $q(t)$ опять стремится к f_* при $t \downarrow 0$. Оценка (6.1.32) в отличие от (6.1.31) справедлива в том и только в том случае, когда $u_* > 0_m$, т.е. если в точке x_* все ограничения активны.

Можно по аналогии с точными внешними штрафными функциями рассмотреть точные внутренние штрафные функции вида (6.1.27),

однако для этого нам потребуется уточнить понятие внутренней свертывающей функции.

Определение 6.1.6. *Функция $B(g) : \mathbb{R}^m \to \mathbb{R}$ называется собственной внутренней свертывающей функцией, если она непрерывна на $\mathbb{R}^m$ и $B(g) < 0$ при $g \in \mathbb{R}^m_{--}$ и $B(g) = 0$ при $g \notin \mathbb{R}^m_{--}$.*

Подчеркивая отличие собственной внутренней свертывающей функции от внутренней свертывающей функции из определения 6.1.5, о последней будем говорить как о *несобственной* внутренней свертывающей функции. Примером собственной внутренней свертывающей функции $B(g)$ является функция

$$B(g) = -\|g\|_p, \quad -\infty \le p \le 0, \tag{6.1.33}$$

причем $B(g) = -\min_{1\le i\le m} |g^i|$, когда $p = -\infty$.

Пусть теперь $X(t)$ множество минимумов функции (6.1.27) по x на всем допустимом множестве X.

Определение 6.1.7. *Функция $P(x,t)$ называется точной внутренней штрафной функцией для задачи (6.1.1), если существует такое множество T значений коэффициента штрафа t, что $X(t) = X_*$ при $t \in T$.*

Теорема 6.1.4. *Пусть в задаче (6.1.1) существует седловая точка $[x_*, u_*]$ функции Лагранжа $L(x,u)$ на $\mathbb{R}^n \times \mathbb{R}^m_+$. Пусть, кроме того, собственная внутренняя свертывающая функция $B(g)$ такова, что*

$$0 < B^0_{R^m_-}(u_*) < +\infty.$$

Тогда $P(x,t)$ — точная внутренняя штрафная функция для задачи (6.1.1) и

$$T = \{0 \le t < t^* : \ t^* = B^0_{R^m_-}(u_*)\}.$$

Доказательство. Так как функция $B(g)$ неположительна на $\mathbb{R}^m_-$, то на основании определения седловой точки и неравенства Минковского–Малера приходим к оценке

$$\begin{aligned} f_* \le f(x) + \langle g(x), u_*\rangle \le f(x) + B(g(x))B^0_{R^m_-}(u_*) \le \\ \le f(x) + tB(g(x)) = P(x,t), \end{aligned} \tag{6.1.34}$$

справедливой для любых $x \in X$ и $t \in T$.

Кроме того, так как $P(x(t),t) \le P(\bar{x},t)$, где $x(t) \in X(t)$ и $\bar{x}$ — произвольная точка из X, то, взяв $\bar{x} = x_* \in X_*$, получаем

$$P(x(t),t) \le P(x_*,t) = f(x_*) + tB(g(x_*)) \le f(x_*) = f_*. \qquad (6.1.35)$$

На основании (6.1.34) и (6.1.35) делаем вывод, что

$$P(x(t),t) = f_*. \qquad (6.1.36)$$

Поскольку $P(x_*,t) \le f_*$, то наряду с (6.1.36) имеет место равенство $P(x_*,t) = f_*$. Следовательно, $X_* \subseteq X(t)$. Аналогично тому, как это делалось при доказательстве теоремы 6.1.1, можно убедиться, что на самом деле $X(t) = X_*$ при $t \in T$. ■

Приведем два примера точных внутренних штрафных функций, которые получаются, если в качестве $B(g)$ брать функцию (6.1.33) при значениях параметра p, равных соответственно $p = 0$ и $p = -\infty$:

$$P(x,t) = f(x) - t\sqrt{m}\sqrt[m]{\prod_{i=1}^{m}(-g^i(x))}, \quad t^* = \sqrt[m]{\prod_{i=1}^{m} u_*^i},$$

$$P(x,t) = f(x) - t\min_{1\le i\le m}(-g^i(x)), \quad t^* = \sum_{i=1}^{m} u_*^i.$$

Понятно, что точные внутренние штрафные функции с вычислительной точки зрения особого смысла не имеют и интересны главным образом лишь как теоретический результат. Они отражают тот факт, что, видоизменяя целевую функцию не очень сильно, мы сохраняем решение задачи. Данное свойство характерно для задач, обладающих острым минимумом.

Отметим также, что в основном методы внутренних штрафных функций применяются в тех случаях, когда целевая функция $f(x)$ не определена за пределами допустимого множества X. В последнее время идеи методов внутренних штрафных функций нашли широкое применение в так называемых *методах внутренних точек*, разработанных для решения задач линейного программирования, а также для решения более общих задач выпуклой минимизации.

6.2. Методы параметризации целевой функции

Метод параметризации целевой функции (называемый также *методом нагруженного функционала, методом внешних центров, ме-*

тодом Моррисона) близок по своим свойствам к методу внешних штрафных функций. Рассмотрим опять для простоты задачу нелинейного программирования с ограничениями типа неравенства

$$f_* = \min_{x \in X} f(x), \quad X = \left\{ x \in \mathbb{R}^n : g(x) \leq 0_n \right\}. \tag{6.2.1}$$

Считаем, что задача имеет решение и обозначим через X_* множество ее решений.

Пусть η — оценка снизу оптимального значения целевой функции f_*. Если $\eta < f_*$, то система неравенств

$$f(x) - \eta \leq 0, \quad g(x) \leq 0_m$$

не имеет решения. Поэтому, беря произвольный внешний штраф $\psi(x)$, например, $\psi(x) = B(g(x))$, где $B(g)$ — внешняя свертывающая функция, а также квадратичный внешний штраф $(f(x) - \eta)_+^2$ для первого неравенства, приходим к тому, что функция

$$M_1(x, \eta) = (f(x) - \eta)_+^2 + B(g(x)) \tag{6.2.2}$$

принимает для всех $x \in \mathbb{R}^n$ только положительные значения как, впрочем, и функция

$$M_2(x, \eta) = (f(x) - \eta)_+^2 + (B(g(x)))_+^2 > 0.$$

Здесь знак «положительной срезки» плюс у функции $B(g(x))$ поставлен формально для единообразия, так как сама функция $B(g(x))$, будучи внешним штрафом, принимает только неотрицательные значения всюду на $\mathbb{R}^n$.

Введем сначала в рассмотрение вспомогательную функцию

$$M(x, \eta) = \sqrt{(f(x) - \eta)_+^2 + (B(g(x)))_+^2}, \tag{6.2.3}$$

где $x \in \mathbb{R}^n$ и $\eta \in \mathbb{R}$. Для данной вспомогательной функции поставим задачу безусловной минимизации:

$$\phi(\eta) = \min_{x \in R^n} M(x, \eta). \tag{6.2.4}$$

Предполагаем, что задача (6.2.4) имеет решение для всех $\eta \in \mathbb{R}$.

Из сказанного выше следует, что $M(x, \eta) > 0$ для любого $x \in \mathbb{R}^n$, если $\eta < f_*$. Поэтому $\phi(\eta) > 0$ для всех таких η. С другой стороны,

беря $\eta \geq f_*$, получаем $\phi(\eta) = 0$. Действительно, функция $M(x, \eta) \geq 0$ неотрицательна для всех $x \in \mathbb{R}^n$. Но, взяв точку $x_* \in X_* \subseteq X$, имеем

$$(f(x_*) - \eta)_+^2 = 0, \quad B(g(x_*)) = 0, \quad M(x_*, \eta) = 0,$$

откуда в силу неотрицательности $M(x, \eta)$ следует, что $\phi(\eta) = 0$.

На основании приведенных рассуждений можно поставить задачу поиска наименьшего корня уравнения

$$\phi(\eta) = 0. \tag{6.2.5}$$

Минимальный корень очевидно равен f_* (см. рис. 6.4). Любая точка x_*, которая решает задачу (6.2.4) при $\eta = f_*$, такова, что $x_* \in X_*$.

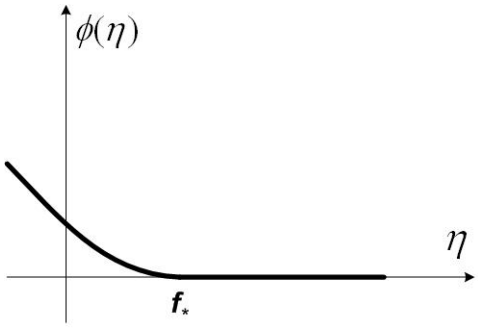

Рис. 6.4. Типичный вид функции $\phi(\eta)$

Встает вопрос, как искать этот минимальный корень? Приведем сначала вариант метода, который основан по существу на использовании метода простой итерации для решения уравнения (6.2.5). Обозначим через $X(\eta)$ множество решений задачи (6.2.4) при фиксированном значении оценки η, т.е. множество

$$X(\eta) = \operatorname{Arg} \min_{x \in R^n} M(x, \eta).$$

Так как мы предположили, что задача (6.2.4) имеет решение при всех $\eta \in \mathbb{R}$, то множество $X(\eta)$ всегда не пусто.

Алгоритм метода параметризации целевой функции. Задаемся начальной оценкой $\eta_0 \leq f_*$ и полагаем $k = 0$.

Общая k-я итерация

Шаг 1. Находим точку $x_k \in X(\eta_k)$.

Шаг 2. Если $M(x_k, \eta_k) = 0$, то останавливаемся. В противном случае вычисляем

$$\eta_{k+1} = \eta_k + M(x_k, \eta_k). \tag{6.2.6}$$

Шаг 3. Полагаем $k := k + 1$ и идем на шаг 1.

В общем случае получаем бесконечную последовательность точек $\{x_k\}$ и бесконечную монотонно возрастающую последовательность оценок $\{\eta_k\}$. Имеет место следующий результат.

Лемма 6.2.1. *Пусть* $\eta_k \le f_*$. *Тогда* $\eta_{k+1} \le f_*$.

Доказательство. Возьмем произвольную точку $x_* \in X_*$. Так как $x_k \in X(\eta_k)$ и $B(g(x_*)) = 0$, то

$$M(x_k, \eta_k) \le M(x_*, \eta_k) = \sqrt{(f(x_*) - \eta_k)_+^2} = (f(x_*) - \eta_k)_+ = f_* - \eta_k.$$

Поэтому $\eta_{k+1} = \eta_k + M(x_k, \eta_k) \le f_*$. ■

Таким образом, действительно получаем монотонно возрастающую последовательность оценок η_k, которая ограничена сверху. Следовательно, существует предел $\lim_{k\to\infty} \eta_k = \eta_* \le f_*$.

Теорема 6.2.1. *Пусть последовательность* $\{x_k\}$, *порождаемая алгоритмом метода параметризации целевой функции, принадлежит компактному множеству* $\tilde{X}$. *Тогда любая ее предельная точка есть решение задачи (6.2.1). При этом*

$$\lim_{k\to\infty} f(x_k) = \lim_{k\to\infty} \eta_k = f_*. \tag{6.2.7}$$

Доказательство. Имеем на основании утверждения леммы 6.2.1

$$\lim_{k\to\infty} (\eta_{k+1} - \eta_k) = \lim_{k\to\infty} M(x_k, \eta_k) = 0.$$

Поэтому существует предел

$$\lim_{k\to\infty} B(g(x_k)) = 0.$$

Возьмем теперь произвольную сходящуюся к точке x_* последовательность $\{x_{k_s}\}$. В силу непрерывности получаем $B(g(x_*)) = 0$, следовательно, $x_* \in X$. Кроме того,

$$M(x_*, \eta_*) = (f(x_*) - \eta_*)_+ = 0.$$

Так как $x_* \in X$, имеет место неравенство $f(x_*) \geq f_*$. Но строгое неравенство $f(x_*) > f_*$ не может иметь места, поскольку иначе с учетом того, что $\eta_* \leq f_*$, получили бы:

$$M(x_*, \eta_*) > 0. \tag{6.2.8}$$

Поэтому обязательно $f(x_*) = f_*$. Отсюда заключаем, что $x_* \in X_*$.

Из вышесказанного следует также, что выполняются предельные равенства (6.2.7). Действительно, случай, когда $\eta_* < f_*$, невозможен, так как тогда приходим к неравенству (6.2.8). ■

Способ пересчета оценок (6.2.6) допускает наглядную геометрическую интерпретацию. Обозначим через $G(f, B)$ функцию

$$G(f, B) = \sqrt{f_+^2 + B_+^2}.$$

Учтем, что для внешней свертывающей функции $B(g) \geq 0$ для всех $g \in \mathbb{R}^m$. Поэтому $B(g) = (B(g))_+$. Тогда вспомогательная функция $M(x, \eta)$ может быть записана как

$$M(x, \eta) = G\left(f(x) - \eta, B(g(x))\right).$$

Введем также в рассмотрение множество

$$\mathcal{W} = \left\{ [f, B] \in \mathbb{R}^2 : \ f = f(x), \ B = B(g(x)), \quad x \in \mathbb{R}^n \right\}.$$

Множество $\mathcal{W}$ есть образ задачи (6.2.1) в пространстве двух переменных, а именно: значения целевой функции и внешнего штрафа, соответствующих каждой точке пространства $\mathbb{R}^n$. Наряду с множеством $\mathcal{W}$ рассмотрим также множество $\mathcal{W}_+ = \mathcal{W} + \mathbb{R}^2_+$.

Упражнение 6. *Пусть (6.2.1) является задачей выпуклого программирования, а $B(g)$ — неубывающей выпуклой функцией. Покажите, что в этом случае множество $\mathcal{W}_+$ выпукло.*

На рис. 6.5 показаны множество $\mathcal{W}$ (вместе с $\mathcal{W}_+$), а также линии уровня функции $G(f - \eta_k, B)$. Решение вспомогательной задачи минимизации есть не что иное, как нахождение такой линии уровня функции $G(f - \eta_k, B)$, которая соответствовала бы минимальному значению этой функции и при этом касалась бы множества $\mathcal{W}$. Последующее значение η_{k+1} получается путем пересечения данной линии уровня с вертикальной осью.

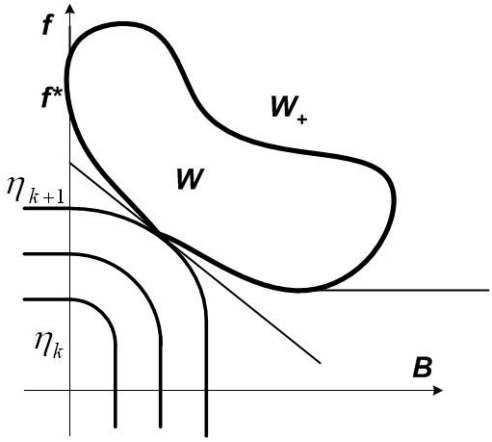

Рис. 6.5. Первый и второй способ пересчета оценок

Предположим, что в задаче (6.2.1) существует седловая точка функции Лагранжа

$$L(x,u) = f(x) + \langle g(x), u\rangle, \quad u \in \mathbb{R}^m_+,$$

т.е. выполняются неравенства

$$L(x_*,u) \le L(x_*,u_*) \le L(x,u_*) \quad \forall x \in \mathbb{R}^n, \quad \forall u \in \mathbb{R}^m_+. \tag{6.2.9}$$

Тогда в силу достаточных условий оптимальности точка x_* является решением задачи (6.2.1) и $f_* = L(x_*,u_*)$. В этом случае из правого неравенства (6.2.9), применяя дважды неравенство Минковского–Малера, получаем для всех $x \in \mathbb{R}^n$ и любого $\eta \le f_*$:

$$\begin{aligned} f_* - \eta &\le f(x) - \eta + \langle g(x), u_*\rangle \le \\ &\le f(x) - \eta + B(g(x))B^0(u_*) = \\ &= \left\langle \begin{bmatrix} f(x)-\eta \\ B(g(x)) \end{bmatrix}, \begin{bmatrix} 1 \\ B^0(u_*) \end{bmatrix} \right\rangle \le \\ &\le G\left(f(x)-\eta, B(g(x))\right) G^0(w_*) = M(x,\eta)G^0(w_*), \end{aligned}$$

где $w_* = [1, B^0(u_*)]^T$.

Поэтому

$$M(x,\eta) \ge \frac{f_* - \eta}{G^0(w_*)}.$$

Так как правая часть не зависит от x, то отсюда, а также из неравенства $\phi(\eta) \le M(x_*, \eta) = f_* - \eta$, где $x_* \in X_*$, приходим к двусторонней оценке:

$$f_* - \eta \ge \phi(\eta) \ge \frac{f_* - \eta}{G^0(w_*)}. \tag{6.2.10}$$

Если в качестве $M(x, \eta)$ берется функция

$$M(x, \eta) = \sqrt{(f(x) - \eta)_+^2 + \sum_{i=1}^{m} (g^i(x))_+^2}, \tag{6.2.11}$$

то $B(g) = \|g_+\|_2$, $B^0(u_*) = \|u_*\|_2$, $G^0(w_*) = \|w_*\|_2$ и оценка (6.2.10) принимает вид

$$f_* - \eta \ge M(x, \eta) \ge \frac{f_* - \eta}{\sqrt{1 + \sum_{i=1}^{m} (u_*^i)^2}}.$$

Отсюда видно, что $M(x_k, \eta_k) \to 0$ при $\eta_k \uparrow f_*$.

Возможен также другой способ пересчета оценок η_k, согласно которому

$$\eta_{k+1} = \eta_k + \frac{(M(x_k, \eta_k))^2}{f(x_k) - \eta_k}. \tag{6.2.12}$$

Данный способ соответствует использованию метода Ньютона для решения уравнения (6.2.5). Он применяется, когда (6.2.1) является задачей выпуклого программирования, т.е. функция $f(x)$ и все функции $g^i(x)$, $1 \le i \le m$, выпуклы. Если снова обратиться к рис. 6.5, то точка η_{k+1} из (6.2.12) получается путем пересечения касательной к соответствующей линии уровня функции $G(f - \eta_k, B)$ с вертикальной осью. Разумеется, она лежит выше, чем та оценка η_{k+1}, которая получается по первому способу (6.2.6), но опять же ниже, чем точка f_*.

Метод параметризации целевой функции может также применяться и для решения задач нелинейного программирования, содержащих ограничения типа равенства. В этом случае, аналогично тому, как это делается в методе внешних штрафных функций, добавляется штраф за нарушение ограничений-равенств. Например, для задачи (6.1.18) функция (6.2.11) преобразуется следующим образом:

$$M(x, \eta) = \sqrt{(f(x) - \eta)_+^2 + \sum_{i=1}^{m} (g^i(x))_+^2 + \sum_{j=1}^{l} (h^j(x))^2}.$$

Заметим также, что при численном решении вспомогательных задач (6.2.3) целесообразно минимизировать не саму функцию $M(x, \eta_k)$, а ее квадрат, что обычно и делают.

Обратимся в заключение к вспомогательной функции $M_1(x, \eta)$ вида (6.2.2). Данная функция при определенных предположениях обладает свойством быть *точной*, т.е. существует целый интервал значений оценки η, для любого значения η из которого множество решений вспомогательной задачи безусловной минимизации

$$\phi(\eta) = \min_{x \in R^n} M_1(x, \eta) \tag{6.2.13}$$

совпадает с множеством решением X_* исходной задачи (6.2.1). Теперь через $X(\eta)$ будем обозначать множество решений задачи (6.2.13).

Теорема 6.2.2. *Пусть в задаче (6.2.1) существует седловая точка $[x_*, u_*]$ функции Лагранжа на множестве $\mathbb{R}^n \times \mathbb{R}^m_+$. Пусть, кроме того, внешняя свертывающая функция $B(g)$ такова, что выполняется неравенство $0 < B^0(u_*) < +\infty$. Тогда $X(\eta) = X_*$ для любого $\eta_* < \eta \leq f_*$, где $\eta_* = f_* - \left(2B^0(u_*)\right)^{-1}$.*

Доказательство. Нам надо показать, что для всех $x \in \mathbb{R}^n$ и любого $\eta \in (\eta_*, f_*]$ имеет место неравенство

$$(f(x) - \eta)^2_+ + B(g(x)) \geq (f_* - \eta)^2. \tag{6.2.14}$$

Оно очевидно, если $f(x) \geq f_*$. Убедимся в его справедливости для остальных x, когда $f(x) < f_*$.

Из наличия у функции Лагранжа седловой точки и из неравенства Минковского–Малера следует, что

$$f_* \leq f(x) + \langle g(x), u_* \rangle \leq f(x) + B(g(x))B^0(u_*),$$

откуда получаем

$$B(g(x)) \geq \frac{f_* - f(x)}{B^0(u_*)}.$$

Учтем теперь, что $(B^0(u_*))^{-1} = 2(f_* - \eta_*)$. Тогда для тех $x \in \mathbb{R}^n$, для которых $f(x) \leq \eta$, имеем

$$\begin{aligned}(f(x) - \eta)^2_+ + B(g(x)) = B(g(x)) \geq \tfrac{f_* - f(x)}{B^0(u_*)} = \\ = 2(f_* - f(x))(f_* - \eta_*) \geq 2(f_* - \eta)^2 \geq (f_* - \eta)^2.\end{aligned}$$

Таким образом, неравенство (6.2.14) оказывается справедливым и для данных x.

Осталось рассмотреть последний случай, когда $\eta < f(x) < f_*$. Оценивая последовательно разность $\Delta = (f(x)-\eta)_+^2 + B(g(x)) - (f_*-\eta)^2$ между левой и правой частями (6.2.14), получаем

$$\begin{aligned}\Delta &\geq (f(x)-\eta)^2 - (f_*-\eta)^2 + \tfrac{f_*-f(x)}{B^0(u_*)} = \\ &= (f(x))^2 - f_*^2 + 2\eta(f_*-f(x)) + \tfrac{f_*-f(x)}{B^0(u_*)} = \\ &= (f(x)-f_*)(f(x)+f_*) + 2\eta(f_*-f(x)) + \tfrac{f_*-f(x)}{B^0(u_*)} = \\ &= (f_*-f(x))[2\eta - f(x) - f_* + \tfrac{1}{B^0(u_*)}].\end{aligned} \tag{6.2.15}$$

Но $\eta > \eta_* = f_* - \frac{1}{2B^0(u_*)}$. Поэтому

$$2\eta - f(x) - f_* + \frac{1}{B^0(u_*)} > f_* - f(x) > 0.$$

Отсюда и из (6.2.15) делаем вывод, что неравенство (6.2.14) имеет место, когда $\eta < f(x) < f_*$.

Если подставить в (6.2.14) $x_* \in X_*$, то неравенство (6.2.14) переходит в равенство. Это означает, что $X_* \subseteq X(\eta)$ для всех $\eta \in (\eta_*, f_*]$. Более того, проводя рассуждения, аналогичные тем, которые делались при доказательстве теоремы 6.1.1, можно убедиться, что на самом деле $X(\eta) = X_*$. ■

Приведем пример вспомогательной функции $M_1(x,\eta)$, для которой выполнены условия теоремы 6.2.2. Возьмем в качестве $B(g)$ функцию $\|g_+\|_p$, в которой $1 \leq p \leq +\infty$. Тогда приходим к следующей вспомогательной функции:

$$M(x,\eta) = (f(x)-\eta)_+^2 + \|g_+(x))\|_p.$$

Для нее $\eta_* = f_* - \frac{1}{2\|u_*\|_{p_*}}$. Поэтому наибольший диапазон допустимых значений η будет в том случае, когда $p = 1$. Тогда p_* принимает предельное значение $p_* = \infty$ и норма $\|u_*\|_\infty$ является чебышевской, т.е. $\|u_*\|_\infty = \max_{1\leq i\leq m} |u_*^i|$.

Отметим также, что для большинства задач оптимальное значение f_* функции $f(x)$ на X, как правило, неизвестно и, следовательно, указать заранее конкретные границы полуинтервала $(\eta_*, f_*]$ достаточно сложно. Кроме того, функция $M_1(x,\eta)$ не является гладкой.

6.3. Методы центров

6.3.1. Вспомогательные функции и классификация методов

Рассмотренные методы внешних и внутренних штрафных функций, а также метод параметризации целевой функции (метод внешних центров) строились примерно по одной и той же схеме. Вводилась вспомогательная функция, зависящая как от основных переменных, входящих в постановку задачи, так и от некоторого дополнительного параметра. В методах штрафных функций это был коэффициент штрафа. В методе внешних центров в качестве параметра бралась оценка снизу оптимального значения целевой функции. Далее решалась последовательность вспомогательных задач безусловной минимизации при разных значениях дополнительных параметров, причем последние целенаправленно изменялись, чтобы обеспечить сходимость решений вспомогательных задач к решению поставленной задачи. Такой подход является достаточно общим и позволяет построить большое семейство методов условной оптимизации, включая и прямые методы. Рассмотрим его на примере задачи нелинейного программирования: найти

$$f_* = \min_{x \in X} f(x), \qquad X = \{x \in \mathbb{R}^n : \ g(x) \le 0_m\}, \tag{6.3.1}$$

где $f(x)$ и $g^i(x)$, $1 \le i \le m$, — непрерывные функции. Через X_* обозначим множество решений задачи (6.3.1), которое предполагается непустым.

Введем функцию $M(x, y)$, зависящую от исходных переменных x и от некоторого вектора y дополнительных параметров из множества Y. Эти дополнительные параметры для разных функций $M(x, y)$ могут иметь разный смысл, поэтому размерность вектора y и вид множества Y пока не конкретизируем. Рассмотрим вспомогательную задачу минимизации:

$$\min_{x \in \Pi} M(x, y), \tag{6.3.2}$$

в которой Π — некоторое замкнутое множество из $\mathbb{R}^n$, содержащее X_*. В частности, в качестве Π может быть взято либо все пространство $\mathbb{R}^n$, либо само допустимое множество X, либо его часть. Определим точечно-множественное отображение

$$X(y) = \operatorname{Arg}\min_{x \in \Pi} M(x, y). \tag{6.3.3}$$

Через $x(y)$ будем обозначать произвольный элемент из множества $X(y)$. Если множество $X(y)$ не пусто для любого $y \in Y$, то отображение (6.3.3) является *строгим в* Y.

Определение 6.3.1. *Функцию $M(x,y)$ назовем точной вспомогательной функцией (ТВФ) для задачи (6.3.1) на множестве $\Pi \times Y$, если $X(y) = X_*$ для любого $y \in Y$.*

Пусть имеется последовательность множеств $\{X_k\}$. Под множеством ее прикосновения понимается такое множество X_T, что если $x \in X_T$, то найдутся подпоследовательность множеств $\{X_{k_j}\}$ и последовательность точек $x_{k_j} \in X_{k_j}$ такие, что $\lim\limits_{j\to\infty} x_{k_j} = x$.

Определение 6.3.2. *Функция $M(x,y)$ называется приближенной вспомогательной функцией (ПВФ) для задачи (6.3.1) на множестве $\Pi \times Y$, если отображение $X(y)$ строгое в Y и существует такое множество Y_*, принадлежащее замыканию $\bar{Y}$ множества Y, что для любой сходящейся к Y_* последовательности $\{y_k\}$ множество прикосновения последовательности $X(y_k)$ принадлежит X_*.*

Таким образом, если известна ТВФ, то решение задачи (6.3.1) сводится к однократной минимизации вспомогательной функции. Если функция $M(x,y)$ является ПВФ, то для решения с ее помощью задачи (6.3.1) достаточно выявить множество Y_* и, взяв произвольную последовательность $y_k \to y_*$, $y_k \in Y$, $y_* \in Y_*$, найти хотя бы один элемент этой последовательности.

Введем дополнительно вектор-функцию $G(x,y)$, размерность которой совпадает с размерностью вектора y. Вместо (6.3.2) рассмотрим совместную задачу, состоящую из задачи минимизации (6.3.2) и задачи отыскания решений системы уравнений

$$G(x,y) = 0. \tag{6.3.4}$$

Точку $[x_*, y_*] \in \Pi \times \bar{Y}$ назовем *особой точкой* пары функций $\{M, G\}$ на множестве $\Pi \times Y$, если выполнены условия

$$x_* \in X(y_*), \qquad G(x_*, y_*) = 0.$$

Соответственно точку $[x_*, y_*] \in \Pi \times \bar{Y}$ назовем *обобщенной особой точкой* пары функций $\{M, G\}$ на множестве $\Pi \times Y$, если можно указать такие последовательности $\{x_k\}$ и $\{y_k\}$, что $x_k \to x_*$, $y_k \to y_*$, $y_k \in Y$ и

$$x_k \in X(y_k), \qquad \lim_{k\to\infty} G(x_k, y_k) = 0.$$

Определение 6.3.3. *Пара функций $\{M, G\}$ согласована с задачей (6.3.1) на множестве $\Pi \times Y$, если всякая ее особая точка $[x_*, y_*]$ на $\Pi \times Y$ (быть может, обобщенная) такова, что $x_* \in X_*$.*

В том случае, когда у пары функций $\{M, G\}$ вторая функция G совпадает с первой M, будем просто говорить о функции M, согласованной с задачей (6.3.1) на множестве $\Pi \times Y$. Точно так же вместо термина *особая точка пары функций* будем употреблять термин *особая точка функции*. Если пара функций $\{M, G\}$ согласована с задачей (6.3.1) на $\Pi \times Y$ и множество ее особых точек на $\Pi \times Y$ не пусто, то функция $M(x, y)$ является ПВФ.

Для решения задачи (6.3.2), (6.3.4) могут быть применены разнообразные численные методы отыскания решений систем уравнений в сочетании с методами безусловной минимизации. Простейший подход основан на применении зависимости $x(y) \in X(y)$, получающейся из решения задачи (6.3.2). Подставляя $x(y)$ в (6.3.4), приходим к уравнению

$$G(x(y), y) = 0. \tag{6.3.5}$$

Использование для решения (6.3.5) метода простой итерации приводит к следующей численной схеме:

$$y_{k+1} = y_k + \alpha G(x_k, y_k), \qquad x_k \in X(y_k), \tag{6.3.6}$$

где $\alpha > 0$ — некоторый коэффициент. Если применить метод Ньютона, то вместо (6.3.6) получаем

$$y_{k+1} = y_k - \left[\frac{d}{dy} G(x(y_k), y_k)\right]^{-1} G(x_k, y_k), \qquad x_k = x(y_k) \in X(y_k). \tag{6.3.7}$$

Методы вида (6.3.6), (6.3.7) назовем *непрямыми* методами решения задачи (6.3.1), так как основные «внешние» итерации в них проводятся по дополнительным параметрам. В отличие от них методы, использующие обратную зависимость $y(x)$, получающуюся из (6.3.4), называются *прямыми*. Примерами прямых методов являются методы, рассмотренные в предыдущей главе. Напротив, методы штрафных функций и параметризации целевой функции относятся к непрямым методам.

Довольно большой класс вспомогательных функций $M(x, y)$ может быть построен на основе двухэтапного процесса с применением специальных свертывающих функций. В этом процессе сначала сворачиваются ограничения, а затем получившееся «единое» ограничение сворачивается с целевой функцией. В результате приходим к вспомогательной функции следующего вида:

$$M(x, y) = H(f(x) - \eta, B(g(x), \beta)), \tag{6.3.8}$$

где $y = [\eta, \beta]$, η — оценка оптимального значения f_* минимизируемой функции, H и B — некоторые свертывающие функции, причем вторая из них зависит дополнительно от параметра β. В качестве свертывающих функций могут, в частности, применяться введенные нами внешние и внутренние свертывающие функции. Еще один очень важный пример свертывающей функции — это *линейная свертывающая функция* $B(g) = \langle e, g \rangle$, где e — вектор со всеми единичными компонентами.

В том случае, когда зависимость функции B от параметра β отсутствует, вид функции (6.3.8) упрощается:

$$M(x, \eta) = H(\tilde{f}(x, \eta), B(g(x))), \tag{6.3.9}$$

где для сокращения записи введено обозначение: $\tilde{f}(x, \eta) = f(x) - \eta$. Именно такие вспомогательные функции $M(x, \eta)$, в которых в качестве H и B брались внешние свертывающие функции, использовались в методах параметризации целевой функции. Эти итерационные процессы являются частными случаями применения метода простой итерации (6.3.6) или метода Ньютона (6.3.7) для решения уравнения (6.3.5).

Заметим также, что в методах внешних и внутренних штрафных функций вспомогательные функции строятся с применением линейной свертывающей функции в качестве H и функции $B(g, t) = t\tilde{B}(g)$, где $\tilde{B}(g)$ — соответственно внешняя или внутренняя свертывающая функция, t — коэффициент штрафа. Поскольку оценка оптимального значения целевой функции η входит как аддитивная константа в такую вспомогательную функцию, то на решение вспомогательных задач минимизации она влияния не оказывает и ее можно опустить.

6.3.2. Методы внутренних центров

Рассмотрим опять задачу (6.3.1), содержащую только ограничения-неравенства. При этом, как и в методе внутренних штрафных функций, считаем, что допустимое множество X в задаче (6.3.1) является *регулярным*, т.е. множество X_0 не пусто и $X_0 = \mathrm{int}X$.

Поступим теперь в отличие от метода параметризации целевой функции по-другому и возьмем в качестве H и B внутренние свертывающие функции. Тогда приходим к вспомогательной функции $M(x, \eta)$ вида (6.3.9), используемой в *методах внутренних центров*. В них предполагается, что $x \in X$ и $\eta > f_*$, т.е. теперь η является оценкой сверху оптимального значения целевой функции $f(x_*)$.

Приведем пример такой вспомогательной функции:

$$M(x,\eta) = -\frac{1}{2}\left[\ln(\eta - f(x))_+ + m^{-1}\sum_{i=1}^{m}\ln(-g^i(x)) + c\right],$$

где c — некоторая константа. Данная функция носит название *логарифмической барьерной функции* метода центров. Она получается, когда в качестве $H(z)$ и $B(g)$ берутся соответственно внутренние свертывающие функции $H(z) = -\ln(\|z_-\|_0) - 0.5$ и $B(g) = -\|g_-\|_0$, в которых $z \in \mathbb{R}^2_{--}$, $g \in \mathbb{R}^m_{--}$.

Другим примером вспомогательной функции является функция

$$M(x,\eta) = (\eta - f(x))_+^{-1} + \sum_{i=1}^{m}(-g^i(x))^{-1},$$

которая называется *обратной барьерной функцией* метода центров. К данной вспомогательной функции можно прийти, взяв внутренние свертывающие функции: $H(z) = \|z_-\|_{-1}^{-1}$ и $B(g) = -\|g_-\|_{-1}$.

Если в качестве H и B использовать собственные внутренние свертывающие функции $H(z) = -\|z_-\|_p$, $B(g) = -\|g_-\|_p$, то можно получить другие вспомогательные функции, имеющие соответственно вид

$$\begin{aligned} M(x,\eta) &= -\left[2\sqrt{m}(\eta - f(x))_+ \prod_{i=1}^{m}\sqrt[m]{-g^i(x)}\right]^{\frac{1}{2}}, \\ M(x,\eta) &= -(\eta - f(x))_+\left[1 + (\eta - f(x))_+\sum_{i=1}^{m}(-g^i(x))^{-1}\right]^{-1}. \end{aligned} \tag{6.3.10}$$

Для первой из функций $p = 0$, для второй — $p = -1$.

Покажем, что при определенных дополнительных требованиях к функциям H и B функция $M(x,\eta)$ типа (6.3.9) может быть согласованной с задачей (6.3.1).

Лемма 6.3.1. *Пусть в задаче (6.3.1) множество X регулярно. Пусть, кроме того, функции H и B являются собственными внутренними свертывающими функциями. Тогда функция (6.3.9) согласована с задачей (6.3.1) на множестве $X \times \mathcal{H}$, где $\mathcal{H} = \{\eta \in \mathbb{R} : \eta > f_*\}$.*

Доказательство. Покажем, что у функции (6.3.9) не существует обычных особых точек. Действительно, пусть $[x_*, \eta_*]$ — обычная особая точка. Это означает, что для $\eta_* > f_*$ выполнено

$$x_* \in X(\eta_*) = \operatorname{Arg}\min_{x\in X} M(x,\eta_*), \quad M(x_*,\eta_*) = 0. \tag{6.3.11}$$

Но поскольку множество X регулярно, то найдется точка $\bar{x} \in X_0$ такая, что $f(\bar{x}) < \eta_*$. Поэтому $M(x_*, \eta_*) \leq M(\bar{x}, \eta_*) < 0$. Данное неравенство противоречит второму равенству из (6.3.11).

Предположим теперь, что $[x_*, \eta_*]$ — обобщенная особая точка, причем $\eta_* > f_*$. Из предположения о регулярности множества X опять получаем, что существует $\bar{x} \in X_0$ и такие константы $\delta_1 \leq \delta_2 < 0$, что при k достаточно больших ($k \geq K$) имеет место двустороннее неравенство $\delta_1 \leq \tilde{f}(\bar{x}, \eta_k) \leq \delta_2$. Поэтому

$$\begin{array}{rcl} M(x_k, \eta_k) & \leq & M(\bar{x}, \eta_k) = H(\tilde{f}(\bar{x}, \eta_k), B(g(\bar{x}))) \leq \\ & \leq & \sup\limits_{\delta_1 \leq \mu \leq \delta_2} H(\mu, B(g(\bar{x}))) \leq \delta_3 < 0. \end{array}$$

Отсюда получаем, что предельное соотношение $M(x_k, \eta_k) \to 0$ в определении обобщенной особой точки не выполняется.

Остается только возможность, когда $[x_*, \eta_*]$ — обобщенная особая точка и $\eta_* = f_*$. Если $f(x_*) > f_*$, то найдется такое число K_1, что для всех $k \geq K_1$ выполнено $\tilde{f}(x_k, \eta_k) > 0$ и, следовательно, равенство $M(x_k, \eta_k) = 0$. С другой стороны, можно указать такую последовательность $\{\bar{x}_k\} \subset X_0$, что $f(\bar{x}_k) < \eta_k$ для k достаточно больших ($k \geq K_2 \geq K_1$). Отсюда для этих k получаем

$$M(x_k, \eta_k) \leq M(\bar{x}_k, \eta_k) = H(\tilde{f}(\bar{x}_k, \eta_k), B(g(\bar{x}_k))) < 0,$$

что противоречит равенствам $M(x_k, \eta_k) = 0$. ■

Положим

$$G(x, \eta) = \tilde{f}(x, \eta) = f(x) - \eta. \tag{6.3.12}$$

Наряду с леммой 6.3.1 имеет место следующее утверждение.

Лемма 6.3.2. *Пусть выполнены предположения леммы 6.3.1. Тогда пара функций (6.3.9), (6.3.12) согласована с задачей (6.3.1) на множестве $X \times \mathcal{H}$, где $\mathcal{H} = \{\eta \in \mathbb{R} : \ \eta > f_*\}$.*

Доказательство. Пара функций (6.3.9), (6.3.12) может иметь только обобщенные особые точки на $X \times \mathcal{H}$. В самом деле, если $[x_*, \eta_*]$ есть обычная особая точка, то $x_* \in X(\eta_*)$, $\eta_* > f_*$, $\tilde{f}(x_*, \eta_*) = 0$ и, следовательно, $M(x_*, \eta_*) = 0$. Но тогда, в силу регулярности множества X, можно указать такое $\bar{x} \in X_0$, что $\tilde{f}(\bar{x}, \eta_*) < 0$. Поэтому

$$M(\bar{x}, \eta_*) = H(\tilde{f}(\bar{x}, \eta_*), B(g(\bar{x}))) < 0. \tag{6.3.13}$$

Сопоставляя это неравенство с равенством $M(x_*, \eta_*) = 0$, приходим к выводу, что $x_* \notin X(\eta_*)$. Мы получили противоречие.

Убедимся теперь, что у пары функций (6.3.9), (6.3.12) не существует обобщенной особой точки $[x_*, \eta_*]$, в которой $\eta_* > f_*$. Действительно, если допустить противное, то найдутся такие последовательности $\{x_k\}$ и $\{\eta_k\}$, что $x_k \to x_*$, $\eta_k \to \eta_*$ и $\tilde{f}(x_k, \eta_k) \to 0$. Так как $\eta_* > f_*$, то всегда можно выбрать такую точку $\bar{x} \in X_0$, для которой имеет место (6.3.13). Кроме того, в силу непрерывности функции H, для любого $\epsilon > 0$ можно указать такое $K(\epsilon)$, что

$$|H(\tilde{f}(\bar{x}, \eta_k), B(g(\bar{x}))) - H(\tilde{f}(\bar{x}, \eta_*), B(g(\bar{x})))| < \epsilon \tag{6.3.14}$$

при $k \geq K(\epsilon)$. Выберем теперь $\epsilon = -\frac{M(\bar{x}, \eta_*)}{2}$. Тогда из (6.3.14) и неравенств $M(x_k, \eta_k) \leq M(\bar{x}, \eta_k)$, справедливых для всех k, получаем, что $M(x_k, \eta_k) \leq \frac{M(\bar{x}, \eta_*)}{2} < 0$, если только $k \geq K(\epsilon)$. С другой стороны, так как $x_k \to x_*$ и $\tilde{f}(x_k, \eta_k) \to 0$, то $\liminf\limits_{k \to \infty} M(x_k, \eta_k) = 0$. Мы опять пришли к противоречию.

Таким образом, у пары функций (6.3.9), (6.3.12) может быть только обобщенная особая точка $[x_*, \eta_*]$, в которой $\eta_* = f_*$. Но в этом случае, в силу непрерывности всех функций, $\tilde{f}(x_*, \eta_*) = 0$, $M(x_*, \eta_*) \leq M(x, \eta_*)$ для любых $x \in X$. Отсюда делаем вывод, что $x_* \in X_*$. ■

Если H не является собственной внутренней свертывающей функцией, то очевидно, что функция (6.3.9) не может быть согласованной с задачей (6.3.1), однако оказывается справедливым следующий результат.

Лемма 6.3.3. *Пусть в задаче (6.3.1) множество X регулярно. Пусть, кроме того, H является несобственной внутренней свертывающей функцией, а B — собственной внутренней свертывающей функцией. Тогда пара функций (6.3.9), (6.3.12) согласована с задачей (6.3.1) на множестве $X \times \mathcal{H}$, где $\mathcal{H} = \{\eta \in \mathbb{R} : \eta > f_*\}$.*

Доказательство. Так как H является несобственной свертывающей функцией, то у пары функций (6.3.9), (6.3.12) могут быть только обобщенные особые точки на $X \times \mathcal{H}$. Покажем, что любая обобщенная особая точка $[x_*, \eta_*]$ такова, что $x_* \in X_*$.

Пусть $x_k \in X(\eta_k)$, $x_k \to x_*$, $\eta_k \to \eta_*$, $\eta_k \geq f_*$ и $\tilde{f}(x_k, \eta_k) \to 0$. Предположим, что $x_* \notin X_*$. Тогда $f(\bar{x}) < f(x_*)$ для некоторого $\bar{x} \in X_0$. Поэтому

$$M(x_k, \eta_k) \leq M(\bar{x}, \eta_k) \leq C < +\infty$$

для k достаточно больших. С другой стороны, $M(x_k, \eta_k) \to +\infty$, так как $\tilde{f}(x_k, \eta_k) \to 0$. Мы пришли к противоречию. Следовательно, имеет место включение $x_* \in X_*$. ■

Теорема 6.3.1. *Пусть выполнены условия леммы 6.3.1 или леммы 6.3.3, множество X компактно. Тогда (6.3.9) — ПВФ для задачи (6.3.1) на множестве $X \times \mathcal{H}$.*

Доказательство. Из компактности множества X и непрерывности функции $M(x,\eta)$ следует, что для любого $\eta \in \mathcal{H}$ решение задачи минимизации функции $M(x,\eta)$ по x на X существует, т.е. отображение $X(\eta)$ — строгое в $\mathcal{H}$.

Покажем, что множество особых точек функции $M(x,\eta)$ непустое. Возьмем $\eta_0 > f_*$ и рассмотрим такие последовательности точек $\{x_k\}$, $\{\eta_k\}$, что $x_k \in X(\eta_k)$, $\eta_{k+1} = f(x_k)$. Понятно, что всегда $f(x_k) < \eta_k$ и, следовательно, $\eta_{k+1} < \eta_k$. Отсюда получаем, что $\{\eta_k\}$ — монотонно убывающая последовательность, ограниченная снизу f_*, поэтому существует предел $\eta_* \geq f_*$.

Пусть x_* — предельная точка последовательности $\{x_k\}$. Из существования предела $\tilde{f}(x_k,\eta_k) \to 0$ вытекает, что $M(x_k,\eta_k) \to 0$ и, следовательно, $[x_*,\eta_*]$ есть обобщенная особая точка функции $M(x,\eta)$. При доказательстве леммы 6.3.1 было установлено, что у этой функции все обобщенные особые точки $[x_*,\eta_*]$ таковы, что $\eta_* = f_*$. Таким образом, предельное множество $\mathcal{H}_*$ из определения 6.3.2 в данном случае не пусто и $f_* \in \mathcal{H}_* \subseteq \bar{\mathcal{H}}$. ■

Рассмотрим методы решения задачи (6.3.1) на основе функции (6.3.9) с внутренними свертывающими функциями H и B. Возьмем в качестве $G(x,\eta)$ функцию (6.3.12) и будем решать уравнение

$$\tilde{f}(x(\eta),\eta) = 0. \tag{6.3.15}$$

Применяя для отыскания особых точек схему (6.3.6) с $\alpha = 1$, приходим к следующему итерационному процессу:

$$x_k = \arg\min_{x \in X_k} M(x,\eta_k), \qquad \eta_{k+1} = f(x_k), \tag{6.3.16}$$

в котором $X_k = \{x \in X_0 : f(x) < \eta_k\}$. В данном методе все текущие точки лежат внутри допустимой области и параметр η_k убывает в процессе расчетов, стремясь к f_*.

Теорема 6.3.2. *Пусть в задаче (6.3.1) допустимое множество X регулярное и ограниченное. Пусть, кроме того, функция H является внутренней свертывающей функцией (собственной или несобственной), а B — собственной внутренней свертывающей функцией. Тогда все предельные точки последовательности $\{x_k\}$, получающиеся в ходе вычислений итерационным процессом (6.3.16), принадлежат X_*.*

Доказательство. Возьмем произвольную сходящуюся последовательность $\{x_k\}$. Пусть $x_k \to x_*$. Так как H является строго внутренней свертывающей функцией, то $f(x_k) < \eta_k$, т.е. последовательность $\{\eta_k\}$ монотонно убывает. Из ее ограниченности снизу следует, что существует предел $\eta_* \geq f_*$. Понятно, что $\eta_k \to \eta_*$ и $\tilde{f}(x_k, \eta_k) \to 0$. Отсюда следует, что $[x_*, \eta_*]$ — обобщенная особая точка пары функций (6.3.9), (6.3.12). Поэтому, согласно утверждениям лемм 6.3.2 или 6.3.3, $x_* \in X_*$. ■

Если $H(z)$ — собственная внутренняя свертывающая функция, то наряду с (6.3.16) можно рассмотреть итерационный процесс

$$x_k \in \operatorname{Arg}\min_{x \in X_k} M(x, \eta_k), \qquad \eta_{k+1} = \eta_k + M(x_k, \eta_k), \tag{6.3.17}$$

основанный на решении уравнения $M(x(\eta), \eta) = 0$. При определенных условиях, наложенных на функцию H, процесс (6.3.17) также позволяет отыскивать решение задачи (6.3.1).

Выведем теперь неравенства, оценивающие минимальное значение вспомогательной функции в зависимости от близости параметра η к f_*. Пусть в задаче (6.3.1) существует седловая точка $[x_*, u_*]$ функции Лагранжа, и пусть обе функции H и B, входящие в функцию (6.3.9), являются собственными внутренними свертывающими функциями. Тогда имеет место оценка сверху $M(x(\eta), \eta) \leq 0$. Кроме того, на основании правого неравенства (6.2.9) и неравенства Минковского–Малера для любых $x \in X$ выполняется следующая цепочка неравенств:

$$\begin{aligned} f_* - \eta &\leq f(x) - \eta + B(g(x)) B^0_{R^m_-}(u_*) \leq \\ &\leq \langle [\tilde{f}(x, \eta), B(g(x))]^\top, [1, B^0_{R^m_-}(u_*)]^\top \rangle \leq \\ &\leq M(x, \eta) H^0_W(w_*), \end{aligned} \tag{6.3.18}$$

где $w_* = [1, B^0_{R^m_-}(u_*)]^\top \in \mathbb{R}^2$, $W \subseteq \mathbb{R}^2$ — некоторое множество, содержащее в себе образ множества $X \times \mathcal{H}$ при отображении

$$\psi(x, \eta) = [\tilde{f}(x, \eta), B(g(x))]^\top.$$

Здесь $\mathcal{H}$ — область рассматриваемых значений параметра η. В случае, когда $H^0_W(w_*) > 0$, согласно (6.3.18)

$$M(x, \eta) \geq \frac{f_* - \eta}{H^0_W(w_*)}. \tag{6.3.19}$$

При $H_W^0(w_*) < 0$ имеет место неравенство, обратное (6.3.19). В частности, для второй функции из (6.3.10) получаем

$$0 \geq M(x(\eta), \eta) \geq \frac{f_* - \eta}{\left(1 + \sum\limits_{i=1}^{m} \sqrt{u_*^i}\right)^2}. \tag{6.3.20}$$

Для первой функции из (6.3.10) имеем соответственно

$$0 \geq M(x(\eta), \eta) \geq \frac{f_* - \eta}{\sqrt{2\sqrt{m} \prod\limits_{i=1}^{m} \sqrt[m]{u_*^i}}}. \tag{6.3.21}$$

Обе оценки (6.3.20) и (6.3.21) стремятся к нулю при $\eta \to f_*$ сверху для любого $u_* \in \mathbb{R}_+^m$, причем вторая оценка имеет смысл только в том случае, когда $u_* > 0_m$.

В рассмотренных нами методах центров (внешних и внутренних) использовалась вспомогательная функция $M(x, \eta)$ вида (6.3.9), в которой обе свертывающие функции H и B имели один и тот же тип — обе были либо внешними, либо внутренними свертывающими функциями. Можно также построить альтернативные методы центров. В них функция $M(x, \eta)$ строится с использованием свертывающих функций разных типов, т.е. функция H является внешней, а функция B — внутренней свертывающей функцией или наоборот.

Глава 7

Методы, использующие функцию Лагранжа или ее модификации

Методы, основанные на нахождении стационарных точек функции Лагранжа, особенно основанные на использовании модифицированных функций Лагранжа (МФЛ), являются одними из наиболее эффективных средств решения задач нелинейного программирования. Обратимся сначала к классической функции Лагранжа.

7.1. Метод Удзавы

Рассмотрим задачу нелинейного программирования, в которой допустимое множество задается только ограничениями типа равенства

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : \ g(x) = 0_m\right\}, \tag{7.1.1}$$

где относительно функции $f(x)$, а также всех компонент вектор-функции $g(x) = [g^1(x), \ldots, g^m(x)]^T$ предполагается, что они по меньшей мере непрерывно дифференцируемы.

