Ф.П.ВАСИЛЬЕВ

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

Ф. П. ВАСИЛЬЕВ

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

Издание второе, переработанное и дополненное

Допущено Государственным комитетом СССР по народному образованию в качестве учебного пособия для студентов вузов, обучающихся по специальности «Прикладная математика»

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1988

Made with ScanKromsator

ББК 22.19
В19
УДК 519.6 (075.8)

Васильев Ф. П. **Численные методы решения экстремальных задач:** Учеб. пособие для вузов.— 2-е изд., перераб. и доп.— М.: Наука. Гл. ред. физ.-мат. лит., 1988.— 552 с.— ISBN 5-02-013796-0.

Содержит основные численные методы решения экстремальных задач. Приводятся теоретическое обоснование и краткие характеристики этих методов. Рассматриваются задачи минимизации функций конечного числа переменных и задачи оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений.

Сохранена структура первого издания, но содержание некоторых глав существенно переработано и дополнено.

1-е издание — в 1980 г.

Для студентов вузов по специальности «Прикладная математика», а также для специалистов, связанных с решением задач оптимизации.

Табл. 11. Ил. 42. Библиогр. 343 назв.

Рецензент

член-корреспондент АН СССР *Л. Д. Кудрявцев*

В $\frac{1702070000-191}{053(02)-88}$ 68-88

ISBN 5-02-013796-0

Scanned with Master Scanner

Издательство «Наука». Главная редакция физико-математической литературы, 1980; с изменениями, 1988

ОГЛАВЛЕНИЕ

Предисловие ко второму изданию 5

Предисловие 6

Глава 1. Методы минимизации функций одной переменной . . 9

§ 1. Постановка задачи 9
§ 2. Классический метод 15
§ 3. Метод деления отрезка пополам 17
§ 4. Метод золотого сечения. Симметричные методы . . . 19
§ 5. Об оптимальных методах 22
§ 6. Метод ломаных 28
§ 7. Методы покрытий 33
§ 8. Выпуклые функции одной переменной 38
§ 9. Метод касательных 45
§ 10. Метод поиска глобального минимума 53
§ 11. Метод парабол 59
§ 12. Другой метод поиска глобального минимума 62
§ 13. О методе стохастической аппроксимации 66

Глава 2. Предварительные сведения о задачах на экстремум . . 68

§ 1. Постановка задачи минимизации. Теорема Вейерштрасса 68
§ 2. Классический метод 78
§ 3. Вспомогательные предложения 91

Глава 3. Элементы линейного программирования 101

§ 1. Постановка задачи 101
§ 2. Геометрическая интерпретация. Угловые точки 106
§ 3. Симплекс-метод 113
§ 4. Антициклин 126
§ 5. Выбор начальной угловой точки 136
§ 6. Об условии разрешимости канонической задачи 145

Глава 4. Элементы выпуклого анализа 148

§ 1. Выпуклые множества 148
§ 2. Выпуклые функции 162
§ 3. Сильно выпуклые функции 181
§ 4. Проекция точки на множество 188
§ 5. Отделимость выпуклых множеств 193
§ 6. Субградиент. Субдифференциал 206
§ 7. Равномерно выпуклые функции 218
§ 8. Правило множителей Лагранжа 223
§ 9. Теорема Куна — Таккера. Двойственная задача . . . 234

Глава 5. Методы минимизации функций многих переменных . . 260

§ 1. Градиентный метод 260
§ 2. Метод проекции градиента 277
§ 3. Метод проекции субградиента 285

§ 4. Метод условного градиента 291
§ 5. Метод возможных направлений 299
§ 6. Метод линеаризации 309
§ 7. Квадратичное программирование 314
§ 8. Метод сопряженных направлений 320
§ 9. Метод Ньютона 329
§ 10. Метод Стеффенсена 338
§ 11. Метод покоординатного спуска 342
§ 12. Метод поиска глобального минимума 347
§ 13. Метод модифицированных функций Лагранжа 356
§ 14. Метод штрафных функций 363
§ 15. Метод барьерных функций 384
§ 16. Метод нагруженных функций 396
§ 17. О методе случайного поиска 410
§ 18. Общие замечания 415

Глава 6. **Принцип максимума Понтрягина** 421
§ 1. Постановка задачи оптимального управления 421
§ 2. Формулировка принципа максимума. Примеры 435
§ 3. Доказательство принципа максимума 461
§ 4. О методах решения краевой задачи принципа максимума 480
§ 5. Связь между принципом максимума и классическим вариационным исчислением 485

Глава 7. **Динамическое программирование** 490
§ 1. Схема Беллмана. Проблема синтеза для дискретных систем 490
§ 2. Схема Моисеева 505
§ 3. Проблема синтеза для систем с непрерывным временем . 513
§ 4. Достаточные условия оптимальности 522

Список литературы 531
Основная литература 531
Дополнительная литература 532

Предметный указатель 546

ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

Во втором издании книга существенно переработана. Добавлен новый материал, посвященный методу линеаризации, методу Стеффенсена, геометрическому и квадратичному программированию, изложены некоторые новые варианты градиентного метода, метода покрытий. Существенно переработаны параграфы, посвященные элементам выпуклого анализа, методу штрафных функций. Приведено простое доказательство принципа максимума Понтрягина для задачи оптимального управления с граничными условиями достаточно общего вида. Из книги исключены параграфы, содержащие доказательство оптимальности метода Фибоначчи. Исправлены замеченные ошибки, неточности.

Автор глубоко признателен С. М. Алиакбарову, А. С. Антипину, А. В. Арутюнову, С. С. Ахиеву, Е. Г. Белоусову, А. И. Беникову, В. А. Березневу, Н. С. Васильеву, О. В. Васильеву, Г. С. Ганшину, Ю. М. Данилину, Д. В. Денисову, Я. И. Заботину, С. К. Завриеву, В. С. Ижуткину, А. С. Ильинскому, А. Д. Искендерову, А. З. Ишмухаметову, А. Г. Коваленко, А. И. Кораблеву, Е. В. Лямину, М. Д. Марданову, Ю. Е. Нестерову, В. Н. Нефедову, В. И. Плотникову, М. М. Потапову, Т. Л. Рудневой, А. Г. Сухареву, А. Г. Тетереву, А. В. Тимохову, А. А. Третьякову, В. Р. Фазылову, Р. Ф. Хабибуллину, Ю. Н. Черемных, Н. Т. Чиричу, которые своими советами, предложениями, замечаниями способствовали улучшению второго издания книги.

ПРЕДИСЛОВИЕ

Первые задачи геометрического содержания, связанные с отысканием наименьших и наибольших величин, появились еще в древние времена. Развитие промышленности в XVII—XVIII веках привело к необходимости исследования более сложных задач на экстремум и к появлению вариационного исчисления. Однако лишь в XX веке при огромном размахе производства и осознании ограниченности ресурсов Земли во весь рост встала задача оптимального использования энергии, материалов, рабочего времени, большую актуальность приобрели вопросы наилучшего в том или ином смысле управления различными процессами физики, техники, экономики и др. Сюда относятся, например, задача организации производства с целью получения максимальной прибыли при заданных затратах ресурсов, задача управления системой гидростанций и водохранилищ с целью получения максимального количества электроэнергии, задача о космическом перелете из одной точки пространства в другую наибыстрейшим образом или с наименьшей затратой энергии, задача о быстрейшем нагреве или остывании металла до заданного температурного режима, задача о наилучшем гашении вибраций и многие другие задачи.

Потребности развития самой вычислительной математики также привели к необходимости исследования таких задач на максимум и минимум, как, например, задачи наилучшего приближения функций, оптимального выбора параметров итерационного процесса или узлов интерполирования, минимизации невязки уравнений и т. д.

На математическом языке такие задачи могут быть сформулированы как задачи отыскания экстремума (максимума или минимума) некоторой функции или функционала $J(u)$, выражающего собой качество (цену) управления u из заданного множества U некоторого пространства. Требование принадлежности управления u некоторому множеству U выражает собой ограничения, обычно вытекающие из законов сохранения, ограниченности наличных ресурсов, возможностей технической реализации управления, нежелательности каких-либо запрещенных (аварийных) состояний и т. п. Задачи отыскания экстремума функции $J(u)$ на множестве U принято называть экстремальными зада-

чами. Заметим, что задача максимизации функционала $J(u)$ на множестве U эквивалентна задаче минимизации функционала $-J(u)$ на том же множестве U, поэтому можно ограничиться рассмотрением задач минимизации.

В настоящее время теория экстремальных задач обогатилась фундаментальными результатами, появились ее новые разделы, такие как линейное, выпуклое, стохастическое программирование, оптимальное управление и др. Потребности практики способствовали бурному развитию методов приближенного решения экстремальных задач. Появление быстродействующих электронных вычислительных машин (ЭВМ) сделало возможным эффективное решение многих важных прикладных экстремальных задач, которые ранее из-за своей сложности представлялись недоступными.

В настоящей книге излагаются элементы теории экстремальных задач, а также основы наиболее часто используемых на практике методов приближенного решения экстремальных задач, теоретическое обоснование и краткая характеристика этих методов. Книга написана как учебное пособие для студентов факультетов и отделений прикладной математики университетов, технических вузов. В основу книги положен курс лекций по численным методам решения экстремальных задач, который автор в течение ряда лет читает на факультете вычислительной математики и кибернетики Московского университета.

В главе 1 излагаются методы минимизации функций одной переменной, в главах 2—5 рассматриваются задачи минимизации функций конечного числа переменных, в главах 6, 7 — задачи оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений. Часть текста, которая содержит материал, дополняющий и расширяющий основное содержание книги, напечатана петитом и при первом чтении может быть опущена.

Заманчиво было бы изложить теорию и методы минимизации сразу в общем виде на языке функционального анализа, охватив при этом как частный случай многие методы минимизации функций конечного числа переменных. Однако такой способ изложения, несмотря на свою привлекательность и удобства для читателя-математика, видимо, все же труден для первого знакомства с предметом, не говоря уже о том, что он не может отразить всю специфику конечномерных задач. Поэтому автор, стремясь сделать книгу доступной читателям, владеющим математикой в объеме программ технических вузов и впервые знакомящихся с теорией и методами решения экстремальных задач, в настоящей книге отобрал материал, не требующий для своего понимания знаний функционального анализа.

За пределами книги остались такие важные разделы теории и методов экстремальных задач, как задачи оптимального уп-

равления процессами, описываемыми уравнениями с частными производными, некорректные экстремальные задачи, аппроксимация и устойчивость экстремальных задач, методы минимизации в функциональных пространствах. Этим разделам теории и методам экстремальных задач, требующим для математически строгого изложения использования аппарата функционального анализа, автор предполагает посвятить отдельную книгу.

По рассматриваемым в книге проблемам имеется обширная библиография, насчитывающая много тысяч названий. Список литературы, который приведен в конце книги, содержит лишь некоторые работы, которые были непосредственно использованы в книге или близко примыкают к ней, дополняя ее содержание.

Нумерация формул, теорем, лемм, определений, упражнений в каждом параграфе самостоятельная; ссылки на материалы, расположенные в пределах данного параграфа, нумеруются одним числом, вне данного параграфа, но в пределах данной главы — двумя числами, вне данной главы — тремя числами. Так, например, теорема 3 из § 2 главы 4 в пределах этого параграфа именуется просто теоремой 3, в других параграфах 4-й главы — теоремой 2.3, в других главах — теоремой 4.2.3. Аналогично параграфы при ссылках на них в пределах данной главы нумеруются одним числом, а вне этой главы — двумя числами: первое число означает номер главы, второе — номер параграфа.

Автор выражает глубокую благодарность академикам А. Н. Тихонову и А. А. Самарскому за внимание и поддержку при написании книги, С. М. Цидилину и Ю. Н. Черемных, прочитавшим книгу в рукописи и сделавшим ряд ценных замечаний, Н. Л. Григоренко, взявшему на себя труд по научному редактированию книги и устранившему многочисленные погрешности изложения, а также Н. С. Бахвалову, И. С. Березину, В. И. Благодатских, В. Г. Карманову, М. Ковач, В. Л. Кулагину, М. С. Никольскому, М. М. Потапову, Н. А. Прохорову, В. Г. Сушко, В. В. Федорову, Б. М. Щедрину, М. Ячимовичу за многочисленные полезные дискуссии и советы, способствовавшие улучшению содержания книги.

В столь бурно развивающейся области, как теория и методы решения экстремальных задач, очень трудно создать учебное пособие, которое обладало бы определенной завершенностью и было бы свободным от недостатков, и поэтому автор будет признателен читателям за критические замечания по содержанию книги.

Ф. П. Васильев

Глава 1

МЕТОДЫ МИНИМИЗАЦИИ ФУНКЦИЙ ОДНОЙ ПЕРЕМЕННОЙ

С задачами минимизации функций одной переменной мы впервые сталкиваемся при изучении начальных глав математического анализа и решаем их методами дифференциального исчисления. Может показаться, что эти задачи относятся к достаточно простым и методы их решения хорошо разработаны и изучены. Однако это не совсем так. Методы дифференциального исчисления находят ограниченное применение и далеко не всегда удобны для реализации на современных ЭВМ. Хотя в последние десятилетия появились другие методы, более удобные для использования на ЭВМ, требующие меньшего объема вычислительного труда, но тем не менее эту область экстремальных задач никак нельзя считать завершенной. Работы, посвященные новым методам минимизации функций одной переменной, продолжают появляться на страницах математических книг и журналов (см., например, [11, 90, 95, 106, 128, 241, 246, 266, 279, 282, 288, 291, 314, 328, 332]). Мы здесь остановимся на некоторых наиболее известных методах, достаточно хорошо проявивших себя на практике.

§ 1. Постановка задачи

Пусть $\mathbf{R}=\{u\colon -\infty<u<\infty\}$ — числовая ось, U — некоторое множество из $\mathbf{R}$, $J(u)$ — функция, определенная на множестве U и принимающая во всех точках $u\in U$ конечные значения. Примерами множеств из $\mathbf{R}$ являются: отрезок $[a, b]=\{u\in\mathbf{R}\colon a\leqslant u\leqslant b\}$, интервал $(a, b)=\{u\in\mathbf{R}\colon a<u<b\}$, полуинтервалы $[a, b)=\{u\in\mathbf{R}\colon a\leqslant u<b\}$, $(a, b]=\{u\in\mathbf{R}\colon a<u\leqslant b\}$, где a, b — заданные числа. Будем рассматривать задачу минимизации функции $J(u)$ на множестве U. Начнем с того, что уточним постановку этой задачи. Для этого сначала напомним некоторые определения из классического математического анализа.

Определение 1. Точку $u_*\in U$ называют *точкой минимума* функции $J(u)$ на множестве U, если $J(u_*)\leqslant J(u)$ для всех $u\in U$; величину $J(u_*)$ называют *наименьшим* или *минимальным значением* $J(u)$ на U и обозначают $\min\limits_{u\in U} J(u)=J(u_*)$. Множество всех точек минимума $J(u)$ на U будем обозначать через U_*.

В зависимости от свойств множества U и функции $J(u)$ множество U_* может содержать одну, несколько или даже бесконечно много точек, а также возможны случаи, когда U_* пусто.

Пример 1. Пусть $J(u)=\sin^2(\pi/u)$ при $u\neq 0$ и $J(0)=0$. На множестве $U=\{u\colon 1\leqslant u\leqslant 2\}$ минимальное значение $J(u)$ равно нулю, множество U_* состоит из единственной точки $u_*=1$. Если $U=\{u\colon 1/3\leqslant u\leqslant 1\}$, то U_* содержит три точки: 1/3, 1/2, 1; если $U=\{u\colon 0<u\leqslant 1\}$, то $U_*=\{u\colon u=1/n,\ n=1, 2, \ldots\}$ — счетное множество. В случае $U=\{u\colon 2\leqslant u<\infty\}$ функция $J(u)$ не имеет наименьшего значения на U. В самом деле, какую бы точку $u\in U$ ни взять, найдется точка $v\in U$ (например, $v=k$ при достаточно большом k) такая, что $J(u)>J(v)$. Это значит, что U_* пусто.

Пример 2. Функция $J(u)=|u|+|u-1|-1$ на $U=\{u\colon |u|\leqslant 1\}$ принимает свое наименьшее значение, равное нулю, во всех точках отрезка $U_*=\{u\colon 0\leqslant u\leqslant 1\}$. Если $U=\{u\colon 1\leqslant u\leqslant 2\}$, то U_* содержит одну точку $u_*=1$; если $U=\{u\colon 1<u\leqslant 2\}$, то $U_*=\varnothing$.

Пример 3. Пусть $J(u)=u$ при $u\neq 0$ и $J(0)=1$. На множествах $U=\{u\colon 0\leqslant u\leqslant 1\}$ или $U=\{u\colon 0<u\leqslant 1\}$ эта функция не имеет наименьшего значения, т. е. $U_*=\varnothing$.

Пример 4. Пусть $J(u)=\ln u$, $U=\{u\colon 0<u\leqslant 1\}$. Здесь $U_*=\varnothing$, так как во всех точках из U функция принимает конечные значения, а для последовательности $u_k=1/k$ $(k=1, 2, \ldots)$ имеем $\lim\limits_{k\to\infty} J(u_k)=-\infty$.

Определение 2. Функция $J(u)$ называется *ограниченной снизу* на множестве U, если существует такое число M, что $J(u)\geqslant M$ для всех $u\in U$. Функция $J(u)$ *не ограничена снизу* на U, если существует последовательность $\{u_k\}\in U$, для которой $\lim\limits_{k\to\infty} J(u_k)=-\infty$.

В примерах 1—3 функции ограничены снизу на рассматриваемых множествах, а в примере 4 функция не ограничена.

В тех случаях, когда $U_*=\varnothing$, естественным обобщением понятия наименьшего значения функции является понятие нижней грани функции.

Определение 3. Пусть функция $J(u)$ ограничена снизу на множестве U. Тогда число J_* называют *нижней гранью* $J(u)$ на U, если: 1) $J_*\leqslant J(u)$ при всех $u\in U$; 2) для любого сколь угодно малого числа $\varepsilon>0$ найдется точка $u_\varepsilon\in U$, для которой $J(u_\varepsilon)<J_*+\varepsilon$. Если функция $J(u)$ не ограничена снизу на U, то в качестве нижней грани $J(u)$ на U принимается $J_*=-\infty$. Нижнюю грань $J(u)$ на U обозначают через $\inf\limits_{u\in U} J(u)=J_*$.

В примерах 1—3 $J_*=0$, а в примере 4 $J_*=-\infty$.

Если $U_*\neq\varnothing$, то, очевидно, нижняя грань $J(u)$ на U совпадает с наименьшим значением этой функции на U, т. е. $\inf\limits_{u\in U} J(u)=\min\limits_{u\in U} J(u)$. В этом случае говорят, что функция $J(u)$

на U достигает своей нижней грани. Подчеркнем, что $\inf\limits_{u\in U} J(u) = J_*$ всегда существует, а $\min\limits_{u\in U} J(u)$, как мы видели из примеров 1—4, не всегда имеет смысл. Введем еще два определения.

Определение 4. Последовательность $\{u_k\} \in U$ называется *минимизирующей* для функции $J(u)$ на множестве U, если

$$\lim_{k\to\infty} J(u_k) = \inf_{u\in U} J(u) = J_*.$$

Из определения и существования нижней грани следует, что минимизирующая последовательность всегда существует.

Определение 5. Скажем, что последовательность $\{u_k\}$ *сходится* к непустому множеству U, если $\lim\limits_{k\to\infty} \rho(u_k, U) = 0$, где $\rho(u_k, U) = \inf\limits_{u\in U} |u_k - u|$ — расстояние от точки u_k до множества U.

Заметим, что если $U_* \neq \varnothing$, то всегда существует минимизирующая последовательность, сходящаяся к U_*; например, можно взять стационарную последовательность $u_k = u_*$ $(k = 1, 2, \ldots)$, где u_* — какая-либо точка из U_*. Однако не следует думать, что при $U_* \neq \varnothing$ любая минимизирующая последовательность будет сходиться к U_*.

Пример 5. Пусть $J(u) = \dfrac{u^2}{1+u^4}$, $U = \mathbf{R}$. Очевидно, здесь $J_* = 0$ и множество U_* состоит из единственной точки $u_* = 0$. Последовательность $u_k = k$ $(k = 1, 2, \ldots)$ является минимизирующей, так как $\lim\limits_{k\to\infty} J(k) = 0$, но $\rho(u_k, U_*) = k$ не стремится к нулю.

Теперь можем перейти к формулировке задачи минимизации функции $J(u)$ на множестве U. В дальнейшем будем различать задачи двух типов. К *первому типу* отнесем задачи, в которых требуется определить величину $J_* = \inf\limits_{u\in U} J(u)$. Сразу же подчеркнем, что в задачах первого типа неважно, будет ли множество U_* точек минимума $J(u)$ на U непустым или оно пусто. Ко *второму типу* задач отнесем те задачи, у которых множество U_* непусто и требуется наряду с J_* найти какую-либо точку $u_* \in U_*$.

Заметим, что получить точное решение задачи первого или второго типа удается лишь в редких случаях. Поэтому на практике при решении задач первого типа обычно строят какую-либо минимизирующую последовательность $\{u_k\}$ для функции $J(u)$ на U и затем в качестве приближения для J_* берут величину $J(u_k)$ при достаточно большом k. Аналогично для приближенного решения задач второго типа достаточно построить минимизирующую последовательность $\{u_k\}$, которая сходится ко множеству U_* в смысле определения 5, и в качестве приближения для J_*

и точки $u_* \in U_*$ взять соответственно величину $J(u_k)$ и точку u_k при достаточно большом k.

Как показывает пример 5, в отличие от задач первого типа не всякая минимизирующая последовательность может быть использована для получения приближенного решения задач второго типа. Построение минимизирующих последовательностей, сходящихся ко множеству U_*, в общем случае требует привлечения специальных методов [6, 22]. В настоящей главе будем рассматривать лишь такие задачи второго типа, у которых любая минимизирующая последовательность сходится к U_*. Один такой класс задач дается следующей теоремой, называемой *теоремой Вейерштрасса.*

Теорема 1. *Пусть U — замкнутое ограниченное множество из* $\mathbf{R}$, *функция $J(u)$ непрерывна на U. Тогда $J(u)$ ограничена снизу на U, множество U_* точек минимума $J(u)$ на U непусто, замкнуто и любая минимизирующая последовательность $\{u_k\}$ сходится к U_*.*

Доказательство этой теоремы можно найти, например, в [10, 160, 165, 233]. Несколько более общий факт будет установлен в § 2.1, из которого также будет следовать теорема 1. Предлагаем читателю вернуться к примерам 1—5 и выяснить, в каких случаях и какое из условий теоремы 1 нарушено и к чему это приводит.

Возможна и более широкая постановка задач минимизации второго типа — когда ищутся не только точки минимума в смысле определения 1, но и точки так называемого локального минимума.

Определение 6. Точка $v_* \in U$ называется *точкой локального минимума* функции $J(u)$ на множестве U со значением $c = J(v_*)$, если существует такое число $\alpha > 0$, что $J(v_*) \leqslant J(u)$ для всех $u \in U \cap \{u\colon |u - v_*| < \alpha\} = O_\alpha(v_*)$. Если при некотором $\alpha > 0$ равенство $J(v_*) = J(u)$ для $u \in O_\alpha(v_*)$ возможно только при $u = v_*$, то v_* называют точкой *строгого локального минимума.*

Для функции, график которой изображен на рис. 1.1, точки u_0, u_2, u_4 являются точками строгого локального минимума, а в точках, удовлетворяющих неравенствам $u_5 < u \leqslant u_6$ и $u_8 \leqslant u \leqslant u_9$, реализуется нестрогий локальный минимум. Функция из примера 1 в точках $u_k = 1/k$ $(k = \pm 1, \pm 2)$ имеет строгий локальный минимум на $U = \mathbf{R}$, а в точке $u_* = 0$ — нестрогий локальный минимум.

Точки локального минимума, в которых минимум достигается в смысле определения 1, часто называют *точками глобального* или *абсолютного минимума* функции $J(u)$ на множестве U.

Выделим класс функций, у которых все точки локального минимума являются точками глобального минимума.

Определение 7. Функцию $J(u)$ назовем *унимодальной* на отрезке $U=[a, b]$, если она непрерывна на $[a, b]$ и существуют числа α, β $(a \leqslant \alpha \leqslant \beta \leqslant b)$ такие, что: 1) $J(u)$ строго монотонно убывает при $a \leqslant u \leqslant \alpha$ (если $a<\alpha$); 2) $J(u)$ строго монотонно возрастает при $\beta \leqslant u \leqslant b$ (если $\beta<b$); 3) $J(u)=J_* = \inf\limits_{u \in U} J(u)$ при $\alpha \leqslant u \leqslant \beta$, так что $U_* = [\alpha, \beta]$. Случаи, когда один или два

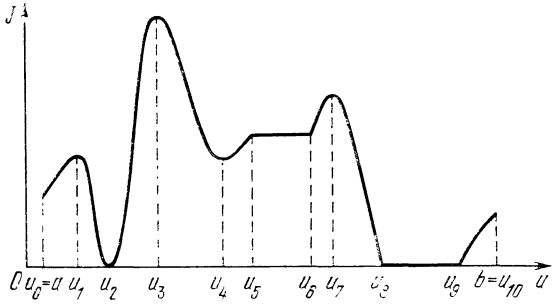

Рис. 1.1

из отрезков $[a, \alpha]$, $[\alpha, \beta]$, $[\beta, b]$ вырождаются в точку, здесь не исключаются. В частности, если $\alpha=\beta$, то $J(u)$ назовем *строго унимодальной* на отрезке $[a, b]$.

Функция из примера 2 унимодальна на любом отрезке $[a, b]$; функция из примера 1 строго унимодальна на $[2/3, 2]$, но не будет унимодальной на $[1/2, 2]$.

Нетрудно видеть, что если функция $J(u)$ унимодальна на $[a, b]$, то она остается унимодальной и на любом отрезке $[c, d] \subseteq [a, b]$.

В заключение кратко остановимся на задаче максимизации функции.

Определение 8. Функция $J(u)$ называется *ограниченной сверху* на множестве U, если существует такое число B, что $J(u) \leqslant B$ при всех $u \in U$. Функция $J(u)$ *не ограничена сверху* на U, если существует последовательность $\{u_k\} \in U$, для которой $\lim\limits_{k \to \infty} J(u_k) = \infty$. Функцию $J(u)$ называют *ограниченной* на U, если она ограничена на U сверху и снизу.

Определение 9. Если функция $J(u)$ ограничена сверху на U, то число J^* называется *верхней гранью* $J(u)$ на U в том случае, когда: 1) $J(u) \leqslant J^*$ для всех $u \in U$; 2) для любого числа $\varepsilon > 0$ найдется такая точка $u_\varepsilon \in U$, что $J(u_\varepsilon) > J^* - \varepsilon$. Если $J(u)$ не ограничена сверху на U, то по определению принимается $J^* = \infty$. Последовательность $\{u_k\} \in U$ называется *максимизирующей* для $J(u)$ на U, если $\lim\limits_{k \to \infty} J(u_k) = J^*$. Если существует такая точка $u^* \in U$, что $J(u^*) = J^*$, то u^* называется *точкой максиму-*

ма $J(u)$ на U, а величина $J(u^*)$ — *наибольшим* или *максимальным значением* $J(u)$ на U. Множество точек максимума $J(u)$ на U будем обозначать через U^*, верхнюю грань — через $J^* = \sup\limits_{u \in U} J(u)$.

Заметим, что верхняя грань и максимизирующая последовательность всегда существуют, а максимальное значение может не существовать. Если выполнены условия теоремы 1, то $J^* < \infty$, $U^* \neq \varnothing$ и любая максимизирующая последовательность $\{u_k\}$ сходится к U^*.

В задачах максимизации также можно различать задачи двух типов: в задачах *первого типа* ищется величина J^*, а в задачах *второго типа* ищется J^* и какая-либо точка $u^* \in U^*$. Нетрудно видеть, что

$$\sup_{u \in U} J(u) = -\inf_{u \in U} (-J(u)),$$

причем любая точка максимума и любая максимизирующая последовательность для $J(u)$ на U являются точкой минимума и соответственно минимизирующей последовательностью для функции $-J(u)$ на U. Это значит, что любая задача максимизации функции $J(u)$ на U равносильна задаче минимизации функции $-J(u)$ на том же множестве U. Поэтому мы можем ограничиться изучением лишь задач минимизации.

Наконец, немного о точках локального максимума.

Определение 10. Точка $v^* \in U$ называется *точкой локального максимума* функции $J(u)$ на множестве U, если существует такое число $\alpha > 0$, что $J(v^*) \geqslant J(u)$ для всех $u \in U \cap \{u\colon |u - v^*| < \alpha\} = O_\alpha(v^*)$. Если при некотором $\alpha > 0$ равенство $J(v^*) = J(u)$ для $u \in O_\alpha(v^*)$ возможно только при $u = v^*$, то v^* называют *точкой строгого локального максимума*.

Для функции, график которой изображен на рис. 1.1, точки u_1, u_3, u_7, u_{10} являются точками строгого локального максимума, а в точках, удовлетворяющих неравенствам $u_5 \leqslant u < u_6$ и $u_8 < u < u_9$, реализуется нестрогий локальный максимум; u_3 — точка глобального максимума.

Множество всех точек локального минимума и максимума функции на множестве U принято называть *точками локального экстремума* функции на этом множестве или, проще, *точками экстремума*.

Упражнения. 1. Построить минимизирующую и максимизирующую последовательности для функции $J(u) = \operatorname{arctg} u$ на $U = \mathbf{R}$. Достигает ли функция своих нижних и верхних граней на $\mathbf{R}$?

2. Пусть $J(u) = |u^2 - 1|$ при $u \neq 1$ и $J(1) = 1$. Найти множество U_* точек минимума $J(u)$ на $U = \mathbf{R}$. Можно ли утверждать, что любая минимизирующая последовательность для этой функции будет сходиться к U_*?

3. Найти все точки локального экстремума функции $J(u) = |||u^2 - 1| - 1| - 1|$ на отрезке $[a, b]$ при различных a, b. При каких a, b эта функция будет унимодальной на $[a, b]$?

4. Выяснить, на каких отрезках будут унимодальными функции $J(u) = e^u$, $J(u) = u^2$, $J(u) = -u^2$, $J(u) = \sqrt{|u|}$, $J(u) = \cos u$.

5. Если функция $G(v)$ унимодальна на отрезке $[c, d]$, то функция $J(u) = G((d-c)(u-a)/(b-a)+c)$ унимодальна на отрезке $[a, b]$. Доказать.

6. Доказать, что линейная функция $J(u) = Au + B$, где A, B — постоянные, $A \neq 0$, достигает своего минимума и максимума на отрезке $[a, b]$ только при $u = a$ или $u = b$.

7. Найти минимум функции $J(u) = \max\limits_{0 \leqslant t \leqslant 1} |t^2 - ut|$ на множествах $U = \mathbf{R}$ и $U = \{u\colon 1 \leqslant u < \infty\}$.

§ 2. Классический метод

Под классическим методом будем подразумевать тот подход к поиску точек экстремума функции, который основан на дифференциальном исчислении и подробно описан в учебниках по математическому анализу [10, 160, 165, 233]. Мы здесь лишь вкратце остановимся на этом методе.

Пусть функция $J(u)$ кусочно непрерывна и кусочно гладка на отрезке $[a, b]$. Это значит, что на $[a, b]$ может существовать лишь конечное число точек, в которых $J(u)$ либо терпит разрыв первого рода, либо непрерывна, но не имеет производной. Тогда, как известно, точками экстремума функции $J(u)$ на $[a, b]$ могут быть лишь те точки, в которых выполняется одно из следующих условий: 1) либо $J(u)$ терпит разрыв; 2) либо $J(u)$ непрерывна, но производная $J'(u)$ не существует; 3) либо производная $J'(u)$ существует и равна нулю; 4) либо $u = a$ или $u = b$. Такие точки принято называть *точками, подозрительными на экстремум.*

Поиск точек экстремума функции начинают с нахождения всех точек, подозрительных на экстремум. После того как такие точки найдены, проводят дополнительное исследование и отбирают среди них те, которые являются точками локального минимума или максимума. Для этого обычно исследуют знак первой производной $J'(u)$ в окрестности (или соответствующей полуокрестности граничных точек $u = a$ или $u = b$) подозрительной точки. Для того чтобы подозрительная точка $v \in [a, b]$ была точкой локального минимума, достаточно, чтобы $\lim\limits_{u \to v-0} J(u) \geqslant J(v)$, $\lim\limits_{u \to v+0} J(u) \geqslant J(v)$ и при некотором $\alpha > 0$ на множествах $[a, b] \cap (v, v+\alpha) = O_\alpha^+(v)$, $[a, b] \cap (v-\alpha, v) = O_\alpha^-(v)$ существовала производная $J'(u)$, причем $J'(u) > 0$ при $u \in O_\alpha^+(v)$ и $J'(u) < 0$ при $u \in O_\alpha^-(v)$. Если же $\lim\limits_{u \to v-0} J(u) \leqslant J(v)$, $\lim\limits_{u \to v+0} J(u) \leqslant J(v)$ и $J'(u) < 0$ при $u \in O_\alpha^+(v)$, $J'(v) > 0$ при $u \in O_\alpha^-(v)$, то v — точка локального максимума.

В тех случаях, когда удается вычислить в подозрительной точке производные второго и более высокого порядков, то их также можно использовать для исследования поведения функции

в окрестности этой точки. А именно, пусть известны производные $J'(v), \ldots, J^{(n)}(v)$, причем $J^{(i)}(v)=0$ $(i=1, \ldots, n-1)$, а $J^{(n)}(v)\neq 0$ $(n\geqslant 1)$. Если n — четное число, то в случае $J^{(n)}(v)>0$ в точке v реализуется локальный минимум, а в случае $J^{(n)}(v)<0$ — локальный максимум. Если n нечетно, то при $a<v<b$ в точке v не может быть локального минимума или максимума; при $v=a$ $(v=b)$ в случае $J^{(n)}(v)>0$ в точке v имеем локальный минимум (максимум), а в случае $J^{(n)}(v)<0$ — локальный максимум (минимум).

Чтобы найти глобальный минимум (максимум) функции $J(u)$ на $[a, b]$, нужно перебрать все точки локального минимума (максимума) на $[a, b]$ и среди них выбрать точку с наименьшим (наибольшим) значением функции, если таковое существует. Если вместо отрезка $[a, b]$ имеем дело со множеством $U=\{u\colon a\leqslant u<\infty\}$, или $U=\{u\colon -\infty<u\leqslant b\}$, или $U=\mathbf{R}$, то наряду с вышеописанными исследованиями нужно также изучить поведение функции при $u\to\infty$ или $u\to-\infty$.

Классический метод исследования функции на экстремум следует использовать во всех тех случаях, когда достаточно просто удается выявить все подозрительные на экстремум точки и реализовать описанную выше схему отбора экстремальных точек. К великому сожалению, классический метод имеет весьма ограниченное применение. Дело в том, что вычисление производной $J'(u)$ в практических задачах зачастую является непростым делом. Например, может оказаться, что значения функции $J(u)$ определяются из наблюдений или каких-либо физических экспериментов, и получить информацию о ее производной крайне трудно. Но даже в тех случаях, когда производную все же удается вычислить, решение уравнения $J'(u)=0$ и выявление других точек, подозрительных на экстремум, может быть связано с серьезными трудностями. Поэтому важно иметь также и другие методы поиска экстремума, не требующие вычисления производных, более удобные для реализации на современных ЭВМ.

Упражнения. 1. Найти точки экстремума функции $J(u)=\sin^3 u+\cos^3 u$ на отрезках $[0, 3\pi/4]$, $[0, 2\pi]$.

2. Пусть $J(u)=(1+e^{1/u})^{-1}$ при $u\neq 0$ и $J(0)=0$. Найти точки экстремума этой функции на отрезках $[0, 1]$, $[-1, 0]$, $[-1, 1]$, $[1, 2]$ и на $\mathbf{R}$.

3. Пусть непрерывная на отрезке $[a, b]$ функция $J(u)$ в точке v $(a<v<b)$ имеет строгий локальный минимум. Можно ли утверждать, что существует число $\alpha>0$ такое, что $J(u)$ монотонно убывает при $v-\alpha<u<v$ и монотонно возрастает при $v<u<v+\alpha$? Рассмотреть функцию $J(u)=2u^2+u^2\sin(1/u)$ $(u\neq 0)$, $J(0)=0$ на $[-1, 1]$. Исследовать случай, когда $J(u)$ имеет на $[a, b]$ конечное число точек локального экстремума.

4. Пусть функция $J(u)$ определена на $[a, b]$ и дважды дифференцируема в точке $v\in[a, b]$. Доказать, что если $a<v<b$ и в точке v реализуется локальный минимум $J(u)$, то необходимо, чтобы $J''(v)\geqslant 0$. Будет ли верным это утверждение, если $v=a$ или $v=b$? Будет ли оно верным, если $v=a$ или $v=b$ и, кроме того $J'(v)=0$? Рассмотреть функции $J(u)=-u^2$, $J(u)=\cos u$ на $[-\pi, \pi]$.

5. Пусть функция $J(u)$ определена на $[a, b]$ и в точке $v \in [a, b]$ имеет n производных ($n \geqslant 2$), причем известно, что $J^{(i)}(v) = 0$ при $i = 1, \ldots, n-1$ и $J^{(n)}(v) \neq 0$. Доказать, что если v — точка локального минимума и $a < v < b$, то n — четное число и $J^{(n)}(v) > 0$. Что изменится, если $v = a$ или $v = b$?

6. Пусть функция $J(u)$ аналитична на отрезке $[a, b]$, т. е. ряд Тейлора этой функции сходится к $J(u)$ во всех точках $[a, b]$. Может ли эта функция иметь на $[a, b]$ бесконечное число точек локального экстремума?

7. Пусть функция $J(u)$ определена на отрезке $[a, b]$ и в точке v имеет производные всех порядков. Можно ли утверждать, что если v — точка локального минимума, то $J^{(n)}(v) \neq 0$ при каком-либо $n \geqslant 1$? Рассмотреть функцию $J(u) = e^{-1/u^2}$ $(u \neq 0)$, $J(0) = 0$, в точке $v = 0$. Что изменится, если функция $J(u)$ аналитична на $[a, b]$?

8. Пусть функция $J(u)$ дифференцируема на отрезке $[a, b]$ и в точке $v \in [a, b]$ достигает своей нижней грани на $[a, b]$. Доказать, что тогда необходимо, чтобы $J'(v)(u - v) \geqslant 0$ при всех $u \in [a, b]$. Будет ли выполнения этого условия достаточно для того, чтобы в точке v достигалась нижняя грань $J(u)$ на $[a, b]$?

§ 3. Метод деления отрезка пополам

Простейшим методом минимизации функции одной переменной, не требующим вычисления производной, является метод деления отрезка пополам. Опишем его, предполагая, что минимизируемая функция $J(u)$ унимодальна на отрезке $[a, b]$. Поиск минимума $J(u)$ на $[a, b]$ начинается с выбора двух точек $u_1 = (a + b - \delta)/2$ и $u_2 = (a + b + \delta)/2$, где δ — постоянная, являющаяся параметром метода, $0 < \delta < b - a$. Величина δ выбирается вычислителем и может определяться целесообразным количеством верных десятичных знаков при задании аргумента u. В частности, ясно, что δ не может быть меньше машинного нуля ЭВМ, используемой при решении рассматриваемой задачи. Точки u_1, u_2 расположены симметрично на отрезке $[a, b]$ относительно его середины и при малых δ делят его почти пополам — этим и объясняется название метода.

После выбора точек u_1, u_2 вычисляются значения $J(u_1)$, $J(u_2)$ и сравниваются между собой. Если $J(u_1) \leqslant J(u_2)$, то полагают $a_1 = a$, $b_1 = u_2$; если же $J(u_1) > J(u_2)$, то полагают $a_1 = u_1$, $b_1 = b$. Поскольку $J(u)$ унимодальна на $[a, b]$, то ясно, что отрезок $[a_1, b_1]$ имеет общую точку с множеством U_* точек минимума $J(u)$ на $[a, b]$ и его длина равна

$$b_1 - a_1 = (b - a - \delta)/2 + \delta.$$

Пусть отрезок $[a_{k-1}, b_{k-1}]$, имеющий непустое пересечение с U_*, уже известен, и пусть $b_{k-1} - a_{k-1} = (b - a - \delta)/2^{k-1} + \delta > \delta$ $(k \geqslant 2)$. Тогда берем точки $u_{2k-1} = (a_{k-1} + b_{k-1} - \delta)/2$, $u_{2k} = (a_{k-1} + b_{k-1} + \delta)/2$, расположенные на отрезке $[a_{k-1}, b_{k-1}]$, симметрично относительно его середины, и вычисляем значения $J(u_{2k-1})$, $J(u_{2k})$. Если $J(u_{2k-1}) \leqslant J(u_{2k})$, то полагаем $a_k = a_{k-1}$, $b_k = u_{2k}$; если же $J(u_{2k-1}) > J(u_{2k})$, то полагаем $a_k = u_{2k-1}$, $b_k = b_{k-1}$. Длина

получившегося отрезка $[a_k, b_k]$ равна $b_k - a_k = (b-a-\delta)/2^k + \delta > \delta$ и $[a_k, b_k] \cap U_* \neq \varnothing$.

Если количество вычислений значений минимизируемой функции ничем не ограничено, то описанный процесс деления отрезка пополам можно продолжать до тех пор, пока не получится отрезок $[a_k, b_k]$ длины $b_k - a_k < \varepsilon$, где ε — заданная точность, $\varepsilon > \delta$. Отсюда имеем, что $k > \log_2((b-a-\delta)/(\varepsilon-\delta))$. Поскольку каждое деление пополам требует двух вычислений значений функции, то для достижения точности $b_k - a_k < \varepsilon$ требуется всего $n = 2k > 2\log_2((b-a-\delta)/(\varepsilon-\delta))$ таких вычислений.

После определения отрезка $[a_k, b_k]$ в качестве приближения ко множеству U_* можно взять точку $\bar{u}_n = u_{2k-1}$ при $J(u_{2k-1}) \leqslant J(u_{2k})$ и $\bar{u}_n = u_{2k}$ при $J(u_{2k-1}) > J(u_{2k})$, а значение $J(\bar{u}_n)$ может служить приближением для $J_* = \inf\limits_{u \in [a,b]} J(u)$. При таком выборе приближения для U_* будет допущена погрешность $\rho(\bar{u}_n, U_*) \leqslant \max\{b_k - \bar{u}_n; \bar{u}_n - a_k\} = (b-a-\delta)/2^k$. Если не требовать того, чтобы значение функции, принимаемое за приближение к J_*, было вычислено непременно в той же точке, которая служит приближением к U_*, то вместо $\bar{u}_n$ можно взять точку $v_n = (a_k + b_k)/2$ с меньшей погрешностью $\rho(v_n, U_*) \leqslant (b_k - a_k)/2 = (b-a-\delta)/2^{k+1} + \delta/2$ (здесь $k = n/2$ и δ достаточно мало).

Конечно, и в этом случае можно бы провести еще одно дополнительное вычисление значения функции в точке v_n и принять $J(v_n) \approx J_*$. Однако заметим, что на практике нередко встречаются функции, нахождение значения которых в каждой точке связано с большим объемом вычислений или дорогостоящими экспериментами, наблюдениями,— понятно, что здесь приходится дорожить каждым вычислением значения минимизируемой функции. В таких ситуациях возможно даже, что число n, определяющее количество вычислений значений функции, заранее жестко задано и превышение его недопустимо.

Из предыдущего следует, что методом деления отрезка пополам с помощью $n = 2k$ вычислений значений функции можно определить точку минимума унимодальной функции на отрезке $[a, b]$ в лучшем случае с точностью $\approx (b-a)2^{-1-n/2}$. Возникает вопрос, не существует ли методов, позволяющих с помощью того же числа вычислений значений функции решить задачу минимизации унимодальной функции поточнее? Оказывается, такие методы есть. Один из них будет описан в § 4.

В заключение отметим, что метод деления отрезка пополам без изменений можно применять для минимизации функций, не являющихся унимодальными. Однако в этом случае нельзя гарантировать, что найденное решение будет достаточно хорошим приближением к глобальному минимуму.

§ 4. Метод золотого сечения. Симметричные методы

Перейдем к описанию метода минимизации унимодальной функции на отрезке, столь же простого, как метод деления отрезка пополам, но позволяющего решить задачу с требуемой точностью при меньшем количестве вычислений значений функции. Речь пойдет о методе золотого сечения.

1. Как известно, *золотым сечением* отрезка называется деление отрезка на две неравные части так, чтобы отношение длины всего отрезка к длине большей части равнялось отношению длины большей части к длине меньшей части отрезка.

Нетрудно проверить, что золотое сечение отрезка $[a, b]$ производится двумя точками $u_1 = a + (3-\sqrt{5})(b-a)/2 = a + 0{,}381966011\ldots(b-a)$ и $u_2 = a + (\sqrt{5}-1)(b-a)/2 = a + 0{,}618033989\ldots(b-a)$, расположенными симметрично относительно середины отрезка, причем $a < u_1 < u_2 < b$, $(b-a)/(b-u_1) = (b-u_1)/(u_1-a) = (b-a)/(u_2-a) = (u_2-a)/(b-u_2) = (\sqrt{5}+1)/2 = 1{,}618033989\ldots$

Замечательно здесь то, что точка u_1 в свою очередь производит золотое сечение отрезка $[a, u_2]$, так как $u_2 - u_1 < u_1 - a = b - u_2$ и $(u_2-a)/(u_1-a) = (u_1-a)/(u_2-u_1)$. Аналогично точка u_2 производит золотое сечение отрезка $[u_1, b]$. Опираясь на это свойство золотого сечения, можно предложить следующий метод минимизации унимодальной функции $J(u)$ на отрезке $[a, b]$.

Положим $a_1 = a$, $b_1 = b$. На отрезке $[a_1, b_1]$ возьмем точки u_1, u_2, производящие золотое сечение, и вычислим значения $J(u_1)$, $J(u_2)$. Далее, если $J(u_1) \leqslant J(u_2)$, то примем $a_2 = a_1$, $b_2 = u_2$, $\bar{u}_2 = u_1$; если же $J(u_1) > J(u_2)$, то примем $a_2 = u_1$, $b_2 = b_1$, $\bar{u}_2 = u_2$. Поскольку функция $J(u)$ унимодальна на $[a, b]$, то отрезок $[a_2, b_2]$ имеет хотя бы одну общую точку с множеством U_* точек минимума $J(u)$ на $[a, b]$. Кроме того, $b_2 - a_2 = (\sqrt{5}-1)(b-a)/2$ и весьма важно то, что внутри $[a_2, b_2]$ содержится точка $\bar{u}_2$ с вычисленным значением $J(\bar{u}_2) = \min\{J(u_1); J(u_2)\}$, которая производит золотое сечение отрезка $[a_2, b_2]$.

Пусть уже определены точки $u_1, \ldots, u_{n-1}$, вычислены значения $J(u_1), \ldots, J(u_{n-1})$, найден отрезок $[a_{n-1}, b_{n-1}]$ такой, что $[a_{n-1}, b_{n-1}] \cap U_* \neq \varnothing$, $b_{n-1} - a_{n-1} = ((\sqrt{5}-1)/2)^{n-2}(b-a)$, и известна точка $\bar{u}_{n-1}$, производящая золотое сечение отрезка $[a_{n-1}, b_{n-1}]$ и такая, что $J(\bar{u}_{n-1}) = \min\limits_{1 \leqslant i \leqslant n-1} J(u_i)$ $(n \geqslant 2)$. Тогда в качестве следующей точки возьмем точку $u_n = a_{n-1} + b_{n-1} - \bar{u}_{n-1}$, также производящую золотое сечение отрезка $[a_{n-1}, b_{n-1}]$, вычислим значение $J(u_n)$.

Пусть для определенности $a_{n-1} < u_n < \bar{u}_{n-1} < b_{n-1}$ (случай $\bar{u}_{n-1} < u_n$ рассматривается аналогично). Если $J(u_n) \leqslant J(\bar{u}_{n-1})$, то

полагаем $a_n = a_{n-1}$, $b_n = \bar{u}_{n-1}$, $\bar{u}_n = u_n$; если же $J(u_n) > J(\bar{u}_{n-1})$, то полагаем $a_n = u_n$, $b_n = b_{n-1}$, $\bar{u}_n = \bar{u}_{n-1}$. Новый отрезок $[a_n, b_n]$ таков, что $[a_n, b_n] \cap U_* \neq \varnothing$, $b_n - a_n = ((\sqrt{5}-1)/2)^{n-1}(b-a)$, точка $\bar{u}_n$ производит золотое сечение $[a_n, b_n]$ и $J(\bar{u}_n) = \min\{J(u_n); J(\bar{u}_{n-1})\} = \min\limits_{1 \leqslant i \leqslant n} J(u_i)$.

Если число вычислений значений $J(u)$ заранее не ограничено, то описанный процесс можно продолжать, например, до тех пор, пока не выполнится неравенство $b_n - a_n < \varepsilon$, где ε — заданная точность. Если же число вычислений значений функции $J(u)$ заранее жестко задано и равно n, то процесс на этом заканчивается и в качестве решения задачи второго типа (см. § 1) можно принять пару $J(\bar{u}_n)$, $\bar{u}_n$, где $J(\bar{u}_n)$ является приближением для $J_* = \inf\limits_{u \in [a,b]} J(u)$, а точка $\bar{u}_n$ служит приближением для множества U_* с погрешностью

$$\rho(\bar{u}_n, U_*) \leqslant \max\{b_n - \bar{u}_n; \bar{u}_n - a_n\} = \\ = \frac{1}{2}(\sqrt{5}-1)(b_n - a_n) = \left(\frac{\sqrt{5}-1}{2}\right)^n (b-a) = A_n.$$

Вспомним, что с помощью метода деления отрезка пополам за $n = 2k$ вычислений значений функции $J(u)$ в аналогичном случае мы получили точку $\bar{u}_n$ с погрешностью

$$\rho(\bar{u}_n, U_*) \leqslant 2^{-n/2}(b - a - \delta) < 2^{-n/2}(b-a) = B_n.$$

Отсюда имеем $A_n/B_n = (2\sqrt{2}/(\sqrt{5}+1))^n \approx (0{,}87\ldots)^n$ — видно, что уже при небольших n преимущество метода золотого сечения перед методом деления отрезка пополам становится ощутимым.

2. Обсудим возможности численной реализации метода золотого сечения на ЭВМ. Заметим, что число $\sqrt{5}$ на ЭВМ неизбежно будет задаваться приближенно, поэтому первая точка $u_1 = a + (3-\sqrt{5})(b-a)/2$ будет найдена с некоторой погрешностью. Посмотрим, как повлияет эта погрешность на результаты последующих шагов метода золотого сечения. Обозначим $\Delta_n = b_n - a_n = ((\sqrt{5}-1)/2)^{n-1}(b-a)$ $(n = 1, 2, \ldots)$. Нетрудно проверить, что Δ_n является решением конечно-разностного уравнения $\Delta_{n-2} = \Delta_{n-1} + \Delta_n$, или

$$\Delta_n = \Delta_{n-2} - \Delta_{n-1}, \qquad n = 3, 4, \ldots, \tag{1}$$

с начальными условиями $\Delta_1 = b - a$, $\Delta_2 = b - u_1$.

Как известно [4, 54], линейно независимые частные решения этого уравнения имеют вид τ_1^n и τ_2^n $(n = 1, 2, \ldots)$, где $\tau_1 = (\sqrt{5}-1)/2$, $\tau_2 = -(\sqrt{5}+1)/2$ — корни характеристического урав-

нения $\tau^2+\tau-1=0$, а любое решение уравнения (1) представимо в виде

$$\Delta_n = A\tau_1^n + B\tau_n^2, \qquad n = 1, 2, \ldots, \tag{2}$$

где постоянные A и B однозначно определяются начальными условиями из линейной системы

$$A\tau_1 + B\tau_2 = \Delta_1, \quad A\tau_1^2 + B\tau_2^2 = \Delta_2. \tag{3}$$

При $\Delta_1 = b-a$, $\Delta_2 = bu_1$ из (3) имеем $A = 2(b-a)/(\sqrt{5}-1)$, $B=0$, и понятно, что формула (2) в этом случае дает уже известное нам решение $\Delta_n = \tau_1^{n-1}(b-a)$. Однако точка u_1 задана с погрешностью, поэтому в системе (3) вместо точного значения Δ_2 придется взять приближенное $\tilde{\Delta}_2 = \Delta_2 + \delta$. Тогда постоянные A, B из (3) определятся с соответствующими погрешностями: $\tilde{A} = A + \delta_1$, $\tilde{B} = B + \delta_2$, и вместо (2) с точными A, B будем иметь $\tilde{\Delta}_n = \tilde{A}\tau_1^n + B\tau_2^n$ $(n = 1, 2, \ldots)$. Поскольку $0 < \tau_1 = 0{,}6\ldots$ $\ldots < 1$, $|\tau_2| = 1{,}6\ldots > 1$, то погрешность $|\Delta_n - \tilde{\Delta}_n| = |\delta_1\tau_1^n + \delta_2\tau_2^n|$ с возрастанием n будет расти очень быстро. Это значит, что уже при не очень больших n отрезок $[a_n, b_n]$ и точки $\bar{u}_n$, $u_{n+1} = a_n + b_n - \bar{u}_n$ будут сильно отличаться от тех, которые получились бы при работе с точными данными. Численные эксперименты на ЭВМ также подтверждают, что метод золотого сечения в описанном выше виде практически неприменим уже при небольших n.

Как же быть? К счастью, имеется достаточно простая модификация метода золотого сечения, позволяющая избежать слишком быстрого возрастания погрешностей при определении точек u_n $(n \geqslant 2)$. А именно, на каждом отрезке $[a_n, b_n]$, содержащем точку $\bar{u}_n$ с предыдущего шага, при выборе следующей точки u_{n+1} нужно остерегаться пользоваться формулой $u_{n+1} = a_n + b_n - \bar{u}_n$, и вместо этого лучше непосредственно произвести золотое сечение отрезка $[a_n, b_n]$ и в качестве u_{n+1} взять ту из точек $a_n + (3-\sqrt{5})(b_n-a)/2$, $a_n + (\sqrt{5}-1)(b_n-a_n)/2$, которая наиболее удалена от $\bar{u}_n$ (здесь под $\sqrt{5}$ подразумевается какое-либо подходящее приближение этого числа). Конечно, после такой модификации метод золотого сечения, вообще говоря, теряет свойство симметричности и, быть может, уже не так красив, но зато вполне годится для приложений. Нетрудно видеть, что этот метод может применяться и без априорного знания о том, что минимизируемая функция унимодальна, но в этом случае полученное решение может оказаться далеким от глобального минимума.

3. Метод золотого сечения относится к классу так называемых *симметричных методов*. Дадим краткое описание произвольного симметричного метода минимизации функции $J(u)$ на отрезке $[a, b]$.

Первый шаг: на $[a, b]$ задается точка u_1 $(a < u_1 < b)$, полагается $a_1 = a$, $b_1 = b$, $\bar{u}_1 = u_1$ и вычисляется $J(u_1)$. Пусть уже сделано $n-1$ шагов $(n \geqslant 2)$ и найдены отрезок $[a_{n-1}, b_{n-1}]$ и точка $\bar{u}_{n-1}(a_{n-1} < \bar{u}_{n-1} < b_{n-1})$ с вычисленным значением $J(\bar{u}_{n-1})$, причем $\bar{u}_{n-1} \neq (a_{n-1} + b_{n-1})/2$. Тогда на следующем n-м шаге берется точка $u_n = a_{n-1} + b_{n-1} - \bar{u}_{n-1}$, расположенная внутри $[a_{n-1}, b_{n-1}]$, симметрично точке $\bar{u}_{n-1}$ относительно середины этого отрезка — отсюда происходит название методов. Затем вычисляется значение $J(u_n)$ и сравнивается с $J(\bar{u}_{n-1})$. Пусть для определенности $\bar{u}_{n-1} < u_n$ (случай $u_n < \bar{u}_{n-1}$ рассматривается аналогично). Тогда при $J(\bar{u}_{n-1}) \leqslant J(u_n)$ полагается $a_n = a_{n-1}$, $b_n = u_n$, $\bar{u}_n = \bar{u}_{n-1}$; если же $J(\bar{u}_{n-1}) > J(u_n)$, то $a_n = \bar{u}_{n-1}$, $b_n = b_{n-1}$, $\bar{u}_n = u_n$. Если $\bar{u}_n \neq (a_n + b_n)/2$, то процесс может быть продолжен дальше. Может оказаться, что $\bar{u}_n = (a_n + b_n)/2$,— в этом случае процесс заканчивается; при необходимости на $[a_n, b_n]$ можно продолжать поиск минимума аналогичным методом, начиная с выбора новой начальной точки $\bar{u}_n \neq (a_n + b_n)/2$.

Из описания симметричного метода видно, что всякий симметричный метод полностью определяется заданием отрезка $[a, b]$ и первой точки u_1 $(a < u_1 < b)$. Отсюда следует, что в качестве другой характеристики симметричного метода можно взять длины Δ_n отрезков $[a_n, b_n]$ $(n = 1, 2, \ldots)$, где $\Delta_1 = b - a$, $\Delta_2 = \max\{b - u_1;\ u_1 - a\}$. Очевидно, $\Delta_{n+1} \geqslant \Delta_n/2$ при всех $n \geqslant 1$. Как видим, симметричные методы весьма просты и, пожалуй, даже изящны. Однако все эти методы страдают тем же недостатком, что и метод золотого сечения: погрешность, допущенная в задании первой точки u_1, приводит к быстрому накапливанию погрешностей на дальнейших шагах, и уже при не очень больших n результаты будут сильно отличаться от тех, которые могли бы получиться при точной реализации симметричного метода с точными исходными данными.

Если симметричный метод таков, что для $\Delta_n = b_n - a_n$ выполнено условие

$$\Delta_n/2 < \Delta_{n+1} \leqslant 2\Delta_n/3, \qquad n = 1, \ldots, N, \tag{4}$$

при некотором $N > 1$, то Δ_n будут удовлетворять конечно-разностному уравнению (1) при $n = 2, \ldots, N$, и исследование поведения погрешностей в этом случае может быть проведено так же, как это было сделано выше для метода золотого сечения. Чтобы избежать слишком быстрого роста погрешностей в симметричных методах со свойством (4), на каждом отрезке $[a_n, b_n]$ $(n = 2, \ldots, N)$, содержащем точку $\bar{u}_n$ с предыдущего шага, следующую точку u_{n+1} нужно определять не по формуле $u_{n+1} = a_n + b_n - \bar{u}_n$, а лучше принять за u_{n+1} ту из точек $a_n + \tau(b_n - a_n)$, $a_n + (1 - \tau)(b_n - a_n)$ $(\tau = (\Delta_2 + \delta)/\Delta_1)$, которая наиболее удалена от $\bar{u}_n$.

Упражнения. 1. Найти наименьшее n, начиная с которого точность метода золотого сечения больше точности метода деления отрезка пополам в 2 раза; в 10 раз.

2. Написать конечно-разностные уравнения для длин Δ_n отрезков $[a_n, b_n]$, получаемых симметричным методом, для случая, когда на каких-то шагах метода нарушается условие (4).

§ 5. Об оптимальных методах

1. В тех случаях, когда вычисление значений функции связано со значительными затратами, большую ценность приобретают экономичные или, как их еще называют, *оптимальные методы*, позволяющие решить задачу минимизации с требуемой точностью на основе вычислений значений минимизируемой функции как можно в меньшем числе точек, а также тесно связанные с ними методы, гарантирующие наилучшую точность при жестко заданном количестве вычислений значений миними-

зируемой функции. В связи с этим возникают вопросы, что такое оптимальные методы, существуют ли такие методы, как их строить? Абсолютно наилучший метод, пригодный для минимизации всех функций, вряд ли существует, и на поставленные вопросы можно попытаться ответить лишь при определенных ограничениях на рассматриваемые методы, функции и постановки задач минимизации.

Предположим, что нам задан некоторый класс функций Q, зафиксирована какая-либо постановка задачи минимизации функций из этого класса (например, задача первого или второго типа из § 1) и указано множество методов P, позволяющих решить поставленную задачу минимизации. Пусть $\Delta(J, p)$ — погрешность решения рассматриваемой задачи минимизации для функции $J = J(u) \in Q$ с помощью метода $p \in P$. Ясно, что, минимизируя одним и тем же методом p различные функции из Q, мы будем получать, вообще говоря, различные погрешности: для некоторых «хороших» функций из Q эта погрешность может оказаться равной нулю, а для других «плохих» функций из Q погрешность может быть значительной. Имеет смысл считать метод $p_1 \in P$ лучше метода $p_2 \in P$, если погрешность метода p_1 даже для самых «плохих» (для p_1) функций из Q будет меньше погрешности метода p_2 для «плохих» (для p_2) функций из Q. В связи с этим представляется разумным ввести величину $\delta(p) = \sup\limits_{J \in Q} \Delta(J, p)$, выражающую собой погрешность метода p при минимизации самой «плохой» (для p) функции из Q.

Определение 1. Величину $\delta(p) = \sup\limits_{J \in Q} \Delta(J, p)$ назовем *гарантированной точностью метода* $p \in P$ *на классе функций* Q. Скажем, что *метод* $p_1 \in P$ *лучше метода* $p_2 \in P$ *на классе* Q, если $\delta(p_1) < \delta(p_2)$. Метод $p_* \in P$ назовем *оптимальным методом на классе* Q, если $\delta(p_*) = \inf\limits_{p \in P} \delta(p) = \delta_*$, а величину δ_* — *наилучшей гарантированной точностью методов* P *на классе* Q. Если для некоторого метода $p_\varepsilon \in P$ выполняется неравенство $\delta(p_\varepsilon) \leqslant \delta_* + \varepsilon$, то метод p_ε назовем *ε-оптимальным на классе* Q.

Вопросы существования оптимальных и ε-оптимальных методов, возможности их построения для различных множеств методов P, классов функций Q и постановок задач минимизации, а также другие возможные подходы к проблеме выбора оптимальных методов изучались, например, в [11, 21, 81, 83, 95, 106, 109, 228, 241, 266, 282, 288, 291, 298, 328, 332].

2. Здесь мы кратко остановимся на оптимальных методах решения задачи минимизации функций из класса Q, состоящего из всех унимодальных функций на отрезке $[a, b]$. Ограничимся рассмотрением множества P методов минимизации, использующих лишь значения функции, считая при этом, что число n вычислений значений минимизируемой функции заранее задано. Будем предполагать, что в описание каждого метода p_n из P входит

задание правила выбора точек $u_1, \ldots, u_n$ из отрезка $[a, b]$, вычисление значений $J(u_1), \ldots, J(u_n)$ минимизируемой функции $J(u) \in Q$, выделение из точек $u_1, \ldots, u_n$ такой точки $\bar{u}_n$, для которой $J(\bar{u}_n) = \min\limits_{1 \leqslant i \leqslant n} J(u_i)$, и определение отрезка $[a_n, b_n]$, где в качестве a_n, b_n берутся ближайшие слева или справа к $\bar{u}_n$ точки среди $u_1, \ldots, u_n$, a, b (возможности $u_n = a$ или $u_n = b$ не исключаются).

Таким образом, применяя конкретный метод $p_n \in P$ к конкретной функции $J(u) \in Q$, в результате получаем отрезок $[a_n, b_n]$ и точку $\bar{u}_n \in [a_n, b_n]$ с вычисленным значением $J(\bar{u}_n) = \min\limits_{1 \leqslant i \leqslant n} J(u_i)$. Из определения унимодальной функции и построения отрезка $[a_n, b_n]$ следует неравенство $J(u) \geqslant J(\bar{u}_n)$ при всех $u \in [a, b] \setminus [a_n, b_n]$, так что

$$J_* = \inf_{a \leqslant u \leqslant b} J(u) = \inf_{a_n \leqslant u \leqslant b_n} J(u), \quad U_* \cap [a_n, b_n] \neq \varnothing. \tag{1}$$

В качестве приближения для J_* обычно берут величину $J(\bar{u}_n)$, а в качестве приближения к множеству U_* можно взять любую точку u_{n_*} из отрезка $[a_n, b_n]$ — на практике часто принимают $u_{n_*} = \bar{u}_n$ или $u_{n_*} = (a_n + b_n)/2$.

Отрезок $[a_n, b_n]$ принято называть *отрезком локализации минимума* функции $J(u)$ на отрезке $[a, b]$. Из (1) следует, что расстояние от любой точки $u_{n_*} \in [a_n, b_n]$ до множества U_* не превышает длины отрезка локализации $b_n - a_n$:

$$\rho(u_{n_*}, U_*) = \inf_{u \in U_*} |u_{n_*} - u| \leqslant b_n - a_n. \tag{2}$$

Величину $\Delta(J, p_n) = b_n - a_n$ можно принять за погрешность решения задачи минимизации функции $J(u) \in Q$ методом $p_n \in P$. Согласно (2) чем меньше погрешность $\Delta(J, p_n)$, тем точнее будет определено приближение u_{n_*} к U_* и, следовательно, тем лучше метод p_n. Для точного определения наилучшего или близкого к нему метода нам остается еще уточнить правило выбора точек $u_1, \ldots, u_n$, в которых вычисляются значения минимизируемой функции. Здесь принято различать два типа методов: пассивные методы и последовательные методы.

Если все точки $u_1, \ldots, u_n$ метода p_n выбираются одновременно до начала вычислений и в дальнейшем уже не меняются, то такой метод называют *пассивным*. Если в методе p_n точки $u_1, \ldots, u_n$ выбираются последовательно отдельными порциями, причем при выборе каждой очередной порции учитываются результаты предыдущих вычислений и проводится уточнение отрезка локализации минимума, то такой метод называется *последовательным*.

Примером пассивного метода является метод равномерного перебора. В этом методе точки $u_1, \ldots, u_n$ выбираются по правилу: $u_i = u_1 + ih$ $(i = 1, \ldots, n)$, где $h > 0$ — шаг метода, u_1 — заданная точка из $[a, b]$, $u_1 - a \leqslant h$ (например, $u_1 = a$ или $u_1 = a + h/2$), и, кроме того, $nh \leqslant b - u_1 < (n+1)h$.

Примерами последовательного метода служат методы деления отрезка пополам, золотого сечения.

Пассивный метод является частным случаем последовательного метода, когда все n точек выбираются сразу в первой же порции. Поэтому нетрудно понять, что последовательные методы, вообще говоря, обладают большей гибкостью и гораздо точнее пассивных методов. Однако отсюда не следует, что пассивные методы вовсе не находят применения. Такие методы весьма полезны, когда можно вести параллельные вычисления, используя, например, многопроцессорные ЭВМ. В тех случаях, когда значения минимизируемой функции определяются из физического эксперимента, условия проведения таких экспериментов также могут сделать необходимым применение пассивных методов.

Таким образом, n-точечные (т. е. использующие вычисление значений функции в n точках) последовательные и пассивные методы описаны. Если теперь в определении 1 принять, что Q — класс унимодальных функций на отрезке $[a, b]$, $P = P_n$ — множество всех n-точечных последовательных (или пассивных) методов, $\Delta(J, p_n) = b_n - a_n$ — длина отрезка локализации минимума, полученного методом p_n для функции $J = J(u) \in Q$, то придем к определениям гарантированной точности, оптимального и ε-оптимального последовательного (пассивного) метода для унимодальных функций. Кратко рассмотрим вопросы существования и построения оптимальных методов для таких функций.

3. Сначала остановимся на пассивных методах. Пусть $p_n = \{u_1, \ldots, u_n\}$ — какой-либо пассивный метод, $a = u_0 \leqslant u_1 < \ldots < u_n \leqslant b = u_{n+1}$. Применяя его к какой-либо функции $J(u) \in Q$, получаем отрезок $[u_{i-1}, u_{i+1}]$ локализации минимума этой функции, так что погрешность метода p_n здесь будет равна $\Delta(J, p_n) = u_{i+1} - u_{i-1}$. Поэтому $\delta(p_n) = \sup\limits_{J \in Q} \Delta(J, p_n) \leqslant \max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1})$. Пусть $\max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) = u_{k+1} - u_{k-1}$. Возьмем функцию $J_k = J_k(u) = |u - u_k|$. Это строго унимодальная функция достигает своей нижней грани на отрезке $[a, b]$ в точке $u_* = u_k \in [u_{k-1}, u_{k+1}]$, причем $\Delta(J_k, p_n) = u_{k+1} - u_{k-1}$. Следовательно, гарантированная погрешность метода p_n на классе Q равна $\delta(p_n) = \max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1})$. Для получения оптимального пассивного метода остается выяснить, достигается ли нижняя грань $\inf\limits_{p_n \in P_n} \delta(p_n) = \delta_*$, где P_n — множество всех пассивных методов, и если достигается, то на каком методе $p_n \in P_n$. Оказывается, здесь нужно различать случаи четного и нечетного n.

Теорема 1. *При всех нечетных $n = 2m + 1$ $(m \geqslant 0)$ существует бесконечно много оптимальных пассивных методов на классе Q; наилучшая гарантированная точность пассивных методов P_{2m+1} на этом классе равна $(b - a)/(m + 1)$.*

Доказательство. Возьмем пассивный метод $p_{n_*} = \{v_1, \ldots, v_n\}$, где $v_{2i} = a + i(b - a)/(m + 1)$ $(i = 1, \ldots, m)$, а точки v_{2i+1} $(i = 0, 1, \ldots, m)$ расположены на отрезке $[a, b]$ произвольно, лишь бы $v_{2i-1} < v_{2i} < v_{2i+1}$, $v_{2i+1} - v_{2i-1} \leqslant (b - a)/(m + 1)$. Очевидно, $\delta(p_{n_*}) = (b - a)/(m + 1)$. С другой стороны, для любого пассивного метода $p_n = \{u_1, \ldots, u_n\}$ $(u_0 = a \leqslant u_1 < \ldots < u_n \leqslant b = u_{n+1}, n = 2m + 1)$ имеем $\delta(p_n) = \max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) \geqslant \max\{b - u_{2m}, u_{2m} - u_{2m-2}, \ldots, u_4 - u_2, u_2 - a\} \geqslant (b - a)/(m + 1)$. Следовательно, методы p_{n_*} оптимальны и $\delta_* = (b - a)/(m + 1)$.

Теорема 2. *При всех четных $n = 2m$ $(m \geqslant 1)$ оптимального пассивного метода на классе Q не существует; наилучшая гарантированная точность пассивных методов P_{2n} на этом классе равна $(b - a)/(m + 1)$. В качестве ε-оптимального метода можно взять $p_{n\varepsilon} = \{v_1, \ldots, v_n\}$, где $v_{2i-1} = a + i(b - a)/(m + 1) - \varepsilon$, $v_{2i} = a + i(b - a)/(m + 1) + \varepsilon$ $(i = 1, \ldots, m, 0 < \varepsilon < (b - a)/(2(m + 1)))$.*

Доказательство. Сначала убедимся в том, что $\delta(p_n) > (b - a)/(m + 1)$ для любого пассивного метода $p_n = \{u_1, \ldots, u_n\}$ $(u_0 = a \leqslant u_1 < \ldots < u_n \leqslant b = u_{n+1}, n = 2m)$. Обозначим $\bar{u} = a + (b - a)/(m + 1)$. Имеются две возможности: либо $u_2 > \bar{u}$, либо $u_2 \leqslant \bar{u}$. Если $u_2 > \bar{u}$, то $\delta(p_n) = \max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) \geqslant u_2 - a > \bar{u} - a = (b - a)/(m + 1)$. Если же $u_2 \leqslant \bar{u}$, то $\delta(p_n) = \max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) > u_2 - a$. В самом деле, если бы $\max\limits_{1 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) \leqslant u_2 - a$, то $u_{2i+1} - u_{2i-1} \leqslant u_2 - a$ $(i = 1, \ldots, m)$ и,

кроме того, $u_1 - a < u_2 - a$. Сложив эти неравенства, придем к противоречивому неравенству $b - a < (m+1)(u_2 - a) \leqslant (m+1)(\bar{u} - a) = b - a$.

Таким образом, при $u_2 < \bar{u}$ имеем $\delta(p_n) = \max\limits_{2 \leqslant i \leqslant n} (u_{i+1} - u_{i-1}) = \delta(p'_{n-1})$, где p'_{n-1} — пассивный метод на отрезке $[u_1, b]$, составленный из точек $u_2, \ldots, u_n$ метода p_n. Но $n - 1 = 2m - 1$ — нечетное число, поэтому, применяя теорему 1 к отрезку $[u_1, b]$, имеем $\delta(p'_{n-1}) \geqslant (b - u_1)/m$. Тогда $\delta(p_n) = \delta(p'_{n-1}) \geqslant (b - u_1)/m > (b - \bar{u})/m = (b - a)/(m+1)$. Тем самым доказано, что $\delta(p_n) > (b - a)/(m+1)$ при всех $p_n \in P_{2m}$. С другой стороны, $\delta(p_{n\varepsilon}) = (b - a)/(m+1) + \varepsilon$ для всех ε $(0 < \varepsilon < (b - a)/(2(m+1))$. Следовательно, $\delta_* = (b - a)/(m+1)$.

Из теорем 1, 2 вытекает, что предпочтительнее пользоваться пассивными методами с четным числом $n = 2m$ точек, поскольку в случае $n = 2m + 1$ наилучшая гарантированная точность остается такой же, как и при $n = 2m$.

4. Перейдем к рассмотрению последовательных методов минимизации унимодальных функций на $[a, b]$. Здесь нам понадобятся знаменитые *числа Фибоначчи*, которые, как известно [95], определяются соотношениями

$$F_{n+2} = F_{n+1} + F_n, \quad n = 1, 2, \ldots; \quad F_1 = F_2 = 1.$$

С помощью индукции легко показать, что n-е число Фибоначчи представимо в виде

$$F_n = \left[\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right]\frac{1}{\sqrt{5}}, \qquad n = 1, 2, \ldots \tag{3}$$

Используя числа F_n, построим n-точечный последовательный метод, который принято называть *методом Фибоначчи*. Этот метод относится к классу симметричных методов, описанных в § 4, и определяется заданием на отрезке $[a, b]$ точки $u_1 = a + (b - a)F_{n-1}/F_{n+1}$ или симметричной ей точки $u_2 = a + b - u_1 = a + (b - a)F_n/F_{n+1}$. С помощью индукции нетрудно показать, что такой симметричный метод обладает свойством (4.4) и на k-м шаге $(k < n)$, когда проведены вычисления значений функции в точках $u_1, \ldots, u_k$, приводит к отрезку локализации минимума $[a_k, b_k]$ длиной

$$\Delta_k = b_k - a_k = (b - a)F_{n-k+2}/F_{n+1},$$

причем точка $\bar{u}_k$ $(a_k < \bar{u}_k < b_k)$ с вычисленным значением $J(\bar{u}_k) = \min\limits_{1 \leqslant i \leqslant k} J(u_i)$ совпадает с одной из точек

$$\begin{aligned} u'_k &= a_k + (b_k - a_k)\frac{F_{n-k}}{F_{n-k+2}} = a_k + (b - a)\frac{F_{n-k}}{F_{n+1}}, \\ u''_k &= a_k + (b_k - a_k)\frac{F_{n-k+1}}{F_{n-k+2}} = a_k + (b - a)\frac{F_{n-k+1}}{F_{n+1}} = a_k + b_k - u'_k, \end{aligned} \tag{4}$$

расположенных на отрезке $[a_k, b_k]$ симметрично относительно его середины.

Как видно из (4), при $k = n - 1$ точки u'_{n-1}, u''_{n-1} совпадают. Это означает, что при $k = n - 1$ первая часть процесса заканчивается вычислением значения функции в точке u_{n-1} и определением отрезка локализации минимума $[a_{n-1}, b_{n-1}]$ длины $b_{n-1} - a_{n-1} = (b - a)F_3/F_{n+1} = 2(b - a)F_{n+1}$, причем точка $u'_{n-1} = u''_{n-1} = \bar{u}_{n-1}$ совпадает с серединой отрезка $[a_{n-1}, b_{n-1}]$. В заключение, несколько нарушая симметричность процесса, последнее n-е вычисление значения минимизуемой функции $J(u)$ проводится в

точке $u_n = \bar{u}_{n-1} + \varepsilon$ (или $u_n = \bar{u}_{n-1} - \varepsilon$), где $0 < \varepsilon < (b-a)/F_{n+1}$, и отрезок $[a_n, b_n]$ локализации минимума определяется по формулам $a_n = a_{n-1}$, $b_n = \bar{u}_{n-1} + \varepsilon$ при $J(\bar{u}_{n-1}) \leqslant J(\bar{u}_{n-1} + \varepsilon)$ и $a_n = \bar{u}_{n-1}$, $b_n = b_{n-1}$ при $J(\bar{u}_{n-1}) > J(\bar{u}_{n-1} + \varepsilon)$, так что в худшем случае $b_n - a_n = (b-a)/F_{n+1} + \varepsilon$. Описанный метод обозначим через Φ_n.

Теорема 3. *При всех $n > 1$ оптимального последовательного метода на классе унимодальных функций не существует; наилучшая гарантированная точность последовательных методов на этом классе равна $(b-a)/F_{n+1}$. В качестве ε-оптимального метода можно взять метод Фибоначчи Φ_n.*

Доказательство этой теоремы можно найти в [11, 15, 83, 291]. Заметим, что число F_{n-1}/F_{n+1}, вообще говоря, является бесконечной периодической десятичной дробью, поэтому первая точка u_1 метода Φ_n будет задаваться на ЭВМ приближенно. Во избежание быстрого роста погрешности из-за неточности задания первой точки на практике нужно пользоваться модификацией метода Φ_n, описанной в § 4 для симметричных методов в общем случае.

Следует подчеркнуть, что метод Φ_n для своей реализации требует, чтобы число n вычислений значений минимизируемой функции было задано заранее — выбор первой точки в этом методе невозможен без знания n. В тех случаях, когда число n по каким-либо причинам не может быть задано заранее, можно применять метод золотого сечения, не требующий для своей реализации априорного значения n.

Для сравнения вспомним, что методом золотого сечения за n вычислений значений функции мы получали отрезок $[a_n, b_n]$ локализации минимума длины $b_n - a_n = ((\sqrt{5}-1)/2)^{n-1}(b-a) = (2/(\sqrt{5}+1))^{n-1}(b-a)$. С учетом формулы $F_{n+1} \approx ((\sqrt{5}+1)/2)^{n+1}/\sqrt{5}$, вытекающей из (3) при больших n, для метода Φ_n получаем отрезок локализации минимума, длина которого близка к $(b-a)/F_{n+1} \approx (2/(\sqrt{5}+1))^{n+1}(b-a)/\sqrt{5}$. Отсюда следует, что метод золотого сечения хуже метода Φ_n при больших n всего в $((\sqrt{5}+1)/2)^2/\sqrt{5} = 1{,}1708\ldots$ раз, т. е. на классе унимодальных функций метод золотого сечения близок к оптимальным методам. Интересно также заметить, что

$$\lim_{n\to\infty} \frac{F_{n-1}}{F_{n+1}} = \frac{3-\sqrt{5}}{2}, \qquad \lim_{n\to\infty} \frac{F_n}{F_{n+1}} = \frac{\sqrt{5}-1}{2},$$

т. е. при достаточно больших n начальные точки u_1, u_2 методов Фибоначчи и золотого сечения практически совпадают.

Упражнения. 1. Найти гарантированную на классе унимодальных функций точность последовательного (или пассивного) n-точечного метода p_n, если в качестве погрешности метода p_n при минимизации функции $J = J(u)$ принята величина $\Delta(J, p_n) = |J_* - J(\bar{u}_n)|$.

2. Сравнить оптимальные и ε-оптимальные пассивные методы на классе унимодальных функций с методом деления отрезка пополам.

3. Указать все точки метода Φ_n на отрезке $[0, 1]$ при $n = 2, 3, 4, 5$.

4. Применить метод Φ_5 к функциям $J(u) = u$, $J(u) = |u-1|$ на отрезке $[0, 2]$.

5. Найти наименьшее n, для которого точность метода золотого сечения хуже точности метода Фибоначчи в 2 раза.

6. Доказать, что число Фибоначчи F_n является ближайшим целым числом к $((1+\sqrt{5})/2)^n/\sqrt{5}$.

7. Доказать, что решение уравнения (4.1) представимо в виде $\Delta_n = (-1)^n F_{n-1}\Delta_2 + (-1)^{n-1}F_{n-2}\Delta_1$ $(n = 3, 4, \ldots)$. Отсюда вывести закон изменения погрешности величины Δ_n, если Δ_1, Δ_2 заданы неточно.

8. Доказать, что последовательность $\{F_{2m}/F_{2m+1}\}$ сходится к $\tau_1 = (\sqrt{5}-1)/2$ монотонно возрастая, а $\{F_{2m-1}/F_{2m}\}$ сходится к τ_1 монотонно убывая.

9. Используя утверждения упражнений 7, 8, доказать, что метод золотого сечения является единственным симметричным методом, удовлетворяющим условию (4.4) при всех $n = 1, 2, \ldots$

10. Пусть дан симметричный метод с начальными отрезками Δ_1, Δ_2, пусть $N \geqslant 2$ — заданное натуральное число. Используя утверждения упражнений 7, 8, указать промежуток изменения отношения Δ_2/Δ_1, чтобы метод удовлетворял условию (4.4) при всех $n = 1, \ldots, N$.

11. Пусть дан некоторый симметричный метод, удовлетворяющий условию (4.4) при $n = 1$. Используя утверждения упражнений 7, 8, указать максимальное число N, при котором условие (4.4) выполняется для всех $n = 2, \ldots, N$.

§ 6. Метод ломаных

Описанные выше методы часто приходится применять без априорного знания о том, что минимизируемая функция является унимодальной. Однако в этом случае погрешности в определении минимального значения и точек минимума функции могут быть значительными. Например, применение этих методов к минимизации непрерывных на отрезке функций приведет, вообще говоря, лишь в окрестность точки локального минимума, в которой значение функции может сильно отличаться от искомого минимального значения на отрезке. Поэтому представляется важной разработка методов поиска глобального минимума, позволяющих строить минимизирующие последовательности и получить приближенное решение задач минимизации первого и второго типов (см. § 1) для функций, не обязательно унимодальных.

Здесь мы рассмотрим один из таких методов для класса функций, удовлетворяющих условию Липшица.

Определение 1. Говорят, что функция $J(u)$ удовлетворяет *условию Липшица* на отрезке $[a, b]$, если существует постоянная $L > 0$ такая, что

$$|J(u) - J(v)| \leqslant L|u - v| \qquad \forall u, v \in [a, b]. \tag{1}$$

Постоянную L называют *постоянной Липшица* функции $J(u)$ на $[a, b]$.

Условие (1) имеет простой геометрический смысл: оно означает, что угловой коэффициент (тангенс угла наклона) $|J(u) - J(v)| \cdot |u - v|^{-1}$ хорды, соединяющей точки $(u, J(u))$ и $(v, J(v))$ графика функции, не превышает постоянной L для всех точек $u, v \in [a, b]$. Из (1) следует, что функция $J(u)$ непрерывна на отрезке $[a, b]$, так что по теореме 1.1 множество U_* точек минимума $J(u)$ на $[a, b]$ непусто.

Теорема 1. *Пусть функция $J(u)$ непрерывна на отрезке $[a, b]$ и на каждом отрезке $[a_i, a_{i+1}]$ $(i = 1, \ldots, m)$, где $a_1 = a$, $a_{m+1} = b$, удовлетворяет условию (1) с постоянной L_i. Тогда $J(u)$ удовлетворяет условию (1) на всем отрезке с постоянной $L = \max\limits_{1 \leqslant i \leqslant m} L_i$.*

Доказательство. Возьмем две произвольные точки $u, v \in [a, b]$. Пусть $a_{p-1} \leqslant u \leqslant a_p$, $a_s \leqslant v \leqslant a_{s+1}$ при некоторых p, s. Тогда

$$|J(u) - J(v)| = \left| J(u) - J(a_p) + \sum_{i=p}^{s-1} (J(a_i) - J(a_{i+1})) + J(a_s) - \right.$$

$$\left. - J(v) \right| \leqslant L_{p-1} |u - a_p| + \left| \sum_{i=p}^{s-1} L_i (a_{i+1} - a_i) \right| + L_s |a_s - v| \leqslant L|u - v|.$$

Теорема 2. *Пусть функция $J(u)$ дифференцируема на отрезке $[a, b]$ и ее производная $J'(u)$ ограничена на этом отрезке. Тогда $J(u)$ удовлетворяет условию* (1) *с постоянной* $L = \sup\limits_{u \in [a,b]} |J'(u)|$.

Доказательство. По формуле конечных приращений для любых $u, v \in [a, b]$ имеем $J(u) - J(v) = J'(v + \theta(u - v))(u - v)$ $(0 < \theta < 1)$. Отсюда и из ограниченности $J'(u)$ следует утверждение теоремы.

Пусть функция $J(u)$ удовлетворяет условию (1) на отрезке $[a, b]$. Зафиксируем какую-либо точку $v \in [a, b]$ и определим функцию $g(u, v) = J(v) - L|u - v|$ переменной u $(a \leqslant u \leqslant b)$. Очевидно, функция $g(u, v)$ кусочно линейна на $[a, b]$ и график ее представляет собой ломаную линию, составленную из отрезков двух прямых, имеющих угловые коэффициенты L и $-L$ и пересекающихся в точке $(v, J(v))$. Кроме того, в силу условия (1)

$$J(u) - g(u, v) \geqslant (L - |J(u) - J(v)|\,|u - v|^{-1})\,|u - v| \geqslant 0, \quad u \neq v,$$

т. е.

$$g(u, v) = J(v) - L|u - v| \leqslant J(u) \qquad \forall u \in [a, b], \tag{2}$$

причем $g(v, v) = J(v)$. Это значит, что график функции $J(u)$ лежит выше ломаной $g(u, v)$ при всех $u \in [a, b]$ и имеет с ней общую точку $(v, J(v))$.

Свойство (2) ломаной $g(u, v)$ можно использовать для построения следующего метода [246], который назовем *методом ломаных*. Этот метод начинается с выбора произвольной точки $u_0 \in [a, b]$ и составления функции $g(u, u_0) = J(u_0) - L|u - u_0| = p_0(u)$. Следующая точка u_1 определяется из условий $p_0(u_1) = \min\limits_{u \in [a,b]} p_0(u)$ $(u_1 \in [a, b])$; очевидно, $u_1 = a$ или $u_1 = b$. Далее берется новая функция $p_1(u) = \max\{g(u, u_1);\ p_0(u)\}$, и очередная точка u_2 находится из условий $p_1(u_2) = \min\limits_{u \in [a,b]} p_1(u)$ $(u_2 \in [a, b])$ и т. д. (рис. 1.2).

Пусть точки $u_0, u_1, \ldots, u_n$ $(n \geqslant 1)$ уже известны. Тогда составляется функция

$$p_n(u) = \max\{g(u, u_n),\ p_{n-1}(u)\} = \max_{0 \leqslant i \leqslant n} g(u, u_i),$$

и следующая точка u_{n+1} определяется условиями

$$p_n(u_{n+1}) = \min_{u \in [a,b]} p_n(u), \quad u_{n+1} \in [a, b]. \tag{3}$$

Если минимум $p_n(u)$ на $[a, b]$ достигается в нескольких точках, то в качестве u_{n+1} можно взять любую из них.

Метод ломаных описан. Очевидно, $p_n(u)$ является кусочно линейной функцией и график ее представляет собой непрерывную ломаную линию, состоящую из отрезков прямых с угловыми наклонами L или $-L$. Из теоремы 1 следует, что $p_n(u)$ удовлетворяет условию (1) с той же постоянной L, что и функция $J(u)$. Ясно также, что

$$p_{n-1}(u) = \max_{0 \leqslant i \leqslant n-1} g(u, u_i) \leqslant \max_{0 \leqslant i \leqslant n} g(u, u_i) = p_n(u), \quad u \in [a, b]. \tag{4}$$

Кроме того, согласно (2) $g(u, u_i) \leqslant J(u)$ $(u \in [a, b])$ для всех $i = 0, 1, \ldots, n$, поэтому

$$p_n(u) \leqslant J(u), \; u \in [a, b], \quad n = 0, 1, \ldots \tag{5}$$

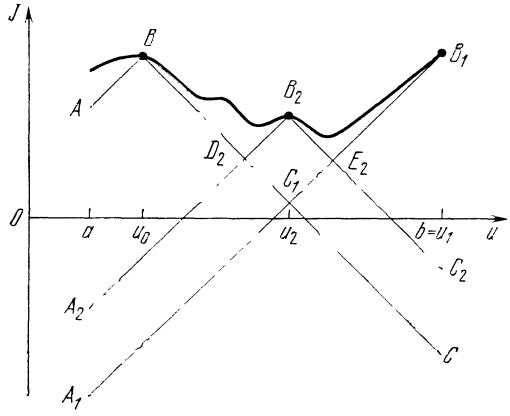

Рис. 1.2. ABC — график $p_0(u) = g(u, u_0)$, A_1B_1 — график $g(u, u_1)$, ABC_1B_1 — график $p_1(u)$, $A_2B_2C_2$ — график $g(u, u_2)$, $ABD_2B_2E_2B_1$ — график $p_2(u)$

Таким образом, на каждом шаге метода ломаных задача минимизации функции $J(u)$ заменяется более простой задачей минимизации кусочно линейной функции $p_n(u)$, которая приближает $J(u)$ снизу, причем согласно (4) $\{p_n(u)\}$ монотонно возрастают. Докажем, что при неограниченном увеличении n метод ломаных сходится.

Теорема 3. *Пусть $J(u)$ — произвольная функция, удовлетворяющая на отрезке $[a, b]$ условию (1). Тогда последовательность $\{u_n\}$, полученная с помощью описанного метода ломаных, такова, что:* 1) $\lim\limits_{n\to\infty} J(u_n) = \lim\limits_{n\to\infty} p_n(u_{n+1}) = J_* = \inf\limits_{u \in [a,b]} J(u)$, *причем справедлива оценка*

$$0 \leqslant J(u_{n+1}) - J_* \leqslant J(u_{n+1}) - p_n(u_{n+1}), \quad n = 0, 1, \ldots; \tag{6}$$

2) *$\{u_n\}$ сходится к множеству U_* точек минимума $J(u)$ на $[a, b]$, т. е.* $\lim\limits_{n\to\infty} \rho(u_n, U_*) = 0$.

Доказательство. Возьмем произвольную точку $u_* \in U_*$. С учетом условий (3) и неравенств (4), (5) имеем

$p_{n-1}(u_n) = \min\limits_{u\in[a,b]} p_{n-1}(u) \leqslant p_{n-1}(u_{n+1}) \leqslant p_n(u_{n+1}) = \min\limits_{u\in[a,b]} p_n(u) \leqslant$ $\leqslant p_n(u_*) \leqslant J(u_*) = J_*$, т. е. последовательность $\{p_n(u_{n+1})\}$ монотонно возрастает и ограничена сверху. Отсюда сразу следует оценка (6) и существование предела $\lim\limits_{n\to\infty} p_n(u_{n+1}) = p_* \leqslant J_*$. Покажем, что $p_* = J_*$.

Последовательность $\{u_n\}$ ограничена и по теореме Больцано — Вейерштрасса обладает хотя бы одной предельной точкой. Пусть v_* — какая-либо предельная точка последовательности $\{u_n\}$. Тогда существует подпоследовательность $\{u_{n_k}\}$, сходящаяся к v_*, причем можем считать, что $n_1 < \ldots < n_{k-1} < n_k < \ldots$ Заметим, что $J(u_i) = g(u_i, u_i) \leqslant p_n(u_i) \leqslant J(u_i)$, т. е. $J(u_i) = p_n(u_i)$ при всех $i = 0, 1, \ldots, n$. Тогда $0 \leqslant p_n(u_i) - \min\limits_{u\in[a,b]} p_n(u) =$ $= J(u_i) - p_n(u_{n+1}) = p_n(u_i) - p_n(u_{n+1}) \leqslant L\,|u_i - u_{n+1}|$ при любом n и $i = 0, 1, \ldots, n$. Принимая здесь $n = n_k - 1$, $i = n_{k-1} \leqslant$ $\leqslant n_k - 1$, получаем $0 \leqslant J(u_{n_{k-1}}) - p_{n_k-1}(u_{n_k}) \leqslant L\,|u_{n_{k-1}} - u_{n_k}|$ $(k \geqslant 2)$. Отсюда при $k \to \infty$ имеем $J_* \leqslant J(v_*) = \lim\limits_{k\to\infty} J(u_{n_{k-1}}) =$ $= \lim\limits_{k\to\infty} p_{n_k-1}(u_{n_k}) = p_* \leqslant J_*$, т. е. $\lim\limits_{k\to\infty} J(u_{n_k}) = \lim\limits_{k\to\infty} p_{n_k-1}(u_{n_k}) =$ $= p_* = J_*$.

Пользуясь тем, что рассуждения проведены для произвольной предельной точки v_* последовательности $\{u_n\}$, убеждаемся в справедливости первого утверждения теоремы. Второе утверждение следует из теоремы 1.1.

Таким образом, с помощью метода ломаных можно получить решение задач минимизации первого и второго типов для функций, удовлетворяющих условию (1). Проста и удобна для практического использования формула (6), дающая оценку неизвестной погрешности $J(u_{n+1}) - J_*$ через известные величины, вычисляемые в процессе реализации метода ломаных. Этот метод не требует унимодальности минимизируемой функции, и более того, функция может иметь сколько угодно точек локального экстремума на рассматриваемом отрезке. На каждом шаге метода ломаных нужно минимизировать кусочно линейную функцию $p_n(u)$, что может быть сделано простым перебором известных вершин ломаной $p_n(u)$, причем здесь перебор существенно упрощается благодаря тому, что ломаная $p_n(u)$ отличается от ломаной $p_{n-1}(u)$ не более чем двумя новыми вершинами. К достоинству метода относится и то, что он сходится при любом выборе начальной точки u_0. В работе [128] показывается, что метод ломаных близок к оптимальным методам на классе функций, удовлетворяющих условию (1). Оптимальные методы поиска минимума строго унимодальных функций, удовлетворяющих условию (1), рассмотрены в [328].

К недостаткам метода ломаных следует отнести то, что с увеличением числа шагов n растет требуемый объем памяти

ЭВМ для хранения координат вершин ломаной $p_n(u)$. В § 7 будет рассмотрен другой метод, по своей идее близкий к методу ломаных, но предъявляющий менее жесткие требования к объему памяти и более удобный для реализации на ЭВМ.

Следует также отметить, что метод ломаных невозможно реализовать без знания постоянной L из условия (1). На практике оценку для L получают, вычисляя угловые коэффициенты некоторого числа хорд, соединяющих точки графика минимизируемой функции. Здесь полезно иметь в виду, что если $u < v < w$, то

$$|J(w) - J(u)|/(w-u) \leqslant$$
$$\leqslant \max\{|J(w) - J(v)|/(w-v);\ |J(v) - J(u)|/(v-u)\}, \qquad (7)$$

т. е. при добавлении новой точки на отрезке $[u, w]$ появляется новая хорда с неменьшим угловым коэффициентом.

Для доказательства (7) нужно рассмотреть два случая, когда неравенство

$$J(v) \geqslant (J(w) - J(u))(v-u)/(w-u) + J(u) \qquad (8)$$

выполняется и когда оно не выполняется. Если $J(w) \geqslant J(u)$ и (8) выполняется, то $(J(v) - J(u))/(v-u) \geqslant (J(w) - J(u))/(w-u) \geqslant 0$; если $J(w) \geqslant J(u)$ и (8) не выполняется, то $(J(w) - J(v))/(w-v) \geqslant (J(w) - J(u))/(w-u) \geqslant 0$. Аналогично доказывается (7) в случае $J(w) < J(u)$.

Пусть $a = v_0 < v_1 < \ldots < v_m = b$; обозначим $L_m = \max\limits_{1 \leqslant i \leqslant m} |J(v_i) - J(v_{i-1})| \cdot |v_i - v_{i-1}|^{-1}$. Ясно, что $L_m \leqslant L$. Пусть при каждом $m \geqslant 1$ величина L_{m+1} вычисляется по точкам $a = w_0 < w_1 < \ldots < w_{m+1} = b$, полученным добавлением к точкам $v_0, v_1, \ldots, v_m$ одной новой точки. Тогда согласно (7) имеем $L_m \leqslant L_{m+1} \leqslant L$ $(m \geqslant 1)$. Это значит, что с возрастанием m величины L_m все лучше и лучше приближают L снизу. Если $\max\limits_{0 \leqslant i \leqslant m} |v_i - v_{i-1}| \to 0$ при $m \to \infty$, то $\lim\limits_{m \to \infty} L_m = L$. Приведенные соображения могут помочь в получении оценки для L. При определении L могут быть полезны теоремы 1, 2.

Следует заметить, что использование завышенной оценки для L ухудшает скорость сходимости метода ломаных, приводит к излишне большому количеству вычислений значений минимизируемой функции. Если же пользоваться заниженной оценкой для L, то метод может привести к неправильному определению приближения минимального значения.

Упражнения. 1. Привести пример функции, удовлетворяющей условию (1), но не являющейся унимодальной.

2. Можно ли утверждать, что всякая унимодальная на отрезке $[a, b]$ функция удовлетворяет условию (1) на $[a, b]$? Рассмотреть пример функции $J(u) = \sqrt{u}$ на $[0, 1]$.

3. Рассмотреть первые шесть шагов метода ломаных для функции $J(u) = ||u^2 - 1| - 1|$ на отрезке $[-2, 2]$ при различном выборе начальной точки u_0.

4. Выяснить, как ведет себя метод ломаных при минимизации функции $J(u) \equiv 1$ на отрезке $[0, 1]$.

5. Пусть $J(u) = a_n u^n + \ldots + a_1 u + a_0$ — многочлен n-й степени на отрезке $[a, b]$, где $0 < a < b$. Обозначим

$$A^+ = \{i\colon 0 \leqslant i \leqslant n,\ a_i > 0\},\quad A^- = \{i\colon 0 \leqslant i \leqslant n,\ a_i < 0\},$$

$$J_+(u) = \sum_{i \in A^+} a_i u^i,\quad J_-(u) = \sum_{i \in A^-} |a_i|\, u^i.$$

Доказать, что $J(u) = J_+(u) - J_-(u)$, а в качестве постоянной L из условия (1) для функции $J(u)$ на $[a, b]$ можно взять величину $\max\{J'_+(b) - J'_-(a);\ |J'_+(a) - J'_-(b)|\}$ [106].

§ 7. Методы покрытий

1. Обозначим через $Q(L)$ класс функций, удовлетворяющих условию Липшица (6.1) на отрезке $[a, b]$ с одной и той же для всех функций этого класса постоянной $L > 0$. Для функций $J = J(u) \in Q(L)$ будем рассматривать задачу минимизации первого типа, когда ищется величина $J_* = \inf\limits_{u \in [a,b]} J(u)$. Для решения этой задачи будем пользоваться методами p_n, которые заключаются в выборе точек $u_1, \ldots, u_n$ $(a \leqslant u_1 < \ldots < u_n \leqslant b)$, вычислении значений функции $J(u_1), \ldots, J(u_n)$ и определении величины $J(u_k) = \min\limits_{1 \leqslant i \leqslant n} J(u_i)$, принимаемой за приближение к J_*.

Возникает вопрос: как выбрать метод $p_n = \{u_1, \ldots, u_n\}$, чтобы

$$\min_{1 \leqslant i \leqslant n} J(u_i) \leqslant J_* + \varepsilon \qquad \forall J(u) \in Q(L), \tag{1}$$

где $\varepsilon > 0$ — заданная точность? Ниже будет изложено несколько методов решения поставленной задачи (1). В каждом из этих методов определенным образом строится некоторая система отрезков, покрывающих исходный отрезок $[a, b]$, и вычисляются значения функции в подходящим образом выбранных точках этих отрезков. Поэтому излагаемые ниже методы принято называть *методами покрытий*.

2. Простейшим методом p_n для решения задачи (1) может служить *метод равномерного перебора*, когда точки $u_1, \ldots, u_n$ выбираются по правилу

$$\begin{gathered} u_1 = a + h/2,\ u_2 = u_1 + h,\ \ldots,\ u_{i+1} = u_i + h = u_1 + ih,\ \ldots, \\ u_{n-1} = u_1 + (n-2)h,\quad u_n = \min\{u_1 + (n-1)h;\ b\}, \end{gathered} \tag{2}$$

где $h = 2\varepsilon/L$ — шаг метода, а число n определяется условием $u_{n-1} < b - h/2 \leqslant u_1 + (n-1)h$.

Теорема 1. *Метод равномерного перебора* (2) *решает задачу* (1) *на классе* $Q(L)$. *Если* $h > 2\varepsilon/L$, *то существует функция* $J(u) \in Q(L)$, *для которой метод* (2) *не решает задачу* (1).

Доказательство. Пусть $J = J(u)$ — произвольная функция из $Q(L)$. С учетом неравенства (6.2) для любого $u \in [u_i - h/2, u_i + h/2]$ имеем $J(u) \geqslant J(u_i) - L|u - u_i| \geqslant J(u_i) - Lh/2 \geqslant \min\limits_{1 \leqslant i \leqslant n} J(u_i) - \varepsilon$ при всех $i = 1, \ldots, n$. Поскольку система отрезков $[u_i - h/2, u_i + h/2]$ $(i = 1, \ldots, n)$ покрывает весь отрезок $[a, b]$, т. е. всякая точка u из $[a, b]$ принадлежит одному из отрезков этой системы, то из предыдущего неравенства следует, что $J(u) \geqslant \min\limits_{1 \leqslant i \leqslant n} J(u_i) - \varepsilon$ для всех $u \in [a, b]$. Поэтому $J_* \geqslant \min\limits_{1 \leqslant i \leqslant n} J(u_i) - \varepsilon$ для любой функции $J = J(u) \in Q(L)$, что равносильно неравенству (1). Если $h > 2\varepsilon/L$, то, например, для функции $J(u) = Lu$ метод (2) дает $\min\limits_{1 \leqslant i \leqslant n} (Lu_i) - La = Lh/2 > \varepsilon$.

3. Метод равномерного перебора (2) относится к пассивным методам, когда точки $u_1, \ldots, u_n$ задаются все одновременно до начала вычислений значений функции. На классе $Q(L)$ можно предложить такой же простой, но более эффективный последовательный метод перебора, когда выбор точки u_i при каждом $i > 2$ производится с учетом вычислений значения функции в предыдущих точках $u_1, \ldots, u_{i-1}$, и задачу (1) удается решить, вообще говоря, за меньшее количество вычислений значений функции, чем методом (2). А именно, следуя работе [139], положим

$$\begin{gathered} u_1 = a + h/2,\ u_{i+1} = u_i + h + (J(u_i) - F_i)/L, \qquad i = 1, \ldots, n-2, \\ u_n = \min\{u_{n-1} + h + (J(u_{n-1}) - F_{n-1})/L;\ b\}, \end{gathered} \tag{3}$$

где $h = 2\varepsilon/L$, $F_i = \min\limits_{1 \leqslant j \leqslant i} J(u_j)$, а число n определяется условием $u_{n-1} < b - h/2 \leqslant u_{n-1} + h + (J(u_{n-1}) - F_{n-1})/L$.

Теорема 2. *Метод последовательного перебора* (3) *решает задачу* (1) *на классе* $Q(L)$.

Доказательство. Пусть $J = J(u)$ — произвольная функция из $Q(L)$. С учетом неравенства (6.2) для всех $u \in [u_i, u_i + h/2 + (J(u_i) - F_i)/L]$ имеем $J(u) \geqslant J(u_i) - L(h/2 + (J(u_i) - F_i)/L) = F_i - Lh/2 \geqslant F_n - \varepsilon$. Аналогично для всех $u \in [u_i - h/2, u_i]$ получим $J(u) \geqslant J(u_i) - Lh/2 \geqslant F_n - \varepsilon$. Поскольку система отрезков $[u_i - h/2, u_i + h/2 + (J(u_i) - F_i)/L]$ $(i = 1, \ldots, n)$ покрывает весь отрезок $[a, b]$, то из предыдущих неравенств следует, что $J(u) \geqslant F_n - \varepsilon$ при всех $u \in [a, b]$. Тогда $J_* \geqslant F_n - \varepsilon$ при всех $J = J(u) \in Q(L)$, что равносильно неравенству (1).

В худшем случае, когда, например, функция $J(u)$ постоянна или монотонно убывает на $[a, b]$ и, следовательно, $F_i = \min\limits_{1 \leqslant j \leqslant i} J(u_j) = J(u_i)$, метод (3) превращается в метод (2), и

для решения задачи (1) тогда потребуется $N_1 \approx (b-a)L/(2\varepsilon)$ вычислений значений функции. В самом лучшем случае, когда $J(u) = A + L(u - B)$, где A, B — постоянные и, следовательно, $F_i = J(u_1)$, $u_i = u_1 + (2^{i-1} - 1)h$ $(i = 1, \ldots, n-2)$, для решения задачи (1) понадобится всего $N_2 \approx 1 + \log_2(b-a)L/(2\varepsilon)$ вычислений значений функции. И вообще, если $J(u_i) > F_i$ при каком-либо i, то $u_{i+1} - u_i > h$, и поэтому число n вычислений значений функции, необходимое для решения задачи (1), будет, вообще говоря, меньше N_1 и больше N_2.

Заметим также, что метод (3) идейно примыкает к методу ломаных из § 6, но метод (3) выгодно отличается простотой реализации и не требует большой машинной памяти. Недостатком метода (3), как и метода ломаных, является необходимость априорного знания постоянной L из условия (6.1).

4. Следуя [141, 231], изложим еще один метод покрытий для решения задачи (1). Сначала определим минимальное целое число n из условия

$$(b-a)/2^{n+1} \leqslant \varepsilon/L, \qquad n \geqslant 0, \tag{4}$$

и на отрезке $[a, b]$ введем точки $c_{ik} = a + i(b-a)/2^k$ $(i = 0, 1, \ldots, 2^k)$, где $0 \leqslant k \leqslant n$. Систему отрезков $\{[c_{ik}, c_{i+1k}], i = 0, 1, \ldots, 2^k - 1\}$ будем называть *разбиением отрезка* $[a, b]$ *k-го уровня*. Таким образом, исходный отрезок будет иметь нулевой уровень. Отрезками первого уровня являются две половины исходного отрезка. И вообще, отрезки $k+1$-го уровня получаются из отрезков k-го уровня делением пополам.

На первом шаге метода берем исходный отрезок $[a_1, b_1] = [a, b]$, полагаем $h_1 = (b_1 - a_1)/2$, $u_1 = (a_1 + b_1)/2$, вычисляем $J(u_1) = F_1$. Если $n = 0$, то процесс закончен и, как увидим ниже, число F_1 — искомое приближение для J_*. Допустим, что $n > 0$. Тогда на втором шаге метода рассматриваем отрезок $[a_2, b_2]$, представляющий собой одну из половин отрезка $[a_1, b_1]$. Полагаем $h_2 = (b_2 - a_2)/2$, $u_2 = (a_2 + b_2)/2$, вычисляем $J(u_2)$, $F_2 = \min\{F_1; J(u_2)\}$. Если $n = 1$, то отрезок $[a_2, b_2]$ исключаем из дальнейшего рассмотрения и на третьем шаге переходим к отрезку $[a_3, b_3] = [a_1, b_1] \backslash [a_2, b_2]$. Если же $n > 1$, то еще проверяем неравенство $F_2 \leqslant J(u_2) - Lh_2 + \varepsilon$. В случае выполнения этого неравенства отрезок $[a_2, b_2]$ исключаем из дальнейшего рассмотрения и на третьем шаге переходим к отрезку $[a_3, b_3] = [a_1, b_1] \backslash [a_2, b_2]$; если же это неравенство не выполняется, то в качестве $[a_3, b_3]$ берем одну из половин отрезка $[a_2, b_2]$ и т. д.

Пусть уже сделано $k-1$ шагов $(k \geqslant 2)$, определена величина $F_{k-1} = \min\limits_{\leqslant i \leqslant k-1} J(u_i)$ и выделен отрезок $[a_k, b_k]$ некоторого j-го уровня $(j \leqslant n)$ для рассмотрения на следующем k-м шаге. Положим $h_k = (b_k - a_k)/2$, $u_k = (a_k + b_k)/2$, вычислим $J(u_k)$,

$F_k = \min\{F_{k-1}; J(u_k)\} = \min\limits_{1 \leqslant i \leqslant k} J(u_i)$. Если $j = n$, то отрезок $[a_k, b_k]$ исключаем из дальнейшего рассмотрения и на $k+1$-м шаге переходим к одному из отрезков $[a_{k+1}, b_{k+1}]$ какого-либо уровня, еще не исключенного из рассмотрения (например, за $[a_{k+1}, b_{k+1}]$ можно принять один из оставшихся отрезков с возможно меньшим уровнем; возможны и другие варианты перебора оставшихся отрезков [141, 231]). Если же $j < n$, то проверяем еще одно неравенство

$$F_k \leqslant J(u_k) - Lh_k + \varepsilon. \tag{5}$$

Если оно выполняется, то отрезок $[a_k, b_k]$ исключаем из дальнейшего рассмотрения, а в качестве следующего отрезка $[a_{k+1}, b_{k+1}]$ берем один из оставшихся отрезков какого-либо уровня (например, самого меньшего уровня). Если (5) не выполняется, то за $[a_{k+1}, b_{k+1}]$ берем одну из половин отрезка $[a_k, b_k]$ и т. д. Процесс исключения продолжается до тех пор, пока не будет исчерпан весь отрезок $[a, b]$.

Теорема 3. *Для каждой функции $J(u) \in Q(L)$ описанный процесс закончится за $N \leqslant 2^n - 1$ шагов, где n взято из (4), исчерпанием всего отрезка $[a, b]$, причем $F_N = \min\limits_{1 \leqslant i \leqslant N} J(u_i) \leqslant$ $\leqslant J_* + \varepsilon$.*

Доказательство. Из описания метода видно, что отрезки $[a_k, b_k]$ n-го уровня дальше не дробятся и исключаются из рассмотрения без проверки условия (5). Исключение отрезка $[a_k, b_k]$ i-го уровня при $i < n$ производится лишь при выполнении условия (5), в противном случае отрезок $[a_k, b_k]$ дробится пополам. Таким образом, в самом худшем случае, когда условие (5) не выполнится ни для одного отрезка i-го уровня $(i < n)$, процесс превратится в последовательное исключение всех отрезков n-го уровня и завершится на шаге $N = 2^n - 1$. Если же хотя бы на одном шаге благодаря условию (5) будет исключен отрезок j-го уровня $(j < n)$, то процесс закончится за $N < 2^n - 1$ шагов. Покажем, что $F_N \leqslant J_* + \varepsilon$. Поскольку отрезки n-го уровня $[c_{in}, c_{i+1,n}]$ $(i = 0, 1, \ldots, 2^n - 1)$ покрывают весь отрезок $[a, b]$, то найдется такой номер m $(0 \leqslant m \leqslant 2^n - 1)$, что $[c_{mn}, c_{m+1,n}] \cap U_* \neq \varnothing$.

Пусть отрезок $[c_{mn}, c_{m+1,n}]$ оказался исключенным на k-м шаге как часть отрезка $[a_k, b_k]$ j-го уровня $(j \leqslant n)$. Имеются две возможности: $j < n$ или $j = n$. Если $j < n$, то исключение произведено из-за выполнения условия (5). Отсюда и из условия Липшица следует, что $J(u) \geqslant J(u_k) - Lh_k \geqslant F_k - \varepsilon \geqslant F_N - \varepsilon$ для всех $u \in [a_k, b_k]$. Тогда, учитывая, что $[c_{mn}, c_{m+1,n}] \cap U_* \subset$ $\subset [a_k, b_k] \cap U_* \neq \varnothing$, имеем $J_* = \inf\limits_{[a_k, b_k]} J(u) \geqslant F_N - \varepsilon$, т. е. $F_N \leqslant J_* + \varepsilon$. Если $j = n$, то отрезок $[a_k, b_k] = [c_{mn}, c_{m+1,n}]$ на k-м шаге исключен из рассмотрения как отрезок n-го уровня.

В этом случае в силу (4) имеем $b_k - a_k = c_{m+1,n} - c_{mn} = (b-a)/2^n \leqslant 2\varepsilon/L$, и из условия Липшица получим $J(u) \geqslant J(u_k) - L(b-a)/2^{n+1} \geqslant J(u_k) - \varepsilon \geqslant F_N - \varepsilon$ для всех $u \in [a_k, b_k]$. Отсюда снова имеем $J_* = \inf\limits_{[a_k, b_k]} J(u) \geqslant F_N - \varepsilon$.

Как видим, описанный метод на классе $Q(L)$ в худшем случае превращается в метод простого перебора отрезков $[c_{in}, c_{i+1,n}]$ $(i = 0, 1, \ldots, 2^n - 1)$ с шагом $h = (b-a)/2^n = 2\varepsilon/L$ с вычислением значений функции в средних точках этих отрезков. В то же время ясно, что для многих функций $J(u) \in Q(L)$ этот метод гораздо эффективнее метода простого перебора.

5. Метод последовательного перебора, аналогичный методу (3), можно предложить и для некоторых других классов функций. Остановимся на классе функций, дважды дифференцируемых на отрезке $[a, b]$, у которых $\sup\limits_{u \in [a,b]} J''(u) \leqslant M$, где M — некоторая фиксированная постоянная. Обозначим этот класс функций через $R(M)$. Заметим, что если $M \leqslant 0$, то $J''(u) \leqslant 0$ и, следовательно, $J'(u)$ монотонно убывает на $[a, b]$. Это значит, что тогда функция достигает своей нижней грани при $u = a$ или $u = b$. Таким образом, задача минимизации функций из класса $R(M)$ в случае $M \leqslant 0$ решается просто. Поэтому имеет смысл рассматривать класс $R(M)$ при $M > 0$. Тогда для решения задачи минимизации первого типа на классе функций $R(M)$ можно предложить следующий метод последовательного перебора [106]:

$$u_1 = a,\ u_{i+1} = u_i + \sqrt{2\varepsilon/M} + h_i, \quad i = 1, \ldots, n-2,\ u_n = b, \tag{6}$$

где $h_i = \sqrt{2(\varepsilon + J(u_i) - F_i)/M}$, $F_i = \min\limits_{1 \leqslant j \leqslant i} J(u_j)$, а число n определяется условием $u_{n-1} < b \leqslant u_{n-1} + \sqrt{2\varepsilon/M} + h_{n-1}$.

Теорема 4. *Применяя метод* (6), *задачу минимизации первого типа для любой функции* $J(u) \in R(M)$ *можно решить с заданной точностью* $\varepsilon > 0$, *т. е.*

$$0 \leqslant \min_{1 \leqslant i \leqslant n} J(u_i) - J_* \leqslant \varepsilon. \tag{7}$$

Доказательство. Пусть u_* — какая-либо точка минимума $J(u)$ на $[a, b]$. Если $u_* = a$ или $u_* = b$, то $J_* = \min\{J(u_1); J(u_n)\}$, и неравенство (7) очевидно справедливо. Поэтому пусть $a < u_* < b$. Тогда $J'(u_*) = 0$ и используя разложение по Тейлору в точке u_*, для $J(u) \in R(M)$ получаем

$$J(u) = J_* + J''(\xi)(u - u_*)^2/2 \leqslant J_* + M(u - u_*)^2/2, \tag{8}$$

где $\xi = u_* + \theta(u - u_*)$ $(0 < \theta < 1)$. Система отрезков $[u_i - \sqrt{2\varepsilon/M},\ u_i + h_i]$ $(i = 1, \ldots, n)$ покрывает отрезок $[a, b]$, поэтому u попадает в один из отрезков этой системы при некотором i. Если $u_i - \sqrt{2\varepsilon/M} \leqslant u_* \leqslant u_i$, то из (8) при $u = u_i$ имеем $J(u_i) - J_* \leqslant (M/2)(2\varepsilon/M) = \varepsilon$. Если $u_i \leqslant u_* \leqslant u_i + h_i$, то аналогично $J(u_i) - J_* \leqslant Mh_i^2/2 = \varepsilon + J(u_i) - F_i$, или $F_i - J_* \leqslant \varepsilon$. Объединяя оба случая, получаем требуемое неравенство (7).

В худшем случае (например, если $J(u) = M(u-b)^2/2$) может оказаться, что $F_i = J(u_i)$ $(i \geqslant 1)$, и тогда метод (6) превратится в метод равномерного перебора с шагом $h = 2\sqrt{2\varepsilon/M}$. Если же $F_i < J(u_i)$ при некоторых i (например, для $J(u) = M(u-a)^2/2$), то методом (6) удается получить неравенство (7), вообще говоря, при меньшем n, чем методом равномерного перебора. Метод (6) можно несколько улучшить, приняв $F_i =$

$= \min \left\{ J(b);\ \min\limits_{1 \leqslant j \leqslant i} J(u_j) \right\}$. Недостатком метода (6) является требование знания постоянной $M \geqslant \sup\limits_{u \in [a,b]} J''(u)$.

Упражнения. 1. Пусть одним из вышеописанных методов покрытий найден $\min\limits_{1 \leqslant i \leqslant n} J(u_i) = J(u_k)$. Можно ли принять u_k за приближение к множеству U_*? Оценить погрешность $\rho(u_k, U_*)$ для метода (2) на классе $Q(L)$; рассмотреть функцию $J(u) = L(u-a) - \varepsilon/2$ при $a \leqslant u \leqslant a + \varepsilon/L$, $J(u) = \varepsilon(b-u)/(2(b-a-\varepsilon/L))$ при $a + \varepsilon/L \leqslant u \leqslant b$, где $\varepsilon > 0$ — малое число. Оценить $\rho(u_k, U_*)$ для методов (3), (6) на классах $Q(L)$ и $R(M)$ соответственно.

2. Найти оптимальный пассивный и оптимальный последовательный методы на классе функций $Q(L)$ [109, 282].

§ 8. Выпуклые функции одной переменной

Рассмотрим класс функций, для которых существует более эффективный вариант метода ломаных, когда ломаные составляются из отрезков касательных и лучше аппроксимируют минимизируемую функцию. Речь идет о выпуклых функциях, играющих важную роль в теории экстремальных задач.

Определение 1. Функция $J(u)$, определенная на отрезке $[a, b]$, называется *выпуклой* на этом отрезке, если

$$J(\alpha u + (1-\alpha)v) \leqslant \alpha J(u) + (1-\alpha) J(v) \tag{1}$$

при всех $u, v \in [a, b]$, $\alpha \in [0, 1]$.

Когда α пробегает отрезок $[0, 1]$, точки $(\alpha u + (1-\alpha)v, \alpha J(u) + (1-\alpha) J(v))$ на плоскости переменных (u, J) пробегают хорду AB, соединяющую точки $A = (u, J(u))$ и $B = (v, J(v))$ на графике функции $J = J(u)$. Поэтому неравенство (1) имеет простой геометрический смысл: график выпуклой функции на любом отрезке $[u, v] \subseteq [a, b]$ находится не выше хорды, соединяющей точки графика $(u, J(u))$ и $(v, J(v))$ (рис. 1.3). Примерами функций, выпуклых на любом отрезке, могут служить функции $J(u) = u^2$, $J(u) = |u|$, $J(u) = u$.

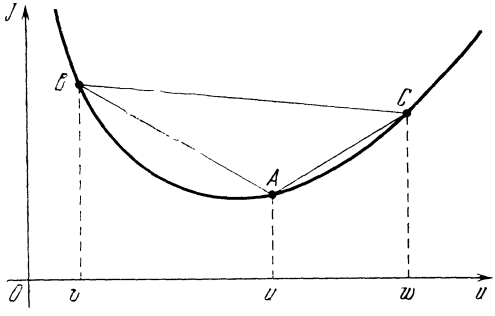

Рис. 1.3

Наряду с выпуклыми функциями в литературе рассматривают вогнутые функции.

Определение 2. Функция $J(u)$ называется *вогнутой* на отрезке $[a, b]$, если

$$J(\alpha u + (1-\alpha)v) \geqslant \alpha J(u) + (1-\alpha) J(v)$$

при всех $u, v \in [a, b]$, $\alpha \in [0, 1]$.

Между выпуклыми и вогнутыми функциями существует тесная связь: если $J(u)$ вогнута на $[a, b]$, то $-J(u)$ выпукла на этом же отрезке. Учитывая эту связь, достаточно ограничиться изучением свойств выпуклых функций.

Теорема 1. *Для выпуклости функции $J(u)$ на отрезке $[a, b]$ необходимо и достаточно, чтобы*

$$(J(u)-J(v))/(u-v)\leqslant(J(w)-J(v))/(w-v)\leqslant \\ \leqslant(J(w)-J(u))/(w-u) \qquad (2)$$

при всех u, v, w $(a\leqslant v<u<w\leqslant b)$.

Доказательство. Необходимость. Пусть функция $J(u)$ выпукла на $[a, b]$. Нетрудно проверить, что $u=\alpha v+(1-\alpha)w$, где $\alpha=(w-u)/(w-v)$ $(0<\alpha<1)$. Отсюда с учетом выпуклости функции $J(u)$ имеем $J(u)\leqslant(w-u)J(v)/(w-v)+(1-(w-u)/(w-v))J(w)$, или

$$(w-v)J(u)\leqslant(w-u)J(v)+(u-v)J(w).$$

Последнее неравенство можно переписать в двоякой форме:

$$(w-v)(J(u)-J(v))\leqslant(u-v)(J(w)-J(v)),$$

или

$$(w-u)(J(w)-J(v))\leqslant(w-v)(J(w)-J(u)),$$

откуда будет следовать (2).

Достаточность. Пусть $J(u)$ удовлетворяет одному из неравенств (2). Отправляясь от этого неравенства и проделав предыдущие преобразования в обратном порядке, убеждаемся в том, что $J(u)$ выпукла на $[a, b]$.

Нетрудно понять геометрический смысл неравенств (2) (см. рис. 1.3), если вспомнить, что $(J(u)-J(v))/(u-v)$ представляет собой угловой коэффициент хорды AB, соединяющей точки $A=(u, J(u))$ и $B=(v, J(v))$ на графике функции $J=J(u)$.

Теорема 2. *Выпуклая на отрезке $[a, b]$ функция $J(u)$ в каждой внутренней точке и отрезка $[a, b]$ непрерывна и имеет конечную правую производную $\lim\limits_{h\to+0}(J(u+h)-J(u))/h=J'(u+0)$, конечную левую производную $\lim\limits_{\tau\to+0}(J(u)-J(u-\tau))/\tau=J'(u-0)$, причем $J'(u-0)\leqslant J'(u+0)$ при всех $u\in(a, b)$.*

Доказательство. Из теоремы 1 следует, что

$$(J(u)-J(u-\tau))/\tau\leqslant(J(u)-J(u-h))/h\leqslant \\ \leqslant(J(u+h)-J(u))/h\leqslant(J(u+\tau)-J(u))/\tau \qquad (3)$$

при всех τ, h, лишь бы $0<h<\tau$ и точки u, $u\pm h$, $u\pm\tau\in(a, b)$ (рис. 1.4). Неравенства (3) означают, что величина $(J(u+h)-J(u))/h$ монотонно убывает при убывании h и ограничена снизу, например, величиной $(J(u)-J(u-\tau))/\tau$, не зависящей от h. Отсюда следует существование правой произ-

водной $J'(u+0)$. Аналогично доказывается существование левой производной $J'(u-0)$. Из (3) при $h \to +0$ получаем неравенство $J'(u-0) \leqslant J'(u+0)$. Из существования левой и правой производных следует непрерывность $J(u)$ при всех $u \in (a, b)$.

Заметим, что на концах отрезка $[a, b]$ выпуклая функция может не иметь соответствующей односторонней производной и, более того, здесь она может терпеть разрыв.

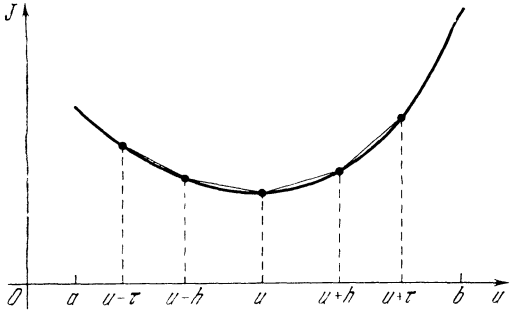

Рис. 1.4

Пример 1. Пусть $J(u) = u$ при $0 < u < 1$, $J(0) = J(1) = 2$. Очевидно, эта функция выпукла на $[0, 1]$, но на концах отрезка терпит разрывы.

Пример 2. Функция $J(u) = -\sqrt{1-u^2}$ выпукла и непрерывна на отрезке $[-1, 1]$, но на концах отрезка не имеет конечных производных $J'(1-0)$, $J'(-1+0)$.

Теорема 3. *Пусть функция $J(u)$ выпукла на отрезке $[a, b]$ и имеет конечные производные $J'(a+0)$, $J'(b-0)$. Тогда*

$$J'(a+0)(u-v) \leqslant J(u) - J(v) \leqslant J'(b-0)(u-v) \qquad (4)$$

при всех u, v $(a \leqslant v \leqslant u \leqslant b)$, так что $J(u)$ на $[a, b]$ удовлетворяет условию Липшица (6.1) с постоянной $L = \max\{|J'(a+0)|; |J'(b-0)|\}$.

Доказательство. Из теоремы 1 имеем

$$(J(a+h) - J(a))/h \leqslant (J(v) - J(a))/(v-a) \leqslant$$
$$\leqslant (J(u) - J(v))/(u-v) \leqslant (J(b) - J(u))/(b-u) \leqslant$$
$$\leqslant (J(b) - J(b-h))/h$$

для всех $h > 0$, $a + h < v < u < b - h$. Отсюда при $h \to +0$ получаем

$$J'(a+0) \leqslant (J(u) - J(v))/(u-v) \leqslant J'(b-0),$$

что равносильно (4) при любых u, v $(a < v < u < b)$. Неравенства (4) остаются верными также и при $v = a$ или $u = b$, так как при выполнении условий теоремы функция $J(u)$ непрерывна во всех точках отрезка $[a, b]$ и в (4) можно совершить предельный переход при $v \to a + 0$ или $u \to b - 0$.

Пример 2 показывает, что конечность величин $J'(a+0)$, $J'(b-0)$ существенна для выполнения условия Липшица (6.1).

Теорема 4. *Пусть функция $J(u)$ выпукла на отрезке $[a, b]$, а $l(v)$ — любая функция, удовлетворяющая неравенствам $J'(v-0) \leqslant l(v) \leqslant J'(v+0)$ при $a < v < b$. Тогда $l(v)$ не убы-*

вает при $v \in (a, b)$ *и справедливо неравенство*

$$J(u) \geqslant J(v) + l(v)(u - v), \quad u \in [a, b]. \tag{5}$$

Если, кроме того, $J(u)$ *дифференцируема во всех точках отрезка* $[a, b]$, *то*

$$J(u) \geqslant J(v) + J'(v)(u - v), \quad u \in [a, b], \tag{6}$$

при любом $v \in [a, b]$. *Если неравенство* (5) (*или* (6)) *обращается в равенство при некотором* $u = c \in [a, b]$ $(c \neq v)$, *то* $J(u) \equiv J(v) + l(v)(u - v)$ *при всех* u *из отрезка* $[c, v]$.

Доказательство. Перепишем неравенство (1) в виде $J(v + \alpha(u - v)) - J(v) \leqslant \alpha(J(u) - J(v))$ $(0 < \alpha < 1)$. Разделив обе части этого неравенства на α и перейдя к пределу при $\alpha \to +0$, получим $J(u) - J(v) \geqslant J'(v + 0)(u - v) \geqslant l(v)(u - v)$ при $u > v$ и $J(u) - J(v) \geqslant J'(v - 0)(u - v) \geqslant l(v)(u - v)$ при $u < v$. Неравенство (5) доказано. Заметим, что при $a < u < b$ переменные u, v в (5) входят равноправно, поэтому, меняя их ролями, получаем $J(v) \geqslant J(u) + l(u)(v - u)$ при всех $v \in [a, b]$. Сложим последнее неравенство с (5) почленно. Будем иметь $(l(u) - l(v))(u - v) \geqslant 0$ при всех $u, v \in (a, b)$, что равносильно монотонному возрастанию $l(v)$.

Пусть теперь $J(u)$ дифференцируема во всех точках $u \in [a, b]$. Тогда $J'(u + 0) = J'(u - 0) = J'(u)$ при всех $u \in [a, b]$. Полагая в (5) $l(v) = J'(v)$, убеждаемся в справедливости неравенства (6) при всех $u \in (a, b)$. Из существования конечных функций $J'(a + 0)$, $J'(b - 0)$ и из (4) следует, что (6) верно и при $v = a$, $v = b$.

Наконец, пусть $J(c) = J(v) + l(v)(c - v)$ при некотором $c \in [a, b]$ $(c \neq v)$. Возьмем произвольную точку $u = \alpha c + (1 - \alpha)v$ из отрезка $[c, v]$. Из выпуклости $J(u)$ тогда следует, что $J(u) \leqslant \alpha J(c) + (1 - \alpha)J(v) = \alpha(J(v) + l(v)(c - v)) + (1 - \alpha)J(v) = J(v) + l(v)(u - v)$ $(u \in [c, v])$. Сравнивая это неравенство с (5), заключаем, что $J(u) = J(v) + l(v)(u - v)$ при всех $u \in [c, v]$.

График линейной функции $J(v) + J'(v)(u - v)$ переменной $u \in [a, b]$ представляет собой касательную к графику $J = J(u)$ в точке v. Поэтому неравенство (6) означает, что график любой выпуклой дифференцируемой функции лежит не ниже любой касательной к нему. Обобщая понятие касательной на случай выпуклых недифференцируемых функций, прямую $g(u, v) = J(v) + l(v)(u - v)$, где $J'(v - 0) \leqslant l(v) \leqslant J'(v + 0)$, также будем называть *касательной* к графику $J = J(u)$ в точке v.

Следствие 1. *Пусть функция* $J(u)$ *выпукла на* $[a, b]$. *Тогда производные* $J'(u + 0)$, $J'(u - 0)$ *монотонно возрастают при* $u \in (a, b)$ (*если существуют конечные функции* $J'(a + 0)$, $J'(b - 0)$, *то утверждение справедливо на всем отрезке* $[a, b]$).

Доказательство этого утверждения непосредственно следует из теоремы 4, если в ней принять $l(v)=J'(v+0)$ или $l(v)=J'(v-0)$.

Теорема 5. *Пусть функция $J(u)$ выпукла на отрезке $[a, b]$ и $\lim\limits_{u\to a+0} J(u)=J(a)$, $\lim\limits_{u\to b-0} J(u)=J(b)$. Тогда множество U_* точек ее глобального минимума на $[a, b]$ непусто и все точки локального минимума $J(u)$ принадлежат U_*. Для того чтобы $u_*\in U_*$, необходимо и достаточно, чтобы*

$$J'(u_*+0)\geqslant 0, \quad J'(u_*-0)\leqslant 0 \tag{7}$$

(если $u_*=a$ или $u_*=b$, то (7) заменяется одним неравенством $J'(a+0)\geqslant 0$ или $J'(b-0)\leqslant 0$ соответственно).

Доказательство. Из условий на функцию $J(u)$ и теоремы 2 следует непрерывность $J(u)$ на $[a, b]$. Согласно теореме 1.1 тогда множество U_* непусто. Пусть u_* — какая-либо точка локального минимума $J(u)$. Тогда $J(u_*+h)-J(u_*)\geqslant 0$ при всех достаточно малых $|h|$, для которых $u_*+h\in[a, b]$. Разделив это неравенство на $h>0$ и $h<0$ и устремив h к нулю, получим условие (7). Заметим, что существование и конечность $J'(u_*\pm 0)$ при $a<u_*<b$ следует из теоремы 2. Если $u_*=a$, то существование и конечность $J'(a+0)$ следует из того, что $(J(a+h)-J(a))/h$ монотонно убывает при $h\to+0$ и ограничена снизу нулем. Аналогично доказывается существование и конечность $J'(b-0)$ при $u_*=b$. Таким образом, показано, что всякая точка локального минимума необходимо удовлетворяет условиям (7).

Пусть теперь некоторая точка $u_*\in(a, b)$ удовлетворяет условию (7). Положим в неравенстве (5) $v=u_*, l(v)=0$ и получим, что $J(u)\geqslant J(u_*)$ при всех $u\in[a, b]$. Это значит, что $u_*\in U_*$. Аналогично с использованием неравенств (4) рассматриваются случаи $u_*=a$ или $u_*=b$ и показывается, что $u_*\in U_*$. Отсюда следует, что всякая точка локального минимума выпуклой и непрерывной на $[a, b]$ функции является точкой ее глобального минимума на $[a, b]$.

Теорема 6. *Пусть функция $J(u)$ выпукла на отрезке $[a, b]$ и $\lim\limits_{u\to a+0} J(u)=J(a)$, $\lim\limits_{u\to b-0} J(u)=J(b)$; пусть U_* — множество точек минимума $J(u)$ на $[a, b]$ и v — некоторая точка $(a<v<b)$. Тогда для того чтобы $U_*\cap[a, v]=\varnothing$ $(U_*\cap[v, b]=\varnothing)$, необходимо и достаточно выполнения неравенства $J'(v+0)<0$ $(J'(v-0)>0)$. Для того чтобы $U_*\cap[a, v]\neq\varnothing$ $(U_*\cap[v, b]\neq\varnothing)$, необходимо и достаточно, чтобы $J'(v+0)\geqslant 0$ $(J'(v-0)\leqslant 0)$.*

Доказательство. Достаточность. Пусть $J'(v+0)<0$. Тогда согласно следствию 1 $J'(u+0)<0$ при всех $u\in[a, v]$. Из теоремы 5 тогда имеем $U_*\cap[a, v]=\varnothing$. Если $J'(v-0)>0$, то аналогично получаем $J'(u-0)>0$ при всех $u\in[v, b]$, так что $U_*\cap[v, b]=\varnothing$.

Необходимость. Пусть $U_* \cap [a, v] = \varnothing$. Допустим, что $J'(v+0) \geqslant 0$. Тогда возможно, что $J'(v-0) \leqslant 0$ или $J'(v-0) > 0$. Если $J'(v-0) \leqslant 0$, то из (7) следует, что $v \in U_*$. Если же $J'(v-0) > 0$, то по доказанному выше $U_* \cap [v, b] = \varnothing$ и, следовательно, $U_* \cap [a, v] \neq \varnothing$. В обоих случаях приходим к противоречию с тем, что $U_* \cap [a, v] = \varnothing$. Это значит, что при $U_* \cap [a, v] = \varnothing$ необходимо, чтобы $J'(v+0) < 0$. Аналогично доказывается, что если $U_* \cap [v, b] = \varnothing$, то необходимо, чтобы $J'(v-0) > 0$.

В справедливости последнего утверждения теоремы 6 легко убедиться рассуждением от противного со ссылкой на уже доказанное первое утверждение.

Теорема 7. *Если функция $J(u)$ выпукла на отрезке $[a, b]$ и $\lim\limits_{u \to a+0} J(u) = J(a)$, $\lim\limits_{u \to b-0} J(u) = J(b)$, то она унимодальна на $[a, b]$.*

Доказательство. Обозначим $u_* = \inf U_*$, $v_* = \sup U_*$. Из непрерывности $J(u)$ на $[a, b]$ и определения верхней и нижней грани множества U_* следует, что $u_*, v_* \in U_*$. Если $u_* = v_*$, то U_* состоит из одной точки u_*. Если $u_* < v_*$, то с учетом выпуклости $J(u)$ имеем $J_* = \inf\limits_{u \in [a,b]} J(u) \leqslant J(\alpha u_* + (1-\alpha) v_*) \leqslant \alpha J(u_*) + (1-\alpha) J(v_*)) = J_*$. Это значит, что $J(\alpha u_* + (1-\alpha) v_*) = J_*$ при всех $\alpha \in [0, 1]$, т. е. $U_* = [u_*, v_*]$.

Далее, так как $U_* \cap [a, v] = \varnothing$ для любого v $(a \leqslant v < u_*)$, то по теореме 6 имеем $J'(v+0) < 0$ при $a \leqslant v < u_*$. А тогда $J'(v+0) \leqslant (J(v+h) - J(v))/h < 0$ при всех достаточно малых h, т. е. $J(u)$ строго монотонно убывает при $a \leqslant u \leqslant u_*$. Аналогично показывается, что при $v_* \leqslant u \leqslant b$ функция $J(u)$ строго монотонно возрастает.

Как показывает пример 1, при нарушении условий теоремы 7 множество U_* может быть пустым. Приведем еще несколько примеров.

Пример 3. Функция $J(u) = u^2$ выпукла на отрезке $[-1, 1]$ и множество U_* состоит из единственной точки $u_* = 0$.

Пример 4. Функция $J(u) = |u| + |u-1|$ выпукла на отрезке $[-1, 2]$ и множество U_* представляет собой отрезок $U_* = [0, 1]$.

Пример 5. Пусть $J(u) = 0$ при $0 < u \leqslant 1$, $J(0) = 1$. Функция $J(u)$ выпукла на $[0, 1]$, но множество $U_* = \{u\colon 0 < u \leqslant 1\}$ не является отрезком. Здесь $\lim\limits_{u \to +0} J(u) \neq J(0)$ — нарушено одно из условий теоремы 7.

Критерий выпуклости функций, приведенный в теореме 1, не очень удобен для практической проверки. Приведем другие, часто более удобные критерии выпуклости функций.

Теорема 8. *Для того чтобы дифференцируемая функция $J(u)$ на отрезке $[a, b]$ была выпуклой, необходимо и достаточно, чтобы ее производная $J'(u)$ не убывала на $[a, b]$.*

Доказательство. Необходимость доказана в теореме 4, так как в рассматриваемом случае $l(v)=J'(v)$ $(v\in[a, b])$.

Достаточность. Пусть $J'(u)$ не убывает на $[a, b]$. Пусть $a\leqslant v<u<w\leqslant b$. Применяя формулу Лагранжа, имеем

$$(J(u)-J(v))/(u-v)=J'(\xi_1), \quad v<\xi_1<u,$$
$$(J(w)-J(u))/(w-u)=J'(\xi_2), \quad u<\xi_2<w.$$

По условию $J'(\xi_1)\leqslant J'(\xi_2)$, поэтому из предыдущих равенств следует одно из неравенств (2), что согласно теореме 1 равносильно выпуклости $J(u)$ на $[a, b]$.

Теорема 9. *Для того чтобы дважды дифференцируемая функция $J(u)$ на отрезке $[a, b]$ была выпуклой, необходимо и достаточно, чтобы $J''(u)\geqslant 0$ на $[a, b]$.*

Доказательство. Условие $J''(u)\geqslant 0$ является необходимым и достаточным для неубывания $J'(u)$ на $[a, b]$. Отсюда и из теоремы 8 следует требуемое.

Используя теоремы 8, 9, легко проверить, что функции $J(u)=a^u$, $J(u)=-\ln u$, $J(u)=u\ln u$ выпуклы на любом отрезке из области своего определения; функции $J(u)=u^r$ при $r\geqslant 1$, $r\leqslant 0$ и $J(u)=-u^r$ при $0<r<1$ выпуклы на любых отрезках $[a, b]$ $(0<a<b<\infty)$. Функция $J(u)=\sin u$ выпукла на отрезке $[-\pi, 0]$, но невыпукла на $[-\pi, \pi]$.

Упражнения. 1. Доказать, что если функция $J(u)$ выпукла на отрезке $[a, b]$, то $J'(u+0)=\inf\limits_{h>0}(J(u+h)-J(u))/h$, $J'(u-0)=\sup\limits_{h>0}(J(u)-J(u-h))/h$ при всех $u\in[a, b]$.

2. Пусть функция $J(u)$ выпукла на отрезке $[a, b]$. Доказать, что тогда $J'(u+0)$ непрерывна слева а $J'(u-0)$ непрерывна справа при всех u $(a<u<b)$.

Указание: воспользоваться непрерывностью $J(u)$, следствием 1 и упражнением 1.

3. Пусть $J(u)$ выпукла на $[a, b]$. Доказать, что $J'(v-0)\leqslant J'(v+0)\leqslant J'(u-0)\leqslant J'(u+0)$ при всех u, v $(a<v<u<b)$. Пользуясь этими неравенствами, показать, что $J(u)$ дифференцируема в точке v $(a<v<b)$ тогда и только тогда, когда одна из функций $J'(u+0)$ или $J'(u-0)$ непрерывна в точке v.

4. Пусть $J(u)$ выпукла на $[a, b]$. Пользуясь упражнениями 2, 3, доказать, что множества точек непрерывности функции $J'(u+0)$ и $J'(u-0)$ совпадают. Вывести отсюда, что множество точек, в которых $J(u)$ недифференцируема, не более чем счетно.

5. Пусть функция $J(u)$ непрерывна на отрезке $[a, b]$, дифференцируема на отрезках $[a, a_1]$, $[a_1, a_2]$, ..., $[a_{n-1}, a_n]$, $[a_n, b]$ $(a<a_1<\ldots<a_n<b)$, причем на каждом таком отрезке производная $J'(u)$ суммируема, не убывает и $J'(a_i-0)\leqslant J'(a_i+0)$ $(i=1, \ldots, n)$. Доказать, что тогда $J(u)$ выпукла на $[a, b]$.

6. Для выпуклости функции $J(u)$ на интервале (a, b) необходимо и достаточно, чтобы существовала функция $l(v)$ $(v\in(a, b))$ такая, что $J(u)\geqslant J(v)+l(v)(u-v)$ при всех $u\in(a, b)$. Необходимость доказана в теореме 4. докажите достаточность. Покажите, что $l(v)=J'(v)$ почти всюду на (a, b).

7. Пользуясь теоремой 3, доказать, что выпуклая на отрезке $[a, b]$ функция $J(u)$ абсолютно непрерывна на любом отрезке $[\alpha, \beta]\subset(a, b)$.

8. Если функция $f(t)$ возрастает на отрезке $[a, b]$ и суммируема на этом отрезке, то функция $J(u) = \int\limits_a^u f(t)\,dt$ выпукла на $[a, b]$. Доказать. Верно ли обратное утверждение?

9. Пусть $J(u)$ выпукла на $[a, b]$ и имеет обратную функцию. Можно ли утверждать, что обратная функция также будет выпуклой? Рассмотреть функции $J(u) = e^u$, $J(u) = e^{-u}$.

10. Пусть $J(u)$ выпукла на $[a, b]$ и $\lim\limits_{u\to a+0} J(u) = J(a)$, $\lim\limits_{u\to b-0} J(u) = J(b)$, а U_* — множество точек минимума $J(u)$ на $[a, b]$. Доказать, что $U_* \cap [\alpha, \beta] \neq \varnothing$ $(U_* \subset \operatorname{int}[\alpha, \beta])$ тогда и только тогда, когда $J'(\alpha - 0) \leqslant 0$, $J'(\beta + 0) \geqslant 0$ $(J'(\alpha - 0) < 0$, $J'(\beta + 0) > 0)$; здесь $a < \alpha < \beta < b$.

11. Доказать, что выпуклая на отрезке $[a, b]$ функция $J(u)$, отличная от постоянной, не может достигать своей верхней грани внутри отрезка $[a, b]$.

12. Пусть функция $J(u)$ выпукла и монотонно возрастает на отрезке $[a, b]$, а функция $z(x)$ выпукла на $[c, d]$, причем $z(x) \in [a, b]$ при всех $x \in [c, d]$. Доказать, что тогда сложная функция $g(x) = J(z(x))$ выпукла на $[c, d]$.

13. Назовем функцию $J(u)$ выпуклой на отрезке $[a, b]$, если $J((u + v)/2) \leqslant (J(u) + J(v))/2$ при всех $u, v \in [a, b]$. Доказать, что для непрерывных функций это определение выпуклой функции равносильно определению 1. Если не требовать непрерывности $J(u)$, то новое определение выделяет более широкий класс функций — см. пример из [312, с. 119].

14. Пусть $J(u)$ — выпуклая функция при $u \geqslant 0$, $J(0) \leqslant 0$. Доказать, что тогда функция $\varphi(u) = J(u)/u$ монотонно возрастает при $u > 0$. На примере функции $J(u) = 1 + u^2$ убедиться, что при $J(0) > 0$ это утверждение неверно.

Указание: воспользоваться равенством

$$J(u) = J\left(\frac{u}{u+h}(u+h) + \frac{h}{u+h}\cdot 0\right) \quad (h > 0).$$

15. Пусть функция $J(u)$ выпукла и дважды дифференцируема при $u \geqslant 0$, причем $\lim\limits_{u\to\infty} (uJ'(u) - J(u)) \leqslant 0$. Доказать, что тогда $\varphi(u) = J(u)/u$ монотонно убывает при $u > 0$.

Указание: вычислить производные функций $\varphi(u)$ и $nJ'(u) - J(u)$.

16. Доказать, что $(a + b)^n \leqslant 2^{n-1}(a^n + b^n)$ при всех $n \geqslant 1$, $a \geqslant 0$, $b \geqslant 0$.

Указание: воспользоваться выпуклостью функции $J(u) = u^n$ при $u \geqslant 0$, $n \geqslant 1$.

§ 9. Метод касательных

1. Пусть функция $J(u)$ выпукла и дифференцируема на отрезке $[a, b]$. Согласно теоремам 8.3, 8.7 такая функция удовлетворяет условию Липшица и унимодальна на $[a, b]$. Поэтому для минимизации $J(u)$ на $[a, b]$ применимы почти все описанные выше методы, в частности, метод ломаных из § 6. Однако если значения функции $J(u)$ и ее производной $J'(u)$ вычисляются достаточно просто, то здесь можно предложить другой, вообще говоря, более эффективный вариант метода ломаных, когда в качестве звеньев ломаных берутся отрезки касательных к графику $J(u)$ в соответствующих точках.

Зафиксируем какую-либо точку $v \in [a, b]$ и определим функцию $g(u, v) = J(v) + J'(v)(u - v)$, $a \leqslant u \leqslant b$.

Согласно теореме 8.4

$$g(u, v) \leqslant J(u) \qquad \forall\, u \in [a, b]. \tag{1}$$

В качестве начального приближения возьмем любую точку $u_0 \in [a, b]$ (например, $u_0 = a$), составим функцию $p_0(u) = g(u, u_0)$ и определим точку $u_1 \in [a, b]$ из условия $p_0(u_1) = \min\limits_{u \in [a,b]}$ (u (ясно, что при $J'(u_0) \neq 0$ будет $u_1 = a$ или $u_1 = b$).

Далее, берем новую функцию $p_1(u) = \max\{p_0(u);\ g(u, u_1)\}$ и следующую точку $u_2 \in [a, b]$ найдем из условия $p_1(u_2) = \min\limits_{u \in [a,b]} p_1(u)$, и т. д. Если точки $u_0, u_1, \ldots, u_n$ $(n \geqslant 1)$ уже известны, то составляем функцию $p_n(u) = \max\{p_{n-1}(u);\ g(u, u_n)\} = \max\limits_{0 \leqslant i \leqslant n} g(u, u_i)$, и следующую точку u_{n+1} определяем из условий $p_n(u_{n+1}) = \min\limits_{u \in [a,b]} p_n(u)$ $(u_{n+1} \in [a, b])$. Если при каком-либо $n \geqslant 0$ окажется, что $J'(u_n + 0) \geqslant 0$, $J'(u_n - 0) \leqslant 0$ (если $a < u_n < b$, то это равносильно условию $J'(u_n) = 0$), то согласно теореме 8.5 $u_n \in U_*$ — в этом случае задача минимизации уже решена и итерации на этом заканчиваются.

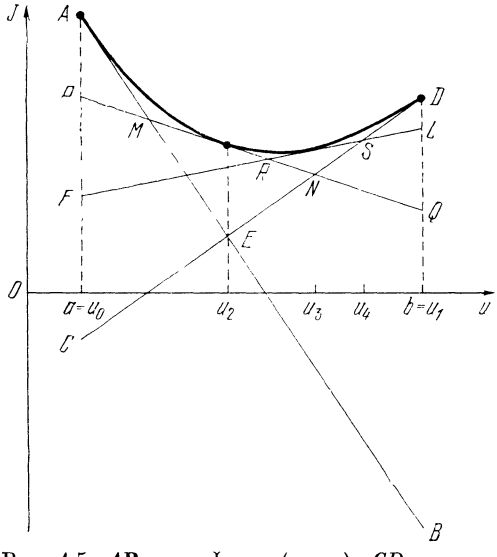

Рис. 1.5. AB — график $g(u, u_0)$, CD — график $g(u, u_1)$, AED — график $p_1(u)$, PQ — график $g(u, u_2)$, $AMND$ — график $p_2(u)$, FL — график $g(u, u_3)$, $AMRSD$ — график $p_3(u)$

Нетрудно видеть, что $p_n(u)$ — непрерывная кусочно линейная функция и ее график представляет собой ломаную, состоящую из отрезков касательных к графику функции $J(u)$ в точках $u_0, u_1, \ldots, u_n$ (рис. 1.5). Поэтому описанный метод естественно назвать *методом касательных*.

Теорема 1. *Пусть функция $J(u)$ на отрезке $[a, b]$ выпукла и дифференцируема, а последовательность $\{u_n\}$ получена описанным выше методом касательных, причем $u_n \notin U$ $(n = 0, 1, \ldots)$. Тогда:*

1) $\lim\limits_{n \to \infty} J(u_n) = \lim p_n(u_{n+1}) = J_*$ *и справедлива оценка*

$$0 \leqslant J(u_{n+1}) - J_* \leqslant J(u_{n+1}) - p_n(u_{n+1}), \qquad n = 1, 2, \ldots;$$

2) $\lim\limits_{n\to\infty} \rho(u_n, U_*) = 0$, *или точнее,* $\{u_n\}$ *имеет не более двух предельных точек, совпадающих с* $u_* = \inf U_*$ *или* $v_* = \sup U_*$.

Доказательство. Поскольку величины $J'(a+0)$, $J'(b-0)$ конечны по условию, то в силу теоремы 8.3 функция $J(u)$ удовлетворяет условию Липшица с постоянной $L = \max\{|J'(a)|; |J'(b)|\}$. Кроме того, согласно (1) и определению функции $p_n(u)$ имеем

$$p_{n-1}(u) \leqslant p_n(u) \leqslant J(u), \quad u \in [a, b], \qquad n = 1, 2, \ldots \tag{2}$$

Тогда $J(u_i) = g(u_i, u_i) \leqslant p_n(u_i) \leqslant J(u_i)$, т. е.

$$J(u_i) = p_n(u_i), \qquad i = 0, 1, \ldots, n. \tag{3}$$

Наконец, угловые коэффициенты касательных $g(u, u_i)$ равны $J'(u_i)$, причем $|J'(u_i)| \leqslant L$. Из теоремы 6.1 тогда следует, что $p_n(u)$ удовлетворяет условию Липшица с постоянной L. Отсюда с учетом (2), (3) с помощью тех же рассуждений, которые применялись при доказательстве теоремы 6.3, нетрудно убедиться в справедливости всех утверждений доказываемой теоремы. Остается лишь заметить, что из того, что функция $J(u)$ унимодальна и $u_n \notin U_* = [u_*, v_*]$ $(n \geqslant 0)$, равенство $\lim\limits_{n\to\infty} \rho(u_n, U_*) = 0$ возможно только в том случае, если предельными для $\{u_n\}$ могут быть лишь точки u_* или v_*.

2. Метод касательных обладает всеми достоинствами метода ломаных из § 6. Недостаток этого метода: он применим лишь в случае, когда минимизируемая функция выпукла и значения функции и ее производных вычисляются достаточно просто.

Можно предложить более удобную для использования на ЭВМ вычислительную схему метода касательных, которая не требует хранения в машинной памяти информации обо всей ломаной $p_n(u)$ при $u \in [a, b]$. А именно, возьмем $a_1 = a$, $b_1 = b$, вычислим $J'(a_1) = J'(a+0)$, $J'(b_1) = J'(b-0)$. Если $J'(a_1) \geqslant 0$ или $J'(b_1) \leqslant 0$, то по теореме 8.5 $a \in U_*$ или $b \in U_*$ — задача решена.

Поэтому, пусть $J'(a_1) < 0$, $J'(b_1) > 0$, что согласно теореме 8.6 означает $U_* \subset (a, b)$. Пусть отрезок $[a_{n-1}, b_{n-1}]$ $(n \geqslant 2)$ уже построен, причем $J'(a_{n-1}) < 0$, $J'(b_{n-1}) > 0$, $U_* \subset (a_{n-1}, b_{n-1})$. Обозначим через u_n точку пересечения касательных $g(u, a_{n-1})$ и $g(u, b_{n-1})$. Ясно, что $a_{n-1} < u_n < b_{n-1}$. Вычислим $J'(u_n)$. Если $J'(u_n) = 0$, то $u_n \in U_*$ — задача решена, итерации на этом заканчиваются. Если $J'(u_n) \neq 0$, то положим

$$a_n = \begin{cases} a_{n-1}, & J'(u_n) > 0, \\ u_n, & J'(u_n) < 0, \end{cases} \qquad b_n = \begin{cases} u_n, & J'(u_n) > 0, \\ b_{n-1}, & J'(u_n) < 0. \end{cases} \tag{4}$$

По построению $J'(a_n) < 0$, $J'(b_n) > 0$, и согласно теореме 8.6 $U_* \subset (a, b)$. Индуктивное описание метода закончено.

Из геометрических построений нетрудно усмотреть (см. рис. 1.5), что этот метод совпадает с описанным выше методом

касательных, в котором за начальную точку берется $u_0 = a$; строгое доказательство этого факта приводится в п. 5. В то же время приведенная схема метода более проста и удобна для реализации на ЭВМ; на каждом шаге метода здесь достаточно хранить в памяти ЭВМ величины a_n, b_n, $J(a_n)$, $J(b_n)$, $J'(a_n)$, $J'(b_n)$. Нетрудно выписать явное выражение для точки u_{n+1}, определяемой условием $g(u, a_n) = g(u, b_n)$ пересечения касательных в точках a_n, b_n при $J'(a_n) < 0$, $J'(b_n) > 0$:

$$u_{n+1} = \frac{J(a_n) - J(b_n) + b_n J'(b_n) - a_n J'(a_n)}{J'(b_n) - J'(a_n)}, \quad n \geqslant 1 \tag{5}$$

3. Поскольку ломаная из отрезков касательных аппроксимирует функцию $J(u)$, вообще говоря, лучше, чем ломаные из § 6, то следует ожидать, что метод касательных для выпуклых функций сходится быстрее метода ломаных из § 6. Исследуем скорость сходимости метода касательных, считая минимизируемую функцию дважды дифференцируемой.

Теорема 2. *Пусть функция $J(u)$ дважды непрерывно дифференцируема на $[a, b]$, $\inf\limits_{u \in [a,b]} J''(u) > 0$, u_* — точка минимума $J(u)$ на $[a, b]$. Пусть последовательность $\{u_n\}$ получена методом касательных при $u_0 = a$ по схеме* (4), (5), $a_1 = a$, $b_1 = b$, *причем $u_n \neq u_*$ $(n = 0, 1, \ldots)$. Тогда для любого числа $\varepsilon > 0$ существует номер $N = N(\varepsilon)$ такой, что*

$$|u_n - u_*| \leqslant ((1 + \varepsilon)/2)^{n-N} (b_N - a_N), \quad n \geqslant N. \tag{6}$$

Доказательство. Из теоремы 8.9 следует выпуклость функции $J(u)$ на $[a, b]$. Кроме того, так как $J'(u)$ строго возрастает, то множество U_* состоит из единственной точки u_*. Тогда из теоремы 1 имеем $\lim\limits_{n \to \infty} u_n = u_*$. Поскольку $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ $n = 1, 2, \ldots)$, то последовательность $\{a_n\}$ монотонно возрастает, а $\{b_n\}$ — монотонно убывает, причем $a_n < u_* < b_n$ $(n = 1, 2, \ldots)$. Покажем, что $\lim\limits_{n \to \infty} a_n = \lim\limits_{n \to \infty} b_n = u_*$.

В силу (4) либо $\{a_n\}$, либо $\{b_n\}$ является подпоследовательностью последовательности $\{u_n\}$ и сходится к u_*. Пусть для определенности $\lim\limits_{n \to \infty} a_n = u_*$. Допустим, что $\{b_n\}$ не сходится к u_*. Тогда согласно (4) последовательность $\{b_n\}$ не может быть подпоследовательностью для $\{u_n\}$, т. е. найдется номер $n_0 \geqslant 1$ такой, что $b_n = b_{n_0}$ при всех $n \geqslant n_0$. При $n \to \infty$ из (5) получим $J(u_*) = J(b_{n_0}) + J'(b_{n_0})(u_* - b_{n_0})$. В силу теоремы 8.4 это возможно только тогда, когда $J(u) \equiv J(b_{n_0}) + J'(b_{n_0})(u - b_{n_0})$ $(u_* \leqslant u \leqslant b_{n_0})$. Отсюда $J'(b_{n_0}) = J'(u_*) = 0$, что противоречит условию $J'(b_{n_0}) > 0$. Тем самым доказано, что обе последовательности $\{a_n\}$, $\{b_n\}$ сходятся к u_*.

Из представления (5) для точки u_{n+1} с помощью формулы Тейлора имеем

$$u_{n+1}-a_n=\frac{J(a_n)-J(b_n)-J'(b_n)(a_n-b_n)}{J'(b_n)-J'(a_n)}=\frac{1}{2}\frac{J''(\xi_n)}{J''(\mu_n)}(b_n-a_n),$$

$$b_n-u_{n+1}=\frac{J(b_n)-J(a_n)-J'(a_n)(b_n-a_n)}{J'(b_n)-J'(a_n)}=\frac{1}{2}\frac{J''(\eta_n)}{J''(\mu_n)}(b_n-a_n),$$

где ξ_n, η_n, μ_n — некоторые точки из отрезка $[a_n, b_n]$. Отсюда получаем, что $b_{n+1}-a_{n+1}\leqslant\max\{u_{n+1}-a_n;\ b_n-u_{n+1}\}\leqslant q_n(b_n-a_n)/2$, где $q_n=\max\{J''(\xi_n)/J''(\mu_n);\ J''(\eta_n)/J''(\mu_n)\}$. Поскольку последовательности $\{\xi_n\}$, $\{\eta_n\}$, $\{\mu_n\}$ вместе с $\{a_n\}$, $\{b_n\}$ стремятся к u_*, то в силу непрерывности функции $J''(u)$ и условия $\inf\limits_{u\in[a,b]}J''(u)>0$ имеем $\lim\limits_{n\to\infty}q_n=1$.

Следовательно, для любого $\varepsilon>0$ найдется номер $N=N(\varepsilon)$ такой, что $q_n\leqslant 1+\varepsilon$ при всех $n\geqslant N$. Тогда $b_{n+1}-a_{n+1}\leqslant q(b_n-a_n)$ $(n\geqslant N)$, где $q=(1+\varepsilon)/2$. Отсюда $b_n-a_n\leqslant q^{n-N}(b_N-a_N)$ $(n\geqslant N)$. Следовательно,

$$|u_n-u_*|\leqslant(b_n-a_n)\leqslant q^{n-N}(b_N-a_N),\qquad n\geqslant N.$$

4. Оценка (6) означает, что метод касательных сходится со скоростью, не меньшей скорости сходимости геометрической прогрессии со знаменателем $q=(1+\varepsilon)/2\approx 1/2$. Конечно, существуют выпуклые функции, для которых этот метод будет сходиться гораздо быстрее (возможно, например, что точка минимума найдется за конечное число шагов). Однако нетрудно привести пример, показывающий, что на классе дважды непрерывно дифференцируемых функций оценка (6) по порядку не может быть улучшена.

Пример 1. Пусть $J(u)=u^2(-1\leqslant u\leqslant 2)$. С помощью формулы (5) легко проверить, что касательные к параболе в точках a_n, b_n пересекаются в точке $u_{n+1}=(a_n+b_n)/2$. Возьмем $u_0=a_1=-1$, $u_1=b_1=2$. Тогда $u_2=1/2$. С помощью индукции нетрудно показать, что $a_n=-1/2^{n-1}$, $b_n=1/2^{n-2}$, $u_{n+1}=1/2^n$ при нечетных n и $a_n=-1/2^{n-2}$, $b_n=1/2^{n-1}$, $u_{n+1}=-1/2^n$ при четных n. Отсюда получается точная оценка $|u_n-u_*|=|n_n|=1/2^{n-1}$, в то время как, используя методику вывода оценки (5), имеем $|u_n-u_*|\leqslant b_n-a_n=3/2^{n-1}$ $(n\geqslant 1)$.

Таким образом, из примера 1 следует, что метод касательных на классе гладких выпуклых функций не лучше метода деления отрезка пополам. Более того, для этого класса функций нетрудно предложить вариант метода деления отрезка пополам, требующий лишь вычисления значений производных минимизируемой функции.

А именно, положим $u_0 = a_1 = a$, $u_1 = b_1 = b$, вычислим значения $J'(a_1) = J'(a+0)$, $J'(b_1) = J'(b-0)$. Если $J'(a_1) \geqslant 0$ или $J'(b_1) \leqslant 0$, то по теореме 8.5 имеем $a \in U$ или $b \in U_*$ — задача решена. Поэтому пусть $J'(a_1) < 0$, $J'(b_1) > 0$. Тогда $U_* \subset (a_1, b_1)$. Пусть отрезок $[a_{n-1}, b_{n-1}]$ $(n \geqslant 2)$ уже построен, причем $J'(a_{n-1}) < 0$, $J'(b_{n-1}) > 0$, так что $U_* \subset (a_{n-1}, b_{n-1})$. Положим $u_n = (a_{n-1} + b_{n-1})/2$ и вычислим $J'(u_n)$. Если $J'(u_n) = 0$, то $u_n \in U_*$ — задача решена. Если $J'(u_n) \neq 0$, то определим точки a_n, b_n по формулам (4), приняв в них $u_n = (a_{n-1} + b_{n-1})/2$. По построению $J'(a_n) < 0$, $J'(b_n) > 0$, и согласно теореме 8.6 $U_* \subset (a_n, b_n)$. Кроме того, ясно, что

$$b_n - a_n = (b_{n-1} - a_{n-1})/2 = (b_1 - a_1)/2^{n-1}, \qquad n = 1, 2, \ldots$$

Описанный метод деления отрезка пополам выгоднее применять при минимизации тех гладких выпуклых функций, у которых значения производных вычисляются проще, чем значения функции. Если же значения и функции, и ее производных вычисляются достаточно просто, то метод касательных может оказаться предпочтительнее — хотя, как мы выше убедились, метод касательных на классе гладких выпуклых функций в целом не лучше метода деления отрезка пополам, но в то же время нетрудно привести примеры таких выпуклых функций, для которых метод касательных сходится гораздо быстрее описанного метода деления отрезка пополам.

5. Метод касательных можно описать и без требования дифференцируемости выпуклой функции, используя лишь односторонние производные во внутренних точках отрезка $[a, b]$ — существование таких производных доказано в теореме 8.2. Чтобы гарантировать непустоту множества U_*, потребуем еще, чтобы $\lim\limits_{u \to a+0} J(u) = J(a)$, $\lim\limits_{u \to b-0} J(u) = J(b)$.

Положим $u_0 = a_1 = a + \delta$, $u_1 = b_1 = b - \gamma$, где δ, γ — достаточно малые положительные числа; если $J'(a+0)$ или $J'(b-0)$ конечны, то здесь можно взять $\delta = 0$ или $\gamma = 0$ соответственно. Вычислим $J'(a_1+0)$, $J'(b_1-0)$. Можем считать, что $J'(a_1+0) < 0$, $J'(b_1-0) > 0$, ибо в противном случае согласно теореме 8.6 либо $U_* \cap [a, a_1] \neq \varnothing$, либо $U_* \cap [b_1, b] \neq \varnothing$ — в обоих случаях точка из U_* находится с требуемой точностью δ или γ, и задача решена.

Абсциссу точки пересечения касательных $g(u, a_1) = J(a_1) + J'(a_1+0) \times (u - a_1)$ и $g(u, b_1) = J(b_1) + J'(b_1-0)(u - b_1)$ обозначим через u_2. Нетрудно видеть, что точка $u_1 = b_1$ доставляет минимум функции $p_0(u) = g(u, u_0)$, а в точке u_2 достигается минимум функции $p_1(u) = \max\{p_0(u); g(u, u_1)\}$ на $[a_1, b_1]$, причем $a_1 < u_2 < b_1$,

$$p_1(u) = \begin{cases} g(u, a_1), & u \in [a_1, u_2], \\ g(u, b_1), & u \in [u_2, b_1], \end{cases}$$

так что на $[a_1, u_2]$ функция $p_1(u)$ строго убывает, а на $[u_2, b_1]$ — строго возрастает (рис. 1.6).

Сделаем индуктивное предположение. Пусть определены точки $u_0, u_1, \ldots, u_{n-1}$ $(n \geqslant 2)$, найден отрезок $[a_{n-1}, b_{n-1}]$ такой, что $J'(a_{n-1}+0) < 0$, $J'(b_{n-1}-0) > 0$, причем a_{n-1}, b_{n-1} совпадают с одной из точек $u_0, u_1, \ldots, u_{n-1}$. Пусть u_n — точка пересечения касательных $g(u, a_{n-1}) = J(a_{n-1}) + J'(a_{n-1}+0)(u - a_{n-1})$ и $g(u, b_{n-1}) = J(b_{n-1}) + J'(b_{n-1}-0) \times$

$\times (u - b_{n-1})$, а функция $p_{n-1}(u) = \max\{p_{n-2}(u);\ g(u,\ u_{n-1})\}$ такова, что

$$p_{n-1}(u) = \begin{cases} g(u,\ a_{n-1}), & u \in [a_{n-1},\ u_n], \\ g(u,\ b_{n-1}), & u \in [u_n,\ b_{n-1}], \end{cases} \tag{7}$$

$p_{n-1}(u)$ на $[a_1,\ u_n]$ строго убывает, а на $[u_n,\ b_1]$ — строго возрастает: это значит, что u_n — точка минимума $p_{n-1}(u)$ на $[a_1,\ b_1]$. Вычислим $J'(u_n + 0)$,

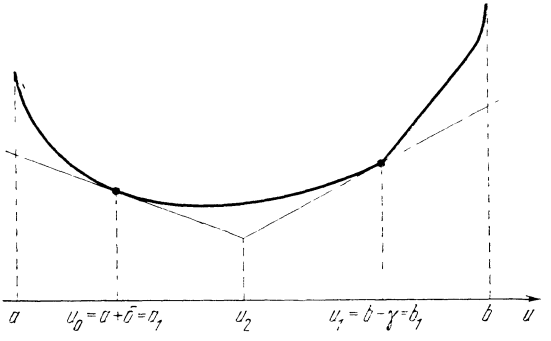

Рис. 1.6

$J'(u_n - 0)$. Если $J'(u_n + 0) \geqslant 0$, $J'(u_n - 0) \leqslant 0$, то по теореме 8.5 имеем $u_n \in U_*$, и задача решена — итерации на этом заканчиваются.

Поэтому пусть выполнено одно из условий

$$J'(u_n + 0) < 0, \quad J'(u_n - 0) > 0. \tag{8}$$

Положим $a_n = a_{n-1}$, $b_n = u_n$, $g(u,\ u_n) = J(u_n) + J'(u_n - 0)(u - u_n)$ при $J'(u_n - 0) > 0$ (рис. 1.7, *а*) и $a_n = u_n$, $b_n = b_{n-1}$, $g(u,\ u_n) = J(u_n) + J'(u_n + 0)(u - u_n)$ при $J'(u_n + 0) < 0$ (рис. 1.7, *б*). Полученный отрезок $[a_n,\ b_n]$ таков, что $[a_n,\ b_n] \subset [a_{n-1},\ b_{n-1}]$, $J'(a_n + 0) < 0$, $J'(b_n - 0) > 0$, и согласно теореме 8.6 тогда $U_* \subset (a_n,\ b_n)$.

Выясним, в какой точке функция $p_n(u) = \max\{p_{n-1}(u);\ g(u,\ u_n)\}$ достигает своего минимума на $[a_1,\ b_1]$. Из рис. 1.7 видно, что

$$p_n(u) = \begin{cases} p_{n-1}(u), & u \in [a_1,\ b_1] \setminus [u_n',\ u_n''], \\ g(u,\ u_n), & u \in [u_n',\ u_n''], \end{cases} \tag{9}$$

где u_n', u_n'' — абсциссы точек пересечения соответственно касательных $g(u,\ a_{n-1})$ и $g(u,\ u_n)$, $g(u,\ b_{n-1})$ и $g(u,\ u_n)$. Дадим строгое доказательство формулы (9).

Из теоремы 8.4 следует, что $g(u,\ u_i) \leqslant J(u)$ при всех $u \in [a_1,\ b_1]$ и $i = 0,\ 1,\ \ldots$, поэтому $p_{n-1}(u) \leqslant J(u)$. В частности, $p_{n-1}(u_n) \leqslant J(u_n)$. Допустим, что $p_{n-1}(u_n) = J(u_n)$. В силу (7) это значит, что точка $(u_n,\ J(u_n))$ лежит на касательных $g(u,\ a_{n-1})$, $g(u,\ b_{n-1})$. Согласно теореме 8.4 тогда $J(u) = p_{n-1}(u)$ при $u \in [a_{n-1},\ b_{n-1}]$. Отсюда имеем

$$J'(u_n - 0) = g'(u_n,\ a_{n-1}) = J'(a_{n-1} + 0) < 0,$$
$$J'(u_n + 0) = g'(u_n,\ b_{n-1}) = J'(b_{n-1} - 0) > 0,$$

что противоречит (8). Таким образом, при выполнении (8) возможен лишь случай $p_{n-1}(u_n) < J(u_n)$.

Далее, в силу теоремы 8.4 $g(a_{n-1}, u_n) \leqslant J(a_{n-1})$. Если $g(a_{n-1}, u_n) = J(a_{n-1})$, то из той же теоремы следует, что $J(u) = g(u, u_n)$ при $u \in [a_{n-1}, u_n]$, и поэтому $g'(a_{n-1}, u_n) = J'(a_{n-1}+0)$ и $J(u) = g(u, a_{n-1})$ $(u \in [a_{n-1}, u_n])$. Тогда $J(u_n) = g(u_n, a_n) = g(u_n, a_{n-1}) = p_{n-1}(u_n)$. Однако выше было показано, что при выполнении (8) это невозможно. Следовательно, $g(a_{n-1}, u_n) < J(a_{n-1}) = g(a_{n-1}, a_{n-1})$. Аналогично доказывается, что $g(b_{n-1}, u_n) < J(b_{n-1}) = g(b_{n-1}, b_{n-1})$.

Таким образом, при выполнении (8) имеем $g(u_n, u_n) = J(u_n) > p_{n-1}(u_n)$, $g(a_{n-1}, u_n) < g(a_{n-1}, a_{n-1})$, $g(b_{n-1}, u_n) < g(b_{n-1}, b_{n-1})$. Это

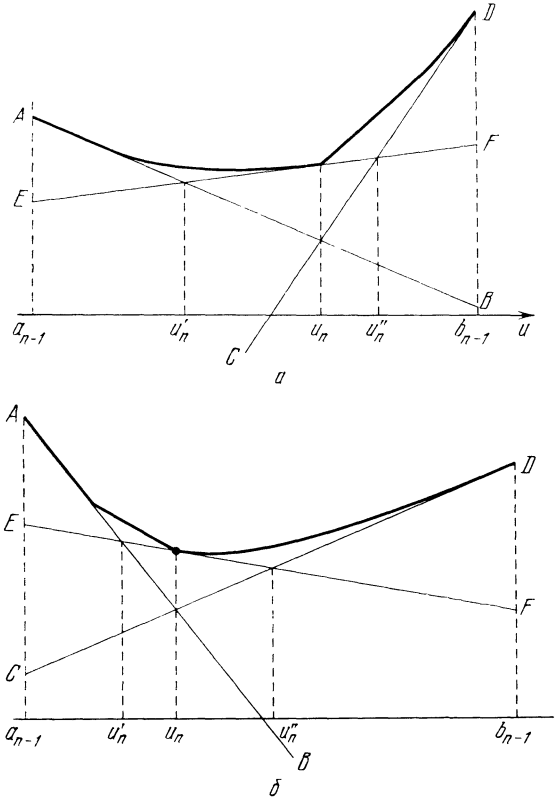

Рис. 1.7. *а*) $J'(u_n - 0) > 0$, *б*) $J'(u_n + 0) < 0$; AB — график $g(u, a_{n-1})$, CD — график $g(u, b_{n-1})$, EF — график $g(u, u_n)$

значит, что касательные $g(u, a_{n-1})$ и $g(u, u_n)$ пересекаются в некоторой точке с абсциссой u'_n $(a_{n-1} < u'_n < u_n)$, а касательные $g(u, u_n)$, $g(u, b_{n-1})$ — в точке с абсциссой u''_n $(u_n < u''_n < b_{n-1})$, причем $g(u, u_n) > p_{n-1}(u)$ при $u'_n < u < u''_n$, $g(u, a_{n-1}) > g(u, u_n)$ при $a_1 \leqslant u < u'_n$, $g(u, b_{n-1}) > g(u, u_n)$ при $u''_n < u \leqslant b_1$, Отсюда следует, что $p_{n-1}(u) = \max\limits_{0 \leqslant i \leqslant n-1} g(u, u_i) \geqslant$

$\geqslant g(u, a_{n-1}) > g(u, u_n)$ при $a_1 \leqslant u < u_n'$, $p_{n-1}(u) \geqslant g(u, b_{n-1}) > g(u, u_n)$ при $u_n'' < u \leqslant b_1$. Таким образом, формула (9) доказана.

Из предположения индукции следует, что $p_{n-1}(u)$ строго возрастает на $[u_n'', b_1]$ и строго убывает на $[a_1, u_n']$, так что минимум $p_n(u)$ на $[a_1, b_1]$ достигается в одной из точек u_n' или u_n''. Нетрудно видеть, что при $J'(u_n - 0) > 0$ минимум функции $p_n(u)$ достигается в точке u_n', а при $J'(u_n + 0) < 0$ — в точке u_n'' (см. рис. 1.7). Вспоминая определение точек u_n', u_n'', a_n, b_n, можем сделать следующие выводы: минимум $p_n(u)$ достигается в точке u_{n+1}, являющейся абсциссой точки пересечения $g(u, a_n)$ и $g(u, b_n)$ $(a_n < u_{n+1} < b_n)$, причем

$$p_n(u) = \begin{cases} g(u, a_n), & u \in [a_n, u_{n+1}], \\ g(u, b_n), & u \in [u_{n+1}, b_n], \end{cases}$$

функция $p_n(u)$ на $[a_1, u_{n+1}]$ строго убывает, а на $[u_{n+1}, b_1]$ — строго возрастает.

Индуктивное описание метода касательных для выпуклых функций закончено. При его реализации будем иметь систему вложенных друг в друга отрезков $[a_n, b_n] \subset \ldots \subset [a_1, b_1]$, причем согласно теоремам 8.5, 8.6 процесс либо завершится при каком-либо n определением точки $u_n \in U_*$, либо $U_* \subset (a_n, b_n)$ при всех $n = 1, 2, \ldots$ Теорема 1 остается верной и для этого метода. Поскольку минимум $p_n(u)$ на $[a_1, b_1]$ достигается на $[a_n, b_n]$ точке u_{n+1}, то информацию о ломаной $p_n(u)$ вне $[a_n, b_n]$ хранить в памяти ЭВМ нет необходимости. Это значит, что на n-м шаге в памяти ЭВМ достаточно хранить величины a_n, b_n, $J(a_n)$, $J(b_n)$, $J'(a_n + 0)$, $J'(b_n - 0)$.

Нетрудно видеть, что для гладкой выпуклой функции описанный вариант метода касательных совпадает с методами п. 1, 2, примененными к отрезку $[a + \delta, b - \gamma]$ с выбором начальной точки $u_0 = a + \delta$; попутно получено строгое доказательство равносильности методов из п. 1, 2.

Упражнение. Применить метод касательных для минимизации функций $J(u) = |u|$, $J(u) = u^2$, $J(u) = |u| + (u-1)^2$ на отрезках $[-1, 1]$, $[-1, 2]$, $[0, 1]$, $[1, 2]$.

§ 10. Метод поиска глобального минимума

В § 6, 7 были рассмотрены методы, пригодные для поиска глобального минимума функций, удовлетворяющих условию Липшица. Эти методы требовали для своей реализации знания постоянной Липшица минимизируемой функции. Опишем другой метод поиска глобального минимума [279], не требующий априорного знания постоянной Липшица и вместо этого использующий угловые коэффициенты хорд минимизируемой функции между определенными точками, получающимися при поиске минимума.

Итак, пусть функция $J(u)$ определена на отрезке $[a, b]$. Поиск минимума функции $J(u)$ на этом отрезке начинается с выбора двух точек $u_0 = v_0 = a$, $u_1 = v_1 = b$ и вычисления величины $L_1 = |J(u_1) - J(u_0)|/(u_1 - u_0)$. После этого полагаем $d_1 = \mu_1 L_1$ при $L_1 > 0$ и $d_1 = \nu$ при $L_1 = 0$; здесь $\nu > 0$, $\mu_1 > 1$ — параметры метода. Затем определяем следующую точку $v_2 = (u_1 + u_0)/2 - (J(u_1) - J(u_0))/(2d_1)$. Нетрудно показать (см. ниже общий случай), что $u_0 < v_2 < u_1$.

Перенумеруем точки u_0, u_1, v_2 в порядке их возрастания, приняв $u_0 = a$, $u_1 = v_2$, $u_2 = b$. Предположим, что уже сделано k шагов и найдены точки $u_0, u_1, \ldots, u_k$ $(k \geqslant 2)$, причем $u_0 = a < u_1 < \ldots < u_{k-1} < u_k = b$.

Опишем $k+1$-й шаг метода. Определим величины

$$L_k = \max_{1 \leqslant i \leqslant k} \frac{|J(u_i) - J(u_{i-1})|}{u_i - u_{i-1}}, \tag{1}$$

$$d_k = \begin{cases} \mu_k L_k, & L_k > 0, \\ \nu, & L_k = 0, \end{cases} \tag{2}$$

где $\nu > 0$, $\mu_k > 1$ — параметры метода, выбираемые вычислителем.

Для каждого отрезка $[u_{i-1}, u_i]$ введем следующую числовую характеристику:

$$R_k(i) = d_k(u_i - u_{i-1}) + \frac{(J(u_i) - J(u_{i-1}))^2}{d_k(u_i - u_{i-1})} - \\ - 2(J(u_i) + J(u_{i-1})), \qquad i = 1, \ldots, k. \tag{3}$$

После этого перебором значений $R_k(i)$ $(i = 1, \ldots, k)$ определим минимальный номер s, на котором достигается $\max\limits_{1 \leqslant i \leqslant k} R_k(i)$, т. е.

$$R_k(s) = \max_{1 \leqslant i \leqslant k} R_k(i), \quad R_k(i) < R_k(s), \qquad i = 1, \ldots, s-1. \tag{4}$$

Затем положим

$$v_{k+1} = (u_s + u_{s-1})/2 - (J(u_s) - J(u_{s-1}))/(2d_k) \tag{5}$$

и перенумеруем точки $u_0, u_1, \ldots, u_k, v_{k+1}$ в порядке их возрастания, чтобы $u_0 = a < u_1 < \ldots < u_k < u_{k+1} = b$. Метод описан. Будем предполагать, что параметры из (2) таковы, что

$$\nu > 0; \quad \mu_k \geqslant \mu > 1, \quad k = 1, 2, \ldots; \quad \sup_{k \geqslant 1} \mu_k < \infty. \tag{6}$$

Точки v_k или u_i, получаемые методом (1)—(6), будем кратко называть *поисковыми точками*. Из описания метода видно, что каждая поисковая точка имеет двойную нумерацию и двойное обозначение: если точка появилась впервые и это случилось на s-м шаге, то она получает номер s и обозначается через v_s; на каждом последующем шаге с номером $k \geqslant s$ та же точка v_s в зависимости от своего местоположения на $[a, b]$ получает один из номеров i $(0 \leqslant i \leqslant k)$ и обозначается через u_i — здесь номер $i = i(k)$ зависит от номера шага. В зависимости от обстоятельств мы будем пользоваться обоими обозначениями поисковых точек.

Покажем, что точка v_{k+1}, определяемая формулой (5), удовлетворяет неравенствам $u_{s-1} < v_{k+1} < u_s$ (здесь $s = s(k)$). В самом деле, из (5) имеем

$$\begin{aligned} u_s - v_{k+1} &= \frac{u_s - u_{s-1}}{2}\left(1 + \frac{J(u_s) - J(u_{s-1})}{d_k(u_s - u_{s-1})}\right), \\ v_{k+1} - u_{s-1} &= \frac{u_s - u_{s-1}}{2}\left(1 - \frac{J(u_s) - J(u_{s-1})}{d_k(u_s - u_{s-1})}\right). \end{aligned} \tag{7}$$

Поскольку в силу (1), (2), (6)

$$\frac{|J(u_i) - J(u_{i-1})|}{d_k(u_i - u_{i-1})} \leqslant \frac{L_k}{d_k} \leqslant \frac{1}{\mu} < 1, \qquad i = 1, \ldots, k, \tag{8}$$

то из (7) следует, что $u_{s-1} < v_{k+1} < u_s$.

Таким образом, вслед за новой поисковой точкой v_{k+1} на $k+1$-м шаге появляются два новых отрезка $[u_{s-1}, u_{k+1}]$ и $[v_{k+1}, u_s]$, длина которых с помощью (7), (8) оценивается так:

$$(u_s - u_{s-1})(1 - 1/\mu)/2 \leqslant \min\{u_s - v_{k+1};\ v_{k+1} - u_{s-1}\} \leqslant$$
$$\leqslant \max\{u_s - v_{k+1};\ v_{k+1} - u_{s-1}\} \leqslant (u_s - u_{s-1})(1 + 1/\mu)/2, \qquad (9)$$

где $0 < q = (1 + 1/\mu)/2 < 1$, $s = s(k)$ $(k = 1, 2, \ldots)$. Кроме того, из неравенства (6.7) имеем

$$\frac{|J(u_s) - J(u_{s-1})|}{u_s - u_{s-1}} \leqslant \max\left\{\frac{|J(u_s) - J(v_{k+1})|}{u_s - v_{k+1}};\ \frac{|J(v_{k+1}) - J(u_{s-1})|}{v_{k+1} - u_{s-1}}\right\}.$$

Это значит, что $L_{k+1} \geqslant L_k$ $(k = 1, 2, \ldots)$. Поэтому если $L_k = 0$ $(k = 1, 2, \ldots)$, то согласно (2), (6) $d_k = \nu > 0$ $(k = 1, 2, \ldots)$; если же $L_k = 0$ при $k = 1, \ldots, k_0 - 1$, $L_{k_0} > 0$, то $d_k \geqslant \min\{\nu;\ \mu_{L_{k_0}}\} > 0$ $(k = 1, 2, \ldots)$. Иначе говоря, существует постоянная d такая, что

$$d_k \geqslant d > 0, \qquad k = 1, 2, \ldots \qquad (10)$$

С другой стороны, если функция $J(u)$ на $[a, b]$ удовлетворяет условию Липшица, т. е.

$$|J(u) - J(v)| \leqslant L|u - v|, \forall u, v \in [a, b], \qquad L = \text{const} > 0, \qquad (11)$$

то $L_k \leqslant L$, и следовательно,

$$d_k \leqslant \max\left\{\nu;\ L \sup_{k \geqslant 1} \mu_k\right\} < \infty, \qquad k = 1, 2, \cdots \qquad (12)$$

Теорема 1. *Пусть функция $J(u)$ на отрезке $[a, b]$ удовлетворяет условию* (11). *Пусть последовательность $\{v_k\}$ построена методом* (1)—(6) *и v_* — какая-либо предельная точка $\{v_k\}$. Тогда $\lim\limits_{k \to \infty} J(v_k) = J(v_*)$ и, кроме того,*

$$J(v_*) \leqslant J(v_i), \qquad i = 0, 1, \cdots \qquad (13)$$

Доказательство. Поскольку $v_i \neq v_k$ при $i \neq k$ и v_* — предельная точка $\{v_k\}$, то существует последовательность отрезков $[u_{m(k)-1}, u_{m(k)}]$ $(k = 1, 2, \ldots)$, каждый из которых содержит точку v_* и бесконечно много поисковых точек, отличных от v_* и, кроме того, $u_{m(k)-1} \leqslant u_{m(k+1)-1} < u_{m(k+1)} \leqslant u_{m(k)}$ $(k = 1, 2, \ldots)$.

Введем множество $N(v_*)$, состоящее из всех тех и только тех номеров $k \geqslant 1$, для которых либо $u_{m(k)-1} < u_{m(k+1)-1}$, либо $u_{m(k+1)} < u_{m(k)}$. Это значит, что при $k \in N(v_*)$ поисковая точка v_{k+1} попадает внутрь отрезка $[u_{m(k)-1}, u_{m(k)}]$, образуя один из концов следующего отрезка $[u_{m(k+1)-1}, u_{m(k+1)}]$. Отсюда в силу условия (4) имеем

$$R_k(m(k)) \geqslant R_k(i), \qquad i = 1, \ldots, k, \qquad k \in N(v_*). \qquad (14)$$

Поскольку каждый из отрезков $[u_{m(k)-1}, u_{m(k)}]$ содержит бесконечно много поисковых точек, то множество $N(v_*)$ также содержит бесконечно много номеров. Согласно оценке (9)

$$0 < u_{m(k+1)} - u_{m(k+1)-1} \leqslant q(u_{m(k)} - u_{m(k)-1}), \qquad k \in N(v_*),$$

где $0 < q < 1$. Отсюда и из счетности множества $N(v_*)$ следует, что

$$\lim_{k \to \infty} (u_{m(k)} - u_{m(k)-1}) = 0, \qquad \lim_{k \to \infty} u_{m(k)-1} = \lim_{k \to \infty} u_{m(k)} = v_*. \qquad (15)$$

Перепишем формулу (3) в виде

$$R_k(i) = d_k(u_i - u_{i-1})\left(1 + \frac{J(u_i) - J(u_{i-1})}{d_k(u_i - u_{i-1})}\right)^2 - 4J(u_i), \qquad i = 1, \ldots, k. \tag{16}$$

Взяв здесь $i = m(k)$, с учетом соотношений (8), (11), (12), (15) получим

$$\lim_{k\to\infty} R_k(m(k)) = -4J(v_*). \tag{17}$$

Далее, из (16) и (14) следует, что $-4J(u_i) < R_k(i) \leqslant R_k(m(k))$ для каждого $k \in N(v_*)$ и всех $i = 1, \ldots, k$. Кроме того, если (3) запишем в виде

$$R_k(i) = d_k(u_i - u_{i-1})\left(1 - \frac{J(u_i) - J(u_{i-1})}{d_k(u_i - u_{i-1})}\right)^2 - 4J(u_{i-1}), \quad i = 1, \ldots, k, \tag{18}$$

и примем здесь $i = 1$, то с учетом (14) будем иметь $-4J(u_0) < R_k(1) \leqslant \leqslant R_k(m(k))$ $(k \in N(v_*))$.

Таким образом, $4J(u_i) > -R_k(m(k))$ при всех $i = 0, 1, \ldots, k$ $(k \in N(v_*))$, т. е. $4 \min\limits_{0\leqslant i\leqslant k} J(u_i) > -R_k(m(k))$ $(k \in N(v_*))$. Поскольку $\min\limits_{0\leqslant i\leqslant k} J(u_i) = \min\limits_{0\leqslant s\leqslant k} J(v_s) \leqslant \min\limits_{0\leqslant s\leqslant p} J(v_s)$ при любом $p \leqslant k$, то из предыдущего неравенства имеем $4 \min\limits_{0\leqslant s\leqslant p} J(v_s) > -R_k(m(k))$ для всех $p \leqslant k$ $(k \in \in N(v_*))$. Перейдем здесь к пределу при $k \to \infty$, $k \in N(v_*)$. С учетом (17) получим $\min\limits_{0\leqslant s\leqslant p} J(v_s) \geqslant J(v_*)$ при каждом фиксированном $p \geqslant 1$. Оценка (13) доказана.

Докажем, что $\lim\limits_{k\to\infty} J(v_k) = J(v_*)$. Пусть c — какая-либо предельная точка последовательности $\{J(v_k)\}$ и пусть $c = \lim\limits_{n\to\infty} J(v_{k_n})$. Выбирая при необходимости подпоследовательность, можем считать, что $\{v_{k_n}\}$ сходится к некоторой точке w_*. Тогда из (13) следует, что $J(v_*) \leqslant J(w_*) = c$. С другой стороны, так как оценка (13) доказана для произвольной предельной точки $\{v_k\}$, то $J(w_*) \leqslant J(v_i)$ $(i = 0, 1, \ldots)$ и, следовательно, $J(w_*) \leqslant J(v_*)$. Таким образом, $J(w_*) = c = J(v_*)$, т. е. $\{J(v_k)\}$ имеет единственную предельную точку $c = J(v_*)$. Это означает, что $\lim\limits_{k\to\infty} J(v_k) = J(v_*)$.

Теорема 2. *Пусть выполнены все условия теоремы 1 и, кроме того, пусть функция $J(u)$ на $[a, b]$ имеет конечное число локальных экстремумов. Тогда любая предельная точка последовательности $\{v_k\}$ является точкой локального минимума функции $J(u)$ со значением $c = \lim\limits_{k\to\infty} J(v_k)$.*

Доказательство. Сначала покажем, что если предельная точка v_* последовательности $\{v_k\}$ такова, что $a < v_* < b$, то из $\{v_k\}$ можно выбрать две подпоследовательности, одна из которых стремится к v_* слева, другая — справа. Возьмем отрезки $[u_{m(k)-1}, u_{m(k)}]$ $(k = 1, 2, \ldots)$ и множество $N(v_*)$, введенные при доказательстве теоремы 1. Если оказалось, что $u_{m(k)-1} < v_* < v_{m(k)}$ $(k = 1, 2, \ldots)$, то с учетом (15) в качестве искомых подпоследовательностей можно взять $\{u_{m(k)-1}\}$ и $\{u_{m(k)}\}$. Остается рассмотреть случаи, когда, начиная с какого-либо s-го шага, $*$ будет совпадать с одним из концов отрезка $[u_{m(k)-1}, u_{m(k)}]$.

Пусть сначала $v_* = u_{m(k)-1}$ $(k \geqslant s)$. Тогда подпоследовательность $\{u_{m(k)}\}$ в силу (15) сходится к v_* справа. Рассмотрим подпоследовательность $\{u_{m(k)-1}\}$ при $k \in N(v_*)$. С помощью формулы (16) при $i = m(k) - 1$

и соотношений (14), (17) имеем

$$d_k(v_* - u_{m(k)-2})\left(1 + \frac{J(v_*) - J(u_{m(k)-2})}{d_k(v_* - u_{m(k)-2})}\right)^2 =$$

$$= R_k(m(k) - 1) + 4J(v_*) \leqslant R_k(m(k)) + 4J(v_*) \to 0, \quad k \to \infty, \qquad k \in N(v_*).$$

Отсюда с учетом (8), (10) получим, что подпоследовательность $\{u_{m(k)-2}\}$ сходится к v_* слева.

Пусть теперь $v_* = u_{m(k)}$ $(k \geqslant s)$. Тогда $\{u_{m(k)-1}\}$ сходится к v_* слева. С другой стороны, из (18) при $i = m(k) + 1$ и из (14), (17) следует, что

$$d_k(u_{m(k)+1} - v_*)\left(1 - \frac{J(u_{m(k)+1}) - J(v_*)}{d_k(u_{m(k)+1} - v_*)}\right)^2 =$$

$$= R_k(m(k) + 1) + 4J(v_*) \leqslant R_k(m(k)) + 4J(v_*) \to 0, \quad k \to \infty, \qquad k \in N(v_*).$$

Отсюда, учитывая (8), (10), имеем $\{u_{m(k)+1}\} \to v$ при $k \to \infty$ $(k \in N(v_*))$.

Искомые подпоследовательности для предельных точек v_* $(a < v_* < b)$ построены. Заметим, что если $v_* = a$, то в силу (15) $\{u_{1(k)}\}$ сходится к v_* справа, а если $v_* = b$, то $\{u_{k(k)-1}\}$ сходится к v_* слева.

Перенумеруем точки θ_i $(i = 0, 1, \ldots, r)$ локального экстремума функции $J(u)$ на $[a, b]$ в порядке возрастания: $\theta_0 = a < \theta_1 < \ldots < \theta_r = b$. Тогда на каждом отрезке $[\theta_i, \theta_{i+1}]$ функция $J(u)$ строго монотонна, причем при переходе через внутреннюю точку θ_i направление монотонности меняется. Пусть v_* — произвольная предельная точка $\{v_k\}$. По доказанному существуют подпоследовательности, сходящиеся к v_* слева и справа (при $v_* = a$ и $v_* = b$ достаточно односторонней сходимости). Отсюда и из (13) следует, что при переходе через точку v_* функция $J(u)$ меняет направление монотонности, причем $J(u)$ вблизи от v_* слева строго убывает, справа строго возрастает. Следовательно, v_* совпадает с одной из точек θ_i и представляет собой точку локального минимума $J(u)$ со значением $J(v_*) = \lim\limits_{k\to\infty} J(v_k)$.

Из теорем 1, 2 и 1.1 непосредственно вытекает

Следствие 1. *Пусть функция $J(u)$ на отрезке $[a, b]$ удовлетворяет условию* (11), *множество $U_* = \{u\colon u \in [a, b],\ J(u) = J_* = \inf\limits_{v\in[a,b]} J(v)\}$ состоит из конечного числа точек и вне U_* других локальных минимумов $J(u)$ на $[a, b]$ не имеет. Тогда последовательность $\{v_k\}$, полученная методом* (1)—(6), *будет минимизирующей для $J(u)$ на $[a, b]$ и* $\lim\limits_{k\to\infty} \rho(v_k, U_*) = 0$.

Сходимость метода (1)—(6) можно доказать и без требования конечности множества точек локального минимума функции.

Теорема 3. *Пусть функция $J(u)$ на отрезке $[a, b]$ удовлетворяет условию* (11); *кроме того, пусть в методе* (1)—(6), *начиная с некоторой итерации, $d_k \geqslant 2L$. Тогда последовательность $\{v_k\}$, полученная этим методом, такова, что:*

1) $\lim\limits_{k\to\infty} J(v_k) = J_*$, $\lim\limits_{k\to\infty} \rho(v_k, U_*) = 0$;

2) *при* $\inf\limits_{k\geqslant k_0} d_k > 2L$ *для некоторого* $k_0 \geqslant 1$ *все точки из U_* будут предельными точками $\{v_k\}$.*

Доказательство. Прежде всего заметим, что случай $d_k \geqslant 2L$ согласно (2) соответствует параметрам $\nu \geqslant 2L$ при $L_k = 0$ и $\mu_k \geqslant 2L/L_k$ при $L_k > 0$. Поскольку если $L_k > 0$, то из $L_{k_0} \leqslant L_k \leqslant L$ следует $1 \leqslant L/L_k \leqslant L/L_{k_0} < \infty$ при всех $k \geqslant k_0$, и поэтому ясно, что имеются возможности удовлетворить всем условиям (6), сохраняя неравенства $d_k \geqslant 2L$.

Возьмем произвольную точку $u_* \in U_*$. Пусть $u_{r(k)-1} \leqslant u_* \leqslant u_{r(k)}$ ($k = 1, 2, \ldots$). Поскольку $J(u_{r(k)}) + J(u_{r(k)-1}) \leqslant L(u_{r(k)} - u_{r(k)-1}) + 2J(u_*)$, то из (3) с учетом условия $d_k \geqslant 2L$ имеем

$$R_k(r(k)) \geqslant (d_k - 2L)(u_{r(k)} - u_{r(k)-1}) - 4J(u_*) \geqslant -4J(u_*), \qquad k = 1, 2, \ldots \tag{19}$$

Далее, пусть v_* — какая-либо предельная точка последовательности $\{v_k\}$. Возьмем отрезки $[u_{m(k)-1}, u_{m(k)}]$ ($k = 1, 2, \ldots$) и множество $N(v_*)$, введенные при доказательстве теоремы 1 для точки v_*. Поскольку $J(u_*) \leqslant J(v_*)$, то из (14) при $i = r(k)$ и из (19) имеем $R_k(m(k)) \geqslant R_k(r(k)) \geqslant -4J(u_*) \geqslant -4J(v_*)$ ($k \in N(v_*)$).

Отсюда с учетом равенства (17) имеем

$$\lim_{k\to\infty} R_k(m(k)) = \lim_{k\to\infty} R_k(r(k)) = -4J(u_*) = -4J(v_*), \qquad k \in N(v_*), \tag{20}$$

т. е. $J(u_*) = J(v_*)$. В силу теоремы 1 тогда $\lim\limits_{k\to\infty} J(v_k) = J(v_*) = J(u_*)$, а из теоремы 1.1 получим $\lim\limits_{k\to\infty} \rho(v_k, U_*) = 0$. Наконец, при $\inf\limits_{k\geqslant k_0} d_k > 2L$ из (19), (20) следует $\lim\limits_{k\to\infty, k\in N(v_*)} (u_{r(k)} - u_{r(k)-1}) = 0$, и поэтому в качестве подпоследовательности, сходящейся к u_*, можно взять $\{u_{r(k)}\}$ или $\{u_{r(k)-1}\}$.

Сравним метод (1)—(6) с методом равномерного перебора (7.2) на классе функций $Q(L)$, удовлетворяющих на $[a, b]$ условию (11) с одной и той же постоянной L. Согласно теореме 7.1 для обеспечения точности

$$\min_i J(w_i) - J_* \leqslant \varepsilon \tag{21}$$

на классе $Q(L)$ перебор на $[a, b]$ нужно осуществить по точкам $w_i = a + (i - 1/2)h$ ($i = 1, 2, \ldots$) с шагом $2\varepsilon/L$.

Теорема 4. *Пусть $J(u) \in Q(L)$, а $[\alpha, \beta] \subseteq \{u\colon u \in [a, b],\ J(u) \geqslant J_* + \gamma\}$, где $\gamma > 0$ — фиксированное число. Тогда на $[\alpha, \beta]$ число поисковых точек метода (1)—(6), реализованного при $d_k = 2L$ ($k \geqslant 1$), по крайней мере в $\gamma/(5\varepsilon)$ раз меньше числа поисковых точек метода равномерного перебора (7.2); здесь число $\varepsilon > 0$ взято из (21), $5\varepsilon < \gamma$.*

Доказательство. Пусть $u_* \in U_*$, $u_{r(k)-1} \leqslant u_* \leqslant u_{r(k)}$ ($k = 1, 2, \ldots$). Тогда из (19) следует, что $R_k(r(k)) \geqslant -4J(u_*) = -4J_*$. Далее, для любого отрезка $[u_{i-1}, u_i] \subseteq [\alpha, \beta]$ ($i = i(k)$) справедливо неравенство $J(u_i) + J(u_{i-1}) \geqslant 2J_* + 2\gamma$. Поэтому из (3) при $d_k = 2L$ получим $R_k(i) < -4J_* - 4\gamma + 5L(u_i - u_{i-1})/2$. Отсюда ясно, что если $u_i - u_{i-1} < 8\gamma/(5L)$, то $R_k(r(k)) - R_k(i(k)) \geqslant 4\gamma - 5L(u_i - u_{i-1})/2 > 0$ и на k-м шаге внутри $[u_{i-1}, u_i]$ новая поисковая точка не появится. Следовательно, для появления новой точки в $[u_{i-1}, u_i]$ на k-м шаге необходимо, чтобы $u_i - u_{i-1} \geqslant 8\gamma/(5L)$.

Однако тогда появление новой поисковой точки v_k в $[u_{i-1}, u_i]$ приведет к двум новым отрезкам $[u_{i-1}, v_k]$ и $[v_k, u_i]$, длина которых согласно (9) при $d_k = 2L$, $\mu_k = 2L/L_k \geqslant 2 = \mu$ оценивается снизу величиной $(u_i - u_{i-1})/4 \geqslant 2\gamma/(5L)$. Это значит, что независимо от номера шагов расстояние между точками u_{i-1}, $u_i \in [\alpha, \beta]$ удовлетворяет условию $u_i - u_{i-1} \geqslant 2\gamma/(5L)$, и следовательно, число поисковых точек, попавших на $[\alpha, \beta]$ при применении метода (1)—(6), не может превышать числа $5L(\beta - \alpha)/(2\gamma)$.

Остается заметить, что число точек метода равномерного перебора на $[\alpha, \beta]$, необходимое для обеспечения точности в смысле неравенства (21), оценивается числом $(\beta - \alpha)/h \geqslant (\alpha - \beta)L/(2\varepsilon)$. Сравнивая полученные числа, убеждаемся в справедливости теоремы.

Из теоремы 4 следует, что на отрезках, на которых значения функции $J(u)$ значительно превышают J_*, число поисковых точек метода (1)—(6), вообще говоря, намного меньше числа точек равномерного перебора.

На практике в методе (1)—(6) часто берут $\nu = 1$, $\mu_k = 2$ и итерации продолжают до тех пор, пока величина $u_s - u_{s-1}$ не станет меньше заданного числа (здесь s определяется условиями (4)). Численные эксперименты показывают [279], что метод (1)—(6) позволяет определить глобальный минимум с достаточно высокой надежностью. Замечено, что при увеличении значений параметров ν, μ_k надежность поиска возрастает, но при этом одновременно может увеличиться и количество вычислений значений минимизируемой функции.

Упражнения. 1. Выяснить поведение поисковых точек метода (1)—(6) для функций $J(u) \equiv 1$, $J(u) = u$ на отрезке $[0, 1]$.

2. Исследовать сходимость метода (1)—(6) для функции $J(u) = |u| + |u - 1|$ на отрезках $[-1, 1]$, $[-1, 2]$, $[1, 2]$.

§ 11. Метод парабол

В рассмотренных выше методах ломаных и касательных минимизируемая функция аппроксимировалась кусочно линейными функциями и исходная задача заменялась последовательностью задач минимизации кусочно линейных функций. Однако существуют и другие достаточно простые классы функций, которыми можно аппроксимировать минимизируемую функцию и для которых задача минимизации легко решается. Выбирая, например, в качестве такого аппроксимирующего класса квадратичные функции, график которых представляет собой параболу, мы придем к методу минимизации, который естественно назвать *методом парабол*.

Определение 1. Пусть функция $J(u)$ определена на отрезке $[a, b]$, а τ, h — положительные постоянные. Тройку точек $v - \tau$, v, $v + h$, принадлежащих $[a, b]$, назовем *выпуклой тройкой* для $J(u)$, если $\Delta^-(v) = J(v - \tau) - J(v) \geqslant 0$, $\Delta^+(v) = J(v + h) - J(v) \geqslant 0$, $\Delta^+ + \Delta^- > 0$.

Если тройка точек $v - \tau$, v, $v + h$ выпукла для $J(u)$, то через точки $(v - \tau, J(v - \tau))$, $(v, J(v))$, $(v + h, J(v + h))$ на плоскости (u, J) можно провести параболу $L_2(u) = pu^2 + qu + r$ со старшим коэффициентом $p > 0$. Эта парабола является графиком обычного интерполяционного многочлена второй степени и имеет вид

$$L_2(u) = \left(\frac{\Delta^+}{h} + \frac{\Delta^-}{\tau}\right)\frac{(u - v)(u - v - h)}{\tau + h} + \frac{\Delta^+}{h}(u - v) + J(v). \quad (1)$$

Поскольку $p = \Delta^+/h + \Delta^-/\tau > 0$, то $L_2(u)$ выпукла на $\mathbf{R}$ и достигает своей нижней грани на $\mathbf{R}$ в точке

$$w = v + \frac{h^2\Delta^- - \tau^2\Delta^+}{2(h\Delta^- + \tau\Delta^+)}. \quad (2)$$

Пользуясь условиями $\Delta^+ \geqslant 0$, $\Delta^- \geqslant 0$, из (2) нетрудно получить, что

$$-\tau/2 \leqslant w - v \leqslant h/2. \quad (3)$$

Перейдем к описанию одного из вариантов метода парабол для минимизации функции $J(u)$, унимодальной на отрезке $[a, b]$. Пусть $h > 0$ — некоторый начальный шаг, $2h \leqslant b - a$, а $u_0 \in [a, b]$ — начальная точка. Поиск минимума начинаем с вычисления значений $J(u_0)$, $J(u_0 + h)$. Если окажется, что $J(u_0 + h) \leqslant J(u_0)$, то продолжаем вычислять значения $J(u_0 + h \cdot 2^{i-1})$ $(i = 2, 3, \ldots)$ (рис. 1.8); если же $J(u_0 + h) > J(u_0)$, то меняем направление поиска: переобозначаем $u_0 + h$ через u_0 и далее вычисляем значения $J(u_0 - h \cdot 2^{i-1})$ $(i = 2, 3, \ldots)$ (рис. 1.9). Перед

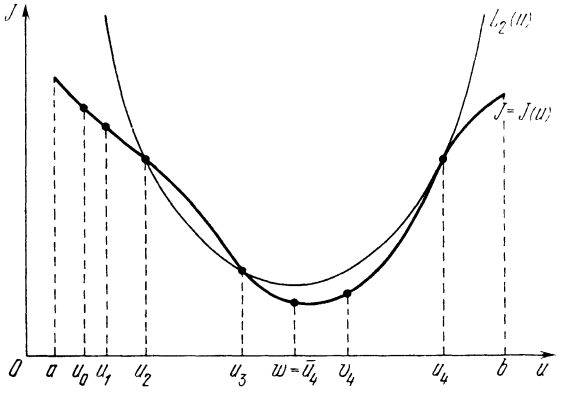

Рис. 1.8

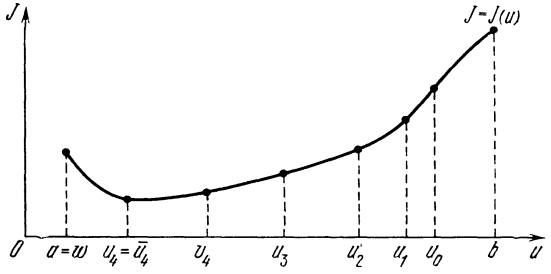

Рис. 1.9

вычислением значения функции в очередной новой точке $u_i = u_0 + h \cdot 2^{i-1}$ (или соответственно $u_i = u_0 - h \cdot 2^{i-1}$) прежде всего выясняем, будет ли $u_i \in [a, b]$. Если $u_i \in [a, b]$, то вычисляем значение $J(u_i)$ и проверяем, образуют ли последние три точки u_{i-2}, u_{i-1}, u_i выпуклую тройку для $J(u)$ или нет. В том случае, когда тройка u_{i-2}, u_{i-1}, u_i невыпукла, переходим к следующей точке u_{i+1} и т. д. Этот процесс закончится отыскиванием такого номера $n \geqslant 1$, что:

1) либо три точки u_{n-2}, u_{n-1}, u_n образуют выпуклую тройку для $J(u)$ — тогда через три точки $(u_i, J(u_i))$ $(i=n-2, n-1, n)$ проводим параболу и находим точку w ее минимума на $\mathbf{R}$; согласно (3) $w \in [u_{n-2}, u_n]$ (см. рис. 1.8, $n=4$);

2) либо ни одна тройка u_{i-2}, u_{i-1}, u_i при $i=2, \ldots, n$ не будет выпуклой, а $u_{n+1} \notin [a, b]$ — тогда полагаем $w=a$ или $w=b$ в зависимости от того, какой из концов отрезка $[a, b]$ ближе всего к точке u_{n+1} (см. рис. 1.9, $n=4$).

Определив указанным образом точку w, вычисляем значение $J(w)$ (если оно еще не было вычислено на предыдущих шагах). Наконец, среди точек w, u_0, u_1, ..., u_n находим точку $\bar{u}_n$ такую, что

$$J(\bar{u}_n) = \min\{J(w), J(u_0), J(u_1), \ldots, J(u_n)\},$$

и за приближение к $J_* = \inf\limits_{u \in [a,b]} J(u)$ берем значение $J(\bar{u}_n)$, а за приближение к множеству U_* точек минимума — точку $\bar{u}_n$. При необходимости для уточнения точки минимума можно взять найденную точку $\bar{u}_n$ за начальную и повторить описанную процедуру поиска с начальным шагом $h/2$, и т. д. Последовательно повторяя этот процесс, можно найти точку, удаленную от множества U_* точек минимума унимодальной функции $J(u)$ на расстояние, не превышающее заданного числа.

Если унимодальная функция хорошо аппроксимируется параболой в некоторой окрестности U_*, то этот метод может оказаться гораздо более эффективным, чем другие методы минимизации. Если функция непрерывна на отрезке $[a, b]$, но не является унимодальной, то применение метода парабол приведет, вообще говоря, лишь к точке локального минимума этой функции.

Существуют различные модификации метода парабол. Иногда после нахождения выпуклой тройки u_{n-2}, u_{n-1}, u_n вычисляют значение функции еще в одной точке $v_n = (u_{n-1} + u_n)/2$ — середине отрезка $[u_{n-1}, u_n]$, затем из получившихся равноотстоящих точек u_{n-2}, u_{n-1}, v_n, u_n исключают либо u_{n-2}, либо u_n, в зависимости от того, какая из них более удалена от точки, соответствующей минимальному значению среди вычисленных, и по оставшейся тройке проводят параболу.

Возможно использование описанного метода парабол и при другом, более общем определении выпуклой тройки. А именно, тройку точек $v-\tau$, v, $v+h \in [a, b]$ можно назвать *выпуклой*, если старший коэффициент $p = \Delta^+/h + \Delta^-/\tau$ параболы (1) положителен, хотя в отличие от определения 1 одно из условий $\Delta^+ \geqslant 0$ или $\Delta^- \geqslant 0$ может быть и нарушено. В этом случае парабола (1) также выпукла на $\mathbf{R}$ и достигает своей нижней грани в точке w, выражаемой той же формулой (2). Однако здесь надо иметь в виду, что w не обязана удовлетворять неравенствам (3), и более того, может оказаться, что $w \notin [a, b]$.

§ 12. Другой метод поиска глобального минимума

1. Описанные выше методы деления отрезка пополам, золотого сечения, парабол и некоторые другие методы приспособлены для поиска глобального минимума унимодальных функций. Если эти методы применить к непрерывным функциям, не являющимся унимодальными на рассматриваемом отрезке, то мы получим, вообще говоря, лишь точку локального минимума. Поэтому такие методы часто называют *локальными методами* минимизации.

Можно предложить следующую схему поиска глобального минимума непрерывной функции $J(u)$ на отрезке $[a, b]$ с помощью локальных методов. А именно, пусть каким-либо локальным методом минимизации найдена точка u_0 локального минимума функции $J(u)$ на $[a, b]$. Если точка u_0 не является точкой глобального минимума, то найдется точка $v_0 \in [a, b]$ такая, что $J(v_0) < J(u_0)$. Допустим, что нам как-то удалось найти хотя бы одну такую точку v_0. Тогда в некоторой окрестности точки v_0 существует новая точка u_1 более глубокого локального минимума со значением $J(u_1) \leqslant \leqslant J(v_0) < J(u_0)$.

Для поиска u_1 можно использовать один из локальных методов минимизации, например метод парабол, беря в качестве начальной точку v_0. Если здесь будет использован метод золотого сечения, то чтобы гарантировать получение точки u_1 локального минимума со значением $J(u_1) \leqslant \leqslant J(v_0)$, на каждом шаге этого метода следует учитывать также и имеющееся вычисленное значение $J(v_0)$; кроме того, сам метод золотого сечения может применяться как на исходном отрезке $[a, b]$, так и на каком-либо другом подходящем отрезке $[\alpha, \beta] \subset [a, b]$, содержащем точку v_0. Если найденная точка u_1 не реализует глобального минимума функции $J(u)$ на $[a, b]$, то можно попытаться найти точку v_1 со значением $J(v_1) < J(u_1)$ и затем, пользуясь одним из локальных методов, перейти к точке u_2 более глубокого минимума со значением $J(u_2) \leqslant J(v_1) < J(u_1) < J(u_0)$, и т. д. В том случае, когда $J(u)$ на $[a, b]$ имеет конечное число локальных минимумов, осуществляя переход от одного локального минимума к другому более глубокому локальному минимуму, за конечное число таких переходов можно отыскать глобальный минимум функции $J(u)$ на $[a, b]$.

Намеченная здесь схема поиска глобального минимума показывает, как в принципе можно использовать локальные методы минимизации для отыскания глобального минимума. Однако эта схема имеет существенный недостаток: в ней отсутствует сколь-нибудь конструктивный способ отыскания точек $v_0, v_1, \ldots$, т. е. точек, в которых функция принимает меньшее значение, чем в найденной точке локального минимума.

2. Опишем один такой способ, следуя работе [90]. Будем предполагать, что функция $J(u)$ непрерывно дифференцируема на отрезке $[a, b]$ и имеет на нем конечное число точек локального минимума. Тогда, очевидно, все локальные минимумы будут строгими. Кроме того, предположим, что в окрестности каждой точки u_i локального минимума имеет место представление

$$J(u) - J(u_i) = (u - u_i)^{n_i} g_i(u), \tag{1}$$

где $g_i(u_i) > 0$ при $a \leqslant u_i < b$ и $(-1)^{n_i} g_i(b) > 0$ при $u_i = b$, $g_i(u)$ непрерывно дифференцируема на $[a, b]$, а n_i — натуральное число. Заметим, что для достаточно гладких функций $J(u)$ представление (1) легко получается из разложения по формуле Тейлора в точке u_i, причем число n здесь равно порядку самой младшей производной функции $J(u)$ в точке u_i, которая отлична от нуля, а условие $J^{(n_i)}(u_i) = n! g_i(u_i) > 0$ при $a \leqslant u_i < b$ и $(-1)^{n_i} J^{(n_i)}(u_i) > 0$ при $u_i = b$ являются необходимыми для того, что-

бы в точке u_i достигался локальный минимум (см. упражнения 2.5—2.7). Заметим также, что при $a < u_i < b$ число n_i обязательно будет четным.

Возьмем одну из точек u_0 локального минимума функции $J(u)$. Тогда $J(u) > J(u_0)$ для всех $u \in O_\alpha(u_0) = \{u\colon u \in [a, b],\ 0 < |u - u_0| < \alpha\}$, если число $\alpha > 0$ достаточно мало. Введем функцию

$$\varphi_m(u, u_0) = (J(u) - J(u_0))/(u - u_0)^{2m}, \tag{2}$$

определенную при всех $u \in [a, b]$, $u \neq u_0$, $m = 1, 2, \ldots$

Напомним, что нам требуется найти хотя бы одну точку $v_0 \in [a, b]$, для которой $J(v_0) < J(u_0)$, что равносильно неравенству $\varphi_m(v_0, u_0) < 0$. Такую точку v_0 можно попытаться искать, решая в свою очередь задачу минимизации функции $\varphi_m(u, u_0)$ на $[a, b]$ с помощью тех же локальных методов, которые упоминались выше. Однако здесь сразу возникает вопрос: чем эта задача минимизации лучше исходной задачи? Ответить на него можно так: во-первых, задача минимизации $\varphi_m(u, u_0)$ играет вспомогательную роль при поиске точки v_0, для которой $\varphi_m(v_0, u_0) < 0$, и сразу же после нахождения такой точки v_0 процесс минимизации $\varphi_m(u, u_0)$ прекращается, во-вторых, функция $\varphi_m(u, u_0)$ обладает рядом интересных свойств, существенно облегчающих поиск точки v_0. Рассмотрим эти свойства функции (2).

1. Перепишем (2) в виде

$$J(u) - J(u_0) = (u - u_0)^{2m}\varphi_m(u, u_0), \quad u \neq u_0, \quad u \in [a, b].$$

Отсюда ясно, что функции $J(u) - J(u_0)$ и $\varphi_m(u, u_0)$ при всех $m = 1, 2, \ldots$ имеют одинаковые знаки при каждом $u \in [a, b]$ $(u \neq u_0)$ и обращаются в нуль в одних и тех же точках, не считая $u = u_0$. Поэтому нетрудно видеть, что точка u_0 будет точкой глобального минимума функции $J(u)$ на $[a, b]$ тогда и только тогда, когда $\varphi_m(u, u_0) \geqslant 0$ при всех $u \in [a, b]$ $(u \neq u_0,\ m = 1, 2, \ldots)$, и следовательно, в точке u_0 не будет достигаться глобальный минимум тогда и только тогда, когда существует точка $v_0 \in [a, b]$, в которой $\varphi_m(v_0, u_0) < 0$ $(m \geqslant 1)$. Очевидно также, что при любом $m = 1, 2, \ldots$ функция $\varphi_m(u, u_0)$ может иметь локальный минимум в точке $w_0 \neq u_0$ со значением $\varphi_m(w_0, u_0) = 0$ тогда и только тогда, когда w_0 является точкой локального минимума функции $J(u)$ со значением $J(w_0) = J(u_0)$.

2. Существуют такие $m \geqslant 1$ и $\alpha > 0$, что функция $\varphi_m(u, u_0)$ не будет иметь локальных минимумов на множестве $O_\alpha(u_0) = \{u\colon u \in [a, b],\ 0 < |u - u_0| < \alpha\}$. В самом деле, эта функция дифференцируема при всех $u \neq u_0$ и ее производная равна

$$\varphi_m'(u, u_0) = [J'(u)(u - u_0) - 2m(J(u) - J(u_0))]/(u - u_0)^{2m+1}. \tag{3}$$

Отсюда с учетом представления (1) имеем

$$\varphi_m'(u, u_0)(u - u_0)^{2m-n_0+1} = (u - u_0)g_0'(u) - (2m - n_0)g_0(u). \tag{4}$$

Если $a \leqslant u_0 < b$ и α достаточно мало, то $g_0(u) \geqslant \mathrm{const} > 0$ при всех $u \in O_\alpha(u_0)$, а тогда, взяв m достаточно большим, можно сделать $\varphi_m'(u, u_0) \times (u - u_0)^{2m-n_0+1} < 0$ при всех $u \in O_\alpha(u_0)$. Учитывая, что при $a < u_0 < b$ число n_0 четно, отсюда получаем, что $\varphi_m(u, u_0)$ в $O_\alpha(u_0)$ строго возрастает слева от u_0 и строго убывает справа от u_0. Случай $a \leqslant u_0 < b$ рассмотрен. Если же $u_0 = b$, то из непрерывности функции $g_0(u)$ и условия $(-1)^{n_0}g_0(b) > 0$ следует существование такого $\alpha > 0$, что $(-1)^{n_0}g_0(u) \geqslant \mathrm{const} > 0$ для всех $u \in O_\alpha(u_0)$. Тогда из (4) при достаточно больших m получим неравенство $\varphi_m'(u, b) > 0$ $(u \in O_\alpha(b))$, означающее, что на $O_\alpha(b)$ функция $\varphi_m(u, b)$ строго возрастает и не может иметь локальных минимумов.

3. Пусть в некоторой точке $v \in (a, b)$ оказалось, что $\varphi_m(v, u_0) > 0$, т. е. $J(v) > J(u_0)$. Тогда найдется столь большой номер $m_0 = m_0(v)$, что функция $\varphi_m(u, u_0)$ не будет достигать локального минимума в точке v при всех $m \geqslant m_0$. В самом деле, возьмем

$$m_0 > |J'(v)|(b-a)/(2(J(v)-J(u_0))).$$

Тогда из формулы (3) при $u = v$ для всех $m \geqslant m_0$ имеем $\varphi'_m(v, u_0) \times \times (v-u_0)^{2m+1} < 0$, т. е. $\varphi'_m(v, u_0) < 0$ при $u_0 < v < b$ и $\varphi'_m(v, u_0) > 0$ при $a < v < u_0$. Это значит, что при $m \geqslant m_0$ функция $\varphi_m(u, u_0)$ не может иметь локального минимума в точке $u = v$.

Опираясь на изложенные свойства функции $\varphi_m(u, u_0)$, можно предложить следующий способ определения точки v_0, для которой $\varphi_m(v_0, u_0) < 0$. Сначала будем рассматривать задачу минимизации $\varphi_m(u, u_0)$ на отрезке $[u_0, b]$ (при $u_0 < b$), затем — на $[a, u_0]$ (при $a < u_0$). Поиск минимума на $[u_0, b]$ начнем с некоторой начальной точки $u_0 + \alpha < b$.

С учетом свойства 2 выберем $\alpha > 0$ столь малым, а m столь большим, чтобы $\varphi'_m(u_0 + \alpha, u_0) < 0$. Тогда непрерывная функция $\varphi_m(u, u_0)$ на отрезке $[u_0 + \alpha, b]$ имеет хотя бы один локальный минимум со значением, меньшим $\varphi_m(u_0 + \alpha, u_0)$. Поиск минимума на этом отрезке будем вести с помощью какого-либо локального метода минимизации. Если в процессе поиска будет найдена точка v_0, для которой $\varphi_m(v_0, u_0) < 0$, то сразу возвращаемся к задаче минимизации исходной функции $J(u)$ в соответствии со схемой, описанной в начале параграфа.

Допустим, что при поиске минимума $\varphi_m(u, u_0)$ на $[u_0 + \alpha, b]$ такая точка v_0 не нашлась, и мы пришли к некоторой точке локального минимума $w_0 (u_0 + \alpha < w_0 < b)$ со значением $\varphi_m(w_0, u_0) \geqslant 0$, причем $\varphi_m(u, u_0) \geqslant 0$ при всех $u \in [u_0 + \alpha, w_0]$.

Здесь имеются две возможности: либо $\varphi_m(w_0, u_0) > 0$, либо $\varphi_m(w_0, u_0) = 0$. Если $\varphi_m(w_0, u_0) > 0$, то, пользуясь свойством 3, увеличиваем m так, чтобы соответствующая функция $\varphi_m(u, u_0)$ в точке w_0 не достигала локального минимума, и продолжаем поиск минимума на отрезке $[w_0, b]$. Заметим, что при минимизации функции $\varphi_m(u, u_0)$, соответствующей новому значению m, нет необходимости возвращаться к отрезку $[u_0 + \alpha, w_0]$, так как неравенство $\varphi_m(u, u_0) \geqslant 0$ $(u \in [u_0 + \alpha, w])$, установленное для какого-то значения m, останется верным при всех $m = 1, 2, \ldots$ и свидетельствует об отсутствии на $[u_0 + \alpha, w]$ искомой точки v_0.

Рассмотрим случай $\varphi_m(w_0, u_0) = 0$. Согласно свойству 1 функция $J(u)$ также будет достигать локального минимума в точке w_0 со значением $J(w_0) = J(u_0)$. Поскольку локальные минимумы функции $J(u)$ строгие, то можем выбрать $\alpha > 0$ столь малым, чтобы $J(u) > J(w_0)$ и, следовательно, $\varphi_m(u, u_0) > 0$ при $w_0 < u \leqslant w_0 + \alpha$. Увеличивая при необходимости m, согласно свойству 3 можно сделать $\varphi'_m(w_0 + \alpha, u_0) < 0$ и затем перейти к минимизации соответствующей функции $\varphi_m(u, u_0)$ на отрезке $[w_0 + \alpha, b]$.

Можно ожидать, что продолжая этот процесс дальше, либо мы найдем искомую точку v_0, для которой $\varphi_m(v_0, u_0) < 0$, либо, конечное число раз увеличивая m и выбирая $\alpha > 0$, доберемся до точки b и выясним, что $\varphi_m(u, u_0) \geqslant 0$ при всех u $(u_0 < u \leqslant b)$. В последнем случае переходим к отрезку $[a, u_0]$ и на нем аналогичным образом ищем точку v_0 со свойством $\varphi_m(v_0, u_0) < 0$. Если и на отрезке $[a, u_0]$ такая точка v_0 не найдется, то это будет означать, что $\varphi_m(u, u_0) \geqslant 0$ при $u \in [a, b]$ $(u \neq u_0)$, т. е. согласно свойству 1 точка u_0 есть точка глобального минимума функции $J(u)$ на $[a, b]$. С другой стороны, если u_0 не является точкой глобального минимума, то в силу того же свойства 1 существует точка v_0, для которой $\varphi_m(v_0, u_0) < 0$, и можно надеяться, что описанным выше способом удастся найти точку и реализовать изложенную в начале параграфа схему поиска глобального минимума функции $J(u)$ на $[a, b]$.

Следует заметить, что для более строгого обоснования предложенного способа поиска точки v_0 надо бы еще доказать, что поиск минимума $\varphi_m(u, u_0)$ на $[u_0, b]$ и $[a, u_0]$ всегда удастся закончить за конечное число изменений величин m, α. Интересно также исследовать, как влияют погрешности, неизбежные при определении точек локального минимума рассматриваемых функций, на весь процесс поиска глобального минимума. Эти вопросы, по-видимому, пока еще должным образом не изучены.

При практическом использовании описанного метода обычно задают некоторое достаточно большое натуральное число M и ограничиваются рассмотрением функций $\varphi_m(u, u_0)$ лишь при $m = 1, \ldots, M$. Численные эксперименты показывают, что удачный выбор M и точность определения точек локального минимума рассматриваемых функций обеспечивают достаточно высокую эффективность этого метода [90].

3. Применим описанный метод поиска глобального минимума к задаче минимизации многочленов $J(u) = p_{2n}(u) = \sum_{i=0}^{2n} a_i u^i$ степени $2n$, где все коэффициенты a_i — действительные числа, старший коэффициент $a_{2n} > 0$. Поскольку $\lim_{|u|\to\infty} p_{2n}(u) = \infty$ и $p_{2n}(u)$ — непрерывная функция, то множество $M(v) = \{u\colon u \in \mathbf{R},\ p_{2n}(u) \leqslant p_{2n}(v)\}$ непусто, ограничено и замкнуто при любом фиксированном v. Кроме того, очевидно $\inf_{u\in\mathbf{R}} p_{2n}(u) = \inf_{u\in M(v)} p_{2n}(u) = J_*$. Поэтому, применяя теорему 1.1 к функции $p_{2n}(u)$ на множестве $M(v)$, получаем, что $J_* > -\infty$ и множество U_* точек минимума $p_{2n}(u)$ на $\mathbf{R}$ непусто. Более того, так как производная $p'_{2n}(u)$, является многочленом степени $2n-1$, то она может иметь не более $2n-1$ действительных корней, и $p_{2n}(u)$ на $\mathbf{R}$ может иметь не более $n-1$ точек локального максимума и не более n точек локального минимума.

Поиск глобального минимума функции $p_{2n}(u)$ на $\mathbf{R}$ можно осуществить следующим образом. Отправляясь от произвольной начальной точки, с помощью какого-либо локального метода минимизации, например, метода парабол, находим точку u_0 локального минимума $p_{2n}(u)$. В этой точке $p'_{2n}(u_0) = 0$ и справедливо представление

$$p_{2n}(u) = p_{2n}(u_0) + p''_{2n}(u_0)(u-u_0)^2/2! + \ldots$$
$$\ldots + p_{2n}^{(2n-1)}(u_0)(u-u_0)^{2n-1}/(2n-1)! + a_{2n}(u-u_0)^{2n}.$$

По аналогии с (2) введем функцию

$$\varphi_2(u, u_0) = (p_{2n}(u) - p_{2n}(u_0))/(u-u_0)^2 = p_{2n-2}(u, u_0).$$

Из предыдущего представления следует, что $p_{2n-2}(u, u_0)$ — многочлен степени $2n-2$ со старшим коэффициентом $a_{2n} > 0$. Поскольку $p_{2n}(u) - p_{2n}(u_0) = (u-u_0)^2 p_{2n-2}(u, u_0)$ при всех $u \in \mathbf{R}$, то ясно, что точка u_0 будет точкой глобального минимума $p_{2n}(u)$ на $\mathbf{R}$ тогда и только тогда, когда $\inf_{u\in\mathbf{R}} p_{2n-2}(u, u_0) \geqslant 0$. Это означает, что исходная задача минимизации многочлена степени $2n$ свелась к задаче минимизации многочлена $p_{2n-2}(u, u_0)$ меньшей степени $2n-2$, которую в свою очередь можно аналогично решать последовательным сведением к задаче минимизации на $\mathbf{R}$ многочленов меньших степеней $2n-4$, $2n-6$, ..., 2 с одним и тем же старшим коэффициентом $a_{2n} > 0$. Остается заметить, что для многочлена второй степени $p_2(u) = b_2u^2 + b_1u + b_0$ $(b_2 = a_{2n} > 0)$ точка минимума $\mathbf{R}$ находится по известной формуле $u_* = -b_1/(2b_2)$.

Если при поиске минимума $p_{2n-2}(u, u_0)$ на $\mathbf{R}$ будет найдена точка v_0, для которой $p_{2n-2}(v_0, u_0) < 0$, то сразу же возвращаемся к исходной задаче минимизации $p_{2n}(u)$: отправляясь от начальной точки v_0, каким-либо ло-

кальным методом находим следующую точку u_1 локального минимума $p_{2n}(u)$ со значением $p_{2n}(u_1) \leqslant p_{2n}(v_0) < p_{2n}(u_0)$ и затем с новой точкой u_1 поступаем так же, как с предыдущей точкой u_0, и т. д. Этот процесс перехода от одной точки u_{i-1} локального минимума к следующей точке u_i с более глубоким локальным минимумом закончится не более чем через n шагов обнаружением того, что $\inf\limits_{u \in \mathbf{R}} p_{2n-2}(u, u_i) \geqslant 0$.

§ 13. О методе стохастической аппроксимации

Выше предполагалось, что значение минимизируемой функции или ее производной в каждой точке вычисляется точно. Между тем из-за погрешностей применяемого здесь метода и ошибок округления значения даже простейших элементарных функций могут быть вычислены, вообще говоря, лишь приближенно. Поэтому при поиске минимума вместо точных значений функции $J(u)$ мы будем иметь дело с приближенными значениями $z(u)$ с некоторой погрешностью $|J(u)-z(u)| \leqslant \varepsilon$. В этом случае уверенно можно различать значения функции в двух точках и выяснить, какое из них меньше, только тогда, когда разность этих значений больше 2ε. Понятно, что это обстоятельство должно быть учтено при использовании описанных выше методов минимизации.

Весьма усложняется решение задачи минимизации в тех случаях, когда на значения функции в каждой точке накладываются случайные ошибки или, как говорят, помехи. Такая ситуация, например, имеет место, если значения функции получаются в результате измерений какой-либо физической величины. В том случае, когда помехи являются случайной величиной и обладают определенными вероятностными характеристиками, для поиска минимума целесообразно использовать метод стохастической аппроксимации.

Опишем один из вариантов этого метода. Будем предполагать, что значения функции $J(u)$ могут быть измерены в любой фиксированной точке $u \in \mathbf{R}$, причем результаты измерений не содержат систематических ошибок. Тогда для поиска минимума функции $J(u)$ на $\mathbf{R}$ может быть использован следующий итерационный метод:

$$u_{n+1} = u_n - a_n(z(u_n + c_n) - z(u_n - c_n))/c_n \quad n = 1, 2, \ldots, \tag{1}$$

где последовательности $\{a_n\}$, $\{c_n\}$ заданы и удовлетворяют условиям

$$a_n > 0, \quad c_n > 0, \quad n = 1, 2, \ldots, \quad \lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = 0, \tag{2}$$

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} \left(\frac{a_n}{c_n}\right)^2 < \infty.$$

Например, здесь можно взять $a_n = 1/n$, $c_n = 1/n^{1/4}$ $(n = 1, 2, \ldots)$.

Следует заметить, что в тех случаях, когда слева от точки минимума график функции имеет крутой спуск, справа — крутой подъем, а на остальных участках функция $J(u)$ изменяется медленно, сходимость метода (1) может ухудшаться. В самом деле, на пологих участках разность $z(u_n+c_n)-z(u_n-c_n)$ может стать очень малой, и тогда шаги поиска $|u_{n+1}-u_n|$ будут малыми, а на крутых участках, наоборот, шаги поиска могут стать очень большими. В результате значительная часть времени на поиск может быть затрачена на чрезмерно медленные спуски на пологих участках и последующие большие скачки через точку минимума с попаданием на другой пологий участок. В таких ситуациях может оказаться полезным следующий вариант метода стохастической аппроксимации:

$$u_{n+1}=u_n-a_n\operatorname{sign}(z(u_n+c_n)-z(u_n-c_n)), \qquad n=1,\ 2,\ \ldots, \quad (3)$$

где последовательности $\{a_n\}$, $\{c_n\}$ по-прежнему удовлетворяют условиям (2), $\operatorname{sign} a$ — знак числа a, т. е. $\operatorname{sign} a=1$ при $a>0$, $\operatorname{sign} a=-1$ при $a<0$, $\operatorname{sign} 0=0$. Сходимость метода (3) можно ускорить, если длину шага a_n менять лишь при изменении знака $z(u_n+c_n)-z(u_n-c_n)$, сохраняя a_n постоянным в остальных случаях.

При некоторых предположениях относительно функции $J(u)$ и вероятностных характеристик случайной величины $z(u)$ можно доказать сходимость по вероятности последовательности $\{u_n\}$, определяемой формулами (1) или (3), к точке глобального минимума $J(u)$ на $\mathbf{R}$. Различные варианты метода стохастической аппроксимации, строгое обоснование этого метода, различные приложения можно найти в [180].

Глава 2

ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ О ЗАДАЧАХ НА ЭКСТРЕМУМ

В этой главе собраны основные факты о задачах на экстремум функций конечного числа переменных, обычно излагаемые в учебниках по математическому анализу [10, 160, 165, 233], а также некоторые другие вспомогательные формулы и оценки, необходимые для дальнейшего изложения.

§ 1. Постановка задачи минимизации. Теорема Вейерштрасса

1. Сначала введем обозначения, напомним некоторые определения из линейной алгебры и математического анализа. Через $\mathbf{R}^n$ будем обозначать n-мерное вещественное линейное пространство, состоящее из вектор-столбцов

$$u = \begin{bmatrix} u^1 \\ \cdots \\ u^n \end{bmatrix}, \quad v = \begin{bmatrix} v^1 \\ \cdots \\ v^n \end{bmatrix}, \quad w = \begin{bmatrix} w^1 \\ \cdots \\ w^n \end{bmatrix}$$

с действительными координатами u^i, v^i, w^i, ... $(i = 1, \ldots, n)$; сумма $u + v$ двух вектор-столбцов и произведение αu вектор-столбца u на действительное число α в $\mathbf{R}^n$ определяется так:

$$u + v = \begin{bmatrix} u^1 + v^1 \\ \cdots\cdots \\ u^n + v^n \end{bmatrix}, \quad \alpha u = \begin{bmatrix} \alpha u^1 \\ \cdots \\ \alpha u^n \end{bmatrix};$$

вектор-столбец

$$0 = \begin{bmatrix} 0 \\ \cdots \\ 0 \end{bmatrix}$$

называется *нулевым*. Вектор-строку, полученную транспонированием вектор-столбца u, обозначим через $u^T = (u^1, \ldots, u^n)$. Там, где не могут возникнуть недоразумения, вектор-столбец или вектор-строку из $\mathbf{R}^n$ для краткости мы часто будем называть просто вектором или точкой, а знак транспонирования «T» будем опускать.

Если в $\mathbf{R}^n$ ввести скалярное произведение двух векторов

$$\langle u, v\rangle = \sum_{i=1}^{n} u^i v^i, \quad u, v \in \mathbf{R}^n,$$

то $\mathbf{R}^n$ превращается в n-мерное евклидово пространство, которое будем обозначать через E^n. Длина вектора или норма вектора в E^n определяется так:

$$|u| = \langle u, u\rangle^{1/2} = \left(\sum_{i=1}^{n} |u^i|^2\right)^{1/2}.$$

Величину

$$\rho(u, v) = |u - v| = \left(\sum_{i=1}^{n} |u^i - v^i|^2\right)^{1/2}$$

называют евклидовым расстоянием между точками $u, v \in E^n$. Для любых точек $u, v, w \in E^n$ справедливо неравенство

$$|u - v| \leqslant |u - w| + |w - v|,$$

называемое неравенством треугольника. Когда важно подчеркнуть, что скалярное произведение, норма, расстояние взяты именно в E^n, мы будем писать $\langle u, v\rangle_{E^n}$, $|u|_{E^n}$, $|u - v|_{E^n}$.

2. Перейдем к постановке задачи минимизации. Пусть U — некоторое непустое множество из E^n, а $J(u)$ — функция, определенная на этом множестве. Всюду ниже, если не оговорено противное, мы будем рассматривать лишь функции, принимающие во всех точках $u \in U$ конечные вещественные значения. Определения таких понятий, как точка минимума и максимума, наименьшее и наибольшее значение, ограниченность снизу и сверху, нижняя и верхняя грань функции $J(u)$ на множестве U, минимизирующая и максимизирующая последовательность, точка локального и строгого локального минимума и максимума функции, сходимость последовательности к заданному множеству в пространстве E^n получаются из определений 1.1.1 — 1.1.6, 1.1.8 — 1.1.10, нужно лишь под u понимать точку $u = (u^1, \ldots, u^n)$ из E^n, под $|u|$ — норму u в E^n. Поэтому здесь мы не будем воспроизводить определения перечисленных понятий. Примеры 1.1.1 — 1.1.5 могут служить иллюстрацией к этим понятиям и в E^n, так как функция одной переменной является частным случаем функции n переменных.

Нижнюю грань функции $J(u)$ на множестве U по-прежнему будем обозначать через

$$\inf_U J(u) = J_*,$$

а множество точек минимума $J(u)$ на U — через

$$U_* = \{u\colon\ u \in U, \quad J(u) = J_*\}.$$

Для обозначения задачи минимизации функции $J(u)$ на множестве U часто будем пользоваться следующей краткой стандартной записью:

$$J(u) \to \inf; \qquad u \in U.$$

Как и в § 1.1, будем различать задачи минимизации двух типов. В задачах первого типа ищется точное или приближенное значение величины J_*, и здесь неважно, будет ли множество U_* пустым или непустым. В задачах второго типа наряду с величиной J_* ищется точка $u \in U$, которая достаточно близка к множеству U_* или даже принадлежит U_*, — здесь естественно требовать, чтобы $J_* > -\infty$, $U_* \neq \varnothing$.

Для приближенного решения задач обоих типов на практике обычно строят какую-либо минимизирующую последовательность $\{u_k\}$:

$$u_k \in U, \qquad k = 1, 2, \ldots, \qquad \lim_{k\to\infty} J(u_k) = J_*$$

(при $U_* \neq \varnothing$ возможно, например, $u_k = u_* \in U_*$, $k = 1, 2, \ldots$). Тогда, как нетрудно видеть, в качестве приближения для J_* можно взять величину $J(u_k)$ при достаточно большом k. В том случае, если $\{u_k\}$ сходится к множеству U_*, т. е. $\rho(u_k, U_*) = \inf\limits_{U_*} |u_k - u| \to 0$ при $k \to \infty$, точку u_k и соответствующее значение функции $J(u_k)$ при достаточно большом k можно принять за приближенное решение задачи второго типа. Однако, как мы видели в примере 1.1.5, условие $\lim\limits_{k\to\infty} \rho(u_k, U_*) = 0$ имеет место не всегда. Поэтому в задачах второго типа построение минимизирующих последовательностей, сходящихся к U_*, в общем случае требует привлечения специальных методов.

В то же время имеются классы задач второго типа, у которых любая минимизирующая последовательность $\{u_k\}$ сходится к U_*. Эти классы задач хороши тем, что для их приближенного решения достаточно построить произвольную минимизирующую последовательность $\{u_k\}$ и затем пару $(u_k, J(u_k))$ при достаточно большом k принять за приближенное решение. Один такой класс задач будет описан ниже в теореме 1. Эту теорему будем называть теоремой Вейерштрасса, поскольку она является некоторым обобщением хорошо известной из учебников по математическому анализу теоремы Вейерштрасса о достижении нижней грани непрерывной функции на замкнутом ограниченном множестве из E^n [10, 160, 165, 233].

3. Для формулировки теоремы Вейерштрасса нам понадобятся понятия компактного множества, полунепрерывности снизу функции и некоторые другие понятия из математического анализа. Кратко напомним их определения.

Пусть $\{u_k\} = (u_1, u_2, \ldots)$ — некоторая последовательность, $\{u_k\} \in E^n$, т. е. $u_k \in E^n$ $(k = 1, 2, \ldots)$. Точка v называется предельной точкой последовательности $\{u_k\}$, если существует подпоследовательность $\{u_{k_m}\}$, сходящаяся к v. Последовательность $\{u_k\}$ называется *ограниченной*, если существует постоянная $M \geqslant 0$ такая, что $|u_k| \leqslant M$ для всех $k = 1, 2, \ldots$

Множество U из E^n называется *ограниченным*, если существует постоянная $R \geqslant 0$ такая, что $|u| \leqslant R$ для всех $u \in U$. Множество $O(v, \varepsilon) = \{u\colon u \in E^n, |u - v| < \varepsilon\}$, представляющее собой открытый шар с центром в точке v и радиусом $\varepsilon > 0$, называется *ε-окрестностью* точки v. Точка $v \in E^n$ называется *предельной точкой* множества $U \subset E^n$, если любая ее ε-окрестность содержит точки из U, отличные от v. Нетрудно видеть, что для любой предельной точки v множества U существует последовательность $\{u_k\} \in U$ $(u_k \neq v)$, сходящаяся к v,— для построения такой последовательности достаточно при каждом $k = 1, 2, \ldots$ взять точку $u_k \in O(v, 1/k)$ $(u_k \neq v)$. Верно и обратное: если в U существует последовательность $\{u_k\}$ $(u_k \neq v)$, сходящаяся к точке v, то v — предельная точка множества U. Множество U из E^n называется *замкнутым*, если оно содержит все свои предельные точки.

Определение 1. Множество U из E^n называется *компактным*, если любая последовательность $\{u_k\} \in U$ имеет хотя бы одну предельную точку v, причем $v \in U$.

Согласно теореме Больцано — Вейерштрасса всякая ограниченная последовательность имеет хотя бы одну предельную точку. Пользуясь этой теоремой, нетрудно доказать, что в E^n компактными являются все замкнутые ограниченные множества и только они.

Числовая последовательность $\{x_k\} = (x_1, x_2, \ldots)$ называется *ограниченной снизу* [*сверху*], если существует число A такое, что $x_k \geqslant A$ $[x_k \leqslant A]$ при всех $k = 1, 2, \ldots$ Если $\{x_k\}$ не ограничена снизу [сверху], то существует подпоследовательность $\{x_{k_m}\}$ такая, что $\lim\limits_{m\to\infty} x_{k_m} = -\infty$ $\left[\lim\limits_{m\to\infty} x_{k_m} = \infty\right]$.

Определение 2. Число a называется *нижним* [*верхним*] пределом ограниченной снизу [сверху] числовой последовательности $\{x_k\}$ и обозначается через $\varliminf\limits_{k\to\infty} x_k = a$ $\left[\varlimsup\limits_{k\to\infty} x_k = a\right]$, если: 1) существует хотя бы одна подпоследовательность $\{x_{k_m}\}$, сходящаяся к a; 2) все предельные точки $\{x_k\}$ не меньше [не больше] числа a, т. е. число a является наименьшей [наибольшей] предельной точкой последовательности $\{x_k\}$. Иначе говоря, a — нижний [верхний] предел $\{x_k\}$, если для любого $\varepsilon > 0$: 1) существует номер N такой, что $x_k \geqslant a - \varepsilon$ $[x_k \leqslant a + \varepsilon]$ для всех $k \geqslant N$; 2) для любого номера m найдется номер $k_m > m$ такой,

что $x_{k_m} \leqslant a + \varepsilon \left[x_{k_m} \geqslant a - \varepsilon\right]$. В том случае, когда $\{x_k\}$ не ограничена снизу [сверху], то по определению принимают $\varliminf\limits_{k\to\infty} x_k = -\infty$ $\left[\varlimsup\limits_{k\to\infty} x_k = \infty\right]$; если $\lim\limits_{k\to\infty} x_k = -\infty$, то полагают $\varlimsup\limits_{k\to\infty} x_k = -\infty$; если $\lim\limits_{k\to\infty} x_k = \infty$, то $\varliminf\limits_{k\to\infty} x_k = \infty$.

Например, если $x_k = (-1)^k$ $(k = 1, 2, \ldots)$, то $\varliminf\limits_{k\to\infty} x_k = -1$, $\varlimsup\limits_{k\to\infty} x_k = 1$; если $x_k = (-1)^k k$ $(k = 1, 2, \ldots)$, то $\varliminf\limits_{k\to\infty} x_k = -\infty$, $\varlimsup\limits_{k\to\infty} x_k = \infty$; если $x_k = \left[1 + (-1)^k\right]k$ $(k=1, 2, \ldots)$, то $\varliminf\limits_{k\to\infty} x_k = 0$, $\varlimsup\limits_{k\to\infty} x_k = \infty$; если $x_k = k^{-1}$ $(k = 1, 2, \ldots)$, то $\varliminf\limits_{k\to\infty} x_k = \varlimsup\limits_{k\to\infty} x_k = 0$.

Для того чтобы числовая последовательность $\{x_k\}$ имела предел, необходимо и достаточно, чтобы $\varliminf\limits_{k\to\infty} x_k = \varlimsup\limits_{k\to\infty} x_k = a$; тогда $\lim\limits_{k\to\infty} x_k = a$.

Определение 3. Пусть функция $J(u)$ определена на множестве $U \subseteq E^n$. Говорят, что функция $J(u)$ *полунепрерывна снизу* [*сверху*] в точке $u \in U$, если для любой последовательности $\{u_k\} \in U$, сходящейся к точке u, имеет место соотношение $\varliminf\limits_{k\to\infty} J(u_k) \geqslant J(u) \left[\varlimsup\limits_{k\to\infty} J(u_k) \leqslant J(u)\right]$. Функцию $J(u)$ называют *полунепрерывной снизу* [*сверху*] на множестве U, если она полунепрерывна снизу [сверху] в каждой точке этого множества.

Предлагаем читателю доказать, что функция $J(u)$ полунепрерывна снизу [сверху] в точке $v \in U$ тогда и только тогда, если для любого $\varepsilon > 0$ существует $\delta > 0$ такое, что для всех $u \in \{u\colon u \in U,\ |u - v| < \delta\}$ справедливо неравенство $J(u) \geqslant J(v) - \varepsilon$ $[J(u) \leqslant J(v) + \varepsilon]$. Нетрудно убедиться, что функция непрерывна в точке v тогда и только тогда, когда она в этой точке полунепрерывна и снизу, и сверху.

Пример 1. Пусть $U = \{u\colon u \in E^n,\ |u| \leqslant 1\}$ — n-мерный единичный шар; пусть $J(u) = |u|$ при $0 < |u| \leqslant 1$ и $J(0) = a$. Тогда при $a \leqslant 0$ функция $J(u)$ будет полунепрерывна снизу на U; при $a \geqslant 0$ — полунепрерывна сверху на U; при $a = 0$ — непрерывна на U.

Пример 2. Пусть $U = \{u\colon u \in E^1,\ -1 \leqslant u \leqslant 1\}$; $J(u) = u$ при $0 < u \leqslant 1$; $J(u) = 1 - u$ $(-1 \leqslant u < 0)$; $J(0) = a$. Нетрудно видеть, что при $a \leqslant 0$ эта функция полунепрерывна снизу на U; при $a \geqslant 1$ — полунепрерывна сверху на U, а при $0 < a < 1$ в точке $u = 0$ она не будет полунепрерывной ни снизу, ни сверху.

Пример 3. Пусть $u = (x, y) \in E^2$; $J(u) = x^2 + y^2$ при $x > 0$, $y \geqslant 0$; $J(u) = 0$ при $x \leqslant 0$, $y \geqslant 0$; $J(u) = 1$ при $x \geqslant 0$, $y < 0$;

$J(u)=-1$ при $x<0$, $y<0$. Нетрудно показать, что эта функция на множестве $U_1=\{(x, y): x>0, y\geqslant 0\}$ непрерывна; на $U_2=\{(x, y): y\geqslant 0\}$ полунепрерывна снизу; на $U_3=\{(x, y): x\leqslant 0\}$ полунепрерывна сверху; на $U_4=\{(x, y): x\geqslant 0\}$ в некоторых точках полунепрерывна снизу, в некоторых — сверху.

Установим связь между свойством полунепрерывности снизу функции и замкнутостью множеств

$$M(c)=\{u: u\in U, J(u)\leqslant c\}, \qquad c=\text{const},$$

называемых множествами Лебега функции $J(u)$ на множестве U.

Лемма 1. *Пусть U — замкнутое множество из E^n. Тогда для того, чтобы функция $J(u)$ была полунепрерывна снизу на U, необходимо и достаточно, чтобы множество Лебега $M(c)$ было замкнутым при всех c (пустое множество считается замкнутым по определению). В частности, если $J(u)$ полунепрерывна снизу на U, то множество U_* точек минимума $J(u)$ на U замкнуто.*

Доказательство. Необходимость. Пусть $J(u)$ полунепрерывна снизу на U. Возьмем произвольное число c. Пусть $M(c)\neq\varnothing$. Возьмем какую-либо предельную точку w множества $M(c)$. Тогда существует последовательность $\{u_k\}\in M(c)$, сходящаяся к w. В силу замкнутости U точка $w\in U$. Из того, что $J(u_k)\leqslant c$ $(k=1, 2, \ldots)$, с учетом полунепрерывности снизу $J(u)$ в точке w имеем $J(w)\leqslant \varliminf J(u_k)\leqslant c$, т. е. $w\in M(c)$. Замкнутость $M(c)$ доказана. В частности, множество $U_*=\left\{u: u\in U, J(u)\leqslant J_*=\inf\limits_U J(u)\right\}$ замкнуто.

Достаточность. Пусть для некоторой функции $J(u)$ множество $M(c)$ замкнуто при любом c. Возьмем произвольные $\varepsilon>0$, $u\in U$ и последовательность $\{u_k\}\in U$, сходящуюся к точке u. Пусть $\varliminf\limits_{k\to\infty} J(u_k)=a=\lim\limits_{m\to\infty} J(u_{k_m})$. Тогда $J(u_{k_m})\leqslant a+\varepsilon$, т. е. $u_{k_m}\in M(a+\varepsilon)$, для всех достаточно больших номеров k_m. Но $M(a+\varepsilon)$ замкнуто по условию, а точка u является пределом для $\{u_{k_m}\}$. Следовательно, $u\in M(a+\varepsilon)$, т. е. $J(u)\leqslant a+\varepsilon$. В силу произвола $\varepsilon>0$ отсюда имеем $J(u)\leqslant a=\varliminf\limits_{k\to\infty} J(u_k)$.

Установим одно интересное свойство расстояния от точки до множества.

Лемма 2. *Пусть U — произвольное непустое множество из E^n. Тогда расстояние $\rho(u, U)=\inf\limits_{w\in U}\rho(u, w)$ от точки u до множества U как функция переменной u непрерывна на E^n и, более того, удовлетворяет условию*

$$|\rho(u, U)-\rho(v, U)|\leqslant\rho(u, v) \quad \forall u, v\in E^n.$$

Доказательство. Прежде всего из $\rho(u, w)=|u-w|\geqslant 0$ и $\rho(u, U)\leqslant|u-w|$ $(w\in U)$ следует, что функция $\rho(u, U)$

неотрицательна и конечна во всех точках $u \in E^n$. Возьмем произвольное число $\varepsilon > 0$. По определению нижней грани (см. определение 1.1.3) для любых u, $v \in E^n$ найдутся точки u_ε, $v_\varepsilon \in U$ такие, что

$$\rho(u, U) \leqslant \rho(u, u_\varepsilon) \leqslant \rho(u, U) + \varepsilon, \quad \rho(v, U) \leqslant \rho(v, v_\varepsilon) \leqslant \rho(v, U) + \varepsilon.$$

Поскольку $\rho(u, U) \leqslant \rho(u, v_\varepsilon)$, то с помощью неравенства треугольника $\rho(u, v_\varepsilon) \leqslant \rho(u, v) + \rho(v, v_\varepsilon)$ имеем $\rho(u, U) - \rho(v, U) \leqslant \leqslant \rho(u, v_\varepsilon) - \rho(v, v_\varepsilon) + \varepsilon \leqslant \rho(u, v) + \varepsilon$. Аналогично получается неравенство $\rho(u, U) - \rho(v, U) \geqslant \rho(u, u_\varepsilon) - \varepsilon - \rho(v, u_\varepsilon) \geqslant -\rho(u, v) - \varepsilon$. Объединяя два последних неравенства, имеем $|\rho(u, U) - \rho(v, U)| \leqslant \rho(u, v) + \varepsilon$. Отсюда при $\varepsilon \to +0$ получим требуемое неравенство.

4. Перейдем к формулировке теоремы Вейерштрасса.

Теорема 1. *Пусть U — компактное множество, а функция $J(u)$ определена, конечна и полунепрерывна снизу на U. Тогда $J_* = \inf\limits_U J(u) > -\infty$, множество $U_* = \{u\colon u \in U, J(u) = J_*\}$ непусто, компактно и любая минимизирующая последовательность сходится к U_*.*

Доказательство. Возьмем произвольную минимизирующую последовательность $\{u_k\}$: $u_k \in U$ $(k = 1, 2, \ldots, \lim\limits_{k\to\infty} J(u_k) = J_*)$. Существование хотя бы одной такой последовательности следует из определения 1.1.3 нижней грани функции. Так как U — компактное множество, то $\{u_k\}$ имеет хотя бы одну предельную точку и все ее предельные точки принадлежат U. Возьмем любую предельную точку u_* этой последовательности. Тогда существует подпоследовательность $\{u_{k_m}\}$, сходящаяся к точке u_*. Пользуясь свойством нижней грани J_* и полунепрерывностью функции $J(u)$ в точке u_*, имеем

$$J_* \leqslant J(u_*) \leqslant \varliminf_{m\to\infty} J(u_{k_m}) = \lim_{k\to\infty} J(u_k) = J_*,$$

т. е. $J(u_*) = J_*$. Отсюда следует, что $J_* > -\infty$, $U_* \neq \varnothing$. Более того, показано, что любая предельная точка любой минимизирующей последовательности принадлежит U_*.

Покажем, что U_* компактно. Возьмем произвольную последовательность $\{v_k\} \in U_*$. Так как $\{v_k\} \in U$ — компактное множество, то существует подпоследовательность $\{v_{k_m}\}$, сходящаяся к некоторой точке $v_* \in U$. Но $\{v_k\}$ — минимизирующая последовательность, так как $J(v_k) = J_*$ $(k = 1, 2, \ldots)$. По вышедоказанному тогда $v_* \in U_*$. Компактность U_* установлена.

Покажем, что любая минимизирующая последовательность $\{u_k\}$ сходится к U_*. Так как $\rho(u_k, U_*) = \inf\limits_{u\in U_*} \rho(u_k, u) \geqslant 0$ $(k = 1, 2, \ldots)$, то ясно, что $\varliminf\limits_{k\to\infty} \rho(u_k, U_*) \geqslant 0$. Пусть $\varlimsup\limits_{k\to\infty} \rho(u_k, U_*) =$

$= \lim\limits_{m\to\infty} \rho(u_{k_m}, U_*) = a \leqslant \infty$. В силу компактности U из $\{u_{k_m}\}$ можно выбрать подпоследовательность, сходящуюся к некоторой точке u_*. Не умаляя общности, можем считать, что сама последовательность $\{u_{k_m}\}$ сходится к u_*. Согласно лемме 2 функция $\rho(u, U_*)$ непрерывна по переменной u, поэтому $\lim\limits_{m\to\infty} \rho(u_{k_m}, U_*) = \rho(u_*, U_*) = a$. Однако по доказанному $u_* \in U_*$. Тогда $a = \rho(u_*, U_*) = 0$. Это значит, что $\varliminf\limits_{k\to\infty} \rho(u_k, U_*) = \varliminf\limits_{k\to\infty} \rho(u_k, U_*) = 0$. Следовательно, предел $\lim\limits_{k\to\infty} \rho(u_k, U_*)$ существует и равен нулю. Теорема 1 доказана.

Предлагаем читателю рассмотреть функции и множества из примеров 1—3 (см. также примеры 1.1.1 — 1.1.5) и проверить, в каких случаях условия теоремы 1 выполнены и, следовательно, нижняя грань достигается, в каких случаях она не достигается и в каких случаях нижняя грань достигается несмотря на то, что условия теоремы 1 нарушены.

5. Заметим, что в теореме 1 условие компактности множества U является довольно жестким. Например, такие часто встречающиеся на практике множества, как $U = E^n$ — все пространство или $U = \{u\colon\ u^1 \geqslant 0,\ \ldots,\ u^n \geqslant 0\}$ — неотрицательный ортант, не являются компактными. Приведем две теоремы, в которых компактность множества U не предполагается, но зато функция, кроме полунепрерывности снизу, удовлетворяет некоторым дополнительным требованиям.

Теорема 2. *Пусть U — непустое замкнутое множество из E^n, функция $J(u)$ конечна, полунепрерывна снизу на U и для некоторой фиксированной точки $v \in U$ множество Лебега*

$$M(v) = \{u\colon\ u \in U,\ J(u) \leqslant J(v)\}$$

ограничено. Тогда $J_* > -\infty$, множество U_* непусто, компактно и любая минимизирующая последовательность $\{u_k\}$, принадлежащая $M(v)$, сходится к U_*.

Доказательство. По определению множества $M(v)$ имеем: $J(u) > J(v)$ при всех $u \in U \backslash M(v)$ и $J(u) \leqslant J(v)$ при всех $u \in M(v)$. Это значит, что на $U \backslash M(v)$ функция $J(u)$ не может достигать своей нижней грани на U и для доказательства теоремы достаточно рассмотреть $J(u)$ на множестве $M(v)$.

Замкнутость множества $M(v)$ вытекает из леммы 1. Из ограниченности и замкнутости $M(v)$ следует его компактность. Применяя теорему 1 к функции $J(u)$ на $M(v)$, получим все утверждения теоремы 2. Попутно установили, что $U_* \subseteq M(v)$.

Заметим, что в теореме 2 утверждается сходимость к U_* только тех минимизирующих последовательностей u_k, которые принадлежат $M(v)$. Если $J(v) > J_*$, то условие $\{u_k\} \in M(v)$ можно не оговаривать, так как в этом случае для любой миними-

зирующей последовательности $\{u_k\}$ найдется номер k_0 такой, что $J(u_k) < J(v)$ для всех $k \geqslant k_0$, т. е. $u_k \in M(v)$ при $k \geqslant k_0$. Если же $J(v) = J_*$, то $U_* = M(v)$ и, как видно из примера 1.1.5, в этом случае могут существовать минимизирующие последовательности, которые не принадлежат $M(v)$ и не сходятся к U_*.

Теорема 3. *Пусть U — непустое замкнутое множество из E^n, функция $J(u)$ конечна, полунепрерывна снизу на U и для любой последовательности $\{u_k\} \in U$, $\lim\limits_{k\to\infty} |u_k| = \infty$ (если такие u_k существуют) имеет место соотношение*

$$\lim_{k\to\infty} J(u_k) = \infty.$$

Тогда $J_* > -\infty$, множество U_* непусто, компактно и любая минимизирующая последовательность $\{u_k\}$ сходится к U_*.

Доказательство. Если множество U ограничено, то все утверждения теоремы следуют из теоремы 1. Поэтому пусть U не ограничено, т. е. существует хотя бы одна последовательность $\{v_k\} \in U$ такая, что $\lim\limits_{k\to\infty} |v_k| = \infty$. Тогда согласно условию теоремы $\lim\limits_{k\to\infty} J(v_k) = \infty$. Возьмем какую-либо точку $v \in U$ такую, что $J(v) > J_*$ (например, можно принять $v = v_k$ при достаточно большом k), и рассмотрим множество Лебега $M(v) = \{u\colon u \in U, J(u) \leqslant J(v)\}$. Покажем, что $M(v)$ ограничено. Допустим противное: пусть существует последовательность $\{w_k\} \in M(v)$ такая, что $\lim\limits_{k\to\infty} |w_k| = \infty$. Тогда $\lim\limits_{k\to\infty} J(w_k) = \infty$, что противоречит неравенству $J(w_k) \leqslant J(v) < \infty$, вытекающему из включения $w_k \in M(v)$ $(k = 1, 2, \ldots)$. Таким образом, множество $M(v)$ ограничено. Отсюда и из теоремы 2 следуют все утверждения теоремы 3.

Следствие 1. *Пусть U — непустое замкнутое множество из E^n. Тогда для любой точки $u \in E^n$ найдется точка $v = v(u) \in U$ такая, что $\rho(u, U) = \inf\limits_{u\in U} |u - w| = |u - v(u)|$, т. е. $v(u)$ — ближайшая к u точка из U.*

Доказательство. Пусть u — произвольная точка из E^n. Рассмотрим функцию $g(w) = |w - u|$ переменной $w \in E^n$. Ясно, что $g(w)$ непрерывна на E^n. Кроме того, $g(w) \geqslant |w| - |u|$, так что $\lim\limits_{|w|\to\infty} g(w) = \infty$. Таким образом, $g(w)$ удовлетворяет условиям теоремы 3 на любом непустом замкнутом множестве $U \subseteq E^n$. Существование искомой точки $v = v(u)$ теперь следует непосредственно из теоремы 3. Заметим, что такая точка $v(u)$, вообще говоря, неединственна.

6. В заключение кратко остановимся на задаче максимизации функции $J(u)$ на множестве U. Так как

$$\sup_U J(u) = -\inf_U (-J(u)),$$

то ясно, что любая точка максимума или любая максимизирующая последовательность для $J(u)$ на U будет соответственно точкой минимума или минимизирующей последовательностью для функции $(-J(u))$ на U. Это значит, что любая задача максимизации функции $J(u)$ на U равносильна задаче минимизации функции $(-J(u))$ на том же множестве U. Поэтому можно ограничиться изучением лишь задач минимизации.

Предлагаем читателю, пользуясь указанной связью между задачами минимизации и максимизации, по аналогии с п. 2 сформулировать задачи максимизации первого и второго типов. Далее, учитывая, что полунепрерывность сверху функции $J(u)$ равносильна полунепрерывности снизу функции $-J(u)$, нетрудно сформулировать и доказать аналоги теорем 1 — 3 для задач максимизации.

Упражнения. 1. Выяснить, будет ли произвольная минимизирующая последовательность сходиться к множеству точек минимума функции $J(u)$ на множестве U, если:

а) $U = \{u = (x, y) \in E^2,\ x \geqslant 0,\ y \geqslant 0,\ x + 2y \leqslant 1\}$, $J(u) = x + y$;

б) $U = E^n$, $J(u) = |u|(1 + |u|^2)^{-1}$;

в) $U = E^n$, $J(u) = |u|^2$.

2. Пусть множество U_* точек минимума функции $J(u)$ на U непусто и ограничено. Доказать, что для сходимости любой минимизирующей последовательности $\{u_k\}$ к U_* необходимо и достаточно, чтобы существовало число $\alpha > 0$ такое, что множество $M_\alpha = \{u\colon\ u \in U,\ J(u) < J_* + \alpha\}$ ограничено.

3. Доказать следующие свойства верхнего и нижнего пределов числовых последовательностей:

а) $\varliminf\limits_{n\to\infty} ca_n = c \varliminf\limits_{n\to\infty} a_n$, $\varlimsup\limits_{n\to\infty} ca_n = c \varlimsup\limits_{n\to\infty} a_n$, $\varliminf\limits_{n\to\infty} (-ca_n) = -c \varlimsup\limits_{n\to\infty} a_n$, $\varlimsup\limits_{n\to\infty} (-ca_n) = -c \varliminf\limits_{n\to\infty} a_n$ для любых $c = \mathrm{const} > 0$;

б) если $a_n \leqslant b_n$ $(n = 1, 2, \ldots)$, то $\varliminf\limits_{n\to\infty} a_n \leqslant \varliminf\limits_{n\to\infty} b_n$, $\varlimsup\limits_{n\to\infty} a_n \leqslant \varlimsup\limits_{n\to\infty} b_n$;

в) $\varliminf\limits_{n\to\infty} a_n + \varlimsup\limits_{n\to\infty} b_n \leqslant \varlimsup\limits_{n\to\infty} (a_n + b_n) \leqslant \varlimsup\limits_{n\to\infty} a_n + \varlimsup\limits_{n\to\infty} b_n$. Рассмотреть пример $a_n = (-1)^n$, $b_n = (-1)^n$ или $b_n = (-1)^{n+1}$ и убедиться, что здесь возможны строгие неравенства;

г) $\varliminf\limits_{n\to\infty} a_n + \varliminf\limits_{n\to\infty} b_n \leqslant \varliminf\limits_{n\to\infty} (a_n + b_n) \leqslant \varliminf\limits_{n\to\infty} a_n + \varlimsup\limits_{n\to\infty} b_n$. Привести примеры последовательностей, когда здесь возможны строгие неравенства;

д) если существует $\lim\limits_{n\to\infty} a_n$, то $\varlimsup\limits_{n\to\infty} (a_n + b_n) = \lim\limits_{n\to\infty} a_n + \varlimsup\limits_{n\to\infty} b_n$, $\varliminf\limits_{n\to\infty} (a_n + b_n) = \lim\limits_{n\to\infty} a_n + \varliminf\limits_{n\to\infty} b_n$;

е) если $a_n \geqslant 0$, $b_n \geqslant 0$ $(n = 1, 2, \ldots)$, то $\varliminf\limits_{n\to\infty} a_n \cdot \varliminf\limits_{n\to\infty} b_n \leqslant \varliminf\limits_{n\to\infty} a_n b_n \leqslant \varliminf\limits_{n\to\infty} a_n \cdot \varlimsup\limits_{n\to\infty} b_n$; $\varliminf\limits_{n\to\infty} a_n \cdot \varlimsup\limits_{n\to\infty} b_n \leqslant \varlimsup\limits_{n\to\infty} a_n b_n \leqslant \varlimsup\limits_{n\to\infty} a_n \cdot \varlimsup\limits_{n\to\infty} b_n$. Привести примеры последовательностей, когда здесь возможны строгие неравенства;

ж) если $a_n \geqslant 0$, $b_n \geqslant 0$ $(n = 1, 2, \ldots)$ и существует $\lim\limits_{n\to\infty} a_n$, то

$$\overline{\lim_{n\to\infty}}\, a_n b_n = \lim_{n\to\infty} a_n \cdot \overline{\lim_{n\to\infty}}\, b_n, \quad \underline{\lim_{n\to\infty}}\, (a_n b_n) = \lim_{n\to\infty} a_n \cdot \underline{\lim_{n\to\infty}}\, b_n.$$

4. Найти верхний и нижний пределы последовательности $a_n = \sin n\alpha$, где α — фиксированное число.

5. Пусть $J(u) = (1 - e^{-|u|})^{-1}$ $(u \neq 0)$. Как надо доопределить эту функцию при $u = 0$, чтобы она стала полунепрерывной снизу или сверху на $E^1 = \mathbf{R}$?

6. Пусть $J_1(u)$, $J_2(u)$ — две функции, полунепрерывные снизу на множестве U. Будут ли полунепрерывными снизу на U следующие функции:

а) $J(u) = \alpha_1 J_1(u) + \alpha_2 J_2(u)$ (рассмотреть случаи положительных и отрицательных α_1, α_2);

б) $J(u) = \min\{J_1(u); J_2(u)\}$;

в) $J(u) = \max\{J_1(u); J_2(u)\}$;

г) $J(u) = |J_1(u)|$?

7. Пусть функция $J(u)$ определена на множестве U. Говорят, что число A является *нижним* [*верхним*] *пределом* этой функции в точке v по множеству U, и обозначают $\underline{\lim\limits_{u\to v}}\, J(u) = A$ $\left[\overline{\lim\limits_{u\to v}}\, J(u) = A\right]$, если:

а) для любой последовательности $\{u_k\} \in U$, сходящейся к v, имеет место неравенство $\underline{\lim\limits_{k\to\infty}}\, J(u_k) \geqslant A$ $\left[\overline{\lim\limits_{k\to\infty}}\, J(u_k) \leqslant A\right]$;

б) существует последовательность $\{u_k\} \in U$, сходящаяся к v и такая, что $\lim\limits_{k\to\infty} J(u_k) = A$.

Доказать, что функция $J(u)$ полунепрерывна снизу [сверху] в точке v, если $\underline{\lim\limits_{u\to v}}\, J(u) \geqslant J(v)$ $\left[\overline{\lim\limits_{u\to v}}\, J(u) \leqslant J(v)\right]$.

8. Пусть $U = E^n$, $J(u) = \sin(\pi|u|^{-1})$ при $u \neq 0$ и $J(0) = a$. При каких a эта функция будет полунепрерывна снизу или сверху на U? Найти $\underline{\lim\limits_{u\to 0}}\, J(u)$, $\overline{\lim\limits_{u\to 0}}\, J(u)$. Что изменится, если $U = \{u\colon |u| = n^{-1}, n = 1, 2, \ldots\}$?

9. Показать, что понятия верхнего и нижнего предела функции в точке обладают свойствами, аналогичными свойствам верхнего и нижнего предела числовых последовательностей, приведенных в упражнении 3.

§ 2. Классический метод

Как и в гл. 1, под классическим методом будем подразумевать тот подход к поиску точек экстремума функции многих переменных (т. е. точек локального минимума и максимума — см. определения 1.1.6, 1.1.10), который основан на дифференциальном исчислении и обычно излагается в учебниках по математическому анализу [10, 160, 165, 233]. Мы здесь лишь вкратце остановимся на этом методе.

1. Сначала напомним некоторые понятия и факты из дифференциального исчисления функций многих переменных.

Определение 1. Пусть функция $J(u)$ определена в некоторой ε-окрестности $O(u, \varepsilon) = \{v\colon v \in E^n, |v - u| < \varepsilon\}$ точки u. Говорят, что функция $J(u)$ *дифференцируема* в точке u, если существует вектор $J'(u) \in E^n$ такой, что приращение функции

$\Delta J(u)=J(u+h)-J(u)$ $(|h|<\varepsilon)$ можно представить в виде

$$\Delta J(u)=\langle J'(u),\ h\rangle+o(h,\ u), \tag{1}$$

где $o(h,\ u)$ — величина, бесконечно малая более высокого порядка, чем $|h|$, т. е. $\lim\limits_{|h|\to 0} o(h, u)|h|^{-1}=0$. Величина $dJ(u)=\langle J'(u),\ h\rangle$ представляет главную линейную относительно h часть приращения (1) и называется *дифференциалом* функции $J(u)$ в точке u, а вектор $J'(u)$ — *градиентом* этой функции в точке u.

Условие (1) однозначно определяет градиент $J'(u)$, причем

$$J'(u)=\left(J'_{u^1}(u),\ \ldots,\ J'_{u^n}(u)\right), \tag{2}$$

где

$$J'_{u^i}(u)=\frac{\partial J(u)}{\partial u^i}=\lim_{\alpha\to 0}\frac{J(u+\alpha e_i)-J(u)}{\alpha}$$

есть частная производная функции $J(u)$ в точке u по переменной u^i, $e_i=(0,\ \ldots,\ 0,\ 1,\ 0,\ \ldots,\ 0)$ — единичный вектор, у которого i-я координата равна 1, остальные координаты равны нулю. Всякая функция, дифференцируемая в точке, непрерывна в этой точке.

Для определения дважды дифференцируемой функции нам понадобится понятие квадратичной формы. Напоминаем, что *квадратичной формой* называют функцию $\sum\limits_{i,j=1}^{n} a_{ij}u^iu^j$ переменных $u=(u^1,\ \ldots,\ u^n)\in E^n$, которая однозначно определяется заданием симметричной числовой матрицы

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{n1} & \cdots & a_{nn}\end{bmatrix}=[a_{ij}]$$

размера n, называемой *матрицей квадратичной формы*. Обозначая через Au вектор-столбец с координатами $(Au)^i=\sum\limits_{j=1}^{n} a_{ij}u^j$ $(i=1,\ \ldots,\ n)$, квадратичную форму можно кратко записать в виде $\langle Au,\ u\rangle$.

Определение 2. Пусть функция $J(u)$ определена в некоторой ε-окрестности точки u. Говорят, что эта *функция дважды дифференцируема в точке* u, если наряду с градиентом $J'(u)$ существует симметричная матрица $J''(u)$ порядка $n\times n$ такая, что приращение функции в точке u можно представить в виде

$$J(u+h)-J(u)=\langle J'(u),\ h\rangle+(1/2)\langle J''(u)h,\ h\rangle+\alpha(h,\ u), \tag{3}$$

где $\alpha(h,\ u)/|h|^2\to 0$ при $|h|\to 0$.

Квадратичную форму $d^2J(u)=\langle J''(u)h,\ h\rangle$ переменной $h=(h^1,\ \ldots,\ h^n)\in E^n$ называют *вторым дифференциалом* функции

$J(u)$ в точке u, а матрицу $J''(u)$ — *второй производной* этой функции в точке u.

Условием (3) матрица $J''(u)$ определяется однозначно, причем

$$J''(u)=\begin{bmatrix} \frac{\partial^2 J(u)}{(\partial u^1)^2} & \frac{\partial^2 J(u)}{\partial u^1 \partial u^2} & \cdots & \frac{\partial^2 J(u)}{\partial u^1 \partial u^n} \\ \frac{\partial^2 J(u)}{\partial u^2 \partial u^1} & \frac{\partial^2 J(u)}{(\partial u^2)^2} & \cdots & \frac{\partial^2 J(u)}{\partial u^2 \partial u^n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 J(u)}{\partial u^n \partial u^1} & \frac{\partial^2 J(u)}{\partial u^n \partial u^2} & \cdots & \frac{\partial^2 J(u)}{(\partial u^n)^2} \end{bmatrix}=\left[J''_{u^i u^j}(u)\right], \tag{4}$$

где $J''_{u^i u^j}(u)=\frac{\partial^2 J(u)}{\partial u^i \partial u^j}=\frac{\partial}{\partial u^i}\left(\frac{\partial J(u)}{\partial u^j}\right)=\frac{\partial}{\partial u^j}\left(\frac{\partial J(u)}{\partial u^i}\right)$ — вторые частные производные функции $J(u)$ по переменным u^i, u^j.

2. Пусть функция $J(u)$ дифференцируема во всех точках пространства E^n. Тогда точками локального минимума или максимума функции $J(u)$ на E^n могут быть лишь те точки v, в которых $J'(v)=0$, что в силу (2) можно записать в виде системы уравнений

$$\frac{\partial J(v)}{\partial u^i}=0, \qquad i=1, \ldots, n. \tag{5}$$

Все точки v, удовлетворяющие системе (5), называются *стационарными точками* функции $J(u)$.

Таким образом, поиск точек экстремума можно начинать с решения системы (5) и определения стационарных точек. Однако, к сожалению, не всякая стационарная точка представляет собой точку локального минимума или максимума. Поэтому после нахождения стационарных точек проводят дополнительное исследование, и из них отбирают те точки, которые в самом деле являются точками экстремума. Если функция $J(u)$ дважды дифференцируема в окрестности стационарной точки и все вторые частные производные этой функции непрерывны в этой точке, то для такого исследования можно использовать второй дифференциал функции. А именно, если в стационарной точке v квадратичная форма $d^2J(v)=\langle J''(v)h, h\rangle$ положительно определена, т. е. $\langle J''(v)h, h\rangle>0$ для всех $h\neq 0$, то v — точка локального минимума функции $J(u)$; если $\langle J''(v)h, h\rangle$ отрицательно определена, т. е. $\langle J''(v)h, h\rangle<0$ при всех $h\neq 0$, то v — точка локального максимума; если же квадратичная форма $\langle J''(v)h, h\rangle$ знакопеременна, т. е. может принимать как положительные, так и отрицательные значения, то в точке v функция $J(u)$ не имеет локального минимума или максимума.

Напомним, что квадратичная форма $\langle Au, u\rangle = \sum_{i,j=1}^{n} a_{ij}u^iu^j$ будет положительно определенной тогда и только тогда, когда все угловые миноры матрицы A, т. е. определители

$$\Delta_i = \det\begin{bmatrix} a_{11} & \dots & a_{1i} \\ \cdot & \cdot & \cdot \\ a_{i1} & \dots & a_{ii} \end{bmatrix}, \qquad i = 1, \dots, n,$$

положительны; форма $\langle Au, u\rangle$ отрицательно определена тогда и только тогда, когда $(-1)^i\Delta_i > 0$ $(i = 1, \dots, n)$, т. е. знаки угловых миноров $\Delta_1, \dots, \Delta_n$ чередуются, причем $\Delta_1 < 0$. Таким образом, для проверки знакоопределенности квадратичной формы $\langle J''(v)h, h\rangle$ достаточно вычислить угловые миноры матрицы $J''(v)$ и исследовать их знаки.

В том случае, когда в стационарной точке квадратичная форма $\langle J''(v)h, h\rangle$ не меняет знака при всех $h \in E^n$, но может равняться нулю при некоторых $h \neq 0$, то для выяснения поведения функции в окрестности точки v можно привлечь старшие производные и связанные с ними формы более высокого порядка:

$$d^mJ(v) = \sum \frac{\partial^m J(v)}{(\partial u^1)^{r_1}\dots(\partial u^n)^{r_n}}(h^1)^{r_1}\dots(h^n)^{r_n},$$

где суммирование проводится по всем целым $r_1, \dots, r_n$ таким, что $0 \leqslant r_i \leqslant m$ $(i = 1, \dots, n,\ r_1 + \dots + r_n = m \geqslant 3)$. Однако на практике исследование характера стационарных точек с помощью форм $d^mJ(v)$ $(m \geqslant 3)$ почти не применяется из-за его громоздкости.

В тех случаях, когда указанным выше путем удается выявить все точки локального минимума [максимума] функции $J(u)$, то для определения глобального минимума [максимума] этой функции на всем пространстве E^n нужно перебрать все точки локального минимума [максимума] и из них выбрать точку с наименьшим [наибольшим] значением функции, если такая точка существует.

Пример 1. Пусть в пространстве E^n даны m точек $u_i = (u_i^1, \dots, u_i^n)$ $(i = 1, \dots, m)$ и требуется найти точку $u \in E^n$, сумма квадратов расстояний от которой до этих данных точек минимальна. Иначе говоря, требуется минимизировать функцию $J(u) = \sum_{i=1}^{m} |u - u_i|^2$ на E^n.

Поскольку $J'(u) = 2\sum_{i=1}^{m}(u - u_i)$, то система (5) здесь выглядит так: $mu - \sum_{i=1}^{m} u_i = 0$. Отсюда получаем стационарную точку

$u_* = \frac{1}{m}\sum_{i=1}^{m} u_i \equiv u_0$, подозрительную на экстремум. Матрица вторых производных равна $J''(u) \equiv 2mE$ при всех $u \in E^n$, где E — единичная матрица размера $n \times n$, т. е. матрица, у которой элементы $a_{ii} = 1$, $a_{ij} = 0$ при $i \neq j$ $(i, j = 1, \ldots, n)$. Отсюда ясно, что $\langle J''(u)h, h\rangle = 2m|h|^2 > 0$ при всех $h \neq 0$. Это значит, что в точке $u_* = u_0$ достигается локальный минимум. Однако здесь можно сказать больше: в точке $u_* = u_0$ функция $J(u)$ достигает своего глобального минимума на E^n. В самом деле, рассматриваемая функция такова, что $\lim\limits_{|u|\to\infty} J(u) = \infty$. Тогда по теореме 1.3 множество U_* точек минимума $J(u)$ на E^n непусто. Кроме того, во всех точках $u_* \in U_*$ должно соблюдаться равенство $J'(u_*) = 0$. Как выяснилось, последнее равенство выполняется только в точке $u_* = u_0$. Следовательно, $U_* = \{u_0\}$, $J_* = J(u_*) = \sum_{i=1}^{m} |u_0 - u_i|^2$.

3. Задачу поиска экстремума функции на всем пространстве E^n, которую мы только что рассмотрели, принято называть *задачей на безусловный экстремум*. В этом названии отражен тот факт, что на переменные $u = (u^1, \ldots, u^n)$ никакие дополнительные условия и ограничения не накладываются. Наряду с задачами на безусловный экстремум имеется немало задач, в которых переменные не могут быть совершенно произвольными и должны удовлетворять некоторым дополнительным условиям, выражающим, например, условие неотрицательности тех или иных переменных, условия ограниченности используемых ресурсов, условия нормировки, ограничение на параметры конструкции системы и т. п. Иначе говоря, переменные $u = (u^1, \ldots, u^n)$ должны принадлежать некоторому заданному множеству U из E^n, причем $U \neq E^n$. Тогда чтобы подчеркнуть, что экстремум функции ищется при условии $u \in U \neq E^n$, часто говорят о *задаче на условный экстремум*.

В классическом математическом анализе традиционно рассматривается следующая задача на условный экстремум: найти экстремумы функции $J(u)$ при условии, что переменные $u = (u^1, \ldots, u^n)$ удовлетворяют ограничениям [10, 160, 165, 233]

$$g_1(u) = 0, \ \ldots, \ g_s(u) = 0; \tag{6}$$

при этом предполагается, что функции $J(u)$, $g_i(u)$ определены и имеют непрерывные частные производные первого порядка на всем пространстве E^n. Ограничения (6) принято называть *ограничениями типа равенств*. Таким образом, здесь мы имеем дело с задачей поиска экстремума функции $J(u)$ на множестве

$$U = \{u\colon\ u \in E^n, \quad g_i(u) = 0, \quad i = 1, \ldots, s\}.$$

В тех случаях, когда систему (6) удается преобразовать к эквивалентному виду

$$u_1 = a_1(u^{p+1}, \ldots, u^n), \ldots, u_p = a_p(u^{p+1}, \ldots, u^n), \tag{7}$$

выразив из (6) первые p переменных через остальные, рассматриваемую задачу на условный экстремум можно свести к задаче на безусловный экстремум функции $\varphi(u^{p+1}, \ldots, u^n) = J(a_1(u^{p+1}, \ldots, u^n), \ldots, a_p(u^{p+1}, \ldots, u^n), u^{p+1}, \ldots, u^n)$ переменных $(u^{p+1}, \ldots, u^n) \in E^{n-p}$, которую можно исследовать, например, по описанной в п. 2 схеме. Однако этот подход имеет ограниченное применение из-за того, что явное выражение вида (7) одной группы переменных через остальные переменные удается получить лишь в редких случаях.

Более общий подход к исследованию задачи поиска экстремума дифференцируемой функции $J(u)$ при ограничениях (6) дает метод множителей Лагранжа. Этот метод заключается в следующем. Вводится функция Лагранжа

$$\mathscr{L}(u, \bar{\lambda}) = \lambda_0 J(u) + \sum_{j=1}^{s} \lambda_j g_j(u) \tag{8}$$

переменных $(u^1, \ldots, u^n, \lambda_0, \lambda_1, \ldots, \lambda_s) = (u, \bar{\lambda}) \in E^{n+s+1}$.

Теорема 1. *Пусть* $*$ — *точка локального минимума или максимума функции* $J(u)$ *на множестве* U, *задаваемом ограничениями* (6), *функции* $J(u)$, $g_1(u), \ldots, g_s(u)$ *непрерывно дифференцируемы в окрестности точки* u_*. *Тогда необходимо существуют числа* $(\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*) = \bar{\lambda}^* \neq 0$, *называемые множителями Лагранжа, такие, что*

$$\left.\frac{\partial \mathscr{L}(u, \bar{\lambda}^*)}{\partial u^i}\right|_{u=u_*} = \lambda_0^* \frac{\partial J(u_*)}{\partial u^i} + \sum_{j=1}^{s} \lambda_j^* \frac{\partial g_j(u_*)}{\partial u^i} = 0, \quad i = 1, \ldots, s. \tag{9}$$

Доказательство. Поскольку не все числа $\lambda_0^*, \ldots, \lambda_s^*$ равны нулю, то условие (9) означает, что векторы $J'(u_*)$, $g_1'(u_*), \ldots, g_s'(u_*)$ линейно зависимы. Допустим противное: пусть эти векторы линейно независимы. Тогда $s+1 \leqslant n$. В случае $s+1 < n$ возьмем какие-либо векторы $e_{s+1}, \ldots, e_{n-1}$ так, чтобы система $J'(u_*), g_1'(u_*), \ldots, g_s'(u_*), e_{s+1}, \ldots, e_{n-1}$ образовала базис в E^n.

Введем функции $f(u, t) = (f_0(u, t), f_1(u, t), \ldots, f_{n-1}(u, t))$: $f_0(u, t) = J(u) - J(u_*) + t$, $f_i(u, t) = g_i(u)$ $(i = 1, \ldots, s)$; $f_i(u, t) = \langle e_i, u - u_* \rangle$ $(i = s+1, \ldots, n-1)$ переменных $(u, t) \in E^{n+1}$ и рассмотрим систему уравнений

$$f_0(u, t) = 0, \ f_1(u, t) = 0, \ \ldots, \ f_{n-1}(u, t) = 0$$

относительно n неизвестных $u=(u^1, \ldots, u^n)$. Для ее исследования воспользуемся теоремой о неявных функциях [10, 160, 165, 233]. Заметим, что $f(u_*, 0)=0$.

Далее, функции $f_i(u, t)$ непрерывно дифференцируемы в окрестности точки $(u_*, 0) \in E^{n+1}$, причем $f'_{0u}(u_*, 0)=J'(u_*)$, $f'_{iu}(u_*, 0)=g'_i(u_*)$ $(i=1, \ldots, s)$; $f'_{iu}(u_*, 0)=e_i$ $(i=s+1, \ldots, n-1)$. Это значит, что в точке $(u_*, 0)$ якобиан системы функций $f_i(u, t)$ $(i=0, 1, \ldots, n-1)$, представляющий собой определитель квадратной матрицы со строками $J'(u_*)$, $g_1(u_*), \ldots, g'_s(u_*), e_{s+1}, \ldots, e_{n-1}$, образующими базис в E^n, отличен от нуля. Тогда по теореме о неявных функциях система $f(u, t)=0$ имеет решение при каждом t $(|t| \leqslant t_0)$, где t_0 — достаточно малое положительное число, или, точнее, существует вектор-функция $u=u(t)=(u^1(t), \ldots, u^n(t))$, которая определена и дифференцируема при всех t $(|t| \leqslant t_0)$ и такая, что

$$u(0)=u_*, \quad J(u(t))=J(u_*)-t, \quad g_i(u(t))=0, \quad i=1, \ldots, s;$$
$$\langle e_i, u(t)-u_* \rangle=0, \quad i=s+1, \ldots, n-1.$$

Это значит, что $u(t) \in U$, $|t| \leqslant t_0$, $u(t) \to u_*$ при $t \to 0$. Отсюда же имеем

$$J(u(t))=J(u_*)-t<J(u_*)<J(u_*)+t=J(u(-t)) \quad \forall t \in (0, t_0),$$

что противоречит тому, что u_* — точка локального экстремума.

Из теоремы 1 следует, что подозрительными на экстремум могут быть лишь те точки u, для которых существуют множители $\bar{\lambda}=(\lambda_0, \lambda_1, \ldots, \lambda_s)$ такие, что точка $(u, \bar{\lambda}) \in E^{n+s+1}$ удовлетворяет системе $n+s$ уравнений

$$\lambda_0 \frac{\partial J(u)}{\partial u^i}+\sum_{j=1}^{s} \lambda_j \frac{\partial g_j(u)}{\partial u^i}=0, \; g_j(u)=0, \; i=1, \ldots, n, \; j=1, \ldots, s, \tag{10}$$

и, кроме того, известно, что $\bar{\lambda}=(\lambda_0, \lambda_1, \ldots, \lambda_s) \neq 0$. Нетрудно видеть, что если $(v, \bar{\lambda})$ — решение системы (10), то $(v, \alpha\bar{\lambda})$ при любом $\alpha \neq 0$ также является решением этой системы, т. е. множители $(\lambda_0, \lambda_1, \ldots, \lambda_s)$ из (10) определяются с точностью постоянного множителя. Поэтому множители $\bar{\lambda}$ можно подчинить какому-либо дополнительному условию нормировки, например,

$$\lambda_0 \geqslant 0, \quad |\bar{\lambda}|^2=\sum_{i=0}^{s} \lambda_i^2=1. \tag{11}$$

Если система (10) имеет решения $(v, \bar{\lambda})$ такие, что $\lambda_0 \neq 0$, то задачу минимизации $J(u)$ при условиях (6) называют *регулярной* (невырожденной) в точке v. В регулярной задаче условие нормировки (11) можно заменить более простым условием $\lambda_0=1$. Нетрудно видеть, что для регулярности задачи в точ-

ке v достаточно, чтобы векторы $g_1'(v), \ldots, g_s'(v)$ были линейно независимы, т. е. чтобы равенство $\alpha_1 g_1'(v) + \ldots + \alpha_s g_s'(v) = 0$ было возможно только при $\alpha_1 = \ldots = \alpha_s = 0$.