Пусть X_* — множество решений задачи (7.1.1) и пусть $x_* \in X_*$. Из необходимых условий экстремума следует, что при выполнении условия регулярности ограничений в точке x_* (линейной независимости градиентов $g^i_x(x_*)$, $1 \le i \le m$) точка x_* вместе с некоторым вектором

множителей Лагранжа u_* образуют *стационарную точку* функции Лагранжа

$$L(x,u) = f(x) + \langle g(x), u\rangle, \qquad u \in \mathbb{R}^m,$$

составленную для задачи (7.1.1). Согласно определению стационарной точки в ней выполняются равенства:

$$L_x(x_*, u_*) = 0_n, \quad L_u(x_*, u_*) = g(x_*) = 0_m. \tag{7.1.2}$$

Если функции, определяющие задачу, дважды непрерывно дифференцируемы и в точке $[x_*, u_*]$ выполняются достаточные условия второго порядка, т.е. в дополнении к стационарности точки $[x_*, u_*]$ оказывается, что

$$\langle y, L_{xx}(x_*, u_*)y\rangle > 0 \tag{7.1.3}$$

для любого ненулевого

$$y \in K(x_*) = \left\{y \in \mathbb{R}^n : \ g_x(x_*)y = 0_m\right\},$$

то x_* — решение задачи (7.1.1).

Допустим, что вместо (7.1.3) выполняется более сильное требование положительной определенности матрицы $L_{xx}(x_*, u_*)$. Из второго условия стационарности (7.1.2) следует допустимость точки x_*. Таким образом, стационарная точка $[x_*, u_*]$ функции Лагранжа $L(x,u)$ оказывается седловой точкой на множестве $\Delta(x_*) \times \mathbb{R}^m$, где $\Delta(x_*)$ — некоторая окрестность точки x_*. Применим для нахождения данной седловой точки метод, который является обобщением метода градиентного спуска. В x-пространстве спускаемся по антиградиенту $-L_x(x,u)$, а в u-пространстве, напротив, — поднимаемся по градиенту $L_u(x,u)$. Соответствующий итерационный процесс запишется в виде

$$x_{k+1} = x_k - \alpha L_x(x_k, u_k), \qquad u_{k+1} = u_k + \alpha g(x_k), \tag{7.1.4}$$

где $\alpha > 0$ — постоянный шаг, который будем брать достаточно малым.

Метод (7.1.4) носит название *метода Удзавы*. Он одним из первых был предложен для решения задачи нелинейного программирования (7.1.1). Покажем, что при определенных условиях метод Удзавы обладает локальной сходимостью. Для этого нам потребуется следующий результат, известный как *теорема А. Островского*.

Теорема 7.1.1. *Пусть x_* является неподвижной точкой отображения $F(x)$, т.е. $x_* = F(x_*)$. Пусть, кроме того, отображение $F(x)$ дифференцируемо в x_* и спектральный радиус матрицы Якоби $F_x(x_*)$ удовлетворяет условию $\rho(F_x(x_*)) < 1$. Тогда итерационный процесс $x_{k+1} = F(x_k)$ локально сходится к x_*.*

В качестве замечания к данной теореме добавим, что скорость сходимости такого итерационного процесса $x_{k+1} = F(x_k)$ к x_* только линейная.

Теорема 7.1.2. *Пусть функции $f(x)$ и $g^i(x)$, $1 \le i \le m$, дважды непрерывно дифференцируемы. Пусть, кроме того, пара $[x_*, u_*]$ является стационарной точкой функции Лагранжа и в точке x_* выполнено условие регулярности ограничений (линейная независимость градиентов $g^i_x(x_*)$, $1 \le i \le m$). Тогда если матрица $L_{xx}(x_*, u_*)$ положительно определена, то можно указать такое $\bar\alpha > 0$, что для любого $0 < \alpha < \bar\alpha$ итерационный процесс (7.1.4) локально сходится с линейной скоростью к $[x_*, u_*]$.*

Доказательство. Объединим обе переменные x и u в единый вектор $z = [x, u]^T$. Объединим также отображения $-L_x(x, u)$ и $g(x)$ в единое отображение $F(z) = [-L_x(z), g(x)]$. Итерационный процесс (7.1.4) в этих обозначениях можно записать как

$$z_{k+1} = z_k + \alpha F(z_k).$$

Согласно теореме Островского 7.1.1 он будет сходиться к стационарной точке $z_* = [x_*, u_*]^T$, если спектральный радиус матрицы Якоби отображения $Q(z) = I_{n+m} + \alpha F(z)$ в этой точке оказывается меньше единицы.

Вычислим сначала матрицу Якоби отображения $F(z)$ в z_*. Имеем

$$F_z(z_*) = \begin{bmatrix} -L_{xx}(x_*, u_*) & -g_x^\top(x_*) \\ g_x(x_*) & 0_{mm} \end{bmatrix}.$$

Пусть λ — произвольное собственное значение матрицы $F_z(z_*)$ и пусть $z = [x, u]$ — соответствующий этому собственному значению собственный вектор. Они связаны между собой соотношением

$$F_z(z_*)z = \lambda z. \tag{7.1.5}$$

Представим (7.1.5) покомпонентно:

$$-L_{xx}(x_*, u_*)x - g_x^\top(x_*)u = \lambda x, \qquad g_x(x_*)x = \lambda u. \tag{7.1.6}$$

Умножим слева первое равенство (7.1.6) на комплексно-сопряженный вектор $\bar x$, а второе равенство — на комплексно-сопряженный вектор $\bar u$, в результате получим

$$-\bar x^\top L_{xx}(x_*, u_*)x - \bar x^\top g_x^\top(x_*)u = \lambda|x|^2, \qquad \bar u^\top g_x(x_*)x = \lambda|u|^2. \tag{7.1.7}$$

Вектор x отличен от нулевого. Действительно, в противном случае имели бы из первого равенства (7.1.6), что $g_x(x_*)^\top u = 0_n$. Так как обязательно $u \neq 0_m$, то данное равенство противоречит условию регулярности.

Складывая оба равенства (7.1.7), приходим к формуле

$$-\bar{x}^\top L_{xx}(x_*, u_*)x - \bar{x}^\top g_x^\top(x_*)u + \bar{u}^\top g_x(x_*)x = \lambda|z|^2.$$

Учтем теперь, что $\mathrm{Re}[\bar{u}^\top g_x(x_*)x - \bar{x}^\top g_x^\top(x_*)u] = 0$. Тогда получим

$$-\mathrm{Re}\langle \bar{x}, L_{xx}(x_*, u_*)x\rangle = \mathrm{Re}\lambda|z|^2. \tag{7.1.8}$$

Пусть $x = a + ib$ и $\lambda = c + id$. Имеем

$$\mathrm{Re}\langle \bar{x}, L_{xx}(x_*, u_*)x\rangle = \langle a, L_{xx}(x_*, u_*)a\rangle + \langle b, L_{xx}(x_*, u_*)b\rangle.$$

Поскольку $x \neq 0_n$ и матрица $L_{xx}(x_*, u_*)$ по предположению положительно определена, то отсюда и из (7.1.8) делаем вывод, что действительная часть c собственного значения λ отрицательна.

Собственные числа μ матрицы $Q_z(z_*)$ связаны с собственными числами λ матрицы $F_z(z_*)$ очевидным соотношением $\mu = 1 + \alpha\lambda$, что приводит к равенству

$$|\mu| = \left[(1 + \alpha c)^2 + \alpha^2 d^2\right]^{\frac{1}{2}} = \left[1 + \alpha\left(2c + \alpha|\lambda|^2\right)\right]^{\frac{1}{2}}.$$

Отсюда видно, что $|\mu| < 1$, когда $0 < \alpha < -2\frac{\mathrm{Re}\,\lambda}{|\lambda|^2}$. Следовательно, существует такое $\bar{\alpha} > 0$, для которого, какое бы $0 < \alpha < \bar{\alpha}$ ни взять, спектральный радиус матрицы $Q_z(z_*)$ оказывается меньше единицы. Поэтому по теореме Островского 7.1.1 итерационный процесс (7.1.4) локально сходится к точке $[x_*, u_*]$, причем скорость сходимости линейная. ■

Для отыскания решения задачи (7.1.1) можно также построить метод Ньютона. Если воспользоваться тем, что условие стационарности записывается в виде уравнения $L_z(z_*) = 0_{n+m}$, то, решая это уравнение с помощью метода Ньютона, приходим к следующему итерационному процессу:

$$z_{k+1} = z_k - L_{zz}^{-1}(z_k)L_z(z_k). \tag{7.1.9}$$

Здесь симметрическая матрица $L_{zz}(z)$ имеет следующий вид:

$$L_{zz}(z) = \begin{bmatrix} L_{xx}(z) & g_x^\top(x) \\ g_x(x) & 0_{mm} \end{bmatrix}.$$

Предполагая, что выполнены достаточные условия второго порядка для задачи (7.1.1) и условие регулярности ограничений в точке x_*, покажем, что матрица $L_{zz}(z_*)$ является неособой. С этой целью рассмотрим однородную систему уравнений $L_{zz}(z_*)z = 0$, где $z = [x, u] \in \mathbb{R}^{n+m}$. Перепишем ее в покомпонентном виде:

$$L_{xx}(z_*)x + g_x^\top(x_*)u = 0_n, \qquad g_x(x_*)x = 0_m. \tag{7.1.10}$$

Если $x = 0_n$, то первое из равенств (7.1.10) влечет: $g_x^\top(x_*)u = 0_n$. Отсюда, согласно условию регулярности ограничений, выполняется $u = 0_m$.

Предположим далее, что $x \neq 0_n$. Тогда в силу второго равенства (7.1.10) обязательно $x \in K(x_*)$. Умножая теперь первое равенство слева на $x^\top$, получим $\langle x, L_{xx}(z_*)x\rangle = 0$. Данное равенство противоречит (7.1.3). Таким образом, однородная система $L_{zz}(z_*)z = = 0_{n+m}$ может иметь только нулевое решение, что означает невырожденность матрицы $L_{zz}(z_*)$. В силу непрерывности данная матрица будет оставаться невырожденной и в некоторой окрестности точки z_*. Отсюда из общих утверждений о сходимости метода Ньютона следует, что итерационный процесс (7.1.9) локально сходится к точке z_*, причем скорость сходимости сверхлинейная.

Обратимся теперь к общей задаче нелинейного программирования, содержащей помимо ограничений типа равенства также ограничения типа неравенства

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : \ g^i(x) = 0, \ 1 \le i \le l, \ \ g^i(x) \le 0, \ l < i \le m\right\}, \tag{7.1.11}$$

где по-прежнему функция $f(x)$ и все компоненты вектор-функции $g(x)$ по крайней мере непрерывно дифференцируемы.

Для задачи (7.1.11) метод Удзавы, основанный на применении классической функции Лагранжа, может оказаться неприемлемым из-за требования неотрицательности двойственных переменных. Поэтому, чтобы получить обобщение метода Удзавы для решения таких задач, воспользуемся простейшей модификацией функции Лагранжа, в которой равенства и неравенства учитываются единым образом:

$$L(x, u) = f(x) + \langle p(u), g(x)\rangle. \tag{7.1.12}$$

Здесь $p(u)$ — m-мерная вектор-функция, зависящая от $u \in \mathbb{R}^m$ и имеющая вид

$$p^i(u) = \begin{cases} u^i, \ 1 \le i \le l, \\ \frac{(u^i)^2}{2}, \ l < i \le m. \end{cases}$$

Будем теперь отыскивать *стационарные точки* $[x_*, u_*]$ функции Лагранжа (7.1.12), в которых

$$L_x(x_*, u_*) = 0_n, \qquad L_u(x_*, u_*) = p_u(u_*)g(x_*) = 0_m.$$

Предположим, что в задаче (7.1.11) выполнены достаточные условия второго порядка теоремы 5.3.1 из [33], согласно которым точка $[x_*, u_*]$ должна быть стационарной и

$$\langle y, L_{xx}(x_*, u_*)y\rangle > 0, \qquad \langle z, L_{uu}(x_*, u_*)z\rangle < 0 \tag{7.1.13}$$

для любых ненулевых $y \in \mathbb{R}^n$ и $z \in \mathbb{R}^m$ таких, что

$$y \in K_1(x_*, u_*) = \left\{y \in \mathbb{R}^n : \ p_u(u_*)g_x(x_*)y = 0_m\right\},$$

$$z \in K_2(u_*) = \left\{z \in \mathbb{R}^m : \ p_u(u_*)z = 0_m\right\}.$$

Как и в случае задачи (7.1.1), усилим требование к матрице вторых производных $L_{xx}(x_*, u_*)$, а именно: будем считать, что она является положительно определенной, т.е. первое неравенство в (7.1.13) выполняется для любого ненулевого $y \in \mathbb{R}^n$, а не только когда $y \in K_1(x_*, u_*)$. При этом предположении $[x_*, u_*]$ оказывается седловой точкой функции Лагранжа (7.1.12), правда, по x лишь локально.

Снова применим обобщение метода градиентного спуска для отыскания седловых точек функции (7.1.12). Итерационный процесс запишется в виде

$$x_{k+1} = x_k - \alpha L_x(x_k, u_k), \qquad u_{k+1} = u_k + \alpha L_u(x_k, u_k). \tag{7.1.14}$$

Пусть в точке x_* выполнено *условие регулярности* ограничений, заключающееся в том, что градиенты всех активных в x_* ограничений (как равенств, так и неравенств) линейно независимы. Тогда можно показать, что итерационный процесс (7.1.14) локально сходится к стационарной точке $[x_*, u_*]$, если шаг $\alpha > 0$ берется достаточно малым. Для этого, как и при доказательстве теоремы 7.1.2, достаточно рассмотреть отображение $F(z) = [-L_x(z), L_u(z)]^T$, где опять обе переменные x и u объединены в единый вектор $z = [x, u]$, и убедиться, что все собственные значения матрицы

$$F_z(z_*) = \begin{bmatrix} -L_{xx}(x_*, u_*) & -g_x^\top(x_*)p_u(u_*) \\ p_u(u_*)g_x(x_*) & D(g(x_*)) \end{bmatrix}, \tag{7.1.15}$$

в которой $D(g)$ — диагональная матрица с вектором g на диагонали, имеют отрицательные действительные части. Тогда у соответствующей матрицы $Q_z(z_*) = I_{n+m} + \alpha F_z(z_*)$ спектральный радиус будет меньше единицы при α достаточно малом.

Упражнение 7. *Покажите, что действительные части всех собственных значений матрицы (7.1.15) отрицательны.*

7.2. Метод модифицированной функции Лагранжа

Рассмотрим сначала задачу нелинейного программирования, содержащую только ограничения-равенства:

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : \ g^i(x) = 0, \ 1 \le i \le m\right\}. \tag{7.2.1}$$

Будем предполагать, что как целевая функция $f(x)$, так и функции-ограничения $g^i(x)$, $1 \le i \le m$, являются по крайней мере непрерывно дифференцируемыми.

Составим для задачи (7.2.1) функцию Лагранжа:

$$L(x,u) = f(x) + \langle g(x), u\rangle, \quad u \in \mathbb{R}^m, \tag{7.2.2}$$

и вместо прибавления квадратичного штрафа к целевой функции, как это делается в методе внешних штрафных функций, прибавим данный штраф к функции Лагранжа. В результате приходим к измененной функции Лагранжа:

$$M(x,u,t) = f(x) + \sum_{i=1}^{m}\left[u^i g^i(x) + \frac{t}{2}\left(g^i(x)\right)^2\right], \tag{7.2.3}$$

где $t > 0$ — некоторый параметр.

Функцию (7.2.3) принято называть *модифицированной функцией Лагранжа* для задачи (7.2.1). Она обладает многими свойствами, присущими обычной функции Лагранжа $L(x,u)$, а именно: если вычислить производную $M(x,u,t)$ по x, то получаем

$$M_x(x,u,t) = f_x(x) + g_x^T(x)u + tg_x^T(x)g(x).$$

Отсюда видно, что

$$M_x(x,u,t) = L_x(x,u),$$

когда $x \in X$ и $u \in \mathbb{R}^m$. Кроме того,

$$M_u(x,u) = L_u(x,u) = g(x)$$

для тех же u и произвольных $x \in \mathbb{R}^n$. Поэтому вместо того, чтобы искать стационарные точки функции Лагранжа $L(x,u)$, т.е. точки $[x_*, u_*]$, в которых

$$L_x(x_*,u_*) = 0_n, \quad L_u(x_*,u_*) = g(x_*) = 0,$$

можно было бы искать стационарные точки модифицированной функции Лагранжа, удовлетворяющие равенствам

$$M_x(x_*, u_*, t) = 0_n, \quad M_u(x_*, u_*) = g(x_*) = 0_m. \tag{7.2.4}$$

Для обеих функций $L(x, u)$ и $M(x, u, t)$ они совпадают.

Однако модифицированная функция Лагранжа $M(x, u, t)$ обладает еще одним очень важным свойством, которое отсутствует у функции $L(x, u)$. Дело в том, что если в стационарной точке функции $L(x, u)$ выполнены достаточные условия изолированного локального минимума второго порядка

$$\langle y, L_{xx}(x_*, u_*)y\rangle > 0 \quad \forall\ y \in K(x_*),\ y \neq 0_n,$$

где $K(x_*) = \left\{y \in \mathbb{R}^n :\ g_x(x_*)y = 0_m\right\}$, то точка x_* может не являться точкой локального минимума функции $L(x, u_*)$ на всем пространстве $\mathbb{R}^n$. Гарантируется только, что точка x_* доставляет локальный минимум функции $L(x, u_*)$ по x лишь на некотором подпространстве, а именно нуль-пространстве матрицы $g_x(x_*)$. Это следствие того, что матрица $L_{xx}(x_*, u_*)$, вообще говоря, не обязана быть положительно полуопределенной.

В отличие от классической функции $L(x, u)$ модифицированная функция Лагранжа $M(x, u, t)$ в этой же самой точке $[x_*, u_*]$ при достаточно больших значениях параметра t уже имеет локальный минимум по x. Действительно, если функции $f(x)$ и $g(x)$, определяющие задачу (7.2.1), дважды непрерывно дифференцируемы, то матрица вторых производных функции $M_{xx}(x, u, t)$ по x в стационарной точке $[x_*, u_*]$ имеет вид

$$M_{xx}(x_*, u_*, t) = L_{xx}(x_*, u_*) + tg_x^T(x_*)g_x(x_*). \tag{7.2.5}$$

Здесь при подсчете производной учтено, что $g(x_*) = 0_m$ в стационарной точке $[x_*, u_*]$. Воспользуемся далее известным из линейной алгебры результатом, называемым иногда *леммой Финслера.*

Лемма 7.2.1. *Пусть A и B — симметричные квадратные матрицы порядка n, причем матрица B положительно полуопределенная. Пусть, кроме того,*

$$\langle y, Ay\rangle > 0,$$

когда $By = 0_n$, $y \neq 0_n$. Тогда можно указать $\bar{t} > 0$ такое, что

$$\langle y, Ay\rangle + t\langle y, By\rangle > 0$$

для всех $t > \bar{t}$ *и любого ненулевого* $y \in \mathbb{R}^n$.

На основании леммы 7.2.1 и представления (7.2.5) заключаем, что в точке $[x_*, u_*]$, в которой выполнены достаточные условия второго порядка, матрица $M_{xx}(x_*, u_*, t)$ положительно определена при t достаточно больших. Так как $M_x(x_*, u_*, t) = L_x(x_*, u_*) = 0_n$, то функция $M(x, u_*, t)$ имеет локальный минимум по x в точке x_* при этих t. В силу непрерывности это свойство сохраняется и в некоторой окрестности $\Delta(u_*)$ точки u_*. Действительно, по теореме о неявной функции в этой окрестности существует решение уравнения $M_x(x, u, t) = 0_n$, т.е. такая непрерывная функция $x(u)$, что $M_x(x(u), u, t) \equiv 0_n$ и $x(u_*) = x_*$. Из положительной определенности матрицы $M_{xx}(x, u, t)$ вблизи точки $[x_*, u_*]$ следует, что

$$M(x(u), u, t) = \min_{x \in \Delta(x_*)} M(x, u, t),$$

где $\Delta(x_*) \subseteq \mathbb{R}^n$ — некоторая окрестность точки x_*. Чтобы найти точку $[x_*, u_*]$ теперь просто следует решить уравнение

$$M_u(x(u), u, t) = g(x(u)) = 0_m. \tag{7.2.6}$$

Указанное свойство позволяет построить *алгоритм решения задачи* (7.2.1), основанный на отыскании соответствующей стационарной точки $[x_*, u_*]$ функции $M(x, u, t)$ путем решения уравнения (7.2.6) методом простой итерации. Пусть задано начальное $u_0 \in \mathbb{R}^m$. Пусть, кроме того, выбрано и зафиксировано достаточно большое значение параметра t, которое гарантировало бы существование решений соответствующих задач безусловной минимизации функции $M(x, u, t)$ по первой переменной x. Итерации в методе проводятся по следующей рекуррентной схеме:

$$\begin{array}{rcl} x_k & = & \operatorname{argmin}_{x \in R^n} M(x, u_k, t), \\ u_{k+1} & = & u_k + tg(x_k), \end{array} \tag{7.2.7}$$

причем для отыскания точки x_k в (7.2.7) могут применяться различные известные методы безусловной минимизации. Отметим также, что решать систему уравнений (7.2.6) можно не только методом простой итерации, но и другими известными методами решения систем нелинейных уравнений, например методом Ньютона.

Рассмотрим теперь вопрос о том, как данный подход можно перенести на задачу с ограничениями-неравенствами:

$$\min_{x \in X} f(x), \quad X = \left\{ x \in \mathbb{R}^n : \ h^j(x) \le 0, \ 1 \le j \le m \right\}, \tag{7.2.8}$$

где по-прежнему предполагается, что все функции, определяющие задачу, по крайней мере непрерывно дифференцируемы.

Задача с ограничениями-неравенствами сводится к задаче с ограничениями-равенствами путем введения дополнительных переменных

$$h^j(x) + (\sigma^j)^2 = 0, \quad j = 1, 2, \ldots, m.$$

Тогда, применяя тот же принцип, что и при построении функции (7.2.3), приходим к соответствующей модифицированной функции Лагранжа:

$$\tilde{M}(x, \sigma, v, t) = f(x) + \sum_{j=1}^{m} \left\{ v^j \left(h^j(x) + (\sigma^j)^2 \right) + \frac{t}{2} \left[h^j(x) + (\sigma^j)^2 \right]^2 \right\}, \tag{7.2.9}$$

где $v \in \mathbb{R}^m$.

Теперь при построении аналога метода (7.2.7) минимизацию придется проводить как по основной переменной x, так и по дополнительной переменной $\sigma = [\sigma^1, \ldots, \sigma^m]^T$. Но минимум функции $\tilde{M}(x, \sigma, v, t)$ по σ при фиксированных x и v легко вычисляется. В самом деле, дифференцируя функцию (7.2.9) по σ^j и приравнивая производную к нулю, получаем

$$\sigma^j \left[v^j + t \left(h^j(x) + (\sigma^j)^2 \right) \right] = 0. \tag{7.2.10}$$

Решением уравнения (7.2.10) является корень $\sigma^j = 0$, а также корни квадратичного уравнения

$$(\sigma^j)^2 + h^j(x) + \frac{v^j}{t} = 0. \tag{7.2.11}$$

Данное уравнение имеет действительный положительный корень, если $h^j(x) + \frac{v^j}{t} < 0$. При этом

$$\sigma^j = \sqrt{-\left(h^j(x) + \frac{v^j}{t} \right)}.$$

Таким образом,

$$h^j(x) + (\sigma^j)^2 = \begin{cases} h^j(x), & h^j(x) + \frac{v^j}{t} \geq 0, \\ -\frac{v^j}{t}, & h^j(x) + \frac{v^j}{t} < 0, \end{cases} \qquad 1 \leq j \leq m,$$

что можно записать также, как

$$h^j(x) + (\sigma^j)^2 = \max \left[h^j(x), -\frac{v^j}{t} \right], \qquad 1 \leq j \leq m. \tag{7.2.12}$$

Подставляя (7.2.12) в (7.2.9), после несложных преобразований приходим к модифицированной функции Лагранжа для задачи с неравенствами (7.2.8), уже избавленной от дополнительной переменной σ:

$$M(x,v,t) = f(x) + \frac{1}{2t}\sum_{j=1}^{m}\left[\left(v^j + th^j(x)\right)_+^2 - (v^j)^2\right], \tag{7.2.13}$$

где, как обычно, знак + внизу означает положительную срезку.

Обратимся теперь к общей задаче нелинейного программирования, в которой присутствуют как ограничения-равенства, так и ограничения-неравенства:

$$\min_{x\in X} f(x), \quad X = \{x \in \mathbb{R}^n : \ g(x) = 0_m, \ h(x) \le 0_l\}. \tag{7.2.14}$$

Тогда, объединяя функции (7.2.3) и (7.2.13), приходим к модифицированной функции Лагранжа для этой задачи:

$$\begin{array}{r} M(x,u,v,t) = f(x) + \sum_{i=1}^{m}\left[u^i g^i(x) + \frac{t}{2}\left(g^i(x)\right)^2\right] + \\ + \frac{1}{2t}\sum_{j=1}^{l}\left[\left(v^j + th^j(x)\right)_+^2 - (v^j)^2\right]. \end{array} \tag{7.2.15}$$

Пусть функции $f(x)$, $g(x)$ и $h(x)$, определяющие задачу (7.2.14), дважды непрерывно дифференцируемы и пусть x_* есть решение этой задачи. В том случае, когда в соответствующей точке Каруша–Куна–Таккера $[x_*, u_*, v_*]$ выполнены достаточные условия второго порядка для (7.2.14), матрица вторых производных $M_{xx}(x,u,v,t)$ модифицированной функции Лагранжа (7.2.15) в данной точке также оказывается положительно определенной для t достаточно больших (превышающих некоторое пороговое значение $\bar{t}$). Поэтому, как и в задаче с равенствами, вблизи точки $[u_*, v_*]$ существует непрерывное решение $x(u,v)$ задачи минимизации:

$$x(u,v) = \arg\min_{x\in R^n} M(x,u,v,t), \tag{7.2.16}$$

причем такое, что $x(u_*, v_*) = x_*$. Таким образом, чтобы найти точку x_*, достаточно решить систему уравнений:

$$M_u(x(u,v),u,v,t) = g(x(u,v)) = 0_m, \tag{7.2.17}$$

$$M_v(x(u,v),u,v,t) = \frac{1}{t}\left[(v + th(x(u,v)))_+ - v\right] = 0_l. \tag{7.2.18}$$

Опишем **алгоритм решения** задачи нелинейного программирования (7.2.14), основанный на использовании метода простой итерации для решения системы (7.2.17), (7.2.18).

Начальная итерация. Выбираем достаточно большое t, чтобы существовали решения задач безусловной минимизации (7.2.16). Выбираем также начальные $u_0 \in \mathbb{R}^m$ и $v_0 \in \mathbb{R}^l$. Полагаем $k = 0$.

Общая k-я итерация

Шаг 1. Решаем задачу безусловной минимизации (7.2.16) и находим

$$x_k = \arg\min_{x \in R^n} M(x, u_k, v_k, t). \tag{7.2.19}$$

Шаг 2. Пересчитываем

$$u_{k+1} = u_k + tg(x_k), \quad v_{k+1} = (v_k + th(x_k))_+ . \tag{7.2.20}$$

Шаг 3. Увеличиваем $k := k + 1$ и идем на шаг 1.

Имеет место следующий результат, касающийся сходимости метода (7.2.19), (7.2.20), который приведем без доказательства.

Теорема 7.2.1. *Пусть функции $f(x)$, $g(x)$ и $h(x)$ дважды непрерывно дифференцируемы и пусть точка $x_* \in X$ является решением задачи (7.2.14). Предположим также, что x_* вместе с $u_* \in \mathbb{R}^m$ и $v_* \in \mathbb{R}^l_+$ образуют точку Каруша–Куна–Таккера, в которой выполнены достаточные условия второго порядка для задачи (7.2.14). Тогда можно указать достаточно большое число $\bar{t} > 0$ такое, что для всех $t > \bar{t}$ и всех достаточно близких к $[u_*, v_*]$ начальных точек $[u_0, v_0]$ для последовательностей $\{u_k\}$ и $\{v_k\}$, вырабатываемых итерационным процессом (7.2.20), имеют место предельные равенства*

$$\lim_{k\to\infty} u_k = u_*, \quad \lim_{k\to\infty} v_k = v_*.$$

При этом последовательность $\{x_k\}$ также оказывается сходящейся, и

$$\lim_{k\to\infty} x_k = x_*.$$

Сделаем несколько замечаний, касающихся метода модифицированных функций Лагранжа (7.2.19), (7.2.20).

1. Метод прост в реализации и не требует выпуклости функций $f(x)$ и $h(x)$, хотя сходимость здесь в случае общих невыпуклых задач лишь локальная.

2. Так как значение параметра t конечно, то овражность функции $M(x, u_k, v_k, t)$ не увеличивается катастрофически с ростом

числа итерации, как это происходит в методе внешних штрафных функций. Поэтому решение задач безусловной минимизации (7.2.16) в методе модифицированных функций Лагранжа оказывается более простой процедурой по сравнению с методом внешних штрафных функций.

3. Рассмотренный метод модифицированной функции Лагранжа, как уже не раз подчеркивалось, локальный. Поэтому начальные значения множителей Лагранжа u_0 и v_0 следует брать достаточно близкими соответственно к u_* и v_*. Отсюда, в частности, следует, что множители v_0^j, $1 \le j \le l$, для обеспечения сходимости целесообразно выбирать неотрицательными, хотя из способа построения метода это не обязательно следует делать. Тогда согласно формуле пересчета множителей, соответствующих ограничениям-неравенствам, и на всех последующих итерациях будем получать $v_k^j \ge 0$.

4. Для ограничений-неравенств $h^j(x) \le 0$, которые являются неактивными в решении, т.е. $h^j(x_*) < 0$, для достаточно больших k выполняется $h^j(x_k) < 0$. Согласно второй из формул пересчета это приводит к тому, что на некоторой итерации соответствующий множитель v_k^j обнуляется и далее остается равным нулю. Фактически данное ограничение перестает влиять на вычислительный процесс.

5. Минимизацию модифицированной функции Лагранжа (7.2.15) по x можно проводить, отбрасывая слагаемые $(-v^j)^2$, $1 \le j \le l$. На точки минимума данные величины влияния не оказывают.

К модифицированной функции Лагранжа (7.2.13) можно прийти, проводя несколько иные рассуждения. Обратимся сначала к задаче (7.2.1) и к равенству $L_u(x_*, u_*) = g(x_*) = 0_m$, входящему в необходимые условия оптимальности для этой задачи. Двойственная переменная u принадлежит всему пространству $\mathbb{R}^m$, конус возможных направлений $\Gamma(u|\mathbb{R}^m)$ относительно $\mathbb{R}^m$ в каждой точке $u \in \mathbb{R}^m$ также совпадает со всем пространством $\mathbb{R}^m$, а сопряженный конус $\Gamma^*(u|\mathbb{R}^m)$ состоит из единственного вектора 0_m. Равенство $L_u(x_*, u_*) = 0_m$ можно записать в виде

$$-L_u(x_*, u_*) \in \Gamma^*(u_*|\mathbb{R}^m), \quad u_* \in \mathbb{R}^m, \tag{7.2.21}$$

что является необходимым условием максимума вогнутой (точнее, линейной) по u функции $L(x, u)$ на $\mathbb{R}^m$. Более того, оно является и достаточным.

Если теперь перейти от задачи с равенствами к задаче с неравенствами (7.2.8), то функция Лагранжа примет вид

$$L(x,v) = f(x) + \langle h(x), v\rangle, \quad v \in \mathbb{R}^m_+, \tag{7.2.22}$$

и необходимое условие $L_u(x_*,u_*) = g(x_*) = 0_m$ заменится на условия:

$$L_v(x_*,v_*) = h(x_*) \le 0_m, \quad v_* \ge 0_m, \quad \langle v_*, L_v(x_*,v_*)\rangle = 0, \tag{7.2.23}$$

из которых вытекает, в частности, условие дополняющей нежесткости, а именно: $v_*^j h^j(x_*) = 0$, $1 \le j \le m$.

Вся совокупность равенств и неравенств (7.2.23) выражает необходимое и достаточное условие максимума функции $L(x_*,v)$ по v на ортанте $\mathbb{R}^m_+$. Это условие также можно переписать в эквивалентной форме, заключающейся в принадлежности антиградиента $-L_v(x_*,v_*)$ конусу $\Gamma^*(v_*|\mathbb{R}^m_+)$, являющимся сопряженным к конусу возможных направлений $\Gamma(v_*|\mathbb{R}^m_+)$ в точке $v_* \in \mathbb{R}^m_+$.

$$-L_v(x_*,v_*) \in \Gamma^*(v_*|\mathbb{R}^m_+), \quad v_* \in \mathbb{R}^m_+. \tag{7.2.24}$$

Как было показано в [33], вид этих конусов полностью подходит для характеризации условий (7.2.23).

Строя модифицированную функцию Лагранжа для задачи с равенствами (7.2.1), мы добавляли внешний штраф к классической функции $L(x,u)$ за нарушение условия (7.2.21). Поступим аналогичным образом и для задачи с неравенствами (7.2.8), а именно добавим внешний штраф за нарушение условия (7.2.24). Но предварительно нам надо построить такую функцию, которая действительно обладала бы «штрафными свойствами» в случае более сложного условия (7.2.24).

Пусть Q — произвольное выпуклое замкнутое множество в $\mathbb{R}^m$ и пусть $q \in Q$. Возьмем произвольную выпуклую функцию $F(p)$ и составим функцию

$$B(p,q|Q) = \left[F(p) + \delta(p|Q-q)\right]^*, \tag{7.2.25}$$

где $\delta(p|Q)$ — индикаторная функция множества Q, звездочка означает сопряжение относительно первой переменной p. Функция $B(p,q|Q)$, как сопряженная функция к сумме двух выпуклых функций, выпукла по p. Считаем, что она определена всюду на $\mathbb{R}^m$.

Лемма 7.2.2. *Пусть $F(p)$ — дифференцируемая выпуклая функция, достигающая своего минимума на $\mathbb{R}^m$ в нуле. Тогда функция $B(p,q|Q)$ при фиксированном $q \in Q$ достигает своего минимума по p в точках множества $-\Gamma^*(q|Q)$.*

Доказательство. Обозначим

$$P_*(q) = \operatorname{Arg}\min_{p\in R^m} B(p,q|Q).$$

Согласно утверждению 3.4.5 из [33] множество $P_*(q)$ совпадает с субдифференциалом сопряженной функции к функции $B(p,q|Q)$ по первому аргументу в нуле. Но эта сопряженная функция, как вторая сопряженная функция, совпадает с выпуклой функцией $F(p)+\delta(p|Q-q)$. Поэтому на основании теоремы Моро–Рокафеллара:

$$P_*(q) = \partial F(0_m) + \partial\delta(0_m|Q-q) = \partial\delta(0_m|Q-q).$$

Здесь учтено, что функция $F(p)$ дифференцируема и $F_p(0_m) = 0_m$.

Остается только заметить, что субдифференциал индикаторной функции $\delta(p|Q-q)$ в нуле совпадает с субдифференциалом индикаторной функции $\delta(p|Q)$ в точке $q \in Q$, а последний, в свою очередь, равен нормальному конусу к множеству Q в точке q. Но данный нормальный конус является обратным к конусу, который является сопряженным к конусу возможных направлений в точке $q \in Q$ относительно множества Q. Отсюда приходим к выводу, что $P_*(q) = -\Gamma^*(q|Q)$. ■

Построим по формуле (7.2.25) функцию $B(p,q|Q)$, используя в качестве $F(p)$ квадратичную функцию $F(p) = \frac{1}{2}\|p\|^2$, где $\|p\|$ — евклидова норма вектора p. В силу утверждения 3.4.7 из [33]

$$[F(p) + \delta(p \mid Q-q)]^* = F^*(p) \oplus \delta^*(p \mid Q-q),$$

где, напомним, $\oplus$ — знак операции инфимальной конволюции. Но квадратичная функция $F(p)$ самосопряженная, т.е. $F^*(p) = F(p)$. Помимо того, для опорной функции $\delta^*(p|Q-q)$ множества $Q-q$ справедливо представление

$$\delta^*(p|Q-q) = \delta^*(p|Q) - \langle p, q\rangle.$$

Тогда, согласно определениям опорной функции и операции инфимальной конволюции

$$\begin{aligned} B(p,q|Q) &= \inf_{p_1+p_2=p}\left[\tfrac{1}{2}\|p_1\|^2 + \sup_{q_1\in Q}\langle q_1,p_2\rangle - \langle q,p_2\rangle\right] = \\ &= \inf_{p_2\in R^n}\sup_{q_1\in Q}\left[\tfrac{1}{2}\|p-p_2\|^2 + \langle p_2, q_1-q\rangle\right]. \end{aligned} \tag{7.2.26}$$

Рассмотрим функцию

$$B_1(p,q|Q) = \sup_{q_1\in Q}\inf_{p_2\in R^n}\left[\frac{1}{2}\|p-p_2\|^2 + \langle p_2, q_1-q\rangle\right], \tag{7.2.27}$$

в которой в отличие от (7.2.26) операции взятия инфимума и супремума переставлены местами. Минимум по p_2 на $\mathbb{R}^n$ для любых фиксированных $q_1 \in Q$ достигается в точке $p_2 = p + q - q_1$. Подставляя это выражение в (7.2.27), получаем

$$\begin{aligned} B_1(p,q|Q) &= \sup_{q_1 \in Q} \left[\tfrac{1}{2}\|q_1 - q\|^2 + \langle q_1 - q, p + q - q_1\rangle\right] = \\ &= \tfrac{1}{2} \sup_{q_1 \in Q} \left[-\|q_1 - q\|^2 + 2\langle p, q_1 - q\rangle\right] = \\ &= \tfrac{1}{2} \sup_{q_1 \in Q} \left[\|p\|^2 - \|p + q - q_1\|^2\right] = \\ &= \tfrac{1}{2} \left[\|p\|^2 - \inf_{q_1 \in Q} \|p + q - q_1\|^2\right]. \end{aligned} \tag{7.2.28}$$

Пусть $\pi_Q(p+q)$ — проекция вектора $p+q$ на множество Q. В силу выпуклости и замкнутости множества Q она существует и единственна. Отсюда следует, что инфимум по q_1 в (7.2.28) всегда достигается, причем в единственной точке. Поэтому по обобщенной теореме фон-Неймана функции (7.2.26) и (7.2.27) совпадают. Таким образом,

$$B(p,q|Q) = \frac{1}{2}\left[\|p\|^2 - \|p + q - \pi_Q(p+q)\|^2\right]. \tag{7.2.29}$$

В частности, для множества $Q = \mathbb{R}^n_+$ получаем $\pi_{R^n_+}(p+q) = (p+q)_+$ и, следовательно,

$$\begin{aligned} B(p,q|\mathbb{R}^n_+) &= \frac{1}{2}\left[\|p\|^2 - \|p + q - (p+q)_+\|^2\right] = \\ &= \frac{1}{2}\left[\|p\|^2 - \|(p+q)_-\|^2\right]. \end{aligned} \tag{7.2.30}$$

Здесь b_+ и b_- — соответственно положительные и отрицательные срезки вектора b.

Обозначим через $(Bt)(p,q|Q)$ функцию $B(p,q|Q)$, умноженную справа на положительную константу t, причем считается, что эта операция относится только к первому аргументу. Согласно определению, это означает, что

$$(Bt)(p,q\,|\,Q) = tB\left(\frac{p}{t}, q\,|\,Q\right).$$

Тогда после прибавления к функции Лагранжа (7.2.22), составленной для задачи с ограничениями-неравенствами, штрафа за нарушение условия (7.2.24), который выражается теперь посредством функции $B(p,q|\mathbb{R}^m_+)$, после некоторых выкладок приходим к следующей модифицированной функции Лагранжа:

$$\begin{aligned} M(x,v;t) &= L(x,v) + (Bt)(h(x),v|\mathbb{R}^m_+) = \\ &= f(x) + \langle h(x), v\rangle + \tfrac{t}{2}\left[\|\tfrac{h(x)}{t}\|^2 - \|(\tfrac{h(x)}{t} + v)_-\|^2\right] = \\ &= f(x) + \tfrac{t}{2}\left[\|(\tfrac{h(x)}{t} + v)_+\|^2 - \|v\|^2\right]. \end{aligned} \tag{7.2.31}$$

Данная функция полностью совпадает с модифицированной функцией Лагранжа (7.2.13), полученной ранее. Используя иные функции $F(p)$, отличные от квадратичной, можно посредством соответствующих функций $B(p,q|\mathbb{R}^m_+)$ по формуле (7.2.31) получить другие модифицированные функции Лагранжа.

Так как проекция точки на выпуклое замкнутое множество единственна, то функция (7.2.29) дифференцируема по первому аргументу. Действительно, тогда функция $\phi(p,q) = \frac{1}{2}\|p+q-\pi_Q(p+q)\|^2$ дифференцируема по первой переменной по всем направлениям $s \in \mathbb{R}^m$ и имеет место формула

$$\phi_p(p,q;s) = \langle p+q-\pi_Q(p+q), s\rangle. \tag{7.2.32}$$

Отсюда следует, что функция $\phi(p,q)$ обладает частными производными по p. Но поскольку функция $\phi(p,q)$ выпукла, то, как можно показать, существование частных производных достаточно, чтобы функция $\phi(p,q)$ была дифференцируемой. Из (7.2.29) и (7.2.32) приходим к выводу, что функция $B(p,q|Q)$ является дифференцируемой по переменной p и ее градиент равен

$$B_p(p,q\,|\,Q) = \pi_Q(p+q) - q.$$

В частности, для множества $Q = \mathbb{R}^m_+$ получаем

$$B_p(p,q\,|\,\mathbb{R}^m_+) = (p+q)_+ - q. \tag{7.2.33}$$

Из этого выражения следует, что функция $B(p,q\,|\,Q)$ в некоторых точках оказывается даже дважды дифференцируемой.

Лемма 7.2.3. *Пусть* $1 \le s \le m$ *и пусть точки* $p_* \in \mathbb{R}^m$ *и* $q_* \in \mathbb{R}^m_+$ *таковы, что*

$$p_* = \left[0,\ldots,0,p_*^{s+1},\ldots,p_*^m\right]^T, \quad q_* = \left[q_*^1,\ldots,q_*^s,0,\ldots,0\right]^T,$$

а компоненты p_*^i, $s < i \le m$, *и* q_*^j, $1 \le j \le s$, *удовлетворяют неравенствам:* $p_*^i < 0$, $q_*^j > 0$. *Тогда* $B_p(p_*,q_*\,|\,\mathbb{R}^m_+) = 0_m$. *Более того, функция* $B(p,q\,|\,\mathbb{R}^m_+)$ *дважды дифференцируема по* p *в точке* $[p_*,q_*]$ *и ее матрица вторых производных имеет вид*

$$B_{pp}(p_*,q_*|\mathbb{R}^m_+) = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix}. \tag{7.2.34}$$

Если $x_* \in X$ есть решение задачи (7.2.8) и в этой точке выполнены достаточные условия второго порядка, то x_* вместе с некоторым вектором $v_* \geq 0_m$ образуют точку Каруша–Куна–Таккера $[x_*, v_*]$. Предположим, не умаляя общности, что ограничения $h^i(x)$, $0 \leq i \leq s$, являются активными в точке x_*, а остальные ограничения — неактивными. В этом случае обязательно $v_*^j = 0$, $s < j \leq m$. Но тогда, используя утверждение леммы 7.2.3 с $p_* = h(x_*)$ и $q_* = v_*$, получаем

$$M_x(x_*, v_*; t) = L_x(x_*, v_*) + h_x^T(x_*) B_p\left(\frac{h(x_*)}{t}, v_* | \mathbb{R}_+^m\right) = L_x(x_*, v_*) = 0_n. \tag{7.2.35}$$

Предположим далее, что в точке $[x_*, v_*]$ выполнено условие строгой дополняющей нежесткости, т.е. $v_*^j > 0$ при $1 \leq j \leq s$. Тогда

$$M_{xx}(x_*, v_*; t) = L_{xx}(x_*, v_*) + t^{-1} h_x^T(x_*) B_{pp}\left(\frac{h(x_*)}{t}, v_* \,|\, \mathbb{R}_+^m\right) h_x(x_*).$$

Учитывая вид (7.2.34) матрицы $B_{pp}\left(\frac{h(x_*)}{t}, v_* \,|\, \mathbb{R}_+^m\right)$, с помощью леммы 7.2.1 убеждаемся, что матрица $M_{xx}(x_*, v_*; t)$ для достаточно малых t будет положительно определенной. Поэтому x_* является точкой локального минимума функции $M(x, v_*; t)$ по x при этих t.

Более того, положительная определенность матрицы $M_{xx}(x_*, v_*; t)$ позволяет нам, как и ранее, утверждать, что в некоторой окрестности точки v_* существует функция $x(v)$, удовлетворяющая тождеству $M_x(x(v), v; t) \equiv 0_n$ и доставляющая локальный минимум функции $M(x, v; t)$ по x.

Если теперь обратиться к задаче отыскания точки v такой, что

$$B_p\left(\frac{h(x(v))}{t}, v \,|\, \mathbb{R}_+^m\right) = 0_m, \tag{7.2.36}$$

то, как видно из (7.2.35), решив ее, мы найдем точку Каруша–Куна–Таккера $[x_*, v_*]$ в задаче (7.2.8). Здесь $x_* = x(v_*)$. Для решения системы (7.2.36) может быть применен метод простой итерации или метод Ньютона. Например, использование метода простой итерации приводит к итерационному процессу:

$$x_k = \arg\min_{x \in R^n} M(x, v_k; t), \quad v_{k+1} = \left[v_k + \frac{h(x_k)}{t}\right]_+,$$

что полностью аналогично итерационному процессу (7.2.20) для задач с ограничениями-неравенствами.

7.3. Другие варианты методов МФЛ

Наряду с рассмотренным основным методом МФЛ существуют и другие его варианты. В качестве примера приведем два из них, предназначенных для решения общей задачи нелинейного программирования с равенствами и неравенствами:

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : \ g^i(x) = 0, \ 1 \le i \le l; \ \ g^i(x) \le 0, \ l < i \le m\right\}. \tag{7.3.1}$$

В отличие от (7.2.2) теперь мы будем модифицировать функцию Лагранжа (7.1.12), т.е. функцию

$$L(x,u) = f(x) + \langle p(u), g(x)\rangle, \tag{7.3.2}$$

где $p^i(u) = u^i$, когда $1 \le i \le l$ и $p^i(u) = \frac{(u^i)^2}{2}$, когда $l < i \le m$.

Предполагается, что функции, определяющие общую задачу нелинейного программирования (7.3.1), являются, по крайней мере, дважды непрерывно дифференцируемыми. Тогда согласно достаточным условиям второго порядка точка $x_* \in \mathbb{R}^n$ будет решением задачи (7.3.1), если x_* вместе с некоторым $u_* \in \mathbb{R}^m$ образуют стационарную точку функции (7.3.2), т.е.

$$L_x(x_*, u_*) = 0_n \qquad L_u(x_*, u_*) = 0_m. \tag{7.3.3}$$

Кроме того:

$$\langle y, L_{xx}(x_*, u_*)y\rangle > 0 \quad \forall\ y \in K_1(x_*, u_*) \quad y \ne 0_n, \tag{7.3.4}$$

$$\langle z, L_{uu}(x_*, u_*)z\rangle < 0 \quad \forall\ z \in K_2(u_*), \quad z \ne 0_m, \tag{7.3.5}$$

где

$$K_1(x,u) = \left\{y \in \mathbb{R}^n : \ p_u(u)g_x(x)y = 0_m\right\},$$
$$K_2(u) = \left\{z \in \mathbb{R}^m : \ p_u(u)z = 0_m\right\}.$$

В этом случае, как было показано в [33], в паре $[x_*, u_*]$ выполнено условие строгой дополняющей нежесткости для ограничений типа неравенства.

Основная идея, которая используется при модификации функции (7.3.2), опять заключается в улучшении ее свойств с помощью «штрафования» отдельных условий оптимальности. Как мы видели, это проще сделать, когда условия оптимальности записываются в виде равенств. В данном случае такими условиями являются условия стационарности (7.3.3).

7.3.1. Двойственные методы МФЛ

Рассмотрим сначала один из возможных подходов к модификации функции Лагранжа (7.3.2), который приводит к классу *двойственных методов*. Термин «двойственный» здесь применяется с целью подчеркнуть, что основные внешние итерации в методе проводятся по дополнительным двойственным переменным. Построенный в предыдущем параграфе основной вариант метода МФЛ также является в этом смысле двойственным методом.

Пусть $\|\cdot\|$ обозначает евклидову норму в пространстве векторов. Возьмем функцию Лагранжа (7.3.2) и составим с ее помощью МФЛ для общей задачи нелинейного программирования с ограничениями типа равенства и неравенства (7.3.1):

$$M(x,u) = L(x,u) + \frac{\alpha}{2}\|L_u(x,u)\|^2, \tag{7.3.6}$$

где $\alpha > 0$ — некоторый фиксированный параметр.

Введем вспомогательную задачу минимизации:

$$\min_{x \in R^n} M(x,u). \tag{7.3.7}$$

Ее решением является, вообще говоря, точечно-множественное отображение $X(u)$. Покажем, что в отличие от исходной функции Лагранжа (7.3.2) решение задачи минимизации (7.3.7) существует в некоторой окрестности стационарной точки $[x_*, u_*]$ для α достаточно больших. Другими словами, отображение $X(u)$ определено в окрестности точки u_*. Если множество $X(u)$ состоит из единственной точки, то будем писать $x(u)$ вместо $X(u)$.

Лемма 7.3.1. *Пусть функции $f(x)$ и $g^i(x)$, $1 \le i \le m$, дважды непрерывно дифференцируемы на $\mathbb{R}^n$. Пусть, кроме того, в стационарной точке $[x_*, u_*]$ функции Лагранжа (7.3.2) выполнено условие второго порядка (7.3.4).*

Тогда можно указать такую константу $\bar{\alpha} > 0$, что для $\alpha > \bar{\alpha}$ и всех u из некоторой окрестности $\Delta(u_*)$ точки u_* существует изолированное локальное решение $x(u)$ задачи минимизации (7.3.7), удовлетворяющее условию $x(u_*) = x_*$. Функция $x(u)$ является непрерывно дифференцируемой на $\Delta(u_*)$.

Доказательство. Если решение задачи (7.3.7) — точка $x(u)$ — существует, то необходимо, чтобы

$$M_x(x,u) = L_x(x,u) + \alpha L_{xu}(x,u) L_u(x,u) = 0_n \tag{7.3.8}$$

при $x = x(u)$. Стационарная точка $[x_*, u_*]$ удовлетворяет этому равенству.

Вычислим матрицу

$$M_{xx}(x_*, u_*) = L_{xx}(x_*, u_*) + \alpha L_{xu}(x_*, u_*) L_{ux}(x_*, u_*). \tag{7.3.9}$$

Здесь мы учли, что $L_u(x_*, u_*) = 0_m$.

Согласно условию (7.3.4), $\langle y, L_{xx}(x_*, u_*)y\rangle > 0$ для любого ненулевого вектора $y \in \mathbb{R}^n$ такого, что $L_{ux}(x_*, u_*)y = 0$. Поэтому по лемме Финслера 7.2.1 матрица $M_{xx}(x_*, u_*)$ положительно определена при достаточно больших α, $\alpha > \bar{\alpha}$. Тогда, согласно теореме о неявной функции, можно указать такую окрестность $\Delta(u_*)$ точки u_*, что на $\Delta(u_*)$ существует непрерывно дифференцируемая функция $x(u)$, удовлетворяющая уравнению (7.3.8) и принимающая при $u = u_*$ значение x_*. Так как всегда можно выбрать окрестность $\Delta(u_*)$ настолько малой, чтобы при $u \in \Delta(u_*)$, $x = x(u)$ матрица $M_{xx}(x, u)$ оставалась положительно определенной, то $x(u)$ является решением задачи минимизации (7.3.7). ■

Воспользуемся решением задачи минимизации (7.3.7) — функцией $x(u)$. Если подставить $x(u)$ в (7.3.8), то данное равенство обращается в тождество по u:

$$M_x(x(u), u) = L_x(x(u), u) + \alpha L_{xu}(x(u), u) L_u(x(u), u) \equiv 0_n. \tag{7.3.10}$$

Обратимся теперь к задаче отыскания корней системы

$$L_u(x(u), u) = 0_m. \tag{7.3.11}$$

Из (7.3.10) видно, что, решив уравнение (7.3.11), мы получим стационарную точку функции Лагранжа (7.3.2).

Применим для решения системы (7.3.11) метод простой итерации:

$$u_{k+1} = u_k + \alpha L_u(x_k, u_k), \tag{7.3.12}$$

где $x_k = x(u_k)$, α — некоторый шаг. Докажем локальную сходимость итерационного процесса (7.3.12). С целью упрощения доказательства предположим, что все ограничения типа неравенства в решении задачи выполняются как равенства, т.е. являются активными. Тогда шаг α в (7.3.12) можно брать совпадающим со значением параметра α в (7.3.6).

Теорема 7.3.1. *Пусть функции $f(x)$ и $g^i(x)$, $1 \le i \le m$, дважды непрерывно дифференцируемы на $\mathbb{R}^n$. Предположим также, что*

1) пара $[x_*, u_*]$ является стационарной точкой функции Лагранжа $L(x, u)$;

2) в $[x_*, u_*]$ выполнены условия (7.3.4) и (7.3.5);

3) матрица $L_{xx}(x_*, u_*)$ неособая;

4) в точке x_* выполнено условие регулярности ограничений;

5) все ограничения типа неравенства в точке x_* являются активными, т.е. $J_0(x_*) = \{l+1, \ldots, m\}$.

Тогда найдется такое $\alpha_* > 0$, что для всех $\alpha > \alpha_*$ процесс (7.3.12) локально сходится к точке u_*. Соответствующая последовательность $\{x_k\}$ сходится к точке x_*.

Доказательство. Для доказательства локальной сходимости метода (7.3.12) воспользуемся теоремой Островского 7.1.1. Согласно этой теореме, процесс (7.3.12) локально сходится к точке u_*, если отображение

$$Q(u) = u + \alpha L_u(x(u), u) \tag{7.3.13}$$

дифференцируемо в u_* и спектральный радиус ρ матрицы $Q_u(u_*)$ удовлетворяет условию

$$\rho(Q_u(u_*)) < 1. \tag{7.3.14}$$

Покажем, что найдется такое $\alpha_* > 0$, для которого неравенство (7.3.14) имеет место для всех $\alpha > \alpha_*$. Прежде всего, используя результат леммы 7.3.1, выберем такое $\bar{\alpha}$, что для всех $\alpha > \bar{\alpha}$ решение задачи (7.3.7) — функция $x(u)$ — существует в некоторой окрестности $\Delta(u_*)$ точки u_* и $x(u_*) = x_*$. Тогда при этих α отображение (7.3.13) полностью определено и непрерывно дифференцируемо на $\Delta(u_*)$.

Продифференцировав тождество (7.3.8) по u, получим

$$M_{xx}(x(u), u)\frac{dx(u)}{du} + M_{xu}(x(u), u) = 0_n.$$

Отсюда, поскольку матрица $M_{xx}(x(u), u)$ при $\alpha > \bar{\alpha}$ положительно определена, находим

$$\frac{dx}{du} = -M_{xx}^{-1}(x(u), u)M_{xu}(x(u), u). \tag{7.3.15}$$

Вычислим теперь матрицу $Q_u(u_*)$. Согласно (7.3.13) и (7.3.15),

$$\begin{aligned} Q_u(u_*) &= I_m + \alpha\left[L_{uu}(x_*, u_*) + L_{ux}(x_*, u_*)\frac{dx(u_*)}{du}\right] = \\ &= I_m + \alpha\left[L_{uu}(x_*, u_*) - L_{ux}(x_*, u_*)M_{xx}^{-1}(x_*, u_*)M_{xu}(x_*, u_*)\right]. \end{aligned} \tag{7.3.16}$$

Найдем матрицы $M_{xx}^{-1}(x_*, u_*)$ и $M_{xu}(x_*, u_*)$. Для сокращения записи будем опускать аргументы у всех встречающихся матриц, считая,

что они вычисляются при $x = x_*$, $u = u_*$. На основании (7.3.9)

$$M_{xx}^{-1} = \left(L_{xx} + \alpha L_{xu} L_{ux}\right)^{-1}.$$

Так как L_{xx} — неособая матрица, то по формуле Шермана–Моррисона–Вудберри получаем

$$M_{xx}^{-1} = L_{xx}^{-1} - \alpha L_{xx}^{-1} L_{xu} \left(I_m + \alpha L_{ux} L_{xx}^{-1} L_{xu}\right)^{-1} L_{ux} L_{xx}^{-1}. \qquad (7.3.17)$$

Матрица M_{xu} имеет вид

$$M_{xu} = L_{xu} + \alpha L_{xu} L_{uu}. \qquad (7.3.18)$$

Здесь учтено, что $L_u(x_*, u_*) = 0_m$.

Если подставить (7.3.18) в (7.3.16), то получаем

$$Q_u(u_*) = \left(I_m - \alpha L_{ux} M_{xx}^{-1} L_{xu}\right)\left(I_m + \alpha L_{uu}\right). \qquad (7.3.19)$$

Воспользуемся теперь предположением, что все ограничения в точке x_* активны. В этом случае матрица L_{uu} является нулевой и выражение (7.3.19) для матрицы $Q_u(u_*)$ упрощается:

$$Q_u(u_*) = I_m - \alpha L_{ux} M_{xx}^{-1} L_{xu}. \qquad (7.3.20)$$

После подстановки (7.3.17) в (7.3.20) приходим к выражению

$$Q_u(u_*) = I_m - \alpha L_{ux}\left[L_{xx}^{-1} - \alpha L_{xx}^{-1} L_{xu}\left(I_m + \alpha L_{ux} L_{xx}^{-1} L_{xu}\right)^{-1} L_{ux} L_{xx}^{-1}\right] L_{xu}.$$

Обозначим $Z = L_{ux} L_{xx}^{-1} L_{xu}$. Тогда матрицу $Q_u(u_*)$ можно представить в следующем виде:

$$Q_u(u_*) = I_m - \alpha Z + \alpha^2 Z \left(I_m + \alpha Z\right)^{-1} Z.$$

Поскольку справедливо тождество

$$I_m - \alpha Z + \alpha^2 Z \left(I_m + \alpha Z\right)^{-1} Z = \left(I_m + \alpha Z\right)^{-1},$$

то окончательно получаем $Q_u(u_*) = \left(I_m + \alpha Z\right)^{-1}$.

Найдем собственные значения матрицы $Q_u(u_*)$. Все они действительны, так как матрица Z симметрическая. Пусть μ — собственные значения матрицы $Q_u(u_*)$, ν — собственные значения матрицы Z. Числа μ удовлетворяют характеристическому уравнению

$$\left|\left(I_m + \alpha Z\right)^{-1} - \mu I_m\right| = 0.$$

Матрица $I_m + \alpha Z$ неособая, поэтому, согласно теореме о произведении определителей, собственные значения μ удовлетворяют также уравнению $|I_m - \mu\,(I_m + \alpha Z)\,| = 0$, которое можно переписать в виде

$$|Z - \frac{1-\mu}{\alpha\mu} I_m| = 0.$$

Отсюда следует, что собственные значения μ связаны с собственными значениями ν следующим соотношением:

$$\mu = \frac{1}{1+\alpha\nu}. \tag{7.3.21}$$

Пусть ν^i — произвольное собственное значение матрицы Z. Оно не равно нулю, поскольку из условия регулярности ограничений и условия строгой дополняющей нежесткости, вытекающего из (7.3.3), следует, что матрица L_{ux} — полного ранга. Предположим сначала, что $\nu^i > 0$. Тогда для соответствующего собственного значения μ^i на основании (7.3.21) получаем, что $0 < \mu^i < 1$ для всех $\alpha > 0$. Если $\nu^i < 0$, то, как видно из (7.3.21), $-1 < \mu^i < 0$ при $\alpha > -\frac{2}{\nu^i}$.

Положим

$$\alpha_* = \begin{cases} \max[\bar\alpha, -\frac{2}{\nu_*}], & \nu_{\min} < 0, \\ \bar\alpha, & \nu_{\min} > 0, \end{cases}$$

где $\nu_{\min}$ — минимальное собственное значение из набора ν, и ν_* — минимальное по модулю отрицательное ν^i. Тогда при $\alpha > \alpha_*$ неравенство (7.3.14) действительно выполняется. ■

Отметим, что шаг α в итерационном процессе (7.3.12) может оказаться достаточно большим, тем не менее имеет место сходимость процесса (7.3.12). Отметим также, что, как уже говорилось, предположение об активности всех ограничений типа неравенства в точке x_* не является существенным и может быть отброшено. В случае, когда не все ограничения являются активными в точке x_*, шаг α в процессе (7.3.12) может отличаться от константы α, используемой в МФЛ (7.3.7). В этом случае он всегда должен быть меньше этой константы.

Упражнение 8. *Покажите, что при наличии неактивных ограничений типа неравенства в точке $x_* \in X_*$ шаг α в итерационном процессе (7.3.12) следует заменить меньшим шагом β и брать β из интервала $\big(0, \bar\beta(\alpha)\big)$, где $\alpha > \alpha_*$ и*

$$\bar\beta(\alpha) = \min\{\alpha, \max_{i \in J_-(x_*)} [-g^i(x_*)]\}, \quad J_-(x_*) = \{l < i \le m : g^i(x_*) < 0\}.$$

Указание. *Убедитесь, что в этом случае все собственные значения матрицы $Q_u(u_*)$ распадаются на две группы. Первая группа состоит из собственных значений, соответствующих ограничениям типа равенства и активным ограничениям типа неравенства, а вторая — из собственных значений, соответствующих неактивным ограничениям типа неравенства. Тем не менее все они остаются вещественными и строго меньшими единицы по модулю.*

Для решения системы уравнений (7.3.11) можно воспользоваться методом Ньютона. Тогда приходим к следующему итерационному процессу:

$$u_{k+1} = u_k - G^{-1}(u_k) L_u(x_k, u_k). \tag{7.3.22}$$

Здесь $x_k = x(u_k)$ и

$$G(u) = \frac{d}{du} L_u(x(u), u)$$

— матрица Якоби отображения $L_u(x(u), u)$.

Теорема 7.3.2. *Пусть $f(x)$ и $g^i(x)$, $1 \le i \le m$, — трижды непрерывно дифференцируемые функции и выполнены предположения теоремы 7.3.1. Тогда существует такое $\alpha_* > 0$, что для всех $\alpha > \alpha_*$ процесс (7.3.22) локально сходится к u_*.*

Доказательство. В соответствии с общими условиями сходимости метода Ньютона надо показать, что матрица $G(u_*)$ неособая для достаточно больших α. Пусть $\alpha_* = \bar{\alpha}$, где константа $\bar{\alpha}$ такова, что выполнено утверждение леммы 7.3.1. Имеем при $\alpha > \alpha_*$

$$G(u_*) = L_{uu}(x_*, u_*) + L_{ux} \frac{dx(u_*)}{du}.$$

Поэтому, согласно (7.3.15) и (7.3.18),

$$\begin{aligned} G(u_*) &= L_{uu} - L_{ux} M_{xx}^{-1} M_{xu} = L_{uu} - L_{ux} M_{xx}^{-1} \left(L_{xu} + \alpha L_{xu} L_{uu}\right) = \\ &= \left(I_m - \alpha \Phi\right) L_{uu} - \Phi, \end{aligned}$$

где введено обозначение

$$\Phi = L_{ux}(x_*, u_*) M_{xx}^{-1}(x_*, u_*) L_{xu}(x_*, u_*).$$

Из активности всех ограничений в точке x_* следует, что матрица L_{uu} нулевая, поэтому вид матрицы $G(u_*)$ упрощается: $G(u_*) = -\Phi$.

При доказательстве теоремы 7.3.1 фактически было получено равенство $I_m - \alpha\Phi = (I_m + \alpha Z)^{-1}$, откуда следует, что

$$\Phi = \frac{1}{\alpha} \left[I_m - (I_m + \alpha Z)^{-1} \right]. \tag{7.3.23}$$

Умножим левую и правую части равенства (7.3.23) на неособую матрицу $I_m + \alpha Z$, тогда оно принимает вид:

$$(I_m + \alpha Z)\,\Phi = -Z.$$

Но симметрическая матрица Z, как отмечалось при доказательстве теоремы 7.3.1, является неособой. Поскольку Z есть произведение Φ на неособую матрицу, матрица Φ должна быть неособой. ■

Как и в итеративном процессе (7.3.12), предположение о том, что все ограничения в точке x_* являются активными, здесь несущественно и введено лишь с целью упрощения доказательства теоремы.

Отметим также, что самая трудоемкая операция при численной реализации метода (7.3.22) — это вычисление матрицы $G(u_k)$, используемой в линейных системах алгебраических уравнений для определения сдвига в итерационном процессе (7.3.22). Возможен вариант метода, когда $G(u_k)$ пересчитывается не на каждой итерации, а лишь на выделенной подпоследовательности номеров.

7.3.2. Прямые методы МФЛ

Рассмотрим теперь прямые методы решения задачи нелинейного программирования (7.3.1), использующие обратную зависимость $u(x)$ вспомогательных двойственных переменных от прямых. Для этого нам потребуется другая модификация функции Лагранжа (7.3.2).

Если взять функцию Лагранжа (7.3.2) и рассмотреть задачу ее максимизации по двойственным переменным, то получаем, что решение этой задачи не обладает свойством непрерывности. Действительно,

$$\max_{u \in R^m} L(x,u) = \begin{cases} f(x), & x \in X, \\ +\infty, & x \notin X. \end{cases} \tag{7.3.24}$$

Чтобы устранить этот недостаток, модифицируем $L(x,u)$, положив

$$M(x,u) = L(x,u) - \frac{\alpha}{2}\|L_x(x,u)\|^2, \tag{7.3.25}$$

где α — фиксированный положительный параметр. Обратим внимание на симметричный вид функции (7.3.25) по отношению к МФЛ (7.3.6), рассмотренной в предыдущем параграфе. Функцию (7.3.25) в отличие от (7.3.6) будем называть *прямой модификацией* функции Лагранжа (7.3.2).

Рассмотрим вспомогательную задачу максимизации

$$\max_{u \in R^m} M(x,u). \tag{7.3.26}$$

Ниже мы покажем, что в отличие от (7.3.24) решение задачи (7.3.26) — точечно-множественное отображение $U(x)$ — существует. Более того, при достаточно естественных предположениях это отображение оказывается однозначным и обладает свойством гладкости. С целью упрощения доказательства приводимых ниже утверждений, как и в предыдущем разделе, везде предполагается, что все ограничения типа неравенства в решении задачи являются активными.

Лемма 7.3.2. *Пусть функции $f(x)$ и $g^i(x)$, $1 \le i \le m$, непрерывно дифференцируемы на $\mathbb{R}^n$. Предположим также, что:*

1) пара $[x_*, u_*]$ является стационарной точкой функции Лагранжа;

2) в $[x_*, u_*]$ выполнено условие (7.3.5);

3) в точке x_* выполнено условие регулярности ограничений;

4) все ограничения типа неравенства в точке x_* являются активными, т.е. $J_0(x_*) = \{l+1, \ldots, m\}$.

Тогда для $\alpha > 0$ и всех x из некоторой окрестности $\Delta(x_*)$ точки x_* существует изолированное локальное решение $u(x)$ задачи максимизации (7.3.26), удовлетворяющее условию $u(x_*) = u_*$. Функция $u(x)$ непрерывно дифференцируема на $\Delta(x_*)$.

Доказательство. Рассмотрим систему уравнений

$$M_u(x,u) = L_u(x,u) - \alpha L_{ux}(x,u) L_x(x,u) = 0_m. \tag{7.3.27}$$

Если $u(x)$ есть решение задачи максимизации (7.3.26), то, в силу необходимых условий экстремума, пара $[x,u]$, где $u = u(x)$, должна удовлетворять этой системе.

Для стационарной точки $[x_*, u_*]$ функции Лагранжа выполнено

$$M_u(x_*, u_*) = L_u(x_*, u_*) - \alpha L_{ux}(x_*, u_*) L_x(x_*, u_*) = 0_m.$$

Вычислим теперь матрицу $M_{uu}(x_*, u_*)$; она имеет вид

$$M_{uu}(x_*, u_*) = L_{uu}(x_*, u_*) - \alpha L_{ux}(x_*, u_*) L_{xu}(x_*, u_*).$$

Воспользуемся предположением, что в точке x_* все ограничения активны. В этом случае $L_{uu}(x_*, u_*)$ является нулевой матрицей, вид матрицы $M_{uu}(x_*, u_*)$ упрощается, а именно:

$$M_{uu}(x_*, u_*) = -\alpha L_{ux}(x_*, u_*) L_{xu}(x_*, u_*).$$

Выясним структуру матрицы $L_{ux}(x,u)$. Непосредственными вычислениями получаем, что $L_{ux}(x,u) = p_u(u) g_x(x)$. Так как матрица

$p_u(u)$ имеет диагональный вид, то

$$L_{ux}(x,u) = \begin{bmatrix} g^1_{x^1}(x) & \dots & g^1_{x^n}(x) \\ & \vdots & \\ g^l_{x^1}(x) & \dots & g^l_{x^n}(x) \\ u^{l+1}g^{l+1}_{x^1}(x) & \dots & u^{l+1}g^{l+1}_{x^n}(x) \\ & \vdots & \\ u^m g^m_{x^1}(x) & \dots & u^m g^m_{x^n}(x) \end{bmatrix}.$$

Из условия регулярности ограничений в точке x_* следует, что градиенты $g^1_x(x_*), \dots, g^m_x(x_*)$ линейно независимы. Кроме того, при сделанных предположениях в паре $[x_*, u_*]$ выполнено условие строгой дополняющей нежесткости. В данном случае это означает, что все множители $u^{l+1}_*, \dots, u^m_*$ отличны от нуля. Отсюда заключаем, что матрица $L_{ux}(x_*, u_*)$ имеет полный ранг, равный m. Произведение матриц $L_{ux}(x_*, u_*)L_{xu}(x_*, u_*)$ есть матрица Грама, составленная для линейно независимых векторов

$$g^1_x(x_*), \dots, g^l_x(x_*),\ u^{l+1}_* g^{l+1}_x(x_*), \dots, u^m_* g^m_x(x_*).$$

Следовательно, данная матрица является положительно определенной, а матрица $M_{uu}(x_*, u_*)$ при $\alpha > 0$ — отрицательно определенной. В силу непрерывности матрица $M_{uu}(x, u)$ будет оставаться отрицательно определенной в некоторой окрестности точки $[x_*, u_*]$.

Таким образом, в точке $u_* = u(x_*)$, а также в других точках $u(x)$ из окрестности $[x_*, u_*]$, выполнены достаточные условия изолированного локального максимума. Из теоремы о неявных функциях следует, что $u(x)$ является непрерывно дифференцируемой функцией. ■

Вдоль решения $u(x)$ задачи максимизации (7.3.26) тождественно по x выполняется равенство

$$L_u(x, u(x)) - \alpha L_{ux}(x, u(x))L_x(x, u(x)) \equiv 0_m.$$

Отсюда видно, что если нами будет найдена такая точка x_*, для которой $L_x(x_*, u(x_*)) = 0_n$, то тем самым будет найдена стационарная точка функции Лагранжа $[x_*, u_*]$, где $u_* = u(x_*)$.

Рассмотрим уравнение

$$L(x, u(x)) = 0_n. \tag{7.3.28}$$

Применим для его решения метод простой итерации:

$$x_{k+1} = x_k - \alpha L_x(x_k, u_k), \quad u_k = u(x_k). \tag{7.3.29}$$

Покажем, что при выполнении достаточных условий второго порядка метод (7.3.29) обладает локальной сходимостью при α положительных, но достаточно малых. Здесь шаг α совпадает со значением соответствующего параметра у МФЛ (7.3.25).

Теорема 7.3.3. *Пусть функции $f(x)$ и $g^1(x), \ldots, g^m(x)$ дважды непрерывно дифференцируемы на $\mathbb{R}^n$. Предположим также, что:*

1) пара $[x_*, u_*]$ является стационарной точкой функции Лагранжа $L(x, u)$;

2) в $[x_*, u_*]$ выполнены условия (7.3.4) и (7.3.5);

3) в точке x_* выполнено условие регулярности ограничений;

4) все ограничения типа неравенства в точке x_* являются активными, т.е. $J_0(x_*) = \{1 + 1, \ldots, m\}$.

Тогда найдется такое $\alpha_* > 0$, что для всех $0 < \alpha < \alpha_*$ процесс (7.3.29) локально сходится к x_*. Соответствующая последовательность $\{u_k\}$ сходится к точке u_*.

Доказательство. Воспользуемся теоремой Островского 7.1.1. Нам надо показать, что отображение

$$Q(x) = x - \alpha L_x(x, u(x)) \tag{7.3.30}$$

дифференцируемо в точке x_* и спектральный радиус ρ матрицы $Q_x(x_*)$ удовлетворяет условию

$$\rho(Q_x(x_*)) < 1. \tag{7.3.31}$$

Дифференцируемость отображения $Q(x)$ в точке x_* следует из непрерывной дифференцируемости функции $u(x)$ в некоторой окрестности точки x_*, установленной в лемме 7.3.2.

Подставляя $u(x)$ в (7.3.27), получаем тождество $M_u(x, u(x)) \equiv 0_m$. Если теперь продифференцировать данное тождество по x, то приходим к уравнению

$$M_{ux}(x, u(x)) + M_{uu}(x, u(x))\frac{du(x)}{dx} = 0_{mn},$$

из которого находим

$$\frac{du(x)}{dx} = -M_{uu}^{-1}(x, u(x))M_{ux}(x, u(x)).$$

При доказательстве леммы 7.3.2 было показано, что

$$M_{uu}(x_*, u_*) = -\alpha L_{ux}(x_*, u_*)L_{xu}(x_*, u_*),$$

где $L_{ux}(x_*, u_*) = p_u(u_*)g_x(x_*)$. Имеем также

$$M_{ux}(x_*, u_*) = L_{ux}(x_*, u_*) - \alpha L_{ux}(x_*, u_*)L_{xx}(x_*, u_*) = \\ = L_{ux}(x_*, u_*)\left[I_n - \alpha L_{xx}(x_*, u_*)\right].$$

Поэтому

$$\frac{du(x_*)}{dx} = \alpha^{-1}\left[L_{ux}(x_*, u_*)L_{xu}(x_*, u_*)\right]^{-1} L_{ux}(x_*, u_*)\left[I_n - \alpha L_{xx}(x_*, u_*)\right]. \tag{7.3.32}$$

Вычислим теперь матрицу $Q_x(x_*)$. Согласно (7.3.30),

$$Q_x(x_*) = I_n - \alpha\left[L_{xx}(x_*, u_*) + L_{xu}(x_*, u_*)\frac{du(x_*)}{dx}\right]. \tag{7.3.33}$$

Подставим (7.3.32) в (7.3.33). Опуская для упрощения записи все аргументы у матриц в правой части, получаем

$$Q_x(x_*) = I_n - \alpha[L_{xx} + \alpha^{-1}L_{xu}(L_{ux}L_{xu})^{-1}L_{ux}(I_n - \alpha L_{xx})]. \tag{7.3.34}$$

Если обозначить через P матрицу

$$P = L_{xu}(L_{ux}L_{ux})^{-1}L_{ux}, \tag{7.3.35}$$

то выражение для $Q_x(x_*)$ можно представить в следующем виде:

$$Q_x(x_*) = (I_n - P)(I_n - \alpha L_{xx}). \tag{7.3.36}$$

Матрица P вида (7.3.35) является матрицей ортогонального проектирования, поскольку она симметрична и идемпотентна ($PP = P$). Подпространство, на которое проектирует матрица P, есть подпространство $\mathcal{L}$, порожденное строками матрицы $L_{ux}(x_*, u_*)$, т.е. градиентами ограничений $g_x^1(x), \ldots, g_x^m(x)$, вычисленными в точке $x = x_*$. Матрица $I_n - P$ также является матрицей ортогонального проектирования, она проектирует на ортогональное дополнение $\mathcal{L}^\perp$ подпространства $\mathcal{L}$.

Оценим собственные значения матрицы $Q_x(x_*)$. Пусть ν — произвольное собственное значение $Q_x(x_*)$, y — соответствующий ему собственный вектор. Согласно определению собственных значений, имеет место равенство $Q_x(x_*)y = \nu y$ или, с учетом (7.3.36),

$$(I_n - P)(I_n - \alpha L_{xx})y = \nu y. \tag{7.3.37}$$

Умножим равенство (7.3.37) поочередно на P и на $(I_n - P)$, получим

$$\nu P y = 0_n, \tag{7.3.38}$$

$$(I_n - P)(I_n - \alpha L_{xx})y = \nu(I_n - P)y. \tag{7.3.39}$$

Возможны два случая:

1. $Py \neq 0_n$. Тогда, согласно (7.3.38), $\nu = 0$.

2. $Py = 0_n$. При этом условии равенство (7.3.39) можно переписать в виде

$$(I_n - P)(I_n - \alpha L_{xx})(I_n - P)y = \nu y, \tag{7.3.40}$$

т.е. ν является одновременно собственным вектором матрицы

$$T = (I_n - P)(I_n - \alpha L_{xx})(I_n - P).$$

Матрица T симметричная и, следовательно, ее собственные значения и собственные векторы вещественны. Но равенство $Py = 0$ для ненулевого собственного вектора y возможно в том и только в том случае, когда

$$L_{ux}(x_*, u_*)y = p_u(u_*)g_x(x_*)y = 0_m,$$

т.е. $y \in K_1(x_*, u_*)$. Поэтому на основании условия (7.3.4)

$$\langle y, L_{xx}(x_*, u_*)y\rangle > 0. \tag{7.3.41}$$

Умножим (7.3.40) слева и справа на $y^\top$. Так как

$$\langle y, Ty\rangle = \|y\|^2 - \alpha\langle y, L_{xx}(x_*, u_*)y\rangle,$$

то приходим к следующему равенству:

$$(1 - \nu)\|y\|^2 = \alpha\langle y, L_{xx}(x_*, u_*)y\rangle. \tag{7.3.42}$$

Отсюда на основании (7.3.41) заключаем, что $\nu < 1$. Имеем также из (7.3.42)

$$\nu = 1 - \frac{\alpha\langle y, L_{xx}(x_*, u_*)y\rangle}{\|y\|^2},$$

или, поскольку $Py = 0_n$,

$$\nu = 1 - \frac{\alpha\langle (I_n - P)y, L_{xx}(x_*, u_*)(I_n - P)y\rangle}{\|y\|^2}.$$

Поэтому, согласно соотношению Релея, справедливо другое неравенство: $\nu > 1 - \alpha\omega_*$, где ω_* — максимальное собственное значение матрицы

$$T_1 = (I_n - P)L_{xx}(x_*, u_*)(I_n - P),$$

причем из неравенства (7.3.41) следует, что ω_* всегда положительно, так как для любого ненулевого вектора $y_1 \in \mathbb{R}^n$ выполняется

включение $(I_n - P)y_1 \in K_1(x_*, u_*)$. Мы получили, что $|\nu| < 1$, если $\alpha < \alpha_* = \frac{2}{\omega_*}$. Таким образом, выполнено (7.3.31). ■

Рассмотрим задачу (7.3.1) с ограничениями только типа равенства, т.е. когда $l = m$, и покажем, что в этом случае метод (7.3.29) сводится к *методу проекции градиента*. Действительно, функция Лагранжа (7.3.2) для такой задачи переходит в обычную классическую функцию Лагранжа (7.2.2). Решение задачи максимизации функции $M(x, u)$ по u сводится к уравнению

$$M_u(x, u) = g(x) - \alpha g_x(x)[f_x(x) + g_x^\top(x)u] = 0_m,$$

решение которого единственно и может быть выписано в явном виде:

$$u(x) = [g_x(x)g_x^\top(x)]^{-1}[\alpha^{-1}g(x) - g_x(x)f_x(x)].$$

Подставляя найденное значение $u(x)$ в (7.3.29), приходим к итерационному процессу:

$$x_{k+1} = x_k - \alpha[(I_n - P(x_k))f_x(x_k) + N(x_k)g(x_k)], \tag{7.3.43}$$

где $P(x) = g_x^\top(x)\left(g_x(x)g_x^\top(x)\right)^{-1} g_x(x)$, $N(x) = g_x^\top(x)(g_x(x)g_x^\top(x))^{-1}$. Если $x_k \in X$, т.е. $g(x_k) = 0_m$, то второе слагаемое в квадратных скобках в правой части (7.3.43) пропадает, и мы получаем

$$x_{k+1} = x_k - \alpha[I_n - P(x_k)]f_x(x_k). \tag{7.3.44}$$

Вектор $[I_n - P(x_k)]f_x(x_k)$ есть проекция градиента функции $f(x)$ на линейное подпространство, касательное к нелинейному многообразию $g(x) = 0_m$ в точке $x = x_k$.

Понятно, что движение вдоль вектора $-[I_n - P(x_k)]f_x(x_k)$ с ненулевым шагом α может в общем случае вывести с данного нелинейного многообразия (допустимого множества X), поэтому вычислительный процесс (7.3.44) в этом случае должен быть дополнен некоторыми процедурами возвращения на допустимое множество.

Отметим также, что если отбросить требование активности всех ограничений типа неравенства в решении задачи, то шаг α в (7.3.29) следует брать меньше, чем соответствующее значение параметра α в МФЛ (7.3.25). При этом величина уменьшения шага будет зависеть от степени близости неактивных ограничений к границе допустимого множества.

Упражнение 9. *Покажите, что при наличии неактивных ограничений типа неравенства в точке $x_* \in X_*$ шаг α в (7.3.29) следует заменить меньшим шагом β из интервала $(0, \bar{\beta}(\alpha))$, где $\alpha > 0$ и*

$\bar{\beta}(\alpha) = 2\min\{\alpha, \frac{1}{\omega_*}\}$. *Здесь величина* ω_* *та же, что и в доказательстве теоремы 7.3.3.*

Применим теперь к решению системы уравнений (7.3.28) метод Ньютона. Тогда приходим к следующему итерационному процессу:

$$x_{k+1} = x_k - G^{-1}(x_k)L_x(x_k, u_k), \tag{7.3.45}$$

где по-прежнему $u_k = u(x_k)$ находится из решения задачи максимизации МФЛ (7.3.26) по u на $\mathbb{R}^m$ при $x = x_k$ и

$$G(x) = \frac{d}{dx}L_x(x, u(x)).$$

Теорема 7.3.4. *Пусть выполнены предположения теоремы 7.3.3 и матрица* $G(x)$ *удовлетворяет в окрестности точки* x_* *условию Липшица. Тогда существует* $\alpha_* > 0$ *такое, что для всех* $0 < \alpha < \alpha_*$ *процесс (7.3.45) локально сходится к* x_**. Соответствующая последовательность* u_k *сходится к* u_**.*

Доказательство. Достаточно показать, что матрица $G(x_*)$ является неособой. Имеем

$$G(x_*) = L_{xx}(x_*, u_*) + L_{xu}(x_*, u_*)\frac{du(x_*)}{dx}.$$

Воспользуемся результатами, полученными при доказательстве леммы 7.3.2 и теоремы 7.3.3, согласно которым

$$\frac{du(x_*)}{dx} = -M_{uu}^{-1}(x_*, u_*)M_{ux}(x_*, u_*),$$

и

$$\begin{aligned} M_{uu}(x_*, u_*) &= -\alpha L_{ux}(x_*, u_*)L_{xu}(x_*, u_*), \\ M_{xu}(x_*, u_*) &= L_{ux}(x_*, u_*)(I_n - \alpha L_{xx}(x_*, u_*)). \end{aligned}$$

Поэтому

$$G(x_*) = (I_n - P)L_{xx}(x_*, u_*) + \alpha^{-1}P,$$

где через P обозначена матрица (7.3.35).

Матрица $G(x_*)$ не является симметричной, однако симметричной будет матрица $G(x_*)+G^\top(x_*)$. В силу условия (7.3.4) для любого ненулевого $y \in K_1(x_*, u_*)$ выполняется $\langle y, L_{xx}(x_*, u_*)y\rangle > 0$. Но включение $y \in K_1(x_*, u_*)$ имеет место в том и только том случае, когда $Py = 0_n$. Поэтому, согласно лемме Финслера 7.2.1, матрица $G(x_*) + G^\top(x_*)$ положительно определена. Отсюда следует, что $G(x_*)$ является неособой матрицей. Таким образом, выполнены условия, гарантирующие локальную сходимость метода (7.3.45). ■

Подчеркнем, что для нахождения матрицы $G(x)$ в методе (7.3.45) достаточно знать только вторые производные функций $f(x)$ и $g(x)$.

При решении задач нелинейного программирования методами, использующими МФЛ вида (7.3.6) или (7.3.25), основной объем вычислений падает на решение внутренних вспомогательных задач — задачи минимизации (7.3.7) в двойственных методах МФЛ или задачи максимизации (7.3.26) в прямых методах МФЛ. Поэтому если размерность n вектора x меньше числа ограничений m, то выгоднее применять двойственные методы. Напротив, в случае, когда n существенно превышает m, более эффективными оказываются прямые методы. При решении задач, в которых числа n и m незначительно отличаются друг от друга, можно использовать как прямые, так и двойственные методы. Однако если вычисление значений функций $f(x)$ и $g(x)$ трудоемко, то некоторое предпочтение можно отдать прямым методам, поскольку они требуют меньшего обращения к вычислению значений функций по сравнению с двойственными.

Глава 8

Методы внутренней точки

К *методам внутренних точек* принято относить те численные алгоритмы решения задач условной оптимизации, в которых все точки соответствующего итерационного процесса принадлежат внутренности допустимого множества. Так, например, рассмотренные нами методы внутренних штрафных функций (барьерных функций) обладают этим свойством и, значит, могут считаться методами внутренних точек. В последнее время понятие метода внутренней точки трактуется в более широком смысле. А именно, если в задаче наряду с равенствами присутствуют ограничения типа неравенства, причем, как правило, простой структуры (базовое допустимое множество), то под методами внутренней точки понимают такие алгоритмы, в которых порождаемые ими траектории не выходят за пределы данного базового множества. Более того, они принадлежат внутренности множества, задаваемого этими неравенствами. Это позволяет, в частности, рассматривать методы внутренних точек для решения задач линейного программирования в канонической форме, в которых в любой текущей точке x_k все компоненты x_k^i, $1 \leq i \leq n$, строго положительны. На самом деле желание отличить методы линейного программирования или более общего выпуклого квадратичного программирования, обладающие этим свойством, от методов симплексного типа и послужило одной из главных причин такого расширенного толкования понятия методов внутренней точки.

Существует много подходов и идей к построению методов внутренних точек. Это приводит к разным вариантам таких методов как для решения задач линейного и квадратичного программирования, так и для более общих задач выпуклого и нелинейного программирования.

Рассмотрим некоторые из них, причем только для линейных задач.

8.1. Мультипликативно-барьерный метод

Пусть имеется задача линейного программирования в канонической форме:

$$\begin{array}{rcl} \langle c, x\rangle & \to & \min, \\ Ax & = & b, \\ x & \geq & 0_n, \end{array} \tag{8.1.1}$$

с двойственной задачей

$$\begin{array}{rcl} \langle b, u\rangle & \to & \max, \\ A^T u & \leq & c. \end{array} \tag{8.1.2}$$

Относительно матрицы A предполагаем, что она размером $m \times n$ и имеет полный ранг, равный m, т.е. строки A линейно независимы. Допустимые множества в задачах (8.1.1) и (8.1.2) будем обозначать соответственно X и U.

Как было показано ранее в [33], необходимые и достаточные условия оптимальности для пары задач (8.1.1) и (8.1.2) сводятся к следующей системе равенств и неравенств:

$$\begin{array}{rcl} D(x)v & = & 0_n, \\ Ax & = & b, \\ c - A^T u & = & v, \\ x \geq 0_n, & & v \geq 0_n. \end{array} \tag{8.1.3}$$

Здесь через $D(x)$ обозначена диагональная матрица с вектором x на диагонали. Первое из равенств $D(x)v = D(v)x = 0_n$ носит название *условия дополнительности* или *условия комплементарности*. Система равенств, входящая в (8.1.3), является «почти линейной». Единственным источником нелинейности является условие комплементарности, которое состоит из n билинейных уравнений.

Чтобы найти решения задач (8.1.1) и (8.1.2), достаточно решить систему (8.1.3). Понятно, что решать эту систему можно разными способами и, по существу, многие известные численные методы решения задачи линейного программирования (8.1.1), включая симплекс-метод, есть не что иное, как разные подходы к решению этой системы. Симплекс-метод, причем как прямой, так и двойственный, характерен тем, что в нем двигаются по таким точкам $[x, u, v]$, в которых условие дополнительности всегда выполняется.

Рассмотрим теперь несколько иной подход к нахождению точек, удовлетворяющих системе (8.1.3), в котором, напротив, условие комплементарности может выполняться только в конце вычислений. Оставим пока в стороне неравенства $x \geq 0_n$, $v \geq 0_n$ и сосредоточим в основном внимание на трех равенствах из (8.1.3):

$$\begin{array}{rcl} D(x)v & = & 0_n, \\ Ax & = & b, \\ c - A^T u & = & v. \end{array} \tag{8.1.4}$$

Воспользуемся выражением для вектора v из третьей строки и подставим его в первое равенство. После умножения получившегося равенства слева на матрицу A приходим к уравнению относительно вектора u:

$$AD(x)A^T u = AD(x)c. \tag{8.1.5}$$

Если матрица $\Gamma(x) = AD(x)A^T$ неособая, то разрешая это уравнение, находим

$$u = u(x) = \left(AD(x)A^T\right)^{-1} AD(x)c.$$

С помощью данного выражения для u можем определить

$$v = v(x) = c - A^T u(x) = \left[I_n - A^T \left(AD(x)A^T\right)^{-1} AD(x)\right] c. \tag{8.1.6}$$

Подставим $v(x)$ в первое равенство в (8.1.4) (в условие комплементарности). В результате приходим к нелинейной системе n уравнений относительно x, а именно:

$$D(x)\left[I_n - A^T \left(AD(x)A^T\right)^{-1} AD(x)\right] c = 0_n. \tag{8.1.7}$$

Для решения получившейся системы могут быть применены различные известные численные методы решения систем нелинейных уравнений. Воспользуемся, например, методом простой итерации. Его применение приводит к следующему итерационному процессу:

$$x_{k+1} = x_k - \alpha_k D(x_k)v_k, \quad v_k = v(x_k), \tag{8.1.8}$$

где $x_0 \in X_0 = \{x \in \mathbb{R}^n : \ Ax = b, \ x > 0_n\}$. Если шаг $\alpha_k > 0$ на каждой итерации брать достаточно малым так, чтобы $x_{k+1} > 0_n$, тогда все точки последовательности $\{x_k\}$ оказываются принадлежащими множеству X_0, т.е. относительной внутренности допустимого множества. В самом деле, обозначая $\Delta x_k = x_{k+1} - x_k$, имеем: $A\Delta x_k = 0_m$. Про итерационный процесс (8.1.8) можно сказать, что он ведет себя как метод внутренней точки.

При выводе метода совсем не учитывалось второе равенство в (8.1.4). Это приводит к тому, что если начальная точка x_0 не удовлетворяет ограничениям типа равенства, т.е. имеется ненулевая невязка $r_0 = Ax_0 - b$, то она сохраняется ненулевой и для всех последующих точек, а именно: $r_k = Ax_k - b = r_0$. Как следствие, все точки x_k также не будут удовлетворять ограничениям типа равенства, причем норма невязки $\|r_k\|$ не уменьшается и остается постоянной в ходе вычислений.

Данный недостаток можно устранить, если при нахождении двойственной переменной $u(x)$ решать не систему линейных уравнений (8.1.5), а, например, возмущенную систему:

$$AD(x)A^T u = AD(x)c + \tau\,(b - Ax)\,, \tag{8.1.9}$$

где τ — некоторый положительный параметр. Фактически здесь к уравнению (8.1.5) прибавлено второе уравнение из (8.1.4), умноженное на коэффициент $\tau > 0$.

Тогда выражение (8.1.6) для $u(x)$ заменится на

$$u(x) = \left(AD(x)A^T\right)^{-1}\left[AD(x)c + \tau\,(b - Ax)\right],$$

а для $v(x)$ имеем

$$\begin{aligned} v(x) = \left[I_n - A^T\left(AD(x)A^T\right)^{-1} AD(x)\right] c - \\ -\tau A^T\left(AD(x)A^T\right)^{-1}(b - Ax)\,. \end{aligned} \tag{8.1.10}$$

Итерационный процесс имеет прежний вид (8.1.8) с той лишь разницей, что $v(x)$ вычисляется согласно формуле (8.1.10) и относительно начальной точки x_0 предполагается только, что $x_0 > 0_n$.

Вектор $\Delta x_k = x_{k+1} - x_k$ распадается в данном случае на сумму двух векторов:

$$\Delta x_k = \Delta x_k^{(1)} + \Delta x_k^{(2)},$$

где

$$\Delta x_k^{(1)} = -\alpha_k D(x_k)\left[I_n - A^T\left(AD(x_k)A^T\right)^{-1} AD(x_k)\right] c,$$

$$\Delta x_k^{(2)} = \tau\alpha_k D(x_k)A^T\left(AD(x_k)A^T\right)^{-1}(b - Ax_k)\,.$$

Первый вектор $\Delta x_k^{(1)}$ по-прежнему принадлежит нуль-пространству матрицы A, т.е. $A\Delta x_k^{(1)} = 0_m$. Для второго вектора $\Delta x_k^{(2)}$ выполняется

$$b - Ax_{k+1} = b - A(x_k + \Delta x_k^{(2)}) = (1 - \tau\alpha_k)\,(b - Ax_k)\,. \tag{8.1.11}$$

Отсюда видно, что можно добиться уменьшения невязки $\|Ax - b\|$ по ограничениями типа равенства на k-й итерации, если взять шаг α_k принадлежащим интервалу $(0, \frac{2}{\tau})$, поскольку согласно (8.1.11)

$$\|Ax_{k+1} - b\| = (1 - \tau\alpha_k)\,\|Ax_k - b\| < \|Ax_k - b\|.$$

При построении метода требовалось, чтобы матрица $\Gamma(x)$ была неособой. Она должна быть неособой как на множестве X, так и на некоторой окрестности X, если зависимость $v(x)$ выбирается согласно (8.1.10). Приведем условие, которое гарантирует выполнение этого требования.

Будем говорить, что в точке $x \in \mathbb{R}^n_+$ выполнено *условие регулярности ограничений*, если строки матрицы A и единичные векторы $e_i \in \mathbb{R}^n$, где $i \in J_0(x) = \{i \in J^n : x^i = 0\}$, линейно независимы.

Утверждение 8.1.1. *Пусть в точке $x \in \mathbb{R}^n_+$ выполнено условие регулярности ограничений. Тогда матрица $\Gamma(x)$ положительно определена.*

Доказательство. Покажем, что ранг матрицы $H(x) = D^{\frac{1}{2}}(x)A^T$ равен m. Тогда из того, что $\Gamma(x) = H^T(x)H(x)$, т.е. является матрицей Грама для столбцов матрицы $H(x)$, будет следовать, что матрица $\Gamma(x)$ положительно определенная. Если x принадлежит внутренности ортанта $\mathbb{R}^n_+$, то это утверждение очевидно, поскольку диагональная матрица $D^{\frac{1}{2}}(x)$ неособая, а матрица A имеет полный ранг, равный m.

Пусть теперь x — граничная точка $\mathbb{R}^n_+$, т.е. множество $J_0(x)$ непустое. Если ранг матрицы $H(x)$ меньше m, то существует такой ненулевой вектор $z \in \mathbb{R}^m$, что

$$H(x)z = D^{\frac{1}{2}}(x)A^T z = 0_n.$$

Данное равенство влечет принадлежность вектора $A^T z$ нуль-пространству матрицы $D^{\frac{1}{2}}(x)$ и, значит, нуль-пространству матрицы $D(x)$. Но вектор $A^T z$ ненулевой, поскольку матрица A имеет полный ранг, равный m. Кроме того, нуль-пространство матрицы $D(x)$ есть подпространство в $\mathbb{R}^n$, порожденное векторами e_i, $i \in J_0(x)$. Отсюда следует, что $A^T z = \sum_{i \in J_0(x)} \beta^i e_i$ для некоторых не равных нулю в совокупности коэффициентов β^i, $i \in J_0(x)$. Выполнение этого равенства означает, что в точке x нарушено условие регулярности ограничений. Мы пришли к противоречию. Таким образом, $H(x)$ имеет полный ранг, равный m. ■

Покажем теперь, что *предположение о невырожденности* задачи линейного программирования (8.1.1), т.е. предположение о невырож-

денности всех угловых точек допустимого множества X, гарантирует выполнение на X условия регулярности ограничений.

Утверждение 8.1.2. *Пусть множество X ограничено и множество X_0 непусто. Тогда невырожденность задачи (8.1.1) эквивалентна условию регулярности ограничений на X.*

Доказательство. Покажем сначала, что выполнение условий регулярности ограничений на X влечет невырожденность задачи (8.1.1). Действительно, во всякой точке $x \in X_0$ активными являются только ограничения типа равенства, поэтому строки матрицы A линейно независимы и, следовательно, ее ранг равен m. Пусть $x \in X \setminus X_0$, тогда часть компонент вектора x равна нулю. В силу условий регулярности ограничений, таких компонент не должно быть больше, чем $d = n - m$. Поэтому каждая точка $x \in X \setminus X_0$ имеет не меньше чем m ненулевых компонент и, значит, всякая угловая точка множества X не вырождена.

Обратно, предполагая невырожденность задачи (8.1.1), покажем, что в любой точке $x \in X$ выполнено условие регулярности ограничений. Пусть x — угловая точка множества X. Не умаляя общности, считаем, что ненулевыми являются первые r компонент вектора x. Тогда имеет место представление

$$x = \begin{bmatrix} x^B \\ x^N \end{bmatrix}, \quad A = [B \mid N], \quad Bx^B = b, \tag{8.1.12}$$

где $x^B > 0_r$, $x^N = 0_{n-r}$, $r \leq m$. Так как точка x не вырождена, то $r = m$ и B есть матрица базиса x. Если рассмотреть квадратную матрицу

$$W = \begin{bmatrix} B & N \\ 0_{dm} & I_d \end{bmatrix}, \tag{8.1.13}$$

то, согласно теореме Фробениуса, матрица W неособая. Тогда ее строки линейно независимы. Отсюда заключаем, что в точке x выполнено условие регулярности ограничений.

Допустим теперь, что x не является угловой точкой X. В этом случае в силу ограниченности X найдутся такие угловые точки x_i и коэффициенты $\alpha_i \geq 0$, $1 \leq i \leq k$, $\sum\limits_{i=1}^{k} \alpha_i = 1$, что $x = \sum\limits_{i=1}^{k} \alpha_i x_i$. Из этого представления следует, что у вектора x больше чем m компонент положительны и матрица B, соответствующая положительным компонентам, содержит невырожденную квадратную подматрицу B_1 порядка m. Поэтому если снова рассмотреть квадратную матрицу W вида (8.1.13),

в которой матрица B заменена на B_1, то она будет неособой и, следовательно, строки матрицы A и единичные векторы, соответствующие нулевым компонентам x, оказываются линейно независимыми. Таким образом, в точке x выполнено условие регулярности ограничений. ■

В утверждении 8.1.2 говорится о выполнении условия регулярности ограничений только на допустимом множестве X. Но в силу непрерывности оно будет выполняться и в некоторой окрестности X. Более того, если A — матрица полного ранга, то условие регулярности имеет место и в любой внутренней точке ортанта $\mathbb{R}^n_+$. Поэтому, согласно утверждению 8.1.1, правые части в (8.1.8) определены во всех этих точках. Точки, в которых нарушается условие регулярности ограничений, могут принадлежать только границе ортанта $\mathbb{R}^n_+$.

Далее считаем, что исходная задача (8.1.1) невырожденная. Тогда любая угловая точка множества X, являющаяся решением задачи (8.1.1), есть стационарная точка для итерационного процесса (8.1.8). Действительно, пусть для определенности базисом B этой точки являются первые m столбцов матрицы A, т.е. имеет место представление (8.1.12). Обозначим через $F(x) = D(x)v(x)$ правую часть в точке x. При этих предположениях вектор $F(x_*)$ также разбивается на две части: на m-мерный вектор $F^B(x_*)$, состоящий из первых m компонент, и на вектор $F^N(x_*)$, содержащий оставшиеся компоненты (последние компоненты $F(x_*)$ в количестве $d = n - m$ штук). Поскольку $x_*^N = 0_d$, то $F^N(x_*) = 0_d$. Проверим, что и $F^B(x_*)$ — нулевой вектор. Имеем с учетом невырожденности матрицы базиса B

$$\begin{aligned} F^B(x_*) &= D(x_*^B)v^B(x_*) = \\ &= D(x_*^B)\left[I_m - B^T\left(BD(x_*^B)B^T\right)^{-1}BD(x_*^B)\right]c^B = \\ &= D(x_*^B)\left[I_m - B^T(B^T)^{-1}D^{-1}(x_*^B)B^{-1}BD(x_*^B)\right]c^B = 0_m. \end{aligned}$$

Таким образом, x_* — стационарная точка для итерационного процесса (8.1.8).

На самом деле не только решение, но и все другие угловые точки X являются стационарными для процесса (8.1.8). В каждой такой точке x выполняется равенство

$$\langle c, x\rangle = \langle b, u(x)\rangle. \tag{8.1.14}$$

Это следует из того, что если B — базис угловой точки x, то

$$\begin{aligned} u(x) &= \left[BD(x^B)B^\top\right]^{-1}BD(x^B)c^B = \\ &= (B^\top)^{-1}D^{-1}(x^B)B^{-1}BD(x^B)c^B = (B^\top)^{-1}c^B. \end{aligned}$$

Отсюда получаем (8.1.14).