Условия (10) с условием нормировки (11) (или $\lambda_0 = 1$) представляют собой систему $n+s+1$ уравнений с $n+s+1$ неизвестными $(u, \bar{\lambda}) = (u^1, \ldots, u^n, \lambda_0, \lambda_1, \ldots, \lambda_s)$. Решив ее, мы найдем точки $v \in U$, подозрительные на экстремум, и соответствующие им множители Лагранжа $\bar{\lambda} = \bar{\lambda}_0 = (\lambda_0, \lambda_1, \ldots, \lambda_s) \neq 0$. Для выяснения того, будет ли в этих точках в действительности реализовываться локальный минимум или максимум, нужно провести дополнительные исследования с привлечением вторых производных функции Лагранжа по переменной u.

Теорема 2. *Пусть*: 1) *функции* $J(u)$, $g_1(u)$, ..., $g_s(u)$ *дважды дифференцируемы в точке* $v \in U = \{u \in E^n\colon g_1(u) = 0, \ldots, g_s(u) = 0\}$; 2) *точка* $(v, \bar{\lambda}_v)$ *удовлетворяет условиям* (10), (11); 3) *квадратичная форма*

$$\langle \mathscr{L}_{uu}(v, \bar{\lambda}_v) h, h \rangle > 0 \quad [<0] \tag{12}$$

при всех h, *для которых*

$$\langle J'(v), h \rangle \leqslant 0 \quad [\geqslant 0], \quad \langle g_i'(v), h \rangle = 0, \quad i = 1, \ldots, s; \ h \neq 0. \tag{13}$$

Тогда в точке v *функция* $J(u)$ *на* U *имеет строгий локальный минимум* [*максимум*].

Доказательство. Допустим противное: точка v удовлетворяет условиям теоремы, но не является точкой строгого локального минимума $J(u)$ на U. Тогда найдется такая последовательность $\{u_k\}$, что

$$\begin{gathered} g_1(u_k) = 0, \ldots, g_s(u_k) = 0, \quad J(u_k) \leqslant J(v), \\ u_k \neq v, \quad k = 1, 2, \ldots; \quad \{u_k\} \to v \end{gathered} \tag{14}$$

(предполагаем, что v не является изолированной точкой множества U; всякая изолированная точка множества U может считаться точкой строгого локального минимума или максимума и без выполнения условий (10)—(13)). Точку u_k можно записать в виде

$$u_k = v + |u_k - v| \frac{u_k - v}{|u_k - v|} = v + t_k e_k, \quad e_k = \frac{u_k - v}{|u_k - v|},$$
$$t_k = |u_k - v|, \quad \{t_k\} \to 0.$$

Поскольку $|e_k| = 1$, то, выбирая при необходимости подпоследовательность, можем считать, что $\{e_k\} \to e_0$, $|e_0| = 1$.

С учетом (14) и дифференцируемости функций $J(u)$, $g_i(u)$ в точке v имеем

$$0 \geqslant J(u_k) - J(v) = \langle J'(v), e_k \rangle t_k + o(t_k),$$
$$0 = g(u_k) - g(v) = \langle g_i'(v), e_k \rangle t_k + o(t_k)$$

для всех $i=1, \ldots, s$, $k=1, 2, \ldots$ Разделив эти соотношения на $t_k>0$ и устремив $k\to\infty$, получим $\langle J'(v), e_0\rangle \leqslant 0$, $\langle g_i'(v), e_0\rangle = 0$ $(i=1, \ldots, s)$, так что e_0 удовлетворяет условиям (13).

Далее, из того, что u_k, $v\in U$, $\lambda_0\geqslant 0$, и из (14) имеем $\mathscr{L}(u_k, \bar{\lambda}_v) = \lambda_0 J(u_k) + \sum_{i=1}^{s} \lambda_i g_i(u_k) = \lambda_0 J(u_k) \leqslant \lambda_0 J(v) = \lambda_0 J(v) + \sum_{i=1}^{s} \lambda_i g_i(v) = \mathscr{L}(v, \bar{\lambda}_v)$. Отсюда с учетом условия (10) и дважды дифференцируемости функции $\mathscr{L}(v, \bar{\lambda}_v)$ в точке v получаем

$$0 \geqslant \mathscr{L}(u_k, \bar{\lambda}_v) - \mathscr{L}(v, \bar{\lambda}_v) = \frac{1}{2} t_k^2 \langle \mathscr{L}_{uu}(v, \bar{\lambda}_v) e_k, e_k\rangle + o(t_k^2),$$
$$k=1, 2, \ldots$$

Разделив это неравенство на $t_k^2>0$ и устремив $k\to\infty$, будем иметь $\langle \mathscr{L}_{uu}(v, \lambda_v) e_0, e_0\rangle \leqslant 0$, что противоречит условиям (12), (13). Аналогично исследуется случай, когда v — точка строгого локального максимума.

Пример 2. Пусть требуется на n-мерной единичной сфере $U=\{u\in E^n\colon |u|^2=\langle u, u\rangle=1\}$ найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \ldots, u_m\in E^n$ была бы минимальной, т. е. нужно минимизировать функцию $J(u) = \sum_{i=1}^{m} |u-u_i|^2$ при условии $\langle u, u\rangle = 1$.

Для решения этой задачи составим функцию Лагранжа $\mathscr{L}(u, \bar{\lambda}) = \lambda_0 J(u) + \lambda(\langle u, u\rangle - 1)$. Система (10) примет вид

$$\mathscr{L}_u(u, \bar{\lambda}) = \lambda_0 \cdot 2m(u-u_0) + 2\lambda u = 0, \quad \langle u, u\rangle = 1, \tag{15}$$

где $u_0 = \frac{1}{m}\sum_{i=1}^{m} u_i$ (см. пример 1). Очевидно, при $\lambda_0=0$ система (15) не имеет решения с $(\lambda_0, \lambda)\neq 0$, поэтому можем принять $\lambda_0=1$. Тогда из (15) при $u_0\neq 0$ получим две точки $v_1=u_0/|u_0|$ и $v_2=-u_0/|u_0|$, подозрительные на минимум, и соответствующие им множители Лагранжа $\lambda_1=m(|u_0|-1)$ и $\lambda_2 = m(-|u_0|-1)$. Матрицы вторых производных функции Лагранжа в найденных точках (v_1, λ_1) и (v_2, λ_2) равны соответственно $2|u_0|mE$ и $-2|u_0|mE$, где E — единичная матрица. Отсюда ясно, что в точке v_1 достигается локальный минимум, а в точке v_2 — локальный максимум функции $J(u)$ при условии $|u|=1$. Поскольку U — компактное множество, $J(u)$ непрерывна на U, то согласно теореме 1.1 на U функция $J(u)$ достигает своего глобального минимума и максимума. Но точки глобального экстремума, конечно же, удовлетворяют системе (15). Как выяснилось, при $u_0\neq 0$ система (15) имеет всего два решения v_1 и v_2. Следовательно, $v_1=u_0/|u_0|$ — точка глобального мини-

мума, $v_2 = -u_0/|u_0|$ — точка глобального максимума при $u_0 \neq 0$. Таким образом, искомая точка есть $u_* = u_0/|u_0|$ при $u_0 \neq 0$.

Рассмотрим случай $u_0 = 0$. Тогда решением системы (15) является точка (v, λ_0, λ), где $\lambda_0 = 1$, $\lambda = -m$, а v — произвольная точка, для которой $|v| = 1$. Это значит, что из необходимых условий экстремума (15) при $u_0 = 0$ не удалось извлечь никакой полезной информации — все точки единичной сферы как были, так и остались подозрительными на экстремум. Тем не менее здесь нетрудно разобраться в происходящем. А именно, при $u_0 = 0$ для всех $u \in U$ имеем $J(u) = \sum_{i=1}^{m} (|u|^2 - 2\langle u, u_i\rangle + |u_i|^2) = m - 2\left\langle u, \sum_{i=1}^{m} u_i \right\rangle + \sum_{i=1}^{m} |u_i|^2 = m - \sum_{i=1}^{m} |u_i|^2 = \text{const}$. Таким образом, при $u_0 = 0$ рассматриваемая задача становится тривиальной: $J(u) = \text{const}$ на U, т. е. можно сказать, что во всех точках $u \in U$ функция достигает глобального минимума (или максимума).

Пример 3. Пусть требуется найти точки экстремума функции $J(u) = x$ на множестве $U = \{u = (x, y) \in E^2,\ x^3 - y^2 = 0\}$. Применим метод множителей Лагранжа. Здесь функция Лагранжа такая: $\mathscr{L}(u, \bar{\lambda}) = \lambda_0 x + \lambda(x^3 - y^2)$. Система (10), (11) имеет вид

$$\lambda_0 + 3\lambda x = 0, \quad 2\lambda y = 0, \quad x^3 - y^2 = 0, \quad \lambda_0^2 + \lambda^2 = 1, \quad \lambda_0 \geqslant 0.$$

Нетрудно видеть, что она совместна лишь при $\lambda_0 = 0$ и выделяет единственную точку $v = (0, 0) = 0$, подозрительную на экстремум. Таким образом, рассматриваемая задача нерегулярна в точке $v = 0$. Она просто решается, если из равенства $x^3 - y^2 = 0$ выразить $x = y^{2/3}$ и заметить, что $J(u) = y^{2/3}$ $(-\infty < y < \infty)$. Ясно, что $v = (0, 0)$ — точка глобального минимума функции $J(u)$ на U.

4. Метод множителей Лагранжа может быть применен для поиска экстремумов функции и в тех случаях, когда на переменные $u = (u^1, \ldots, u^n)$ наряду с ограничениями (6) типа равенств накладываются еще ограничения

$$h_1(u) \leqslant 0, \ \ldots, \ h_m(u) \leqslant 0, \tag{16}$$

называемые *ограничениями типа неравенств*.

Будем предполагать, что функции $J(u)$, $g_i(u)$, $h_j(u)$ $(i = 1, \ldots, s,\ j = 1, \ldots, m)$ определены и дифференцируемы во всех точках $u \in E^n$. Оказывается, если ввести новые вспомогательные переменные $w = (w^1, \ldots, w^m)$, связанные с исходными переменными $u = (u^1, \ldots, u^n)$ соотношениями

$$(w^1)^2 + h_1(u) = 0, \ \ldots, \ (w^m)^2 + h_m(u) = 0, \tag{17}$$

то задача поиска экстремумов функции $J(u)$ при ограничениях (6), (16) сводится к равносильной задаче поиска экстремумов

той же функции в пространстве переменных $\bar{u}=(u^1, \ldots, u^n, w^1, \ldots, w^m)=(u, w)$ при ограничениях типа равенств (6), (17). Равносильность этих двух задач понимается в следующем смысле: если u_* — точка локального минимума [максимума] функции $J(u)$ при ограничениях (6), (16), то $\bar{u}_*=(u_*, w_*)$, где $w_*=(w_*^1, \ldots, w_*^m)$, $w_*^j=(-h_j(u_*))^{1/2}$ $(j=1, \ldots, m)$ будет точкой локального минимума [максимума] функции $J(u)$ при ограничениях (6), (17), и наоборот, если $\bar{u}_*=(u_*, w_*)$ — точка локального минимума [максимума] функции $J(u)$ при ограничениях (6), (17), то u_* — точка локального минимума [максимума] $J(u)$ при ограничениях (6), (16).

Таким образом, для поиска экстремумов функции $J(u)$ при ограничениях (6), (17) можно ввести функцию Лагранжа

$$\mathscr{L}(u, w, \bar{\lambda}, \mu)=\lambda_0 J(u)+\sum_{i=1}^{s} \lambda_i g_i(u)+\sum_{j=1}^{m} \mu_j\left((w^j)^2+h_j(u)\right)$$

и в соответствии с изложенным в п. 3 написать необходимые условия экстремума в виде системы (10), найти точки, подозрительные на экстремум, выявить среди них точки локального минимума или максимума, а затем, исключив переменные $w^1, \ldots, w^m$, получить точки экстремума функции $J(u)$ при ограничениях (6), (16).

Пример 4. Пусть требуется в n-мерном единичном шаре $U=\{u\colon u \in E^n, |u|^2=\langle u, u\rangle \leqslant 1\}$ найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \ldots, u_m \in E^n$ была бы минимальной, т. е. требуется минимизировать функцию $J(u)=\sum_{i=1}^{m}|u-u_i|^2$ при ограничениях $\langle u, u\rangle \leqslant 1$.

Из примера 1 следует, что глобальный минимум функции $J(u)$ на всем пространстве E^n достигается в точке $u_*=u_0=\frac{1}{m}\sum_{i=1}^{m} u_i$. Поэтому если $|u_0| \leqslant 1$, то эта точка u_0 будет решением рассматриваемой задачи.

Остается рассмотреть случай $|u_0|>1$. Введем новую переменную w соотношением

$$w^2+\langle u, u\rangle-1=0 \tag{18}$$

и рассмотрим задачу минимизации функции $J(u)$ в пространстве переменных $\bar{u}=(u, w)=(u^1, \ldots, u^n, w) \in E^{n+1}$ при ограничениях (18). Составим функцию Лагранжа $\mathscr{L}(\bar{u}, \bar{\lambda})=\lambda_0 \sum_{i=1}^{m}|u-u_i|^2+\lambda(w^2+|u|^2-1)$. Система (10) будет иметь вид

$$\mathscr{L}_u=2\lambda_0 m(u-u_0)+2\lambda u=0,$$
$$\mathscr{L}_w=2\lambda w=0, \quad w^2+|u|^2-1=0,$$

где $\bar{\lambda}=(\lambda_0, \lambda)\neq 0$. Напоминаем, что $|u_0|>1$. Очевидно, при $\lambda_0=0$ эта система не имеет решения. Поэтому можем принять $\lambda_0=1$ и переписать систему в виде

$$(m+\lambda)u=mu_0, \quad \lambda w=0, \quad w^2+|u|^2=1, \quad |u_0|>1. \tag{19}$$

Заметим, что здесь случай $\lambda=0$ невозможен, так как тогда $u=u_0$ и $|u|^2=|u_0|^2=1-w^2\leqslant 1$, что противоречит условию $|u_0|>1$. Если же $\lambda\neq 0$, то из (19) получаем два решения $\bar{v}_1=(u_0/|u_0|, 0)$, $\lambda_1=m(|u_0|-1)>0$ и $\bar{v}_2=(-u_0/|u_0|, 0)$, $\lambda_2=-m(|u_0|+1)<0$. Это значит, что экстремум функции $J(u)$ при ограничениях (18) может достигаться лишь в точках $\bar{v}_1$ и $\bar{v}_2$.

Матрица вторых производных функции Лагранжа в найденных точках $(\bar{v}_1, \lambda_1)$, $(\bar{v}_2, \lambda_2)$ соответственно имеет вид

$$\begin{bmatrix} m|u_0|E & 0 \\ 0 & m(|u_0|-1) \end{bmatrix}, \quad \begin{bmatrix} -m|u_0|E & 0 \\ 0 & -m(|u_0|+1) \end{bmatrix}.$$

Отсюда ясно, что при $|u_0|>1$ в точках $\bar{v}_1=(u_0/|u_0|, 0)$ достигается локальный минимум, а в точке $\bar{v}_2=(-u_0/|u_0|, 0)$ — локальный максимум при ограничениях (18). Исключив переменную w, отсюда получаем, что $u_1=u_0/|u_0|$ — точка локального минимума $J(u)$ на единичном шаре, а $u_2=-u_0/|u_0|$ — точка локального максимума. Используя теорему 1.1, как и в примере 2, нетрудно убедиться, что в действительности в этих точках достигается глобальный минимум и соответственно глобальный максимум функции $J(u)$ на единичном шаре. Таким образом, искомой точкой будет $u_*=u_0$ при $|u_0|\leqslant 1$ и $u_*=u_0/|u_0|$ при $|u_0|>1$.

Мы еще не раз будем возвращаться к задаче поиска экстремума функций при ограничениях типа (6), (16); в частности, в § 4.8, 4.9 с других позиций будут получены необходимые и достаточные условия минимума, также использующие множители Лагранжа.

В заключение заметим, что изложенный выше классический метод исследования задач на условный и безусловный экстремум может быть широко использован во всех тех случаях, когда достаточно просто удается выявить все подозрительные на экстремум точки и отобрать из них точки локального минимума и максимума. Однако, как и в случае функции одной переменной, классический метод, к сожалению, имеет весьма ограниченное применение. В общем случае отыскание точек, подозрительных на экстремум, из условий (5) или (10) само по себе представляет весьма серьезную задачу, сравнимую по трудности, быть может, с исходной задачей на экстремум. Дело осложняется также и тем, что в практических задачах иногда бывает непросто выписать даже сами системы уравнений (5) или (10), так как не всегда ясно, существуют ли требуемые для этого

производные и как их вычислить. Из сказанного ясно, что кроме классического метода необходимы также и другие численные методы поиска экстремумов функций многих переменных.

Упражнения. 1. Найти экстремумы функций:

а) $J(u) = (x+y-1)\exp\{-(x^2-xy+y^2)\}$, где $u=(x, y)\in E^2$;

б) $J(u) = xy^2z^3(1-x-2y-3z)$, где $u=(x, y, z)\in E^3$;

2. Найти точки экстремума функции $J(u) = \sin(x+y) - \sin x - \sin y$ на множестве U, если:

а) $U = E^2$;

б) $U = \{u=(x, y): x\geqslant 0, y\geqslant 0, x+y\leqslant 2\pi\}$.

3. Найти точки экстремума функций $J(u) = x+y$, $J(u) = |x| + |y-1|$, $J(u) = x^2+2y^2$ на множествах U, если:

а) $U = \{u=(x, y): 0\leqslant x\leqslant 1, 0\leqslant y\leqslant 1\}$;

б) $U = \{u=(x, y): x\geqslant 0, y\geqslant 0, ax+by=1\}$, где a, b — неотрицательные числа;

в) $U = \{u=(x, y): x-y^2\geqslant 0, x^2+y^2\leqslant 1\}$;

г) $U = \{u=(x, y): x^4-4x^3+4x^2+1\leqslant y\leqslant 1\}$.

Указание: на плоскости нарисовать графики функций $J(u) = c$ для различных значений постоянной c (линий уровня).

4. Среди всех вписанных в данный круг радиуса R треугольников найти тот, площадь которого наибольшая.

5. Из всех параллелепипедов, имеющих ребра данной длины, найти параллелепипед наибольшего объема.

6. Среди треугольных пирамид с данным основанием и высотой найти ту, которая имеет наименьшую боковую поверхность.

7. Пусть $\Delta = \Delta(u_1, \ldots, u_n)$ — определитель матрицы $(u_1, \ldots, u_n)$, столбцами которой являются вектор-столбцы u_i с координатам $(u_i^1, \ldots, u_i^n)$ $(i=1, \ldots, n)$. Найти наибольшее и наименьшее значение величины определителя Δ при условии, что $|u_i| = a_i$, где a_i — заданные числа $(i=1, \ldots, n)$. Доказать неравенство Адамара $|\Delta|\leqslant |u_1|\cdot\ldots\cdot|u_n|$. Дать геометрическую интерпретацию задачи при $n=2, 3$ (ср. с упражнением 5) ([165, ч. I, с. 554—557]).

8. Найти наименьшее и наибольшее значение квадратичной формы $J(u) = \langle Au, u\rangle = \sum\limits_{i,j=1}^{n} a_{ij}u^iu^j$ при условии $u\in U = \{u: u\in E^n, \langle u, u\rangle = 1\}$, где A — симметричная матрица. Показать, что $J_* = \min\limits_U J(u)$ и $J^* = \max\limits_U J(u)$ представляют собой соответственно наименьшее и наибольшее собственное число матрицы A ([164, с. 209]).

9. В множестве $U = \{u: u\in E^n, \langle c, u\rangle = 1\}$ или $U = \{u: u\in E^n, \langle c, u\rangle\leqslant 1\}$, где c — заданный вектор из E^n, найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \ldots, u_m\in E^n$ была бы минимальной (ср. с задачами из примеров 1, 2, 4).

10. Может ли функция двух переменных на плоскости иметь бесконечно много точек локального минимума и ни одной точки локального максимума? Рассмотреть функцию $J(u) = xe^x - (1+e^x)\cos y$.

11. Пусть в некоторой точке $u_0 = (x_0, y_0)\in E^2$ функция $J(u)$ переменных $u=(x, y)$ имеет локальный минимум вдоль каждой прямой, проходящей через точку u_0. Можно ли утверждать, что в точке u_0 реализуется локальный минимум функции $J(u)$? Рассмотреть функцию $J(u) = (x-y^2)(2x-y^2)$ в точке $u_0 = (0, 0)$.

12. Доказать, что если функция $J(u)$ имеет первые частные производные, непрерывные в окрестности точки v, то $J(u)$ дифференцируема в точке v. На примере функции $J(u) = |xy|^{1/2}$ $(u=(x, y), v=(0, 0))$ показать, что одно лишь существование частных производных в точке v еще не гарантирует ее дифференцируемость в этой точке.

13. Доказать, что если v — точка локального минимума [максимума] дважды дифференцируемой функции $J(u)$ на всем пространстве E^n, то необходимо $\langle J''(v)h, h\rangle \geqslant 0$ [$\leqslant 0$] при всех $h \in E^n$.

14. Доказать, что если u_* — точка локального минимума [максимума] дважды дифференцируемой функции $J(u)$ при ограничениях (6), а $\bar{\lambda}^* = (\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*)$ — соответствующие множители Лагранжа, определенные из системы (10), причем $\lambda_0^* = 1$, то $\langle \mathscr{L}_{uu}(u_*, \bar{\lambda}^*)h, h\rangle \geqslant 0$ [$\leqslant 0$] для всех h, удовлетворяющих условиям (13) (ср. с условиями (12)).

15. Применить метод множителей Лагранжа для поиска экстремума функции $J(u) = x$ на множестве $U = \{u = (x, y) \in E^2,\ x^p + y^q = 0\}$, где p, q — натуральные числа. Выяснить, при каких p, q задача является вырожденной в точках экстремума.

16. Применить метод множителей Лагранжа для поиска экстремума $J(u) = x$ на множествах $U = \{u = (x, y) \in E^2\colon\ y - x^{3/2} = 0\}$, $U = \{u\colon x \geqslant 0,\ y - x^{3/2} = 0\}$.

17. Пусть $J(u) = x$, $U = \{u = (x, y) \in E^2\colon g_1(u) = x^2 - y = 0,\ g_2(u) = x^2 + y = 0,\ g_3(u) = x = 0\}$. Показать, что для точки минимума $u_* = (0, 0)$ множителями Лагранжа $\bar{\lambda} = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ могут быть точки $\bar{\lambda}_1 = (0, 1, -1, 0)$, $\bar{\lambda}_2 = (1, 0, 0, -1)$, а также точки $\alpha\bar{\lambda}_1 + \beta\bar{\lambda}_2$, где α, β — любые действительные числа.

§ 3. Вспомогательные предложения

Ниже приводятся некоторые формулы и различные другие сведения, которые будут использованы в дальнейшем при описании и исследовании методов минимизации.

1. Начнем с того, что введем классы функций $C^1(U)$, $C^2(U)$.

Определение 1. Функцию $J(u)$ называют *непрерывно дифференцируемой* или *гладкой* на множестве $U \subseteq E^n$, если $J(u)$ дифференцируема во всех точках $u \in U$ и, кроме того, $|J'(u+h) - J'(u)| \to 0$ при $|h| \to 0$ для всех u, $u + h \in U$. Множество таких функций принято обозначать через $C^1(U)$.

Определение 2. Функцию $J(u)$ называют *дважды непрерывно дифференцируемой* или *дважды гладкой* на множестве $U \subseteq E^n$, если $J(u)$ дважды дифференцируема во всех точках $u \in U$ и $\|J''(u+h) - J''(u)\| \to 0$ при $|h| \to 0$ для всех u, $u + h \in U$. Множество дважды гладких на U функций обозначают через $C^2(U)$.

В определении 2 $\|A\| = \sup\limits_{|e| \leqslant 1} |Ae|$ — норма матрицы A. Заметим, что в определениях 2.1, 2.2 требуется, чтобы дифференцируемая в точке функция была определена в некоторой окрестности этой точки, причем радиус $\varepsilon > 0$ этой окрестности может зависеть от рассматриваемой точки, самой функции и может быть очень малым. Это значит, что если $J(u) \in C^1(U)$ или $J(u) \in C^2(U)$, то либо множество U открыто, либо функция $J(u)$ определена на открытом множестве, содержащем U. Это обстоятельство мы всегда будем подразумевать, говоря о классах функций $C^1(U)$, $C^2(U)$.

Возьмем какую-либо функцию $J(u)$, определенную на множестве $U \subseteq E^n$. Пусть точки u, $u+h \in U$ таковы, что $h \neq 0$, $u+th \in U$ при всех t $(0 \leqslant t \leqslant 1)$. Тогда можно рассматривать функцию одной переменной $f(t) = J(u+th)$ при $t \in [0, 1]$. Оказывается, если $J(u) \in C^p(U)$ при $p=1$ или $p=2$, то $f(t) \in C^p[0, 1]$, причем

$$f'(t) = \langle J'(u+th), h\rangle, \quad f''(t) = \langle J''(u+th)h, h\rangle, \quad 0 \leqslant t \leqslant 1. \tag{1}$$

В самом деле, если, например, $J(u) \in C^2(U)$, то, заменив в формуле (2.3) u на $u+th$, h на $\Delta t h$, получим

$$f(t+\Delta t) - f(t) = \Delta t \langle J'(u+th), h\rangle + \\ + \frac{1}{2}(\Delta t)^2 \langle J''(u+th)h, h\rangle + o(|\Delta t|^2).$$

Такое разложение означает, что $f(t) \in C^2[0, 1]$, и указывает на справедливость формул (1).

Для функций одной переменной имеют место формулы

$$f(t) - f(0) = f'(\theta_1 t)\, t = \int_0^t f'(\tau)\, d\tau = f'(0)\, t + \frac{1}{2} f''(\theta_2 t)\, t^2,$$

$$f'(t) - f'(0) = f''(\theta_3 t)\, t, \qquad 0 \leqslant \theta_1, \theta_2, \theta_3 \leqslant 1.$$

Полагая в этих формулах $t=1$ и пользуясь равенствами (1), получаем различные формулы для конечных приращений функции многих переменных:

$$J(u+h) - J(u) = \langle J'(u+\theta_1 h), h\rangle = \int_0^1 \langle J'(u+th), h\rangle\, dt, \tag{2}$$

$$J(u+h) - J(u) = \langle J'(u), h\rangle + \frac{1}{2}\langle J''(u+\theta_2 h)h, h\rangle, \tag{3}$$

$$\langle J'(u+h) - J'(u), h\rangle = \langle J''(u+\theta_3 h)h, h\rangle, \tag{4}$$

где $0 \leqslant \theta_1, \theta_2, \theta_3 \leqslant 1$. Далее, так как

$$\frac{d}{dt}(J'(u+th)) = J''(u+th)h, \qquad 0 \leqslant t \leqslant 1,$$

то, интегрируя это равенство по t на отрезке $[0, 1]$, получаем

$$J'(u+h) - J'(u) = \int_0^1 J''(u+th)h\, dt = \left(\int_0^1 J''(u+th)\, dt\right) h. \tag{5}$$

Подчеркнем еще раз, что в формулах (1)—(5) подразумевается, что точки u, $u+h$ принадлежат множеству U вместе с отрезком $u+th$ $(0 \leqslant t \leqslant 1)$. В частности, эти формулы верны на любых выпуклых множествах — множествах, которые содер-

жат вместе с любыми двумя своими точками u и v и отрезок $[u, v] = \{u_\alpha = \alpha u + (1-\alpha) v,\ 0 \leqslant \alpha \leqslant 1\}$, соединяющий эти точки (подробнее о выпуклых множествах см. § 4.1).

2. При описании и исследовании методов минимизации нам часто придется иметь дело с функциями, градиент которых удовлетворяет условию Липшица.

Определение 3. Пусть $J(u) \in C^1(U)$. Скажем, что градиент $J'(u)$ этой функции удовлетворяет *условию Липшица* на множестве U с постоянной $L \geqslant 0$, если

$$|J'(u) - J'(v)| \leqslant L|u - v|, \quad u,\ v \in U. \tag{6}$$

Класс таких функций будем обозначать через $C^{1,1}(U)$.

Лемма 1. *Пусть U — выпуклое множество, $J(u) \in C^{1,1}(U)$. Тогда*

$$|J(u) - J(v) - \langle J'(v),\ u - v\rangle| \leqslant L|u - v|^2/2 \tag{7}$$

при всех u, $v \in U$.

Доказательство. С помощью формулы (2) имеем

$$J(u) - J(v) - \langle J'(v), u - v\rangle = \int_0^1 \langle J'(v + t(u - v)) - J'(v), u - v\rangle dt.$$

С учетом условия (6) получим

$$J(u) - J(v) - \langle J'(v), u - v\rangle| \leqslant$$

$$\leqslant \int_0^1 |J'(v + t(u - v)) - J'(v)|\,|u - v|\,dt \leqslant \int_0^1 L|u - v|^2 t\,dt = \frac{L|u - v|^2}{2}.$$

3. Приведем несколько лемм о числовых последовательностях, которые нам пригодятся при доказательстве сходимости методов минимизации, при оценке скорости их сходимости.

Лемма 2. *Пусть числовая последовательность $\{a_k\}$ такова, что*

$$a_{k+1} \leqslant a_k + \delta_k, \quad \delta_k \geqslant 0, \quad k = 0, 1, \ldots, \quad \sum_{k=0}^{\infty} \delta_k < \infty. \tag{8}$$

Тогда существует $\lim\limits_{k\to\infty} a_k < \infty$. Если $\{a_k\}$ ограничена еще и снизу, то $\lim\limits_{k\to\infty} a_k$ конечен.

Заметим, что если $\delta_k = 0$ $(k = 0, 1, \ldots)$, то последовательность $\{a_k\}$ не возрастает, и лемма 1 превращается в хорошо известное утверждение о пределе монотонной последовательности.

Доказательство. Суммируя первое из неравенств (8), имеем

$$a_m \leqslant a_k + \sum_{i=k}^{m-1} \delta_i \leqslant a_k + \sum_{i=k}^{\infty} \delta_i \tag{9}$$

при всех $m > k \geqslant 0$. Пусть $\varliminf\limits_{k\to\infty} a_k = \lim\limits_{n\to\infty} a_{k_n}$ $(k_n < k_{n+1},\ n = 0, 1, \ldots)$; $\lim\limits_{n\to\infty} k_n = \infty$. Положим в (9) $k = k_n$. Получим $a_m \leqslant a_{k_n} + \sum\limits_{i=k_n}^{\infty} \delta_i$ $(m > k_n)$. Следовательно, $\varlimsup\limits_{m\to\infty} a_m \leqslant a_{k_n} + \sum\limits_{i=k_n}^{\infty} \delta_i$ для всех $n = 1, 2, \ldots$ Отсюда при $n \to \infty$ имеем $\varlimsup\limits_{m\to\infty} a_m \leqslant \lim\limits_{n\to\infty} a_{k_n} = \varliminf\limits_{n\to\infty} a_m$. Однако всегда $\varliminf\limits_{n\to\infty} a_m < \varlimsup\limits_{m\to\infty} a_m$. Поэтому $\varliminf\limits_{m\to\infty} a_m = \varlimsup\limits_{m\to\infty} a_m$. Отсюда следует существование предела $\{a_k\}$. Далее, при $k = 0$ из (9) следует ограниченность $\{a_k\}$ сверху. Поэтому, если $\{a_k\}$ ограничена еще и снизу, то $\lim\limits_{k\to\infty} a_k$ конечен.

Лемма 3. *Пусть числовая последовательность $\{b_k\}$ такова, что*

$$b_{k+1} \geqslant b_k - \delta_k, \quad \delta_k \geqslant 0, \quad k = 0, 1, \ldots, \quad \sum_{k=0}^{\infty} \delta_k < \infty.$$

Тогда существует $\lim\limits_{k\to\infty} b_k > -\infty$. Если $\{b_k\}$ ограничена еще и сверху, то $\lim\limits_{k\to\infty} b_k$ конечен.

Эта лемма сводится к лемме 2, если принять $b_k = -a_k$ $(k = 0, 1, \ldots)$.

Лемма 4. *Пусть числовая последовательность $\{a_k\}$ такова, что*

$$a_k \geqslant 0, \quad k = 0, 1, \ldots; \quad a_k - a_{k+1} \geqslant A a_k^2, \quad k \geqslant k_0 \geqslant 0, \tag{10}$$
$$A = \text{const} > 0.$$

Тогда $a_k = O(k^{-1})$ $(k = 1, 2, \ldots)$, т. е. найдется постоянная $B > 0$ такая, что

$$0 \leqslant a_k \leqslant B k^{-1}, \quad k = 1, 2, \ldots \tag{11}$$

Доказательство. Если $a_m = 0$ при некотором $m \geqslant k_0$, то из (10) следует, что $a_k = 0$ при всех $k \geqslant m$, и оценка (11) становится тривиальной — в (11) достаточно взять $B = m \max\limits_{1 \leqslant i \leqslant m} a_i$. Поэтому пусть $a_n > 0$ при всех $n \geqslant k_0$. Тогда из (10) имеем

$$\frac{1}{a_{n+1}} - \frac{1}{a_n} = \frac{a_n - a_{n+1}}{a_n a_{n+1}} \geqslant \frac{a_n}{a_{n+1}} A \geqslant A > 0, \quad n \geqslant k_0.$$

Суммируя эти неравенства по n от k_0 до некоторого $k - 1 \geqslant k_0$, получаем $a_k^{-1} - a_{k_0}^{-1} \geqslant A(k - k_0)$ или $a_k \leqslant A^{-1}(k - k_0)^{-1}$ $(k > k_0)$.

Но $(k-k_0)^{-1} \leqslant (k_0+1)k^{-1}$ при $k > k_0$, поэтому $0 < a_k < (k_0+1)A^{-1}k^{-1}$. Если $1 \leqslant k \leqslant k_0$, то $0 \leqslant a_k = ka_k k^{-1} \leqslant k_0\left(\max\limits_{1\leqslant k\leqslant k_0} a_k\right)k^{-1}$. Остается в (11) принять $B = \max\left\{(k_0+1)A^{-1};\ k_0 \max\limits_{1\leqslant k\leqslant k_0} a_k\right\}$.

Лемма 5. *Пусть числовая последовательность $\{a_k\}$ удовлетворяет условиям*

$$a_k \geqslant 0, \qquad k \in N = \{1, 2, \ldots\}; \tag{12}$$

$$a_{k+1} \leqslant a_k - \frac{a_k^2}{A} + \frac{A}{k^{2\rho}}, \qquad k \in I_0; \tag{13}$$

$$a_k \leqslant Bk^{-2\rho}, \qquad k \in I_1; \tag{14}$$

$$a_{k+1} \leqslant a_k + Ck^{-2\rho}, \qquad k \in I_1, \qquad k+1 \in I_0, \tag{15}$$

где A, B, C, ρ — положительные постоянные, $\rho \leqslant 1$, а множества индексов I_0, I_1 таковы, что $I_0 \cup I_1 = N$, $I_0 \cap I_1 = \varnothing$ (случаи $I_0 = \varnothing$ или $I_1 = \varnothing$ не исключаются). Тогда существует постоянная $D > 0$ такая, что

$$0 \leqslant a_k \leqslant Dk^{-\rho}, \quad k = 1, 2, \ldots \tag{16}$$

Доказательство. Можем считать, что $A \geqslant B + C$, так как если неравенство (13) верно для некоторого $A = A_0 > 0$, то оно верно для всех $A > A_0$. Выберем натуральное число k_0 так, чтобы

$$4 \leqslant k_0^\rho < (k_0+1)^\rho \leqslant 6. \tag{17}$$

Убедимся в том, что такое число существует. Для этого перепишем (17) в равносильном виде: $4^\varepsilon \leqslant k_0 \leqslant 6^\varepsilon - 1$, где $\varepsilon = \rho^{-1} \geqslant 1$. Существование такого числа k_0 будет доказано, если покажем, что длина отрезка $[4^\varepsilon, 6^\varepsilon - 1]$ при любом $\varepsilon \geqslant 1$ не меньше 1, т. е. $6^\varepsilon - 4^\varepsilon - 1 \geqslant 1$ или $g(\varepsilon) = 6^\varepsilon - 4^\varepsilon \geqslant 2$ при всех $\varepsilon \geqslant 1$. Но $g'(\varepsilon) = 6^\varepsilon \ln 6 - 4^\varepsilon \ln 4 \geqslant \ln 6(6^\varepsilon - 4^\varepsilon) > 0$ $(\varepsilon \geqslant 1)$, так что $g(\varepsilon)$ строго монотонно возрастает при $\varepsilon \geqslant 1$. Следовательно, $g(\varepsilon) \geqslant g(1) = 2$ для всех $\varepsilon \geqslant 1$. Таким образом, при каждом ρ $(0 < \rho \leqslant 1)$ число k_0, удовлетворяющее условиям (17), существует.

Покажем, что

$$a_{k_0+1} \leqslant 2A(k_0+1)^{-\rho}. \tag{18}$$

Может случиться, что $k_0 \in I_0$. Тогда воспользуемся неравенством (13). Заметим, что функция $f_k(a) = a - a^2A^{-1} + Ak^{-2\rho}$ достигает своего максимума на числовой оси при $a = A/2$, и поэтому $f_k(a) \leqslant f_k(A/2) = A/4 + Ak^{-2\rho}$ для всех $a \geqslant 1$, $k = 1, 2, \ldots$

Тогда из (13) с учетом неравенств (17) имеем

$$a_{k_0+1} \leqslant f_k(a_{k_0}) \leqslant A/4 + Ak_0^{-2\rho} \leqslant A/4 + A/16 \leqslant$$
$$\leqslant (5A/16)\, 6\,(k_0+1)^{-\rho} < 2A\,(k_0+1)^{-\rho}.$$

Если же $k_0 \in I_1$, то возможно и $k_0+1 \in I_0$. Тогда из (14), (17) следует, что

$$a_{k_0+1} \leqslant B\,(k_0+1)^{-2\rho} \leqslant B\,(k_0+1)^{-\rho}/4 < 2A\,(k_0+1)^{-\rho}.$$

Если $k_0 \in I_1$, но $k_0+1 \in I_0$, то из (14), (15), (17) получим

$$a_{k_0+1} \leqslant Bk_0^{-2\rho} + Ck_0^{-2\rho} \leqslant Ak_0^{-2\rho} < 2A\,(k_0+1)^{-\rho}.$$

Оценка (18) доказана. Далее, сделаем индуктивное предположение: пусть при некотором $k \geqslant k_0+1$ верна оценка $a_k \leqslant 2Ak^{-\rho}$. Возможно, что $k \in I_0$. Тогда с учетом (17) имеем $a_k \leqslant 2Ak^{-\rho} \leqslant 2Ak_0^{-\rho} \leqslant A/2$. Поскольку $f_k(a)$ монотонно возрастает на отрезке $[0, A/2]$, то из (13) следует, что $a_{k+1} \leqslant f_k(a_k) \leqslant f_k(2Ak^{-\rho}) = 2Ak^{-\rho} - 3Ak^{-2\rho} < 2A\,(k^{-\rho} - k^{-2\rho})$. Но при $0 < \rho \leqslant 1$ справедливы соотношения

$$k^{-\rho} - k^{-2\rho} < (k^\rho + 1)^{-1} < (k^\rho + k^{\rho-1})^{-1} =$$
$$= k^{-\rho+1}(k+1)^{-1} < (k+1)^{-\rho}, \qquad (19)$$

поэтому $a_{k+1} < 2A\,(k+1)^{-\rho}$. Если же $k \in I_1$ и $k+1 \in I_1$, то из (14), (17) получим

$$a_{k+1} \leqslant B\,(k+1)^{-2\rho} \leqslant B\,(k+1)^{-\rho}(k_0+1)^{-\rho} < 2A\,(k+1)^{-\rho}.$$

Если $k \in I_1$, но $k+1 \in I_0$, то из (14), (15), (17), (19) имеем

$$a_{k+1} \leqslant (B+C)\,k^{-2\rho} \leqslant Ak^{-2\rho} < A\,(k_0^0 - 1)\,k^{-2\rho} < A\,(k^\rho - 1)\,k^{-2\rho} =$$
$$= A\,(k^{-\rho} - k^{-2\rho}) < 2A\,(k+1)^{-\rho}.$$

Тем самым показано, что $a_k \leqslant 2Ak^{-\rho}$ при всех $k \geqslant k_0+1$. Если $1 \leqslant k \leqslant k_0$, то $a_k = k^\rho a_k k^{-\rho} \leqslant k_0 k^{-\rho} \max\limits_{1 \leqslant k \leqslant k_0} a_k$. Остается в (16) принять $D = \max\left\{2A;\ k_0 \max\limits_{1 \leqslant k \leqslant k_0} a_k\right\}$.

Лемма 6. *Пусть числовая последовательность* $\{\omega_k\}$ *такова, что*

$$0 \leqslant \omega_{k+1} \leqslant (1 - s_k)\,\omega_k + d_k, \quad k = 1, 2, \ldots, \quad \omega_1 \geqslant 0, \qquad (20)$$

где

$$0 < s_k \leqslant 1, \quad d_k \geqslant 0, \quad k = 1, 2, \ldots, \quad \sum_{k=1}^{\infty} s_k \to \infty,$$
$$\lim_{k\to\infty} d_k/s_k = 0. \qquad (21)$$

Тогда $\lim\limits_{k\to\infty} \omega_k = 0$.

Доказательство. Поскольку $1-x\leqslant e^{-x}$ при $0\leqslant x\leqslant 1$, то $1-s_k\leqslant e^{-s_k}$. Из неравенства (20) тогда имеем $0\leqslant\omega_{k+1}\leqslant\leqslant\omega_k e^{-s_k}+d_k$ $(k=1, 2, \ldots)$. Отсюда с помощью индукции нетрудно получить, что

$$0\leqslant\omega_{k+1}\leqslant\left(\omega_1+\sum_{i=1}^{n}d_i\exp\left\{\sum_{j=1}^{i}s_j\right\}\right)\exp\left\{-\sum_{j=1}^{k}s_j\right\},\qquad k=1, 2, \ldots \tag{22}$$

Далее воспользуемся известной теоремой Штольца ([165, ч. I, с. 88]), представляющей собой разностный аналог правила Лопиталя и гласящей, что если последовательность $\{y_k\}$ монотонно возрастает, предел $\lim\limits_{k\to\infty}(x_k-x_{k-1})/(y_k-y_{k-1})$ существует, $\lim\limits_{k\to\infty}y_k=\infty$, то существует и предел $\lim\limits_{k\to\infty}x_k/y_k$, причем

$$\lim_{k\to\infty}x_k/y_k=\lim_{k\to\infty}(x_k-x_{k-1})/(y_k-y_{k-1}).$$

Положим $y_k=\exp\left\{\sum\limits_{j=1}^{k}s_j\right\}$, $x_k=\omega_1+\sum\limits_{i=1}^{k}d_i\exp\left\{\sum\limits_{j=1}^{i}s_j\right\}$ $(k=1, 2, \ldots)$. Из условий (21) следует, что $\{y_k\}$ монотонно возрастает и стремится к бесконечности. Кроме того,

$$\lim_{k\to\infty}\frac{x_k-x_{k-1}}{y_k-y_{k-1}}=\lim_{k\to\infty}d_k\exp\left\{\sum_{j=1}^{k}s_j\right\}\left(\exp\left\{\sum_{j=1}^{k}s_j\right\}-\exp\left\{\sum_{j=1}^{k-1}s_j\right\}\right)^{-1}=$$
$$=\lim_{k\to\infty}\frac{d_k}{1-e^{-s_k}}=\lim_{k\to\infty}\left(\frac{d_k}{s_k}\right)\frac{s_k}{1-e^{-s_k}}=0,$$

так как функция $x/(1-e^{-x})$ ограничена на множестве $0<x\leqslant\leqslant 1$. По теореме Штольца с учетом неравенства (22) получим

$$\lim_{k\to\infty}\omega_k=\lim_{k\to\infty}x_k/y_k=\lim_{k\to\infty}(x_k-x_{k-1})/(y_k-y_{k-1})=0.$$

Заметим, что неравенство (22) по существу представляет собой оценку скорости сходимости последовательности $\{\omega_k\}$. Однако правая часть оценки (22) трудно обозрима. Поэтому полезно иметь другие, быть может, более грубые, но более обозримые оценки. Здесь может быть полезна следующая простая

Лемма 7. *Пусть числовая последовательность* $\{\omega_k\}$ *такова, что*

$$0\leqslant\omega_{k+1}\leqslant(1-s_k)\omega_k+d_k,\quad k=1, 2, \ldots,\quad \omega_1\geqslant 0, \tag{23}$$

где

$$0\leqslant s_k\leqslant 1,\quad d_k\geqslant 0,\quad k=1, 2, \ldots,\ \sup_{k\geqslant 1}d_k/s_k=c<\infty. \tag{24}$$

Тогда

$$0\leqslant\omega_k\leqslant\omega_1+c,\quad k=1, 2, \ldots \tag{25}$$

Доказательство легко проводится по индукции. При $k=1$ оценка (25) очевидна. Если (25) верно для некоторого $k \geqslant 1$, то из (23), (24) следует $0 \leqslant \omega_{k+1} \leqslant (1-s_k)(\omega_1+c)+d_k \leqslant \leqslant (1-s_k)\omega_1+(1-s_k)c+cs_k \leqslant \omega_1+c$, что и требовалось.

Покажем, как может быть применена лемма 7 для оценки конкретных последовательностей.

Лемма 8. *Пусть числовая последовательность* $\{a_k\}$ *такова, что*

$$0 \leqslant a_{k+1} \leqslant (1-1/k)a_k + c_1/k^2, \quad k=1, 2, \ldots, \qquad c_1 = \text{const} > 0, \quad a_1 \geqslant 0. \tag{26}$$

Тогда справедлива оценка

$$0 \leqslant a_k \leqslant c_2 \ln(k+1)/k, \quad k=1, 2, \ldots, \quad c_2 = \text{const} > 0. \tag{27}$$

Доказательство. Сделаем замену $\omega_k = a_k k(\ln(k+1))^{-1}$ и, пользуясь леммой 7, докажем ограниченность $\{\omega_k\}$. Из (26) имеем

$$0 \leqslant \omega_{k+1} \leqslant \left(1-\frac{1}{k}\right)\frac{k+1}{k}\frac{\ln(k+1)}{\ln(k+2)}\omega_k + c_1\frac{k+1}{k^2\ln(k+2)}, \qquad k=1, 2, \ldots$$

Таким образом, $\{\omega_k\}$ удовлетворяет условиям (23) при

$$s_k = 1-\left(1-\frac{1}{k^2}\right)\frac{\ln(k+1)}{\ln(k+2)}, \quad d_k = c_1\frac{k+1}{k^2\ln(k+2)}.$$

Нетрудно видеть, что $0 < s_k < 1$ и $\lim\limits_{k\to\infty} d_k/s_k = c_1$, так что $\sup\limits_{k\geqslant 1} d_k/s_k = c_3 < \infty$. Из леммы 7 имеем $0 \leqslant \omega_k \leqslant \omega_1 + c_3$ $(k=1, 2, \ldots)$, что равносильно оценке (27) с $c_2 = c_3 + a_1/\ln 2$.

Лемма 9. *Пусть числовая последовательность* $\{a_k\}$ *такова, что*

$$0 \leqslant a_{k+1} \leqslant (1-1/k^\beta)a_k + c_1/k^{2\beta}, \quad k=1, 2, \ldots, \qquad c_1 = \text{const} > 0, \quad 0 < \beta < 1, \quad a_1 \geqslant 0. \tag{28}$$

Тогда

$$0 \leqslant a_k \leqslant \left(a_1 + \frac{c_1}{1-\beta}\right)\frac{1}{k^\beta}, \quad k=1, 2, \ldots \tag{29}$$

Доказательство. Сделаем замену $\omega_k = k^\beta a_k$. Тогда из (28) имеем

$$0 \leqslant \omega_{k+1} \leqslant \left(1-\frac{1}{k^\beta}\right)\frac{(k+1)^\beta}{k^\beta}\omega_k + c_1\frac{(k+1)^\beta}{k^{2\beta}}.$$

Это значит, что $\{\omega_k\}$ удовлетворяет условиям (23) при

$$s_k = 1-\left(1-\frac{1}{k^\beta}\right)\left(1+\frac{1}{k}\right)^\beta < 1,$$

$$d_k = c_1\left(1+\frac{1}{k}\right)^\beta\frac{1}{k^\beta}, \quad k=1, 2, \ldots$$

Поскольку $(1+1/k)^{-\beta} \geqslant 1-\beta/k \quad (k=1, 2, \ldots)$, то

$$s_k = \left(1+\frac{1}{k}\right)^{\beta}\left[\left(1+\frac{1}{k}\right)^{-\beta} - 1 + \frac{1}{k^{\beta}}\right] \geqslant$$

$$\geqslant \left(1+\frac{1}{k}\right)^{\beta}\left[1-\frac{\beta}{k} - 1 + \frac{1}{k^{\beta}}\right] =$$

$$= \left(1+\frac{1}{k}\right)^{\beta}\frac{1}{k^{\beta}}\left(1-\frac{\beta}{k^{1-\beta}}\right) \geqslant d_k\frac{1}{c_1}(1-\beta) > 0, \quad k=1, 2, \ldots;$$

отсюда же следует, что $d_k/s_k \geqslant c_1/(1-\beta) \quad (k=1, 2, \ldots)$. По лемме 7 тогда $0 \leqslant \omega_k \leqslant \omega_1 + c_1/(1-\beta)$, что равносильно оценке (29).

Лемма 10. *Пусть $\{z_k\}$, $\{w_k\}$ — некоторые последовательности из евклидова пространства E^n, z_* — точка из E^n такая, что*

$$a|w_{k+1}-z_k|^2 + |w_{k+1}-z_*|^2 \leqslant |z_k - z_*|^2, \tag{30}$$

$$|z_{k+1}-w_{k+1}| \leqslant b\delta_k, \quad \delta_k \geqslant 0, \quad k=0, 1, \ldots, \quad \sum_{k=0}^{\infty}\delta_k < \infty, \tag{31}$$

где a, b — положительные постоянные. Тогда существует конечный предел

$$\lim_{k\to\infty}|z_k - z_*| = \lim_{k\to\infty}|w_k - z_*| \tag{32}$$

и справедливо равенство

$$\lim_{k\to\infty}|w_{k+1}-z_k| = 0. \tag{33}$$

Если, кроме того, точка z_* из (30) является предельной для $\{z_k\}$, то обе последовательности $\{z_k\}$, $\{w_k\}$ сходятся к z_*.

Доказательство. Из (30) следует, что $|w_{k+1}-z_*| \leqslant |z_k - z_*|$. Тогда с помощью (31) имеем

$$|z_{k+1}-z_*| \leqslant |z_{k+1}-w_{k+1}| + |w_{k+1}-z_*| \leqslant b\delta_k + |z_k - z_*|, \tag{34}$$

или

$$|z_{k+1}-z_*| \leqslant |z^k - z_*| + b\delta_k, \quad k=0, 1, \ldots$$

Отсюда и из леммы 2 при $a_k = |z_k - z_*|$ вытекает существование конечного предела $\lim\limits_{k\to\infty}|z_k - z_*|$. Из (34) следует (32), что в свою очередь гарантирует выполнение равенства (33). Наконец, пусть в (30) z_* — предельная точка $\{z_k\}$, пусть $\{z_{k_m}\} \to z_*$. Тогда $\lim\limits_{k\to\infty}|z_k - z_*| = \lim\limits_{m\to\infty}|z_{k_m} - z_*| = 0$, т. е. $\{z_k\} \to z_*$. Отсюда и из (32) получим, что $\{\omega_k\} \to z_*$.

Лемма 11. *Пусть неотрицательное число z таково, что*

$$0 \leqslant z^p \leqslant bz + d, \tag{35}$$

где b, d — неотрицательные числа, $p > 1$. Тогда

$$0 \leqslant z \leqslant (b^q + qd)^{1/p}, \tag{36}$$

где q определяется равенством $p^{-1} + q^{-1} = 1$.

Доказательство. Если $d = 0$, то из (35) имеем $0 \leqslant z^{p-1} \leqslant b$, или $0 \leqslant z \leqslant b^{1/(p-1)} = b^{q/p}$, что совпадает с оценкой (36) при $d = 0$. Поэтому можем считать, что $d > 0$. Рассмотрим функцию

$$\varphi(x) = x^p - bx - d, \qquad x \geqslant 0.$$

Нетрудно проверить, что эта функция имеет единственную точку минимума $x_* (b/p)^{1/(p-1)} \geqslant 0$, $\varphi(x_*) \leqslant \varphi(0) = -d < 0$, $\lim\limits_{x\to\infty} \varphi(x) = \infty$, строго монотонно убывает при $0 \leqslant x \leqslant x_*$ (если $x_* > 0$), строго монотонно возрастает при $x \geqslant x_*$. Отсюда следует, что уравнение $\varphi(x) = 0$ имеет единственное решение $x = \gamma$, так что $\varphi(x) < 0$ при $0 \leqslant x < \gamma$ и $\varphi(x) > 0$ при $x > \gamma$. Согласно (35) $\varphi(z) \leqslant 0$, поэтому справедлива оценка $0 \leqslant z \leqslant \gamma$. Однако получить явное выражение для γ в общем случае не удается, поэтому в приложениях удобнее оценка (36). Для доказательства оценки (36) достаточно установить, что $\gamma \leqslant a = (b^q + qd)^{1/p}$. Пользуясь известным неравенством [10, 160, 179]

$$|ab| \leqslant |a|^p/p + |b|^q/q,$$

справедливым для всех действительных чисел a, b, $p > 1$, $q > 1$, $p^{-1} + q^{-1} = 1$, имеем $\varphi(a) = a^p - ab - d \geqslant a^p - a^p p^{-1} - b^q q^{-1} - d = a^{pq-1} - (b^q + qd) q^{-1} = 0$. Следовательно, $a \geqslant \gamma$ и $0 \leqslant z \leqslant a = (b^q + qd)^{1/p}$.

Глава 3

ЭЛЕМЕНТЫ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Изучение методов минимизации функций многих переменных начнем с методов решения сравнительно простых и достаточно хорошо изученных задач линейного программирования. Под линейным программированием понимается раздел теории экстремальных задач, в котором изучаются задачи минимизации (или максимизации) линейных функций на множествах, задаваемых системами линейных равенств и неравенств. Различные аспекты теории и методов линейного программирования, его приложения к технико-экономическим задачам изложены, например в [3, 7, 11, 12, 15, 21—23, 25, 30, 33, 35, 40, 41, 48, 57, 71, 79, 102, 106, 130, 144, 146, 150, 155, 183, 190, 194, 202, 211, 224, 225, 234, 250, 261, 265, 274, 290, 292, 297, 307, 313, 320—324, 333, 340].

§ 1. Постановка задачи

1. *Общая задача линейного программирования* может быть сформулирована следующим образом: минимизировать функцию

$$J(u) = c_1u^1 + \ldots + c_nu^n \tag{1}$$

при условиях

$$u^k \geqslant 0, \quad k \in I; \tag{2}$$

$$\begin{gathered} a_{11}u^1 + \ldots + a_{1n}u^n \leqslant b^1, \\ \cdots\cdots\cdots\cdots \\ a_{m1}u^1 + \ldots + a_{mn}u^n \leqslant b^m, \\ a_{m+1,\,1}u^1 + \ldots + a_{m+1,\,n}u^n = b^{m+1}, \\ \cdots\cdots\cdots\cdots \\ a_{s1}u^1 + \ldots + a_{sn}u^n = b^s, \end{gathered} \tag{3}$$

где c_j, a_{ij}, b^i $(i = 1, \ldots, s,\ j = 1, \ldots, n)$ — заданные числа, причем не все из чисел c_j и a_{ij} равны нулю; I — заданное подмножество индексов из множества $\{1, \ldots, n\}$ (в частности, здесь возможно, что $I = \varnothing$ или $I = \{1, \ldots, n\}$; в (3) не исключаются случаи, когда отсутствуют ограничения типа равенств $(m = s)$ или типа неравенств $(m = 0)$). Если ввести векторы $c = (c_1, \ldots, c_n)$, $a_i = (a_{i1}, \ldots, a_{in})$, $u = (u^1, \ldots, u^n)$, то задачу (1)—(3) можно кратко записать так:

$$J(u) = \langle c, u\rangle \to \inf;\ u \in U = \{u \in E^n:\ u^k \geqslant 0,\ k \in I;\ \langle a_i, u\rangle \leqslant b^i,\ i = 1, \ldots, m;\ \langle a_i, u\rangle = b^i,\ i = m+1, \ldots, s\}. \tag{4}$$

Если для каких-либо двух векторов $x=(x^1, \ldots, x^p)$, $y=(y^1, \ldots, y^p)$ справедливы неравенства $x^i \geqslant y^i$ при всех $i=1, \ldots, p$, то будем кратко писать: $x \geqslant y$. Тогда, например, неравенство $x \geqslant 0$ означает, что $x^i \geqslant 0$ для всех $i=1, \ldots, p$. Используя принятые обозначения, задачу (1)—(3), или (4), можно записать и в таком виде:

$$\begin{gathered} J(u)=\langle c, u\rangle \to \inf; \\ u \in U=\{u \in E^n\colon\ u^k \geqslant 0,\ k \in I,\ Au \leqslant b,\ \bar{A}u=\bar{b}\}, \end{gathered} \tag{5}$$

где

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad \bar{A}=\begin{bmatrix} a_{m+1,1} & \cdots & a_{m+1,n} \\ \cdot & \cdots & \cdot \\ a_{s1} & \cdots & a_{sn} \end{bmatrix},$$

$$b=\begin{bmatrix} b^1 \\ \cdot \\ b^m \end{bmatrix}, \quad \bar{b}=\begin{bmatrix} b^{m+1} \\ \cdot \\ b^s \end{bmatrix}.$$

Точку $u_* \in U$ назовем *точкой минимума* функции $\langle c, u\rangle$ на множестве U или, короче, *решением* задачи (4) или (5), если $\langle c, u_*\rangle = \inf\limits_U \langle c, u\rangle$.

2. Приведем примеры прикладных задач, приводящих к задачам линейного программирования.

Задача оптимального планирования производства. Пусть на некотором предприятии изготовляются n видов продукции из s видов сырья. Известно, что на изготовление одной единицы продукции j-го вида нужно a_{ij} единиц сырья i-го вида. В распоряжении предприятия имеется b_i единиц сырья i-го вида. Известно также, что на каждой единице продукции j-го вида предприятие получает c_j единиц прибыли. Требуется определить, сколько единиц $u^1, \ldots, u^n$ каждого вида продукции должно изготовить предприятие, чтобы обеспечить себе максимальную прибыль.

Если предприятие наметит себе план производства $u=\{u^1, \ldots, u^n\}$, то оно израсходует $a_{i1}u^1+\ldots+a_{in}u^n$ единиц сырья i-го вида и получит $c_1u^1+\ldots+c_nu^n$ единиц прибыли. Ясно также, что все величины u^i $(i=1, \ldots, n)$ неотрицательны. Поэтому мы приходим к следующей задаче линейного программирования: максимизировать функцию $J(u)=c_1u^1+\ldots+c_nu^n$ при ограничениях $u^1 \geqslant 0, \ldots, u^n \geqslant 0$, $a_{i1}u^1+\ldots+a_{in}u^n \leqslant b^i$ $(i=1, \ldots, s)$. Поскольку задача максимизации функции $J(u)$ равносильна задаче минимизации функции $-J(u)$, то с учетом введенных выше обозначений сформулированную задачу линейного программирования можно кратко записать в виде

$$\langle -c, u\rangle \to \inf;\ u \in U=\{u \in E^n\colon\ u \geqslant 0,\ Au \leqslant b\}. \tag{6}$$

Ясно, что задача (6) является частным случаем задачи (5).

Задача об оптимальном использовании посевной площади. Пусть под посев p культур отведено r земельных участков пло-

щадью соответственно в $b_1, \ldots, b_r$ гектаров. Известно, что средняя урожайность i-й культуры на j-м участке составляет a_{ij} центнеров с гектара, а прибыль за один центнер i-й культуры составляет c_i рублей. Требуется определить, какую площадь на каждом участке следует отвести под каждую из культур, чтобы получить максимальную прибыль, если по плану должно быть собрано не менее d_i центнеров i-й культуры.

Обозначим через u_{ij} площадь, которую планируется отвести под i-ю культуру на j-м участке. Тогда

$$u_{1j} + \ldots + u_{pj} = b_j, \quad j = 1, \ldots, r. \tag{7}$$

Ожидаемый средний урожай i-й культуры со всех участков равен $a_{i1}u_{i1} + \ldots + a_{in}u_{in}$ центнеров. Поскольку согласно плану должно быть произведено не менее d_i центнеров i-й культуры, то

$$a_{i1}u_{i1} + \ldots + a_{in}u_{ir} \geqslant d_i, \quad i = 1, \ldots, p. \tag{8}$$

Ожидаемая прибыль за урожай i-й культуры равна $c_i(a_{i1}u_{i1} + \ldots + a_{in}u_{in})$, а за урожай всех культур —

$$\sum_{i=1}^{p} c_i (a_{i1}u_{i1} + \ldots + a_{in}u_{ir}) = J(u). \tag{9}$$

Таким образом, приходим к задаче максимизации функции (9) (или минимизации функции $-J(u)$) при условиях (7), (8) и естественных ограничениях

$$u_{ij} \geqslant 0, \quad i = 1, \ldots, p, \quad j = 1, \ldots, r.$$

Если умножить соотношения (8) на -1 и переменные $\{u_{ij}\}$ переобозначить через $u^1, \ldots, u^n$, то придем к задаче вида (1)—(3).

Транспортная задача. Пусть имеется r карьеров, где добывается песок, и p потребителей песка (например, кирпичные заводы). В i-м карьере ежесуточно добывается a_i тонн песка, а j-му потребителю ежесуточно требуется b_j тонн песка. Пусть c_{ij} — стоимость перевозки одной тонны песка с i-го карьера j-му потребителю. Требуется составить план перевозок песка так, чтобы общая стоимость перевозок была минимальной.

Обозначим через u_{ij} планируемое количество тонн песка из i-го карьера j-му потребителю. Тогда с i-го карьера будет вывезено

$$u_{i1} + \ldots + u_{ip} = a_i, \quad i = 1, \ldots, r, \tag{10}$$

тонн песка, j-му потребителю доставлено

$$u_{1j} + \ldots + u_{rj} = b_j, \quad j = 1, \ldots, p, \tag{11}$$

тонн песка, а стоимость перевозок будет равна

$$J(u) = \sum_{j=1}^{p} \sum_{i=1}^{r} c_{ij}u_{ij}. \tag{12}$$

Естественно требовать, чтобы

$$u_{ij} \geqslant 0, \quad i=1, \ldots, r, \quad j=1, \ldots, p. \tag{13}$$

В результате получили задачу минимизации функции (12) при условиях (10), (11), (13), которая, очевидно, является частным случаем общей задачи линейного программирования (1)—(3).

К задачам типа (1)—(3) сводятся также и многие другие прикладные задачи технико-экономического содержания.

Следует заметить, что приведенные выше примеры задач линейного программирования, вообще говоря, представляют лишь приближенную, упрощенную математическую модель реальных задач, и некритическое использование получаемых на основе анализа этих моделей результатов может привести иногда к парадоксам. К чему может привести пренебрежение существенными факторами реальных задач при составлении математических моделей, хорошо иллюстрируют следующие строки из брошюры Б. В. Гнеденко [114, с. 4]: «В пятидесятых годах, когда методы линейного программирования только начали входить в широкий обиход, наши центральные газеты обошла статья, в которой сообщалось, что перераспределение снабжения строек Москвы песком путем прикрепления их к пристаням и карьерам позволило снизить каждодневные транспортные расходы на 20 000 рублей. Действительно, этим приемом удалось добиться минимального суммарного пути и, значит, минимального расхода топлива. Однако уже через несколько лет мне на глаза попалась небольшая статья из газеты «Труд», в которой сообщалось, что недоучет пропускной способности карьеров, отсутствие подъездных путей и погрузочных устройств приводят к огромным простоям грузовых машин на местах погрузки песка. Таким образом, мы видим, что недостаточно только оптимизировать расход бензина. Минимизации требует вся операция в целом».

Эти критические замечания можно, конечно, отнести не только к задачам линейного программирования, но и к другим математическим моделям. Вполне может оказаться, что принятая математическая модель, обычно составляемая на основе приближенных данных о реальном моделируемом явлении (объекте, процессе), не охватывает какие-либо важные существенные стороны исследуемого явления и приводит к результатам, существенно расходящимся с реальностью. В этом случае математическая модель должна быть изменена, доработана с учетом вновь поступившей информации, а получаемые при анализе совершенствованной модели данные должны снова и снова критически сопоставляться с реальными данными и использоваться для выяснения границ применимости модели. Математическая модель лишь при высокой степени адекватности моделируемому явлению может быть использована для более глубокого анализа явления

и проникновения в его сущность, для выработки целенаправленного управления.

Было бы ошибочным, основываясь на приведенных выше критических строках, делать вывод о том, что линейные модели, приводящие к задачам вида (1)—(3), слишком просты и вовсе непригодны для исследования реальных задач. Наоборот, практика показала, что линейное программирование может быть успешно применено для исследования и анализа широкого класса реальных технико-экономических задач. Линейное программирование является одним из наиболее изученных разделов теории экстремальных задач с достаточно богатым арсеналом методов. Ниже мы увидим, что задачи линейного программирования используются также и в качестве вспомогательных во многих методах решения более сложных нелинейных задач минимизации.

3. Из общей задачи линейного программирования обычно выделяют и исследуют два ее подкласса: *каноническую задачу*

$$\langle c,\ u\rangle \to \inf;\quad u\in U=\{u\in E^n:\ u\geqslant 0,\ Au=b\} \tag{14}$$

и *основную задачу*

$$\langle c,\ u\rangle \to \inf;\quad u\in U=\{u\in E^n:\ u\geqslant 0,\ Au\leqslant b\}. \tag{15}$$

Здесь c, b — заданные векторы, $c\in E^n$, $c\neq 0$, $b\in E^m$, A — матрица размера $m\times n$, $A\neq 0$. Несмотря на внешнее различие задач (14) и (15) (в одной из них ограничения $Au=b$, в другой $Au\leqslant b$), эти задачи, оказывается, в определенном смысле эквивалентны. В самом деле, если ограничения типа равенств $Au=b$ заменить на равносильную систему двух неравенств: $Au\leqslant b$, $Au\geqslant b$, то каноническую задачу (14) можно записать в виде основной задачи

$$\langle c,\ u\rangle \to \inf;\ u\in U=\{u\in E^n:\ u\geqslant 0,\ Au\leqslant b,\ -Au\leqslant -b\}. \tag{16}$$

Ясно, что задачи (14) и (16) эквивалентны, т. е. всякое решение задачи (14) является решением задачи (16) и наоборот (или, возможно, обе задачи не имеют решения).

Основную задачу линейного программирования (15) также можно записать в виде канонической задачи. А именно, введем дополнительные переменные $v=(v^1,\ \ldots,\ v^m)$ посредством соотношений $v=b-Au$ $(v\geqslant 0)$ и в пространстве E^{n+m} переменных $z=(u,\ v)=(u^1,\ \ldots,\ u^n,\ v^1,\ \ldots,\ v^m)$ рассмотрим каноническую задачу

$$\langle d,\ z\rangle \to \inf;\ z\in Z=\{z=(u,\ v):\ z\in E^{n+m},\ z\geqslant 0,\ Cz=Au+v=b\}, \tag{17}$$

где $d=(c,\ 0)\in E^{n+m}$, $C=(A,\ E)$, E — единичная матрица размера $m\times m$. Нетрудно видеть, что если u_* — решение задачи (15), т. е. $u_*\in U$, $\langle c, u_*\rangle=\inf\limits_U\langle c,u\rangle$, то $z_*=(u_*,\ v_*)$, $v_*=b-Au_*$ — ре-

шение задачи (17), т. е. $z_* \in Z$, $\langle d, z_*\rangle = \inf\limits_U \langle d, z\rangle$, и обратно, если $z_* = (u_*, v_*)$ — решение задачи (17), то u_* — решение задачи (15). Таким образом, путем введения новых переменных или увеличением числа ограничений всякую задачу вида (15) можно привести к виду (14), а задачу (14) — к задаче вида (15).

Оказывается, общая задача линейного программирования (5) также может быть записана в виде канонической (или основной) задачи линейного программирования. В самом деле, положим

$$v = b - Au,\ w^i = \max\{0,\ u^i\},\ \overline{w}^i = \max\{0;\ -u^i\},\ i \notin I, \quad (18)$$

и в пространстве переменных $z = (v,\ u^i,\ i \in I;\ w^i,\ \overline{w}^i,\ i \notin I)$ рассмотрим задачу

$$J(z) = \sum_{i \in I} c_i u^i + \sum_{i \notin I} c_i (w^i - \overline{w}^i) \to \inf;$$

$$z \in Z = \Big\{z\colon\ v \geqslant 0,\quad u^i \geqslant 0,\quad i \in I;\quad w^i \geqslant 0,\quad \overline{w}^i \geqslant 0,\quad i \notin I;$$
$$\sum_{i \in I} A_i u^i + \sum_{i \notin I} A_i (w^i - \overline{w}^i) + v = b,\ \sum_{i \in I} \overline{A}_i u^i + \sum_{i \notin I} \overline{A}_i (w^i - \overline{w}^i) = \overline{b}\Big\}, \quad (19)$$

где A_i — i-й столбец матрицы A, $\overline{A}_i$ — i-й столбец $\overline{A}$. Как видим, в задаче (19) функция $J(z)$ линейна, все переменные неотрицательны, остальные ограничения имеют только вид равенств, т. е. (19) является канонической задачей линейного программирования. Учитывая обозначения (18) и равенство $u^i = w^i - \overline{w}^i$ $(i \neq 0)$, остается лишь заметить, что точка $z_* = (v_*, u^i_*, i \in I;\ w^i_*, \overline{w}^i_*, i \notin I)$ будет решением задачи (19) тогда и только тогда, когда точка u_* с координатами u^i_* $(i \in I)$, $u^i_* = w^i_* - \overline{w}^i_*$ $(i \notin I)$ будет решением задачи (5).

Заметим, что изложенные приемы сведения задач (5), (14), (15) к канонической или основной задаче линейного программирования на практике без особой необходимости не применяют, так как это может привести к чрезмерному увеличению размерности переменных или числа ограничений. Поэтому методы решения задач линейного программирования обычно разрабатывают для задачи (14) или (15), а затем, учитывая указанную выше связь между задачами (5), (14), (15), модифицируют полученные методы применительно к другим классам задач линейного программирования.

§ 2. Геометрическая интерпретация. Угловые точки

1. Кратко остановимся на геометрическом смысле задачи линейного программирования. Рассмотрим задачу (1.15) при $n = 2$. Для краткости обозначим $u^1 = x$, $u^2 = y$, $u = (x, y)$ и перепишем

задачу (1.15) в виде

$$c_1x + c_2y \to \inf;$$
$$u \in U = \{u = (x,\ y):\ x \geqslant 0,\ y \geqslant 0,\ a_{i1}x + a_{i2}y \leqslant b^i,\ i = 1,\ \ldots,\ m\}. \tag{1}$$

Введем множества: $U_0 = \{u = (x,\ y):\ x \geqslant 0,\ y \geqslant 0\}$ — положительный квадрант плоскости $(x,\ y)$, $U_i = \{u = (x,\ y):\ a_{i1}x + a_{i2}y \leqslant b^i\}$ — полуплоскость, образуемая прямой $a_{i1}x + a_{i2}y = b^i$ $(i = 1,\ \ldots,\ m)$. Ясно, что множество U является пересечением множеств $U_0,\ U_1,\ \ldots,\ U_m$. Может случиться, что это пересечение пусто (рис. 3.1) — тогда задача (1) теряет смысл. Если множество U непусто, то оно, образованное пересечением конечного числа полуплоскостей, представляет собой выпуклое многоугольное множество, границей которого является ломаная,

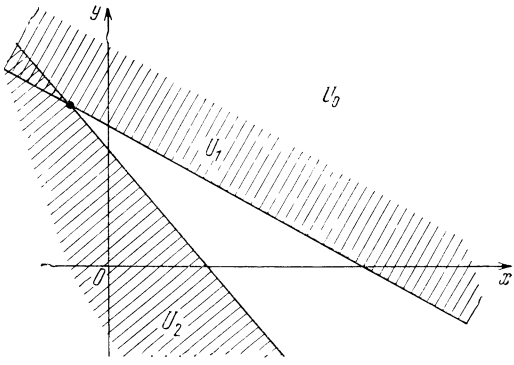

Рис. 3.1

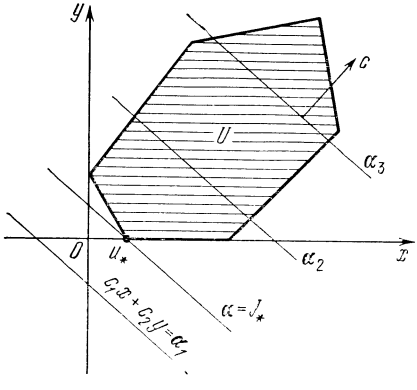

Рис. 3.2

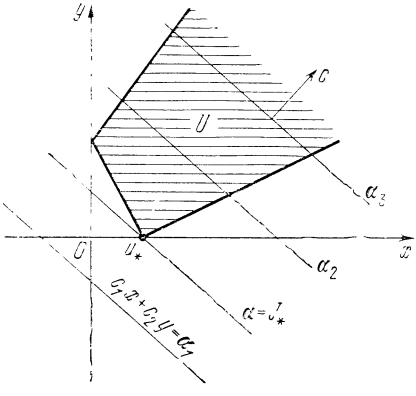

Рис. 3.3

составленная из отрезков каких-либо координатных осей и прямых $a_{i1}x + a_{i2}y = b^i$ $(i = 1,\ \ldots,\ m)$. Это многоугольное множество может быть как ограниченным (рис. 3.2) — тогда U представляет собой выпуклый многоугольник, так и неограниченным (рис. 3.3).

Пусть α — какое-либо значение функции $J(u)=\langle c, u\rangle = c_1x + c_2y$. Тогда уравнение

$$c_1x + c_2y = \alpha \tag{2}$$

задает линию уровня функции $J(u)$, соответствующую ее значению α, и на плоскости определяет прямую, перпендикулярную вектору $c=(c_1, c_2)\neq 0$. При изменении α от $-\infty$ до ∞ прямая (2), смещаясь параллельно самой себе, «зачертит» («заметет») всю плоскость. При этом вектор c указывает направление, в котором следует смещать прямую (2), чтобы увеличивать значение функции $J(u)=\langle c, u\rangle$. Если U — многоугольник (см. рис. 3.2), то при изменении α от $-\infty$ до ∞ прямая (2) при некотором значении $\alpha = J_*$ впервые коснется U и будет иметь с U общую точку u_* (на рис. 3.2—3.5 прямая (2) представлена при $\alpha=\alpha_1<J_*<\alpha_2<\alpha_3$). Ясно, что $\langle c, u_*\rangle = J_* = \inf\limits_U \langle c, u\rangle$, т. е. u_* — решение задачи (1). Возможен случай, когда прямая (2) при первом касании с многоугольником U будет иметь не одну общую точку

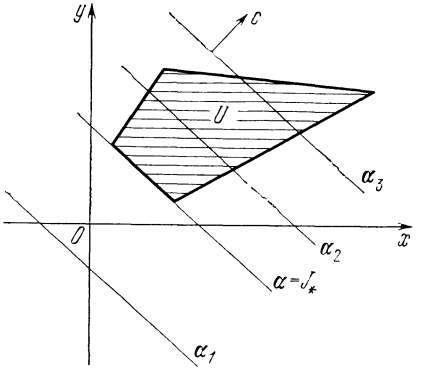

Рис. 3.4

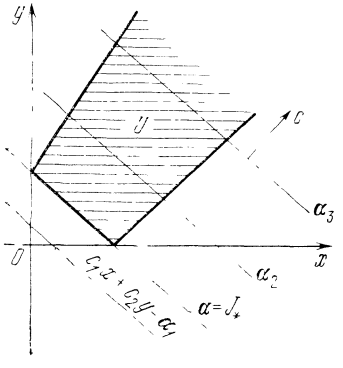

Рис. 3.5

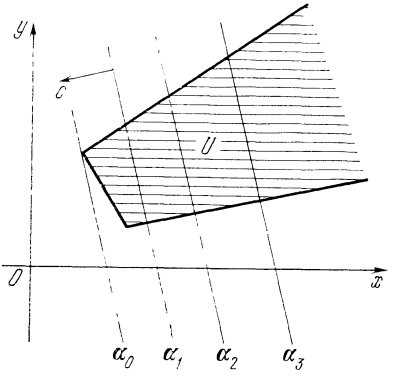

Рис. 3.6

с U, а целую сторону (рис. 3.4) — это может случиться, если U имеет сторону, перпендикулярную вектору c.

Если многоугольное множество U не ограничено, то наряду со случаями, когда при первом касании прямая (2) будет иметь с U одну общую точку u_* (см. рис. 3.3) или сторону (рис. 3.5),

возможна ситуация, когда прямая (2) при всех α $(-\infty < \alpha \leqslant \alpha_0 \leqslant \infty)$ имеет общую точку с U (рис. 3.6) — тогда $\inf\limits_U (c, u) = -\infty$ (первого касания прямой (2) с U нет), т. е. задача (1) не имеет решения.

Из рассмотренных случаев задачи (1) видно, что задача линейного программирования может не иметь ни одного решения (см. рис. 3.1, 3.6), может иметь лишь одно решение (см. рис. 3.2, 3.3), может иметь бесконечно много решений (см. рис. 3.4, 3.5).

Аналогично можно показать, что множество U в задаче (1.15) при $n = 3$ является многогранным множеством, и дать геометрическую интерпретацию этой задачи. Предлагаем читателю самостоятельно рассмотреть этот случай, а также исследовать задачу (1.14) при $n = 2, 3$.

2. На примере рассмотренной выше задачи (1) нетрудно усмотреть, что если задача (1) имеет решение, то среди решений найдется хотя бы одна угловая точка (вершина) многоугольного множества U. Ниже мы увидим, что это не случайно: и в более общей задаче линейного программирования, оказывается, нижняя грань функции $\langle c, u\rangle$ на U достигается в угловой точке множества.

Определение 1. Точка v множества U называется *угловой точкой* (вершиной, крайней точкой, экстремальной точкой) множества U, если представление $v = \alpha v_1 + (1-\alpha) v_2$ при $v_1, v_2 \in U$ и $0 < \alpha < 1$ возможно лишь при $v_1 = v_2$. Иначе говоря, v — *угловая точка* множества U, если она не является внутренней точкой никакого отрезка, принадлежащего множеству U.

Например, угловыми точками многоугольника на плоскости или параллелепипеда в пространстве являются их вершины; все граничные точки шара будут его угловыми точками; замкнутое полупространство или пересечение двух замкнутых полупространств не имеют ни одной угловой точки.

В задачах линейного программирования понятие угловой точки играет фундаментальную роль и лежит в основе многих методов решения таких задач. В дальнейшем мы будем подробно исследовать каноническую задачу (1.14). Поэтому начнем с изучения свойств угловых точек множества

$$U = \{u \in E^n:\ u \geqslant 0,\ Au = b\}, \tag{3}$$

где A — матрица размера $m \times n$, $A \neq 0$, b — вектор из E^m. Ниже будет показано, что множество (3), если оно непусто, имеет хотя бы одну угловую точку (см. теорему 5.1). Возникает вопрос, как узнать, будет ли та или иная точка множества (3) угловой точкой? Приведем один достаточно простой алгебраический критерий угловой точки множества (3). Для этого вначале обозначим j-й столбец матрицы A через A_j и запишем систему уравнений

$Au = b$ в следующей эквивалентной форме:

$$A_1u^1 + \ldots + A_nu^n = b. \tag{4}$$

Теорема 1. *Пусть множество U определено условиями* (3), $A \neq 0$, $r = \operatorname{rang} A$ — *ранг матрицы A. Для того чтобы точка $v = (v^1, \ldots, v^n) \in U$ была угловой точкой множества U, необходимо и достаточно, чтобы существовали номера $j_1, \ldots, j_r$ $(1 \leqslant j_l \leqslant n$, $l = 1, \ldots, r)$ такие, что*

$$A_{j_1}v^{j_1} + \ldots + A_{j_r}v^{j_r} = b; \quad v^j = 0, \quad j \neq j_l, \quad l = 1, \ldots, r, \tag{5}$$

причем столбцы $A_{j_1}, \ldots, A_{j_r}$ линейно независимы.

Доказательство. Необходимость. Пусть v — угловая точка множества U. Если $v = 0$, то из условия $0 \in U$ следует, что $b = 0$. Поскольку $A \neq 0$, то $r = \operatorname{rang} A \geqslant 1$ и существуют линейно независимые столбцы $A_{j_1}, \ldots, A_{j_r}$. Отсюда имеем $A_{j_1} \cdot 0 + \ldots + A_{j_r} \cdot 0 = 0$. Для случая $v = 0$ соотношения (5) доказаны.

Пусть теперь $v \neq 0$ и пусть $v^{j_1}, \ldots, v^{j_k}$ — все положительные координаты точки v. Отсюда и из условия $Av = b$ с учетом представления (4) имеем

$$A_{j_1}v^{j_1} + \ldots + A_{j_k}v^{j_k} = b, \quad v^j = 0, \quad j \neq j_l, \quad l = 1, \ldots, k. \tag{6}$$

Покажем, что столбцы $A_{j_1}, \ldots, A_{j_k}$ линейно независимы.

Пусть при некоторых $\alpha_1, \ldots, \alpha_k$ имеет место равенство

$$\alpha_1 A_{j_1} + \ldots + \alpha_k A_{j_k} = 0. \tag{7}$$

Возьмем точку $v_+ = (v_+^1, \ldots, v_+^n)$ с координатами $v_+^{j_p} = v^{j_p} + \varepsilon\alpha_p$, $v_+^j = 0$ при $j \neq j_p$ $(p = 1, \ldots, k)$ и точку $v_- = (v_-^1, \ldots, v_-^n)$ с координатами $v_-^{j_p} = v^{j_p} - \varepsilon\alpha_p$, $v_-^j = 0$ при $j \neq j_p$ $(p = 1, \ldots, k)$. Поскольку $v^{j_p} > 0$ $(p = 1, \ldots, k)$, то при достаточно малых $\varepsilon > 0$ будем иметь $v_+ \geqslant 0$, $v_- \geqslant 0$. Кроме того, умножая (7) на ε или $-\varepsilon$ и складывая с (6), приходим к равенствам $Av_+ = b$, $Av_- = b$. Таким образом, $v_+, v_- \in U$. Очевидно, $v = (v_+ + v_-)/2$, т. е. $v = \alpha v_+ + (1 - \alpha)v_-$ при $\alpha = 1/2$. По определению угловой точки это возможно лишь при $v_+ = v_- = v$, что в свою очередь означает, что $\alpha_1 = \ldots = \alpha_k = 0$. Таким образом, равенство (7) возможно только при $\alpha_1 = \ldots = \alpha_k = 0$. Линейная независимость столбцов $A_{j_1}, \ldots, A_{j_k}$ доказана. Отсюда следует, что $k \leqslant r$.

Если $k = r$, то соотношения (6) равносильны (5). Если $k < r$, то добавим к столбцам $A_{j_1}, \ldots, A_{j_k}$ новые столбцы $A_{j_{k+1}} \ldots, A_{j_s}$ матрицы A так, чтобы система $A_{j_1}, \ldots, A_{j_k}, A_{j_{k+1}}, \ldots, A_{j_s}$ была линейно независимой, а при добавлении любого другого столбца A_j эта система становилась линейно зависимой. Тогда система $A_{j_1}, \ldots, A_{j_s}$ образует некоторый базис линейной оболочки векто-

ров $A_1, \ldots, A_n$. Размерность линейной оболочки векторов $A_1, \ldots$ $\ldots, A_n$ равна рангу матрицы A, так что $s = r = \operatorname{rang} A$. Добавив к первому равенству (6) столбцы $A_{j_{k+1}}, \ldots, A_{j_r}$, умноженные соответственно на $v^{j_{k+1}} = 0, \ldots, v^{j_r} = 0$, из (6) получим соотношения (5). Необходимость доказана.

Достаточность. Пусть некоторая точка $v = (v^1, \ldots, v^n)$ удовлетворяет условиям (5), где $A_{j_1}, \ldots, A_{j_r}$ — линейно независимы, $r = \operatorname{rang} A$. Пусть $v = \alpha v_1 + (1-\alpha) v_2$ при некоторых $v_1, v_2 \in U$, $0 < \alpha < 1$. Покажем, что такое представление возможно только при $v_1 = v_2 = v$. Сразу же заметим, что если $v^j = 0$, то из этого представления с учетом неравенств $0 < \alpha < 1$, $v_1^j \geqslant 0$, $v_2^j \geqslant 0$ получим $0 \leqslant \alpha v_1^j + (1-\alpha) v_2^j = v^j = 0$, что возможно лишь при $v_1^j = v_2^j = v^j = 0$. Таким образом, для получения равенства $v = v_1 = v_2$ остается еще доказать, что $v_1^j = v_2^j = v^j$ и при тех j, для которых $v^j > 0$.

По условию (5) у точки v положительными могут быть лишь координаты $v^{j_1}, \ldots, v^{j_r}$. Произведя при необходимости перенумерацию переменных, можем считать, что $v^{j_1} > 0, \ldots, v^{j_k} > 0$, $v^{j_{k+1}} = 0, \ldots, v^{j_r} = 0$ (случаи $k = 0$ или $k = r$ здесь не исключаются). Тогда (4) можно переписать в виде $A_{j_1} v^{j_1} + \ldots + A_{j_k} v^{j_k} = b$. Кроме того, учитывая, что по доказанному $v_1^j = v_2^j = 0$ при всех $j \neq j_p$ $(p = 1, \ldots, k)$, равенства $Av_i = b$ также можно записать в виде $A_{j_1} v_i^{j_1} + \ldots + A_{j_k} v_i^{j_k} = b$ $(i = 1, 2)$. Вспомним, что векторы $A_{j_1}, \ldots, A_{j_k}$ линейно независимы. Поэтому вектор b может линейно выражаться через $A_{j_1}, \ldots, A_{j_k}$ единственным способом. Это значит, что $v^{j_p} = v_1^{j_p} = v_2^{j_p}$ для $p = 1, \ldots, k$. Тем самым установлено, что $v = v_1 = v_2$. Следовательно, v — угловая точка множества U.

Определение 2. Систему векторов $A_{j_1}, \ldots, A_{j_r}$, входящих в первое из равенств (5), называют *базисом угловой точки* v, а соответствующие им переменные $v^{j_1}, \ldots, v^{j_r}$ — *базисными координатами* угловой точки v. Если все базисные координаты угловой точки положительны, то такую угловую точку называют *невырожденной*. Если же среди базисных координат $v^{j_1}, \ldots, v^{j_r}$ — хотя бы одна равна нулю, то такая угловая точка называется *вырожденной*. При фиксированном базисе $A_{j_1}, \ldots, A_{j_r}$ переменные $u^{j_1}, \ldots, u^{j_r}$ называются *базисными переменными* угловой точки, а остальные переменные u^j — *небазисными* (*свободными*) *переменными*.

Из теоремы 1 следует, что невырожденная угловая точка обладает единственным базисом — ее базис составляют столбцы с теми номерами, которым соответствуют положительные коорди-

наты угловой точки. Если угловая точка вырожденная, то она может обладать несколькими базисами. В самом деле, если $v^{j_1} > 0, \ldots, v^{j_k} > 0$ $(k < r = \operatorname{rang} A)$, а остальные координаты v^j угловой точки v равны нулю, то, как видно из доказательства теоремы 1, в базис такой точки обязательно войдут столбцы $A_{j_1}, \ldots, A_{j_k}$, а остальные базисные столбцы $A_{j_{k+1}}, \ldots, A_{j_r}$, входящие в представление (5), могут быть выбраны, вообще говоря, различными способами.

Пример 1. Пусть $U = \{u = (u^1, u^2, u^3, u^4) \in E^4\colon u^j \geqslant 0, j = 1, \ldots, 4, u^1 + u^2 + 3u^3 + u^4 = 3, u^1 - u^2 + u^3 + 2u^4 = 1\}$. Обозначим

$$A_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Нетрудно видеть, что точки $u_1 = (2, 1, 0, 0)$ и $u_2 = (0, 5/3, 0, 4/3)$ являются невырожденными угловыми точками множества U, причем точке u_1 соответствует базис A_1, A_2, а точке u_2 — базис A_2, A_4; угловая точка $u_3 = (0, 0, 1, 0)$ вырожденная и ей соответствуют базисы A_1, A_3 или A_2, A_3 или A_3, A_4; точка $u_4 = (5, 0, 0, -2)$ не является угловой для множества U, так как $u_4 \notin U$.

Поскольку из n столбцов матрицы A можно выбрать r линейно независимых столбцов не более чем C_n^r способами (C_n^r — число сочетаний из n элементов по r), то из теоремы 1 следует, что число угловых точек множества (3) конечно.

Кроме того, как увидим ниже (см. теоремы 6.1, 6.2), если $U \neq \varnothing$ и $\inf\limits_U \langle c, u\rangle > -\infty$, то эта нижняя грань достигается хотя бы в одной угловой точке множества (3). Это значит, что каноническую задачу (1.14) можно попытаться решить следующим образом: 1) найти все угловые точки множества (3) — для этого можно, например, рассмотреть C_n^r систем вида (5) и проверить, будут ли выбранные столбцы $A_{j_1}, \ldots, A_{j_r}$ линейно независимы и будут ли соответствующие решения системы (5) неотрицательными; 2) вычислить значение функции $J(u) = \langle c, u\rangle$ в каждой из угловых точек, число которых, как мы знаем, конечно, и определить наименьшее из них. Однако такой подход к решению задачи (1.14) практически не применяется, так как даже в задачах не очень большой размерности число угловых точек может быть столь большим, что простой перебор всех угловых точек множества (3) может оказаться невозможным за разумное время даже при использовании самых быстродействующих ЭВМ.

Тем не менее идея перебора угловых точек множества оказалась весьма плодотворной и послужила основой ряда методов решения канонической и других задач линейного программирования. Одним из таких методов является так называемый *симп-*

лекс-метод. Название этого метода связано с тем, что он впервые разрабатывался применительно к задачам линейного программирования, в которых множество U представлял собой симплекс в E^n: $U=\left\{u=(u^1, \ldots, u^n): u\geqslant 0, \sum_{i=1}^{n} u^i=1\right\}$; затем метод был обобщен на случай более общих множеств U, но первоначальное название за ним так и сохранилось; в литературе этот метод также называют *методом последовательного улучшения плана*.

Оказывается, с помощью симплекс-метода можно осуществить упорядоченный (направленный) перебор угловых точек множества (3), при котором значение функции $\langle c, u\rangle$ убывает при переходе от одной угловой точки к другой, и при этом, рассмотрев лишь относительно небольшое число угловых точек, удается выяснить, имеет ли задача (1.14) решение, и если имеет, то найти его. Кроме того, как увидим ниже, симплекс-метод может быть использован также и для определения угловой точки множества (3).

§ 3. Симплекс-метод

1. Перейдем к описанию симплекс-метода для решения канонической задачи (1.14):

$$\langle c, u\rangle \to \inf;\ u\in U=\{u\in E^n:\ u\geqslant 0,\ Au=b\}. \tag{1}$$

В (1.14) предполагалось, что матрица A имеет размер $m\times n$. Тогда $r=\operatorname{rang} A\leqslant \min\{m;\ n\}$. Предполагая, что из системы $(Au)^i=b^i$ $(i=1, \ldots, m)$ исключены линейно зависимые уравнения, можем считать, что матрица A имеет размер $r\times n$, где $r=\operatorname{rang} A$. Тогда $r\leqslant n$. Если $r=n$, то система $Au=b$ будет иметь единственное решение u и множество U будет либо пустым (если не соблюдается ограничение $u\geqslant 0$), либо будет состоять из одной точки (если $u\geqslant 0$) — в этом случае задача минимизации функции $J(u)$ на U становится малосодержательной. Поэтому будем считать $r<n$. Тогда система $Au=b$ запишется в виде

$$\begin{aligned} a_{11}u^1+\ldots+a_{1n}u^n&=b^1,\\ \cdot\ \cdot\ \cdot

а переменные $u^1, \ldots, u^r$ — базисными переменными. Обозначим

$$\bar{u} = \begin{bmatrix} u^1 \\ \cdot\ \cdot\ \cdot \\ u^r \end{bmatrix}, \quad \bar{v} = \begin{bmatrix} v^1 \\ \cdot\ \cdot\ \cdot \\ v^r \end{bmatrix}, \quad \bar{c} = \begin{bmatrix} c^1 \\ \cdot\ \cdot\ \cdot \\ c^r \end{bmatrix}, \quad A_j = \begin{bmatrix} a_{1j} \\ \cdot\ \cdot\ \cdot \\ a_{rj} \end{bmatrix},$$

$$B = \begin{bmatrix} a_{11} & \ldots & a_{1r} \\ \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{r1} & \ldots & a_{rr} \end{bmatrix} = (A_1 \ldots A_r).$$

Тогда систему (2) можно переписать кратко так:

$$A_1u^1 + \ldots + A_ru^r + A_{r+1}u^{r+1} + \ldots + A_nu^n = \\ = B\bar{u} + A_{r+1}u^{r+1} + \ldots + A_nu^n = b. \quad (3)$$

Поскольку столбцы $A_1, \ldots, A_r$ линейно независимы, то $\det B \neq 0$ и, следовательно, существует обратная матрица B^{-1}. Кроме того, согласно теореме 2.1 небазисные координаты $v^{r+1}, \ldots, v^n$ угловой точки v равны нулю, так что $v = \begin{bmatrix} \bar{v} \\ 0 \end{bmatrix}$, где $\bar{v} \geqslant 0$. Тогда из (3) следует, что базисные координаты $\bar{v}$ удовлетворяют системе $B\bar{v} = b$, откуда имеем $\bar{v} = B^{-1}b \geqslant 0$.

Учитывая последнее равенство, умножим систему (3) на B^{-1} слева и получим следующее соотношение между базисными переменными $\bar{u}$ и небазисными переменными $u^{r+1}, \ldots, u^n$:

$$\bar{u} + \sum_{k=r+1}^{n} B^{-1}A_ku^k = B^{-1}b = \bar{v} \geqslant 0. \quad (4)$$

Обозначим: $(B^{-1}A_k)^s = \gamma_{sk}$ — s-я координата вектор-столбца $B^{-1}A_k$ ($s = 1, \ldots, r$, $k = r+1, \ldots, n$). Тогда уравнение (4) можно записать в покоординатной форме

$$\begin{array}{l} u^1 \quad + \gamma_{1,r+1}u^{r+1} + \ldots + \gamma_{1n}u^n = v^1, \\ u^2 \quad + \gamma_{2,r+1}u^{r+1} + \ldots + \gamma_{2n}u^n = v^2, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \quad u^i \quad + \gamma_{i,r+1}u^{r+1} + \ldots + \gamma_{in}u^n = v^i, \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \qquad u^r + \gamma_{r,r+1}u^{r+1} + \ldots + \gamma_{rn}u^n = v^n. \end{array} \quad (5)$$

Перенося в (4) и (5) слагаемые с небазисными переменными вправо, получаем выражение базисных переменных через небазисные в векторной форме:

$$\bar{u} = \bar{v} - \sum_{k=r+1}^{n} B^{-1}A_ku^k \quad (6)$$

и соответственно в покоординатной форме:

$$\begin{aligned} u^1 &= v^1 - \gamma_{1,\,r+1}u^{r+1} - \ldots - \gamma_{1k}u^k - \ldots - \gamma_{1n}u^n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ u^i &= v^i - \gamma_{i,\,r+1}u^{r+1} - \ldots - \gamma_{ik}u^k - \ldots - \gamma_{in}u^n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ u^r &= v^r - \gamma_{r,\,r+1}u^{r+1} - \ldots - \gamma_{rk}u^k - \ldots - \gamma_{rk}u^n. \end{aligned} \tag{7}$$

Подчеркнем, что системы (3)—(7) являются эквивалентными переформулировками исходной системы (2). Заметим также, что если заранее знать номера базисных переменных, то для получения из (2) соотношений (7) можно воспользоваться известными методами [4, 54] (например, схемой Жордана). О том, как определить угловую точку и номера ее базисных переменных и как привести систему (2) к виду (7), будет рассказано в § 5.

Пользуясь соотношениями (6) или (7), функцию $J(u) = \langle c, u\rangle = \sum_{i=1}^{n} c_i u^i$ можно выразить через небазисные переменные:

$$J(u) = \left\langle \bar{c}, \bar{v} - \sum_{k=r+1}^{n} B^{-1}A_i u^i \right\rangle + \sum_{i=r+1}^{n} c_i u^i =$$

$$= \langle \bar{c}, \bar{v}\rangle - \sum_{i=r+1}^{n} \left(\langle \bar{c}, B^{-1}A_i\rangle - c_i\right) u^i.$$

Поскольку $\langle \bar{c}, \bar{v}\rangle = \langle c, v\rangle = J(v)$, то

$$J(u) = J(v) - \sum_{i=r+1}^{n} \Delta_i u^i, \tag{8}$$

где

$$\Delta_i = \langle \bar{c}, B^{-1}A_i\rangle - c_i = \sum_{s=1}^{r} c_s \gamma_{si} - c_i. \tag{9}$$

Кстати заметим, что величины Δ_i имеют смысл и при $i = 1, \ldots, r$; тогда $\Delta_i = \langle \bar{c}_i, e_i\rangle - c_i = 0$, так как по определению обратной матрицы $B^{-1}A_i = e_i$ $(i = 1, \ldots, r)$, где e_i — i-й столбец единичной матрицы порядка $r \times r$.

Входящие в (5), (8) величины γ_{sk}, v^i, Δ_i удобно записать в виде табл. 1, которую принято называть *симплекс-таблицей*, соответствующей угловой точке v. Для дальнейшего полезно заметить, что величины $\gamma_{sk} = (B^{-1}A_k)^s$, $\Delta_i = \langle \bar{c}, B^{-1}A_i\rangle - c_i$ полностью определяются заданием A, c и базиса угловой точки и не зависят от b.

Таким образом, каноническая задача (1) может быть теперь сформулирована в равносильной форме через небазисные переменные: минимизировать функцию (8) при условиях (6) (или (7)) и, кроме того, $u \geqslant 0$. Конечно, от такой переформулировки

задача (1) проще не стала, но в новой ее формулировке, оказывается, легче проследить за тем, как изменяется функция $J(u)$ при изменении небазисных переменных, и можно попытаться выбрать эти переменные так, чтобы в новой точке $w \in U$ было $J(w) < J(v)$. Однако если мы начнем изменять все небазисные

Таблица 1

Базисные переменные	u^1	...	u^i	...	u^s	...	u^r	u^{r+1}	...	u^k	...	u^j	...	u^n	Свободные члены
u^1	1	...	0	...	0	...	0	$\gamma_{1,\ r+1}$	...	γ_{1k}	...	γ_{1j}	...	γ_{1n}	v^1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^i	0	...	1	...	0	...	0	$\gamma_{i,\ r+1}$	...	γ_{ik}	...	γ_{ij}	...	γ_{in}	v^i
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^s	0	...	0	...	1	...	0	$\gamma_{s,\ r+1}$	...	γ_{sk}	...	γ_{sj}	...	γ_{sn}	v^s
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^r	0	...	0	...	0	...	1	$\gamma_{r,\ r+1}$	...	γ_{rk}	...	γ_{rj}	...	γ_{rn}	v^r
Функция	0	...	0	...	0	...	0	Δ_{r+1}	...	Δ_k	...	Δ_j	...	Δ_n	$J(v)$

переменные снизу, то вряд ли сможем проследить и за изменением функции $J(u)$ и за соблюдением ограничений $u \geqslant 0$. Поэтому мы попробуем изменить лишь одну из небазисных переменных, скажем, переменную u^k $(r+1 \leqslant k \leqslant n)$, а остальные небазисные переменные положим равными нулю. Тогда из соотношений (7) получим

$$u^1 = v^1 - \gamma_{1k}u^k,\ \ldots,\ u^i = v^i - \gamma_{ik}u^k,\ \ldots,\ u^s = v^s - \gamma_{sk}u^k,\ \ldots$$
$$\ldots,\ u^r = v^r - \gamma_{rk}u^k,\ u^{r+1} = \ldots = u^{k-1} = u^{k+1} = \ldots = u^n = 0, \qquad (10)$$

или короче

$$\bar{u} = \bar{v} - (B^{-1}A_k)u^k, \quad u^j = 0, \quad j = r+1,\ \ldots,\ n, \quad j \neq k, \qquad (11)$$

а функция (8) запишется в виде

$$J(u) = J(v) - \Delta_k u^k. \qquad (12)$$

Наша ближайшая задача: выбрать номер k $(r+1 \leqslant k \leqslant n)$ и величину $u^k \geqslant 0$ так, чтобы новая точка w, у которой k-я координата равна u^k, а остальные координаты определяются равенствами (10), удовлетворяла требованиям $Aw = b$ $(w \geqslant 0)$, $J(w) \leqslant J(v)$ (будет еще лучше, если удастся получить $J(w) < J(v)$). Что касается первого требования $Aw = b$, то как видно из соотношений (7), равносильных (2), оно будет выполняться при любом выборе u^k. Анализируя знаки величин Δ_k, γ_{sk}, нетрудно выяснить можно ли удовлетворить оставшимся двум требова-

ниям: $w \geqslant 0$, $J(w) \leqslant J(v)$, и указать правило выбора нужного номера k и нужной величины $u^k \geqslant 0$. А именно, рассмотрим следующие три взаимоисключающих случая I—III.

Случай I. Справедливы неравенства

$$\Delta_i = \langle \bar{c}, B^{-1}A_i \rangle - c_i \leqslant 0, \quad i = r+1, \ldots, n, \tag{13}$$

т. е. в нижней строке симплекс-таблицы 1 все Δ_i неположительны. Как видно из (12), тогда невозможно добиться неравенства $J(u) < J(v)$ при взятом выше специальном способе изменения переменных (10) ни при каких k $(r+1 \leqslant k \leqslant n)$ и $u^k > 0$. Однако это обстоятельство не должно огорчать нас, так как, оказывается, при выполнении условий (13) рассматриваемая точка является решением задачи (1). В самом деле, для любой точки $u \in U = \{u\colon u \geqslant 0,\ Au = b\}$ с учетом представления (4) и неравенств (13) имеем $J(u) = \langle c, u \rangle = \sum_{i=1}^{r} c_i u^i + \sum_{i=r+1}^{n} c_i u^i \geqslant \langle \bar{c}, \bar{u} \rangle + \sum_{i=r+1}^{n} \langle \bar{c}, B^{-1}A_i \rangle u^i = \left\langle \bar{c}, \ \bar{u} + \sum_{i=r+1}^{n} B^{-1}A_k u^i \right\rangle = \langle \bar{c}, \ \bar{v} \rangle = J(v)$. Таким образом, $J(u) \geqslant J(v)$ при всех $u \in U$, т. е. v — решение задачи (1).

Случай II. Существует номер k $(r+1 \leqslant k \leqslant n)$ такой, что

$$\Delta_k > 0, \ \gamma_{ik} \leqslant 0, \ i = 1, \ldots, r, \text{ т. е. } B^{-1}A_k \leqslant 0. \tag{14}$$

Это значит, что в симплекс-таблице 1 в k-м столбце, где находится $\Delta_k > 0$, нет ни одного положительного числа γ_{ik}. В этом случае при всех $u^k \geqslant 0$ точка u, определяемая формулой (10) (или (11)), будет иметь неотрицательные координаты и, следовательно, будет принадлежать множеству U. Тогда, как видно из (12), $J(u) = J(v) - \Delta_k u^k \to -\infty$ при $u^k \to \infty$. Это значит, что $\inf\limits_U \langle c, u \rangle = -\infty$, т. е. задача (1) не имеет решения.

Случай III. Существуют номера k, i $(r+1 \leqslant k \leqslant n,\ 1 \leqslant i \leqslant r)$ такие, что

$$\Delta_k > 0, \quad \gamma_{ik} > 0. \tag{15}$$

Это значит, что в k-м столбце симплекс-таблицы 1, где находится $\Delta_k > 0$, имеется хотя бы одно положительное число γ_{ik}. В этом случае множество индексов $I_k = \{i\colon 1 \leqslant i \leqslant r,\ \gamma_{ik} > 0\}$ непусто. Тогда для точки u, координаты которой определяются формулами (10), согласно (12) при любом $u^k > 0$ будем иметь $J(u) = J(v) - \Delta_k u^k < J(v)$. Остается лишь позаботиться о выполнении условия $u \geqslant 0$. Как видно из формул (10) при $\gamma_{ik} \leqslant 0$, т. е. $i \notin I_k$, будем иметь $u^i = v^i - \gamma_{ik} u^k \geqslant v^i \geqslant 0$ при любом выборе $u^k \geqslant 0$. Если же $\gamma_{ik} > 0$, т. е. $i \in I_k$, то при слишком больших значениях u^k, а именно, при $u^k > \min\limits_{i \in I_k} v^i/\gamma_{ik}$ величина $u^i = v^i - \gamma_{ik} u^k$ станет отрицательной хотя бы для одного номера точек,

определяемых формулами (10), нужно u^k взять так, чтобы $0 \leqslant u^k \leqslant \min\limits_{i \in I_k} v^i/\gamma_{ik}$. Пусть

$$\min_{i \in I_k} v^i/\gamma_{ik} = v^s/\gamma_{sk}, \qquad s \in I_k. \tag{16}$$

Поскольку множество I_k непусто и конечно, то хотя бы один такой номер s существует. Величину γ_{sk}, где номера k, s определяются из (15), (16), называют *разрешающим элементом* симплекс-таблицы 1.

Зафиксируем один из разрешающих элементов γ_{sk} и в формулах (10), (12) положим $u^k = v_s/\gamma_{sk}$. Получим точку $w = (w^1, \ldots, w^n)$ с координатами

$$w^1 = v^1 - \gamma_{1k}\frac{v^s}{\gamma_{sk}}, \ldots, w^i = v^i - \gamma_{ik}\frac{v^s}{\gamma_{sk}}, \ldots, w^{s-1} = v^{s-1} - \gamma_{s-1,k}\frac{v^s}{\gamma_{sk}},$$

$$w^s = v^s - \gamma_{sk}\frac{v^s}{\gamma_{sk}} = 0, \quad w^{s+1} = v^{s+1} - \gamma_{s+1,k}\frac{v^s}{\gamma_{sk}}, \ldots, w^r = v^r - \gamma_{rk}\frac{v^s}{\gamma_{sk}},$$

$$w^{r+1} = 0, \ldots, w^{k-1} = 0, \quad w^k = \frac{v^s}{\gamma_{sk}}, \quad w^{k+1} = 0, \ldots, w^n = 0 \tag{17}$$

и с соответствующим значением функции

$$J(w) = \langle c, w\rangle = J(v) - \Delta_k v^s/\gamma_{sk} \leqslant J(v). \tag{18}$$

В силу формул (7), (10) и условия (16) ясно, что $w \in U$. Покажем, что w — угловая точка множества U с базисом

$$A_1, \ldots, A_{s-1}, \quad A_{s+1}, \ldots, A_r, A_k, \tag{19}$$

получающимся из базиса точки v заменой столбца A_s на A_k. Учитывая, что $w^s = w^{r+1} = \ldots = w^{k-1} = w^{k+1} = \ldots = w^n = 0$, условие $Aw = b$ можно записать в виде $A_1w^1 + \ldots + A_{s-1}w^{s-1} + A_{s+1}w^{s+1} + \ldots + A_rw^r + A_kw^k = b$. Согласно теореме 2.1 остается показать, что система векторов (19) линейно независима.

Пусть для некоторых чисел $\alpha_1, \ldots, \alpha_{s-1}, \alpha_{s+1}, \ldots, \alpha_r, \alpha_k$ оказалось, что

$$\alpha_1A_1 + \ldots + \alpha_{s-1}A_{s-1} + \alpha_{s+1}A_{s+1} + \ldots + \alpha_rA_r + \alpha_kA_k = 0. \tag{20}$$

Поскольку $A_k = BB^{-1}A_k = \sum\limits_{i=1}^{r} A_i\left(B^{-1}A_k\right)^i = \sum\limits_{i=1}^{r} \gamma_{ik}A_i$, то из (20) следует

$$\sum_{i=1, i\neq s}^{r} \alpha_iA_i + \alpha_k\sum_{i=1}^{r}\gamma_{ik}A_i = \sum_{i=1, i\neq s}^{r}(\alpha_i + \alpha_k\gamma_{ik})A_i + \alpha_k\gamma_{sk}A_s = 0.$$

Но система $A_1, \ldots, A_s, \ldots, A_r$ является базисом точки v и, следовательно, линейно независима. Тогда последнее равенство

возможно лишь при $\alpha_i + \alpha_k\gamma_{ik} = 0$ $(i = 1, \ldots, r,\ i \neq s,\ \alpha_k\gamma_{sk} = 0)$. Но $\gamma_{sk} > 0$, как разрешающий элемент, поэтому $\alpha_k = 0$. А тогда все остальные $\alpha_i = 0$ $(i = 1, \ldots, r,\ i \neq s)$. Таким образом, равенство (20) возможно лишь при $\alpha_1 = \ldots = \alpha_{s-1} = \alpha_{s+1} = \ldots = \alpha_r = \alpha_k = 0$, т. е. система (19) линейно независима. Тем самым показано, что w — угловая точка множества U, причем система (19) является ее базисом, а $u^1, \ldots, u^{s-1}, u^{s+1}, \ldots, u^r, u^k$ — ее базисными переменными.

Пользуясь соотношениями (7), нетрудно выразить новые базисные переменные $u^1, \ldots, u^{s-1}, u^{s+1}, \ldots, u^r, u^k$ через остальные небазисные переменные. Для этого из s-го уравнения системы (7) нужно выразить переменную u^k:

$$u^k = \frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n}{}' \frac{\gamma_{sj}}{\gamma_{sk}} u^j \tag{21}$$

(здесь и ниже $\sum'$ означает, что суммирование ведется по всем $j = r+1, \ldots, n$, исключая $j = k$) и затем подставить ее во все остальные уравнения этой системы. С учетом равенств (17) получим

$$\begin{aligned} u^i &= v^i - \sum_{j=r+1}^{n} \gamma_{ij} u^j = \\ &= v^i - \sum_{j=r+1}^{n}{}' \gamma_{ij} u^j - \gamma_{ik}\left(\frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n}{}' \frac{\gamma_{sj}}{\gamma_{sk}} u^j\right) = \\ &= w^i - \left(-\frac{\gamma_{ik}}{\gamma_{sk}}\right) u^s - \sum_{j=r+1}^{n}{}' \left(\gamma_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}}\gamma_{sj}\right) u^j \end{aligned} \tag{22}$$

для всех $i = 1, \ldots, s-1, s+1, \ldots, r$.

Аналогично, подставляя выражение (21) для u^k в (8), функцию $J(u)$ также можно выразить через новые небазисные переменные:

$$\begin{aligned} J(u) &= J(v) - \sum_{j=r+1}^{n}{}' \Delta_j u^j - \Delta_k\left(\frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n}{}' \frac{\gamma_{sj}}{\gamma_{sk}} u^j\right) = \\ &= J(w) - \left(-\frac{\Delta_k}{\gamma_{sk}}\right) u^s - \sum_{j=r+1}^{n}{}' \left(\Delta_j - \Delta_k \frac{\gamma_{sj}}{\gamma_{sk}}\right) u^j; \end{aligned} \tag{23}$$

здесь мы учли равенство (18). Таким образом, базисные переменные угловой точки w и значение функции $J(u)$ выражаются через небазисные переменные этой точки по следующим

формулам:

$$u^1 = w^1 - \gamma'_{1s}u^s - \gamma'_{1,r+1}u^{r+1} - \ldots - \gamma'_{1,k-1}u^{k-1} - \\ - \gamma'_{1,k+1}u^{k+1} - \ldots - \gamma'_{1n}u^n,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$u^i = w^i - \gamma'_{is}u^s - \gamma'_{i,r+1}u^{r+1} - \ldots - \gamma'_{i,k-1}u^{k-1} - \\ - \gamma'_{i,k+1}u^{k+1} - \ldots - \gamma'_{in}u^n,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$u^{s-1} = w^{s-1} - \gamma'_{s-1,s}u^s - \gamma'_{s-1,r+1}u^{r+1} - \ldots - \gamma'_{s-1,k-1}u^{k-1} - \\ - \gamma'_{s-1,k+1}u^{k+1} - \ldots - \gamma'_{s-1,n}u^n,$$

$$u^{s+1} = w^{s+1} - \gamma'_{s+1,s}u^s - \gamma'_{s+1,r+1}u^{r+1} - \ldots - \gamma'_{s+1,k-1}u^{k-1} - \\ - \gamma'_{s+1,k+1}u^{k+1} - \ldots - \gamma'_{s+1,n}u^n,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$u^r = w^r - \gamma'_{rs}u^s - \gamma'_{r,r+1}u^{r+1} - \ldots - \gamma'_{r,k-1}u^{k-1} - \\ - \gamma'_{r,k+1}u^{k+1} - \ldots - \gamma'_{rn}u^n,$$

$$u^k = w^k - \gamma'_{ks}u^s - \gamma'_{k,r+1}u^{r+1} - \ldots - \gamma'_{k,k-1}u^{k-1} - \\ - \gamma'_{k,k+1}u^{k+1} - \ldots - \gamma'_{kn}u^n,$$

$$J(u) = J(w) - \Delta'_s u^s - \Delta'_{r+1}u^{r+1} - \ldots - \Delta'_{k-1}u^{k-1} - \\ - \Delta'_{k+1}u^{k+1} - \ldots - \Delta'_n u^n,$$

где $w^1, \ldots, w^{s-1}, w^{s+1}, \ldots, w^r, w^k$ и $J(w)$ определяются формулами (17), (18), а γ'_{ij}, Δ'_j получаются из (21)—(23) приравниванием коэффициентов при соответствующих неизвестных:

$$\gamma'_{is} = -\frac{\gamma_{ik}}{\gamma_{sk}}, \quad i = 1, \ldots, s-1, s+1, \ldots, r, \quad \gamma'_{ks} = \frac{1}{\gamma_{sk}}; \tag{24}$$

$$\gamma'_{ij} = \gamma_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}}\gamma_{sj}, \quad i = 1, \ldots, s-1, s+1, \ldots, r, \quad \gamma'_{kj} = \frac{\gamma_{sj}}{\gamma_{sk}}, \tag{25}$$

$$j = r+1, \ldots, k-1, k+1, \ldots, n;$$

$$\Delta'_s = -\frac{\Delta_k}{\gamma_{sk}},$$

$$\Delta'_j = \Delta_j - \Delta_k\frac{\gamma_{sj}}{\gamma_{sk}}, \quad j = r+1, \ldots, k-1, \quad k+1, \ldots, n. \tag{26}$$

Табл. 2 представляет собой симплекс-таблицу новой угловой точки w. Таким образом, один шаг симплекс-метода, заключающийся в переходе от одной угловой точки v множества U к другой угловой точке w, описан. Этот шаг формально можно истолковать как переход от одной симплекс-таблицы 1 к следующей

симплекс-таблице 2 по формулам (17), (18), (24)—(26), в которых разрешающий элемент γ_{sk} определяется условиями (15), (16).

Эти формулы перехода были получены в предположении, что базисные переменные угловой точки v имеют номера $1, \ldots, r$. Нетрудно видеть, что если базисные переменные точки v имеют, скажем, номера $j_1, \ldots, j_r$ $(1 \leqslant j_1 < \ldots < j_k \leqslant n)$, то можно выписать формулы

$$u^{j_i} = v^{j_i} - \sum_{k \notin I_v} \gamma_{ik} u^k, \quad i = 1, \ldots, r, \quad J(u) = J(v) - \sum_{k \notin I_v} \Delta_k u^k,$$

выражающие базисные переменные и функцию $J(u)$ через небазисные переменные и аналогичные формулам (5)—(9). Здесь

$$I_v = \{j_1, \ldots, j_r\}, \ \gamma_{ik} = (B^{-1}A_k)^i, \quad i = 1, \ldots, r, \quad k \notin I_v,$$

$$B = (A_{j_1} \ldots A_{j_r}), \quad v^{j_i} = (B^{-1}b)^i, \quad i = 1, \ldots, r;$$

$$\Delta_k = \sum_{i=1}^{r} c_{j_i}\gamma_{ik} - c_k, \quad k \notin I_v; \quad \Delta_k = 0, \quad i \in I_v;$$

табл. 3 представляет собой симплекс-таблицу точки v. Как и выше, здесь нужно рассмотреть три случая.

Случай I.

$$\Delta_k \leqslant 0, \quad k \notin I_v. \tag{13'}$$

Случай II. Существует номер $k \notin I_v$ такой, что

$$\Delta_k > 0, \quad B^{-1}A_k \leqslant 0. \tag{14'}$$

Случай III. Существуют номера $k \notin I_v$, $j_i \in I_v$ такие, что

$$\Delta_k > 0, \quad (B^{-1}A_k)^i = \gamma_{ik} > 0. \tag{15'}$$

Как и выше, нетрудно убедиться, что в случае I точка v является решением задачи (1), в случае II — $\inf\limits_U \langle c, u\rangle = -\infty$ и задача (1) не имеет решения, а в случае III можно выбрать разрешающий элемент γ_{sk} из условий

$$\Delta_k > 0, \quad \min_{i \in I_{vk}} v^{j_i}/\gamma_{ik} = v^{j_s}/\gamma_{sk}, \quad I_{vk} = \{i\colon\ j_i \in I_v,\ \gamma_{ik} > 0\}, \tag{16'}$$

аналогичных условиям (15), (16), и затем осуществить переход к следующей угловой точке w по формулам, аналогичным формулам (17), (18), (24)—(26).

2. Итак, описаны правила перехода от одной угловой точки v множества U к другой угловой точке w, в которой $J(w) \leqslant J(v)$. Возникает вопрос, когда $J(w) = J(v)$ и когда $J(w) < J(v)$? Учитывая, что поскольку по определению номеров s и k согласно (15), (16) $\Delta_k > 0$, $\gamma_{sk} > 0$, а $v^s \geqslant 0$ из-за $v \in U$, то из соотношений (18) нетрудно усмотреть, что $J(w) = J(v)$ тогда и только

Таблица 2

Базисные переменные	u^1	...	u^{s-1}	u^s	u^{s+1}	...	u^r	u^{r+1}	...	u^{k-1}	u^k	u^{k+1}	...	u^n	Свободные члены
u^1	1	...	0	γ'_{1s}	0	...	0	$\gamma'_{1,r+1}$	...	$\gamma'_{1,k-1}$	0	$\gamma'_{1,k+1}$	...	γ'_{1n}	w^1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^{s-1}	0	...	1	$\gamma'_{s-1,s}$	0	...	0	$\gamma'_{s-1,r+1}$	...	$\gamma'_{s-1,k-1}$	0	$\gamma'_{s-1,k+1}$	...	$\gamma'_{s-1,n}$	w^{s-1}
u^{s+1}	0	...	0	$\gamma'_{s+1,s}$	1	...	0	$\gamma'_{s+1,r+1}$	...	$\gamma'_{s+1,k-1}$	0	$\gamma'_{s+1,k+1}$	...	$\gamma'_{s+1,n}$	w^{s+1}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^r	0	...	0	γ'_{rs}	0	...	1	$\gamma'_{r,r+1}$	...	$\gamma'_{r,k-1}$	0	$\gamma'_{r,k+1}$	...	γ'_{rn}	w^r
u^k	0	...	0	γ'_{ks}	0	...	0	$\gamma'_{k,r+1}$	...	$\gamma'_{k,k-1}$	1	$\gamma'_{k,k+1}$	...	γ'_{kn}	w^k
Функция	0	...	0	Δ'_s	0	...	0	Δ'_{r+1}	...	Δ'_{k-1}	0	Δ'_{k+1}	...	Δ'_n	$J(w)$

Таблица 3

Базисные переменные	u^1	...	u^{j_1}	...	u^{j_i}	...	u^k	...	u^{j_s}	...	u^j	...	u^{j_r}	...	u^n	Свободные члены
u^{j_1}	γ_{11}	...	1	...	0	...	γ_{1k}	...	0	...	γ_{1j}	...	0	...	γ_{1n}	v^{j_1}
...	...	...	.	...	.	...	...	...	.	...	...	...	.	...	...	...
u^{j_i}	γ_{i1}	...	0	...	1	...	γ_{ik}	...	0	...	γ_{ij}	...	0	...	γ_{in}	v^{j_i}
...	...	...	.	...	.	...	...	...	.	...	...	...	.	...	...	...
u^{j_s}	γ_{s1}	...	0	...	0	...	γ_{sk}	...	1	...	γ_{sj}	...	0	...	γ_{sn}	v^{j_s}
...	...	...	.	...	.	...	...	...	.	...	...	...	.	...	...	...
u^{j_r}	γ_{r1}	...	0	...	0	...	γ_{rk}	...	0	...	γ_{rj}	...	1	...	γ_{rn}	v^{j_r}
Функция	Δ_1	...	0	...	0	...	Δ_k	...	0	...	Δ_j	...	0	...	Δ_n	$J(v)$

тогда, когда $v^s = 0$, т. е. угловая точка v вырожденная. Таким образом, среди канонических задач имеет смысл выделять задачи вырожденные и невырожденные.

Определение 1. Задача (1) называется *невырожденной*, если все угловые точки множества U невырожденные. Если хотя бы одна угловая точка множества U вырожденная, то задачу (1) называют *вырожденной*.

В том случае, когда задача (1) невырожденная, все базисные координаты угловой точки v будут положительны. В частности, тогда $v^s > 0$ и согласно (18) $J(w) < J(v)$. Далее, из формул (17) видно, что следующая угловая точка w уже имеет $n - r$ нулевых координат: $w^s = w^{r+1} = \ldots = w^{k-1} = w^{k+1} = \ldots = w^n = 0$. Поэтому в невырожденной задаче все остальные r координат точки w, являющиеся базисными, должны быть положительными. А тогда для всех $i \in I_k$ $(i \neq s)$ из формул (17) будем иметь $w^i = \gamma_{ik}(v^i/\gamma_{ik} - v^s/\gamma_{sk}) > 0$, $v^i/\gamma_{ik} > v^s/\gamma_{sk}$, так что минимум в левой части (16) достигается на единственном номере $s \in I_k$. Это значит, что если номер k из условий (15) зафиксирован, то номер переменной u^s, которая выводится из числа базисных переменных, и разрешающий элемент γ_{sk} симплекс-таблицы в невырожденных задачах определяются однозначно.

Итак, мы выяснили, что в невырожденной задаче с помощью симплекс-метода можно осуществить переход от угловой точки, не являющейся решением задачи (1), к другой угловой точке со строгим уменьшением значения функции $J(u) = \langle c, u \rangle$. Учитывая, что число угловых точек множества (2.3) конечно, заключаем, что случай III (условие 15′) бесконечно повторяться не может и процесс закончится тем, что на каком-то шаге симплекс-метода либо реализуется случай II (условие (14′)) и выяснится, что $\inf\limits_U \langle c, u \rangle = -\infty$ — задача (1) не имеет решения, либо реализуется случай I (условие (13′)), означающий, что рассматриваемая угловая точка является решением задачи (1).

Таким образом, показано, что если в невырожденной канонической задаче известна какая-либо угловая точка множества, то, отправляясь от нее, с помощью симплекс-метода за конечное число шагов можно выяснить, разрешима ли эта задача, и если разрешима, то найти ее решение.

3. Перейдем к рассмотрению вырожденных задач (1). Посмотрим внимательнее, к чему может привести применение симплекс-метода, если v — вырожденная угловая точка. Если в этой точке реализуется случай I или II, то выводы останутся прежними: в случае I v — точка минимума функции $\langle c, u \rangle$ на U, а в случае II $\inf\limits_U \langle c, u \rangle = -\infty$; в обоих случаях процесс прекращается. Остается рассмотреть случай III. Конечно, и в случае III может оказаться, что базисные координаты точки v с номерами $i \in I_k$ будут положительными, и тогда, как видно из фор-

мулы (18), симплекс-метод приведет к новой угловой точке w со значением функции $J(w) < J(v)$. Однако здесь возможна и худшая ситуация, когда $\gamma_{sk} > 0$, $v^s = 0$. Ясно, что тогда $s \in I_k$ и минимум в (16) или (16′) будет достигаться именно на этом номере s. Из формул (17), (18) в этом случае получим $w = v$, $J(w) = J(v)$. Это значит, что в результате применения симплекс-метода мы оказались в прежней же угловой точке и лишь заменили один ее базис $A_1, \ldots, A_r$ другим базисом (19).

Возникает вопрос: не приведет ли дальнейшее применение симплекс-метода к бесконечному перебору базисов угловой точки v? Поскольку угловая точка имеет конечное число базисов, то при фиксированном правиле выбора разрешающего элемента (например, часто выбирают разрешающим элементом ту величину γ_{sk}, у которой номера k, s являются наименьшими среди всех номеров, удовлетворяющих условиям (15), (16) или (16′)) это привело бы к так называемому *зацикливанию*, т. е. к бесконечному циклическому перебору базисов точки v. Оказывается, действительно существуют примеры задач (1), в которых возможно зацикливание (см. ниже упражнение 6.7). Можно ли избежать зацикливания? Каким для этого должно быть правило выбора разрешающего элемента?

Любое правило выбора разрешающего элемента, с помощью которого можно избежать зацикливания в задаче (1), назовем *антициклином*. К счастью, имеются достаточно простые антициклины. Остановимся на одном из них. Пусть v_0 — некоторая угловая точка множества (2.3) с базисом $A_{j_1}, \ldots, A_{j_r}$. К симплекс-таблице точки v_0 добавим еще r столбцов $e_1, \ldots, e_r$ единичной матрицы порядка $r \times r$ и в результате придем к табл. 4. На каждом шаге симплекс-метода вновь добавленные столбцы будем преобразовывать по тем же правилам, по которым преобразовываются столбцы небазисных переменных, остающихся небазисными и на очередном шаге симплекс-метода (см. (25), а также (4.11)).

Пусть уже сделано несколько шагов симплекс-метода с расширенной симплекс-таблицей и найдена очередная угловая точка v_l, пусть в результате преобразований в дополнительных столбцах первоначальных $e_1, \ldots, e_r$ появились столбцы $d_1, \ldots, d_r$, где d_j имеет координаты $d_{1j}, \ldots, d_{rj}$ $(j = 1, \ldots, r)$. Пусть табл. 5 представляет собой расширенную симплекс-таблицу точки v_l — в ней для простоты изложения предполагается, что базисом v_l являются столбцы $A_1, \ldots, A_r$ (как уже отмечалось, этого всегда можно добиться перенумерацией переменных). Пусть $\Delta_k > 0$, $I_k = \{i\colon 1 \leqslant i \leqslant r, \gamma_{ik} > 0\} \neq \varnothing$.

Образуем множество $I_{k1} = \left\{s\colon s \in I_k, \min\limits_{i \in I_k} v^i/\gamma_{ik} = v^s/\gamma_{sk}\right\}$. В невырожденной задаче, как было замечено выше, множество I_{k1} всегда состоит из единственного номера s. В вырожденных за-

Таблица 4

Базисные переменные	u^1	...	u^k	...	u^{j_s}	...	u^n	Свободные члены	e_1	...	e_s	...	e_r
u^{j_1}	γ_{11}	...	γ_{1k}	...	0	...	γ_{1n}	v^{j_1}	1	...	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...;	...
u^{j_s}	γ_{s1}	...	γ_{sk}	...	1	...	γ_{sn}	v^{j_s}	0	...	1	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^{j_r}	γ_{r1}	...	γ_{rk}	...	0	...	γ_{rn}	v^{j_r}	0	...	0	...	1
Функция	Δ_1	...	Δ_k	...	0	...	Δ_n	$J(v)$					

Таблица 5

Базисные переменные	u^1	...	u^i	...	u^s	...	u^r	u^{r+1}	...	u^k	...	u^n	Свободные члены	d_1	d_2	...	d_m	...	d_r
u^1	1	...	0	...	0	...	0	$\gamma_{1,r+1}$	...	γ_{1k}	...	γ_{1n}	v^1	d_{11}	d_{12}	...	d_{1m}	...	d_{1r}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^i	0	...	1	...	0	...	0	$\gamma_{i,r+1}$	...	γ_{ik}	...	γ_{in}	v^i	d_{i1}	d_{i2}	...	d_{im}	...	d_{ir}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^s	0	...	0	...	1	...	0	$\gamma_{s,r+1}$	...	γ_{sk}	...	γ_{sn}	v^s	d_{s1}	d_{s2}	...	d_{sm}	...	d_{sr}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^r	0	...	0	...	0	...	1	$\gamma_{r,r+1}$	...	γ_{rk}	...	γ_{rn}	v^r	d_{r1}	d_{r2}	...	d_{rm}	...	d_{rr}
Функция	0	...	0	...	0	...	0	Δ_{r+1}	...	Δ_k	...	Δ_n	$J(v)$						

дачах это не всегда так, и множество I_{k1} может состоять из двух и более номеров. В последнем случае образуем множество $I_{k2} = \left\{s\colon\ s \in I_{k1},\ \min\limits_{i \in I_{k1}} d_{i1}/\gamma_{ik} = d_{s1}/\gamma_{sk}\right\}$. Если уже определено множество I_{km} $(m \geqslant 2)$ и оно содержит не менее двух номеров, то образуем множество

$$I_{k,m+1} = \left\{s\colon\ s \in I_{km},\ \min_{i \in I_{km}} d_{im}/\gamma_{ik} = d_{sm}/\gamma_{sk}\right\}.$$

Ниже будет показано, что в конце концов мы получим множество I_{kl} $(1 \leqslant l \leqslant r+1)$, состоящее из единственного номера s, причем все предыдущие множества I_{kl} $(i = 1, \ldots, l-1)$ будут содержать не менее двух номеров.

Выведем переменную с полученным таким образом номером s из базисных и вместо нее в базисные введем переменную u^k. Оказывается, применение на каждом шаге симплекс-метода описанного правила выбора номера s позволяет избежать зацикливания, и в результате через конечное число шагов симплекс-метода с расширенной симплекс-таблицей процесс закончится реализацией случая I или II. Антициклин для задачи (1) описан. Строгое обоснование этого антициклина дается в § 4.

Следует заметить, что хотя среди прикладных задач линейного программирования вырожденные задачи встречаются довольно часто, но зацикливание бывает крайне редко. Кроме того, использование антициклина на каждом шаге симплекс-метода приводит к заметному увеличению машинного времени ЭВМ, требующегося для решения задачи. Поэтому на практике чаще всего пользуются упрощенным правилом выбора номеров k и s из условий (15), (16) или (16′), причем если выбор здесь неоднозначный, то берут наименьшие номера, удовлетворяющие указанным условиям. И лишь в том случае, когда обнаруживается зацикливание, принимают описанный выше или какой-нибудь другой антициклин, а после того, как зацикливание будет преодолено, снова возвращается к упрощенному правилу выбора разрешающего элемента. Любопытно отметить, что длина циклов в задачах линейного программирования меньше шести не бывает.

§ 4. Антициклин

1. Как было выше замечено, с практической точки зрения проблема зацикливания, по-видимому, не представляет особой важности. Но с теоретической точки зрения указание хотя бы одного правила выбора разрешающего элемента, позволяющего избежать зацикливания, принципиально важно для полного обоснования симплекс-метода для канонической задачи

$$J(u) = \langle c,\ u\rangle \to \inf;\quad u \in U = \{u \in E^n\colon\ u \geqslant 0,\ Au = b\}. \tag{1}$$

Покажем, что описанное в конце § 3 правило выбора разрешающего элемента в самом деле является антициклином. С этой целью рассмот-

рим так называемую *возмущенную каноническую задачу*

$$J(u) = \langle c, u\rangle \to \inf;\ u \in U_\varepsilon = \left\{u \in E^n:\ u \geqslant 0,\ Au = b(\varepsilon) = b + \sum_{i=1}^{r} \varepsilon^i R_i\right\}, \tag{2}$$

получаемую из задачи (1) возмущением правой части уравнения $Au = b$. Здесь ε^i — i-я степень числа $\varepsilon > 0$, R_i — i-й столбец какой-либо невырожденной матрицы R порядка $r \times r$; как и в § 3, предполагаем, что A — матрица размера $r \times n$ ($r = \operatorname{rang} A < n$). Ниже мы покажем, что при малых $\varepsilon > 0$ и удачном выборе матрицы R множество U_ε будет непустым, а задача (2) — невырожденной. Затем исследуем задачу (2) при малых $\varepsilon > 0$ с помощью симплекс-метода и на основе этих исследований выработаем антициклин для задачи (1).

Лемма 1. *Пусть матрица R невырожденная и выбрана так, что множество U_ε непусто при всех ε ($0 \leqslant \varepsilon \leqslant \varepsilon_0$, $\varepsilon_0 > 0$). Тогда найдется число ε_1 ($0 < \varepsilon_1 \leqslant \varepsilon_0$) такое, что: 1) при всех ε ($0 < \varepsilon < \varepsilon_1$) задача (2) невырожденная; 2) если $v(\gamma) = (v^1(\gamma), \ldots, v^n(\gamma))$ — какая-либо угловая точка множества U_γ ($0 < \gamma < \varepsilon_1$) с базисом $A_{j_1}, \ldots, A_{j_r}$, то условия*

$$A_{j_1} v^{j_1} + \ldots + A_{j_r} v^{j_r} = b(\varepsilon), \qquad v^i = 0, \qquad i \neq j_l, \qquad l = 1, \ldots, r, \tag{3}$$

при всех ε ($0 \leqslant \varepsilon < \varepsilon_1$) однозначно определяют угловую точку $v(\varepsilon) = (v^1(\varepsilon), \ldots, v^n(\varepsilon))$ множества U_ε с тем же базисом; в частности, при $\varepsilon = 0$ из (3) получим угловую точку $v = v(0)$ множества (2.3) с базисом $A_{j_1}, \ldots, A_{j_r}$.

Доказательство. Возьмем произвольную линейно независимую систему столбцов $A_{j_1}, \ldots, A_{j_r}$, где $r = \operatorname{rang} A$. Обозначим через $B = (A_{j_1} \ldots A_{j_r})$ квадратную невырожденную матрицу. Тогда система $Au = b(\varepsilon)$ определяет единственную точку $u = u(\varepsilon) = (u^1(\varepsilon), \ldots u^n(\varepsilon))$ с координатами $u^i(\varepsilon) = 0$ при $i \neq j_l$ ($l = 1, \ldots, r$), а остальные координаты $\bar{u}(\varepsilon) = (u^{j_1}(\varepsilon), \ldots, u^{j_r}(\varepsilon))$ определяются равенствами $u^{j_l}(\varepsilon) = (B^{-1}b)^l + \sum_{i=1}^{r} \varepsilon^i (B^{-1}R_i)^l$ ($l = 1, \ldots, r$) и представляют собой многочлены переменной ε степени не выше r; последние равенства можно записать в векторной форме

$$\bar{u}(\varepsilon) = B^{-1}b(\varepsilon) = B^{-1}b + \sum_{i=1}^{r} \varepsilon^i B^{-1} R_i.$$

Поскольку $\det R \neq 0$, $\det B \neq 0$, то $\det(B^{-1}R) = \det B^{-1} \cdot \det R \neq 0$. Это значит, что все строки матрицы $B^{-1}R$ ненулевые. Поскольку именно элементы l-й строки $(B^{-1}R_1)^l, \ldots, (B^{-1}R_r)^l$ этой матрицы служат коэффициентами многочлена $u^{j_l}(\varepsilon)$, то ясно, что ни один из многочленов $u^{j_l}(\varepsilon)$ ($l = 1, \ldots, r$) не является тождественным нулю. Тогда каждый многочлен $u^{j_l}(\varepsilon)$ имеет не более r положительных корней или же таковых вовсе не имеет. Пусть η — наименьший из всех положительных корней всевозможных многочленов $u^{j_i}(\varepsilon)$, построенных по всевозможным линейно независимым наборам $A_{j_1}, \ldots, A_{j_r}$ столбцов матрицы A. Поскольку таких многочленов конечное число, то величина η имеет смысл и $\eta > 0$, если хотя бы один из этих многочленов имеет хотя бы один положительный корень; если же ни один из этих многочленов не имеет ни одного положительного корня, то положим $\eta = \infty$.

Возьмем произвольное ε $(0 < \varepsilon < \varepsilon_1 = \min\{\varepsilon_0; \eta\})$ и произвольную угловую точку $v(\varepsilon)$ множества U_ε с базисом $A_{j_1}, \ldots, A_{j_r}$. Тогда базисные координаты $\bar{v}(\varepsilon) = \left(v^{j_1}(\varepsilon), \ldots, v^{j_r}(\varepsilon)\right)$ точки $v(\varepsilon)$, определяемые условиями $A_{j_1}v^{j_1}(\varepsilon) + \ldots + A_{j_r}v^{j_r}(\varepsilon) = B\bar{v}(\varepsilon) = b(\varepsilon)$, представимы в виде

$$\bar{v}(\varepsilon) = B^{-1}b(\varepsilon) = B^{-1}b + \sum_{i=1}^{r} \varepsilon^i \left(B^{-1}R_i\right)$$

и, следовательно, принадлежат рассмотренному выше множеству многочленов, а небазисные координаты, как им и полагается, равны нулю. Из того, что $v(\varepsilon) \in U_\varepsilon$, следует, что $\bar{v}(\varepsilon) \geqslant 0$. Но тогда в силу выбора ε из $0 < \varepsilon < \varepsilon_1$ имеем $v(\varepsilon) > 0$, т. е. $v(\varepsilon)$ — невырожденная угловая точка множества U_ε. Отсюда следует невырожденность задачи (2) при всех ε $(0 < \varepsilon < \varepsilon_1)$. Утверждение 1) леммы доказано.

Пусть $v(\gamma)$ — какая-либо угловая точка множества U_γ $(0 < \gamma < \varepsilon_1)$ с базисом $A_{j_1}, \ldots, A_{j_r}$, т. е.

$$A_{j_1}v^{j_1}(\gamma) + \ldots + A_{j_r}v^{j_r}(\gamma) = B\bar{v}(\gamma) = b(\gamma), \qquad v^i(\gamma) = 0,\ i \neq j_l, \quad l = 1, \ldots, r,$$

где $B = \left(A_{j_1} \ldots A_{j_r}\right)$, $\bar{v}_{(\gamma)} = \left(v^{j_1}(\gamma), \ldots, v^{j_r}(\gamma)\right)$. По доказанному выше $\bar{v}(\gamma) > 0$. Покажем, что точка $v(\varepsilon)$, определяемая условиями (3), является угловой точкой множества U_ε при каждом ε $(0 \leqslant \varepsilon < \varepsilon_1)$. Для этого прежде всего заметим, что условия (3) означают, что $Av(\varepsilon) = b(\varepsilon)$ при $0 \leqslant \varepsilon < \varepsilon_1$. Перепишем условия (3) в виде

$$\bar{v}(\varepsilon) = B^{-1}b + \sum_{i=1}^{r} \varepsilon^i \left(B^{-1}R_i\right), \quad v^i(\varepsilon) = 0, \quad i \neq j_l, \quad l = 1, \ldots, r. \tag{4}$$

Отсюда и из определения ε_1 видно, что ни одна из координат вектора $\bar{v}(\varepsilon)$ не может обратиться в нуль ни при каком ε $(0 < \varepsilon < \varepsilon_1)$. Но нам уже известно, что $\bar{v}(\gamma) > 0$. С учетом непрерывности $\bar{v}(\varepsilon)$ тогда заключаем, что $\bar{v}(\varepsilon) > 0$ при всех ε $(0 < \varepsilon < \varepsilon_1)$. Отсюда и из (4) при $\varepsilon \to +0$ имеем $\lim\limits_{\varepsilon \to +0} v(\varepsilon) = B^{-1}b = \bar{v}(0) = \bar{v} \geqslant 0$.

Таким образом, показано, что точка $v(\varepsilon)$, определяемая условиями (3), при всех ε $(0 \leqslant \varepsilon < \varepsilon_1)$ имеет неотрицательные координаты. Следовательно, $v(\varepsilon) \in U_\varepsilon$ $(0 \leqslant \varepsilon < \varepsilon_1)$. Из (3) с помощью теоремы 2.1 заключаем, что $v(\varepsilon)$ — угловая точка множества U_ε с базисом $A_{j_1}, \ldots, A_{j_r}$ при каждом ε $(0 \leqslant \varepsilon < \varepsilon_1)$.

2. Предполагая, что условия леммы 1 выполнены, рассмотрим подробнее один шаг симплекс-метода для задачи (2) при фиксированном ε $(0 < \varepsilon < \varepsilon_1)$. Пусть $v(\varepsilon) = (v^1(\varepsilon), \ldots, v^n(\varepsilon))$ — какая-либо угловая точка множества U_ε. Без ограничения общности можем считать, что столбцы $A_1, \ldots, A_r$ являются базисом, а координаты $(v^1(\varepsilon), \ldots, v^r(\varepsilon)) = \bar{v}(\varepsilon)$ — базисными координатами этой точки. Согласно лемме 1 задача (2) невырожденная при $0 < \varepsilon < \varepsilon_1$, поэтому

$$\bar{v}(\varepsilon) > 0, \quad v^{r+1}(\varepsilon) = \ldots = v^n(\varepsilon) = 0.$$

Обозначим $B = (A_1 \ldots A_r)$. Поскольку $Av(\varepsilon) = B\bar{v}(\varepsilon) = b(\varepsilon)$, то $\bar{v}(\varepsilon) = B^{-1}b(\varepsilon) > 0$. Умножая уравнение $Au = B\bar{u} + A_{r+1}u^{r+1} + \ldots + A_n u^n = b(\varepsilon) = b + \sum\limits_{j=1}^{r} \varepsilon^j R_j$ на B^{-1} слева, получаем следующее соотношение

между базисными переменными $\bar{u} = (u^1, \ldots, u^r)$ и небазисными переменными $u^{r+1}, \ldots, u^n$ точки $v(\varepsilon)$:

$$\bar{u} + \sum_{k=r+1}^{n} B^{-1}A_k u^k = B^{-1}b(\varepsilon) = \bar{v}(\varepsilon) = B^{-1}b + \sum_{j=1}^{r} \varepsilon^j \left(B^{-1}R_j\right). \tag{5}$$

Обозначим

$$B^{-1}A_k = \begin{bmatrix} \gamma_{1k} \\ \cdots \\ \gamma_{rk} \end{bmatrix}, \quad d_j = B^{-1}R_j = \begin{bmatrix} d_{1j} \\ \cdots \\ d_{rj} \end{bmatrix}, \quad \bar{v} = B^{-1}b = \begin{bmatrix} v^1 \\ \cdots \\ v^r \end{bmatrix}, \quad \bar{c} = \begin{bmatrix} c_1 \\ \cdots \\ c_r \end{bmatrix}.$$

Тогда соотношение (5) можно переписать в следующей покоординатной форме:

$$\begin{array}{l} u^1 \quad + \gamma_{1,r+1}u^{r+1} + \ldots + \gamma_{1k}u^k + \ldots + \gamma_{1n}u^n = v^1 + d_{11}\varepsilon + \ldots + d_{1r}\varepsilon^r, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ u^s \quad + \gamma_{s,r+1}u^{r+1} + \ldots + \gamma_{sk}u^k + \ldots + \gamma_{sn}u^n = v^s + d_{s1}\varepsilon + \ldots + d_{sr}\varepsilon^r, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ u^r + \gamma_{r,r+1}u^{r+1} + \ldots + \gamma_{rk}u^k + \ldots + \gamma_{rn}u^n = v^n + d_{r1}\varepsilon + \ldots + d_{rr}\varepsilon^r, \end{array} \tag{6}$$

а для базисных координат $\bar{v}(\varepsilon) = B^{-1}b + \sum_{j=1}^{r} \varepsilon^j \left(B^{-1}R_j\right)$ получим

$$v^i(\varepsilon) = v^i + \sum_{j=1}^{r} d_{ij}\varepsilon^j, \qquad i = 1, \ldots, r. \tag{7}$$

Далее, по аналогии с (3.8), с помощью равенств (6) или (7) нетрудно выразить функцию $J(u) = \langle c, u\rangle$ через базисные переменные:

$$J(u) = J(v(\varepsilon)) - \sum_{i=r+1}^{n} \Delta_i u^i, \quad \Delta_i = \left\langle \bar{c}, B^{-1}A_i \right\rangle - c_i. \tag{8}$$

Учитывая, что $\Delta_i = \langle \bar{c}, B^{-1}A_i\rangle - c_i = 0$ при $i = 1, \ldots, r$, на основе представлений (6), (8) выпишем симплекс-таблицу для угловой точки $v(\varepsilon)$, выделив в ней для векторов $d_1, \ldots, d_r$ дополнительные столбцы. В результате получим табл. 6.