Метод (8.1.8) не имеет общепринятого устоявшегося названия. Используемому здесь термину *мультипликативно-барьерный метод* можно дать следующее объяснение. Если точка x_k принадлежит границе ортанта $\mathbb{R}^n_+$, т.е. для нее множество индексов $J_0(x_k)$ не пусто, то у вектора $\Delta x_k = -\alpha_k D(x_k)v_k$ из-за присутствия диагональной матрицы $D(x_k)$ все компоненты $(\Delta x_k)^i$, $i \in J_0(x_k)$, нулевые, т.е. данное направление Δx_k принадлежит грани ортанта $\mathbb{R}^n_+$, содержащей точку x_k. Это не позволяет следующей точке x_{k+1} при надлежащем выборе шага α_k выйти за пределы $\mathbb{R}^n_+$, т.е. появляется эффект «прилипания» к границе ортанта $\mathbb{R}^n_+$. Диагональная матрица $D(x_k)$ играет как-бы роль мультипликативного барьера, ответственного за то, чтобы траектория не покидала ортант $\mathbb{R}^n_+$.

Сходимость метода. Дадим обоснование сходимости метода (8.1.8) к решению задачи (8.1.1) с помощью теоремы Островского 7.1.1. Считаем, что зависимость $v(x)$ выбирается согласно (8.1.10).

Теорема 8.1.1. *Пусть прямая и двойственная задачи (8.1.1) и (8.1.2) не вырождены. Тогда для любого $\tau > 0$ можно указать $\bar{\alpha} > 0$ такое, что какое бы $0 < \alpha < \bar{\alpha}$ ни взять, метод (8.1.8) с постоянным шагом $\alpha_k = \alpha$ локально сходится к x_* с линейной скоростью.*

Доказательство. В силу предположения о невырожденности прямой и двойственной задач точка x_* оказывается единственным решением задачи (8.1.1), причем является угловой точкой допустимого множества X. В нашем случае упомянутое в теореме 7.1.1 отображение $Q(x)$ имеет вид: $Q(x) = x - \alpha R(x)$, где $R(x) = D(x)v(x)$. Поскольку $R(x_*) = 0_n$, то x_* — неподвижная точка отображения $Q(x)$.

Вычислим матрицу Якоби отображения $Q(x)$ в точке x_*. Так как $Q_x(x) = I_n - \alpha R_x(x)$, то нам достаточно найти матрицу Якоби $R_x(x)$ отображения $R(x)$ в этой точке. Имеем

$$R_x(x) = D(v(x)) + D(x)\frac{dv(x)}{dx} = D(v(x)) - D(x)A^\top \frac{du(x)}{dx}. \qquad (8.1.15)$$

Согласно (8.1.9), вектор-функция $u(x)$ удовлетворяет тождеству

$$AD(x)(c - A^\top u(x)) \equiv \tau(Ax - b),$$

дифференцируя которое получаем

$$AD(v(x)) - AD(x)A^\top \frac{du}{dx} = \tau A.$$

Отсюда находим

$$\frac{du}{dx} = (AD(x)A^\top)^{-1}[AD(v(x)) - \tau A]. \tag{8.1.16}$$

Пусть $v_* = v(x_*)$. После подстановки (8.1.16) в (8.1.15) приходим к следующему выражению для матрицы $R_x(x_*)$:

$$R_x(x_*) = (I_n - P(x_*))\, D(v_*) + \tau P(x_*), \tag{8.1.17}$$

где

$$P(x) = D(x)A^\top \left(AD(x)A^\top\right)^{-1} A.$$

Предположим для определенности, что базис точки x_* образуют первые m столбцов матрицы A. Тогда имеет место представление (8.1.12). Вектор v_* также разбивается на две части: $v_*^B \in \mathbb{R}^m$ и $v_*^N \in \mathbb{R}^d$ в соответствии с разбиением x_*. Здесь, как и ранее, $d = n - m$. Из оптимальности x_* и предположения о невырожденности двойственной задачи следует, что $v_*^B = 0$, $v_*^N > 0$.

Так как $x_*^N = 0$, то матрицу $R_x(x_*)$ можно представить в следующем блочном виде:

$$R_x(x_*) = \begin{bmatrix} R_1 & R_3 \\ 0_{dm} & R_2 \end{bmatrix}.$$

Здесь

$$R_1 = \tau D(x_*^B)B^\top \left(BD(x_*^B)B^\top\right)^{-1} B, \qquad R_2 = D(v_*^N).$$

Собственные значения матрицы $R_x(x_*)$ определяются собственными значениями матриц R_1 и R_2. Используя невырожденность матриц B и $D(x_*^B)$, получаем, что $R_1 = \tau I_m$. Матрица R_2 также является диагональной, все ее диагональные элементы, совпадающие с элементами вектора v_*^N, положительны.

Таким образом, собственные значения матрицы $Q_x(x_*)$ оказываются равными либо величине $1 - \alpha\tau$, либо $1 - \alpha v_*^j$, где $m < j \leq n$. Отсюда следует, что спектральный радиус матрицы $Q_x(x_*)$ будет строго меньше единицы, когда

$$0 < \alpha < \bar{\alpha} = \frac{2}{\max\{\tau, \max_{m<j\leq n} v_*^j\}}. \tag{8.1.18}$$

По теореме Островского для любого постоянного шага α, удовлетворяющего неравенству (8.1.18), метод локально сходится к x_* с линейной скоростью. ■

Теорема 8.1.1 гарантирует локальную сходимость метода (8.1.8) к решению задачи линейного программирования (8.1.1), когда слабая двойственная переменная $v(x)$ выбирается согласно (8.1.10). Но если начальная точка x_0 удовлетворяет равенству $Ax = b$, то последовательность точек x_k, генерируемых методом (8.1.8) с правилом выбора $v(x)$ согласно (8.1.10), полностью совпадает с той последовательностью, которая получается, когда применяется правило (8.1.6). Из отмеченного обстоятельства и из теоремы 8.1.1 сразу приходим к следующему результату.

Теорема 8.1.2. *Пусть выполнены условия теоремы 8.1.1. Тогда можно указать $\bar{\alpha} > 0$ такое, что какое бы $0 < \alpha < \bar{\alpha}$ ни взять, метод (8.1.8), (8.1.6) с постоянным шагом $\alpha_k = \alpha$ локально сходится к x_* с линейной скоростью на множестве $X_1 = \{x \in \mathbb{R}^n : \ Ax = b\}$.*

В качестве замечания к теоремам 8.1.1 и 8.1.2 скажем, что локальная сходимость не препятствует возможности выбрать начальную точку x_0 таким образом, чтобы у нее часть компонент были отрицательными величинами. Более того, даже если $x_0 > 0_n$, то в ходе вычислительного процесса у последующих точек x_k некоторые компоненты могут оказаться отрицательными, т.е. сами точки выходят за пределы ортанта $\mathbb{R}^n_+$. Это очень важно с вычислительной точки зрения, так как делает метод устойчивым по отношению к нарушению ограничений типа неравенства, а именно $x \geq 0_n$. Чтобы итерационный процесс (8.1.8) вел себя как метод внутренней точки, следует начальные точки x_0 брать из внутренности ортанта $\mathbb{R}^n_+$ и за счет выбора шагов α_k следить за тем, чтобы точки x_k оставались в этой внутренности.

Метод спуска. Обратимся снова к варианту метода (8.1.8), в котором зависимость $v(x)$ выбирается согласно (8.1.6). Как уже не раз отмечалось, если начальная точка x_0 допустима по ограничениям типа равенства, т.е. $Ax_0 = b$, то и все последующие точки также будут допустимыми по этим ограничениям при любом выборе шаге α_k. Имеем

$$\langle c, x_{k+1} \rangle = \langle c, x_k \rangle - \alpha_k \delta(x_k), \tag{8.1.19}$$

где $\delta(x) = \langle c, D(x)v(x) \rangle$. Так как, согласно (8.1.6),

$$D(x)v(x) = \\ = D^{\frac{1}{2}}(x) \left[I - D^{\frac{1}{2}}(x) A^\top (AD(x)A^\top)^{-1} AD^{\frac{1}{2}}(x) \right] D^{\frac{1}{2}}(x)c,$$

и матрица, стоящая в квадратных скобках, есть матрица ортогональ-

ного проектирования, то

$$\delta(x) = \left\| \left[I - D^{\frac{1}{2}}(x) A^\top (AD(x)A^\top)^{-1} AD^{\frac{1}{2}}(x) \right] D^{\frac{1}{2}}(x) c \right\|^2 = \\ = \|D^{\frac{1}{2}}(x) v(x)\|^2 \geq 0.$$

Поэтому, в силу (8.1.19), минимизируемая функция монотонно убывает в ходе итерационного процесса.

Потребуем дополнительно, чтобы при начальном $x_0 > 0_n$ на всех последующих шагах выполнялось условие $x_k > 0$. Обозначим

$$\mu(x) = \|v_+(x)\|_\infty, \qquad v_+(x) = [c - A^\top u(x)]_+.$$

Из (8.1.8) видно, что если $x_0 > 0_n$ и

$$0 < \alpha_k < \frac{1}{\mu(x_k)}, \qquad k \geq 0, \tag{8.1.20}$$

то и на всех итерациях будем иметь $x_k > 0$. Метод (8.1.8), (8.1.20) в этом случае является релаксационным методом внутренней точки.

Выполнения условий (8.1.20) можно добиться разными способами. Простейший из них заключается в выборе постоянного шага $\alpha_k = \alpha$, где

$$0 < \alpha < \frac{1}{\mu^*}, \qquad \mu^* = \sup_{x \in X_0} \mu(x). \tag{8.1.21}$$

Если множество X ограничено, то величина μ^* обязательно конечна.

Другой способ выбора шага α_k можно получить, если провести нормировку шага, положив

$$\alpha_k = \frac{\beta}{\mu(x_k)}, \tag{8.1.22}$$

где $0 < \beta < 1$. Тогда метод (8.1.8) запишется в виде

$$x_{k+1} = x_k - \beta \mu^{-1}(x_k) D(x_k) v(x_k). \tag{8.1.23}$$

Данный метод обладает не только локальной сходимостью.

Теорема 8.1.3. *Пусть выполнены условия утверждений 8.1.1 и 8.1.2. Пусть, кроме того, функция $c^\top x$ принимает разные значения во всех угловых точках множества X. Тогда для любого $x_0 \in X_0$ процесс (8.1.23) сходится к решению задачи (8.1.1) — точке x_*, соответствующий вектор $u_* = u(x_*)$ является решением двойственной задачи (8.1.2).*

Доказательство. Согласно (8.1.19),

$$\langle c, x_{k+1}\rangle = \langle c, x_0\rangle - \sum_{s=0}^{k} \alpha_s \delta_s, \qquad \delta_s = \|D^{\frac{1}{2}}(x_s)v(x_s)\|^2 > 0.$$

Поскольку монотонно убывающая последовательность $\{\langle c, x_k\rangle\}$ ограничена снизу величиной $\langle c, x_*\rangle$, то $\lim\limits_{k\to\infty} \alpha_k\delta_k = 0$. В силу сделанных предположений $\mu^* < +\infty$, поэтому для любого k имеет место оценка $\alpha_k \geq \frac{\beta}{\mu^*} > 0$, используя которую, получаем

$$\lim_{k\to\infty} \delta_k = 0. \tag{8.1.24}$$

Из ограниченности множества X следует, что последовательность $\{x_k\}$ имеет предельные точки. На основании (8.1.24) в любой предельной точке $\bar{x}$ выполнено условие $D^{\frac{1}{2}}(\bar{x})v(\bar{x}) = 0_n$, т.е. $\bar{x}$ является стационарной точкой процесса (8.1.23) и, следовательно, угловой точкой множества X.

Во всех предельных точках функция $\langle c, x\rangle$ должна принимать одно и то же значение, а поскольку ее значения во всех угловых точках различны, то на самом деле последовательность $\{x_k\}$ сходится к единственной точке $\bar{x}$, которой соответствует двойственный вектор $\bar{u} = u(\bar{x})$. Покажем, что пара $[\bar{x}, \bar{u}]$ является точкой Каруша–Куна–Таккера. Пусть для $\bar{x}$ имеет место представление $\bar{x} = [\bar{x}^B, \bar{x}^N]$, где $\bar{x}^B > 0_m$, $\bar{x}^N = 0_d$, $d = n - m$. Пусть $A = [B \mid N]$ — соответствующее разбиение матрицы A, а c^B и c^N — разбиение вектора c. Тогда

$$\bar{v}^B = c^B - B^\top \bar{u} = 0_m, \quad \bar{v}^N = c^N - N^\top \bar{u}.$$

Достаточно показать, что $\bar{v}^N \geq 0_d$. От противного: предположим, что у вектора $\bar{v}^N$ есть компонента $\bar{v}^i < 0$. Тогда для всех k, превышающих некоторый номер s, выполняется неравенство $v_k^i = c^i - a_i^\top u_k < 0$. Здесь a_i — i-й столбец матрицы A. Так как

$$x_k^i = x_s^i - \sum_{j=s}^{k-1} \alpha_j x_j^i v_j^i, \qquad x_s^i > 0,$$

то, переходя к пределу при $k \to \infty$, получаем $\lim\limits_{k\to\infty} x_k^i = 0$. Отсюда

$$\lim_{k\to\infty} \sum_{j=s}^{k-1} \alpha_j x_j^i v_j^i = x_s^i - \lim_{k\to\infty} x_k^i = x_s^i > 0, \tag{8.1.25}$$

что невозможно в силу отрицательности всех членов ряда в левой части (8.1.25). Таким образом, $\bar{v}^N \geq 0_d$ и, следовательно, $\bar{x}$ и $\bar{u}$ — решения задач (8.1.1) и (8.1.2), т.е. $\bar{x} = x_*$. ■

Отметим, что предположение о разных значениях целевой функции в разных угловых точках не является существенным и используется только с целью упрощения доказательства теоремы. С его помощью устанавливается единственность предельной точки последовательности $\{x_k\}$.

Рассмотренный вариант мультипликативно-барьерного метода можно трактовать как *метод наискорейшего спуска*. Он является релаксационным методом внутренней точки, т.е. все точки x_k, включая начальную точку x_0, принадлежат относительной внутренности допустимого множества X — множеству X_0.

Метод Ньютона. В основе построения итерационного процесса (8.1.8) лежал метод простой итерации, примененный к решению нелинейной системы уравнений $D(x)v(x) = 0_n$. Разумеется, для решения этой системы могут быть использованы и другие известные методы, например метод Ньютона.

Пусть вектор слабых двойственных переменных $v(x)$ определяется согласно (8.1.10) и пусть $R(x) = D(x)v(x)$. Решая систему $R(x) = 0_n$ с помощью метода Ньютона, приходим к следующему итерационному процессу:

$$x_{k+1} = x_k - R_x^{-1}(x_k)R(x_k). \tag{8.1.26}$$

От начальной точки x_0 потребуем, чтобы она находилась вблизи решения задачи (8.1.1) — точки x_*.

Найдем более подробный вид матрицы Якоби $R_x(x)$. Как следует из формул, полученных при доказательстве теоремы 8.1.1, справедливо следующее представление этой матрицы:

$$R_x(x) = (I_n - P(x))\, D(v(x)) + \tau P(x),$$

где

$$P(x) = D(x)A^\top \left(AD(x)A^\top\right)^{-1} A.$$

Таким образом, итерационный процесс (8.1.26) может быть переписан в следующем виде:

$$\begin{aligned} x_{k+1} = x_k - \\ - \left[D(x_k)A^\top \left(AD(x_k)A^\top\right)^{-1} A\left(\tau I_n - D(v_k)\right) + D(v_k)\right]^{-1} D(x_k)v_k. \end{aligned} \tag{8.1.27}$$

Здесь $v_k = v(x_k)$.

Если выполнены предположения теоремы 8.1.1, то матрица $R_x(x_*)$ оказывается неособой. В этом случае метод (8.1.26) локально сходится к x_* со сверхлинейной скоростью. Вместо классического варианта метода Ньютона с единичным шагом можно также рассмотреть и демпфированный вариант метода, в котором вводится переменный шаг. Но тогда начальную точку x_0 следует брать из внутренности неотрицательного ортанта $\mathbb{R}^n_+$ и за счет выбора шага следить за тем, чтобы последующие точки x_k также не покидали этой внутренности. Дополнительно надо добиваться уменьшения невязки по ограничениям типа равенства.

8.2. Аффинно-масштабирующий метод

Решение системы (8.1.3) не изменится, если в первом равенстве вместо матрицы $D(x)$ взять какую-нибудь ее положительную степень, например, квадрат этой матрицы (нуль-пространство матрицы совпадает с нуль-пространством ее квадрата). Тогда появится условие

$$D^2(x)v = 0$$

и можно построить аналоги методов (8.1.8), (8.1.27), в которых везде в правых частях вместо матрицы $D(x)$ присутствует ее квадрат — матрица $D^2(x)$. В частности, метод (8.1.8), (8.1.6) примет вид

$$x_{k+1} = x_k - \alpha_k D^2(x_k)\left[I_n - A^T\left(AD^2(x_k)A^T\right)^{-1}AD^2(x_k)\right]c, \quad (8.2.1)$$

где $x_0 \in X_0$. Итерационный метод (8.2.1) был впервые предложен в 1967 году И.И. Дикиным. Впоследствии в 80-х годах прошлого века он был вновь переоткрыт на Западе и получил название аффинно-масштабирующего метода. В настоящее время, говоря об аффинно-масштабирующих методах, имеют в виду целое семейство методов типа (8.2.1).

К итерационной схеме (8.2.1) можно прийти и из несколько иных соображений. Предположим, что текущая точка $x_k \in X$ такова, что $x_k > 0_n$, т.е. $x_k \in X_0$. Мы хотим найти направление Δx, которое, с одной стороны, не выводило бы за пределы аффинного многообразия $Ax = b$, а с другой, давало бы наибольшее уменьшение значения целевой функции (требование, чтобы у последующей точки x_{k+1} все компоненты были бы неотрицательными пока оставим в стороне).

Объединение двух таких пожеланий дает следующую вспомогательную задачу:

$$\begin{array}{rcl} \langle c, \Delta x \rangle & \to & \min, \\ A\Delta x & = & 0_m, \\ \|\Delta x\|^2 & = & 1, \end{array} \tag{8.2.2}$$

где последнее ограничение введено для нормировки.

Решим задачу (8.2.2), используя принцип Лагранжа. Для этого составим функцию Лагранжа (регулярную):

$$L(\Delta x, u, v) = \langle c, \Delta x \rangle - \langle A\Delta x, u \rangle + v\left(\|\Delta x\|^2 - 1\right),$$

где $u \in \mathbb{R}^m$, $v \in \mathbb{R}$. Дифференцируя ее по Δx и приравнивая производную нулю, получаем

$$L_{\Delta x}(\Delta x, u, v) = c - A^T u + 2v\Delta x = 0_n. \tag{8.2.3}$$

К этому условию следует добавить также условия допустимости:

$$A\Delta x = 0_m, \quad \|\Delta x\|^2 = 1. \tag{8.2.4}$$

Поскольку нас интересует только направление Δx в пространстве $\mathbb{R}^n$, а не его длина, то требование $\|\Delta x\|^2 = 1$ можно отбросить. Важно только, чтобы длина Δx была конечной величиной. Тогда, беря $v = \frac{1}{2}$, получаем из равенства (8.2.3) $\Delta x = -c + A^T u$. Подставляя данный вектор Δx в первое равенство (8.2.4), приходим к уравнению относительно двойственной переменной u:

$$AA^T u = Ac.$$

Если A — матрица полного ранга, то разрешая данное уравнение, находим: $u = (AA^T)^{-1}Ac$. Само направление Δx теперь может быть записано, как $\Delta x = -P(x)c$, где $P(x) = I_n - A^T(AA^T)^{-1}A$ есть матрица ортогонального проектирования на нуль-пространство матрицы A.

Найденное направление Δx является в некотором смысле наилучшим. О нем говорят как об *условном направлении наискорейшего спуска* (условном, поскольку оно удовлетворяет условию $A\Delta x = 0_m$). Однако эффект от такого выбора направления сильно теряется или даже сходит на нет, когда текущая точка x_k находится вблизи границы ортанта $\mathbb{R}^n_+$. Может оказаться, что шаг следует взять очень маленьким, чтобы не покинуть ортант, идя вдоль направления Δx. Это приводит к незначительному перемещению из точки x_k в новую точку x_{k+1} и, как следствие, к малому уменьшению значения целевой функции. Другое дело, если точка x_k находится «равномерно вдали» от границ

ортанта $\mathbb{R}^n_+$. Тогда не запрещается делать большой шаг и добиваться значительного уменьшения значений целевой функции.

Основная идея аффинно-масштабирующего метода заключается в переходе от текущей точки x_k с помощью масштабирования переменных к другой точке, которая в новом пространстве оказывается равномерно удаленной от границ ортанта $\mathbb{R}^n_+$. Делается это путем введения переменных

$$y^i = \frac{x^i}{x^i_k}, \quad 1 \le i \le n. \tag{8.2.5}$$

Тогда старые переменные x^i связаны с новыми переменными y^i линейной зависимостью $x^i = x^i_k y^i$, $1 \le i \le n$ или, в векторном виде, $x = D(x_k)y$. Такое преобразование называют *аффинным масштабированием*. Оно переводит текущую точку x_k в вектор $\bar{e} \in \mathbb{R}^n$, состоящий из единиц.

Теперь в «отмасштабированном пространстве» вспомогательная задача (8.2.2) принимает вид

$$\begin{array}{rcl} \langle c, D(x_k)\Delta y\rangle & \to & \min, \\ AD(x_k)\Delta y & = & 0_m, \\ \|\Delta y\|^2 & = & 1. \end{array}$$

Опять, не обращая внимания на длину вектора Δy, получаем, что

$$\Delta y = -\left[I_n - D(x_k)A^T\left(AD^2(x_k)A^T\right)^{-1}AD(x_k)\right]D(x_k)c,$$

а поскольку согласно (8.2.5) $\Delta x = D(x_k)\Delta y$, то

$$\Delta x_k = -D^2(x_k)\left[I_n - A^T\left(AD^2(x_k)A^T\right)^{-1}AD^2(x_k)\right]c. \tag{8.2.6}$$

Мы пришли к итерационному процессу с тем же самым направлением Δx, что и в итерационном процессе (8.2.1). Шаг α_k в данном случае следует выбирать таким образом, чтобы новая точка x_{k+1} была бы строго внутренней относительно ортанта $\mathbb{R}^n_+$, т.е. $x_{k+1} > 0_n$. Направление (8.2.6) называют *аффинно-масштабированным направлением*. Отсюда и название метода.

Рассмотрим эллипсоид в пространстве $\mathbb{R}^n$:

$$\mathcal{E}(x_k) = \left\{x \in \mathbb{R}^n : \ \|D^{-1}(x_k)x\|^2 \le 1\right\}, \tag{8.2.7}$$

который получил название *эллипсоида Дикина*.

Упражнение 10. *Покажите, что полученное нами направление (8.2.6) есть одновременно решение следующей оптимизационной задачи:*

$$\begin{array}{rcl} \langle c, \Delta x\rangle & \to & \min, \\ A\Delta x & = & 0_m, \\ \Delta x & \in & \mathcal{E}(x_k). \end{array}$$

Аффинно-масштабирующий метод (8.2.1) не обладает локальной сходимостью. По крайней мере, это нельзя доказать с помощью теоремы Островского, так как у соответствующей матрицы появляются нулевые собственные значения. Однако доказано, что этот метод сходящийся, причем не только локально, но стартовые точки обязательно должны принадлежать множеству X_0. Все угловые точки допустимого множества, как и в методе (8.1.8), являются стационарными.

8.3. Метод Кармаркара

Рассмотрим снова аффинно-масштабирующий метод (8.2.1). Один из основных вопросов при его применении (как, впрочем, и при применении других методов внутренней точки) — это вопрос о выборе шага α_k. С одной стороны, желательно шаг взять как можно больше, чтобы добиться наибольшего уменьшения значения целевой функции. С другой стороны, близкий подход к границе допустимого множества усиливает эффект «прилипания». Поэтому стараются при выборе шага сбалансировать эти две противоположные тенденции. В 1984 году Н. Кармаркар предложил метод, в котором проблема выбора шага решается довольно своеобразно. Приведем описание данного метода, базируясь на рекуррентной формуле (8.2.1).

Метод Кармаркара предназначен для решения задачи линейного программирования специального вида:

$$\begin{array}{c} \min \langle c, x\rangle, \\ Ax = 0_m, \\ x \in \Lambda^n, \end{array} \tag{8.3.1}$$

где $\Lambda^n = \left\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^i = 1\right\}$ — вероятностный симплекс в $\mathbb{R}^n$. Пусть e — n-мерный вектор, все компоненты которого равны единице. Предполагается, что $(m \times n)$-матрица A такова, что точка

$$x_0 = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]^\top = n^{-1}e \tag{8.3.2}$$

удовлетворяет однородным ограничениям $Ax = 0_m$, т.е. $Ax_0 = 0_m$. Кроме того, считается, что оптимальное значение целевой функции в (8.3.1) равно нулю. Несмотря на достаточно специфическую форму задачи (8.3.1), оказывается, что задачу линейного программирования (8.1.1) можно рядом последовательных преобразований свести к задаче вида (8.3.1).

В методе Кармаркара в качестве начальной берется точка (8.3.2). Все последующие точки x_k являются допустимыми и, следовательно, принадлежат симплексу Λ^n. Более того, эти точки принадлежат *относительной внутренности* $\Lambda_0^n = \{x \in \Lambda^n : \ x > 0_n\}$ симплекса Λ^n. В силу однородности системы ограничений $Ax = 0_m$ гарантированный шаг, не выводящий за пределы допустимого множества, — это гарантированный шаг, не выводящий за пределы симплекса. Последний зависит от того, насколько текущая точка x_k близка к границе симплекса.

Минимальное расстояние $\rho(x_k)$ от x_k до точек границы симплекса оказывается максимально возможным, когда точка x_k совпадает с центром симплекса Λ^n — точкой x_0. Данное расстояние равно

$$\rho(x_0) = \|\bar{x}_i - x_0\|, \tag{8.3.3}$$

где $\bar{x}_i$ — точка на грани $\Lambda_i^n = \{x \in \Lambda^n : x^i = 0\}$ симплекса Λ^n, ближайшая к x_0. В силу симметрии можно взять любую грань Λ_i^n, $1 \le i \le n$, при этом расстояние от $\bar{x}_i$ до x_0 будет одним и тем же. Как несложно подсчитать, решая задачу минимизации:

$$\begin{array}{c} \min \dfrac{1}{2}\|x - x_0\|^2, \\ x \in \Lambda^n, \\ x^i = 0, \end{array} \tag{8.3.4}$$

у точки $\bar{x}_i$ все компоненты равны $\frac{1}{n-1}$, за исключением i-й компоненты, равной нулю. Подставляя данную точку в (8.3.3), получаем

$$\rho(x_0) = \frac{1}{\sqrt{n(n-1)}}. \tag{8.3.5}$$

Величина $\rho(x_0)$ определяет радиус наибольшего $(n-1)$-мерного шара с центром в точке x_0, который может быть вписан в симплекс Λ^n.

Упражнение 11. *Решите оптимизационную задачу (8.3.4) и убедитесь, что расстояние $\rho(x_0)$ равно величине (8.3.5).*

Если точка x не совпадает с x_0, то расстояние $\rho(x)$ убывает, стремясь к нулю, когда x приближается к границе симплекса. Это уменьшает возможный гарантированный шаг в методах типа (8.2.1) с нормированными правыми частями, который не выводил бы за пределы симплекса. Чтобы исправить данную ситуацию, можно «повернуть» симплекс вокруг текущей точки.

Пусть $x_k \in \Lambda_0^n$. Наряду с симплексом Λ^n рассмотрим «повернутый» симплекс

$$\bar{\Lambda}^n(x_k) = \{x \in \mathbb{R}_+^n : \langle D^{-1}(x_k)e, x\rangle = n\}. \tag{8.3.6}$$

Нетрудно видеть, что всегда $x_k \in \bar{\Lambda}^n(x_k)$. Отметим также, что $\bar{\Lambda}^n(x_0)$ совпадает с основным симплексом Λ^n, т.е. для точки (8.3.2) поворота симплекса не происходит.

Дадим теперь описание основной итерации метода Кармаркара. Допустим, что в результате работы алгоритма получена точка x_k, которая принадлежит Λ_0^n. Вычисления, проводимые на k-й итерации, разбиваются на два этапа.

Этап 1. Делается поворот основного симплекса Λ^n, т.е. вместо Λ^n рассматривается симплекс $\bar{\Lambda}^n(x_k)$, определяемый согласно (8.3.6). Система ограничений в задаче (8.3.1) заменяется на следующую:

$$\begin{gathered} Ax = 0_m, \\ \langle D^{-1}(x_k)e, x\rangle = n. \end{gathered}$$

Если объединить последнее линейное ограничение с первыми m однородными ограничениями, то их можно записать в виде

$$\bar{A}x = \bar{b}, \tag{8.3.7}$$

где $\bar{A}$ — расширенная матрица A размером $(m+1)\times n$, $\bar{b}$ — расширенный $(m+1)$-мерный вектор,

$$\bar{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \\ \dfrac{1}{x_k^1} & \dfrac{1}{x_k^2} & \dots & \dfrac{1}{x_k^n} \end{bmatrix}, \qquad \bar{b} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ n \end{bmatrix}.$$

Отметим, что $\bar{A}x_k = \bar{b}$, т.е. точка x_k удовлетворяет системе линейных равенств (8.3.7).

Используя формулу пересчета аффинно-масштабирующего метода (8.2.1), вычисляем новую точку:

$$\bar{x}_{k+1} = x_k - \alpha_k D^2(x_k)\left[I_n - \bar{A}^\top\left(\bar{A}D^2(x_k)\bar{A}^\top\right)^{-1}\bar{A}D^2(x_k)\right]c. \tag{8.3.8}$$

Здесь шаг α_k полагается равным

$$\alpha_k = \beta\sqrt{\frac{n-1}{n}}\|F(x_k)\|^{-1}, \tag{8.3.9}$$

где $0 < \beta < 1$ и

$$F(x) = D(x)\left[I_n - \bar{A}^\top\left(\bar{A}D^2(x)\bar{A}^\top\right)^{-1}\bar{A}D^2(x)\right]c.$$

В силу основных свойств метода (8.2.1), $\bar{A}\bar{x}_{k+1} = \bar{b}$. Более того, как будет показано ниже, $\bar{x}_{k+1}$ принадлежит относительной внутренности $\bar{\Lambda}^n_0(x_k) = \{x \in \bar{\Lambda}^n(x_k) : x > 0_n\}$ «повернутого» симплекса $\bar{\Lambda}^n(x_k)$.

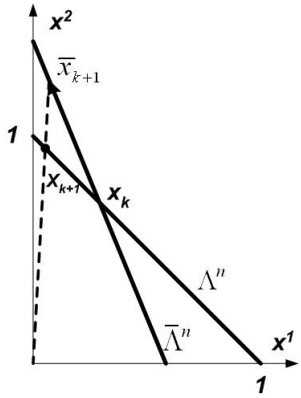

Рис. 8.1. Коническое проектирование

Этап 2. Проводим "коническое проектирование" точки $\bar{x}_{k+1}$ на исходный симплекс Λ^n и получаем точку $x_{k+1} \in \Lambda^n_0$ (см. рис. 8.1). Другими словами, полагаем

$$x_{k+1} = \frac{1}{\langle \bar{x}_{k+1}, e\rangle}\bar{x}_{k+1}. \tag{8.3.10}$$

На этом итерация заканчивается.

Поясним выбор шага (8.3.9). Обозначим

$$\bar{\alpha}_k = \|F(x_k)\|\alpha_k = \beta\sqrt{\frac{n-1}{n}}, \qquad \bar{F}(x_k) = \frac{F(x_k)}{\|F(x_k)\|}. \tag{8.3.11}$$

Вектор $\bar{F}(x_k)$ имеет единичную длину. С его помощью формулу (8.3.8) для нахождения промежуточного вектора $\bar{x}_{k+1}$ можно записать в виде

$$\bar{x}_{k+1} = D(x_k)\left[e - \bar{\alpha}_k \bar{F}(x_k)\right]. \tag{8.3.12}$$

Так как вектор $D(x_k)\bar{F}(x_k)$ принадлежит нуль-пространству матрицы $\bar{A}$, а последняя строка матрицы $\bar{A}$ равна $D^{-1}(x_k)e$, то

$$\langle D^{-1}(x_k)e, D(x_k)\bar{F}(x_k)\rangle = \langle e, \bar{F}(x_k)\rangle = 0.$$

Мы получили, что сумма всех компонент вектора $\bar{F}(x_k)$ равна нулю.

Имеет место следующий результат.

Утверждение 8.3.1. *Пусть вектор $c \in \mathbb{R}^n$ удовлетворяет равенствам: $\|c\| = 1$ и $\langle e, c\rangle = 0$. Тогда максимальная компонента вектора c не превосходит значение $\sqrt{\frac{n-1}{n}}$.*

Доказательство. Не умаляя общности, считаем, что максимальной является компонента c^n. Рассмотрим задачу максимизации:

$$\begin{gathered} \max x^n, \\ \sum_{i=1}^{n} x^i = 0, \quad \sum_{i=1}^{n} (x^i)^2 = 1. \end{gathered} \tag{8.3.13}$$

Составим для нее функцию Лагранжа:

$$L(x, u, v) = x^n + u\sum_{i=1}^{n} x^i + v\left(1 - \sum_{i=1}^{n}(x^i)^2\right), \qquad u, v \in \mathbb{R}.$$

Из условий оптимальности следует, что в решении задачи (8.3.13) должны выполняться равенства

$$\begin{aligned} &L_{x^i} = u - 2vx^i = 0, && 1 \le i < n, \\ &L_{x^n} = 1 + u - 2vx^n = 0. \end{aligned}$$

Считаем, что $v \neq 0$. Тогда

$$x^n = \frac{1+u}{2v}, \qquad x^i = \frac{u}{2v}, \quad 1 \le i < n. \tag{8.3.14}$$

После подстановки найденных значений компонент вектора x в первое ограничение задачи (8.3.13) получаем

$$\frac{(n-1)u}{2v} + \frac{1+u}{2v} = \frac{1}{2v}(1+nu) = 0,$$

откуда $u = -\frac{1}{n}$. Тогда на основании второго ограничения в (8.3.13):

$$\frac{(n-1)^2}{4n^2v^2} + \frac{n-1}{4n^2v^2} = 1.$$

Отсюда находим: $v = 0.5\sqrt{\frac{n-1}{n}}$. Таким образом, согласно (8.3.14), $x^n = \sqrt{\frac{n-1}{n}}$. ■

Используя утверждение 8.3.1, на основании (8.3.11) и (8.3.12) приходим к выводу, что $\bar{x}_{k+1} > 0_n$, следовательно, и $x_{k+1} > 0_n$, т.е. $x_{k+1} \in \Lambda_0^n(x_k)$. Отметим также, что, в силу однородности системы ограничений типа равенства $Ax = 0_m$, «коническое проектирование» на симплекс Λ^n на втором этапе не приводит к их нарушению.

Существует много способов перехода от задачи в канонической форме (8.1.1) к задаче вида (8.3.1). Приведем один из них.

Предположим, что имеется оценка сверху всех компонент вектора x, например, их сумма не превышают величину σ. Тогда, добавив дополнительную неотрицательную переменную x^{n+1}, приходим к равенству $\sum\limits_{i=1}^{n} x^i + x^{n+1} = \sigma$, которое после деления на σ запишется в виде

$$\sum_{i=1}^{n} \bar{x}^i + \bar{x}^{n+1} = 1. \tag{8.3.15}$$

Здесь $\bar{x}^i = \frac{x^i}{\sigma}$. Разделим также на σ все уравнения, входящие в систему линейных ограничений $Ax = b$; получим

$$A\bar{x} - \sigma^{-1}b = 0_m. \tag{8.3.16}$$

Введем еще одну переменную $\bar{x}^{n+2}$, связав ее со столбцом $\sigma^{-1}b$, который обозначим для сокращения записи $\bar{b}$. Если положить переменную $\bar{x}^{n+2}$ равной единице, то система равенств (8.3.15), (8.3.16) запишется как

$$\begin{array}{rcl} A\bar{x} - \bar{b}\bar{x}^{n+2} & = & 0_m, \\ \bar{x}^{n+2} & = & 1, \\ \sum\limits_{i=1}^{n} \bar{x}^i + \bar{x}^{n+1} & = & 1. \end{array} \tag{8.3.17}$$

Вычитая из третьего равенства второе, а также складывая оба эти равенства, получаем

$$\begin{array}{rcl} \sum\limits_{i=1}^{n} \bar{x}^i + \bar{x}^{n+1} - \bar{x}^{n+2} & = & 0, \\ \sum\limits_{i=1}^{n} \bar{x}^i + \bar{x}^{n+1} + \bar{x}^{n+2} & = & 2. \end{array} \tag{8.3.18}$$

Перейдем теперь к переменным $\tilde{x}^i = \frac{\bar{x}^i}{2}$. Тогда, согласно (8.3.18), $\sum\limits_{i=1}^{n+2} \tilde{x}^i = 1$. Кроме того, из (8.3.17) и (8.3.18) получаем, что $\tilde{A}\tilde{x} = 0$, где $\tilde{x} = [\tilde{x}^1, \ldots, \tilde{x}^{n+2}]^\top$ и $\tilde{A}$ — матрица размером $(m+1) \times (n+2)$ следующего вида:

$$\tilde{A} = \begin{bmatrix} A & 0 & -\bar{b} \\ e^\top & 1 & -1 \end{bmatrix}.$$

Чтобы удовлетворить условию $\tilde{A}x_0 = 0$, где x_0 — начальная точка вида (8.3.2), можно добавить еще один столбец в матрицу $\tilde{A}$ и еще одну неотрицательную переменную $\tilde{x}^{n+3}$, которую будем стремить к нулю. Тогда приходим к системе ограничений:

$$\begin{array}{rcl} \tilde{A}\tilde{x} + \tilde{a}_{n+3}\tilde{x}^{n+3} & = & 0_{m+1}, \\ \sum\limits_{i=1}^{n+2} \tilde{x}^i + \tilde{x}^{n+3} & = & 1. \end{array}$$

Если положить

$$\tilde{a}_{n+3} = \begin{bmatrix} \bar{b} - Ae \\ -n \end{bmatrix},$$

то точка $\tilde{x}_0 \in \mathbb{R}^{n+3}_+$ со всеми компонентами $\tilde{x}^i_0 = \frac{1}{n+3}$, $1 \le i \le n+3$, может быть взята в качестве начальной.

Метод Кармаркара в отличие от симплекс-метода является *полиномиальным*. Это означает, что для решения задачи (8.3.1) ему требуется количество операций, которое оценивается сверху полиномом от числа битов, используемых для представления всей информации о задаче в компьютере.

8.4. Прямодвойственные методы центрального пути

Обратимся снова к задаче линейного программирования (8.1.1). В рассмотренных методах внутренней точки главное внимание уделялось условию комплементарности $D(x)v = D(v)x = 0_n$. В ходе итерационного процесса это условие на каждом шаге нарушалось, однако методы строились с таким расчетом, чтобы оно выполнялось в пределе. В методах *центрального пути* стараются добиться более регулярного по сравнению с другими методами релаксации условия комплементарности путем его замены на следующее условие:

$$D(x)v = D(v)x = \mu e, \tag{8.4.1}$$

где μ — положительный параметр, e — n-мерный вектор, состоящий из единиц. Данное условие назовем *возмущенным условием комплементарности*. Сами условия (8.1.3) теперь примут вид

$$\begin{array}{rcl} D(x)v & = & \mu e, \\ Ax & = & b, \\ c - A^T u & = & v, \\ x \geq 0_n, & & v \geq 0_n. \end{array} \tag{8.4.2}$$

При $\mu = 0$ они переходят в условия (8.1.3).

Пусть тройка переменных $[x, u, v] \in \mathbb{R}^{2n+m}$, входящих в прямую и двойственную задачи, допустима, т.е. $x \in X$, $u \in U$ и $v = c - A^T u$. Если, кроме того, $x > 0_n$, $v > 0_n$ и для $[x, v]$ выполнено возмущенное условие комплементарности (8.4.1), то про такую тройку переменных $[x, u, v]$ говорят, что она *μ-центральная*. Совокупность всех μ-центральных точек при различных $\mu > 0$ образует *центральный путь*, точнее, *прямо-двойственный центральный путь*. В μ-центральных точках «зазор двойственности», т.е. разность между значениями целевых функций в прямой и двойственной задаче, равняется величине $n\mu$. В самом деле,

$$\langle c, x\rangle - \langle b, u\rangle = \langle c, x\rangle - \langle Ax, u\rangle = \langle c - A^T u, x\rangle = \langle v, x\rangle = n\mu.$$

Основная идея заключается в том, чтобы брать начальную пару $[x_0, v_0]$, лежащую на центральном пути, и далее переходить к новым точкам $[x_k, v_k]$, которые либо лежат на центральном пути, либо находятся вблизи него, но при меньшем значении параметра μ. Если относительное уменьшение параметра μ от итерации к итерации оказывается постоянным, то такой метод становится *полиномиальным* и обладает сходимостью. В настоящее время общепризнанно, что прямо-двойственные методы центрального пути являются наиболее эффективными алгоритмами решения задач линейного программирования, особенно большой размерности.

Перейдем к более подробному рассмотрению метода центрального пути, точнее, к одному из вариантов целого семейства таких методов. В англоязычной литературе их называют *path-following methods*. Но сначала обратим внимание на связь, которая возникает между условиями оптимальности (8.4.2) и решением вспомогательной задачи в методе внутренних штрафных функций с логарифмической барьерной функцией. Предположим, что данная барьерная функция применяется для освобождения от ограничений типа неравенства $x^i \geq 0$, где $1 \leq i \leq n$. Обозначая через μ штрафной коэффициент, приходим к

следующей вспомогательной задаче для определения точки минимума внутренней штрафной функции при заданном $\mu > 0$: найти

$$\min_{x \in X_0} \langle c, x \rangle - \mu \sum_{i=1}^{n} \ln x^i, \quad X_0 = \{x \in \mathbb{R}^n : \ Ax = b, \ x > 0_n\}.$$

Если составить для этой задачи функцию Лагранжа:

$$L(x, u) = \langle c, x \rangle - \mu \sum_{i=1}^{n} \ln x^i + \langle b - Ax, u \rangle,$$

где $u \in \mathbb{R}^m$, то условия оптимальности первого порядка запишутся как

$$L_x(x, u) = c - A^T u - \mu D^{-1}(x) e = 0_n, \quad Ax = b.$$

Вводя переменную $v = c - A^T u$, первое равенство можно представить как $D(x)v = \mu e$. При этом обязательно $v > 0_n$. Тогда, ослабляя требования к x и v (переходя от строгих неравенств $x > 0_n$ и $v > 0_n$ к нестрогим $x \geq 0_n$, $v \geq 0_n$), получаем условия оптимальности, вид которых полностью совпадают с условиями (8.4.2).

Для упрощения выкладок преобразуем систему (8.4.2), освободившись от переменной u. С этой целью введем матрицу K размером $d \times n$, где $d = n - m$. Строки K образуют некоторый базис в нульпространстве матрицы A. Если, например, для матрицы A справедливо разбиение $A = [B \ N]$, где B — неособая матрица порядка m, то в качестве матрицы K можно взять $K = [N^T (B^T)^{-1} | - I_d]$. Тогда после умножения левой и правой частей третьего равенства на матрицу K приходим вместо (8.4.2) к системе

$$\begin{array}{rcl} D(x)v = D(v)x & = & \mu e, \\ Ax & = & b, \\ Kv & = & \bar{c}, \\ x \geq 0_n, & & v \geq 0_n, \end{array} \tag{8.4.3}$$

где $\bar{c} = Kc$. Теперь *μ-центральными* будут пары $[x, v] \in \mathbb{R}^{2n}$ с компонентами $x > 0_n$ и $v > 0_n$, удовлетворяющими (8.4.3). *Прямо-двойственным решением* задачи (8.1.1) и двойственной к ней будем называть точку $[x_*, v_*]$, которая удовлетворяет (8.4.3) при $\mu = 0$. Данная точка является предельной для μ-центральных точек при $\mu \downarrow 0$.

Наше стремление брать точки, принадлежащие центральному пути или близкие к нему, реализуем с помощью метода Ньютона, который применим для решения системы равенств, входящих в (8.4.3). Однако решение проводим не до конца, а делая на каждой итерации при

фиксированном μ только один шаг методом Ньютона. Поскольку одновременно желательно уменьшать значение параметра μ, мы будем осуществлять такое изменение μ путем умножения текущего значения μ на некоторый понижающий коэффициент $0 < \sigma < 1$, т.е. фактически на итерации делается шаг методом Ньютона при значении параметра μ в (8.4.3), равном $\sigma\mu$. Отметим также, что для простоты строится *допустимый прямо-двойственный метод*. Это означает, что все точки x и v, задействованные в итерационном процессе, допустимы, т.е. они удовлетворяют второму и третьему равенствам из (8.4.3).

Пусть у нас имеется допустимая пара точек $x > 0_n$ и $v > 0_n$. Ньютоновские направления Δx и Δv находим, линеаризуя в этой паре равенства из (8.4.3) и приравнивая их нулю, что приводит к следующей системе линейных алгебраических уравнений:

$$\begin{bmatrix} D(v) & D(x) \\ A & 0_{mn} \\ 0_{dn} & K \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta v \end{bmatrix} = - \begin{bmatrix} r_c(x,v) \\ 0_m \\ 0_d \end{bmatrix},$$

где $r_c(x,v) = D(x)v - \sigma\mu e$ — невязка относительно условия комплементарности. В покомпонентном виде эта система уравнений запишется как

$$\begin{aligned} D(v)\Delta x + D(x)\Delta v &= \sigma\mu e - D(x)v, \\ A\Delta x &= 0_m, \\ K\Delta v &= 0_d. \end{aligned} \tag{8.4.4}$$

Отсюда видно, что направления Δx и Δv принадлежат нуль-пространствам матриц A и K соответственно. Поэтому $\Delta x = K^T z_1$, $\Delta v = A^T z_2$ для некоторых векторов $z_1 \in \mathbb{R}^d$, $z_2 \in \mathbb{R}^m$. Следовательно,

$$\langle \Delta x, \Delta v \rangle = \langle K^T z_1, A^T z_2 \rangle = \langle AK^T z_1, z_2 \rangle = 0. \tag{8.4.5}$$

Кроме того, беря в (8.4.4) параметр μ, равным $\mu(x,v) = \frac{\langle x,v \rangle}{n}$, приходим к равенству

$$\langle D(v)\Delta x + D(x)\Delta v, e \rangle = \langle \sigma\mu e - D(x)v, e \rangle = (\sigma - 1)\langle x, v \rangle. \tag{8.4.6}$$

Здесь учтено, что $\langle e, e \rangle = n$. Величину $\mu(x,v)$ принято называть *двойственной мерой*.

Движение вдоль ньютоновских направлений Δx и Δv будем осуществлять с некоторым шагом $\alpha > 0$, который, в принципе, может оказаться и меньше единицы, например, чтобы не нарушать условия неотрицательности переменных x и v. Поэтому если ввести в рассмотрение функцию $g(x,v) = n\mu(x,v) = \langle x, v \rangle$, значения которой на объединении допустимых множеств прямой и двойственных задач (его так

и называют *прямо-двойственным допустимым множеством*)

$$\mathcal{F}_{PD} = \left\{ [x, v] \in \mathbb{R}^{2n}_{+} : \; Ax = b, \; Kv = \bar{c} \right\},$$

являются «двойственным зазором», то при переходе из точки $[x, v]$ с шагом $\alpha > 0$ в новую точку $[\bar{x}, \bar{v}] = [x + \alpha\Delta x, v + \alpha\Delta v]$ согласно (8.4.5) и (8.4.6) выполняется равенство

$$\begin{aligned} g(\bar{x}, \bar{v}) &= g(x, v) + \alpha \left[\langle v, \Delta x \rangle + \langle x, \Delta v \rangle \right] + \alpha^2 \langle \Delta x, \Delta v \rangle = \\ &= [1 - \alpha(1 - \sigma)] \; g(x, v). \end{aligned} \tag{8.4.7}$$

Отсюда немедленно следует, что значение двойственной меры в новой паре $\bar{\mu} = \mu(\bar{x}, \bar{v})$ связано со значением в старой паре $\mu = \mu(x, v)$ соотношением

$$\bar{\mu} = [1 - \alpha(1 - \sigma)]\mu. \tag{8.4.8}$$

Функция $g(x, v)$ неотрицательна всюду на множестве $\mathcal{F}_{PD}$, причем равна нулю в том и только в том случае, когда выполнено условие комплементарности $D(x)v = 0_n$, т.е. когда x и v являются решением прямой и двойственных задач (8.1.1) и (8.1.2) (оптимальному v соответствует оптимальное u такое, что $v = c - A^T u \geq 0_n$). Формулы (8.4.7), (8.4.8) показывают, что происходит уменьшение двойственной меры и связанного с ней «двойственного зазора», в некотором смысле мы оказываемся ближе к решению.

Как уже отмечалось, наша цель состоит в том, чтобы брать точки вблизи центрального пути и стараться так подобрать параметры метода, чтобы новые точки не слишком отходили от этого пути. Прежде всего уточним понятие окрестности центрального пути, положив

$$\mathcal{N}(\theta) = \{ [x, v] \in \mathcal{F}_{PD} : \; \| D(x)D(v)e - \mu e \| \leq \theta\mu \},$$

где $0 < \theta < 1$ — некоторый параметр, а в качестве нормы берется евклидова норма.

Все μ-центральные точки входят, разумеется, в эту окрестность. Но и другие точки $[x, v] \in \mathcal{F}_{PD}$, для которых разброс значений произведений $x^i v^i$ не очень велик, также принадлежат этой окрестности. Из неравенства между нормами $\|z\|_\infty \leq \|z\|_2$, имеющего место для любого вектора z, вытекает, что для каждой точки $[x, v] \in \mathcal{N}(\theta)$ выполняются неравенства

$$|x^i v^i - \mu| \leq \| D(x)D(v)e - \mu e \| \leq \theta\mu, \quad 1 \leq i \leq n.$$

Следовательно, справедливы следующие оценки на произведения $x^i v^i$, $1 \leq i \leq n$, снизу и сверху:

$$(1 - \theta)\,\mu \leq x^i v^i \leq (1 + \theta)\,\mu, \quad 1 \leq i \leq n, \tag{8.4.9}$$

которые должны выполняться для всех точек $[x, v]$ из $\mathcal{F}_{PD}$ для некоторого μ, в частности для $\mu = \mu(x, v)$.

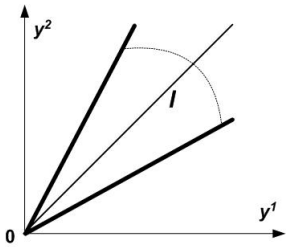

Рис. 8.2. Окрестность $\mathcal{N}(\theta)$

Окрестность $\mathcal{N}(\theta)$ для случая $n = 2$ показана на рис. 8.2. По координатным осям здесь указываются величины $y^1 = x^1 v^1$ и $y^2 = x^2 v^2$. Всем μ-центральным точкам $[x, v] \in \mathbb{R}^4$ на рисунке соответствуют точки $y \in \mathbb{R}^2$, лежащие на луче l, который делит неотрицательный ортант $\mathbb{R}^2_+$ пополам. Это фактически изображение центрального пути в пространстве переменных y. Окрестность $\mathcal{N}(\theta)$, опять же в этом пространстве переменных, является выпуклым конусом с центральным лучом l. Начало конуса совпадает с началом координат. Насколько конус «узок», определяется параметром θ. Конус расширяется при удалении от своего начала , что соответствует увеличению μ.

Если $\mu = \mu(x, v)$, то имеем

$$\langle (D(x)v - \mu e), e \rangle = \langle x, v \rangle - \mu n = 0. \tag{8.4.10}$$

Поэтому в этом случае

$$\begin{aligned} \|D(x)v - \sigma\mu e\|^2 &= \|D(x)v - \mu e + (1-\sigma)\mu e\|^2 = \\ &= \|D(x)v - \mu e\|^2 + n(1-\sigma)^2\mu^2 \leq \\ &\leq \theta^2\mu^2 + n(1-\sigma)^2\mu^2 = \left[\theta^2 + n(1-\sigma)^2\right]\mu^2, \end{aligned} \tag{8.4.11}$$

когда $[x, v] \in \mathcal{N}(\theta)$.

Далее предполагаем, что параметр σ зависит от размерности n векторов x и v, а также от параметра θ следующим образом:

$$\sigma = 1 - \frac{\theta}{\sqrt{n}}. \tag{8.4.12}$$

Понятно, что в этом случае σ оказывается близким к единице.

Лемма 8.4.1. *Пусть* $[x, v] \in \mathcal{N}(\theta)$ *и* $0 < \alpha \leq 1$. *Пусть, кроме того, выполняется (8.4.12). Тогда можно указать достаточно малое* $\theta_* > 0$ *такое, что для любого* $0 < \theta < \theta_*$ *для новых точек* $\bar{x} = x + \alpha\Delta x$, $\bar{v} = v + \alpha\Delta v$ *и новой величины* $\bar{\mu} = \mu(\bar{x}, \bar{v})$ *имеет место неравенство*

$$\|D(\bar{x})\bar{v} - \bar{\mu}e\| \leq \theta\bar{\mu}. \tag{8.4.13}$$

Доказательство. Для каждого индекса $1 \leq i \leq n$ в силу системы (8.4.4), а также формул (8.4.7) и (8.4.8) выполняется равенство

$$\begin{aligned}\bar{x}^i\bar{v}^i - \bar{\mu} &= (x^i + \alpha\Delta x^i)(v^i + \alpha\Delta v^i) - \bar{\mu} = \\ &= x^i v^i + \alpha(v^i\Delta x^i + x^i\Delta v^i) + \alpha^2\Delta x^i\Delta v^i - \bar{\mu} = \\ &= x^i v^i + \alpha(\sigma\mu - x^i v^i) + \alpha^2\Delta x^i\Delta v^i - [1 - \alpha(1-\sigma)]\mu = \\ &= (1-\alpha)(x^i v^i - \mu) + \alpha^2\Delta x^i\Delta v^i.\end{aligned}$$

Отсюда следует неравенство

$$\|D(\bar{x})\bar{v} - \bar{\mu}e\| \leq (1-\alpha)\|D(x)v - \mu e\| + \alpha^2\|D(\Delta x)\Delta v\|. \tag{8.4.14}$$

Оценим $\|D(\Delta x)\Delta v\|$. Для этого умножим первое равенство в (8.4.4) на диагональную матрицу $D^{-\frac{1}{2}}(x)D^{-\frac{1}{2}}(v)$. Если обозначить

$$q_1 = D^{\frac{1}{2}}(v)D^{-\frac{1}{2}}(x)\Delta x, \quad q_2 = D^{\frac{1}{2}}(x)D^{-\frac{1}{2}}(v)\Delta v,$$

то получившееся равенство можно записать как

$$q_1 + q_2 = D^{-\frac{1}{2}}(x)D^{-\frac{1}{2}}(v)(\sigma\mu e - D(x)D(v)e). \tag{8.4.15}$$

Так как направления Δx и Δv ортогональны друг другу, то векторы q_1 и q_2 также ортогональны друг другу. Следовательно, $\sum_{i=1}^n q_1^i q_2^i = 0$.

Обозначим через J_+ множество индексов i из $[1:n]$, для которых $q_1^i q_2^i \geq 0$. Через J_- обозначим множество оставшихся индексов, т.е. множество таких индексов i, для которых $q_1^i q_2^i < 0$. Имеем

$$0 = \sum_{i=1}^n q_1^i q_2^i = \sum_{i\in J_+} |q_1^i q_2^i| - \sum_{i\in J_-} |q_1^i q_2^i|.$$

Поэтому

$$\begin{aligned}\|D(q_1)D(q_2)e\| &= \left(\textstyle\sum_{i\in J_+} |q_1^i q_2^i|^2 + \sum_{i\in J_-} |q_1^i q_2^i|^2\right)^{\frac{1}{2}} \leq \\ &\leq \left(\left(\textstyle\sum_{i\in J_+} |q_1^i q_2^i|\right)^2 + \left(\sum_{i\in J_-} |q_1^i q_2^i|\right)^2\right)^{\frac{1}{2}} = \\ &= \left(2\left(\textstyle\sum_{i\in J_+} |q_1^i q_2^i|\right)^2\right)^{\frac{1}{2}} = \sqrt{2}\textstyle\sum_{i\in J_+} |q_1^i q_2^i| \leq \\ &\leq \tfrac{\sqrt{2}}{4}\textstyle\sum_{i\in J_+}\left(q_1^i + q_2^i\right)^2 \leq 2^{-\frac{3}{2}}\|q_1 + q_2\|^2.\end{aligned} \tag{8.4.16}$$

Здесь первое неравенство следует из соотношения между евклидовой и октаэдрической нормами $\|\cdot\|_2 \le \|\cdot\|_1$. Второе неравенство основывается на неравенстве между геометрическим и арифметическим среднем $\sqrt{q_1^i q_2^i} \le \frac{|q_1^i + q_2^i|}{2}$, которое справедливо для любых чисел q_1^i и q_2^i таких, что $q_1^i q_2^i \ge 0$.

Примем далее во внимание, что

$$D(\Delta x)\Delta v = D(\Delta x)D(\Delta v)e = D(q_1)D(q_2)e.$$

Из неравенства (8.4.16) и равенства (8.4.15), а также из левого неравенства (8.4.9) и неравенства (8.4.11) следует

$$\begin{aligned}
&\|D(\Delta x)\Delta v\| \le 2^{-\frac{3}{2}}\|q_1 + q_2\|^2 = \\
&= 2^{-\frac{3}{2}}\|D^{-\frac{1}{2}}(x)D^{-\frac{1}{2}}(v)\,(\sigma\mu e - D(x)v)\,\|^2 = \\
&= 2^{-\frac{3}{2}}\textstyle\sum_{i=1}^n (x^i v^i)^{-1}(\sigma\mu - x^i v^i)^2 \le \\
&\le 2^{-\frac{3}{2}}(1-\theta)^{-1}\mu^{-1}\|D(x)v - \sigma\mu e\|^2 \le c(\sigma,\theta)\mu,
\end{aligned}$$

где $c(\sigma,\theta) = \left(2\sqrt{2}(1-\theta)\right)^{-1}\left(\theta^2 + (1-\sigma)^2 n\right)$.

Воспользуемся теперь зависимостью (8.4.12) и оценим $c(\sigma,\theta)$ сверху. Имеем

$$c(\sigma,\theta) = \frac{\theta^2 + (1-\sigma)^2 n}{2\sqrt{2}(1-\theta)} = \frac{\theta^2}{\sqrt{2}(1-\theta)}.$$

Но если взять θ достаточно малым, то

$$\frac{\theta}{\sqrt{2}(1-\theta)} < \frac{1}{2}(1-\theta).$$

Действительно, для этого следует взять θ из интервала $(0,\theta_*)$, где θ_* — минимальный корень квадратного уравнения $(1-\theta)^2 = \sqrt{2}\theta$. Очевидно также, что $1 - \theta < 1 - \frac{\theta}{\sqrt{n}} = \sigma$ для всех n, больших единицы. Следовательно:

$$\|D(\Delta x)\Delta v\| \le \frac{1}{2}\theta\sigma\mu.$$

Отсюда и из (8.4.14), (8.4.8) получаем

$$\|D(\bar{x})\bar{v} - \bar{\mu}e\| \le (1-\alpha)\theta\mu + \alpha\theta\sigma\mu = \theta(1-\alpha(1-\sigma))\mu = \theta\bar{\mu}.$$

Здесь учтено дополнительно, что $\frac{\alpha^2}{2} < \alpha$ при $0 < \alpha \le 1$. Таким образом, выполнено (8.4.13). ■

Результат леммы 8.4.1 показывает, что при θ достаточно малом, если σ выбирается согласно (8.4.12), то даже при единичном шаге α

новая точка $[\bar{x}, \bar{v}]$ вновь оказывается в окрестности $\mathcal{N}(\theta)$, если, конечно, исходная точка $[x, v]$ лежала в этой окрестности. Это означает, что можно брать единичные шаги, не нарушая требования неотрицательности переменных x и v.

Опишем теперь итерационный процесс, соответствующий так называемому *короткошаговому методу центрального пути*. В этом методе хотя и используются единичные шаги, но траектория проходит достаточно далеко от границы неотрицательного ортанта $\mathbb{R}^{2n}_+$, а точнее, вблизи центрального пути. В нем задается $0 < \theta < \theta_*$ и полагается $\sigma = 1 - \frac{\theta}{\sqrt{n}}$. Стартовая точка $[x_0, v_0]$ берется из окрестности $\mathcal{N}(\theta)$. Итерационный процесс описывается следующей рекуррентной схемой:

$$x_{k+1} = x_k + \Delta x_k, \quad v_{k+1} = v_k + \Delta v_k,$$

где Δx_k и Δv_k находятся из решения системы (8.4.4) при $x = x_k$, $v = v_k$.

Теорема 8.4.1. *Пусть задано* $0 < \varepsilon < 1$. *Тогда для любой стартовой точки* $[x_0, v_0] \in \mathcal{N}(\theta)$, *в которой* $\mu_0 = \mu(x_0, v_0)$, *можно указать достаточно большое* $k_* = O(\sqrt{n} \ln \frac{1}{\varepsilon})$ *такое, что* $\mu_k = \mu(x_k, v_k) \le \varepsilon$ *при* $k \ge k_*$.

Доказательство. Как следует из утверждения леммы 8.4.1, все точки $[x_k, v_k]$ остаются вблизи центрального пути. Кроме того, в силу (8.4.7) между двойственными мерами μ_k и μ_{k-1} имеет место связь, которая при $\alpha = 1$ принимает вид $\mu_k = \sigma \mu_{k-1}$. После логарифмирования приходим к $\ln \mu_k = \ln \sigma + \ln \mu_{k-1}$. Применяя это равенство последовательно для $k = 1, 2, \ldots$, получаем

$$\ln \mu_k = k \ln \sigma + \ln \mu_0. \tag{8.4.17}$$

Но $\mu_0 \le \frac{1}{\varepsilon^l}$ для некоторого целого $l \ge 1$. Поэтому (8.4.17) можно переписать как

$$\ln \mu_k \le k \ln \sigma + l \ln \frac{1}{\varepsilon}.$$

Воспользуемся далее неравенством

$$\ln \sigma = \ln \left(1 - \frac{\theta}{\sqrt{n}}\right) \le -\frac{\theta}{\sqrt{n}},$$

справедливым для логарифмической функции. Тогда нужная оценка $\mu_k \le \varepsilon$ выполняется, если только

$$-\frac{k\theta}{\sqrt{n}} + l \ln \frac{1}{\varepsilon} \le \ln \varepsilon.$$

Для этого следует взять $k \geq k_*$, где k_* — произвольное целое число такое, что

$$k_* \geq \frac{l+1}{\theta}\sqrt{n}\ln\left(\frac{1}{\varepsilon}\right).$$

Теорема доказана. ■

Результат теоремы 8.4.1 важен по крайней мере с двух точек зрения. Во-первых, он показывает, что короткошаговый метод центрального пути обладает в некотором смысле глобальной сходимостью — стартовые точки могут быть сколь угодно далеко удалены от прямо-двойственного решения $[x_*, v_*]$, необходимо только, чтобы они принадлежали окрестности центрального пути. Во-вторых, метод принадлежит к так называемым *полиномиальным методам* решения задач линейного программирования (8.1.1), (8.1.2). Чтобы добиться прямо-двойственного решения с нужной точностью, необходимо совершить число итераций, которое зависит главным образом от размерности задачи, а именно пропорционально величине $\sqrt{n}\ln\left(\frac{1}{\varepsilon}\right)$.

Можно выписать явные выражения для ньютоновских направлений Δx и Δv, удовлетворяющих системе (8.4.4). В самом деле, из второго и третьего уравнений следует, что $\Delta x = K^T z_1$, $\Delta v = A^T z_2$ для некоторых $z_1 \in \mathbb{R}^d$, $z_2 \in \mathbb{R}^m$. После подстановки данных Δx и Δv в первое уравнение получаем

$$D(v)K^T z_1 + D(x)A^T z_2 = \sigma\mu e - D(x)v. \tag{8.4.18}$$

Воспользуемся несколько упрощенным обозначением $D\left(\frac{x}{v}\right)$ для диагональной матрицы $D(x)D^{-1}(v)$. Аналогично пусть $D\left(\frac{v}{x}\right)$ обозначает диагональную матрицу $D(v)D^{-1}(x)$. Если умножить равенство (8.4.18) слева на матрицу $AD^{-1}(v)$, то приходим к уравнению относительно вектора z_2:

$$AD\left(\frac{x}{v}\right)A^T z_2 = \sigma\mu AD^{-1}(v)e - Ax. \tag{8.4.19}$$

Отсюда, если матрица $AD\left(\frac{x}{v}\right)A^T$ неособая, находим

$$z_2 = \left(AD(\frac{x}{v})A^T\right)^{-1}\left(\sigma\mu AD^{-1}(v)e - Ax\right),$$

поэтому

$$\Delta v = A^T\left(AD\left(\frac{x}{v}\right)A^T\right)^{-1}A\left(\sigma\mu D^{-1}(v)e - x\right). \tag{8.4.20}$$

Кроме того, согласно первому равенству в (8.4.4)

$$\Delta x = -D\left(\frac{x}{v}\right)\Delta v + \sigma\mu D^{-1}(v)e - x.$$

Следовательно,

$$\Delta x = \left[I_n - D(\frac{x}{v}) A^T \left(AD\left(\frac{x}{v}\right) A^T \right)^{-1} A \right] \left(\sigma\mu D^{-1}(v)e - x \right). \quad (8.4.21)$$

Можно поступить и другим образом, а именно, умножить равенство (8.4.18) слева на матрицу $KD^{-1}(x)$. Тогда вместо (8.4.19) приходим к уравнению относительно z_1:

$$KD\left(\frac{v}{x}\right) K^T z_1 = \sigma\mu KD^{-1}(x)e - Kv, \quad (8.4.22)$$

разрешая которое при неособой матрице $KD\left(\frac{v}{x}\right) K^T$ находим

$$z_1 = \left(KD\left(\frac{v}{x}\right) K^T \right)^{-1} K \left(\sigma\mu D^{-1}(x)e - v \right).$$

Зная z_1, аналогично вышесказанному определяем

$$\Delta x = K^T \left(KD\left(\frac{v}{x}\right) K^T \right)^{-1} K \left(\sigma\mu D^{-1}(x)e - v \right), \quad (8.4.23)$$

$$\Delta v = \left[I_n - D(\frac{v}{x}) K^T \left(KD\left(\frac{v}{x}\right) K^T \right)^{-1} K \right] \left(\sigma\mu D^{-1}(x)e - v \right). \quad (8.4.24)$$

Ньютоновские направления Δx и Δv, выраженные через формулы (8.4.20) и (8.4.21), полностью совпадают с теми, которые задаются формулами (8.4.23) и (8.4.24). Какую из систем (8.4.19) или (8.4.22) предпочтительнее решать для определения Δx и Δv зависит от нескольких факторов, в частности от размерности z_1 и z_2. Напомним, что вектор z_2 имеет размерность m, а вектор z_1 — размерность $n - m$. Системы линейных алгебраических уравнений (8.4.19) и (8.4.22) в отличие от основной ньютоновской системы (8.4.4) носят названия *нормализованных систем*. В этих системах матрицы перед неизвестными симметрические, что позволяет использовать для их решения разложение Холесского.

Глава 9

Сложность задач и эффективность методов

9.1. Сложность задачи для методов

При рассмотрении численных методов решения оптимизационных задач встает вопрос об их сравнительной эффективности и выборе наилучшего метода для решения конкретной задачи. Сопоставление методов проводят на основе разных характеристик. Одной из основных является уже знакомая нам *скорость сходимости*. Но скорость сходимости, как правило, определяется в пределе, когда итерации находятся вблизи решения, т.е. указывает на финальное локальное поведение методов. Поэтому для многих методов, например для симплекс-метода решения задач линейного программирования, являющегося конечным методом, такое понятие не имеет смысла.

Существует общая *теория сложности задачи для методов*. Согласно этой теории при сопоставлении численных методов прежде всего выделяют *класс задач*, для решения которых они предназначены. Метод, решая конкретную задачу из данного класса, пользуется некоторой информацией о ней. Та часть информации о задаче, которая доступна методу, называется *моделью решаемой задачи*. Под моделью понимаются постановка задачи, описание свойств функций, входящих в данную постановку и т.д. Фактически метод при решении задачи общается не с самой задачей, а с моделью. Конкретная задача из данного класса может быть описана разными моделями. Например, при решении задач условной оптимизации ограничения простого вида могут

выделяться в отдельное подмножество ограничений, образуя базовое допустимое множество, а могут учитываться единым образом с другими функциональными ограничениями. Для задачи линейной условной оптимизации возможно задание допустимого множества как через ограничения типа равенства, так и через ограничения-неравенства и тому подобное.

Стоит сразу отметить, что получить точное решение оптимизационной задачи, особенно нелинейной, как правило, невозможно. Поэтому ограничиваются нахождением некоторого *приближенного решения с точностью* $\varepsilon > 0$, и именно оно считается приемлемым решением задачи. Отсюда возникает понятие точности, которое, вообще говоря, может иметь разный смысл для разных задач.

Далее вводят понятие *оракула*, удобное для описания взаимодействия метода с моделью задачи. Под оракулом понимают некоторый «механизм» или процедуру, с помощью которых метод получает конкретную информацию о задаче, причем обычно предполагается, что ответы оракула являются локальными. Это соответствует так называемой *концепции черного ящика*, согласно которой нам не доступна информация о задаче вдали от текущей точки. Метод в процессе вычисления указывает точку x, а оракул выдает информацию о задаче в этой точке.

Для используемых обычно *функциональных моделей* оптимизационных задач, в которых задача задается с помощью целевой функции и функций-ограничений, весьма важными являются предположения о гладкости функций, входящих в данную постановку, и о возможности использовании в ходе вычислений производных. Поэтому выделяют оракулы нулевого, первого и второго порядков. В принципе, возможны также оракулы более высоких порядков. *Оракул нулевого порядка* возвращает только значения функций. *Оракул первого порядка* возвращает дополнительно значения производных. *Оракул второго порядка* добавляет к этим значениям также матрицы вторых производных. Производные могут быть приближенными, если они считаются численно.

Таким образом, класс решаемых задач формально задается как совокупность используемой модели, заданного критерия точности и доступного оракула. Это и есть представление задачи, с которым работает метод.

Обратимся теперь к методам. Многие методы, применяемые для решения оптимизационных задач, носят итерационный характер. Они строят последовательности точек, анализируя информацию о задаче в текущей точке и, быть может, в предыдущих точках траектории. Вся

эта информация на каждой итерации и выдается оракулом. Используя эту информацию, метод указывает, какую точку взять следующей. Например, следующая точка определяется с помощью выбора направления сдвига и шага вдоль этого направления.

Под *эффективностью метода* или его *трудоемкостью* на рассматриваемом классе задач понимаются *вычислительные затраты*, необходимые для решения любой задачи из этого класса. Эти вычислительные затраты возникают прежде всего при обращении к оракулу для собирания нужной информации о задаче в текущей точке и передачи ее методу. Кроме того, они возникают при обработке полученной информации и построении последующей точки (собственно работа метода). Поэтому вводят два понятия *сложности задачи для метода* — аналитическую и арифметическую.

Аналитическая сложность задачи определяется как число обращений к оракулу, необходимое для решения задачи с требуемой точностью $\varepsilon > 0$. *Арифметическая сложность* есть общее число всех вычислений, необходимых для решения задачи с точностью $\varepsilon > 0$. Она включает работу оракула и работу метода. Хотя арифметическая сложность более точно характеризует работу метода, однако, как правило, зная аналитическую сложность, можно определить и арифметическую сложность.

Возьмем некоторый конкретный метод, предназначенный для решения заданного класса задач, и определим для него аналитическую сложность задачи. Имея такую аналитическую сложность задачи для выбранного метода, мы тем самым имеем некоторую *верхнюю оценку аналитической сложности* для класса задач. Но есть и другие методы. Подсчитаем для каждого другого метода аналитическую сложность для «наихудшей» задачи из данного класса (эти «наихудшие» задачи могут быть разными для разных методов). Тем самым будут получены ряд оценок и мы можем выбрать среди них наименьшую. О такой оценке говорят, как о *нижней оценке сложности* для класса задач. Если окажется, что верхняя оценка сложности для конкретного метода с точностью до константы совпадает с нижней оценкой, то такой метод называется *асимптотически оптимальным* на рассматриваемом классе задач. Для выбора для конкретного метода «наихудшей» задачи из класса задач используется понятие *сопротивляющегося оракула*. В процессе работы метода он постепенно строит такую «наихудшую» задачу, выдавая каждый раз «наихудшую» с точки зрения метода информацию.

Для многих методов, особенно итерационного типа, верхняя оценка сложности может быть определена через количество итераций, ко-

торое необходимо проделать методу, чтобы получить решение задачи с заданной точностью. Конечно, такая характеристика связана определенным образом со скоростью сходимости итерационных методов, хотя, быть может, и не всегда. Количество итераций зависит дополнительно от выбранной стартовой точки и от ряда других факторов, в частности от свойств функций, описывающих задачу.

Сравнительно просто можно оценить необходимое количество итераций для задач, имеющих специальную структуру, например, задач линейного или квадратичного программирования. При решении задач линейного программирования такой оценкой является число угловых точек у допустимого множества. Для общих же нелинейных задач ситуация с оценкой количества необходимых итераций становится гораздо более трудной и она может быть разрешима лишь при дополнительных предположениях о свойствах функций, определяющих задачу. К таким предположениям, упрощающим нахождение данных оценок, относится *предположение о выпуклости задачи*. Поэтому возьмем задачу выпуклой безусловной минимизации:

$$f_* = \min_{x \in R^n} f(x), \tag{9.1.1}$$

где $f(x)$ — дважды непрерывно дифференцируемая сильно выпуклая функция с константой $\theta > 0$. Задача (9.1.1) обязательно имеет решение, причем единственное.

Из необходимых и достаточных условий сильной выпуклости первого порядка следует, что для любых x и y из $\mathbb{R}^n$ справедливо неравенство

$$f(y) \geq f(x) + \langle f_x(x), y - x \rangle + \theta \|y - x\|^2. \tag{9.1.2}$$

Отсюда получаем

$$\begin{array}{rcl} f(y) & \geq & f(x) + \langle f_x(x), y - x \rangle + \theta \|y - x\|^2 \geq \\ & \geq & f(x) + \min_{z \in R^n} \left[\langle f_x(x), z - x \rangle + \theta \|z - x\|^2 \right] = \\ & = & f(x) - (4\theta)^{-1} \|f_x(x)\|^2. \end{array} \tag{9.1.3}$$

Последнее равенство следует из того, что минимум сильно выпуклой функции, стоящей в квадратных скобках, достигается согласно необходимым условиям оптимальности при $z = x - (2\theta)^{-1} f_x(x)$. Беря теперь в качестве y в левой части неравенства (9.1.3) точку x_*, являющуюся решением задачи (9.1.1), приходим к следующей оценке:

$$f(x) - f_* \leq (4\theta)^{-1} \|f_x(x)\|^2. \tag{9.1.4}$$

Данное неравенство показывает, что норма градиента целевой функции может служить хорошей оценкой близости к решению задачи по функционалу.

Для сильно выпуклой целевой функции норма градиента также является оценкой близости к решению x_* в пространстве исходных переменных x. В самом деле, если положить в неравенстве (9.1.2) $y = x_*$ и воспользоваться неравенством Коши–Буняковского, то имеем

$$\begin{aligned} f_* &\geq f(x) + \langle f_x(x), x_* - x\rangle + \theta\|x_* - x\|^2 \geq \\ &\geq f(x) - \|f_x(x)\|\|x_* - x\| + \theta\|x_* - x\|^2. \end{aligned}$$

Поскольку $f(x) \geq f_*$, то отсюда следует

$$\|x - x_*\| \leq \theta^{-1}\|f_x(x)\|.$$

Константа θ сильной выпуклости функции $f(x)$ может служить оценкой минимального собственного значения $\lambda_{\min}$ матрицы вторых производных $f_{xx}(x)$, причем для любого x из области определения функции $f(x)$. Действительно, согласно критерию второго порядка, дважды непрерывно дифференцируемая функция $f(x)$ будет сильно выпуклой с константой $\theta > 0$ на выпуклом множестве X тогда и только тогда, когда в любой точке x, принадлежащей внутренности множества X,

$$\langle s, f_{xx}(x)s\rangle \geq 2\theta\|s\|^2$$

для всех $s \in \mathbb{R}^n$. Отсюда следует, что $\lambda_{\min} \geq m$, где $m = 2\theta$. Если максимальное собственное значение $\lambda_{\max}$ также равномерно ограничено на X, например числом M, то выполняются неравенства

$$m\|s\|^2 \leq \langle s, f_{xx}(x)s\rangle \leq M\|s\|^2. \tag{9.1.5}$$

Из правого неравенства следует, что

$$f(y) \leq f(x) + \langle f_x(x), y - x\rangle + \frac{M}{2}\|y - x\|^2 \tag{9.1.6}$$

для любого $y \in \mathbb{R}^n$. В частности, можно указать то y, при котором выпуклая квадратичная функция в правой части в (9.1.6) принимает минимальное значение. Дифференцируя эту функцию по y и приравнивая производную нулю, получаем $y = -M^{-1}f_x(x) + x$. Поэтому

$$f(x) - f(y) \geq \frac{1}{2M}\|f_x(x)\|^2$$

для такого y. Поскольку $f(x) - f_* \geq f(x) - f(y)$, то отсюда и из (9.1.4) приходим к двусторонней оценке:

$$\frac{1}{2M}\|f_x(x)\|^2 \leq f(x) - f_* \leq \frac{1}{2m}\|f_x(x)\|^2. \tag{9.1.7}$$

Определим теперь число итераций, которое необходимо проделать методами первого или второго порядков для решения задачи (9.1.1) с точностью $\varepsilon > 0$ по функционалу, т.е. для нахождения такой точки x_k, что $f(x_k) - f_* \leq \varepsilon$. Рассматриваемый класс задач, для решения которых привлекаются эти методы, — задачи минимизации выпуклой функции со второй производной, удовлетворяющей неравенствам (9.1.5), с критерием точности по функционалу и с использованием оракулов первого или второго порядков. Согласно вышесказанному, для оценки близости точки x_k к решению можно воспользоваться нормой градиента функции. В силу (9.1.7) достаточно, чтобы выполнялось неравенство $\|f_x(x_k)\| \leq \sqrt{2m\varepsilon}$.

Проведем сначала анализ *метода градиентного спуска*. Для простоты предполагаем, что шаг в методе выбирается на основе одномерной минимизации, т.е., другими словами, рассматривается метод наискорейшего спуска (метод Коши):

$$x_{k+1} = x_k - \alpha_k f_x(x_k), \quad \alpha_k = \arg\min_{\alpha \geq 0} f(x_k - \alpha f_x(x_k)).$$

Полагая в неравенстве (9.1.6) $x = x_k$ и $y = x_k - \alpha f_x(x_k)$, получаем

$$f\left(x_k - \alpha f_x(x_k)\right) \leq f(x_k) - \alpha \|f_x(x_k)\|^2 + \frac{M\alpha^2}{2}\|f_x(x_k)\|^2.$$

Если рассматривать правую часть этого неравенства как квадратичную функцию относительно α, то ее минимум достигается при значении $\tilde{\alpha} = M^{-1}$. Поэтому

$$f(x_{k+1}) = f\left(x_k - \alpha_k f_x(x_k)\right) \leq f\left(x_k - \tilde{\alpha} f_x(x_k)\right) \leq f(x_k) - \frac{1}{2M}\|f_x(x_k)\|^2.$$

Отсюда приходим к неравенству

$$f(x_{k+1}) - f_* \leq f(x_k) - f_* - \frac{1}{2M}\|f_x(x_k)\|^2.$$

Но, согласно (9.1.7), $\|f_x(x_k)\|^2 \geq 2m\left(f(x_k) - f_*\right)$. Поэтому

$$f(x_k) - f_* \leq C\left(f(x_{k-1}) - f_*\right) \leq C^k\left(f(x_0) - f_*\right), \tag{9.1.8}$$

где $C = 1 - \frac{m}{M} < 1$, если, конечно, $m < M$. Из (9.1.8) следует оценка на число итераций K, требуемых для достижения точности ε по функционалу, а именно K — наименьшее целое, превышающее величину

$$K_\varepsilon = \frac{\ln \frac{f(x_0) - f_*}{\varepsilon}}{\ln \frac{1}{C}}. \tag{9.1.9}$$

Данное количество итераций зависит от коэффициента C, т.е. фактически от отношения $\frac{M}{m}$, являющегося верхней границей *числа обусловленности* матрицы вторых производных $f_{xx}(x)$ целевой функции $f(x)$.

Можно также определить соответствующее количество итераций, которое необходимо проделать методу градиентного спуска для нахождения решения, когда шаг выбирается по способу Армихо или берется постоянным. Как было отмечено ранее, возможно также получение соответствующих оценок и по аргументу.

Обратимся теперь к *методу Ньютона*. Ранее в главе 2 мы рассмотрели два основных варианта метода Ньютона, отличающиеся выбором шага. В первом варианте шаг берется постоянным, равным единице. Данный вариант метода обладает локальной сверхлинейной скоростью сходимости. Более того, скорость сходимости является квадратичной, если матрица вторых производных функции $f(x)$ удовлетворяет условию Липшица.

Во втором варианте шаг подбирается путем дробления начального единичного шага по правилу Армихо. Это делает метод глобально сходящимся, причем, как было установлено, при приближении к решению метод начинает вести себя как первый вариант, т.е. единичный шаг перестает дробиться.

Пусть матрица вторых производных $f_{xx}(x)$ сильно выпуклой функции $f(x)$ удовлетворяет условию Липшица:

$$\|f_{xx}(x) - f_{xx}(y)\| \leq L\|x - y\| \tag{9.1.10}$$

для любых x и y из области определения функции $f(x)$. Считаем также для простоты, что начальная точка x_0 и все последующие точки x_k находятся достаточно близко к решению, так что шаг α_k оказывается равным единице. В этом случае, беря $s_k = -f_{xx}^{-1}(x_k)f_x(x_k)$, на основании формул (9.1.10) и (9.1.5) получаем

$$\begin{aligned} \|f_x(x_{k+1})\| &= \|\int_0^1 [f_{xx}(x_k + \tau s_k) - f_{xx}(x_k)]s_k \, d\tau\| \leq \\ &\leq \int_0^1 \|f_{xx}(x_k + \tau s_k) - f_{xx}(x_k)\|\|s_k\|d\tau \leq \\ &\leq \tfrac{L}{2}\|s_k\|^2 = \tfrac{L}{2}\|f_{xx}(x_k)f_x(x_k)\|^2 \leq \tfrac{L}{2m^2}\|f_x(x_k)\|^2. \end{aligned} \tag{9.1.11}$$

Отсюда видно, что если $\|f_x(x_k)\| \leq \frac{m^2}{L}$, то

$$\|f_x(x_{k+1})\| \leq \frac{1}{2}\|f_x(x_k)\| \leq \frac{m^2}{L},$$

т.е. градиент удовлетворяет этой же оценке и в последующей точке.

Более того, согласно (9.1.11) справедливо следующее неравенство:

$$\frac{L}{2m^2}\|f_x(x_{k+1})\| \le \left(\frac{L}{2m^2}\|f_x(x_k)\|\right)^2. \tag{9.1.12}$$

Далее предполагаем, что начальная точка x_0 такова, что градиент $f_x(x_0)$ удовлетворяет неравенству $\|f_x(x_0)\| \le \frac{m^2}{L}$. Тогда неравенство (9.1.12), связывающие нормы градиентов, будет выполняться для всех итераций. Применяя данное неравенство рекуррентным образом, получаем оценку

$$\frac{L}{2m^2}\|f_x(x_k)\| \le \left(\frac{L}{2m^2}\|f_x(x_0)\|\right)^{2^k} \le \left(\frac{1}{2}\right)^{2^k}. \tag{9.1.13}$$

С помощью (9.1.7) и (9.1.13) приходим к оценке на близость решения по функционалу:

$$f(x_k) - f_* \le \frac{1}{2m}\|f_x(x_k)\|^2 \le \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{k+1}}. \tag{9.1.14}$$

Неравенство (9.1.14) позволяет указать нижнюю оценку на количество итераций, после выполнения которых задача (9.1.1) решена с точностью $\varepsilon > 0$ по функционалу, а именно: теперь вместо (9.1.9) следует брать величину

$$K_\varepsilon = \log_2 \log_2 \frac{\varepsilon_0}{\varepsilon},$$

где $\varepsilon_0 = \frac{2m^3}{L^2}$. Можно получить оценку и в общем случае, когда сначала используется второй вариант метода, а при приближении к решению задачи он заменяется на первый вариант. При этом в оценку дополнительно войдет константа M.

Итак, нами получены верхние оценки аналитической сложности задачи (9.1.1) для метода градиентного спуска и для метода Ньютона.

9.2. Самосогласованные функции

Одним из важных полезных качеств метода Ньютона является его инвариантность по отношению к линейным преобразованиям переменных. В самом деле, предположим, что $x = Qy$, где Q — неособая квадратная матрица порядка n. Тогда, заменяя x на y, приходим к следующей задаче: найти

$$f_* = \min_{y \in R^n} \tilde{f}(y), \quad \tilde{f}(y) = f(Qy). \tag{9.2.1}$$

Применяя метод Ньютона с постоянным единичным шагом к минимизации функции $\tilde{f}(y)$, получаем

$$y_{k+1} = y_k - \tilde{f}_{yy}^{-1}(y_k)\tilde{f}_y(y_k). \tag{9.2.2}$$

Но

$$\tilde{f}_y(y) = Q^T f_x(Qy) = Q^T f_x(x), \quad \tilde{f}_{yy}(y) = Q^T f_{xx}(Qy)Q = Q^T f_{xx}(x)Q.$$

Поэтому формула пересчета (9.2.2) переписывается как

$$\begin{array}{rcl} y_{k+1} & = & y_k - Q^{-1} f_{xx}^{-1}(x_k)(Q^T)^{-1}Q^T f_x(x_k) = \\ & = & y_k - Q^{-1} f_{xx}^{-1}(x_k) f_x(x_k) = Q^{-1}\left[x_k - f_{xx}^{-1}(x_k) f_x(x_k)\right], \end{array}$$

из которой следует, что $x_{k+1} = Qy_{k+1}$. Таким образом, можно менять систему координат, при этом расчеты по методу Ньютона не меняются.

В предыдущем разделе были получены оценки снизу на необходимое количество итераций метода Ньютона, после выполнения которых достигается нужная точность решения по функционалу. Эти оценки зависят от констант, определяющих функцию $f(x)$, а именно: от m, L и, быть может, M. Однако если сменить систему координат, то поменяются и эти константы, что повлечет изменение на оценки количества итераций.

С другой стороны, как только что отмечено, метод Ньютона инвариантен по отношению к замене системы координат. Поэтому анализ метода Ньютона, касающийся оценок снизу количества итераций и опирающийся на такие характеристики функции $f(x)$, как неравенства (9.1.5) и (9.1.10) для вторых производных, представляется не совсем удачным.

Встает вопрос: существуют ли выпуклые функции, удовлетворяющие некоторому условию, которое остается инвариантным относительно линейных преобразований переменных. Тогда, опираясь на это условие, можно проводить теоретический анализ метода Ньютона, включая скорость сходимости и оценку числа итераций, необходимых для решения задачи. Ответ на такой вопрос положительный и он приводит к выделению среди всей совокупности выпуклых функций целого класса функций, обладающих этим полезным фундаментальным свойством.

Определение 9.2.1. *Трижды непрерывно дифференцируемая выпуклая замкнутая функция одной переменной $f(x)$ называется самосогласованной, если*

$$|f'''(x)| \leq 2\left(f''(x)\right)^{\frac{3}{2}} \tag{9.2.3}$$

для всех $x \in \mathbb{R}$ из области определения функции $f(x)$, которая считается открытой.

Прежде всего отметим, что условие самосогласованности (9.2.3) может быть переписано в виде

$$\left| \frac{d}{dx} \left(f''(x) \right)^{-\frac{1}{2}} \right| \leq 1. \tag{9.2.4}$$

Понятно, что класс выпуклых самосогласованных функций не пустой, так как к нему заведомо относятся линейные функции и выпуклые квадратичные функции. Из других важных примеров укажем следующие две функции: *отрицательный логарифм* $f(x) = -\ln x$ и его *сумму с отрицательной энтропией* $f(x) = x \ln x - \ln x$. Обе функции определены при $x > 0$.

Упражнение 12. *Покажите, что функция $f(x) = -\ln x$ и функция $f(x) = x \ln x - \ln x$ являются самосогласованными в смысле определения 9.2.1.*

Упражнение 13. *Пусть a — отличная от нуля константа. Покажите, что функция $\tilde{f}(y) = f(ay)$ будет самосогласованной, если самосогласованной является функция $f(x)$.*

Дадим теперь обобщение понятия самосогласованной функции одной переменной на функции с несколькими переменными.

Определение 9.2.2. *Трижды непрерывно дифференцируемая выпуклая замкнутая функция $f(x)$, определенная на открытом множестве $X \subseteq \mathbb{R}^n$, называется самосогласованной, если для любых $x \in X$ и любых ненулевых $s \in \mathbb{R}^n$ функция одной переменной $\phi(\alpha) = f(x+\alpha s)$ является самосогласованной на своей области определения.*

Самосогласованные функции обладают свойством, что они стремятся к бесконечности при подходе к границе множества X. Их также можно умножать на коэффициент $\alpha \geq 1$ и складывать между собой. При этом свойство самосогласованности не нарушается. Действительно, пусть $f_1(x)$ и $f_2(x)$ — две самосогласованные функции, $x \in \mathbb{R}$. Тогда, используя неравенство

$$(a_1^\beta + a_2^\beta)^{\frac{1}{\beta}} \leq a_1 + a_2,$$

справедливое для любых $a_1, a_2 \geq 0$ и $\beta \geq 1$, получаем

$$\begin{aligned} |f_1'''(x) + f_2'''(x)| &\leq |f_1'''(x)| + |f_2'''(x)| \leq \\ &\leq 2\left[(f_1''(x))^{\frac{3}{2}} + (f_2''(x))^{\frac{3}{2}}\right] \leq \\ &\leq 2\left[f_1''(x) + f_2''(x)\right]^{\frac{3}{2}}. \end{aligned}$$

Таким образом, функция $f(x) = f_1(x) + f_2(x)$ является самосогласованной.

Как мы знаем, суперпозиция выпуклой функции f с аффинной функцией $Ax + b$ является выпуклой функцией. Если дополнительно функция f оказывается самосогласованной, то и результирующая функция $f(Ax + b)$ также будет самосогласованной.

Предположим, что имеется система линейных неравенств

$$Ax - b \leq 0_m, \tag{9.2.5}$$

где A — матрица размера $m \times n$, $b \in \mathbb{R}^m$. Обозначим через a_i i-ю строку матрицы A, $1 \leq i \leq m$. Тогда в силу вышесказанного функция

$$f(x) = -\sum_{i=1}^{m} \ln(b_i - \langle a_i, x\rangle) \tag{9.2.6}$$

определена на множестве всех $x \in \mathbb{R}^n$, таких, что $Ax < b$, и является самосогласованной. Это следует из того, что каждое слагаемое в (9.2.6) есть суперпозиция отрицательного логарифма и аффинной функции $b^i - \langle a_i, x\rangle$.

Функция (9.2.6) есть не что иное, как логарифмический внутренний штраф для допустимого множества X, задаваемого системой неравенств (9.2.5). Если минимум функции (9.2.6) существует, то точка x_*, в которой он достигается, носит название *аналитического центра* системы линейных неравенств (9.2.5).

Сделаем одно замечание. В определении 9.2.1 самосогласованной функции стоит константа «двойка». В принципе, она может быть заменена на любую другую неотрицательную константу, т.е. вместо неравенства (9.2.3) можно взять

$$|f'''(x)| \leq L_f \left(f''(x)\right)^{\frac{3}{2}},$$

где $L_f \geq 0$. В том случае, когда $L_f = 2$, о функции $f(x)$ говорят как о *стандартной самосогласованной функции*. Любая самосогласованная функция $f(x)$ с константой $L_f > 0$ может быть сведена к стандартной самосогласованной функции $\tilde{f}(x)$ путем умножения $f(x)$ на

коэффициент $c = (\frac{L_f}{2})^2$. В свете такого расширенного определения самосогласованной функции получаем, что произведение самосогласованной функции на произвольную положительную константу (а не только большую единицы) дает нам самосогласованную функцию.

Проведем теперь анализ метода Ньютона для решения задачи безусловной минимизации (9.1.1), предполагая, что $f(x)$ является сильно выпуклой стандартной самосогласованной функцией. Наша цель по-прежнему состоит в том, чтобы найти оценку необходимого числа итераций для получения нужной точности $\varepsilon > 0$ по функционалу. Для простоты рассмотрим только классический вариант метода Ньютона с единичным шагом. Это предполагает, что начальная точка x_0 взята достаточно близко к решению задачи (9.1.1) — точке x_*.

Пусть $\delta_f(x)$ обозначает *ньютоновское убывание* в точке x, которое, напомним, определяется как

$$\delta_f(x) = \langle\, f_x(x), f_{xx}^{-1}(x) f_x(x) \,\rangle^{\frac{1}{2}}.$$

Для исследования поведения метода Ньютона при минимизации самосогласованной функции такая величина оказывается более подходящей, чем квадрат нормы градиента. Ньютоновское убывание можно трактовать также как норму градиента, но относительно квадратичной нормы $\|z\|_Q = \langle z, Qz\rangle$, в которой в качестве симметричной положительно определенной матрицы Q берется матрица $Q = f_{xx}^{-1}(x)$. Поэтому о величине $\delta_f(x)$ будем говорить как о локальной норме градиента $f_x(x)$ и использовать дополнительно обозначение: $\delta_f(x) = \|f_x(x)\|_{f_{xx}^{-1}(x)}$.

Наряду с квадратично нормой $\|\cdot\|_{f_{xx}^{-1}(x)}$ нам потребуется другая квадратичная норма $\|\cdot\|_{f_{xx}(x)}$, задаваемая с помощью обратной матрицы $Q = f_{xx}(x)$. Данная норма является двойственной по отношению к норме $\|\cdot\|_{f_{xx}^{-1}(x)}$. Имеет место неравенство

$$|\langle s_1, s_2\rangle| \le \|s_1\|_{f_{xx}^{-1}(x)} \|s_2\|_{f_{xx}(x)} \quad \forall\ s_1, s_2 \in \mathbb{R}^n. \tag{9.2.7}$$

Покажем сначала, как с помощью ньтоновского убывания можно оценить невязку по функционалу. Для этого нам потребуются некоторые вспомогательные результаты, касающиеся самосогласованных функций.

Лемма 9.2.1. *Пусть $f(x)$ — самосогласованная функция, определенная на открытом множестве $X \subseteq \mathbb{R}^n$. Тогда для любых $x, y \in X$ имеет место неравенство*

$$f(y) \ge f(x) + \langle f_x(x), y - x\rangle + \omega\left(\, \|y - x\|_{f_{xx}(x)} \,\right), \tag{9.2.8}$$

где $\omega(t) = t - \ln(1 + t)$.

Доказательство. В силу определения самосогласованной функции $f(x)$, функция одной переменной $\phi(\alpha) = f(x + \alpha s)$ также является самосогласованной для любого ненулевого $s \in \mathbb{R}^n$ на своей области определения. Вторая производная этой функции имеет вид $\phi''(\alpha) = \langle s, f_{xx}(x + \alpha s)s\rangle$.

Возьмем теперь $s = y - x$ и обозначим $x(\alpha) = x + \alpha s$, $h = \|s\|_{f_{xx}(x)}$. Рассмотрим дополнительно функцию одного аргумента:

$$\psi(\alpha) = \|s\|^{-1}_{f_{xx}(x(\alpha))} = \left(\phi''(\alpha)\right)^{-\frac{1}{2}}. \tag{9.2.9}$$

Область её определения — это такие α, для которых $\psi(\alpha) > 0$.

В силу (9.2.4) имеет место неравенство $|\psi'(\alpha)| \leq 1$. Поэтому область определения $\psi(\alpha)$ заведомо содержит интервал $(-\psi(0), \psi(0))$, и для всех α из этого интервала выполняются неравенства

$$\psi(0) - |\alpha| \leq \psi(\alpha) \leq \psi(0) + |\alpha|. \tag{9.2.10}$$

Правое неравенство в (9.2.10) при $\alpha > 0$ переписывается как

$$\|s\|_{f_{xx}(x(\alpha))} \geq \frac{\|s\|_{f_{xx}(x)}}{1 + \alpha\|s\|_{f_{xx}(x)}} = \frac{h}{1 + \alpha h}. \tag{9.2.11}$$

Далее на основании (9.2.11) получаем

$$\begin{array}{l} \langle f_x(y) - f_x(x), y - x\rangle = \int_0^1 \langle f_{xx}(x(\alpha))s, s\rangle d\alpha = \\ \quad = \int_0^1 \|s\|^2_{f_{xx}(x(\alpha))} d\alpha \geq \\ \quad \geq \int_0^1 \frac{h^2}{(1+\alpha h)^2} d\alpha = h\int_0^h \frac{1}{(1+\alpha)^2} d\alpha = \frac{h^2}{1+h}. \end{array} \tag{9.2.12}$$

Неравенство (9.2.12) сохранится, если вместо y взять $x(\alpha)$ и учесть, что $\|x(\alpha) - x\|_{f_{xx}(x)} = \alpha h$. Тогда это неравенство примет вид

$$\langle f_x(x(\alpha)) - f_x(x), x(\alpha) - x\rangle \geq \frac{\alpha^2 h^2}{1 + \alpha h}. \tag{9.2.13}$$

Применяя теперь неравенство (9.2.13), приходим к

$$\begin{array}{l} f(y) - f(x) - \langle f_x(x), y - x\rangle = \int_0^1 \langle f_x(x(\alpha)) - f_x(x), y - x\rangle d\alpha = \\ \quad = \int_0^1 \alpha^{-1}\langle f_x(x(\alpha)) - f_x(x), x(\alpha) - x\rangle d\alpha \geq \\ \quad \geq \int_0^1 \frac{\alpha h^2}{1+\alpha h} d\alpha = \int_0^h \frac{\alpha}{1+\alpha} d\alpha = \omega(h). \end{array}$$

Таким образом, неравенство (9.2.8) действительно выполняется. ■

Лемма 9.2.2. *Пусть $f(x)$ — самосогласованная функция, определенная на открытом множестве $X \subseteq \mathbb{R}^n$. Тогда для любых $x, y \in X$ таких, что $\|y - x\|_{f_{xx}(x)} < 1$, имеет место неравенство*

$$f(y) \le f(x) + \langle f_x(x), y - x\rangle + \omega^*\left(\|y - x\|_{f_{xx}(x)} \right), \qquad (9.2.14)$$

где $\omega^*(\tau) = -\tau - \ln(1 - \tau)$.

Доказательство этой леммы проводится примерно по той же самой схеме, что и доказательство предыдущей леммы. ■

Упражнение 14. *Докажите лемму 9.2.2 с помощью левого неравенства (9.2.10).*

Сделаем два замечания, касающиеся функций и величин, входящих в неравенства (9.2.8) и (9.2.14).

Замечание 1. Функции $\omega(t)$ и $\omega^*(\tau)$ являются строго выпуклыми соответственно при $t > -1$ и $\tau < 1$. В этом легко убедиться с помощью критерия второго порядка, поскольку

$$\frac{d^2}{dt^2}\omega(t) = \frac{1}{(1+t)^2} > 0, \quad \frac{d^2}{d\tau^2}\omega^*(\tau) = \frac{1}{(1-\tau)^2} > 0.$$

Более того, они являются сопряженными друг к другу. В самом деле, по определению сопряженной функции

$$\omega^*_{T_+}(\tau) = \sup_{t \in T_+}[t\tau - \omega(t)],$$

где $T_+ = (-1, +\infty)$. Продифференцировав функцию $\varphi_1^\tau(t) = t\tau - \omega(t)$ по t и приравняв производную нулю, получаем, что максимум по t достигается в точке $t^*(\tau) = \frac{\tau}{1-\tau}$. После подстановки данной точки в функцию $\varphi_1^\tau(t)$ приходим к

$$\omega^*_{T_+}(\tau) = \varphi_1^\tau(t^*(\tau)) = \omega^*(\tau).$$

Данная функция определена на множестве $T_- = (-\infty, 1)$.

Если строить к функции $\omega^*(\tau)$ ее сопряженную функцию, т.е. вторую сопряженную функцию $\omega^{**}_{T_-}(t)$, то, в силу выпуклости, функция $\omega^{**}_{T_-}(t)$ совпадет с исходной функцией $\omega(t)$, определенной на множестве T_+. Действительно, соответствующая точка $\tau^*(t)$, в которой достигается максимум функции $\varphi_2^t(\tau) = \tau t - \omega^*(\tau)$, равна $\tau^*(t) = \frac{t}{(1+t)}$. Поэтому

$$\omega^{**}_{T_-}(t) = \sup_{\tau \in T_-} \varphi_2^t(\tau^*(t)) = \omega(t), \qquad t \in T_+.$$

Обратим внимание, что при $0 \le \tau < 1$ и $t \ge 0$

$$\sup_{t \in T_+} \varphi_1^\tau(t) = \sup_{t \ge 0} \varphi_1^\tau(\tau), \quad \sup_{\tau \in T_-} \varphi_2^t(\tau) = \sup_{\tau \in [0,1)} \varphi_2^t(\tau). \tag{9.2.15}$$

Замечание 2. Введем в рассмотрение эллипсоид

$$W(x;r) = \|y \in \mathbb{R}^n : \ \|y - x\|_{f_{xx}(x)} < r\}.$$

Такой эллипсоид называется *эллипсоидом Дикина функции* $f(x)$ в точке x (не путать с эллипсоидом (8.2.7)). Имеет место включение $W(x;1) \subseteq X$, справедливое для любого $x \in X$. Это следует из того, что область определения функции $\phi(\tau)$ вида (9.2.9), как уже отмечалось, содержит интервал $(-\phi(0), \phi(0))$. Но $\psi(0) = \|y - x\|^{-1}_{f_{xx}(x)}$. Поэтому, беря $0 \le \alpha < \psi(0)$, получаем, что $\|\alpha(y - x)\|_{f_{xx}(x)} < 1$. Данное неравенство означает принадлежность y множеству X. Ниже нам потребуется подмножество эллипсоида $W(x;1)$, определяемое как

$$W_\tau(x;1) = \left\{y \in W^0(x,1) : \ \|y - x\|_{f_{xx}(x)} = \tau\right\},$$

где $0 \le \tau < 1$ — фиксированный параметр.