Заметим, что точка $v = (v^1, \ldots, v^n) = v(0)$ согласно лемме 1 является угловой для множества U с тем же базисом $A_1, \ldots, A_r$, что и точка $v(\varepsilon)$. Ясно также, что выражения (3.5), (3.8) базисных переменных $\bar{u} = (u^1, \ldots, u^r)$ угловой точки v и функции $J(u)$ через небазисные переменные $u^{r+1}, \ldots, u^n$ получаются соответственно из (6), (8) при $\varepsilon = 0$. Это значит, что выписанная ранее табл. 5 является симплекс-таблицей точки v, расширенной дополнительными столбцами $d_1, \ldots, d_r$; кстати, теперь выясняется, что элементы $d_{i1}, \ldots, d_{ir}$ i-й строки табл. 5 представляют собой коэффициенты многочленов (7).

Опираясь на симплекс-таблицу 6, сделаем один шаг симплекс-метода из угловой точки $v(\varepsilon)$ для задачи (2) и посмотрим, чему это будет соответствовать в задаче (1). Как и в § 3, рассмотрим три случая.

Случай I. В нижней строке симплекс-таблицы 6 величины $\Delta_i \leqslant 0$ при всех $i = r+1, \ldots, n$. Тогда, как и в § 3, нетрудно показать, что $v(\varepsilon)$ — решение задачи (2), т. е. $\langle c, v(\varepsilon)\rangle = \inf\limits_{U_\varepsilon} \langle c, u\rangle > -\infty$. Теперь

Таблица 6

Базисные переменные	u^1	...	u^i	...	u^s	...	u^r	u^{r+1}	...	u^k	...	u^n	Свободные члены	d_1	d_2	...	d_m	...	d_r
u^1	1	...	0	...	0	...	0	$\gamma_{1,r+1}$	...	γ_{1k}	...	γ_{1n}	v^1	d_{11}	d_{12}	...	d_{1m}	...	d_{1r}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^i	0	...	1	...	0	...	0	$\gamma_{i,r+1}$	...	γ_{ik}	...	γ_{in}	v^i	d_{i1}	d_{i2}	...	d_{im}	...	d_{ir}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^s	0	...	0	...	1	...	0	$\gamma_{s,r+1}$	...	γ_{sk}	...	γ_{sn}	v^s	d_{s1}	d_{s2}	...	d_{sm}	...	d_{sr}
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
u^r	0	...	0	...	0	...	1	$\gamma_{r,r+1}$	...	γ_{rk}	...	γ_{rn}	v^r	d_{r1}	d_{r2}	...	d_{rm}	...	d_{rr}
Функция	0	...	0	...	0	...	0	Δ_{r+1}	...	Δ_k	...	Δ_n	$J(v(\varepsilon))$						

заметим, что в нижней строке симплекс-таблицы 5 угловой точки $v = v(0)$ находятся те же величины $\Delta_i \leqslant 0$ $(i = r+1, \ldots, n)$. Следовательно, $v = v(0)$ — решение задачи (1), т. е. $\langle c, v\rangle = \inf\limits_U \langle c, u\rangle > -\infty$.

Случай II. Существует номер k $(r+1 \leqslant k \leqslant n)$ такой, что $\Delta_k > 0$ и $\gamma_{ik} \leqslant 0$ при всех $i = 1, \ldots, r$. Тогда, как и в § 3, нетрудно показать, что $\inf\limits_{U_s} \langle c, u\rangle = -\infty$. Поскольку в симплекс-таблице 5 угловой точки $v = v(0)$ в столбце переменной u^k находятся те же величины $\Delta_k > 0$, $\gamma_{ik} \leqslant 0$ $(i = 1, \ldots, r)$, то и в задаче (1) будем иметь $\inf\limits_U \langle c, u\rangle = -\infty$. Таким образом, в случае II задачи (1) и (2) не имеют решения.

Случай III. Существуют номера k, i $(r+1 \leqslant k \leqslant n,\ 1 \leqslant i \leqslant r)$ такие, что $\Delta_k > 0$, $\gamma_{ik} > 0$. Тогда множество $I_k = \{i\colon 1 \leqslant i \leqslant r,\ \gamma_{ik} > 0\}$ непусто и из условия

$$\min_{i \in I_k} \frac{v^i(\varepsilon)}{\gamma_{ik}} = \frac{v^s(\varepsilon)}{\gamma_{sk}}, \qquad s \in I_k, \tag{9}$$

где $v^i(\varepsilon)$ задано выражением (7), определяем номер s. Поскольку задача (2) невырожденная, то номер s и, следовательно, разрешающий элемент γ_{sk} из (9) определяются однозначно. Тогда, рассуждая так же, как в § 3, выведем из числа базисных переменную u^s, а переменную u^k введем в число базисных и придем к следующей угловой точке $w(\varepsilon) = (w^1(\varepsilon), \ldots, w^n(\varepsilon))$ с координатами

$$w^i(\varepsilon) = v^i(\varepsilon) - \frac{\gamma_{ik}}{\gamma_{sk}} v^s(\varepsilon) = v^i + \sum_{j=1}^{r} d_{ij}\varepsilon^j - \frac{\gamma_{ik}}{\gamma_{sk}}\left(v^s + \sum_{j=1}^{r} d_{sj}\varepsilon^j\right) =$$

$$= v^i - \frac{\gamma_{ik}}{\gamma_{sk}} v^s + \sum_{j=1}^{r}\left(d_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}} d_{sj}\right)\varepsilon^j > 0, \quad i = 1, \ldots, s-1, s+1, \ldots, r, \tag{10}$$

$$w^k(\varepsilon) = \frac{v^s(\varepsilon)}{\gamma_{sk}} = \frac{v^s}{\gamma_{sk}} + \sum_{j=1}^{r} \frac{d_{sj}}{\gamma_{sk}}\varepsilon^j,$$

$$w^i(\varepsilon) = 0,\ i = s,\ r+1, \ldots, k-1, k+1, \ldots, n.$$

Таким образом, в симплекс-таблице угловой точки $w(\varepsilon)$ в столбце свободных членов будут находиться числа

$$w^1 = v^1 - \gamma_{1k}\frac{v^s}{\gamma_{sk}}, \ldots, w^i = v^i - \gamma_{ik}\frac{v^s}{\gamma_{sk}}, \ldots, w^{s-1} = v^{s-1} - \gamma_{s-1,k}\frac{v^s}{\gamma_{sk}},$$

$$w^{s+1} = v^{s+1} - \gamma_{s+1,k}\frac{v^s}{\gamma_{sk}}, \ldots, w^r = v^r - \gamma_{rk}\frac{v^s}{\gamma_{sk}}, \quad w^k = \frac{v^s}{\gamma_{sk}},$$

а в j-м дополнительном столбце d'_j $(j = 1, \ldots, r)$ появятся величины

$$d'_{ij} = d_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}} d_{sj}, \quad i = 1, \ldots, s-1, s+1, \ldots, r, \quad d'_{kj} = \frac{d_{sj}}{\gamma_{sk}}. \tag{11}$$

Это значит, что дополнительные столбцы d_i симплекс-таблицы в задаче (2) преобразуются по тем же правилам, по которым преобразуются столбцы, соответствующие небазисным переменным, остающимися небазисными и

на следующем шаге симплекс-метода (ср. с (3.25)). Далее, из (8), (10) следует, что

$$J(w(\varepsilon)) = J(v(\varepsilon)) - \Delta_k \frac{v^s(\varepsilon)}{\gamma_{sk}} < J(v(\varepsilon)).$$

Наконец, исходя из равенств (6) и пользуясь формулами, аналогичными (3.21) — (3.23), нетрудно показать, что остальные элементы γ_{ij}, Δ_j симплекс-таблицы 6 при переходе к следующей угловой точке $w(\varepsilon)$ преобразуются по прежним же формулам (3.24) — (3.26). Описание одного шага симплекс-метода в задаче (2) закончено.

Теперь вернемся к задаче (1). В угловой точке $v = v(0)$ в качестве разрешающего элемента выберем тот же элемент γ_{sk}, который обеспечил выше переход от точки $v(\varepsilon)$ к точке $w(\varepsilon)$. Такой выбор разрешающего элемента в задаче (1) вполне возможен, так как в симплекс-таблице 5 угловой точки $v = v(0)$ в столбце переменной u^k находятся те же величины $\Delta_k > 0$, $\gamma_{ik} > 0$ при $i \in I_k$, $\gamma_{ik} \leqslant 0$ при $i \notin I_k$, что и в симплекс-таблице 6. Кроме того, ниже в лемме 2 будет показано, что для того же номера s, который был однозначно определен из условия (9), справедливо равенство (3.16).

Конечно, в отличие от условия (9) в условии (3.16) минимум может достигаться и на некоторых других, не совпадающих с s номерах из I_k, поскольку невырожденность задачи (1) мы здесь не предполагаем. С выбранным разрешающим элементом γ_{sk} сделаем один шаг симплекс-метода по формулам (3.17), (3.18), (3.24) — (3.26) и из точки $v = v(0)$ придем к следующей угловой точке w. Сравнение формул (3.17) и (10) показывает, что $w = w(0)$; ясно также, что расширенная симплекс-таблица угловой точки $w = w(0)$ может быть получена из симплекс-таблицы точки $w(\varepsilon)$ при $\varepsilon = 0$. Отсюда следует весьма важный для дальнейшего вывод: если в возмущенной задаче (2) с помощью симплекс-метода совершается переход от угловой точки $v(\varepsilon)$ к другой угловой точке $w(\varepsilon)$ с помощью разрешающего элемента γ_{sk} из (9), то в задаче (1) при выборе того же разрешающего элемента γ_{sk} произойдет переход от угловой точки $v = v(0)$ к угловой точке $w = w(0)$. Рассмотрение случая III закончено.

3. Теперь уже нетрудно дать обоснование антициклина, изложенного в конце § 3. Пусть $v_1 = (v_1^1, \ldots, v_1^n)$ — некоторая угловая точка множества U с базисом $A_{j_1}, \ldots, A_{j_r}$. Обозначим $B = (A_{j_1} \ldots A_{j_r})$, $\bar{v}_1 = (v_1^{j_1}, \ldots, v_1^{i_r})$. Согласно теореме 2.1

$$Av_1 = A_{j_1} v^{j_1} + \ldots + A_{j_r} v^{j_r} = B\bar{v}_1 = b, \quad v_1^j = 0, \quad j \neq j_l, \quad l = 1, \ldots, r,$$

так что $\bar{v}_1 = B^{-1}b \geqslant 0$.

Рассмотрим следующую возмущенную задачу:

$$\langle c, u\rangle \to \inf; \quad u \in U_\varepsilon = \left\{u\colon\ u \geqslant 0,\ Au = b(\varepsilon) = b + \sum_{i=1}^{r} \varepsilon^i A_{j_i}\right\}, \tag{12}$$

получающуюся из задачи (2) при $R = (A_{j_1} \ldots A_{j_r}) = B$. Возьмем точку $v_1(\varepsilon) = (v_1^1(\varepsilon), \ldots, v_1^n(\varepsilon))$ с координатами

$$v_1^{j_i}(\varepsilon) = v_1^{j_i} + \varepsilon^i, \quad v_1^j(\varepsilon) = 0, \quad j \neq j_i, \quad i = 1, \ldots, r.$$

Напоминаем, что здесь и в (12) ε^i — i-я степень числа $\varepsilon > 0$. Поскольку $v_1 \geqslant 0$, то ясно, что $v_1(\varepsilon) \geqslant 0$, причем $\bar{v}_1(\varepsilon) = (v^{j_1}(\varepsilon), \ldots, v^{j_r}(\varepsilon)) > 0$ при

всех $\varepsilon > 0$. Кроме того,

$$Av_1(\varepsilon) = \sum_{i=1}^{r} A_{j_i} v_1^{j_i}(\varepsilon) = \sum_{i=1}^{r} A_{j_i}\left(v^{j_i} + \varepsilon^i\right) = b + \sum_{i=1}^{r} A_{j_i}\varepsilon^i = b(\varepsilon).$$

Следовательно, $v_1(\varepsilon) \in U_\varepsilon$ при всех $\varepsilon > 0$, причем согласно теореме 2.1 $v_1(\varepsilon)$ — угловая точка множества U_ε с тем же базисом $A_{j_1}, \ldots, A_{j_r}$, что и точка v_1. Полезно также заметить, что $v_1 = v_1(0)$.

Таким образом, множество U_ε, образованное из множества U с помощью невырожденной матрицы $R = B$, непусто при всех $\varepsilon > 0$. Согласно лемме 1 тогда задача (12) невырожденная при каждом ε $(0 < \varepsilon < \varepsilon_1)$. Применим к задаче (12) симплекс-метод, беря в качестве начальной угловой точки множества U_ε введенную выше точку $v_1(\varepsilon)$ и считая, что $0 < \varepsilon < \varepsilon_1$. В силу невырожденности этой задачи мы получим конечную последовательность угловых точек $v_1(\varepsilon), \ldots, v_m(\varepsilon)$ из множества U_ε со значениями функции $J(v_1(\varepsilon)) > \ldots > J(v_m(\varepsilon))$, причем в последней точке реализуется либо случай I, либо случай II.

Из проведенного выше исследования следует, что если применить симплекс-метод к задаче (1), беря в качестве начальной точки v_1 и выбирая на каждом шаге те же разрешающие элементы, которые в задаче (12) привели к последовательности $v_1(\varepsilon), \ldots, v_m(\varepsilon)$, то получим последовательность угловых точек $v_1 = v_1(0), \ldots, v_m = v_m(0)$, причем в последней точке v_m будет реализовываться соответственно либо случай I, либо случай II. Выше было выяснено, что в случае I $v_m(\varepsilon)$ — решение задачи (12), а $v_m = v_m(0)$ — решение задачи (1), а в случае II задачи (12) и (1) не имеют решения.

Тем самым показано, что если задачу (1) решать симплекс-методом, выбирая на каждом шаге разрешающий элемент с помощью возмущенной задачи (12) указанным выше способом, то зацикливания не будет и за конечное число шагов этого метода выяснится, имеет ли задача (1) решение, и если имеет, то оно будет найдено.

4. Однако изложенный способ, позволяющий избежать зацикливания и имеющий принципиальное значение для обоснования симплекс-метода, пока что не слишком удобен для практического использования, поскольку требует рассмотрения возмущенной задачи (12) при достаточно малых и заранее неизвестных $\varepsilon > 0$. Нельзя ли все-таки определить разрешающий элемент на каждом шаге симплекс-метода, явно не прибегая к возмущенной задаче (12)? Оказывается, можно.

Чтобы понять, как это делается, вспомним, как выбирается разрешающий элемент γ_{sk} в задаче (2) (или (12)): сначала фиксируется номер k небазисной переменной с величиной $\Delta_k > 0$ (если таких k несколько, то можно из них выбрать наименьший), а затем из условия (9) однозначно определяется номер s и тем самым находится разрешающий элемент γ_{sk}. Важно заметить, что коэффициенты многочленов $v^i(\varepsilon)$, определяемые формулами (7), можно узнать, явно не привлекая возмущенную задачу (2),— эти коэффициенты расположены в столбце свободных членов и дополнительных столбцах расширенной симплекс-таблицы (см. табл. 4, 5). В частности, для начальной точки v_1 с базисом $A_{j_1}, \ldots, A_{j_r}$ в силу выбора матрицы $R = \left(A_{j_1} \ldots A_{j_r}\right) = B$ имеем $d_j = B^{-1}A_j = e_j$, что и отражено в табл. 4, а в дальнейшем при переходе от одной угловой точки к следующей величины j-го дополнительного столбца преобразуются по формулам (11), аналогичным формулам (3.25).

Таким образом, коэффициенты многочленов

$$\frac{v_i(\varepsilon)}{\gamma_{ik}} \equiv f_i(\varepsilon) = c_{i0} + c_{i1}\varepsilon + \ldots + c_{ir}\varepsilon^r$$

на каждом шаге симплекс-метода могут быть определены без явного привлечения возмущенной задачи по формулам: $c_{i0} = v^i/\gamma_{ik}$, $c_{ij} = d_{ij}/\gamma_{ik}$

($j = 1, \ldots, r$). Заметим, что все эти многочлены различны. В самом деле, если бы $f_i(\varepsilon) \equiv f_m(\varepsilon)$ при $i \neq m$, то коэффициенты $\{c_{ij}\}$ и $\{c_{mj}\}$ были бы равны: $c_{ij} = c_{mj}$ или $d_{ij} = d_{mj}\gamma_{ik}/\gamma_{mk}$ ($j = 1, \ldots, r$). Это значило бы, что пропорциональны i-я и m-я строки в матрице $D = (d_1 \ldots d_r)$ из дополнительных столбцов расширенной симплекс-таблицы. Однако это невозможно, так как на каждом шаге матрица D невырожденная — она является произведением невырожденных матриц вида $\left(A_{j_1} \ldots A_{j_r}\right)^{-1}$ на R.

Таким образом, приходим к следующей задаче: задано конечное число различных многочленов $f_i(\varepsilon) = v^i(\varepsilon)/\gamma_{ik}$ ($i \in I_k$) и требуется найти $\min\limits_{i \in I_k} f_i(\varepsilon)$ при достаточно малом, но заранее неизвестном $\varepsilon > 0$. Здесь нам будет полезна

Лемма 2. *Пусть дано множество многочленов*

$$f_i(\varepsilon) = c_{i0} + c_{i1}\varepsilon + \ldots + c_{ir}\varepsilon^r, \quad 0 < \varepsilon < \varepsilon_1, \quad i \in S_0,$$

степени не выше r, среди которых нет совпадающих; S_0 — конечное множество номеров. Тогда существует число ε_2 ($0 < \varepsilon_2 \leqslant \varepsilon_1$) такое, что $\min\limits_{i \in S_0} f_i(\varepsilon)$ будет достигаться на единственном многочлене, номер которого не зависит от ε ($0 < \varepsilon < \varepsilon_2$). Для определения номера этого многочлена нужно рекуррентным образом строить множества $S_0, S_1 = \left\{s\colon s \in S_0,\ c_{s0} = \min\limits_{i \in S_0} c_{i0}\right\}, \ldots$ $\ldots, S_{m+1} = \left\{s\colon s \in S_m, c_{sm} = \min\limits_{i \in S_m} c_{im}\right\}, \ldots$ до тех пор, пока не будет обнаружено множество S_l ($0 \leqslant l \leqslant r+1$), состоящее из единственного номера s, который и будет искомым номером.

Доказательство. Если множество S_0 состоит из единственного номера s, то утверждение леммы тривиально. Если же S_0 состоит более чем из одного номера, то перейдем к рассмотрению множества S_1. Из определения множества S_1 следует, что все многочлены $f_s(\varepsilon)$ с номерами $s \in S_1$ имеют один и тот же младший коэффициент $c_{s0} = \min\limits_{i \in S_0} c_{i0}$. Возможно, что $S_1 = S_0$. Тогда положим $\varepsilon_{01} = \varepsilon_1$ и заметим, что $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_1} f_i(\varepsilon)$ при всех ε ($0 < \varepsilon < \varepsilon_{01}$). Если $S_1 \neq S_0$, т. е. $S_0 \setminus S_1 \neq \varnothing$, то положим

$$\varepsilon_{01} = \min\left\{1;\ \varepsilon_1;\ \frac{1}{2c}\left(\min_{i \in S_0 \setminus S_1} c_{i0} - c_{s0}\right),\ s \in S_1\right\} > 0,$$

где $c = \max\limits_{i \in S_0} \sum\limits_{j=1}^{r} |c_{ij}|$. Тогда для всех $s \in S_1$, $p \in S_0 \setminus S_1$, $0 < \varepsilon < \varepsilon_{01}$ имеем $f_s(\varepsilon) \leqslant c_{s0} + c\varepsilon < c_{p0} - c\varepsilon \leqslant f_p(\varepsilon)$, т. е. $f_s(\varepsilon) < f_p(\varepsilon)$. Это значит, что $\min\limits_{i \in S_0} f_i(\varepsilon)$ при всех ε ($0 < \varepsilon < \varepsilon_{01}$) может достигаться лишь на многочленах с номерами $s \in S_1$, т. е. $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_1} f_i(\varepsilon) < f_p(\varepsilon)$ при всех ε ($0 < \varepsilon < \varepsilon_{01}$) и всех $p \notin S_1$. В частности, если S_1 состоит из единственного номера s, то этот номер будет искомым для всех ε ($0 < \varepsilon < \varepsilon_{01} = \varepsilon_2$) и лемма будет доказана. Если же S_1 состоит более чем из одного номера, то перейдем к рассмотрению множества S_2 и т. д.

Пусть уже рассмотрены множества $S_0 \supseteq S_1 \supseteq \ldots \supseteq S_m$ ($1 \leqslant m \leqslant r$), найдено число $\varepsilon_{0m} > 0$ и показано, что: 1) $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_m} f_i(\varepsilon) < f_p(\varepsilon)$ для всех ε ($0 < \varepsilon < \varepsilon_{0m}$) и всех $p \notin S_m$; 2) все многочлены $f_s(\varepsilon)$ с номерами

$s \in S_m$ имеют одинаковые коэффициенты при степенях $\varepsilon^0, \varepsilon^1, \ldots, \varepsilon^{m-1}$; 3) S_m состоит более чем из одного номера.

Рассмотрим множество $S_{m+1} = \left\{s\colon s \in S_m, c_{sm} = \min\limits_{i \in S_m} c_{im}\right\}$. Отсюда с учетом индуктивного предположения 2) заключаем, что все многочлены $f_s(\varepsilon)$ с номерами $s \in S_{m+1}$ имеют одинаковые коэффициенты при степенях $\varepsilon^0, \varepsilon^1, \ldots, \varepsilon^m$. Возможно, что $S_{m+1} = S_m$. Тогда положим $\varepsilon_{0,\,m+1} = \varepsilon_{0m} > 0$ и с учетом индуктивного предположения 1) заметим, что $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_{m+1}} f_i(\varepsilon) < f_p(\varepsilon)$ при всех $0 < \varepsilon < \varepsilon_{0,\,m+1}$, $p \notin S_{m+1}$.

Если же $S_{m+1} \neq S_m$, то положим

$$\varepsilon_{0,m+1} = \min\left\{\varepsilon_{0m};\ \frac{1}{2c}\left(\min_{i \in S_m \setminus S_{m+1}} c_{im} - c_{sm}\right),\ s \in S_{m+1}\right\} > 0.$$

Тогда для всех $s \in S_{m+1}$, $p \in S_m \setminus S_{m+1}$, $0 < \varepsilon < \varepsilon_{0,\,m+1}$ имеем

$$f_s(\varepsilon) \leqslant c_{s0} + c_{s1}\varepsilon + \ldots + c_{s,\,m-1}\varepsilon^{m-1} + \varepsilon^m(c_{sm} + \varepsilon c) <$$

$$< c_{s0} + c_{s1}\varepsilon + \ldots + c_{s,\,m-1}\varepsilon^{m-1} + \varepsilon^m(c_{pm} - \varepsilon c) = c_{p0} + c_{p1}\varepsilon + \ldots$$

$$\ldots + c_{p,\,m-1}\varepsilon^{m-1} + c_{pm}\varepsilon^m - \varepsilon^{m+1}c \leqslant f_p(\varepsilon),$$

т. е. $f_s(\varepsilon) < f_p(\varepsilon)$. Это значит, что $\min\limits_{i \in S_m} f_i(\varepsilon)$ при всех ε $(0 < \varepsilon < \varepsilon_{0,\,m+1})$ может достигаться лишь на многочленах с номерами $s \in S_{m+1}$, т. е. $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_{m+1}} f_i(\varepsilon) < f_p(\varepsilon)$ при всех ε $(0 < \varepsilon < \varepsilon_{0,\,m+1})$ и всех $p \notin S_{m+1}$.

Если S_{m+1} состоит из единственного номера s, то он и будет искомым номером для всех ε $(0 < \varepsilon < \varepsilon_{0,\,m+1} = \varepsilon_2)$ и лемма будет доказана. Если же S_{m+1} состоит более чем из одного номера и $m < r$, то переходим к рассмотрению множества S_{m+2} и т. д. В самом худшем случае мы доберемся до последнего множества $S_{r+1} = \left\{s\colon s \in S_r, c_{sr} = \min\limits_{i \in S_r} c_{ir}\right\}$, зная, что $\min\limits_{i \in S_0} f_i(\varepsilon) = \min\limits_{i \in S_{r+1}} f_i(\varepsilon) < f_p(\varepsilon)$, $p \notin S_{r+1}$, $0 < \varepsilon < \varepsilon_{0,\,r+1}$, и что все многочлены $f_s(\varepsilon)$ при $s \in S_{r+1}$ имеют одинаковые коэффициенты при степенях $\varepsilon^0, \varepsilon^1, \ldots, \varepsilon^r$. Однако по условию леммы среди многочленов $\{f_i(\varepsilon),\ i \in S_0\}$ нет двух одинаковых. Следовательно, множество S_{r+1} будет состоять из одного номера s. Остается принять $\varepsilon_2 = \varepsilon_{0,\,r+1}$.

Если в лемме 2 взять $f_i(\varepsilon) = v^i(\varepsilon)/\gamma_{ik}$ $(i \in I_k = S_0)$ и применить изложенное в ней правило определения номера s, на котором достигается $\min\limits_{i \in I_k} v^i(\varepsilon)/\gamma_{ik}$, то мы получим именно тот самый антициклин, который был описан в конце § 3. Этот антициклин в литературе часто называют *лексикографическим правилом выбора разрешающего элемента*. Поясним это название.

Определение 1. Вектор $a = (a^1, \ldots, a^m)$ называют *лексикографически положительным* и пишут $a \succ 0$, если $a \neq 0$ и первая из отличных от нуля координат вектора a положительна. Вектор $a \in E^m$ называют *лексикографически большим* вектора $b \in E^m$ и обозначают $a \succ b$, если $a - b \succ 0$.

Другими словами, $a \succ b$ означает, что $a^1 > b^1$ или же $a^1 = b^1$, но $a^2 > b^2$, или же $a^1 = b^1$, $a^2 = b^2$, но $a^3 > b^3$ и т. д. Для любых a, $b \in E^m$ выполнено одно и только одно из соотношений: $a \succ b$, $b \succ a$ или $a = b$. Ясно, что если $a \succ b$, $b \succ c$, то $a \succ c$. Упорядочение множества векторов в их лексикографическом убывании вполне аналогично упорядочению слов в словарях, что

и объясняет присутствие слова «лексикографический» в приведенном определении.

Нетрудно видеть, что в лемме 2 с помощью цепочки множеств $S_0 \supseteq S_1 \supseteq \supseteq \ldots \supseteq S_m \supseteq \ldots \supseteq S_l$ приводится лексикографическое упорядочение векторов $\{c_i = (c_{i0}, c_{i1}, \ldots, c_{ir}),\ i \in S_0\}$: $c_p \succ c_i$ для всех $i \in S_m$, $p \notin S_m$, а вектор c_s с номером s из одноэлементного множества S_l представляет собой лексикографический минимум на множестве $\{c_i,\ i \in S_0\}$, т. е. $c_p \succ c_s$ для всех $p \in S_0$, $p \neq s$. Таким образом, согласно лемме 2 минимум конечного числа различных многочленов при всех достаточно малых положительных значениях аргумента достигается на том многочлене, вектор из коэффициентов которого является лексикографически наименьшим среди всех векторов из коэффициентов рассматриваемых многочленов.

§ 5. Выбор начальной угловой точки

Выше при описании симплекс-метода для канонической задачи

$$J(u) = \langle c,\ u\rangle \to \inf;\quad u \in U = \{u \in E^n:\ u \geqslant 0,\ Au = b\} \tag{1}$$

мы предполагали, что нам уже известна некоторая угловая точка множества U и что система $Au = b$ уже записана в виде (3.5), где $r = \operatorname{rang} A$. Возникают вопросы: как узнать, не пусто ли множество

$$U = \{u \in E^n:\ u \geqslant 0,\ Au = b\}, \tag{2}$$

где A — матрица размера $m \times n$, b — вектор из E^m? Если это множество непусто, то имеет ли оно хотя бы одну угловую точку и как найти ее? Как исключить из системы $Au = b$ линейно зависимые уравнения и привести ее к виду (3.5)? Ниже будут даны ответы на все эти вопросы. Замечательно то, что для ответа на них может быть использован симплекс-метод.

1. Можем считать, что $b \geqslant 0$, так как если $b^i < 0$ при некотором i $(1 \leqslant i \leqslant m)$, то соответствующее i-е уравнение $(Au)^i = b^i$ из (2) можно умножить на -1. Наряду с основными переменными $u = (u^1, \ldots, u^n)$ введем вспомогательные (искусственные) переменные $w = (u^{n+1}, \ldots, u^{n+m})$ и в пространстве E^{n+m} переменных $z = (u^1, \ldots, u^n, \ldots, u^{n+m}) = (u, w)$ рассмотрим следующую каноническую задачу линейного программирования:

$$\begin{gathered} J_1(z) = u^{n+1} + \ldots + u^{n+m} \to \inf; \\ z \in Z = \left\{z:\ z = \begin{bmatrix} u \\ w \end{bmatrix} \in E^{n+m},\ z \geqslant 0,\ Cz \equiv Au + w = b\right\}. \end{gathered} \tag{3}$$

Систему $Cz \equiv Au + w = b$ перепишем в покоординатной форме

$$\begin{aligned} a_{11}u^1 + \ldots + a_{1n}u^n + u^{n+1} &= b^1, \\ a_{21}u^1 + \ldots + a_{2n}u^n + u^{n+2} &= b^2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}u^1 + \ldots + a_{mn}u^n + u^{n+m} &= b^m. \end{aligned} \tag{4}$$

При $u^{n+1}=\ldots=u^{n+m}=0$ система (4) очевидно равносильна системе $Au=b$. Множество Z непусто: оно содержит, например, точку $z_0=(0,\ b)\geqslant 0$. Более того, нетрудно видеть, что точка z_0 является угловой точкой множества Z с базисом из последних m столбцов $e_1,\ \ldots,\ e_m$ матрицы $C=(A,\ E)$, где E — единичная матрица размера $m\times m$ $(m=\operatorname{rang} C\geqslant \operatorname{rang} A=r)$. Важно также заметить, что из системы (4) легко выразить базисные переменные $w=(u^{n+1},\ \ldots,\ u^{n+m})$ угловой точки z_0 через небазисные:

$$w=b-Au=b-A_1u^1-\ldots-A_nu^n.$$

Таким образом, задача (3) уже записана в форме, удобной для применения симплекс-метода с начальной угловой точкой z_0.

Поскольку $J_1(z)\geqslant 0$ при всех $z\in Z$, то случай $\inf\limits_Z J_1(z)=-\infty$ здесь невозможен. Поэтому, взяв в качестве начальной точку z_0, с помощью симплекс-метода, описанного в § 3, за конечное число шагов найдем угловую точку $z_*=(v_1,\ w_1)$ множества Z, являющуюся решением задачи (3): $v_1=(v_1^1,\ldots,v_1^n)$, $w_1=(v_1^{n+1},\ldots,v_1^{n+m})$, $\inf\limits_Z J_1(z)=J_1(z_*)=v_1^{n+1}+\ldots+v_1^{n+m}\geqslant 0$.

Имеются две возможности: или $J_1(z_*)>0$, или $J_1(z_*)=0$. Если $J_1(z_*)>0$, то $w_1\neq 0$ и, оказывается, множество U, определяемое условиями (2), будет пустым. В самом деле, если бы существовала хотя бы одна точка $u\in U$, то точка $z=(u,\ 0)$ принадлежала бы множеству Z и, кроме того, мы имели бы $J(z)=0$, что противоречит тому, что $J(z)\geqslant J(z_*)>0$ при всех $z\in Z$. Таким образом, при $J_1(z_*)>0$ множество U пусто и задача (1) не имеет смысла.

Пусть теперь $J_1(z_*)=v_1^{n+1}+\ldots+v_1^{n+m}=0$. Тогда $w_1=(v_1^{n+1},\ldots,v_1^{n+m})=0$, $z_*=(v_1,\ 0)$. Кроме того, по построению $z_*=(v_1,\ 0)$— угловая точка множества Z. Покажем, что тогда v_1 — угловая точка множества (2). Прежде всего ясно, что из $z_*\geqslant 0$ следует $v_1\geqslant 0$, а из $Cz_*=b$ имеем $Av_1=b$. Это значит, что $v_1\in U$.

Рассмотрим представление

$$v_1=\alpha u_1+(1-\alpha)u_2,\quad 0<\alpha<1,\quad u_1\in U,\quad u_2\in U, \tag{5}$$

и покажем, что оно возможно лишь при $v_1=u_1=u_2$. Точки $z_1=(u_1,\ 0)$, $z_2=(u_2,\ 0)$, очевидно, принадлежат Z. Тогда (5) можно переписать в виде $z_*=\alpha z_1+(1-\alpha)z_2$ $(0<\alpha<1)$. Но z_* — угловая точка множества Z. Поэтому последнее равенство для z_* возможно лишь при $z_*=z_1=z_2$. Отсюда следует, что $v_1=u_1=u_2$. Таким образом, v_1 — угловая точка множества U. Тем самым доказана важная

Теорема 1. *Если множество* (2) *непусто, то оно имеет хотя бы одну угловую точку.*

2. Итак, исходная угловая точка v_1 множества U найдена. Для того чтобы, отправляясь от этой точки, можно было решать задачу (1) с помощью симплекс-метода, остается еще найти $\operatorname{rang} A$, указать базис точки v_1 и записать систему $Au=b$ в виде, аналогичном (3.5), т. е. выразить базисные переменные через небазисные. Как это сделать? Обратимся к симплекс-таблице 7 точки z_*, которая получена при решении задачи (3) симплекс-методом.

Поскольку $m=\operatorname{rang} C \geqslant \operatorname{rang} A$, то в базис точки z_* могут входить не только столбцы матрицы A, но и столбцы матрицы E (кстати, при $m>\operatorname{rang} A$ в базис обязательно войдут столбцы матрицы E). Не ограничивая общности можем считать, что базисом точки z_* являются столбцы $A_1, \ldots, A_r, e_1, \ldots, e_{m-r}$ матрицы $C=(A, E)$, соответствующие основным переменным $u^1, \ldots, u^r$ и вспомогательным переменным $u^{n+1}, \ldots, u^{n+m-r}$ — именно эта ситуация отражена в симплекс-таблице 7. Подчеркнем еще раз, что из $z_*=(v_1, 0)$ следует, что $v_1^{n+1}=\ldots=v_1^{n+m}=0$ — эти равенства также нашли отражение в столбце свободных членов табл. 7. Нижняя строка симплекс-таблицы 7, содержащая величины $\Delta_i \geqslant 0$, $J_1(z_*)=0$, характеризующие решение z_* задачи (3), опущена — это обстоятельство не повлияет на дальнейшие преобразования табл. 7. Рассмотрим возможные здесь случаи.

Случай I. В табл. 7 отсутствуют строки, соответствующие вспомогательным переменным $u^{n+1}, \ldots, u^{n+m}$, т. е. $r=m \leqslant n$ и столбцы $A_1, \ldots, A_r$ составляют базис точки $z_*=(v_1, 0)$. Тогда согласно правилу составления симплекс-таблиц j-й строке табл. 7

Базисные переменные	u^1	...	u^j	...	u^r	u^{r+1}	...	u^k	...	u^n
u^1	1	...	0	...	0	$\gamma_{1,r+1}$	...	γ_{1k}	...	γ_{1n}
...	...	...	...	...	...	...	...	...	...	...
u^j	0	...	1	...	0	$\gamma_{j,r+1}$	...	γ_{jk}	...	γ_{jn}
...	...	...	...	...	...	...	...	...	...	...
u^r	0	...	0	...	1	$\gamma_{r,r+1}$	...	γ_{rk}	...	γ_{rn}
u^{n+1}	0	...	0	...	0	$\gamma_{n+1,r+1}$	...	$\gamma_{n+1,k}$	...	$\gamma_{n+1,n}$
...	...	...	...	...	...	...	...	...	...	...
u^{n+i}	0	...	0	...	0	$\gamma_{n+i,r+1}$	...	$\gamma_{n+i,k}$	...	$\gamma_{n+i,n}$
...	...	...	...	...	...	...	...	...	...	...
u^{n+m-r}	0	...	0	...	0	$\gamma_{n+m-r,r+1}$	...	$\gamma_{n+m-r,k}$	...	$\gamma_{n+m-r,n}$

будет соответствовать уравнение

$$u^j + \gamma_{j,r+1}u^{r+1} + \ldots + \gamma_{jn}u^n + \gamma_{j,n+1}u^{n+1} + \ldots + \gamma_{j,n+m}u^{n+m} = v_1^j, \tag{6}$$

$$j = 1, \ldots, r = m.$$

Пользуясь этим уравнением, исключим вспомогательные переменные $u^{n+1}, \ldots, u^{n+m}$ из дальнейших рассмотрений. Для этого заметим, что система (6) является другой равносильной формой записи системы (4) — система (6) может быть получена из $Au + w = b$ умножением слева на матрицу B^{-1}, обратную матрице $B = (A_1, \ldots, A_r)$. Поэтому, положив в уравнении (6) $u^{n+1} = \ldots = u^{n+m} = 0$, придем к системе

$$u^j + \gamma_{j,r+1}u^{r+1} + \ldots + \gamma_{jn}u^n = v_1^j, \qquad j = 1, \ldots, r = m, \tag{7}$$

которая также может быть получена из $Au = b$ умножением слева на B^{-1} и поэтому равносильна системе $Au = b$. Это значит, что угловая точка v_1 имеет своим базисом столбцы $A_1, \ldots, A_r$ $(r = \operatorname{rang} A = m)$, а система $Au = b$ уже записана в виде (7), удобном для применения симплекс-метода. Резюмируя, можно сказать, что если в табл. 7 отсутствуют строки, соответствующие вспомогательным переменным, то для получения симплекс-таблицы угловой точки v_1 нужно вычеркнуть из табл. 7 все столбцы для $u^{n+1}, \ldots, u^{n+m}$ и добавить нижнюю строку из $\Delta_1 = \ldots = \Delta_r = 0$, $\Delta_i = \sum_{s=1}^{r} c_s\gamma_{si} - c_i$, $i = r+1, \ldots, n$, $J(u_*)$.

Таблица 7

u^{n+1}	...	u^{n+i}	...	u^{n+m-r}	$u^{n+m-r+1}$	...	u^{n+m}	Свободные члены
0	...	0	...	0	$\gamma_{1,n+m-r+1}$	...	$\gamma_{1,n+m}$	$v_1^1 \geqslant 0$
...	...	...	...	...	...	...	...	...
0	...	0	...	0	$\gamma_{j,n+m-r+1}$	...	$\gamma_{j,n+m}$	$v_1^j \geqslant 0$
...	...	...	...	...	...	...	...	...
0	...	0	...	0	$\gamma_{r,n+m-r+1}$	...	$\gamma_{r,n+m}$	$v_1^r \geqslant 0$
1	...	0	...	0	$\gamma_{n+1,n+m-r+1}$	...	$\gamma_{n+1,n+m}$	$v_1^{n+1} = 0$
...	...	...	...	...	...	...	...	...
0	...	1	...	0	$\gamma_{n+i,n+m-r+1}$	...	$\gamma_{n+i,n+m}$	$v_1^{n+i} = 0$
...	...	...	...	...	...	...	...	...
0	...	0	...	1	$\gamma_{n+m-r,n+m-r+1}$		$\gamma_{n+m-r,n+m}$	$v_1^{n+m} = 0$

Случай II. В табл. 7 имеются строки, соответствующие вспомогательным переменным, т. е. $r < m$, и среди коэффициентов $\gamma_{n+p,j}$ $(p = 1, \ldots, m-r,\ j = r+1, \ldots, n)$ имеются положительные. Пусть, например, $\gamma_{n+i,k} > 0$ $(1 \leqslant i \leqslant m-r,\ r+1 \leqslant k \leqslant n)$. Тогда

$$I_k = \{j\colon\ 1 \leqslant j \leqslant r \text{ или } n+1 \leqslant j \leqslant n+m-r,\ \gamma_{jk} > 0\} \neq \varnothing.$$

Поскольку $v_1^j/\gamma_{jk} \geqslant 0$ для всех $j \in I_k$, а $v_1^{n+i}/\gamma_{n+i,k} = 0$, то

$$\min_{j \in I_k} v_1^j/\gamma_{ik} = v_1^{n+i}/\gamma_{n+i,k} = 0.$$

Это значит, что любой коэффициент $\gamma_{n+i,k} > 0$ $(1 \leqslant i \leqslant m-r,\ r+1 \leqslant k \leqslant n)$ можно взять за разрешающий элемент симплекс-таблицы 7 и с его помощью вывести вспомогательную переменную u^{n+i} из базисных и вместо нее ввести в базисные основную переменную u^k. Из формул преобразования, аналогичных формулам (3.17), (3.24), (3.25), нетрудно увидеть, что из-за $v_1^{n+i} = 0$ в результате такого преобразования симплекс-таблицы мы останемся в той же угловой точке z_*, но ее базис $A_1, \ldots, A_r, e_1, \ldots, e_{m-r}$ заменится новым базисом $A_1, \ldots, A_r, e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_{m-r}, A_k$. Если и в новой симплекс-таблице, соответствующей новому базису точки z_*, на пересечении столбцов $u^{r+1}, \ldots, u^{k-1}, u^{k+1}, \ldots, u^n$ со строками, отвечающими вспомогательным переменным, еще останутся положительные коэффициенты, то, взяв любой из них за разрешающий элемент, можно вывести из базиса еще один из вспомогательных столбцов $e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_{m-r}$ и заменить его столбцом исходной матрицы A и т. д.

Может оказаться, что за конечное число таких операций нам удастся вывести из базиса точки z_* все вспомогательные столбцы $e_1, \ldots, e_{m-r}$ — это будет означать, что после упомянутых операций мы пришли к симплекс-таблице, соответствующей рассмотренному выше первому случаю при $m = \operatorname{rang} A$. Может, однако, получиться и так (при $\operatorname{rang} A < m$ так это и будет), что в очередной симплекс-таблице на пересечении столбцов основных переменных, не являющихся базисными для точки z_*, со строками, соответствующими вспомогательным базисным переменным, не найдется ни одного положительного коэффициента, и тогда описанный процесс вывода столбцов $\{e_i\}$ из базиса оборвется. Это будет означать, что реализовался случай III, к рассмотрению которого мы сейчас переходим.

Случай III. В табл. 7 имеются строки, соответствующие вспомогательным переменным, т. е. $r < m$, но $\gamma_{n+p,j} \leqslant 0$ при всех $p = 1, \ldots, m-r,\ j = r+1, \ldots, n$. Согласно правилу составления симплекс-таблиц строке переменной u^j $(j = 1, \ldots, r)$

в табл. 7 соответствует уравнение

$$u^j + \gamma_{j,r+1}u^{r+1} + \ldots + \gamma_{jn}u^n + \ldots + \gamma_{j,n+m-r+1}u^{n+m-r+1} + \ldots$$
$$\ldots + \gamma_{j,n+m}u^{n+m} = v_1^j, \quad j = 1, \ldots, r, \quad (8)$$

а строке переменной u^{n+i} — уравнение

$$\gamma_{n+i,r+1}u^{r+1} + \ldots + \gamma_{n+i,n}u^n + u^{n+i} + \gamma_{n+i,n+m-r+1}u^{n+m-r+1} + \ldots$$
$$\ldots + \gamma_{n+i,n+m}u^{n+m} = v_1^{n+i} = 0, \quad i = 1, \ldots, m-r. \quad (9)$$

Заметим, что система (8), (9) может быть получена из $Au + w = b$ умножением слева на матрицу, обратную матрице $(A_1, \ldots, A_m, e_1, \ldots, e_{m-r})$, где $A_1, \ldots, A_r, e_1, \ldots, e_{m-r}$ — базис точки z_*, поэтому системы (4) и (8), (9) равносильны. С другой стороны, напоминаем, что при $u^{n+1} = \ldots = u^{n+m} = 0$ система (4) превращается в систему $Au = b$. Поэтому, полагая в (8), (9) $u^{n+1} = \ldots = u^{n+m} = 0$, получаем систему

$$u^j + \gamma_{j,r+1}u^{r+1} + \ldots + \gamma_{jk}u^k + \ldots + \gamma_{jn}u^n = v_1^j, \quad j = 1, \ldots, r, \quad (10)$$

$$\gamma_{n+i,r+1}u^{r+1} + \ldots + \gamma_{n+i,k}u^k + \ldots + \gamma_{n+i,n}u^n = 0, \quad i = 1, \ldots, m-r, \quad (11)$$

которая также равносильна системе $Au = b$.

Заметим, что в рассматриваемом случае в (11) все $\gamma_{n+i,j} \leqslant 0$ $(j = r+1, \ldots, n, i = 1, \ldots, m-r)$. Возможно, что в некотором из уравнений (11), например, в $n+i$-м уравнении $\gamma_{n+i,j} = 0$ при всех $j = r+1, \ldots, n$. Ясно, что такие уравнения будут тривиально удовлетворяться при всех $u \in E^n$ и их можно без ущерба вычеркнуть из системы (10), (11). Поэтому можем считать, что в каждом из уравнений (11) имеется хотя бы один отрицательный коэффициент.

Пусть в $n+i$-м уравнении (11) $\gamma_{n+i,s} < 0, \ldots, \gamma_{n+i,s_q} < 0$, а остальные коэффициенты равны нулю. Тогда это уравнение запишется в виде

$$\gamma_{n+i,s_1}u^{s_1} + \ldots + \gamma_{n+i,s_q}u^{s_q} = 0.$$

Поскольку $u^{s_j} \geqslant 0$ для всех $u \in U$, а $\gamma_{n+i,s_j} < 0$ $(j = 1, \ldots, q)$, то последнее уравнение будет удовлетворяться только при $u^{s_1} = 0, \ldots, u^{s_q} = 0$.

Координату u^k $(r+1 \leqslant k \leqslant n)$ назовем *отмеченной*, если $\gamma_{n+i,k} < 0$ хотя бы для одного номера i $(1 \leqslant i \leqslant m-r)$. Таким образом, показано, что у системы (10), (11) и, следовательно, у равносильной ей исходной системы $Au = b$ возможны лишь такие неотрицательные решения, у которых отмеченные координаты равны нулю. Поскольку отмеченные координаты уже од-

нозначно определены (они равны нулю), то их можно исключить из дальнейшего рассмотрения. Для этого в (10), (11), а также в (1) нужно вычеркнуть все коэффициенты c_j, a_{ij}, γ_{ij}, которые умножаются на отмеченные координаты. Кстати, после исключения отмеченных координат все уравнения (11) в рассматриваемом случае превратятся в тривиальные тождества, и их также следует вычеркнуть.

В результате вместо исходной задачи (1) получим новую каноническую задачу линейного программирования: минимизировать функцию

$$\sum_{j=1}^{r} c_i u^i + \sum_{i \in I} c_i u^i \tag{12}$$

при ограничениях

$$u^1 \geqslant 0, \ \ldots, \ u^r \geqslant 0; \quad u^i \geqslant 0, \quad i \in I, \tag{13}$$

$$u^j + \sum_{k \in I} \gamma_{jk} u^k = v_1^j, \qquad j = 1, \ldots, r, \tag{14}$$

где I — множество номеров тех из координат $u^{r+1}, \ldots, u^n$, которые не являются отмеченными. Задача (12)—(14) уже приведена к виду, удобному для использования симплекс-метода: исходная угловая точка $(v_1^1, \ldots, v_1^r, v_1^i = 0 \quad (i \in I))$ известна, r — ранг матрицы ограничений (14), базисные переменные $u^1, \ldots, u^r$ легко выражаются из (14) через небазисные.

Решая задачу (12)—(14) симплекс-методом, можно найти ее решение $u_*^1, \ldots, u_*^r, u_*^i \quad (i \in I)$. Добавив сюда нулевые значения отмеченных координат, получим решение исходной задачи (1).

Рассмотрение всех трех возможных случаев закончено. Заметим лишь, что хотя каждый из этих случаев формально охватывает возможность $r = n \leqslant m$, следует ее выделить особо. Из симплекс-таблицы 7 при $r = n$, полагая $u^{n+1} = \ldots = u^{n+m} = 0$, получаем $u^1 = v_1^1, \ldots, u^n = v_1^n$. Это значит, что при $r = n$ множество U состоит из одной точки $u = v_1$ и задача (1) становится малосодержательной.

Из приведенного анализа вытекает следующее полезное правило заполнения симплекс-таблиц задачи (3): если в процессе применения симплекс-метода к этой задаче какая-либо вспомогательная переменная u^{n+i} перешла в небазисные, то в дальнейшем из столбца u^{n+i} разрешающий элемент брать не следует; более того, поскольку в дальнейшем все равно будет принято $u^{n+i} = 0$, то ясно, что столбец для u^{n+i} не окажет никакого влияния на окончательную таблицу с основными переменными и поэтому его можно сразу же вычеркнуть из симплекс-таблицы. Кстати, в табл. 7 столбцы для вспомогательных переменных $u^{n+m-r+1}, \ldots, u^{n+m}$ приведены лишь для пояснения дальнейших

преобразований — их нужно было вычеркивать по мере того, как соответствующая вспомогательная переменная переходила из базисной в небазисную. Заметим, что в симплекс-таблицах столбцы, соответствующие базисным переменным, также часто вычеркивают, так как в каждом таком столбце всегда находится одна и та же заранее известная информация: в строке, номер которой равен номеру базисной переменной, находится единица, а все остальные элементы такого столбца равны нулю.

Таким образом, метод построения начальной угловой точки и ее симплекс-таблицы для канонической задачи описан. Симплекс-метод для решения таких задач полностью обоснован. Существуют и другие методы поиска начальной угловой точки.

3. Для иллюстрации изложенного выше приведем один пример.

Пример 1. Минимизировать функцию

$$J(u) = u^1 + 3u^2 + 2u^3 + u^4 - 3u^5 \tag{15}$$

на множестве U, определяемом условиями

$$u^1 \geqslant 0, \ldots, u^5 \geqslant 0; \tag{16}$$

$$\begin{aligned} u^1 + u^2 - 4u^3 + u^4 - 3u^5 &= 3, \\ u^1 \qquad - 4u^3 + 2u^4 - 5u^5 &= 6, \\ -2u^1 - 5u^2 + 8u^3 + u^4 \qquad &= 3. \end{aligned} \tag{17}$$

Для определения начальной угловой точки множества U введем вспомогательные переменные u^6, u^7, u^8 и в пространстве E^8 переменных $z = (u^1, \ldots, u^8)$ рассмотрим задачу: минимизировать функцию

$$J_1(z) = u^6 + u^7 + u^8 \tag{18}$$

на множестве Z, определяемом условиями

$$u^1 \geqslant 0, \ldots, u^8 \geqslant 0; \tag{19}$$

$$\begin{aligned} u^1 + u^2 - 4u^3 + u^4 - 3u^5 + u^6 &= 3, \\ u^1 \qquad - 4u^3 + 2u^4 - 5u^5 + u^7 &= 6, \\ -2u^1 - 5u^2 + 8u^3 + u^4 \qquad + u^8 &= 3. \end{aligned} \tag{20}$$

Заполним симплекс-таблицу 8 для угловой точки $z_0 = (0, 0, 0, 0, 0, 3, 6, 3)$ множества Z; в этой таблице столбцы, соответствующие базисным переменным u^6, u^7, u^8, опущены; в нижней строке выписаны коэффициенты функции (18) после замены базисных переменных u^6, u^7, u^8 через небазисные согласно равенствам (20) — эти коэффициенты, очевидно, получаются суммированием величин в соответствующих столбцах небазисных переменных $u^1, \ldots, u^5$ точки z_0. В табл. 8 в качестве разрешающего элемента возьмем величину 2 из столбца u^4 и совершим симплекс-преобразование (см. (3.17), (3.18), (3.24)—(3.26)). После такого преобразования переменная u^4 перейдет

в базисные, а переменная u^7 станет небазисной, и, как было замечено выше, в следующей симплекс-таблице столбец для u^7 можно вычеркнуть.

В результате получим табл. 9. Из этой таблицы видно, что в полученной угловой точке $z_1 = (0, 0, 0, 3, 0, 0, 0, 0)$ множества

Таблица 8

Базисные переменные	u^1	u^2	u^3	u^4	u^5	Свободные члены
u^6	1	1	—4	1	—3	3
u^7	1	0	—4	2	—5	6
u^8	—2	—5	8	1	0	3
Функция	0	—4	0	4	—8	12

Таблица 9

Базисные переменные	u^1	u^2	u^3	u^4	Свободные члены
u^6	1/2	1	—2	—1/2	0
u^4	1/2	0	—2	—5/2	3
u^8	—5/2	—5	10	5/2	0
Функция	—2	—4	8	2	0

Z значение функции $J_1(z_1) = 0$. Поскольку $J_1(z) \geqslant 0$ при всех $z \in Z$, а $J_1(z_1) = 0$, то ясно, что $\inf\limits_Z J_1(z) = J_1(z_1) = 0$, т. е. вспомогательная задача (18)—(20) решена.

Тогда точка $v_1 = (0, 0, 0, 3, 0)$ является угловой для множества U. Для получения базиса точки v_1 вспомогательные переменные u^6, u^8 в табл. 9 нужно вывести из числа базисных. В строках переменных u^6, u^8 на пересечении со столбцами небазисных переменных имеются положительные величины, любую из которых можно взять в качестве разрешающего элемента для вывода одной из вспомогательных переменных из числа базисных. Возьмем, например, величину 1/2 из столбца u^1 и строки u^6 разрешающим элементом и осуществим симплекс-преобразование табл. 9. Получим табл. 10. В этой таблице в строке для u^8 все величины равны нулю. Как было замечено выше, такую строку без ущерба можно вычеркнуть из симплекс-таблицы. В результате получим табл. 11. В ней строки переменных u^1, u^4 соответствуют системе уравнений

$$u^1 = -2u^2 + 4u^3 + u^5, \quad u^4 = u^2 + 2u^5 + 3, \tag{21}$$

которая эквивалентна системе (17). В нижней строке табл. 11 представлены коэффициенты, полученные подстановкой выражений (21) для базисных переменных u^1, u^4 в (15):

$$J(u) = 2u^2 + 6u^3 + 3 = -(-2)u^2 - (-6)u^3 + 3.$$

Таким образом, табл. 11 является симплекс-таблицей угловой точки (0, 0, 0, 3, 0) для задачи минимизации функции (15)

Таблица 10

Базисные переменные	u^2	u^3	u^5	Свободные члены
u^1	2	—4	—1	0
u^4	—1	0	—2	3
u^5	0	0	0	0

Таблица 11

Базисные переменные	u^2	u^3	u^5	Свободные члены
u^1	2	—4	—1	0
u^4	—1	0	—2	3
Функция	—2	—6	0	3

при ограничениях (16), (21) (ср. с задачей (12)—(14)). Остается решить эту задачу симплекс-методом. Впрочем, в данном случае из табл. 11 сразу видно, что в нижней строке нет положительных величин — реализовался случай I (см. (3.13′)). Это значит, что $u_* = (0, 0, 0, 3, 0)$ — решение задачи (15)—(17).

§ 6. Об условии разрешимости канонической задачи

Как показывает теорема 5.1, с помощью описанного выше симплекс-метода можно не только численно решать задачи линейного программирования, но и можно доказывать основные факты теории линейного программирования. Приводим еще две такие теоремы.

Теорема 1. *Для того чтобы каноническая задача* (1.14) *была разрешима, т. е. существовала точка* $u_* \in U$ *такая, что* $\langle c, u_* \rangle = \inf\limits_U \langle c, u \rangle > -\infty$, *необходимо и достаточно, чтобы:* 1) *множество* U *было непустым*; 2) *функция* $J(u) = \langle c, u \rangle$ *была ограничена снизу на* U.

Доказательство. Необходимость очевидна. Докажем достаточность. Из того, что $U \neq \varnothing$, по теореме 5.1 следует суще-

ствование угловой точки множества U. Принимая эту точку за начальную, будем решать задачу (1.14) с помощью симплекс-метода. Поскольку по условию $\inf\limits_U J(u) > -\infty$, то случай II (условие (3.14)) здесь невозможен, и за конечное число шагов симплекс-метода процесс завершится отысканием точки u_*, являющейся решением задачи (1.14).

Теорема 2. *Если задача* (1.14) *разрешима, то среди ее решений найдется хотя бы одна угловая точка множества* U.

Доказательство. По условию теоремы $U \neq \varnothing$ и существует точка $v_* \in U$ такая, что $\langle c, v_* \rangle = \inf\limits_U \langle c, u \rangle > -\infty$. По теореме 5.1 тогда множество U имеет хотя бы одну угловую точку. Отправляясь от одной из этих угловых точек, с помощью симплекс-метода за конечное число шагов придем к угловой точке u_*, являющейся решением задачи (1.14).

На этом заканчиваем изложение симплекс-метода для решения и исследования канонической задачи (1.14). Учитывая связь между общей задачей линейного программирования (1.5) и задачами (1.14) и (1.15), установленную в § 1, можно сказать, что симплекс-метод является универсальным методом решения задач линейного программирования. Однако это не значит, что симплекс-метод выгодно применять во всех случаях. Существуют и другие общие методы, а также ряд частных методов, приспособленных для решения специальных классов задач линейного программирования, лучше учитывающих конкретные особенности задачи. Например, для транспортных задач, у которых матрица A в ограничении $Au = b$ имеет ряд особенностей, разработаны специальные методы.

Отметим, что известен пример задачи линейного программирования с n переменными и $2n$ ограничениями, для решения которой требуется не менее $2^n - 1$ итераций симплекс-метода и, следовательно, количество необходимых вычислений оценивается экспоненциальной функцией параметров задачи. Это значит, что существуют задачи линейного программирования не слишком большой размерности, решение которых симплекс-методом невозможно за обозримое время. Однако вопреки этому пессимистическому выводу в практических задачах симплекс-метод показывает поразительную эффективность, и число необходимых итераций отнюдь не растет экспоненциально с ростом размерности. Причина этого удивительного явления пока еще не выяснена. Хотя и в последнее время появились методы [313], в которых объем вычислений при решении задач линейного программирования выражается полиномом от параметров задачи, тем не менее симплекс-метод по-прежнему остается основным методом в линейном программировании.

В заключение подчеркнем, что всюду выше предполагалось, что исходные данные задачи линейного программирования —

матрица A, векторы b, c — известны точно и, кроме того, все промежуточные вычисления в симплекс-методе проводятся без погрешностей. Такая идеализация позволила нам дать строгое обоснование симплекс-метода, доказать ряд важных теорем линейного программирования. Однако на практике исходные данные задаются, как правило, неточно, промежуточные вычисления проводятся с округлениями и применение симплекс-метода или других методов в конкретных задачах линейного программирования может привести к большим погрешностям, неверным выводам. В частности, наличие погрешностей может сделать задачу линейного программирования некорректной, и для ее решения могут понадобиться специальные методы регуляризации [6, 22]. К задачам линейного программирования мы еще вернемся в § 4.9.

Упражнения. 1. Минимизировать функцию $u^1 - u^2 - u^3 - u^4 + 2u^5$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 5)$, $u^1 + 3u^2 + u^3 + u^4 - 2u^5 = 10$, $2u^1 + 6u^2 + u^3 + 3u^4 - 4u^5 = 20$, $3u^1 + 10u^2 + u^3 + 6u^4 - 7u^5 = 30$.

2. Максимизировать функцию $u^1 - 4u^2 + 3u^3 + 10u^4$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 4)$, $u^1 + u^2 - u^3 + u^4 = 0$, $u^1 + 14u^2 + 10u^3 - 10u^4 = 11$.

3. Минимизировать функцию $u^1 + 2u^2 - 2u^3 + 5u^4$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 4)$, $u^1 + 2u^2 - u^3 - u^4 = 1$, $-u^1 + 2u^2 + 3u^3 + u^4 = 2$, $u^1 + 5u^2 + u^3 - u^4 = 5$.

4. Минимизировать функцию $u^1 + u^2 + u^3$ при условиях $u^1 \geqslant 0$, $u^2 \geqslant 0$, $u^3 \geqslant 0$, $u^1 + u^2 = 1$, $u^1 - u^2 = 1$, $3u^1 - u^2 = 3$.

5. Минимизировать функцию $u^1 + 2u^2 + 3u^3 + 4u^4$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 4)$, $u^1 + u^2 + u^3 + u^4 \geqslant 1$.

6. Минимизировать функцию $u^1 + u^2 + u^3 + u^4$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 5)$, $u^1 - u^2 \geqslant 0$, $u^1 + u^2 - u^3 + u^4 - u^5 \geqslant 1$. Рассмотреть эту же задачу при дополнительном ограничении $1 \leqslant u^1 \leqslant 2$.

7. Минимизировать функцию $u^3 - u^4 + u^5 - u^6$ при условиях $u^i \geqslant 0$ $(i = 1, \ldots, 7)$, $u^1 + u^3 - 2u^4 - 3u^5 + 4u^6 = 0$, $u^2 + 4u^3 - 3u^4 - 2u^5 + u^6 = 0$, $u^3 + u^4 + u^5 + u^6 + u^7 = 1$. Показать, что в этой задаче можно получить зацикливание, если с помощью симплекс-метода организовать перебор базисов в следующем порядке: $(A_3, A_2, A_7) \to (A_3, A_4, A_7) \to (A_5, A_4, A_7) \to (A_5, A_6, A_7) \to (A_1, A_6, A_7) \to (A_1, A_2, A_7) \to (A_3, A_2, A_7)$. Применить для решения этой задачи симплекс-метод с антициклином, изложенный в конце § 3.

8. Обобщить симплекс-метод на случай задачи

$$\langle c, u\rangle \to \inf;\ u \in U = \{u\colon\ u \in E^n,\ Au = b,\ \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\},$$

где α_i, β_i — заданные величины, $\alpha_i \leqslant \beta_i$ $(i = 1, \ldots, n)$ (возможно, некоторые из $\alpha_i = \infty$ и некоторые из $\beta_j = \infty$).

9. Пусть $U = \{u\colon\ u \in E^n,\ \langle a_i, u\rangle \leqslant b^i,\ i = 1, \ldots, m\}$, $m \geqslant n$. Показать, что точка $v \in U$ является угловой для U тогда и только тогда, если в точке v обращаются в точные равенства не менее чем n из неравенств $\langle a_i, v\rangle \leqslant b^i$, среди которых есть n линейно независимых.

10. Исследовать возможность обобщения теорем 5.1 и 6.1, 6.2 на случай задач (1.5), (1.15).

11. Требуется составить наиболее дешевую смесь, содержащую не менее b^i единиц i-го вещества $(i = 1, \ldots, m)$ при условии, что для изготовления смеси имеется n видов продуктов, причем в одной единице j-го продукта содержится a_{ij} единиц i-го вещества, а цена одной единицы j-го продукта равна c_j рублей (задача о смесях). Сформулировать эту задачу в виде задачи (1.15).

Глава 4

ЭЛЕМЕНТЫ ВЫПУКЛОГО АНАЛИЗА

Прежде чем переходить к рассмотрению задач, более сложных, чем задачи линейного программирования, остановимся на элементах выпуклого анализа — области математики, в которой изучаются свойства выпуклых множеств и выпуклых функций, и которая играет фундаментальную роль в теории и методах решения экстремальных задач [1, 2, 7, 8, 11, 12, 15—19, 21, 24, 33, 41, 52, 66, 67, 92, 132, 137, 144, 146, 156, 166, 167, 195, 197, 255, 264, 283, 290, 321, 337, 338, 341].

§ 1. Выпуклые множества

1. Начнем с рассмотрения конкретных примеров выпуклых множеств. Сначала напомним

Определение. 1. Множество U называется *выпуклым*, если для любых u, $v \in U$ точка $u_\alpha = v + \alpha(u - v) = \alpha u + (1 - \alpha) v$ принадлежит U при всех α $(0 \leqslant \alpha \leqslant 1)$. Иначе говоря, множество U выпукло, если отрезок $[v, u] = \{u_\alpha\colon u_\alpha = v + \alpha(u - v),\ 0 \leqslant \alpha \leqslant 1\}$, соединяющий любые две точки u, v из U, целиком лежит в U (рис. 4.1).

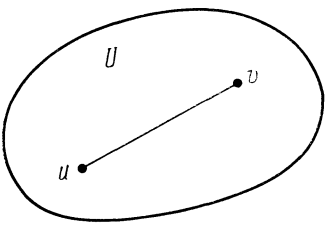

Рис. 4.1

Все пространство E^n, очевидно, образует выпуклое множество. Пустое множество и множество, состоящее из одной точки, удобно считать выпуклыми. Тогда из определения 1 непосредственно следует, что пересечение любого числа выпуклых множеств является выпуклым множеством. Приведем другие примеры выпуклых множеств.

Пример 1. Шар

$$S(u_0, R) = \{u\colon u \in E^n,\ |u - u_0| \leqslant R\} \tag{1}$$

радиуса $R > 0$ с центром в точке u_0 является выпуклым множеством. В самом деле, если u, $v \in S(u_0, R)$, то, пользуясь нера-

венством треугольника, имеем

$$|\alpha u+(1-\alpha)v-u_0|=|\alpha(u-u_0)+(1-\alpha)(v-u_0)|\leqslant$$
$$\leqslant\alpha|u-u_0|+(1-\alpha)|v-u_0|\leqslant\alpha R+(1-\alpha)R=R,$$

т. е. $u_\alpha=\alpha u+(1-\alpha)v\in S(u_0, R)$ для всех $\alpha\in[0, 1]$.

Пример 2. *Гиперплоскостью* в E^n называется множество

$$\Gamma=\Gamma(c, \gamma)=\{u\colon u\in E^n, \langle c, u\rangle=\gamma\}, \tag{2}$$

где $c=(c_1, \ldots, c_n)\neq 0$ — вектор из E^n, γ — действительное число. Это множество всегда непусто: если, например, $c_i\neq 0$, то точка u_0 с координатами $u^i=\gamma/c_i$, $u^j=0$ $(j=1, \ldots, n, j\neq i)$ удовлетворяет равенству $\langle c, u_0\rangle=\gamma$, т. е. $u_0\in\Gamma$. Если u_0 — какая-либо точка из Γ, т. е. $\langle c, u_0\rangle=\gamma$, то гиперплоскость Γ можно представить в виде

$$\Gamma=\Gamma(c, u_0)=\{u\colon u\in E^n, \langle c, u-u_0\rangle=0\}.$$

Напоминаем, что два вектора a, $b\in E^n$, называются *ортогональными*, если $\langle a, b\rangle=0$. Предыдущее представление для Γ означает, что гиперплоскость состоит из тех и только тех точек u, для которых вектор $u-u_0$ ортогонален вектору c. Вектор c называют *нормальным* вектором гиперплоскости Γ.

Возьмем произвольные точки u, $v\in\Gamma$, т. е. $\langle c, u\rangle=\langle c, v\rangle=\gamma$. Тогда $\langle c, \alpha u+(1-\alpha)v\rangle=\alpha\langle c, u\rangle+(1-\alpha)\langle c, v\rangle=\gamma$ при всех $\alpha\in[0, 1]$. Следовательно, Γ — выпуклое множество.

Пример 3. Пусть $\Gamma=\{u\colon \langle c, u\rangle=\gamma\}$ — некоторая гиперплоскость. Тогда множества

$$\Gamma^+=\{u\colon \langle c, u\rangle>\gamma\},\quad \Gamma^-=\{u\colon \langle c, u\rangle<\gamma\}$$

называются *открытыми* полупространствами, а множества

$$\bar\Gamma^+=\{u\colon \langle c, u\rangle\geqslant\gamma\},\quad \bar\Gamma^-=\{u\colon \langle c, u\rangle\leqslant\gamma\}$$

называются *замкнутыми* полупространствами. Нетрудно видеть, что Γ^+, Γ^-, $\bar\Gamma^+$, $\bar\Gamma^-$ — выпуклые множества. Например, если u, $v\in\Gamma^+$, то $\langle c, \alpha u+(1-\alpha)v\rangle=\alpha\langle c, u\rangle+(1-\alpha)\langle c, v\rangle>\alpha\gamma+(1-\alpha)\gamma=\gamma$ для всех $\alpha\in[0, 1]$.

Пример 4. Множество $M_1=\{u\in E^n\colon u=\alpha w+(1-\alpha)v=v+\alpha(w-v), \alpha\in R\}$ называется *прямой*, проходящей через точки v, w, $v\neq w$. Прямую в E^n можно задать также и в виде $M_1=\{u\in E^n\colon u=v+td, t\in\mathbf{R}\}$, где вектор $d\neq 0$ называют *направляющим вектором* прямой, v — фиксированная точка из E^n. Множество $M_1^+=\{u\in E^n\colon u=v+td, t\geqslant 0\}$ называют *лучом*, выходящим из точки v в направлении вектора $d\neq 0$. Нетрудно видеть, что прямая и луч — выпуклые множества.

Пример 5. Важным примером выпуклого множества являются аффинные множества (или линейные многообразия).

Определение 2. Множество M из E^n называется *аффинным*, если $\alpha u+(1-\alpha)v\in M$ при всех u, $v\in M$, $\alpha\in\mathbf{R}$, т. е.

прямая, проходящая через любые две точки u, $v \in M$, целиком лежит в M.

Все пространство E^n является аффинным множеством. Пустое множество и множество, состоящее из одной точки, удобно считать аффинными. Любое подпространство пространства E^n представляет собой аффинное множество. Множество $M = L + u_0$, получаемое сдвигом подпространства L на произвольный фиксированный вектор u_0, также является аффинным.

Верно и обратное: всякое аффинное множество M может быть получено сдвигом некоторого подпространства L на некоторый вектор u_0. В самом деле, возьмем произвольную точку $u_0 \in M$ и положим $L = M - u_0$. Ясно, что L — аффинное множество, причем $0 \in L$. Тогда для каждого $u \in L$ имеем $\alpha u = \alpha u + (1-\alpha)\cdot 0 \in L$ при всех $\alpha \in \mathbf{R}$. Кроме того, если u, $v \in L$, то $(u+v)/2 = u/2 + (1-1/2)v \in L$ и, следовательно, $u + v = 2((u+v)/2) \in L$. Таким образом, сумма двух векторов из L и произведение вектора из L на любое число принадлежат L, т. е. L — подпространство.

Убедимся, что подпространство $L = M - u_0$ не зависит от выбора точки $u_0 \in M$. В самом деле, пусть $L_1 = M - u_1$, где $u_1 \in M$. Возьмем любую точку $u \in L$. Поскольку $u_1 - u_0 \in L$, то $u + (u_1 - u_0) \in L$ и, следовательно, $u \in L - (u_1 - u_0) = (L + u_0) - u_1 = M - u_1 = L_1$. Это значит, что $L \subset L_1$. Обратное включение $L_1 \subset L$ доказывается совершенно так же. Следовательно, $L_1 = L$.

Таким образом, всякое аффинное множество M из E^n представимо в виде

$$M = L + u_0, \tag{3}$$

где L — подпространство, однозначно определяемое множеством M, u_0 — произвольная точка из M. Подпространство из этого представления называют *параллельным* аффинному множеству M.

Опираясь на полученное представление (3), можно дать алгебраическое описание аффинных множеств из E^n. Как известно из линейной алгебры [93, 164], всякое подпространство L из E^n представимо в виде $L = \{u \in E^n\colon Au = 0\} = \{u \in E^n\colon \langle a_i, u\rangle = 0,\ i = 1, \ldots, m\}$, где A — некоторая матрица размера $m \times n$, a_i — i-я строка матрицы A (например, в качестве векторов a_i $(i = 1, \ldots, m)$ можно взять базис ортогонального дополнения к L в E^n). Отсюда и из (3) следует, что всякое аффинное множество из E^n может быть задано в виде

$$\begin{aligned} M &= \{u \in E^n\colon Au = 0\} + u_0 = \{u \in E^n\colon u = u_0 + v,\ Av = 0\} = \\ &= \{u \in E^n\colon A(u - u_0) = 0\} = \{u \in E^n\colon Au = b\} = \\ &= \{u \in E^n\colon \langle a_i, u\rangle = b_i,\ i = 1, \ldots, m\}, \end{aligned} \tag{4}$$

где $b = Au_0 = (b^1, \ldots, b^m)$. Нетрудно проверить, что верно и обратное: всякое множество вида (4) является аффинным. В са-

мом деле, если $u, v \in M$, т. е. $Au = b$, $Av = b$, то $A(\alpha u + (1-\alpha)v) = \alpha Au + (1-\alpha)Av = b$ или, иначе, $\alpha u + (1-\alpha)v \in M$ при всех $\alpha \in \mathbf{R}$. Таким образом, множества вида (4) и только они являются аффинными.

Согласно теореме Кронекера — Капелли [93, 164] множество (4) непусто тогда и только тогда, когда матрица A и расширенная матрица $B = (A, b)$ имеют один и тот же ранг. Если $\operatorname{rang} A < \operatorname{rang} B$ (например, $A = 0$, $b \neq 0$), то $M = \varnothing$. Если $A = 0$, $b = 0$, то $M = E^n$. Рассмотрим случай $A \neq 0$, $\operatorname{rang} A = \operatorname{rang} B = r$. Тогда множество (4) состоит из тех и только тех точек, которые представимы в виде [93, 164]

$$u = u_0 + \sum_{i=1}^{n-r} t_i u_i, \tag{5}$$

где u_0 — какое-либо частное решение неоднородной системы линейных алгебраических уравнений $Au = b$, а $u_1, \ldots, u_{n-r}$ — линейно независимые решения однородной системы $Au = 0$; $t_1, \ldots, t_{n-r}$ — действительные числа. Векторы $u_1, \ldots, u_{n-r}$ образуют базис подпространства

$$L = \{u \in E^n\colon\ Au = 0\} = \left\{u \in E^n\colon\ u = \sum_{i=1}^{n-r} t_i u_i,\ t_i \in \mathbf{R}\right\},$$

так что размерность L равна $n - r$. С помощью введенного подпространства L равенство (5) можно переписать в виде $M = u_0 + L$ — мы снова пришли к представлению (3).

Размерность аффинного множества M по определению принимается равной размерности подпространства L, параллельного M. Таким образом, размерность аффинного множества (4) равна $n - r$, где $r = \operatorname{rang} A = \operatorname{rang} B$. Аффинное множество размерности p часто называют *гиперплоскостью* размерности p. В тех случаях, когда нужно подчеркнуть размерность p аффинного множества M и соответствующего параллельного подпростства L, будем писать M_p, L_p. В частности, если в (4), (5) $r = n$, то $L_0 = \{0\}$ и $M_0 = \{u_0\}$ состоит из одной точки. Если $r = n - 1$, то $M_1 = \{u \in E^n\colon\ u = u_0 + tu_1,\ t \in \mathbf{R}\}$ — прямая (см. пример 4). Далее, гиперплоскость (2) также является аффинным множеством: в этом случае в (4) нужно принять $A = c$, $b = \gamma$. Поскольку $c \neq 0$, то $\operatorname{rang} A = \operatorname{rang}(A, b) = 1$ и, следовательно, гиперплоскость имеет размерность $n - 1$. Согласно (5) тогда $\Gamma = M_{n-1} = \left\{u \in E^n\colon\ u = u_0 + \sum_{i=1}^{n-1} t_i u_i,\ t_i \in \mathbf{R}\right\}$, где $u_1, \ldots, u_{n-1}$ — базис параллельного подпространства $L_{n-1} = \{u \in E^n\colon\ \langle c, u\rangle = 0\}$. Как видим, вектор c ортогонален к L_{n-1} и является базисом ортогонального дополнения к L_{n-1} до E^n, а векторы $u_1, \ldots, u_{n-1}$, c образуют базис в E^n.

Заметим, что пересечение любого числа аффинных множеств само является аффинным множеством и, следовательно, представимо в виде (4).

Определение 3. Пересечение всех аффинных множеств, содержащих множество U из E^n, называется *аффинной оболочкой* множества U и обозначается через $\operatorname{aff} U$; подпространство L, параллельное $\operatorname{aff} U$, называется *несущим подпространством* множества U и обозначается через $\operatorname{Lin} U$.

Таким образом, $\operatorname{aff} U$ представляет собой минимальное аффинное множество, содержащее U. Пользуясь (3)—(5), нетрудно показать, что:

1) $\operatorname{aff} U = u_0 + \operatorname{Lin} U$, где u_0 — произвольная точка из U;

2) если $0 \in U$, то $\operatorname{aff} U = \operatorname{Lin} U$;

3) $u = v - w \in \operatorname{Lin} U$ для всех $v, w \in \operatorname{aff} U$, в частности, для $v, w \in U$;

4) если $c \in \operatorname{Lin} U$, $v \in \operatorname{aff} U$, то $u = v + \varepsilon c \in \operatorname{aff} U$ при всех $\varepsilon \in \mathbf{R}$.

Определение 4. *Размерностью произвольного множества* U из E^n называется размерность его аффинной оболочки; размерность множества U обозначают $\dim U$.

Согласно этому определению отрезок $[u, v] = \{u_\alpha = v + \alpha(u - v),\ 0 \leqslant \alpha \leqslant 1\}$, соединяющий две точки $u, v \in E^n$ $(u \neq v)$, имеет размерность 1, так как его аффинной оболочкой является прямая $\Gamma = \{w\colon w = v + t(u - v),\ -\infty < t < \infty\}$. Размерность шара (1) равна n.

Пример 6. Множество $U = \{u \in E^n\colon Au \leqslant b\}$, где A — заданная матрица размера $m \times n$, b — заданный вектор из E^m, выпукло. Это множество называют *многогранным множеством* или *полиэдром*. Напоминаем, что неравенство $Au \leqslant b$ означает, что $\langle a_i, u\rangle = (Au)^i \leqslant b^i$ при всех $i = 1, \ldots, m$, a_i — i-я строка матрицы A. Тогда для любых $u, v \in U$ имеем $A(\alpha u + (1 - \alpha)v) = \alpha Au + (1 - \alpha)Av \leqslant b$ при всех $\alpha \in [0, 1]$.

Пример 7. Множество $U = \{u = (u^1, \ldots, u^n)\colon \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\}$, где α_i, β_i — заданные величины, $\alpha_i < \beta_i$ (возможно, что некоторые из $\alpha_i = \infty$ и некоторые из $\beta_j = \infty$), выпукло и имеет размерность n. В частности, неотрицательный ортант пространства E^n — это множество $E^n_+ = \{u\colon u \in E^n, u \geqslant 0\}$ — выпукло, причем $\dim E^n_+ = n$. Если в определении множества U величины α_i, β_i конечны при всех $i = 1, \ldots, n$, то это множество называют *n-мерным параллелепипедом*.

Пример 8. Множество $U = \{u = (u^1, \ldots, u^n)\colon u^i \geqslant 0,\ i \in I;\ Au \leqslant b,\ \bar{A}u = \bar{b}\}$, взятое из общей задачи линейного программирования (3.1.5), выпукло. Это нетрудно проверить, исходя из определения 1. Впрочем, выпуклость U следует также из того, что U является пересечением множеств $U_0 = \{u\colon u^i \geqslant 0,\ i \in I\}$, $U_1 = \{u\colon Au \leqslant b\}$, $U_2 = \{u\colon \bar{A}u = \bar{b}\}$, каждое из которых выпукло. Очевидно, U — многогранное множество. Аналогично прове-

ряется выпуклость множеств из основной и канонической задач линейного программирования (3.1.14), (3.1.15).

2. Выше было отмечено, что пересечение любого числа выпуклых множеств выпукло. Нетрудно видеть, что объединение двух выпуклых множеств, вообще говоря, невыпукло (рис. 4.2). Посмотрим, как влияют на выпуклость другие операции над множествами: сложение, вычитание, умножение множества на число, замыкание и т. п.

Определение 5. *Суммой множеств* $A_1, \ldots, A_m$ называется множество $A = A_1 + \ldots + A_m = \sum_{i=1}^{m} A_i$, состоящее из тех и только тех точек a, которые представимы в виде $a = \sum_{i=1}^{m} a_i$ $(a_i \in A_i,\ i = 1, \ldots, m)$. *Разностью двух множеств* A и B называется множество $C = A - B$, состоящее из тех и только тех точек c, которые представимы в виде $c = a - b$ $(a \in A,\ b \in B)$. *Произведением множества* A *на действительное число* λ называется множество $B = \lambda A$, состоящее из всех точек вида $b = \lambda a$ $(a \in A)$.

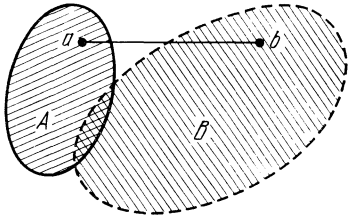

Рис. 4.2

Теорема 1. *Если* $A_1, \ldots, A_m$, A, B — *выпуклые множества, то множества* $C = A_1 + \ldots + A_m$, $C = A - B$, $C = \lambda A$ *выпуклы.*

Доказательство. Проведем его, например, для множества $C = A - B$. Пусть c_1, c_2 — произвольные точки из $C = A - B$. Это значит, что существуют $a_i \in A$, $b_i \in B$ такие, что $c_i = a_i - b_i$ $(i = 1, 2)$. Из выпуклости A и B следует, что $a_\alpha = \alpha a_1 + (1 - \alpha) a_2 \in A$, $b_\alpha = \alpha b_1 + (1 - \alpha) b_2 \in B$ при всех $\alpha \in [0, 1]$. Тогда $c_\alpha = \alpha c_1 + (1 - \alpha) c_2 = \alpha (a_1 - b_1) \in (1 - \alpha)(a_2 - b_2) = a_\alpha - b_\alpha$, так что $c_\alpha \in C$ при всех $\alpha \in [0, 1]$. Выпуклость $C = A - B$ доказана. Аналогично доказывается выпуклость множеств $C = A_1 + \ldots + A_m$ и $C = \lambda A$.

Определение 6. *Замыканием множества* U называется множество, являющееся объединением множества U и всех его предельных точек. Замыкание множества U будем обозначать через $\overline{U}$.

Для любой точки v и любого множества U из E^n имеет место одна и только одна из следующих трех возможностей.

1. Найдется ε-окрестность точки v, которая целиком принадлежит множеству U — тогда точка v называется *внутренней* точкой множества U. Совокупность всех внутренних точек множества U будем обозначать через $\operatorname{int} U$. Множество U, все точки которого являются внутренними, называют *открытым множеством.*

Примером открытого множества является открытое полупространство из примера 3.

2. Найдется ε-окрестность точки v, которая не содержит ни одной точки множества U — такая точка называется *внешней* по отношению ко множеству U.

3. Любая ε-окрестность точки v содержит как точки из U, так и точки из $E^n \backslash U$ — тогда точка v называется *граничной* точкой множества U. Совокупность всех граничных точек множества U будем обозначать через Гр U.

Всякая внутренняя точка множества, очевидно, является его предельной точкой. Однако не всякая граничная точка множества будет его предельной точкой — исключение здесь составляют изолированные точки множества. Точку $v \in U$ называют *изолированной* точкой этого множества, если существует ε-окрестность этой точки, не содержащая ни одной точки множества U, отличной от v.

Таким образом, замыкание $\overline{U}$ множества U состоит, вообще говоря, из точек четырех типов: 1) внутренние точки множества U; 2) изолированные точки множества U; 3) предельные граничные точки множества U, принадлежащие U; 4) предельные граничные точки множества U, не принадлежащие U. Отсюда ясно, что замыкание любого множества является замкнутым множеством.

Очевидно, выпуклое множество не может иметь изолированных точек.

Шар (1) замкнут и его замыкание состоит из внутренних точек $\operatorname{int} S(u_0, R) = \{u\colon |u - u_0| < R\}$ и граничных точек Гр $S(u_0, R) = \{u\colon |u - u_0| = R\}$. Множество $U = \{u = (x, y) \in E^2\colon 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1,\ x + y < 1\}$ выпукло, но не замкнуто — точки прямой $x + y = 1$ при $0 \leqslant x,\ y \leqslant 1$ являются предельными граничными для U, но не принадлежат U; $\overline{U} = \{u = (x, y)\colon 0 \leqslant x, y \leqslant 1,\ x + y \leqslant 1\}$. Множество $U = \{u = (x, y) \in E^2\colon -1 < x < 1,\ y = 0\}$ выпукло и не имеет внутренних точек; $\overline{U} = \{u = (x, y)\colon -1 \leqslant x \leqslant 1,\ y = 0\}$.

Для любого аффинного множества M из E^n имеем $\overline{M} = M$, так что M — замкнутое множество: это видно, например, из представления (4) аффинного множества; для множеств из примеров 6—8 также $\overline{U} = U$. В частности, $\operatorname{aff} U = \overline{\operatorname{aff} U}$, откуда будет следовать, что $\operatorname{aff} \overline{U} = \operatorname{aff} U$ для любого множества U из E^n.

Теорема 2. *Если A — выпуклое множество, то его замыкание тоже выпукло.*

Доказательство. Пусть a и b — произвольные точки множества $\overline{A}$. Поскольку выпуклое множество не имеет изолированных точек, то точки a и b будут предельными для A. Тогда существуют последовательности $\{a_k\}, \{b_k\} \in A$, сходящиеся соответственно к a, b. Возьмем произвольные $\alpha \in [0, 1]$. В силу выпуклости A тогда $c_k = \alpha a_k + (1 - \alpha) b_k \in A$. Отсюда при $k \to \infty$

получим $\lim\limits_{k\to\infty} c_k = c_\alpha = \alpha a + (1-\alpha) b$. Таким образом, точка c_α является предельной для A и, следовательно, принадлежит $\bar{A}$ при любом $\alpha \in [0, 1]$.

Теорема 3. *Пусть U — выпуклое множество и $\operatorname{int} U \neq \varnothing$. Пусть $u_0 \in \operatorname{int} U$, $v \in \bar{U}$. Тогда $v_\alpha = v + \alpha(u_0 - v) \in \operatorname{int} U$ при всех α $(0 < \alpha \leqslant 1)$. Если $u \in \operatorname{int} U$, $y \notin \operatorname{int} U$, $y \in \bar{U}$, то $w_\lambda = u + \lambda(y - u) \notin \bar{U}$ при всех $\lambda > 1$.*

Доказательство. Поскольку точка $u_0 \in \operatorname{int} U$, то найдется ее δ-окрестность $O(u_0, \delta) = \{u\colon |u - u_0| < \delta\}$, целиком принадлежащая U. Сначала рассмотрим случай, когда $v \in U$. Возьмем произвольное α $(0 < \alpha < 1)$. Покажем, что окрестность

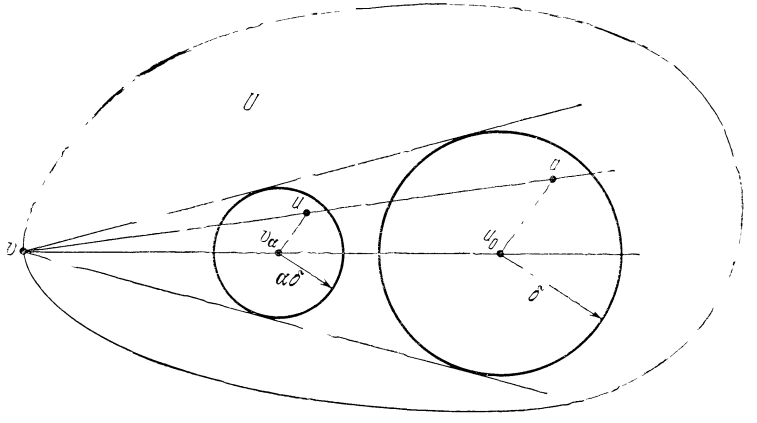

Рис. 4.3

$O(v_\alpha, \alpha\delta) = \{u\colon |u - u_\alpha| < \alpha\delta\}$ точки v_α принадлежит U. С этой целью возьмем произвольную точку $u \in O(v_\alpha, \alpha\delta)$ и положим $a = u_0 + (u - v_\alpha)/\alpha$ (рис. 4.3). Поскольку $|a - u_0| = |u - v_\alpha|/\alpha < \delta\alpha/\alpha = \delta$, то $a \in O(u_0, \delta) \subset U$. Из определения точки a имеем представление $u = v_\alpha + \alpha(a - u_0) = v + \alpha(u_0 - v) + \alpha(a - u_0) = \alpha a + (1 - \alpha) v$, где a, $v \in U$ и $0 < \alpha < 1$. Тогда $u \in U$ в силу выпуклости U. Тем самым показано, что произвольная точка u из $O(v_\alpha, \alpha\delta)$ принадлежит U. Следовательно, $O(v_\alpha, \alpha\delta) \subset U$, т. е. v_α — внутренняя точка U.

Пусть теперь $v \in \bar{U} \setminus U$. Поскольку v — предельная точка U, то найдется точка $w \in U$ такая, что $|v - w| < \alpha(1 - \alpha)^{-1}\delta$. Возьмем точку $w_\alpha = w + \alpha(u_0 - w)$ (рис. 4.4). В силу только что доказанного точка w_α принадлежит множеству U вместе со своей окрестностью $O(w_\alpha, \alpha\delta)$. Но $|v_\alpha - w_\alpha| = |v + \alpha(u_0 - v) - w - \alpha(u_0 - w)| = (1 - \alpha)|v - w| < (1 - \alpha)\alpha(1 - \alpha)^{-1}\delta = \alpha\delta$. Следовательно, $v_\alpha \in O(w_\alpha, \alpha\delta) \subset U$. Нетрудно видеть, что окрестность $O(v_\alpha, \beta)$ $(\beta = \delta\alpha - |v_\alpha - w_\alpha|)$ точки v_α также принадле-

жит U. В самом деле, если $u \in O(v_\alpha, \beta)$, то $|u - w_\alpha| \leqslant \leqslant |u - v_\alpha| + |v_\alpha - w_\alpha| < \beta + |v_\alpha - w_\alpha| = \delta\alpha$, т. е. $O(v_\alpha, \beta) \in \in O(w_\alpha, \alpha\delta) \in U$.

Наконец, пусть $w_\lambda = u + \lambda(y - u)$ $(\lambda > 1)$, где $u \in \operatorname{int} U$, $y \notin \operatorname{int} U$, $y \in \overline{U}$. Допустим, что $w_\lambda \in \overline{U}$ при каком-либо $\lambda > 1$. Из представления для w_λ имеем $y = u + (w_\lambda - u)/\lambda = w_\lambda + + (1 - 1/\lambda)(u - w_\lambda) = w_\lambda + \alpha(u - w_\lambda)$, где $\alpha = 1 - 1/\lambda \in (0, 1)$,

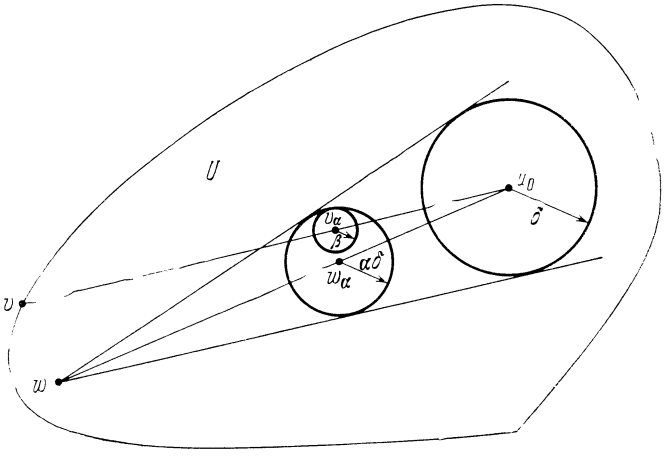

Рис. 4.4

$w_\lambda \in \overline{U}$, $u \in \operatorname{int} U$. По доказанному выше тогда $y \in \operatorname{int} U$, что противоречит условию. Следовательно, $w_\lambda \notin \overline{U}$ при всех $\lambda > 1$.

Теорема 4. *Если U — выпуклое множество, то $\operatorname{int} U$ тоже выпукло.*

Доказательство. Пусть u, v — произвольные точки из $\operatorname{int} U$. В теореме 3 было показано, что $u_\alpha = \alpha u + (1 - \alpha) v \in \in \operatorname{int} U$ при всех α $(0 < \alpha < 1)$. Это и означает выпуклость $\operatorname{int} U$.

3. В тех случаях, когда рассматриваемое множество U невыпукло, часто бывает полезно расширить его до выпуклого множества. Посмотрим, как это делается.

Определение 7. Точка u называется *выпуклой комбинацией* точек $u_1, \ldots, u_m$, если существуют числа $\alpha_1 \geqslant 0, \ldots, \alpha_m \geqslant 0$, $\alpha_1 + \ldots + \alpha_m = 1$ такие, что $u = \alpha_1 u_1 + \ldots + \alpha_m u_m$.

Теорема 5. *Множество выпукло тогда и только тогда, когда оно содержит все выпуклые комбинации любого конечного числа своих точек.*

Доказательство. Необходимость. Пусть U — выпуклое множество. Тогда по определению 1 множество U содержит выпуклые комбинации любых двух своих точек. Сделаем индуктивное предположение: пусть множество U содержит выпуклые комбинации любых $m - 1$ своих точек. Рассмотрим выпуклую комбинацию $\alpha_1 u_1 + \ldots + \alpha_m u_m$ произвольных m точек из U. Можем считать, что $\alpha_i > 0$ $(i = 1, \ldots, m)$. Поскольку $\alpha_1 + \ldots + \alpha_m = 1$, то $0 < \alpha_i < 1$ $(i = 1, \ldots, m)$. Следовательно, точка $v = \sum_{i=1}^{m-1} \alpha_i (1 - \alpha_m)^{-1} u_i$ является выпуклой комбинацией точек $u_1 \ldots, u_{m-1}$

и по предположению индукции принадлежит U. Однако $u = (1-\alpha_m)\sum_{i=1}^{m-1}\alpha_i(1-\alpha_m)^{-1}u_i + \alpha_m u_m = (1-\alpha_m)v + \alpha_m u_m \in U$ в силу выпуклости U.

Достаточность. Если множество U содержит все выпуклые комбинации любого конечного числа своих точек, то оно содержит, в частности, выпуклые комбинации любых двух своих точек и, следовательно, выпукло.

Определение 8. Пересечение всех выпуклых множеств, содержащих множество U, называется *выпуклой оболочкой* множества U и обозначается через $\operatorname{co} U$.

Ясно, что $\operatorname{co} U$, как пересечение выпуклых множеств, является выпуклым множеством. Кроме того, $\operatorname{co} U$ содержится в любом выпуклом множестве, содержащем U. Так что $\operatorname{co} U$ — это минимальное выпуклое множество, содержащее U.

Теорема 6. *Выпуклая оболочка множества U состоит из тех и только тех точек, которые являются выпуклой комбинацией конечного числа точек из U.*

Доказательство. Пусть W — множество всех точек, являющихся выпуклыми комбинациями любого конечного числа точек из U. Нам надо показать, что $\operatorname{co} U = W$. Поскольку $U \subseteq \operatorname{co} U$ и $\operatorname{co} U$ — выпуклое множество, то по теореме 5 $\operatorname{co} U$ содержит все выпуклые комбинации точек из $\operatorname{co} U$ и, в частности, точек из U. Следовательно, $W \subseteq \operatorname{co} U$.

Покажем, что W — выпуклое множество. В самом деле, пусть $u, v \in W$, т. е. $u = \alpha_1 u_1 + \ldots + \alpha_m u_m$ $(u_i \in U, \alpha_i \geqslant 0, i = 1, \ldots m, \alpha_1 + \ldots + \alpha_m = 1)$, $v = \beta_1 v_1 + \ldots + \beta_p v_p$ $(v_i \in U, \ \beta_i \geqslant 0, \ i = 1, \ldots, p, \ \beta_1 + \ldots + \beta_p = 1)$. Тогда $u_\alpha = \alpha u + (1-\alpha)v = \sum_{i=1}^{m}\alpha\alpha_i u_i + \sum_{j=1}^{p}(1-\alpha)\beta_j v_j$, где $\alpha\alpha_i \geqslant 0$, $(1-\alpha)\beta_j \geqslant 0$, $\sum_{i=1}^{m}\alpha\alpha_i + \sum_{j=1}^{p}(1-\alpha)\beta_j = 1$ для каждого $\alpha \in [0, 1]$. Следовательно, u_α является выпуклой комбинацией точек $u_1, \ldots, u_m, v_1, \ldots, v_p \in U$ и принадлежит W при всех $\alpha \in [0, 1]$. Таким образом, W — выпуклое множество, содержащее U. Но $\operatorname{co} U$ по своему определению принадлежит всем выпуклым множествам, содержащим U, и поэтому $\operatorname{co} U \subseteq W$. Сравнивая с ранее доказанным включением $W \subset \operatorname{co} U$, заключаем, что $\operatorname{co} U = W$, что и требовалось.

Заметим, что выпуклая оболочка двух точек на плоскости представляет собой отрезок, выпуклая оболочка трех точек, не лежащих на одной прямой,— треугольник. В общем случае выпукая оболочка конечного числа точек на плоскости образует выпуклый многоугольник, а в пространстве — выпуклый многогранник.

Определение 9. Выпуклая оболочка множества точек $u_0, u_1, \ldots, u_m$ из E^n таких, что система векторов $\{u_i - u_0, \ i = 1, \ldots, m\}$ линейно независима, называется *симплексом*, натянутым на эти точки, и обозначается через $s_m = s_m(u_0, u_1, \ldots, u_m)$. Точки $u_0, u_1, \ldots, u_m$ называются *вершинами симплекса*.

В случае $m = 0, 1, 2, 3$ симплекс представляет собой соответственно точку, отрезок, треугольник, тетраэдр.

Согласно теореме 6 симплекс s_m представим в виде

$$S_m = \left\{u\colon\ u = \sum_{i=0}^{m}\alpha_i u_i, \alpha_i \geqslant 0, \quad i = 1, \ldots, m, \sum_{i=0}^{m}\alpha_i = 1\right\}.$$

По теореме 6 любая точка выпуклой оболочки множества U является выпуклой комбинацией конечного, но, быть может, довольно большого

числа точек из U. Замечательно, однако, то, что в E^n для получения множества $\operatorname{co} U$ достаточно ограничиться рассмотрением выпуклых комбинаций не более чем $n+1$ точек из U. Точнее, верна

Теорема 7. *Пусть U — произвольное непустое множество из E^n. Тогда любая точка $u \in \operatorname{co} U$ представима в виде выпуклой комбинации не более чем $n+1$ точек из U.*

Доказательство. Согласно теореме 6 любая точка $u \in \operatorname{co} U$ представима в виде $u = \alpha_1 u_1 + \ldots + \alpha_m u_m$, где $u_i \in U$, $\alpha_i \geqslant 0$ $(i = 1, \ldots, m)$, $\alpha_1 + \ldots + \alpha_m = 1$. Пусть $m > n+1$ и все $\alpha_i < 0$ (если $\alpha_i = 0$, то число m может быть уменьшено). В $n+1$-мерном пространстве рассмотрим векторы $\bar{u}_i = (u_i, 1)$ $(i = 1, \ldots, m)$. Поскольку $m > n+1$, то эти векторы линейно зависимы, т. е. существуют числа $\gamma_1, \ldots, \gamma_m$, не все равные нулю и такие, что $\gamma_1 \bar{u}_1 + \ldots + \gamma_m \bar{u}_m = 0$. Это равенство эквивалентно следующим двум равенствам:

$$\gamma_1 u_1 + \ldots + \gamma_m u_m = 0, \quad \gamma_1 + \ldots + \gamma_m = 0.$$

Тогда точка u представима другими выпуклыми комбинациями тех же точек $u_1, \ldots, u_m$: $\sum\limits_{i=1}^{m} (\alpha_i - t\gamma_i) u_i = \sum\limits_{i=1}^{m} \alpha_i u_i - t \sum\limits_{i=1}^{m} \gamma_i u_i = u$. В самом деле, здесь $\alpha_i > 0$ и, следовательно, $\alpha_i - t\gamma_i \geqslant 0$ $(i = 1, \ldots, m)$ при всех достаточно малых t и, кроме того, $\sum\limits_{i=1}^{m} (\alpha_i - t\gamma_i) = \sum\limits_{i=1}^{m} \alpha_i - t \sum\limits_{i=1}^{m} \gamma_i = 1$. Поскольку не все γ_i равны нулю, но $\gamma_1 + \ldots + \gamma_m = 0$, то среди $\{\gamma_i\}$ найдутся положительные.

Пусть $\alpha_s \gamma_s^{-1} = \min\limits_{\gamma_i > 0} \alpha_i \gamma_i^{-1}$. Положим $t = \alpha_s \gamma_s^{-1}$. При таком выборе t все $\alpha_i - t\gamma_i$ останутся неотрицательными, причем $\alpha_s - t\gamma_s = 0$. Это значит, что точку удалось представить в виде выпуклой комбинации меньшего числа точек $u_1, \ldots, u_{s-1}, u_{s+1}, \ldots, u_m$. Ясно, что, последовательно применяя описанный прием далее, число точек, участвующих в выпуклой комбинации, можно уменьшить до $n+1$.

Теорема 8. *Если U — замкнутое ограниченное множество из E^n, то $\operatorname{co} U$ замкнуто и ограничено.*

Доказательство. По условию существует число $R > 0$ такое, что $|u| \leqslant R$ для всех $u \in U$, т. е. $U \subseteq S(0, R)$ — шар радиуса R с центром в точке 0. Но шар — выпуклое множество. Согласно определению 8 тогда $\operatorname{co} U \subseteq S(0, R)$, так что $|u| \leqslant R$ для всех $u \in \operatorname{co} U$. Ограниченность $\operatorname{co} U$ доказана.

Докажем замкнутость $\operatorname{co} U$. Пусть u — предельная точка $\operatorname{co} U$, $\{u_k\} \in \operatorname{co} U$ и $\lim\limits_{k \to \infty} u_k = u$. Согласно теореме 7 существуют точки $u_{ki} \in U$, числа $\alpha_{ki} \geqslant 0$ $(i = 1, \ldots, n+1)$, $\alpha_{k1} + \ldots + \alpha_{k,\,n+1} = 1$ такие, что $u_k = \alpha_{k1} u_{k1} + \ldots + \alpha_{k,\,n+1} u_{k,\,n+1}$. Заметим, что $|u_{ki}| \leqslant R$, $0 \leqslant \alpha_{ki} \leqslant 1$ при всех $i = 1, \ldots, n+1$ и всех $k = 1, 2, \ldots$ Пользуясь теоремой Больцано — Вейерштрасса, сначала из $\{u_{k1}\}$, $\{\alpha_{k1}\}$ выберем подпоследовательности $\{u_{k_1 1}\}$, $\{\alpha_{k_1 1}\}$, сходящиеся соответственно к некоторым u_1, α_1; затем из $\{u_{k_1 2}\}$, $\{\alpha_{k_1 2}\}$ — подпоследовательности $\{u_{k_2 2}\} \to u_2$, $\{\alpha_{k_2 2}\} \to \alpha_2$ и т. д., наконец, $\{u_{k_{n+1} n+1}\} \to u_{n+1}$, $\{\alpha_{k_{n+1} n+1}\} \to \alpha_{n+1}$. Тогда из $u_{k_{n+1}} = \sum\limits_{i=1}^{n+1} \alpha_{k_{n+1} i} u_{k_{n+1} i}$ $\left(u_{k_{n+1} i} \in U, \alpha_{k_{n+1} i} \geqslant 0, \quad i = 1, \ldots, n+1, \quad \sum\limits_{i=1}^{n+1} \alpha_{k_{n+1} i} = 1\right)$ предельным переходом при $k_{n+1} \to \infty$ получим $u = \sum\limits_{i=1}^{n+1} \alpha_i u_i$ $\left(\alpha_i \geqslant 0, \quad \sum\limits_{i=1}^{n+1} \alpha_i = 1\right)$,

где $u_i \in U$ $(i = 1, \ldots, n+1)$ в силу замкнутости U. Следовательно, по теореме 6 $u \in \operatorname{co} U$, что и требовалось.

Заметим, что требование ограниченности множества U в теореме 8 существенно. Например, множество $U = \{u = (x, y) \in E^2\colon\ x \geqslant 0,\ y = \sqrt{x}\}$ замкнуто, но $\operatorname{co} U = \{u = (x, y)\colon\ x \geqslant 0,\ 0 < y \leqslant \sqrt{x}\}$ незамкнуто. Если U — выпуклое замкнутое множество из E^n, то $\operatorname{co} U$ будет замкнутым и без требования ограниченности U, поскольку в этом случае в силу теорем 5, 6 $\operatorname{co} U = U = \overline{U} = \overline{\operatorname{co} U}$.

Теорема 9. *Пусть A — произвольное непустое множество из E^n. Тогда*

$$\sup_{a \in A} \langle c, a\rangle = \sup_{a \in \overline{A}} \langle c, a\rangle = \sup_{a \in \operatorname{co} A} \langle c, a\rangle = \sup_{a \in \overline{\operatorname{co} A}} \langle c, a\rangle \quad \forall c \in E^n.$$

Доказательство. Поскольку $A \subset \overline{A}$, то $\sup\limits_{A} \langle c, a\rangle \leqslant \sup\limits_{\overline{A}} \langle c, a\rangle$. С другой стороны, для любого $a \in \overline{A}$ существуют $a_k \in A$, $\{a_k\} \to a$. Поэтому из $\langle c, a_k\rangle \leqslant \sup\limits_{A} \langle c, a\rangle$ при $k \to \infty$ получим $\langle c, a\rangle \leqslant \sup\limits_{A} \langle c, a\rangle$ для всех $a \in \overline{A}$. Следовательно, $\sup\limits_{\overline{A}} \langle c, a\rangle \leqslant \sup\limits_{A} \langle c, a\rangle$, так что $\sup\limits_{\overline{A}} \langle c, a\rangle = \sup\limits_{A} \langle c, a\rangle$. Отсюда же имеем $\sup\limits_{\operatorname{co} A} \langle c, a\rangle = \sup\limits_{\overline{\operatorname{co} A}} \langle c, a\rangle$. Далее, так как $A \subset \operatorname{co} A$, то $\sup\limits_{A} \langle c, a\rangle \leqslant \sup\limits_{\operatorname{co} A} \langle c, a\rangle$. С другой стороны, для любого $a \in \operatorname{co} A$ согласно теореме 6 найдутся $a_i \in A$, $\alpha_i \geqslant 0$ $(i = 1, \ldots, r)$, $\alpha_1 + \ldots + \alpha_r = 1$ такие, что $a = \sum\limits_{i=1}^{r} \alpha_i a_i$. Поэтому

$$\langle c, a\rangle = \sum_{i=1}^{r} \alpha_i \langle c, a_i\rangle \leqslant \sum_{i=1}^{r} \alpha_i \sup_{a \in A} \langle c, a\rangle = \sup_{A} \langle c, a\rangle$$

для каждого $a \in \operatorname{co} A$. Следовательно, $\sup\limits_{\operatorname{co} A} \langle c, a\rangle \leqslant \sup\limits_{A} \langle c, a\rangle$, так что $\sup\limits_{\operatorname{co} A} \langle c, a\rangle = \sup\limits_{A} \langle c, a\rangle$.

4. Приведем условия существования внутренней точки выпуклого множества.

Теорема 10. *Пусть U — непустое выпуклое множество из E^n. Для того чтобы $\operatorname{int} U \neq \varnothing$, необходимо и достаточно, чтобы $\dim U = n$.*

Доказательство. Необходимость. Пусть $\operatorname{int} U \neq \varnothing$. Тогда для любой внутренней точки v множества U существует ε-окрестность $O(v, \varepsilon) = \{u\colon\ |u - v| < \varepsilon\}$, также принадлежащая U. Отсюда вытекает, что минимальным аффинным множеством, содержащим множество U и, следовательно, шар $O(v, \varepsilon)$, является все пространство E^n. Это значит, что $\operatorname{aff} U = E^n$, $\dim U = n$.

Достаточность. Пусть $\dim U = n$. Тогда $\operatorname{aff} U = E^n$ и найдутся точки $u_0, u_1, \ldots, u_n \in U$ такие, что векторы $u_1 - u_0, \ldots, u_n - u_0$ линейно независимы. Натянем на эти точки симплекс

$$S_n = \left\{u \in E^n\colon\ u = \sum_{i=0}^{n} \alpha_i u_i,\ \alpha_i \geqslant 0,\ i = 0, 1, \ldots, n,\ \sum_{i=0}^{n} \alpha_i = 1\right\}.$$

Согласно теореме 5 $S_n \subset U$, а по теореме 6 S_n выпукло.

Рассмотрим систему линейных алгебраических уравнений

$$\sum_{i=1}^{n} (u_i - u_0) x_i = u - u_0, \quad u \in E^n. \tag{6}$$

Матрица этой системы $A = (u_1 - u_0, \ldots, u_n - u_0)$, столбцами которой являются линейно независимые векторы $u_i - u_0$ $(i = 1, \ldots, n)$, имеет размеры $n \times n$ и невырождена. Поэтому система (6) для каждого $u \in E^n$ имеет и притом единственное решение $x = x(u) = (x_1(u), \ldots, x_n(u))$. Из известных формул Крамера [93, 164] видно, что функции $x_i(u)$ непрерывно (и даже линейно) зависят от u.

Опираясь на это свойство решений системы (6), покажем, что любая точка $w = \sum_{i=0}^{n} \alpha_i u_i$ симплекса S_n при $\alpha_i > 0$ $(i = 0, 1, \ldots, n)$ является внутренней точкой S_n. В самом деле, для такой точки w имеем $w - u_0 = \sum_{i=0}^{n} \alpha_i u_i - \sum_{i=0}^{n} \alpha_i u_0 = \sum_{i=1}^{n} \alpha_i (u_i - u_0)$. Сравнение с (6) показывает, что $\alpha_i = x_i(w)$ $(i = 1, \ldots, n)$. В силу непрерывности $x_i(u)$ тогда $x_i(u) > 0$ $(i = 1, \ldots, n)$ для всех $u \in O(w, \varepsilon) = \{u\colon |u - w| < \varepsilon\}$, где $\varepsilon > 0$ — достаточно малое число. Функция $x_0(u) = 1 - \sum_{i=1}^{n} x_i(u)$ также непрерывна, причем $x_0(w) = 1 - \sum_{i=1}^{n} \alpha_i = \alpha_0 > 0$. Взяв $\varepsilon > 0$ достаточно малым, можем считать, что $x_0(u) > 0$ для всех $u \in O(w, \varepsilon)$.

Таким образом, для каждой точки $u \in O(w, \varepsilon)$ в силу (6) имеем представление $u = u_0 + \sum_{i=1}^{n} x_i(u)(u_i - u_0) = \sum_{i=0}^{n} x_i(u) u_i$, где $x_i(u) > 0$ $\sum_{i=0}^{n} x_i(u) = 1$. Это означает, что $O(w, \varepsilon) \subset S_n$. Следовательно, $w \in \operatorname{int} S_n$. Отсюда и из включения $S_n \subset U$ следует, что $O(w, \varepsilon) \subset U$, т. е. $w \in \operatorname{int} U$ и $\operatorname{int} U \neq \varnothing$.

5. Выпуклое множество $U = \{u = (x, y, z) \in E^3\colon x^2 + y^2 \leqslant 1, z = 0\}$, представляющее собой единичный круг в плоскости $\Gamma = \{u = (x, y, z) \in E^3\colon z = 0\}$, не имеет внутренних точек. Кстати, здесь плоскость Γ представляет собой аффинную оболочку множества U. В то же время, если это множество U рассматривать лишь относительно плоскости Γ (т. е не «признавать» точки E^3, лежащие вне Γ), то U — единичный круг — конечно же, имеет внутренние точки. Приводимая ниже теорема 11 показывает, что это не случайно. Для ее формулировки нам понадобится

Определение 10. Точка $v \in U$ называется *относительно внутренней точкой* множества U, если существует ε-окрестность $O(v, \varepsilon) = \{u \in E^n\colon |u - v| < \varepsilon\}$ точки v такая, что пересечение $O(v, \varepsilon) \cap \operatorname{aff} U$ целиком принадлежит U. Множество всех относительно внутренних точек множества U обозначается через $\operatorname{ri} U$ (иногда обозначают $\operatorname{relint} U$).

Например, если $U = \{u = (x, y, z) \in E^3\colon x^2 + y^2 \leqslant 1, z = 0\}$, то $\operatorname{ri} U = \{u = (x, y, z) \in E^3\colon x^2 + y^2 < 1, z = 0\}$.

Если множество $U \subseteq E^n$ имеет размерность n, т. е. $\dim \operatorname{aff} U = n$, то понятия внутренней и относительно внутренней точки для множества U совпадают и $\operatorname{ri} U = \operatorname{int} U$. Нетрудно указать множества U (например, множество, состоящее из двух различных точек E^n), у которых $\operatorname{ri} U = \varnothing$. Однако для выпуклых множеств верна

Теорема 11. *Если U — непустое выпуклое множество из E^n, то* $\operatorname{ri} U$ *непусто, выпукло. При этом если* $u_0 \in \operatorname{ri} U$, $v \in \overline{U}$, *то* $v_\alpha = v + \alpha(u_0 - v) \in \operatorname{ri} U$ *при всех* α $(0 < \alpha \leqslant 1)$. *Если* $u \in \operatorname{ri} U$, $y \notin \operatorname{ri} U$, $y \in \overline{U}$, *то* $w_\lambda = u + \lambda(y - u) \notin \overline{U}$ *при всех* $\lambda > 1$.

Доказательство. Можем считать, что точка $0 \in U$, так как в противном случае вместо множества U мы рассмотрели бы множество $U - \{v\} = \{w \in E^n;\ w = u - v,\ u \in U\}$, где v — какая-либо точка из U. Тогда $0 \in \operatorname{aff} U = \operatorname{Lin} U$ — подпространство в E^n. Пусть множество U имеет размерность $\dim U = m\,(1 \leqslant m < n)$ (в случае $m = 0$, когда U состоит из единственной точки, утверждения теоремы тривиальны; случай $m = n$ рассмотрен в теоремах 3, 4, 10). Тогда найдутся такие точки $u_0, u_1, \ldots, u_m \in U$, что векторы $e_1 = u_1 - u_0, \ldots, e_m = u_m - u_0$ линейно независимы и образуют базис $\operatorname{aff} U$. Можно дополнить систему $e_1, \ldots, e_m$ векторами $e_{m+1}, \ldots, e_n$ до базиса E^n, причем можно считать, что $\langle e_i, e_j \rangle = 0$ $(i = 1, \ldots, m,\ j = m + 1, \ldots, n)$. В этом базисе $\operatorname{aff} U = \{u = (u^1, \ldots, u^n) \in E^n\colon u^{m+1} = \langle e_{m+1}, u \rangle = 0, \ldots, u^n = \langle e_n, u \rangle = 0\} = \{u = (x, u^{m+1} = 0, \ldots, u^n = 0)\colon x \in E^m\}$, так что $\operatorname{aff} U$ можно отождествить с пространством E^m. Повторив в этом пространстве соответствующие рассуждения из доказательств теорем 3, 4, 10, убеждаемся в справедливости утверждений доказываемой теоремы.

Теорема 12. *Пусть U — непустое выпуклое множество из E^n и $y \in \overline{U}$, $y \notin \operatorname{ri} U$. Тогда существует последовательность* $\{y_k\}$ $(y_k \notin \overline{U},\ y_k \in \operatorname{aff} \overline{U},\ k = 1, 2, \ldots)$, *сходящаяся к y.*

Доказательство. Возьмем какую-либо точку $u \in \operatorname{ri} U$. Согласно теореме 11 тогда $w_\lambda = u + \lambda(y - u) \notin \overline{U}$ при всех $\lambda > 1$. Кроме того, поскольку $\operatorname{aff} \overline{U} = \operatorname{aff} U$, то $u, y \in \operatorname{aff} U$ и, следовательно, $w_\lambda \in \operatorname{aff} U$. Тогда $y_k = w_{\lambda_k} = u + \lambda_k(y - u)$, где $\lambda_k > 1$, $\{\lambda_k\} \to 1$, — искомая точка.

Заметим, что если множество U не является выпуклым, то утверждение теоремы 12 может оказаться неверным. Например, пусть U — множество точек на числовой оси E^1, имеющих рациональные координаты. Тогда $\operatorname{aff} U = E^1$, и любая точка $y \in E^1$ является граничной для U. Таким образом, $\overline{U} = E^1$, и последовательности $\{y_k\} \notin \overline{U}$, сходящейся к y, здесь не существует.

Теорема 13. *Пусть U — выпуклое множество из E^n. Тогда* $\operatorname{ri} U = \operatorname{ri} \overline{U}$, $\overline{U} = \overline{\operatorname{ri} U}$.

Доказательство. Возьмем любую точку $v \in \operatorname{ri} U$. Согласно определению 10 тогда существует такое $\varepsilon > 0$, что $O(v, \varepsilon) \cap \operatorname{aff} U = O(v, \varepsilon) \cap \operatorname{aff} \overline{U} \subset U \subset \overline{U}$. Это значит, что $v \in \operatorname{ri} \overline{U}$. Следовательно, $\operatorname{ri} U \subset \operatorname{ri} \overline{U}$. Докажем обратное включение. Возьмем $w \in \operatorname{ri} \overline{U}$. Тогда существует такое $\varepsilon > 0$, что $O(w, \varepsilon) \cap \operatorname{aff} \overline{U} \subset \overline{U}$.

Возьмем какую-либо точку $v \in \operatorname{ri} U$ и положим $w_\lambda = w + \lambda(w - v)$ $(\lambda \in \mathbf{R})$. Поскольку $v, w \in \operatorname{aff} \overline{U} = \operatorname{aff} U$, то $w_\lambda \in \operatorname{aff} \overline{U}$ при всех $\lambda \in \mathbf{R}$. Кроме того, существует такое $\lambda_0 > 0$, что $w_\lambda \in O(w, \varepsilon)$ для всех $\lambda\,(|\lambda| \leqslant \lambda_0)$. Следовательно, $w_\lambda \in O(w, \varepsilon) \cap \operatorname{aff} \overline{U} \subset \overline{U}\,(|\lambda| \leqslant \lambda_0)$. Из выражения для w_λ следует, что $w = w_\lambda + \dfrac{-\lambda}{1 - \lambda}(v - w_\lambda)$. Возьмем здесь $|\lambda|$ столь малым, чтобы $-\lambda_0 < \lambda < 0$ и $\alpha = (-\lambda)/(1 - \lambda) = |\lambda|/(1 + |\lambda|) \in (0, 1)$. Согласно теореме 11 тогда $w \in \operatorname{ri} U$. Это значит, что $\operatorname{ri} \overline{U} \subset \operatorname{ri} U$. Тем самым установлено, что $\operatorname{ri} \overline{U} = \operatorname{ri} U$.

Далее, так как $\operatorname{ri} U \subset U$, то $\overline{\operatorname{ri} U} \subset \overline{U}$. Возьмем любую точку $u \in \overline{U}$ и $v \in \operatorname{ri} U$. По теореме 11 $u_\alpha = u + \alpha(v - u) \in \operatorname{ri} U$ при всех $\alpha \in (0, 1]$, причем $u_\alpha \to u$ при $\alpha \to 0$. Следовательно, $u \in \overline{\operatorname{ri} U}$. Это значит, что $\overline{U} \subset \overline{\operatorname{ri} U}$. Таким образом, показано, что $\overline{U} = \overline{\operatorname{ri} U}$.

Некоторые другие свойства выпуклых множеств будут рассмотрены ниже.

Упражнения. 1. Пусть U — некоторое множество из E^n, $\overline{U}$ — замыкание множества U. Если $\overline{U}$ выпукло, то можно ли утверждать, что U также выпукло?

2. Существует ли невыпуклое множество, удалив из которого одну точку (или несколько точек), можно получить выпуклое множество? Рассмотреть пример $U = \{u = (x, y) \in E^2\colon x \geqslant 0, y \geqslant 0, x + y < 1\} \cup \{(0, 1)\} \cup \{(1, 0)\}$.

3. Показать, что равенство $\operatorname{ri} U = \operatorname{ri} \overline{U}$ для невыпуклых множеств, вообще говоря, неверно (рассмотреть круг с выколотым центром).

4. Доказать, что если $A \subset B$, то $\overline{A} \subset \overline{B}$, $\operatorname{int} A \subset \operatorname{int} B$, но, вообще говоря, не будет включения $\operatorname{ri} A \subset \operatorname{ri} B$ даже для выпуклых A и B. Рассмотреть пример B — куб в E^3, A — одна из его граней.

5. Если A — выпуклое множество из E^n, то $\operatorname{aff} A = \operatorname{aff}(\operatorname{ri} A)$. Доказать.

6. Доказать, что размерность множества U совпадает с максимальной размерностью симплексов, содержащихся в U.

7. Если A, B — выпуклые множества из E^n, то $\overline{A} + \overline{B} \subset \overline{A + B}$, $\operatorname{ri} A + \operatorname{ri} B = \operatorname{ri}(A + B)$, $\operatorname{ri}(\lambda A) = \lambda \operatorname{ri} A$ для любых действительных чисел λ. Доказать.

8. Доказать, что если A, B — выпуклые множества из E^n, $\operatorname{ri} A \cap \operatorname{ri} B \neq \varnothing$, то $\operatorname{ri} A \cap \operatorname{ri} B = \operatorname{ri}(A \cap B)$. Существенно ли здесь требование $\operatorname{ri} A \cap \operatorname{ri} B \neq \varnothing$? Рассмотреть пример $A = \{u = (x, y) \in E^2\colon x \geqslant 0\}$, $B = \{u = (x, y) \in E^2\colon x \leqslant 0\}$.

9. Если U — открытое множество, то $\operatorname{co} U$ открыто. Доказать.

10. Доказать, что $\overline{\operatorname{co}(A + B)} = \operatorname{co} A + \operatorname{co} B$.

11. Доказать, что $\overline{\operatorname{co}\, U} = \overline{\operatorname{co}}\, U$, где $\overline{\operatorname{co}}\, U$ — пересечение всех выпуклых замкнутых множеств, содержащих U.

12. Доказать, что вершины $u_0, u_1, \ldots, u_m$ m-мерного симплекса $S_m = S_m(u_0, u_1, \ldots, u_m)$ являются его угловыми точками (см. определения 9, 3.2.1).

13. Доказать, что аффинная оболочка множества U состоит из точек вида $\alpha_1 u_1 + \ldots + \alpha_m u_m$ при всевозможных $u_1, \ldots, u_m \in U$, α_i — действительные числа $(i = 1, \ldots, m)$, $\alpha_1 + \ldots + \alpha_m = 1$, и только из них.

14. Пусть A, B — выпуклые замкнутые множества из E^n, причем хотя бы одно из них ограничено. Доказать, что тогда $A + B$ — выпуклое замкнутое множество. Будет ли $A + B$ замкнутым, если A, B не ограничены? Рассмотреть пример $A = \{a = (x, y, z) \in E^3\colon x = z = 0, y \leqslant 0\}$, $B = \{b = (x, y, z) \in E^3\colon x^2 + y^2 \leqslant 2yz, y \geqslant 0\}$.

§ 2. Выпуклые функции

1. В гл. 2 были рассмотрены некоторые свойства выпуклых функций одной переменной. Здесь мы продолжим изучение свойств выпуклых функций многих переменных.

Определение 1. Функция $J(u)$, определенная на выпуклом множестве U, называется *выпуклой* на этом множестве, если

$$J(\alpha u + (1 - \alpha) v) \leqslant \alpha J(u) + (1 - \alpha) J(v) \tag{1}$$

при всех $u, v \in U$, всех α $(0 \leqslant \alpha \leqslant 1)$. Если в (1) при $u \neq v$ равенство возможно только при $\alpha = 0$ и $\alpha = 1$, то функция $J(u)$ называется *строго выпуклой* на U. Функцию $J(u)$ называют

вогнутой [*строго вогнутой*] на выпуклом множестве U, если $-J(u)$ выпукла [строго выпукла] на U.

Если множество U пусто или состоит из одной точки, то функцию на таком множестве нам будет удобно считать выпуклой (или вогнутой) по определению. Подчеркнем также, что всюду, если не оговорено противное, будем рассматривать лишь функции, принимающие конечные значения во всех точках области определения.

Примерами выпуклой функции на всем пространстве E^n служат линейная функция $J(u)=\langle c, u\rangle$ и норма $J(u)=|u|$. Кстати, линейная функция $J(u)=\langle c, u\rangle$ одновременно является и вогнутой на E^n.

В теореме 1.5 было показано, что выпуклое множество U содержит выпуклые комбинации $\sum\limits_{i=1}^{m}\alpha_i u_i$ $\left(\alpha_i \geqslant 0,\ i=1,\ldots,m,\ \sum\limits_{i=1}^{m}\alpha_i=1\right)$ любых своих точек $u_1, \ldots, u_m$ при любых $m=2, 3, \ldots$ Пользуясь индукцией по той же схеме, какая была использована при доказательстве теоремы 1.5, нетрудно показать, что для любой выпуклой функции $J(u)$ на выпуклом множестве имеет место *неравенство Йенсена*

$$J\left(\sum_{i=1}^{m}\alpha_i u_i\right)\leqslant\sum_{i=1}^{m}\alpha_i J(u_i) \tag{2}$$

для любых $m=1, 2, \ldots$, любых $u_i\in U$, $\alpha_i\geqslant 0$ $(i=1, \ldots, m)$, $\sum\limits_{i=1}^{m}\alpha_i=1$.

2. Как и в случае выпуклых функций одной переменной, выпуклые функции многих переменных на выпуклом множестве не могут иметь локальных минимумов. Точнее, верна

Теорема 1. *Пусть U — выпуклое множество, а функция $J(u)$ определена и выпукла на U. Тогда всякая точка локального минимума $J(u)$ одновременно является точкой ее глобального минимума на U, причем множество*

$$U_*=\left\{u\colon\ u\in U,\ J(u)=J_*=\inf_U J(u)\right\}$$

выпукло. Если $J(u)$ строго выпукла на U, то U_* содержит не более одной точки.

Доказательство. Пусть u_* — точка локального минимума функции $J(u)$ на множестве U. Это значит, что существует окрестность $O(u_*, \varepsilon)=\{u\colon |u-u_*|<\varepsilon\}$ точки u_* такая, что $J(u_*)\leqslant J(v)$ для всех $v\in O(u_*, \varepsilon)\cap U$. Возьмем произвольную точку $u\in U$ и число $\alpha>0$, столь малое, что $\alpha|u-u_*|<\varepsilon$. Тогда $u_*+\alpha(u-u_*)\in O(u_*, \varepsilon)\cap U$, и с учетом выпуклости функции $J(u)$ имеем $J(u_*)\leqslant J(u_*+\alpha(u-u_*))\leqslant J(u_*)+\alpha(J(u)-J(u_*))$

или $0 \leqslant \alpha (J(u) - J(u_*))$. Сокращая на $\alpha > 0$, отсюда получаем $J(u) \geqslant J(u_*)$ при любом $u \in U$. Следовательно, $u_* \in U_*$.

Пусть теперь $u, \quad v \in U_*$, т. е. $J(u) = J(v) = J_* \quad (u, v \in U)$. Тогда

$$J_* \leqslant J(\alpha u + (1-\alpha) v) \leqslant \alpha J(u) + (1-\alpha) J(v) = J_*, \tag{3}$$

т. е. $J(\alpha u + (1-\alpha) v) = J_*$ при всех α $(0 \leqslant \alpha \leqslant 1)$. Следовательно, $\alpha u + (1-\alpha) v \in U_*$ $(0 \leqslant \alpha \leqslant 1)$. Выпуклость U_* доказана.

Если $u \neq v$, то для строго выпуклых функций неравенства (3) не могут обратиться в равенства при $0 < \alpha < 1$. Следовательно, строго выпуклая функция может достигать своей нижней грани на выпуклом множестве не более чем в одной точке.

Примеры 1.11.1, 1.11.3—1.11.5 показывают, что у выпуклых функций множество U_* может быть пустым, может содержать одну или бесконечно много точек.

3. Прежде чем переходить к формулировке критерия оптимальности для выпуклых функций, остановимся на одном характеристическом свойстве гладких выпуклых функций.

Теорема 2. *Пусть U — выпуклое множество, функция $J(u) \in C^1(U)$. Тогда для того чтобы $J(u)$ была выпукла на U, необходимо и достаточно выполнения неравенства*

$$J(u) \geqslant J(v) + \langle J'(v), u - v \rangle \qquad \forall u, v \in U. \tag{4}$$

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U. Перепишем неравенство (1) в виде

$$J(v + \alpha(u - v)) - J(v) \leqslant \alpha [J(u) - J(v)], \quad 0 \leqslant \alpha \leqslant 1, \quad u, v \in U.$$

Применяя к левой части формулу конечных приращений (2.3.2), имеем

$$\alpha \langle J'(v + \theta\alpha(u - v)), u - v \rangle \leqslant \alpha [J(u) - J(v)], \qquad 0 \leqslant \theta \leqslant 1.$$

Деля обе части этого неравенства на $\alpha > 0$ и устремляя $\alpha \to +0$, с учетом гладкости функции получим требуемое неравенство (4).

Достаточность. Пусть для некоторой гладкой функции на выпуклом множестве выполняется неравенство (4) при всех $u, v \in U$. Покажем, что тогда $J(u)$ выпукла на U. Возьмем произвольные точки $u, v \in U$ и число α $(0 \leqslant \alpha \leqslant 1)$. Положим $u_\alpha = \alpha u + (1-\alpha) v$. Из (4) получим

$$J(u) - J(u_\alpha) \geqslant \langle J'(u_\alpha), u - u_\alpha \rangle,$$
$$J(v) - J(u_\alpha) \geqslant \langle J'(u_\alpha), v - u_\alpha \rangle.$$

Умножим первое из этих неравенств на α, а второе — на $1-\alpha$ и сложим. Получим $\alpha J(u) + (1-\alpha) J(v) - J(u_\alpha) \geqslant \langle J'(u_\alpha), u_\alpha \rangle$, что равносильно неравенству (1).

Неравенство (4) имеет простой геометрический смысл. Как известно [10, 160, 165, 233], гиперплоскость $\Gamma = \{(u, \gamma) \in E^{n+1}$:

$u \in E^n$, $\gamma = J(v) + \langle J'(v), u - v\rangle\}$ является касательной плоскостью к графику функции $\gamma = J(u)$ в точке v. Поэтому неравенство (4) означает, что график выпуклой функции лежит не ниже касательной плоскости к этому графику в любой точке $v \in U$ (ср. с теоремой 1.8.4).

4. Следующая теорема, называемая *критерием оптимальности* для выпуклых функций, дает необходимые и достаточные условия минимума гладких выпуклых функций на выпуклом множестве.

Теорема 3. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$ и пусть U_* — множество точек минимума $J(u)$ на U. Тогда в любой точке $u_* \in U_*$ необходимо выполняется неравенство*

$$\langle J'(u_*), u - u_*\rangle \geqslant 0 \qquad \forall u \in U, \tag{5}$$

а в случае $u_* \in \operatorname{int} U$ неравенство (5) *превращается в равенство $J'(u_*) = 0$. Если, кроме того, $J(u)$ выпукла на U, то условие* (5) *является достаточным для того, чтобы $u_* \in U_*$.*

Доказательство. Необходимость. Пусть $u_* \in U_*$. Тогда при любых $u \in U$ и $\alpha \in [0, 1]$ с помощью формулы (2.2.1), определяющей градиент функции, имеем $0 \leqslant J(u_* + \alpha(u - u_*)) - J(u_*) = \alpha\langle J'(u_*), u - u_*\rangle + o(\alpha)$ или

$$0 \leqslant \langle J'(u_*), u - u_*\rangle + o(\alpha)/\alpha, \qquad 0 < \alpha < 1.$$

Отсюда при $\alpha \to +0$ получим условие (5). Если $u_* \in \operatorname{int} U$, то для любого $e \in E^n$ найдется $\varepsilon_0 > 0$ такое, что $u = u_* + \varepsilon e \in U$ при всех ε $(|\varepsilon| \leqslant \varepsilon_0)$. Полагая в (5) $u = u_* + \varepsilon e$, получаем $\varepsilon\langle J'(u_*), e\rangle \geqslant 0$ при всех ε $(|\varepsilon| \leqslant \varepsilon_0)$, что возможно только при $\langle J'(u_*), e\rangle = 0$. В силу произвола e отсюда имеем $J'(u_*) = 0$.

Заметим, что если u_* — граничная точка множества U, то равенство $J'(u_*) = 0$ может выполняться, может и не выполняться. Например, если $J(u) = u^2$, $U = \{u \in E^1\colon 1 \leqslant u \leqslant 2\}$, то $u_* = 1$, $J'(u_*) = 2$ и условие (5) в точке u_*, конечно, выполняется. Если ту же функцию $J(u) = u^2$ рассматривать на отрезке $U = \{u \in E^1\colon 0 \leqslant u \leqslant 2\}$, то $u_* = 0$, $J'(u_*) = 0$, хотя $u_* = 0$ — граничная точка U. Таким образом, условие (5) является естественным обобщением условия стационарности (2.2.5) на задачи минимизации гладких функций на выпуклых множествах $U \neq E^n$.

Достаточность. Пусть функция $J(u) \in C^1(U)$ является выпуклой на U, пусть для некоторой точки $u_* \in U$ выполнено условие (5). Тогда из неравенства (4) при $v = u_*$ получим $J(u) - J(u_*) \geqslant \langle J'(u_*), u - u_*\rangle \geqslant 0$ или $J(u) \geqslant J(u_*)$ при всех $u \in U$, т. е. $u_* \in U_*$.

5. Сформулируем и докажем два критерия выпуклости для гладких функций.

Теорема 4. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$. Тогда для выпуклости функции $J(u)$ на U необходимо и доста-*

точно, чтобы

$$\langle J'(u) - J'(v), u - v\rangle \geqslant 0 \quad \forall u, \quad v \in U. \tag{6}$$

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U. Тогда для любых u, $v \in U$ имеет место неравенство (4). Поменяв в (4) переменные u и v ролями, получим

$$J(v) \geqslant J(u) + \langle J'(u), v - u\rangle.$$

Сложив это неравенство с (4), придем к условию (6).

Достаточность. Пусть для некоторой функции $J(u) \in C^1(U)$ выполнено условие (6). Тогда с помощью формулы конечных приращений (2.3.2) для любых u, $v \in U$ и $\alpha \in [0, 1]$ имеем

$$\begin{aligned}
&\alpha J(u) + (1-\alpha) J(v) - J(\alpha u + (1-\alpha) v) = \\
&\qquad = \alpha [J(u) - J(\alpha u + (1-\alpha) v)] + \\
&\qquad + (1-\alpha)[J(v) - J(\alpha u + (1-\alpha) v)] = \\
&= \alpha \int_0^1 \langle J'(\alpha u + (1-\alpha) v + t(u - \alpha u - (1-\alpha) v)), \quad u - \alpha u - \\
&- (1-\alpha) v\rangle\, dt + (1-\alpha) \int_0^1 \langle J'(\alpha u + (1-\alpha) v + t(v - \alpha u - \\
&- (1-\alpha) v)), \quad v - \alpha u - (1-\alpha) v\rangle\, dt = \\
&= \alpha(1-\alpha) \int_0^1 \langle J'(\alpha u + (1-\alpha) v + t(1-\alpha)(u - v)) - \\
&\qquad - J'(\alpha u + (1-\alpha) v + t\alpha(v - u)), u - v\rangle\, dt,
\end{aligned}$$

или

$$\begin{aligned}
&\alpha J(u) + (1-\alpha) J(v) - J(\alpha u + (1-\alpha) v) = \\
&\qquad = \alpha(1-\alpha) \int_0^1 \langle J'(z_1) - J'(z_2), z_1 - z_2\rangle \frac{1}{t}\, dt,
\end{aligned} \tag{7}$$

где $z_1 = \alpha u + (1-\alpha) v + t(1-\alpha)(u - v)$, $z_2 = \alpha u + (1-\alpha) v + t\alpha(v - u)$. Из условия (6) имеем $\langle J'(z_1) - J'(z_2), z_1 - z_2\rangle \geqslant 0$ при всех t $(0 < t \leqslant 1)$. Это значит, что правая часть (7) и, следовательно, левая часть (7) неотрицательна при любом выборе u, $v \in U$, $\alpha \in [0, 1]$, т. е. $J(u)$ выпукла на U.

Заметим, что для функций одной переменной неравенство (6) равносильно неубыванию производной $J'(u)$. Это значит, что доказанная теорема 4 является естественным обобщением теоремы 1.8.8 на случай гладких функций многих переменных.

Следующий критерий выпуклости обобщает теорему 1.8.9.

Теорема 5. *Пусть U — выпуклое множество из E^n, $J(u) \in C^2(U)$. Тогда для выпуклости $J(u)$ на U необходимо и достаточно, чтобы*

$$\langle J''(u)\xi, \xi\rangle \geqslant 0 \tag{8}$$

при всех $u \in U$ и всех $\xi = (\xi^1, \ldots, \xi^n)$, принадлежащих подпространству $L = \operatorname{Lin} U$, параллельному аффинной оболочке множества U (в частности, если $\operatorname{int} U \neq \varnothing$, то (8) выполняется при всех $\xi \in E^n$).

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U. Пусть $\operatorname{aff} U = \{u \in E^n\colon Au = b\}$, где A — некоторая матрица размера $m \times n$, а $b \in E^n$ (см. пример 1.4). Тогда подпространство L, параллельное $\operatorname{aff} U$, имеет вид $L = \{\xi \in E^n\colon A\xi = 0\}$. Далее, согласно теореме 1.10 $\operatorname{ri} U \neq \varnothing$. Возьмем произвольные $\xi \in L$, $u \in \operatorname{ri} U$. Тогда $A(u + \varepsilon\xi) = Au + \varepsilon A\xi = Au = b$, т. е. $u + \varepsilon\xi \in \operatorname{aff} A$ при всех ε.

По определению 1.10 относительно внутренней точки множества найдется такое число $\varepsilon_0 > 0$, что $u + \varepsilon\xi \in U$ при всех ε ($|\varepsilon| \leqslant \varepsilon_0$). Поскольку для гладкой выпуклой функции справедливо неравенство (6), то из него с учетом формулы (2.3.4) имеем $\langle J'(u + \varepsilon\xi) - J'(u), \xi\rangle\varepsilon = \langle J''(u + \theta\varepsilon\xi)\xi, \xi\rangle\varepsilon^2 \geqslant 0$, $0 \leqslant \theta \leqslant 1$, или $\langle J''(u + \theta\varepsilon\xi)\xi, \xi\rangle \geqslant 0$ для всех ε ($0 < |\varepsilon| \leqslant \varepsilon_0$). Отсюда, пользуясь непрерывностью $J''(u)$, при $\varepsilon \to 0$ получим условие (8) для всех $u \in \operatorname{ri} U$. Если $u \in U \setminus \operatorname{ri} U$, то, как следует из теоремы 1.10, существует последовательность $\{u_k\} \in \operatorname{ri} U$, сходящаяся к u. По доказанному $\langle J''(u_k)\xi, \xi\rangle \geqslant 0$ при всех $\xi \in L$. Отсюда при $k \to \infty$ получим неравенство (8) и для точек $u \in U \setminus \operatorname{ri} U$.

Достаточность. Пусть $J(u) \in C^2(U)$ и выполнено условие (8). Возьмем произвольные точки u, $v \in U$. Тогда $\xi = u - v \in L$. Пользуясь формулой (2.3.4) и неравенством (8) при $\xi = u - v$, получим

$$\langle J'(u) - J'(v), u - v\rangle = \\ = \langle J''(v + \theta(u - v))(u - v), u - v\rangle \geqslant 0 \qquad \forall u, \quad v \in U.$$

Таким образом, для функции $J(u)$ выполняется условие (6). Из теоремы 4 следует выпуклость $J(u)$ на U.

Замечание 1. Если $\operatorname{int} U \neq \varnothing$, то $L = E^n$ и условие (8) должно выполняться при всех $\xi \in E^n$. Следующий пример показывает, что при $\operatorname{int} U = \varnothing$ условие (8) может и не выполняться при каждом $\xi \in E^n$.

Пример 1. Пусть $J(u) = x^2 - y^2$, $U = \{u = (x, y) \in E^2\colon y = 0\}$. Ясно, что $J(u)$ выпукла на U. Но условие $\langle J''(u)\xi, \xi\rangle = 2(\xi^1)^2 - 2(\xi^2)^2 \geqslant 0$ не выполняется, например, для $\xi = (0, 1)$. Здесь $\operatorname{int} U = \varnothing$, $\operatorname{aff} U = U = L$.

Однако если требовать, чтобы условие (8) выполнялось лишь для тех ξ, которые принадлежат подпространству, параллельному

$\operatorname{aff} U$ (а не для всех $\xi \in E^n$), то теорема 5 остается справедливой для любых выпуклых множеств $U \subseteq E^n$. В этом случае доказательство необходимости проводится по той же схеме, что и выше, нужно лишь сначала рассмотреть точки $u \in \operatorname{ri} U$, а для точек $u \in U \setminus \operatorname{ri} U$ воспользоваться последовательностью $\{u_m\} \in \operatorname{ri} U$, сходящейся к u. Доказательство достаточности остается без изменений, так как вектор $\xi = u - v$ при любых $u, v \in U$ принадлежит подпространству, параллельному $\operatorname{aff} U$.

Замечание 2. Условие (8) представляет собой условие неотрицательности квадратичной формы

$$\langle J''(u)\xi, \xi\rangle = \sum_{i,j=1}^{n} \frac{\partial^2 J(u)}{\partial u^i \partial u^j} \xi^i \xi^j$$

на E^n. Имеется следующий простой алгебраический критерий неотрицательности квадратичной формы [105]: для того чтобы $\langle A\xi, \xi\rangle = \sum_{i,j=1}^{n} a_{ij}\xi^i\xi^j \geqslant 0$ при всех $\xi = (\xi^1, \ldots, \xi^n)$, необходимо и достаточно, чтобы все главные миноры матрицы $A = [a_{ij}]$ были неотрицательны.

Напоминаем, что *главными минорами* матрицы $A = [a_{ij}]$ называются всевозможные определители

$$\Delta_{i_1 \ldots i_k} = \det \begin{bmatrix} a_{i_1 i_1} & \ldots & a_{i_1 i_k} \\ \cdot & \cdot & \cdot \\ a_{i_k i_1} & \ldots & a_{i_k i_k} \end{bmatrix},$$

где $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$ $(k = 1, \ldots, n)$. Симметричную матрицу A называют *неотрицательно определенной*, если она является матрицей неотрицательно определенной квадратичной формы, и обозначают $A \geqslant 0$. Отметим также, что неотрицательность квадратичной формы $\langle J''(u)\xi, \xi\rangle$ равносильна тому, что собственные числа $\lambda_1(u), \ldots, \lambda_n(u)$ матрицы $J''(u)$ (т. е. решения уравнения $\det|J''(u) - \lambda I| = 0$, I — единичная матрица размера $n \times n$) неотрицательны при всех $u \in U$.