Лемма 9.2.3. *Пусть $f(x)$ — самосогласованная функция, определенная на множестве $X \subseteq \mathbb{R}^n$. Тогда для любых $x, y \in X$ таких, что $\|f_x(y) - f_x(x)\|_{f_{xx}^{-1}(y)} < 0.5$, справедливы неравенства*

$$f(y) \ge f(x) + \langle f_x(x), y - x\rangle + \omega\left(\ \|f_x(y) - f_x(x)\|_{f_{xx}^{-1}(y)} \ \right), \tag{9.2.16}$$

$$f(y) \le f(x) + \langle f_x(x), y - x\rangle + \omega^*\left(\ \|f_x(y) - f_x(x)\|_{f_{xx}^{-1}(y)} \ \right). \tag{9.2.17}$$

Доказательство. Докажем неравенство (9.2.16). Для этого зафиксируем некоторое $x \in X$ и рассмотрим функцию

$$\phi(z) = f(z) - \langle f_x(x), z\rangle,$$

определенную на множестве X. Эта функция является самосогласованной и $\phi_z(x) = 0_n$, т.е. x есть точка минимума данной выпуклой функции. Имеем

$$f(x) - \langle f_x(x), x\rangle = \phi(x) = \min_{z \in X} \phi(z). \tag{9.2.18}$$

Кроме того, матрица вторых производных функции $\phi(z)$ в любой точке $z \in X$ совпадает с матрицей вторых производных функции $f(z)$ в этой точке.

Оценим теперь минимальное значение функции $\phi(z)$ с помощью утверждения леммы 9.2.2. Учитывая неравенства (9.2.14) и (9.2.7), а также совпадение вторых производных у функций $\phi(z)$ и $f(z)$, получаем

$$\begin{aligned}
&\min_{z\in X}\phi(z) \le \min_{z\in X}\left[\phi(y)+\langle\phi_z(y),z-y\rangle+\omega^*\left(\|z-y\|_{f_{xx}(y)}\right)\right] = \\
&\quad = \phi(y)+\min_{z\in X}\left[\langle\phi_z(y),z-y\rangle+\omega^*\left(\|z-y\|_{f_{xx}(y)}\right)\right] \le \\
&\quad \le \phi(y)+\min_{z\in W(y;1)}\left[\langle\phi_z(y),z-y\rangle+\omega^*\left(\|z-y\|_{f_{xx}(y)}\right)\right] = \\
&\quad = \phi(y)+\min_{0\le\tau<1}\min_{z\in W_\tau(y;1)}\left[\langle\phi_z(y),z-y\rangle+\omega^*\left(\|z-y\|_{f_{xx}(y)}\right)\right].
\end{aligned} \tag{9.2.19}$$

Примем теперь во внимание, что нормы $\|\cdot\|_{f_{xx}(y)}$ и $\|\cdot\|_{f_{xx}^{-1}(y)}$ двойственные, и для них выполняется неравенство (9.2.7) . С учетом того, что $\|z-y\|_{f_{xx}(y)} = \tau$ при $z \in W_\tau(y;1)$, вычисляем внутренний минимум:

$$\begin{aligned}
&\min_{z\in W_\tau(y;1)}\left[\langle\phi_z(y),z-y\rangle+\omega^*\left(\|z-y\|_{f_{xx}(y)}\right)\right] = \\
&\quad = \min_{z\in W_\tau(y;1)}\langle\phi_z(y),z-y\rangle+\omega^*(\tau) = \\
&\quad = -\tau\|\phi_z(y)\|_{f_{xx}^{-1}(y)}+\omega^*(\tau).
\end{aligned} \tag{9.2.20}$$

Используя далее определение сопряженной функции и правое равенство (9.2.15), получаем

$$\min_{0\le\tau<1}\left[-\tau\|\phi_z(y)\|_{f_{xx}^{-1}(y)}+\omega^*(\tau)\right] = -\omega\left(\|\phi_z(y)\|_{f_{xx}^{-1}(y)}\right). \tag{9.2.21}$$

Остается только учесть, что $\phi_z(y) = f_x(y) - f_x(x)$. Тогда из (9.2.19)–(9.2.21) приходим к (9.2.16).

Неравенство (9.2.17) доказывается аналогично с применением утверждения леммы 9.2.1. ■

Теорема 9.2.1. *Пусть выполнены предположения леммы 9.2.3. Тогда*

$$\omega\left(\delta_f(x)\right) \le f(x) - f_* \le \omega^*\left(\delta_f(x)\right), \tag{9.2.22}$$

где $\delta_f(x) = \|f_x(x)\|_{f_{xx}^{-1}(x)}$.

Доказательство. Для доказательства достаточно воспользоваться неравенствами (9.2.16) и (9.2.17). Если положить в них $y = x$, $x = x_*$ и учесть, что $f_x(x_*) = 0_n$, то приходим к (9.2.22). ■

Результат теоремы 9.2.1 весьма примечательный. Он показывает, что в случае самосогласованных функций можно оценивать отклонение по функционалу через ньютоновское убывание, т.е. через локальную норму градиента. Таким образом, для того, чтобы найти оценку

уменьшения невязки по функционалу за одну итерацию метода Ньютона, достаточно вычислить ньютоновское убывание.

9.3. Ньютоновские методы в выпуклой оптимизации

Пусть выпуклая самосогласованная функция $f(x)$ определена на открытом выпуклом множестве X, не содержащем прямых линий. Пусть, кроме того, существует точка $x \in X$ такая, что $\delta_f(x) < 1$. Тогда, как можно показать, у задачи

$$\min_{x \in X} f(x) \tag{9.3.1}$$

обязательно имеется решение, причем единственное.

Для решения такой задачи может быть применен метод Ньютона, состоящий из двух вариантов. Первый вариант — это стандартный классический метод Ньютона с единичным шагом. Второй вариант — это *демпфированный метод* Ньютона, в котором итерации выполняются согласно следующей рекуррентной схеме:

$$x_{k+1} = x_k - \alpha_k f''^{-1}_{xx}(x_k) f'_x(x_k), \quad \alpha_k = (1 + \delta_f(x_k))^{-1}. \tag{9.3.2}$$

Таким образом, от рассмотренного нами ранее демпфированного метода Ньютона здесь шаг определяется через ньютоновское убывание.

Рассмотрим сначала классический метод Ньютона с единичным шагом для минимизации самосогласованной функции $f(x)$ и оценим его скорость сходимости. Для этого нам потребуется следующий вспомогательный результат, который приведем без доказательства. Напомним, знак $\succeq$ обозначает порядок Левнера на множестве симметричных матриц порядка n. По определению, $A \succeq B$, если матрица $A - B$ положительно полуопределена.

Утверждение 9.3.1. *Пусть функция $f(x)$, определенная на открытом множестве $X \subseteq \mathbb{R}^n$, является самосогласованной. Тогда для любых $x \in X$, $s \in \mathbb{R}^n$ и $0 \le \alpha \le 1$*

$$c_1 f''_{xx}(x) \;\succeq\; f''_{xx}(x + \alpha s) \;\succeq\; c_2 f''_{xx}(x), \tag{9.3.3}$$

если только $x + \alpha s \in X$. Здесь

$$c_1 = \frac{1}{\left(1 - \alpha \|s\|_{f''_{xx}(x)}\right)^2}, \quad c_2 = \frac{1}{c_1} = \left(1 - \alpha \|s\|_{f''_{xx}(x)}\right)^2.$$

Замечание. *Так как обе матрицы $f_{xx}(x)$ и $f_{xx}(x+\alpha s)$ в (9.3.3) симметричные и положительно определенные, то неравенства (9.3.3) сохраняются, если перейти от этих матриц к обратным.*

Теорема 9.3.1. *Пусть $f(x)$ — самосогласованная функция, определенная на открытом множестве X. Тогда если $x \in X$ и $\delta_f(x) < 1$, то в точке $\bar{x} = x - f_{xx}^{-1}(x) f_x(x)$ для ньютоновского убывания справедлива оценка*

$$\delta_f(\bar{x}) \le \left(\frac{\delta_f(x)}{1 - \delta_f(x)} \right)^2. \tag{9.3.4}$$

Доказательство. Обозначим для сокращения записи: $s = \bar{x} - x$. Имеем: $s = -f_{xx}^{-1}(x) f_x(x)$ и

$$\|s\|_{f_{xx}(x)} = \|f_x(x)\|_{f_{xx}^{-1}(x)} = \delta_f(x) < 1.$$

Но единичный эллипсоид Дикина $W(x;1)$, как уже отмечалось, принадлежит множеству X. Поэтому $\bar{x} \in X$, и в силу утверждения 9.3.1 и замечания к нему

$$\langle f_x(\bar{x}), f_{xx}^{-1}(\bar{x}) f_x(\bar{x}) \rangle^{\frac{1}{2}} \le \left(1 - \|s\|_{f_{xx}(x)}\right)^{-1} \langle f_x(\bar{x}), f_{xx}^{-1}(x) f_x(\bar{x}) \rangle^{\frac{1}{2}}.$$

Данное неравенство переписывается как

$$\delta_f(\bar{x}) \le (1 - \delta_f(x))^{-1} \|f_x(\bar{x})\|_{f_{xx}^{-1}(x)}. \tag{9.3.5}$$

Оценим теперь $\|f_x(\bar{x})\|_{f_{xx}^{-1}(x)}$. Так как $f_x(x) + f_{xx}(x)s = 0_n$, то

$$\begin{array}{rcl} f_x(\bar{x}) & = & f_x(x) + \int_0^1 f_{xx}(x+\alpha s) s \, d\alpha = \\ & = & f_x(x) + f_{xx}(x)s + Gs = Gs, \end{array} \tag{9.3.6}$$

где $G = \int_0^1 [f_{xx}(x+\alpha s) - f_{xx}(x)] d\alpha$. Из (9.3.6), если обозначить через H матрицу $f_{xx}^{-\frac{1}{2}}(x) G f_{xx}^{-\frac{1}{2}}(x)$, следует, что

$$\begin{array}{l} \|f_x(\bar{x})\|_{f_{xx}^{-1}(x)} = \langle Gs, f_{xx}^{-1}(x) Gs \rangle^{\frac{1}{2}} = \\ \quad = \langle H f_{xx}^{-\frac{1}{2}}(x) f_x(x), H f_{xx}^{-\frac{1}{2}}(x) f_x(x) \rangle^{\frac{1}{2}} = \|H f_{xx}^{-\frac{1}{2}}(x) f_x(x)\|_2 \ \le \\ \quad \le \|H\|_2 \|f_{xx}^{-\frac{1}{2}}(x) f_x(x)\|_2 = \|H\|_2 \|f_x(x)\|_{f_{xx}^{-1}(x)} = \|H\|_2 \delta_f(x). \end{array} \tag{9.3.7}$$

Здесь $\|H\|_2$ — матричная норма, подчиненная евклидовой норме в пространстве $\mathbb{R}^n$ (спектральная норма).

Пусть $h = \|s\|_{f_{xx}(x)} = \delta_f(x)$. В силу правого неравенства (9.3.3)

$$\begin{aligned} G &= \int_0^1 [f_{xx}(x+\alpha s) - f_{xx}(x)]d\alpha \succeq \left[\int_0^1 (1-\alpha h)^2\, d\alpha - 1\right] f_{xx}(x) = \\ &= \left(\tfrac{1}{3}h^2 - h\right) f_{xx}(x). \end{aligned}$$

С другой стороны, на основании левого неравенства (9.3.3) выполняется

$$G \preceq \left[\int_0^1 (1-\alpha h)^{-2} d\alpha - 1\right] f_{xx}(x) = \frac{h}{1-h} f_{xx}(x).$$

Поэтому

$$\|H\|_2 \leq \max\left\{-h + \frac{1}{3}h^2, \frac{h}{1-h}\right\} = \frac{h}{1-h}.$$

Отсюда, а также из (9.3.5) и (9.3.7) приходим к (9.3.4). ■

Согласно оценке (9.3.4) существует окрестность решения x_* задачи (9.3.1) такая, что, стартуя из данной окрестности, траектория ее не покидает, и ньютоновское убывание уменьшается с квадратичной скоростью. Действительно, для этого надо потребовать, чтобы

$$\frac{\delta_f(x)}{(1-\delta_f(x))^2} < 1,$$

что действительно имеет место, когда $\delta_f(x) < \bar{\delta} = \frac{3-\sqrt{5}}{2}$.

Обратимся теперь к демпфированному методу Ньютона.

Теорема 9.3.2. *В демпфированном методе Ньютона на каждой итерации выполняется неравенство:* $f(x_{k+1}) \leq f(x_k) - \omega\left(\delta_f(x_k)\right)$.

Доказательство. Согласно лемме 9.2.2:

$$f(x_{k+1}) \leq f(x_k) + \langle f_x(x_k), x_{k+1} - x_k\rangle + \omega^*\left(\|x_{k+1} - x_k\|_{f_{xx}(x_k)}\right). \quad (9.3.8)$$

Но

$$\langle f_x(x_k), x_{k+1} - x_k\rangle = -\frac{\delta_f^2(x_k)}{1+\delta_f(x_k)}, \quad \|x_{k+1} - x_k\|_{f_{xx}(x_k)} = \frac{\delta_f(x_k)}{1+\delta_f(x_k)}. \quad (9.3.9)$$

Кроме того, поскольку $\omega'(t) = \frac{t}{1+t}$, получаем из (9.3.8), (9.3.9)

$$f(x_{k+1}) \leq f(x_k) - \delta_f(x_k)\omega'(\delta_f(x_k)) + \omega^*(\omega'(\delta_f(x_k)) = f(x_k) - \omega(\delta_f(x_k)).$$

Последнее равенство следует из соотношения

$$t\omega'(t) = \omega(t) + \omega^*\left(\omega'(t)\right), \quad (9.3.10)$$

имеющего место для взаимосопряженных функций $\omega(t)$ и $\omega^*(t)$. ■

Упражнение 15. *Убедитесь в справедливости равенства (9.3.10).*

Собственно метод Ньютона для минимизации самосогласованной функции $f(x)$ состоит из двух этапов.

Этап 1. Выбирается $\beta \in (0, \bar{\delta})$ и до тех пор, пока $\delta_f(x_k) \geq \beta$, применяется демпфированный метод Ньютона.

Этап 2. Когда на некоторой k-й итерации выполняется неравенство $\delta_f(x_k) < \beta$, то в дальнейшем применяется классический метод Ньютона с единичным шагом. Неравенство (9.3.4) гарантирует, что на всех последующих итерациях мы не покинем области, где $\delta_f(x) < \beta$. В самом деле:

$$\delta_f(x_{k+1}) \leq \left(\frac{\delta_f(x_k)}{1 - \delta_f(x_k)} \right)^2 \leq \frac{\beta}{(1-\beta)^2} \delta_f(x_k) < \delta_f(x_k).$$

Понятно, что не любая выпуклая функция, минимум которой требуется найти, может оказаться самосогласованной. Но важность самосогласованной функции для выпуклой минимизации заключается в том, что нам часто приходится строить вспомогательные функции, например, в методах последовательной безусловной минимизации, и здесь при построении таких вспомогательных функций целесообразно использовать самосогласованные функции.

Среди таких вспомогательных функций особый интерес вызывают барьерные функции, применяемые в методах внутренних штрафных функций. Дело в том, что самосогласованные функции согласно своему определению обладают барьерными свойствами, а именно: при подходе к границе своей области определения их значения бесконечно увеличиваются. Если целевая функция в задаче является самосогласованной, то вспомогательная функция в методе внутренних штрафных функций оказывается самосогласованной. Для минимизации такой функции можно использовать метод Ньютона, который, как было показано, является весьма эффективным. Более того, можно сохранить и квадратичную скорость сходимости метода Ньютона для всей задачи условной минимизации (а не только для вспомогательной задачи при фиксированном значении параметра внутреннего штрафа). Делается это за счет специальной политики пересчета коэффициента штрафа. Метод называют *барьерным*, и он относится к классу так называемых *методов отслеживания пути*. Теория самосогласованных

функций дает возможность оценить сложность задач выпуклого программирования для таких методов.

Мы уже рассматривали прямодвойственный метод центрального пути для решения задачи линейного программирования. Однако концепция центрального пути с привлечением метода Ньютона, позволяет построить эффективные методы решения и в нелинейном случае. Рассмотрим задачу условной минимизации:

$$f_* = \min_{x \in X} f(x), \quad X = \left\{ x \in \mathbb{R}^n : \; g^i(x) \leq 0, \; 1 \leq i \leq m \right\}, \qquad (9.3.11)$$

где $f(x)$ и $g^i(x)$, $1 \leq i \leq m$, выпуклые дважды дифференцируемые функции. Предположим, что допустимое множество в (9.3.11) имеет непустую внутренность — множество

$$X_0 = \left\{ x \in \mathbb{R}^n : \; g^i(x) < 0, \; 1 \leq i \leq m \right\}.$$

Составим для задачи (9.3.11) логарифмическую барьерную функцию:

$$P(x,t) = f(x) - t \sum_{i=1}^{m} \ln \left(-g^i(x) \right), \quad t > 0.$$

Так как функция $-\ln(-y)$ является выпуклой и неубывающей при $y < 0$, то, согласно теореме о суперпозиции двух выпуклых функций, функция $P(x,t)$ выпукла по x на X_0. *Центральным путем* в данном случае называется совокупность точек $x(t)$, определенных при $t > 0$, и таких, что

$$x(t) \in X(t) = \operatorname{Arg} \min_{x \in X_0} P(x,t).$$

Точки $x(t)$, лежащие на центральном пути, называются *центральными точками*. Они допустимы и удовлетворяют системе уравнений

$$f_x(x) + t \sum_{i=1}^{m} \frac{1}{-g^i(x)} g^i_x(x) = 0_n. \qquad (9.3.12)$$

Обозначим: $v^i(t) = -t(g^i(x(t))^{-1}$, $1 \leq i \leq m$. Имеем $v^i(t) > 0$. Если подставить данные множители в функцию Лагранжа

$$L(x,v) = f(x) + \langle v, g(x) \rangle, \quad v \in \mathbb{R}^m_+,$$

составленную для задачи (9.3.11), то из (9.3.12) следует, что

$$L_x(x(t), v(t)) = f_x(x(t)) + \sum_{i=1}^{m} v^i(t) g^i_x(x(t) = 0_n.$$

Это означает, что точка $v(t)$ для любого $t > 0$ является допустимой в задаче, двойственной к (9.3.11).

Если вычислить значение целевой функции в двойственной задаче $\phi(v) = \inf_x L(x, v)$ в этой точке, то данное значение равно:

$$\phi(v(t)) = L(x(t), v(t)) = f(x(t)) - tm.$$

Таким образом, для разности значений целевой функции $f(x)$ в прямой задаче (9.3.11) и целевой функции $\phi(v)$ в двойственной задаче соответственно в точках $x(t)$ и $v(t)$ получаем: $f(x(t)) - \phi(v(t)) = mt$. Отсюда приходим к оценке: $f(x(t)) - f_* \leq mt$, из которой следует, что $f(x(t)) \to f_*$ при $t \downarrow 0$, что полностью согласуется с рассмотренными ранее свойствами методов внутренних штрафных функций.

Рассмотрим барьерный метод решения задачи (9.3.11), заключающийся в решении с помощью метода Ньютона серии задач минимизации функции $P(x, t_k)$ при последовательно уменьшающихся значениях коэффициента t_k.

Пусть задано $t_0 > 0$ и параметр $0 < \theta < 1$. Предположим также, что известна начальная стартовая точка $\bar{x}_0 \in X_0$, достаточно близкая к центральной точке $x(t_0)$.

На каждой k-й итерации, начиная с известной точки $\bar{x}_k$, делается несколько итераций методом Ньютона с целью найти приближенное решение задачи минимизации функции $P(x, t_k)$, т.е. найти точку x_k, близкую к центральной точке $x(t_k)$. После этого переходим к следующей $(k+1)$-й итерации, полагая $\bar{x}_{k+1} = x_k$ и $t_{k+1} = \theta t_k$.

Итерации, которые выполняет метод Ньютона по вычислению точки x_k, называются *внутренними*. Напротив, итерации, в которых меняется коэффициент штрафа t_k, называются *внешними*. Вопрос заключается в том, сколько следует делать внутренних итераций на каждом шаге или, другими словами, с какой точностью минимизировать функцию $P(x, t_k)$? Понятно, что при больших значениях t_k нет особого смысла добиваться очень точного решения вспомогательной задачи. Однако, с другой стороны, при приближении к решению задачи (9.3.11) целесообразно увеличивать точность решения этих задач.

Еще один вопрос возникает в связи с выбором параметра θ. Слишком близкое к единице значение θ приводит к тому, что коэффициент внутреннего штрафа t_k изменяется мало и, как следствие, стартовая точка $\bar{x}_{k+1}$ на последующей $(k+1)$-й итерации оказывается достаточно хорошей, что приводит к малому числу внутренних итераций на этой внешней итерации. Однако, разумеется, общее число внешних итераций в этом случае увеличивается. Правильный выбор здесь достигается компромиссом между этими двумя факторами.

Глава 10

Методы многокритериальной оптимизации

10.1. Основные подходы к нахождению решений

Рассмотрим задачу многокритериальной оптимизации, в которой требуется найти

$$\min_{x \in X} f(x), \tag{10.1.1}$$

где $X \subseteq \mathbb{R}^n$ — допустимое множество, $f(x)$ — векторный критерий, состоящий из r критериев $f^1(x), \dots f^r(x)$. В таких задачах, как об этом говорилось в [33] в главе 8, существует целое параметрическое множество решений в критериальном пространстве, причем каждому такому решению в критериальном пространстве соответствует свое множество решений в пространстве исходных переменных (в пространстве переменных x). В настоящее время разработано достаточно много методик и подходов к выбору конкретных решений из всей совокупности решений, а также к обоснованию разумности сделанного выбора. Весьма условно их можно разбить на два направления.

Пусть $F = \left\{ f \in \mathbb{R}^r : \ f = f(x), \, x \in X \right\}$ — множество достижимых оценок в критериальном пространстве $\mathbb{R}^r$. В рамках первого из этих направлений главное внимание уделяется нахождению одного подходящего решения из множества оптимальных по Парето оценок $F_*^P \subseteq F$ или из множества слабо оптимальных по Парето (оптимальных по

Слейтеру) оценок $F_*^S \subseteq F$. При этом считается, что выбор конкретной оценки осуществляется *лицом принимающим решение* (ЛПР), у которого на самом деле имеются некоторые «скрытые» предпочтения среди критериев, например, их относительная важность. Руководствуясь данными предпочтениями ЛПР и принимает решение. Формально это описывается наличием некоторого *бинарного отношения предпочтения* и нахождением решений, которые являются *недоминируемыми* с точки зрения этих отношений, т.е. неулучшаемыми. Для выявления этих отношений предпочтения могут применяться, в частности, *интерактивные процедуры*.

Конкретные решения из множества оптимальных оценок могут быть получены также и с применением *метода целевой точки* или *метода главного критерия*. Согласно первому методу выделяется некоторая точка в критериальном пространстве, не принадлежащая множеству F, и ищется точка из F, ближайшая к выделенной точке в заданной метрике. В методе главного критерия все критерии ранжируются по степени важности и далее находится множество точек, которые доставляют оптимум наиболее важному критерию, затем на найденном множестве определяется его подмножество, доставляющее минимум следующему по важности критерию и т.д.

Второе направление в решении задач многокритериальной оптимизации заключается в предъявлении ЛПР всего множества оптимальных оценок, и далее ЛПР сам решает, какая из этих оценок его более устраивает. Имеются разные подходы к нахождению множеств оптимальных оценок или, по крайней мере, их аппроксимации конечным количеством точек. Одним из наиболее простых является подход, основанный на сведении задачи многокритериальной оптимизации к параметрической задаче нелинейного программирования. Осуществляется подобный переход путем *скаляризации* векторного критерия, т.е. свертывания всех частных критериев в единый критерий. Например, использование *линейной свертки* критериев приводит к параметрической скалярной функции:

$$f(x;u) = \sum_{i=1}^{r} u^i f^i(x), \quad u \in \Lambda^r, \tag{10.1.2}$$

где $\Lambda^r = \left\{u \in \mathbb{R}_+^r : \sum_{i=1}^r u^i = 1\right\}$ — единичный (вероятностный) симплекс. Тогда задача нахождения разных парето-оптимальных оценок сводится к минимизации получившейся функции $f(x;u)$ на X при всевозможных значениях $u \in \Lambda^r$, т.е. к задаче

$$\min_{x \in X} f(x;u).$$

Если все критерии $f^i(x)$ неотрицательны на X, то вместо (10.1.2) можно воспользоваться следующей функцией (сверткой Гермейера [18]):

$$f(x;u) = \max_{1\le i\le r} u^i f^i(x), \quad u \in \Lambda^r.$$

Данный способ сворачивания критериев применим также, когда задача многокритериальной оптимизации (10.1.1) не является выпуклой.

Возможно, однако, и непосредственное обобщение рассмотренных ранее методов условной минимизации для решения задач многокритериальной оптимизации.

Ниже в качестве примеров указанных обобщений рассматриваются обобщения метода параметризации целевой функции, метода возможных направлений и метода модифицированной функции Лагранжа. Разумеется, все эти обобщения проводятся таким путем, чтобы в частном случае, когда в задаче имеется один критерий, они переходили в известные численные схемы. В этих обобщениях также присутствуют оценки оптимальных значений критериев, но они не фиксированы, а меняются в ходе вычислительного процесса. Поэтому условно данные обобщения можно назвать *методами подвижной целевой точки.*

10.2. Метод внешних центров для многокритериальной оптимизации

Рассмотрим обобщение метода внешних центров (метода параметризации целевой функции) на примере задачи многокритериальной минимизации (10.1.1), в которой допустимое множество X задается с помощью ограничений типа неравенства

$$\min_{x\in X} f(x), \quad X = \left\{x \in \mathbb{R}^n : g(x) \le 0_m\right\}. \tag{10.2.1}$$

Компонентами вектор-функции $g(x)$ являются функции $g^j(x)$, $j \in J^m$.

Предположим, что нам известен вектор $\eta \in \mathbb{R}^r$, который не принадлежит множеству $F_+ = F + \mathbb{R}^r_+$. Вектор η по смыслу полностью аналогичен оценке снизу оптимального значения целевой функции, которая использовалась в методе внешних центров для задач с одним критерием. Составим вспомогательную функцию:

$$M(x,\eta) = \sqrt{\sum_{i=1}^{r} (f^i(x) - \eta^i)_+^2 + (B(g(x)))_+^2},$$

где $B(g)$ — внешняя свертывающая функция. Обозначим через $X(\eta)$ множество минимумов функции $M(x,\eta)$ на всем пространстве $\mathbb{R}^n$ при фиксированном значении вектора η, т.е. множество решений задачи

$$\min_{x\in R^n} M(x,\eta). \tag{10.2.2}$$

Опишем теперь сам **алгоритм метода параметризации целевых функций**, обобщенный для решения задачи многокритериальной минимизации.

Начальная итерация. Пусть задан вектор $\eta_0 \notin F_+$ и направление $e \in \mathbb{R}$ такое, что $e > 0_r$ и $\|e\|_2 = 1$. Полагаем $k = 0$.

Общая k-я итерация

Шаг 1. Решаем задачу (10.2.2) при $\eta = \eta_k$ и находим точку x_k такую, что $x_k \in X(\eta_k)$. Если $M(x_k,\eta_k) = 0$, то останавливаемся. Если нет, то идем на шаг 2.

Шаг 2. Пересчитываем вектор η_k по формуле

$$\eta_{k+1} = \eta_k + \lambda_k e, \tag{10.2.3}$$

где $\lambda_k = M(x_k,\eta_k)$. Полагаем $k := k+1$ и идем на шаг 1.

Убедимся, что при таком способе пересчета векторов η_k ни один из них не попадает в множество $\mathrm{int}F_+$.

Лемма 10.2.1. *Пусть* $\eta_k \notin F_+$. *Тогда* $\eta_{k+1} \notin \mathrm{int}\, F_+$.

Доказательство. Предположим противное, что $\eta_{k+1} \in \mathrm{int}F_+$. Тогда обязательно найдется точка $\bar{x} \in X$ такая, что для векторной оценки $\bar{f} = f(\bar{x})$ выполняется $\bar{f} < \eta_{k+1}$. Имеем для данной точки $\bar{x}$ в силу ее допустимости и правила пересчета (10.2.3):

$$\begin{aligned} M(\bar{x},\eta_k) &= \sqrt{\sum_{i=1}^r \left(f^i(\bar{x}) - \eta_k^i\right)_+^2 + (B(g(\bar{x})))_+^2} = \\ &= \sqrt{\sum_{i=1}^r \left(f^i(\bar{x}) - \eta_k^i\right)_+^2} = \|(f(\bar{x}) - \eta_k)_+\|_2 = \\ &= \|(f(\bar{x}) - \eta_{k+1} + M(x_k,\eta_k)e)_+\|_2 < \\ &< \|M(x_k,\eta_k)e\|_2 = M(x_k,\eta_k)\|e\|_2 = M(x_k,\eta_k). \end{aligned}$$

Мы пришли к противоречию с тем, что $x_k \in X(\eta_k)$. Таким образом, $\eta_{k+1} \notin \mathrm{int}F_+$. ■

Так как на каждой итерации $M(x_k,\eta_k) \geq 0$ и все компоненты вектора e строго положительны, то $\{\eta_k\}$ является последовательностью векторов со все увеличивающимися компонентами, расположенными

на луче, который выходит из точки $\eta_0 \notin F_+$ и имеет направление e. Из-за того, что ни один из векторов η_k не принадлежит $\mathrm{int}F_+$, следует ограниченность этой последовательности сверху и, значит, существование предела $\lim_{k\to\infty}\eta_k = \eta_* \notin \mathrm{int}F_+$. Очевидно, что данный предельный вектор η_* также принадлежит указанному лучу.

Убедимся теперь, что в результате работы алгоритма получаются оптимальные по Слейтеру решения задачи (10.2.1).

Теорема 10.2.1. *Пусть последовательность $\{x_k\}$, порождаемая алгоритмом метода параметризации целевых функций, принадлежит компактному множеству $\tilde{X} \subset \mathbb{R}^n$. Тогда любая ее предельная точка является оптимальной по Слейтеру, а предельный вектор η_* принадлежит границе множества F_+.*

Доказательство. Обозначим $\lambda_* = \inf\{\lambda \geq 0 : \eta_0 + \lambda e \in F_+\}$. Какое бы $\eta_0 \notin F_+$ и направление e ни взять, луч $\{\eta = \eta_0 + \lambda e : \ \lambda \geq 0\}$ обязательно пересечет множество F_+, поэтому величина λ_* конечна. Имеет место представление

$$\eta_{k+1} = \eta_0 + \sum_{s=0}^{k} \lambda_s e, \quad \lambda_k = M(x_k, \eta_k) \geq 0. \tag{10.2.4}$$

Поскольку согласно лемме 10.2.1 все $\eta_k \notin \mathrm{int}F_+$, то на основании (10.2.4) $\sum_{s=0}^{k}\lambda_s \leq \lambda_*$ для всех $k \geq 0$ и, следовательно, ряд $\sum_{k=0}^{\infty}\lambda_k$ сходится. Отсюда приходим к выводу, что

$$\lim_{k\to\infty} M(x_k, \eta_k) = \lim_{k\to\infty} \lambda_k = 0. \tag{10.2.5}$$

Пусть x_* — предельная точка последовательности $\{x_k\}$, т.е. для некоторой подпоследовательности $\{x_{k_l}\}$ выполняется: $\lim_{l\to\infty} x_{k_l} = x_*$. Согласно равенству (10.2.5) $M(x_*, \eta_*) = 0$, где $\eta_* = \lim_{k\to\infty}\eta_k$. Поэтому обязательно $x_* \in X$ и, кроме того, $f(x_*) - \eta_* \leq 0_r$. Но η_*, как уже отмечалось, не принадлежит множеству $\mathrm{int}F_+$. Отсюда приходим к выводу, что $x_* \in X_*^S$. Действительно, иначе для оценки $f_* = f(x_*) \in F$ можно было бы указать точку $\bar{x} \in X$ и оценку $\bar{f} = f(\bar{x}) \in F$ такие, что $\bar{f} < f_*$. Но тогда $\bar{f} < f_* \leq \eta_*$. Данное неравенство указывает на то, что справедливо включение $\eta_* \in \mathrm{int}F_+$. Мы пришли к противоречию. Таким образом, $x_* \in X$, а из неравенства $f(x_*) \leq \eta_*$ вытекает, что на самом деле предельный вектор η_* принадлежит границе множества F_+. ■

В предложенном обобщении метода параметризации целевой функции выбор конкретной оптимальной оценки f_* из множества всех оп-

тимальных оценок F_*^S сводился к выбору начального вектора η_0 и направления e. Разные $\eta_0 \notin \mathrm{int}F_+$ и разные $e > 0_r$ приводят к разным $f_* \in F_*^S$. Чтобы получить конечную аппроксимацию всего множества F_*^S, можно поступить следующим образом. Обратимся к так называемой *идеальной точке*:

$$f_{\min} = [f_{\min}^1, \dots f_{\min}^r], \qquad f_{\min}^i = \min_{x \in X} f^i(x), \quad 1 \le i \le r,$$

т.е. это такая точка в критериальном пространстве $\mathbb{R}^r$, каждая компонента которой является минимальным значением соответствующего отдельного критерия на допустимом множестве X. Выберем начальный вектор η_0 таким образом, чтобы он лежал строго «юго-западнее» идеальной точки. Другими словами, чтобы он удовлетворял неравенству $\eta_0 < f_{\min}$. Возьмем также «пучок единичных направлений» e_p, принадлежащих внутренности неотрицательного ортанта $\mathbb{R}_+^r$ и заполняющих этот неотрицательный ортант достаточно равномерно. Тогда, проводя вычисления с каждым из этих направлений (в том числе, используя параллельные вычисления), можно получить достаточно полную конечную аппроксимацию множества F_*^S. Разные лучи будут приводить к разным оптимальным решениям (см. рис. 10.1).

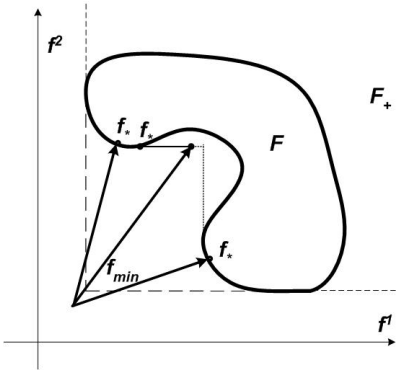

Рис. 10.1. Разные случаи пересечения лучами границы множества F_+

Отметим также, что даже если луч, выходящий из точки η_0 и имеющий направление e, пересекает «юго-западную» границу множества F_+ в точке $\eta_* = \eta_*(e)$, которая не принадлежит множеству оптимальных оценок F_*^S, тем не менее будет найдена оценка, входящая в F_*^S и ближайшая в некотором смысле к точке η_*. Если точка пересечения η_*

принадлежит F_*^S, то будет найдена оптимальная оценка, совпадающая с η_* (оба эти случая показаны на рис. 10.1).

10.3. Метод возможных направлений для многокритериальной оптимизации

Рассмотрим обобщение метода возможных направлений для решения задачи многокритериальной оптимизации (10.2.1). Относительно функций $f^i(x)$, $1 \le i \le r$, и $g^j(x)$, $1 \le j \le m$, задающих критерии и ограничения, предполагаем, что они непрерывно дифференцируемы и выпуклы на $\mathbb{R}^n$. Предполагаем также, что ограничения в задаче удовлетворяют условию регулярности Слейтера, т.е. множество $X_0 = \{x \in \mathbb{R}^n : \ g(x) < 0_m\}$ непусто.

Пусть $\eta \in \mathbb{R}^r$ и пусть $\phi(x, \eta)$ и $\psi(x)$ — функции, определяемые следующим образом:

$$\phi(x, \eta) = \max_{1 \le i \le r} \left[f^i(x) - \eta^i \right], \qquad \psi(x) = \max_{1 \le j \le m} g^j(x).$$

Из выпуклости вектор-функций $f(x)$ и $g(x)$ следует, что $\phi(x, \eta)$ и $\psi(x)$ — выпуклые относительно x функции на $\mathbb{R}^n$. Составим с помощью $\phi(x, \eta)$ и $\psi(x)$ вспомогательную функцию

$$M(x, \eta) = \max\{\phi(x, \eta), \psi(x)\}, \tag{10.3.1}$$

которая также выпукла по x. Через $X(\eta)$ обозначим множество

$$X(\eta) = \operatorname{Arg} \min_{x \in R^n} M(x, \eta).$$

Данное множество, если непустое, обязательно выпуклое.

Утверждение 10.3.1. *Пусть существуют $x_* \in \mathbb{R}^n$ и $\eta_* \in \mathbb{R}^r$ такие, что*

$$M(x_*, \eta_*) = 0, \qquad x_* \in X(\eta_*). \tag{10.3.2}$$

Тогда $x_* \in X_*^S$.

Доказательство. Из равенства $M(x_*, \eta_*) = 0$ вытекает неравенство $\psi(x_*) \le 0$. Поэтому $g^j(x_*) \le 0$ для всех $1 \le j \le m$. Следовательно, $x_* \in X$.

Проверим теперь, что $x_* \in X_*^S$. На основании (10.3.2) для любых $x \in X_0$ выполнено неравенство $M(x, \eta_*) \ge M(x_*, \eta_*) = 0$. Если предположить, что $\eta_* \ge f(x_*)$, то отсюда, поскольку $\psi(x) < 0$, получаем

$$0 \le M(x, \eta_*) = \phi(x, \eta_*) \le \max_{1 \le i \le r} [f^i(x) - f^i(x_*)]. \tag{10.3.3}$$

В силу непрерывности $f(x)$, неравенство (10.3.3) остается справедливым и для всех $x \in X$. Таким образом, $x_* \in X_*^S$.

Покажем, что случай иного отношения η_* и $f(x_*)$ невозможен. Действительно, предположение, что существует индекс $i_0 \in [1:r]$ такой, что $\eta_*^{i_0} < f^{i_0}(x_*)$, приводит к неравенству $M(x_*, \eta_*) \geq f^{i_0}(x_*) - \eta_*^{i_0} > 0$. Данное неравенство противоречит первому равенству в (10.3.2). ■

Согласно утверждению 10.3.1, мы можем найти слабо оптимальную по Парето точку $x_* \in X_*^S$, минимизируя функцию $M(x, \eta)$ по x. Разумеется, надо одновременно менять вектор η так, чтобы добиться равенства $M(x_*, \eta_*) = 0$.

Рассмотрим вспомогательную задачу по нахождению направления $s \in \mathbb{R}^n$, вдоль которого функция $M(x, \eta)$ уменьшается. Данная задача полностью аналогична той, которая строилась в методе возможных направлений для задачи с одним критерием. Только теперь приходится учитывать наличие нескольких критериев.

Пусть задано $\varepsilon > 0$. Обозначим через $J_\varepsilon^{(1)}(x, \eta)$ и $J_\varepsilon^{(2)}(x, \eta)$ индексные множества

$$J_\varepsilon^{(1)}(x, \eta) = \left\{ i \in J^r : \ f^i(x) - \eta^i \geq M(x, \eta) - \varepsilon \right\},$$

$$J_\varepsilon^{(2)}(x, \eta) = \left\{ j \in J^m : \ g^j(x) \geq M(x, \eta) - \varepsilon \right\}.$$

Вспомогательная задача является задачей линейного программирования и заключается в следующем: требуется найти

$$\begin{gathered} \min_{s,\sigma} \sigma, \\ \langle f_x^i(x), s \rangle \leq \sigma, \quad i \in J_\varepsilon^{(1)}(x, \eta), \\ \langle g_x^j(x), s \rangle \leq \sigma, \quad j \in J_\varepsilon^{(2)}(x, \eta), \\ \|s\|_\infty \leq 1. \end{gathered} \tag{10.3.4}$$

Решение задачи (10.3.4) в точке $[x, \eta] \in \mathbb{R}^{n+r}$ при заданном ε обозначим

$$q_\varepsilon(x, \eta) = [s_\varepsilon(x, \eta), \sigma_\varepsilon(x, \eta)].$$

В тех случаях, когда оно не единственное, $q_\varepsilon(x, \eta)$ будет обозначать произвольное решение задачи (10.3.4). Поскольку нулевая точка является допустимой в (10.3.4), обязательно $\sigma_\varepsilon(x, \eta) \leq 0$. Приведем ряд свойств, которым обладает решение $q_\varepsilon(x, \eta)$ задачи (10.3.4) при $\varepsilon = 0$.

Утверждение 10.3.2. *Пусть точка* $[x, \eta] \in \mathbb{R}^{n+r}$ *такова, что* $x \in X$ *и* $M(x, \eta) = 0$. *Тогда если* $x \notin X_*^S$, *то* $\sigma_0(x, \eta) < 0$.

Доказательство. Прежде всего отметим, что выполнение равенства $M(x,\eta)=0$ влечет выполнение неравенства $\eta \geq f(x)$. Кроме того, если $i \in J_0^{(1)}(x,\eta)$ или $j \in J_0^{(2)}(x,\eta)$, то $f^i(x)=\eta^i$ и $g^j(x)=0$ для этих индексов i и j.

Поскольку, по предположению, $x \notin X_*^S$, то найдется такая точка $x' \in X$, что $f(x') < f(x)$. Кроме того, так как выполнено условие регулярности ограничений, всегда x' можно выбрать таким образом, что $x' \in X_0$. Поэтому $f^i(x') < \eta^i$ для $i \in J_0^{(1)}(x,\eta)$ и $g^j(x') < 0$ для $j \in J_0^{(2)}(x,\eta)$.

Из выпуклости функций $f^i(x)$ вытекают неравенства

$$\langle f_x^i(x), x'-x\rangle \leq f^i(x')-f^i(x) < 0, \quad i \in J_0^{(1)}(x,\eta). \tag{10.3.5}$$

Точно так же из выпуклости функции $g^j(x)$ следует, что

$$\langle g_x^j(x), x'-x\rangle \leq g^j(x')-g^j(x) = g^j(x') < 0, \quad j \in J_0^{(2)}(x,\eta). \tag{10.3.6}$$

Взяв вектор $s'=\frac{1}{\lambda}(x'-x)$, где λ — максимальная по модулю компонента вектора $x'-x$, получим, что s' удовлетворяет последнему нормирующему ограничению в (10.3.4). Кроме того, согласно (10.3.5) и (10.3.6)

$$\langle f_x^i(x), s'\rangle < 0, \quad \langle g_x^j(x), s'\rangle < 0$$

для всех $i \in J_0^{(1)}(x,\eta)$ и $j \in J_0^{(2)}(x,\eta)$. Таким образом, обязательно $\sigma_0(x,\eta) < 0$. ■

Утверждение 10.3.3. *Пусть точка $[x,\eta] \in \mathbb{R}^{n+r}$ такова, что $x \notin X$ и $\eta \geq f(x)$. Тогда $\sigma_0(x,\eta) < 0$.*

Доказательство аналогично доказательству предыдущего утверждения. ■

Упражнение 16. *Докажите утверждение 10.3.3.*

Утверждение 10.3.4. *Пусть точка $[x_*,\eta_*] \in \mathbb{R}^{n+r}$ такова, что $M(x_*,\eta_*)=0$. Тогда если $\sigma_0(x_*,\eta_*)=0$, то $x_* \in X_*^S$.*

Доказательство. Из равенства $\sigma_0(x,\eta)=0$ следует, что для любого направления $s \in \mathbb{R}^n$ найдутся либо индекс $i_0 \in J_0^{(1)}(x,\eta)$, либо индекс $j_0 \in J_0^{(2)}(x,\eta)$ такие, что $\langle f_x^{i_0}(x), s\rangle \geq 0$ или $\langle g_x^{j_0}(x), s\rangle \geq 0$. Поэтому для производной функции $M(x,\eta)$ по направлению относительно первого аргумента выполняется

$$M'_x(x_*,\eta_*) = \max\left\{ \max_{i \in J_0^{(1)}(x,\eta)} \langle f_x^i(x_*), s\rangle, \max_{j \in J_0^{(2)}(x,\eta)} \langle g_x^j(x_*), s\rangle \right\} \geq 0,$$

причем направление $s \in \mathbb{R}^n$ произвольно. В силу выпуклости функций $f^i(x)$ и $g^j(x)$ это означает, что x_* — точка минимума функции $M(x, \eta_*)$ по x. Тогда, в силу утверждения 10.3.1, $x_* \in X_*^S$. ■

Согласно утверждениям 10.3.2 и 10.3.3 направления $s = s(x, \eta)$, получающиеся из решения задачи (10.3.4) при $\varepsilon = 0$, для которых $\sigma(x, \eta) < 0$, являются направлениями убывания функции $M(x, \eta)$ по x. При положительном значении параметра ε они оказываются направлениями ε-наискорейшего спуска.

Опишем теперь **алгоритм метода возможных направлений**.

Начальная итерация. Пусть задано $x_0 \in \mathbb{R}^n$ и $\eta_0 \in F_+$. Пусть, кроме того, выбраны начальное значение параметра $\varepsilon = \varepsilon_0 > 0$ и направление $e > 0_n$. Полагаем $k = 0$.

Общая k-я итерация

Шаг 1. Вычисляем

$$\eta_{k+1} = \eta_k - \beta(x_k, \eta_k, e)e, \tag{10.3.7}$$

где

$$\beta(x, \eta, e) = \min_{1 \le i \le r} \frac{\eta^i - f^i(x)}{e^i}, \tag{10.3.8}$$

e^i — i-я компонента вектора e.

Шаг 2. Решаем вспомогательную задачу (10.3.4) в точке $[x_k, \eta_{k+1}]$ при $\varepsilon = \varepsilon_k$ и определяем направление $s_k = s(x_k, \eta_{k+1})$, а также величину $\sigma_k = \sigma(x_k, \eta_{k+1})$.

Шаг 3. Проверяем выполнение неравенства $|\sigma(x_k, \eta_k)| > \varepsilon_k$. Если оно нарушается, то уменьшаем ε_k в два раза и идем на шаг 2.

Шаг 4. Находим новую точку x_{k+1} по формуле

$$x_{k+1} = x_k + \alpha_k s_k.$$

Шаг α_k находится путем дробления пополам некоторого начального шага α до тех пор, пока не будет выполнено условие

$$M(x_k + \alpha_k s_k, \eta_{k+1}) \le M(x_k, \eta_{k+1}) + \frac{\alpha_k \sigma_k}{2}. \tag{10.3.9}$$

Шаг 5. Полагаем $k := k + 1$ и идем на шаг 1.

В описанном алгоритме имеется процедура выбора шага α_k. Если наложить на функции $f(x)$ и $g(x)$, входящие в постановку задачи, некоторые дополнительные условия, то можно показать, что эта процедура является корректной, т.е. за конечное число делений пополам шаг α_k будет найден.

Для произвольного $x_0 \in \mathbb{R}^n$ определим множество

$$\tilde{X}(x_0) = \left\{x \in \mathbb{R}^n : \psi_+(x) \leq \psi_+(x_0)\right\},$$

где $\psi_+(x) = \max\{\psi(x), 0)\}$. В силу выпуклости функции $\psi_+(x)$, множество $\tilde{X}(x_0)$ всегда выпукло. Кроме того, $X \subseteq \tilde{X}(x_0)$ для любого $x_0 \in \mathbb{R}^n$.

Потребуем, чтобы градиенты всех функций $f^i(x)$, $1 \leq i \leq r$, и $g^j(x)$, $1 \leq j \leq m$, удовлетворяли на $\mathbb{R}^n$ условию Липшица, т.е.

$$\|f^i_x(x_1) - f^i_x(x_2)\| \leq L\|x_1 - x_2\|, \quad \|g^j_x(x_1) - g^j_x(x_2)\| \leq L\|x_1 - x_2\| \tag{10.3.10}$$

для любых $x_1, x_2 \in \mathbb{R}^n$ и для всех $1 \leq i \leq r$, $1 \leq j \leq m$.

Предположим также, что точка x_0 такова, что

$$\max_{1\leq i\leq r} \max_{x\in\tilde{X}(x_0)} \|f^i_x(x)\| \leq K, \quad \max_{1\leq i\leq m} \max_{x\in\tilde{X}(x_0)} \|g^j_x(x)\| \leq K. \tag{10.3.11}$$

Неравенства (10.3.11) заведомо имеют место, если функции $f(x)$ и $g(x)$ непрерывно дифференцируемы и множество $\tilde{X}(x_0)$ ограничено.

Утверждение 10.3.5. *Пусть выполнены условия (10.3.10), (10.3.11) и пусть $x_k \in \tilde{X}(x_0)$. Тогда шаг α_k удовлетворяет неравенству*

$$\alpha_k \geq \bar{\alpha}_k = \min\left\{\alpha,\ \frac{-\sigma_k}{2Ln},\ \frac{\varepsilon_k}{-\sigma_k + 2\sqrt{n}K}\right\}. \tag{10.3.12}$$

Доказательство. Если $i \in J^{(1)}_{\varepsilon_k}(x_k, \eta_{k+1})$, то в силу (10.3.10) имеет место неравенство

$$f^i(x_k + \alpha_k s_k) \leq f^i(x_k) + \alpha_k\langle f^i_x(x_k), s_k\rangle + \frac{1}{2}L\alpha_k^2\|s_k\|^2.$$

Но из ограничений вспомогательной задачи (10.3.4) следует, что

$$\langle f^i_x(x_k), s_k\rangle \leq \sigma_k.$$

Кроме того, опять же согласно последнему ограничению задачи (10.3.4) $\|s_k\| \leq \sqrt{n}$. Отсюда приходим к неравенству

$$f^i(x_k + \alpha_k s_k) \leq f^i(x_k) + \alpha_k\left[\sigma_k + \frac{1}{2}Ln\alpha_k\right],$$

из которого вытекает, что при $\alpha_k \leq -\sigma_k(Ln)^{-1}$ выполнено неравенство

$$f^i(x_k + \alpha_k s_k) - \eta^i_{k+1} \leq M(x_k, \eta_{k+1}) + \frac{1}{2}\alpha_k\sigma_k. \tag{10.3.13}$$

Предположим теперь, что $i \notin J^{(1)}_{\varepsilon_k}(x_k, \eta_{k+1})$. В этом случае имеет место представление

$$f^i(x_k + \alpha_k s_k) = f^i(x_k) + \alpha_k \langle f^i_x(x_k + \alpha_k \theta^i_k s_k), s_k \rangle,$$

где $0 \le \theta^i_k \le 1$, и, значит, справедливо неравенство

$$\begin{aligned} f^i(x_k + \alpha_k s_k) - \eta^i_{k+1} &\le M(x_k, \eta_{k+1}) - \varepsilon_k + \alpha_k \sqrt{n} \| f^i_x(x_k + \alpha_k \theta^i_k s_k) \| \le \\ &\le M(x_k, \eta_{k+1}) - \varepsilon_k + \alpha_k \sqrt{n} K. \end{aligned}$$

Из него видно, что при $\alpha_k \le 2\varepsilon_k \left(-\sigma_k + 2\sqrt{n}K\right)^{-1}$ опять будет выполнено (10.3.13).