Учитывая замечание 2, можно сказать, что условие (8) является достаточно удобным средством проверки выпуклости дважды гладких функций небольшого числа переменных.

Пример 2. Определим, при каких a, b, c функция

$$J(u) = x^2 + 2axy + by^2 + cz^2$$

будет выпуклой на E^3. Здесь

$$J''(u) = \begin{bmatrix} 2 & 2a & 0 \\ 2a & 2b & 0 \\ 0 & 0 & 2c \end{bmatrix}.$$

Условие неотрицательности всех главных миноров этой матрицы дает искомые условия на a, b, c: $b - a^2 \geqslant 0$, $c \geqslant 0$.

Пример 3. Пусть

$$J(u)=\frac{1}{2}\langle Au, u\rangle-\langle b, u\rangle, \qquad u\in E^n, \tag{9}$$

где A — симметричная неотрицательно определенная матрица размера $n\times n$, $b\in E^n$. В частности, если $A=2I$ — единичная матрица, $b=0$, то $J(u)=\langle u, u\rangle=|u|^2$.

Приращение функции (9) нетрудно записать в виде

$$J(u+h)-J(u)=\langle Au-b, h\rangle+\frac{1}{2}\langle Ah, h\rangle \tag{10}$$

при любых u, $h\in E^n$. Из (10) имеем

$$J'(u)=Au-b, \quad J''(u)=A.$$

По условию $A\geqslant 0$. Отсюда и из теоремы 5 следует выпуклость $J(u)$ на E^n. Согласно теореме 3 для того, чтобы функция (9) достигала своей нижней грани на E^n в точке u, необходимо и достаточно, чтобы u_* являлась решением линейной алгебраической системы $Au=b$. Указанная связь между задачей минимизации функции (9) на E^n и системой $Au=b$ с матрицей $A\geqslant 0$ лежит в основе ряда численных методов линейной алгебры [4, 13, 54].

Пример 4. Пусть

$$J(u)=|Au-b|^2, \qquad u\in E^n, \tag{11}$$

где A — матрица порядка $m\times n$, $b\in E^n$. Покажем, что такая функция выпукла на E^n. Для этого вычислим ее производные.

Пользуясь следующей просто проверяемой формулой

$$\langle Ax, y\rangle=\langle x, A^Ty\rangle, \qquad x\in E^n, \quad y\in E^m,$$

где A^T — матрица, полученная транспонированием матрицы A, нетрудно представить приращение функции (11) в виде

$$J(u+h)-J(u)=2\left\langle A^T(Au-b), h\right\rangle+\frac{1}{2}\left\langle 2A^TAh, h\right\rangle$$

при всех u, $h\in E^n$. Отсюда имеем

$$J'(u)=2A^T(Au-b), \quad J''(u)=2A^TA.$$

Но $\langle J''(u)\xi, \xi\rangle=2\langle A^TA\xi, \xi\rangle=2\langle A\xi, A\xi\rangle=2|A\xi|^2\geqslant 0$ при всех $\xi\in E^n$. В силу теоремы 5 функция (11) выпукла на E^n.

Согласно теореме 3 для того, чтобы функция (11) достигала своей нижней грани на E^n в точке u, необходимо и достаточно, чтобы u удовлетворяла системе линейных алгебраических уравнений

$$A^TAu=A^Tb.$$

6. Посмотрим, как влияют на выпуклость сложение, умножение на число и некоторые другие операции над выпуклыми функциями. Легко доказывается

Теорема 6. *Если функции* $J_i(u)$ $(i=1, \ldots, m)$ *выпуклы на выпуклом множестве, то функция*

$$J(u)=\alpha_1 J_1(u)+\ldots+\alpha_m J_m(u)$$

выпукла на этом множестве при любых $\alpha_i \geqslant 0$ $(i=1, \ldots, m)$.

Теорема 7. *Пусть* $J_i(u)$ $(i \in I)$ — *произвольное семейство функций, выпуклых на выпуклом множестве* U, *пусть*

$$J(u)=\sup_{i \in I} J_i(u), \qquad u \in U.$$

Тогда функция $J(u)$ *выпукла на* U.

Доказательство. Возьмем произвольные точки u, $v \in U$ и числа $\alpha \in [0, 1]$, $\varepsilon > 0$. Положим $u_\alpha = \alpha u + (1-\alpha) v$. По определению верхней грани найдется индекс $i = i(\varepsilon, \alpha) \in I$ такой, что $J(u_\alpha) - \varepsilon \leqslant J_i(u_\alpha)$. С учетом выпуклости $J_i(u)$ тогда имеем

$$J(u_\alpha) \leqslant J_i(u_\alpha)+\varepsilon \leqslant \alpha J_i(u)+(1-\alpha) J_i(v)+\varepsilon \leqslant$$
$$\leqslant \alpha J(u)+(1-\alpha) J(v)+\varepsilon$$

или $J(u_\alpha) \leqslant \alpha J(u)+(1-\alpha) J(v)+\varepsilon$ при любом $\varepsilon > 0$. Отсюда при $\varepsilon \to +0$ следует выпуклость $J(u)$, что и требовалось.

Следствие 1. *Пусть функция* $g(u)$ *выпукла на выпуклом множестве* U. *Тогда функция*

$$g^+(u)=\max\{g(u);\ 0\}$$

выпукла на U.

Теорема 8. *Пусть функция* $\varphi(t)$ *одной переменной выпукла и не убывает на отрезке* $[a, b]$ (*возможность* $a=-\infty$ *или* $b=\infty$ *здесь не исключается*), *пусть функция* $g(u)$ *выпукла на выпуклом множестве* $U \subseteq E^n$, *причем* $g(u) \in [a, b]$ *при всех* $u \in U$. *Тогда функция*

$$J(u)=\varphi(g(u))$$

выпукла на U.

Доказательство. Возьмем произвольные u, $v \in U$ и $\alpha \in [0, 1]$. Тогда

$$J(\alpha u+(1-\alpha) v)=$$
$$=\varphi(g(\alpha u+(1-\alpha) v)) \leqslant \varphi(\alpha g(u)+(1-\alpha) g(v)) \leqslant$$
$$\leqslant \alpha \varphi(g(u))+(1-\alpha) \varphi(g(v))=\alpha J(u)+(1-\alpha) J(v),$$

что и требовалось.

Иногда удобнее пользоваться другим вариантом этой теоремы: если функция $\varphi(t)$ выпукла и не возрастает на отрезке $[a, b]$, а $g(u)$ вогнута на выпуклом множестве $U \subseteq E^n$, $g(u) \in [a, b]$ при $u \in U$, то функция $J(u)=\varphi(g(u))$ выпукла на U.

Следствие 1. *Если функция $g(u)$ выпукла и неотрицательна на выпуклом множестве U, то функция*

$$J(u)=(g(u))^p$$

выпукла на U при всех $p\geqslant 1$.

Следствие 2. *Если функция $g(u)$ выпукла на выпуклом множестве U, то функция*

$$J(u)=(\max\{0;\ g(u)\})^p=(g^+(u))^p$$

выпукла на U при всех $p\geqslant 1$.

Следствие 3. *Если функция $g(u)$ выпукла на выпуклом множестве U, причем $g(u)<0$ при всех $u\in U$, то функции*

$$J(u)=-1/g(u),\quad J(u)=\max\{-\ln(-g(u));\ 0\}^p,\quad p\geqslant 1,$$

выпуклы на U.

Как увидим ниже, функции, указанные в следствиях к теоремам 7, 8, будут использованы при описании различных методов минимизации (например, в методах штрафных и барьерных функций и др.).

7. Выпуклые функции являются удобным средством задания выпуклых множеств. Это связано с тем, что надграфик всякой выпуклой функции является выпуклым множеством.

Определение 2. *Надграфиком* (или *эпиграфом*) всякой функции $J(u)$, определенной на множестве $U\subseteq E^n$, называется множество (рис. 4.5)

$$\operatorname{epi} J=\{(u,\ \gamma)\in E^{n+1}:\ u\in U,\quad \gamma\geqslant J(u)\}.$$

Теорема 9. *Для того чтобы функция $J(u)$, определенная на выпуклом множестве U, была выпуклой на U, необходимо и достаточно, чтобы ее надграфик был выпуклым множеством.*

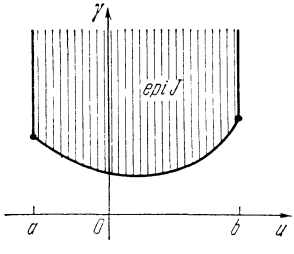

Рис. 4.5. (Надграфик)

Доказательство. Необходимость. Пусть функция $J(u)$ выпукла на выпуклом множестве U. Возьмем две произвольные точки $z_1=(u_1,\ \gamma_1)$, $z_2=(u_2,\ \gamma_2)\in\operatorname{epi} J$ и составим их выпуклую комбинацию $z_\alpha=\alpha z_1+(1-\alpha)z_2=(\alpha u_1+(1-\alpha)u_2,\ \alpha\gamma_1+(1-\alpha)\gamma_2)$ $(0\leqslant\alpha\leqslant 1)$. Из выпуклости U следует, что $u_\alpha=\alpha u_1+(1-\alpha)u_2\in U$. Из выпуклости функции $J(u)$, учитывая, что $z_1,\ z_2\in\operatorname{epi} J$, имеем $J(u_\alpha)\leqslant\alpha J(u_1)+(1-\alpha)J(u_2)\leqslant\alpha\gamma_1+(1-\alpha)\gamma_2$. Следовательно, $z_\alpha\in\operatorname{epi} J$ при всех $\alpha\in[0,\ 1]$. Выпуклость $\operatorname{epi} J$ доказана.

Достаточность. Пусть $\operatorname{epi} J$ — выпуклое множество. Возьмем произвольные $u_1,\ u_2\in U$ и $\alpha\in[0,\ 1]$. Тогда $z_1=(u_1,\ J(u_1))$, $z_2=(u_2,\ J(u_2))\in\operatorname{epi} J$. В силу выпуклости $\operatorname{epi} J$

точка $z_\alpha = \alpha z_1 + (1-\alpha) z_2 \in \operatorname{epi} J$. Это значит, что $\alpha J(u_1) + (1-\alpha) J(u_2) \leqslant J(\alpha u_1 + (1-\alpha) u_2)$. Выпуклость $J(u)$ доказана.

Теорема 10. *Пусть U — выпуклое множество, а функция $J(u)$ выпукла на U. Тогда множество $M(c) = \{u\colon u \in U,\ J(u) \leqslant c\}$ выпукло при любом c.*

Доказательство. Возьмем произвольные u, $v \in M(c)$, $\alpha \in [0, 1]$. Используя выпуклость множества U и функции $J(u)$, имеем $J(\alpha u + (1-\alpha) v) \leqslant \alpha J(u) + (1-\alpha) J(v) \leqslant c$, т. е. $\alpha u + (1-\alpha) v \in M(c)$, что и требовалось.

Заметим, что обратное утверждение здесь неверно: из выпуклости множества $M(c)$ при любом c, вообще говоря, не следует выпуклость функции $J(u)$. Например, множество $M(c) = \{u\colon u \in E^1,\ u^3 \leqslant c\}$ выпукло при любом c, а функция $J(u) = u^3$ невыпукла на E^1 (см. упражнение 33).

Теорема 11. *Пусть U_0 — выпуклое множество, функции $g_i(u)$ $(i = 1, \ldots, m)$ выпуклы на U_0, а $g_i(u) = \langle a_i, u\rangle - b_i$ $(i = m+1, \ldots, s)$, где a_i — заданные векторы из E^n, b_i — заданные числа $(i = m+1, \ldots, s)$. Тогда выпукло множество*

$$U = \{u\colon u \in U_0,\ g_i(u) \leqslant 0,\ i = 1, \ldots, m;\ g_i(u) = 0,\ i = m+1, \ldots, s\}. \tag{12}$$

Доказательство. В силу теоремы 10 множество $U_i = \{u\colon u \in U_0,\ g_i(u) \leqslant 0\}$ выпукло при всех $i = 1, \ldots, m$. Выпукло также множество $M = \{u \in E^n\colon \langle a_i, u\rangle - b_i = 0,\ i = m+1, \ldots, s\}$ — см. пример 1.4. Тогда множество (12), являющееся пересечением выпуклых множеств $U_1, \ldots, U_m$, M, само будет выпуклым.

8. Рассмотренное в теореме 3 условие оптимальности сформулировано для непрерывно-дифференцируемых функций. Однако аналогичное условие можно получить при гораздо меньших ограничениях на функцию, используя лишь существование производных по направлениям. Напоминаем, что *производной функции $J(u)$ в точке u по направлению e* ($|e| = 1$) называется число

$$\frac{dJ(u)}{de} = \lim_{t \to +0} \frac{J(u + te) - J(u)}{t}. \tag{13}$$

Заметим, что для определения производной по направлению в точке u нужно, чтобы $u + te$ принадлежало области определения $J(u)$ при $0 \leqslant t \leqslant t_0$ хотя бы при малом $t_0 > 0$.

Определение 3. Пусть U — некоторое множество из E^n, пусть $u \in U$. Направление $e \neq 0$ называется *возможным* в точке u, если существует число $t_0 > 0$ такое, что $u + te \in U$ при всех t $(0 \leqslant t \leqslant t_0)$. Иначе говоря, достаточно малое перемещение из точки u по возможному направлению не выводит за пределы множества U.

Очевидно, если $u \in \operatorname{int} U$, то любое направление $e \neq 0$ является возможным в этой точке. В граничных точках множества возможное направление может и не существовать.

Пример 5. Пусть $U = \{x = (x, y) \in E^2\colon x \geqslant 0,\ x^2 \leqslant y \leqslant 2x^2\}$. Нетрудно видеть, что в граничной точке (0, 0) нет ни одного возможного направления.

Для выпуклых множеств U, содержащих не менее двух точек, приведенная в примере 5 ситуация невозможна: в любой точке u такого выпук-

лого множества U имеется хотя бы одно возможное направление, причем направление $e \neq 0$ будет возможным в точке u тогда и только тогда, когда существуют точка $v \in U$ $(v \neq u)$ и число $\gamma > 0$ такие, что $e = \gamma(v - u)$.

Таким образом, если функции $J(u)$ определена на множестве U, а направление $e(|e| = 1)$ является возможным в точке $u \in U$, то функция $f(t) = J(u + te)$ определена на отрезке $[0, t_0]$, где $t_0 > 0$, и $dJ(u)/de = f'(+0)$ — правая производная $f(t)$ в точке $t = 0$.

Заметим, что если функция $J(u)$ определена в некоторой ε-окрестности точки u и дифференцируема в этой точке, то $J(u)$ имеет производные по всем направлениям, причем

$$\frac{dJ(u)}{de} = \langle J'(u), e\rangle, \quad |e| = 1 \tag{14}$$

(ср. с (2.3.1) при $t \to +0$). Однако обратное неверно: из того, что функция в некоторой точке имеет производные по всем направлениям, вообще говоря, не следует ее дифференцируемость в этой точке, и более того, нельзя гарантировать даже ее непрерывность.

Пример 6. Пусть

$$J(u) = J(x, y) = \begin{cases} \dfrac{x^2 y}{x^4 + y^2}, & u = (x, y) \neq 0, \\ 0, & u = (0, 0) = 0. \end{cases}$$

Возьмем произвольное направление $e = (\cos\alpha, \sin\alpha)$. Тогда

$$\frac{J(0 + te) - J(0)}{t} \equiv \frac{1}{t} J(t\cos\alpha, t\sin\alpha) = \frac{\cos^2\alpha \sin\alpha}{t^2\cos^4\alpha + \sin^2\alpha}, \quad t > 0.$$

Отсюда имеем

$$\frac{dJ(0)}{de} = \begin{cases} \cos^2\alpha/\sin\alpha, & \sin\alpha \neq 0, \\ 0, & \sin\alpha = 0. \end{cases}$$

Однако эта функция разрывна в точке $u = 0$. В самом деле, устремим точку $u = (x, y)$ к нулю по параболе $y = x^2$. Тогда $J(x, x^2) \equiv 1/2 \not\to J(0, 0) = 0$.

Таким образом, требование существования производных по направлению существенно менее жесткое, чем требование дифференцируемости. В связи с этим представляет интерес получить условия оптимальности в терминах производных по направлению.

Теорема 12. *Пусть U — выпуклое множество, U_* — множество точек минимума функции $J(u)$ на U, пусть в точке $u_* \in U$ функция $J(u)$ имеет производные по всем возможным направлениям. Тогда необходимо выполняется условие*

$$\frac{dJ(u_*)}{de} \geqslant 0 \tag{15}$$

для всех возможных направлений $e(|e| = 1)$ в точке u_*. Если, кроме того, функция $J(u)$ выпукла на U, то условие (15) *достаточно для того, чтобы $u_* \in U$.*

Доказательство. Необходимость. Пусть $u_* \in U$ и e $(|e| = 1)$ — возможное направление в точке u_*. Тогда $J(u_* + te) \geqslant J(u_*)$ или $(J(u_* + te) - J(u_*))/t \geqslant 0$ при всех достаточно малых $t > 0$. Отсюда при $t \to +0$ получим условие (15).

Достаточность. Пусть $J(u)$ — выпуклая функция на U, пусть в некоторой точке $u_* \in U$ выполняется условие (15). Возьмем любую точку $u \in U$ $(u \neq u_*)$ и положим $e = (u - u_*)/|u - u_*|$. Направление e — возможное в точке u_*, так как $u_* + te \in U$ при всех t $(0 \leqslant t \leqslant t_0 = |u - u_*|$, $t_0 > 0)$. Из условия (15) тогда имеем $f'(+0) \geqslant 0$, где $f(t) = J(u_* + te)$.

Ниже в теореме 13 будет показано, что $f(t)$ выпукла на $[0, t_0]$. Из неравенства (1.8.5) тогда следует, что $f(t) - f(0) \geqslant f'(+0)t$ или $f(t) \geqslant f(0)$ при всех $f \in [0, t_0]$. В частности, при $t = t_0 = |u - u_*|$ отсюда имеем $J(u) \geqslant J(u_*)$, что и требовалось.

В частности, если в точке u_* существует градиент $J'(u_*)$, то для $e = (u - u_*)/|u - u_*|$ $(u \in U,\ u \neq u_*)$ согласно формуле (14) имеем $J(u_*)/de = \langle J'(u_*), u - u_* \rangle / |u - u_*|$, и в этом случае условие (15) превращается в условие (5). Таким образом, теорема 12 является обобщением теоремы 3 на существенно более широкий класс функций. Более того, условие (15) является наиболее естественным для класса выпуклых функций. Дело в том, что, оказывается, всякая выпуклая функция в любой внутренней точке множества имеет производные по всем направлениям. Это вытекает из следующих двух теорем.

Теорема 13. *Пусть U — выпуклое множество, функция $J(u)$ определена на U. Для того чтобы $J(u)$ была выпуклой на U, необходимо и достаточно, чтобы для любой точки $u \in U$ и любого возможного направления e в точке u функция $f(t) = J(u + te)$ одной переменной t была выпукла на отрезке $[a, b]$, где $a = \inf\{t\colon u + te \in U\}$, $b = \sup\{t\colon u + te \in U\}$ (ясно, что $a \leqslant 0 < b$; если $u + ae \notin U$ или $u + be \notin U$, то функцию $g(t)$ не следует рассматривать соответственно при $t = a$ или $t = b$).*

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U. Возьмем произвольную точку $u \in U$, какое-либо возможное направление e в этой точке и составим функцию $f(t) = J(u + te)$ $(a \leqslant t \leqslant b)$. Пусть t_1, t_2 — произвольные точки из $[a, b]$ и $\alpha \in [0, 1]$. Тогда $f(\alpha t_1 + (1 - \alpha)t_2) = J(\alpha(u + t_1 e) + (1 - \alpha)(u + t_2 e)) \leqslant \alpha J(u + t_1 e) + (1 - \alpha)J(u + t_2 e) = \alpha f(t_1) + (1 - \alpha)f(t_2)$, что и требовалось.

Достаточность. Пусть для всех $u \in U$ и всех возможных направлений e в точке u функция $f(t) = J(u + te)$ выпукла на соответствующем отрезке $[a, b]$. Возьмем любые точки u, $v \in U$ и положим $e = v - u$ — это возможное направление в точке u, так как $u + t(v - u) \in U$ при $0 \leqslant t \leqslant 1$. Тогда из выпуклости $f(t) = J(u + te)$ получим $J(\alpha v + (1 - \alpha)u) = f(\alpha) = f(\alpha \cdot 1 + (1 - \alpha) \cdot 0) \leqslant \alpha f(1) + (1 - \alpha)f(0) = \alpha J(v) + (1 - \alpha)J(u)$ при всех $\alpha \in [0, 1]$.

Теорема 14. *Пусть U — выпуклое множество, функция $J(u)$ выпукла на U. Тогда в любой точке $u \in \operatorname{ri} U$ функция $J(u)$ имеет производные по всем направлениям $e \in \operatorname{Lin} U$. В частности, если $\operatorname{int} U \neq \varnothing$, то в точке $u \in \operatorname{int} U$ существуют производные функции $J(u)$ по всем направлениям $e \in E^n$, $|e| = 1$.*

Доказательство. Зафиксируем какое-либо направление $e \in \operatorname{Lin} U$ $(|e| = 1)$ и точку $u \in \operatorname{ri} U$. Согласно определению 1.10 существует ε-окрестность $O(u, \varepsilon) = \{v \in E^n\colon |v - u| < \varepsilon\}$ точки u такая, что пересечение $O(u, \varepsilon) \cap \operatorname{aff} U$ целиком принадлежит U. Учитывая, что $-e$ также принадлежит $\operatorname{Lin} U$, можем сказать, что $u + te \in U$ для всех t $(|t| \leqslant t_0$, $0 < t_0 < \varepsilon)$. Это значит, что функция $f(t) = J(u + te)$ определена на отрезке $[-t_0, t_0]$ и согласно теореме 13 она выпукла на этом отрезке. Поскольку $t = 0$ — внутренняя точка отрезка $[-t_0, t_0]$, то по теореме 1.8.2 существует

$$f'(+0) = \lim_{t \to +0} \frac{f(t) - f(0)}{t} = \lim_{t \to +0} \frac{J(u + te) - J(u)}{t} = \frac{dJ(u)}{de}.$$

Если $u \in U \setminus \operatorname{ri} U$, то в такой точке у выпуклой функции производные по возможным направлениям могут и не существовать — об этом свидетельствует пример 1.8.2.

9. Приведенный выше пример 6 показывает, что существование производных по всем направлениям не гарантирует непрерывности функции. Но для выпуклых функций такая ситуация, оказывается, невозможна.

Теорема 15. *Пусть множество U выпукло и $\operatorname{int} U \neq \varnothing$. Тогда выпуклая функция $J(u)$ во всех внутренних точках множества U непрерывна. В частности, функция, выпуклая на всем пространстве E^n, непрерывна во всех точках.*

Доказательство. Возьмем произвольные $u \in \operatorname{int} U$ и $\varepsilon > 0$. По определению внутренней точки существует число $\delta > 0$ такое, что $u + h \in U$, $u + nh^i e_i \in U$ для всех $h = (h^1, \ldots, h^n)$, $|h| < \delta/n$; здесь $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ $(i = 1, \ldots, n)$ — базис в E^n. Поскольку по теореме 14 функция $J(u)$ в точке u имеет производные по направлениям e_i, то она непрерывна в этой точке по направлениям e_i $(i = 1, \ldots, n)$. Поэтому можно взять число δ столь малым, чтобы $|J(u + nh^i e_i) - J(u)| < \varepsilon$ при всех h $(|h| < \delta/n,\ i = 1, \ldots, n)$. Тогда, пользуясь неравенством (2), получаем

$$J(u+h) - J(u) = J\left(\frac{1}{n}\sum_{i=1}^{n}\left(u + nh^i e_i\right)\right) - J(u) \leqslant$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}\left(J\left(u + nh^i e_i\right) - J(u)\right) < \varepsilon \qquad (16)$$

для всех h $(|h| < \delta/n)$. В частности, для $-h$, удовлетворяющих неравенству $|-h| < \delta/n$, из (16) следует $J(u-h) - J(u) < \varepsilon$. Но в силу выпуклости $J(u)$ имеем $J(u) = J((u+h)/2 + (u-h)/2) \leqslant (J(u+h) + J(u-h))/2$, поэтому $J(u) - J(u+h) \leqslant J(u-h) - J(u) < \varepsilon$. Отсюда и из (16) следует $|J(u+h) - J(u)| < \varepsilon$ при всех h $(|h| < \delta/n)$.

Заметим, что если $\operatorname{int} U = \varnothing$, то, рассматривая лишь точки из $\operatorname{aff} U$, аналогично можно доказать непрерывность выпуклой на U функции во всех точках $u \in \operatorname{ri} U$. В качестве базиса $\{e_i\}$, участвующего в доказательстве, в этом случае нужно взять базис подпространства $\operatorname{Lin} U$. В точках $u \in U \setminus \operatorname{ri} U$ выпуклая функция может терпеть разрыв — об этом говорит пример 1.8.1.

10. Рассмотрим выпуклые функции на выпуклом множестве U, принадлежащие классу $C^{1,1}(U)$ (см. определение 2.3.3), т. е. гладкие выпуклые функции, градиент которых удовлетворяет условию

$$|J'(u) - J'(v)| \leqslant L(u - v) \qquad \forall u, v \in U, \quad L = \text{const} \geqslant 0. \qquad (17)$$

Для таких функций имеют место неравенства

$$0 \leqslant \langle J'(u) - J'(v), u - v\rangle \leqslant L|u - v|^2 \qquad \forall u, v \in U. \qquad (18)$$

В самом деле, левое неравенство следует из теоремы 4, а правое — из условия (17). Оказывается, эти два неравенства можно записать в виде одного равносильного неравенства [29], полностью характеризующего класс выпуклых функций из $C^{1,1}(U)$ с данной постоянной $L \geqslant 0$.

Теорема 16. *Пусть U — выпуклое множество из E^n. Для того чтобы функция $J(u)$ из класса $C^1(U)$ была выпуклой и удовлетворяла условию* (17) *с постоянной L, необходимо и достаточно, чтобы*

$$|J'(u) - J'(v)|^2 \leqslant L\langle J'(u) - J'(v), u - v\rangle \qquad \forall u, v \in U. \qquad (19)$$

Из (19) *следует неравенство*

$$\langle J'(u) - J'(v), v - w\rangle \leqslant \frac{1}{4}L|u - w|^2 \qquad \forall u, v, w \in U. \qquad (20)$$

Доказательство. Достаточность. Если выполняется неравенство (19), то из него, во-первых, следует, что $\langle J'(u) - J'(v), u - v\rangle \geqslant 0$ $(u, v \in U)$, и выпуклость $J(u)$ гарантируется теоремой 4, и, во-вторых, применяя к правой части (19) неравенство Коши — Буняковского и деля на

$|J'(u)-J'(v)|$, получаем условие (17). Из выпуклости $J(u)$ и условия (17) имеем неравенства (18). Таким образом, из (19) следует (18).

Кроме того, из (19) имеем

$$\langle J'(u)-J'(v),\ v-w\rangle = \langle J'(u)-J'(v),\ u-w\rangle - \langle J'(u)-J'(v),\ u-v\rangle \leqslant$$
$$\leqslant \langle J'(u)-J'(v),\ u-w\rangle - \frac{1}{L}\,|J'(u)-J'(v)|^2 = -\Big|L^{-1/2}(J'(u)-J'(v)) -$$
$$-\frac{1}{2}L^{1/2}(u-w)\Big|^2 + \frac{1}{4}L\,|u-w|^2 \leqslant \frac{1}{4}L\,|u-w|^2 \qquad \forall u,\ v,\ w\in U.$$

Неравенство (20) установлено.

Необходимость. Пусть функция $J(u)$ выпукла и удовлетворяет условию (17). Тогда, как было показано выше, справедливы неравенства (18). Остается из (18) получить (19). Сначала рассмотрим случай, когда $\operatorname{int} U=\varnothing$ и $J(u)\in C^2(U)$. Тогда

$$0\leqslant \langle J''(u)\,\xi,\ \xi\rangle \leqslant L\,|\xi|^2 \qquad \forall u\in U \tag{21}$$

при всех $\xi\in E^n$. В самом деле, из неравенств (18) с помощью формулы (2.3.4) в случае $u\in\operatorname{int} U$ имеем

$$0\leqslant \langle J'(u+\varepsilon\xi)-J'(u),\ \varepsilon\xi\rangle = \varepsilon^2\langle J''(u+\theta\varepsilon\xi)\,\xi,\ \xi\rangle \leqslant L|\xi|^2\varepsilon^2$$

или $0\leqslant \langle J''(u+\theta\varepsilon\xi)\,\xi,\ \xi\rangle \leqslant L|\xi|^2$ $(0\leqslant\theta\leqslant 1)$ для всех ε $(|\varepsilon|\leqslant\varepsilon_0$, $\varepsilon_0>0)$. Отсюда при $\varepsilon\to+0$ получим (21) для точек $u\in\operatorname{int} U$. Если $u\in\operatorname{Гр} U$, то оценка (21) доказывается с помощью предельного перехода от внутренних точек так же, как это делалось при доказательстве теоремы 5.

Далее, пользуясь формулой (2.3.5), имеем

$$J'(u+h)-J'(u) = Ah,\quad A=\int_0^1 J''(u+th)\,dt,\quad h=v-u. \tag{22}$$

Разумеется, матрица A зависит от u, v, но эту зависимость мы для краткости не будем явно указывать. Согласно (21) $0\leqslant\langle J''(u+th)\xi,\ \xi\rangle\leqslant L|\xi|^2$ $(0\leqslant t\leqslant 1)$, откуда, интегрируя по t, получаем

$$0\leqslant\langle A\xi,\ \xi\rangle\leqslant L|\xi|^2,\qquad \xi\in E^n. \tag{23}$$

Таким образом, симметричная матрица A неотрицательно определена. Тогда существует симметричная неотрицательно определенная матрица $A^{1/2}$ такая, что $(A^{1/2})^2=A$ [93, 164]. Пользуясь оценкой (23) при $\xi=A^{1/2}h$, с помощью формул (22) имеем

$$|J'(u)-J'(v)|^2 = \langle Ah,\ Ah\rangle = \langle AA^{1/2}h,\ A^{1/2}h\rangle \leqslant L|A^{1/2}h|^2 =$$
$$= L\langle Ah,\ h\rangle = L\langle J'(u)-J'(v),\ u-v\rangle.$$

Неравенство (19) доказано при дополнительных предположениях $\operatorname{int} U\neq\varnothing$ и $J(u)\in C^2(U)$.

Наметим схему доказательства для случая, когда $\operatorname{int} U\neq\varnothing$, но $J(u)\in C^1(U)$. Построим последовательность функций $\{J_k(u)\}$ $(u\in U)$ и последовательность $\{U_k\}$ строго внутренних и выпуклых подмножеств множества U таких, что $U=\bigcup\limits_{k\geqslant 1}U_k$, $U_k\subset U_{k+1}$ $(k=1,\ 2,\ \ldots)$. Для всех $k\geqslant 1$ и всех $m\geqslant k$ функция $J_m(u)$ выпукла на U_k, $J_m(u)\in C^2(U_k)$, $\Big|J'_m(u)-J'_m(v)\Big|\leqslant L\,|u-v|$ для любых $u,\ v\in U_k$, $\lim\limits_{m\to\infty}J_m(u)=J(u)$, $\lim\limits_{m\to\infty}J'_m(u)=J'(u)$ при всех $u\in U_k$. В силу доказанного тогда $\Big|J'_m(u)-J'_m(v)\Big|^2\leqslant L\Big\langle J'_m(u)-J'_m(v),\ u-v\Big\rangle$ при всех $u,\ v\in U_k$ и всех $m\geqslant k$. Отсюда при

$m \to \infty$ получим неравенство (19) на множестве U_k. Далее, при $k \to \infty$ убеждаемся в справедливости (19) для всех $u, v \in \operatorname{int} U$. Наконец, для граничных точек множества U неравенство (19) доказывается с помощью предельного перехода от внутренних точек.

Как построить упомянутые последовательности $\{U_k\}$ и $\{J_k(u)\}$? В качестве $\{U_k\}$ может быть взята последовательность подмножеств всех внутренних точек множества U, удаленных от границы U на расстояние не менее чем $3\delta_k$, где $\lim \delta_k = 0$, $\delta_k > \delta_{k+1} > 0$ $(k = 1, 2, \ldots)$. В качестве $J_k(u)$ могут быть взяты средние функции Стеклова — Соболева [233], например,

$$J_k(u) = \int\limits_{E^n} J(w)\,\omega_k(w-u)\,dw, \quad \omega_k(u) = \delta_k^{-n}\varkappa^{-1}\omega\left(|u|\,\delta_k^{-1}\right),$$

где $\omega(r) = \exp\{-1/(1-r^2)\}$ при $|r| < 1$, $\omega(r) = 0$ при $|r| \geqslant 1$, $\varkappa = \int\limits_{-1}^{1} \omega(r)\,dr$ и функция $J(u)$ вне U доопределена тождественным нулем. Если $\operatorname{int} U = \varnothing$, то аналогичные построения надо провести в $\operatorname{ri} U$.

11. Остановимся на одном замечательном свойстве выпуклых множеств, задаваемых ограничениями $g(u) \leqslant c$, где $g(u)$ — выпуклая функция.

Теорема 17. *Пусть U_0 — непустое выпуклое замкнутое множество из E^n, функция $g(u)$ выпукла и полунепрерывна снизу на U_0, пусть*

$$M(c) = \{u\colon u \in U_0,\ g(u) \leqslant c\}.$$

Тогда для ограниченности множества $M(c)$ при каждом c необходимо и достаточно, чтобы при некотором a множество $M(a)$ было непустым и ограниченным.

Доказательство. Необходимость. Она очевидна.

Достаточность. Пусть $M(a) \neq \varnothing$ и это множество ограничено. Поскольку $M(c) \subset M(a)$ при всех $c \leqslant a$, то $M(c)$ ограничено при всех $c \leqslant a$ (пустое множество ограничено по определению). Остается рассмотреть случай $c > a$. Предположим, что при некотором $c > a$ множество $M(c)$ не ограничено. Заметим, что $M(c)$ выпукло и замкнуто, что следует из леммы 2.1.1 и теоремы 10.

Покажем, что существует вектор $e \neq 0$ такой, что $u + te \in M(c)$ при всех $t \geqslant 0$ и всех $u \in M(c)$ (такое направление, задаваемое вектором e, принято называть *рецессивным направлением* неограниченного выпуклого множества).

Поскольку множество $M(c)$ не ограничено по предположению, то существует последовательность $\{u_k\} \in M(c)$ такая, что $|u_k| \to \infty$ при $k \to \infty$. Возьмем какую-либо точку $\bar{u} \in M(c)$ и построим вектор $e_k = (u_k - \bar{u}) \times |u_k - \bar{u}|^{-1}$ $(k = 1, 2, \ldots)$. По теореме Больцано — Вейерштрасса из последовательности $\{e_k\}$ можно выбрать подпоследовательность $\{e_{k_m}\}$, сходящуюся к некоторому вектору e $(|e| = 1)$.

Возьмем произвольное $t < 0$. Поскольку $0 < t/|u_k - \bar{u}| < 1$ при всех $k \geqslant k_0$, то в силу выпуклости $M(c)$ имеем

$$\bar{u} + te_k = \frac{t}{|u_k - \bar{u}|}\,u_k + \left(1 - \frac{t}{|u_k - \bar{u}|}\right)\bar{u} \in M(c), \qquad k \geqslant k_0.$$

Отсюда при $k_m \to \infty$ получим $\bar{u} + te \in M(c)$, так как $M(c)$ замкнуто. В силу произвольности $t > 0$ заключаем, что $\bar{u} + te \subset M(c)$ при всех $t \geqslant 0$.

Теперь возьмем любую точку $u \in M(c)$ и покажем, что $u + te \in M(c)$ при каждом $t \geqslant 0$. По доказанному $\bar{u} + \mu e \in M(c)$ $(\mu \geqslant 0)$. В силу выпуклости $M(c)$ тогда

$$\frac{t}{\mu}(\bar{u} + \mu e) + \left(1 - \frac{t}{\mu}\right)u = u + te + \frac{t(\bar{u} - u)}{\mu} \in M(c)$$

при всех $\mu > t$. Отсюда при $\mu \to \infty$ с учетом замкнутости $M(c)$ получим $u + te \in M(c)$ при каждом $t \geqslant 0$.

Зафиксируем какую-либо точку $v \in M(a) \subset M(c)$. В силу построения вектора e тогда $v + te \in M(c)$ $(t \geqslant 0)$. По условию множество $M(a)$ ограничено, поэтому луч $\{v + te,\ t \geqslant 0\}$ пересекает границу выпуклого множества $M(a)$ в некоторой точке, или, точнее говоря, найдется число $t_0 = \sup\{t\colon\ v + te \in M(a)\}$ такое, что $v + te \notin M(a)$ при всех $t > t_0$. Это значит, что $v + te \in U_0$, но $c \geqslant g(v + te) > a \geqslant g(v)$ для всех $t > t_0$.

Зафиксируем какое-либо $t > t_0$. Тогда, пользуясь представлением

$$v + te = \lambda\left(v + \frac{t}{\lambda}e\right) + (1 - \lambda)v, \qquad 0 < \lambda < 1,$$

и выпуклостью функции $g(u)$, имеем $g(v + te) \leqslant \lambda g(v + (t/\lambda)e) + (1 - \lambda)g(v)$, или $g(v + (t/\lambda)e) \geqslant (g(v + te) - g(v))/\lambda + g(v)$ $(0 < \lambda < 1)$. Поскольку $g(v + te) > g(v)$, то при $\lambda \to +0$ отсюда получим $g(v + (t/\lambda)e) \to \infty$. Тогда найдется число $\lambda_0 > 0$ такое, что $g(v + (t/\lambda)e) > c$ при всех λ $(0 < \lambda < \lambda_0)$. С другой стороны, по построению вектора e имеем $v + (t/\lambda)e \in M(c)$ или $g(v + (t/\lambda)e) \leqslant c$ при всех λ $(0 < \lambda < 1)$. Полученное противоречие доказывает теорему.

Отметим, что требование полунепрерывности снизу функции в теореме 17 существенно (см. ниже упражнение 21).

12. Наконец, получим оценку снизу скорости роста выпуклой функции для случая, когда множество точек ее минимума ограничено.

Теорема 18. *Пусть $J(u)$ — выпуклая функция на $U = E^n$, $J_* = \inf\limits_{E^n} J(u) > -\infty$, $U_* = \{u \in E^n\colon\ J(u) = J_*\} \neq \varnothing$, причем U_* — ограниченное множество, т. е. существует такое число $R > 0$, что $U_* \subset S_* = \{u \in E^n\colon\ |u - u_*| < R\}$, где u_* — какая-либо фиксированная точка из U_*. Тогда $\lim\limits_{|u| \to \infty} J(u) = \infty$ и, более того, верна оценка*

$$J(u) \geqslant |u - u_*|\,\frac{J_{*R} - J_*}{R} + J_* \qquad \forall u \notin S_*, \tag{24}$$

где $J_{*R} = \inf\limits_{u \in \operatorname{Гр} S_*} J(u) > J_*$, $\operatorname{Гр} S_* = \{u \in E^n\colon\ |u - u_*| = R\}$.

Доказательство. Возьмем любую точку $u \notin S_*$. Поскольку

$$v = u_* + R\,\frac{u - u_*}{|u - u_*|} = \frac{R}{|u - u_*|}\,u + \left(1 - \frac{R}{|u - u_*|}\right)u_* \in \operatorname{Гр} S_*,$$

то с учетом выпуклости $J(u)$ имеем

$$J_{*R} \leqslant J(v) \leqslant \frac{R}{|u - u_*|}\,J(u) + \left(1 - \frac{R}{|u - u_*|}\right)J(u_*).$$

Отсюда сразу получаем требуемое неравенство (24). Согласно теореме 15 функция $J(u)$ непрерывна на E^n, и в силу теоремы 2.1.1 на замкнутом ограниченном множестве $\operatorname{Гр} S_*$ она достигает своей нижней грани хотя бы в одной точке. Отсюда и из того, что замкнутые множества U_* и $\operatorname{Гр} S_*$ не пересекаются, следует, что $J_{*R} > J_*$. Тогда из оценки (24) имеем, что $\lim\limits_{|u| \to \infty} J(u) = \infty$.

Отметим, что для функции $J(u) = |u|$ $(u \in E^n)$ неравенство (24) превращается в тождественное равенство. Это значит, что оценка (24) на классе выпуклых функций является точной.

Некоторые другие свойства выпуклых функций и множеств будут рассмотрены ниже.

Упражнения. 1. При каких a, b, c функция $J(u) = ax^2 + 2bxy + cy^2$ переменных $u = (x, y) \in E^2$ будет выпуклой на E^2? Вогнутой на E^2?

2. Найти области выпуклости и вогнутости функций $J(u) = \sin(x + y + z)$, $J(u) = \sin(x^2 + y^2 + z^2)$.

3. При каких p, q функция $J(u) = x^p y^q$ будет выпуклой (или вогнутой) на множестве $U = \{u = (x, y) \in E^2\colon x > 0,\ y > 0\}$? Аналогичное исследование провести для функции $J(u) = x^p y^q z^r$ на $U = \{u = (x, y, z)\colon x > 0,\ y > 0,\ z > 0\}$.

4. Если функция $J(u)$ выпукла, то будет ли выпуклой функция $|J(u)|$?

5. Если функция $J(v)$ выпукла на E^m, а A — матрица размера $m \times n$, то функция $g(u) = J(Au)$ выпукла на E^n. Доказать.

6. Если $J_1(u)$, $J_2(u)$ выпуклы, то будет ли их произведение выпуклой функцией? Рассмотреть пример $J_1(u) = u$, $J_2(u) = u^2$. Что изменится, если от функций $J_1(u)$, $J_2(u)$ потребовать неотрицательности? Или монотонности?

7. Пусть функции $J_k(u)$ выпуклы на выпуклом множестве U при всех $k = 0, 1, \ldots$ и пусть существует предел $\lim\limits_{k\to\infty} J_k(u) = J(u)$ или сходится ряд $\sum\limits_{k=0}^{\infty} J_k(u) = J(u)$ при всех $u \in U$. Доказать, что функция $J(u)$ выпукла на U.

8. Выяснить, когда в неравенстве (2) возможно равенство, если $J(u)$ — строго выпуклая функция.

9. Доказать неравенства:

а) $\sqrt[m]{x_1 \ldots x_m} \leqslant \frac{1}{m}(x_1 + \ldots + x_m)$, $x_1 \geqslant 0, \ldots, x_m \geqslant 0$;

б) $(x_1 + \ldots + x_m)^n \leqslant m^{n-1}(x_1^n + \ldots + x_m^n)$, $x_1 \geqslant 0, \ldots, x_m \geqslant 0$, $n \geqslant 1$;

в) $(x_1 + \ldots + x_m)(x_1^{-1} + \ldots + x_m^{-1}) \geqslant m^2$, $x_1 > 0, \ldots, x_m > 0$.

Указание: воспользоваться неравенством (2) для выпуклых функций $J(u) = -\ln u$ $(u > 0)$; $J(u) = u^n$ $(u \geqslant 0,\ n \geqslant 1)$; $J(u) = u^{-1}$ $(u > 0)$. Выяснить, при каких условиях в этих неравенствах возможно равенство.

10. Пусть функция $J(u)$ $(u \in E^n)$ такова, что $J(\alpha u + (1-\alpha)v) \leqslant \alpha J(u) + (1-\alpha)J(v)$ при всех $u, v \in E^n$, $\alpha \in \mathbf{R}$. Проверить, что аффинная функция $J(u) = \langle a, u\rangle + b$ $(a \in E^n,\ b \in E^1)$ удовлетворяет этому неравенству. Существуют ли другие функции, обладающие этим свойством?

11. Для того чтобы функция $J(u)$ была строго выпуклой на выпуклом множестве U, необходимо и достаточно выполнения неравенства (4) (а в случае $J(u) \in C^1(v)$ — неравенства (6)), которое может обратиться в равенство лишь при $u = v$. Доказать.

12. Доказать, что если U — выпуклое множество, $J(u) \in C^2(U)$ и неравенство (8) является строгим при всех $\xi \in \operatorname{Lin} U$ $(\xi \neq 0)$, то функция $J(u)$ строго выпукла на U. Верно ли обратное утверждение? Рассмотреть пример $J(u) = u^4$ $(u \in E^1)$.

13. Пусть U — выпуклое множество, функция $J(u)$ выпукла на U, $J(u) \in C^1(U)$. Доказать, что тогда критерий оптимальности (5) равносилен неравенству $\langle J'(u), u - u_*\rangle \geqslant 0$ при всех $u \in U$.

14. Пусть $J(u)$ — выпуклая функция на выпуклом множестве U, $J(u) \in C^1(U)$ и $J_* = \inf\limits_U J(u) > -\infty$. Доказать, что для того чтобы некоторая последовательность $\{u_k\} \in U$ была минимизирующей, т. е. $\lim\limits_{k\to\infty} J(u_k) = J_*$, необходимо и достаточно выполнения условия $\varliminf\limits_{k\to\infty} \langle J'(u_k), u - u_k\rangle \geqslant 0$ при всех $u \in U$ (ср. с теоремой 3).

15. Пусть строго выпуклая функция $J(u)$ достигает на E^n своей нижней грани. Доказать, что тогда $\lim\limits_{|u|\to\infty} J(u) = \infty$.

Указание: воспользоваться теоремами 1, 18.

16. Пусть функция $J(u)$ выпукла и полунепрерывна снизу на выпуклом замкнутом множестве U из E^n, $J_* > -\infty$, $U_* \neq \varnothing$, причем $U_* \subset S_* = \{u \in U\colon |u - u_*| < R\}$, где u_* — какая-либо фиксированная точка из U_*. Тогда

$$J(u) \geqslant |u - u_*| \frac{J_{*R} - J_*}{R} + J_* \qquad \forall u \in U, \quad u \notin S_*,$$

где $J_{*R} = \inf\limits_{\operatorname{Гр} S_*} J(u) > J_*$, $\operatorname{Гр} S_* = \{u \in U\colon |u - u_*| = R\}$. Доказать, пользуясь схемой доказательства теоремы 18.

17. Выпуклая функция, отличная от постоянной, может достигать своей верхней грани на выпуклом множестве лишь в его граничных точках. Доказать.

18. Для того чтобы функция $\rho(u, U) = \inf\limits_{v \in U} |u - v|$ была выпуклой на E^n, необходимо и достаточно, чтобы замыкание множества U было выпуклым. Доказать.

19. Пусть U — ограниченное множество из E^n. Доказать, что функция $\delta(c, U) = \sup\limits_{u \in U} \langle c, u\rangle$ переменной $c \in E^n$, называемая *опорной функцией* множества U, выпукла на E^n.

20. Пусть U — выпуклое множество из E^n, $0 \in \operatorname{int} U$. Доказать, что функция $\mu(u, U) = \inf\limits_{\alpha \in A_u} \alpha$, $A_u = \{\alpha\colon \alpha > 0,\ u/\alpha \in U\}$, называемая *функцией Минковского*, выпукла на E^n.

21. Пусть $U_0 = \{u = (x, y)\colon x \geqslant 0,\ y \geqslant 0\} = E^2_+$. Показать, что функция

$$g(u) = \begin{cases} y, & x \geqslant 0, \quad y > 0 \quad \text{или} \quad 0 \leqslant x \leqslant 1, \quad y = 0, \\ x - 1, & x > 1, \quad y = 0, \end{cases}$$

выпукла и полунепрерывна сверху, но не является полунепрерывной снизу на U_0. Убедиться, что множество $M(c) = \{u \in U_0\colon g(u) \leqslant c\}$ ограничено при $c = 0$ и не ограничено при всех $c > 0$ (ср. с теоремой 17). Показать, что $M(c)$ не замкнуто при каждом $c > 0$.

22. Множество $U_c = \{u \in E^n\colon g_i(u) \leqslant c,\ i = 1, \ldots, m\}$, где $g_i(u)$ — выпуклая функция на E^n, будет ограничено при любых c тогда и только тогда, когда U_c ограничено хотя бы при одном значении $c = c_0$. Доказать.

23. Пусть U — неограниченное замкнутое множество из E^n. Доказать, что:

1) для любой точки $v \in U$ существует ненулевой вектор e такой, что луч $\{u = v + te,\ t \geqslant 0\} \in U$;

2) если луч $\{u = v + te,\ t \geqslant 0\} \in U$ при некотором $v \in U$, то луч $\{u = w + te,\ t \geqslant 0\} \in U$ при всех $w \in U$. Показать, что требование замкнутости U существенно для обоих утверждений, рассмотрев множество $U = \{u = (x, y)\colon 0 < x < 1\} \cup \{(0, 0)\}$.

Указание: воспользоваться рассуждениями из доказательства теоремы 17.

24. Доказать, что функция

$$J(u) = \begin{cases} x^2/y, & y \neq 0, \\ 0, & y = 0, \end{cases}$$

выпукла на множестве $U = \{u = (x, y)\colon y > 0\} \cup \{(0, 0)\}$ и полунепрерывна снизу на U. Убедиться, что $J(u)$ не является полунепрерывной сверху в точке $u_0 = (0, 0)$, и, более того, показать, что для любого числа $A \geqslant 0$

существует такая последовательность $\{u_k\} \in U$, $\{u_k\} \to 0$, что $\lim\limits_{k\to\infty} J(u_k) = A$.

25. Пусть $U = \{u \in E^n\colon Au \leqslant b\}$ — многогранное множество, функция $J(u)$ выпукла на U. Доказать, что $J(u)$ полунепрерывна сверху на U.

26. Пусть функция $J(u)$ выпукла и ограничена сверху на $E^n_+ = \{u = (u^1, \ldots, u^n) \in E^n\colon u^1 \geqslant 0, \ldots, u^n \geqslant 0\}$. Доказать, что $J(u)$ монотонна и не возрастает на E^n_+ по каждой переменной.

27. Доказать, что если выпуклая функция $J(u)$ на E^n ограничена сверху, то $J(u)$ постоянна.

28. Пусть $J(u)$ — выпуклая дифференцируемая функция на открытом выпуклом множестве W из E^n. Доказать, что тогда градиент $J'(u) = (\partial J(u)/\partial u^1, \ldots, \partial J(u)/\partial u^n)$ непрерывен на W.

29. Пусть $J(u)$ — выпуклая функция на выпуклом множестве U из E^n. Доказать, что $J(u)$ удовлетворяет условию Липшица на каждом ограниченном множестве V, замыкание которого принадлежит ri U.

30. Пусть $J(u)$ — выпуклая функция на открытом выпуклом множестве W из E^n. Доказать, что $J(u)$ почти всюду на W дифференцируема.

31. Пусть U_0 — выпуклое замкнутое множество из E^n, $g(u)$ — выпуклая функция на U_0, $U = \{u \in U_0\colon g(u) \leqslant 0\}$. Пусть $\{u_k\} \in U_0$, $\{g(u_k)\} \to 0$. Можно ли утверждать, что $\{\rho(u_k, U)\} \to 0$? Рассмотреть пример $U_0 = \{u = (x, y) \in E^2\colon x \geqslant 1\}$, $g(u) = y^2/x$, $u_k = \left(k, \sqrt[4]{k}\right)$ $(k = 1, 2, \ldots)$.

32. Пусть U — выпуклое замкнутое множество из E^n, $J(u)$ — выпуклая непрерывная функция на U, $J_* > -\infty$, $U_* \neq \varnothing$. Пусть $\{u_k\} \in U$, $\{J(u_k)\} \to J_*$. Можно ли ожидать, что $\{\rho(u_k, U_*)\} \to 0$? Рассмотреть пример $U = \{u = (x, y) \in E^2\colon x \geqslant 1\}$, $J(u) = y^2/x$, $u_k = \left(k, \sqrt[4]{k}\right)$ $(k = 1, 2, \ldots)$.

33. Функция $J(u)$, определенная на выпуклом множестве U, называется *квазивыпуклой* на U, если $J(\alpha u + (1-\alpha)v) \leqslant \max\{J(u); J(v)\}$ при всех $u, v \in U$, $\alpha \in [0, 1]$. Доказать, что $J(u)$ квазивыпукла на U тогда и только тогда, когда множество $M(v) = \{u \in U\colon J(u) \leqslant J(v)\}$ выпукло при всех $v \in U$ (ср. с теоремой 10).

§ 3. Сильно выпуклые функции

1. Непрерывная выпуклая функция на выпуклом замкнутом множестве может не достигать своей нижней грани на этом множестве. Например, если $J(u) = 1/u$, $U = \{u \in E^1\colon u \geqslant 1\}$, то $J_* = \inf\limits_U J(u) = 0$, но $J(u) > 0$ при всех $u \in U$. Однако можно выделить подкласс выпуклых функций, для которых подобная ситуация невозможна.

Определение 1. Функция $J(u)$, определенная на выпуклом множестве U, называется *сильно выпуклой* на U, если существует постоянная $\varkappa > 0$ такая, что

$$J(\alpha u + (1-\alpha)v) \leqslant \alpha J(u) + (1-\alpha)J(v) - \alpha(1-\alpha)\varkappa|u-v|^2 \quad (1)$$

при всех $u, v \in U$ и всех α $(0 \leqslant \alpha \leqslant 1)$. Постоянную $\varkappa$ называют *постоянной сильной выпуклости* функции $J(u)$ на множестве U.

Очевидно, сильно выпуклая на U функция будет выпуклой и даже строго выпуклой на U. Примером сильно выпуклой функ-

ции на всем пространстве E^n может служить функция

$$\Omega(u) = \langle u, u\rangle = |u|^2, \qquad u \in E^n.$$

Для этой функции неравенство (1) превращается в тождественное равенство с постоянной $\varkappa = 1$:

$$|\alpha u + (1-\alpha)v|^2 = \alpha|u|^2 + (1-\alpha)|v|^2 - \alpha(1-\alpha)|u-v|^2 \qquad (2)$$

при всех u, $v \in E^n$, $\alpha \in [0, 1]$. Линейная функция $J(u) = \langle c, u\rangle$ выпукла на E^n, но не сильно выпукла. Упомянутая выше функция $J(u) = 1/u$ при $u \geqslant 1$ выпукла, но не сильно выпукла при $u \geqslant 1$.

Нетрудно видеть, что сумма двух сильно выпуклых функций на выпуклом множестве U будет сильно выпуклой функцией на U с той же постоянной $\varkappa$. Если $J(u)$ сильно выпукла на U с постоянной $\varkappa$, то $g(u) = cJ(u)$ при любом $c = \text{const} > 0$ будет сильно выпуклой на U с постоянной $c\varkappa$.

Теорема 1. *Пусть множество U выпукло и замкнуто, а функция $J(u)$ сильно выпукла и полунепрерывна снизу на U. Тогда*:

1) *множество Лебега*

$$M(v) = \{u\colon\ u \in U,\ J(u) \leqslant J(v)\}$$

выпукло, замкнуто и ограничено при всех $v \in U$;

2) $J_* = \inf\limits_U J(u) > -\infty$, *множество* $U_* = \{u\colon\ u \in U,\ J(u) = J_*\}$ *непусто и, более того, состоит из единственной точки u_**;

3) *имеет место неравенство*

$$\varkappa|u - u_*|^2 \leqslant J(u) - J(u_*) \qquad \forall u \in U; \qquad (3)$$

4) *любая минимизирующая последовательность $\{u_k\}$: $u_k \in U$, $k = 1, 2, \ldots$, $\lim\limits_{k\to\infty} J(u_k) = J_*$, сходится к точке u_*.*

Сформулированная теорема является обобщением теоремы Вейерштрасса 2.1.1. В отличие от теоремы 2.1.1, здесь на функцию накладывается более жесткое ограничение, но зато от множества U не требуется ограниченности. В частности, в теореме 1 возможно $U = E^n$.

Доказательство. Если множество U ограничено, замкнуто, т. е. U компактно, то все утверждения теоремы 1, кроме неравенства (3), следуют из теорем 2.1.1, 2.10. Поэтому пусть U — неограниченное множество. Возьмем произвольную точку $v \in U$ и рассмотрим шар

$$S = S(v, 1) = \{u\colon\ u \in U,\ |u - v| \leqslant 1\}.$$

Из теоремы 2.1.1 следует, что $\inf\limits_S J(u) = J_{*S} > -\infty$, так что

$$J(u) \geqslant J_{*S} = J(v) - \nu \qquad \forall u \in S, \quad \nu = J(v) - J_{*S} \geqslant 0. \qquad (4)$$

Возьмем произвольную точку $u \in U \setminus S$, т. е. $u \in U$, $|u-v| > 1$. Тогда

$$0 < \alpha_0 = 1/|u-v| < 1. \tag{5}$$

При $\alpha = \alpha_0$ из (1) получаем

$$\alpha_0 J(u) \geqslant J(v + \alpha_0(u-v)) - (1-\alpha_0) J(v) + \alpha_0(1-\alpha_0)\varkappa|u-v|^2. \tag{6}$$

В силу (5) $\alpha_0|u-v| = 1$, поэтому $v + \alpha_0(u-v) \in S$. Согласно (4) тогда $J(v+\alpha_0(u-v)) \geqslant J(v) - \nu$. Учитывая эту оценку, из (6) получаем

$$\alpha_0 J(u) \geqslant \alpha_0 J(v) - \nu + \alpha_0(1-\alpha_0)\varkappa|u-v|^2.$$

Отсюда, сокращая на $\alpha_0 > 0$ и вспоминая определение (5) величины α_0, получаем

$$J(u) \geqslant J(v) + (1-\alpha_0)\varkappa|u-v|^2 - \nu/\alpha_0 =$$
$$= J(v) + \varkappa|u-v|^2 - (|u-v|\sqrt{\varkappa})(\sqrt{\varkappa} + \nu/\sqrt{\varkappa}).$$

Применяя к последнему слагаемому неравенство $ab \leqslant (a^2+b^2)/2$, будем иметь

$$J(u) \geqslant J(v) + \varkappa|u-v|^2/2 - (\sqrt{\varkappa} + \nu/\sqrt{\varkappa})^2/2 \tag{7}$$

для всех $u \in U \setminus S$. Нетрудно видеть, что неравенство (7) справедливо и при $u \in S$. В самом деле, если $u \in S$, т. е. $|u-v| \leqslant 1$, то $\nu < (\sqrt{\varkappa} + \nu/\sqrt{\varkappa})^2/2 - \varkappa|u-v|^2/2$. Отсюда и из (4) следует справедливость (7) и для $u \in U$.

Таким образом, неравенство (7) имеет место для всех $u \in U$. Для любых $u \in M(v)$ из (7) следует

$$\varkappa|u-v|^2/2 - (\sqrt{\varkappa} + \nu/\sqrt{\varkappa})^2/2 \leqslant J(u) - J(v) \leqslant 0$$

или

$$|u-v| \leqslant 1 + \nu/\varkappa \qquad \forall u \in M(v).$$

Ограниченность $M(v)$ доказана. Выпуклость $M(v)$ следует из теоремы 2.10, а замкнутость $M(v)$ — из леммы 2.1.1. Из теоремы 2.1.2 имеем $J_* > -\infty$, $U_* \neq \varnothing$. Поскольку сильно выпуклая функция строго выпукла, то в силу теоремы 2.1 множество U_* состоит из единственной точки u_*.

Докажем неравенство (3). Учитывая, что $J(u_*) \leqslant J(u)$ при всех $u \in U$, из (1) имеем $0 \leqslant J(\alpha u + (1-\alpha)u_*) - J(u_*) \leqslant \alpha(J(u) - J(u_*)) - \alpha(1-\alpha)\varkappa|u-u_*|^2$, или $\alpha(1-\alpha)\varkappa|u-u_*|^2 \leqslant \alpha(J(u) - J(u_*))$ при всех $\alpha \in [0, 1]$ и $u \in U$. Деля на $\alpha > 0$ и устремляя $\alpha \to +0$, отсюда получаем неравенство (3).

Наконец, пусть $\{u_k\} \in U$, $\lim\limits_{k\to\infty} J(u_k) = J_*$. Полагая в (3) $u = u_k$, получаем $\varkappa|u_k - u_*|^2 \leqslant J(u_k) - J_*$ $(k = 1, 2, \ldots)$. Отсюда при $k \to \infty$ следует, что $|u_k - u_*| \to 0$.

Замечание 1. При выполнении условий теоремы 1 утверждение о $J_* > -\infty$, $U_* \neq \varnothing$ остается верным для любого замкнутого (необязательно выпуклого) подмножества $W \subseteq U$ — это следует из замкнутости и ограниченности $M(v) = \{u\colon u \in W, J(u) \leqslant J(v)\}$ и теоремы 2.1.2.

2. Укажем критерии сильной выпуклости для гладких функций, аналогичные теоремам 2.2, 2.4, 2.5. Сначала докажем одно вспомогательное утверждение.

Лемма 1. *Пусть U — выпуклое множество. Функция $J(u)$ сильно выпукла на U с постоянной сильной выпуклости $\varkappa > 0$ тогда и только тогда, когда функция $g(u) = J(u) - \varkappa|u|^2$ выпукла на U.*

Доказательство. Необходимость. Пусть функция $J(u)$ сильно выпукла на U, т. е. выполнено неравенство (1). Умножим равенство (2) на $\varkappa > 0$ и почленно вычтем получившееся равенство из (1). Будем иметь неравенство, которое с помощью функции $g(u)$ можно записать в виде

$$g(\alpha u + (1-\alpha)v) \leqslant \alpha g(u) + (1-\alpha) g(v) \qquad \forall u, v \in U, \quad \alpha \in [0, 1]. \tag{8}$$

Это значит, что $g(u)$ выпукла на U.

Достаточность. Пусть функция $g(u)$ выпукла на U, т. е. выполняется (8). Сложив (8) с равенством (2), умноженным на $\varkappa > 0$, придем к неравенству (1). Это значит, что $J(u)$ сильно выпукла на U.

Теорема 2. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$. Тогда для того чтобы функция $J(u)$ была сильно выпуклой на U, необходимо и достаточно существования такой постоянной $\varkappa > 0$, что*

$$J(u) \geqslant J(v) + \langle J'(v), u - v\rangle + \varkappa|u - v|^2 \qquad \forall u, v \in U. \tag{9}$$

Доказательство. Заметим, что для сильно выпуклой функции $\Omega(u) = |u|^2$ неравенство (9) превращается в легко проверяемое тождественное равенство с постоянной $\varkappa = 1$:

$$|u|^2 = |v|^2 + \langle 2v, u - v\rangle + |u - v|^2 \qquad \forall u, v \in U. \tag{10}$$

Теперь нетрудно доказать, что условие (9) равносильно неравенству

$$g(u) \geqslant g(v) + \langle g'(v), u - v\rangle \qquad \forall u, v \in U, \tag{11}$$

где $g(u) = J(u) - \varkappa|u|^2$, $g'(u) = J'(u) - 2\varkappa u$. В самом деле, если умножить равенство (10) на $\varkappa$ и сложить с (11), то получим неравенство (9). И обратно: вычитая из (9) равенство (10), умноженное на $\varkappa$, приходим к (11). Из равносильности неравенств (9) и (11), из леммы 1 и теоремы 2.2 следует утверждение теоремы.

Теорема 3. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$. Тогда для сильной выпуклости функции $J(u)$ на U необходимо и достаточно существования такой постоянной $\mu > 0$, что*

$$\langle J'(u) - J'(v), u - v\rangle \geqslant \mu |u - v|^2, \quad \forall u, v \in U. \tag{12}$$

Доказательство. Легко проверить, что неравенство (12) равносильно неравенству

$$\langle g'(u) - g'(v), u - v\rangle \geqslant 0 \quad \forall u, v \in U$$

для функции $g(u) = J(u) - \varkappa |u|^2$, где $\varkappa = \mu/2$. Отсюда и из леммы 1, теоремы 2.4 следует утверждение теоремы.

Теорема 4. *Пусть U — выпуклое множество из E^n, $J(u) \in C^2(U)$. Тогда для сильной выпуклости функции $J(u)$ на U необходимо и достаточно существования такой постоянной $\mu > 0$, что*

$$\langle J''(u)\xi, \xi\rangle \geqslant \mu |\xi|^2 \tag{13}$$

при всех $u \in U$ и всех $\xi = (\xi^1, \ldots, \xi^n)$, принадлежащих подпространству $L = \operatorname{Lin} U$, параллельному афинной оболочке множества U (в частности, если $\operatorname{int} U \neq \varnothing$, то (13) выполняется при всех $\xi \in E^n$).

Доказательство. Для функции $g(u) = J(u) - \varkappa |u|^2$ ее вторая производная равна $g''(u) = J''(u) - 2\varkappa I$, где I — единичная матрица. Отсюда ясно, что неравенство (13) с $\mu = 2\varkappa$ равносильно неравенству

$$\langle g''(u)\xi, \xi\rangle \geqslant 0 \quad \forall u \in U, \quad \forall \xi \in L.$$

Отсюда и из леммы 1, теоремы 2.5 следует справедливость теоремы.

Замечания. 1. Если $\operatorname{int} U \neq \varnothing$, то $L = E^n$ и условие (13) должно выполняться при всех $\xi \in E^n$. Пример 2.1 показывает, что при $\operatorname{int} U = \varnothing$ условие (13) может и не выполняться при каждом $\xi \in E^n$.

2. Для функций одной переменной неравенство (13) имеет вид $J''(u) \geqslant \mu > 0$ при всех $u \in U$. Отсюда нетрудно вывести, что функция $J(u) = u^p$ при любом $p > 1$ будет сильно выпукла на множестве $U_\varepsilon = \{u \in E^1\colon u \geqslant \varepsilon\}$ при всех $\varepsilon > 0$; если здесь $\varepsilon \leqslant 0$, то такая функция сильно выпукла лишь при $p = 2$. Функция $J(u) = \sin u$ сильно выпукла на $U_\varepsilon = [-\pi + \varepsilon, -\varepsilon]$ при всех ε $(0 < \varepsilon < \pi/2)$, но не сильно выпукла на $[-\pi, 0]$.

3. Условие (13) в теореме 4 не может быть заменено условием

$$\langle J''(u)\xi, \xi\rangle > 0 \quad \forall u \in U, \quad \forall \xi \in L, \quad \xi \neq 0. \tag{14}$$

Например, для функции $J(u) = e^u$ $(u \in U = E^1 = L)$ имеем $\langle J''(u)\xi, \xi\rangle = e^u \xi^2 > 0$ при всех $u \in U$, $\xi \in L$, $\xi \neq 0$, но эта функция не является сильно выпуклой. Однако если $\operatorname{int} U \neq \varnothing$, $J''(u) = A = \{a_{ij}\}$ — постоянная матрица, то условие (14), означающее положительную определенность квадратичной формы

$\langle A\xi, \xi\rangle = \sum_{i,j=1}^{n} a_{ij}\xi^i\xi^j$, влечет за собой условие $\langle A\xi, \xi\rangle \geqslant \mu|\xi|^2$ $(\xi \in E^n)$, где $\mu = \inf_{|\xi|=1} \langle A\xi, \xi\rangle > 0$.

Как известно [93, 164], для положительной определенности квадратичной формы $\langle A\xi, \xi\rangle$ необходимо и достаточно, чтобы все угловые миноры

$$\Delta_i = \det \begin{bmatrix} a_{11} \dots a_{1i} \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{i1} \dots a_{ii} \end{bmatrix}, \qquad i = 1, \dots, n,$$

были положительны. Пользуясь этим условием, для функции $J(u) = x^2 + 2axy + by^2 + cz^2$ из примера 2.2 находим, что $J(u)$ будет сильно выпукла на E^3 тогда и только тогда, когда $b - a^2 > 0$, $c > 0$.

Далее, для функции

$$J(u) = \langle Au, u\rangle/2 - \langle b, u\rangle, \quad u \in E^n,$$

из примера 2.3 сильная выпуклость на E^n будет тогда и только тогда, когда A — положительно определенная матрица. Аналогично для функции

$$J(u) = |Au - b|^2$$

из примера 2.4 сильная выпуклость на E^n будет тогда и только тогда, когда матрица A^TA — невырожденная.

3. Рассмотрим сильно выпуклые функции $J(u)$ из класса $C^{1,1}(U)$, т. е. гладкие сильно выпуклые функции, градиент которых удовлетворяет условию

$$|J'(u) - J'(v)| \leqslant L|u - v|, \qquad u, v \in U. \tag{15}$$

Полезно установить связь между постоянными $\varkappa$, μ, L из (1), (9), (12), (13), (15). Из условия (12) с помощью неравенства Коши — Буняковского и условия (15) имеем $\mu \leqslant L$. При доказательстве теорем 3, 4 было показано, что $\mu = 2\varkappa$. Поэтому $2\varkappa \leqslant L$.

Из определения 1 видно, что если неравенство (1) имеет место при некотором $\varkappa$, то оно будет иметь место и для всех меньших положительных значений $\varkappa$. Можно поставить вопрос об определении самой большой, точной постоянной $\varkappa$ в (1). Очевидно, такой постоянной в (1) будет

$$\varkappa_0 = \inf_{0<\alpha<1} \inf_{u,v \in U} \frac{\alpha J(u) + (1-\alpha) J(v) - J(\alpha u + (1-\alpha) v)}{\alpha(1-\alpha)|u - v|^2}.$$

Аналогично самые большие постоянные в (12), (13) соответственно имеют вид

$$\mu_0 = \inf_{u,v \in U} \frac{\langle J'(u) - J'(v), u - v\rangle}{|u - v|^2},$$

$$\mu_1 = \inf_{u \in U} \inf_{\xi \in \operatorname{Lin} U, \xi \neq 0} \frac{\langle J'(u)\xi, \xi\rangle}{|\xi|^2}.$$

Из доказательства теорем 3, 4 следует, что $\mu_0 = \mu_1 = 2\varkappa_0$, причем для функций из $C^{1,1}(U)$ все эти постоянные не больше L. Заметим, что для функции $\Omega(u) = |u|^2$ на E^n имеем $\mu_0 = \mu_1 = 2\varkappa_0 = L = 2$.

4. Продолжим рассмотрение сильно выпуклых функций $J(u) \in C^{1,1}(U)$. Для таких функций из (12) и (15) имеем неравенства

$$\mu|u-v|^2 \leqslant \langle J'(u) - J'(v),\ u-v\rangle \leqslant L|u-v|^2, \quad u, v \in U. \tag{16}$$

Оказывается, как в теореме 2.16, два неравенства (16) можно записать в виде одного равносильного (16) неравенства [29], полностью характеризующего класс сильно выпуклых функций из $C^{1,1}(U)$ с данными постоянными L, μ $(L \geqslant \mu > 0)$.

Теорема 5. *Пусть U — выпуклое множество из E^n и $J(u) \in C^1(U)$. Тогда для того чтобы функция $J(u)$ была сильно выпуклой с постоянной $\varkappa = \mu/2 > 0$ и удовлетворяла условию* (15) *с постоянной $L > 0$, необходимо и достаточно, чтобы*

$$|J'(u) - J'(v)|^2 + L\mu|u-v|^2 \leqslant (L+\mu)\langle J'(u) - J'(v),\ u-v\rangle \tag{17}$$

при всех $u, v \in U$.

Доказательство. Необходимость. Пусть $J(u) \in C^{1,1}(U)$ и эти функции сильно выпуклы на U. Тогда справедливы неравенства (16). Введем функцию

$$g(u) = J(u) - \mu|u|^2/2, \quad u \in U.$$

Имеем $g'(u) = J'(u - \mu u)$. Тогда из левого неравенства (16) следует

$$\langle g'(u) - g'(v),\ u-v\rangle = \langle J'(u) - J'(v),\ u-v\rangle - \mu\,|u-v|^2 \geqslant 0, \quad \forall u, v \in U,$$

а из правого неравенства (16) получим

$$\langle g'(u) - g'(v),\ u-v\rangle \leqslant (L-\mu)\,|u-v|^2 \quad \forall u, v \in U.$$

Объединяя оба полученных неравенства, имеем

$$0 \leqslant \langle g'(u) - g'(v),\ u-v\rangle \leqslant (L-\mu)\,|u-v|^2 \quad \forall u, v \in U.$$

Таким образом, функция $g(u)$ удовлетворяет неравенствам вида (2.18). Согласно теореме 2.16 эти два неравенства равносильны одному неравенству

$$|g'(u) - g'(v)|^2 \leqslant (L-\mu)\langle g'(u) - g'(v),\ u-v\rangle \quad \forall u, v \in U.$$

Подставляя сюда $g'(u) = J(u) - \mu u$, после несложных тождественных преобразований получаем неравенство (17).

Достаточность. Пусть некоторая функция $J(u) \in C^1(U)$ и удовлетворяет неравенству (17). Покажем, что тогда функция $J(u)$ сильно выпукла с постоянной $\varkappa = \mu/2$ и удовлетворяет условию (15) с L, где μ, L взяты из (17). С помощью неравенства Коши — Буняковского из (17) имеем

$$|J'(u) - J'(v)|^2 + L\mu|u-v|^2 \leqslant (L+\mu|J'(u) - J'(v)|\cdot|u-v|.$$

Приняв $x = |J'(u) - J'(v)|$, последнее неравенство можно переписать в виде $x^2 - (L+\mu)|u-v|x + L\mu|u-v|^2 \leqslant 0$. Квадратный трехчлен в левой части этого неравенства имеет корни $x_1 = \mu|u-v|$, $x_2 = L|u-v|$. Поэтому $x_1 \leqslant x \leqslant x_2$, т. е.

$$\mu\,|u-v| \leqslant |J'(u) - J'(v)| \leqslant L\,|u-v| \quad \forall u, v \in U. \tag{18}$$

Тогда $\langle J'(u) - J'(v),\ u-v\rangle \leqslant L|u-v|^2$ — правое неравенство (16) получено.

Используя левое неравенство (18), из (17) имеем $\mu^2|u-v|^2 + L\mu|u-v|^2 \leqslant (L+\mu)\langle J'(u)-J'(v),\ u-v\rangle$. Поделив обе части этого неравенства на $L+\mu>0$, придем к левому неравенству (16). Таким образом, из (17) получили неравенства (18) и (16). Левое неравенство (16) согласно теореме 3 означает сильную выпуклость $J(u)$ с постоянной $\varkappa = \mu/2$, а правое неравенство (16) (или (18)) дает условие (15).

Из (17) вытекает неравенство

$$\langle J'(u)-J'(v),\ v-w\rangle \leqslant \frac{1}{4}(L+\mu)\,|u-w|^2 - \frac{L\mu}{L+\mu}\,|u-v|^2 \quad \forall u,v,w\in U.$$

Оно доказывается так же, как и подобное неравенство (2.20).

Упражнения. 1. При каких a, b, c функция $J(u) = ax^2+2bxy+cy^2$ переменных $u=(x,y)\in E^2$ будет сильно выпукла на E^2?

2. Найти области сильной выпуклости функций $J(u)=\sin(x+y+z)$, $J(u)=\sin(x^2+y^2+z^2)$.

3. Можно ли утверждать, что сильно выпуклая функция обладает более лучшими дифференциальными свойствами по сравнению с выпуклыми функциями? Рассмотреть функцию $J(u)=\langle u,\ u\rangle + g(u)$, где $g(u)$ — выпуклая функция.

4. Доказать, что функция $J(u)$ сильно выпукла на выпуклом множестве U тогда и только тогда, когда функция $g(t)=J(v+t(u-v))$ переменной t $(0\leqslant t\leqslant 1)$ является сильно выпуклой на $[0,\ 1]$ с одной и той же для всех $u,\ v\in U$ постоянной сильной выпуклости.

5. Пусть функция $J(u)$ сильно выпукла и дифференцируема на выпуклом множестве U. Доказать, что:

а) $J'(u)\neq J'(v)$ $(u,\ v\in U,\ u\neq v)$;

б) $|u-v|\leqslant \dfrac{1}{\varkappa}\,|J'(v)|^2$ при всех $u\in M(v)=\{u\in U\colon J(u)\leqslant J(v)\}$, $v\in U$;

в) $0\leqslant J(u)-J_*\leqslant \dfrac{1}{4\varkappa}\,|J'(u)|^2$, $|u-u_*|\leqslant \dfrac{1}{2\varkappa}\,|J'(u)|$ $(u\in U)$.

6. Для того чтобы симметричная матрица A порядка $n\times n$ была положительно определенной, необходимо и достаточно существования постоянных L, μ $(0<\mu\leqslant L)$ таких, что

$$|Ae|^2 + L\mu\,|e|^2 \leqslant (L+\mu)\,\langle Ae,\ e\rangle \qquad \forall e\in E^n.$$

Доказать. Убедиться, что в приведенном неравенстве в качестве μ можно взять минимальное собственное число матрицы A, в качестве L — максимальное собственное число.

7. Пусть U — выпуклое множество, $J(u)\in C^1(U)$. Показать, что для того, чтобы функция $J(u)$ была сильно выпуклой и удовлетворяла условию (15), необходимо и достаточно выполнения неравенств (18) при каких-нибудь постоянных L, μ $(0<\mu\leqslant L)$.

§ 4. Проекция точки на множество

1. При описании и исследовании некоторых методов минимизации ниже нам понадобится понятие проекции точки на множество.

Определение 1. Пусть U — некоторое множество из E^n. *Проекцией точки* u из E^n называется ближайшая к u точка w множества U, т. е. точка $w\in U$, удовлетворяющая условию

$$|u-w| = \inf_{v\in U}|u-v|.$$

Проекцию точки u на множество U будем обозначать через $\mathcal{P}_U(u) = w$.

Поскольку $\rho(u, U) = \inf\limits_{v \in U} |u - v|$ — расстояние от точки u до множества U, то из определения 1 следует, что

$$\rho(u, U) = |u - \mathcal{P}_U(u)| \leqslant |u - v| \quad \forall v \in U, \quad \forall u \in E^n.$$

Если $u \in U$, то, очевидно, всегда $\mathcal{P}_U(u) = u$. Однако проекция на множество существует не всегда. Например, если $U = \{u \in E^n\colon |u| < 1\}$ — открытый единичный шар в E^n, то ни одна точка $u \notin U$ не будет иметь проекции на это множество. Однако если множество U замкнуто, то любая точка $u \in E^n$ имеет проекцию на U — это было доказано в следствии 1 к теореме 2.1.3. Проекция точки на множество может определяться неоднозначно

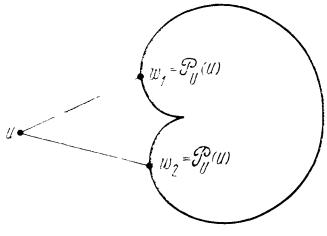

Рис. 4.6

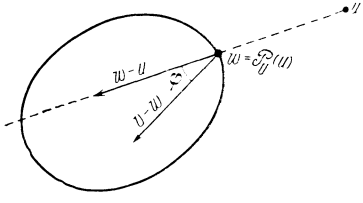

Рис. 4.7

(рис. 4.6). Однако, как показывает следующая теорема, для выпуклых множеств такая ситуация невозможна (рис. 4.7).

Теорема 1. *Пусть U — выпуклое замкнутое множество из E^n. Тогда*:

1) *всякая точка $u \in E^n$ имеет, и притом единственную, проекцию на это множество*;

2) *для того чтобы точка $w \in U$ была проекцией точки u на множество U, необходимо и достаточно выполнения неравенства* (*см. рис.* 4.7)

$$\langle w - u, v - w \rangle \geqslant 0 \quad \forall v \in U. \tag{1}$$

При этом если U — аффинное множество (*см. пример* 1.4), *то вместо* (1) *можно писать*

$$\langle w - u, v - w \rangle = 0 \quad \forall v \in U. \tag{2}$$

Доказательство. Рассмотрим функцию $g(v) = |v - u|^2$ переменной $v \in E^n$ при произвольной фиксированной $u \in E^n$. Поскольку $g(v)$ сильно выпукла на E^n, то по теореме 3.1 эта функция достигает своей нижней грани на U в единственной точке $w \in U$. Это означает, что $|v - u|^2 \geqslant |w - u|^2$ или $|v - u| \geqslant |w - u|$ при всех $v \in U$, причем равенство здесь возможно только при $v = w$. Остается принять $\mathcal{P}_U(u) = w$.

Докажем второе утверждение теоремы. Согласно теореме 2.3 для того, чтобы функция $g(v)$ достигала минимума на U в точке w, необходимо и достаточно, чтобы $\langle g'(w), v-w\rangle = 2\langle w-u, v-w\rangle \geqslant 0$ при всех $v \in U$, что равносильно неравенству (1).

Наконец, пусть $U = \{u \in E^n\colon Au = b\}$ — аффинное множество. Поскольку это множество выпукло и замкнуто, то неравенство (1) сохраняет силу и здесь. Аффинное множество обладает следующим замечательным свойством: если $v, v_0 \in U$ $(v \neq v_0)$, то и $2v_0 - v \in U$, что проверяется непосредственно. Поэтому если здесь взять $v_0 = \mathcal{P}_U(u) = w \in U$, то $2w - v \in U$ при любом выборе $v \in U$. Подставим в (1) вместо v точку $2w - v$. Получим $\langle w-u, 2w-v-w\rangle = \langle w-u, w-v\rangle \geqslant 0$ при всех $v \in U$. Сравнивая полученное неравенство с (1), приходим к равенству (2).

Покажем, что оператор проектирования на выпуклое множество обладает сжимающим свойством.

Теорема 2. *Если U — выпуклое замкнутое множество из E^n, то*

$$|\mathcal{P}_U(u) - \mathcal{P}_U(v)| \leqslant |u - v| \qquad \forall u, v \in E^n. \tag{3}$$

Доказательство. Из неравенства (1) имеем

$$\langle \mathcal{P}_U(u) - u, \mathcal{P}_U(v) - \mathcal{P}_U(u)\rangle \geqslant 0.$$

Поменяв ролями точки u и v в последнем неравенстве, получим

$$\langle \mathcal{P}_U(v) - v, \mathcal{P}_U(u) - \mathcal{P}_U(v)\rangle \geqslant 0.$$

Сложим эти два неравенства. Имеем

$$\langle \mathcal{P}_U(u) - u - \mathcal{P}_U(v) + v, \mathcal{P}_U(v) - \mathcal{P}_U(y)\rangle \geqslant 0.$$

Отсюда следует

$$|\mathcal{P}_U(u) - \mathcal{P}_U(v)|^2 \leqslant \langle \mathcal{P}_U(u) - \mathcal{P}_U(v), u - v\rangle \leqslant$$
$$\leqslant |\mathcal{P}_U(u) - \mathcal{P}_U(v)| \cdot |u - v| \qquad \forall u, v \in E^n.$$

Разделив на $|\mathcal{P}_U(u) - \mathcal{P}_U(v)| \neq 0$, получим требуемое неравенство (3). Если $|\mathcal{P}_U(u) - \mathcal{P}_U(v)| = 0$, то (3) очевидно.

2. Приведем примеры множеств, проекция на которые может быть выписана явно.

Пример 1. Пусть $U = S(u_0, R) = \{u \in E^n\colon |u - u_0| \leqslant R\}$ — шар радиуса $R > 0$ с центром в точке u_0. Из геометрических соображений (рис. 4.8) ясно, что проекцией точки $u \notin U$ является точка

$$w = u_0 + R(u - u_0)/|u - u_0|.$$

Для строгого доказательства этого факта достаточно проверить выполнение неравенства (1). Имеем

$$\langle w - u, v - w\rangle = (R/|u - u_0| - 1)(\langle u - u_0, v - u_0\rangle -$$
$$- R|u - u_0|) \geqslant 0,$$

так как $|u-u_0|>R$, а $\langle u-u_0,\ v-u_0\rangle \leqslant |u-u_0|\cdot|v-u_0| \leqslant |u-u_0|R$ в силу неравенства Коши — Буняковского для всех $v\in U$.

Пример 2. Пусть $U=\Gamma=\{u\in E^n\colon \langle c,\ u\rangle=\gamma\}$ — гиперплоскость; здесь $c\in E^n$, $c\neq 0$, $\gamma=\mathrm{const}$. Пользуясь геометрическими соображениями (рис. 4.9), проекцию точки $u\notin U$ на U

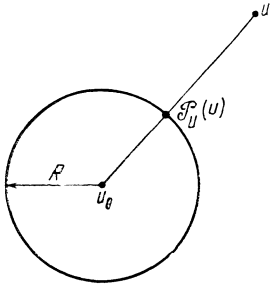

Рис. 4.8

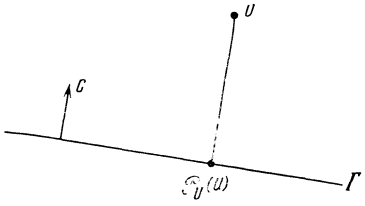

Рис. 4.9

будем искать в виде $w=u+\alpha c$. Определяя число α из условия $w\in U$, имеем

$$w=u+(\gamma-\langle c,\ u\rangle)c/|c|^2.$$

Поскольку $\langle w-u,\ v-w\rangle=(\gamma-\langle c,\ u\rangle)/|c|^2\cdot\langle c,\ v-w\rangle=0$ при всех $v\in U$, то согласно теореме 1 найденная точка w представляет собой проекцию точки u на U.

Пример 3. Пусть $U=\{u\in E^n\colon \langle a_i,\ u\rangle=b^i,\ i=1,\ \ldots,\ m\}$ — аффинное множество; здесь $a_i\in E^n$, $b^i=\mathrm{const}$ $(i=1,\ \ldots,\ m)$. Можем считать, что векторы $a_1,\ \ldots,\ a_m$ линейно независимы и $m<n$ (если $m=n$, то U будет состоять из одной точки). Проекцию точки u на множество U будем искать в виде

$$w=u+\sum_{j=1}^{m}\alpha_j a_j. \tag{4}$$

Из требования $w\in U$ имеем систему линейных алгебраических уравнений

$$\sum_{j=1}^{m}\alpha_j\langle a_i,\ a_j\rangle=b^i-\langle a_i,\ u\rangle,\qquad i=1,\ \ldots,\ m, \tag{5}$$

для определения коэффициентов $\alpha_1,\ \ldots,\ \alpha_m$. Определителем этой системы является определитель Грама [54, 93, 164], который для линейно независимых $a_1,\ \ldots,\ a_m$ будет отличным от нуля. Поэтому искомые $\alpha_1,\ \ldots,\ \alpha_m$ существуют и однозначно определяются из системы (5). Для точки w из (4) будем иметь

$$\langle w-u,\ v-w\rangle=\sum_{j=1}^{m}\alpha_j\langle a_j,\ v-u\rangle-\sum_{i=1}^{m}\alpha_i\left(\sum_{j=1}^{m}\alpha_j\langle a_i,\ a_j\rangle\right)=0$$

для всех $v \in U$. Следовательно, по теореме 1 найденная из (4), (5) точка w будет проекцией точки u на множество U. Если ввести матрицу A, строками которой являются векторы a_i $(i = 1, \ldots, m)$, то точку (4) можно записать в виде

$$w = u - A^T(AA^T)^{-1}(Au - b).$$

Предлагаем читателю провести проверку того, что такая точка w принадлежит U, т. е. $Aw = b$, и выполняется условие (1).

Пример 4. Пусть $U = \{u \in E^n\colon \langle c, u\rangle \leqslant \gamma\}$ — замкнутое полупространство, определяемое гиперплоскостью $\langle c, u\rangle = \gamma$. Пусть $u \notin U$, т. е. $\langle c, u\rangle > \gamma$. Как и в примере 2, попробуем представить проекцию точки u на U в виде

$$w = u + (\gamma - \langle c, u\rangle)c/|c|^2.$$

Имеем $\langle w - u, v - w\rangle = (\gamma - \langle c, u\rangle)|c|^{-2}(\langle c, v\rangle - \gamma) \geqslant 0$ при всех $v \in U$. Следовательно, точка w — искомая проекция.

Пример 5. Пусть $U = \{u = (u^1, \ldots, u^n) \in E^n\colon \alpha_i \leqslant u^i \leqslant \beta_i, i = 1, \ldots, n\}$ — n-мерный параллелепипед, где α_i, β_i $(\alpha_i < \beta_i)$ — заданные числа, $i = 1, \ldots, n$. Пусть $u \notin U$. Положим $w = (w^1, \ldots, w^n)$, где

$$w^i = \begin{cases} \alpha_i, & u^i < \alpha_i, \\ \beta_i, & u^i > \beta_i, \\ u^i, & \alpha_i \leqslant u^i \leqslant \beta_i, \qquad i = 1, \ldots, n. \end{cases}$$

Тогда $(w^i - u^i)(v^i - w^i) \geqslant 0$ для всех v^i $(\alpha_i \leqslant v^i \leqslant \beta_i, i = 1, \ldots, n)$. Отсюда, суммируя по i от 1 до n, получаем $\langle w - u, v - w\rangle \geqslant 0$ для всех $v \in U$. Следовательно, построенная точка w является проекцией точки u на множество U.

Пример 6. Пусть $U = E^n_+ = \{u = (u^1, \ldots, u^n)\colon u^i \geqslant 0, i = 1, \ldots, n\}$ — неотрицательный октант пространства E^n. Легко проверить, что проекцией точки u на U является точка $u^+ = ((u^1)^+, \ldots, (u^n)^+)$, где $(u^i)^+ = \max\{0; u^i\}$ $(i = 1, \ldots, n)$.

3. Критерий оптимальности, сформулированный в теореме 2.3, с помощью оператора проектирования может быть переформулирован следующим образом.

Теорема 3. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$, U_* — множество точек минимума функции $J(u)$ на U. Если $u_* \in U_*$, то необходимо выполняется равенство*

$$u_* = \mathscr{P}_U(u_* - \alpha J'(u_*)) \qquad \forall \alpha > 0. \tag{6}$$

Если, кроме того, $J(u)$ выпукла на U, то всякая точка u_*, удовлетворяющая уравнению (6), принадлежит U_*.

Доказательство. Согласно теореме 1 равенство (6) эквивалентно неравенству $\langle u_* - (u_* - \alpha J'(u_*)), v - u_*\rangle \geqslant 0$ $(v \in U)$, откуда имеем $\alpha\langle J'(u_*), v - u_*\rangle \geqslant 0$ $(v \in U)$. Поскольку $\alpha > 0$, то отсюда получим неравенство $\langle J'(u_*), v - u_*\rangle \geqslant 0$ при

всех $v \in U$. Таким образом, условия (6) и (2.5) эквивалентны. Отсюда и из теоремы 2.3 следует утверждение теоремы 3.

Таким образом, если ввести отображение A из E^n в E^n по формуле

$$Au = \mathscr{P}_U(u - \alpha J'(u)), \qquad \alpha > 0,$$

то условие (6) перепишется в виде $u_* = Au_*$, т. е. u_* — неподвижная точка отображения A. Ниже мы увидим, что при некоторых условиях на функцию $J(u)$ отображение будет сжимающим и для определения точки u_* могут быть использованы свойства сжимающих отображений [54, 179].

Упражнения. 1. Найти проекцию точки $u \in E^n$ на множество

$$U = \{u \in E^n\colon \langle a_1, u\rangle \leqslant b^1, \langle a_2, u\rangle \leqslant b^2\}.$$

2. Найти проекцию точки $u \in E^n$ на множество

$$U = \{u = (u^1, \ldots, u^n)\colon \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\}$$

(здесь $\alpha_i \leqslant \beta_i$, причем возможно, что $\alpha_i = \beta_i$, или $\alpha_j = -\infty$, или $\beta_k = \infty$ при некоторых i, j, k).

3. Выяснить геометрический смысл равенства (2).

4. Будут ли верными неравенства (1) или равенство (2), если U — невыпуклое множество?

5. Охарактеризовать все множества U из E^n, для которых существует точка $u \notin U$ такая, что $\mathscr{P}_U(u) = v$ для всех $v \in U$.

6. Для того чтобы точка $w \in U$ была проекцией точки u на выпуклое множество U, необходимо и достаточно, чтобы $\langle v - u, v - w\rangle \geqslant 0$ при всех $v \in U$. Доказать. Выяснить геометрический смысл этого условия.

7. Доказать, что для любого замкнутого множества U имеет место неравенство $\big||u - \mathscr{P}_U(u)| - |v - \mathscr{P}_U(v)|\big| \leqslant |u - v|$ для всех $u, v \in U$ (ср. с леммой 2.1.2).

8. Пусть U — выпуклое замкнутое множество из E^n. Доказать, что тогда

$$|v - \mathscr{P}_U(u)|^2 \leqslant \langle v - u, v - \mathscr{P}_U(u)\rangle \qquad \forall v \in U, \quad \forall u \in E^n,$$

$$|v - \mathscr{P}_U(u)|^2 + |u - \mathscr{P}_U(u)|^2 \leqslant |v - u|^2 \qquad \forall v \in U, \quad \forall u \in E^n.$$

9. Пусть $J(u) = |Au - b|^2$, где A — матрица порядка $m \times n$, $b \in E^m$ (см. пример 2.4). Доказать, что

$$U_* = \left\{u \in E^n\colon J(u) = \inf_{E^n} J(u) = J_*\right\} \neq \varnothing.$$

Указание: взять проекцию точки b на множество $U = \{v \in E^m\colon v = Au,\ u \in E^n\}$ и показать, что $J(u) = |Au - \mathscr{P}_U(b)|^2 + |b - \mathscr{P}_U(b)|^2$, $J_* = |b - \mathscr{P}_U(b)|^2$; замкнутость U см. в лемме 9.3.

§ 5. Отделимость выпуклых множеств

1. В теории экстремальных задач важную роль играют теоремы, называемые *теоремами отделимости*. Основное содержание этих теорем сводится к тому, что для некоторых двух множеств A и B утверждается существование гиперплоскости такой, что множество A находится в одном из открытых или замкнутых полупространств, определяемых этой гиперплоскостью, а множе-

ство B — в другом открытом или замкнутом полупространстве (см. пример 1.3), т. е. гиперплоскости, которая отделяет эти два множества.

Определение 1. Пусть A и B — два множества из E^n. Говорят, что *гиперплоскость* $\langle c, u\rangle = \gamma$ *с нормальным вектором* $c \neq 0$ *отделяет* (*разделяет*) *множества* A и B, если $\langle c, a\rangle \geqslant \gamma$ при всех $a \in A$ и $\langle c, b\rangle \leqslant \gamma$ при всех $b \in B$, или, иначе говоря, выполняются неравенства

$$\sup_{b\in B} \langle c, b\rangle \leqslant \gamma \leqslant \inf_{a\in A} \langle c, a\rangle. \tag{1}$$

Если $\sup\limits_{b\in B} \langle c, b\rangle < \inf\limits_{a\in A} \langle c, a\rangle$, то говорят, что множества A и B *сильно отделены*. Если $\langle c, b\rangle < \langle c, a\rangle$ при всех $a \in A$, $b \in B$, то говорят о *строгом отделении* этих множеств. Если выполнено (1), причем существуют такие точки $a_0 \in A$, $b_0 \in B$, что $\langle c, b_0\rangle < \langle c, a_0\rangle$, то говорят, что множества A, B *собственно отделимы*.

Понятие собственной отделимости введено для того, чтобы исключить из (1) вырожденный случай, когда оба множества A, B лежат в разделяющей гиперплоскости и, возможно, даже имеют общие относительно внутренние точки.

Заметим, что в определение 1 множества A и B входят несколько несимметрично. Однако симметрию здесь нетрудно восстановить: нужно лишь взять вектор $-c$ вместо c, постоянную $-\gamma$ вместо γ и записать уравнение отделяющей гиперплоскости в виде $\langle -c, u\rangle = -\gamma$. Очевидно, если гиперплоскость $\langle c, u\rangle = \gamma$ отделяет множества A и B, то гиперплоскость $\langle \mu c, u\rangle = \mu\gamma$ при $\mu \neq 0$ также отделяет эти множества. Поэтому при необходимости можно считать, что $|c| = 1$.

На рис. 4.10—4.14 изображены случаи, когда два множества собственно отделимы, на рис. 4.13 — сильно отделимы, на рис. 4.14 — строго отделимы. Однако ясно, что не всякие два множества можно отделить гиперплоскостью (рис. 4.15). Ниже приводятся теоремы об отделимости выпуклых множеств.

Теорема 1. *Пусть X — непустое выпуклое множество из E^n. Тогда для любой точки $y \notin \operatorname{ri} X$ существует гиперплоскость $\langle c, u\rangle = \gamma$, собственно отделяющая множество X и точку y, или, точнее,*

$$\begin{aligned} &\langle c, x\rangle \geqslant \gamma \quad \forall x \in X, \quad \gamma \geqslant \langle c, y\rangle, \\ &\langle c, x\rangle > \gamma \quad \forall x \in \operatorname{ri} X. \end{aligned} \tag{2}$$

Если точка y не принадлежит $\overline{X}$ — замыканию X, то множество X (а также и $\overline{X}$) сильно отделимо от y.

Доказательство. Сначала рассмотрим случай $y \notin \overline{X}$. Напомним, что если X — выпуклое множество, то $\overline{X}$ также выпукло (см. теорему 1.2). Согласно теореме 4.1 тогда существует проекция $z = \mathscr{P}_{\overline{X}}(y)$ точки y на множество $\overline{X}$, причем $\langle z - y, x - z\rangle \geqslant 0$ для всех $x \in \overline{X}$. Положим $c = z - y$. С учетом

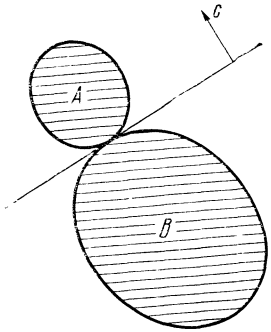

Рис. 4.10

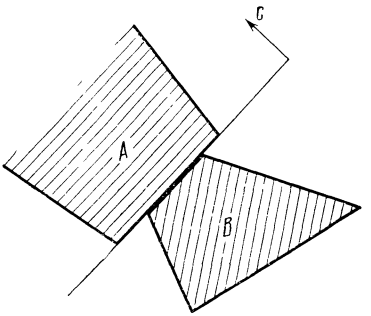

Рис. 4.11

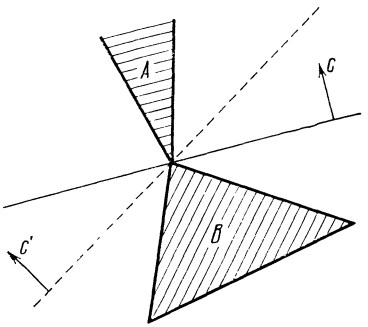

Рис. 4.12

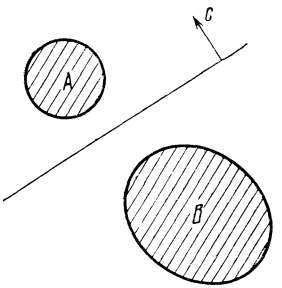

Рис. 4.13

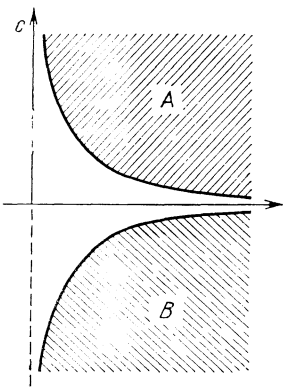

Рис. 4.14

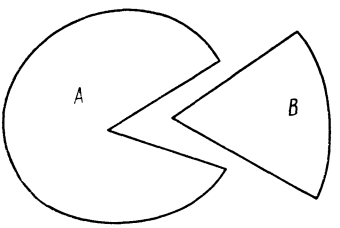

Рис. 4.15

предыдущего неравенства будем иметь $\langle c, x-y\rangle = \langle z-y, x-z\rangle + \langle z-y, z-y\rangle \geqslant |c|^2 > 0$ или $\langle c, x\rangle \geqslant \langle c, y\rangle + |c|^2 > \langle c, y\rangle$ $(x \in \overline{X})$. Это значит, что гиперплоскость $\langle c, u\rangle = \langle c, y\rangle = \gamma$ сильно отделяет точку y от множества $\overline{X}$ и, тем более, множества X. Нетрудно видеть, что такая гиперплоскость не единственна. Например, гиперплоскость $\langle c, u\rangle = \langle c, z\rangle = \gamma$ также сильно отделяет $\overline{X}$ и y, так как $\langle c, x-z\rangle \geqslant 0$ при всех $x \in \overline{X}$, $\langle c, y-z\rangle = \langle z-y, y-z\rangle = -|c|^2 < 0$.

Теперь пусть $y \in \overline{X}$, $y \notin \operatorname{ri} X$. Согласно теореме 1.11 тогда существует последовательность $\{y_k\} \to y$: $y_k \notin \overline{X}$, $y_k \in \operatorname{aff} \overline{X}$ $(k = 1, 2, \ldots)$. По доказанному гиперплоскость $\langle c_k, u\rangle = \langle c_k, y_k\rangle$, где $c_k = (z_k - y_k)/|z_k - y_k|$, $z_k = \mathscr{P}_{\overline{X}}(y_k)$, сильно отделяет $\overline{X}$ и y_k, так что $\langle c_k, x\rangle > \langle c_k, y_k\rangle$ при всех $x \in \overline{X}$. По теореме Больцано — Вейерштрасса из последовательности $\{c_k\}$ $(|c_k| = 1)$ можно выбрать подпоследовательность, сходящуюся к некоторому вектору c $(|c| = 1)$. Без умаления общности можем считать, что вся последовательность $\{c_k\} \to c$. Переходя к пределу при $k \to \infty$ в неравенстве $\langle c_k, x\rangle > \langle c_k, y_k\rangle$ $(x \in \overline{X})$, получаем $\langle c, x\rangle \geqslant \langle c, y\rangle = \gamma$ для всех $x \in \overline{X}$. Отсюда следует первое из неравенств (2).

Далее, возьмем любую точку $x \in \operatorname{ri} X$. Согласно определению 1.10 существует такое $\varepsilon > 0$, что $O(x, 2\varepsilon) \cap \operatorname{aff} X \in X$. Тогда $u_k = x - \varepsilon c_k \in O(x, 2\varepsilon) \cap \operatorname{aff} X$ $(k = 1, 2, \ldots)$. Ясно, что $\{u_k\} \to u = x - \varepsilon c$. Покажем, что $u = x - \varepsilon c \in \operatorname{ri} X$. Поскольку $z_k = \mathscr{P}_{\overline{X}}(y_k) \in \overline{X} \subset \operatorname{aff} \overline{X} = \operatorname{aff} X$, $y_k \in \operatorname{aff} X$, то $z_k - y_k \in \operatorname{Lin} X$ — подпространство, параллельное $\operatorname{aff} X$. Тогда $c_k = (z_k - y_k)/|z_k - y_k| \in \operatorname{Lin} X$. Поскольку $\operatorname{Lin} X$ замкнуто в E^n, $\{c_k\} \to c$, то $c \in \operatorname{Lin} X$. Отсюда следует, что $u = x - \varepsilon c \in \operatorname{aff} X$. Кроме того, так как $|u_k - x| = \varepsilon$ $(k = 1, 2, \ldots)$, то при $k \to \infty$ имеем $|u - x| = \varepsilon$, так что $u = x - \varepsilon c \in O(x, 2\varepsilon)$. Следовательно, $u = x - \varepsilon c \in \operatorname{ri} X \subset X \subset \overline{X}$. Тогда $\gamma = \langle c, y\rangle \leqslant \langle c, u\rangle = \langle c, x - \varepsilon c\rangle = \langle c, x\rangle - \varepsilon < \langle c, x\rangle$, или $\gamma < \langle c, x\rangle$ для каждой точки $x \in \operatorname{ri} X$, т. е. $\operatorname{ri} X$ и y строго отделимы.

Теорема 2. *Пусть A и B — непустые выпуклые множества из E^n, $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$ (например, $A \cap B = \varnothing$). Тогда существует гиперплоскость $\langle c, u\rangle = \gamma$, строго отделяющая множества $\operatorname{ri} A$ и $\operatorname{ri} B$, собственно отделяющая множества A и B, а также их замыкания $\overline{A}$, $\overline{B}$, причем если $\overline{A}$ и $\overline{B}$ имеют общую граничную точку y, то $\gamma = \langle c, y\rangle$. Верно и обратное: если два выпуклых множества A и B собственно отделимы, то $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$.*

Доказательство. Введем множество $X = \operatorname{ri} A - \operatorname{ri} B = \{x \in E^n\colon x = a - b,\ a \in \operatorname{ri} A,\ b \in \operatorname{ri} B\}$. По теоремам 1.1, 1.10 множество X выпукло. Поскольку $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$, то $0 \notin X$. Возможно два случая: $0 \notin \overline{X}$ или $0 \in \overline{X}$. Если $0 \notin \overline{X}$, то согласно теореме 1 точка 0 и множество X сильно отделимы, т. е. существуют такие $c \neq 0$, $\varepsilon > 0$, что $\langle c, x\rangle \geqslant \langle c, 0\rangle + \varepsilon = \varepsilon$ при всех $x \in X$. Если $0 \in \overline{X}$, $0 \notin X$, то по той же теореме 1 найдется такой вектор $c \neq 0$, что

$\langle c, x\rangle \geqslant 0$ для всех $x \in X$, причем $\langle c, x\rangle > 0$ при $x \in \operatorname{ri} X$. Таким образом, в обоих случаях существует такой вектор $c \neq 0$, что

$$\langle c, x\rangle \geqslant 0 \quad \forall x \in X, \qquad \langle c, x\rangle > 0 \quad \forall x \in \operatorname{ri} X. \tag{3}$$

Из первого неравенства (3) с учетом определения множества X имеем

$$\langle c, a\rangle \geqslant \langle c, b\rangle \quad \forall a \in \operatorname{ri} A, \quad b \in \operatorname{ri} B. \tag{4}$$

Далее, по теореме 1.10 $\operatorname{ri} X \neq \varnothing$. Это значит, что существуют $a_0 \in \operatorname{ri} A$, $b_0 \in \operatorname{ri} B$ такие, что $x_0 = a_0 - b_0 \in \operatorname{ri} X$. Из второго неравенства (3) тогда имеем $\langle c, x_0\rangle = \langle c, a_0 - b_0\rangle > 0$, или

$$\langle c, a_0\rangle > \langle c, b_0\rangle, \quad a_0 \in \operatorname{ri} A, \quad b_0 \in \operatorname{ri} B. \tag{5}$$

Неравенство (4) остается справедливым для всех предельных точек множеств $\operatorname{ri} A$, $\operatorname{ri} B$, т. е. для всех $a \in \overline{\operatorname{ri} A}$, $b \in \overline{\operatorname{ri} B}$. Но по теореме 1.12 $\overline{\operatorname{ri} A} = \bar{A}$, $\overline{\operatorname{ri} B} = \bar{B}$, так что $\langle c, a\rangle \geqslant \langle c, b\rangle$ при любых $a \in \bar{A}$, $b \in \bar{B}$. Отсюда $\inf\limits_{a \in \bar{A}} \langle c, a\rangle \geqslant \sup\limits_{b \in \bar{B}} \langle c, b\rangle$. Возьмем гиперплоскость $\langle c, u\rangle = \gamma$, где γ — произвольное число, удовлетворяющее неравенству $\inf\limits_{a \in \bar{A}} \langle c, a\rangle \geqslant \gamma \geqslant \sup\limits_{b \in \bar{B}} \langle c, b\rangle$. Тогда

$$\langle c, a\rangle \geqslant \gamma \geqslant \langle c, b\rangle \quad \forall a \in \bar{A}, \quad b \in \bar{B}. \tag{6}$$

Из (5), (6) следует собственно отделимость множеств $\bar{A}$ и $\bar{B}$ и, тем более, множеств A и B. Если $y \in \bar{A} \cap \bar{B}$, то $\gamma = \langle c, y\rangle$.

Покажем, что построенная гиперплоскость $\langle c, u\rangle = \gamma$ строго отделяет множества $\operatorname{ri} A$ и $\operatorname{ri} B$. Из (5), (6) следует, что либо $\langle c, a_0\rangle > \gamma$, либо $\langle c, b_0\rangle < \gamma$. Пусть $\langle c, a_0\rangle > \gamma$ (случай $\langle c, b_0\rangle < \gamma$ рассматривается аналогично). Возьмем произвольную точку $a \in \operatorname{ri} A$. Тогда $a - a_0 \in \operatorname{Lin} A$, $a + \varepsilon(a - a_0) \in \operatorname{aff} A$ при всех $\varepsilon \in \mathbf{R}$ и по определению относительно внутренней точки найдется такое $\varepsilon > 0$, что $b = a + \varepsilon(a - a_0) \in \operatorname{ri} A$. Отсюда $a = \alpha a_0 + (1 - \alpha) b$ ($\alpha = \varepsilon/(1 + \varepsilon) \in (0, 1)$). Умножим неравенство $\langle c, a_0\rangle > \gamma$ на α, $\langle c, b\rangle \geqslant \gamma$ на $1 - \alpha$ и сложим. Получим $\alpha\langle c, a_0\rangle + (1 - \alpha)\langle c, b\rangle = \langle c, a\rangle > \gamma$. Таким образом, $\langle c, a\rangle > \gamma$ при всех $a \in \operatorname{ri} A$. Отсюда и из (6) следует, что множества $\operatorname{ri} A$ и $\operatorname{ri} B$ строго отделимы.

Докажем вторую часть теоремы. Пусть множества A и B собственно отделимы, но пусть тем не менее $\operatorname{ri} A \cap \operatorname{ri} B \neq \varnothing$. Возьмем какую-либо точку $u \in \operatorname{ri} A \cap \operatorname{ri} B$. Тогда при достаточно малом $\varepsilon > 0$ имеем $a_\varepsilon = u - \varepsilon(a_0 - u) \in \operatorname{ri} A$, $b_\varepsilon = u - \varepsilon(b_0 - u) \in \operatorname{ri} B$, где a_0, b_0 взяты из (5). В силу (4) $\langle c, a_\varepsilon\rangle = \langle c, u\rangle(1 + \varepsilon) - \varepsilon\langle c, a_0\rangle \geqslant \langle c, b_\varepsilon\rangle = \langle c, u\rangle(1 + \varepsilon) - \varepsilon\langle c, b_0\rangle$. Отсюда получаем $\langle c, a_0\rangle \leqslant \langle c, b_0\rangle$, что противоречит (5). Следовательно, $\operatorname{ri} A \cap \operatorname{ri} B = \varnothing$.

Приведем одну теорему о сильной отделимости двух выпуклых множеств.

Теорема 3. *Пусть A и B — два выпуклых замкнутых множества, не имеющие общих точек, причем хотя бы одно из этих множеств ограничено. Тогда множества A и B сильно отделимы.*

Доказательство. Введем множество $X = A - B$. Покажем, что оно замкнуто. Пусть x — некоторая предельная точка множества X, пусть последовательность $\{x_k\} \in X$ сходится к x. Поскольку $x_k \in X$, то найдутся $a_k \in A$, $b_k \in B$ такие, что $x_k = a_k - b_k$ $(k = 1, 2, \ldots)$. По условию одно из множеств A или B ограничено. Пусть для определенности ограничено множество A. Тогда последовательность $\{a_k\} \in A$ ограничена. По теореме Больцано — Вейерштрасса найдется подпоследовательность $\{a_{k_n}\}$, сходящаяся к некоторой точке a. В силу замкнутости A точка a принадлежит A. Тогда $b_{k_n} = a_{k_n} - x_{k_n} \to b = a - x$ при $k_n \to \infty$, причем $b \in B$ в силу замкнутости B. Таким образом, для точки x получили представление $x = a - b$, где $a \in A$, $b \in B$. Это значит, что $x \in X$. Замкнутость X доказана.

Далее, по условию множества A и B не имеют общих точек. Поэтому $0 \notin X = \overline{X}$. По теореме 1 тогда существует гиперплоскость $\langle c, u\rangle = 0$ такая, что $\langle c, x\rangle \geqslant |c|^2 > 0$ для всех $x \in X$. Отсюда имеем $\langle c, a - b\rangle \geqslant |c|^2$, или $\langle c, a\rangle \geqslant \langle c, b\rangle + |c|^2$ для всех $a \in A$, $b \in B$. Следовательно, $\inf\limits_{a \in A} \langle c, a\rangle \geqslant \sup\limits_{b \in B} \langle c, b\rangle + |c|^2$. Любая гиперплоскость $\langle c, u\rangle = \gamma$, где $\sup\limits_{b \in B} \langle c, b\rangle \leqslant \gamma \leqslant \inf\limits_{a \in B} \langle c, a\rangle$, будет сильно отделять множества A и B, что и требовалось.

Заметим, что требование ограниченности хотя бы одного из множеств в теореме 3 не может быть ослаблено (см. рис. 4.14).

2. Теоремы отделимости являются одним из важных интрументов исследования свойств выпуклых функций и множеств, экстремальных задач. Ряд приложений этих теорем будут даны в последующих параграфах. Здесь же мы воспользуемся ими для получения представления любого выпуклого замкнутого множества из E^n в виде пересечения некоторого семейства полупространств.

Определение 2. Гиперплоскость $\Gamma = \{u \in E^n\colon \langle c, u\rangle = \gamma\}$ называют *опорной* к множеству X, если $\langle c, x\rangle \geqslant \gamma$ при всех $x \in X$ и $\langle c, y\rangle = \gamma$ для некоторой точки $y \in \overline{X}$. Опорную к X гиперплоскость Γ называют *собственно опорной* к X, если X не содержится в Γ, т. е. $\langle c, x_0\rangle > \gamma$ при некотором $x_0 \in X$. Вектор c, являющийся нормальным вектором опорной [собственно опорной] к X гиперплоскости, проходящей через точку $y \in \overline{X}$, называют *опорным* [*собственно опорным*] вектором множества X в точке y (рис. 4.16).

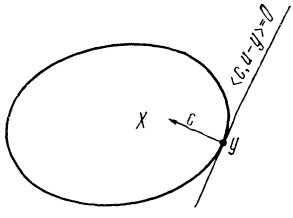

Рис. 4.16

Отметим, что через любую граничную точку y выпуклого множества X из E^n может быть проведена хотя бы одна опорная к X гиперплоскость. В самом деле, если $\operatorname{int} X \neq \varnothing$, то граничными для X будут только точки $y \in \overline{X}$, $y \notin \operatorname{int} X$, и согласно теореме 1 через каждую такую точку y можно провести собственно опорную к X гиперплоскость. Если $\operatorname{int} X = \varnothing$, то $\operatorname{aff} X \neq E^n$, $\operatorname{Гр} X = \overline{X}$, и через каждую точку $y \in \overline{X}$ можно провести гиперплоскость Γ: $\langle c, x - y\rangle = 0$, где c — любой ненулевой вектор из ортогонального дополнения к $\operatorname{Lin} X$. Тогда $X \subset \operatorname{aff} X \subset \Gamma$, так что Γ — опорная

к X гиперплоскость, не являющаяся собственно опорной. Теорема 1 уточняет, что если $y \in \overline{X}$, $y \notin \operatorname{ri} X$, то среди опорных к X гиперплоскостей, проходящих через точку y, можно найти собственно опорную.

Заметим, что выпуклое множество с непустой внутренностью не может иметь опорных гиперплоскостей, не являющихся собственно опорными. Это вытекает из следующего несколько более общего утверждения.

Теорема 4. *Пусть X — выпуклое множество из E^n, $\operatorname{int} X \neq \varnothing$. Пусть вектор $c \neq 0$ и число γ таковы, что $\langle c, x\rangle \geqslant \gamma$ при всех $x \in X$. Тогда*

$$\langle c, x\rangle > \gamma \qquad \forall x \in \operatorname{int} X.$$

Доказательство. Допустим противное: пусть существует такая точка $x_0 \in \operatorname{int} X$, что $\langle c, x_0\rangle = \gamma$. По определению внутренней точки найдется такое $\varepsilon_0 > 0$, что $z = x_0 - \varepsilon c/|c| \in X$ при всех ε $(0 < \varepsilon < \varepsilon_0)$. Тогда $\gamma \leqslant \langle c, z\rangle = \langle c, x_0\rangle - \varepsilon|c| = \gamma - \varepsilon|c| < \gamma$. Получилось противоречивое неравенство, из которого и следует утверждение теоремы.

Следствие 1. *Пусть X — выпуклое множество из E^n, $\operatorname{int} X \neq \varnothing$. Тогда любая гиперплоскость $\langle c, x - y\rangle = 0$, опорная к множеству X в какой-либо точке $y \in \operatorname{Гр} X$, является собственно опорной к X, или, точнее,*

$$\langle c, x - y\rangle > 0 \qquad \forall x \in \operatorname{int} X.$$

Доказательство. Из определения опорной гиперплоскости к X в точке y следует, что $\langle c, x\rangle \geqslant \langle c, y\rangle = \inf\limits_{x \in X} \langle c, x\rangle = \inf\limits_{x \in \overline{X}} \langle c, x\rangle = \gamma$ при всех $x \in X$. В силу теоремы 4 тогда $\langle c, x\rangle > \gamma = \langle c, y\rangle$ для любой точки $x \in \operatorname{int} X$.

В следующей теореме показывается, что выпуклое замкнутое множество полностью характеризуется своими опорными гиперплоскостями.

Теорема 5. *Всякое непустое выпуклое замкнутое множество X из E^n $(X \neq E^n)$ является пересечением замкнутых полупространств, образованных всевозможными опорными гиперплоскостями к множеству X, содержащими X.*

Доказательство. Поскольку $X \neq E^n$, то $\operatorname{Гр} X \neq \varnothing$. Возьмем любую точку $y \in \operatorname{Гр} X$. Множество всех опорных векторов множества X в точке y обозначим через C_y. Выше было замечено, что $C_y \neq \varnothing$ при всех $y \in \operatorname{Гр} X$. Обозначим $A = \bigcap\limits_{y \in \operatorname{Гр} X} \bigcap\limits_{c \in C_y} \{x\colon \langle c, x - y\rangle \geqslant 0\}$. Нам надо показать, что $A = X$. Если $x \in X$, то для всех $y \in \operatorname{Гр} X$ и всех $c \in C_y$ имеем $\langle c, x - y\rangle \geqslant 0$, т. е. $x \in A$. Следовательно, $X \subset A$.

Докажем обратное включение $A \subset X$. Допустим противное: пусть существует точка $a \in A$, $a \notin X$. Поскольку X — замкнутое множество, то по теореме 1 множество X и точка a сильно отделимы. Точнее, при доказательстве теоремы 1 было показано, что гиперплоскость $\langle c_z, u - z\rangle = 0$, где $z = \mathscr{P}_X(a)$, $c_z = z - a$, такова, что $\langle c_z, x - z\rangle \geqslant 0$ при всех $x \in X = \overline{X}$, а $\langle c_z, a - z\rangle < 0$. Это значит, что $c_z \in C_z$ $(z \in \operatorname{Гр} X)$, и поэтому для точки $a \in A$ должно бы быть $\langle c_z, a - z\rangle \geqslant 0$ в силу определения A. Полученное противоречие показывает, что $A \subset X$. Требуемое равенство $A = X$ доказано.

Согласно теореме 5 выпуклое замкнутое множество характеризуется системой неравенств $\langle c, x\rangle \geqslant \langle c, y\rangle$ $(x \in X)$, которые можно записать в виде $\inf\limits_X \langle c, x\rangle = \langle c, y\rangle$, $c \in C_y$, $y \in \operatorname{Гр} X$. Если заменить c на $-e$, то эти условия приводят к равенствам $\sup\limits_X \langle e, x\rangle = \langle e, y\rangle$, $e \in -C_y$, $y \in \operatorname{Гр} X$. Таким образом, всякое выпуклое замкнутое множество X характеризуется значениями функции $\delta(e, X) = \sup\limits_X \langle e, x\rangle$, называемой *опорной функцией* множества X (здесь возможны значения $\delta(e, X) = \infty$ для некоторых $e \in E^n$). Это обстоятельство отражено также и в следующей теореме.

Теорема 6. *Пусть для двух множеств A, B из E^n известно, что*

$$\sup_{a\in A}\langle e, a\rangle \leqslant \sup_{b\in B}\langle e, b\rangle \qquad \forall e\in E^n, \quad |e|=1.$$

Тогда $\overline{\operatorname{co} A}\subset\overline{\operatorname{co} B}$; *в частности, если A, B выпуклы и замкнуты, то $A\subset B$. Если*

$$\sup_{a\in A}\langle e, a\rangle = \sup_{b\in B}\langle e, b\rangle \qquad \forall e\in E^n, \quad |e|=1,$$

то $\overline{\operatorname{co} A}=\overline{\operatorname{co} B}$; *в частности, если A, B выпуклы и замкнуты, то $A=B$.*

Доказательство. Допустим, что $\overline{\operatorname{co} A}\not\subset\overline{\operatorname{co} B}$. Тогда существует точка $a_0\in\overline{\operatorname{co} A}$, но $a_0\notin\overline{\operatorname{co} B}$. По теореме 5.1 множество $\overline{\operatorname{co} B}$ и точка a_0 сильно отделимы, т. е. существуют такое $e_0\in E^n$ ($|e_0|=1$), $\varepsilon_0>0$, что $\langle e_0, b\rangle\leqslant\langle e_0, a_0\rangle-\varepsilon_0$ при всех $b\in\overline{\operatorname{co} B}$. Отсюда, пользуясь теоремой 1.9, имеем $\langle e_0, b\rangle\leqslant\sup\limits_{a\in\overline{\operatorname{co} A}}\langle e_0, a\rangle-\varepsilon_0=\sup\limits_{a\in A}\langle e_0, a\rangle-\varepsilon_0$ при всех $b\in\overline{\operatorname{co} B}$, так что $\sup\limits_{b\in B}\langle e_0, b\rangle=\sup\limits_{b\in\overline{\operatorname{co} B}}\langle e_0, b\rangle\leqslant\sup\limits_{a\in A}\langle e_0, a\rangle-\varepsilon_0<\sup\limits_{a\in A}\langle e_0, a\rangle$. Пришли к противоречию с условием теоремы. Следовательно, $\overline{\operatorname{co} A}\subset\overline{\operatorname{co} B}$. Если A, B выпуклы и замкнуты, то в силу теорем 1.5, 1.6 $A=\operatorname{co} A=\overline{A}=\overline{\operatorname{co} A}$, $\overline{\operatorname{co} B}=B$, и поэтому $A\subset B$. Справедливость последнего утверждения теоремы следует из того, что равенство $\delta(e, A)=\delta(e, B)$ эквивалентно двум неравенствам $\delta(e, A)\leqslant\delta(e, B)$, $\delta(e, B)\leqslant\delta(e, A)$ $(e\in E^n)$.

3. Теорему 2 можно истолковать как необходимое условие пустоты пересечения двух выпуклых множеств A и B; если $A\cap B=\varnothing$, A и B выпуклы (тогда $\operatorname{ri} A\cap\operatorname{ri} B=\varnothing$), то необходимо существуют вектор $c\in E^n$ $(c\neq 0)$ и число γ такие, что $\langle c, a\rangle\geqslant\gamma$ при всех $a\in A$ и $\langle c, b\rangle\leqslant\gamma$ при всех $b\in B$. Положим $c_1=c$, $c_2=-c$, $\gamma_1=\gamma$, $\gamma_2=-\gamma$. Тогда приведенное необходимое условие пустоты пересечения двух выпуклых множеств A и B может быть записано в следующей симметричной форме:

$$\langle c_1, u\rangle\geqslant\gamma_1 \qquad \forall u\in A,$$
$$\langle c_2, u\rangle\geqslant\gamma_2 \qquad \forall u\in B,$$

$$c_1+c_2=0, \quad \gamma_1+\gamma_2=0,$$

где хотя бы один из векторов c_1 или c_2 не равен нулю.

Следующая теорема обобщает это утверждение и дает необходимое условие пустоты пересечения любого конечного числа выпуклых множеств [52].

Теорема 7. *Пусть непустые множества $A_0, A_1, \ldots, A_m$ из E^n выпуклы и $A_0\cap A_1\cap\ldots\cap A_m=\varnothing$. Тогда необходимо существуют векторы $c_0, c_1, \ldots, c_m\in E^n$, не все равные нулю, и числа $\gamma_0, \gamma_1, \ldots, \gamma_m$ такие, что*

$$\langle c_i, u\rangle\geqslant\gamma_i \qquad \forall u\in A_i, \qquad i=0, 1, \ldots, m. \tag{7}$$

$$c_0+c_1+\ldots+c_m=0, \tag{8}$$

$$\gamma_0+\gamma_1+\ldots+\gamma_m=0. \tag{9}$$

Для доказательства этой теоремы нам понадобится прямое (декартово) произведение конечного числа множеств, а также прямое произведение евклидовых пространств. Напомним соответствующие определения.

Определение 3. Пусть $A_1, \ldots, A_m$ — какие-либо множества. Множество A, состоящее из всевозможных упорядоченных наборов (точек) $a=(a_1, \ldots, a_m)$, где $a_i=A_i$ $(i=1, \ldots, m)$, называется *прямым произведением множеств* $A_1, \ldots, A_m$ и обозначается через $A_1\times\ldots\times A_m=A$.

Пусть $L^{n_1}, \ldots, L^{n_m}$ — вещественные линейные пространства. Положим $L = L^{n_1} \times \ldots \times L^{n_m}$. Для элементов (точек) $a = (a_1, \ldots, a_m)$, $b = (b_1, \ldots, b_m) \in L$ определим сумму $a + b = (a_1 + b_1, \ldots, a_m + b_m)$ и произведение на вещественное число $\alpha a = (\alpha a_1, \ldots, \alpha a_m)$, где под $a_i + b_i$ и αa_i понимаются соответствующие операции в L^{n_i} $(i = 1, \ldots, m)$. В результате получим вещественное линейное пространство L, называемое *прямым произведением линейных пространств* $L^{n_1}, \ldots, L^{n_m}$. Если $L^{n_i} = E^{n_i}$ — евклидовы пространства размерности n_i $(i = 1, \ldots, m)$, то в прямом произведении $E = E^{n_1} \times \ldots \times E^{n_m}$ также можно ввести скалярное произведение $\langle a, b\rangle = \langle a_1, b_1\rangle + \ldots + \langle a_m, b_m\rangle$ и норму $|a| = \langle a, a\rangle^{1/2} = (|a_1|^2 + \ldots + |a_m|^2)^{1/2}$, где $\langle a_i, b_i\rangle$ и $|a_i|$ — соответственно скалярное произведение и норма в E^{n_i}. Полученное евклидово пространство E называют *прямым произведением евклидовых пространств* $E^{n_1}, \ldots, E^{n_m}$; размерность пространства E равна $n_1 + \ldots + n_m$. Например, само евклидово пространство E^n является прямым произведением n одномерных евклидовых пространств: $E^n = E^1 \times \ldots \times E^1$. Параллелепипед $\{u = (u^1, \ldots, u^n)\colon \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\}$ представляет собой прямое произведение отрезков $[\alpha_i, \beta_i]$ $(i = 1, \ldots, n)$.

Прямое произведение выпуклых множеств, очевидно, само выпукло.

Доказательство теоремы 7. Пусть $E = E^n \times \ldots \times E^n$ — прямое произведение m n-мерных евклидовых пространств. Тогда $A = A_1 \times \ldots \times A_m \in E$. Введем в E «диагональное» множество $B = \{b = (b_1, \ldots, b_m)\colon b_1 = \ldots = b_m = a_0,\ a_0 \in A_0\}$. Нетрудно видеть, что пересечение $\bigcap\limits_{i=0}^{m} A_i$ пусто тогда и только тогда, когда $A \cap B$ пусто. Далее, A и B — выпуклые множества. По теореме 2 множества A и B отделимы, т. е. существует $c = \{c_1, \ldots, c_m\} \in E$, не все c_i равны нулю и $\langle c, a\rangle \geqslant \langle c, b\rangle$ при всех $a \in A$, $b \in B$, или

$$\sum_{i=1}^{m} \langle c_i, a_i\rangle \geqslant \sum_{i=1}^{m} \langle c_i, b_i\rangle = \left\langle \sum_{i=1}^{m} c_i, a_0 \right\rangle \qquad \forall a_i \in A_i, \quad i = 0, 1, \ldots, m. \tag{10}$$

Положим $c_0 = -(c_1 + \ldots + c_m)$, так что равенство (8) будет выполнено. Тогда неравенство (10) принимает вид

$$\sum_{i=0}^{m} \langle c_i, a_i\rangle \geqslant 0 \qquad \forall a_i \in A_i, \quad i = 0, 1, \ldots, m. \tag{11}$$

Если в этом неравенстве зафиксируем какие-либо $a_i = \bar{a}_i \in A_i$ при всех $i = 0, 1, \ldots, m$, кроме $i = k$, то получим $\langle c_k, a_k\rangle \geqslant -\sum\limits_{i \neq k} \langle c_i, \bar{a}_i\rangle = \mathrm{const}$ для всех $a_k \in A_k$. Следовательно,

$$\langle c_k, u\rangle \geqslant \gamma_k = \inf_{a \in A_k} \langle c_k, a\rangle > -\infty \qquad \forall u \in A_k, \quad k = 1, \ldots, m. \tag{12}$$

Положим

$$\gamma_0 = -(\gamma_1 + \ldots + \gamma_m). \tag{13}$$

Тогда, переходя в (11) к нижней грани по всем $a_i \in A_i$ $(i = 1, \ldots, m)$, получаем $\langle c_0, a_0\rangle + \sum\limits_{i=1}^{m} \gamma_i = \langle c_0, a_0\rangle - \gamma_0 \geqslant 0$ для каждого $a_0 \in A_0$, или

$$\langle c_0, u\rangle \geqslant \gamma_0 \qquad \forall u \in A_0.$$

Все соотношения (7)—(9) получены.

При некоторых дополнительных ограничениях на множества $A_0, A_1, \ldots, A_m$ теорема 7 обратима. А именно, верна

Теорема 8. *Пусть $A_0, A_1, \ldots, A_m$ — непустые выпуклые множества из E^n, пусть все эти множества, кроме, быть может, одного, открыты. Тогда для того чтобы $A_0 \cap A_1 \cap \ldots \cap A_m = \varnothing$, необходимо и достаточно, чтобы существовали векторы $c_0, c_1, \ldots, c_m \in E^n$, не все равные нулю, и числа $\gamma_0, \gamma_1, \ldots, \gamma_m$, для которых выполнены соотношения* (7)—(9).

Доказательство. Необходимость доказана в теореме 7. Достаточность докажем, рассуждая от противного. Допустим, что условия (7)—(9) выполнены, но тем не менее существует точка $v \in \bigcap_{i=0}^{m} A_i$. Поскольку не все c_i равны нулю, то из (8) вытекает существование по крайней мере двух векторов c_i, c_j $(i \neq j)$, отличных от нуля. По условию все множества $A_0, A_1, \ldots, A_m$, кроме, быть может, одного, открыты. Поэтому можем считать $c_i \neq 0$, A_i — открытое множество, т. е. $A_i = \operatorname{int} A_i$.

Согласно условию (7) $\langle c_i, u\rangle \geqslant \gamma_i$ при всех $u \in A_i$. В силу теоремы 4 тогда $\langle c_i, u\rangle > \gamma_i$ для всех $u \in A_i = \operatorname{int} A_i$. В частности, для точки $v \in \bigcap_{j=0}^{m} A_j \subset A_i$ также имеем $\langle c_i, v\rangle > \gamma_i$. Кроме того, для всех остальных номеров $j \neq i$ также $v \in A_j$ и в силу (7) $\langle c_j, v\rangle \geqslant \gamma_j$. Сложим все эти неравенства. С учетом равенства (9) получим $\langle c_0, v\rangle + \langle c_1, v\rangle + \ldots + \langle c_m, v\rangle > \gamma_0 + \gamma_1 + \ldots + \gamma_m = 0$, т. е. $\langle c_0 + c_1 + \ldots + c_m, v\rangle > 0$. Однако это невозможно в силу равенства (8). Полученное противоречие показывает, что $\bigcap_{i=0}^{m} A_i = \varnothing$.

Приведенное выше доказательство теоремы 7 принадлежит В. И. Плотникову. Оно привлекает своей простотой и тем, что позволяет убедиться в справедливости теоремы 7 и в бесконечномерных гильбертовых (и более общих) пространствах — ее доказательство при этом остается неизменным, нужно лишь уточнить ссылки на соответствующие теоремы отделимости в бесконечномерных пространствах.

4. С помощью теорем 7, 8 можно получить условия совместности или несовместности систем неравенств [321, 324]. Приведем некоторые из них.

Лемма 1. *Пусть $A = \{u \in E^n\colon \langle e, u\rangle < \mu\}$ — открытое полупространство, $\bar{A} = \{u \in E^n\colon \langle e, u\rangle \leqslant \mu\}$ — замыкание A; здесь $e \in E^n$ $(e \neq 0)$, $\mu \in \mathbf{R}$. Тогда для того чтобы линейная функция $\langle c, u\rangle$ была ограничена снизу на $\bar{A}$ (или A), т. е.*

$$\langle c, u\rangle \geqslant \gamma > -\infty \qquad \forall u \in \bar{A} \ \ (\text{или } \forall u \in A), \tag{14}$$

необходимо и достаточно, чтобы существовало такое число $\lambda \geqslant 0$, что

$$c = -\lambda e, \qquad \gamma \leqslant -\lambda\mu. \tag{15}$$

Доказательство. В силу теоремы 1.9 ограниченность функции $\langle c, u\rangle$ на A следует из ее ограниченности на $\bar{A}$, и наоборот. Поэтому лемму достаточно доказать для множества $\bar{A}$.

Необходимость. Пусть выполнено (14). Если $c = 0$, то из (14) следует, что $\gamma \leqslant 0$ и в (15) можно взять $\lambda = 0$. Пусть $c \neq 0$. Возьмем какую-либо точку $u_0 \in \bar{A}$, $\langle e, u_0\rangle = \mu$. Прямая $u_t = u_0 + t\left(\frac{\langle c, e\rangle}{|e|^2}e - c\right)$ $(t \in \mathbf{R})$ принадлежит $\bar{A}$, так как $\langle e, u_t\rangle = \langle e, u_0\rangle + t\frac{\langle c, e\rangle}{|e|^2}\langle e, e\rangle - t\langle c, e\rangle = \langle e, u_0\rangle = \mu$ $(t \in \mathbf{R})$. В силу (14) $\langle c, u_t\rangle = \langle c, u_0\rangle + t\left(\frac{\langle c, e\rangle^2}{|e|^2} - |c|^2\right) \geqslant \gamma$ при всех $t \in \mathbf{R}$. Разделим это неравенство на $t > 0$ и перейдем к пределу при $t \to \infty$. Получим $|c|^2 \leqslant \langle c, e\rangle^2/|e|^2$. С другой стороны, в силу неравенст-

ва Коши — Буняковского $\langle c, e\rangle^2 \leqslant |c|^2|e|^2$. Отсюда и из предыдущего неравенства следует равенство $|\langle c, e\rangle| = |c|\cdot|e|$. Однако при $c \neq 0$, $e \neq 0$ в неравенстве Коши — Буняковского равенство возможно лишь тогда, когда векторы c, e коллинеарны, т. е. $c = \alpha e$ $(\alpha \neq 0)$. Покажем, что $\alpha < 0$.

Возьмем луч $v_t = u_0 - te$ $(t \geqslant 0)$. Поскольку $\langle e, v_t\rangle = \langle e, u_0\rangle - t|e|^2 = \mu - t|e|^2 \leqslant \mu$ при всех $t \geqslant 0$, то луч принадлежит $\bar{A}$. Согласно (14) тогда $\gamma \leqslant \langle c, v_t\rangle = \langle \alpha e, v_t\rangle = \alpha\mu - \alpha t|e|^2$ $(t \geqslant 0)$. Разделим это неравенство на $t > 0$ и перейдем к пределу при $t \to \infty$. Получим $0 \leqslant -\alpha|e|^2$, что возможно только при $\alpha < 0$. Положим $\lambda = -\alpha$, так что $c = -\lambda e$ $(\lambda > 0)$. Тогда $\langle c, u\rangle = -\lambda\langle e, u\rangle \geqslant -\lambda\mu$ при всех $u \in \bar{A}$, причем при $u = u_0$ здесь достигается равенство. Следовательно, $\inf\limits_{u\in\bar{A}} \langle c, u\rangle = -\lambda\mu$. Переходя в (14) к нижней грани при $u \in \bar{A}$, получаем $-\lambda\mu = \inf\limits_{u\in\bar{A}} \langle c, u\rangle \geqslant \gamma$, т. е. $\gamma \leqslant -\lambda\mu$. Соотношения (15) получены.

Достаточность. Пусть выполнены условия (15). Тогда для любых $u \in \bar{A}$ имеем $\langle c, u\rangle = -\lambda\langle e, u\rangle \geqslant -\lambda\mu \geqslant \gamma$, т. е. выполняется неравенство (14).

Теорема 9. *Пусть заданы открытые или замкнутые полупространства* $A_i = \{u \in E^n\colon \langle e_i, u\rangle < \mu_i\}$ *или* $A_i = \{u \in E^n\colon \langle e_i, u\rangle \leqslant \mu_i\}$ $(i = 0, 1, \ldots, m)$; *пусть* $A_0 \cap A_1 \cap \ldots \cap A_m = \varnothing$. *Тогда необходимо существуют такие числа* $\lambda_0, \lambda_1, \ldots, \lambda_m$, *что*

$$\lambda_0 \geqslant 0, \quad \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0, \quad \sum_{i=0}^{m} \lambda_i > 0, \quad \sum_{i=0}^{m} \lambda_i e_i = 0, \quad \sum_{i=0}^{m} \lambda_i\mu_i \leqslant 0. \tag{16}$$

Доказательство. Поскольку пересечение выпуклых множеств $A_0, A_1, \ldots, A_m$ пусто, то согласно теореме 7 существуют векторы $c_0, c_1, \ldots, c_m$, не все равные нулю, и числа $\gamma_0, \gamma_1, \ldots, \gamma_m$, для которых справедливы соотношения (7)—(9). Согласно лемме 1 условия (7) могут выполняться тогда и только тогда, когда $c_i = -\lambda_i e_i$, $\gamma_i \leqslant -\lambda_i\mu_i$ при некоторых $\lambda_i \geqslant 0$ $(i = 0, 1, \ldots, m)$. Поскольку $e_i \neq 0$ и не все c_i равны нулю, то не все λ_i равны нулю. Далее, из (8) следует, что $\sum\limits_{i=0}^{m} \lambda_i e_i = 0$, а из (9) имеем $-\sum\limits_{i=0}^{m} \lambda_i\mu_i \geqslant \sum\limits_{i=0}^{m} \gamma_i = 0$. Все соотношения (16) получены.

Теорема 10. *Пусть* $A_i = \{u \in E^n\colon \langle e_i, u\rangle < \mu_i\}$ $(i = 1, \ldots, m)$, *а* $A_0 = \langle u \in E^n\colon \langle e_0, u\rangle < \mu_0\}$ *или* $A_0 = \{u \in E^n\colon \langle e_0, u\rangle \leqslant \mu_0\}$. *Тогда для того чтобы* $A = \bigcap\limits_{i=0}^{m} A_i = \varnothing$, *необходимо и достаточно существования таких чисел* $\lambda_0, \lambda_1, \ldots, \lambda_m$, *которые удовлетворяют соотношениям* (16).

Доказательство. Необходимость доказана в теореме 9. Достаточность докажем, как и в аналогичной теореме 8, рассуждая от противного. Пусть выполнены (16), но пусть тем не менее существует точка $v \in A$. Поскольку не все λ_i равны нулю, $e_i \neq 0$ $(i = 0, 1, \ldots, m)$, то из условия $\sum\limits_{i=0}^{m} \lambda_i e_i = 0$ следует существование по крайней мере двух чисел $\lambda_i, \lambda_j > 0$ $(i \neq j)$. Тогда либо $i \neq 0$, либо $j \neq 0$. Для определенности пусть $i > 0$. Из условия $v \in A$ тогда следует, что $v \in A_i$, т. е. $\langle e_i, v\rangle < \mu_i$. Для остальных $j \neq i$ имеем $\langle e_j, v\rangle \leqslant \mu_j$ (впрочем, равенство здесь возможно лишь при $j = 0$). Умножая эти неравенства на соответствующие $\lambda_j \geqslant 0$ и суммируя по $j = 0, 1, \ldots, m$, получаем $\sum\limits_{i=0}^{m} \lambda_i \langle e_i, v\rangle = \left\langle \sum\limits_{i=0}^{m} \lambda_i e_i, v \right\rangle < \sum\limits_{i=0}^{m} \lambda_i\mu_i \leqslant 0$,

что противоречит равенству $\sum_{i=0}^{m} \lambda_i e_i = 0$. Следовательно, $A = \varnothing$, что и требовалось доказать.

Нетрудно видеть, что в теоремах 9, 10 говорится об условиях несовместности систем линейных неравенств вида $\langle e_i, u\rangle \leqslant \mu_i$ или $\langle e_i, u\rangle < \mu_i$. Например, из теоремы 10 следует, что для несовместности систем неравенств

$$\langle e_i, u\rangle < \mu_i, \quad e_i \neq 0, \quad i = 0, 1, \ldots, m \tag{17}$$

(в (17) одно из неравенств может быть нестрогим), необходимо и достаточно выполнения соотношений (16).

Опираясь на теорему 10 и рассуждая от противного, нетрудно доказать следующий критерий совместности системы (17).

Теорема 11. *Для того чтобы система неравенств* (17) *была совместной* (*или, иначе, пересечение множеств* $A_0, A_1, \ldots, A_m$ *из теоремы* 9 *было непустым*), *необходимо и достаточно, чтобы для любых чисел* $\lambda_0 \geqslant 0$, $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0$, *не все из которых равны нулю, из равенства* $\sum_{i=0}^{m} \lambda_i e_i = 0$ *следовало неравенство* $\sum_{i=0}^{m} \lambda_i \mu_i > 0$.

5. Переформулируем теоремы 7, 8 для случая, когда $A_0, A_1, \ldots, A_m$ являются выпуклыми конусами в E^n.

Определение 4. *Конусом* (с вершиной в нуле) называется множество K, содержащее вместе с любой своей точкой u и точки λu при всех $\lambda > 0$. Если множество K выпукло, то K называют *выпуклым конусом*, если K замкнуто — *замкнутым конусом*, если K открыто — *открытым конусом*.

Рассмотрим множество

$$K^* = \{c \in E^n\colon \langle c, u\rangle \geqslant 0 \quad \forall u \in K\}. \tag{18}$$

Это множество всегда непусто, так как $0 \in K^*$. Далее, если $c \in K^*$, то для λc при любом $\lambda > 0$ имеем $\langle \lambda c, u\rangle = \lambda\langle c, u\rangle \geqslant 0$ для всех $u \in K$, т. е. $\lambda c \in K^*$. Следовательно, K^* — конус.

Определение 5. Конус K^*, определенный посредством (18), называется *двойственным* (*сопряженным*) *конусом* к конусу K (рис. 4.17).

Например, если $K = \{u \in E^n\colon \langle a, u\rangle = 0\}$ — гиперплоскость, то $K^* = \{c \in E^n\colon c = \lambda a, \lambda \in \mathbf{R}\}$; если $K = \{u \in E^n\colon \langle a, u\rangle \leqslant 0\}$ — замкнутое полупространство или $K = \{u \in E^n\colon \langle a, u\rangle < 0\}$ — открытое полупространство, то $K^* = \{c \in E^n\colon c = -\lambda a, \lambda \geqslant 0\}$; если $K = E^n$, то $K^* = \{0\}$; если $K = \{0\}$, то $K^* = E^n$; если $K = \{u \in E^n\colon u \geqslant 0\}$, то $K^* = \{c \in E^n\colon c \geqslant 0\}$.

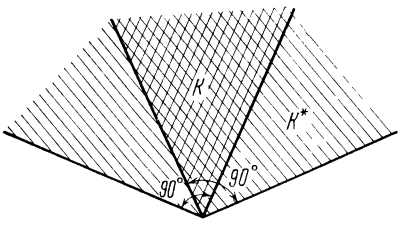

Рис. 4.17

С помощью двойственных конусов удобно переформулировать теорему 7 для случая, когда множества $A_0, A_1, \ldots, A_m$ являются конусами.

Теорема 12. *Пусть* $K_0, K_1, \ldots, K_m$ — *непустые выпуклые конусы из* E^n (*с вершиной в нуле*), *пусть* $K_0 \cap K_1 \cap \ldots \cap K_m = \varnothing$. *Тогда необходимо существуют векторы* $c_0, c_1, \ldots, c_m$, *не все равные нулю*, $c_i \in K_i^*$ $(i = 0, 1, \ldots, m)$, *и такие, что*

$$c_0 + c_1 + \ldots + c_m = 0. \tag{19}$$

Доказательство. Согласно теореме 7 существуют векторы c_0, $c_1, \ldots, c_m$, не все равные нулю, и числа $\gamma_0, \gamma_1, \ldots, \gamma_m$, удовлетворяющие условиям (7)—(9). Воспользуемся тем, что рассматриваемые множества K_0, $K_1, \ldots, K_m$ являются именно конусами, и покажем, что тогда $\gamma_0 = \gamma_1 = \ldots = \gamma_m = 0$. В самом деле, если $\langle c_i, u\rangle \geqslant \gamma_i$ при всех $u \in K_i$, то $\langle c_i, \lambda u\rangle \geqslant \gamma_i$ или $\langle c_i, u\rangle \geqslant \gamma_i/\lambda$ для любых $\lambda > 0$ и $u \in K_i$. Отсюда при $\lambda \to +0$ получим $\langle c_i, u\rangle \geqslant 0$ при всех $u \in K_i$, т. е. $c_i \in K_i^*$ $(i = 0, 1, \ldots, m)$.

Кроме того, если $u \in K_i$, то, взяв в неравенстве $\langle c_i, u\rangle \geqslant 0$ вместо u точку λu при малых $\lambda > 0$, получим сколь угодно малые значения функции $\langle c_i, u\rangle$ на K_i и придем к равенству $\inf\limits_{u \in K_i} \langle c_i, u\rangle = 0$. Согласно (12) это означает, что все величины γ_i $(i = 1, \ldots, m)$, участвующие в неравенствах (7), равны нулю. Из (13) тогда имеем $\gamma_0 = 0$. Таким образом, если в теореме 7 множества $A_0, A_1, \ldots, A_m$ являются выпуклыми конусами, то условие (9) выполняется тривиально, так как все $\gamma_i = 0$ $(i = 0, 1, \ldots, m)$, условия (7) означают, что $c_i \in K_i^*$ $(i = 0, 1, \ldots, m)$, а из (8) следует (19).

При некоторых дополнительных ограничениях на конусы K_0, $K_1, \ldots, K_m$ теорема 12 обратима. А именно верна

Теорема 13. *Пусть $K_0, K_1, \ldots, K_m$ — непустые выпуклые конусы из E^n (с вершиной в нуле), пусть все эти конусы, кроме, быть может, одного, открыты. Тогда для того чтобы $K_0 \cap K_1 \cap \ldots \cap K_m = \varnothing$, необходимо и достаточно, чтобы существовали векторы $c_0, c_1, \ldots, c_m$, не все равные нулю, $c_i \in K_i^*$ $(i = 0, 1, \ldots, m)$, и такие, что*

$$c_0 + c_1 + \ldots + c_m = 0. \tag{19}$$

Доказательство. Необходимость доказана в теореме 12. Достаточность вытекает из теоремы 8, если заметить, что условие $c_i \in K_i^*$ равносильно неравенству $\langle c_i, u\rangle \geqslant 0$ $(u \in K_i)$, и отсюда и из (19) следуют условия (7)—(9) при $\gamma_0 = \gamma_1 = \ldots = \gamma_m = 0$.

Упражнения. 1. Пусть A и B — выпуклые множества, не имеющие общих внутренних точек. Можно ли утверждать, что A и B отделимы? Рассмотреть пример $A = \{u = (x, y)\colon y = 0, |x| \leqslant 1\}$, $B = \{u = (x, y)\colon x = 0, |y| \leqslant 1\}$ в E^2.

2. Пусть X — выпуклое множество из E^n, $\operatorname{int} X = \varnothing$. Доказать, что любая гиперплоскость, опорная к X и проходящая через точку $y \in \operatorname{ri} X$, содержит X, т. е. не является собственно опорной.

3. Пусть A — выпуклое множество из E^n, причем $A \cap \operatorname{int} E^n_+ = \varnothing$. Доказать, что существует такой вектор $c = (c_1, \ldots, c_n) \neq 0$ $(c_1 \geqslant 0, \ldots, c_n \geqslant 0)$, что $\langle c, a\rangle \leqslant 0$ при всех $a \in A$.

4. Пусть A — выпуклое множество из E^n, M — аффинное или многогранное множество из E^n. Для того чтобы A и M были собственно отделимы, необходимо и достаточно, чтобы $M \cap \operatorname{ri} A = \varnothing$. Доказать.

5. Пусть $\rho(A, B) = \inf\limits_{a \in A} \inf\limits_{b \in B} |a - b|$ — расстояние между множествами A и B. Доказать, что два непустых выпуклых множества A, B из E^n сильно отделимы тогда и только тогда, когда $\rho(A, B) > 0$.

6. Доказать, что всякое выпуклое замкнутое ограниченное множество из E^n имеет хотя бы одну угловую точку (см. определение 3.2.1).

7. Пусть U — выпуклое замкнутое множество из E^n. Доказать, что U имеет хотя бы одну угловую точку тогда и только тогда, когда U не содержит прямых (см. пример 1.4).

8. Доказать, что выпуклое замкнутое ограниченное множество A из E^n является выпуклой оболочкой своих угловых точек. Показать, что без требования ограниченности множества A это утверждение неверно. Рассмотреть пример $A = \{a = (x, y) \in E^2\colon y \geqslant |x|\}$.

9. Если $A_1, \ldots, A_m$ — выпуклые множества из E^n, причем $\bigcap_{i=1}^{m} \operatorname{ri} A_i \neq \varnothing$, то

$$\overline{\bigcap_{i=1}^{m} A_i} = \bigcap_{i=1}^{m} \overline{A}_i; \quad \operatorname{ri}\left(\bigcap_{i=1}^{m} A_i\right) = \bigcap_{i=1}^{m} (\operatorname{ri} A_i); \quad \operatorname{aff}\left(\bigcap_{i=1}^{m} A_i\right) = \bigcap_{i=1}^{m} (\operatorname{aff} A_i).$$

Доказать.

10. Пусть K — произвольный конус из E^n. Доказать, что конус K^* — замкнутый и выпуклый.

11. Доказать, что если K — выпуклый конус, то $K^* = (\overline{K})^*$.

12. Доказать, что если K — замкнутый выпуклый конус, то $(K^*)^* = K$.

13. Пусть K — выпуклый конус. Доказать, что для ограниченности снизу линейной функции $\langle c, u\rangle$ на K необходимо и достаточно, чтобы $c \in K^*$.

14. Пусть K — выпуклый конус, $\operatorname{int} K \neq \varnothing$. Тогда $\langle c, u\rangle > 0$ для всех $u \in \operatorname{int} K$ при любом выборе $c \in K^*$ $(c \neq 0)$. Доказать.

15. Доказать, что $(K + \ldots + K_m)^* = K_1^* \cap \ldots \cap K_m^*$, где $K_1, \ldots, K_m$ — конусы из E^n.

16. Пусть K_1, K_2 — выпуклые замкнутые конусы. Доказать, что $(K_1 \cap K_2)^* = \overline{K_1^* + K_2^*}$.

17. Пусть $K_0, K_1, \ldots, K_m$ — выпуклые конусы, пусть $K_0 \cap \operatorname{int} K_1 \cap \ldots \cap \operatorname{int} K_m \neq \varnothing$. Тогда $\left(\bigcap_{i=0}^{m} K_i\right)^* = K_0^* + K_1^* + \ldots + K_m^*$. Доказать.

18. Пусть $K_0, K_1, \ldots, K_m$ — выпуклые конусы. Тогда либо $\left(\bigcap_{i=0}^{m} K_i\right)^* = K_0^* + K_1^* + \ldots + K_m^*$, либо существуют не все равные нулю векторы $c_i \in K_i^*$ $(i = 0, 1, \ldots, m)$ такие, что $c_0 + c_1 + \ldots + c_m = 0$. Доказать.

19. Для того чтобы выпуклые конусы K_0, K_1 были неотделимы, необходимо и достаточно, чтобы $0 \in \operatorname{int}(K_0 - K_1)$. Доказать.

20. Доказать, что два выпуклых конуса K_0, K_1 неотделимы тогда и только тогда, когда одновременно выполнены два условия: $\operatorname{ri} K_0 \cap \operatorname{ri} K_1 \neq \varnothing$, $\operatorname{Lin} K_0 + \operatorname{Lin} K_2 = E^n$.

21. Множество $A^* = \{c \in E^n\colon \langle c, u\rangle \leqslant 1 \quad \forall u \in A\}$ называется *полярой* множества A. Найти поляры множеств A, если $A = \{0\}$; $A = [a, b] \subset E^1$; $A = \{u \in E^n\colon u = te,\ 0 < t < \infty,\ e \neq 0\}$; $A = \{u \in E^n\colon \langle c, u\rangle \leqslant \gamma\}$; A — шар; A — конус с вершиной в нуле. Выяснить связь между полярой конуса и двойственным конусом.

§ 6. Субградиент. Субдифференциал

1. Для выпуклых дифференцируемых функций на выпуклом множестве выше было доказано неравенство (см. теорему 2.2)

$$J(u) \geqslant J(v) + \langle J'(v),\ u - v\rangle \qquad \forall u \in U. \tag{1}$$

К сожалению, выпуклая функция может не быть дифференцируемой даже на внутренних точках множества, и в этом случае полезное во многих случаях неравенство (1) не будет иметь смысла. Тем не менее, оказывается, для выпуклых функций это неравенство можно сохранить, если надлежащим образом обобщить понятие градиента.

Определение 1. Пусть функция $J(u)$ определена на множестве U из E^n. Вектор $c = c(v) \in E^n$ называется *субградиентом* функции $J(u)$ в точке $v \in U$, если

$$J(u) \geqslant J(v) + \langle c(v),\ u - v\rangle \qquad \forall u \in U. \tag{2}$$

Множество всех субградиентов функции $J(u)$ в точке v называют *субдифференциалом* этой функции в точке v и обозначают через $\partial J(v)$.

Неравенство (2) имеет простой геометрический смысл и означает, что график функции $\gamma = J(u)$ $(u \in U)$ в пространстве переменных (u, γ) лежит не ниже графика линейной функции $\gamma = J(v) + \langle c(v), u - v\rangle$ $(u \in U)$, причем в точке $u = v$ оба графика пересекаются (рис. 4.18).

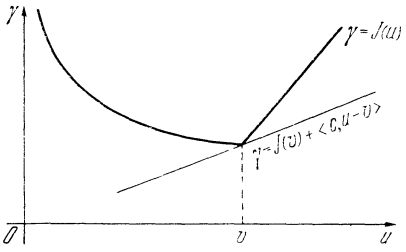

Рис. 4.18

Для гладких выпуклых функций, как показывает неравенство (1), субдифференциал непуст и градиенты этих функций являются их субградиентами. Во внутренних точках множества гладкая функция, оказывается, других субградиентов, кроме градиента, иметь не может. В самом деле, пусть $v \in \operatorname{int} U$, $c(v) \in \partial J(v)$. Поскольку $J(u) \in C^1(U)$, то $J(u) = J(v) + \langle J'(v), u - v\rangle + o(|u - v|)$ $(u \in U)$. Отсюда и из (2) следует, что $\langle J'(v) - c(v), u - v\rangle \geqslant o(|u - v|)$ $(u \in U)$. Поскольку $v \in \operatorname{int} U$, то $u = v - \varepsilon(J'(v) - c(v)) \in U$ при всех ε $(0 < \varepsilon < \varepsilon_0)$. Подставив эту точку в предыдущее неравенство, получим $-\varepsilon|J'(v) - c(v)|^2 \geqslant o(\varepsilon)$ $(0 < \varepsilon < \varepsilon_0)$. Деля на $\varepsilon > 0$ и устремляя $\varepsilon \to +0$, отсюда будем иметь $-|J'(v) - c(v)|^2 \geqslant 0$, т. е. $c(v) = J'(v)$.

Тем самым показано, что для гладкой выпуклой функции $\partial J(v) = \{J'(v)\}$ при всех $v \in \operatorname{int} U$. Существуют функции, которые недифференцируемы в точке, но тем не менее субдифференциал в этой точке непуст.

Пример 1. Функция $J(u) = |g(u)|$ $(u \in U)$ в точке v, где $g(v) = 0$, всегда имеет субградиент $c(v) = 0$, так как $|g(u)| - |g(v)| = |g(u)| \geqslant 0 = \langle 0, u - v\rangle$ для всех $u \in U$. В то же время в точках v, где $g(v) \neq 0$, эта функция может быть недифференцируемой и не имеющей субградиента.

Пример 2. Пусть $J(u) = |u|$ $(u \in E^n)$. В точке $v = 0$ эта функция недифференцируема, но для нее верно соотношение

$$J(u) - J(0) = |u| \geqslant \langle c, u - 0\rangle = \langle c, u\rangle, \qquad u \in E^n,$$

для всех c $(|c| \leqslant 1)$. Это значит, что $\partial(|u|)|_{u=0} = \partial J(0) = \{c \in E^n\colon |c| \leqslant 1\}$ — единичный шар с центром в нуле. Если $v \neq 0$, то $\partial J(v) = \{v/|v| = J'(v)\}$.

Заметим, что в примере 2 функция $J(u) = |u|$ выпукла на E^n. Оказывается, если совсем отказаться от выпуклости функции, то даже гладкая функция может не иметь субградиента ни в одной точке. Например, для функции $J(u) = u^3$ на E^1 субдифференциал пуст во всех точках. В то же время эта функция $J(u) = u^3$ на множестве $U = \{u \in E^1\colon u \geqslant 0\}$ выпукла и во всех точках $v \in U$ имеет на U непустой субдифференциал. Ниже увидим, что это не случайно.

2. Следующая теорема показывает, что понятия субградиента и субдифференциала являются естественными для выпуклых функций.

Теорема 1. *Пусть U — открытое выпуклое множество на E^n (например, возможно, $U \equiv E^n$). Тогда для того чтобы функция $J(u)$, определенная на U, имела непустой субдифференциал во всех точках U, необходимо и достаточно, чтобы $J(u)$ была выпукла на U.*

Доказательство. Необходимость. Пусть для некоторой функции $J(u)$ субдифференциал $\partial J(u) \neq \varnothing$ при всех $u \in U$. Покажем, что $J(u)$ выпукла на U. Возьмем произвольные $u, v \in U$, $\alpha \in [0, 1]$ и положим $u_\alpha = \alpha u + (1 - \alpha)v$. Пусть $c = c(u_\alpha) \in \partial J(u_\alpha)$. Тогда

$$J(u) - J(u_\alpha) \geqslant \langle c, u - u_\alpha\rangle, \quad J(v) - J(u_\alpha) \geqslant \langle c, v - u_\alpha\rangle.$$

Умножим первое из этих неравенств на α, второе на $1-\alpha$ и сложим. Получим $\alpha J(u)+(1-\alpha)J(v)-J(u_\alpha)\geqslant\langle c,\ u_\alpha-u_\alpha\rangle=0$ при всех $u,\ v\in U$, $\alpha\in[0,\ 1]$. Выпуклость $J(u)$ на U доказана.

Достаточность. Пусть $J(u)$ выпукла на открытом выпуклом множестве U. Пусть v — произвольная точка на U. Покажем, что $\partial J(v)\neq\varnothing$. Возьмем некоторый единичный вектор e. Поскольку U — открытое множество, то $v+te\in U$ при всех t $(0\leqslant t<t_0,\ t_0>0)$. По теореме 2.14 существует производная $dJ(v)/de$ по направлению e.

В пространстве E^{n+1} переменных $(u,\ \gamma)$ введем два множества

$$A=\{(u,\ \gamma)\in E^{n+1}:\ u\in U,\ \gamma>J(u)\},$$

$$B=\left\{(u,\ \gamma)\in E^{n+1}:\ u=v+te,\ \gamma=J(v)+t\frac{dJ(v)}{de},\ 0\leqslant t<t_0\right\}.$$

Нетрудно показать, что множество A выпукло — это делается так же, как доказывалась выпуклость надграфика выпуклой функции в теореме 2.9 (кстати, в данном случае $A=\operatorname{int}(\operatorname{epi}J)$). Множество B является отрезком прямой в E^{n+1} и тоже выпукло. Покажем, что множества A и B не имеют общих точек.

В самом деле, пусть $(u,\ \gamma)\in A$. Имеются две возможности: 1) $u\neq v+te$ при всех t $(0\leqslant t<t_0)$ — тогда заведомо $(u,\ \gamma)\notin B$; 2) при некотором t $(0\leqslant t<t_0)$ оказалось, что $u=v+te$ — тогда с учетом неравенства (1.11.5) и $\gamma>J(u)=J(v+te)$ имеем $\gamma-J(v)>J(v+te)-J(v)\geqslant tdJ(v)/de$, т. е. $\gamma>J(v)+tdJ(v)/de$, и снова $(u,\ \gamma)\notin B$.

Итак, множества A и B выпуклы, $A\cap B=\varnothing$. По теореме 5.2 тогда существует гиперплоскость с нормальным вектором $(d,\ \nu)\neq 0$, отделяющая A и B, т. е.

$$\langle d,\ u\rangle+\nu\gamma\geqslant\langle d,\ v+te\rangle+\nu\left(J(v)+t\frac{dJ(v)}{de}\right)\tag{3}$$

при всех $\gamma>J(u)$, $u\in U$, $0\leqslant t<t_0$. В частности, при $u=v$, $t=0$ из (3) имеем $\nu(\gamma-J(v))\geqslant 0$ для всех $\gamma>J(v)$. Отсюда следует, что $\nu\geqslant 0$.

Допустим, что $\nu=0$. Тогда из (3) имеем $\langle d,\ u\rangle\geqslant\langle d,\ v+te\rangle$ для всех $u\in U$, $0\leqslant t<t_0$. Положим здесь $u=v+\varepsilon d$, $t=0$ — так можно делать, ибо $v+\varepsilon d\in U$ при всех ε $(0<|\varepsilon|<\varepsilon_0)$ в силу открытости U. Получим $\langle d,\ v+\varepsilon d\rangle\geqslant\langle d,\ v\rangle$, или $\varepsilon|d|^2\geqslant 0$ при всех ε $(0<|\varepsilon|\leqslant\varepsilon_0)$, что возможно только при $d=0$. Однако $(d,\ \nu)\neq 0$ по построению. Полученное противоречие показывает, что $\nu=0$ не может быть.

Итак, $\nu>0$. Поделим (3) на $\nu>0$. Обозначая $c=-d/\nu$ и устремляя $\gamma\to J(u)+0$, из (3) получаем

$$J(u)-J(v)+t\langle c,\ e\rangle\geqslant\langle c,\ u-v\rangle+t\frac{dJ(v)}{de}\tag{4}$$

при всех $u\in U$ и всех t $(0\leqslant t<t_0)$. Полагая здесь $t=0$, будем иметь

$$J(u)-J(v)\geqslant\langle c,\ u+v\rangle\qquad\forall u\in U.$$

Это означает, что $c\in\partial J(v)$, т. е. $\partial J(v)\neq\varnothing$.

В следующей теореме изучаются некоторые свойства субдифференциала выпуклой функции.

Теорема 2. *Пусть U — открытое выпуклое множество из E^n (например, $U=E^n$), $J(u)$ — выпуклая функция на U. Тогда субдифференциал $\partial J(u)$ при всех $v\in U$ является непустым выпуклым замкнутым и ограниченным множеством.*

Доказательство. Непустота субдифференциала доказана в теореме 1. Покажем выпуклость $\partial J(v)$. Пусть $c_1,\ c_2\in\partial J(v)$, т. е.

$$J(u)-J(v)\geqslant\langle c_1,\ u-v\rangle,\quad J(u)-J(v)\geqslant\langle c_2,\ u-v\rangle,\quad u\in U.$$

Возьмем $\alpha \in [0, 1]$. Умножая первое неравенство на α, второе на $1-\alpha$ и, складывая, получаем $J(u) - J(v) \geqslant (\alpha c_1 + (1-\alpha)c_2, u - v\rangle$ при всех $u \in U$. Это значит, что $\alpha c_1 + (1-\alpha)c_2 \in \partial J(v)$ для любых $\alpha \in [0, 1]$. Выпуклость $\partial J(u)$ доказана.

Пусть c — предельная точка множества $\partial J(v)$, пусть $\{c_k\} \in \partial J(v)$ и $c_k \to c$ при $k \to \infty$. Из $c_k \in \partial J(v)$ следует, что $J(u) - J(v) \geqslant \langle c_k, u - v\rangle$ $(u \in U)$. При $k \to \infty$ отсюда получим $c \in \partial J(v)$. Замкнутость $\partial J(v)$ доказана.

Покажем ограниченность $\partial J(v)$. Возьмем любой вектор $c \in \partial J(v)$. Поскольку U — открытое множество, то $S(v, \varepsilon) = \{u \in E^n: |u - v| \leqslant \varepsilon\} \subset U$ при достаточно малом $\varepsilon > 0$. Далее, в силу теоремы 2.15, функция $J(u)$ непрерывна на U, поэтому $\sup\limits_{S(v,\varepsilon)} J(u) = J^*(S) < \infty$. Положим в неравенстве (2) $u = v + \varepsilon c/|c| \in S(v, \varepsilon)$. Получим $|c| \leqslant (J(v + \varepsilon c/|c|) - J(v))/\varepsilon < (J^*(S) - J(v))/\varepsilon < \infty$ при всех $c \in \partial J(v)$.

3. Теоремы 1, 2 не дают конструктивного описания субдифференциала выпуклой функции. Такое описание удается получить лишь в немногих случаях.

Пример 3. Пусть $J(u) = \max\limits_{i \in I} u^i$, $u = (u^1, \ldots, u^n) \in E^n$, $I = \{1, \ldots, n\}$. Согласно теореме 2.7 функция $J(u)$ выпукла на E^n. Покажем, что

$$\partial J(v) = \{c = (c_1, \ldots, c_n): c_i \geqslant 0,\ i \in I(v),\quad c_i = 0,\quad i \notin I(v);\quad c_1 + \ldots + c_n = 1\}, \tag{5}$$

где $I(v) = \left\{i \in I: \max\limits_{j \in I} v^j = v^i\right\}$. Множество, определяемое правой частью формулы (5), обозначим через $A(v)$. Пусть $c \in A(v)$. Умножим неравенство $J(u) - J(v) = \max\limits_{j \in I} u^j - v^i \geqslant u^i - v^i$, верное при всех $i \in I(v)$, $u \in E^n$ на $c_i \geqslant 0$ и сложим по всем $i \in I(v)$. С учетом равенств $c_i = 0$ при $i \notin I(v)$, $c_1 + \ldots + c_n = 1$ получим $J(u) - J(v) \geqslant \langle c, u - v\rangle$ $(u \in E^n)$, т. е. $c \in \partial J(v)$. Это значит, что $A(v) \subset \partial J(v)$.

Докажем включение $\partial J(v) \subset A(v)$. Пусть $c \in \partial J(v)$, т. е.

$$J(u) - J(v) = \max_{i \in I} u^i - \max_{i \in I} v^i \geqslant \langle c, u - v\rangle \qquad \forall u \in E^n. \tag{6}$$

Возьмем в (6) $u = u_\pm = (v^1 \pm 1, \ldots, v^n \pm 1)$. Тогда $J(u_\pm) - J(v) = \pm 1$ и из (6) получим $\pm 1 \geqslant \langle c, u_\pm - v\rangle = \pm \sum\limits_{i=1}^n c_i$, что возможно лишь при $\sum\limits_{i=1}^n c_i = 1$. Далее, в (6) возьмем $u_\varepsilon = (u^1, \ldots, u^n)$, где $u^i = v^i - \varepsilon$ при некотором $i \in I$, $u^j = v^j$ при $j \neq i$, $\varepsilon > 0$. Тогда $J(u_\varepsilon) \leqslant J(v)$ и из (6) получим $0 \geqslant \langle c, u_\varepsilon - v\rangle = c_i(-\varepsilon)$, так что $c_i \geqslant 0$ $(i = 1, \ldots, n)$.

Далее, пусть $i \notin I(v)$. Тогда $v^i < J(v)$ и можно выбрать $\varepsilon > 0$ так, что $v^i + \varepsilon < J(v)$. Положим $u_\varepsilon = (u^1, \ldots, u^n)$, где $u^i = v^i + \varepsilon$, $u^j = v^j$ при $j \neq i$. Тогда $J(v) = J(u_\varepsilon)$, и из (6) получим $0 \geqslant \langle c, u_\varepsilon - v\rangle = \varepsilon c_i$, т. е. $c_i \leqslant 0$ $(i \notin I(v))$. Сравнивая это неравенство с уже доказанным $c_i \geqslant 0$, получаем $c_i = 0$ $(i \notin I(v))$. Это значит, что $c \in A(v)$, так что $\partial J(v) \subset A(v)$. Равенство (5) установлено.

4. Установим связь между производными по направлению и субдифференциалом выпуклой функции.

Теорема 3. *Пусть U — открытое выпуклое множество из E^n, $J(u)$ — выпуклая функция на U. Тогда во всех точках $v \in U$ производная функции $J(u)$ по любому направлению e $(|e| = 1)$ существует, причем*

$$\frac{dJ(v)}{de} = \max_{c \in \partial J(v)} \langle c, e\rangle. \tag{7}$$

Доказательство. Существование производной $dJ(v)/de$ установлено в теореме 2.14. Докажем формулу (7). Из (2) имеем $(J(v+te)-J(v))/t \geqslant \langle c, e\rangle$ при всех $c \in \partial J(v)$ и всех достаточно малых $t > 0$. Отсюда при $t \to +0$ получим $dJ(v)/de \geqslant \langle c, e\rangle$ для любого $c \in \partial J(v)$, так что $dJ(v)/de \geqslant \sup\limits_{c \in \partial J(v)} \langle c, e\rangle$. С другой стороны, при доказательстве теоремы 1 был построен специальный субградиент c, для которого выполняется неравенство (4). Полагая в (4) $u = v$, будем иметь $\partial J(v)/\partial e \leqslant \langle c, e\rangle$ $(c \in \partial J(v))$. Сравнивая это неравенство с предыдущим, приходим к формуле (7). Попутно показали, что максимум в правой части (7) достигается именно на том субградиенте, который был построен в теореме 1.

Формула (7) обобщает известную формулу $dJ(v)/de = \langle J'(v), e\rangle$ для гладких функций.

5. С помощью субдифференциала можно сформулировать критерий оптимальности, обобщающий теорему 2.3 на случай негладких выпуклых функций.

Теорема 4. *Пусть $J(u)$ — выпуклая функция на открытом выпуклом множестве W из E^n, U — выпуклое подмножество множества W. Тогда для того чтобы функция $J(u)$ достигала своей нижней грани на множестве U в точке $u_* \in U$, необходимо и достаточно, чтобы существовал субградиент $c_* = c(u_*) \in \partial J(u_*)$ такой, что*

$$\langle c_*, u - u_*\rangle \geqslant 0 \qquad \forall u \in U. \tag{8}$$

Если $u_* \in \operatorname{int} U$, то в (8) $c_* = 0$.

Необходимость. Пусть $u_* \in U_* = \left\{u \in U\colon J(u) = \inf\limits_U J(u) = J_* > -\infty\right\}$. В пространстве E^{n+1} введем множества

$$A = \{a = (u, a_0) \in E^{n+1}\colon u \in W,\ a_0 \geqslant J(u) - J_*\},$$

$$B = \{b = (v, b_0)\colon v \in U,\ b_0 < 0\}.$$

Эти множества не имеют общих точек.

В самом деле, пусть $a = (u, a_0) \in A$. Тогда возможно либо $u \in U$, либо $u \in W \setminus U$. В первом случае $a_0 \geqslant J(u) - J_* \geqslant 0$ и заведомо $a \notin B$. Если $u \in W \setminus U$, то $a \notin B$ по определению множества B. Выпуклость множеств A и B доказывается так же, как и выпуклость надграфика выпуклой функции в теореме 2.9. По теореме 5.2 тогда существует гиперплоскость с нормальным вектором $g = (d, \nu) \neq 0$, отделяющая A и $\overline{B}$ — замыкание B, т. е. $\langle g, a\rangle \leqslant \gamma \leqslant \langle g, b\rangle$, $a \in A$, $b \in B$. Поскольку $y = (u_*, 0) \in A \cap \overline{B}$ и согласно теореме 5.2 $\gamma = \langle g, y\rangle = \langle d, u_*\rangle$, то предыдущие неравенства запишутся в виде

$$\begin{gathered}\langle d, u\rangle + \nu a_0 \leqslant \langle d, u_*\rangle \leqslant \langle d, v\rangle + \nu b_0 \\ \forall u \in W,\quad a_0 \geqslant J(u) - J_*,\quad v \in U,\quad b_0 \leqslant 0.\end{gathered} \tag{9}$$

Из правого неравенства (9) при $b = (u_*, -1) \in B$ имеем $\nu \cdot (-1) \geqslant 0$, т. е. $\nu \leqslant 0$. Если $\nu = 0$, то из (9) получим $\langle d, u\rangle \leqslant \langle d, u_*\rangle$, $u \in W$. Однако W — открытое множество и $u = u_* + \varepsilon d \in W$ при всех ε $(0 < |\varepsilon| < \varepsilon_0)$. Поэтому из предыдущего неравенства получаем $\langle d, \varepsilon d\rangle = \varepsilon |d|^2 \leqslant 0$ $(0 < |\varepsilon| < \varepsilon_0)$, что возможно лишь при $d = 0$. Однако это противоречит тому, что $g = (d, \nu) \neq 0$. Следовательно, $\nu < 0$.

Разделим (9) на $\nu < 0$. Полагая $c_* = -d/\nu$, получаем

$$\begin{gathered}-\langle c_*, u - u_*\rangle + a_0 \geqslant 0 \geqslant -\langle c_*, u - u_*\rangle + b_0, \\ \forall u \in W,\quad a_0 \geqslant J(u) - J_*,\quad v \in U,\quad b_0 \leqslant 0.\end{gathered} \tag{10}$$

Отсюда при $a = (u, a_0 = J(u) - J(u_*))$ имеем $J(u) - J(u_*) \geqslant \langle c_*, u - u_* \rangle$ при всех $u \in W$, т. е. $c_* \in \partial J(u_*)$. При $b = (v, 0)$ $(v \in U)$ из (10) следует неравенство (8).

Если $u_* \in \operatorname{int} U$, то $u = u_* + \varepsilon c_* \in U$ при всех ε $(0 < |\varepsilon| < \varepsilon_0)$, и из (8) получим $\langle c_*, \varepsilon c_* \rangle = \varepsilon |c_*|^2 \geqslant 0$ $(0 < |\varepsilon| < \varepsilon_0)$. Это возможно лишь при $c_* = 0$.

Достаточность. Пусть для некоторой точки $u_* \in U$ выполнено неравенство (8) при каком-либо $c_* \in \partial J(u_*)$. По определению субградиента тогда $J(u) - J(u_*) \geqslant \langle c_*, u - u_* \rangle \geqslant 0$ при всех $u \in U$, т. е. $u_* \in U_*$.

Замечание. Как видно из доказательства теоремы 4, множества A, B и, следовательно, вектор $g = (d, \nu)$, а также и $c_* = -d/\nu$ из (8) не зависят от выбора $u_* \in U_*$. Таким образом, в (8) для всех $u_* \in U_*$ можно выбрать один и тот же субградиент c_*, который, конечно, является общим для всех $\partial J(u_*)$, $u_* \in U_*$. Это значит, что, в частности, когда $J(u)$ — гладкая выпуклая функция, в теореме 2.3 $c_* = J'(u_*)$ не зависит от $u_* \in U_*$.

Следующие примеры показывают, что субградиент c_* из (8) в общем случае определяется неоднозначно.

Пример 4. Пусть $J(u) = |u|$ $(u \in W = E^1)$. Если $U = E^1$, то $U_* = \{0\}$, $\partial J(0) = [-1, 1]$ и неравенство (8) выполняется лишь при $c_* = 0$. Если $U = \langle u \in E^1 \colon u \geqslant 0 \}$, то $U_* = \{0\}$ и (8) выполняется для всех $c_* \in [0, 1] \subset \partial J(0)$.

Пример 5. Пусть $J(u) = \max\{u; 0\}$ $(u \in W = E^1)$. Если $U = E^1$, то $U_* = \{0\}$, $\partial J(0) = [0, 1]$ и (8) имеет место лишь для $c_* = 0$. Если $U = \{u \in E^1 \colon u \geqslant 0\}$, то по-прежнему $U_* = \{0\}$, но неравенство (8) здесь выполняется для всех $c_* \in \partial J(0) = [0, 1]$.

6. Определение 2. Пусть E^n, E^m — евклидовы пространства, $W \subset E^n$, $\Pi(E^m)$ — множество всех непустых множеств из E^m. Говорят, что на W задано *многозначное отображение* $F \colon W \to \Pi(E^m)$, если каждой точке $u \in W$ поставлено в соответствие некоторое множество $F(u) \subset \Pi(E^m)$.

Определение 3. Многозначное отображение, которое каждой точке u из открытого выпуклого множества $W \subset E^n$ ставит в соответствие субдифференциал $\partial J(u)$ некоторой выпуклой на W функции $J(u)$, называется *субдифференциальным* отображением и обозначается через ∂J.

Субдифференциальное отображение обладает рядом замечательных свойств [18, 21, 132, 255, 264]; на некоторых из них мы здесь кратко остановимся.

Определение 4. Пусть W — множество из E^n. Многозначное отображение $F \colon W \to \Pi(E^m)$ называется:

1) *компактным*, если для любого компактного множества $U \subset W$ множество $F(U) = \bigcup\limits_{u \in U} F(u)$ компактно;

2) *монотонным*, если $\langle c(u) - c(v), u - v \rangle \geqslant 0$ при всех $u, v \in W$, $c(u) \in F(u)$, $c(v) \in F(v)$;

3) *выпуклозначным*, если $F(u)$ — выпуклое множество при каждом $u \in W$;

4) *полунепрерывным сверху* в точке $v \in W$, если из того, что $\{v_k\} \to v$ $(v_k \in W)$ и $\{c_k\} \to c$ $(c_k \in F(v_k)$, $k = 1, 2, \ldots)$, следует $c \in F(v)$;

5) *полунепрерывным снизу* в точке $v \in W$, если для всякого элемента $c \in F(v)$ и любой последовательности $\{v_k\} \to v$ $(v_k \in W)$ найдутся $c_k \in F(v_k)$ такие, что $\{c_k\} \to c$.

Теорема 5. *Пусть $J(u)$ — выпуклая функция на открытом выпуклом множестве W из E^n. Тогда субдифференциальное отображение $\partial J \colon W \to \Pi(E^n)$ выпуклозначно, монотонно, полунепрерывно сверху, компактно.*

Доказательство. Выпуклозначность отображения ∂J следует из теоремы 2. Возьмем произвольные $u, v \in W$, $c(u) \in \partial J(u)$, $c(v) \in \partial J(v)$. Тогда согласно (2) $J(u) - J(v) \geqslant \langle c(v), u - v \rangle$, $J(v) - J(u) \geqslant \langle c(u), v - u \rangle$. Сложив эти два неравенства, получим $\langle c(u) - c(v), u - v \rangle \geqslant 0$. Монотон-

ность ∂J установлена. Далее, пусть $v \in W$, $\{v_k\} \to v$, $v_k \in W$, пусть $\{c_k\} \to c$, $c_k \in \partial J(v_k)$. Это значит, что $J(u) - J(v_k) \geqslant \langle c_k, u - v_k \rangle$ при всех $u \in W$. Поскольку функция $J(u)$ непрерывна на W (см. теорему 2.15), то, переходя к пределу в этом неравенстве при $k \to \infty$, приходим неравенству (2). Это значит, что $c \in \partial J(v)$. Полунепрерывность сверху отображения ∂J доказана.

Наконец, возьмем произвольное ограниченное замкнутое множество $U \subset W$. Поскольку W — открытое множество, то все точки U являются внутренними для W и поэтому найдется такое число $\delta > 0$, что ограниченное замкнутое множество $U_\delta = \{u \in E^n\colon |u - v| \leqslant \delta,\ v \in U\}$, представляющее собой δ-раздутие множества U, принадлежит W.

В самом деле, если $W = E^n$, то $U_\delta \subset E^n$ при любом $\delta > 0$. Если же $W \neq E^n$, то граница $\operatorname{Гр} W$ выпуклого множества W непуста и $\rho(v, \operatorname{Гр} W) = \inf\limits_{w \in \operatorname{Гр} W} |v - w| > 0$ при всех $v \in U$. В силу леммы 2.1.2 функция $\rho(v, \operatorname{Гр} W)$ непрерывна на компактном множестве U и согласно теореме 2.1.1 найдется такая точка $v_* \in U$, что $\inf\limits_{v \in U} \rho(v, \operatorname{Гр} W) = \rho(v_*, \operatorname{Гр} W) = 2\delta > 0$. Это значит, что $U_\delta \subset W$. Функция $J(u)$ непрерывна на компактном множестве U_δ, поэтому $\sup\limits_{U_\delta} J(u) = J_\delta^* < \infty$.

Возьмем любые $v \in U$, $c \in \partial J(v)$. В неравенстве (2) положим $u = v + \delta c/|c| \subset U_\delta \subset W$. Получим $|c| \leqslant [J(v + \delta c/|c|) - J(v)]/\delta \leqslant 2J_\delta^*/\delta = M < \infty$ для всех $c \in \partial J(v)$, $v \in U$. Таким образом, $\sup\limits_{v \in U} \sup\limits_{c \in \partial J(v)} |c| = \sup\limits_{c \in \partial J(U)} |c| \leqslant M < \infty$, т. е. множество $\partial J(U)$ ограничено. Докажем замкнутость $\partial J(U)$. Пусть $\{c_k\} \to c$ $(c_k \in \partial J(U))$. Это значит, что существует такая точка $v_k \in U$, что $c_k \in \partial J(v_k)$. Поскольку U — компактное множество, то без ограничения общности можем считать, что $\{v_k\} \to v \subset U$. В силу полунепрерывности сверху отображения ∂J тогда $c \in \partial J(v) \subset \partial J(U)$. Следовательно, $\partial J(V)$ — замкнутое множество. Компактность отображения ∂J установлена.

Отметим, что субдифференциальное отображение ∂J, вообще говоря, не является полунепрерывным снизу. Например, функция $J(u) = |u|$, $u \in W = E^1$, имеет субдифференциал $\partial J(0) = [-1, 1]$, но точки $c \in (-1, 1) \in \partial J(0)$ не могут быть получены как предел какой-либо последовательности $\{c_k\}$, $c_k \in \partial J(v_k)$, где $v_k \neq 0$, $k = 1, 2, \ldots$, $\{v_k\} \to 0$.

Используя компактность отображения ∂J, нетрудно получить обобщение теоремы 1.8.3 на многомерный случай.

Теорема 6. *Пусть $J(u)$ — выпуклая функция на открытом выпуклом множестве W из E^n. Тогда на любом ограниченном множестве $G \subset W$ функция $J(u)$ удовлетворяет условию Липшица, т. е. существует такая постоянная $L = L(G) \geqslant 0$, что $|J(u) - J(v)| \leqslant L|u - v|$, u, $v \in G$.*

Доказательство. Возьмем $U = \operatorname{co} \bar{G}$ — это выпуклая оболочка замыкания множества G. В силу теоремы 1.8 U — выпуклое компактное множество $U \subset W$. Тогда $J(u) - J(v) \geqslant \langle c(v), u - v \rangle \geqslant -L|u - v|$ $(u, v \in U)$, где $L = \sup\limits_{\partial J(U)} |c| < \infty$ в силу теоремы 5. Поменяв здесь u, v местами, имеем $J(v) - J(u) \geqslant -L|u - v|$ $(u, v \in U)$. Отсюда следует утверждение теоремы.

Для получения интересных экстремальных свойств субдифференциального отображения нам понадобится

Теорема 7. *Пусть W — открытое множество из E^n, многозначные отображения A и B: $W \to \Pi(E^n)$ таковы, что A полунепрерывно сверху и компактно, B монотонно, причем $A(u) \cap B(u) \neq \varnothing$ при всех $u \in W$. Тогда $\operatorname{co}(B(u) \subset \operatorname{co} A(u)$ $(u \in W)$.*

Доказательство. Зафиксируем любые $u \in W$ и $e \in E^n$ $(|e| = 1)$. Поскольку W — открытое множество, то $S = \{v \in E^n\colon |v - u| \leqslant \varepsilon_0\} \subset W$ при некотором $\varepsilon_0 > 0$. Возьмем последовательность $\{\varepsilon_k\} \to 0$ $(0 < \varepsilon_k < \varepsilon_0)$.

Тогда $v_k = u + \varepsilon_k e \subset S \subset W$, $\{v_k\} \to u$. По условию существуют $a_k \in A(v_k) \cap \cap B(v_k)$. В силу монотонности B для всех $b \in B(u)$ имеем $\langle a_k - b, v_k - u\rangle = \langle a_k - b, \varepsilon_k e\rangle \geqslant 0$ или $\langle a_k - b, e\rangle \geqslant 0$. Поскольку отображение A компактно, то множество $A(S)$ является компактным. Поэтому, учитывая включение $a_k \in A(v_k) \subset A(S)$, можем считать, что $\{a_k\} \to a_0$. В силу полунепрерывности сверху отображения A тогда $a_0 \in A(u)$. Поэтому, переходя к пределу в неравенстве $\langle a_k - b, e\rangle \geqslant 0$, получаем $\langle a_0 - b, e\rangle \geqslant 0$, или $\langle e, b\rangle \leqslant \langle e, a_0\rangle \leqslant \sup\limits_{a \in A(u)} \langle e, a\rangle$ при всех $b \in B(u)$. Отсюда $\sup\limits_{b \in B(u)} \langle e, b\rangle \leqslant \sup\limits_{a \in A(u)} \langle e, a\rangle$, и требуемое утверждение следует из теоремы 4.6.

Теорема 8. *Пусть $J(u)$ — выпуклая функция на открытом выпуклом множестве W из E^n, пусть $F\colon W \to \Pi(E^n)$ — какое-либо многозначное отображение. Тогда:*

а) *если F монотонно, $\partial J(u) \in F(u)$ при всех $u \in W$, то $\partial J(u) = F(u)$ $(u \in W)$, т. е. субдифференциальное отображение максимально в классе монотонных отображений;*

б) *если F полунепрерывно сверху, выпуклозначно и $F(u) \subset \partial J(u)$ при всех $u \in W$, то $\partial J(u) = F(u)$, т. е. субдифференциальное отображение минимально в классе полунепрерывных сверху выпуклозначных отображений.*

Доказательство. В случае а), пользуясь теоремой 7 при $A(u) = \partial J(u)$, $B(u) = F(u)$, получаем $F(u) \subset \overline{\mathrm{co}\, F(u)} \subset \overline{\mathrm{co}\, \partial J(u)} = \partial J(u)$, так что $\partial J(u) = F(u)$ $(u \in W)$. В случае б) в теореме 7 возьмем $A(u) = F(u)$, $B(u) = \partial J(u)$. Нужно проверить компактность отображения F. Пусть U — какой-либо компакт из W. Из включения $F(u) \subset \partial J(u)$ $(u \in W)$ следует $F(U) \subset \partial J(U)$. В силу теоремы 5 множество $\partial J(U)$ компактно. Это значит, что его подмножество $F(U)$ ограничено.

Далее, пусть $\{c_k\} \to c$ $(c_k \in F(U))$. Тогда найдутся такие точки $u_k \in U$, что $c_k \in F(u_k)$ $(k = 1, 2, \ldots)$. В силу компактности U можем считать, что $\{u_k\} \to u \in U$. Из полунепрерывности сверху отображения F следует, что $c \in F(u) \subset F(U)$, т. е. $F(U)$ замкнуто. Компактность F установлена. В частности, взяв здесь одноточечный компакт $U = \{u\}$, заключим, что $F(u)$ — компактное, т. е. ограниченное и замкнутое множество при каждом $u \in W$. Отсюда и из теорем 1.5, 1.8 с учетом выпуклости $F(u)$ имеем равенство $\overline{\mathrm{co}\, F(u)} = F(u)$. Из теоремы 7 теперь получаем $\overline{\mathrm{co}\, \partial J(u)} = \partial J(u) \subset \overline{\mathrm{co}\, F(u)} = F(u)$ $(u \in U)$. Отсюда и из включения $F(u) \subset \partial J(u)$ $(u \in W)$ следует утверждение б) теоремы.

7. Субдифференциал для выпуклых функций играет роль, аналогичную той, какую играет градиент для дифференцируемых функций. Как для работы с градиентами полезно иметь некоторый набор правил дифференцирования, так и для работы с субдифференциалами нужно иметь некоторые правила субдифференцирования. Предлагаем читателю самостоятельно доказать следующие правила субдифференцирования 1—4:

1. Если $g(u) = J(u + u_0)$, то $\partial g(u) = \partial J(u + u_0)$.
2. Если $g(u) = \lambda J(u)$ $(\lambda > 0)$, то $\partial g(u) = \lambda \partial J(u)$.
3. Если $g(u) = J(\lambda u)$, то $\partial g(u) = \lambda \partial J(\lambda u)$.
4. Если функция $J(u)$ выпукла на E^n, а A — матрица порядка $n \times n$, $\det A \neq 0$, $b \in E^n$, то функция

$$g(u) = J(Au + b), \qquad u \in E^n,$$

выпукла на E^n, причем

$$\partial g(u) = A^T \partial J(v)\,|_{v = Au + b}.$$

5. Справедлива следующая теорема [21], обобщающая известную теорему о производной сложной функции.

Теорема 9. *Пусть $J_1(u), \ldots, J_m(u)$ — выпуклые функции, определенные на открытом выпуклом множестве W из E^n, функция $\varphi(x) = \varphi(x^1, \ldots, x^m)$ — выпуклая функция на открытом выпуклом множестве X из E^m причем $J(u) = (J_1(u), \ldots, J_m(u)) \in X$ при всех $u \in W$, $\varphi(x)$ моно-*

тонно возрастает на X, т. е. $\varphi(x) \geqslant \varphi(y)$ для всех $x = (x^1, \ldots, x^m)$, $y = (y^1, \ldots, y^m) \in X$, $x^i \geqslant y^i$, $i = 1, \ldots, m$. Тогда функция $\Phi(u) = \varphi(J(u))$ выпукла на W и ее субдифференциал имеет вид

$$\partial\Phi(u) = \bigcup_{p=(p_1,\ldots,p_m)\in\partial\varphi(J(u))} \left\{\sum_{i=1}^{m} p_i \partial J_i(u)\right\}, \quad u \in W. \tag{11}$$

Для доказательства этой теоремы нам понадобится

Лемма 1. *Пусть $A_1, \ldots, A_m$ — выпуклые множества из E^n, P — выпуклое множество из E^n_+, тогда множество $A = \bigcup_{p=(p_1,\ldots,p_m)\in P} \left\{\sum_{i=1}^{m} p_i A_i\right\}$ выпукло.*

Доказательство. Возьмем произвольные $c_1, c_2 \in A$, $\alpha \in (0, 1)$. По определению A существуют такие $p_i = (p_{i1}, \ldots, p_{im}) \in P \subset E^n_+$, $a_{ij} \in A_j$ $(j = 1, \ldots, m)$, что $c_i = \sum_{j=1}^{m} p_{ij} a_{ij}$ $(i = 1, 2)$. Тогда $\alpha c_1 + (1-\alpha) c_2 = \sum_{j=1}^{m} (\alpha p_{1j} a_{1j} + (1-\alpha) p_{2j} a_{2j})$. По условию $p_{ij} \geqslant 0$. Обозначим через I множество всех номеров $j = 1, \ldots, m$, для которых $p_{1j} > 0$ или $p_{2j} > 0$. Тогда $\alpha p_{1j} + (1-\alpha) p_{2j} > 0$, $\gamma_{\alpha j} = \dfrac{\alpha p_{1j}}{\alpha p_{1j} + (1-\alpha) p_{2j}} \in (0, 1)$, $j \in I$.

Положим $a_j^\alpha = \gamma_{\alpha j} a_{1j} + (1 - \gamma_{\alpha j}) a_{2j}$ при $j \in I$, $a_j^\alpha = a_{1j}$ при $j \notin I$. В силу выпуклости A_j точки a_j^α принадлежит A_j $(j = 1, \ldots, m)$. Кроме того, $p^\alpha = (p_1^\alpha, \ldots, p_m^\alpha) = \alpha p_1 + (1-\alpha) p_2 \in P$ из-за выпуклости P, причем здесь $p_j^\alpha = 0$ при $j \notin I$. Тогда $\alpha c_1 + (1-\alpha) c_2 = \sum_{j\in I} (\alpha p_{1j} a_{1j} + (1-\alpha) p_{2j} a_{2j}) = \sum_{j\in I} (\alpha p_{1j} + (1-\alpha) p_{2j}) (\gamma_{\alpha j} a_{1j} + (1-\gamma_{\alpha j}) a_{2j}) = \sum_{j\in I} p_j^\alpha a_j^\alpha = \sum_{j=1}^{m} p_j^\alpha a_j^\alpha$, где $a_j^\alpha \in A_j$ $(j = 1, \ldots, m)$, $p^\alpha \in P$. Это значит, что $\alpha c_1 + (1-\alpha) c_2 \in A$ при всех $\alpha \in (0, 1)$, т. е. A — выпуклое множество.

Доказательство теоремы 9. Из выпуклости функций $J_i(u)$, $\varphi(x)$ и монотонности $\varphi(x)$ следует выпуклость сложной функции $\Phi(u) = \varphi(J(u))$ на открытом выпуклом множестве W — это доказывается так же, как и теорема 2.8. Согласно теореме 2 тогда субдифференциал $\partial\Phi(u)$ при каждом $u \in W$ представляет собой непустое выпуклое компактное множество.

Докажем формулу (11). Обозначим $F(u) = \bigcup_{p\in\partial\varphi(J(u))} \left\{\sum_{i=1}^{m} p_i \partial J_i(u)\right\}$. По теореме 2 субдифференциалы $\partial J_i(u)$, $\partial\varphi(x)$ также непусты, выпуклы, компактны и поэтому $F(u) \neq \varnothing$ $(u \in W)$. Отметим, что $\partial\varphi(x) \in E^m_+$ при всех $x \in X$. В самом деле, возьмем любые $x = (x^1, \ldots, x^m) \in X$, $p = (p_1 \ldots, p_m) \in \partial\varphi(x)$. Поскольку множество X открыто, то при достаточно малом $\varepsilon > 0$ точка $y = (y^1, \ldots, y^m)$, где $y^i = x^i - \varepsilon$, $y^j = x^j$ при $j \neq i$, принадлежит X. С учетом монотонности $\varphi(x)$ тогда $0 \geqslant \varphi(y) - \varphi(x) \geqslant \langle p, y - x\rangle = p_i(-\varepsilon)$, так что $p_i \geqslant 0$ $(i = 1, \ldots, m)$. Следовательно, $\partial\varphi(x) \in E^m_+$. По лемме 1 тогда множество $F(u)$ выпукло при каждом $u \in W$.

Покажем, что $F = F(u)$, как многозначное отображение $W \to \Pi(E^n)$, полунепрерывно сверху. Пусть $u \in W$, $\{u_k\} \to u$, $\{c_k\} \to c$, $c_k \in F(u_k)$. Тогда

найдутся $p_k \in \partial\varphi(J(u_k))$, $c_{ik} \in \partial J_i(u_k)$ такие, что $c_k = \sum_{i=1}^{m} p_{ik}c_{ik}$. Поскольку сходящаяся последовательность $\{u_k\}$ ограничена, то найдется компактное множество $G \subset U$, содержащее все точки $u, u_1, u_2, \ldots$ Аналогично, поскольку в силу непрерывности выпуклых функций $J_i(u)$ последовательность $\{x_k = J(u_k)\} \to J(u) \in X$, то существует компактное множество $Y \subset X$, содержащее все точки $J(u), J(u_1), J(u_2), \ldots$ (можно взять $Y = J(G) = J_1(G) \times \ldots \times J_m(G)$). По теореме 5 множества $\partial J_i(G)$, $\partial\varphi(Y)$ компактны. Поскольку $c_{ik} \in \partial J_i(u_k) \subset \partial J_i(G)$, $p_k \in \partial\varphi(J(u_k)) \in \partial\varphi(Y)$ $(k = 1, 2, \ldots)$, то не теряя общности можем считать, что $\{c_{ik}\} \to c_i$, $\{p_k\} \to p$. Из полунепрерывности сверху отображений $\partial J_i(u)$, $\partial\varphi(x)$ имеем $c_i \in \partial J_i(u)$, $p \in \partial\varphi(J(u))$. Переходя к пределу в равенстве $c_k = \sum_{i=1}^{m} p_{ik}c_{ik}$, получаем $c = \sum_{i=1}^{m} p_i c_i$, т. е. $c \in F(u)$. Это значит, что отображение F полунепрерывно сверху.

Возьмем любые $u \in W$ и $c \in F(u)$. Тогда $c = \sum_{i=1}^{m} p_i c_i$ при некоторых $c_i \in \partial J_i(u)$, $p = (p_1, \ldots, p_m) \in \partial\varphi(J(u))$. Учитывая определение субградиента и неотрицательность p_i, получим

$$\Phi(v) - \Phi(u) = \varphi(J(v)) - \varphi(J(u)) \geqslant \langle p, J(v) - J(u)\rangle =$$

$$= \sum_{i=1}^{m} p_i (J_i(v) - J_i(u)) \geqslant \sum_{i=1}^{m} p_i \langle c_i, v - u\rangle =$$

$$= \left\langle \sum_{i=1}^{m} p_i c_i, v - u \right\rangle = \langle c, v - u\rangle \quad \forall v \in W.$$

Это значит, что $c \in \partial\Phi(u)$ и, следовательно, $F(u) \subset \partial\Phi(u)$ при всех $u \in W$. Отсюда, пользуясь утверждением б) теоремы 8, заключаем, что $\partial\Phi(u) = F(u)$ $(u \in W)$. Формула (11) доказана.

С помощью теоремы 9 можно получить более сложные правила субдифференцирования, дополняющие приведенные выше правила 1—4. Ниже при ссылках на формулу (11) предполагается, что выполнены условия теоремы 9.

6. Если $\varphi(x)$ — дифференцируемая функция, то $\partial\varphi(x) = \{\varphi'(x)\} = \{(\partial\varphi/\partial x^1, \ldots, \partial\varphi/\partial x^m)\}$, $\partial\varphi(J(u)) = \{\varphi'(J(u))\}$, и из формулы (11) имеем

$$\partial\Phi(u) = \sum_{i=1}^{m} \frac{\partial\varphi\ (J(u))}{\partial x^i}\, \partial J_i(u), \qquad u \in W.$$

В частности, если $J_i(u)$ дифференцируема и $\partial J_i(u) = \{J_i'(u)\}$, отсюда получаем классическое правило дифференцирования сложной функции.

7. Если $\varphi(x) = \sum_{i=1}^{m} \alpha_i x^i$ $(\alpha_i \geqslant 0)$, то $\partial\varphi(x) = \{(\alpha_1, \ldots, \alpha_m)\} \in E_+^m$ и для функции $\Phi(u) = \sum_{i=1}^{m} \alpha_i J_i(u)$ $(u \in W)$ из (11) имеем $\partial\Phi(u) = \sum_{i=1}^{m} \alpha_i \partial J_i(u)$ $(u \in W)$.

8. Если $\varphi(x) = \max\limits_{1 \leqslant i \leqslant m} x^i$, то согласно формуле (5) $\partial\varphi(x) = \{p = (p_1, \ldots, p_m)\colon p_i \geqslant 0,\ i \in I(x);\ p_i = 0,\ i \notin I(x),\ p_1 + \ldots + p_m = 1\}$,

где $I(x) = \left\{i: 1 \leqslant i \leqslant m, \max\limits_{1 \leqslant j \leqslant m} x^j = x^i\right\} (x \in E^m)$. Отсюда и из (11) для функции $\Phi(u) = \max\limits_{1 \leqslant i \leqslant m} J_i(u)$ имеем

$$\partial\Phi(u) = \left\{c: c = \sum_{i \in I(J(u))} p_i c_i,\ c_i \in \partial J_i(u),\ p_i \geqslant 0,\ i \in I(J(u)),\right.$$
$$\left.\sum_{i \in I(J(u))} p_i = 1\right\} = \operatorname{co}\left(\bigcup_{i \in I(J(u))} \partial J_i(u)\right), \quad I(J(u)) =$$
$$= \left\{i: 1 \leqslant i \leqslant m, \max_{1 \leqslant j \leqslant m} J_j(u) = J_i(u)\right\}, \quad u \in W. \quad (12)$$

9. Если $\varphi(x) = \max\{0; x\}$ $(x \in E^1)$, то согласно (12) $\partial\varphi(0) = \{c: c = p_1 \cdot 0 + p_2 \cdot 1 = p_2,\ p_1 + p_2 = 1,\ p_1 \geqslant 0,\ p_2 \geqslant 0\} = [0, 1]$, $\partial\varphi(x) = \{1\}$ при $x > 0$, $\partial\varphi(x) = \{0\}$ при $x < 0$, и для функции $\Phi(u) = \max\{0; J(u)\}$ $(u \in W)$ из (11) имеем $\partial\Phi(u) = p\partial J(u)$ $(0 \leqslant p \leqslant 1)$ при $J(u) = 0$, $\partial\Phi(u) = \partial J(u)$ при $J(u) > 0$, $\partial\Phi(u) = 0$ при $J(u) < 0$.

10. Если $\varphi(x) = (\max\{0; x\})^p$ $(p > 1,\ x \in E^1)$, то $\partial\varphi(x) = \{\varphi'(x) = p(\max\{0; x\})^{p-1}\}$ и для функции $\Phi(u) = (\max\{0; J(u)\})^p$ $(u \in W)$ имеем

$$\partial\Phi(u) = p(\max\{0; J(u)\})^{p-1}\partial J(u), \quad u \in W, \quad p > 1.$$

11. Приведем еще одну теорему, в которой дается обобщение формулы (12).

Теорема 10. *Пусть A — компактное множество из E^m, W — открытое выпуклое множество из E^n, функция $G(u, a)$ определена на $W \times A$, полунепрерывна сверху по a при каждом $u \in W$, выпукла по переменной $u \in W$ при каждом $a \in A$. Тогда функция $\Phi(u) = \max\limits_{a \in A} G(u, a)$ выпукла на W и ее субдифференциал имеет вид*

$$\partial\Phi(u) = \operatorname{co}\left(\overline{\bigcup_{a \in R(u)} \partial G(u, a)}\right), \quad R(u) = \{a: a \in A,\ G(u, a) = \Phi(u)\}, \quad u \in W. \quad (13)$$

Доказательство может быть проведено по той же схеме, как и теорема 9, и представляется читателю.

12. Если A — выпуклое замкнутое ограниченное множество из E^n, то функция $\Phi(u) = \max\limits_{a \in A} \langle a, u\rangle$ $(u \in E^n)$ выпукла на E^n, причем, как следует из (13), при $G(u, a) = \langle a, u\rangle$ имеем $\partial\Phi(u) = \{a: a \in A, \langle a, u\rangle = \Phi(u)\}$.

Более подробно о перечисленных и других правилах субдифференцирования, о различных свойствах субдифференциала, о различных обобщениях понятий субградиента и субдифференциала, о применении этих понятий для исследования и приближенного решения экстремальных задач см., например [2, 15, 18, 21, 27, 123, 132, 133, 148, 156, 166, 195, 214, 219, 235, 255, 264, 306, 334].

Упражнения. 1. Найти субдифференциалы функций:

а) $J(u) = |u - 1|$ $(u \in E^1)$;
б) $J(u) = |u - 1| + |u + 1|$ $(u \in E^1)$;
в) $J(u) = |x + y| + |x - y|$ $(u = (x, y) \in E^2)$;
г) $J(u) = \max\{u^2, u + 2\}$ $(u \in E^1)$;
д) $J(u) = \max\{|u|; |u - 1|\}$ $(u \in E^1)$;
е) $J(u) = |\langle a, u\rangle - b|$ $(u \in E^n)$.

2. Пусть функции $J_1(u), \ldots, J_m(u)$ $(u \in E^n)$ непрерывно дифференцируемы в некоторой окрестности точки v. Доказать, что тогда функция $J(u) = \max\limits_{1 \leqslant i \leqslant m} J_i(u)$ в точке v имеет производные по любому направ-

лению e ($|e| = 1$), причем

$$\frac{dJ(v)}{de} = \max_{i \in I(v)} \langle J_i'(v), e \rangle, \quad I(v) = \{i\colon 1 \leqslant i \leqslant m,\ J_i(v) = J(v)\}.$$

Установить связь между этой формулой и формулами (7), (12).

3. Найти субдифференциалы функций $J(u) = \max\limits_{|t| \leqslant 1} |t^2 + xt + y|$, $J(u) = \max\limits_{|t| \leqslant 1} |xt^2 + yt|$, $J(u) = \max\limits_{0 \leqslant t \leqslant 1} |x + ty|$ $(u = (x, y) \in E^2)$.

4. Пусть A — замкнутое ограниченное множество из E^m, функция $g(u, a)$ непрерывна по совокупности переменных (u, a) на $E^n \times A$ вместе с производной $\partial g(u, a)/\partial u$. Доказать, что тогда функция $J(u) = \max\limits_{a \in A} g(u, a)$ во всех точках $v \in E^n$ имеет производную по любому направлению e, $|e| = 1$, причем

$$\frac{dJ(v)}{de} = \max_{a \in A_0(v)} \left\langle \frac{\partial g(v, a)}{\partial u}, e \right\rangle, \quad A_0(v) = \{a\colon a \in A,\ g(v, a) = J(v)\}.$$

Установить связь между этой формулой и формулами (7), (13).

5. Пусть $J(u)$ — выпуклая функция одной переменной на отрезке $[a, b]$. Доказать, что $\partial J(u) = [J'(u-0),\ J'(u+0)]$ при всех $u \in (a, b)$, где $J'(u-0)$, $J'(u+0)$ — левая и правая производные в точке u. Показать, что в точках $u = a$ или $u = b$ субдифференциал может быть пустым (рассмотреть пример $J(u) = -\sqrt{1-u^2}$ $(|u| \leqslant 1)$).

6. Пусть $J(u)$ — выпуклая функция на выпуклом множестве U из E^n. Доказать, что при всех $u \in \operatorname{ri} U$ множество $\partial J(u)$ непусто, выпукло, компактно, причем $\dfrac{dJ(u)}{de} = \max\limits_{c \in \partial J(u)} \langle c, e \rangle$ для всех $e \in \operatorname{Lin} U$.

7. Пусть функция $J(u)$ определена на открытом выпуклом множестве $W \subset E^n$ и такова, что функция $\Phi(u) = |J(u)|$ выпукла на W. Описать множество $\partial\Phi(u)$ $(u \in W)$.

8. Описать субдифференциалы функций $\rho(u, U)$, $\delta(c, U)$, $\mu(u, U)$ из упражнений 18—20 к § 4.2.

9. Пусть $J(u)$ — выпулая функция на открытом выпуклом множестве W из E^n, пусть субдифференциал $\partial J(u)$ в некоторой точке $u \in W$ состоит из единственного элемента c. Доказать, что $J(u)$ дифференцируема в точке u, причем $J'(u) = c$.

10. Пусть выпуклая функция $J(u)$ дифференцируема в каждой точке открытого выпуклого множества W. Доказать, что ее градиент $J'(u)$ непрерывен на W.

11. Пусть $J(u)$, $G(u)$ — выпуклые функции на открытом выпуклом множестве W из E^n, причем $\partial J(u) = \partial G(u)$ при всех $u \in W$. Доказать, что тогда $J(u) = G(u) + \text{const}$ $(u \in W)$.

12. Пусть функция $J(u)$ выпукла на открытом выпуклом множестве W из E^n. Доказать, что для того чтобы $J(u)$ была сильно выпуклой на W, необходимо и достаточно, чтобы для каждой точки $v \in W$ существовал субградиент $c(v) \in \partial J(v)$ такой, что

$$J(u) - J(v) \geqslant \langle c(v), u - v \rangle + \varkappa |u - v|^2 \qquad \forall u \in W, \quad \varkappa = \text{const} > 0.$$

13. Пусть функция $J(u)$ сильно выпукла на открытом выпуклом множестве W из E^n. Доказать:

а) $\langle c(u) - c(v),\ u - v \rangle \geqslant 2\varkappa |u - v|^2$ для всех $\forall u, v \in W$, $c(u) \in \partial J(u)$, $c(v) \in \partial J(v)$;

б) $\partial J(u) \cap \partial J(v) = \varnothing$ для всех $u, v \in W$, $u \neq v$;

в) для любой точки $v \in W$ справедливо неравенство $|u-v| \leqslant \leqslant \frac{1}{\varkappa} \min\limits_{c \in \partial J(v)} |c|$ для всех $u \in M(v) = \{u \in U\colon J(u) \leqslant J(v)\}$.

Опираясь на это утверждение, доказать теорему 3.1 для любого выпуклого замкнутого множества $U \subset W$.

14. Пусть функция $J(u)$ выпукла на открытом выпуклом множестве $W \subset E^n$ и сильно выпукла на выпуклом замкнутом подмножестве $U \subset W$. Доказать, что тогда

$$0 \leqslant J(u) - J_* \leqslant \frac{1}{4\varkappa} \min_{c \in \partial J(u)} |c|^2, \quad |u - u_*| \leqslant \frac{1}{2\varkappa} \min_{c \in \partial J(u)} |c|,$$

где u_* — точка минимума $J(u)$ на U, $J_* = J(u_*)$.

15. Пусть функция $J(u)$ сильно выпукла на E^n. Доказать, что для любого $c \in E^n$ существует такая единственная точка $u(c) \in E^n$, что $c \in \partial J(u(c))$.

Указание: рассмотреть точку минимума функции $g(u) = J(u) - \langle c, u\rangle$ на E^n.

§ 7. Равномерно выпуклые функции

1. Рассмотренный в § 3 класс сильно выпуклых функций обладает замечательным свойством — для функций этого класса имеет место теорема 3.1. Однако этот подкласс выпуклых функций недостаточно широк и не содержит, например, такую выпуклую функцию, как $J(u) = u^4$ $(u \in E^1)$, которая, между прочим, достигает своей нижней грани на любом выпуклом замкнутом множестве из E^1, причем в единственной точке. Хотелось бы выделить такой подкласс выпуклых функций, для которого была бы верна теорема типа теоремы 3.1 и который был бы шире класса сильно выпуклых функций. Оказывается, таким классом является класс равномерно выпуклых функций.

Определение 1. Функцию $J(u)$, определенную на выпуклом множестве U, называют *равномерно выпуклой* на U, если существует неотрицательная функция $\delta(t)$, определенная при всех $t\left(0 \leqslant t \leqslant \operatorname{diam} U = \sup\limits_{u,v \in U} |u-v|\right)$, $\delta(0) = 0$, $\delta(t_0) > 0$ при некотором t_0 $(0 < t_0 < \operatorname{diam} U)$ и такая, что

$$J(\alpha u + (1-\alpha)v) \leqslant \alpha J(u) + (1-\alpha) J(v) - \alpha(1-\alpha)\delta(|u-v|) \qquad (1)$$

при всех u, $v \in U$, $\alpha \in [0, 1]$. Функцию $\delta(t)$ называют *модулем выпуклости* функции $J(u)$ на U, а функцию

$$\mu(t) = \inf_{0<\alpha<1} \inf_{|u-v|=t,\, u,v \in U} \frac{\alpha J(u) + (1-\alpha) J(v) - J(\alpha u + (1-\alpha) v)}{\alpha(1-\alpha)}$$

точным модулем выпуклости $J(u)$ на U. Если равномерно выпуклая функция имеет модуль выпуклости $\delta(t) > 0$ при всех t $(0 < t < \operatorname{diam} U)$, то такую функцию называют *строго равномерно выпуклой* на U [92].

Очевидно, всякая сильно выпуклая функция является строго равномерно выпуклой с модулем $\delta(t) = \varkappa t^2$. Сумма равномерно

выпуклой функции с модулем $\delta(t)$ и выпуклой функции будет равномерно выпуклой с модулем $\delta(t)$. Если $J(u)$ равномерно выпукла с модулем $\delta(t)$, то функция $g(u)=cJ(u)$ при любом $c=\text{const}>0$ также будет равномерно выпуклой с модулем $c\delta(t)$. Если $\mu(t)$ — точный модуль выпуклости равномерно выпуклой функции $J(u)$ на U, то любая функция $\delta(t)\leqslant\mu(t)$ $(0\leqslant t\leqslant\operatorname{diam}U)$ неотрицательная, нетождественно равная нулю, $\delta(0)=0$, будет модулем выпуклости функции $J(u)$ на U.

Следующая теорема является обобщением теоремы 3.1.

Теорема 1. *Пусть U — выпуклое замкнутое множество из E^n (например, $U=E^n$), а функция $J(u)$ равномерно выпукла и полунепрерывна снизу на U. Тогда:*

1) *множество Лебега $M(v)=\{u\colon u\in U,\ J(u)\leqslant J(v)\}$ выпукло, замкнуто и ограничено при всех $v\in U$;*

2) $J_*=\inf\limits_U J(u)>-\infty$, $U_*=\{u\colon u\in U,\ J(u)=J_*\}\neq\varnothing$;

3) *имеет место неравенство*

$$\delta(|u-u_*|)\leqslant J(u)-J(u_*) \tag{2}$$

при всех $u\in U$, $u_*\in U_*$;

4) *если, кроме того, $J(u)$ строго равномерно выпукла на U, то U_* состоит из единственной точки u_* и всякая минимизирующая последовательность $\{u_k\}$: $\{u_k\}\in U$, $\lim\limits_{k\to\infty}J(u_k)=J_*$, сходится к точке u_*.*

Для доказательства этой теоремы нам понадобятся следующие две леммы о свойствах точного модуля выпуклости.

Лемма 1. *Пусть $\mu(t)$ — точный модуль выпуклости равномерно выпуклой функции $J(u)$ на выпуклом множестве U. Тогда*

$$\mu(ct)\geqslant c^2\mu(t) \tag{3}$$

для всех $c\geqslant 1$, $t\geqslant 0$, $0\leqslant ct\leqslant\operatorname{diam}U$.

Доказательство. Сначала рассмотрим случай $1\leqslant c<2$. По определению $\mu(ct)$ для любого $\varepsilon>0$ существуют точки u_1, $u_2\in U$ и число α $(0<\alpha<1)$ такие, что $|u_1-u_2|=ct$ и

$$\mu(ct)\leqslant\frac{\alpha J(u_1)+(1-\alpha)J(u_2)-J(u_\alpha)}{\alpha(1-\alpha)}\leqslant\mu(ct)+\varepsilon,$$

где $u_\alpha=\alpha u_1+(1-\alpha)u_2$. Отсюда имеем

$$\alpha J(u_1)+(1-\alpha)J(u_2)-J(u_\alpha)\leqslant\alpha(1-\alpha)\mu(ct)+\alpha(1-\alpha)\varepsilon. \tag{4}$$

Можем считать, что $0<\alpha\leqslant 1/2$, так как в противном случае в (4) точки u_1 и u_2 можно поменять ролями. Тогда с учетом $1\leqslant c<2$ можем сказать, что $0<\alpha\leqslant\alpha c<1$. Кроме того, $1/2<1/c\leqslant 1$, поэтому $u_3=(1/c)u_1+(1-1/c)u_2\in U$, причем $|u_3-u_2|=|u_1-u_2|/c=t$. Заметим также, что $u_\alpha=\alpha c u_3+(1-\alpha c)u_2$. Тогда

$$J(u_3)\leqslant(1/c)J(u_1)+(1-1/c)J(u_2)-(1/c)(1-1/c)\mu(ct),$$
$$J(u_\alpha)\leqslant\alpha cJ(u_3)+(1-\alpha c)J(u_2)-\alpha c(1-\alpha c)\mu(t).$$

Умножим первое из этих неравенств на αc и сложим со вторым. Учитывая неравенство (4), получаем

$$\alpha c(1/c)(1-1/c)\mu(ct)+\alpha c(1-\alpha c)\mu(t)\leqslant$$
$$\leqslant \alpha J(u_1)+(\alpha c-\alpha)J(u_2)-(1-\alpha c)J(u_2)-J(u_\alpha)=$$
$$=\alpha J(u_1)+(1-\alpha)J(u_2)-J(u_\alpha)\leqslant \alpha(1-\alpha)\mu(ct)+\alpha(1-\alpha)\varepsilon$$

или

$$\alpha(1-1/c)\mu(ct)+\alpha c(1-\alpha c)\mu(t)\leqslant \alpha(1-\alpha)\mu(ct)+\alpha(1-\alpha)\varepsilon.$$

Поскольку здесь $\varepsilon>0$ — произвольное число, то можем ε устремить к $+0$. Будем иметь

$$\alpha c(1-\alpha c)\mu(t)\leqslant(\alpha/c)(1-\alpha c)\mu(ct) \quad \text{или} \quad c^2\mu(t)\leqslant\mu(ct).$$

Неравенство (3) для случая $1\leqslant c<2$ доказано. Пусть теперь $c\geqslant 1$ — произвольное число, $0\leqslant ct\leqslant \operatorname{diam} U$. Поскольку $\lim\limits_{n\to\infty}\sqrt[n]{c}=1$, то найдется b $(1\leqslant b<2)$ такое, что $c=b^n$ при некотором натуральном $n\geqslant 1$. Учитывая, что по доказанному $\mu(bt)\geqslant b^2\mu(t)$, получаем $\mu(ct)=\mu(b^nt)=$ $=\mu(b\cdot b^{n-1}t)\geqslant b^2\mu(b^{n-1}t)\geqslant b^4\mu(b^{n-2}t)\geqslant\ldots\geqslant b^{2n}\mu(t)=c^2\mu(t)$.

Лемма 2. *Пусть $\mu(t)$ — точный модуль выпуклости равномерно выпуклой функции $J(u)$ на выпуклом множестве U. Тогда:*

1) $\mu(t)=O(t^2)$ *при* $t\to+0$;

2) $\mu(t)\equiv 0$ *при* $0\leqslant t<\tau=\inf\{t\colon \mu(t)>0\}$, $\mu(t)$ *строго монотонна при* $\tau<t\leqslant\operatorname{diam} U$;

3) *если* $\operatorname{diam} U\to\infty$, *то* $\lim\limits_{t\to\infty}\mu(t)=\infty$.

Доказательство. Из определения 1 следует, что $\mu(t_0)>0$ при некотором t_0 $(0<t_0<\operatorname{diam} U)$. Поэтому $0\leqslant\tau<\operatorname{diam} U$. Если $\tau>0$, то $\mu(t)\equiv 0$ при $0\leqslant t<\tau$ по определению τ. Пусть $\tau<t<a<\operatorname{diam} U$. Тогда с помощью неравенства (3) имеем $\mu(a)=\mu((a/t)t)\geqslant(a/t)^2\mu(t)>\mu(t)>0$, т. е. $\mu(t)$ строго монотонна при $\tau<t\leqslant\operatorname{diam} U$.

Далее, если $\tau>0$, то условие $\mu(t)=O(t^2)$ при $t\to+0$ выполняется тривиально. Поэтому пусть $\tau=0$. Тогда, фиксируя какое-либо t_0 $(0<t_0\leqslant$ $\leqslant\operatorname{diam} U)$, для всех $0<t<t_0$ имеем $\mu(t_0)=\mu((t_0/t)t)\geqslant(t_0/t)^2\mu(t)$ или $\mu(t)\leqslant\mu(t_0)\,t^2/t_0^2=\text{const}\cdot t^2$. Это и означает, что $\mu(t)=O(t^2)$ при $t\to+0$.

Наконец, пусть $\operatorname{diam} U\to\infty$. Тогда $\mu(t)$ определена при всех $t\geqslant 0$. Пусть $t\geqslant t_0>\tau$. Тогда $\mu(t)=\mu(t/t_0)t_0)\geqslant(t/t_0)^2\mu(t_0)=\text{const}\cdot t^2$. Это значит, что $\mu(t)\to\infty$ при $t\to\infty$ со скоростью не медленнее, чем t^2.

Заметим, что из неравенства $0\leqslant\delta(t)\leqslant\mu(t)$, справедливого для любого модуля выпуклости равномерно выпуклой функции, и из леммы 2 следует, что условие $\delta(t)=O(t^2)$ при $t\to+0$ является необходимым для того, чтобы некоторая функция $\delta(t)$ могла служить модулем выпуклости для какой-либо равномерно выпуклой функции.

Доказательство теоремы 1. Если множество U ограничено, замкнуто, т. е. U компактно, то утверждения 1), 2) теоремы следуют из теорем 2.1.1, 2.10. Остается рассмотреть случай, когда U — неограниченное множество. Тогда $\operatorname{diam} U\to\infty$ и точный модуль выпуклости $\mu(t)$ функции $J(u)$ будет определен при всех $t\geqslant 0$.

Пусть $t_0>0$ и $\mu(t_0)>0$. Возьмем произвольную точку $v\in U$ и рассмотрим шар

$$S=S(v,t_0)=\{u\colon u\in U,\ |u-v|\leqslant t_0\}.$$

Из теоремы 2.1.1 следует, что $\inf\limits_S J(u)=J_S^*>-\infty$, так что

$$J(u)\geqslant J_S^*=J(v)-\nu,\quad \nu=J(v)-J_S^*\geqslant 0, \tag{5}$$

при всех $u\in S$. Возьмем произвольную точку $u\in U\setminus S$, т. е. $|u-v|>t_0$.

Тогда, учитывая доказанную в лемме 2 строгую монотонность $\mu(t)$ при $t > \tau$, имеем

$$0 < \alpha_0 = (\mu(t_0)/\mu(|u-v|))^{1/2} < 1. \tag{6}$$

При $\alpha = \alpha_0$ из (1) получаем

$$\alpha_0 J(u) \geqslant J(v + \alpha_0(u-v)) - (1-\alpha_0)J(v) + \alpha_0(1-\alpha_0)\mu(|u-v|). \tag{7}$$

Из (6) и леммы 1 следует $\mu(t_0) = \alpha_0^2(|u-v|) = \alpha_0^2\mu((1/\alpha_0)\alpha_0|u-v|) \geqslant \mu(\alpha_0|u-v|)$ или $\mu(t_0) \geqslant \mu(\alpha_0|u-v|)$. В силу монотонности $\mu(t)$ это означает, что $\alpha_0|u-v| \leqslant t_0$. Тогда $v + \alpha_0(u-v) \in S$ и согласно (5) $J(v+\alpha_0(u-v)) \geqslant J(v) - \nu$. Учитывая эту оценку, из (7) имеем

$$\alpha_0 J(u) \geqslant \alpha_0 J(v) - \nu + \alpha_0(1-\alpha_0)\mu(|u-v|).$$

Отсюда, сокращая на $\alpha_0 > 0$ и вспоминая определение (6) величины α_0, получаем

$$J(u) \geqslant J(v) + (1-\alpha_0)\mu(|u-v|) - \nu/\alpha_0 =$$
$$= J(v) + \mu(|u-v|) - \sqrt{\mu(|u-v|)}\,(\sqrt{\mu(t_0)} + \nu/\sqrt{\mu(t_0)}).$$

Применяя к последнему слагаемому неравенство $ab \leqslant (a^2+b^2)/2$, будем иметь

$$J(u) \geqslant J(v) + \mu(|u-v|)/2 - (\sqrt{\mu(t_0)} + \nu/\sqrt{\mu(t_0)})^2/2 \tag{8}$$

для всех $u \in U \setminus S$. На самом деле, неравенство (8) имеет место для всех $u \in U$. Действительно, если $u \in S$, то $\mu(|u-v|) \leqslant \mu(t_0)$, а тогда $\nu < (\sqrt{\mu(t_0)} + \nu/\sqrt{\mu(t_0)})^2/2 - \mu(|u-v|)/2$. Отсюда и из (5) следует справедливость (8) и для $u \in S$.

Для всех $u \in M(v)$ из (8) имеем $\mu(|u-v|)/2 - (\sqrt{\mu(t_0)} + \nu/\sqrt{\mu(t_0)})^2/2 \leqslant J(u) - J(v) \leqslant 0$, т. е. $\mu(|u-v|) \leqslant (\sqrt{\mu(t_0)} + \nu/\sqrt{\mu(t_0)})^2$ при любом $u \in M(v)$. Поскольку $\mu(t) \to \infty$ при $t \to \infty$ и только в этом случае, то из последнего неравенства следует ограниченность множества $M(v)$. Выпуклость $M(v)$ следует из теоремы 2.10, а замкнутость $M(v)$ — из леммы 2.1.1. Из теоремы 2.1.2 имеем, что $J_* > -\infty$, $U_* \neq \varnothing$.

Докажем неравенство (2). Поскольку любой модуль выпуклости $\delta(t) \leqslant \mu(t)$, то неравенство (2) достаточно доказать для $\mu(t)$. Возьмем любую точку $u_* \in U_*$. Тогда $0 \leqslant J(\alpha u + (1-\alpha)u_*) - J(u_*) \leqslant \alpha(J(u) - J(u_*)) - \alpha(1-\alpha)\mu(|u-v|)$ или $\alpha(1-\alpha)\mu(|u-u_*|) \leqslant \alpha(J(u) - J(u_*))$ $(0 < \alpha < 1,\ u \in U)$. Деля на $\alpha > 0$ и устремляя $\alpha \to +0$, отсюда получаем неравенство (2).

Наконец, пусть функция $J(u)$ строго равномерно выпукла на U. Тогда она строго выпукла на U и согласно теореме 2.1 множество U_* будет состоять из единственной точки u_*. Возьмем произвольную минимизирующую последовательность $\{u_k\}$. Полагая в (2) $u = u_k$ и устремляя $k \to \infty$, получаем $\delta(|u_k - u_*|) \to 0$. Это возможно только при $|u_k - u_*| \to 0$, так как $J(u)$ строго равномерно выпукла.

2. Остановимся на некоторых необходимых, а также достаточных условиях равномерной выпуклости функции.

Теорема 2. *Пусть U — открытое выпуклое множество из E^n, пусть функция $J(u)$ равномерно выпукла на U с модулем выпуклости $\delta(t)$. Тогда необходимо выполняются неравенства*

$$J(u) \geqslant J(v) + \langle c(v),\ u-v\rangle + \delta(|u-v|), \tag{9}$$
$$\langle c(u) - c(v),\ u-v\rangle \geqslant 2\delta(|u-v|) \tag{10}$$

при всех $c(v) \in \partial J(v)$, $c(u) \in \partial J(u)$ и всех $u, v \in U$.

Доказательство. Поскольку равномерно выпуклая функция является и просто выпуклой, то из теоремы 6.1 следует, что $\partial J(u) \neq \varnothing$ при всех $u \in U$. Возьмем произвольные $u, v \in U$, $c(v) \in \partial J(v)$. Из определения субградиента и из (1) при всех α $(0 < \alpha < 1)$ имеем

$$\alpha\langle c(v), u - v\rangle + J(v) \leqslant J(\alpha u + (1-\alpha)v) \leqslant$$
$$\leqslant \alpha J(u) + (1-\alpha)J(v) - \alpha(1-\alpha)\delta(|u-v|)$$

или $(1-\alpha)\delta(|u-v|) + \langle c(v), u - v\rangle \leqslant J(u) - J(v)$. Отсюда при $\alpha \to +0$ получим неравенство (9). Поменяя в (9) переменные u и v ролями; будем иметь

$$J(v) \geqslant J(u) + \langle c(u), v - u\rangle + \delta(|u-v|), \quad c(u) \in \partial J(u).$$

Складывая это неравенство с (9), приходим к (10).

Приведем одно достаточное условие равномерной выпуклости функции.

Теорема 3. *Пусть U — выпуклое множество, $J(u) \in C^1(U)$, и пусть для некоторой непрерывной неотрицательной функции $\xi(t)$ $(0 \leqslant t \leqslant \operatorname{diam} U)$, $\xi(t) = O(t)^2$ при $t \to +0$, $\xi(t) \not\equiv 0$, выполняется неравенство*

$$\langle J'(u) - J'(v), u - v\rangle \geqslant \xi(|u-v|) \tag{11}$$

при всех $u, v \in U$. Тогда функция $J(u)$ равномерно выпукла на U с модулем выпуклости $\delta(t) = \int_0^1 (\xi(\tau t)/\tau)\, d\tau$.

Доказательство. Из формулы (2.7) и условия (11) имеем

$$\alpha J(u) + (1-\alpha) J(v) - J(\alpha u + (1-\alpha) v) =$$
$$= \alpha(1-\alpha)\int_0^1 \langle J'(z_1) - J'(z_2), z_1 - z_2\rangle \frac{d\tau}{\tau} \geqslant \alpha(1-\alpha)\int_0^1 \xi(|z_1 - z_2|)\frac{d\tau}{\tau} =$$
$$= \alpha(1-\alpha)\int_0^1 \xi(\tau|u-v|)\frac{d\tau}{\tau} = \alpha(1-\alpha)\,\delta(|u-v|), \quad u, v \in U, \quad \alpha \in [0, 1],$$

что и требовалось

Доказанная теорема 3 может быть использована для установления равномерной выпуклости конкретных функций.

Теорема 4. *Функция $J(u) = |u|^p$ строго равномерно выпукла на E^n при всех $p \geqslant 2$.*

Доказательство. Покажем, что

$$\langle J'(u) - J'(v), u - v\rangle \geqslant \frac{p}{2}\min\{1; 2^{3-p}\}\,|u-v|^p, \quad u, v \in E^n. \tag{12}$$

Здесь $J'(u) = p|u|^{p-2}$. Тогда

$$\langle J'(u) - J'(v), u - v\rangle = \langle p|u|^{p-2}u - p|v|^{p-2}v, u - v\rangle =$$
$$= p[|u|^p + |v|^p - \langle u, v\rangle(|u|^{p-2} + |v|^{p-2})] =$$
$$= p\left[|u|^p + |v|^p - \frac{|u|^2 + |v|^2 - |u-v|^2}{2}(|u|^{p-2} + |v|^{p-2})\right] =$$
$$= \frac{p}{2}[(|u|^{p-2} - |v|^{p-2})(|u|^2 - |v|^2) + |u-v|^2(|u|^{p-2} + |v|^{p-2})] \geqslant$$
$$\geqslant \frac{p}{2}|u-v|^2(|u|^{p-2} + |v|^{p-2}), \quad u, v \in E^n. \tag{13}$$

Покажем, что

$$|u|^{p-2}+|v|^{p-2}\geqslant|u-v|^{p-2}\min\{1;\ 2^{3-p}\},\qquad u,\ v\in E^n. \tag{14}$$

Рассмотрим функцию $\varphi(x)=(x^\alpha+1)/(x+1)^\alpha$ при $x\geqslant 1$, $\alpha>0$. Имеем $\varphi'(x)=\alpha(x^{\alpha-1}-1)(x+1)^{-\alpha-1}$. Если $\alpha\geqslant 1$, то $\varphi'(x)\geqslant 0$, и $\varphi(x)\geqslant\varphi(1)=2^{1-\alpha}$ для всех $x\geqslant 1$. Если $0<\alpha<1$, то $\varphi'(x)<0$ и $\varphi(x)\geqslant\varphi(\infty)=1$ при всех $x\geqslant 1$. Следовательно, $\varphi(x)\geqslant A_\alpha=\min\{1;\ 2^{1-\alpha}\}$ $(x\geqslant 1)$, или

$$A_\alpha(x+1)^\alpha\leqslant x^\alpha+1,\qquad x\geqslant 1,\qquad \alpha>0. \tag{15}$$

Далее имеем $|u-v|^{p-2}\leqslant(|u|+|v|)^{p-2}$. Без ограничения общности можем считать $|u|\geqslant|v|$. Тогда с помощью неравенства (15) получим

$$|u-v|^{p-2}\leqslant|v|^{p-2}(|u|/|v|+1)^{p-2}\leqslant A_{p-2}^{-1}\left((|u|/|v|)^{p-2}+1\right)|v|^{p-2},$$

что равносильно (14). Из (13) и (14) следует неравенство (12). С помощью теоремы 3 отсюда заключаем, что функция $J(u)=|u|^p$ при $p\geqslant 2$ равномерно выпукла на E^n с модулем $\delta(t)=t^p\min\{1;\ 2^{3-p}\}/2$.

Более тонкие оценки показывают, что функция $J(u)=|u|^p$ при $p\geqslant 2$ имеет точный модуль выпуклости $\mu(t)=t^p/2^{p-2}$ $(t\geqslant 0)$.

Будет ли функция $J(u)=|u|^p$ равномерно выпуклой на E^n при $1<p<2$? Оказывается, не будет. Чтобы убедиться в этом, достаточно показать, что функция $\varphi(x)=x^p$ одной переменной при $1<p<2$ не будет равномерно выпуклой на полуоси $x\geqslant 0$, поскольку функция $J(u)=|u|^p$ вдоль лучей $u=te$ $(|e|=1)$ ведет себя как функция t^p одной переменной.

Если бы функция $\varphi(x)=x^p$ $(1<p<2)$ была равномерно выпуклой при $x\geqslant 1$ с некоторым модулем выпуклости $\delta(t)$, то согласно теореме 2 необходимо выполнялось бы неравенство (10). В данном случае неравенство (10) имеет вид

$$2\delta(t)\leqslant tp\,[(x+t)^{p-1}-x^{p-1}],\qquad x\geqslant 0,\qquad t\geqslant 0.$$

С помощью формулы конечных превращений отсюда имеем

$$2\delta(t)\leqslant t^2p(p-1)(x+\theta t)^{p-2}\leqslant t^2p(p-1)x^{p-2},\ x\geqslant 0,\ t\geqslant 0.$$

Зафиксируем здесь произвольное $t>0$, а x устремим к ∞. Получим $\delta(t)\equiv 0$ при всех $t>0$.

Таким образом, функция $J(u)=|u|^p$ при $1<p<2$ не является равномерно выпуклой на E^n. Можно, однако, показать, что эта функция строго равномерно выпукла на любом выпуклом ограниченном множестве из E^n [92].

§ 8. Правило множителей Лагранжа

1. Пользуясь теоремами отделимости выпуклых множеств, можно получить условия экстремума для более общих задач на условный экстремум, чем задачи из § 2.2. А именно, рассмотрим задачу

$$J(u)\to\inf;\qquad u\in U, \tag{1}$$

$$U=\{u\in U_0:\ g_i(u)\leqslant 0,\ i=1,\ \ldots,\ m;\ g_i(u)=0,\ i=m+1,\ \ldots,\ s\}, \tag{2}$$

где U_0 — заданное множество из E^n, функции $J(u)$, $g_1(u)$, ..., $g_s(u)$ определены на U_0 и принимают конечные значения.

Здесь не исключаются возможности, когда отсутствуют либо ограничения $g_i(u) \leqslant 0$ типа неравенств ($m=0$), либо ограничения $g_i(u)=0$ типа равенств ($s=m$), либо оба вида ограничений ($m=s=0$, $U=U_0$). Разумеется, и само множество U_0 здесь может задаваться ограничениями типа равенств и неравенств.

При выделении множества U_0 в (2) обычно руководствуются тем, чтобы U_0 имело простую структуру, чтобы легко (без трудоемкой вычислительной работы) можно было проверить включение $u \in U_0$, указать какую-либо конкретную точку из U_0, чтобы легко было проектировать точку на U_0 и т. д. В задачах линейного программирования роль U_0 обычно играет неотрицательный ортант E^n_+. Часто множество U_0 представляет собой параллелепипед

$$U_0 = \{u = (u^1, \ldots, u^n) \in E^n\colon \alpha_i \leqslant u^i \leqslant \beta_i, \quad i = 1, \ldots, s\}, \tag{3}$$

где α_i, β_i — заданные числа, $\alpha_i < \beta_i$ (возможно, некоторые $\alpha_i = -\infty$, $\beta_j = \infty$). Конечно, в (2) не исключается возможность $U_0 = E^n$. Здесь мы можем вспомнить теорему 2.2.1, в которой с помощью функции Лагранжа было сформулировано необходимое условие оптимальности для задачи (1), (2) в частном случае, когда $U_0 = E^n$, $m=0$. При исследовании задачи (1), (2) в общем случае также важную роль играет функция Лагранжа

$$\mathscr{L}(u, \bar{\lambda}) = \lambda_0 J(u) + \lambda_1 g_1(u) + \ldots + \lambda_s g_s(u) \tag{4}$$

переменных $u = (u^1, \ldots, u^n) \in U_0$, $\bar{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_s) \in E^{s+1}$.

Теорема 1. *Пусть $u_* \in U$ — точка локального минимума в задаче* (1), (2) *функции $J(u)$, $g_1(u)$, …, $g_m(u)$ дифференцируемы в точке u_*, функции $g_{m+1}(u)$, …, $g_s(u)$ непрерывно дифференцируемы в некоторой окрестности точки u_*, U_0 — выпуклое множество. Тогда существуют числа $\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*$ такие, что*

$$\bar{\lambda}^* = (\lambda_0^*, \ldots, \lambda_s^*) \neq 0, \quad \lambda_0^* \geqslant 0, \quad \lambda_1^* \geqslant 0, \ldots, \lambda_m^* \geqslant 0, \tag{5}$$

$$\langle \mathscr{L}_u(u_*, \bar{\lambda}^*), u - u_* \rangle \geqslant 0 \qquad \forall u \in U_0, \tag{6}$$

$$\lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, s; \tag{7}$$

здесь $\mathscr{L}_u(u_*, \bar{\lambda}^*) = \lambda_0^* J'(u_*) + \lambda_1^* g_1'(u_*) + \ldots + \lambda_s^* g_s'(u_*)$ — градиент функции $\mathscr{L}(u, \bar{\lambda}^*)$ переменной $u \in U_0$ в точке $u = u_*$.

Числа $\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*$ называют *множителями Лагранжа*. Согласно условию (5) λ_0^* и множители, соответствующие ограничениям типа неравенств, неотрицательны, а множители $\lambda_{m+1}^*, \ldots, \lambda_s^*$, соответствующие ограничениям типа равенств, могут иметь любой знак.

Ограничение $g_i(u) \leqslant 0$, где $1 \leqslant i \leqslant m$, называют *активным* [*пассивным*] в точке u_*, если $g_i(u_*) = 0$ [$g_i(u_*) < 0$]. Из условий (7), часто называемых *условием дополняющей нежесткости*,

следует, что множители Лагранжа, соответствующие пассивным ограничениям типа неравенств, равны нулю.

Задачу (1), (2) называют *регулярной* (*невырожденной*) в точке u_*, если существуют множители Лагранжа $\bar{\lambda}^*$ с координатой $\lambda_0^* > 0$; в противном случае задача (1), (2) называется *нерегулярной* (*вырожденной*).

Простейший класс регулярных задач получается из (1), (2) при $m = s = 0$, $U = U_0$. В этом случае ограничения типа равенств и неравенств в (2) отсутствуют и нет необходимости вводить множители $\lambda_1, \ldots, \lambda_s$, поэтому $\mathscr{L}(u, \bar{\lambda}) = \lambda_0 J(u)$, условия (7) исчезают; кроме того, из (5) следует, что $\lambda_0^* > 0$, а тогда неравенство (6) превратится в условие $\langle J'(u_*), u - u_* \rangle \geqslant 0$ $(u \in U)$, известное нам из теоремы 2.3. Регулярность задачи (1), (2) гарантируется также и в том случае, когда $U_0 = E^n$ и градиенты $g_i'(u_*)$ $(i \in I = \{i\colon 1 \leqslant i \leqslant s, g_i(u_*) = 0\})$ линейно независимы.

В самом деле, для пассивных ограничений $\lambda_i^* = 0$ и условие (6) при $U_0 = E^n$ будет иметь вид

$$\lambda_0^* J'(u) + \sum_{i \in I} \lambda_i^* g_i'(u_*) = 0.$$

Если бы здесь было $\lambda_0^* = 0$, в силу линейной независимости $g_i'(u_*)$ $(i \in I)$ получили бы $\lambda_i^* = 0$ для всех $i \in I$, а тогда $\lambda_0^* = \lambda_1^* = \ldots = \lambda_s^* = 0$, что противоречит условию (5). Другие классы регулярных задач будут рассмотрены в § 9.

2. Из теоремы 1 следует, что в задаче (1), (2) с гладкими функциями $J(u)$, $g_i(u)$ на выпуклом множестве U_0 для поиска точек минимума (локального или глобального) нужно решить систему

$$\langle \lambda_0 J'(u) + \lambda_1 g_1'(u) + \ldots + \lambda_s g_s'(u), v - u \rangle \geqslant 0 \quad \forall v \in U_0, \tag{8}$$

$$\begin{gathered} \lambda_i g_i(u) = 0, \quad i = 1, \ldots, m; \\ g_i(u) = 0, \quad i = m+1, \ldots, s; \\ u \in U_0, \ g_1(u) \leqslant 0, \ldots, g_m(u) \leqslant 0, \end{gathered} \tag{9}$$

$$\lambda_0 \geqslant 0, \ \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0, \quad \bar{\lambda} \neq 0, \tag{10}$$

относительно $n + s + 1$ переменных $(u^1, \ldots, u^n, \lambda_0, \lambda_1, \ldots, \lambda_s) = (u, \bar{\lambda})$.

Заметим, что если какие-либо $(u, \bar{\lambda})$ получены из системы (8)—(10), то $(u, \mu\bar{\lambda})$ при любом $\mu > 0$ также удовлетворяют этой системе. Это значит, что множители Лагранжа из (8)—(10) определяются с точностью до постоянного положительного множителя. Поэтому множители Лагранжа можно подчинить како-

му-либо дополнительному условию нормировки, например,

$$|\bar{\lambda}|^2 = \lambda_0^2 + \lambda_1^2 + \ldots + \lambda_s^2 = 1. \tag{11}$$

В регулярной задаче вместо (10) можно взять $\lambda_0 = 1$. Отсюда следует, что систему (8)—(10) достаточно исследовать для двух значений λ_0: при $\lambda_0 = 0$ и при $\lambda_0 = 1$.

Условия (8)—(10) вместе с условием нормировки (11) (или $\lambda_0 = 1$ в регулярной задаче) дают «полную» систему соотношений для определения основных переменных $u = (u^1, \ldots, u^n)$ и соответствующих множителей Лагранжа $\bar{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_s)$. Для пояснения сказанного рассмотрим возможность $u \in \operatorname{int} U_0$ (это случится, например, при $U_0 = E^n$). Тогда неравенство (8) эквивалентно равенствам

$$\frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} = \lambda_0 \frac{\partial J(u)}{\partial u^i} + \lambda_1 \frac{\partial g_1(u)}{\partial u^i} + \ldots + \lambda_s \frac{\partial g_s(u)}{\partial u^i} = 0, \quad i = 1, \ldots, n. \tag{12}$$

Условия (12) вместе с (9), (11) представляют систему $n + s + 1$ уравнений с $n + s + 1$ неизвестными $(u, \bar{\lambda})$; решив ее и отобрав среди решений те, которые удовлетворяют неравенствам (10) и включению $u \in \operatorname{int} U_0$, получим набор $(u, \bar{\lambda})$, удовлетворяющий необходимому условию оптимальности. Для выяснения того, будет ли в отобранных точках в действительности реализовываться локальный минимум, нужно провести дополнительное исследование. Если же точка u принадлежит границе $\operatorname{Гр} U_0$ множества U_0, то неравенство (8) вместе с условием $u \in \operatorname{Гр} U_0$, вообще говоря, также дополняют уравнения (9), (11) до системы $n + s + 1$ уравнений. Так, например, если $U_0 = E^n_+$, то $\operatorname{Гр} E^n_+ = \{u = (u^1, \ldots, u^n)\colon u^i = 0,\ i \in I_1;\ u^i > 0,\ i \in I_2\}$, где $I_1 \cup I_2 = \{1, \ldots, n\}$, $I_1 \cap I_2 = \varnothing$, $I_1 \neq \varnothing$, и неравенство (8) при $u \in \operatorname{Гр} E^n_+$ приводит к соотношениям

$$\frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} \geqslant 0, \quad u^i = 0; \qquad \frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} = 0, \quad u^i > 0, \qquad i = 1, \ldots n. \tag{13}$$

Если же U_0 имеет вид (3), то (8) равносильно условиям

$$\frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} = 0, \quad \alpha_i < u^i < \beta_i; \qquad \frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} \geqslant 0, \quad u^i = \alpha_i > -\infty;$$

$$\frac{\partial \mathscr{L}(u, \bar{\lambda})}{\partial u^i} \leqslant 0, \quad u^i = \beta_i < \infty, \qquad i = 1, \ldots, n. \tag{14}$$

Для иллюстрации теоремы 1 приведем несколько примеров.

Пример 1. Пусть $J(u) = x + \cos y \to \inf$ $(u \in U = \{u = (x, y) \in U_0 = E^2\colon g(u) = -x \leqslant 0\})$. Тогда $\mathscr{L}(u, \bar{\lambda}) = \lambda_0(x + \cos y) + \lambda(-x)$, $\mathscr{L}_u = (\mathscr{L}_x, \mathscr{L}_y) = (\lambda_0 - \lambda, -\lambda_0 \sin y)$. Для опре-

деления подозрительных на минимум точек $u=(x, y)$ и соответствующих им множителей $\bar{\lambda}=(\lambda_0, \lambda)$ согласно (9), (10), (12) имеем систему

$$\lambda_0 \geqslant 0, \quad \lambda \geqslant 0, \quad \bar{\lambda}=(\lambda_0, \lambda) \neq 0, \quad \lambda(-x)=0, \quad -x \leqslant 0,$$
$$\mathscr{L}_x=\lambda_0-\lambda=0, \quad \mathscr{L}_y=-\lambda_0 \sin y=0.$$

Отсюда следует, что $\lambda_0=\lambda>0$. Учитывая условия нормировки вида (11), можем считать $\lambda_0=\lambda=1$. Тогда из предыдущей системы получаем точки $u=(0, \pi k)$ $(k=0, \pm 1, \pm 2, \ldots)$, подозрительные на оптимальность. Поскольку теорема 1 дает только необходимые условия оптимальности, то без дополнительного исследования нельзя сказать, будут ли отобранные точки $(0, \pi k)$ точками минимума $J(u)$ на U или нет. В данном случае такое исследование проводится легко. Точки $(0, \pi(2m+1))$ $(m=0, \pm 1, \pm 2, \ldots)$ будут точками минимума, так как $J(0, \pi(2m+1))=-1 \leqslant x+\cos y$ при любых $x \geqslant 0$ и любых y. В точках $(0, 2\pi m)$ $(m=0, \pm 1, \pm 2, \ldots)$ функция $J(x, y)$ не может достигать ни глобального, ни локального минимума на U, так как $J(0, 2\pi m)=1>J(0, y)=\cos y$ для всех y $(0<|y-2\pi m|<2\pi)$.

Пример 2. Пусть $J(u)=x \to \inf(u \in U=\{u=(x, y) \in U_0=E^2\colon g_1(u)=-x \leqslant 0, \quad g_2(u)=x^2-y \leqslant 0, \quad g_3(u)=y-2x^2 \leqslant 0\})$. Здесь $\mathscr{L}(u, \bar{\lambda})=\lambda_0 x+\lambda_1(-x)+\lambda_2(x^2-y)+\lambda_3(y-2x^2)$, $\mathscr{L}_u=(\mathscr{L}_x, \mathscr{L}_y)=(\lambda_0-\lambda_1+2\lambda_2 x-4\lambda_3 x, -\lambda_2+\lambda_3)$. Для определения подозрительных на оптимальность точек $u=(x, y)$ и соответствующих им множителей $\bar{\lambda}=(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ из (9)—(12) имеем систему

$$\lambda_0 \geqslant 0, \quad \lambda_1 \geqslant 0, \quad \lambda_2 \geqslant 0, \quad \lambda_3 \geqslant 0, \quad \bar{\lambda} \neq 0,$$
$$-x \leqslant 0, \quad x^2-y \leqslant 0, \quad y-2x^2 \leqslant 0,$$
$$\lambda_1(-x)=0, \quad \lambda_2(x^2-y)=0, \quad \lambda_3(y-2x^2)=0,$$
$$\lambda_0-\lambda_1+2x(\lambda_2-2\lambda_3)=0, \quad -\lambda_2+\lambda_3=0.$$

Отсюда ясно, что если $x>0$, то $\lambda_0=\lambda_1=\lambda_2=\lambda_3=0$, что противоречит условию $\bar{\lambda} \neq 0$. Следовательно, $x=0$. А тогда $y=0$, так что на минимум претендует всего одна точка $u_*=(0, 0)$. Нетрудно видеть, что в ней и достигается минимум $J(u)$ на U. В рассматриваемой задаче множителями Лагранжа будут любые $\bar{\lambda}=(\lambda_0, \lambda_1, \lambda_2, \lambda_3) \neq 0$, лишь бы $\lambda_0=\lambda_1 \geqslant 0$, $\lambda_2=\lambda_3 \geqslant 0$. Можно, например, взять $\bar{\lambda}=(1, 1, 0, 0)$ или $\bar{\lambda}=(0, 0, 1, 1)$ — это линейно независимые наборы, их нельзя получить друг из друга никакой нормировкой вида (11).

Пример 3. Пусть $J(u)=\sum_{i=1}^{m}|u-u_i|^2 \to \inf(u \in U=\{u \in U_0=E^n\colon g(u)=|u|^2-1 \leqslant 0\})$. Здесь $u_1, \ldots, u_m$ — заданные

точки из E^n. Эта задача рассматривалась в примере 2.2.4 и решалась сведением к задаче с ограничениями типа равенств путем введения дополнительной переменной. Применим к ней теорему 1. Здесь $\mathscr{L}(u, \bar{\lambda}) = \lambda_0 \sum_{i=1}^{m} |u - u_i|^2 + \lambda(\langle u, u\rangle - 1)$, $\mathscr{L}_u(u, \bar{\lambda}) = 2\lambda_0 \sum_{i=1}^{m} (u - u_i) + 2\lambda u = 2\lambda_0 m(u - u_0) + 2\lambda u$, $u_0 = \frac{1}{m}\sum_{i=1}^{m} u_i$.

Из (9)—(12) имеем систему

$$\lambda_0 m(u - u_0) + \lambda u = 0, \quad \lambda(|u|^2 - 1) = 0, \quad |u| \leqslant 1, \quad \lambda_0 \geqslant 0,$$
$$\lambda \geqslant 0, \quad \lambda_0^2 + \lambda^2 > 0.$$

Решением этой системы при $|u_0| > 1$ является набор $u = u_0/|u_0|$, $\lambda = m(|u_0| - 1)$, $\lambda_0 = 1$; если же $|u_0| \leqslant 1$, то $u = u_0$, $\lambda = 0$, $\lambda_0 = 1$. Как было показано в § 2.2, найденные точки действительно являются точками глобального минимума $J(u)$ на U.

Пример 4. Рассмотрим задачу линейного программирования $J(u) = -x - y \to \inf(u \in U = \{u = (x, y) \in U_0\colon g(u) = x - y = 0\})$, где $U_0 = \{u = (x, y) \in E^2\colon 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 2\}$ — прямоугольник. Здесь $\mathscr{L}(u, \bar{\lambda}) = \lambda_0(-x - y) + \lambda(x - y)$, $\mathscr{L}_x = -\lambda_0 + \lambda$, $\mathscr{L}_y = -\lambda_0 - \lambda$. Из (8)—(10) с учетом (14) имеем

$$-\lambda_0 + \lambda = 0, \ 0 < x < 1; \quad -\lambda_0 + \lambda \geqslant 0, \ x = 0; \quad -\lambda_0 + \lambda \leqslant 0, \ x = 1;$$
$$-\lambda_0 - \lambda = 0, \ 0 < y < 2; \quad -\lambda_0 - \lambda \geqslant 0, \ y = 0; \quad -\lambda_0 - \lambda \leqslant 0, \ y = 2;$$
$$x - y = 0, \quad 0 \leqslant x \leqslant 1; \quad 0 \leqslant y \leqslant 2; \quad \lambda_0 \geqslant 0, \quad \lambda_0^2 + \lambda^2 \neq 0.$$

При $\lambda_0 = 0$, как нетрудно проверить, система не имеет решения, поэтому можем положить $\lambda_0 = 1$. Последовательно перебирая возможности $0 < x = y < 1$, $x = y = 0$, $x = y = 1$, находим единственную точку $u = (1, 1)$, подозрительную на оптимальность, и соответствующие множители Лагранжа $\lambda_0 = 1$, $\lambda = -1$. Легко проверить, что в точке $u = (1, 1)$ функция $J(u)$ достигает минимума на U.

Пример 5. Задачу из примера 4 можно задать в эквивалентной форме, заменив ограничение типа равенств двумя ограничениями типа неравенств: $J(u) = -x - y \to \inf(u \in U = \{u = (x, y) \in U_0\colon g_1(u) = x - y \leqslant 0,\ g_2(u) = -x + y \leqslant 0\})$, где $U_0 = \{u = (x, y) \in E^2\colon 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 2\}$. Тогда $\mathscr{L}(u, \bar{\lambda}) = \lambda_0(-x - y) + \lambda_1(x - y) + \lambda_2(-x + y)$, $\mathscr{L}_x = -\lambda_0 + \lambda_1 - \lambda_2$, $\mathscr{L}_y = -\lambda_0 - \lambda_1 + \lambda_2$. С помощью (8)—(10) с учетом (14) придем

к системе

$$-\lambda_0+\lambda_1-\lambda_2=0,\ \ 0<x<1;\ \ -\lambda_0+\lambda_1-\lambda_2\geqslant 0\,[\leqslant 0],\ \ x=0\ \ [x=1];$$
$$-\lambda_0-\lambda_1+\lambda_2=0,\ \ 0<y<2;\ \ -\lambda_0-\lambda_1+\lambda_2\geqslant 0\,[\leqslant 0],\ \ y=0\,[y=2];$$
$$\lambda_1(x-y)=0,\ \ \lambda_2(-x+y)=0;\ \ 0\leqslant x\leqslant 1,\ \ 0\leqslant y\leqslant 2,$$
$$x-y\leqslant 0,\ \ -x+y\leqslant 0;\ \ \lambda_0\geqslant 0,\ \ \lambda_1\geqslant 0,\ \ \lambda_2\geqslant 0,$$
$$\lambda_0^2+\lambda_1^2+\lambda_2^2\neq 0.$$

При $\lambda_0=0$ этой системе удовлетворяют все точки $u=(x,\ y)\in\in U$ с множителями Лагранжа $\bar\lambda=(0,\ \lambda_1=\lambda,\ \lambda_2=\lambda)$, где $\lambda>0$; в этом случае теорема 1 не позволила сузить исходное множество точек, подозрительных на оптимальность. При $\lambda_0=1$, как и в примере 4, получаем единственную точку $u=(1,\ 1)$, подозрительную на оптимальность, но в отличие от примера 4 здесь множителей Лагранжа много: $\bar\lambda=(\lambda_0=1,\ \lambda_1=\lambda,\ \lambda_2=1+\lambda)$ $(\lambda\geqslant 0)$.

В § 2.2 был приведен пример 2.2.3 нерегулярной задачи с ограничением типа равенств. Приведем пример такой задачи с ограничениями типа неравенств.

Пример 6. Рассмотрим задачу $J(u)=-u\to\inf(u\in U=\{u\in U_0:\ g(u)=u^2\leqslant 0\})$, где $U_0=\{u\in E^1:\ 0\leqslant u\leqslant a\}$, $a>0$ — фиксированное число (возможно, $a=\infty$). Здесь $U=\{0\}=U_*$, $J_*=0$, $\mathscr{L}(u,\ \bar\lambda)=-\lambda_0u+\lambda u^2$, $\mathscr{L}_u(u,\ \bar\lambda)=-\lambda_0+2\lambda u$. Система (8)—(10) запишется в виде

$$(-\lambda_0+2\lambda u)(v-u)\geqslant 0\qquad \forall v\in[0,\ a],$$
$$\lambda u^2=0,\ \ u^2\leqslant 0,\ \ \lambda_0\geqslant 0,\ \ \lambda\geqslant 0,\ \ \lambda_0^2+\lambda^2\neq 0.$$

Отсюда видно, что $u=0$. Тогда первое неравенство системы дает $-\lambda_0v\geqslant 0$ при всех $0\leqslant v\leqslant a$, что возможно только при $-\lambda_0\geqslant 0$. С другой стороны, $\lambda_0\geqslant 0$. Следовательно, $\lambda_0=0$, т. е. рассматриваемая задача не является регулярной. В качестве множителей Лагранжа здесь можно взять $\bar\lambda=(0,\ \lambda)$ при любом $\lambda>0$.

3. Доказательство теоремы 1. Проведем его с помощью принятого в [21, гл. 4, § 2] метода, простого и очень изящного. Введем множество $I_*=\{i:\ 1\leqslant i\leqslant m,\ g_i(u_*)=0\}$ номеров активных в точке u_* ограничений (возможность $I_*=\varnothing$ не исключается). Определим множества

$$A=\{a=(a_0;\ a_i,\ i\in I_*;\ a_{m+1},\ \ldots,\ a_s):\ a_0=\langle J'(u_*),\ u-u_*\rangle;$$
$$a_i=\langle g_i'(u_*),\ u-u_*\rangle,\quad i\in I_*,\quad i=m+1,\ \ldots\ s;\quad u\in \operatorname{ri}U_0\},$$
$$B=\{b=(b_0;\ \ b_i,\ \ i\in I_*;\ b_{m+1},\ \ldots,\ b_s):\ b_0<0;\ b_i<0,\ i\in I_*;$$
$$b_{m+1}=0,\ \ldots,\ b_s=0\}.$$

Нетрудно видеть, что эти множества выпуклы.

Покажем, например, выпуклость A. Пусть $a_j = (a_{0j};\ a_{ij},\ i \in I_*;\ a_{m+1j},\ \ldots,\ a_{sj})$ $(j = 1,\ 2)$ — две любые точки из A. Это значит, что существуют точки $u_1,\ u_2 \in \operatorname{ri} U_0$ такие, что $a_{0j} = \langle J'(u_*),\ u_j - u_* \rangle$, $a_{ij} = \langle g_i'(u_*),\ u_j - u_* \rangle$ $(i \in I_*,\ i = m+1,\ \ldots,\ s,\ j = 1,\ 2)$.

Возьмем любое $\alpha \in [0,\ 1]$ и положим $a_\alpha = \alpha a_1 + (1-\alpha) a_2$, $u_\alpha = \alpha u_1 + (1-\alpha) u_2$. Из выпуклости $\operatorname{ri} U_0$ (см. теорему 1.11) следует, что $u_\alpha \in \operatorname{ri} U_0$. Далее, из линейности функций $\langle J'(u_*),\ u - u_* \rangle$, $\langle g_i'(u_*),\ u - u_* \rangle$ переменной u имеем $\langle J'(u_*),\ u_\alpha - u_* \rangle = \alpha \langle J'(u_*),\ u_1 - u_* \rangle + (1-\alpha) \langle J'(u_*),\ u_2 - u_* \rangle = \alpha a_{01} + (1-\alpha) \times a_{02} = a_{0\alpha}$; аналогично $\langle g_i'(u_*),\ u_\alpha - u_* \rangle = \alpha a_{i1} + (1-\alpha) a_{i2} = a_{i\alpha}$ $(i = m+1,\ \ldots,\ s)$. Это означает, что $a_\alpha \in A$. Выпуклость A доказана. Аналогично доказывается выпуклость B.

Покажем, что $A \cap B = \varnothing$. Возьмем несущее подпространство $L = \operatorname{Lin} U_0$ множества U_0 (см. определение 1.3). Пусть $e_1,\ \ldots,\ e_p$ — базис подпространства $L^\perp$, ортогонального L. Тогда $L = \{h \in E^n\colon \langle e_i,\ h \rangle = 0,\ i = 1,\ \ldots,\ p\}$. Может оказаться, что система векторов $g_{m+1}'(u_*),\ \ldots,\ g_s'(u_*),\ e_1,\ \ldots,\ e_p$ линейно зависима, т. е. существуют числа $\lambda_{m+1}^*,\ \ldots,\ \lambda_s^*,\ \alpha_1,\ \ldots,\ \alpha_p$, не все равные нулю и такие, что

$$\lambda_{m+1}^* g_{m+1}'(u_*) + \ldots + \lambda_s^* g_s'(u_*) + \alpha_1 e_1 + \ldots + \alpha_p e_p = 0. \quad (15)$$

Среди чисел $\lambda_{m+1}^*,\ \ldots,\ \lambda_s^*$ найдутся отличные от нуля числа, так как в противном случае из (15) следовала бы линейная зависимость векторов $e_1,\ \ldots,\ e_p$.

Кроме того, из (15) следует, что

$$\sum_{i=m+1}^{s} \lambda_i^* \langle g_i'(u_*),\ h \rangle = -\sum_{i=1}^{p} \alpha_i \langle e_i,\ h \rangle = 0 \qquad \forall h \in \operatorname{Lin} U_0.$$

Но $\operatorname{Lin} U_0 = \operatorname{aff} U_0 - u_*$, поэтому предыдущее равенство можно переписать в виде

$$\sum_{i=m+1}^{s} \lambda_i^* \langle g_i'(u_*),\ u - u_* \rangle = 0 \qquad \forall u \in U_0 \subset \operatorname{aff} U_0.$$

Отсюда следует, что набор чисел $\bar{\lambda}^* = (\lambda_0^* = 0,\ \lambda_1^* = 0,\ \ldots,\ \lambda_m^* = 0,\ \lambda_{m+1}^*,\ \ldots,\ \lambda_s^*)$ удовлетворяет условиям (5) — (7).

Таким образом, равенство $A \cap B = \varnothing$ можем доказывать, предполагая, что система векторов $g_{m+1}'(u_*),\ \ldots,\ g_s'(u_*),\ e_1,\ \ldots,\ e_p$ линейно независима. Более того, можем считать, что эта система образует базис в E^n, так как в противном случае дополним ее до базиса каким-либо способом; тогда $p = n - s + m$.

Допустим, что $A \cap B \neq \varnothing$. Это значит, что найдется такая точка $\bar{u} \in \operatorname{ri} U_0$, что

$$a_0 = \langle J'(u_*), \bar{u} - u_* \rangle < 0; \quad a_i = \langle g_i'(u_*), \bar{u} - u_* \rangle < 0, \quad i \in I_*; \tag{16}$$

$$a_i = \langle g_i'(u_*), \bar{u} - u_* \rangle = 0, \qquad i = m+1, \ldots, s.$$

Обозначим $h = \bar{u} - u_*$; введем функции $f_i(r, t) \equiv g_{m+i}(u_* + th + r)$ $(i = 1, \ldots, s-m)$, $f_i(r, t) = \langle e_{i-s+m}, r \rangle$ $(i = s-m+1, \ldots, n = p + s - m)$, $f(r, t) = (f_1(r, t), \ldots, f_n(r, t))$ и рассмотрим систему n уравнений

$$f_1(r, t) = 0, \ \ldots, \ f_n(r, t) = 0 \tag{17}$$

относительно n неизвестных $r = (r_1, \ldots, r_n)$.

Для исследования системы (17) воспользуемся известной теоремой о неявных функциях [10, 160, 165, 233]. Прежде всего заметим, что $f_1(0, 0) = 0, \ldots, f_n(0, 0) = 0$. Далее, функции $f_i(r, t)$ непрерывно дифференцируемы в окрестности точки (0, 0), причем с учетом (16)

$$\frac{\partial f_i(0, 0)}{\partial r} = g_{m+i}'(u_*), \quad \frac{\partial f_i(0, 0)}{\partial t} = \langle g_{m+i}'(u_*), h \rangle = 0,$$

$$i = 1, \ldots, \ s - m;$$

$$\frac{\partial f_i(0, 0)}{\partial r} = e_{i-s+m}, \quad \frac{\partial f_i(0, 0)}{\partial t} = 0, \quad i = s - m + 1, \ldots, n.$$

Таким образом, якобиан $\partial(f_1, \ldots, f_n)/\partial(r_1, \ldots, r_n)$ системы (17) в точке (0, 0), представляющий собой определитель квадратной матрицы $\partial f(0, 0)/\partial r$ со строками $g_{m+1}'(u_*), \ldots, g_s'(u_*), e_1, \ldots, e_p$, образующими базис в E^n, отличен от нуля.

Все условия теоремы о неявных функциях выполнены. Согласно этой теореме существуют непрерывно дифференцируемые функции $r = r(t) = (r_1(t), \ldots, r_n(t))$, определенные при всех $t (|t| \leqslant t_0)$, где t_0 — достаточно малое положительное число, и удовлетворяющие системе

$$r(0) = 0, \ f(r(t), t) \equiv 0, \quad |t| \leqslant t_0.$$

Дифференцируя последнее тождество по t, получаем $\frac{\partial f(r(t), t)}{\partial r} r'(t) + \frac{\partial f(r(t), t)}{\partial t} = 0$ $(|t| \leqslant t_0)$. Отсюда при $t = 0$ с учетом равенства $\frac{\partial f(0, 0)}{\partial t} = 0$ будем иметь $\frac{\partial f(0, 0)}{\partial r} r'(0) = 0$. Однако матрица $\partial f(0, 0)/\partial r$ невырожденная, поэтому $r'(0) = 0$. Это значит, что $r(t) = r(0) + tr'(0) + o(t) = o(t)$, т. е. $\lim\limits_{t \to 0} r(t)/t = 0$. Таким образом, найдена вектор-функция $r(t) = (r_1(t), \ldots$

$\ldots, r_n(t))$ $(|t| \leqslant t_0)$, для которой

$$g_i(u_* + t(\bar{u} - u_*) + r(t)) = 0, \quad i = m+1, \ldots, s, \quad \langle e_i, r(t)\rangle = 0,$$
$$i = 1, \ldots, p = n - s + m, \quad |t| \leqslant t_0, \quad \lim_{t\to 0} r(t)/t = 0. \tag{18}$$

Покажем, что по кривой $u(t) = u_* + t(\bar{u} - u_*) + r(t)$ можно двигаться, оставаясь в множестве U при всех t, $0 < t < t_1$, где t_1 — достаточно малое число. Вторая группа равенств (18) означает, что $r(t) \in \operatorname{Lin} U_0$, поэтому $\bar{u} + r(t)/t \in \operatorname{aff} U_0$. Напоминаем, что $\bar{u} \in \operatorname{ri} U_0$, а поскольку $\lim\limits_{t\to 0} r(t)/t = 0$, то $\bar{u} + r(t)/t \in U_0$ при всех малых t. Тогда, учитывая выпуклость U_0, имеем $u(t) = u_* + t(\bar{u} - u_*) + r(t) = t(\bar{u} + r(t)/t) + (1-t)u_* \in U_0$ при всех малых t $(0 < t < 1)$.

Далее, первая группа равенств (18) означает, что $g_i(u(t)) = 0$ $(i = m+1, \ldots, s,\ 0 \leqslant t \leqslant t_0)$. Пусть $1 \leqslant i \leqslant m$. Если $i \in I_*$, то $g_i(u_*) = 0$ и с учетом (16) имеем

$$g_i(u(t)) = g_i(u_*) + \langle g_i'(u_*), t(\bar{u} - u_*) + r(t)\rangle + o_i(t) =$$
$$= t\left[\langle g_i'(u_*), \bar{u} - u_*\rangle + \langle g_i'(u_*), r(t)/t\rangle + o_i(t)/t\right] < 0$$

при всех малых $t > 0$. Если $i \notin I_*$, $1 \leqslant i \leqslant m$, то $g_i(u_*) < 0$ и в силу непрерывности $g_i(u)$ неравенство $g_i(u(t)) = g_i(u_* + t(\bar{u} - u_*) + r(t)) < 0$ сохранится при всех малых t. Таким образом, существует достаточно малое число t_1 $(0 < t_1 < \min\{t_0;\ 1\})$ такое, что $u(t) \in U$ при всех t $(0 \leqslant t \leqslant t_1)$.

Беря при необходимости t_1 еще меньшим, с учетом (16) имеем

$$J(u(t)) - J(u_*) =$$
$$= t\left[\langle J'(u_*), \bar{u} - u_*\rangle + \langle J'(u_*), r(t)/t\rangle + o(t)/t\right] < 0, \quad 0 < t < t_1.$$

Однако $u(t) \to u_*$ при $t \to 0$ и $u(t) \in U$ $(0 < t < t_1)$ и последнее неравенство противоречит тому, что u_* — точка локального минимума в задаче (1), (2). Следовательно, $A \cap B = \varnothing$. По теореме 5.2 тогда существует гиперплоскость $\langle c, a\rangle = \gamma$ с нормальным вектором $c = (\lambda_0^*;\ \lambda_i^*,\ i \in I_*;\ \lambda_{m+1}^*, \ldots, \lambda_s^*) \neq 0$, отделяющая множества A и B, а также A и $\bar{B} = \{b = (b_0;\ b_i,\ i \in I_*;\ b_{m+1}, \ldots, b_s)\colon b_0 \leqslant 0;\ b_i \leqslant 0,\ i \in I_*,\ b_{m+1} = 0, \ldots, b_s = 0\}$. Это значит, что

$$\langle c, b\rangle = \lambda_0^* b_0 + \sum_{i\in I_*} \lambda_i^* b_i + \sum_{i=m+1}^{s} \lambda_i^* b_i \leqslant \gamma \leqslant \langle c, a\rangle =$$
$$= \lambda_0^* a_0 + \sum_{i\in I_*} \lambda_i^* a_i + \sum_{i=m+1}^{s} \lambda_i^* a_i \tag{19}$$

при всех $a \in A$, $b \in \bar{B}$.

Разделив (19) почленно на $b_j < 0$, где $j = 0$ или $j \in I_*$, и устремив $b_j \to -\infty$ при фиксированных остальных b_i, a, получим $\lambda_j^* \geqslant 0$ при $j = 0$ или $j \in I_*$. Далее, полагая в (19) $a_0 = \langle J'(u_*), u - u_* \rangle$, $a_i = \langle g_i'(u_*), u - u_* \rangle$ $(i \in I_*,\ i = m+1, \ldots, s)$, где $u \in \operatorname{ri} U_0$, $b = 0 \in \overline{B}$, будем иметь

$$\lambda_0^* \langle J'(u_*), u - u_* \rangle + \sum_{i \in I_*} \lambda_i^* \langle g_i'(u_*), u - u_* \rangle + \\ + \sum_{i=m+1}^{s} \lambda_i^* \langle g_i'(u_*), u - u_* \rangle \geqslant 0 \qquad \forall u \in \operatorname{ri} U_0.$$

Отсюда, взяв $\lambda_i^* = 0$ при $i \notin I_*$, $1 \leqslant i \leqslant m$, получим

$$\left\langle \lambda_0^* J'(u_*) + \sum_{i=1}^{s} \lambda_i^* g_i'(u_*), u - u_* \right\rangle \geqslant 0 \qquad \forall u \in \operatorname{ri} U_0.$$

Для получения неравенства (6) здесь остается совершить предельные переходы с учетом того, что $U_0 \subset \overline{U}_0 = \overline{\operatorname{ri} U_0}$ (теорема 1.13). Справедливость условий (5), (7) следует из определения множества I_*, построения $\overline{\lambda}^* = (\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*)$, включения $u_* \in U$.

4. Заметим, что если функция $\mathscr{L}(u, \overline{\lambda}^*)$ переменной $u \in U_0$ выпукла на U_0, то согласно теореме 2.3 из условия (6) следует, что $\mathscr{L}(u, \overline{\lambda}^*)$ достигает своей нижней грани на U_0 в точке u_*, и условия (5)—(7) можно переписать в виде

$$\overline{\lambda}^* \neq 0, \quad \lambda_0^* \geqslant 0, \ldots, \lambda_m^* \geqslant 0, \quad \mathscr{L}(u_*, \lambda^*) \leqslant \mathscr{L}(u, \lambda^*) \quad \forall u \in U_0; \\ \lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, s. \tag{20}$$

В § 9 будет показано, что для выпуклых регулярных задач (1), (2) условия (20) являются необходимыми и достаточными условиями оптимальности. Условия оптимальности (необходимые и достаточные), использующие вторые производные функции Лагранжа $\mathscr{L}(u, \overline{\lambda})$, рассмотрены в [21]. Различные обобщения и модификации правила множителей Лагранжа см. в [1, 2, 8, 17, 21, 24, 29, 66, 67, 91, 116, 137, 146, 152, 156, 166, 167, 201, 219, 254, 255, 264, 278, 283, 290, 299, 308, 330, 341].

Упражнения. 1. Сформулировать правило множителей Лагранжа для задачи $g(u) \to \sup$ $(u \in U)$, где множество U определено посредством (2).

Указание: рассмотреть задачу $J(u) = -g(u) \to \inf$ $(u \in U)$ и воспользоваться теоремой 1.

2. С помощью правила множителей Лагранжа исследовать задачи:

а) $J(u) = x \to \inf$ $(u \in U)$, где $U = \{u = (x, y) \in E^2 = U_0\colon x^2 + y^2 \leqslant 1,\ x^2 \leqslant y,\ x + y \leqslant 0\}$, или $U = \{u \in E^2\colon x^2 + y^2 \leqslant 1,\ x^3 + y^3 = 1\}$, или $U = \{u \in E^2\colon x^2 + y^2 \leqslant 1,\ -x^3 \leqslant y \leqslant x^3\}$;

б) $J(u) = |u-a|^2 \to \inf$ [$\to \sup$] $(u \in U)$, где $U = \{u = (u^1, \ldots, u^n) \in E^n$: $u^1 + u^2 + \ldots + u^n = 0\}$, или $U = \{u \in E^n_+ : |u|^2 \leqslant 1\}$ (a — заданная точка из E^n);

в) $J(u) = 2x^{-2} + 4x^5y^{-2} \to \inf$ $(u \in U = \{u = (x, y) \in E^2$: $x > 0$, $y > 0$, $x^{-4}y^2 \leqslant 1\})$;

г) $J(u) = x + y^{-1}z^{-1/2} \to \inf$ $(u \in U = \{u = (x, y, z) \in E^3$: $x > 0$, $y > 0$, $z > 0$, $x^{-1}y + x^{-1}z \leqslant 1\})$.

3. Исследовать задачи из примеров и упражнений к § 2.2, к гл. 3, пользуясь правилом множителей Лагранжа.

4. Доказать равносильность условий (13) для $U_0 = E^n_+$ и (14) для U_0 из (3) условию (8).

5. Пусть в теореме 1 U_0 — аффинное множество. Доказать, что тогда условие (8) равносильно условию $\langle \mathscr{L}_u(u_*, \bar{\lambda}^*), h\rangle = 0$ при всех $h \in \operatorname{Lin} U_0$.

6. Пусть $u_* \in U$ — точка локального минимума в задаче (1), (2). U_0 — выпуклое множество, $u_* \in \operatorname{int} U_0$; функции $J(u)$, $g_1(u)$, ..., $g_s(u)$ дважды дифференцируемы в точке u_*; функции $g_i(u)$ $(i \in I_{**} = \{i: 1 \leqslant i \leqslant s,\ g_i(u_*) = 0\})$ непрерывно дифференцируемы в некоторой окрестности точки u_*, причем градиенты $g_i'(u_*)$ $(i \in I_{**})$ линейно независимы. Доказать, что тогда необходимо $\langle \mathscr{L}_{uu}(u_*, \bar{\lambda}^*)h, h\rangle \geqslant 0$ при всех $\bar{\lambda}^*$, удовлетворяющих условиям (5) — (7), и всех $h \in H(u_*) = \{h \in E^n$: $\langle J'(u_*), h\rangle \leqslant 0$; $\langle g_i'(u_*), h\rangle \leqslant 0$, $i \in I_* = \{i$: $1 \leqslant i \leqslant m$, $g_i(u_*) = 0$; $\langle g_i'(u_*), h\rangle = 0$, $i = m+1, \ldots, s\}$ ([21, с. 143]).

7. Пусть в задаче (1), (2) функции $J(u)$, $g_1(u)$, ..., $g_s(u)$ дважды дифференцируемы в точке $u_* \in U$, пусть для некоторого $\bar{\lambda}^*$ выполнены условия (5) — (7) и, кроме того, $\langle \mathscr{L}_{uu}(u_*, \bar{\lambda}^*)h, h\rangle > 0$ для всех ненулевых $h \in \overline{K(u_*)} \cap H(u_*)$, где $\overline{K(u_*)}$ — замыкание множества $K(u_*) = \{h \in E^n$: $h = \lambda(u - u_*)$, $\lambda > 0$, $u \in U_0\}$, $H(u_*)$ определено в упражнении 6. Тогда u_* — точка строгого локального минимума в задаче (1), (2) ([21, с. 141]).

§ 9. Теорема Куна — Таккера. Двойственная задача

1. Перейдем к рассмотрению условий оптимальности для задач выпуклого программирования. Под *выпуклым программированием* понимается раздел теории экстремальных задач, в котором изучаются задачи минимизации (или максимизации) выпуклых функций на выпуклых множествах. Точнее, под задачей выпуклого программирования понимается следующая задача:

$$J(u) \to \inf; \quad u \in U, \tag{1}$$

$$U = \{u \in U_0: g_i(u) \leqslant 0,\ i = 1, \ldots, m;$$
$$g_i(u) = 0,\ i = m+1, \ldots, s\}, \tag{2}$$

где U_0 — заданное выпуклое множество из E^n, функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ определены и выпуклы на U_0, а $g_i(u) = \langle a_i, u\rangle - b^i$ при $i = m+1, \ldots, s$ — линейные функции, b^i — заданные числа, a_i — заданные векторы из E^n. Здесь не исключаются возможности, когда отсутствуют либо ограничения

$g_i(u) \leqslant 0$ типа неравенств ($m=0$), либо ограничения $g_i(u)=0$ типа равенств ($s=m$), либо оба эти вида ограничений ($m=s=0$, $U=U_0$). При сделанных предположениях по теореме 2.11 множество (2) выпукло.

Важное место в теории выпуклого программирования занимает теорема о седловой точке функции Лагранжа, известная в литературе под названием теоремы Куна — Таккера в честь американских математиков Куна и Таккера, впервые сформулировавших и доказавших некоторые варианты этой теоремы. Эта теорема дает необходимое и достаточное условие оптимальности в задаче (1), (2), т. е. условие принадлежности той или иной точки множеству

$$U_* = \{u \in U\colon\ J(u) = \inf_{v \in U} J(v) = J_*\},$$

и выражает собой правило множителей Лагранжа для регулярной задачи (1), (2). Для формулировки теоремы Куна — Таккера введем функцию

$$L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i g_i(u), \tag{3}$$

называемую в отличие от (8.4) *регулярной функцией Лагранжа* задачи (1), (2), где $u \in U_0$, а переменные $\lambda = (\lambda_1, \ldots, \lambda_s)$ принадлежат множеству

$$\Lambda_0 = \{\lambda = (\lambda_1, \ldots, \lambda_s) \in E^s\colon\ \lambda_1 \geqslant 0, \ \ldots, \ \lambda_m \geqslant 0\}. \tag{4}$$

Определение 1. Точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ называют *седловой точкой* функции Лагранжа (3), если

$$L(u_*, \lambda) \leqslant L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \qquad \forall u \in U_0, \quad \lambda \in \Lambda_0. \tag{5}$$

Прежде чем переходить к выяснению связи между седловой точкой функции Лагранжа и решением задачи (1), (2), дадим другую равносильную (5) формулировку для седловой точки.

Лемма 1. *Для того чтобы точка $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ была седловой точкой функции Лагранжа, необходимо и достаточно, чтобы выполнялись следующие условия:*

$$L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \qquad \forall u \in U_0, \tag{6}$$

$$\lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, s, \quad u_* \in U. \tag{7}$$

Доказательство. Необходимость. Пусть $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка. Тогда условие (6) представляет собой правое неравенство (5). Остается получить условия (7). Для этого перепишем левое неравенство (5) с учетом конкретного вида (3) функции Лагранжа:

$$J(u_*) + \sum_{i=1}^{s} \lambda_i g_i(u_*) \leqslant J(u_*) + \sum_{i=1}^{s} \lambda_i^* g_i(u_*) \qquad \forall \lambda \in \Lambda_0. \tag{8}$$

Отсюда имеем

$$\sum_{i=1}^{s} (\lambda_i^* - \lambda_i)\, g_i(u_*) \geqslant 0 \qquad \forall \lambda \in \Lambda_0. \tag{9}$$

Покажем, что $u_* \in U$. Возьмем точку $\lambda = (\lambda_1, \ldots, \lambda_s)$, где $\lambda_j = \lambda_j + 1$ при некотором j $(1 \leqslant j \leqslant m)$ и $\lambda_i = \lambda_i^*$ при всех остальных $i = 1, \ldots, s$ $(i \neq j)$. Из определения (4) множества Λ_0 и из того, что $\lambda^* \in \Lambda_0$, следует, что выбранная точка $\lambda \in \Lambda_0$. Из (9) при таком λ получим $(-1)\, g_j(u_*) \geqslant 0$, т. е. $g_j(u_*) \leqslant 0$ при всех $j = 1, \ldots, m$.

Далее, пусть $\lambda = (\lambda_1, \ldots, \lambda_s)$ — точка с координатами $\lambda_j = \lambda_j^* + g_j(u_*)$ при некотором j $(m+1 \leqslant j \leqslant s)$ и $\lambda_i = \lambda_i^*$ при всех $i = 1, \ldots, s$ $(i \neq j)$. Ясно, что $\lambda \in \Lambda_0$. Поэтому из (9) имеем $-|g_j(u_*)|^2 \geqslant 0$, т. е. $g_j(u_*) = 0$ при всех $j = m+1, \ldots, s$. Таким образом, доказано, что $u_* \in U$.

Из того, что $g_i(u_*) = 0$ при $i = m+1, \ldots, s$, следует, что $\lambda_i^* g_i(u_*) = 0$ $(i = m+1, \ldots, s)$. Остается получить равенства (7) при $i = 1, \ldots, m$. Возьмем точку $\lambda = (\lambda_1, \ldots, \lambda_s)$ с координатами $\lambda_j = 0$ при некотором j $(1 \leqslant j \leqslant m)$ и $\lambda_i = \lambda_i^*$ при всех остальных $i = 1, \ldots, s$ $(i \neq j)$. Такая точка принадлежит Λ_0, поэтому из (9) получим $0 \leqslant \lambda_j^* g_j(u_*)$. Но $\lambda_j^* \geqslant 0$, $g_j(u_*) \leqslant 0$ при $j = 1, \ldots, m$, поэтому последнее неравенство возможно лишь при $\lambda_j^* g_j(u_*) = 0$ $(j = 1, \ldots, m)$. Все соотношения (6), (7) получены.

Достаточность. Пусть для некоторой точки $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ выполнены соотношения (6), (7). Покажем, что тогда (u_*, λ^*) — седловая точка. Из (6) следует правое неравенство (5). Остается доказать левое неравенство (5). По условию (7) $u_* \in U$, т. е. $g_i(u_*) \leqslant 0$ $(i = 1, \ldots, m)$, $g_i(u_*) = 0$ $(i = m+1, \ldots, s)$. Тогда

$$(\lambda_i^* - \lambda_i)\, g_i(u_*) = 0 \tag{10}$$

при всех $i = m+1, \ldots, s$ и всех тех i $(1 \leqslant i \leqslant m)$, для которых $g_i(u_*) = 0$. Если $g_i(u_*) < 0$ при некотором i $(1 \leqslant i \leqslant m)$, то из равенства (7) следует, что $\lambda_i^* = 0$. Поэтому $(\lambda_i^* - \lambda_i)\, g_i(u_*) = -\lambda_i g_i(u_*) \geqslant 0$ для всех $\lambda_i \geqslant 0$ $(1 \leqslant i \leqslant m)$, для которых $g_i(u_*) < 0$. Складывая полученные неравенства с (10), будем иметь $\sum_{i=1}^{s} (\lambda_i^* - \lambda_i)\, g_i(u_*) \geqslant 0$ для всех $\lambda \in \Lambda_0$. Отсюда $\sum_{i=1}^{s} \lambda_i g_i(u_*) \leqslant \sum_{i=1}^{s} \lambda_i^* g_i(u_*)$ при всех $\lambda \in \Lambda_0$. Добавляя к обеим частям этого неравенства $J(u_*)$, придем к неравенству (8), представляющему собой левое неравенство (5).

Если сделать дополнительные предположения о выпуклости и гладкости задачи (1), (2), то лемму 1 можно переформулировать в следующей так называемой *дифференциальной форме.*

Лемма 2. *Пусть* (1), (2) *представляет собой задачу выпуклого программирования и* $J(u)$, $g_1(u)$, ..., $g_m(u) \in C^1(U_0)$. *Тогда для того чтобы точка* $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ *была седловой точкой функции Лагранжа, необходимо и достаточно, чтобы*

$$\langle L_u(u_*, \lambda^*), u - u_* \rangle = \left\langle J'(u_*) + \sum_{i=1}^{s} \lambda_i^* g_i'(u_*), u - u_* \right\rangle \geqslant 0 \quad \forall u \in U_0, \tag{6'}$$

$$\lambda_i^* g_i(u_*) = 0, \quad i = 1, \ldots, s, \quad u_* \in U. \tag{7'}$$

Доказательство. При сделанных предположениях функция Лагранжа (3) выпукла и дифференцируема по переменной $u \in U_0$ при каждом $\lambda \in \Lambda_0$. Поэтому условие (6) согласно теореме 2.3 равносильно условию (6′).

Как видим, соотношения (6′), (7′) напоминают нам условия (8.5) — (8.7) при $\lambda_0 = 1$, а соотношениям (6), (7) соответствуют аналогичные (8.20) с $\lambda_0 = 1$. Эти аналогии подчеркивают тесную связь между правилом множителей Лагранжа из § 8 и следующими ниже теоремами.

Теперь выясним, как связаны между собой седловая точка функции Лагранжа и решение задачи (1), (2).

Теорема 1. *Пусть* $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — *седловая точка функции Лагранжа. Тогда* $u_* \in U_0$, $J_* = L(u_*, \lambda^*) = J(u_*)$, *т. е.* u_* *является решением задачи* (1), (2).

Доказательство. Из условия (7) имеем $u_* \in U$ и $L(u_*, \lambda^*) = J(u_*)$. Тогда неравенство (6) перепишется в виде

$$J(u_*) \leqslant L(u, \lambda^*) = J(u) + \sum_{i=1}^{s} \lambda_i^* g_i(u), \quad u \in U_0. \tag{11}$$

В частности, (11) верно и для всех $u \in U$. Но $\sum_{i=1}^{s} \lambda_i^* g_i(u) \leqslant 0$ при $u \in U$, так как тогда $g_i(u) \leqslant 0$ и $\lambda_i^* \geqslant 0$ при $i = 1, \ldots, m$, так что $\lambda_i^* g_i(u) \leqslant 0$ $(i = 1, \ldots, m)$ и $g_i(u) = 0$ при $i = m+1, \ldots, s$, так что $\lambda_i g_i(u) = 0$ $(i = m+1, \ldots, s)$. Поэтому из (11) следует, что $J(u_*) \leqslant L(u, \lambda^*) \leqslant J(u)$ при всех $u \in U$, т. е. $u_* \in U_*$.

Заметим, что теорема 1, как и лемма 1, доказаны без каких-либо ограничений на функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, s)$ и на множество U_0; в частности, никакие предположения о выпуклости, сделанные выше при формулировке задачи (1), (2), мы пока не использовали.

2. Возникает вопрос: во всякой ли задаче вида (1), (2) функция Лагранжа имеет седловую точку? Ответ здесь, конечно, от-

рицательный: если в задаче (1), (2) $U_* = \varnothing$, то, как следует из теоремы 1, функция Лагранжа такой задачи не может иметь седловую точку. Более того, даже в задачах выпуклого программирования (1), (2) с $U_* \neq \varnothing$ в общем случае нельзя ожидать, что функция Лагранжа будет иметь седловую точку.