Неравенство, аналогичное (10.3.13), получается и для функций $g^j(x)$, но в виде

$$g^j(x_k + \alpha_k s_k) \le M(x_k, \eta_{k+1}) + \frac{1}{2} \alpha_k \sigma_k, \tag{10.3.14}$$

причем для всех $1 \le j \le m$. Соответствующее ограничение на α_k зависит от того, принадлежит ли индекс j множеству $J^{(2)}_{\varepsilon_k}(x_k, \eta_{k+1})$ или нет.

Поскольку согласно схеме алгоритма шаг α_k подбирается путем деления пополам начального шага α, то из (10.3.13) и (10.3.14) приходим к выводу, что справедлива оценка (10.3.12). ■

Таким образом, если точка $x_k \in \tilde{X}(x_0)$ такова, что $\sigma_k < 0$, то шаг α_k отличен от нуля и может быть найден путем дробления α конечное число раз.

Метод возможных направлений относится к *прямым методам* решения задачи (10.2.1). В нем безусловная минимизации вспомогательной функции $M(x, \eta)$ по x сочетается с одновременным решением на каждой итерации уравнения $\phi(x_k, \eta) = 0$. В тех случаях, когда точка x_k допустима, любое решение η этого уравнения вместе с x_k удовлетворяет равенству $M(x_k, \eta) = 0$.

Лемма 10.3.1. *На каждой итерации имеют место равенства*

$$\phi(x_k, \eta_{k+1}) = 0, \quad M(x_k, \eta_{k+1}) = \psi_+(x_k). \tag{10.3.15}$$

Доказательство. Согласно формуле пересчета (10.3.7) и (10.3.8)

$$f^i(x_k) - \eta^i_{k+1} = f^i(x_k) - \eta^i_k + \beta(x_k, \eta_k, e) e^i \le 0 \tag{10.3.16}$$

для любых $1 \le i \le r$. Поскольку хотя бы для одного индекса i неравенство (10.3.16) переходит в равенство, то из (10.3.16) следует, что

на каждой итерации справедливо левое равенство (10.3.15), влекущее также правое равенство. ■

Из правого равенства (10.3.15), в частности, вытекает, что в случае, когда точка x_k допустима и $\sigma_k < 0$, величина $\beta_{k+1} = \beta(x_{k+1}, \eta_{k+1})$ также оказывается строго положительной, т.е. новая оценка η_{k+2} отлична от предыдущей оценки η_{k+1} и $\eta_{k+2} < \eta_{k+1}$.

Утверждение 10.3.6. *Пусть точка $x_{k-1} \in X$ и пусть e_* — максимальная компонента вектора e. Тогда*

$$\beta(x_k, \eta_k, e) \geq \frac{\sigma_{k-1}\bar{\alpha}_{k-1}}{2e_*}, \tag{10.3.17}$$

где величина $\bar{\alpha}_k$ определена в (10.3.12).

Доказательство. Так как $x_{k-1} \in X$, то в силу предыдущей леммы $M(x_{k-1}, \eta_k) = 0$. В этом случае из (10.3.9) следует неравенство $M(x_k, \eta_k) \leq 2^{-1}\sigma_{k-1}\alpha_{k-1}$ и, значит, $f^i(x_k) - \eta_k^i \leq 2^{-1}\sigma_{k-1}\alpha_{k-1}$ для любого $1 \leq i \leq r$. Поэтому

$$\beta(x_k, \eta_k, e) = \min_{1 \leq i \leq r} \frac{\eta_k^i - f^i(x_k)}{e^i} \geq -\frac{1}{2}\sigma_{k-1}\alpha_{k-1}e_*^{-1}.$$

Отсюда и из утверждения 10.3.5 приходим к оценке (10.3.17). ■

Приведем еще одно важное свойство алгоритма метода возможных направлений.

Утверждение 10.3.7. *На каждой итерации $x_{k+1} \in \tilde{X}(x_k)$.*

Доказательство. Если $x_{k+1} \in X$, то заведомо $x_{k+1} \in \tilde{X}(x_k)$, поскольку, как было сказано выше, всегда $X \subseteq \tilde{X}(x)$, какое бы ни было $x \in \mathbb{R}^n$.

Предположим теперь, что $x_{k+1} \notin X$. Из неравенства (10.3.9) получаем, что $M(x_{k+1}, \eta_{k+1}) \leq M(x_k, \eta_{k+1})$. Но согласно лемме 10.3.1 $M(x_k, \eta_{k+1}) = \psi_+(x_k)$. Кроме того, из формулы (10.3.1) вытекает, что $\psi(x_{k+1}) \leq M(x_{k+1}, \eta_{k+1})$. Поэтому $\psi(x_{k+1}) \leq \psi(x_k)$ и, следовательно, $x_{k+1} \in \tilde{X}(x_k)$. ■

Утверждение 10.3.7 позволяет нам заключить, что независимо от выбора начальной стартовой точки $x_0 \in \mathbb{R}^n$ вся генерируемая алгоритмом последовательность $\{x_k\}$ остается в множестве $\tilde{X}(x_0)$. Если допустимое множество ограничено, то в силу выпуклости функции $\psi_+(x)$ множество $\tilde{X}(x_0)$ также будет ограничено. Поэтому у последовательности $\{x_k\}$ найдутся предельные точки. Можно показать, сделав дополнительные предположения о задаче, что все эти предельные точки

принадлежат множеству X_*^S. Последовательность оценок $\{\eta_k\}$ оказывается сходящейся и сходится она к оценке η_*, принадлежащей границе оболочки Эджворта–Парето F_+. Сформулируем это утверждение в виде теоремы.

Теорема 10.3.1. *Пусть начальная точка $x_0 \in \mathbb{R}^n$ такова, что множество $\tilde{X}(x_0)$ ограничено. Пусть, кроме того, выполнены условия (10.3.10) и (10.3.11). Тогда множество предельных точек последовательности $\{x_k\}$ принадлежит X_*^S.*

В результате работы метода получается одна оптимальная по Слейтеру оценка $f_* \in F_*^S$. Выбор разных начальных векторов η_0 приводит к разным оптимальным оценкам из F_*^S и к соответствующим им точкам из X_*^S. Возможен также другой подход к получению разных оценок $f_* \in F_*^S$, аналогичный тому, о котором говорилось при рассмотрении метода внешних центров, а именно: зафиксировать начальный вектор η_0 и менять направления e. Другими словами, строить некоторую радиальную сетку в виде пучка направлений. Отличие состоит в том, что теперь для построения сетки, охватывающей все множество F_*^S, целесообразно начальный вектор η_0 взять «северо-восточнее» точки, которую условно называют «точкой надира», т.е. положить $\eta_0 > f_{\max}$, где

$$f_{\max} = \left[f_{\max}^1, \ldots, f_{\max}^r\right], \quad f_{\max}^i = \max_{x \in X} f^i(x).$$

На рис. 10.2 показано типичное поведение траекторий в пространстве критериев $\mathbb{R}^r$ в зависимости от выбранных направлений. Рассмат-

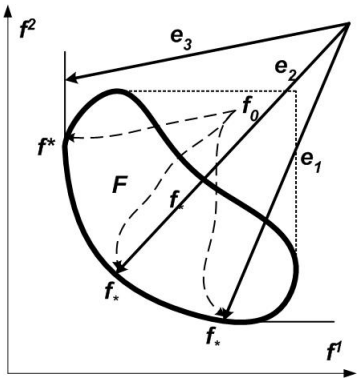

Рис. 10.2. Получение разных оптимальных оценок

ривается случай двух критериев. Все траектории выходят из одной точки x_0 со значением векторного критерия $f_0 = f(x_0)$. Если луч, задаваемый направлением e, пересекает множество оптимальных решений F_*^S в точке f_*, то последовательность $\{f(x_k)\}$ сходится к этой точке. В противном случае, когда оценки η_k сходятся к точке η_*, которая не принадлежит F_*^S, соответствующая последовательность $\{f(x_k)\}$ сходится к ближайшей точке из F_*^S.

10.4. Метод модифицированной функции Лагранжа для многокритериальной оптимизации

Приведем обобщение метода модифицированной функции Лагранжа для задачи многокритериальной оптимизации (10.2.1). Для построения такой функции воспользуемся подходом, который ранее применялся при построении модифицированной функции Лагранжа в случае задачи с одним критерием и с ограничениями-неравенствами. Наша цель здесь состоит в том, чтобы просто перенести его на задачу многокритериальной оптимизации (10.2.1).

Обратимся к введенной в [33] для этой задачи функции Лагранжа:

$$L(z) = \langle u, f(x) - \eta \rangle + \langle v, g(x) \rangle,$$

где $z = [x, u, v, \eta]$, причем $u \in \Lambda^r$, $v \in \mathbb{R}_+^m$ и $\eta \in \mathbb{R}^r$. Согласно необходимым условиям оптимальности точка Каруша–Куна–Таккера $z_* = [x_*, u_*, v_*, \eta_*]$ в задаче (10.2.1) определяется через равенства

$$\begin{array}{l} L_x(z_*) = 0_n, \quad L(z_*) = 0, \\ \pi_{\Lambda^r}\left(u_* + \alpha L_u(z_*)\right) = u_*, \\ \pi_{R_+^m}\left(v_* + \beta L_v(z_*)\right) = v_*, \end{array} \tag{10.4.1}$$

где α и β — произвольные положительные константы.

Если в задаче с одним критерием в методе модифицированной функции Лагранжа «штрафовалось» только последнее условие (10.4.1), касающееся ограничений $g(x) \le 0_m$, то в задаче с несколькими критериями надо дополнительно штрафовать предпоследнее условие (10.4.1), в которое входят целевые функции. Если это проделать, то приходим к следующей модифицированной функции Лагранжа:

$$M(z;t) = L(z) + (Bt)\left(L_u(z), u \,|\, \Lambda^r\right) + (Bt)\left(L_v(z), v \,|\, \mathbb{R}_+^m\right). \tag{10.4.2}$$

Здесь, напомним, $B(p,q\,|\,Q)$ — специальная функция, определяемая равенством (7.2.25) и обладающая «штрафными» свойствами. Она задается формулой

$$B(p,q\,|\,Q) = [\varphi(p) + \delta(p\,|\,Q-q)]^*\,, \tag{10.4.3}$$

в которой $\varphi(p)$ является гладкой выпуклой кофинитной функцией, а $\delta(p\,|\,Q-q)$ — индикаторная функция множества $Q-q$, звездочка означает сопряжение относительно переменной p. Далее предполагается, что множество Q выпукло и замкнуто.

Возьмем в качестве $\varphi(p)$ квадратичную функцию $\varphi(p) = \frac{\|p\|^2}{2}$. Она является самосопряженной, т.е.: $\varphi^*(p) = \varphi(p)$. В этом случае для множества $Q = \mathbb{R}^m_+$, как было показано в главе 7, функция (10.4.3) принимает вид

$$B(p,q\,|\,\mathbb{R}^m_+) = \frac{1}{2}\left[\|p\|^2 - \|p+q\|^2_-\right]. \tag{10.4.4}$$

Здесь a_- — отрицательная срезка вектора a.

Пусть теперь Q — вероятностный симплекс Λ^r и пусть по-прежнему $\varphi(p) = \frac{\|p\|^2}{2}$. Согласно формуле

$$B(p,q\,|\,Q) = \frac{1}{2}\left[\|p\|^2 - \|p+q-\pi_Q(p+q)\|^2\right], \tag{10.4.5}$$

являющейся частным случаем общей формулы (10.4.3), чтобы вычислить функцию $B(p,q\,|\,\Lambda^r)$, нам надо знать проекцию вектора $p+q$ на симплекс Λ^r. Но вопрос о нахождении такой проекции, был уже рассмотрен в [33], где показано, что всё определяется величинами μ_k, $1 \le k \le r$, которые подсчитываются специальным образом. Сначала упорядочиваются величины $p^i + q^i$ в порядке их убывания, т.е. $p^{i_1} + q^{i_1} \ge \ldots \ge p^{i_r} + q^{i_r}$, по ним находятся величины

$$\mu_k = \frac{1}{k}\left(k(p^{i_k} + q^{i_k}) - \sum_{j=1}^{k}(p^{i_j} + q^{i_j}) + 1\right), \quad k = 1,2,\ldots,r.$$

Всегда $\mu_1 = 1$.

Пусть k_* — максимальный индекс из $1,2,\ldots,r$ такой, что $\mu_{k_*} > 0$. Определим с его помощью индексное множество $J(p+q) = \{i_1,\ldots,i_{k_*}\}$ и величину

$$\lambda(p+q) = \frac{1}{k_*}\left(\sum_{i\in J(p+q)}(p^i + q^i) - 1\right).$$

Тогда

$$\pi_{\Lambda^r}(p+q) = \begin{cases} p^i + q^i - \lambda(p+q), & i \in J(p+q), \\ 0, & i \notin J(p+q) \end{cases} \tag{10.4.6}$$

и, подставляя это выражение для проекции в формулу (10.4.5), получаем

$$B(p,q\,|\,\Lambda^r) = \frac{1}{2}\left\{ \|p\|^2 - \sum_{i \notin J(p+q)} (p^i + q^i)^2 - \left[\sum_{i \in J(p+q)} (p^i + q^i) - 1 \right]^2 \right\}. \tag{10.4.7}$$

В силу единственности проекции точки на выпуклое замкнутое множество Λ^r функция (10.4.7) оказывается дифференцируемой по первому аргументу и

$$B_p(p+q\,|\,\Lambda^r) = \pi_{\Lambda^r}(p+q) - q, \tag{10.4.8}$$

причем, как и функция (10.4.4), в некоторых точках она оказывается дважды дифференцируемой.

Лемма 10.4.1. *Пусть $1 \le s \le r$ и пусть точки $p_* \in \mathbb{R}^r$ и $q_* \in \Lambda^r$ таковы, что*

$$p_* = \left[0, \ldots, 0, p_*^{s+1}, \ldots, p_*^r\right]^T, \quad q_* = \left[q_*^1, \ldots, q_*^s, 0, \ldots, 0\right]^T,$$

а компоненты p_*^i, $s < i \le r$, и q_*^j, $1 \le j \le s$, удовлетворяют неравенствам: $p_*^i < 0$, $q_*^j > 0$. Тогда $B_p(p_*, q_*\,|\,\Lambda^r) = 0_r$. Более того, функция $B(p,q\,|\,\Lambda^r)$ дважды дифференцируема по p в точке $[p_*, q_*]$ и ее матрица вторых производных имеет вид

$$B_{pp}(p_*, q_*\,|\,\Lambda^r) = \begin{bmatrix} I_s & 0 \\ 0 & 0 \end{bmatrix} - \frac{1}{s}\begin{bmatrix} E_s & 0 \\ 0 & 0 \end{bmatrix}, \tag{10.4.9}$$

где E_s — квадратная матрица порядка s, все элементы которой равны единице.

Доказательство. Воспользуемся формулой (10.4.8) для производной функции $B(p,q\,|\,\Lambda^r)$ по переменной p.

Вычислим $\pi_{\Lambda^r}(p_*+q_*)$. В силу сделанных предположений $p_*^i + q_*^i > 0$ при $1 \le i \le s$ и $p_*^j + q_*^j < 0$ при $s < j \le r$. Пусть для определенности $q_*^1 \ge \ldots \ge q_*^s$ и $p_*^{s+1} \ge \ldots p_*^r$. Тогда с учетом того, что $q_*^1 + \ldots + q_*^s = 1$, получаем

$$\mu_k = \frac{1}{k}\left[kq_*^k + 1 - \sum_{j=1}^{k} q_*^j \right] > 0, \quad 1 \le k \le s, \tag{10.4.10}$$

$$\mu_k = \frac{1}{k}\left[kp_*^k - \sum_{j=s+1}^{k} p_*^j\right] < 0, \quad s < k \le r. \tag{10.4.11}$$

Отсюда приходим к выводу, что $J(p_* + q_*) = \{1, \ldots, s\}$ и, следовательно,

$$\lambda(p_* + q_*) = \frac{1}{s}\left[\sum_{i=1}^{s}(p_*^i + q_*^i) - 1\right] = \frac{1}{s}\sum_{i=1}^{s} p_*^i = 0.$$

Поэтому, согласно (10.4.6), $\pi_{\Lambda^r}(p_* + q_*) = q_*$ и, значит, выполняется равенство $B_p(p_*, q_* \,|\, \Lambda^r) = 0_r$.

Убедимся теперь, что у функции $B(p, q \,|\, \Lambda^r)$ в точке $[p_*, q_*]$ имеется вторая производная. Из-за того, что неравенства (10.4.10) и (10.4.11) строгие, найдется достаточно малая окрестность точки p_* такая, что для всех p из этой окрестности строгие неравенства сохраняются. Поэтому множество $J(p + q_*)$ для этих p также будет состоять из первых s индексов и справедлива формула

$$B_p^i(p, q_*) \,|\, \Lambda^r) = \begin{cases} p^i - \lambda(p + q_*), & 1 \le i \le s, \\ 0, & s < i \le r, \end{cases}$$

в которой $\lambda(p + q_*) = s^{-1}\sum_{i=1}^{s} p^i$. Таким образом, вектор-функция $B_p(p, q_* \,|\, \Lambda^r)$ дифференцируема по p в окрестности точки p_* и ее матрица Якоби $B_{pp}(p, q_* \,|\, \Lambda^r)$ имеет вид (10.4.9). ■

Используя вид матрицы $B_{pp}(p, q_* \,|\, \Lambda^r)$, получаем, что для любого $d \in \mathbb{R}^r$ выполнено

$$\langle d,\, B_{pp}(p, q_* \,|\, \Lambda^r)\, d\rangle = \sum_{i=1}^{s}(d^i)^2 - \frac{1}{s}\left(\sum_{i=1}^{s} d^i\right)^2.$$

Отсюда на основании неравенства между арифметическим и квадратичным средним приходим к выводу, что матрица $B_{pp}(p, q_* \,|\, \Lambda^r)$ неотрицательно определена, причем равенство $\langle d,\, B_{pp}(p, q_* \,|\, \Lambda^r)\, d\rangle = 0$ возможно только в том случае, когда первые s компонент вектора d равны между собой.

Далее рассматриваем модифицированную функцию (10.4.2), в которой функции $B(p, q \,|\, \mathbb{R}_+^m)$ и $B(p, q \,|\, \Lambda^r)$ имеют соответственно вид (10.4.4) и (10.4.7), t — фиксированный положительный параметр. Добавление к функции Лагранжа штрафных членов значительно улучшает ее поведение, а именно, можно показать, что при некоторых дополнительных предположениях вблизи точек Каруша–Куна–Таккера

всегда существуют локальные решения задачи безусловной минимизации:

$$\min_{x\in R^n} M(x,u,v,\eta;t). \tag{10.4.12}$$

Для этого потребуются достаточные условия второго порядка из [33] для задачи (10.2.1). Кроме того, следует также уточнить понятие дополняющей нежесткости и строгой дополняющей нежесткости для случая задачи с несколькими критериями.

Определение 10.4.1. *В точке $z_* = [x_*, u_*, v_*, \eta_*]$ выполнено условие дополняющей нежесткости, если $u_*^i(f^i(x_*) - \eta_*^i) = 0$, $1 \le i \le r$, и $v_*^j g^j(x_*) = 0$, $1 \le j \le m$. Если, кроме того, для всех $1 \le i \le r$ и всех $1 \le j \le m$ равенства $f^i(x_*) - \eta_*^i = 0$ и $g^j(x_*) = 0$ возможны в том и только в том случае, когда $u_*^i > 0$, $v_*^j > 0$, то в точке z_* выполнено условие строгой дополняющей нежесткости.*

Утверждение 10.4.1. *Пусть в точке z_* выполнены достаточные условия второго порядка и условие строгой дополняющей нежесткости. Тогда можно указать $t_* > 0$ такое, что при $0 < t < t_*$ для всех $[u, v, \eta]$ из некоторой окрестности $\Delta(u_*, v_*, \eta_*)$ точки $[u_*, v_*, \eta_*]$, где $u \in \Lambda^r$ и $v \in \mathbb{R}^m_+$, существует изолированное локальное решение $x(u, v, \eta)$ задачи (10.4.12), удовлетворяющее условию $x_* = x(u_*, v_*, \eta_*)$.*

Доказательство. Допустим, для определенности, что первые s компонент вектора u_* и первые l компонент вектора v_*, и только они, больше нуля. Из предположения о строгой дополняющей нежесткости вытекает, что точка $[p_*, q_*] = [L_u(z_*), u_*]$ удовлетворяет условиям леммы 10.4.1, а точка $[p_*, q_*] = [L_v(z_*), v_*]$ — условиям леммы 7.2.2. Поэтому

$$M_x(z_*;t) = L_x(z_*) + f_x^T(x_*)(Bt)_p(L_u(z_*), u_* \,|\, \Lambda^r) + \\ + g_x^T(x_*)(Bt)_p(L_v(z_*), v_* \,|\, \mathbb{R}^m_+) = 0_n.$$

Кроме того, функция $M(z;t)$ дважды дифференцируема по x в z_* и

$$M_{xx}(z_*;t) = L_{xx} + f_x^T(x_*)(Bt)_{pp}(L_u(z_*), u_* \,|\, \Lambda^r) f_x(x_*) + \\ + g_x^T(x_*)(Bt)_{pp}(L_v(z_*), v_* \,|\, \mathbb{R}^m_+) g_x(x_*).$$

Матрица $(Bt)_{pp}(L_v(z_*), v_* \,|\, \mathbb{R}^m_+)$ неотрицательно определена, причем

$$\langle d,\, g_x^T(x_*)(Bt)_{pp}(L_v(z_*), v_* \,|\, \mathbb{R}^m_+) g_x(x_*) d\rangle = 0 \tag{10.4.13}$$

в том и только в том случае, когда $\langle g_x^j(x_*), d\rangle = 0$, $1 \le j \le l$.

Матрица $(Bt)_{pp}(L_u(z_*), u_* \,|\, \Lambda^r)$, как уже отмечалось, также неотрицательно определена, причем равенство

$$\langle d,\, f_x^T(x_*)(Bt)_{pp}(L_u(z_*), u_* \,|\, \Lambda^r) f_x(x_*) d\rangle = 0 \tag{10.4.14}$$

имеет место, когда величины $\langle f_x^i(x_*), d\rangle$, $1 \le i \le s$, равны между собой и, в частности, совпадают с нулем. Обозначим множество векторов d, удовлетворяющих равенствам (10.4.13) и (10.4.14), через $\mathcal{D}$.

Отсюда согласно требованию

$$\langle d,\, L_{xx}(z_*) d\rangle > 0 \quad d \in \mathcal{D}, \quad d \neq 0_n,$$

входящему в достаточные условия второго порядка, и лемме Финслера 7.2.1 получаем, что матрица $M_{xx}(z_*; t)$ положительно определена при t достаточно малых, $t < t_*$. В силу непрерывности она будет оставаться положительно определенной и вблизи точки z_*. Тогда по теореме о неявной функции можно указать окрестность $\Delta(u_*, v_*, \eta_*)$ точки $[u_*, v_*, \eta_*]$ такую, что на ней существует непрерывно дифференцируемая функция $x(u, v, \eta)$, определяемая уравнением $M_x(z; t) = 0_n$ и являющаяся решением задачи (10.4.12). ■

Пусть задан начальный вектор $\eta_0 \in \mathbb{R}^r$ и выбрано направление $e > 0_r$ такое, что $\sum_{i=1}^r e^i = 1$. Положим $\eta(\eta_0, e, \sigma) = \eta_0 + \sigma e$, и построим итерационный процесс для отыскания точек z_*, удовлетворяющих условиям (10.4.1):

$$x_k = \arg \min_{x \in R^n} M(x, u_k, v_k, \eta_k; t), \quad \eta_k = \eta(\eta_0, e, \sigma_k), \tag{10.4.15}$$

$$u_{k+1} = \pi_{\Lambda^r}\left(u_k + \frac{L_u(z_k)}{t}\right), \quad v_{k+1} = \pi_{R_+^m}\left(v_k + \frac{L_v(z_k)}{t}\right), \tag{10.4.16}$$

$$\sigma_{k+1} = \sigma_k + M(z_k; t). \tag{10.4.17}$$

Здесь $\sigma_0 = 0$, $u_0 \in \Lambda^r$, $v_0 \in \mathbb{R}_+^m$, $z_k = [x_k, u_k, v_k, \eta_k]$.

Положим $w = [u, v, \sigma]$, $x(w; t) = x(u, v, \eta(\eta_0, e, \sigma); t)$, где $x(u, v, \eta; t)$ — непрерывно дифференцируемая вектор-функция, определяемая из решения задачи (10.4.12). Итерационный процесс (10.4.15)–(10.4.17), — это, по существу, метод простой итерации, примененный для решения системы уравнений

$$(Bt)_p(L_u(x(w; t), u, v, \eta(\eta_0, e, \sigma)), u \,|\, \Lambda^r) = 0_r, \tag{10.4.18}$$

$$(Bt)_p(L_v(x(w; t), u, v, \eta(\eta_0, e, \sigma)), v \,|\, \mathbb{R}_+^m) = 0_m, \tag{10.4.19}$$

$$M(x(w; t), u, v, \eta(\eta_0, e, \sigma)) = 0. \tag{10.4.20}$$

Пусть $\eta_* = \eta_*(\eta_0, e)$ — точка, лежащая на прямой

$$\gamma(\eta_0, e) = \{\eta \in \mathbb{R}^r : \ \eta = \eta_0 + \sigma e, \ \sigma \in \mathbb{R}\},$$

и такая, что $\eta_*(\eta_0, e)$ принадлежит границе множества F_+ (оболочки Эджворта–Парето множества F). Понятно, что для каждого $\eta_0 \in \mathbb{R}^r$ и $e > 0_r$ существует только одна такая точка. Величину σ, для которой $\eta(\eta_0, e, \sigma) = \eta_*(\eta_0, e)$, обозначим через $\sigma_* = \sigma_*(\eta_0, e)$.

При определенных дополнительных предположениях описанный метод обладает локальной сходимостью в том смысле, что последовательность $\{[u_k, v_k, \sigma_k]\}$ сходится к точке $[u_*, v_*, \sigma_*]$. Соответствующая последовательность $\{\eta_k\}$ при этом сходится к точке η_*.

Пусть $h(x, \eta)$ обозначает $(r+m)$-мерную вектор-функцию, первыми r компонентами которой являются функции $f^i(x) - \eta^i$, а последующими m компонентами — функции $g^j(x)$. Через $J_0(x, \eta)$ обозначим индексное множество $\{i \in J^{r+m} : \ h^i(x, \eta) = 0\}$. Символом $h_x(x, \eta)$ будем обозначать матрицу первых производных функции $h(x, \eta)$ относительно первого аргумента.

Определение 10.4.2. *В точке $[x_*, \eta_*]$ выполнено условие регулярности, если $h_x(x_*, \eta_*)d \neq 0$ для любого ненулевого вектора $d \in \mathbb{R}^{r+m}$, у которого сумма первых r компонент равна нулю, и такого, что $d^i = 0$, если $i \notin J_0(x_*, \eta_*)$.*

Таким образом, теперь условие регулярности относится не только к ограничениям задачи, но и к целевым функциям. При его выполнении оказывается справедливым следующий результат, касающийся локальной сходимости метода (10.4.15) – (10.4.17). Он может быть доказан с помощью теоремы Островского 7.1.1.

Теорема 10.4.1. *Пусть в точке $z_* = [x_*, u_*, v_*, \eta_*]$ выполнены достаточные условия второго порядка и условие строгой дополняющей нежесткости. Пусть, кроме того, в точке $[x_*, \eta_*]$ выполнено условие регулярности и $L_{xx}(z_*)$ — неособая матрица. Тогда можно указать $t_* > 0$ такое, что для всех $0 < t < t_*$ итерационный процесс (10.4.15) – (10.4.17) локально сходится к точке $[u_*, v_*, \sigma_*]$. Соответствующая последовательность $\{x_k\}$ сходится к точке x_*.*

Метод целесообразно применять на последнем этапе расчетов, когда известно хорошее начальное приближение и надо только получить более точное решение. Выбор различных начальных η_0 и различных направляющих векторов $e > 0_r$ приводит к различным решениям из F_*^S, однако теперь оценки η_k в ходе итерационного процесса могут оказываться по обе стороны границы множества F_+.

Глава 11

Методы глобальной оптимизации

11.1. Постановка задачи глобальной оптимизации

Под глобальной оптимизацией понимают решение таких оптимизационных задач, в которых целевая функция имеет много локальных минимумов, но требуется найти именно глобальный минимум. Функции такого типа принято называть *многоэкстремальными* и они характерны для большей части современных прикладных оптимизационных задач. Задачи глобальной оптимизации являются одними из наиболее трудных с точки зрения нахождения решений среди всех задач оптимизации. Часто помимо невыпуклости они обладают негладкостью или вообще описываются моделью «черного ящика».

В качестве примера задачи глобальной оптимизации, которой в последнее время уделяется повышенное внимание, приведем задачу определения декартовых координат m атомов в атомном кластере. Для описания энергии такого кластера используют, в частности, *функцию Морса*:

$$f(z_1,\ldots,z_m,\rho)=\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left[\left(e^{\rho(1-\|z_i-z_j\|)}-1\right)^2-1\right], \qquad (11.1.1)$$

где ρ — скалярный параметр, z_i и z_j — трехмерные векторы, задающие координаты центров атомов i и j. Поэтому размерность задачи (общее число переменных в ней) равняется $n=3m$, вектор x является объ-

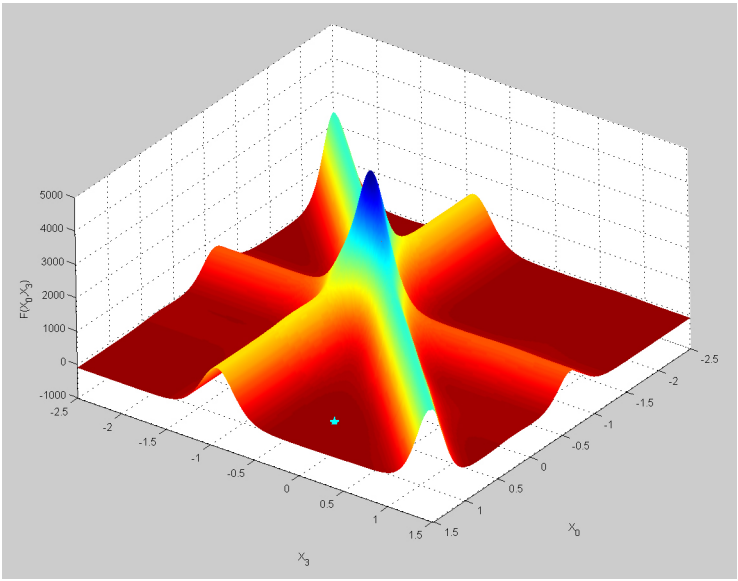

Рис. 11.1. Сечение функции Морса по двум координатам

единением всех трехмерных векторов: $z_i = \left[z_i^{(1)}, z_i^{(2)}, z_i^{(3)}\right]$, $1 \le i \le m$. На рис. 11.1 в качестве примера показано сечение по двум координатам функции Морса (11.1.1), когда в кластере присутствуют 75 атомов.

Численные методы решения задач глобальной оптимизации существенно отличаются от рассмотренных ранее методов, направленных главным образом на отыскание локальных минимумов. Для них трудно или практически невозможно покинуть *зону притяжения* одного локального минимума и перейти в зону притяжения другого локального минимума. Кроме того, сами задачи глобальной оптимизации могут существенно отличаться по своему виду, что ведет к разработке разных подходов к их решению. При этом предлагаются как специальные методы, направленные на решение узкого класса задач, так и более универсальные, которые могут применяться для решения задач глобальной оптимизации в более широкой постановке. Здесь будут рассмотрены только три метода, основанные на разных идеях и являющиеся достаточно универсальными.

Эти три метода можно отнести к классу так называемых *последовательных или адаптивных алгоритмов*. В методах из этого класса при

выборе точек вычисления значения функции учитывается не только априорная информация о задаче, но и та информация, которая была получена в ходе вычислений. Проводится анализ этой информации и делается вывод о том, в каких новых точках вычислять значение целевой функции.

Что касается априорной информации о задаче, то, как правило, она заключается в характере поведения целевой функции на всем множестве поиска или на каких-то отдельных участках. Наиболее популярными являются предположения о *липшицевой непрерывности* либо самой функции, либо ее производных. В последнем случае минимизируемая функция должна быть достаточно гладкой.

В дальнейшем для простоты речь будет идти только о целевых функциях $f(x)$, которые непрерывны по Липшицу. Согласно определению для таких функций существует константа L такая, что

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\| \qquad \forall\ x_1, x_2 \in X, \tag{11.1.2}$$

причем в качестве нормы может браться не только евклидова норма, но и другие нормы, в частности октаэдрическая или чебышевская норма. Для простоты считаем также, что допустимое множество X является *множеством параллелепипедного типа* или, как еще говорят, *гиперинтервалом*:

$$X = P, \qquad P = \left\{x \in \mathbb{R}^n :\ a \leq x \leq b\right\}, \tag{11.1.3}$$

где $a, b \in \mathbb{R}^n$ и $a < b$. О задаче

$$f_* = \operatorname{glob} \min_{x \in X} f(x), \tag{11.1.4}$$

в которой функция $f(x)$ удовлетворяет (11.1.2), а X имеет вид (11.1.3), иногда говорят как о задаче *липшицевой глобальной оптимизации*. Ее решением является множество

$$X_* = \{x \in X :\ f(x) = f_*\}.$$

Для большинства задач глобальной оптимизации (не только липшицевой) найти хотя бы одну точку из множества X_* не представляется возможным. Поэтому ищут приближенные решения. Зададимся $\varepsilon > 0$ и вместо X_* рассмотрим множество ε-оптимальных решений:

$$X_*^\varepsilon = \{x \in X :\ f(x) \leq f_* + \varepsilon\}.$$

Теперь решение задачи (11.1.4) означает нахождение некоторой произвольной точки x_*^ε из X_*^ε. Ей соответствует приближенное оптимальное значение целевой функции $f_*^\varepsilon = f(x_*^\varepsilon)$.

11.2. Метод ломанных

В этом методе строится вспомогательная кусочно-линейная функция, являющаяся аппроксимацией минимизируемой функции снизу, и в качестве новой точки вычисления значения функции (*точки испытания*) берется точка минимума вспомогательной функции.

Идея данного метода проста и давно используется в *методе касательных*, предназначенном для минимизации выпуклых функций. Метод касательных основан на том свойстве выпуклых функций, что если вычислено значение выпуклой функции $f(x)$ в некоторой точке y и найден субградиент $a \in \partial f(y)$, то аффинная относительно x функция

$$h(x) = f(y) + \langle a, x - y \rangle \tag{11.2.1}$$

есть *опорная функция* к $f(x)$ снизу. Другими словами, она является *минорантой* $f(x)$, т.е. выполняется неравенство $f(x) \geq h(x)$, причем $f(y) = h(y)$. Это касается произвольных точек y, поэтому если у нас есть несколько точек $y_1, \ldots, y_k$ и известны субградиенты $a_1, \ldots, a_k$, принадлежащие соответственно субдифференциалам $\partial f(y_1), \ldots, \partial f(y_k)$, то обязательно $f(x) \geq h_k(x)$, где

$$h_k(x) = \max_{1 \leq i \leq k} \left[f(y_i) + \langle a_i, x - y_i \rangle \right]. \tag{11.2.2}$$

Функция $h_k(x)$ является кусочно-линейной.

В качестве следующей точки y_{k+1} берут ту точку, которая доставляет минимум миноранте $h_k(x)$ на X. Функции-миноранты $h_k(x)$ с каждым вычислением новой точки y_{k+1} все более точнее аппроксимируют $f(x)$ в том смысле, что

$$h_k(x) \leq h_{k+1}(x) \leq f(x)$$

всюду на X. Это следует из очевидного представления для $h_{k+1}(x)$, а именно:

$$h_{k+1}(x) = \max \left[h_k(x),\ f(y_{k+1}) + \langle a_{k+1}, x - y_{k+1} \rangle \right].$$

Если ввести рекордные значения

$$f_k^* = \min_{1 \leq i \leq k} f(y_i), \qquad h_k^* = \min_{x \in X} h_k(x) = h_k(y_{k+1}),$$

то $f_k^* \geq h_k^*$ для всех $k \geq 1$. Более того,

$$f_{k+1}^* \leq f_k^*, \qquad h_{k+1}^* \geq h_k^*. \tag{11.2.3}$$

Таким образом, зазор между рекордными значениями $\Delta_k = f_k^* - h_k^*$ лишь убывает от итерации к итерации, т.е. $\Delta_{k+1} \leq \Delta_k$. Выполнение неравенства $\Delta_k \leq \varepsilon$, где ε — некоторая наперед заданная точность, может служить критерием останова в таком итерационном процессе.

Перейдем теперь от минимизации выпуклой функции к минимизации липшицевой функции $f(x)$. Теперь вместо линейной функции (11.2.1) нам приходится пользоваться функцией

$$h(x) = f(y) - L\|x - y\|,$$

а вместо миноранты (11.2.2) — соответственно функцией

$$h_k(x) = \max_{1 \leq i \leq k} \left[f(y_i) - L\|x - y_i\| \right].$$

Последующая точка y_{k+1} по-прежнему выбирается исходя из минимизации миноранты $h_k(x)$. Неравенства (11.2.3), характеризующие рекордные значения f_k^* и h_k^* и зазор Δ_k между ними, также сохраняются.

График функции $h_k(x)$ определяется множеством пересекающихся конусов в пространстве $\mathbb{R}^{n+1}$ с вершинами в точках $[y_k, f(y_k)]$. Все эти конусы направлены вниз. Понятно, что нахождение точки минимума такой функции в общем случае n-мерного пространства представляется очень сложной вычислительной задачей. Она существенно упрощается, если $f(x)$ — функция одного аргумента, т.е. $x \in X$, где X — отрезок $[a, b]$, принадлежащий действительной прямой $\mathbb{R}$. Примерный график многоэкстремальной функции $f(x)$ и соответствующей миноранты $h_k(x)$ показаны на рис. 11.2. На этом рисунке проиллюстрирован также выбор новой точки испытания y_3 по найденным предыдущим точкам y_1 и y_2.

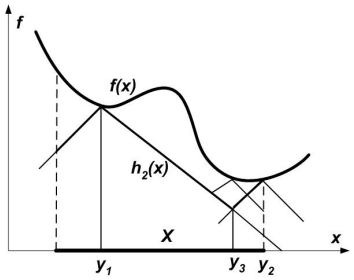

Рис. 11.2. Графики функций $f(x)$ и $h_2(x)$

Одномерный метод ломанных, который предназначен для минимизации функции одной переменной, может быть разными способами перенесен на многомерный случай. Одним из таких полезных способов является редукция размерности задачи с помощью *кривых Пеано–Гильберта*. Эти кривые являются фракталами и обладают важным свойством: они однозначно и непрерывно отображают отрезок $[0,1]$ на гиперинтервал P. Их еще называют *кривыми, заполняющими пространство*.

Первым кривую со свойством заполнения пространства предложил в 1890 г. Дж. Пеано. Это кривая на плоскости, заполняющая единичный квадрат и проходящая через каждую его точку по меньшей мере один раз. Строится она следующим образом. Каждая сторона единичного квадрата делится на три равные части и в результате получаются девять меньших квадратов. Кривая проходит эти девять квадратов в определенном порядке. Затем каждый из девяти маленьких квадратов разбивается аналогичным образом на девять еще меньших квадратов и кривая проходит вновь появившиеся маленькие квадратики, охватывая все из них и сохраняя при этом непрерывность, и т.д.

Затем в 1891 г. Д. Гильберт дал свой способ построения кривой, заполняющей пространство, согласно которому единичный квадрат разбивается на четыре равных подквадрата, что соответствует делению каждой стороны квадрата пополам. Каждый из подквадратов в свою очередь делится опять на четыре части и т.д. На всех стадиях такого деления может быть построена кривая кусочно-линейного вида, проходящая через центры всех квадратов. На рис. 11.3 показана кривая Гильберта в случае, когда проведено два этапа разбиения исходного единичного квадрата, т.е. когда количество подквадратов равно $4^2 = 16$.

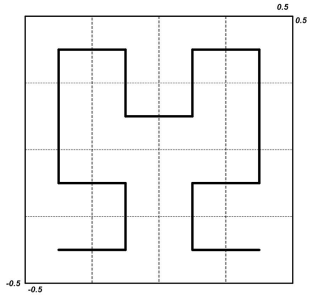

Рис. 11.3. Пример кривой Гильберта

Кривые Пеано–Гильберта могут быть построены и в многомерном случае. Если воспользоваться такой кривой или, точнее, ее приближением $x(t)$, найденным на какой-то стадии процесса заполнения допустимого множества, то многомерная задача глобальной оптимизации редуцируется к одномерной задаче

$$f_* = \text{glob} \min_{t\in[0,1]} f(x(t)).$$

При этом если многомерная функция $f(x)$, $x \in \mathbb{R}^n$, удовлетворяла условию Липшица с константой L, то одномерная функция $f(x(t))$ удовлетворяет на отрезке $[0,1]$ условию Гёльдера с показателем степени $\frac{1}{n}$ и с коэффициентом $C = 2L\sqrt{n+3}$, т.е.

$$|f(x(t_1)) - f(x(t_2))| \le \mathrm{C}|t_1 - t_2|^{\frac{1}{n}}.$$

Хотя функция $f(x(t))$ не является уже липшицевой, однако многие алгоритмы липшицевой оптимизации и, в частности, описанный алгоритм ломанных могут быть обобщены на случай минимизации гельдеровской функции.

11.3. Метод неравномерных покрытий

Метод неравномерных покрытий, предложенный Ю.Г. Евтушенко, предназначен для отыскания ε-оптимальных решений задачи глобальной оптимизации (11.1.4), причем с гарантией.

Пусть имеется набор точек $Y_k = [y_1, \ldots, y_k]$, где все $y_i \in X$. По точкам из Y_k последовательно определяем рекордные значения:

$$f_i^* = \min_{1\le j\le i} f(y_j), \qquad 1 \le i \le k,$$

а также ту точку x_i^* из $[y_1, \ldots, y_i]$, в которой реализуется это рекордное значение (назовем ее i-м *рекордным решением*).

Свяжем также с каждой точкой y_i из Y_k ее некоторую окрестность P_i. В качестве P_i берут обычно n-мерный параллелепипед или шар с центром в точке y_i. Введем дополнительно множества

$$P_i^+ = \{x \in P_i : \ f(x) \ge f_i^* - \varepsilon\}, \qquad i = 1, 2, \ldots, k. \tag{11.3.1}$$

Через X_k обозначим объединение всех множеств $P_1^+, \ldots, P_k^+$. Таким образом, $X_k = \cup_{i=1}^k P_i^+$. Будем говорить, что множество X_k *покрывает множество* X, если $X \subseteq X_k$. На рис. 11.4 показаны множества P_2 и P_2^+ для точки y_2. Рекордные значения f_2^* и f_1^* в данном случае совпадают.

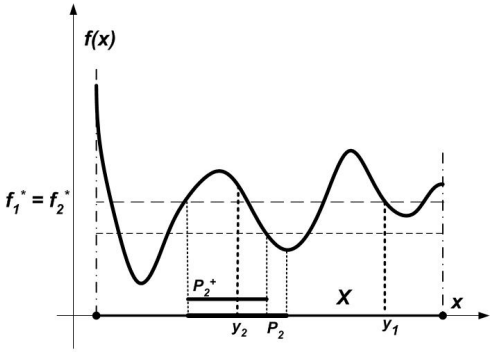

Рис. 11.4. Множества P_2 и P_2^+

Теорема 11.3.1. *Пусть набор точек Y_k такой, что соответствующее множество X_k покрывает допустимое множество X. Тогда k-е рекордное решение принадлежит множеству X_*^ε. Для рекордного значения f_k^* выполняется неравенство $f_* \le f_k^* \le f_* + \varepsilon$.*

Доказательство. Предположим для определенности, что k-м рекордным решением является точка $y_r \in Y_k$, т.е. $f(y_r) = f_k^*$. Далее, так как множество X_k покрывает X, то каждая оптимальная точка x_* из X_* принадлежит по крайней мере одному из множеств P_i^+, $1 \le i \le k$. Пусть это будет множество P_j^+, где $1 \le j \le k$. Рекордные значения $f_1^*, \ldots, f_k^*$ образуют набор монотонно невозрастающих величин. Поэтому $f_j^* \ge f_k^*$, и на основании (11.3.1) получаем

$$f_* = f(x_*) \ge f_j^* - \varepsilon \ge f_k^* - \varepsilon = f(y_r) - \varepsilon.$$

Отсюда следует, что $y_r \in X_\varepsilon^*$. ■

Идея метода неравномерных покрытий вытекает из утверждения теоремы 11.3.1. Она заключается в последовательном пополнении набора точек Y_k и в использовании все более расширяющегося набора множеств $\mathcal{P}_k^+ = \left[P_1^+, \ldots, P_k^+\right]$ так, чтобы, в конце концов, соответствующее множество X_k покрыло допустимое множество X.

Проводить такое расширение множеств X_k можно разными способами, рассмотрим один из них. Пусть в итерационном процессе последовательно выбираются новые точки y_{k+1} и определяются новые множества P_k^+.

Введем в рассмотрение лебеговское множество:

$$\mathcal{L}(f_k^*) = \{x \in X : \ f(x) \geq f_k^* - \varepsilon\}.$$

Понятно, что возникающее на k-шаге множество $\mathcal{L}(f_k^*)$ не представляет особого интереса с точки зрения глобальной оптимизации, так как, минимизируя функцию $f(x)$ на этом множестве, мы в лучшем случае уменьшим рекордное значение лишь на ε. Отсюда следует, что новую точку y_{k+1} целесообразно выбирать из дополнения множества $\mathcal{L}(f_k^*)$, т.е. из множества $X \setminus \mathcal{L}(f_k^*)$. Но множество $\mathcal{L}(f_k^*)$ нам фактически неизвестно, однако нам известно его подмножество — множество X_k. Поэтому выбираем точку y_{k+1} таким образом, чтобы $y_{k+1} \in X \setminus X_k$. Вычисляем $f(y_{k+1})$ и, если $f(y_{k+1}) < f_k^*$, обновляем рекордное значение f_{k+1}^*, полагая $f_{k+1}^* = f(y_{k+1})$. Берем далее окрестность P_{k+1} точки y_{k+1} и определяем множество P_{k+1}^+. Добавляем множество P_{k+1}^+ к множеству X_k и получаем новое множество $X_{k+1} = X_k \cup P_{k+1}^+$. Процесс заканчивается, когда будет покрыто все множество X, т.е. когда на некотором k-м шаге окажется, что $X \subseteq X_k$. Другими словами, это означает, что лебеговское множество $\mathcal{L}(f_k^*)$ становится наибольшим из возможных и совпадает с X.

Эффективность такого процесса в сильной степени зависит от рекордных значений f_k^*. Она наибольшая, когда известно оптимальное значение f_*, например из каких-то дополнительных физических соображений. Если же априори величина f_* не известна, то можно ускорить процесс с помощью методов локального поиска. В тех точках y_{k+1}, в которых $f(y_{k+1}) < f_k^*$, т.е. происходит улучшение рекорда, есть основание полагать, что, запуская из этой точки какой-то метод локальной оптимизации, можно будет найти новую точку $\bar{y}_{k+1}$, в которой $f(\bar{y}_{k+1}) < f(y_{k+1})$, и положить уже $f_{k+1}^* = f(\bar{y}_{k+1})$.

Если $P_{k+1}^+ = P_{k+1}$, то все множество P_{k+1} может быть исключено из дальнейшего рассмотрения. В противном случае надо дополнительно исследовать множество $P_{k+1}^- = P_{k+1} \setminus P_{k+1}^+$.

Рассмотренная схема метода носит, скорее, теоретический характер. При практической реализации этой схемы возникает ряд проблем, которые, однако, могут быть успешно разрешены.

Во-первых, это проблема нахождения множеств P_i^+, $1 \leq i \leq k$. Использование минорант функции $f(x)$ существенно упрощает решение данной проблемы.

Пусть $f(x)$ удовлетворяет условию Липшица, т.е. выполняется неравенство (11.1.2). Беря в качестве y точку y_i, получаем следующую ми-

норанту функции $f(x)$:

$$h(x; y_i) = f(y_i) - L\|x - y_i\|. \tag{11.3.2}$$

Для всех $x \in X$ оказывается справедливым неравенство $f(x) \geq h(x; y_i)$. Если теперь подставить в (11.3.1) вместо $f(x)$ ее миноранту $h(x; y_i)$, то приходим к множеству

$$\tilde{P}_i^+ = \{x \in P_i : \ h(x; y_i) \geq f_i^* - \varepsilon\}.$$

Понятно, что $\tilde{P}_i^+ \subseteq P_i^+$ для всех $1 \leq i \leq k$.

Для множества $\tilde{P}_i^+$ можно дать и другое представление, а именно:

$$\tilde{P}_i^+ = P_i \cap B_i, \qquad B_i = \left\{x \in \mathbb{R}^n : \ \|x - y_i\| \leq \rho_i\right\},$$

где $\rho_i = \frac{f(y_i) - f_i^* + \varepsilon}{L}$. Если $\|\cdot\|$ — евклидова норма, то B_i — шар с центром в точке y_i и с радиусом ρ_i. Если $\|\cdot\|$ — чебышевская норма $\|\cdot\|_\infty$, то B_i — n-мерный куб с центром в точке y_i и с главной диагональю, равной $2\rho_i$. Данные шар или куб будут наименьшими, когда $f_i^* = f(y_i)$, т.е. когда в точке y_i реализуется рекордное значение за все предыдущие испытания, включая текущее испытание. В этом случае $\rho_i = \frac{\varepsilon}{L}$.

Но в в противоположном случае, когда $f(y_i) \gg f_i^*$, радиус шара будет большим, и в принципе мы можем исключить из области дальнейшего рассмотрения не только множество $P_i \cap B_i$, но и весь шар B_i. Если объединение шаров B_i, $1 \leq i \leq k$, покрыло все множество X, то, как и в теореме 11.3.1, можно утверждать, что k-е рекордное решение x_k^* принадлежит множеству ε-оптимальных решений X_ε^* и $f_k^* \leq f_* + \varepsilon$.

Процесс неравномерного покрытия существенно ускоряется, если предположить, что минимизируемая функция $f(x)$ гладкая, а ее градиент $f_x(x)$ или матрица вторых производных $f_{xx}(x)$ удовлетворяют условию Липшица. В этом случае имеется возможность построить более точные миноранты и тем самым получить соответствующие шары B_i увеличенных размеров.

Упражнение 17. *Постройте миноранту $h(x; y_i)$ в случае, когда градиент функции $f(x)$ удовлетворяет условию Липшица с константой L. Считая, что множество P_i есть n-мерный параллелепипед, найдите диаметр шара B_i.*

Перейдем теперь ко второй проблеме, которая появляется при практической реализации алгоритма. Она заключается в способе выбора точек y_i и связанных с ними множеств P_i.