Пример 1. Рассмотрим задачу из примера 8.6: $J(u) = -u \to \inf$ $(u \in U = \{u \in U_0\colon g(u) = u^2 \leqslant 0\})$, где $U_0 = \{u \in E^1\colon 0 \leqslant u \leqslant a\}$, $0 < a \leqslant \infty$. Здесь множество U_0 выпукло, функции $J(u)$, $g(u)$ выпуклы на U_0. Множество U состоит из одной точки $u = 0$, так что $J_* = J(0) = 0$, $U_* = \{0\}$. Функция Лагранжа

$$L(u, \lambda) = -u + \lambda u^2, \qquad 0 \leqslant u \leqslant a, \qquad \lambda \geqslant 0,$$

рассматриваемой задачи не имеет седловой точки.

Таким образом, для существования седловой точки на задачу (1), (2), кроме условий выпуклости, должны быть наложены какие-то дополнительные ограничения. Начнем с рассмотрения случая, когда в (2) ограничения типа равенств отсутствуют $(m = s)$, т. е. множество U имеет вид

$$U = \{u \in U_0\colon g_i(u) \leqslant 0, \quad i = 1, \ldots, m\}. \tag{12}$$

Определение 2. Множество (12) называют *регулярным*, если существует точка $\bar{u} \in U$ такая, что

$$g_1(\bar{u}) < 0, \ \ldots, \ g_m(\bar{u}) < 0. \tag{13}$$

Если U_0 — выпуклое множество, функции $g_i(u)$ выпуклы на U_0, то вместо (13) достаточно потребовать для каждого i существования точки $\bar{u}_i \in U$ такой, что $g_i(\bar{u}_i) < 0$ $(i = 1, \ldots, m)$. Тогда в качестве $\bar{u}$ из (13) можно взять $\bar{u} = \sum_{i=1}^{m} \alpha_i \bar{u}_i$, $\alpha_i > 0$, $\alpha_1 + \alpha_2 + \ldots + \alpha_m = 1$, поскольку $\bar{u} \in U_0$ и в силу неравенства (2.2) $g_j(\bar{u}) \leqslant \sum_{i=1}^{m} \alpha_i g_j(\bar{u}_i) \leqslant \alpha_j g_j(\bar{u}_j) < 0$ $(j = 1, \ldots, m)$. Условие (13) часто называют *условием Слейтера*.

Для наших целей подошло бы и несколько иное более общее, чем (13), определение регулярного множества: множество (12) назовем *регулярным*, если для любого вектора $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*) \neq 0$ $(\lambda_1^* \geqslant 0, \ldots, \lambda_m^* \geqslant 0)$ существует такая точка $\bar{u} = \bar{u}(\lambda^*)$, что

$$\bar{u} \in U, \quad \langle \lambda^*, g(\bar{u}) \rangle = \sum_{i=1}^{m} \lambda_i^* g_i(\bar{u}) < 0. \tag{14}$$

Существуют и другие определения регулярности множеств вида (12), (2), используемые в теоремах Куна — Таккера [24, 41, 299].

Теорема 2. *Пусть множество U_0 выпукло, функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, m)$ выпуклы на U_0, а множество* (12) *регуляр-*

но. Пусть множество U_* точек минимума функции $J(u)$ на множестве (12) *непусто. Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа* $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*) \in \Lambda_0 = \{\lambda \in E^m\colon \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\}$ *такие, что пара* (u_*, λ^*) *образует седловую точку функции Лагранжа на множестве* $U_0 \times \Lambda_0$.

Доказательство. В пространстве E^{m+1} переменных $a = (a_0, a_1, \ldots, a_m)$ введем множества

$$A = \{a = (a_0, a_1, \ldots, a_m) \in E^{m+1}\colon a_0 \geqslant J(u), \\ a_1 \geqslant g_1(u), \ldots, a_m \geqslant g_m(u),\ u \in U_0\},$$

$$B = \{b = (b_0, b_1, \ldots, b_m) \in E^{m+1}\colon b_0 < J_*,\ b_1 < 0, \ldots, b_m < 0\}.$$

Покажем, что A и B не имеют общих точек. В самом деле, пусть $a \in A$. Тогда найдется точка $u \in U_0$ такая, что $a_0 \geqslant J(u)$, $a_1 \geqslant g_1(u)$, ..., $a_m \geqslant g_m(u)$. Возможно, что $u \in U$. Тогда $a_0 \geqslant J(u) \geqslant J_*$ и заведомо $a \notin B$. Если же $u \in U_0 \backslash U$, то найдется номер i $(1 \leqslant i \leqslant m)$ такой, что $g_i(u) > 0$. Тогда $a_i \geqslant g_i(u) > 0$ и снова $a \notin B$. Итак, $A \cap B = \varnothing$.

Далее, нетрудно видеть, что A и B — выпуклые множества. Покажем, например, что A выпукло. Пусть a, c — две произвольные точки из A. Тогда существуют точки u, $v \in U_0$ такие, что $a_0 \geqslant J(u)$, $c_0 \geqslant J(v)$, $a_i \geqslant g_i(u)$, $c_i \geqslant g_i(v)$ $(i = 1, \ldots, m)$. Возьмем произвольное $\alpha \in [0, 1]$ и положим $a_\alpha = \alpha a + (1 - \alpha) c$, $u_\alpha = \alpha u + (1 - \alpha) v$. Из выпуклости U_0 следует $u_\alpha \in U_0$. Далее, из выпуклости функций $J(u)$, $g_i(u)$ имеем

$$J(u_\alpha) \leqslant \alpha J(u) + (1 - \alpha) J(v) \leqslant \alpha a_0 + (1 - \alpha) c_0,$$
$$g_i(u_\alpha) \leqslant \alpha g_i(u) + (1 - \alpha) g_i(v) \leqslant \alpha a_i + (1 - \alpha) c_i, \quad i = 1, \ldots, m.$$

Это означает, что $a_\alpha \in A$. Выпуклость A доказана. Аналогично доказывается выпуклость B.

В силу теоремы 5.2 тогда существует гиперплоскость $\langle c, a\rangle = \gamma$ с нормальным вектором $c = (\lambda_0^*, \lambda_1^*, \ldots, \lambda_m^*) \neq 0$, отделяющая A и B, а также A и $\overline{B} = \{b = (b_0, b_1, \ldots, b_m) \in E^{m+1}\colon b_0 \leqslant J_*, b_1 \leqslant 0, \ldots, b_m \leqslant 0\}$. Это значит, что

$$\langle c, b\rangle = \sum_{i=0}^{m} \lambda_i^* b_i \leqslant \gamma \leqslant \langle c, a\rangle = \sum_{i=0}^{m} \lambda_i^* a_i \quad \forall a \in A, \quad b \in \overline{B}. \tag{15}$$

Заметим, что $y = (J_*, 0, \ldots, 0) \in A \cap \overline{B}$. В самом деле, возьмем какую-либо точку $u_* \in U_*$. Тогда $J(u_*) = J_*$, $g_i(u_*) \leqslant 0$ $(i = 1, \ldots, m)$, что означает $y \in A$. Включение $y \in \overline{B}$ очевидно. Тогда по теореме 5.2 величина γ из (15) равна $\gamma = \langle c, y\rangle = \lambda_0^* J_*$, и (15) можно переписать в виде

$$\lambda_0^* b_0 + \sum_{i=1}^{m} \lambda_i^* b_i \leqslant \lambda_0^* J_* \leqslant \lambda_0^* a_0 + \sum_{i=1}^{m} \lambda_i^* a_i \quad \forall a \in A, \quad b \in \overline{B}. \tag{16}$$

Возьмем точку $b=(J-1,\ 0,\ \ldots,\ 0)\in\bar{B}$. Из левого неравенства (16) получим $\lambda_0^*(J_*-1)\leqslant\lambda_0^*J_*$, откуда $\lambda_0^*\geqslant 0$. Далее, беря $b=(J_*,\ 0,\ \ldots,\ 0,\ -1,\ 0,\ \ldots,\ 0)$, из левого неравенства (16) имеем $\lambda_0^*J_*-\lambda_i^*\leqslant\lambda_0^*J_*$, т. е. $\lambda_i^*\geqslant 0$ $(i=1,\ \ldots,\ m)$. Таким образом, показано, что $\lambda^*=(\lambda_1^*,\ \ldots,\ \lambda_m^*)\geqslant 0$, $\lambda_0^*\geqslant 0$.

Далее, возьмем произвольную точку $u_*\in U_*$. Тогда $a=(J(u_*)=J_*,\ 0,\ \ldots,\ 0,\ g_i(u_*),\ 0,\ \ldots,\ 0)\in A\cap\bar{B}$. Подставляя эту точку в левое и правое неравенства (16), получаем $\lambda_0^*J_*+\lambda_i^*g_i(u_*)\leqslant\lambda_0^*J_*\leqslant\lambda_0^*J_*+\lambda_i^*g_i(u_*)$, откуда $\lambda_i^*g_i(u_*)\leqslant 0\leqslant\lambda_i^*g_i(u_*)$ или $\lambda_i^*g(u_*)=0$ $(i=1,\ \ldots,\ m)$. Равенства (7) доказаны.

Покажем, что $\lambda_0^*>0$. В самом деле, в (16) подставим $a=(J(\bar{u}),\ g_1(\bar{u}),\ \ldots,\ g_m(\bar{u}))\in A$, где $\bar{u}$ взято из (13) или (14). Получим $\lambda_0^*J_*\leqslant\lambda_0^*J(\bar{u})+\sum\limits_{i=0}^{m}\lambda_i^*g_i(\bar{u})$. Допустим, что $\lambda_0^*=0$. Поскольку не все λ_i^* $(i=0,1,\ldots,m)$ равны нулю, то $\lambda^*\neq 0$. Тогда из предыдущего неравенства при $\lambda_0^*=0$ с учетом условия (13) или (14) имеем $0\leqslant\sum\limits_{i=1}^{m}\lambda_i^*g_i(\bar{u})<0$. Полученное противоречие показывает, что $\lambda_0^*>0$. Неравенство (16) и все последующие рассуждения сохраняют силу, если (16) поделить на $\lambda_0^*>0$. Это значит, что в (16) можно принять $\lambda_0^*=1$.

Наконец, возьмем произвольную точку $u\in U_0$. Тогда $a=(J(u),\ g_1(u),\ \ldots,\ g_m(u))\in A$. Подставим эту точку в правое неравенство (16). С учетом того, что $\lambda_0^*=1$, получим $J_*\leqslant J(u)+\sum\limits_{i=1}^{m}\lambda_i^*g_i(u)=L(u,\ \lambda^*)$ $(u\in U_0)$. Но в силу (7) $J_*=L(u_*,\ \lambda^*)$ при любом выборе $u_*\in U_*$. Отсюда и из предыдущего неравенства следует условие (6). Согласно лемме 1 тогда $(u_*,\ \lambda^*)$ — седловая точка.

3. Приведенный выше пример 1 показывает, что без дополнительного требования регулярности множества теорема 2, вообще говоря, неверна. Однако если (2) — многогранное множество, то, оказывается, теорема о существовании седловой точки будет верна без каких-либо дополнительных условий типа условий регулярности. Важную роль при установлении этого факта, а также во многих других вопросах выпуклого анализа играет следующая теорема, известная в литературе под названием *теоремы Фаркаша.*

Теорема 3. *Пусть дано множество*

$$K=\{e\in E^n:\ \langle c_i,\ e\rangle\leqslant 0,\quad i=1,\ \ldots,\ m;\ \langle c_i,\ e\rangle<0,\quad i=m+1,\ \ldots,\ p;\ \langle c_i,\ e\rangle=0,\ i=p+1,\ \ldots,\ s\},\qquad(17)$$

где $c_1,\ \ldots,\ c_s$ — некоторые векторы из E^n. Тогда для того чтобы

некоторый вектор $c \in E^n$ *удовлетворял неравенству*

$$\langle c, e\rangle \geqslant 0 \qquad \forall e \in K, \tag{18}$$

необходимо и достаточно, чтобы существовали такие числа $\lambda_1, \ldots, \lambda_s$ $(\lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0)$, *что*

$$c = -\lambda_1 c_1 - \ldots - \lambda_s c_s. \tag{19}$$

Нетрудно видеть, что K — конус с вершиной в нуле, а всякий вектор c, удовлетворяющий условию (18), принадлежит двойственному конусу K^* (см. определения 5.4, 5.5). Поэтому теорему Фаркаша можно переформулировать на языке конусов следующим образом.

Теорема 3. *Пусть* K — *конус, определенный условиями* (17). *Тогда двойственный ему конус имеет вид*

$$K^* = \left\{c \in E^n\colon\ c = -\sum_{i=1}^{s} \lambda_i c_i,\ \ \lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0\right\}. \tag{19}$$

Заметим, что в (17) не исключаются случаи, когда какие-либо виды ограничений (ограничения типа нестрогого или строгого неравенства, или типа равенства) отсутствуют, т. е. возможно $m = 0$ при $m = p$, или $p = s$, или $m = p = 0$, или $m = 0$, $p = s$ или $m = p = s$.

Для доказательства теоремы 3 нам понадобится

Лемма 3. *Пусть* $a_1, \ldots, a_p$ — *некоторое конечное множество векторов из* E^n. *Тогда*

$$Q = \left\{a \in E^n\colon\ a = \sum_{i=1}^{p} \alpha_i a_i, \quad \alpha_i \geqslant 0, \quad i = 1, \ldots, p\right\}$$

есть выпуклый замкнутый конус.

Доказательство. Если $a \in Q$, то $\lambda a = \sum_{i=1}^{p} \lambda\alpha_i a_i\ (\lambda\alpha_i \geqslant 0,$ $i = 1, \ldots, p)$ при любом $\lambda > 0$. Это значит, что Q — конус. Далее, пусть a, b — две произвольные точки на Q, т. е. $a = \sum_{j=1}^{p} \alpha_i a_i$, $b = \sum_{i=1}^{p} \mu_i a_i$ $(\alpha_i \geqslant 0,\ \mu_i \geqslant 0,\ i = 1, \ldots, p)$. Тогда для всех $\alpha \in [0, 1]$

$$\alpha a + (1-\alpha) b = \sum_{i=1}^{p} [\alpha\alpha_i + (1-\alpha)\mu_i]\, a_i,$$

$$\alpha\alpha_i + (1-\alpha)\mu_i \geqslant 0, \quad i = 1, \ldots, p,$$

так что Q — выпуклый конус. Кстати, тогда из a, $b \in Q$ следует $a + b \in Q$. В самом деле, в силу выпуклости конуса Q имеем $a + b = \alpha(a/\alpha) + (1-\alpha)(b/(1-\alpha)) \in Q$ при всех α $(0 < \alpha < 1)$, поскольку a/α, $b/(1-\alpha) \in Q$.

Замкнутость Q докажем индукцией по числу образующих векторов $a_1, \ldots, a_p$. При $p=1$ $Q=\{a \in E^n\colon\ a=\alpha a_1,\ \alpha \geqslant 0\}$ — полупрямая (луч) — замкнутое множество. Пусть известно, что конус с любыми $p-1$ образующими замкнут. Покажем, что тогда конус Q с p образующими $a_1, \ldots, a_p$ также замкнут. Рассмотрим два случая.

1. Конус Q наряду с векторами $a_1, \ldots, a_p$ содержит также и векторы $-a_1, \ldots, -a_p$. Тогда наряду с точками $a=\sum\limits_{i=1}^{p} \alpha_i a_i$ $(\alpha_i \geqslant 0,\ i=1, \ldots, p)$ выпуклый конус Q содержит все точки $b=\sum\limits_{i=1}^{p} \beta_i(-a_i)$ $(\beta_i \geqslant 0,\ i=1, \ldots, p)$, а также точки $a+b$. Покажем, что в рассматриваемом случае Q является подпространством (линейной оболочкой), натянутым на векторы $a_1, \ldots, a_p$.

В самом деле, пусть $d=\sum\limits_{i=1}^{p} \gamma_i a_i$, где γ_i $(i=1, \ldots, p)$ — произвольные числа. Положим $\alpha_i=\max\{0;\ \gamma_i\}$, $\beta_i=\max\{0;\ -\gamma_i\}$. Тогда $\gamma_i=\alpha_i-\beta_i$ $(i=1, \ldots, p)$. Поэтому $d=\sum\limits_{i=1}^{p} \gamma_i a_i=\sum\limits_{i=1}^{p} \alpha_i a_i+$ $+\sum\limits_{i=1}^{p} \beta_i(-a_i) \in Q$ при всех действительных γ_i $(i=1, \ldots, p)$. Таким образом, Q — подпространство, натянутое на векторы $a_1, \ldots, a_p$, размерность которого не превышает числа $\min\{p;\ n\}$. Но в конечномерном пространстве E^n любое подпространство замкнуто [93]. Следовательно, Q замкнут.

2. Хотя бы один из векторов $-a_1, \ldots, -a_p$ не принадлежит Q. Пусть для определенности $-a_p \notin Q$. Обозначим $Q_{p-1}=$ $=\left\{b \in E^n\colon\ b=\sum\limits_{i=1}^{p-1} \alpha_i a_i,\ \alpha_i \geqslant 0,\ i=1, \ldots, p-1\right\}$ — это конус, порожденный векторами $a_1, \ldots, a_{p-1}$. По предположению индукции конус Q_{p-1} замкнут. Далее, из представления $a=\sum\limits_{i=1}^{p-1} \alpha_i a_i+\alpha a_p$ $\left.(\alpha_i \geqslant 0,\ i=1, \ldots, p-1,\ \alpha \geqslant 0\right\}$ следует, что

$$Q=Q_{p-1}+\alpha a_p, \quad \alpha \geqslant 0. \tag{20}$$

Пусть d — какая-либо предельная точка множества Q. Тогда существует последовательность $\{d_m\} \in Q$, сходящаяся к d. Из (20) следует существование $b_m \in Q_{p-1}$ и чисел $\mu_m \geqslant 0$ таких, что $d_m=b_m+\mu_m a_p$ $(m=1, 2, \ldots)$. Покажем, что $\{\mu_m\}$ ограничена сверху. Допустим, что $\{\mu_m\}$ не ограничена сверху. Тогда существует подпоследовательность $\{\mu_{m_k}\} \to \infty$. Поскольку $\{d_m\}$, как сходящаяся последовательность, ограничена, то $d_{m_k}/\mu_{m_k} \to 0$ при

$k \to \infty$. А тогда $\{b_{m_k}/\mu_{m_k} = (d_{m_k}/\mu_{m_k}) - a_p\}$ имеет предел при $k \to \infty$, равный $-a_p$. Но так как $\{b_{m_k}/\mu_{m_k}\} \in Q_{p-1}$, а конус Q_{p-1} замкнут, то предел $-a_p$ будет принадлежать Q_{p-1}. Из (20) при $\alpha = 0$ следует $Q_{p-1} \subseteq Q$, так что $-a_p \in Q$. Получили противоречие с рассматриваемым случаем.

Это означает, что $0 \leqslant \mu_m \leqslant \text{const}$ $(m = 1, 2, \ldots)$. Тогда существует подпоследовательность $\{\mu_{m_k}\}$, сходящаяся к некоторому числу $\mu \geqslant 0$. Из равенства $b_{m_k} = d_{m_k} - \mu_{m_k} a_p$ $(k = 1, 2, \ldots)$ при $k \to \infty$ следует, что предел $b = \lim\limits_{k\to\infty} b_{m_k}$ существует, причем $b = d - \mu a_p \in Q_{p-1}$ в силу замкнутости Q_{p-1}. Таким образом, показано, что $d = b + \mu a_p$, где $b \in Q_{p-1}$, $\mu \geqslant 0$. Из (20) тогда следует, что $d \in Q$. Замкнутость Q доказана.

Доказательство теоремы 3. Введем множество

$$Q = \left\{c \in E^n\colon c = -\sum_{i=1}^{s} \lambda_i c_i,\ \lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0\right\}. \tag{21}$$

Заметим, что Q — конус, порожденный векторами $-c_1, \ldots, -c_p, -c_{p+1}, \ldots, -c_s, c_{p+1}, \ldots, c_s$. В самом деле, с одной стороны, все точки $c = \sum\limits_{i=1}^{p} \lambda_i(-c_i) + \sum\limits_{i=p+1}^{s} \alpha_i(-c_i) + \sum\limits_{i=p+1}^{s} \beta_i c_i$ $(\lambda_i \geqslant 0, i = 1, \ldots, p;\ \alpha_i \geqslant 0,\ \beta_i \geqslant 0,\ i = p+1, \ldots, s)$ принадлежат Q. С другой стороны, любая точка $c = -\lambda_1 c_1 - \ldots - \lambda_s c_s$ представима в виде предыдущего равенства при $\alpha_i = \max\{0;\ \lambda_i\}$, $\beta_i = \max\{0;\ -\lambda_i\}$ $(i = m+1, \ldots, s)$, так как $\lambda_i = \alpha_i - \beta_i$. В силу леммы 3 тогда множество (21) является выпуклым замкнутым множеством.

Для доказательства теоремы 3 достаточно установить, что $K^* = Q$. Возьмем произвольный вектор $c \in Q$, т. е. $c = -\sum\limits_{i=1}^{s} \lambda_i c_i$ $(\lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0)$. Тогда для любого $e \in K$ имеем $\langle c, e\rangle = -\sum\limits_{i=1}^{s} \lambda_i \langle c_i, e\rangle \geqslant 0$. Это значит, что $c \in K^*$. Тем самым показано, что $Q \subseteq K^*$.

Покажем обратное включение $K^* \subseteq Q$. Предположим противное: пусть существует $y \in K^*$, но $y \notin Q$. Поскольку Q — замкнутое выпуклое множество, то по теореме 5.1 это множество сильно отделимо от точки y. Это означает, что существует гиперплоскость $\langle d, u - y\rangle = 0$ с нормальным вектором $d \neq 0$ такая, что $\langle d, c - y\rangle > 0$ или $\langle d, c\rangle > \langle d, y\rangle$ для всех $c \in Q$. Согласно (21) это значит, что

$$\left\langle d, -\sum_{j=1}^{s} \lambda_j c_j \right\rangle = -\sum_{j=1}^{s} \lambda_j \langle c_j, d\rangle > \langle d, y\rangle \tag{22}$$

при всех $\lambda_1, \ldots, \lambda_s$, лишь бы $\lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0$. Покажем, что тогда $d \in \overline{K}$ — замыкание K.

Зафиксируем некоторый номер i, $1 \leqslant i \leqslant p$, и в (22) примем $\lambda_j = 0$ при всех $j \neq i$. Получим $-\lambda_i \langle c_i, d\rangle > \langle d, y\rangle$ для любого $\lambda_i \geqslant 0$. Отсюда, деля на $\lambda_i > 0$ и устремляя $\lambda_i \to \infty$, получаем $\langle c_i, d\rangle \geqslant 0$ или $\langle c_i, d\rangle \leqslant 0$ $(i = 1, \ldots, p)$. Далее, зафиксируем номер i $(p+1 \leqslant i \leqslant s)$ и в (22) примем $\lambda_j = 0$ при всех $j \neq i$, $\lambda_i = t\langle c_i, d\rangle$ $(t > 0)$. Получим $-t|\langle c_i, d\rangle|^2 > \langle d, y\rangle$. Отсюда, деля на $t > 0$ и устремляя $t \to \infty$, получаем $-|\langle c_i, d\rangle|^2 \geqslant 0$ или $\langle c_i, d\rangle = 0$ $(i = p+1, \ldots, s)$. Таким образом, показано, что $d \in \overline{K}$.

Теперь вспомним, что $y \in K^*$. Это значит, что $\langle y, e\rangle \geqslant 0$ для всех $e \in K$. Отсюда с помощью предельного перехода нетрудно получить, что $\langle y, e\rangle \geqslant 0$ для всех $e \in \overline{K}$. В частности, для $d \in \overline{K}$ имеем $\langle y, d\rangle \geqslant 0$. Но, с другой стороны, если в (22) принять $\lambda_i = 0$ $(i = 1, \ldots, s)$, то получим $\langle y, d\rangle < 0$. Пришли к противоречию. Следовательно, $K^* \subseteq Q$. Сравнивая с ранее доказанным включением $Q \subseteq K^*$, заключаем, что $K^* = Q$. Равенство (19) и, тем самым, теорема 3 доказаны.

4. Рассмотрим задачу минимизации функции $J(u)$ на множестве

$$U = \{u \in U_0\colon\ g_i(u) = \langle a_i, u\rangle - b^i \leqslant 0, \quad i = 1, \ldots, m;$$
$$g_i(u) = \langle a_i, u\rangle - b^i = 0, \quad i = m+1, \ldots, s\}, \qquad (23)$$

где $U_0 = \{u \in E^n\colon\ \langle d_i, u\rangle \leqslant f^i,\ i = 1, \ldots, p;\ \langle d_i, u\rangle = f^i,\ i = p+1, \ldots, q\}$ — многогранное множество, a_i, $d_i \in E^n$ — заданные векторы, b^i, f^i — заданные числа. В частности, здесь может быть $U_0 = E^n_+$; $U_0 = \{u = (u^1, \ldots, u^n)\colon\ u^i \geqslant 0,\ i \in I\}$, I — некоторое подмножество номеров $\{1, \ldots, n\}$; $U_0 = \{u = (u^1, \ldots, u^n)\colon\ \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\}$, α_i, β_i — заданные величины, $\alpha_i \leqslant \beta_i$, причем, возможно, некоторые $\alpha_i \to -\infty$, $\beta_j \to \infty$.

Теорема 4. *Пусть функция $J(u)$ выпукла на U_0, $J(u) \in C^1(U_0)$, множество U определено согласно* (23), $U_* \{u \in U\colon J(u) = \inf\limits_{v \in U} J(v) = J_* > -\infty\} \neq \varnothing$. *Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*) \in \Lambda_0 = \{\lambda = (\lambda_1, \ldots, \lambda_s) \in E^s\colon\ \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\}$ такие, что пара (u_*, λ^*) образует седловую точку функции Лагранжа на множестве $U_0 \times \Lambda_0$.*

Из этой теоремы, в частности, следует, что в любой задаче линейного программирования, имеющей решение, функция Лагранжа всегда имеет седловую точку.

Доказательство. В силу теоремы 2.11 множество (23) выпукло. Возьмем любую точку $u_* \in U_*$. Введем множества индексов

$$I_1^* = \{i\colon 1 \leqslant i \leqslant m,\ \langle a_i, u_*\rangle = b^i\}, \quad I_2^* = \{i\colon 1 \leqslant i \leqslant p,$$
$$\langle d_i, u_*\rangle = f^i\}$$

и составим конус

$$K = \{e \in E^n\colon \langle a_i, e\rangle \leqslant 0,\ i \in I_1^*,\ \langle a_i, e\rangle = 0,\ i = m+1, \ldots, s;$$
$$\langle d_i, e\rangle \leqslant 0,\ i \in I_2^*,\ \langle d_i, e\rangle = 0,\ i = p+1, \ldots, q;\ e \neq 0\}. \quad (24)$$

Покажем, что множество возможных направлений множества (23) в точке u_* совпадает с конусом (24).

Пусть $e = (e^1, \ldots, e^n) \neq 0$ — произвольное возможное направление в точке u_*. Согласно определению 2.3 тогда существует такое число $t_0 > 0$, что $u = u_* + te \in U$ или

$$\begin{aligned} &\langle a_i, u_* + te\rangle \leqslant b^i, \quad i = 1, \ldots, m; \quad \langle a_i, u_* + te\rangle = b^i, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = m+1, \ldots, s; \\ &\langle d_i, u_* + te\rangle \leqslant f^i, \quad i = 1, \ldots, p; \quad \langle d_i, u_* + te\rangle = f^i, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = p+1, \ldots, q, \end{aligned} \quad (25)$$

при всех t $(0 < t \leqslant t_0)$. С учетом $u_* \in U_* \subset U$ из (25) сразу получаем $e \in K$. Верно и обратное: если $e \in K$, то e — возможное направление в точке u_*. В самом деле, пусть $e \in K$. Тогда для $i \in I_1^*$ имеем $\langle a_i, u_* + te\rangle = b^i + t\langle a_i, e\rangle \leqslant b^i$ при всех $t \geqslant 0$, а если $i \notin I_1^*$ $(1 \leqslant i \leqslant m)$, то $\langle a_i, u_*\rangle < b^i$ и найдется такое $t_0 > 0$, что $\langle a_i, u_* + te\rangle \leqslant b^i$ при $0 \leqslant t \leqslant t_0$. Если $m+1 \leqslant i \leqslant s$, то $\langle a_i, u_* + te\rangle = b^i$ при всех t. Аналогично, взяв при необходимости $t_0 > 0$ еще меньшим, убедимся, что выполняются и остальные соотношения (25), так что $u_* + te \in U$ $(0 \leqslant t \leqslant t_0)$. Тем самым показано, что для множества (23) множество возможных направлений в точке u_* совпадает с конусом (24).

Согласно теореме 2.3 для того, чтобы $u_* \in U_*$, необходимо и достаточно выполнения неравенства

$$\langle J'(u_*), u - u_*\rangle \geqslant 0 \qquad \forall u \in U. \quad (26)$$

Возьмем любое $e \in K$. Тогда $u = u_* + te \in U$ $(0 \leqslant t \leqslant t_0,\ t_0 > 0)$. Подставим такую точку u в (26). Получим $\langle J'(u_*), e\rangle t \geqslant 0$ или $\langle J'(u_*), e\rangle \geqslant 0$ при всех $e \in K$. Это значит, что $J'(u_*) \in K^*$. По теореме 3 тогда найдутся числа $\lambda_i^* \geqslant 0$ $(i \in I_1^*)$, $\lambda_{m+1}^*, \ldots, \lambda_s^*$, $\mu_i^* \geqslant 0$ $(i \in I_2^*)$, $\mu_{p+1}^*, \ldots, \mu_q^*$ такие, что

$$J'(u_*) = -\sum_{i \in I_1^*} \lambda_i^* a_i - \sum_{i=m+1}^{s} \lambda_i^* a_i - \sum_{i \in I_2^*} \mu_i^* d_i - \sum_{i=p+1}^{q} \mu_i^* d_i. \quad (27)$$

Если доопределим $\lambda_i^* = 0$ при $i \in \{1, \ldots, m\} \setminus I_1^*$, то получим точку $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*) \in \Lambda_0$. Отсюда, учитывая определение множества I_1^* и условие $u_* \in U_* \subseteq U$, имеем

$$\lambda_i^* (\langle a_i, u_*\rangle - b^i) = \lambda_i^* g_i(u_*) = 0, \quad i = 1, \ldots, s, \quad (28)$$

а равенство (27) можем переписать в виде

$$J'(u_*) + \sum_{i=1}^{s} \lambda_i^* a_i = -\sum_{i \in I_2^*} \mu_i^* d_i - \sum_{i=p+1}^{q} \mu_i^* d_i. \tag{29}$$

Функция Лагранжа в рассматриваемой задаче такая:

$$L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i (\langle a_i, u\rangle - b^i), \quad u \in U_0, \quad \lambda \in \Lambda_0.$$

Тогда, используя неравенство $J(u) - J(u_*) \geqslant \langle J'(u_*), u - u_*\rangle$ $(u \in U)$ (см. теорему 2.2), определение множества I_2^*, условие $\mu_i^* \geqslant 0$ $(i \in I_2^*)$ и равенство (29), для каждого $u \in U_0$ получаем

$$L(u, \lambda^*) - L(u_*, \lambda^*) = J(u) - J(u_*) + \sum_{i=1}^{s} \lambda_i^* \langle a_i, u - u_*\rangle \geqslant$$

$$\geqslant \left\langle J'(u_*) + \sum_{i=1}^{s} \lambda_i^* a_i, u - u_* \right\rangle = -\sum_{i \in I_2^*} \mu_i^* \langle d_i, u - u_*\rangle -$$

$$- \sum_{i=p+1}^{q} \mu_i^* \langle d_i, u - u_*\rangle = -\sum_{i \in I_2^*} \mu_i^* (\langle d_i, u\rangle - f^i) \geqslant 0,$$

или

$$L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \qquad \forall u \in U_0.$$

Отсюда и из (28) с помощью леммы 1 заключаем, что (u_*, λ^*) — седловая точка функции Лагранжа.

5. Наконец, приведем следующий более общий вариант теоремы Куна — Таккера.

Теорема 5. *Пусть U_0 — выпуклое множество из E^n, функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, m)$ выпуклы на U_0, $g_i(u) = \langle a_i, u\rangle - b^i$ $(i = m+1, \ldots, s)$ — линейные функции. Пусть множество U_* точек минимума функции $J(u)$ на множестве*

$$U = \{u \in U_0:\ g_i(u) \leqslant 0,\ i = 1, \ldots, m;$$
$$g_i(u) = \langle a_i, u\rangle - b^i \leqslant 0,\ i = m+1, \ldots, p;$$
$$g_i(u) = \langle a_i, u\rangle - b^i = 0,\ i = p+1, \ldots, s\}$$

непусто. Кроме того, пусть выполняется хотя бы одно из следующих условий:

а) *U_0 — многогранное множество, функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ выпуклы на выпуклом множестве W, открытом в* $\operatorname{aff} W$ (т. е. $W = \operatorname{ri} W$), *$U_0 \subset W$ и существует точка $\bar{u} \in U$ такая, что $g_i(\bar{u}) < 0$ $(i = 1, \ldots, m)$;*

б) *существует точка $\bar{u} \in \operatorname{ri} U_0 \cap U$ такая, что $g_i(\bar{u}) < 0$ $(i = 1, \ldots, m)$.*

Тогда для каждой точки $u_* \in U_*$ *необходимо существуют множители Лагранжа* $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*) \in \Lambda_0 = \{\lambda \in E^n\colon \lambda_1 \geqslant 0, \ldots, \lambda_p \geqslant 0\}$ *такие, что пара* (u_*, λ^*) *образует седловую точку функции Лагранжа на множестве* $U_0 \times \Lambda_0$.

В этой теореме не исключаются возможности, когда отсутствуют какие-либо из ограничений $g_i(u) \leqslant 0$ или $g_i(u) = 0$, т. е. $p = 0$ или $m = 0$, или $s = 0$, или $m = p$, или $m = s$, или $p = m = s$. Доказательство теоремы 5 требует весьма тонкого использования теоремы отделимости 5.2; за подробностями отсылаем читателя к [21] (ср. с [264, § 28]).

Условия а), б) теоремы 5 представляют собой обобщения условия регулярности (13) на случай более общей задачи (1), (2). Нарушение этих условий может привести к отсутствию седловой точки.

Пример 2. Пусть $U_0 = \{u = (x, y) \in E^2\colon x \geqslant 0,\ y \geqslant 0\} = E^2_+$, $J(u) = \sqrt{xy}$, $g(u) = x$, $U = \{u \in U_0\colon g(u) \leqslant 0\}$. Здесь U_0 выпукло, $J(u)$, $g(u)$ выпуклы на U_0, $J_* = 0$, $U_* = U = \{u_* = (0, y), y \geqslant 0\}$. Функция Лагранжа $L(u, \lambda) = -\sqrt{xy} + \lambda x$ $(x \geqslant 0,\ y \geqslant 0,\ \lambda \geqslant 0)$ не имеет седловой точки. Нарушено условие а): функция $J(u)$ выпукла лишь на U_0, требуемой точки $\bar{u}$ нет.

В примере 1 $\operatorname{ri} U_0 \cap U = \varnothing$ — нарушено условие б). С другой стороны, нетрудно привести примеры выпуклых задач, в которых условия регулярности а), б) нарушены, но седловая точка существует.

Пример 3. Пусть $U_0 = \{u \in E^1\colon u \geqslant 0\}$, $J(u) = u$, $g(u) = u^2$, $U = \{u\colon u \in U_0,\ g(u) \leqslant 0\}$. Здесь U_0 выпукло, функции $J(u)$, $g(u)$ выпуклы на U_0. Множество U состоит из единственной точки $u = 0$, так что $J_* = 0$, $U_* = \{0\}$. Функция Лагранжа $L(u, \lambda) = u + \lambda u^2$ $(u \geqslant 0,\ \lambda \geqslant 0)$ имеет седловую точку $(u_* = 0, \lambda^* = 0)$, хотя $\operatorname{ri} U_0 \cap U = \varnothing$. Этот же пример показывает, что множители Лагранжа, вообще говоря, определяются неоднозначно — здесь точки $(u_* = 0, \lambda^*)$ при любом $\lambda^* \geqslant 0$ являются седловыми точками.

Теоремы 2, 4, 5 дают достаточные условия существования седловой точки в задачах выпуклого программирования. Однако существуют и невыпуклые задачи, в которых функция Лагранжа имеет седловую точку [91].

Пример 4. Пусть $U_0 = \{u \in E^1\colon u \leqslant 1\}$, $J(u) = u^3$, $g(u) = -u^3 - 1$, $U = \{u\colon u \in U_0,\ g(u) \leqslant 0\}$. Здесь U_0 выпукло, но функции $J(u)$, $g(u)$ не являются выпуклыми на U_0. Множество U представляет собой отрезок $-1 \leqslant u \leqslant 1$, так что $J_* = -1$, $u_* = -1$. Функция Лагранжа $L(u, \lambda) = u^3 + \lambda(-u^3 - 1)$ имеет единственную седловую точку $(u_* = -1, \lambda^* = 1)$ на множестве $U_0 \times \Lambda_0$ $(\Lambda_0 = \{\lambda \in E^1\colon \lambda \geqslant 0\})$.

6. С помощью функции Лагранжа $L(u, \lambda)$ $(u \in U_0,\ \lambda \in \Lambda_0)$ задачу (1), (2) можно переформулировать следующим образом.

Введем функцию

$$\chi(u) = \sup_{\lambda \in \Lambda_0} L(u, \lambda), \quad u \in U_0. \tag{30}$$

Заметим, что если $u \in U$, то $\sum_{i=1}^{s} \lambda_i g_i(u) \leqslant 0$ при всех $\lambda \in \Lambda_0$, причем равенство получается при $\lambda = 0 \in \Lambda_0$. Если же $u \in U_0 \setminus U$, то найдется номер i такой, что либо $1 \leqslant i \leqslant m$ и $g_i(u) > 0$, либо $m+1 \leqslant i \leqslant s$ и $g_i(u) \neq 0$, так что сумму $\sum_{i=1}^{s} \lambda_i g_i(u)$ выбором $\lambda \in \Lambda_0$ можно сделать сколь угодно большой. Поэтому функция $\chi(u)$, определяемая условием (30), имеет вид

$$\chi(u) = \begin{cases} J(u) & \forall u \in U, \\ \infty & \forall u \in U_0 \setminus U. \end{cases}$$

Отсюда ясно, что $\inf_{U_0} \chi(u) = \inf_{U} J(u) = J_*$, и задачу (1), (2) можно переписать в равносильном виде

$$\chi(u) \to \inf; \quad u \in U_0. \tag{31}$$

Как и выше, в задаче (1), (2) будем предполагать, что $J_* > -\infty$, $U_* \neq \varnothing$. Тогда задача (31) будет иметь то же множество решений U_* с тем же минимальным значением J_*, т. е.

$$\inf_{U_0} \chi(u) = J_*, \quad U_* = \{u\colon u \in U_0, \ \chi(u) = J_*\}. \tag{32}$$

Наряду с функцией (30) введем функцию

$$\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda), \quad \lambda \in \Lambda_0, \tag{33}$$

и рассмотрим задачу

$$\psi(\lambda) \to \sup; \quad \lambda \in \Lambda_0. \tag{34}$$

Задачу (34) принято называть *двойственной задачей* к задаче (31) (или к исходной, основной задаче (1), (2)), а переменные $\lambda = (\lambda_1, \ldots, \lambda_s)$ называют *двойственными переменными* в отличие от исходных, основных переменных $u = (u^1, \ldots, u^n)$. Обозначим

$$\sup_{\Lambda_0} \psi(\lambda) = \psi^*, \quad \Lambda^* = \{\lambda \in \Lambda_0\colon \psi(\lambda) = \psi^*\}. \tag{35}$$

Оказывается, задачи (31) и (34) тесно связаны между собой. Прежде всего всегда выполняются неравенства

$$\psi(\lambda) \leqslant \psi^* \leqslant J_* \leqslant \chi(u), \quad u \in U_0, \quad \lambda \in \Lambda_0. \tag{36}$$

В самом деле, $\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) \leqslant L(u, \lambda)$ при всех $\lambda \in \Lambda_0$ и $u \in U_0$. Поэтому $\psi^* = \sup_{\Lambda_0} \psi(\lambda) \leqslant \sup_{\lambda \in \Lambda_0} L(u, \lambda) = \chi(u)$ для любого

$u \in U_0$. Переходя к нижней грани по $u \in U_0$ в этом неравенстве, получаем $\psi^* \leqslant J_*$, откуда следуют неравенства (36).

Интересно выяснить, когда $\psi^* = J_*$, и обе задачи (31) и (34) имеют решение, т. е.

$$U_* \neq \varnothing, \quad \Lambda^* \neq \varnothing, \quad J_* = \psi^*. \tag{37}$$

Оказывается, соотношения (37) тесно связаны с седловой точкой функции Лагранжа.

Теорема 6. *Для того чтобы имели место соотношения* (37), *необходимо и достаточно, чтобы функция* $L(u, \lambda)$ $(u \in U_0, \lambda \in \Lambda_0)$ *имела седловую точку на* $U_0 \times \Lambda_0$ *в смысле определения* 1. *Множество седловых точек функции* $L(u, \lambda)$ *на* $U_0 \times \Lambda_0$ *совпадает с множеством* $U_* \times \Lambda^*$.

Доказательство. Необходимость. Пусть выполнены соотношения (37). Возьмем произвольные $u_* \in U_*$ и $\lambda^* \in \Lambda^*$ и покажем, что (u_*, λ^*) — седловая точка. Имеем

$$\psi^* = \psi(\lambda^*) = \inf_{u \in U_0} L(u, \lambda^*) \leqslant L(u_*, \lambda^*) \leqslant$$
$$\leqslant \sup_{\lambda \in \Lambda_0} L(u_*, \lambda) = \chi(u_*) = J_*.$$

По условию $\psi^* = J_*$. Поэтому предыдущие неравенства превращаются в равенства:

$$L(u_*, \lambda^*) = \inf_{u \in U_0} L(u, \lambda^*) = \sup_{\lambda \in \Lambda_0} L(u_*, \lambda) = J_*.$$

Отсюда имеем неравенства (5), т. е. (u_*, λ^*) — седловая точка. Тем самым показано, что $U_* \times \Lambda^*$ принадлежит множеству седловых точек функции $L(u, \lambda)$ на $U_0 \times \Lambda_0$.

Достаточность. Пусть $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка функции $L(u, \lambda)$ на $U_0 \times \Lambda_0$. Согласно (5) это значит, что $L(u_*, \lambda) \leqslant L(u_*, \lambda^*)$ $(\lambda \in \Lambda_0)$. Отсюда имеем

$$\sup_{\lambda \in \Lambda_0} L(u_*, \lambda) = \chi(u_*) = L(u_*, \lambda^*).$$

Кроме того, $L(u_*, \lambda^*) \leqslant L(u, \lambda^*)$ $(u \in U_0)$, так что

$$L(u_*, \lambda^*) = \inf_{u \in U_0} L(u, \lambda^*) = \psi(\lambda^*),$$

откуда и из неравенств (36) следует

$$L(u_*, \lambda^*) = \psi(\lambda^*) \leqslant \psi^* \leqslant J_* \leqslant \chi(u_*) = L(u_*, \lambda^*),$$

т. е. $\psi(\lambda^*) = \psi^* = J_* = \chi(u_*)$. Это значит, что $\psi^* = J_*$, $\lambda^* \in \Lambda^*$, $u_* \in U_*$. Тем самым установлено, что множество седловых точек функции $L(u, \lambda)$ на $U_0 \times \Lambda_0$ принадлежит множеству $U_* \times \Lambda^*$.

Следствие 1. *Следующие четыре утверждения равносильны:*

1) $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — *седловая точка функции* $L(u, \lambda)$ *на* $U_0 \times \Lambda_0$;

2) *выполняются соотношения* (37);

3) *существуют точки* $u_* \in U_0$, $\lambda^* \in \Lambda_0$ *такие, что*

$$\chi(u_*) = \psi(\lambda^*);$$

4) *справедливо равенство*

$$\max_{\lambda \in \Lambda_0} \inf_{u \in U_0} L(u, \lambda) = \min_{u \in U_0} \sup_{\lambda \in \Lambda_0} L(u, \lambda)$$

(*напоминаем, что когда пишут* max *или* min, *то достижение соответствующей верхней или нижней грани предполагается*).

Следствие 2. *Если* (u_*, λ^*) *и* $(a_*, b^*) \in U_0 \times \Lambda_0$ — *седловые точки функции* $L(u, \lambda)$ *на* $U_0 \times \Lambda_0$, *то* (u_*, b^*), (a_*, λ^*) *также являются седловыми точками этой функции на* $U_0 \times \Lambda_0$, *причем*

$$L(u_*, b^*) = L(a_*, \lambda^*) = L(u_*, \lambda^*) = L(a_*, b^*) = J_* = \psi^*.$$

Отсюда и из теоремы 1 вытекает, что в теоремах 2, 4, 5 можно выбрать одни и те же множители Лагранжа λ^* для всех $u_* \in U_*$ сразу.

Полезно заметить, что в доказательстве теоремы 6 нигде не использовано то, что $L(u, \lambda)$ является функцией Лагранжа какой-либо задачи вида (1), (2), а множества U_0, Λ_0 выпуклы — там были важны лишь функции (30), (33), задачи (31), (34) и множества (32), (35), которые могут быть введены для любой функции $L(u, \lambda)$ на любых множествах U_0, Λ_0. Это значит, что теорема 6 и следствие 1, 2 к ней верны для произвольных функций $L(u, \lambda)$ и множеств U_0, Λ_0.

Заметим, что равенство $\psi^* = J_*$ может выполняться и в том случае, когда одно из множеств U_* или Λ^* пусто.

Пример 5. Рассмотрим задачу из примера 1 при $a = \infty$. Было показано, что $J_* = 0$, $U_* = \{0\}$. Поскольку $L(u, \lambda) = -u + \lambda u^2$ $(u \geqslant 0,\ \lambda \geqslant 0)$, то $\psi(\lambda) = \inf\limits_{u \geqslant 0} L(u, \lambda) = -1/(4\lambda)$ при $\lambda > 0$ и $\psi(0) = -\infty$. Следовательно, $\sup\limits_{\lambda \geqslant 0} \psi(\lambda) = \psi^* = 0 = J_*$, но $\Lambda^* = \varnothing$.

Однако при отсутствии седловой точки возможно строгое неравенство $\psi^* < J_*$ даже в том случае, когда $U_* \neq \varnothing$, $\Lambda^* \neq \varnothing$.

Пример 6. Пусть $J(u) = e^{-u}$, $U_0 = E^1$, $g(u) = ue^{-u}$, $U = \{u\colon u \in U_0,\ g(u) = 0\}$. Здесь множество U состоит из единственной точки $u = 0$, так что $J_* = J(0) = 1$, $U_* = \{0\}$. Посколь-

ку $L(u, \lambda) = e^{-u} + \lambda u e^{-u} = e^{-u}(1 + \lambda u)$ $(u \in E^1,\ \lambda \in \Lambda_0 = E^1)$, то

$$\psi(\lambda) = \inf_{u \in E^1} L(u, \lambda) = \begin{cases} 0, & \lambda = 0, \\ -\infty, & \lambda > 0, \\ \lambda e^{-1+1/\lambda}, & \lambda < 0, \end{cases}$$

$$\chi(u) = \sup_{\lambda \in E^1} L(u, \lambda) = \begin{cases} 1, & u = 0, \\ \infty, & u \neq 0. \end{cases}$$

Отсюда $\psi^* = \sup\limits_{E^1} \psi(\lambda) = 0 = \psi(0)$, $\Lambda^* = \{0\}$, $J_* = \inf\limits_{E^1} \chi(u) = 1 = \chi(0)$, $U_* = \{0\}$. Имеем $\psi^* < J_*$.

Не следует думать, что если $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка функции $L(u, \lambda)$, то и точки $(a, b) \in U_0 \times \Lambda_0$, для которых $L(a, b) = L(u_*, \lambda^*)$, также будут седловыми точками. Далее, если ввести множества

$$U_*(\lambda^*) = \{u\colon u \in U_0,\ L(u, \lambda^*) = L(u_*, \lambda^*)\},$$
$$\Lambda(u_*) = \{\lambda\colon \lambda \in \Lambda_0,\ L(u_*, \lambda) = L(u_*, \lambda^*)\},$$

где (u_*, λ^*) — седловая точка функции $L(u, \lambda)$ на $U_0 \times \Lambda_0$, то для множеств U_* и Λ^* из (32), (35) в общем случае можно утверждать, что

$$U_* \subset U_*(\lambda^*), \quad \Lambda^* \subset \Lambda(u_*). \tag{38}$$

Пример 7. Функция $L(u, \lambda) = \lambda u$ при $u \in U_0 = E^1$, $\lambda \in \Lambda_0 = E^1$ имеет единственную седловую точку $(u_* = 0,\ \lambda^* = 0$ $L(u_*, \lambda^*) = 0$. Но $U(\lambda^*) = E^1$, $\Lambda(u_*) = E^1$, так что в рассматриваемом случае включения (38) являются строгими.

7. Заметим, что двойственная задача (34) равносильна задаче выпуклого программирования независимо от того, была ли основная задача (1), (2) выпуклой или нет. В самом деле, функция $L(u, \lambda)$ линейна по λ, поэтому согласно теореме 2.7 функция $-\psi(\lambda) = \sup\limits_{u \in U_0} (-L(u, \lambda))$ выпукла на выпуклом множестве Λ_0. Тогда задача (34), записанная в виде

$$-\psi(\lambda) \to \inf; \quad \lambda \in \Lambda_0,$$

представляет собой задачу выпуклого программирования (здесь мы допускаем и значения $\psi(\lambda) = -\infty$). Благодаря этому обстоятельству в задаче вида (1), (2), имеющей седловую точку, бывает удобнее сначала исследовать двойственную к ней задачу, а затем, пользуясь теоремой 6, возвращаться к исходной задаче. Особенно плодотворным оказывается этот подход в задачах линейного программирования, поскольку в этом случае двойственную задачу удается выписать в явном виде.

Рассмотрим каноническую задачу линейного программирования (см. задачу (3.1.14)):

$$\langle c,\ u\rangle \to \inf;\quad u \in U = \{u \in E^n:\ u \geqslant 0,\ Au - b = 0\}, \tag{39}$$

где A — матрица порядка $s \times n$, $b \in E^s$, $c \in E^n$. Здесь $U_0 = \{u \in E^n:\ u \geqslant 0\}$, $\Lambda_0 = E^s$; функция Лагранжа имеет вид $L(u,\ \lambda) = \langle c,\ u\rangle + \langle \lambda,\ Au - b\rangle = \langle c + A^T\lambda,\ u\rangle - \langle b,\ \lambda\rangle$. Функция $\psi(\lambda)$, определяемая согласно (33), сразу выписывается в явном виде

$$\psi(\lambda) = \inf_{u \in U_0} L(u,\ \lambda) = \begin{cases} -\langle b,\ \lambda\rangle, & c + A^T\lambda \geqslant 0, \\ -\infty, & (c + A^T\lambda)^i < 0, \quad \lambda \in \Lambda_0 = E^s. \end{cases}$$

Отсюда ясно, что точку $\lambda^* \in \Lambda_0$, в которой может достигаться верхняя грань $\psi(\lambda)$ на Λ_0, достаточно искать среди тех $\lambda \in \Lambda_0$, для которых $c + A^T\lambda \geqslant 0$. Поэтому двойственную задачу (34) для задачи (39) можно сформулировать так: $-\langle b,\ \lambda\rangle \to \sup$ или

$$\langle b,\ \lambda\rangle \to \inf;\ \lambda \in \Lambda = \{\lambda \in E^s:\ c + A^T\lambda \geqslant 0\}. \tag{40}$$

Как видим, двойственная к (39) задача также представляет собой задачу линейного программирования.

Любопытно заметить, что если для задачи (40), в свою очередь, написать двойственную к ней задачу, то мы вернемся к исходной задаче (39). Чтобы показать это, составим функцию Лагранжа для задачи (40), взяв в качестве множителя Лагранжа переменные $u = (u^1,\ \ldots,\ u^n)$ и переписав ограничения $c + A^T\lambda \geqslant 0$ в стандартном виде: $-c - A^T\lambda \leqslant 0$. Получим

$$L_1(\lambda,\ u) = \langle b,\ \lambda\rangle + \langle u, -c - A^T\lambda\rangle = \langle b - Au,\ \lambda\rangle - \langle c, u\rangle = -L(u, \lambda);$$

здесь $\lambda \in \Lambda_0 = E^s$, $u \in U_0 = \{u \in E^n:\ u \geqslant 0\}$. Тогда

$$\psi_1(u) = \inf_{\lambda \in \Lambda_0} L_1(\lambda,\ u) = \begin{cases} -\langle c,\ u\rangle, & b - Au = 0, \\ -\infty, & b - Au \neq 0,\ u \in U_0. \end{cases}$$

Ясно, что $\sup\limits_{u \in U_0} \psi_1(u)$ имеет смысл искать на множестве лишь тех $u \in U_0$, для которых $b - Au = 0$. В результате придем к задаче: $-\langle c,\ u\rangle \to \sup$, или

$$\langle c,\ u\rangle \to \inf;\ u \in U = \{u \in E^n:\ u \in U_0,\ b - Au = 0\},$$

совпадающей с исходной канонической задачей (39).

Перейдем к рассмотрению основной задачи линейного программирования (см. задачу (3.1.15)):

$$\langle c,\ u\rangle \to \inf;\quad u \in U = \{u \in E^n:\ u \geqslant 0,\ Au - b \leqslant 0\}, \tag{41}$$

где A — матрица порядка $m \times n$, $b \in E^m$, $c \in E^n$. Здесь $U_0 = \{u\colon u \geqslant 0\}$, $\Lambda_0 = \{\lambda \in E^m\colon \lambda \geqslant 0\}$, функция Лагранжа имеет вид

$$L(u, \lambda) = \langle c, u\rangle + \langle \lambda, Au - b\rangle = \langle c + A^T\lambda, u\rangle - \langle b, \lambda\rangle.$$

Отсюда

$$\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) = \begin{cases} -\langle b, \lambda\rangle, & c + A^T\lambda \geqslant 0, \\ -\infty, & (c + A^T\lambda)^i < 0, \ \lambda \in \Lambda_0. \end{cases}$$

Двойственная к (41) задача запишется в виде

$$\langle b, \lambda\rangle \to \inf;\ \lambda \in \Lambda = \{\lambda \in E^m\colon \lambda \geqslant 0,\ c + A^T\lambda \geqslant 0\}. \qquad (42)$$

Это тоже задача линейного программирования. Нетрудно проверить, что двойственная к (42) задача совпадает с исходной задачей (41).

Наконец, рассмотрим общую задачу линейного программирования (см. задачу (3.1.5)):

$$\langle c, u\rangle \to \inf;$$
$$u \in U = \{u = (u^1, \ldots, u^n)\colon u^j \geqslant 0,\ j \in I;\ Au - b \leqslant 0,\ \bar{A}u - \bar{b} = 0\}, \qquad (43)$$

где I — заданное множество номеров из $\{1, \ldots, n\}$, A — матрица порядка $m \times n$, $\bar{A}$ — матрица порядка $s \times n$, $b \in E^m$, $\bar{b} \in E^s$, $c \in E^n$. Здесь $U_0 = \{u = (u^1, \ldots, u^n)\colon u^j \geqslant 0,\ j \in I\}$. Множители Лагранжа удобно представить в виде $\lambda = (\mu, \bar{\mu})$ ($\mu \in E^m$, $\bar{\mu} \in E^s$). Поскольку множители, отвечающие ограничениям типа неравенств, неотрицательны, то

$$\lambda \in \Lambda_0 = \{\lambda = (\mu, \bar{\mu}) \in E^m \times E^s\colon \mu \geqslant 0\}.$$

Функция Лагранжа имеет вид

$$L(u, \lambda) = \langle c, u\rangle + \langle \mu, Au - b\rangle + \langle \bar{\mu}, \bar{A}u - \bar{b}\rangle =$$
$$= \langle c\colon A^T\mu + \bar{A}^T\bar{\mu}, u\rangle + \langle b, \mu\rangle - \langle \bar{b}, \bar{\mu}\rangle,\ u \in U_0,\ \lambda \in \Lambda_0.$$

Отсюда $\psi(\lambda) = \inf\limits_{u \in U_0} L(u, \lambda) = -\langle b, \mu\rangle - \langle \bar{b}, \bar{\mu}\rangle$ при $(c + A^T\mu + \bar{A}^T\bar{\mu})_i \geqslant 0$ $(i \in I)$ и $(c + A^T\mu + \bar{A}^T\bar{\mu})_i = 0$ $(i \notin I)$, а $\psi(\lambda) = -\infty$ при остальных $\lambda = (\mu, \bar{\mu}) \in \Lambda_0$. Тогда двойственная к (43) задача запишется в виде

$$\langle b, \mu\rangle + \langle \bar{b}, \bar{\mu}\rangle \to \inf,\ \lambda = (\mu, \bar{\mu}) \in \Lambda = \{\lambda = (\mu, \bar{\mu}) \in E^s\colon \mu \geqslant 0,$$
$$(c + A^T\mu + \bar{A}^T\bar{\mu})_i \geqslant 0,\ i \in I;\ (c + A^T\mu + \bar{A}^T\bar{\mu})_i = 0,\ i \notin I\}. \qquad (44)$$

В результате снова получили задачу линейного программирования. Предлагаем читателю проверить, что двойственная к (44) задача будет совпадать с исходной задачей (43).

Если задача линейного программирования имеет решение, т. е. $J_* > -\infty$, $U_* \neq \varnothing$, то согласно теореме 4 функция Лагранжа этой задачи всегда имеет седловую точку. Отсюда и из установленной в теореме 6 связи между основной и двойственной задачами вытекают следующие теоремы, занимающие одно из центральных мест в теории линейного программирования.

Теорема 7. *Для того чтобы точка $u_* \in U$ была решением задачи* (43), *необходимо и достаточно существование такой точки $\lambda^* = (\mu^*, \bar{\mu}^*) \in \Lambda$ (здесь Λ определяется условиями* (44)), *для которой*

$$\langle c, u_* \rangle = -\langle b, \mu^* \rangle - \langle \bar{b}, \bar{\mu}^* \rangle. \tag{45}$$

Теорема 8. *Для того чтобы точка $u_* \in E^n$ была решением задачи* (43), *необходимо и достаточно существование точки $\lambda^* = (\mu^*, \bar{\mu}^*) \in E^m \times E^s$ такой, что*

$$u_*^j \geqslant 0, \quad j \in I; \quad Au_* \leqslant b, \quad \bar{A}u_* = \bar{b},$$

$$\mu^* \geqslant 0, \ (c + A^T\mu^* + \bar{A}^T\bar{\mu}^*)_i \geqslant 0, \ i \in I;$$

$$(c + A^T\mu^* + \bar{A}^T\bar{\mu}^*)_i = 0, \ i \notin I;$$

$$\mu_i^* (Au_* - b)_i = 0, \quad i = 1, \ldots, m; \quad u_*^i (c + A^T\mu^* + \bar{A}^T\bar{\mu}^*)_i = 0,$$

$$i = 1, \ldots, s.$$

Теорема 9. *Задачи* (43) *и* (44) *либо обе не имеют решения, либо обе имеют решение, причем в последнем случае выполняется равенство* (45), *где u_* — решение задачи* (43), $\lambda^* = (\mu^*, \bar{\mu}^*)$ — *решение задачи* (44).

Теоремы 7, 8 являются переформулировкой теорем 1, 4, 6 применительно к задаче (43) и в отдельном доказательстве не нуждаются. Теорема 9 специфична для задач линейного программирования — об этом говорит пример 5, в котором основная задача имеет решение, а двойственная задача — не имеет. Поэтому теорема 9 требует отдельного доказательства.

Доказательство теоремы 9. Составим функцию Лагранжа для двойственной задачи (44), взяв в качестве множителей Лагранжа $u = (u^1, \ldots, u^n)$,

$$L_1(\lambda, u) = \langle b, \mu \rangle + \langle \bar{b}, \bar{\mu} \rangle + \sum_{i \in I} u^i (-c - A^T\mu - \bar{A}^T\bar{\mu})_i +$$

$$+ \sum_{i \notin I} u^i (-c - A^T\mu - \bar{A}^T\bar{\mu})_i =$$

$$= \langle b, \mu \rangle + \langle \bar{b}, \bar{\mu} \rangle - \langle u, c + A^T\mu + \bar{A}^T\bar{\mu} \rangle,$$

$$\lambda \in \Lambda_0 = \{\lambda = (\bar{\mu}, \mu) \in E^m \times E^n: \ \mu \geqslant 0\},$$

$$u \in U_0 = \{u \in E^n: \ u^j \geqslant 0, \ j \in I\}.$$

Как видим, функции Лагранжа задач (43) и (44) отличаются лишь знаком. Согласно теореме 4 задача (44) имеет решение $\lambda^* = (\mu^*, \bar{\mu}^*)$ тогда и только тогда, когда функция $L_1(\lambda, u)$ на $\Lambda_0 \times U_0$ имеет седловую точку $(\lambda^*, u_*) \in \Lambda_0 \times U_0$, т. е.

$$L_1(\lambda^*, u) \leqslant L_1(\lambda^*, u_*) \leqslant L_1(\lambda, u_*), \quad u \in U_0, \quad \lambda \in \Lambda_0.$$

Но $L_1(\lambda, u) = -L(u, \lambda)$, поэтому из последних неравенств следует, что если (λ^*, u_*) — седловая точка функции $L_1(\lambda, u)$ на $\Lambda_0 \times U_0$, то (u_*, λ^*) — седловая точка функции $L(u, \lambda)$ на $U_0 \times \Lambda_0$. Таким образом, функции $L_1(\lambda, u)$ и $L(u, \lambda)$ одновременно либо имеют седловую точку, либо ее не имеют.

Согласно теореме 4 это означает, что задачи (43) и (44) либо обе не имеют решения, либо обе имеют решение. В том случае, когда обе эти задачи имеют решение, равенство (45) вытекает из теоремы 6 и следствия 1 из нее.

Таким образом, в теоремах 7—9 установлена определенная эквивалентность исходной и двойственной к ней задач линейного программирования. Такая эквивалентность широко используется при разработке и исследовании различных методов решения задач линейного программирования. Так, например, если к двойственной задаче применить симплекс-метод и затем его истолковать в терминах исходной задачи, то придем к так называемому двойственному симплекс-методу. Об этом и других методах решения задач линейного программирования, основанных на теории двойственности, см., например, [3, 11, 12, 33, 71, 130, 225, 265, 340].

8. Выше было замечено, что в любой задаче линейного программирования двойственная задача, написанная для двойственной задачи, совпадает с исходной. В общем случае это не так. В самом деле, двойственная задача всегда равносильна задаче выпуклого программирования. Поэтому в невыпуклых задачах (1), (2) задача, двойственная к двойственной задаче, заведомо не может совпадать с исходной. Любопытно заметить, что такое явление возможно и в задачах выпуклого программирования.

Пример 8. Пусть $J(u) = |u|^2 \to \inf (u \in U = \{u \in E^n\colon g(u) = |u|^2 - 1 \leqslant 0\})$. Здесь $U_0 = E^n$, $J_* = 0$, $u_* = 0$, $\Lambda_0 = \{\lambda \in E^1\colon \lambda \geqslant 0\}$. Функция Лагранжа $L(u, \lambda) = |u|^2 + \lambda(|u|^2 - 1) = (1 + \lambda)|u|^2 - \lambda$. Отсюда $\psi(\lambda) = \inf\limits_{u \in E^n} L(u, \lambda) = -\lambda$ при $\lambda + 1 \geqslant 0$ и $\psi(\lambda) = -\infty$ при $\lambda + 1 < 0$. Следовательно, двойственная задача имеет вид $\psi(\lambda) = -\lambda \to \sup$; $\lambda \in \Lambda = \{\lambda, \lambda \in \Lambda_0, \lambda + 1 \geqslant 0\}$. Эта задача линейного программирования, поэтому двойственная к ней задача не может совпасть с исходной задачей. Заметим, что в этой задаче $(u_* = 0, \lambda^* = 0)$ — седловая точка.

9. Кратко остановимся еще на одном важном классе задач, называемых *задачами геометрического программирования*, в кото-

рых переход к двойственной задаче весьма плодотворен. Речь идет о задачах минимизации следующего вида:

$$g(x) = \sum_{i=1}^{n} c_i x_1^{a_{i1}} \ldots x_r^{a_{ir}} \to \inf; \quad x \in X = \operatorname{int} E_+^r, \tag{46}$$

где $c_i > 0$, a_{ij} — заданные числа, $\operatorname{int} E_+^r = \{x = (x_1, \ldots, x_r)\colon x_1 > 0, \ldots, x_r > 0\}$. Функция $g(x)$ из (46) называется *позиномом.*

Для исследования задачи (46) удобнее перейти к новым переменным $u = (u_1, \ldots, u_{r+n})$ по формулам $u_i = \ln x_i$ $(i = 1, \ldots, r)$;

$$u_{r+1} = -b_i + \sum_{j=1}^{r} a_{ij} u_j, \quad b_i = -\ln c_i, \quad i = 1, \ldots, n, \tag{47}$$

и переписать ее в эквивалентном виде:

$$\begin{gathered} J(u) = \sum_{i=1}^{n} e^{u_{r+i}} \to \inf; \\ u \in U = \left\{ u \in E^{r+n}\colon \sum_{j=1}^{r} a_{ij} u_j - u_{r+i} - b_i = 0, \quad i = 1, \ldots, n \right\}. \end{gathered} \tag{48}$$

Отметим, что функция $J(u)$ выпукла на E^{r+n}, U — многогранное множество, и поэтому к задаче (48) применима теорема 4. Составим функцию Лагранжа (3) для этой задачи:

$$\begin{aligned} L(u, \lambda) &= \sum_{i=1}^{n} e^{u_{r+i}} + \sum_{i=1}^{n} \lambda_i \left(\sum_{j=1}^{r} a_{ij} u_j - u_{r+i} - b_i \right) = \\ &= \sum_{i=1}^{n} \left(e^{u_{r+i}} - \lambda_i u_{r+i} - \lambda_i b_i \right) + \sum_{j=1}^{r} \left(\sum_{i=1}^{n} a_{ij} \lambda_i \right) u_j, \quad u \in E^{r+n} = U_0, \\ &\qquad \lambda \in E^n = \Lambda_0. \end{aligned}$$

С помощью классического метода (§ 1.2) нетрудно показать, что нижняя грань функции $\varphi(z) = e^z - \lambda_i z - \lambda_i b_i$ переменной z на числовой оси равна $\varphi_* = \lambda_i - \lambda_i \ln \lambda_i - \lambda_i b_i$, причем при $\lambda_i > 0$ она достигается в точке $z_* = -\ln \lambda_i$; функция $\lambda_i \ln \lambda_i$ при $\lambda_i = 0$ здесь считается доопределенной по непрерывности нулем. Отсюда, опираясь на линейность функции $L(u, \lambda)$ по переменным $u_1, \ldots, u_r$, получаем

$$\begin{aligned} \psi(\lambda) &= \inf_{u \in E^{r+n}} L(u, \lambda) = \\ &= \begin{cases} \sum_{i=1}^{n} (\lambda_i - \lambda_i \ln \lambda_i - \lambda_i b_i), & \lambda \in E_+^n, \quad \sum_{i=1}^{n} a_{ij} \lambda_i = 0, \quad j = 1, \ldots, r, \\ -\infty & \text{при других } \lambda. \end{cases} \end{aligned}$$

Поэтому двойственная задача (34) здесь будет иметь вид

$$\psi(\lambda) = \sum_{i=1}^{n} (\lambda_i - \lambda_i \ln \lambda_i - \lambda_i b_i) \to \sup; \quad \lambda \in \Lambda,$$
$$\Lambda = \left\{\lambda = (\lambda_1, \ldots, \lambda_n) \in E_+^n: \sum_{i=1}^{n} a_{ij}\lambda_i = 0, \quad j = 1, \ldots, r\right\}. \tag{49}$$

Если здесь верхняя грань достигается в точке $\lambda^* \neq 0$, то задачу (49) можно записать в более простой форме. А именно, учитывая, что любую точку $\lambda = (\lambda_1, \ldots, \lambda_n) \neq 0$ можно представить в виде $\lambda = \alpha\nu$, где $\alpha = \sum_{i=1}^{n} \lambda_i$, $\nu = (\nu_1, \ldots, \nu_n)$, $\nu_i = \lambda_i/\alpha$, $\nu_1 + \ldots + \nu_n = 1$, задачу (49) перепишем сначала в терминах переменных (α, ν):

$$\psi_1(\alpha, \nu) = \psi(\alpha\nu) = \sum_{i=1}^{n} \alpha(\nu_i - \nu_i \ln \alpha\nu_i - \nu_i b_i) =$$
$$= \alpha\left[1 - \ln\alpha + \sum_{i=1}^{n} (\nu_i \ln c_i - \nu_i \ln \nu_i)\right] \to \sup;$$
$$(\alpha, \nu) \in \Lambda_1 =$$
$$= \left\{(\alpha, \nu): \alpha > 0, \quad \nu \in E_+^n, \quad \sum_{i=1}^{n} \nu_i = 1, \quad \sum_{i=1}^{n} a_{ij}\nu_i = 0, \quad j = 1, \ldots, r\right\}.$$

Далее, пользуясь классическим методом (§ 1.2), убеждаемся, что точка $\alpha^* = \prod_{i=1}^{n} \left(\frac{c_i^{\nu_i}}{\nu_i^{\nu_i}}\right) > 0$ (здесь принято $0^0 = 1$) доставляет функции $\psi_1(\alpha, \nu)$ максимальное значение по $\alpha > 0$ при фиксированном $\nu \in E_+^n$, причем $\psi_1(\alpha^*, \nu) = \prod_{i=1}^{n} \left(\frac{c_i^{\nu_i}}{\nu_i^{\nu_i}}\right)$.

Тогда двойственная задача (49) перепишется в следующем виде:

$$\psi_2(\nu) = \frac{c_1^{\nu_1} \ldots c_n^{\nu_n}}{\nu_1^{\nu_1} \ldots \nu_n^{\nu_n}} \to \sup; \quad \nu \in \Lambda_2,$$
$$\Lambda_2 = \left\{\nu = (\nu_1, \ldots, \nu_n) \in E_+^n: \sum_{i=1}^{n} \nu_i = 1, \ \sum_{i=1}^{n} a_{ij}\nu_i = 0, \ j = 1, \ldots, r\right\}. \tag{50}$$

Если $\nu^* = (\nu_1^*, \ldots, \nu_n^*) \in \operatorname{int} E_+^n$ — решение задачи (50), то, полагая $\lambda^* = \alpha^*\nu^*$, где $\alpha^* = \prod_{i=1}^{n} \left(\frac{c_i}{\nu_i^*}\right)^{\nu_i^*}$, $u_{r+i,*} = \lambda_i^*$ $(i = 1, \ldots, n)$, из системы линейных алгебраических уравнений (47) можно

получить $u_{1*}, \ldots, u_{r*}$, откуда имеем решение $x_* = (x_{1*} = e^{u_{1*}}, \ldots, x_{r*} = e^{u_{r*}})$ исходной задачи (46). Задача (50) часто бывает проще задачи (46). Переход к двойственной задаче особенно эффективным оказывается тогда, когда множество Λ_2 в задаче (50) состоит из единственной точки ν^*, которая и будет решением этой задачи.

Аналогично может быть исследована и более общая задача геометрического программирования

$$g_0(x) \to \inf; \quad x \in X = \{x \in \operatorname{int} E^r_+ \colon g_1(x) \leqslant 1, \ldots, g_m(x) \leqslant 1\},$$

где $g_0(x), \ldots, g_m(r)$ — позиномы. Подробнее о геометрическом программировании, его приложениях см., например, в [7, 131, 234].

Читателей, желающих подробнее ознакомиться с красивой и богатой результатами теорией двойственности, с различными ее приложениями, отсылаем к [2, 3, 18, 21, 67, 116, 137, 146, 156, 166, 167, 201, 255, 264, 290, 293]. Здесь мы лишь отметим, что параллельное рассмотрение задачи минимизации и двойственной к ней задачи, с одной стороны, приводит к важным теоретическим результатам, с другой стороны, служит источником различных методов минимизации.

Заметим также, что в последнее время растет интерес к задачам, в которых нарушены соотношения двойственности (37),— такие задачи возникают при исследовании объектов, описываемых противоречивыми системами ограничений, и имеют интересные приложения [146].

Упражнения. 1. Сформулировать аналоги теорем Куна — Таккера для задачи максимизации: $g(u) \to \sup$ $(u \in U)$, где множество U определено посредством (2).

Указание: рассмотреть задачу: $J(u) = -g(u) \to \inf$ $(u \in U)$ и воспользоваться теоремами 2, 4, 5.

2. С помощью теорем Куна — Таккера исследовать задачу: $J(u) = \sum_{i=1}^{n} |u^i - a_i| \to \inf$ $(u \in U)$, где $U = \{u \in E^n \colon |u| \leqslant 1\}$ или $U = \{u \in E^n \colon u^1 + \ldots + u^n = 0\}$, или $U = E^n_+$; $a_1, \ldots, a_n$ — заданные числа.

3. Применить теорему Куна — Таккера к задаче квадратичного программирования: $J(u) = (u^1)^2 + \ldots + n(u^n)^2 \to \inf$ $(u \in U)$, где $U = \{u \in E^n \colon u^1 + \ldots + u^n = 1\}$, или $U = \{u \in E^n \colon u^1 + \ldots + u^n \leqslant 1\}$, или $U = \{u \in E^n \colon -1 \leqslant u^1 + \ldots + u^n \leqslant 1\}$, или U является пересечением предыдущих множеств с E^n_+.

4. Решить задачи геометрического программирования:

а) $g(x) = c_1 x^{a_1} + c_2 x^{-a_2} \to \inf$ при $x > 0$, где $c_i > 0$, $a_i > 0$ — заданные числа;

б) $g(x, y) = x^{-1}y + 2x^2y + 3x^{-1}y^{-2} \to \inf$ при $x > 0$, $y > 0$;

в) $g(x, y, z) = 4x^{-1}y^{-1}z^{-1} + xy + 4xz + 2yz \to \inf$ при $x > 0$, $y > 0$, $z > 0$;

г) $g(x, y) = y \to \inf$ при $x > 0$, $y > 0$, $x^4y^{-4} + x^{-1}y^{1/2} \leqslant 1$;

д) $g(x, y, z) = x + y^{-1}z^{-1/2} \to \inf$ при $x > 0$, $y > 0$, $z > 0$, $x^{-1}y + x^{-1}z \leqslant 1$.

5. Выяснить, существуют ли седловые точки функции Лагранжа в задачах из примеров и упражнений к § 2.2, к гл. 3, к § 4.8; найти эти точки.

6. Доказать, что выпуклая квадратичная функция $J(u) = \frac{1}{2}\langle Cu, u\rangle + \langle c, u\rangle$ либо достигает своей нижней грани на E^n, либо не ограничена снизу.

7. Сформулировать и доказать аналог леммы 2 с использованием субградиента. Сформулировать субдифференциальные аналоги теорем 2, 4, 5.

8. Пусть в задачах (43), (44) $U \neq \varnothing$, $\Lambda \neq \varnothing$. Доказать, что тогда в этих задачах существуют такие решения $u_* \in U_*$, $\lambda^* \in \Lambda^*$, что в правых частях равенств $\mu_i^*(Au_* - b)_i = 0$ $(i = 1, \ldots, m)$, $u_*^i(c + A^T\mu^* + \overline{A}^T\overline{\mu}^*)_i = 0$ $(i = 1, \ldots, s)$ (см. теорему 8) один из сомножителей отличен от нуля.

9. Пусть U_0 — выпуклое множество из E^n, функции $g_1(u), \ldots, g_m(u)$ выпуклы на U_0, $g_i(u) = \langle a_i, u\rangle - b^i$ $(i = m + 1, \ldots, s)$. Доказать, что если система неравенств $g_i(u) < 0$ $(i = 1, \ldots, m)$, $g_i(u) = 0$ $(i = m + 1, \ldots, s)$ не имеет решения на U_0, то существуют числа $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0, \lambda_{m+1}, \ldots, \lambda_s$ такие, что $\lambda_1 g_1(u) + \ldots + \lambda_s g_s(u) \geqslant 0$ при всех $u \in U_0$.

Указание: построить множества, аналогичные множествам A и B из доказательства теоремы 2, и применить к ним теорему отделимости 5.2.

10. Пользуясь теоремой 3 (Фаркаша), доказать, что для несовместности системы линейных неравенств $\langle e_i, u\rangle \leqslant \mu_i$ $(i = 0, 1, \ldots, m)$ необходимо и достаточно, чтобы существовали такие числа $\lambda_0 \geqslant 0$, $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0$, что $\lambda_0 e_0 + \lambda_1 e_1 + \ldots + \lambda_m e_m = 0$, $\lambda_0\mu_0 + \lambda_1\mu_1 + \ldots + \lambda_m\mu_m < 0$ (ср. с теоремами 5.9, 5.10) [321].

11. Доказать, что два непустых многогранных множества $A = \{u \in E^n: \langle e_i, u\rangle \leqslant \mu_i, i = 0, 1, \ldots, k\}$ и $B = \{u \in E^n: \langle e_i, u\rangle \leqslant \mu_i, i = k + 1, \ldots, m\}$, не имеющие общих точек, сильно отделимы.

Указание: рассмотреть гиперплоскость $\langle c, u\rangle = \gamma$, где $c = \sum_{i=0}^{k} \lambda_i e_i$, $\gamma = \sum_{i=0}^{k} \lambda_i \mu_i$, числа $\lambda_0, \ldots, \lambda_m$ взяты из упражнения 10.

12. Доказать, что если система линейных неравенств $\langle e_0, u\rangle < 0$, $\langle e_1, u\rangle \leqslant 0, \ldots, \langle e_m, u\rangle \leqslant 0$ несовместна, то существуют такие числа $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0$, что $e_0 = -\lambda_1 e_1 - \ldots - \lambda_m e_m$.

Указание: воспользоваться теоремой 3 (Фаркаша).

13. Пусть система $\langle e_0, u\rangle < \mu_0$, $\langle e_i, u\rangle \leqslant \mu_i$ $(i = 1, \ldots, m)$ несовместна, а подсистема $\langle e_i, u\rangle \leqslant \mu_i$ $(i = 1, \ldots, m)$ совместна. Доказать, что тогда существуют числа $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0$ такие, что $e_0 = -\lambda_1 e_1 - \ldots - \lambda_m e_m$, $\mu_0 + \lambda_1\mu_1 + \ldots + \lambda_m\mu_m \leqslant 0$ [21].

Указание: в пространстве переменных $(u, t) \in E^{n+1}$ рассмотреть систему $\langle e_0, u\rangle - t\mu_0 < 0$, $\langle e_i, u\rangle - t\mu_i \leqslant 0$ $(i = 1, \ldots, m)$, $\langle 0, u\rangle - t \leqslant 0$ и воспользоваться утверждением из упражнения 12.

Глава 5

МЕТОДЫ МИНИМИЗАЦИИ ФУНКЦИЙ МНОГИХ ПЕРЕМЕННЫХ

Выше в гл. 3 был рассмотрен симплекс-метод для решения задач линейного программирования. Перейдем к изложению других методов минимизации функций конечного числа переменных, не предполагая линейности рассматриваемых задач.

К настоящему времени разработано и исследовано большое число методов минимизации функций многих переменных. Мы ниже остановимся лишь на некоторых наиболее известных и часто используемых на практике методах минимизации. Будет дано краткое описание каждого из рассматриваемых методов, исследованы вопросы сходимости, обсуждены некоторые вычислительные аспекты этих методов. При этом мы ограничимся рассмотрением лишь одного — двух основных вариантов излагаемых методов, чтобы ознакомить читателя с основами этих методов, полагая, что знание основ методов облегчит читателю изучение литературы, позволит ему без особого труда понять суть того или иного метода и выбрать подходящий вариант метода или самому разработать более удобные его модификации, лучше приспособленные для решения интересующего читателя класса задач. В конце главы будут высказаны некоторые общие замечания по методам минимизации.

Из обширной литературы, посвященной методам минимизации функций конечного числа переменных и их приложениям, мы можем здесь упомянуть лишь весьма незначительную ее часть [3, 4, 7, 8, 10—13, 15, 16, 19—23, 25—27, 29, 30, 35, 38, 40—42, 44, 46—49, 52—58, 71, 72, 76. 78, 79, 81—89, 94, 96, 103, 106, 107, 109—113, 115, 116, 119, 122, 123, 126, 127, 129—135, 139—141, 143—159, 162, 163, 165—171, 173, 174, 176—178, 180, 183, 190, 191, 194, 196, 198, 200—203, 207—209, 213—218, 221, 224, 225, 227—231, 234, 235, 237, 238, 240—242, 245, 247, 248, 250, 254—259, 261, 262, 265—267, 269—272, 274, 276—279, 282, 284, 287, 288, 291—294, 296—299, 301, 302, 306, 307, 313, 314, 316, 318, 320—324, 329—340, 343].

§ 1. Градиентный метод

1. Будем рассматривать задачу

$$J(u) \to \inf; \quad u \in U \equiv E^n, \tag{1}$$

предполагая, что функция $J(u)$ непрерывно дифференцируема на E^n, т. е. $J(u) \in C^1(E^n)$. Согласно определению 2.2.1 дифференцируемой функции

$$J(u+h) - J(u) = \langle J'(u),\ h\rangle + o(h;\ u), \tag{2}$$

где $\lim\limits_{|h|\to 0} o(h; u)|h|^{-1} = 0$. Если $J'(u) \neq 0$, то при достаточно малых $|h|$ главная часть приращения (2) будет определяться дифференциалом функции $dJ(u) = \langle J'(u), h\rangle$. Справедливо неравенство Коши — Буняковского

$$-|J'(u)| \cdot |h| \leqslant \langle J'(u), h\rangle \leqslant |J'(u)| \cdot |h|,$$

причем если $J'(u) \neq 0$, то правое неравенство превращается в равенство только при $h = \alpha J'(u)$, а левое неравенство — только при $J'(u) \neq 0$, где $\alpha = \text{const} \geqslant 0$. Отсюда ясно, что при $J'(u) \neq 0$ направление наибыстрейшего возрастания функции $J(u)$ в точке u совпадает с направлением градиента $J'(u)$, а направление наибыстрейшего убывания — с направлением антиградиента $(-J'(u))$.

Это замечательное свойство градиента лежит в основе ряда итерационных методов минимизации функций. Одним из таких методов является градиентный метод, к описанию которого мы переходим. Этот метод, как и все итерационные методы, предполагает выбор начального приближения — некоторой точки u_0. Общих правил выбора точки u_0 в градиентном методе, как, впрочем, и в других методах, к сожалению, нет. В тех случаях, когда из геометрических, физических или каких-либо других соображений может быть получена априорная информация об области расположения точки (или точек) минимума, то начальное приближение u_0 стараются выбрать поближе к этой области.

Будем считать, что некоторая начальная точка u_0 уже выбрана. Тогда градиентный метод заключается в построении последовательности $\{u_k\}$ по правилу

$$u_{k+1} = u_k - \alpha_k J'(u_k), \quad \alpha_k > 0, \quad k = 0, 1, \ldots \tag{3}$$

Число α_k из (3) часто называют длиной шага или просто шагом градиентного метода. Если $J'(u_k) \neq 0$, то шаг $\alpha_k > 0$ можно выбрать так, чтобы $J(u_{k+1}) < J(u_k)$. В самом деле, из равенства (2) имеем

$$J(u_{k+1}) - J(u_k) = \alpha_k \left[-|J'(u_k)|^2 + o(\alpha_k)\alpha_k^{-1}\right] < 0$$

при всех достаточно малых $\alpha_k > 0$. Если $J'(u_k) = 0$, то u_k — стационарная точка. В этом случае процесс (3) прекращается, и при необходимости проводится дополнительное исследование поведения функции в окрестности точки u_k для выяснения того, достигается ли в точке u_k минимум функции $J(u)$ или не достигается. В частности, если $J(u)$ — выпуклая функция, то согласно теореме 4.2.3 в стационарной точке всегда достигается минимум.

Существуют различные способы выбора величины α_k в методе (3). В зависимости от способа выбора α_k можно получить различные варианты градиентного метода. Укажем несколько наиболее употребительных на практике способов выбора α_k.

1) На луче $\{u \in E^n\colon\ u = u_k - \alpha J'(u_k),\ \alpha \geqslant 0\}$, направленном по антиградиенту, введем функцию одной переменной

$$f_k(\alpha) = J(u_k - \alpha J'(u_k)), \quad \alpha \geqslant 0,$$

и определим α_k из условий

$$f_k(\alpha_k) = \inf_{\alpha \geqslant 0} f_k(\alpha) = f_{k*}, \quad \alpha_k > 0. \tag{4}$$

Метод (3), (4) принято называть методом скорейшего спуска. При $J'(u_k) \neq 0$ согласно формуле (2.3.1) $f_k'(0) = -|J'(u_k)|^2 < 0$, поэтому нижняя грань в (4) может достигаться лишь при $\alpha_k > 0$. Приведем пример, когда величина α_k, определяемая условием (4), существует и может быть выписана в явном виде.

Пример 1. Пусть дана квадратичная функция

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle, \tag{5}$$

где A — симметричная положительно определенная матрица порядка $n \times n$, b — вектор из E^n. Выше было показано, что эта функция сильно выпукла и ее производные вычисляются по формулам

$$J'(u) = Au - b; \quad J''(u) = A.$$

Поэтому метод (3) в данном случае будет выглядеть так:

$$u_{k+1} = u_k - \alpha_k(Au_k - b), \qquad k = 0, 1, \ldots$$

Таким образом, градиентный метод для функции (5) представляет собой хорошо известный итерационный метод решения системы линейных алгебраических уравнений $Au = b$. Определим α_k из условий (4). Пользуясь формулой (4.2.10), имеем

$$f_k(\alpha) = J(u_k) - \alpha|J'(u_k)|^2 + (\alpha^2/2)\langle AJ'(u_k), J'(u_k)\rangle, \ \alpha \geqslant 0.$$

При $J'(u_k) \neq 0$ условие $f_k'(\alpha) = -|J'(u_k)|^2 + \alpha\langle AJ'(u_k), J'(u_k)\rangle = 0$ дает

$$\alpha_k = \frac{|J'(u_k)|^2}{\langle AJ'(u_k), J'(u_k)\rangle} = \frac{|Au_k - b|^2}{\langle A(Au_k - b), Au_k - b\rangle} > 0.$$

Поскольку функция $f_k(\alpha)$ выпукла, то в найденной точке α_k эта функция достигает своей нижней грани при $\alpha \geqslant 0$. Метод скорейшего спуска для функции (5) описан.

Однако точное определение величины α_k из условий (4) не всегда возможно. Кроме того, нижняя грань в (4) при некоторых k может и не достигаться. Поэтому на практике ограничиваются нахождением величины α_k, приближенно удовлетворяющей условиям (4). Здесь возможен, например, выбор α_k из условий

$$f_{k*} \leqslant f_k(\alpha_k) \leqslant f_{k*} + \delta_k, \quad \delta_k \geqslant 0, \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty, \tag{6}$$

или из условий (9)

$$f_{k*} \leqslant f_k(\alpha_k) \leqslant (1-\lambda_k) f_k(0) + \lambda_k f_{k*}, \quad 0 < \bar{\lambda} \leqslant \lambda_k \leqslant 1. \qquad (7)$$

Величины δ_k, λ_k из (6), (7) характеризуют погрешность выполнения условия (4): чем ближе δ_k к нулю или λ_k к единице, тем точнее выполняется условие (4). При поиске α_k из условий (6), (7) можно пользоваться различными методами минимизации функций одной переменной (например, методами гл. 1).

Следует также заметить, что антиградиент $(-J'(u_k))$ указывает направление быстрейшего спуска лишь в достаточно малой окрестности точки u_k. Это означает, что если функция $J(u)$ меняется быстро, то в следующей точке u_{k+1} направление антиградиента $(-J'(u_{k+1}))$ может сильно отличаться от направления $(-J'(u_k))$. Поэтому слишком точное определение величины α_k из условий (4) не всегда целесообразно.

2) На практике нередко довольствуются нахождением какого-либо $\alpha_k > 0$, обеспечивающего условие монотонности: $J(u_{k+1}) < J(u_k)$. С этой целью задаются какой-либо постоянной $\alpha > 0$ и в методе (3) на каждой итерации берут $\alpha_k = \alpha$. При этом для каждого $k \geqslant 0$ проверяют условие монотонности, и в случае его нарушения $\alpha_k = \alpha$ дробят до тех пор, пока не восстановится монотонность метода. Время от времени полезно пробовать увеличить α с сохранением условия монотонности.

3) Если функция $J(u) \in C^{1,1}(E^n)$, т. е. $J(u) \in C^1(E^n)$, и градиент $J'(u)$ удовлетворяет условию

$$|J'(u) - J'(v)| \leqslant L|u-v|, \quad u, v \in E^n,$$

причем константа L известна, то в (3) в качестве α_k может быть взято любое число, удовлетворяющее условиям

$$0 < \varepsilon_0 \leqslant \alpha_k \leqslant 2/(L+2\varepsilon), \qquad (8)$$

где ε_0, ε — положительные числа, являющиеся параметрами метода. В частности, при $\varepsilon = L/2$, $\varepsilon_0 = 1/L$ получим метод (3) с постоянным шагом $\alpha_k = 1/L$. Отсюда ясно, что если константа L большая или получена с помощью слишком грубых оценок, то шаг α_k в (3) будет маленьким. Метод (3), (8) подробнее рассмотрим в следующем параграфе.

4) Возможен выбор α_k из условия [11, 19, 56]

$$J(u_k) - J(u_k - \alpha_k J'(u_k)) \geqslant \varepsilon\alpha_k |J'(u_k)|^2, \quad \varepsilon > 0. \qquad (9)$$

Для удовлетворения условия (9) сначала обычно берут некоторое число $\alpha_k = \alpha > 0$ (одно и то же на всех итерациях; например, $\alpha_k = 1$), а затем при необходимости дробят его, т. е. изменяют по закону $\alpha_k = \lambda^i \alpha$ $(i = 0, 1, \ldots, 0 < \lambda < 1)$ до тех пор, пока впервые не выполнится условие (9).

5) Возможно априорное задание величин α_k из условий

$$\alpha_k > 0, \quad k = 0, 1, \ldots; \qquad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{10}$$

Например, в качестве α_k можно взять $\alpha_k = c(k+1)^{-\alpha}$, где $c = \text{const} > 0$, а число α таково, что $1/2 < \alpha \leqslant 1$. В частности, если $\alpha = 1$, $c = 1$, то получим $\alpha_k = (k+1)^{-1}$ $(k = 0, 1, \ldots)$. Такой выбор $\{\alpha_k\}$ в (3) очень прост для реализации, но не гарантирует выполнения условия монотонности $J(u_{k+1}) < J(u_k)$ и, вообще говоря, сходится медленно. Более подробно о методе (3), (10) см. ниже в § 3.

6) В тех случаях, когда заранее известна величина $J_* = \inf\limits_{E^n} J(u) > -\infty$, в (3) можно принять

$$\alpha_k = (J(u_k) - J_*) \, | J'(u_k) |^{-2}$$

— это абсцисса точки пересечения прямой $J = J_*$ с касательной к кривой $J = f_k(\alpha) = J(u_k - \alpha J'(u_k))$ в точке $(0, f_k(0))$.

Допустим, что какой-либо способ выбора α_k в (3) (например, один из перечисленных выше способов) уже выбран. Тогда на практике итерации (3) продолжают до тех пор, пока не выполнится некоторый критерий окончания счета. Здесь часто используются следующие критерии:

$$|u_k - u_{k+1}| \leqslant \varepsilon, \text{ или } |J(u_k) - J(u_{k+1})| \leqslant \delta, \text{ или } |J'(u_k)| \leqslant \gamma,$$

где ε, δ, γ — заданные числа. Иногда заранее задают число итераций; возможны различные сочетания этих и других критериев. Разумеется, к этим критериям окончания счета надо относиться критически, поскольку они могут выполняться и вдали от искомой точки минимума. К сожалению, надежных критериев окончания счета, которые гарантировали бы получение решения задачи (1) с требуемой точностью, и применимых к широкому классу задач, пока нет. Сделанное замечание о критериях окончания счета относится и к другим излагаемым ниже методам.

В теоретических вопросах, когда исследуется сходимость метода, предполагается, что процесс (3) продолжается неограниченно и приводит к последовательности $\{u_k\}$. Здесь возникают вопросы, будет ли полученная последовательность $\{u_k\}$ минимизирующей для задачи (1), будет ли она сходиться к множеству точек минимума

$$U_* = \left\{ u \in E^n, \; J(u) = J_* = \inf_{E^n} J(u) \right\},$$

или, иначе говоря, выполняются ли соотношения

$$\lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} \rho(u_k, U_*) = 0? \tag{11}$$

Для положительного ответа на эти вопросы на функцию $J(u)$, кроме условия $J(u)\in C^1(E^n)$, приходится накладывать дополнительные более жесткие ограничения.

2. Подробнее рассмотрим эти вопросы для метода скорейшего спуска, когда в (3) величина α_k выбирается из условия (6).

Теорема 1. *Пусть* $J_*=\inf\limits_{E^n} J(u)>-\infty$, $J(u)\in C^{1,1}(E^n)$. *Тогда последовательность* $\{u_k\}$, *полученная методом* (3), (6), *при произвольном начальном приближении* u_0 *такова, что* $\lim\limits_{k\to\infty} J'(u_k)=0$. *Если при этом множество* $M_\delta(u_0)=\{u\in E^n\colon J(u)\leqslant J(u_0)+\delta\}$, *где* δ *взято из* (6), *ограничено, то* $\lim\limits_{k\to\infty}\rho(u_k, S_*)=0$, *где* $S_*=\{u\in M_\delta(u_0)\colon J'(u)=0\}$ — *множество стационарных точек функции* $J(u)$ *на* $M_\delta(u_0)$.

Доказательство. Если при некотором $k\geqslant 0$ окажется, что $J'(u_k)=0$, то из (3), (6) формально получаем $u_k=u_{k+1}=\ldots$ и утверждения теоремы становятся тривиальными. Поэтому будем считать, что $J'(u_k)\neq 0$ при всех $k=0, 1, \ldots$ Так как $J(u_{k+1})=f_k(\alpha_k)\leqslant\inf\limits_{\alpha\geqslant 0} f_k(\alpha)+\delta_k\leqslant J(u_k-\alpha J'(u_k))+\delta_k$ при всех $\alpha\geqslant 0$, то из неравенства (2.3.7) при $v=u_k$, $u=u_k-\alpha J'(u_k)$ имеем

$$J(u_k)-J(u_{k+1})\geqslant J(u_k)-J(u_k-\alpha J'(u_k))-\delta_k\geqslant$$
$$\geqslant\alpha|J'(u_k)|^2-L\alpha^2|J'(u_k)|^2/2-\delta_k\geqslant\alpha(1-\alpha L/2)|J'(u_k)|^2-\delta_k$$

при всех $\alpha\geqslant 0$ и $k=0, 1, \ldots$ Следовательно,

$$J(u_k)-J(u_{k+1})\geqslant\max_{\alpha\geqslant 0}\alpha(1-\alpha L/2)|J'(u_k)|^2-\delta_k=$$
$$=(1/(2L))|J'(u_k)|^2-\delta_k,\qquad k=0,1,\ldots \tag{12}$$

Отсюда получаем

$$J(u_{k+1})\leqslant J(u_k)+\delta_k,\qquad k=0,\ 1,\ \ldots \tag{13}$$

Так как $J(u_k)\geqslant J_*>-\infty$ $(k=0,1,\ldots)$, то из леммы 2.3.2 и (13) следует существование предела $\lim\limits_{k\to\infty} J(u_k)\geqslant J_*$. Тогда $\lim\limits_{k\to\infty}(J(u_k)-J(u_{k+1}))=0$ и из (12) будем иметь $\lim\limits_{k\to\infty} J'(u_k)=0$.

Наконец, пусть множество $M_\delta(u_0)$ ограничено. Суммируя неравенства (13) по k от 0 до $m-1$, получим

$$J(u_m)\leqslant J(u_0)+\sum_{k=0}^{m-1}\delta_k\leqslant J(u_0)+\delta,\qquad m=1,2,\ldots,$$

т. е. $\{u_k\}\in M_\delta(u_0)$. По теореме Больцано — Вейерштрасса ограниченная последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$ и $\{u_{k_m}\}\to u_*$. Пользуясь непрерывностью $J'(u)$, отсюда имеем

$\lim\limits_{m\to\infty} J'(u_{k_m}) = J'(u_*) = 0$, т. е. $u_* \in S_*$. Так как расстояние $\rho(u, S_*)$ непрерывно (см. лемму 2.1.2), то $\lim\limits_{m\to\infty} \rho(u_{k_m}, S_*) = \rho(u_*, S_*) = 0$. Отсюда следует, что числовая последовательность $\{\rho(u_k, S_*)\}$ имеет единственную предельную точку, равную нулю, т. е. $\lim\limits_{k\to\infty} \rho(u_k, S_*) = 0$. Теорема 1 доказана.

Теорема 2. *Пусть выполнены все условия теоремы* 1 *и, кроме того, функция* $J(u)$ *выпукла на* E^n. *Тогда для последовательности* $\{u_k\}$, *определяемой условиями* (3), (6), *имеют место соотношения* (11). *Если, кроме того, в* (6) $\{\delta_k\} = O(k^{-2})$, *то справедлива оценка*

$$0 \leqslant J(u_k) - J_* \leqslant c_0 k^{-1}, \qquad c_0 = \text{const} > 0. \tag{14}$$

Доказательство. Из ограниченности $M_\delta(u_0)$, непрерывности $J(u)$ согласно теореме 2.1.2 имеем $J_* > -\infty$, $U_* \neq \varnothing$, $U_* \subset M_\delta(u_0)$. Тогда для любой точки $u_* \in U_*$ с помощью неравенства (4.2.4) получаем

$$0 \leqslant J(u_k) - J_* = J(u_k) - J(u_*) \leqslant \langle J'(u_k), u_k - u_* \rangle \leqslant$$
$$\leqslant |J'(u_k)| \cdot |u_k - u_*| \leqslant d\,|J'(u_k)|, \qquad k = 0, 1, \ldots, \tag{15}$$

где $d \geqslant \operatorname{diam} M_\delta(u_0) = \sup\limits_{u, v \in M_\delta(u_0)} |u - v|$ — диаметр множества $M_\delta(u_0)$. В теореме 1 было доказано, что $\lim\limits_{k\to\infty} J'(u_k) = 0$. Отсюда и из (15) следует, что $\lim\limits_{k\to\infty} J(u_k) = J_*$. Учитывая включение $\{u_k\} \in M_\delta(u_0)$, тогда с помощью теоремы 2.1.2 получаем второе из равенств (11).

Докажем оценку (14). Обозначим $a_k = J(u_k) - J_*$. Из неравенств (12), (15) имеем $a_k - a_{k+1} = J(u_k) - J(u_{k+1}) \geqslant (1/(2L)) \times |J'(u_k)|^2 - \delta_k \geqslant a_k^2/(2Ld^2) - \delta_k$. По условию $\delta_k = O(k^{-2})$, т. е. $0 \leqslant \delta_k \leqslant c_1 k^{-2}$ $(k = 1, 2, \ldots, \quad c_1 = \text{const} > 0)$. Полагая $A = \max\{c_1; 2Ld^2\}$, получим

$$a_{k+1} \leqslant a_k - a_k/A + Ak^{-2}, \qquad k = 1, 2, \ldots$$

Отсюда и из леммы 2.3.5 при $I_0 = \{1, 2, \ldots\}$, $I_1 = \varnothing$ следует оценка (14). Если $\delta_k = 0$ $(k = 0, 1, \ldots)$, то оценка (14) вытекает из неравенств (12), (15) и леммы 2.3.4. Теорема 2 доказана.

Теорема 3. *Пусть* $J(u) \in C^{1,1}(E^n)$, $J(u)$ *сильно выпукла на* E^n. *Тогда для последовательности* $\{u_k\}$, *получаемой из* (3), (6) *при любом начальном приближении* u_0, *справедливы соотношения* (11). *Если при этом* $\delta_k = O(k^{-2})$, *то имеет место оценка* (14). *Если* $\delta_k = 0$ $(k = 0, 1, \ldots)$, *то верна более сильная, чем* (14), *оценка*

$$0 \leqslant J(u_k) - J_* \leqslant (J(u_0) - J_*)\, q^k, \tag{16}$$

$$|u_k - u_*|^2 \leqslant (2/\mu)(J(u_0) - J_*)\, q^k, \qquad k = 0, 1, \ldots, \tag{17}$$

где u_* — точка минимума $J(u)$ на E^n, $q=1-\mu/L$, $0\leqslant q<1$, μ — постоянная из теоремы 4.3.3.

Доказательство. Согласно теореме 4.3.1 множество $M_\delta(u_0)$ ограничено, $J_*>-\infty$, U_* состоит из единственной точки u_*. Поэтому равенства (11) и оценка (14) следуют из теорем 1, 2. Докажем оценки (15), (16). Из (4.3.7) при $v=u_k$, $u=u_*$ имеем

$$a_k=J(u_k)-J(u_*)\leqslant\langle J'(u_k),u_k-u_*\rangle-\varkappa|u_k-u_*|^2\leqslant$$
$$\leqslant|J'(u_k)|\cdot|u_k-u_*|-\varkappa|u_k-u_*|^2\leqslant$$
$$\leqslant\sup_{z\geqslant0}(|J'(u_k)|z-\varkappa z^2)=|J'(u_k)|^2/(4\varkappa),$$

т. е.

$$a_k=J(u_k)-J(u_*)\leqslant|J'(u_k)|^2/(4\varkappa),\qquad k=0,1,\ldots \tag{18}$$

Подставив неравенство (18) в правую часть (12) при $\delta_k=0$, получим

$$a_k-a_{k+1}\geqslant\frac{4\varkappa}{2L}a_k=\frac{2\varkappa}{L}a_k,\qquad k=0,1,\ldots$$

В § 4.3 было установлено, что $2\varkappa=\mu\leqslant L$. Поэтому $0\leqslant q=1-(\mu/L)<1$, и предыдущее неравенство можно переписать в виде $0\leqslant a_{k+1}\leqslant a_k(1-\mu/L)=qa_k$. Отсюда имеем $a_k\leqslant qa_{k-1}\leqslant q^2a_{k-2}\leqslant\ldots\leqslant q^ka_0$, что равносильно оценке (16). Наконец, из неравенства (4.3.2) следует

$$\varkappa|u_k-u_*|^2\leqslant J(u_k)-J(u_*)=a_k,\qquad k=0,1,\ldots$$

Отсюда и из (16) получим оценку (17). Теорема 3 доказана.

Метод скорейшего спуска имеет простой геометрический смысл: оказывается, точка u_{k+1}, определяемая условиями (3), (4), лежит на луче $L_k=\{u\colon u=u_k-\alpha J'(u_k),\ \alpha\geqslant0\}$ в точке его касания линии уровня (при $n\geqslant3$ — поверхности уровня) $\Gamma_{k+1}=\{u\in E^n\colon J(u)=J(u_{k+1})\}$, а сам луч L_k перпендикулярен к линии уровня $\Gamma_k=\{u\in E^n\colon J(u)=J(u_k)\}$ — см. рис. 5.1 и 5.2. В самом деле, пусть $u=u(t)$, $a\leqslant t\leqslant b$ — некоторое параметрическое уравнение линии уровня Γ_k, т. е. $J(u(t))=J(u_k)=\mathrm{const}$, $a\leqslant t\leqslant b$, причем $u(t_0)=u_k$. Тогда $\frac{d}{dt}J(u(t))=\langle J'(u(t)),\ \dot u(t)\rangle=0$, $a\leqslant t\leqslant b$. В частности, при $t=t_0$ имеем $\langle J'(u_k),\ \dot u(t_0)\rangle=0$. Это означает, что градиент (или антиградиент) $J'(u_k)$ перпендикулярен к касательному направлению поверхности уровня Γ_k в точке u_k, или, иначе говоря, луч L_k перпендикулярен к Γ_k. Далее, из условия (4) при $\alpha_k>0$ получаем $f_k'(\alpha_k)=-\langle J'(u_k-\alpha_kJ'(u_k)),\ J'(u_k)\rangle=-\langle J'(u_{k+1}),\ J'(u_k)\rangle=0$. Но вектор $J'(u_{k+1})$ перпендикулярен к Γ_{k+1} в точке u_{k+1}, поэтому последнее равенство означает, что направление $J'(u_k)$ и, следовательно, луч L_k являются касательными к линии уровня Γ_{k+1} в точке u_{k+1}.

3. Из рис. 5.1 и 5.2 можно понять, что чем ближе линии уровня $J(u)=\text{const}$ к окружности, тем лучше сходится метод скорейшего спуска. Это же явление можно усмотреть и из оценок (16), (17) — чем ближе μ/L к единице (для функции $J(u)=|u|^2$, у которой линиями уровня являются окружности (сферы), как раз имеем $\mu/L=1$), тем ближе q к нулю и тем лучше сходимость.

Те же рис. 5.1 и 5.2 показывают, а теоретические исследования и численные эксперименты подтверждают, что метод скорейшего спуска и другие варианты градиентного метода медленно

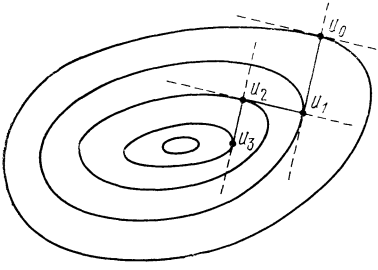

Рис. 5.1

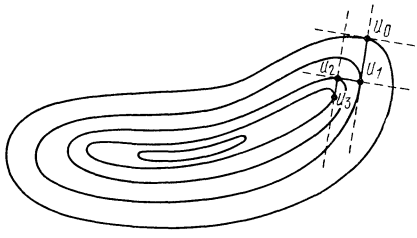

Рис. 5.2

сходятся в тех случаях, когда поверхности уровня функции $J(u)$ сильно вытянуты и функция имеет так называемый «овражный» характер. Это означает, что небольшое изменение некоторых переменных приводит к резкому изменению значений функции — эта группа переменных характеризует «склон оврага», а по остальным переменным, задающим направление «дна оврага», функция меняется незначительно (на рис. 5.2 и 5.3 изображены

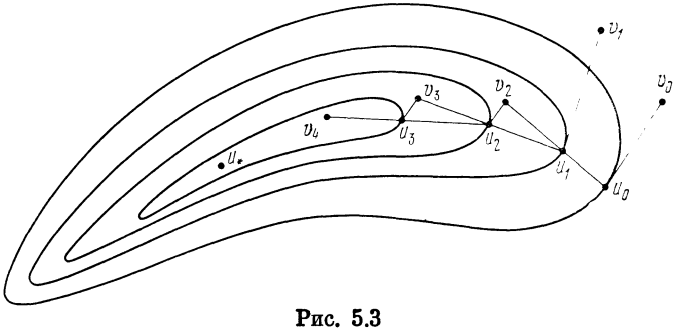

Рис. 5.3

линии уровня «овражной» функции двух переменных). Если точка лежит на «склоне оврага», то направление спуска из этой точки будет почти перпендикулярным к направлению «дна оврага», и в результате приближения $\{u_k\}$, получаемые градиентным методом, будут поочередно находиться то на одном, то на дру-

гом «склоне оврага». Если «склоны оврага» достаточно круты, то такие скачки «со склона на склон» точек u_k могут сильно замедлить сходимость градиентного метода.

Для ускорения сходимости этого метода при поиске минимума «овражной» функции можно предложить следующий эвристический прием, называемый овражным методом. Сначала опишем простейший вариант этого метода. В начале поиска задаются две точки v_0, v_1, из которых производят спуск с помощью какого-либо варианта градиентного метода, и получают две точки u_0, u_1 на «дне оврага». Затем полагают

$$v_2 = u_1 - (u_1 - u_0)\,|u_1 - u_0|^{-1} h \operatorname{sign}(J(u_1) - J(u_0)),$$

где h — положительная постоянная, называемая овражным шагом. Из точки v_2, которая, вообще говоря, находится на «склоне оврага», производят спуск с помощью градиентного метода и определяют следующую точку u_2 на «дне оврага».

Если уже известны точки $u_0, u_1, \ldots, u_k$ $(k \geqslant 2)$, то из точки

$$v_{k+1} = u_k - (u_k - u_{k-1})\,|u_k - u_{k-1}|^{-1} h \operatorname{sign}[J(u_k) - J(u_{k-1})] \quad (19)$$

совершают спуск с помощью градиентного метода и находят следующую точку u_{k+1} на «дне оврага» (см. рис. 5.3; спуск из точки v_k в точку u_k, состоящий, быть может, из нескольких итерационных шагов градиентного метода, условно изображен отрезком прямой, соединяющей точки v_k, u_k, $k = 0, 1, \ldots$).

Величина овражного шага h подбирается эмпирически с учетом информации о минимизируемой функции, получаемой в ходе поиска минимума. От правильного выбора h существенно зависит скорость сходимости метода. Если шаг h велик, то на крутых поворотах «оврага» точки v_k могут слишком удаляться от «дна оврага» и спуск из точки v_k в точку u_k может потребовать большого объема вычислений. Кроме того, при больших h на крутых поворотах может произойти выброс точки v_k из «оврага», и правильное направление поиска точки минимума будет потеряно. Если шаг h слишком мал, то поиск может очень замедлиться и эффект от применения овражного метода может стать незначительным.

Эффективность овражного метода может существенно возрасти, если величину овражного шага выбирать переменной, реагирующей на повороты «оврага» с тем, чтобы: 1) по возможности быстрее проходить прямолинейные участки на «дне оврага» за счет увеличения овражного шага; 2) на крутых поворотах «оврага» избежать выброса из «оврага» за счет уменьшения овражного шага; 3) добиться по возможности меньшего отклонения точек v_k от «дна оврага» и тем самым сократить объем вычислений, требуемый для градиентного спуска из точки v_k в точку u_k $(k = 0, 1, \ldots)$. Интуитивно ясно, что для правильной реакции на поворот «оврага» надо учитывать «кривизну дна оврага», причем

информацию о «кривизне» желательно получить, опираясь на результаты предыдущих итераций овражного метода.

В работе [276] предлагается следующий способ выбора овражного шага:

$$h_{k+1} = h_k \cdot c^{\cos\alpha_k - \cos\alpha_{k-1}}, \qquad k = 2, 3, \ldots, \tag{20}$$

где α_k — угол между векторами $v_k - u_{k-1}$, $u_k - u_{k-1}$, определяемый условием

$$\cos\alpha_k = \langle v_k - u_{k-1}, u_k - u_{k-1}\rangle |v_k - u_{k-1}|^{-1} |u_k - u_{k-1}|^{-1},$$

а постоянная $c > 1$ является параметром алгоритма. Точка v_{k+1} определяется из (19) при $h = h_{k+1}$. Разность $\cos\alpha_k - \cos\alpha_{k-1}$ в равенстве (20) связана с «кривизной дна оврага» и, кроме того, обладает важным свойством указывать направление изменения «кривизны». А именно, при переходе с участков «дна оврага» с малой «кривизной» на участки с большей «кривизной» будем иметь $\cos\alpha_k - \cos\alpha_{k-1} < 0$ (см. рис. 5.4). Тогда в силу (19) $h_{k+1} < h_k$, т. е. овражный шаг уменьшается, приспосабливаясь к повороту «дна оврага», что в свою очередь приводит к уменьшению выбросов точки v_{k+1} на «склоны оврага». При переходе с участков «дна оврага» с большой «кривизной» на участки с меньшей «кривизной», наоборот, $\cos\alpha_k - \cos\alpha_{k-1} > 0$, поэтому овражный шаг увеличится и появится возможность сравнительно быстро пройти участки с малой «кривизной», в частности, прямолинейные участки на «дне оврага». Если «кривизна дна оврага» на некоторых участках остается постоянной, то разность $\cos\alpha_k - \cos\alpha_{k-1}$ будет близка к нулю, и поиск минимума на таких участках будет проводиться с почти постоянным шагом, сформированным с учетом величины «кривизны» при выходе на рассматриваемый участок.

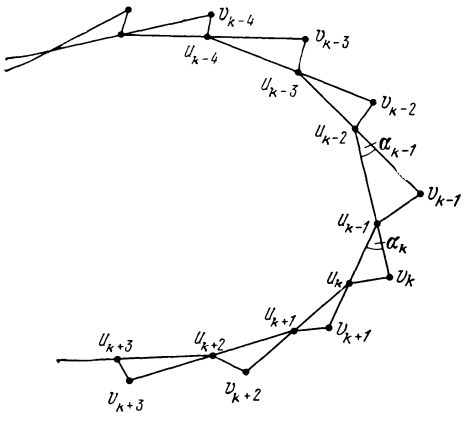

Рис. 5.4

Параметр c в равенстве (20) регулирует «чувствительность» метода к изменению «кривизны дна оврага», и правильный выбор этого параметра во многом определяет скорость движения по «оврагу». Некоторые эвристические соображения по поводу выбора c и другие аспекты применения овражного метода обсуждены в [276]. Выражение (20) для овражного шага удобнее пре-

образовать так:

$$h_{k+1} = h_k c^{\cos\alpha_k - \cos\alpha_{k-1}} = h_{k-1} c^{\cos\alpha_k - \cos\alpha_{k-2}} = \ldots = h_2 c^{\cos\alpha_k - \cos\alpha_1},$$

откуда имеем

$$h_{k+1} = A c^{\cos\alpha_k}, \quad A = h_2 c^{-\cos\alpha_1} = \text{const} > 0, \quad k = 2, 3, \ldots$$

Другой способ ускорения сходимости градиентного метода заключается в выборе подходящей замены переменных $u = g(\xi) = (g_1(\xi), \ldots, g_n(\xi))$ с тем, чтобы поверхности уровня функции $J(g(\xi)) = G(\xi)$ в пространстве переменных $\xi = (\xi^1, \ldots, \xi^n)$ были близки к сферам. Заметим, что $G'(\xi) = (g'(\xi))^T J'(g(\xi))$, где $g'(\xi) = \left\{g_{i\xi^j}(\xi)\right\}$ — матрица, i-я строка которой представляет собой $g_i'(\xi) = \left(g_{i\xi^1}(\xi), \ldots, g_{i\xi^n}(\xi)\right)$, а $(g'(\xi))^T$ — матрица, полученная из $g'(\xi)$ транспонированием. В пространстве переменных ξ градиентный метод выглядит так:

$$\xi_{k+1} = \xi_k - \beta_k (g'(\xi_k))^T J'(g(\xi_k)), \quad \beta_k > 0, \quad k = 0, 1, \ldots$$

В пространстве исходных переменных $u = (u^1, \ldots, u^n)$ этот подход можно трактовать как итерационный процесс вида

$$u_{k+1} = u_k - \alpha_k A_k J'(u_k), \quad \alpha_k > 0, \quad k = 0, 1, \ldots,$$

где A_k — некоторая невырожденная матрица порядка $n \times n$, представляющая собой параметр метода. То, что на этом пути можно добиться существенного ускорения скорости сходимости итераций, подтверждается, например, излагаемым ниже методом Ньютона, в котором полагается $A_k = (J''(u_k))^{-1}$ $(k = 0, 1, \ldots)$. О методах минимизации овражных функций и различных приемах ускорения сходимости итерационных методов см. [4, 19, 54, 111, 229, 238, 250, 258, 284, 307, 314, 336].

4. Исследуем сходимость другого варианта градиентного метода (3), в котором параметр α_k определяется из условия (9) с помощью дробления. А именно, пусть $1 < \varepsilon < 2$, $\alpha > 0$ — фиксированные числа, а $i \geqslant 0$ — наименьший номер, для которого выполняется неравенство [11, 19, 56]

$$J(u_k) - J(u_k - 2^{-i}\alpha J'(u_k)) \geqslant 2^{-i-1}\alpha\varepsilon |J'(u_k)|^2, \tag{21}$$

и пусть

$$\alpha_k = \alpha/2^i. \tag{22}$$

Теорема 4. *Пусть в задаче* (1) $J_* > -\infty$, $U_* \neq \varnothing$, *функция* $J(u)$ *выпукла на* E^n, $J(u) \in C^{1,1}(E^n)$. *Тогда для последовательности* $\{u_k\}$, *определяемой методом* (3), (21), (22), *имеют место соотношения* (11) *и, более того, существует точка* $v_* \in U_*$ *такая, что* $\{u_k\} \to v_*$,

$$|u_{k+1} - v_*| \leqslant |u_k - v_*|, \quad \rho(u_{k+1}, U_*) \leqslant \rho(u_k, U_*), \quad k = 0, 1, \ldots, \tag{23}$$

причем равенство в (23) *возможно лишь при* $u_k = u_{k+1} = \ldots = v_*$; *справедлива оценка*

$$0 \leqslant J(u_k) - J_* \leqslant \{\min\{(2-\varepsilon)/(2L); \alpha\}\}^{-1} (2/\varepsilon) |u_0 - v_*|^2 k^{-1} = O(1/k),$$
$$k = 1, 2, \ldots, \tag{24}$$

и если U — аффинное множество, то $v_* = \mathcal{P}_{U_*}(u_0)$, т. е. v_* — ближайшая к u_0 точка из U_*.

Доказательство. Сначала покажем возможность выбора α_k из условий (21), (22). Пусть $j \geqslant 0$ — наименьший номер, для которого

$$L \cdot 2^{-j}\alpha \leqslant 2 - \varepsilon; \tag{25}$$

здесь $L > 0$ — константа Липшица для $J'(u)$. Из неравенства (2.3.7) при $v = u_k$, $u = u_k - 2^{-j}\alpha J'(u_k)$ с учетом (25) имеем

$$J(u_k) - J(u_k - 2^{-j}\alpha J'(u_k)) \geqslant \langle J'(u_k),\ 2^{-j}\alpha J'(u_k)\rangle - L \cdot 2^{-j-1}\alpha |J'(u_k)|^2 = \\ = 2^{-j-1}\alpha(2 - 2^{-j}\alpha L)|J'(u_k)|^2 \geqslant 2^{-j-1}\alpha\varepsilon |J'(u_k)|^2. \tag{26}$$

Это значит, что при $i = j$ неравенство (21) выполняется, и, следовательно, минимальный номер $i \geqslant 0$, при котором справедливо (21), существует и не превышает номера j из (25). Покажем, что для α_k из (21), (22) справедлива оценка

$$\alpha_k \geqslant \min\{(2-\varepsilon)/(2L);\ \alpha\}, \qquad k = 0, 1, \ldots \tag{27}$$

Сначала рассмотрим случай $\alpha > (2-\varepsilon)/(2L)$. Тогда оказывается, $\alpha_k > (2-\varepsilon)/(2L)$ при всех $k = 0, 1, \ldots$ В самом деле, для номера j из (25) в этом случае имеем $2^{-j}\alpha \leqslant (2-\varepsilon)/L < 2^{-j+1}\alpha$ $(j \geqslant 0)$. Поэтому с учетом правила выбора номера i, определения α_k из (22) и неравенства $i \leqslant j$ получим $\alpha_k = \alpha/2^i \geqslant \alpha/2^j > (2-\varepsilon)/(2L)$. Пусть теперь $\alpha \leqslant (2-\varepsilon)/(2L)$. Тогда неравенство (25) и, следовательно, (26) выполняется при $j = 0$. Отсюда и из (21) следует, что $i = 0$. Согласно (22) тогда $\alpha_k = \alpha/2^0 = \alpha$ $(k = 0, 1, \ldots)$. Объединяя оба рассмотренных случая, приходим к оценке (27).

Далее, возьмем любую точку $u_* \in U_*$. Из (3), (21), (22) и теоремы 4.2.2 имеем

$$(\varepsilon/2)\,\alpha_k\,|J'(u_k)|^2 \leqslant J(u_k) - J(u_{k+1}) \leqslant J(u_k) - J(u_*) \leqslant \langle J'(u_k),\ u_k - u_*\rangle. \tag{28}$$

Кроме того, из (3) следует $|u_{k+1} - u_*|^2 = |u_k - \alpha_k J'(u_k) - u_*|^2 = |u_k - u_*|^2 - 2\alpha_k\langle J'(u_k), u_k - u_*\rangle + \alpha_k^2 |J'(u_k)|^2$. Отсюда с учетом оценки (28) получаем

$$|u_{k+1} - u_*|^2 \leqslant |u_k - u_*|^2 - (\varepsilon - 1)\,\alpha_k^2\,|J'(u_k)|^2, \qquad 1 < \varepsilon < 2. \tag{29}$$

Следовательно,

$$|u_{k+1} - u_*|^2 \leqslant |u_k - u_*|^2 \leqslant \ldots \leqslant |u_0 - u_*|^2 \qquad \forall u_* \in U_*. \tag{30}$$

Из (30) вытекает существование предела $\lim\limits_{k\to\infty} |u_k - u_*|^2$ и ограниченность последовательности $\{u_k\}$. Тогда найдется подпоследовательность $\{u_{k_m}\}$, сходящаяся к некоторой точке v_*. Из (27), (29) следует, что $\{J'(u_{k_m})\} \to J'(v_*) = 0$. По теореме 4.2.3 тогда $v_* \in U_*$. Приняв $u_* = v_*$, из (30) получаем $\lim\limits_{k\to\infty} |u_k - v_*| = \lim\limits_{m\to\infty} |u_{k_m} - v_*| = 0$, т. е. вся последовательность $\{u_k\}$ сходится к точке v_*. Отсюда и из (29), (30) следуют неравенства (23). Как видно из (29), равенство в (23) возможно лишь при $J'(u_k) = 0$. Тогда в силу теоремы 4.2.3 $u_k = v_* \in U_*$, и процесс (3), (21), (22) на этом заканчивается.

Докажем оценку (24). Обозначим $a_k = J(u_k) - J_*$. Из (28), (30) при $u_* = v_*$ имеем

$$(\varepsilon/2)\,\alpha_k\,|J'(u_k)|^2 \leqslant a_k - a_{k+1}, \qquad a_k \leqslant |J'(u_k)|\,|u_0 - v_*|, \quad k = 0, 1, \ldots$$

Отсюда с учетом (27) получаем $a_k - a_{k+1} \geqslant (\varepsilon/2)\min\{(2-\varepsilon)/(2L);\ \alpha\} \times |u_0 - v_*|^{-2} a_k^2$ $(k = 0, 1, \ldots)$. Из леммы 2.3.4 тогда следует оценка (24).

Наконец, пусть U_* — аффинное множество. При $k \to \infty$ из (30) имеем $|v_* - u_*|^2 \leqslant |u_0 - u_*|^2$ при любом $u_* \in U_*$. В частности, в этом неравенстве можно взять $u_* = v_* + \alpha(\mathcal{P}_{U_*}(u_0) - v_*) = v_\alpha \in U_*$, $\alpha > 0$. Получим $|u_0 - v_\alpha|^2 \geqslant |v_* - v_\alpha|^2 = |(v_* - u_0) - (v_\alpha - u_0)|^2 = |v_* - u_0|^2 + |v_\alpha - u_0|^2 - 2\langle v_* - u_0, v_\alpha - u_0\rangle = |v_\alpha - u_0|^2 - |v_* - u_0|^2 - 2\alpha\langle v_* - u_0, \mathcal{P}_{U_*}(u_0) - v_*\rangle$ или

$$|v_* - u_0|^2 \geqslant 2\alpha\,|v_* - \mathcal{P}_{U_*}(u_0)|^2 + 2\alpha\langle \mathcal{P}_{U_*}(u_0) - u_0,\ v_* - \mathcal{P}_{U_*}(u_0)\rangle.$$

Отсюда с учетом равенства (4.4.2) имеем $|v_* - u_0|^2 \geqslant 2\alpha\,|v_* - \mathcal{P}_{U_*}(u_0)|^2$ при всех $\alpha > 0$. Разделив это неравенство на $\alpha > 0$ и устремив $\alpha \to \infty$, получим $v_* = \mathcal{P}_{U_*}(u_0)$. Теорема 4 доказана.

5. Следуя [229], рассмотрим метод, представляющий собой комбинацию несколько модифицированного метода (3), (21), (22) и овражного метода.

Возьмем начальные приближения: $v_0 \in E^n$, $b_0 = 1$, $\alpha_{-1} > 0$, положим $u_{-1} = v_0$. Пусть для некоторого $k \geqslant 0$ уже известны $v_k \in E^n$, $b_k \geqslant 1$, $\alpha_{k-1} > 0$, $u_{k-1} \in E^n$. Определим наименьший номер $i \geqslant 0$, для которого выполняется неравенство

$$J(v_k) - J(v_k - 2^{-i}\alpha_{k-1}J'(v_k)) \geqslant 2^{-i-1}\alpha_{k-1}\,|J'(v_k)|^2. \tag{31}$$

Далее, положим

$$\alpha_k = \alpha_{k-1}/2^i, \qquad u_k = v_k - \alpha_k J'(v_k), \tag{32}$$

$$b_{k+1} = \frac{1}{2}\left(1 + \sqrt{4b_k^2 + 1}\right), \qquad v_{k+1} = u_k + \frac{b_k - 1}{b_{k+1}}(u_k - u_{k-1}). \tag{33}$$

Таким образом, в описанном методе (31)—(33) спуск из точки v_k на «дно оврага» осуществляется по формулам (31), (32) с помощью одного шага градиентного метода (3) с правилом выбора параметра α_k, близким к (21), (22). Здесь возможно использование некоторых других вариантов градиентного метода, например, по аналогии с (8) в (32) можно взять $\alpha_k = 1/L$. Как видно из (33), пересчет точки v_k осуществляется с помощью овражного метода по формуле, близкой к (19). Первое из равенств (33) представляет собой правило пересчета длины овражного шага; величина b_{k+1} является положительным корнем квадратного уравнения $x^2 - x - b_k^2 = 0$, так что

$$b_{k+1}^2 - b_{k+1} = b_k^2, \qquad b_0 = 1, \qquad b_k > 0, \quad k = 0, 1, \ldots \tag{34}$$

С помощью индукции нетрудно получить оценку

$$b_k > k, \qquad k = 1, 2, \ldots \tag{35}$$

Теорема 5. *Пусть функция $J(u)$ выпукла на E^n, $J(u) \in C^{1,1}(E^n)$, $J_* > -\infty$, $U_* \neq \varnothing$, последовательность $\{u_k\}$ определена методом*

(31) — (33). *Тогда*

$$0 \leqslant J(u_k) - J_* \leqslant (\min\{1/(2L);\ \alpha_{-1}\})^{-1} (2\alpha_0 (J(u_0) - J_*) + \\ + \inf_{u \in U_*} |u - u_0|^2)/(2k^2) = O(1/k^2), \qquad k = 1, 2, \ldots \tag{36}$$

Доказательство. Пусть $j \geqslant 0$ — наименьший номер, для которого

$$2^{-j}\alpha_{k-1} \leqslant 1/L. \tag{37}$$

Нетрудно видеть, что тогда

$$J(v_k) - J(v_k - 2^{-j}\alpha_{k-1}J'(v_k)) \geqslant 2^{-j-1}\alpha_{k-1}|J'(v_k)|^2; \tag{38}$$

это неравенство доказывается так же, как (26). Отсюда следует существование номера $i \leqslant j$, удовлетворяющего неравенству (31). Рассуждая так же, как при доказательстве оценки (27), из (31), (32), (37), (38) с помощью индукции получаем

$$\alpha_k \geqslant \min\{1/(2L);\ \alpha_{-1}\}, \qquad k = 0, 1, \ldots \tag{39}$$

Обозначим $p_k = (b_k - 1)(u_{k-1} - u_k)$. Тогда из (33) следует

$$v_{k+1} = u_k - p_k/b_{k+1}, \qquad p_k = b_{k+1}(u_k - v_{k+1}). \tag{40}$$

Далее, с учетом (32), (40) имеем

$$p_{k+1} - u_{k+1} = (b_{k+1} - 1)(u_k - u_{k+1}) - u_{k+1} = b_{k+1}(u_k - u_{k+1}) - u_k = \\ = b_{k+1}(u_k - v_{k+1} + \alpha_{k+1}J'(v_{k+1})) - u_k = p_k - u_k + \alpha_{k+1}b_{k+1}J'(v_{k+1}).$$

Тогда для любого $u_* \in U_*$ получаем

$$p_{k+1} - u_{k+1} + u_*|^2 = |p_k - u_k + u_*|^2 + 2\alpha_{k+1}b_{k+1}\langle J'(v_{k+1}),\ p_k - u_k + u_*\rangle + \\ + \alpha_{k+1}^2 b_{k+1}^2 |J'(v_{k+1})|^2,$$

или с учетом (40)

$$|p_{k+1} - u_{k+1} + u_*|^2 - |p_k - u_k + u_*|^2 \leqslant 2\alpha_{k+1}\langle J'(v_{k+1}),\ (b_{k+1}p_k - p_k) + \\ + (p_k - b_{k+1}u_k) + b_{k+1}u_*\rangle + \alpha_{k+1}^2 b_{k+1}^2 |J'(v_{k+1})|^2 = 2\alpha_{k+1}(b_{k+1} - 1) \times \\ \times \langle J'(v_{k+1}),\ p_k\rangle + 2\alpha_{k+1}b_{k+1}\langle J'(v_{k+1}),\ u_* - v_{k+1}\rangle + \alpha_{k+1}^2 b_{k+1}^2 |J'(v_{k+1})|^2, \\ k = 0, 1, \ldots \tag{41}$$

Заметим, что (31) с учетом (32) можно переписать в виде $J(v_k) - J(u_k) \geqslant (\alpha_k/2)|J'(v_k)|^2$. Для $k+1$-й итерации это неравенство имеет вид

$$J(v_{k+1}) \geqslant J(u_{k+1}) + \frac{1}{2}\alpha_{k+1}|J'(v_{k+1})|^2. \tag{42}$$

Из теоремы 4.2.2 с учетом (42) получаем

$$\langle J'(v_{k+1}),\ u_* - v_{k+1}\rangle \leqslant J(u_*) - J(v_{k+1}) \leqslant J_* - J(u_{k+1}) - \frac{1}{2}\alpha_{k+1}|J'(v_{k+1})|^2. \tag{43}$$

Далее, из теоремы 4.2.2 и из (40), (42) следует

$$(\alpha_{k+1}/2)|J'(v_{k+1})|^2 \leqslant J(v_{k+1}) - J(u_{k+1}) \leqslant J(u_k) - \langle J'(v_{k+1}),\ u_k - v_{k+1}\rangle - \\ - J(u_{k+1}) = J(u_k) - J(u_{k+1}) - \langle J'(v_{k+1}),\ p_k\rangle/b_{k+1},$$

откуда

$$\langle J'(v_{k+1}),\ p_k\rangle \leqslant b_{k+1}((J(u_k) - J(u_{k+1})) - (\alpha_{k+1}/2)|J'(v_{k+1})|^2). \tag{44}$$

Обозначим $a_k = J(u_k) - J_*$. Подставим оценки (43), (44) в (41). С учетом (32), (34) получим

$$|p_{k+1} - u_{k+1} + u_*|^2 - |p_k - u_k + u_*|^2 \leqslant$$
$$\leqslant 2\alpha_{k+1}(b_{k+1} - 1)\, b_{k+1}(a_k - a_{k+1} - (\alpha_{k+1}/2)\,|J'(v_{k+1})|^2) +$$
$$+ 2\alpha_{k+1} b_{k+1}(-a_{k+1} - (\alpha_{k+1}/2)\,|J'(v_{k+1})|^2) + \alpha_{k+1}^2 b_{k+1}^2 |J'(v_{k+1})|^2 =$$
$$= 2\alpha_{k+1} b_k^2 a_k - 2\alpha_{k+1} a_{k+1}(b_k^2 + b_{k+1}) \leqslant 2\alpha_k b_k^2 a_k - 2\alpha_{k+1} b_{k+1}^2 a_{k+1}.$$

Таким образом,

$$|p_{m+1} - u_{m+1} + u_*|^2 - |p_m - u_m + u_*|^2 \leqslant 2\alpha_m b_m^2 a_m - 2\alpha_{m+1} b_{m+1}^2 a_{m+1},$$
$$m = 0, 1, \ldots$$

Суммируя эти неравенства по m от 0 до некоторого $m = k - 1$, получим

$$|p_k - u_k + u_*|^2 + 2\alpha_k b_k^2 a_k \leqslant 2\alpha_0 b_0^2 a_0 + |p_0 - u_0 + u_*|^2.$$

Отсюда с учетом равенств $b_0 = 1$, $p_0 = 0$, оценок (35), (39), произвольности выбора точки u_* из U_* приходим к оценке (36). Теорема 5 доказана.

Отметим, что метод (31)—(33) не обеспечивает монотонное убывание функции $J(u)$ на последовательностях $\{u_k\}$, $\{v_k\}$. Сравнение оценок (24) и (36) показывает, что для выпуклых гладких задач овражный метод имеет более высокую скорость сходимости, чем градиентный метод (3), (9). В [228] показано, что оценка $J(u_k) - J_* = O(1/k^2)$ является неулучшаемой на этом классе функций среди всех методов, использующих лишь значения $J(u)$, $J'(u)$.

6. Кратко остановимся на непрерывном аналоге градиентного метода. Для этого перепишем формулу (3) в безындексной форме, приняв $u_k = u(t)$, $u_{k+1} = u(t + \Delta t)$, $\alpha_k = \Delta t \beta(t)$, $\Delta t > 0$, $\beta(t) > 0$. Получим

$$(u(t + \Delta t) - u(t))/\Delta t = -\beta(t) J'(u(t)), \quad t \geqslant 0; \quad u(0) = u_0.$$

Отсюда, формально переходя к пределу при $\Delta t \to +0$, придем к следующей задаче Коши:

$$\dot{u}(t) = -\beta(t) J'(u(t)), \quad t \geqslant 0; \quad u(0) = u_0. \tag{45}$$

Задача (45) представляет собой непрерывный аналог градиентного метода, а исходный процесс (3) является методом ломаных Эйлера для решения задачи (45). Понятно, что задачу Коши (45) можно решать и другими численными методами, которые, возможно, будут сходиться быстрее метода Эйлера и лучше приспособлены для минимизации овражных функций [4, 13, 39, 54, 258].

Определение 1. Траекторию (решение) $u(t)$ задачи (45) будем называть *минимизирующей*, если $u(t)$ определена при всех $t \geqslant 0$ и $\lim\limits_{t\to\infty} J(u(t)) = J_*$.

Ограничимся следующей теоремой о сходимости метода (45).

Теорема 6. *Пусть функция $J(u)$ сильно выпукла на E^n и $J(u) \in C^{1,1}(E^n)$, функция $\beta(t)$ определена, непрерывна и $\beta(t) \geqslant \beta_0 > 0$ при всех $t \geqslant 0$. Тогда траектория задачи (45) при любом выборе начальной точки u_0 является минимизирующей и сходится к точке минимума u_* с оценкой*

$$|u(t) - u_*| \leqslant |u_0 - u_*| \exp\{-\mu\beta_0 t\}, \quad t \geqslant 0, \tag{46}$$

где постоянная μ взята из теоремы 4.3.3.

Доказательство. Прежде всего заметим, что по теореме 4.3.1 точка минимума u_* функции $J(u)$ на E^n существует и единственна, а по

теореме 4.2.3 $J'(u_*)=0$. Далее, из доказываемой ниже теоремы 6.1.1 следует, что траектория задачи (45) определена при всех $t \geqslant 0$ для любой начальной точки u_0. Положим

$$V(t) = |u(t) - u_*|^2/2, \quad t \geqslant 0. \tag{47}$$

Тогда с учетом условий (45) и теоремы 4.3.3 имеем

$$\dot V(t) = \langle u(t) - u_*, \dot u(t)\rangle = -\beta(t)\langle J'(u(t)) - J'(u_*), u(t) - u_*\rangle \leqslant$$
$$\leqslant -\mu\beta_0 |u(t) - u_*|^2 = -2\mu\beta_0 V(t), \quad t \geqslant 0; \quad V(0) = |u_0 - u_*|^2/2.$$

Отсюда следует $\frac{d}{dt}(V(t)\exp\{+2\mu\beta_0 t\}) \leqslant 0$ $(t \geqslant 0)$. Интегрируя это неравенство, получим

$$0 \leqslant V(t) \leqslant V(0)\exp\{-2\mu\beta_0 t\} = |u_0 - u_*|^2 \exp\{-2\mu\beta_0 t\}/2,$$

что равносильно оценке (46). В силу непрерывности функции $J(u)$ тогда $\lim\limits_{t\to\infty} J(u(t)) = J_*$, что и требовалось.

Пользуясь терминологией, принятой в теории устойчивости обыкновенных дифференциальных уравнений [9, 172, 251, 295], можно сказать, что в теореме 6 доказана асимтотическая устойчивость системы (45) относительно точки равновесия u_* этой системы. Для доказательства этого факта использован второй метод Ляпунова, в качестве функции Ляпунова была взята функция (47). В связи с этим полезно заметить, что при исследовании многих методов минимизации явно или неявно используется второй метод Ляпунова или его дискретный аналог; в качестве функции Ляпунова наряду с (47) часто используются также функции $V(t) = J(u(t)) - J_*$, $V(t) = |J'(u(t))|^2$ и др. Систематическое исследование сходимости методов минимизации с помощью метода Ляпунова проведено в [49].

Существуют и другие дифференциальные уравнения, траектории которых являются минимизирующими. Например, так называемый метод тяжелого шарика [4] заключается в рассмотрении системы дифференциальных уравнений

$$\ddot u(t) + \gamma\dot u(t) + J'(u(t)) = 0, \quad t \geqslant 0, \quad \gamma = \text{const} > 0.$$

Оказывается, траектории этой системы при довольно широких предположениях сходятся к точке минимума функции $J(u)$ на E^n, причем скорость сходимости, вообще говоря, выше, чем у траекторий системы (45).

7. В заключение отметим, что градиентный метод, вообще говоря, хорошо работает лишь на первых этапах поиска минимума, когда точки u_k из (3) не слишком близки к точке минимума u_*, а вблизи точки u_* расстояние $|u_k - u_*|$ часто перестает уменьшаться, сходимость метода ухудшается. Это связано с тем, что в окрестности точки минимума градиент $J'(u_k)$ близок к нулю, главная линейная часть приращения $J(u_k) - J(u_*)$, на базе которой выбирается направление спуска в методе (3), становится малой, учитывается влияние квадратичной части приращения, метод (3) становится слишком чувствительным к неизбежным погрешностям вычислений. Поэтому вблизи точки минимума при необходимости пользуются более точными и, вообще говоря, более трудоемкими методами, лучше учитывающими не только линейные, но и квадратичные части приращения.

Упражнения. 1. Описать различные варианты градиентного метода для задачи из примера 2.2.1.

2. Установить сходимость метода скорейшего спуска для функции (5); описать другие варианты градиентного метода для этой функции.

3. Рассмотреть метод скорейшего спуска и другие варианты градиентного метода для задачи минимизации функции $J(u) = |Au - b|^2$, $u \in E^n$, где A — матрица порядка $m \times n$, $b \in E^m$; исследовать их сходимость.

4. Рассмотреть метод скорейшего спуска для минимизации функций $J(u) = x^2 + ay^2$, $u = (x, y) \in E^2$, и $J(u) = x^2 + y^2 + az^2$, $u = (x, y, z) \in E^3$, при различном начальном приближении u_0, считая коэффициент a намного больше единицы.

5. Доказать теоремы 1, 2 для метода (3), (7).

§ 2. Метод проекции градиента

1. Будем рассматривать задачу

$$J(u) \to \inf; \quad u \in U \subseteq E^n, \tag{1}$$

где множество U необязательно совпадает со всем пространством E^n, а функция $J(u) \in C^1(U)$. Непосредственное применение описанного выше градиентного метода в случае $U \neq E^n$ может привести к затруднениям, так как точка u_{k+1} из (1.3) при каком-то k может не принадлежать U. Однако эту трудность можно преодолеть, если полученную с помощью формулы (1.3) точку $u_k - \alpha_k J'(u_k)$ при каждом k проектировать на множество U (см. определение 4.4.1). В результате мы придем к так называемому методу проекции градиента.

А именно, пусть $u_0 \in U$ — некоторое начальное приближение. Далее будем строить последовательность $\{u_k\}$ по правилу

$$u_{k+1} = \mathcal{P}_U(u_k - \alpha_k J'(u_k)), \qquad k = 0, 1, \ldots, \tag{2}$$

где α_k — положительная величина. Если U — выпуклое замкнутое множество и способ выбора $\{\alpha_k\}$ в (2) задан, то в силу теоремы 4.4.1 последовательность $\{u_k\}$ будет однозначно определяться условием (2). В частности, при $U = E^n$ метод (2) превратится в градиентный метод.

Если в (2) на некоторой итерации оказалось $u_{k+1} = u_k$ (например, это случится при $J(u_k) = 0$), то процесс (2) прекращают. В этом случае точка u_k удовлетворяет необходимому условию оптимальности $u_k = \mathcal{P}_U(u_k - \alpha_k J'(u_k))$ (см. теорему (4.4.3), и для выяснения того, является ли в действительности u_k решением задачи (1) или нет, при необходимости нужно провести дополнительное исследование поведения функции $J(u)$ в окрестности точки u_k. В частности, если $J(u)$ — выпуклая функция, то такая точка u_k является решением задачи (1).

В зависимости от способа выбора α_k в (2) можно получить различные варианты метода проекции градиента. Укажем несколько наиболее употребительных на практике способов выбора α_k.

1) Введем функцию одной переменной $f_k(\alpha) = J(\mathcal{P}_U(u_k - \alpha J'(u_k)))$ $(\alpha \geqslant 0)$ и определим α_k из условий

$$f_k(\alpha_k) = \inf_{\alpha \geqslant 0} f_k(\alpha) = f_{k_*}, \quad \alpha_k > 0. \tag{3}$$

Очевидно, при $U=E^n$ метод (2), (3) превратится в метод скорейшего спуска. Поскольку величину α_k из условий (3) удается найти точно лишь в редких случаях (возможно также, что нижняя грань в (3) не всегда достигается), то α_k на практике определяют приближенно из условий типа (1.6) или (1.7).

2) Иногда приходится довольствоваться нахождением какого-либо $\alpha_k>0$, обеспечивающего условие монотонности: $J(u_{k+1})<J(u_k)$. Для этого обычно выбирают какую-либо постоянную $\alpha>0$ и в методе (2) на каждой итерации берут $\alpha_k=\alpha$, а затем проверяют условие монотонности и при необходимости дробят величину $\alpha_k=\alpha$, добиваясь выполнения условия монотонности.

3) Если функция $J(u)$ принадлежит $C^{1,1}(U)$ и константа Липшица L для градиента $J'(u)$ известна, то в (2) в качестве α_k можно взять любое число, удовлетворяющее условиям

$$0<\varepsilon_0\leqslant\alpha_k\leqslant 2/(L+2\varepsilon), \tag{4}$$

где ε_0, ε — положительные числа, являющиеся параметрами метода.

4) Возможен выбор α_k из условия

$$J(u_k)-J(\mathscr{P}_U(u_k-\alpha_kJ'(u_k)))\geqslant\varepsilon|u_k-\mathscr{P}_U(u_k-\alpha_kJ'(u_k))|^2, \tag{5}$$

где $\varepsilon>0$ — параметр метода. Для определения такого α_k можно взять какое-либо число $\alpha_k=\alpha$ (например, $\alpha=1$) и затем дробить его до тех пор, пока не выполнится условие (5). Если $J(u)\in C^{1,1}(U)$, то можно показать, что выполнения условия (5) можно добиться за конечное число дроблений.

5) Возможно априорное задание величин α_k из условий

$$\alpha_k>0,\quad k=0,1,\ldots;\qquad \sum_{k=0}^{\infty}\alpha_k=\infty,\quad \sum_{k=0}^{\infty}\alpha_k^2<\infty, \tag{6}$$

например, $\alpha_k=(k+1)^{-1}$ $(k=0,1,\ldots)$. Сходимость метода (2), (6) будет исследована в § 3.