Считаем далее, что множество X есть n-мерный параллелепипед:

$$X = P = \left\{x \in \mathbb{R}^n : \ a \leq x \leq b\right\},$$

где $a, b \in \mathbb{R}^n$, $a < b$. В процессе работы алгоритма будем пользоваться вспомогательными параллелепипедами:

$$P_i = \left\{ x \in \mathbb{R}^n : \ a_i \le x \le b_i \right\}, \quad a \le a_i < b_i \le b.$$

В качестве точек y_i будем брать центры этих параллелепипедов. Обозначим: $d_i = b_i - a_i$. Тогда радиус шара, в который вписывается этот параллелепипед P_i, равен: $\Delta_i = \frac{\|d_i\|_2}{2}$. Отметим также, что чебышевская норма вектора d_i определяет длину максимального ребра параллелепипеда P_i. Пусть h_i^* — минимальное значение миноранты $h(x; y_i)$ на параллелепипеде P_i. Согласно (11.3.2) имеем

$$h_i^* = \min_{x \in P_i} h(x; y_i) = f(y_i) - L\Delta_i.$$

Пусть на k-м шаге берется параллелепипед P_k. Если оказывается, что $\Delta_k \le \rho_k$, то $P_k = P_k^+$ и весь этот параллелепипед P_k можно удалить из дальнейшего рассмотрения, т.е., другими словами, включить его в множество X_k. В противном случае производится последовательное дробление P_k на части по какому-нибудь ребру. При этом сечение параллелепипеда по максимальному ребру пополам представляется наиболее разумным. Если главные диагонали одной или обеих частей опять оказываются меньше ρ_k, то они исключаются из дальнейшего рассмотрения. Если нет, то в оставшихся параллелепипедах вычисляются новые точки y_{k+1} и т.д., пока параллелепипед P_k не будет полностью покрыт. Фактически здесь реализуется схема *метода ветвей и границ.*

Процесс половинных делений параллелепипедов можно интерпретировать как рост двоичного дерева. Вершинам дерева соответствуют параллелепипеды, полученные при начальном разбиении. Дуги соединяют данный параллелепипед с параллелепипедами, полученными из него в результате деления. Параллелепипеды, отвечающие концевым вершинам дерева, образуют текущий набор параллелепипедов, с которыми работает алгоритм. Некоторые концевые вершины могут быть исключены из этого набора, если согласно алгоритму они уже не представляют интереса с точки зрения нахождения оптимума в задаче.

Скажем несколько слов об эффективности метода неравномерных покрытий, а также о выборе константы Липшица. Как правило, точное значение константы Липшица не известно. Можно, конечно, взять ее верхнюю оценку. Но такая завышенная константа Липшица приводит к увеличению числа шагов и, следовательно, к возрастанию времени решения задачи. Поэтому поступают иначе: сначала берут заниженную константу Липшица с целью получить какое-то приближенное

значение, которое давало бы достаточно хорошее рекордное значение, а далее, используя это рекордное значение, увеличивают константу Липшица. Знание хорошего начального приближения, или, что одно и то же, хорошего рекордного значения, существенно увеличивает эффективность метода покрытия. Положительно влияет на эффективность метода и использование локальных констант Липшица, т.е. разных констант Липшица в различных областях.

11.4. Метод секущих углов

Метод секущих углов основан на идее, близкой к идее метода касательных. Отличие заключается в типе используемых опорных функций. Если в классическом методе касательных, предназначенном для минимизации выпуклой функции, используются линейные опорные функции, то в методе секущих углов, предназначенном для минимизации абстрактной выпуклой функции, в качестве опорных функций берутся абстрактные линейные функции. Это приводит к некоторому упрощению задачи нахождения локальных минимумов у минорант и выбора среди них наименьшего.

Рассмотрим задачу отыскания глобального минимума липшицевой функции на допустимом множестве, принадлежащему симплексу,

$$f_* = \operatorname{glob}\min_{x\in X} f(x), \qquad X \subseteq \Lambda^n, \tag{11.4.1}$$

где $\Lambda^n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x^i = 1\}$ — вероятностный симплекс (его размерность равна $n-1$). Несмотря на специальный вид допустимого множества в задаче (11.4.1), данная постановка достаточно общая, и задача глобальной минимизации на n-мерном параллелепипеде P сводится к задаче вида (11.4.1), причем несколькими способами. Рассмотрим два из них. При этом в обоих способах размерность симплекса увеличивается на единицу.

Способ 1. Проведем замену переменных:

$$x_1^i = n^{-1}\frac{x^i - a^i}{b^i - a^i}, \qquad 1 \le i \le n. \tag{11.4.2}$$

Тогда для точек x_1 выполняются неравенства: $0 \le x_1^i \le n^{-1}$, а сумма всех компонент вектора x_1 не превосходит единицу. Преобразование, обратное к (11.4.2), имеет вид

$$x^i = a^i + n(b^i - a^i)x_1^i, \qquad 1 \le i \le n, \tag{11.4.3}$$

или в векторном виде $x = a + nD(b-a)x_1$, где $D(b-a)$ — диагональная матрица с вектором $b-a$ на диагонали.

Обозначим

$$f_1(x_1) = f\left(a + nD(b-a)x_1\right).$$

Если ввести дополнительную переменную x_1^{n+1}, положив

$$x_1^{n+1} = 1 - \sum_{i=1}^{n} x_1^i,$$

то $0 \le x_1^{n+1} \le 1$, и для расширенного вектора $\bar{x} = [x_1, x_1^{n+1}] \in \mathbb{R}^{n+1}$ получаем $\bar{x} \in \Lambda^{n+1}$. Мы приходим к задаче глобальной минимизации на симплексе:

$$\operatorname{glob}\min_{\bar{x}\in\bar{X}} \bar{f}(\bar{x}), \qquad \bar{X} = \left\{\bar{x} \in \Lambda^{n+1} :\ \bar{x}^i \le n^{-1},\ 1 \le i \le n\right\},$$

где $\bar{f}(\bar{x}) = f_1(x_1)$. Формально эта задача имеет вид (11.4.1), однако фактически целевая функция $\bar{f}(\bar{x})$ от последней компоненты $\bar{x}^{n+1}$ не зависит. Графически данное преобразование переменных для случая, когда $n = 1$, показано на рис. 11.5.

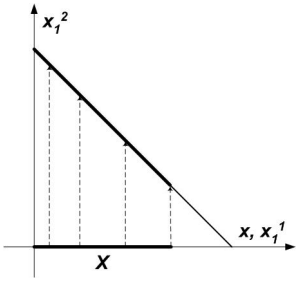

Рис. 11.5. Переход от множества X к множеству $\bar{X}$

Способ 2. Сделаем опять замену переменных, подобную (11.4.2):

$$x_1^i = c\frac{x^i - a^i}{b^i - a^i}, \qquad 1 \le i \le n,$$

где $c > 0$ — некоторая константа. Кроме того, введем дополнительную переменную x_1^{n+1} и зафиксируем ее, взяв $x_1^{n+1} = d$, где d — произвольное положительное число. Если рассмотреть расширенный вектор

$\bar{x}_1 = [x_1, x_1^{n+1}]$, то получаем

$$\bar{x}_1 \in \bar{X}_1, \qquad \bar{X}_1 = \left\{ \bar{x}_1 \in \mathbb{R}_+^{n+1} : \ \bar{x}_1^i \le c, \ 1 \le i \le n, \ \bar{x}_1^{n+1} = d \right\}.$$

Далее, каждой точке $\bar{x}_1$ из $\bar{X}_1$ поставим в соответствие точку $\bar{x}$ из симплекса Λ^{n+1}, определяя ее как $\bar{x} = \lambda \bar{x}_1$. Положительный параметр λ подбираем таким образом, чтобы $\sum_{i=1}^{n+1} \bar{x}^i = 1$. Подставляя в данное равенство выражение $\bar{x} = \lambda \bar{x}_1$, получаем $\lambda = \left(\sum_{i=1}^{n} x_1^i + d \right)^{-1}$. Следовательно, преобразование $\bar{x}_1 \to \bar{x}$ переписывается в виде

$$\bar{x}^i = \left(\sum_{i=1}^{n} \bar{x}_1^i + d \right)^{-1} \bar{x}_1^i, \quad 1 \le i \le n,$$
$$\bar{x}^{n+1} = \left(\sum_{i=1}^{n} \bar{x}_1^i + d \right)^{-1} d.$$

Нетрудно видеть, что при такой замене переменных всегда $\bar{x} \in \Lambda^{n+1}$, причем $\bar{x}^{n+1} > 0$. Геометрически переход от переменной $\bar{x}_1$ к переменной $\bar{x}$ для случая $n = 1$ и $d = 1$ показан на рис. 11.6.

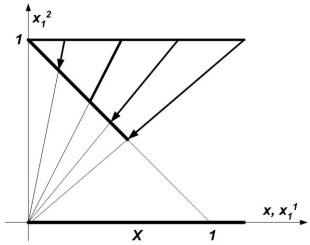

Рис. 11.6. Переход от множества $\bar{X}_1$ к множеству $\bar{X}$

Обратное преобразование $\bar{x} \to \bar{x}_1$ имеет вид $x_1 = \frac{d}{\bar{x}^{n+1}} \bar{x}$. Учтем далее вид обратного преобразования $x_1 \to x$, которое аналогично (11.4.3) следующее:

$$x = a + \frac{1}{c} D(b-a) x_1.$$

Подстановка сюда x_1 и подстановка x в функцию $f(x)$ дает нам новую функцию:

$$\bar{f}(\bar{x}) = f\left(a + \frac{d}{c \bar{x}^{n+1}} D(b-a) \bar{x} \right).$$

Задача сводится к минимизации функции $\bar{f}(\bar{x})$ на множестве $\bar{X}$, принадлежащему симплексу Λ^{n+1}, т.е. к задаче вида (11.4.1).

В связи с этим рассмотрим простейшую задачу *симплексного программирования*:

$$\operatorname{glob} \min_{x \in \Lambda^n} f(x). \tag{11.4.4}$$

Метод секущих углов позволяет решать задачи такого типа. Он основан на использовании возрастающих положительно однородных на $\mathbb{R}^n_+$ функций (IPH-функций). Данные функции являются частным случаем более общих возрастающих выпуклых вдоль лучей функций (ICAR-функций) опять же определенных на $\mathbb{R}^n_+$ (см. [33]).

Напомним, что если ввести на $\mathbb{R}^n_+$ семейство $\mathcal{L}_0$ *абстрактных линейных функций*:

$$l(x) = \min_{i \in I_+(l)} l^i x^i, \tag{11.4.5}$$

где $l \in \mathbb{R}^n_+$ и $I_+(l) = \{i \in J^n : \ l^i > 0\}$, то функция $f(x)$ будет IPH-функцией на $\mathbb{R}^n_+$ в том и только том случае, когда существует подмножество линейных функций $\mathcal{L}(f) \subset \mathcal{L}_0$ такое, что

$$f(x) = \max_{l \in \mathcal{L}(f)} l(x), \qquad x \in \mathbb{R}^n_+. \tag{11.4.6}$$

При этом несложно указать ту функцию $l(x)$, которая доставляет максимум в (11.4.6) в произвольной точке $x = x_0$ из $\mathbb{R}^n_+$, а именно: такой функцией будет $l(x)$, удовлетворяющая равенству $l(x_0) = \frac{f(x_0)}{x_0}$. Здесь считается, что $l^i = \frac{f(x_0)}{x_0^i}$, если $x_0^i > 0$. В противном случае полагается $l^i = 0$. Тогда $f(x_0) = l(x_0)$ и $f(x) \geq l(x)$ для всех $x \in \mathbb{R}^n_+$. Вектор l при этом называют *опорным вектором* к функции $f(x)$ в точке x_0. Любая IPH-функция принимает на $\mathbb{R}^n_+$ неотрицательные значения, так как в силу монотонности $f(x) \geq f(0_n) = 0$ для любого $x \in \mathbb{R}^n_+$.

Предположим, что функция $f(x)$ является липшицевой на Λ^n и пусть L есть липшицева константа этой функции на Λ^n относительно первой гельдеровской (октаэдрической) нормы, т.е.

$$|f(x_1) - f(x_2)| \leq L\|x_1 - x_2\|_1$$

для любых $x_1, \ x_2 \in \Lambda^n$. Тогда если $f(x) \geq 2L$ для всех $x \in \Lambda^n$, то существует такая IPH-функция $g(x)$, определенная на $\mathbb{R}^n_+$, что $f(x) = g(x)$ всюду на Λ^n. Выполнение данного неравенства $f(x) \geq 2L$ всегда можно добиться, прибавив к $f(x)$ достаточно большую константу. Поэтому $f(x)$ может быть расширена до IPH-функции, определенной на $\mathbb{R}^n_+$. Далее считаем, не умаляя общности, что сама $f(x)$ есть IPH-функция. Более того, считаем также, что $f_* = \min_{x \in \Lambda^n} f(x) > 0$.

Метод секущих углов для минимизации $f(x)$ строится аналогично методу касательных и его обобщению — методу ломанных. Опишем данный алгоритм.

Алгоритм метода секущих углов

Начальные итерации. Берем n единичных ортов e_i и последовательно полагаем $x_i = e_i$, $1 \le i \le n$. По точкам x_i вычисляем опорные векторы $l_i = \frac{f(x_i)}{x_i}$ и составляем функцию

$$h_n(x) = \max_{1 \le i \le n} l_i(x).$$

Функция $h_n(x)$ является минорантой функции $f(x)$ на Λ^n. Полагаем $k = n$ и $h_k(x) = h_n(x)$.

Общая k-я итерация ($k \ge n$)

Шаг 1. Находим решение x_* задачи

$$h_k^* = \min_{x \in \Lambda_n} h_k(x). \tag{11.4.7}$$

Шаг 2. Увеличиваем $k := k + 1$ и полагаем $x_k = x_*$.

Шаг 3. Вычисляем $l_k = \frac{f(x_k)}{x_k}$ и по формуле (11.4.5) строим новую абстрактную линейную функцию $l_k(x)$. Составляем функцию

$$h_k(x) = \max_{1 \le j \le k} l_j(x) = \max\left\{h_{k-1}(x), l_k(x)\right\},$$

являющуюся обновленной минорантой $f(x)$, и идем на шаг 1.

Можно показать, что при $k \ge n$ у всех минорант $h_k(x)$ их точки минимума x_* на симплексе Λ^n принадлежат относительной внутренности этого симплекса, т.е. $x_* > 0_n$. Из неравенства $f_* > 0$ следует, что все векторы l_k, $k \ge 1$, ненулевые, и поэтому абстрактные линейные функции $l_k(x)$ также отличны от функций, тождественно равных нулю. Кроме того, добавление новых функций $l_k(x)$ приводит к тому, что $h_k^* \ge h_{k-1}^*$.

Утверждение 11.4.1. *Пусть на некоторой k-й итерации алгоритма для точки x_k выполняется неравенство $f(x_k) \le h_k^* + \varepsilon$, где ε — заданная точность. Тогда x_k есть ε-оптимальное решение задачи симплексного программирования* (11.4.4).

Доказательство. Так как каждый из векторов l_k является опорным к функции $f(x)$ в точке x_k, то $f(x) \ge l_k(x)$ для всех $k \ge 1$. Поэтому $h_k^* \le f_*$. Имеем, с одной стороны, $f(x_k) \ge f_* \ge h_k^*$, а с другой —

$f(x_k) \le h_k^* + \varepsilon$. Таким образом, $f_* \le f(x_k) \le f_* + \varepsilon$. Отсюда приходим к требуемому результату. ■

Рассмотрим вспомогательную задачу (11.4.7) в методе секущих углов. Пусть на некоторой k-й общей итерации имеется N опорных векторов, которым соответствуют функции $l_j(x)$, $1 \le j \le N$. Тогда вспомогательная задача состоит в нахождении

$$\min_{x \in \Lambda^n} h(x), \tag{11.4.8}$$

где функция $h(x)$ имеет вид

$$h(x) = \max_{1 \le j \le N} l_j(x) = \max_{1 \le j \le N} \min_{i \in I_+(l_j)} l_j^i x^i. \tag{11.4.9}$$

Векторы l_j являются опорными к липшицевой функции $f(x)$. Первые n векторов l_j направлены вдоль единичных ортов, т.е.

$$l_j = c_j e_j, \quad c_j > 0, \quad 1 \le j \le n.$$

Остальные векторы l_{n+1}, $l_{n+2}, \ldots, l_N$, как это следует из алгоритма, принадлежат внутренности ортанта $\mathbb{R}^n_+$. В дальнейшем считаем, что все векторы l_1, $l_2, \ldots, l_N$ различны между собой и образуют недоминируемую систему векторов, т.е. для любых двух различных векторов l_s и l_t неравенства $l_s^i \le l_t^i$, $1 \le i \le n$, невозможны.

Пусть r_j — вектор, «обратный» по отношению к вектору l_j. Он определяется следующим образом:

$$r_j = \left[r_j^1, \ldots, r_j^n\right], \quad r_j^i = \begin{cases} \left(l_j^i\right)^{-1}, & \text{если} \quad i \in I(l_j), \\ 0, & \text{если} \quad i \notin I(l_j), \end{cases} \quad 1 \le i \le n.$$

Рассмотрим конечное множество векторов

$$\mathcal{P} = \{r_1, r_2, \ldots, r_N\},$$

где каждый из векторов r_j, $1 \le j \le N$, является «обратным» по отношению к вектору l_j. Из предположения о недоминируемости векторов $l_1, \ldots, l_N$ следует, что все векторы $r_1, \ldots, r_N$, входящие в множество $\mathcal{P}$, также отличны друг от друга и являются недоминируемыми, т.е. для любых двух различных векторов $r_s \in \mathcal{P}$ и $r_t \in \mathcal{P}$ противоположные неравенства $r_s^i \ge r_t^i$, $1 \le i \le n$, невозможны.

Обозначим через $\mathcal{P}_+$ множество

$$\mathcal{P}_+ = \mathcal{P} + \mathbb{R}^n_+ = \bigcup_{j=1}^{N} \left(r_j + \mathbb{R}^n_+\right).$$

Так как $r_j \in \mathbb{R}^n_+$, $1 \le j \le N$, то $\mathcal{P}_+ \subset \mathbb{R}^n_+$. Пусть ∂A — граница множества A. Для произвольного $\delta > 0$ введем в рассмотрение δ-окрестность начала координат, определив ее следующим образом:

$$\Delta_\delta = \left\{ x \in \partial\mathbb{R}^n_+ : \|x\|_1 \le \delta \right\}.$$

Здесь $\|x\|_1$ — первая гельдеровская норма вектора x.

Определение 11.4.1. *Точка $x \in \partial\mathcal{P}_+$ называется нижней (верхней) вершиной множества $\mathcal{P}_+$, если можно указать $\delta > 0$ такое, что*

$$x + \Delta_\delta \subseteq \partial\mathcal{P}_+, \qquad (x - \Delta_\delta \subseteq \partial\mathcal{P}_+). \tag{11.4.10}$$

На рис. 11.7 нижние и верхние вершины множества $\mathcal{P}_+$ показаны соответственно заполненными и незаполненными кружочками. Как можно видеть, нижние вершины $\mathcal{P}_+$ совпадают с самими векторами r_j, $1 \le j \le N$.

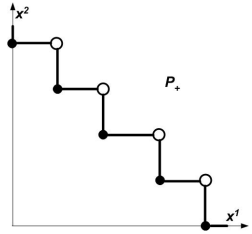

Рис. 11.7. Верхние и нижние вершины множества $\mathcal{P}_+$

Рассмотрим «юго-западную» часть границы $\partial\mathcal{P}_+$ множества $\mathcal{P}_+$:

$$\mathcal{P}_0 = \mathrm{cl}\left\{ x \in \partial\mathcal{P}_+ : \ \lambda x \notin \partial\mathcal{P}_+ \text{ для всех } \lambda > 1 \right\}.$$

Непосредственно из определения следует, что все нижние и верхние вершины множества $\mathcal{P}_+$ принадлежат $\mathcal{P}_0$. Более того, любая верхняя вершина является внутренней точкой ортанта $\mathbb{R}^n_+$.

Функция $h(x)$ через «обратные» векторы $r_1, \ldots, r_N$ может быть переписана в виде

$$h(x) = \max_{1 \le j \le N} \min_{i \in I_+(r_j)} \frac{x^i}{r^i_j}. \tag{11.4.11}$$

Если точка $x \in \partial\mathcal{P}_+$, то для нее найдется по крайней мере один индекс $1 \le j \le N$ такой, что $x \ge r_j$, причем в силу того, что x является

граничной точкой множества $\mathcal{P}_+$, выполняется равенство $x^i = r_j^i$ для некоторых компонент x^i, $1 \le i \le n$, в том числе, и для $i \in I_+(r_j)$. Отсюда следует, что

$$\min_{i \in I_+(r_j)} \frac{x^i}{r_j^i} = 1. \tag{11.4.12}$$

Для остальных векторов r_j, для которых $x \notin r_j + \mathbb{R}^n_+$, получаем, что $x^i < r_j^i$ хотя бы для одного индекса $i \in I_+(r_j)$. Поэтому

$$\min_{i \in I_+(r_j)} \frac{x^i}{r_j^i} < 1. \tag{11.4.13}$$

Сравнивая (11.4.12) и (11.4.13), приходим к выводу, что

$$h(x) = 1, \quad x \in \partial\mathcal{P}_+. \tag{11.4.14}$$

В отличие от (11.4.14) для внутренних точек множества $\mathcal{P}_+$ получаем

$$h(x) > 1, \quad x \in \operatorname{int}\mathcal{P}_+.$$

Таким образом, если $x \in \mathcal{P}_+$, то $x \in \partial\mathcal{P}_+$ в том и только том случае, когда выполнено равенство (11.4.14).

Утверждение 11.4.2. *Пусть точка $x_* \in \mathbb{R}^n_{++}$ есть нижняя вершина множества P_+. Тогда точка $\bar{x} = \lambda x_*$, где $\lambda = \left(\sum_{i=1}^n x_*^i\right)^{-1}$, принадлежит симплексу Λ^n и является точкой локального максимума функции $h(x)$ на Λ^n.*

Доказательство. Из вида (11.4.11) функции $h(x)$ следует, что имеет место тождество $h(x) \equiv 1$ для всех $x \in \mathcal{P}_0$. Но $h(x)$ является IPH-функцией. Поэтому $h(\lambda x) = \lambda h(x)$ для всех $x \in \mathcal{P}_0$. Согласно определению, $x_* + \Delta_\delta \subseteq \mathcal{P}_0$. Но тогда $\bar{x} + \Delta_{\delta_1} \subseteq \lambda\mathcal{P}_0$, где $\delta_1 = \lambda\delta$. Так как $(\bar{x} + \Delta_{\delta_1}) \bigcap \Lambda^n = \bar{x}$, то отсюда приходим к выводу, что $\bar{x}$ есть точка локального максимума функции $h(x)$ на симплексе Λ^n. ■

Из предположения о недоминированности всех векторов $r_1, \ldots, r_N$ и из определения (11.4.10) следует, что точка x может быть нижней вершиной множества $\mathcal{P}_+$ в том и только том случае, когда она совпадает с одним из векторов $r_1, \ldots, r_N$. Таким образом, у множества $\mathcal{P}_+$ имеется ровно N нижних вершин.

Утверждение 11.4.3. *Пусть точка $x_* \in \mathbb{R}^n_{++}$ есть верхняя вершина множества $\mathcal{P}_+$. Тогда точка $\bar{x} = \lambda x_*$, где $\lambda = \left(\sum_{i=1}^n x_*^i\right)^{-1}$,*

принадлежит симплексу Λ^n *и является точкой локального минимума функции* $h(x)$ *на* Λ^n.

Доказательство аналогично доказательству предыдущего утверждения. ■

На самом деле всем локальным минимумам функции $h(x)$ на симплексе Λ^n и только им соответствуют верхние вершины множества $\mathcal{P}_+$, принадлежащие границе $\mathcal{P}_0$. Отсюда следует, что для того, чтобы найти глобальный минимум функции $h(x)$ на Λ^n, другими словами, решить задачу (11.4.8), следует сначала перебрать все верхние вершины множества P_+, затем вычислить соответствующие точки $\bar{x}$, принадлежащие симплексу Λ^n, и сравнить значения функции $h(x)$ в этих точках. Любая точка $\bar{x}$, в которой это значение минимально, и есть глобальное решение задачи (11.4.8).

Утверждение 11.4.4. *Пусть точка* x_* *является верхней вершиной множества* $\mathcal{P}_+$. *Тогда можно указать* n *различных векторов* $r_{k_1}, \ldots, r_{k_n}$ *таких, что*

$$r^i_{k_i} > r^i_{k_j} \tag{11.4.15}$$

для всех $i \in [1:n]$ *и любого* $j \in [1:n]$, $j \neq i$.

Данный результат есть не что иное, как необходимые условия локального минимума функции $h(x)$ на симплексе Λ^n. Справедливо и обратное, а именно: если точка $x_* \in P_0$ такова, что

$$x^* = \left[r^1_{k_1}, r^2_{k_2}, \ldots, r^n_{k_n}\right]^T,$$

где все числа $k_1, k_2, \ldots, k_n$ различны и для $r_{k_1}, r_{k_2}, \ldots, r_{k_n}$ выполняются неравенства (11.4.15), то x_* есть верхняя вершина множества $\mathcal{P}_+$. Согласно этому утверждению, для нахождения верхней вершины достаточно взять n различных векторов из множества $r_1, \ldots, r_N$ и проверить для них выполнение неравенств (11.4.15). Разумеется, дополнительно надо убедиться в том, что $x_* \in \mathcal{P}_0$.

В дальнейшем векторы $r_{k_1}, r_{k_2}, \ldots, r_{k_n}$ называются *базисом* верхней вершины x_*. Если составить из векторов $r_{k_1}, \ldots, r_{k_n}$ квадратную матрицу $\mathcal{R}$ порядка n, поместив их в виде столбцов,

$$\mathcal{R} = \begin{bmatrix} r^1_{k_1} & r^1_{k_2} & \ldots & r^1_{k_n} \\ r^2_{k_1} & r^2_{k_2} & \ldots & r^2_{k_n} \\ \ldots & \ldots & \ldots & \ldots \\ r^n_{k_1} & r^n_{k_2} & \ldots & r^n_{k_n} \end{bmatrix},$$

то эта матрица такова, что любой ее диагональный элемент является максимальным в строке, причем единственным. Верхняя вершина x_* совпадает с диагональю этой матрицы.

Вспомогательная задача (11.4.8) является наиболее трудоемкой в методе секущих углов, и от эффективности ее решения в сильной степени зависит эффективность самого метода. Поэтому для решения вспомогательной задачи было разработано несколько достаточно искусных рекурсивных алгоритмов, позволяющих находить либо одну верхнюю вершину, либо всю совокупность верхних вершин.

Ссылки на литературу и комментарии

Литература, посвященная численным методам решения оптимизационных задач, весьма обширна. Даже список только книг и учебных пособий состоит из нескольких сотен наименований. Среди них имеются как фундаментальные монографии, охватывающие методы, решения разных классов задач: [2], [4], [7], [10], [13], [15], [17], [21], [19], [28], [31], [36], [39], [42], [46], [48], [49], [52], [55], [57], [58], [63], [64], [65], [68], [71], [76], [80], [82], [83], [84], [86], [90], так и книги, посвященные методам решения отдельных специальных классов задач, например задачам линейного программирования или задачам безусловной минимизации и т.д.: [5], [14], [23], [25], [26], [30], [45], [54], [85], [93], [97]. Поскольку охватить их в достаточно полном объеме не представляется возможным, то, как и в первой части, приводятся ссылки лишь на некоторые монографии и учебные пособия, причем главным образом на русском языке. В них наряду с более детальными описаниями методов можно найти указания на первоисточники. Ссылки на статьи даются только при отсутствии описания метода в книгах. Рассмотрим далее каждый раздел отдельно. Материал большинства разделов взят из [13], [28], [39], [64], [76], [86].

Глава 1. Основные определения, касающиеся сходимости итерационных методов, а также скорости сходимости, как отмечается в этой главе, заимствованы из вычислительной математики. Более общие сведения об этих понятиях можно найти во многих книгах по методам оптимизации, например в [28], а также в литературе по численным методам решения систем нелинейных уравнений: [8], [59], [62].

Глава 2. Методы минимизации унимодальных функций на отрезке выделяются своей спецификой. Важным качеством таких методов является их эффективность, понимаемая как число обращений к вы-

числению значений функции для достижения заданной точности. Приведенное здесь определение унимодальной функции взято из [76]. Известны также другие определения унимодальных функций и их обобщений, например [83]. Более подробное описание методов минимизации унимодальных функций, а также других приближенных методов одномерной минимизации дано, например, в [4], [39], [49], [76], [83].

В настоящее время известно много различных методов решения задач безусловной минимизации. Среди них метод градиентного спуска, метод Ньютона, метод сопряженных градиентов, которые рассматриваются в пособии, занимают особое место. Они приводятся практически во всех монографиях и учебных пособиях, посвященных численным методам, например в [4], [7], [13], [21], [28], [26], [39], [42], [46], [49], [64], [65], [76], [86]. Особенно обширна литература по методу Ньютона, см., например, [88]. Поскольку метод Ньютона для задач безусловной минимизации тесно связан со своим аналогом, предназначенным для решения систем нелинейных уравнений, то подробное описание метода Ньютона и его различных вариантов приводится и в литературе по вычислительной математике [8],[59], [62]. Другие методы безусловной минимизации, главным образом эвристического характера, можно найти, например, в [80]. В [25] рассмотрены методы минимизации негладких функций, из которых в настоящем пособии упомянут лишь субградиентный метод.

Глава 3. Для решения задач линейного программирования разработано большое количество численных методов, написаны специальные пакеты программ. Литература по методам линейного программирования наиболее обширна. Среди них методам симплексного типа уделяется особое внимание. Описание симплекс-метода приводится во многих книгах и учебных пособиях, где рассматриваются задачи линейного программирования: [2], [13], [16], [42]. Кроме того, имеется целый ряд монографий, специально посвященных методам этого типа: [5], [11], [14], [85], [23], [54], [67], [90], [97]. Укажем также учебники и задачники, где решение задач линейного программирования симплекс-методом иллюстрируется на простейших примерах: [5], [61].

Описание методов решения задач выпуклого квадратичного программирования с линейными ограничениями можно найти, например, в книгах [7], [24], [45].

Глава 4. Методы проекции градиента и условного градиента, предназначенные для минимизации гладких целевых функций на допустимых множествах простого вида, являются естественным обобщением метода градиентного спуска. Поэтому описание этих методов можно найти во многих книгах, где рассматриваются методы спуска, напри-

мер в [4], [28], [39], [42], [52], [64], [65], [82]. Изложение метода приведенных направлений взято из книг [28] и [82].

Глава 5. Идея использования линейных или квадратичных приближений для решения задач условной оптимизации с функциональными ограничениями вполне понятна. На основе этих идей были разработаны, в частности, различные методы градиентного типа. К их числу относится метод возможных направлений. Рассматриваемый в пособии вариант метода принадлежит Г. Зойтендейку [38]. Независимо он был переоткрыт С.И. Зуховицким, Р.А. Поляком и М.Е. Примаком [37]. Другой вариант метода, рассмотренный Д. Топкисом и А. Вейнотом можно найти, например в [63]. Метод линеаризации был предложен и обоснован Б.Н. Пшеничным [66], его можно трактовать как метод минимизации негладкой точной штрафной функции. Приведенный в пособии вариант метода линеаризации отличается от метода Б.Н. Пшеничного лишь использованием другой нормировки. Достаточно полное изложение методов последовательного квадратичного программирования можно найти, например, в книге [39].

Глава 6. Методы последовательной безусловной минимизации, особенно методы штрафных функций, являются одними из основных практических методов, предназначенных для нахождения приближенных решений задач нелинейного программирования. Впоследствии эти приближенные решения могут быть уточнены с помощью других более эффективных методов локального характера. Поэтому методы последовательной безусловной минимизации достаточно хорошо изучены, им посвящена специальная литература, например [77], [22]. Описания методов можно найти также в многих книгах как учебного, так и научного характера [7], [13], [28], [39], [42], [49], [63], [64], [76].

Глава 7. Литература по численным методам, основанным на использовании функции Лагранжа и ее модификациям, весьма обширна. Здесь сошлемся лишь на монографии [9], [20], в которых приведена подробная библиография работ в этом направлении. Среди других книг, где рассматриваются эти методы, следует указать [28], [39], [42]. Изложение прямых и двойственных методов МФЛ, представленных в настоящей книге и использующих простейшую модификацию функции Лагранжа, в основном следует работе [32]. Близкие варианты прямых методов МФЛ рассматривались также в [3]. Описание метода Удзавы дается, например, в книге [84]. Известно, однако, что даже в случае выпуклой целевой функции он не всегда обладает сходимостью. Более подробные сведения о теореме А. Островского можно найти в [59], [60].

Глава 8. Под методами внутренней точки в широком смысле понимают численные методы, предназначенные для решения задач услов-

ной оптимизации, в которых помимо функциональных ограничений присутствует допустимое базовое множество простой структуры. Методы строятся таким образом, чтобы все точки в ходе вычислений принадлежали внутренности этого базового множества. К числу таких задач относится и задача линейного программирования, рассмотренная в данной главе. В качестве базового множества здесь выступает неотрицательный ортант пространства. По-видимому, первый метод внутренней точки для задачи линейного программирования был предложен И.И. Дикиным [27].

Возможны разные подходы к построению методов внутренней точки для задач линейного программирования. В главе рассматриваются лишь некоторые из них. В частности, описание мультипликативно-барьерного метода и метода Кармаркара взято из [32]. Метод Кармаркара оказался первым полиномиальным методом решения задач линейного программирования, практически сопоставимый с симплекс-методом по эффективности на задачах большой размерности. Он приводится также в [93], [97]. Прямо-двойственные методы центрального пути оказались наиболее эффективными полиномиальными методами решения задач линейного программирования. Среди монографий, где рассматриваются такие методы, сошлемся, в частности, на [93], [98]. Приведенное описание прямо-двойственного метода следует [87].

Глава 9. Среди первых публикаций, посвященных вопросам сложности непрерывных оптимизационных задач и трудоемкости методов их решения следует указать [56]. В дальнейшем эта теория получила существенное развитие. В 1994 году А. Немировский и Ю.Е. Нестеров опубликовали фундаментальную монографию [92], где, в частности, было введено понятие самосогласованной функции и предложены эффективные методы минимизации таких функций. Материал, вошедший в данную главу, полностью заимствован из [57].

Глава 10. Численные методы решения задач многокритериальной оптимизации основаны обычно на предварительном сведе́нии таких задач к задачам линейного или нелинейного программирования путем сворачивания с помощью линейной свертки или свертки Ю.Б. Гермейера [18] нескольких критериев в единый критерий. В дальнейшим получившаяся задача решается с помощью известных методов линейного и нелинейного программирования. Выбор разных решений из множества Парето-оптимальных решений при этом осуществляется за счет двойственных множителей. В пособии показывается, что возможен и другой подход, когда методы нелинейного программирования обобщаются на случай задач с несколькими критериями. За выбор конкретных решений теперь отвечают специальные оценки, которые пересчи-

тываются в ходе итерационного процесса и стремятся вдоль заданных направлений к конкретным решениям в критериальном пространстве. Впервые данный способ нахождения Парето-оптимальных решений был рассмотрен в [34]. В [89] излагается подход к диалоговому построению и визуализации всей паретовской границы, что дает возможность лицу, принимающему решение, выбрать нужное единственное решение. Многие другие подходы к нахождению компромиссных решений можно найти в [44], [47] [43], [50], [70], [81], [72].

Глава 11. Для решения задач глобальной оптимизации предложены разнообразные численные методы, многие из которых существенным образом учитывают специфику решаемых задач. Приведенные в данной главе методы, предназначенные для минимизации липшицевых функций, являются сравнительно универсальными. Описание метода ломанных, а также кривых Пеано–Гильберта можно найти в [69], [74]. Метод неравномерных покрытий был предложен Ю.Г. Евтушенко. Более эффективные варианты этого метода, использующие, в частности, параллельные вычисления, излагаются в [29]. Метод секущих углов разработан А.М. Рубиновым и М.Ю. Андрамоновым. Наиболее полное изложение метода дается в [94]. О некоторых других подходах к отысканию глобальных экстремумов функций, отличных от затронутых в данном пособии, можно прочесть, например, в книгах [35], [51], [69], [75], [74], [78], [96].

В заключение приведенного краткого обзора укажем список рекомендованной литературы, где дается описание методов как вошедших в данное пособие, так и многих других.

1) Васильев Ф.П. Методы оптимизации. Часть I. Конечномерные задачи оптимизации. Принцип максимума. Динамическое программирование [13].

2) Евтушенко Ю.Г. Методы решения экстремальных задач и их применение в системах оптимизации [28].

3) Измаилов А.Ф., Солодов М.В. Численные методы оптимизации [39].

4) Нестеров Ю.Е. Введение в выпуклую оптимизацию [57].

5) Поляк Б.Т. Введение в оптимизацию [64].

6) Сухарев А.Г., Тимохов А.В., Федоров В.В. Курс методов оптимизации [76].

Обратим также внимание на сборники задач и упражнений [1], [6], [12], [40], [53], [61], в которых можно найти ряд дополнительных полезных сведений о численных методах оптимизации.

ЛИТЕРАТУРА

1. Алексеев В.М., Галеев Е.М., Тихомиров В.М. Сборник задач по оптимизации. Теория. Примеры. Задачи. М.: Физматлит, 2005. 256 с.
2. Андреева Е.А., Цирулева В.М. Вариационное исчисление и методы оптимизации: учебное пособие. Тверь: Тверской государственный университет, 2001. 576 с.
3. Антипин А.С. Методы нелинейного программирования, основанные на прямой и двойственной модификации функции Лагранжа: препринт ВНИИСИ. М., 1979. 74 с.
4. Аоки М. Введение в методы оптимизации. М.: Наука, 1977. 344 с.
5. Ашманов С.А. Линейное программирование: учебное пособие. М.: Наука, 1981. 340 с.
6. Ашманов С.А., Тимохов А.В. Теория оптимизации в задачах и упражнениях. М.: Наука, 1991. 448 с.
7. Базара М., Шетти К. Нелинейное программирование. Теория и алгоритмы. М.: Мир, 1982. 583 с.
8. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. М.: БИНОМ. Лаборатория знаний, 2004. 637 с.
9. Бертсекас Д. Условная оптимизация и методы множителей Лагранжа. М.: Радио и связь, 1987. 400 с.
10. Бирюков С.И. Оптимизация. Элементы теории и численные методы: учебное пособие. М.: МЗ-Пресс, 2003. 248 с.
11. Булавский В.А., Звягина Р.А., Яковлева М.А. Численные методы линейного программирования. М.: Наука, 1977.
12. Васильев О.В., Аргучинцев А.В. Методы оптимизации в задачах и упражнениях. М.: Физматлит, 1999. 208 с.
13. Васильев Ф.П. Методы оптимизации. Часть I. Конечномерные задачи оптимизации. Принцип максимума. Динамическое программирование. М.: Изд-во МЦМНО, 2011. 620 с.
14. Васильев Ф.П., Иваницкий А.Ю. Линейное программирование. М.: Факториал Пресс, 2008. 347 с.
15. Габасов Р.Ф., Кириллова Ф.М. Методы оптимизации. Минск: Изд-во БГУ, 1981. 352 с.
16. Габасов Р., Кириллова Ф.М. Методы линейного программирования. Часть 1. Общие задачи. М.: URSS, 2010. 174 с.
17. Галеев Э.М. Оптимизация. Теория, примеры, задачи. М.: URSS, 2006. 336 с.
18. Гермейер Ю.Б. Введение в теорию исследования операций. М.: Наука, 1971. 383 с.

19. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация. М.: Мир, 1985. 509 с.

20. Гольштейн Е.Г., Третьяков Н.В. Модифицированные функции Лагранжа. М.: Наука, 1989. 400 с.

21. Гороховик В.В. Конечномерные задачи оптимизации. Минск: Издательский центр БГУ, 2007. 240 с.

22. Гроссман К., Каплан А.А. Нелинейное программирование на основе безусловной минимизации. Новосибирск: Наука, 1981. 183 с.

23. Данциг Дж. Б. Линейное программирование, его применение и обобщения. М.: Прогресс, 1966. 600 с.

24. Даугавет В.А. Численные методы квадратичного программирования. СПб.: Изд-во С.-Петербургского университета, 2004. 128 с.

25. Демьянов В.Ф., Васильев Л.В. Недифференцируемая оптимизация. М.: Наука, 1981. 384 с.

26. Деннис Дж., Шнабель Р. Численные методы безусловной оптимизации и решения нелинейных уравнений. М.: Мир, 1988. 249 с.

27. Дикин И.И. Метод внутренних точек в линейном и нелинейном программировании. М.: Красанд, 2010. 120 с.

28. Евтушенко Ю.Г. Методы решения экстремальных задач и их применение в системах оптимизации. М.: Наука, 1982. 432 с.

29. Евтушенко Ю.Г., Малкова В.У., Станевичюс А.А. Параллельный поиск глобального экстремума функций многих переменных // Журнал вычислительной математики и математической физики. 2009. Т. 49, № 2. С. 255–269.

30. Еремин И.И. Теория линейной оптимизации. Екатеринбург: Институт математики и механики УрО РАН, 1999. 312 с.

31. Еремин И.И., Мазуров В.Д., Скарин В.Д., Хачай М.Ю. Математические методы в экономике. Екатеринбург: Уральский государственный университет, 2000. 280 с.

32. Жадан В.Г. Численные методы линейного и нелинейного программирования. Вспомогательные функции в условной оптимизации. М.: ВЦ РАН, 2002. 160 с.

33. Жадан В.Г. Методы оптимизации. Часть I. Введение в выпуклый анализ и теорию оптимизации: учебное пособие. М.: МФТИ, 2014. 272 с.

34. Жадан В.Г. Метод параметризации целевых функций в условной многокритериальной оптимизации // Журнал вычислительной математики и математической физики. 1986. Т.26, № 2. С. 177–189.

35. Жиглявский А.А., Жилинскас А.Г. Методы поиска глобального экстремума. М.: Наука, 1991. 248 с.

36. Зангвилл У.И. Нелинейное программирование. Единый подход. М.: Советское радио, 1973. 312 с.

37. Зуховицкий С.И., Авдеева Л.И. Линейное и выпуклое программирование. М.: Наука, 1967. 460 с.

38. Зойтендейк Г. Методы возможных направлений. М.: ИЛ, 1963. 176 с.

39. Измаилов А.Ф., Солодов М.В. Численные методы оптимизации. М.: Физматлит, 2003. 304 с.

40. Ириарт-Уррути Ж.-Б. Оптимизация и выпуклый анализ. Сборник задач и упражнений. Киев: Изд. комп. «Кит», 2004. 376 с.

41. Канторович Л.В. Математические методы в организации и планировании производства. Ленинград: ЛГУ, 1939. 67 с.

42. Карманов В.Г. Математическое программирование. М.: Наука, 1986. 286 с.

43. Кини Р.Л., Райфа Х. Принятие решений при многих критериях: предпочтения и замещения. М.: Радио и связь, 1981. 560 с.

44. Краснощеков П.С., Морозов В.В., Попов Н.М. Оптимизация в автоматизированном проектировании. М.: МАКС Пресс, 2008. 323 с.

45. Кюнци Г.П., Крелле В. Нелинейное программирование. М.: Советсткое радио, 1965. 304 с.

46. Лесин В.В., Лисовец Ю.П. Основы методов оптимизации. М.: Лань, 2011. 352 с.

47. Лотов А.В., Поспелова И.И. Многокритериальные задачи принятия решений. М.: МАКС Пресс, 2008. 200 с.

48. Лэсдон Л.С. Оптимизация больших систем. М.: Наука, 1975. 432 с.

49. Мину М. Математическое программирование. Теория и алгоритмы. М.: Наука, 1990. 578.

50. Михалевич В.С., Волкович В.А. Вычислительные методы исследования и проектирования сложных систем. М.: Наука, 1982. 287 с.

51. Михалевич В.С., Гупал А.М., Норкин В.И. Методы невыпуклой оптимизации. М.: Наука, 1987. 280 с.

52. Моисеев Н.Н., Иванилов Ю.П., Столярова Е.М. Методы оптимизации. М.: Наука, 1978. 352 с.

53. Морозов В.В., Сухарев А.Г., Федоров В.В. Исследование операций в примерах и задачах. М.: Высшая школа, 1986. 287 с.

54. Муртаф Б. Современное линейное программирование. Теория и практика. М.: Мир, 1984. 224 с.

55. Мухачева Э.А., Рубинштейн Г.Ш. Математическое программирование. Новосибирск: Наука, 1987. 274 с.

56. Немировский А.С., Юдин Д.Б. Сложность задач и эффективность методов оптимизации. М.: Наука, 1979. 384 с.

57. Нестеров Ю.Е. Введение в выпуклую оптимизацию. М.: Изд-во МЦНМО, 2010. 280 с.

58. Нурминский Е.А. Численные методы выпуклой оптимизации. М.: Наука, 1991. 168 с.

59. Ортега Дж., Рейнболдт В. Итерационные методы решения систем нелинейных уравнений со многими неизвестными. М.: Мир, 1975. 560 с.

60. Островский А. Решение уравнений и систем уравнений. М.: ИЛ, 1963. 220 с.

61. Пантелеев А.В., Летова Т.А. Методы оптимизации в примерах и задачах: учебное пособие. М.: Высшая школа, 2002. 544 с.

62. Петров И.Б., Лобанов А.И. Лекции по вычислительной математике: учебное пособие. М.: Интернет-университет информационных технологий; БИНОМ. Лаборатория знаний, 2009. 523 с.

63. Полак Э. Численные методы оптимизации. Единый подход. М.: Мир, 1974. 376 с.

64. Поляк Б.Т. Введение в оптимизацию. Изд. 2-е, испр. и доп. М.: ЛЕНАНД, 2014. 392 с.

65. Пшеничный Б.Н., Данилин Ю.М. Численные методы в экстремальных задачах. М.: Наука, 1975. 320 с.

66. Пшеничный Б.Н. Метод линеаризации. М.: Наука, 1983. 136 с.

67. Романовский И.В. Алгоритмы решения экстремальных задач. М.: Наука, 1977. 352 с.

68. Сеа Ж. Оптимизация. Теория и алгоритмы. М.: Мир, 1973. 244 с.

69. Сергеев Я.Д., Квасов Д.Е. Диагональные методы глобальной оптимизации. М.: Физматлит, 2008. 352 с.

70. Соболь И.М., Статников Р.Б. Выбор оптимальных параметров в задачах со многими критериями. М.: Дрофа, 2006. 178 с.

71. Соколов А.В., Токарев В.В. Методы оптимальных решений. Том 1. Общие положения. Математическое программирование. М.: Физматлит, 2011. 564 с.

72. Соколов А.В., И.И. Токарев В.В. Методы оптимальных решений. Том 2. Многокритериальность. Динамика. Неопределенность. М.: Физматлит, 2011. 418 с.

73. Стрекаловский А.С. Элементы невыпуклой оптимизации. Новосибирск: Наука, 2003. 356 с.

74. Стронгин Р.Г. Численные методы в многоэкстремальных задачах. Информационно-статистический подход. М.: Наука, 1978. 242 с.

75. Сухарев А.Г. Глобальный экстремум и методы его отыскания / Математические методы в исследовании операций. М.: Изд-во МГУ, 1981. С. 4–37.

76. Сухарев А.Г., Тимохов А.В., Федоров В.В. Курс методов оптимизации. М.: Физматлит, 2008. 328 с.

77. Фиакко А., Мак-Кормик Г. Нелинейное программирование. Методы последовательной безусловной минимизации. М.: Мир, 1972. 240 с.

78. Хансен Э., Уолстер Дж. Глобальная оптимизация с помощью методов интервального анализа. М.: Изд-во УдГУ, 2012. 516 с.

79. Хачиян Л.Г. Избранные труды. М.: МЦНМО, 2009. 520 с.

80. Химмельблау Д. Прикладное нелинейное программирование. М.: Мир, 1975. 236 с.

81. Штойер Р. Многокритериальная оптимизация. Теория, вычисления и приложения. М.: Радио и связь, 1992. 504 с.

82. Численные методы условной оптимизации / под ред. Ф. Гилл, У. Мюррей. М.: Мир, 1977. 290 с.

83. Эльстер К.-Х., Рейнгардт Р., Шойбле М., Донат Г. Введение в нелинейное программирование. М.: Наука, 1985. 264 с.

84. Эрроу К.Дж., Гурвиц Л., Удзава Х. Исследования по линейному и нелинейному программированию. М.: ИЛ, 1962. 335 с.

85. Юдин Д.Б., Гольштейн Е.Г. Задачи и методы линейного программирования. Конечные методы. М.: URSS, 2010. 185 с.

86. Boyd S., Vandenberge L. Convex Optimization. Cambridge University Press, 2004. 716 p.

87. Gondzio J. Interior point methods 25 years later // European Journal of Operational Research. V. 218, № 3, 2012. P. 587–601.

88. Izmailov A.F., Solodov M.V. Newton-Type methods for Optimization and Variational Problems. Cham: Springer, 2014. 573 p.

89. Lotov A.V., Bushenkov V.A., Kamenev G.K. Interactive decision maps. Approximation and Visualization of Pareto frontier. Kluwer Academic Publishers, 2004. 310 p.

90. Luenberger D.G. Linear and nonlinear programming. London: Addison-Wesle, 1984. 551 p.

91. Nemirovski A. Five lectures on modern convex optimization // CORE Summer School on Modern Convex Optimization. August 26–30, 2002. 240 p.

92. Nesterov Yu., Nemirovskii A. Interior-Point Polynomial Algjrithms in Convex Programming. SIAM Publications. Philadelphia: SIAM, 1994. 405 p.

93. Roos C., Terlaky T., Vial J.-Ph. Theory and Algorithms for Linear Optimization. An Interior Point Approach. Chichester, New York, Weinheim: John Wiley @ Sons, 1997. 483 p.

94. Rubinov A. Abstract Convexity and Global Optimization. Springer and Business Media, 2000. 490 p.

95. Sawaragi Yo., Nakayama H., Tanino T. Theory of multiobjective optimization. Orlando, San Diego, New-York, London: Academic Press, Inc., 1985. 291 p.

96. Strongin R.G., Sergeev Ya.D. Global optimization with non-convex constraints: Sequential and parallel algorithms. Dordrecht, Kluwer Academic Publishers, 2000. 728 p.

97. Vanderbei R.J. Linear programming. Foundations and extensions. Boston, London, Dordrecht: Kluwer Academic Publishers, 1997. 418 p.

98. Wright S.J. Primal-Dual Interior-Point Methods. Philadelphia: SIAM, 1997. 289 p.

Учебное издание

Жадан Виталий Григорьевич

МЕТОДЫ ОПТИМИЗАЦИИ

ЧАСТЬ II

ЧИСЛЕННЫЕ АЛГОРИТМЫ

Редактор *О. П. Котова*. Корректор *Н. Е. Кобзева*
Компьютерная верстка *Н. Е. Кобзева*

Подписано в печать 20.08.2015. Формат 60×84 $^{1}/_{16}$.
Усл. печ. л. 20,0. Уч.-изд. л. 19,0. Тираж 300 экз. Заказ № 329.

Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Московский физико-технический институт (государственный университет)»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. (495) 408-58-22, e-mail: rio@mail.mipt.ru

Отдел оперативной полиграфии «Физтех-полиграф»
141700, Московская обл., г. Долгопрудный, Институтский пер., 9
Тел. (495) 408-84-30, e-mail: polygraph@mail.mipt.ru