Заметим, что описанные здесь варианты метода (2) при $U=E^n$ переходят в соответствующие варианты градиентного метода.

На практике для ускорения сходимости вместо (2) часто пользуются более общим вариантом метода проекции градиента

$$u_{k+1}=u_k+\beta_k(\mathscr{P}_U(u_k-\alpha_kJ'(u_k))-u_k)=$$
$$=\beta_k\mathscr{P}_U(u_k-\alpha_kJ'(u_k))+(1-\beta_k)u_k,\quad 0<\beta_k\leqslant 1,\quad \alpha_k>0, \tag{2'}$$

где параметры α_k, β_k могут выбираться различными способами.

Заметим, что в методах (2) или (2′) на каждой итерации, кроме выбора параметров α_k, β_k, нужно еще проектировать точку на множество U или, иначе говоря, решить задачу минимизации

$$\Phi_k(u)=|u-(u_k-\alpha_kJ'(u_k))|^2\to\inf,\quad u\in U; \tag{7}$$

здесь возможно использование функции $\Phi_k(u) = |u - u_k|^2 + 2\alpha_k\langle J'(u_k),\ u - u_k\rangle$, отличающейся от предыдущей функции постоянным слагаемым. Задачу (7) можно решать приближенно и вместо точки $u_{k+1} \in U$, $\Phi_k(u_{k+1}) = \inf\limits_U \Phi_k(u) = \Phi_{k*}$ определить ее приближение z_{k+1} из условий

$$z_{k+1} \in U:\ \Phi_k(z_{k+1}) \leqslant \Phi_{k*} + \delta_k^2. \tag{8}$$

Предполагая, что U — выпуклое замкнутое множество, из (8) с помощью неравенства (4.3.3) имеем $|z_{k+1} - u_{k+1}|^2 \leqslant \Phi_k(z_{k+1}) - \Phi_k(u_{k+1}) \leqslant \delta_k^2$ или

$$z_{k+1} \in U:\ |z_{k+1} - u_{k+1}| \leqslant \delta_k.$$

Конечно, задачи (7), (8) далеко не всегда просто решаются. Поэтому методом проекции градиента обычно пользуются лишь в тех случаях, когда проекция точки на множество легко определяется. Например, когда множество U представляет собой шар в E^n, параллелепипед, гиперплоскость, полупространство или положительный октант (см. примеры 4.4.1—4.4.6), задача проектирования точки решается просто и в явном виде, и реализация каждой итерации метода проекции градиента в этом случае не вызывает особых затруднений. Если же задача проектирования для своего решения в свою очередь требует применения тех или иных итерационных методов, то эффективность метода проекции градиента, вообще говоря, значительно снижается.

2. Остановимся на вопросах сходимости метода (2), (4).

Теорема 1. *Пусть U — выпуклое замкнутое множество из E^n, функция $J(u) \in C^{1,1}(U)$, $\inf\limits_U J(u) = J_* > -\infty$. Тогда для последовательности $\{u_k\}$, полученной методом* (2), (4) *при любом начальном приближении u_0, имеет место соотношение $\lim\limits_{k\to\infty} |u_{k+1} - u_k| = 0$. Если при этом множество $M(u_0) = \{u:\ u \in U,\ J(u) \leqslant J(u_0)\}$ ограничено, то $\lim\limits_{k\to\infty} \rho(u_k, S_*) = 0$, где $S_* = \{u:\ u \in M(u_0),\ \langle J'(u),\ v - u\rangle \geqslant 0$ при всех $v \in U\}$ — множество стационарных точек функции $J(u)$ на $M(u_0)$.*

Доказательство. Из неравенства (2.3.7) при $v = u_k$, $u = u_{k+1}$ имеем

$$J(u_k) - J(u_{k+1}) \geqslant \langle J'(u_k),\ u_k - u_{k+1}\rangle - (L/2)|u_k - u_{k+1}|^2, \quad k = 0, 1, \ldots \tag{9}$$

Из (2) и теоремы 4.4.1 следует, что

$$\langle u_{k+1} - [u_k - \alpha_k J'(u_k)],\ u - u_{k+1}\rangle \geqslant 0 \qquad \forall u \in U.$$

Перепишем это неравенство в виде

$$\langle J'(u_k),\ u - u_{k+1}\rangle \geqslant \langle u_k - u_{k+1},\ u - u_{k+1}\rangle/\alpha_k, \quad k = 0, 1, \ldots \tag{10}$$

Отсюда при $u = u_k$ с учетом условия (4) получим

$$\langle J'(u_k),\ u_k - u_{k+1}\rangle \geqslant |u_k - u_{k+1}|^2/\alpha_k \geqslant (L/2 + \varepsilon)\,|u_k - u_{k+1}|^2.$$

Подставим эту оценку в (9):

$$J(u_k) - J(u_{k+1}) \geqslant \varepsilon |u_k - u_{k+1}|^2, \qquad k = 0,\ 1,\ \ldots \tag{11}$$

Так как $J(u_k) \geqslant J_* > -\infty$ и последовательность $\{J(u_k)\}$ — убывающая, то существует конечный предел $\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$ и, следовательно, $\lim\limits_{k\to\infty} (J(u_k) - J(u_{k+1})) = 0$. Отсюда и из (11) сразу получим $\lim\limits_{k\to\infty} |u_k - u_{k+1}| = 0$.

Пусть теперь множество $M(u_0)$ ограничено. Так как согласно (11) $J(u_{k+1}) \leqslant J(u_k) \leqslant \ldots \leqslant J(u_0)$, то $\{u_k\} \in M(u_0)$. По теореме Больцано — Вейерштрасса ограниченная последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$ и $\{u_{k_m}\} \to u_*$. По доказанному $\lim\limits_{k\to\infty} |u_{k+1} - u_k| = 0$, поэтому $\{u_{k_m+1}\} \to u_*$. Переходя в (10) к пределу при $k = k_m \to \infty$, с учетом условий (4) и непрерывности $J'(u)$ получим $\langle J'(u_*),\ u - u_*\rangle \geqslant 0$ при любом $u \in U$, т. е. $u_* \in S_*$. По лемме 2.1.2 расстояние $\rho(u, S_*)$ непрерывно по u, поэтому $\lim\limits_{m\to\infty} \rho(u_{k_m}, S_*) = \rho(u_*, S_*) = 0$. Отсюда следует, что $\{\rho(u_k, S_*)\}$ имеет единственную предельную точку, равную нулю, т. е. $\lim\limits_{k\to\infty} \rho(u_k, S_*) = 0$. Теорема 1 доказана.

Теорема 2. *Пусть выполнены все условия теоремы* 1 *и, кроме того, функция* $J(u)$ *выпукла на* U. *Тогда для последовательности* $\{u_k\}$ *из* (2), (4) *имеем*

$$\lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} \rho(u_k, U_*) = 0, \tag{12}$$

причем справедлива оценка

$$0 \leqslant J(u_k) - J_* \leqslant C_0 k^{-1}, \quad C_0 = \mathrm{const} > 0, \qquad k = 1,\ 2,\ \ldots \tag{13}$$

Доказательство. Из ограниченности $M(u_0)$, непрерывности $J(u)$ согласно теореме 2.1.2 следует $J_* > -\infty$, $U_* = \{u\colon u \in U,\ J(u) = J_*\} \neq \varnothing$, $U_* \subset M(u_0)$. Возьмем произвольную точку $u_* \in U_*$. Из неравенства (4.2.4) тогда имеем

$$0 \leqslant a_k = J(u_k) - J(u_*) \leqslant \langle J'(u_k),\ u_k - u_*\rangle =$$
$$= \langle J'(u_k),\ u_k - u_{k+1}\rangle - \langle J'(u_k),\ u_* - u_{k+1}\rangle, \qquad k = 0,\ 1,\ \ldots$$

Пользуясь неравенством (10) при $u=u_*$ и условием (4) выбора α_k, отсюда получим

$$0\leqslant a_k\leqslant\langle J'(u_k),\ u_k-u_{k+1}\rangle-\langle u_k-u_{k+1},\ u_*-u_{k+1}\rangle/\alpha_k\leqslant$$
$$\leqslant|u_k-u_{k+1}|\Big(\sup_{M(u_0)}|J'(u)|+D/\varepsilon_0\Big)=C_1|u_k-u_{k+1}|,\quad k=0,1,\ldots \tag{14}$$

Здесь мы учли ограниченность множества $M(u_0)$, поэтому $D=\sup\limits_{u,v\in M(u_0)}|u-v|<\infty$ и, кроме того, $|J'(u)|\leqslant|J'(u)-J'(u_0)|+|J'(u_0)|\leqslant L|u-u_0|+|J'(u_0)|\leqslant LD+|J'(u_0)|$ при любом $u\in M(u_0)$, так что $\sup\limits_{M(u_0)}|J'(u)|<\infty$. Из (11), (14) следует $a_k-a_{k+1}\geqslant\varepsilon C_1^{-2}a_k^2=Aa_k^2$ $(k=0,1,\ldots)$. Отсюда с помощью леммы 2.3.4 придем к оценке (13), из которой также следует первое из равенств (12). Второе равенство (12) является следствием теоремы 2.1.2.

Рассмотрим случай сильно выпуклой функции, предполагая, что в методе (2) величина α_k выбирается постоянной.

Теорема 3. *Пусть U — выпуклое замкнутое множество, функция $J(u)\in C^{1,1}(U)$ и сильно выпукла на U. Пусть $0<\alpha<2\mu L^{-2}$, где постоянные μ, L, $\mu\leqslant L$, взяты из* (2.3.6), (4.3.8). *Тогда последовательность $\{u_k\}$, получаемая из* (2) *при $\alpha_k=\alpha$ $(k=0,1,\ldots)$, сходится к точке минимума u_*, причем справедлива оценка*

$$|u_k-u_*|\leqslant|u_0-u_*|(q(\alpha))^k,\qquad k=0,1,\ldots, \tag{15}$$

где $q(\alpha)=(1-2\mu\alpha+\alpha^2L^2)^{1/2}$, $0<q(\alpha)<1$.

Доказательство. Введем отображение

$$Au=\mathscr{P}_U(u-\alpha J'(u)),$$

действующее из U в U. Покажем его сжимаемость при $0<\alpha<2\mu L^{-2}$. С помощью теоремы 4.4.2 имеем

$$|Au-Av|^2=|\mathscr{P}_U(u-\alpha J'(u))-\mathscr{P}_U(v-\alpha J'(v))|^2\leqslant$$
$$\leqslant|u-\alpha J'(u)-v+\alpha J'(v)|^2=|u-v|^2+\alpha^2|J'(u)-J'(v)|^2-$$
$$-2\alpha\langle J'(u)-J'(v),\ u-v\rangle\leqslant|u-v|^2(1+\alpha^2L^2-2\mu\alpha)=$$
$$=q^2(\alpha)|u-v|^2,$$

т. е.

$$|Au-Av|\leqslant q(\alpha)|u-v|,\quad u,\ v\in U. \tag{16}$$

Так как $0<\alpha<2\mu L^{-2}$, то $0<q(\alpha)<1$. Это значит, что отображение A — сжимающее. Заметим также, что замкнутое множество $U\subseteq E^n$ представляет собой полное метрическое пространство с метрикой $\rho(u,v)=|u-v|$. Следовательно, можно пользоваться принципом сжимающих отображений [179]. Метод (2) при $\alpha_k=\alpha$, записанный в виде $u_{k+1}=Au_k$, представляет собой

известный процесс поиска неподвижной точки u_* сжимающего отображения A, т. е. точки u_*, для которой $u_* = Au_*$. Известно [179], что такая точка u_* существует, единственна и $\lim\limits_{k\to\infty} |u_k - u_*| = 0$. Из (16) следует, что

$$|u_k - u_m| \leqslant (q(\alpha))^k |u_0 - u_{m-k}| \qquad \forall m \geqslant k.$$

Отсюда при $m \to \infty$ получим оценку (15). Так как $u_* = \mathcal{P}_U(u_* - \alpha J'(u_*))$, то из теоремы 4.4.3 следует, что u_* — точка минимума функции $J(u)$ на множестве U. Теорема 3 доказана.

Заметим, что наименьшее значение $q(\alpha)$ из (15) при $0 < \alpha < 2\mu L^{-2}$ достигается при $\alpha_* = \mu L^{-2}$ и равно $q(\alpha_*) = (1 - (\mu/L)^2)^{1/2}$.

3. Следуя [29], рассмотрим сходимость метода (2), (4), не требуя, в отличие от теоремы 1, 2, ограниченности множества $M(u_0)$. Кроме того, будем считать, что вычисление градиента функции и проектирование на множество на каждой итерации проводятся с погрешностями.

Теорема 4. *Пусть U — выпуклое замкнутое множество из E^n, функция $J(u)$ выпукла на U, $J(u) \in C^{1,1}(U)$, $J_* > -\infty$, $U_* \neq \varnothing$. Пусть вместо точного значения градиента $J'(u)$ и проекции $\mathcal{P}_U(u) \equiv \mathcal{P}(u)$ известны их приближения $J'_k(u)$ и соответственно $\mathcal{P}_k(u)$ с погрешностью*

$$\begin{gathered} |J'(u) - J'_k(u)| \leqslant \delta_k, \quad u \in U; \quad |\mathcal{P}(u) - \mathcal{P}_k(u)| \leqslant C_0\delta_k, \\ u \in E^n, \quad C_0 = \text{const} > 0, \quad \delta_k \geqslant 0, \quad k = 0, 1, \ldots; \quad \sum_{k=0}^{\infty} \delta_k = \delta < 0. \end{gathered} \tag{17}$$

Наконец, пусть последовательность $\{u_k\}$ определяется условиями

$$u_{k+1} = \mathcal{P}_k\left(u_k - \alpha_k J'_k(u_k)\right), \quad u_0 \in U, \quad k = 0, 1, \ldots, \tag{18}$$

где α_k выбирается так:

$$0 < \varepsilon_0 \leqslant \alpha_k \leqslant 2(1-\varepsilon)/L, \quad k = 0, 1, \ldots, 0 < \varepsilon < 1. \tag{19}$$

Тогда $\{u_k\}$ сходится к некоторой точке $v_* \in U_*$.

Доказательство. Наряду с $\{u_k\}$ введем вспомогательную последовательность $\{v_k\}$, определяемую следующим образом:

$$v_k = \mathcal{P}(u_k - \alpha_k J'(u_k)), \quad k = 0, 1, \ldots; \quad v_0 = u_0. \tag{20}$$

Отсюда и из (18) с помощью теоремы 4.4.2 и условий (17) получаем

$$\begin{aligned} |u_{k+1} - v_k| &\leqslant \left|\mathcal{P}_k\left(u_k - \alpha_k J'_k(u_k)\right) - \mathcal{P}\left(u_k - \alpha_k J'_k(u_k)\right)\right| + \\ &+ \left|\mathcal{P}\left(u_k - \alpha_k J'_k(u_k)\right) - \mathcal{P}\left(u_k - \alpha_k J'(u_k)\right)\right| \leqslant C_0\delta_k + \alpha_k \left|J'_k(u_k) - J'(u_k)\right| \leqslant \\ &\leqslant C_0\delta_k + (2(1-\varepsilon)/L)\,\delta_k = C_1\delta_k, \quad k = 0, 1, \ldots \end{aligned} \tag{21}$$

Возьмем произвольную точку $u_* \in U_*$. Согласно теореме 4.2.3 тогда

$$\langle J'(u_*), u - u_* \rangle \geqslant 0, \quad u \in U. \tag{22}$$

Из (20) и неравенства (4.4.1) получаем

$$\langle v_k - u_k + \alpha_k J'(u_k), u - v_k \rangle \geqslant 0 \qquad \forall u \in U. \tag{23}$$

Положим в (23) $u = u_*$, а (22) умножим на $\alpha_k > 0$ и примем $u = v_k$. Сложим получившиеся неравенства

$$\langle v_k - u_k,\ u_* - v_k\rangle + \alpha_k \langle J'(u_k) - J'(u_*),\ u_* - v_k\rangle \geqslant 0, \qquad (24)$$
$$k = 0, 1, \ldots$$

Преобразуем каждое слагаемое в правой части (24). Прежде всего имеем

$$2\langle v_k - u_k,\ u_* - v_k\rangle = |u_k - u_*|^2 - |u_k - v_k|^2 - |v_k - u_*|^2. \qquad (25)$$

Далее, воспользуемся неравенством (4.2.20) при $u = u_k$, $v = u_*$, $w = v_k$; получим

$$\langle J'(u_k) - J'(u_*),\ u_* - v_k\rangle \leqslant (L/4)\,|u_k - v_k|^2, \qquad k = 0, 1, \ldots \qquad (26)$$

Подставив (25), (26) в (24), получим

$$|u_k - u_*|^2 - |v_k - u_*|^2 - (1 - \alpha_k L/2)\,|u_k - v_k|^2 \geqslant 0.$$

Отсюда, учитывая условие (19), имеем

$$|u_k - u_*|^2 \geqslant |v_k - u_*|^2 + \varepsilon\,|u_k - v_k|^2, \qquad k = 0, 1, \ldots \qquad (27)$$

Далее, воспользуемся леммой 2.3.10 при $z_k = u_k$, $z_* = u_*$, $w_k = v_k$; из (17), (21), (27) получим

$$\lim_{k\to\infty} |u_k - u_*| = \lim_{k\to\infty} |v_k - u_*| < \infty, \qquad \lim_{k\to\infty} |v_k - u_k| = 0. \qquad (28)$$

Отсюда следует, что последовательность $\{u_k\}$ ограничена. Тогда существует хотя бы одна предельная точка v_* этой последовательности и подпоследовательность $\{u_{k_m}\}$, сходящаяся к v_*. Из (28) имеем $\lim\limits_{m\to\infty} v_{k_m} = v_*$.

Согласно (23) с учетом (19) получаем

$$\langle J'(u_k),\ u - v_k\rangle \geqslant -\langle v_k - u_k,\ u - v_k\rangle\,\alpha_k^{-1} \geqslant -|v_k - u_k|\,|u - v_k|\,\varepsilon_0^{-1}.$$

Отсюда при $k = k_m \to \infty$ будем иметь $\langle J'(v_*),\ u - v_*\rangle \geqslant 0$ при всех $u \in U$. По теореме 4.2.3 тогда $v_* \in U_*$. Вспомним, что неравенство (27) было получено для любой точки $u_* \in U_*$. В частности, (27) верно и для $u_* = v_*$. Но v_* — предельная точка последовательности $\{u_k\}$. Из леммы 2.3.10 тогда следует, что $\{u_k\}$ сходится к v_*. Теорема 4 доказана.

Замечание 1. Если в (17) $\delta_k = 0$ $(k = 0, 1, \ldots)$, то из (18)—(20) следует, что $u_{k+1} = v_k$ $(k = 0, 1, \ldots)$. Тогда из (27) имеем

$$|u_k - u_*|^2 \geqslant |u_{k+1} - u_*|^2 + \varepsilon\,|u_k - u_{k+1}|^2, \qquad k = 0, 1, \ldots, \quad \forall u_* \in U_*.$$

Пользуясь произволом в выборе $u_* \in U_*$, отсюда получаем

$$|u_k - v_*| \geqslant |u_{k+1} - v_*|, \qquad \rho(u_k, U_*) \geqslant \rho(u_{k+1}, U_*), \quad k = 0, 1, \ldots,$$

причем равенство здесь возможно лишь при $u_{k+1} = u_k$, что в силу теоремы 4.4.3 означает $u_k \in U_*$. Таким образом, при точной реализации метода (17)—(19) расстояние от точки u_k до множества U_* или до точки v_* монотонно убывает. Как мы видели, таким же свойством обладает градиентный метод (1.3), (1.21), (1.22).

4. Опираясь на неравенства, полученные при доказательстве теоремы 4, можно оценить скорость сходимости метода (2), (4) для сильно выпуклых функций, причем, в отличие от теоремы 3, новая оценка оказывается неулучшаемой на классе сильно выпуклых функций, принадлежащих $C^{1,1}(U)$.

Теорема 5. *Пусть U — выпуклое замкнутое множество из E^n, а функция $J(u)$ сильно выпукла на U и принадлежит $C^{1,1}(U)$. Пусть последовательность $\{u_k\}$ построена методом* (2) *при* $\alpha_k = \alpha$ $(k = 0, 1, \ldots, 0 < \alpha < 2/L)$. *Тогда*

$$|u_k - u_*| \leqslant (q(\alpha))^k |u_0 - u_*|, \qquad k = 0, 1, \ldots, \tag{29}$$

где

$$q(\alpha) = \begin{cases} 1 - \mu\alpha, & 0 < \alpha < 2(L+\mu)^{-1}, \\ L\alpha - 1, & 2(L+\mu)^{-1} \leqslant \alpha < 2L^{-1}, \end{cases}$$

$0 < q(\alpha) < 1$, *постоянные* μ, L *взяты из* (2.3.6), (4.3.8), *а* u_* — *точка минимума* $J(u)$ *на* U. *Наименьшее значение* $q(\alpha)$ *при* $0 < \alpha < 2L^{-1}$ *достигается при* $\alpha = \alpha_* = 2(L+\mu)^{-1}$ *и равно* $q(\alpha_*) = (L-\mu)(L+\mu)^{-1}$.

Доказательство. Из теоремы 4.3.1 следует, что $J_* > -\infty$, U_* состоит из единственной точки u_*. Тогда из теоремы 4 имеем $\lim\limits_{k\to\infty} |u_k - u_*| = 0$. Здесь мы предполагаем, что в (17) $\delta_k = 0$ $(k = 0, 1, \ldots)$, поэтому из (18), (20), (21) следует $v_k = u_{k+1}$ $(k = 0, 1, \ldots)$. Учитывая последнее равенство и условие $\alpha_k = \alpha$, подставим (25) в (24). Получим

$$\begin{aligned} |u_{k+1} - u_*|^2 &\leqslant |u_k - u_*|^2 - |u_{k+1} - u_k|^2 + 2\alpha \langle J'(u_k) - J'(u_*), \\ &u_* - u_{k+1}\rangle = |u_k - u_*|^2 - |u_k - u_{k+1} - \alpha(J'(u_k) - J'(u_*))|^2 + \\ &+ \alpha^2 |J'(u_k) - J'(u_*)|^2 - 2\alpha \langle J'(u_k) - J'(u_*), u_k - u_*\rangle, \qquad k = 0, 1, \ldots \end{aligned} \tag{30}$$

Вспомним неравенства (4.3.17), (4.3.18), из которых имеем

$$|J'(u_k) - J'(u_*)|^2 + L\mu |u_k - u_*|^2 \leqslant$$
$$\leqslant (L+\mu) \langle J'(u_k) - J'(u_*), u_k - u_*\rangle, \qquad k = 0, 1, \ldots, \tag{31}$$

$$\mu |u_k - u_*| \leqslant |J'(u_k) - J'(u_*)| \leqslant L |u_k - u_*|, \qquad k = 0, 1, \ldots \tag{32}$$

Из (30), (31) следует

$$\begin{aligned} |u_{k+1} - u_*|^2 &\leqslant |u_k - u_*|^2 + \alpha^2 |J'(u_k) - J'(u_*)|^2 - \\ &- 2\alpha(L+\mu)^{-1} |J'(u_k) - J'(u_*)|^2 - 2\alpha L\mu (L+\mu)^{-1} |u_k - u_*|^2 = \\ &= \alpha[\alpha - 2(L+\mu)^{-1}] |J'(u_k) - J'(u_*)|^2 + \\ &+ [1 - 2\alpha L\mu (L+\mu)^{-1}] |u_k - u_*|^2, \qquad k = 0, 1, \ldots \end{aligned} \tag{33}$$

Рассмотрим два случая:

1) если $0 < \alpha < 2(L+\mu)^{-1} \leqslant \mu^{-1}$, то из (33) и левого неравенства (32) имеем $|u_{k+1} - u_*|^2 \leqslant |u_k - u_*|^2 [\alpha(\alpha - 2(L+\mu)^{-1})\mu^2 + 1 - 2\alpha L\mu (L+\mu)^{-1}] = (1 - \alpha\mu)^2 |u_k - u_*|^2$;

2) если $L^{-1} \leqslant 2(L+\mu)^{-1} \leqslant \alpha < 2L^{-1}$, то из (33) и правого неравенства (32) получим $|u_{k+1} - u_*|^2 \leqslant |u_k - u_*|^2 [\alpha(\alpha - 2(L+\mu)^{-1}) L^2 + 1 - 2\alpha L\mu (L+\mu)^{-1}] = (L\alpha - 1)^2 |u_k - u_*|^2$. Объединяя оба случая, имеем $|u_{k+1} - u_*| \leqslant q(\alpha) |u_k - u_*|$ $(k = 0, 1, \ldots)$, откуда следует оценка (29). Из графика функции $q(\alpha)$ (рис. 5.5) видно, что функция $q(\alpha)$ достигает минимума при $0 < \alpha < 2L^{-1}$ в точке $\alpha_* = 2(L+\mu)^{-1}$, причем $q(\alpha_*) = (L-\mu)(L+\mu)$. Теорема 5 доказана.

Приведем пример, показывающий, что оценка (29) неулучшаема на классе сильно выпуклых функций из $C^{1,1}(U)$.

Пример 1. Пусть $u = (x, y) \in U = E^2$, $J(u) = (Lx^2 + \mu y^2)/2$ $(0 < \mu \leqslant L)$. Ясно, что это функция сильно выпукла с константой $\varkappa = \mu/2$, принадлежит $C^{1,1}(E^2)$ с константой L и достигает минимума на E^2 в точке $u_* = (0, 0)$. Процесс (2) при $\alpha_k = \alpha$ $(0 < \alpha < 2L^{-1})$ имеет вид

$$x_{k+1} = x_k - \alpha L x_k = (1 - \alpha L) x_k,$$

$$y_{k+1} = y_k - \alpha\mu y_k = (1 - \alpha\mu) y_k,$$

$$k = 0, 1, \ldots$$

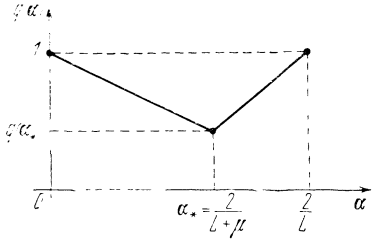

Рис. 5.5

Положим здесь $\alpha = 2(L + \mu)^{-1}$, $q = (L - \mu)(L + \mu)^{-1}$. Тогда $x_{k+1} = -qx_k$, $y_{k+1} = qy_k$. Следовательно, $|u_{k+1} - u_*| = q|u_k| = q^{k+1}|u_0|$, т. е. оценка (29) неулучшаема. Заметим, что если в теоремах 1—5 $U = E^n$, то мы получим сходимость соответствующих вариантов градиентного метода (1.3).

Упражнения. 1. Вычислить несколько итераций метода проекции градиента при различных способах выбора α_k в (2) для функции

$$J(u) = (x - 1)^2 + (y + 1)^2,$$

$$u \in U = E^2_+ = \{u = (x, y) \in E^2\colon\ x \geqslant 0,\quad y \geqslant 0\}.$$

Рассмотреть начальные приближения $u_0 = (0, 0)$, $u_0 = (0, 1)$, $u_0 = (1, 0)$, $u_0 = (1, 1)$.

2. Описать одну итерацию метода проекции градиента для функции (1.5), считая, что множество U представляет собой шар, гиперплоскость, параллелепипед, полупространство или положительный октант (см. примеры 4.4.1—4.4.6). Исследовать сходимость метода.

3. Рассмотреть метод проекции градиента для функции $J(u) = |Au - b|^2$, где A — матрица порядка $m \times n$, $b \in E^m$, считая, что множество U имеет вид, описанный в примерах 4.4.1—4.4.6. Исследовать сходимость.

§ 3. Метод проекции субградиента

1. В рассмотренных выше градиентном методе и методе проекции градиента требовалась дифференцируемость минимизируемой функции. Однако для выпуклых функций указанные методы более естественно описать на языке субградиентов (см. § 4.6). А именно, пусть U — выпуклое замкнутое множество из E^n, функция $J(u)$ выпукла на U и ее субдифференциал $\partial J(u)$ непуст при всех $u \in U$. Тогда для приближенного решения задачи

$$J(u) \to \inf; \quad u \in U, \tag{1}$$

можно предложить следующий итерационный метод:

$$u_{k+1} = \mathscr{P}_U(u_k - \alpha_k c_k), \quad \alpha_k > 0, \quad c_k \in \partial J(u_k), \quad k = 0, 1, \ldots, \tag{2}$$

где u_0 — некоторая точка из U, а субградиент c_k выбирается из $\partial J(u_k)$ произвольным образом. Если при некотором k окажется, что $u_{k+1} = u_k$, то процесс (2) прекращается, так как в этом случае u_k — решение задачи (1). В самом деле, при $u_k = \mathscr{P}_U(u_k - \alpha_k c_k)$ согласно теореме 4.4.1 $\langle u_k - (u_k - \alpha_k c_k), u - u_k\rangle = \langle c_k, u - u_k\rangle \alpha_k \geqslant 0$ или $\langle c_k, u - u_k\rangle \geqslant 0$ при всех $u \in U$. Отсюда из теоремы 4.6.4 следует, что $u_k \in U_*$.

В том случае, когда функция $J(u)$ дифференцируема во всех точках $u \in U$, метод (2) превращается в метод проекции градиента, а при $U = E^n$ — в градиентный метод. При выборе длины шага α_k в (2) можно руководствоваться теми же соображениями, которые были описаны выше в § 1, 2. Мы здесь ограничимся рассмотрением случая, когда α_k в (2) выбирается из условий

$$\alpha_k > 0, \quad k = 0, 1, \ldots; \qquad \sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \tag{3}$$

Как уже отмечалось в § 1, в качестве α_k можно взять $\alpha_k = C(k+1)^{-\alpha}$, где $C = \text{const} > 0$, $1/2 < \alpha \leqslant 1$; например, $\alpha_k = (k+1)^{-1}$ $(k = 0, 1, \ldots)$.

Метод (2), (3) не гарантирует выполнения условия монотонности $J(u_k) > J(u_{k+1})$ на каждой итерации и сходится, вообще говоря, медленно, но если проекция точки на множество U и субградиент $c_k \in \partial J(u_k)$ находятся несложно, то этот метод очень прост для реализации на ЭВМ. Докажем его сходимость.

Теорема 1. *Пусть U — выпуклое замкнутое множество из E^n, функция $J(u)$ определена и выпукла на некотором открытом выпуклом множестве W, содержащем U (например, $W = E^n$). Пусть $J_* > -\infty$, множество U_* точек минимума $J(u)$ на U непусто и ограничено, и пусть, кроме того,*

$$\sup_{u \in U} \sup_{c \in \partial J(u)} |c| = A < \infty. \tag{4}$$

Тогда последовательность $\{u_k\}$, определяемая условиями (2), (3), такова, что

$$\lim_{k \to \infty} J(u_k) = J_*, \qquad \lim_{k \to \infty} \rho(u_k, U_*) = 0. \tag{5}$$

Доказательство. При сделанных предположениях функция $J(u)$ непрерывна на U, субдифференциалы $\partial J(u)$ непусты, выпуклы, замкнуты и ограничены при всех $u \in U$ (см. теоремы 4.2.15, 4.6.1, 4.6.2). Из ограниченности множества U_* и теоремы 4.2.17 следует, что множество $M(C) = \{u \in U\colon J(u) \leqslant C\}$ ограничено при любом $C \geqslant J_*$. Множество U_* выпукло и замкнуто в силу теоремы 4.2.1 и леммы 2.1.1.

Согласно определению 4.4.1 проекции точки на множество и теореме 4.4.2 имеем

$$\begin{aligned}\rho^2(u_{k+1}, U_*) &= |u_{k+1} - \mathcal{P}_{U_*}(u_{k+1})|^2 \leqslant |u_{k+1} - \mathcal{P}_{U_*}(u_k)|^2 = \\ &= |\mathcal{P}_U(u_k - \alpha_k c_k) - \mathcal{P}_U(\mathcal{P}_{U_*}(u_k))|^2 \leqslant |u_k - \alpha_k c_k - \mathcal{P}_{U_*}(u_k)|^2 = \\ &= \rho^2(u_n, U_*) + \alpha_k^2 |c_k|^2 - 2\alpha_k \langle c_k, u_k - \mathcal{P}_{U_*}(u_k)\rangle\end{aligned}$$

или

$$2\alpha_k \langle c_k, u_k - \mathcal{P}_{U_*}(u_k)\rangle \leqslant \alpha_k^2 |c_k|^2 + \rho^2(u_k, U_*) - \rho^2(u_{k+1}, U_*), \qquad k = 0, 1, \ldots \tag{6}$$

Суммируя неравенства (6) по k от 0 до некоторого $m \geqslant 1$, с учетом условий (3), (4) получим

$$\begin{aligned}\sum_{k=0}^{m} 2\alpha_k \langle c_k, u_k - \mathcal{P}_{U_*}(u_k)\rangle &\leqslant A^2 \sum_{k=0}^{m} \alpha_k^2 + \rho^2(u_0, U_*) - \rho^2(u_{m+1}, U_*) \leqslant \\ &\leqslant A^2 \sum_{k=0}^{\infty} \alpha_k^2 + \rho^2(u_0, U_*) = B < \infty, \qquad m = 1, 2, \ldots\end{aligned} \tag{7}$$

Далее, по определению субградиента имеем

$$0 \leqslant J(u_k) - J_* = J(u_k) - J(\mathscr{P}_{U_*}(u_k)) \leqslant \langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle, \quad k = 0, 1, \ldots \tag{8}$$

Из (7), (8) следует, что числовой ряд

$$\sum_{k=0}^{\infty} \alpha_k \langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle$$

с неотрицательными членами сходится. Но согласно (3) $\sum\limits_{k=0}^{\infty} \alpha_k = \infty$. Поэтому сходимость предыдущего ряда возможна лишь при $\varliminf\limits_{k\to\infty} \langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle = 0$. Это значит, что существуют номера $k_1 < k_2 < \ldots < k_m < \ldots$ такие, что

$$\lim_{m\to\infty} \langle c_{k_m}, u_{k_m} - \mathscr{P}_{U_*}(u_{k_m})\rangle = 0. \tag{9}$$

Тогда из (8) при $k = k_m \to \infty$ получим $\lim\limits_{m\to\infty} J(u_{k_m}) = J_*$. Кроме того, из (8), (9) следует, что $J(u_{k_m}) \leqslant J_* + \sup\limits_m \langle c_{k_m}, u_{k_m} - \mathscr{P}_{U_*}(u_{k_m})\rangle = C < \infty$, т. е. $\{u_{k_m}\} \in M(C)$. Но $M(C)$ ограничено, а $\{u_{k_m}\}$ — минимизирующая последовательность, поэтому из теоремы 2.1.2 имеем $\lim\limits_{m\to\infty} \rho(u_{k_m}, U_*) = 0$.

Тем самым показано, что для подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (9), справедливы равенства (5). Опираясь на это, покажем, что равенства (5) имеют место для всей последовательности $\{u_k\}$. Из (3), (4), (6), (8) получаем, что

$$\rho^2(u_{k+1}, U_*) \leqslant \rho^2(u_k, U_*) + \alpha_k^2 |c_k|^2 \leqslant \rho^2(u_k, U_*) + \alpha_k^2 A^2, \qquad k = 0, 1, \ldots$$

Это значит, что числовая последовательность $a_k = \rho^2(u_k, U_*)$ $(k = 0, 1, \ldots)$ удовлетворяет условиям леммы 2.3.2, из которой следует существование предела $\lim\limits_{k\to\infty} \rho^2(u_k, U_*)$. Так как подпоследовательность $\left\{\rho^2(u_{k_m}, U_*)\right\}$ сходится к нулю, то этот предел может равняться лишь нулю. Следовательно, $\lim\limits_{k\to\infty} \rho(u_k, U_*) = 0$. Покажем, что тогда $\lim\limits_{k\to\infty} J(u_k) = J_*$. По условию множество U_* ограничено. Тогда последовательность $\{u_k\}$, сходящаяся к U_*, также ограничена. Возьмем любую предельную точку u_* этой последовательности. Пусть $\{u_{k_r}\} \to u_*$. Так как U_* — замкнутое множество и $\lim\limits_{r\to\infty} \rho(u_{k_r}, U_*) = \rho(u_*, U_*) = 0$, то $u_* \in U_*$. А тогда $\lim\limits_{r\to\infty} J(u_{k_r}) = J(u_*) = J_*$. Это означает, что числовая последовательность $\{J(u_k)\}$ имеет единственную предельную точку, равную J_*, т. е. $\lim\limits_{k\to\infty} J(u_k) = J_*$. Теорема 1 доказана.

Замечание 1. В условии теоремы 1 предполагается выполнение условия (4). В том случае, когда U ограничено, то, как следует из теоремы 4.6.5, условие (4) всегда выполняется. Заметим также, что в теореме 1 вместо (4) можно потребовать сходимость ряда $\sum\limits_{k=0}^{\infty} \alpha_k^2 |c_k|^2$.

2. Описанный выше метод проекции субградиента после некоторой модификации можно использовать для решения следующей задачи

выпуклого программирования:

$$J(u) \to \inf; \quad u \in U = \{u \in E^n\colon u \in U_0,\ g_i(u) \leqslant 0,\ \ i = 1, \ldots, m\}. \tag{10}$$

Заметим, что система неравенств $g_i(u) \leqslant 0$ $(i = 1, \ldots, m)$ равносильна одному неравенству $g(u) \leqslant 0$, где $g(u) = \max\limits_{1 \leqslant i \leqslant m} g_i(u)$ $(u \in U)$. Кроме того, из выпуклости функций $g_i(u)$ $(i = 1, \ldots, m)$ на U_0 следует выпуклость $g(u)$ на U_0 (см. теорему 4.2.7). Поэтому задачу (10) можно переформулировать в виде эквивалентной задачи

$$J(u) \to \inf; \quad u \in U = \{u \in U_0,\ g(u) \leqslant 0\}, \tag{11}$$

также являющейся задачей выпуклого программирования.

Предположим, что субдифференциалы $\partial J(u)$, $\partial g(u)$ непусты при всех $u \in U_0$. Следуя [334], рассмотрим метод

$$u_{k+1} = \mathscr{P}_{U_0}(u_k - \alpha_k c_k), \qquad k = 0, 1, \ldots; \qquad u_0 \in U_0, \tag{12}$$

где $\{\alpha_k\}$ выбирается из условий

$$\alpha_k > 0, \quad k = 0, 1, \ldots, \quad \sum_{k=0}^{\infty} \alpha_k^{1+\gamma} = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \quad 0 < \gamma < 1, \tag{13}$$

а субградиенты $\{c_k\}$ таковы, что

$$c_k \in \partial J(u_k) \ \text{ при } \ g(u_k) \leqslant \alpha_k^{\gamma} \ \text{ и } \ c_k \in \partial g(u_k) \ \text{ при } \ g(u_k) > \alpha_k^{\gamma}. \tag{14}$$

Таким образом, метод (12)—(14) работает так: если ограничение $g(u) \leqslant 0$ при $u = u_k$ не нарушено или нарушено немного, то минимизируем функцию $J(u)$, а если нарушение этого ограничения велико, то минимизируем функцию $g(u)$. Если функции $J(u)$, $g(u)$ дифференцируемы на U_0, то в (12), (14) вместо c_k нужно брать соответствующие градиенты $J'(u_k)$ или $g'(u_k)$. В качестве последовательности $\{\alpha_k\}$, удовлетворяющей условиям (13), можно взять, например, $\alpha_k = C(k+1)^{-\alpha}$, где $C = \text{const} > 0$, а число α таково, что $1/2 < \alpha < (1+\gamma)^{-1}$. В частности, при $\alpha = 3/5$, $\gamma = 1/2$, $C = 1$ получим $\alpha_k = (k+1)^{-3/5}$ $(k = 0, 1, \ldots)$.

Теорема 2. *Пусть U_0 — выпуклое замкнутое множество из E^n, функции $J(u)$, $g(u)$ определены и выпуклы на некотором открытом выпуклом множестве W, содержащем U_0* (например, $W = E^n$). *Пусть $J_* = \inf\limits_U J(u) > -\infty$, множество U_* точек минимума задачи* (11) *непусто, ограничено и, кроме того,*

$$\sup_{u \in U_0} \sup_{c \in \partial J(u) \cup \partial g(u)} |c| = A < \infty.$$

Тогда для последовательности $\{u_k\}$, определяемой условиями (12)—(14), *справедливы равенства* (5).

Доказательство. При выполнении условий теоремы функции $J(u)$, $g(u)$ непрерывны на U_0, субдифференциалы $\partial J(u)$, $\partial g(u)$ непусты, выпуклы, замкнуты и ограничены при всех $u \in U_0$ (см. теоремы 4.2.15, 4.6.1, 4.6.2), а множество U_* выпукло и замкнуто (см. теорему 4.2.1 и лемму 2.1.1). Покажем, что множество

$$M(C_1, C_2) = \{u \in U_0\colon J(u) \leqslant C_1,\ g(u) \leqslant C_2\}$$

ограничено при всех $C_1 > \inf\limits_{U_0} J(u)$, $C_2 > \inf\limits_{U_0} g(u)$. В самом деле, $M(J_*, 0) = U_*$ ограничено по условию. Тогда по теореме 4.2.17 множество $M(C_1, 0) = \{u\colon\ u \in U_0,\ g(u) \leqslant 0,\ J(u) \leqslant C_1\}$ ограничено при всех $C_1 > \inf\limits_{U_0} J(u)$.

Теперь, фиксируя любое C_1, по той же теореме 4.2.17 получаем ограниченность $M(C_1, C_2)$ при каждом $C_2 > \inf\limits_{U_0} g(u)$.

Нетрудно видеть, что неравенства (6), (7) сохраняют силу и для метода (12)—(14). Из (7) имеем

$$\sum_{k=0}^{r} \alpha_k^{1+\gamma} \alpha_k^{-\gamma} \langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle \leqslant B < \infty, \qquad r = 1, 2, \ldots \tag{15}$$

Отсюда следует существование номеров $k_1 < k_2 < \ldots < k_m < \ldots$ таких, что

$$\langle c_{k_m}, u_{k_m} - \mathscr{P}_{U_*}(u_{k_m})\rangle \leqslant \alpha\gamma_{k_m}, \qquad m = 1, 2, \ldots \tag{16}$$

В самом деле, допустим, что (16) не имеет места. Тогда $\langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle > \alpha_k^{\gamma}$ при всех $k = 0, 1, \ldots$ Отсюда и из (15) имеем

$$\sum_{k=0}^{r} \alpha_k^{1+\gamma} \leqslant \sum_{k=0}^{r} \alpha_k \langle c_k, u_k - \mathscr{P}_{U_*}(u_k)\rangle \leqslant B < \infty, \qquad r = 1, 2, \ldots,$$

что противоречит расходимости ряда $\sum\limits_{k=0}^{\infty} \alpha_k^{1+\gamma}$.

Тем самым показано существование подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (16). Докажем, что

$$\lim_{m\to\infty} J(u_{k_m}) = J_*, \quad \lim_{m\to\infty} \rho(u_{k_m}, U_*) = 0. \tag{17}$$

Сначала убедимся в том, что

$$c_{k_m} \in \partial J(u_{k_m}), \qquad m = 1, 2, \ldots \tag{18}$$

Для этого достаточно показать, что $g(u_{k_m}) \leqslant \alpha_{k_m}^{\gamma}$ $(m = 1, 2, \ldots)$, и вспомнить условия (14). Допустим, что $g(u_{k_m}) > \alpha_{k_m}^{\gamma}$ при некотором $m \geqslant 1$. Учитывая, что тогда $c_{k_m} \in \partial g(u_{k_m})$ и, кроме того, $\mathscr{P}_{U_*}(u_{k_m}) \in U_* \subset U$, т. е. $g(\mathscr{P}_{U_*}(u_{k_m})) \leqslant 0$, из (16) имеем

$$\alpha_{k_m}^{\gamma} < g(u_{k_m}) \leqslant g(u_{k_m}) - g(\mathscr{P}_{U_*}(u_{k_m})) \leqslant$$
$$\leqslant \langle c_{k_m}, u_{k_m} - \mathscr{P}_{U_*}(u_{k_m})\rangle \leqslant \alpha_{k_m}^{\gamma}.$$

Получили противоречивое неравенство. Включения (18) доказаны, и попутно установлено, что

$$g(u_{k_m}) \leqslant \alpha_{k_m}^{\gamma}, \qquad m = 1, 2, \ldots \tag{19}$$

Множество номеров $\{k_m\}$, удовлетворяющих условиям (16), представим в виде объединения непересекающихся множеств $I_1 = \{k_m\colon J(u_{k_m}) \geqslant J_*\}$ и $I_2 = \{k_m\colon J(u_{k_m}) < J_*\}$.

Сначала рассмотрим случай, когда множество I_1 бесконечно. Из (16), (18) имеем

$$0 \leqslant J\left(u_{k_m}\right) - J_* = J\left(u_{k_m}\right) - J\left(\mathcal{P}_{U_*}\left(u_{k_m}\right)\right) \leqslant$$
$$\leqslant \left\langle c_{k_m}, u_{k_m} - \mathcal{P}_{U_*}\left(u_{k_m}\right)\right\rangle \leqslant \alpha_{k_m}^{\gamma}, \quad k_m \in I_1.$$

Отсюда следует, что $J\left(u_{k_m}\right) \to J_*$ при $m \to \infty$, $k_m \in I_1$. Тогда $J\left(u_{k_m}\right) \leqslant C_1 < \infty$, и, кроме того, согласно (13), (19) $g\left(u_{k_m}\right) \leqslant \alpha_{k_m}^{\gamma} \leqslant \sup\limits_{k \geqslant 0} \alpha_k^{\gamma} = C_2 < \infty$, т. е. $\{u_{k_m}\} \in M(C_1, C_2)$ $(k_m \in I_1)$. Так как $M(C_1, C_2)$ ограничено, то $\{u_{k_m}, k_m \in I_1\}$ имеет хотя бы одну предельную точку. Не умаляя общности, можем считать, что $\{u_{k_m}\} \to u_*$ $(k_m \in I_1)$. Из замкнутости U_0, неравенства (19) и непрерывности $g(u)$ следует, что $u_* \in U$. Но $J\left(u_{k_m}\right) \to J(u_*) = J_*$ при $k_m \to \infty, k_m \in I_1$, так что $u_* \in U_*$. Отсюда и из непрерывности $\rho(u, U_*)$ имеем $\rho\left(u_{k_m}, U_*\right) \to \rho(u_*, U_*) = 0$ при $k_m \to \infty$, $k_m \in I_1$.

Теперь рассмотрим случай, когда множество I_2 бесконечно. Если $k_m \in I_2$, то $J\left(u_{k_m}\right) < J_*$, а $g\left(u_{k_m}\right) \leqslant \alpha_{k_m}^{\gamma} < \sup\limits_{k \geqslant 0} \alpha_k^{\gamma} = C_2$ в силу (19). Это значит, что $\{u_{k_m}\} \in M(J_*, C_2)$ $(k_m \in I_2)$. Поскольку множество $M(J_*, C_2)$ ограничено, то $\{u_{k_m}, k_m \in I_2\}$ имеет предельную точку. Не умаляя общности, можем считать, что $\{u_{k_m}\} \to u_*$, $k_m \in I_2$. Из (19) и замкнутости U_0 следует, что $u_* \in U$. Поэтому $J(u_*) \geqslant J_*$. С другой стороны, $J\left(u_{k_m}\right) < J_*$ $(k_m \in I_2)$, так что $\lim\limits_{k_m \in I_2, m \to \infty} J\left(u_{k_m}\right) = J(u_*) \leqslant J_*$. Следовательно, $J(u_*) = J_*$, т. е. $u_* \in U_*$. Отсюда и из непрерывности $\rho(u, U_*)$ получаем $\rho\left(u_{k_m}, U_*\right) \to \rho(u_*, U_*) = 0$ при $k_m \to \infty$, $k_m \in I_2$.

Объединяя оба рассмотренных случая, заключаем, что для подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (16), справедливы равенства (17). Отсюда, повторив заключительные рассуждения из доказательства теоремы 1, убеждаемся в справедливости равенств (5) и для метода (12)—(14). Замечание 1 сохраняет силу и здесь.

На этом закончим рассмотрение методов минимизации негладких выпуклых функций. Отметим, что негладкие задачи в последние годы интенсивно исследуются, продолжается разработка различных методов их решения [27, 113, 123, 127, 132, 133, 148, 156, 158, 214, 219, 235, 306, 334, 336, 339].

Упражнения. 1. Рассмотреть возможность применения метода проекции субградиента к задачам из упражнений 4.6.1 и 4.6.3.

2. Описать метод (12)—(14) применительно к задаче

$$J(u) = |x + y| + |x - y| \to \inf, \quad u \in U = \{u = (x, y) \in E^2\colon u \in U_0,$$
$$g(u) = u^2 - 1 \leqslant 0\}, \quad U_0 = \{u = (x, y)\colon x \geqslant 0, y \geqslant 0\}.$$

3. Проверить условия теоремы 2 для задачи

$$J(u) = |\langle c, u\rangle| \to \inf; \quad u \in U = \{u \in E^n\colon u \geqslant 0, g(u) = |\langle a, u\rangle| - 1 \leqslant 0\},$$

где $a, c \in E^n$. Описать метод (12)—(14) применительно к этой задаче.

4. Пользуясь формулой (4.6.12), модифицировать метод (12)—(14) так, чтобы его можно было применять к задаче (10) непосредственно, не сводя ее к задаче (11).

5. Пусть W — открытое выпуклое множество, $J(u)$ — выпуклая функция на W. Показать, что вектор c_*, удовлетворяющий условиям

$$|c_*| = \inf_{c \in \partial J(u)} |c| > 0, \quad c_* \in \partial J(u),$$

является направлением убывания функции $J(u)$ в точке u.

§ 4. Метод условного градиента

1. Этот метод приспособлен для решения задачи

$$J(u) \to \inf; \quad u \in U, \tag{1}$$

где U — выпуклое замкнутое ограниченное множество из E^n, функция $J(u) \in C^1(U)$. Опишем его. Пусть $u_0 \in U$ — некоторое начальное приближение. Если известно k-е приближение $u_k \in U$ $(k \geqslant 0)$, то приращение функции $J(u)$ в точке u_k можем представить в виде

$$J(u) - J(u_k) = \langle J'(u_k),\ u - u_k \rangle + o(|u - u_k|).$$

Возьмем главную линейную часть этого приращения

$$J_k(u) = \langle J'(u_k),\ u - u_k \rangle, \tag{2}$$

и определим вспомогательное приближение $\bar{u}_k$ из условий

$$\bar{u}_k \in U, \quad \inf_U J_k(u) = J_k(\bar{u}_k) = \langle J'(u_k),\ \bar{u}_k - u_k \rangle. \tag{3}$$

Так как множество U замкнуто и ограничено, а линейная функция $J_k(u)$ непрерывна, то точка $\bar{u}_k$ из (3) всегда существует. Если функция $J_k(u)$ достигает своей нижней грани на U более чем в одной точке, то в качестве точки $\bar{u}_k$ возьмем любую из них.

Заметим, что если

$$U = \{u \in E^n:\ u \geqslant 0, \quad \langle a_i, u \rangle \leqslant b^i, \\ i = 1, \ldots, m; \quad \langle a_i, u \rangle = b^i, \quad i = m+1, \ldots, s\},$$

то задача (3) превратится в задачу линейного программирования, которая может быть решена известными методами (например, описанным в гл. 3 симплекс-методом).

Укажем случаи, когда решение задачи (3) выписывается в явном виде. Если

$$U = \{u = (u^1, \ldots, u^n):\ \alpha_i \leqslant u^i \leqslant \beta_i, \quad i = 1, \ldots, n\}$$

— n-мерный параллелепипед, то функция $J_k(u) = \sum_{i=1}^{n} J_{u^i}(u_k)(u^i - u_k^i)$ или $\sum_{i=1}^{n} J_{u^i}(u_k)\, u^i$, очевидно, достигает своей нижней

грани на U в точке $\bar{u}_k = (\bar{u}_k^1, \ldots, \bar{u}_k^n)$, где

$$\bar{u}_k^i = \begin{cases} \alpha_i, & J_{u^i}(u_k) > 0, \\ \beta_i, & J_{u^i}(u_k) < 0; \end{cases}$$

в случае $J_{u^i}(u_k) = 0$ здесь возникает неопределенность и в качестве $\bar{u}_k^i$ можно взять любое число из отрезка $[\alpha_i, \beta_i]$ (обычно берут $\bar{u}_k^i = \alpha_i$, или $\bar{u}_k^i = \beta_i$, или $\bar{u}_k^i = (\alpha_i + \beta_i)/2$).

Если

$$U = \{u \in E^n\colon\ |u - v_0| \leqslant r\}$$

— шар радиуса r с центром в точке v_0, то с помощью неравенства Коши — Буняковского $\langle J'(u_k),\ u\rangle = \langle J'(u_k),\ u - v_0\rangle + \langle J'(u_k),\ v_0\rangle \geqslant -|J'(u_k)|r + \langle J'(u_k),\ v_0\rangle$, $u \in U$, получаем, что

$$\bar{u}_k = v_0 - rJ'(u_k)\,|J'(u_k)|^{-1}.$$

Разумеется, так просто получить вспомогательное приближение $\bar{u}_k$ удается далеко не всегда, и вместо точного решения задачи (3) часто приходится довольствоваться определением какого-либо приближенного решения. А именно, будем предполагать, что оно определяется из следующих условий:

$$\bar{u}_k \in U, \quad J_k(\bar{u}_k) \leqslant \min_U J_k(u) + \varepsilon_k, \quad \varepsilon_k \geqslant 0, \quad \lim_{k\to\infty} \varepsilon_k = 0. \tag{4}$$

Допустим, что точка $\bar{u}_k$, удовлетворяющая условиям (4) (или (3)), уже найдена. Тогда следующее $(k+1)$-е приближение будем искать в виде

$$u_{k+1} = u_k + \alpha_k(\bar{u}_k - u_k), \quad 0 \leqslant \alpha_k \leqslant 1. \tag{5}$$

В силу выпуклости множества U всегда $u_{k+1} \in U$.

Заметим, что при $\bar{u}_k = u_k$ (это может случиться, например, когда $J'(u_k) = 0$) имеем $u_{k+1} = u_k$ независимо от способа выбора α_k в (5). Если при этом $\bar{u}_k$ было определено точно из условий (3), то имеем

$$J_k(\bar{u}_k) = J_k(u_k) = 0 = \min_U J_k(u) \text{ или } \langle J'(u_k),\ u - u_k\rangle \geqslant 0$$

при всех $u \in U$. Согласно теореме 4.2.3 это означает, что точка u_k удовлетворяет необходимому условию минимума в задаче (1). В этом случае итерации прекращаются, и для выяснения того, будет ли $u_k \in U_*$, при необходимости проводится дополнительное исследование поведения функции $J(u)$ в окрестности точки u_k. В частности, если $J(u)$ выпукла, то согласно теореме 4.2.3 $u_k \in U_*$, т. е. задача (1) решена. Если случай $\bar{u}_k = u_k$ реализовался при определении $\bar{u}_k$ из условия (4), то будем иметь $-\varepsilon_k \leqslant \min_U J_k(u) \leqslant J_k(\bar{u}_k) = J_k(u_k) = 0$, и при $\varepsilon_k > 0$ здесь тео-

рему 4.2.3 применять нельзя. В этом случае согласно (5) полагаем $u_{k+1}=u_k$ и переходим к проверке условия (4) для номера $k+1$ и т. д.

В зависимости от способа выбора величины α_k в (5) можно получить различные варианты описанного метода, часто именуемого в литературе методом условного градиента. Укажем некоторые наиболее употребительные способы выбора α_k в (5).

1) Величина α_k может выбираться из условий

$$0\leqslant\alpha_k\leqslant 1,\quad f_k(\alpha_k)=\min_{0\leqslant\alpha\leqslant 1} f_k(\alpha)=f_{k*},\quad f_k(\alpha)=J\left(u_k+\alpha(\bar{u}_k-u_k)\right). \tag{6}$$

Для некоторых классов задач удается получить из (6) явное выражение для α_k.

Пример 1. Пусть

$$J(u)=\frac{1}{2}\langle Au,u\rangle-\langle b,u\rangle,$$

где A — симметричная положительно определенная матрица порядка $n\times n$, $b\in E^n$. Тогда $J'(u_k)=Au_k-b$. Пользуясь формулой (4.2.10), имеем

$$f_k(\alpha)=J(u_k)+\alpha\langle J'(u_k),\bar{u}_k-u_k\rangle+ \\ +(\alpha^2/2)\langle A(\bar{u}_k-u_k),\bar{u}_k-u_k\rangle. \tag{7}$$

Если $\langle A(\bar{u}_k-u_k),\bar{u}_k-u_k\rangle=0$, то $u_k=\bar{u}_k$ и, как было указано выше, тогда полагаем $u_{k+1}=u_k$. Поэтому пусть $\langle A(\bar{u}_k-u_k),\bar{u}_k-u_k\rangle>0$. Тогда функция (7) представляет собой квадратный трехчлен, достигающий своего наименьшего значения на числовой оси $-\infty<\alpha<+\infty$ при

$$\alpha_k^*=-\langle J'(u_k),\bar{u}_k-u_k\rangle\left(\langle A(\bar{u}_k-u_k),\bar{u}_k-u_k\rangle\right)^{-1}.$$

Рассматривая возможные случаи $\alpha_k^*<0$, $0\leqslant\alpha_k^*\leqslant 1$, $\alpha_k^*>1$, из условий (6) тогда получаем

$$\alpha_k=\begin{cases}0, & \alpha_k^*<0,\\ \alpha_k^*, & 0\leqslant\alpha_k^*\leqslant 1,\\ 1, & \alpha_k^*>1.\end{cases} \tag{8}$$

Кстати, если точка $\bar{u}_k$ в (7) найдена из условий (3), то $J_k(\bar{u}_k)\leqslant J_k(u_k)=0$ и, следовательно, $\alpha_k^*\geqslant 0$ — в этом случае формула (8) для α_k запишется в виде $\alpha_k=\min\{1;\alpha_k^*\}$.

Однако точное определение α_k из условия (6) возможно далеко не всегда. Поэтому вместо (6) можно ограничиться определением величины α_k из условий

$$0\leqslant\alpha_k\leqslant 1;\quad f_k(\alpha_k)\leqslant f_{k*}+\delta_k,\quad \delta_k\geqslant 0,\quad \sum_{k=0}^{\infty}\delta_k=\delta<\infty \tag{9}$$

или

$$0 \leqslant \alpha_k \leqslant 1, \quad f_k(\alpha_k) \leqslant (1-\lambda_k) f_k(0) + \lambda_k f_{k*}, \quad 0 < \bar{\lambda} \leqslant \lambda_k \leqslant 1.$$

Здесь могут быть использованы известные методы минимизации функций одной переменной (например, методы из гл. 1).

2) Если $J(u) \in C^{1,1}(U)$ и константа Липшица L для $J'(u)$ известна, то возможен выбор α_k в (5) из условий

$$\alpha_k = \begin{cases} \min\{1;\ \rho_k |J_k(\bar{u}_k)| |\bar{u}_k - u_k|^{-2}\}, & J_k(\bar{u}_k) \leqslant 0, \\ 0, & J_k(\bar{u}_k) > 0, \end{cases} \tag{10}$$

где

$$0 < \varepsilon_0 \leqslant \rho_k \leqslant 2(1-\varepsilon)/L, \tag{11}$$

ε_0, ε — параметры метода, $0 < \varepsilon < 1$.

3) Другой способ выбора α_k: при $J_k(\bar{u}_k) > 0$ полагают $\alpha_k = 0$, а если $J_k(\bar{u}_k) \leqslant 0$, то $\alpha_k = \lambda^{i_0}$, где i_0 — минимальный номер среди номеров $i \geqslant 0$, удовлетворяющих условию

$$J(u_k) - J(u_k + \lambda^i(\bar{u}_k - u_k)) \geqslant \lambda^i \varepsilon |J_k(\bar{u}_k)|,$$

где λ, ε — параметры метода, $0 < \lambda$; $\varepsilon < 1$.

4) Величины α_k в (5) можно априорно задавать из условий

$$0 < \alpha_k \leqslant 1, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \tag{12}$$

например, $\alpha_k = (k+1)^{-1}$ $(k = 0, 1, \ldots)$ (предложен М. Ячимовичем). Такой выбор α_k очень прост для реализации на ЭВМ, но, вообще говоря, не гарантирует выполнение условия монотонности $J(u_{k+1}) < J(u_k)$.

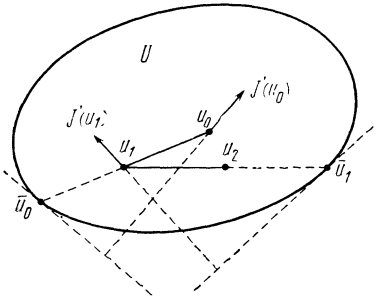

Рис. 5.6

5) Возможны и другие способы выбора α_k в (5). Например, можно задавать $\alpha_k = 1$ и проверять условие монотонности $J(u_{k+1}) < J(u_k)$, а затем при необходимости дробить α_k до тех пор, пока не выполнится условие монотонности.

На рис. 5.6 поясняется геометрический смысл метода (3), (5) в двумерном случае.

2. Рассмотрим теперь сходимость метода (4), (5), (9).

Теорема 1. *Пусть U — выпуклое замкнутое ограниченное множество из E^n, функция $J(u) \in C^{1,1}(U)$. Тогда при любом выборе $u_0 \in U$ для последовательности $\{u_k\}$, определяемой условиями* (4), (5), (9), *справедливы равенства*

$$\lim_{k\to\infty} \langle J'(u_k), \bar{u}_k - u_k \rangle = 0, \quad \lim_{k\to\infty} \rho(u_k, S_*) = 0, \tag{13}$$

где $S_* = \{u\colon u \in U, \langle J'(u), v-u\rangle \geqslant 0$ *при всех* $v \in U\}$ — *множество стационарных точек функции* $J(u)$ *на* U.

Если, кроме перечисленных условий, $J(u)$ *выпукла на* U *и*

$$\varepsilon_k + \delta_k \leqslant C_0 k^{-2\rho}, \quad C_0 = \text{const} > 0, \quad 1/2 < \rho \leqslant 1, \tag{14}$$

то

$$\lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} \rho(u_k, U_*) = 0, \tag{15}$$

и справедлива оценка

$$0 \leqslant J(u_k) - J_* \leqslant C_1 k^{-\rho}, \quad k = 1, 2, \ldots; \quad C_1 = \text{const} \geqslant 0. \tag{16}$$

Наконец, если, кроме того, $J(u)$ *сильно выпукла на* U, *то*

$$|u_k - u_*|^2 \leqslant (C_1/\varkappa) k^{-\rho}, \quad k = 1, 2, \ldots \tag{17}$$

Доказательство. При сделанных предположениях $J_* > -\infty$, $U_* \neq \varnothing$. Так как множество U ограничено, то $\sup\limits_{u,v\in U} |u-v| \leqslant d < \infty$. Из условия (9) следует $J(u_{k+1}) = f_k(\alpha_k) \leqslant f_{k*} + \delta_k \leqslant J(u_k + \alpha(\bar{u}_k - u_k)) + \delta_k$ при всех α $(0 \leqslant \alpha \leqslant 1)$. Поэтому пользуясь неравенством (2.3.7), имеем

$$\begin{aligned} J(u_k) - J(u_{k+1}) + \delta_k &\geqslant J(u_k) - J(u_k + \alpha(\bar{u}_k - u_k)) \geqslant \\ &\geqslant -\alpha\langle J'(u_k), \bar{u}_k - u_k\rangle - \alpha^2 L|\bar{u}_k - u_k|^2/2 \geqslant \\ &\geqslant -\alpha J_k(\bar{u}_k) - \alpha^2 L d^2/2, \quad 0 \leqslant \alpha \leqslant 1, \quad k = 0, 1, \ldots \end{aligned} \tag{18}$$

Множество $N = \{0, 1, 2, \ldots\}$ разобьем на два множества $N^+ = \{k\colon k \in N, \langle J'(u_k), \bar{u}_k - u_k\rangle > 0\}$ и $N^- = N \setminus N^+$. Так как $\inf\limits_U J_k(u) \leqslant J_k(u_k) = 0$, то из (4) получаем $0 \leqslant J_k(\bar{u}_k) \leqslant \varepsilon_k$ при всех $k \in N^+$. Поэтому если N^+ — бесконечное множество, то $J_k(\bar{u}_k) \to 0$ при $k \to \infty$, $k \in N^+$.

Теперь пусть $k \in N^-$. Тогда из (18) имеем

$$0 \leqslant -J_k(\bar{u}_k) \leqslant (J(u_k) - J(u_{k+1}) + \delta_k)/\alpha + \alpha L d^2/2 \tag{19}$$

при всех α $(0 < \alpha < 1)$, $k \in N^-$. Далее, из (9) следует, что $J(u_{k+1}) \leqslant J(u_k) + \delta_k$ $(k = 0, 1, \ldots)$. Так как $J(u_k) \geqslant J_* > -\infty$ $(k = 0, 1, \ldots)$, то из леммы 2.3.2 вытекает существование конечного предела $\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$. Следовательно, $\lim\limits_{k\to\infty} (J(u_k) - J(u_{k+1})) = 0$. Если N^- — бесконечное множество, то при $k \to \infty$, $k \in N^-$, из (19) имеем

$$0 \leqslant \varliminf_{k\to\infty} |J_k(\bar{u}_k)| \leqslant \varlimsup_{k\to\infty} |J_k(\bar{u}_k)| \leqslant \alpha L d^2/2$$

при всех α $(0 < \alpha < 1)$. Устремляя $\alpha \to +0$, отсюда получим $J_k(\bar{u}_k) \to 0$ при $k \to \infty$, $k \in N^-$. Объединяя оба случая $k \in N^+$ и $k \in N^-$, приходим к первому равенству (13). Так как U ограничено и $\{u_k\} \in U$, то последовательность $\{u_k\}$ имеет хотя бы одну

предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$, пусть $\{u_{k_m}\} \to u_*$. Согласно (4) $J_k(\bar{u}_k) - \varepsilon_k \leqslant \inf\limits_U J_k(u) \leqslant \langle J'(u_k), u - u_k\rangle$ при всех $u \in U$ и $k = 0, 1, \ldots$ Отсюда при $k = k_m \to \infty$ с учетом первого равенства (13) получим, что $\langle J'(u_*), u - u_*\rangle \geqslant 0$ при всех $u \in U$. Тем самым показано, что любая предельная точка последовательности $\{u_k\}$ принадлежит S_*. Отсюда следует второе равенство (13).

Пусть теперь $J(u)$ выпукла на U и u_* — произвольная точка из U_*. Тогда из теоремы 4.2.2 и условия (4) имеем

$$0 \leqslant a_k = J(u_k) - J(u_*) \leqslant \langle J'(u_k), u_k - u_*\rangle =$$
$$= -J_k(u_*) \leqslant -\min_U J_k(u) \leqslant -J_k(\bar{u}_k) + \varepsilon_k, \quad k = 0, 1, \ldots \tag{20}$$

Отсюда и из первого равенства (13) следует $\lim\limits_{k\to\infty} J(u_k) = J_*$, т. е. $\{u_k\}$ — минимизирующая последовательность. Из теоремы 2.1.1 тогда получаем $\lim\limits_{k\to\infty} \rho(u_k, U_*) = 0$. Равенства (15) доказаны. Заметим, что неравенство (20) может служить полезной апостериорной оценкой при практическом использовании метода (4), (5), (9).

Остается получить оценку (16). Для этого множество $N = \{0, 1, 2, \ldots\}$ разобъем на два множества $I_0 = \{k\colon k \in N,\ a_k \geqslant \varepsilon_k\}$, $I_1 = \{k\colon k \in N,\ 0 \leqslant a_k < \varepsilon_k\}$. Из оценки

$$0 \leqslant a_k - \varepsilon_k \leqslant -J_k(\bar{u}_k), \quad k \in I_0, \tag{21}$$

являющейся следствием неравенства (20), следует, что $I_0 \subseteq N^-$. Поэтому (18) можно переписать в виде

$$a_k - a_{k+1} \geqslant \alpha |J_k(\bar{u}_k)| - \alpha^2 L d^2/2 - \delta_k, \quad 0 \leqslant \alpha \leqslant 1, \quad k \in I_0. \tag{22}$$

Так как в силу (13) $\{|J_k(\bar{u}_k)|\}$ ограничена, то взяв при необходимости d еще большим, можем сделать $0 \leqslant \bar{\alpha}_k = |J_k(\bar{u}_k)| d^{-2} L^{-1} \leqslant 1$ при всех $k = 0, 1, \ldots$ Принимая в (22) $\alpha = \bar{\alpha}_k$, получим

$$a_k - a_{k+1} \geqslant 1/(2Ld^2)\,|J_k(\bar{u}_k)|^2 - \delta_k, \quad k \in I_0.$$

Отсюда и из (21) с учетом условия (14) имеем

$$a_{k+1} \leqslant a_k - (a_k - \varepsilon_k)^2/(2Ld^2) + \delta_k - a_k^2/(2Ld^2) +$$
$$+ \left(\sup_{k\geqslant 0} a_k\right) L^{-1} d^{-2} \varepsilon_k + \delta_k \leqslant a_k - a_k^2/A + Ak^{-2\rho}, \quad k \in I_0, \tag{23}$$

где $A = \max\left\{2Ld^2;\ \left(\sup\limits_{k\geqslant 0} a_k\right) L^{-1} d^{-2} C_0;\ C_0\right\}$.

Если $k \in I_1$, то $0 \leqslant a_k < \varepsilon_k \leqslant C_0 k^{-2\rho}$. Кроме того, из (18) при $\alpha \to +0$ получим $a_k - a_{k+1} + \delta_k \geqslant 0$ или $a_{k+1} \leqslant a_k + \delta_k \leqslant a_k + C_0 k^{-2\rho}$ для всех $k = 0, 1, \ldots$ Таким образом, последовательность $\{a_k\}$

удовлетворяет условиям леммы 2.3.5, из которой следует оценка (16).

Наконец, оценка (17) вытекает из неравенства (4.3.2) и оценки (16). Теорема 1 доказана.

3. Исследуем сходимость метода (4), (5), (10).

Теорема 2. *Пусть U — выпуклое замкнутое ограниченное множество из E^n, функция $J(u)$ принадлежит $C^{1,1}(U)$. Тогда при любом $u_0 \in U$ для последовательности $\{u_k\}$, определяемой условиями* (4), (5), (10), *справедливы равенства* (13). *Если, кроме того, $J(u)$ выпукла на U, то имеют место равенства* (15), *а при $\varepsilon_k \leqslant C_0 k^{-2\rho}$, $C_0 = \text{const} > 0$, $0 < \rho \leqslant 1$, верна оценка* (16). *Для сильно выпуклой функции справедлива оценка* (17).

Доказательство. Так же, как неравенство (18), нетрудно показать, что

$$J(u_k) - J(u_{k+1}) \geqslant -\alpha_k J_k(\bar{u}_k) - \alpha_k^2 L\,|\bar{u}_k - u_k|^2/2, \qquad k = 0, 1, \ldots \tag{24}$$

В соответствии с формулой (10), определяющей величину α_k, рассмотрим три возможных случая:

1) Если $J_k(\bar{u}_k) \leqslant 0$, $\alpha_k = 1 \leqslant \rho_k |J_k(\bar{u}_k)|\,|\bar{u}_k - u_k|^{-2}$, то из (24) с учетом (11) имеем

$$J(u_k) - J(u_{k+1}) \geqslant |J_k(\bar{u}_k)| - L\rho_k |J_k(\bar{u}_k)|/2 \geqslant \varepsilon |J_k(\bar{u}_k)|. \tag{25}$$

2) Если $J_k(\bar{u}_k) \leqslant 0$, $\alpha_k = \rho_k |J_k(\bar{u}_k)|\,|\bar{u}_k - u_k|^{-2} < 1$, то из (24) с учетом (11) получаем

$$\begin{aligned} J(u_k) - J(u_{k+1}) &\geqslant \rho_k |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2} - L\rho_k^2 |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2}/2 = \\ &= |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2} \rho_k (1 - L\rho_k/2) \geqslant \\ &\geqslant |J_k(\bar{u}_k)|^2 d^{-2} \varepsilon_0 \varepsilon, \quad d \geqslant \sup_{u,v \in U} |u - v|. \end{aligned} \tag{26}$$

3) Наконец, если $J_k(\bar{u}_k) > 0$, то согласно (10) и из (24) имеем

$$J(u_k) - J(u_{k+1}) \geqslant 0, \tag{27}$$

а из (4) следует

$$0 < J_k(\bar{u}_k) \leqslant \varepsilon_k. \tag{28}$$

Из (25)—(27) вытекает, что последовательность $\{J(u_k)\}$ не возрастает. Так как $J(u_k) \geqslant J_* > -\infty$, то существует $\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$ и, следовательно, $\lim\limits_{k\to\infty} (J(u_k) - J(u_{k+1})) = 0$. Отсюда и из (25), (26), (28) имеем $0 \leqslant |J_k(\bar{u}_k)| \leqslant \max\{\varepsilon_k;\ \text{const}\cdot(J(u_k) - J(u_{k+1}))^{1/2}\} \to 0$ при всех $k \to \infty$. Первое из равенств (13) доказано. Второе равенство (13) устанавливается так же, как в теореме 1.

Пусть теперь функция $J(u)$ выпукла на U. Тогда справедлива цепочка неравенств (20), из которой следуют равенства (15). Предполагая, что $\varepsilon_k \leqslant C_0 k^{-2\rho}$ $(0 < \rho \leqslant 1)$, докажем оценку (16). Предварительно заметим, что $0 \leqslant a_k = J(u_k) - J_* \leqslant \sup\limits_U J(u) - J_* = C_2 < \infty$, поэтому

$$a_k^2 \leqslant \sup_{k \geqslant 0} a_k \cdot a_k \leqslant C_2 a_k, \qquad k = 0, 1, \ldots \tag{29}$$

Еще раз переберем рассмотренные выше три возможности.

1) Если $J_k(\bar{u}_k) \leqslant 0$, $\alpha_k = 1 \leqslant \rho_k |J_k(\bar{u}_k)|\,|\bar{u}_k - u_k|^2$, то из (20), (25), (29) имеем $a_k - a_{k+1} \geqslant \varepsilon a_k - \varepsilon\varepsilon_k$, если

$$a_{k+1} \leqslant a_k - \varepsilon a_k + \varepsilon\varepsilon_k \leqslant a_k - a_k^2 (\varepsilon/C_2) + \varepsilon C_0 k^{-2\rho}. \tag{30}$$

2) Пусть $J_k(\bar{u}_k) \leqslant 0$, $\alpha_k = \rho_k |J_k(\bar{u}_k)|\ |\bar{u}_k - u_k|^{-2} < 1$. Здесь, в свою очередь, имеются две возможности: $a_k \geqslant \varepsilon_k$ или $0 \leqslant a_k < \varepsilon_k$. Если $a_k \geqslant \varepsilon_k$, то из (20), (26), (29) получим $a_k - a_{k+1} \geqslant (a_k - \varepsilon_k)^2 d^{-2}\varepsilon_0\varepsilon \geqslant a_k^2 d^{-2}\varepsilon_0\varepsilon - 2C_2 \cdot \varepsilon_k d^{-2}\varepsilon_0\varepsilon$ или

$$a_{k+1} \leqslant a_k - a_k^2\left(\varepsilon_0\varepsilon/d^2\right) + 2C_2\varepsilon_0\varepsilon d^{-2}C_0 k^{-2\rho}. \tag{31}$$

Если же $0 \leqslant a_k < \varepsilon_k$, то достаточно воспользоваться более простым следствием (26): $a_{k+1} \leqslant a_k$. Последние два неравенства можно переписать в виде

$$0 \leqslant a_k \leqslant C_0 k^{-2\rho}, \quad a_{k+1} \leqslant a_k \leqslant a_k + C_0 k^{-2\rho}. \tag{32}$$

3) Наконец, пусть $J_k(\bar{u}_k) > 0$, $\alpha_k = 0$. Тогда из (20), (27) получим $0 \leqslant a_k \leqslant \varepsilon_k$, $a_{k+1} \leqslant a_k$, что снова приведет к неравенствам (32).

Из (30)—(32) следует, что последовательность $\{a_k\}$ удовлетворяет условиям леммы 2.3.5, из которой получаем оценку (16). Теорема 2 доказана.

4. Наконец, рассмотрим вариант метода условного градиента (4), (5), (12).

Теорема 3. *Пусть U — выпуклое замкнутое ограниченное множество из E^n, функция $J(u) \in C^{1,1}(U)$ и выпукла на U. Тогда при любом $u_0 \in U$ для последовательности $\{u_k\}$, определяемой условиями* (4), (5), (12), *справедливы равенства* (15). *Если при этом $\alpha_k = (k+1)^{-1}$, $\varepsilon_k = C_3(k+1)^{-1}$ $(k = 0, 1, \ldots)$, то*

$$0 \leqslant J(u_k) - J_* \leqslant C_4 \ln(k+1)/k, \qquad k = 1, 2, \ldots, \tag{33}$$

а если $\alpha_k = (k+1)^{-\beta}$, $\varepsilon_k = C_3(k+1)^{-\beta}$ $(k = 0, 1, \ldots, 0 < \beta < 1)$, то

$$0 \leqslant J(u_k) - J_* \leqslant C_4 k^{-\beta}, \qquad k = 1, 2, \ldots; \tag{34}$$

здесь C_3, C_4 — некоторые положительные постоянные.

Доказательство. Заметим, что неравенства (20), (24) не зависят от способа выбора α_k $(0 \leqslant \alpha_k \leqslant 1)$ в (5), поэтому сохраняют силу и в рассматриваемом случае. Из них имеем $a_k - a_{k+1} \geqslant \alpha_k(a_k - \varepsilon_k) - \alpha_k^2 L d^2/2$ или

$$a_{k+1} \leqslant (1 - \alpha_k)\, a_k + \alpha_k^2 L d^2/2 + \alpha_k \varepsilon_k, \qquad k = 0, 1, \ldots$$

Отсюда с учетом свойств последовательностей $\{\alpha_k\}$, $\{\varepsilon_k\}$ из (4), (12) заключаем, что $\{a_k\}$ удовлетворяет условиям леммы 2.3.6. Поэтому $\lim\limits_{k\to\infty} a_k = 0$ или $\lim\limits_{k\to\infty} J(u_k) = J_*$. Отсюда и из теоремы 2.1.1 получаем равенства (15). Оценки (33), (34) следуют из лемм 2.3.8, 2.3.9.

Упражнения. 1. Вычислить несколько итераций метода (3), (5), (6) для функции $J(u) = x^2 + xy + y^2$ при $u \in U = \{u = (x, y) \in E^2\colon 0 \leqslant x \leqslant 1, -1 \leqslant y \leqslant 0\}$, выбирая $u_0 = (1, -1)$, $(-1, 0)$, $(1, 0)$ или $(0, 0)$.

2. Для функции из примера 1 проверить выполнение условий теорем 1—3 и сформулировать условия сходимости соответствующих вариантов метода условного градиента.

3. Дать описание различных вариантов метода условного градиента для функции $J(u) = |Au - b|^2$, где A — матрица $m \times n$, $b \in E^m$, а множество U является шаром или параллелепипедом. Опираясь на теоремы 1—3, доказать сходимость метода.

§ 5. Метод возможных направлений

1. Продолжим рассмотрение задачи минимизации гладкой функции $J(u)$ на заданном множестве $U \subseteq E^n$. Напомним, что направление $e \neq 0$ называется *возможным* в точке $u \in U$, если $u + te \in U$ при всех t, $0 \leqslant t \leqslant t_0$, где t_0 — положительное число, зависящее от точки u, направления e и от структуры множества U (см. определение 4.2.3).

Определение 1. Направление $e \neq 0$ назовем *возможным направлением убывания* функции $J(u)$ в точке u на множестве U, если e — возможное направление в точке u и $J(u + \alpha e) < J(u)$ при всех α, $0 < \alpha < \beta$, где $0 < \beta \leqslant t_0$.

Метод возможных направлений основан на следующей естественной и прозрачной идее: на каждой итерации этого метода определяется возможное направление убывания функции и по этому направлению осуществляется спуск с некоторым положительным шагом. Собственно говоря, эта идея для нас уже не новая — именно на ней были основаны многие варианты изложенных в § 1, 2, 4 методов. В самом деле, если $U = E^n$, $J'(u) \neq 0$, то возможное направление убывания функции легко находится — это направление антиградиента $e = -J'(u)$. Более трудным был выбор возможного направления убывания в методах § 2, 4: в методе проекции градиента (см. формулы (2.2) и (2.2′)) для этого нужно было проектировать точку на исходное множество U, а в методе условного градиента — решать задачу минимизации линейной функции на множестве U (см. задачу (4.3)).

Понятно, что если задача выбора возможного направления убывания на каждой итерации слишком сложна и требует решения вспомогательных задач минимизации, сравнимых по трудности, быть может, с исходной задачей, то такой метод минимизации будет малоэффективным. Возникает вопрос: нельзя ли указать простые и достаточно удобные для реализации на ЭВМ способы выбора возможных направлений убывания? Оказывается, для достаточно широких классов гладких задач такие способы существуют. Покажем это на примере следующей задачи:

$$J(u) \to \inf; \quad u \in U = \{u \in E^n\colon\ g_i(u) \leqslant 0,\ i = 1, \ldots, m\}, \qquad (1)$$

где функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, m)$, определены на всем пространстве E^n и $J(u), g_i(u) \in C^1(U)$.

Чтобы проще было пояснить суть метода возможных направлений для задачи (1), сначала опишем более простой вариант этого метода. Пусть $u_0 \in U$ — некоторое начальное приближение. Пусть известно k-е приближение $u_k \in U$ $(k \geqslant 0)$. Введем множество номеров

$$I_k = \{i\colon\ 1 \leqslant i \leqslant m,\ g_i(u_k) = 0\}.$$

Возможно, что $I_k = \varnothing$, — это будет означать, что $g_i(u_k) < 0$ при всех $i = 1, \ldots, m$, т. е. $u_k \in \operatorname{int} U$ — такая возможность ниже не

исключается. В пространстве переменных

$$z=(e,\ \sigma)=(e^1,\ \ldots,\ e^n,\ \sigma)\in E^{n+1}$$

рассмотрим вспомогательную задачу

$$\sigma\to\inf,\quad z=(e,\ \sigma)\in W_k=\{(e,\ \sigma):\ \langle J'(u_k),\ e\rangle\leqslant\sigma,$$
$$\langle g_i'(u_k),\ e\rangle\leqslant\sigma,\quad i\in I_k;\quad |e^j|\leqslant 1,\quad j=1,\ \ldots,\ n\}.\qquad (2)$$

Заметим, что задача (2) является задачей линейного программирования, причем минимизируемая функция $\langle c,\ z\rangle=\langle 0,\ e\rangle+1\cdot\sigma$, $c=(0,\ 1)\in E^{n+1}$, явно не зависит от переменных $e=(e^1,\ \ldots,\ e^n)$. Далее, ясно, что точка $z=(e=0,\ \sigma=0)=(0,\ 0)\in W_k$, так что $W_k\neq\varnothing$ и $\inf\limits_{W_k}\sigma=\sigma_k\leqslant 0$ при всех $k=0,\ 1,\ \ldots$ Очевидно, множество W_k замкнуто. Наконец, условия $|e^j|\leqslant 1$ $(j=1,\ \ldots,\ n)$, называемые условиями нормировки, гарантируют ограниченность множества W_k. Тогда из теоремы 2.1.1 следует, что задача (2) имеет хотя бы одно решение. Для получения решения задачи (2) могут быть использованы известные конечные методы линейного программирования (например, симплекс-метод, описанный в гл. 3).

Предположим, что задача (2) решена и найдены $(e_k,\ \sigma_k)\in W_k$ такие, что $\sigma_k=\inf\limits_{W_k}\sigma$. Выше было замечено, что $\sigma_k\leqslant 0$.

Сначала рассмотрим случай $\sigma_k<0$. Оказывается, в этом случае направление e_k, полученное из задачи (2), является возможным направлением убывания функции $J(u)$ в точке u_k. В самом деле, из условия $(e_k,\ \sigma_k)\in W_k$ следует, что

$$\langle J'(u_k),\ e_k\rangle\leqslant\sigma_k<0,\quad \langle g_i'(u_k),\ e_k\rangle\leqslant\sigma_k<0,\quad i\in I_k.$$

Отсюда ясно, что $e_k\neq 0$. Кроме того, для любого номера $i\in I_k$ имеем

$$(g_i(u_k+\alpha e_k)=g_i(u_k+\alpha e_k)-g_i(u_k)=\langle g_i'(u_k),\ e_k\rangle\alpha+o(\alpha)\leqslant$$
$$\leqslant\alpha\,[\sigma_k+o(\alpha)/\alpha]<0\quad\text{при всех}\quad\alpha,\quad 0<\alpha<\alpha_i,\ \alpha_i>0.$$

Если $i\notin I_k$, т. е. $g_i(u_k)<0$, то в силу непрерывности функции $g_i(u)$ неравенство $g_i(u_k+\alpha e_k)<0$ сохранится при всех α $(0<\alpha<\alpha_i)$, где $\alpha_i>0$ — достаточно малое число. Положим $\alpha_0=\min\{\alpha_1,\ \ldots,\ \alpha_m\}>0$. Тогда

$$g_i(u_k+\alpha e_k)<0,\quad i=1,\ \ldots,\ m;\quad 0<\alpha<\alpha_0,$$

т. е. e_k — возможное направление множества U в точке u_k.

Далее, взяв при необходимости число $\alpha_0>0$ еще меньшим, можно добиться выполнения неравенства

$$J(u_k+\alpha e_k)-J(u_k)=\langle J'(u_k),\ e_k\rangle\alpha+o(\alpha)\leqslant$$
$$\leqslant\alpha[\sigma_k+o(\alpha)/\alpha]<0\ \text{при всех}\ \alpha,\ 0<\alpha<\alpha_0.$$

Тем самым показано, что если (e_k, σ_k) — решение задачи (2), причем $\sigma_k < 0$, то e_k — возможное направление убывания функции $J(u)$ в точке u_k на множестве U.

Используя найденное таким образом направление e_k, следующее $(k+1)$-е приближение определим так:

$$u_{k+1} = u_k + \alpha_k e_k, \quad 0 < \alpha_k \leqslant \beta_k, \tag{3}$$

где

$$\beta_k = \sup\{\alpha\colon\ u_k + te_k \in U, \quad 0 \leqslant t \leqslant \alpha\} > 0. \tag{4}$$

Выбирая α_k в (3) различными способами, будем получать различные варианты метода возможных направлений. Перечислим некоторые способы выбора α_k.

1) Величина α_k может выбираться из условий

$$0 < \alpha_k \leqslant \beta_k, \quad f_k(\alpha_k) = \inf_{0<\alpha\leqslant\beta_k} f_k(\alpha) = f_{k*};$$

$$f_k(\alpha) = J(u_k + \alpha e_k). \tag{5}$$

Для минимизации функции $f_k(\alpha)$ могут быть использованы известные методы (см., например, гл. 1). Точное решение задачи (5) удается найти лишь в редких случаях; возможно также, что на некоторых направлениях e_k величина $\beta_k = \infty$ и нижняя грань функции $f_k(\alpha)$ при $\alpha > 0$ не достигается. Поэтому вместо (5) на практике целесообразно употреблять такой способ выбора α_k:

$$0 < \alpha_k \leqslant \beta_k, \quad f_k(\alpha_k) \leqslant f_k + \delta_k, \quad \delta_k \geqslant 0, \quad \sum_{k=0}^{\infty} \delta_k = \delta < 0 \tag{6}$$

или

$$J(u_k + \alpha_k e_k) \leqslant (1 - \lambda_k) J(u_k) + \lambda_k f_{k*}, \quad 0 < \lambda \leqslant \lambda_k \leqslant 1.$$

2) Если функция $J(u) \in C^{1,1}(U)$ и константа Липшица L для градиента $J'(u)$ известна, то в (3) в качестве α_k можно принять

$$\alpha_k = \min\{\beta_k;\ |\sigma_k| L^{-1}\}.$$

3) Возможен выбор α_k из условий

$$J(u_k) - J(u_k + \alpha_k e_k) \geqslant \varepsilon \alpha_k |\sigma_k|, \quad 0 < \alpha_k \leqslant \beta_k, \quad 0 < \varepsilon < 1/2.$$

Для определения такого α_k сначала можно положить $\alpha_k = \beta_k$, а затем при необходимости дробить эту величину.

4) В тех случаях, когда трудно оценить величину β_k из (4), приходится довольствоваться нахождением какого-либо $\alpha_k > 0$, обеспечивающего включение $u_k + \alpha_k e_k \in U$ и условие монотонности $J(u_k + \alpha_k e_k) < J(u_k)$. Для этого обычно выбирают какую-либо постоянную $\alpha > 0$, полагают $\alpha_k = \alpha$ и проверяют условие монотонности и принадлежность точки u_{k+1} множеству U; при необходимости дробят величину $\alpha_k = \alpha$, добиваясь выполнения упомянутых условий.

Один шаг метода возможных направлений для задачи (1) в случае $\sigma_k < 0$ описан. Попутно выяснен смысл вспомогательной задачи (2): минимизируя σ, мы добиваемся того, чтобы направление e_k было как можно ближе к направлению антиградиента (это обеспечивается условием $\langle J'(u_k), e_k\rangle \leqslant \sigma$) и в то же время оставалось возможным направлением для множества U в точке u_k (это обеспечивается условиями $\langle g_i'(u_k), e_k\rangle \leqslant \sigma$, $i \in I_k$), причем чем меньше σ, тем ярче выражены указанные свойства направления e_k. Кстати, если $I_k = \varnothing$, т. е. $u_k \in \operatorname{int} U$, то $e_k = -\alpha J'(u_k)$, $\alpha = \left(\max\limits_{1 \leqslant j \leqslant n} |J_{u^j}(u_k)|\right)^{-1} > 0$ — направление антиградиента.

Теперь рассмотрим случай, когда в решении (e_k, σ_k) задачи (2) координата $\sigma_k = 0$. Как видно из (2), это может случиться, например, при $J'(u_k) = 0$ или $g_i'(u_k) = 0$ для некоторого номера $i \in I_k$. При $\sigma_k = 0$ уже нельзя гарантировать, что e_k будет возможным направлением убывания. В этом случае итерационный процесс (2)—(4) прекращается. Оказывается, при $\sigma_k = 0$ точка u_k является стационарной точкой задачи (1), или иначе говоря, в точке u_k выполняются необходимые условия минимума, выраженные в теореме 4.8.1. Для выпуклой регулярной задачи (1) условие $\sigma_k = 0$ гарантирует, что $u_k \in U_*$. Покажем это.

Теорема 1. *Пусть функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, m)$ определены на E^n, $J(u)$, $g_i(u) \in C^1(U)$, где множество U задано условиями* (1), *и пусть задача* (1) *имеет решение, т. е. $J_* > -\infty$, $U_* \neq \varnothing$. Тогда для любой точки $u_* \in U_*$ задача*

$$\sigma \to \inf; \quad z = (e, \sigma) \in W_* = \{(e, \sigma)\colon \langle J'(u_*), e\rangle \leqslant \sigma,$$
$$\langle g_i'(u_*), e\rangle \leqslant \sigma,\ i \in I_*,\ |e^j| \leqslant 1,\ j = 1, \ldots, n\}, \quad (7)$$

где

$$I_* = \{i\colon 1 \leqslant i \leqslant m,\ g_i(u_*) = 0\},$$

необходимо имеет решение (e_*, σ_*) с $\sigma_* = \min\limits_{W_*} \sigma = 0$. Если, кроме того, $J(u)$, $g_i(u)$ выпуклы на E^n, а множество U регулярно (*см. определение* 4.9.2), *то всякая точка $u_* \in U$, для которой задача* (7) *определяет величину $\sigma_* = \inf\limits_{W_*} \sigma = 0$, является решением задачи* (1).

Доказательство. Необходимость. Пусть $u_* \in U_*$. По теореме 4.8.1 тогда существуют множители Лагранжа $\lambda_0^*, \ldots, \lambda_m^*$, неотрицательные и не все равные нулю, такие, что

$$\lambda_0^* J'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i'(u_*) = 0, \quad \lambda_i^* g_i(u_*) = 0, \quad i = 1, \ldots, m. \quad (8)$$

Если $i \notin I_*$, то из второго равенства (8) следует $\lambda_i^* = 0$, поэтому

первое равенство (8) можно переписать в виде

$$\lambda_0^* J'(u_*) + \sum_{i \in I_*} \lambda_i^* g_i'(u_*) = 0. \tag{9}$$

Возьмем любую точку $(e, \sigma) \in W_*$. Тогда $\langle J'(u_*), e\rangle \leqslant \sigma$, $\langle g'(u_*), e\rangle \leqslant \sigma$ $(i \in I_*)$. Умножим первое из этих неравенств на $\lambda_0^* \geqslant 0$, остальные — на соответствующие $\lambda_i^* \geqslant 0$ и сложим. С учетом равенства (9) получим

$$\left\langle \lambda_0^* J'(u_*) + \sum_{i \in I_*} \lambda_i^* g_i'(u_*), e \right\rangle = 0 \leqslant \sigma(\lambda_0^* + \lambda_1^* + \ldots + \lambda_m^*).$$

Достаточность. Пусть теперь $J(u)$, $g_i(u)$ выпуклы на U, а множество U регулярно, пусть для некоторой точки $u_* \in U$ задача (7) определяет величину $\sigma_* = \inf\limits_{W_*} \sigma = 0$. Покажем, что тогда $u_* \in U_*$. С этой целью введем конус

$$K = \{z = (e, \sigma) \in E^{n+1}\colon \langle J'(u_*), e\rangle + (-1)\sigma \leqslant 0,$$
$$\langle g_i'(u_*), e\rangle + (-1)\sigma \leqslant 0,\ i \in I_*\},$$

образующими которого являются векторы $c_0 = (J'(u_*), -1)$, $c_i = (g_i'(u_*), -1)$ $(i \in I_*)$. Покажем, что вектор $d = (0, 1) \in K^*$ — двойственный к K конус. Из (7) с учетом $\inf\limits_{W_*} \sigma = \sigma_* = 0$ имеем

$$\langle d, z\rangle = \langle 0, e\rangle + 1 \cdot \sigma = \sigma \geqslant \sigma_* = 0 \tag{10}$$

для всех $z = (e, \sigma) \in K$, для которых $|e^j| \leqslant 1$ $(j = 1, \ldots, n)$. Однако условие $|e^j| \leqslant 1$ $(j = 1, \ldots, n)$ здесь можно отбросить, и неравенство (10) на самом деле верно для всех $z \in K$. В самом деле, пусть $z = (e, \sigma) \in K$ и $|e^j| > 1$ для некоторого номера j $(1 \leqslant j \leqslant n)$. Тогда $\|e\| = \max\limits_{1 \leqslant j \leqslant n} |e^j| > 1$. Положим

$$\bar{z} = (\bar{e}, \bar{\sigma}), \quad \bar{e} = e/\|e\|, \quad \bar{\sigma} = \sigma/\|e\|.$$

Ясно, что $\bar{z} \in W_*$. Следовательно, $\langle d, \bar{z}\rangle = \langle d, z\rangle/\|e\| = \bar{\sigma} = \sigma/\|e\| \geqslant 0$, так что $\langle d, z\rangle = \sigma \geqslant 0$. Тем самым показано, что неравенство (10) верно для всех $z \in K$, т. е. $d \in K^*$.

По теореме 4.9.3 тогда существуют неотрицательные числа $\lambda_0^*, \ldots, \lambda_m^*$ такие, что

$$d = (0, 1) = -\lambda_0^* c_0 - \sum_{i \in I_*} \lambda_i^* c_i.$$

Вспоминая определения векторов c_0, c_i $(i \in I_*)$, отсюда имеем

$$0 = -\lambda_0^* J'(u_*) - \sum_{i \in I_*} \lambda_i^* g_i'(u_*) = 0, \quad 1 = \lambda_0^* + \sum_{i \in I_*} \lambda_i^*. \tag{11}$$

Кроме того, из определения множества I_* следует, что $g_i(u_*) = 0$, поэтому $\lambda_i^* g_i(u_*) = 0$ $(i \in I_*)$. Доопределим $\lambda_i^* = 0$ при всех

$i \notin I_*$. В результате с учетом первого равенства (11) получим

$$\lambda_0^* J'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i'(u_*) = 0, \quad \lambda_i^* g_i(u_*) = 0, \quad i = 1, \ldots, m, \quad (12)$$

а из второго равенства (11) следует, что не все числа λ_0^*, λ_i^* ($i \in I_*$), равны нулю.

Покажем, что $\lambda_0^* > 0$. Если $I_* = \varnothing$, то из (11) сразу имеем $\lambda_0^* = 1$. Допустим, что $I_* \neq \varnothing$, но тем не менее $\lambda_0^* = 0$. Тогда среди неотрицательных чисел λ_i^* ($i \in I_*$) найдется хотя бы одно положительное число. Пусть $\lambda_p^* > 0$, $p \in I_*$. По условию множество U регулярно, поэтому существует точка $\bar{u} \in U$ такая, что $g_i(\bar{u}) < 0$ для всех $i = 1, \ldots, m$. Поскольку $I_* \neq \varnothing$, то $\bar{u} \neq u_*$. В силу выпуклости множества U тогда $\alpha\bar{u} + (1-\alpha)u_* = u_* + \alpha(\bar{u}_* - u_*) \in U$ при всех α $(0 \leqslant \alpha \leqslant 1)$. Это значит, что направление $e = \bar{u} - u_* \neq 0$ является возможным для множества U в точке u_*. Из выпуклости функций $g_i(u)$ для всех $i \in I_*$ имеем $0 > g_i(\bar{u}) = g_i(\bar{u}) - g_i(u_*) \geqslant \langle g_i'(u_*), \bar{u} - u_* \rangle = \langle g_i'(u_*), e \rangle$. Поэтому $\sum_{i=1}^{m} \lambda_i^* \langle g_i'(u_*), e \rangle \leqslant \lambda_p^* \langle g_p'(u_*), e \rangle < 0$. Но с другой стороны, из первого равенства (12) при $\lambda_0^* = 0$ получим $\sum_{i=1}^{m} \lambda_i^* \langle g_i'(u_*), e \rangle = 0$. Полученное противоречие показывает, что $\lambda_0^* > 0$. Разделив первое равенство (12) на $\lambda_0^* > 0$ и сделав очевидные переобозначения, придем к равенству $J'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i(u_*) = 0$. Функция Лагранжа $L(u, \lambda^*) = J(u) + \sum_{i=1}^{m} \lambda_i^* g_i(u)$ выпукла по $u \in E^n$ из-за выпуклости $J(u)$, $g_i(u)$ и неотрицательности λ_i^* ($i = 1, \ldots, m$). Поэтому предыдущее равенство в силу теоремы 4.2.3 равносильно условию $L(u_*, \lambda^*) \geqslant L(u, \lambda^*)$ при всех $u \in E^n$. Отсюда и из второго равенства (12) с помощью леммы 4.9.1 и теоремы 4.9.1 получим, что $u_* \in U_*$. Теорема 1 доказана.

В невыпуклых задачах условие $\sigma_* = 0$ не является достаточным для оптимальности точки u_*. Это показывает следующий

Пример 1. Пусть $J(u) = x + \cos y$, $u \in U = \{u = (x, y) \in E^2\colon g(u) = -x \leqslant 0\}$ (ср. с примером 4.8.1). Возьмем точку $u_* = (0, 0)$. Тогда $J'(u_*) = (1, 0)$, $g'(u_*) = (-1, 0)$, $W_* = \{(e, \sigma) = (e^1, e^2, \sigma)\colon e^1 \leqslant \sigma, -e^1 \leqslant \sigma, |e^1| \leqslant 1, |e^2| \leqslant 1\}$. Отсюда $|e^1| \leqslant \sigma$ при всех $(e, \sigma) \in W_*$. Это значит, что $\inf\limits_{W_*} \sigma = \sigma_* = 0$, причем нижняя грань достигается при $e_* = (0, 1)$ или $e_* = (0, -1)$, $\sigma_* = 0$. Но здесь $u_* = (0, 0)$ не является точкой минимума $J(u)$ на U. Любо-

пытно заметить, что векторы $e_* = (0,1)$ или $(0, -1)$ в данном случае являются возможными направлениями убывания.

2. Описанный выше вариант метода возможных направлений (2)—(4) на практике применяют редко. Дело в том, что когда в решении (e_k, σ_k) задачи (2) координата $\sigma_k < 0$ мала по абсолютной величине, направление e_k, теоретически являясь возможным направлением убывания в точке u_k, практически может обладать указанными свойствами в весьма слабой форме. Это означает, что либо $\langle g_i'(u_k), e_k\rangle \approx \sigma_k \approx 0$ при некотором $i \in I_k$ и направление e_k почти «касается» множества U, не ведет «вглубь» U, а величина β_k из (4) может оказаться очень малой, либо $\langle J'(u_k), e_k\rangle \approx \sigma_k \approx 0$, т. е. вдоль e_k функция $J(u)$ в точке u_k убывает слишком медленно. В результате длина шага α_k в (3) может получиться очень малой даже вдали от стационарной точки, и сходимость метода может оказаться очень медленной.

Чтобы как-то избежать таких неприятных явлений, можно попытаться несколько варьировать множество номеров I_k в (2) и осуществлять спуск из точки u_k только в том случае, когда получаемое из (2) направление e_k обладает достаточно ярко выраженными свойствами возможного направления убывания.

Опишем один из таких подходов. Пусть $u_0 \in U$, $\varepsilon_0 > 0$ — некоторое начальное приближение. Допустим, что k-е приближение (u_k, ε_k), $u_k \in U$, $\varepsilon_k > 0$, при каком-то $k \geqslant 0$ уже известно. Определим множество номеров

$$I_k = \{i : 1 \leqslant i \leqslant m, \ -\varepsilon_k \leqslant g_i(u_k) \leqslant 0\} \tag{13}$$

и решим вспомогательную задачу (2) при таком I_k. Задача (2) по-прежнему будет задачей линейного программирования и будет обладать хотя бы одним решением (e_k, σ_k) с $\sigma_k = \inf\limits_{W_k} \sigma \leqslant 0$. Имеются две возможности:

1) $\sigma_k \leqslant -\varepsilon_k$. В этом случае считаем, что e_k является достаточно хорошим возможным направлением убывания в точке u_k, и полагаем

$$u_{k+1} = u_k + \alpha_k e_k, \quad 0 < \alpha_k \leqslant \beta_k, \quad \varepsilon_{k+1} = \varepsilon_k, \tag{14}$$

где β_k определяется из (4), а выбор α_k может быть осуществлен одним из описанных выше способов.

2) $-\varepsilon_k < \sigma_k \leqslant 0$. В этом случае считаем, что направление e_k не обладает ясно выраженным свойством возможного направления в точке u_k, полагаем

$$u_{k+1} = u_k, \quad \varepsilon_{k+1} = \varepsilon_k/2 \tag{15}$$

и снова переходим к рассмотрению задачи (2) с заменой множества I_k на множество $I_{k+1} = \{i: 1 \leqslant i \leqslant m, \ -\varepsilon_{k+1} = -\varepsilon_k/2 \leqslant g_i(u_k) \leqslant 0\}$, надеясь на то, что на более широком множестве (при сужении I_k множество W_k, вообще говоря, расширяется) удастся найти лучшее возможное направление убывания и т. д.

Описание одной итерации метода возможных направлений для задачи (1) закончено. В методе (2), (13)—(15) имеются параметры $\varepsilon_0, \varepsilon_1, \ldots$, удачным выбором которых, вообще говоря, можно улучшить выбор направлений e_k на каждой итерации, ускорить сходимость метода. Кстати, в (15) вместо деления пополам можно принять иной способ дробления ε_k, например, $\varepsilon_{k+1} = 0{,}9\varepsilon_k$.

3. Следуя [11], изучим сходимость метода (2), (4), (6), (13)—(15). Предварительно докажем несколько лемм.

Лемма 1. Пусть $J(u), g_i(u) \in C^{1,1}(U)$ $(i = 1, \ldots, m)$ *и* I — *некоторое фиксированное множество номеров, взятых из* $\{1, 2, \ldots, m\}$ *(возможности* $I = \varnothing$ *или* $I = \{1, \ldots, m\}$ *не исключаются). Для каждого* $u \in U$ *положим* $\sigma(u) = \min\limits_{G(u)} \sigma$, *где* $G(u) = \{(e, \sigma) = (e^1, \ldots, e^n, \sigma) \in E^{n+1}: \langle J'(u), e\rangle \leqslant \sigma, \langle g_i'(u), e\rangle \leqslant \sigma, \quad i \in I; |e^j| \leqslant 1, j = 1, n\}$. *Тогда*

$$|\sigma(u) - \sigma(v)| \leqslant L\sqrt{n}|u - v|, \quad u, v \in U, \tag{16}$$

где L — *константа Липшица для градиентов* $J'(u), g_i'(u)$ $(i = 1, \ldots, m)$.

Доказательство. Возьмем произвольные точки $u, v \in U$. Пусть $(e, \sigma) \in C(v)$, т. е.

$$\langle J'(v), e\rangle \leqslant \sigma, \quad \langle g_i'(v), e\rangle \leqslant \sigma, \quad i \in I, \quad |e^j| \leqslant 1, \quad j = 1, \ldots, n.$$

Тогда

$$\langle J'(u), e\rangle = \langle J'(v), e\rangle + \langle J'(u) - J'(v), e\rangle \leqslant \sigma + L|u - v|\,|e| \leqslant \leqslant \sigma + L|u - v|\sqrt{n}$$

и, аналогично,

$$\langle g_i'(u), e\rangle \leqslant \sigma + L|u - v|\sqrt{n}, \quad i \in I.$$

Это значит, что при каждом $(e, \sigma) \in G(v)$ точка $(e, \sigma + L\sqrt{n}|u - v|)$ принадлежит множеству $G(u)$. Тогда $\sigma(u) = \min\limits_{G(u)} \sigma \leqslant \sigma + L\sqrt{n}|u - v|$ для любых $(e, \sigma) \in G(v)$. Следовательно, $\sigma(u) \leqslant \sigma(v) + L\sqrt{n}|u - v|$. Поменяв в этих рассуждениях точки u, v ролями, получим $\sigma(v) \leqslant \sigma(u) + L\sqrt{n}|u - v|$. Из последних двух неравенств следует неравенство (16).

Лемма 2. *Пусть*

$$J(u), g_1(u), \ldots, g_m(u) \in C^{1,1}(U), \quad \max_{1 \leqslant i \leqslant m} \sup_{u \in U} |g_i'(u)| < \infty,$$

а последовательности $\{u_k\}$, $\{\alpha_k\}$, $\{\beta_k\}$, $\{\varepsilon_k\}$, $\{\sigma_k\}$ *определены условиями* (2), (4), (6), (13)—(15). *Тогда*

$$\beta_k \geqslant A_1 \min\{\varepsilon_k, |\sigma_k|\}, \quad k = 0, 1, \ldots, \tag{17}$$

где $A_1 = \min\{1/(A_0\sqrt{n}); 1/(nL)\} > 0$, L — *константа Липшица для градиентов* $J'(u)$, $g_1'(u), \ldots, g_m'(u)$.

Доказательство. Если $\beta_k = \infty$, то неравенство (17) верно. Поэтому пусть $\beta_k < \infty$. Из определения (4) величины β_k и замкнутости U следует, что $u_k + \beta_k e_k \in U$ и $g_i(u_k + \beta_k e_k) = 0$ хотя бы для одного номера i. Зафиксируем один из таких номеров i. Может оказаться, что $g_i(u_k) < -\varepsilon_k$. Тогда $\varepsilon_k < -g_i(u_k) = g_i(u_k + \beta_k e_k) - g_i(u_k) = \langle g_i'(u_k + \theta\beta_k e_k), \beta_k e_k\rangle \leqslant \leqslant A_0\beta_k|e_k| \leqslant A_0\sqrt{n}\beta_k$, т. е. $\beta_k \geqslant \varepsilon_k/(A_0\sqrt{n})$.

Если же оказалось, что $-\varepsilon_k \leqslant g_i(u_k) \leqslant 0$, то $i \in I_k$ и $\langle g_i'(u_k), e_k\rangle \leqslant \sigma_k \leqslant 0$. Допустим, что $\sigma_k < 0$. Тогда направление e_k является возможным в точке u_k и заведомо $\beta_k > 0$. По определению β_k имеем $g_i(u_k + \alpha e_k) \leqslant 0$ при всех $\alpha (0 < \alpha < \beta_k)$. Кроме того, $g_i(u_k + \beta_k e_k) = 0$ по выбору номера i. Тогда $0 \geqslant g_i(u_k + \alpha e_k) - g_i(u_k + \beta_k e_k) = \langle g_i'(u_k + \beta_k e_k), e_k\rangle (\alpha - \beta_k) + o(|\alpha - \beta_k|)$ или $\langle g_i'(u_k + \beta_k e_k), e_k\rangle \geqslant o(|\alpha - \beta_k|)(\beta_k - \alpha)^{-1}$ при всех α $(0 < \alpha < \beta_k)$. Отсюда при $\alpha \to \beta_k - 0$ получим $\langle g_i'(u_k + \beta_k e_k), e_k\rangle \geqslant 0$. Тогда $|\sigma_k| = -\sigma_k \leqslant \langle -g_i'(u_k), e_k\rangle \leqslant \langle g_i'(u_k + \beta_k e_k) - g_i'(u_k), e_k\rangle \leqslant L\beta_k |e_k|^2 \leqslant Ln\beta_k$, т. е. $\beta_k \geqslant |\sigma_k|/(nL)$. Если $\sigma_k = 0$, то последнее неравенство также остается верным, так как согласно (4) всегда $\beta_k \geqslant 0$. Объединяя обе полученные оценки для β_k, приходим к оценке (17). Лемма 2 доказана.

Лемма 3. *Пусть $J(u)$, $g_i(u) \in C^{1,1}(U)$ $(i = 1, \ldots, m)$. Пусть, кроме того, в процессе* (2), (4), (6), (13)—(15) *на некоторой k-й итерации оказалось $\sigma_k \leqslant -\varepsilon_k$. Тогда*

$$J(u_k) - J(u_{k+1}) \geqslant A_2 \min\{\beta_k |\sigma_k|; \sigma_k^2\} - \delta_k, \tag{18}$$

где $A_2 = \min\{1/2; 1/(2nL)\} > 0$.

Доказательство. Из неравенства $\sigma_k \leqslant -\varepsilon_k$ и определения e_k, σ_k следует, что $\langle J'(u_k), e_k\rangle \leqslant \sigma_k \leqslant -\varepsilon_k < 0$. Кроме того, e_k является возможным направлением в точке u_k и, следовательно, $\beta_k > 0$. Из (6) и леммы 2.3.1 имеем

$$J(u_k) - J(u_{k+1}) \geqslant J(u_k) - \inf_{0 \leqslant \alpha \leqslant \beta_k} f_k(\alpha) - \delta_k \geqslant J(u_k) - J(u_k + \alpha e_k) - \delta_k \geqslant$$

$$\geqslant -\alpha \langle J'(u_k), e_k\rangle - \alpha^2 L |e_k|^2/2 - \delta_k \geqslant -\alpha\sigma_k - \alpha^2 nL/2 - \delta_k \tag{19}$$

при всех α $(0 \leqslant \alpha \leqslant \beta_k)$. Положим здесь $\alpha = \min\{\beta_k; |\sigma_k|/(nL)\}$.

Может случиться, что $\alpha = \beta_k \leqslant |\sigma_k|/(nL)$. Тогда из (19) получаем

$$J(u_k) - J(u_{k+1}) \geqslant \alpha|\sigma_k| - \alpha\cdot\alpha\cdot nL/2 - \delta_k \geqslant \beta_k|\sigma_k| - \beta_k(|\sigma_k|/(nL))(nL/2) - \delta_k = \beta_k|\sigma_k|/2 - \delta_k.$$

Если же $\alpha = |\sigma_k|/(nL) < \beta_k$, то из (19) следует

$$J(u_k) - J(u_{k+1}) \geqslant \sigma_k^2/(nL) - (\sigma_k^2/(nL)^2)(nL/2) - \delta_k = \sigma_k^2/(2nL) - \delta_k.$$

Объединяя оба рассмотренных случая, приходим к оценке (18).

Теорема 2. *Пусть функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, m)$, определены и выпуклы на E^n; множество U из* (1) *регулярно; $J(u)$, $g_i(u) \in C^{1,1}(U)$, $A_0 = \max\limits_{1 \leqslant i \leqslant m} \sup\limits_U |g_i'(u)| < \infty$. Пусть задача* (1) *имеет решение, т. е. $J_* > -\infty$, $U_* \neq \varnothing$, и начальная точка $u_0 \in U$ такова, что множество $M_\delta(u_0) = \{u\colon u \in U, J(u) \leqslant J(u_0) + \delta\}$ ограничено. Тогда при любом выборе $\varepsilon_0 > 0$ для последовательности $\{u_k\}$, определяемой условиями* (2), (4), (6), (13)—(15), *справедливы равенства*

$$\lim_{k\to\infty} J(u_k) = J_*, \qquad \lim_{k\to\infty} \rho(u_k, U_*) = 0. \tag{20}$$

Доказательство. Сначала установим, что

$$\lim_{k\to\infty} (J(u_k) - J(u_{k+1})) = 0. \tag{21}$$

Если $\sigma_k \leqslant -\varepsilon_k$, то из (14), (6) имеем $J(u_{k+1}) = f_k(\alpha_k) \leqslant f_k(0) + \delta_k = J(u_k) + \delta_k$. Если же $-\varepsilon_k < \sigma_k \leqslant 0$, то из (15) следует $J(u_{k+1}) = J(u_k) \leqslant J(u_k) + \delta_k$.

Таким образом, $J(u_k) \geqslant J_* > -\infty$ и, кроме того,

$$J(u_{k+1}) \leqslant J(u_k) + \delta_k, \quad k=0, 1, \ldots; \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty. \tag{22}$$

Согласно лемме 2.3.2 тогда существует $\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$. Отсюда следует равенство (21).

Далее, покажем, что $\lim\limits_{k\to\infty} \varepsilon_k = 0$. Согласно (14), (15) последовательность $\{\varepsilon_k\}$ получается дроблением и не возрастает. Допустим, что $\lim\limits_{k\to\infty} \varepsilon_k = \varepsilon > 0$. Это значит, что в процессе построения $\{u_k\}$ было конечное число дроблений и $\varepsilon_k = \varepsilon > 0$ при всех $k \geqslant k_0$. Из (14) тогда имеем $\sigma_k \leqslant -\varepsilon_k = -\varepsilon$, т. е. $|\sigma_k| \geqslant \varepsilon$, $k \geqslant k_0$. В этом случае согласно лемме 2 $\beta_k \geqslant A_1\varepsilon$, $k \geqslant k_0$. Поэтому из леммы 3 получим $J(u_k) - J(u_{k+1}) \geqslant A_2 \min\{A_1\varepsilon^2; \varepsilon^2\} - \delta_k$ $(k \geqslant k_0)$, что противоречит равенству (21). Итак, показано, что $\lim\limits_{k\to\infty} \varepsilon_k = 0$.

Пусть $k_1 < k_2 < \ldots < k_r < \ldots$ — номера тех итераций, когда происходит дробление ε_k. Согласно (14), (15) тогда $-\varepsilon_{k_r} \leqslant \sigma_{k_r} \leqslant 0$ $(r = 1, 2, \ldots)$. Следовательно, $\lim\limits_{r\to\infty} \sigma_{k_r} = 0$. Тем самым установлено, что существует хотя бы одна подпоследовательность $\{\sigma_{k_r}\}$, сходящаяся к нулю.

Возьмем произвольную подпоследовательность $\{\sigma_{k_r}\} \to 0$. Покажем, что тогда любая предельная точка соответствующей подпоследовательности $\{u_{k_r}\}$ принадлежит множеству U_*. Из (22) следует, что $J(u_{k+1}) \leqslant J(u_0) + \delta$ $(k = 0, 1, \ldots)$, т. е. $\{u_k\} \in M_\delta(u_0)$. По условию множество $M_\delta(u_0)$ ограничено. Поэтому можем считать, что взятая выше подпоследовательность $\{u_{k_r}\}$ сходится к некоторой точке u_*. Далее, множество номеров I_k, определяемое согласно (13), представляет собой подмножество конечного числа номеров $\{1, 2, \ldots, m\}$, поэтому число различных множеств I_k конечно. Это значит, что среди $\{I_{k_r}, r = 1, 2, \ldots\}$ найдется хотя бы одно множество $I_{k_r} = I$, которое повторяется бесконечно много раз. Выбирая при необходимости подпоследовательности, можем, таким образом, считать, что

$$\{\sigma_{k_r}\} \to 0, \quad \{u_{k_r}\} \to u_*, \quad I_{k_r} = I, \quad r = 1, 2, \ldots$$

Согласно лемме 1 при $G(u_{k_r}) = W_{k_r}$ имеем

$$|\sigma_{k_r} - \sigma(u_*)| = |\sigma(u_{k_r}) - \sigma(u_*)| \leqslant L\sqrt{n}\,|u_{k_r} - u_*| \to 0,$$

т. е. $\lim\limits_{r\to\infty} \sigma(u_{k_r}) = \lim\limits_{r\to\infty} \sigma_{k_r} = 0 = \sigma(u_*)$, где $\sigma(u_*) = \inf\limits_{G(u_*)} \sigma$, $G(u_*) = \{(e, \sigma) \in E^{n+1}: \langle J'(u_*), e\rangle \leqslant \sigma, \langle g_i'(u_*), e\rangle \leqslant \sigma, i \in I; |e^j| \leqslant 1, j = 1, \ldots, n\}$.

Рассмотрим задачу (7), соответствующую точке $u_* = \lim\limits_{r\to\infty} u_{k_r}$. Покажем, что $W_* \subseteq G(u_*)$. По определению множеств $I = I_{k_r} = \{i: 1 \leqslant i \leqslant r, -\varepsilon_{k_r} \leqslant g_i(u_{k_r}) \leqslant 0\}$ $(r = 1, 2, \ldots)$. Отсюда при $r \to \infty$ получим $g_i(u_*) = 0$ для всех $i \in I$. Это значит, что $I \subseteq I_*$, т. е. в определении множества W_* число ограничений типа неравенств не меньше, чем число таких ограничений в определении $G(u_*)$. Тем самым установлено, что $W_* \subseteq G(u_*)$. А то-

гда, замечая, что одна и та же функция на более широком множестве имеет меньшую нижнюю грань, получаем $\sigma_* = \inf\limits_{W_*} \sigma \geqslant \inf\limits_{G(u_*)} \sigma = \sigma(u_*) = 0$. С другой стороны, $(0, 0) \in W_*$, поэтому $\sigma_* \leqslant 0$. Следовательно, $\sigma_* = 0$. Поскольку задача (1) по условию выпукла и множество U регулярно, то согласно теореме 1 $u_* \in U_*$.

Выше было доказано существование предела $\lim\limits_{k\to\infty} J(u_k)$. Теперь можем сказать, чему равен этот предел: $\lim\limits_{k\to\infty} J(u_k) = \lim\limits_{r\to\infty} J(u_{k_r}) = J(u_*) = J_*$.

Таким образом, построенная последовательность $\{u_k\}$ минимизирует функцию $J(u)$ на множестве U. Поскольку $\{u_k\} \in M_\delta(u_0)$ — ограниченное множество, то из теоремы 2.1.2 следует, что $\lim\limits_{k\to\infty} \rho(u_k, U_*) = 0$. Равенства (20) и, тем самым, теорема 2 доказаны.

4. Для задачи

$$J(u) \to \inf; \quad u \in U = \{u \in E^n\colon g_i(u) \leqslant 0, \quad i = 1, \ldots, m,$$
$$g_i(u) = \langle a_i, u\rangle - b^i = 0, \quad i = m+1, \ldots, s\},$$

содержащей линейные ограничения типа равенств, метод возможных направлений описывается так же, как выше, лишь в задаче (2) нужно добавить еще ограничения $\langle a_i, e\rangle = 0$ $(i = m+1, \ldots, s)$.

Можно заметить, что описанный в гл. 3 симплекс-метод для решения канонической задачи линейного программирования по существу является вариантом метода возможных направлений. Более того, опираясь на идеи метода возможных направлений, можно получить симплекс-метод непосредственно для основной задачи линейного программирования (без ее сведения к канонической задаче).

Выше во вспомогательной задаче (2) было принято условие нормировки $|e^j| \leqslant 1$ $(j = 1, \ldots, n)$. Возможны и другие условия нормировки, например, $|e|^2 \leqslant 1$ или $|B_k e| \leqslant 1$, где B_k — специально выбираемая матрица. Заметим, что при такой нормировке задача (2) уже не будет задачей линейного программирования. Тем не менее удачный выбор B_k может облегчить выбор возможного направления убывания, ускорить сходимость метода. О других способах нормировки, о сходимости различных вариантов метода возможных направлений и других аспектах этого метода см., например, [11, 159, 163, 338].

Упражнения. 1. Сделать несколько итераций метода возможных направлений для задачи минимизации $J(u) = x + y$ на множестве $U = \{u = (x, y)\colon g_1(u) = x^2 - y \leqslant 0,\ g_2(u) = y - 1 \leqslant 0\}$ при различном выборе начальной точки u_0.

2. Вычислить несколько приближений по методу возможных направлений для задачи из примера 1 при различном начальном приближении u_0.

§ 6. Метод линеаризации

Этот метод на каждой итерации использует линейные аппроксимации минимизируемой функции и функций, задающих ограничения. Опишем его для задачи

$$J(u) \to \inf, \quad u \in U = \{u \in U_0\colon g_1(u) \leqslant 0, \ \ldots, \ g_m(u) \leqslant 0\}, \qquad (1)$$

предполагая, что U_0 — выпуклое замкнутое множество из E^n и функции $J(u)$, $g_i(u) \in C^1(U_0)$. Пусть u_0 — начальное приближение, $u_0 \in U_0$. Предположим, что k-е приближение $u_k \in U_0$ при

некотором $k \geqslant 0$ уже известно. Введем функцию

$$\Phi_k(u) = \frac{1}{2}|u - u_k|^2 + \beta_k \langle J'(u_k),\ u - u_k\rangle, \quad \beta_k > 0, \tag{2}$$

и множество

$$W_k = \{u \in U_0:\ g_i(u_k) + \langle g_i'(u_k), u - u_k\rangle \leqslant 0,\ i = 1, \ldots, m\}. \tag{3}$$

Пусть $W_k \neq \varnothing$. В качестве $k+1$-го приближения u_{k+1} возьмем решение следующей задачи минимизации:

$$\Phi_k(u) \to \inf, \quad u \in W_k. \tag{4}$$

Поскольку функция (2) сильно выпукла, множество (3) выпукло и замкнуто, то согласно теореме 4.3.1 задача (4) имеет, притом единственное, решение. Задачу (4) необязательно решать точно: достаточно найти точку u_{k+1} из условий

$$u_{k+1} \in W_k:\ \Phi_k(u_{k+1}) \leqslant \inf_{W_k} \Phi_k(u) + \varepsilon_k, \quad \varepsilon_k \geqslant 0. \tag{5}$$

Если U_0 многогранное множество, то задача (4) представляет собой задачу квадратичного программирования и может быть решена конечношаговым методом (см. ниже § 7). Если W_k — ограниченное множество, то для решения задачи (4) может быть использован, например, метод условного градиента, который будет сходиться и при $\varepsilon_k > 0$ позволит определить точку u_{k+1} из (5) за конечное число шагов. В общем случае задача (5), конечно, не всегда просто решается. Метод линеаризации (5) обычно используют лишь в тех случаях, когда определение точки u_{k+1} из (5) не требует большого объема вычислений. Полезно заметить, что задача (4) равносильна задаче

$$\varphi_k(u) = \frac{1}{2}|u - (u_k - \beta_k J'(u_k))|^2 \to \inf, \quad u \in W_k,$$

так как $\varphi_k(u) - \Phi_k(u) = \beta_k^2 |J'(u_k)|^2 = \text{const}$, $u \in E^n$. Это значит, что точное решение v_k задачи (4) представляет собой проекцию точки $u_k - \beta_k J'(u_k)$ на множество W_k, а точка u_{k+1} из (5) является приближением для v_k. Отсюда следует, что если в (1) ограничения $g_i(u) \leqslant 0$ отсутствуют ($m = 0$), то $U = U_0 = W_k$ и метод линеаризации превратится в метод проекции градиента.

Теорема 1. *Пусть U_0 — выпуклое замкнутое множество из E^n (в частности, возможно $U_0 = E^n$); функции $J(u)$, $g_i(u) \in C^1(U_0)$, выпуклы на U_0 и*

$$\max\left\{|J'(u) - J'(v)|;\ \max_{1 \leqslant i \leqslant m} |g_i'(u) - g_i'(v)|\right\} \leqslant L|u - v| \quad \forall u, v \in U_0;$$

существует такая точка $\bar{u} \in U$, что

$$g_1(\bar{u}) < 0, \ \ldots, \ g_m(\bar{u}) < 0; \tag{6}$$

$J_* > -\infty$, $U_* \neq \varnothing$; *числа* β_k, ε_k *в* (2), (5) *таковы, что*

$$\varepsilon_k \geqslant 0, \quad \sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty, \quad 0 < \gamma_0 \leqslant \beta_k \leqslant \beta, \tag{7}$$

где β *определяется ниже формулой* (22). *Тогда множество* (3) *непусто при всех* $k \geqslant 0$, *последовательность* $\{u_k\}$, *определяемая методом* (5), *сходится к некоторой точке* $v_* \in U_*$.

Доказательство. Согласно теореме 4.2.2

$$g_i(u_k) + \langle g_i'(u_k), u - u_k \rangle \leqslant g_i(u) \qquad \forall u \in U_0. \tag{8}$$

Отсюда следует, что если $u \in U$, то $u \in W_k$, так что $U \subset W_k$. По условию $U \neq \varnothing$, поэтому $W_k \neq \varnothing$ $(k = 0, 1, \ldots)$. Таким образом, при каждом $k \geqslant 0$, $\varepsilon_k \geqslant 0$ существует точка u_{k+1}, удовлетворяющая условиям (5): например, можно взять $u_{k+1} = v_k$, где v_k — точное решение задачи (4). Применяя теорему 4.3.1 к задаче (4) с учетом (5) имеем $|u_{k+1} - v_k|^2/2 \leqslant \Phi_k(u_{k+1}) - \Phi_k(v_k) \leqslant \varepsilon_k$, так что

$$|u_{k+1} - v_k| \leqslant \sqrt{2\varepsilon_k}. \tag{9}$$

Возьмем произвольную точку $u_* \in U_*$. При сделанных предположениях о выпуклости и регулярности задачи (1) по теореме 4.9.2 и лемме 4.9.2 найдутся такие числа $\lambda_1^* \geqslant 0, \ldots, \lambda_m^* \geqslant 0$, что

$$\left\langle J'(u_*) + \sum_{i=1}^{m} \lambda_i^* g_i'(u_*), u - u_* \right\rangle \geqslant 0 \qquad \forall u \in U_0, \tag{10}$$

$$\lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, m, \quad u_* \in U_*. \tag{11}$$

Подчеркнем, что в силу замечания, сделанного после формулировки следствия 2 к теореме 4.9.6, числа $\lambda_1^*, \ldots, \lambda_m^*$ в (10), (11) могут быть выбраны одни и те же для всех $u_* \in U_*$.

Далее, из условия регулярности (6) и неравенств (8) следует

$$g_i(u_k) + \langle g_i'(u_k), \bar{u} - u_k \rangle \leqslant g_i(\bar{u}) < 0, \qquad i = 1, \ldots, m. \tag{12}$$

Это значит, что множество (3) также регулярно и к задаче (4) также применима теорема 4.9.2, из которой следует, что функция Лангража этой задачи $L_1(u, \xi) = \frac{1}{2}|u - u_k|^2 + \beta_k \langle J'(u_k), u - u_k \rangle + \sum_{i=1}^{m} \xi_i \big(g_i(u_k) + \langle g_i'(u_k), u - u_k \rangle\big)$, $u \in U_0$, $\xi = (\xi_1, \ldots, \xi_m) \in \Lambda_0 = E_+^m$, имеет седловую точку (v_k, ξ^k), v_k — решение задачи (4), $\xi^k = (\xi_{1k}, \ldots, \xi_{mk}) \in E_+^m$. В силу леммы 4.9.2

$$\left\langle v_k - u_k + \beta_k J'(u_k) + \sum_{i=1}^{m} \xi_{ik} g_i'(u_k), u - v_k \right\rangle \geqslant 0 \qquad \forall u \in U_0, \tag{13}$$

$$\xi_{ik}\big(g_i(u_k) + \langle g_i'(u_k), v_k - u_k \rangle\big) = 0, \qquad i = 1, \ldots, m, \tag{14}$$

$$g_i(u_k) + \langle g_i'(u_k), v_k - u_k \rangle \leqslant 0, \qquad i = 1, \ldots, m. \tag{15}$$

Возьмем в (10) $u = v_k$, умножим на $\beta_k > 0$ и сложим с (13) при $u = u_*$.

Получим

$$0 \leqslant \langle v_k - u_k, u_* - v_k \rangle + \beta_k \langle J'(u_*) - J'(u_k), v_k - u_* \rangle + \\ + \beta_k \sum_{i=1}^{m} \lambda_i^* \langle g_i'(u_*), v_k - u_* \rangle + \sum_{i=1}^{m} \xi_{ik}^* \langle g_i'(u_k), u_* - v_k \rangle. \quad (16)$$

Преобразуем и оценим каждое слагаемое в правой части (16). Для первого слагаемого имеем

$$\langle v_k - u_k, u_* - v_k \rangle = \frac{1}{2} | u_k - u_* |^2 - \frac{1}{2} | v_k - u_k |^2 - \frac{1}{2} | u_* - v_k |^2. \quad (17)$$

Пользуясь неравенством (4.2.20) при $u = u_k$, $w = v_k$, $v = u_*$, получаем оценку для второго слагаемого

$$\beta_k \langle J'(u_*) - J'(u_k), v_k - u_* \rangle \leqslant \beta_k L | u_k - v_k |^2/4. \quad (18)$$

Далее, из леммы 2.3.1 при $J(u) = g_i(u)$, $u = v_k$, $v = u_k$ с учетом неравенств (15) имеем $g_i(v_k) \leqslant g_i(u_k) + \langle g_i'(u_k), v_k - u_k \rangle + L | u_k - v_k |^2/2 \leqslant \leqslant L | u_k - v_k |^2/2$ $(i = 1, \ldots, m)$. Отсюда для третьего слагаемого из (16) с помощью равенств (11), неравенств (8) при $u = v_k$ и $\lambda_k^* \geqslant 0$ получаем

$$\beta_k \sum_{i=1}^{m} \lambda_i^* \langle g_i'(u_*), v_k - u_* \rangle = \beta_k \sum_{i=1}^{m} \lambda_i^* \left(g_i(u_*) + \langle g_i'(u_*), v_k - u_* \rangle \right) \leqslant \\ \leqslant \beta_k \sum_{i=1}^{m} \lambda_i^* g_i(v_k) \leqslant \beta_k | \lambda^* |_1 L | u_k - v_k |^2/2, \quad | \lambda^* |_1 = \sum_{i=1}^{m} | \lambda_i^* |. \quad (19)$$

Наконец, для четвертого слагаемого из (16) с учетом равенств (14), неравенств (8) при $u = u_*$, включений $u_* \in U$, $\xi^k \in E_+^m$ имеем

$$\sum_{i=1}^{m} \xi_{ik} \langle g_i'(u_*), u_k - v_k \rangle = \sum_{i=1}^{m} \xi_{ik} \left(g_i(u_k) + \langle g_i'(u_k), u_* - u_k \rangle \right) - \\ - \xi_{ik} \left(g_i(u_k) + \langle g_i'(u_k), v_k - u_k \rangle \right) \leqslant \sum_{i=1}^{m} \xi_{ik} g_i(u_*) \leqslant 0. \quad (20)$$

Сложим оценки (17)—(20); с помощью (16) получим

$$0 \leqslant \frac{1}{2} | u_k - u_* |^2 - \frac{1}{2} | v_k - u_* |^2 - \frac{1}{2} | v_k - u_k |^2 \times \\ \times \left(1 - \frac{1}{2} L \beta_k - L | \lambda^* |_1 \beta_k \right). \quad (21)$$

Выберем β_k из условий

$$0 < \gamma_0 \leqslant \beta_k \leqslant \frac{2(1-\gamma)}{L(1 + 2|\lambda^*|_1)} = \beta, \quad (22)$$

где γ_0, γ — настолько малые положительные числа, что $\gamma_0 \leqslant \leqslant 2(1-\gamma)/(L(1 + 2|\lambda^*|_1))$. Из (21), (22) тогда имеем

$$| v_k - u_* |^2 + \gamma | v_k - u_k |^2 \leqslant | u_k - u_* |^2, \qquad k = 0, 1, \ldots \quad (23)$$

Из неравенств (7), (9), (23) и леммы 2.3.10 следует существование конеч-

ных пределов

$$\lim_{k\to\infty}|u_k-u_*|=\lim_{k\to\infty}|v_k-u_*|,\quad \lim_{k\to\infty}|u_k-v_k|=0. \tag{24}$$

Это значит, что последовательности $\{u_k\}$, $\{v_k\}$ ограничены. Покажем ограниченность последовательности $\{\xi^k\}$ из (13)—(15). С помощью (12), (14) из (13) при $u=\bar u$ имеем

$$\langle v_k-u_k+\beta_k J'(u_k),\ \bar u-v_k\rangle\geqslant-\sum_{i=1}^m \xi_{ik}\langle g_i'(u_k),\ \bar u-v_k\rangle=$$

$$=\sum_{i=1}^m\left[\xi_{ik}(-g_i(u_k)-\langle g_i'(u_k),\ \bar u-u_k\rangle)+\xi_{ik}(g_i(u_k)+\langle g_i'(u_k),\ v_k-u_k\rangle)\right]\geqslant$$

$$\geqslant\sum_{i=1}^m\xi_{ik}(-g_i(\bar u))\geqslant\xi_{jk}\min_{1\leqslant i\leqslant m}|g_i(\bar u)|,\qquad j=1,\ldots,m.$$

Отсюда и из неравенств (22), ограниченности $\{u_k\}$ и $\{v_k\}$ получаем

$$0\leqslant\xi_{jk}\leqslant\frac{1}{\min\limits_{1\leqslant i\leqslant m}|g_i(\bar u)|}[\langle v_k-u_k+\beta_k J'(u_k),\ \bar u-v_k\rangle]\leqslant\mathrm{const}<\infty,$$

$$j=1,\ldots,m.$$

Таким образом, последовательность $\{\xi^k\}$ ограничена. Отсюда и из (22) следует ограниченность $\{\xi^k/\beta_k\}$. Перепишем (13), (14) в виде

$$\left\langle\frac{v_k-u_k}{\beta_k}+J'(u_k)+\sum_{i=1}^m\frac{\xi_{ik}}{\beta_k}g_i'(u_k),\ u-v_k\right\rangle\geqslant0\qquad\forall u\in U_0,$$
$$\frac{\xi_{ik}}{\beta_k}(g_i(u_k)+\langle g_i'(u_k),\ v_k-u_k\rangle)=0,\qquad i=1,\ldots,m. \tag{25}$$

Выбирая при необходимости подпоследовательности из ограниченных последовательностей $\{u_k\}$, $\{v_k\}$, $\{\xi^k/\beta_k\}$, можем считать, что эти последовательности сходятся. С учетом (24) тогда

$$\lim_{k\to\infty}u_k=\lim_{k\to\infty}v_k=v_*,\qquad\lim_{k\to\infty}\xi^k/\beta_k=\mu_i^*\geqslant0,\qquad i=1,\ldots,m.$$

Из замкнутости U_0 следует, что $v_*\in U_0$, а из (15) при $k\to\infty$ получим $g_i(u_*)\leqslant0$ $(i=1,\ldots,m)$. Следовательно, $v_*\in U$. Далее, из (25) при $k\to\infty$ с учетом (22) имеем

$$\left\langle J'(v_*)+\sum_{i=1}^m\mu_i^*g_i'(v_*),\ u-v_*\right\rangle\geqslant0\qquad\forall u\in U_0,$$
$$\mu_i^*g_i(v_*)=0,\qquad i=1,\ldots,m. \tag{26}$$

Как следует из леммы 4.9.2 и теоремы 4.9.2, соотношения (26) означают, что $v_*\in U_*$. Вспомним, что неравенство (23) было получено для любых $u_*\in U_*$. В частности, (23) верно и при $u_*=v_*$. Но v_* — предельная точка последовательности $\{u_k\}$. Согласно лемме 2.3.10 тогда $\{u_k\}\to v_*$. Теорема доказана.

Замечание 1. Если в (5) $\varepsilon_k=0$, то согласно (9) $u_{k+1}=v_k$ $(k=0,1,\ldots)$, и неравенство (23) можно записать в виде

$$|u_{k+1}-u_*|^2+\gamma|u_{k+1}-u_k|^2\leqslant|u_k-u_*|^2,\qquad k=0,1,\ldots,\qquad\forall u_*\in U_*.$$

Пользуясь произволом в выборе $u_* \in U_*$, отсюда имеем

$$|u_k - v_*| \geqslant |u_{k+1} - v_*|, \quad \rho(u_k, U_*) \geqslant \rho(u_{k+1}, U_*), \qquad k = 0, 1, \ldots,$$

причем равенство здесь возможно лишь при $u_{k+1} = u_k = v_k \in U_*$. Таким образом, при точной реализации описанного метода линеаризации расстояние от точки u_k до множества U_* или до точки v_* монотонно убывает. В то же время можно отметить, что хотя и $\{J(u_k)\} \to J(v_*) = J_*$, но $\{J(u_k)\}$ не обязательно монотонно убывает и не обязательно $u_k \in U$.

Различные варианты метода линеаризации описаны и исследованы в [8, 19, 21, 29, 115, 132, 150, 250, 314, 338]; регуляризованные формы метода линеаризации для задач с неточно заданными исходными данными исследованы в [86, 329].

Упражнения. 1. Доказать, что если в (4) окажется $v_k = u_k$ при некотором $k \geqslant 0$, то точка u_k удовлетворяет необходимым условиям оптимальности. Указание: применить теорему 4.8.1 к задаче (4), затем принять $u_k = v_k$.

2. Доказать, что если выполнены условия теоремы 1, $\varepsilon_k = 0$, $k = 0, 1, \ldots$, и в (4) $v_k = u_k$ при некотором $k \geqslant 0$, то $u_k \in U_*$. Указание: положить в (25), (15) $v_k = u_k$ и воспользоваться леммой 4.9.2 и теоремой 4.9.2.

3. Рассмотреть метод линеаризации для задачи (1) при $U_0 = E^n_+$, $m = 0$.

4. Описать метод линеаризации для задачи (1) с дополнительными линейными ограничениями $\langle a_i, u\rangle = b^i$ $(i = m+1, \ldots, s)$.

§ 7. Квадратичное программирование

1. Рассмотрим задачу

$$J(u) = \frac{1}{2}\langle Cu, u\rangle + \langle c, u\rangle \to \inf, \quad u \in U, \tag{1}$$

$$U = \{u \in E^n\colon \langle a_i, u\rangle \leqslant b^i, \quad i = 1, \ldots, m;$$
$$\langle a_i, u\rangle = b^i, \quad i = m+1, \ldots, s\}, \tag{2}$$

где C — симметричная неотрицательно определенная матрица размера $n \times n$, т. е. $C \geqslant 0$; c, $a_i \in E^n$, $b^i \in \mathbf{R}$ $(i = 1, \ldots, s)$ (возможности $m = 0$, или $s = m$, или $s = m = 0$ не исключаются). Задачу (1), (2) принято называть *задачей квадратичного программирования*: в ней квадратичная выпуклая функция минимизируется на многогранном множестве. Такие задачи возникают в различных приложениях. Задачи определения расстояния от точки до многогранного множества, проектирования на такое множество также представляют примеры задачи квадратичного программирования, когда в (1) $C = I$ — единичная матрица. Задачи вида (1), (2) часто возникают как вспомогательные при описании различных методов минимизации (см., например, § 6). Поэтому важно иметь достаточно простые методы решения задачи квадратичного программирования. Оказывается, для задачи (1), (2), как и для задачи линейного программирования, существуют конечные (конечношаговые) методы их решения. Для построения таких методов сначала нужно выявить некоторые специфические особенности этой задачи. В частности, здесь полезно рассмотреть двойственную к (1), (2) задачу.

Введем функцию Лагранжа задачи (1), (2):

$$L(u, \lambda) = \frac{1}{2}\langle Cu, u\rangle + \langle c, u\rangle + \langle \lambda, Au - b\rangle = \frac{1}{2}\langle Cu, u\rangle + \langle c + A^T\lambda, u\rangle - \langle \lambda, b\rangle, \quad u \in U_0 = E^n, \ \lambda \in \Lambda_0 = \{\lambda = (\lambda_1, \ldots, \lambda_s) \in E^s: \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\},$$

где A — матрица размера $s \times n$ со строками $a_1, \ldots, a_s$, $b = (b^1, \ldots, b^s)$. Если $J_* > -\infty$, $U_* \neq \varnothing$, то согласно теореме 4.9.4 функция $L(u, \lambda)$ имеет седловую точку (u_*, λ^*), причем в силу леммы 4.9.2

$$\frac{\partial L(u_*, \lambda^*)}{\partial u} = Cu_* + c + A^T\lambda^* = 0, \tag{3}$$

$$\lambda_i^* (\langle a_i, u_*\rangle - b^i) = 0, \quad i = 1, \ldots, s; \quad u_* \in U_*, \quad \lambda^* \in \Lambda_0. \tag{4}$$

Тогда двойственная к (1), (2) задача

$$\psi(\lambda) = \inf_{u \in E^n} L(u, \lambda) \to \sup, \quad \lambda \in \Lambda_0, \tag{5}$$

согласно теореме 4.9.6 также имеет решение, причем

$$J_* = \psi^* = \sup_{\Lambda_0} \psi(\lambda) = \psi(\lambda^*) = J(u_*), \quad u_* \in U_*, \quad \lambda^* \in \Lambda^* = \{\lambda \in \Lambda_0: \psi(\lambda) = \psi^*\}.$$

При дополнительном предположении положительной определенности матрицы C, т. е. $C > 0$, функция $\psi(\lambda)$ может быть выписана явно. В самом деле, тогда C невырождена и точка минимума $u = u(\lambda)$ функции $L(u, \lambda)$ при $u \in E^n$ однозначно определяется из системы $Cu + c + A^T\lambda = 0$, так что $u(\lambda) = -C^{-1}(c + A^T\lambda)$. Поэтому

$$\psi(\lambda) = L(u(\lambda), \lambda) = -\frac{1}{2}\langle c + A^T\lambda, C^{-1}(c + A^T\lambda)\rangle - \langle \lambda, b\rangle = -\frac{1}{2}\langle (AC^{-1}A^T)\lambda, \lambda\rangle - \langle AC^{-1}c + b, \lambda\rangle - \frac{1}{2}\langle C^{-1}c, c\rangle, \quad \lambda \in \Lambda_0,$$

где $AC^{-1}A^T \geqslant 0$. Таким образом, при $C > 0$ двойственная задача (5), записанная в виде

$$-\psi(\lambda) \to \inf, \quad \lambda \in \Lambda_0, \tag{6}$$

также является задачей квадратичного программирования вида (1), (2), но множество Λ_0 по сравнению с (2) имеет более простую структуру. Зная какое-либо решение λ^* задачи (6), можно записать решение исходной задачи (1), (2) в виде

$$u_* = -C^{-1}(c + A^T\lambda^*). \tag{7}$$

В самом деле, при $C > 0$ функция $J(u)$ сильно выпукла и согласно теореме 4.3.1 задача (1), (2) имеет единственное решение

u_*, которое обязательно будет решением системы (3), (4), где λ^* — решение задачи (6). И поскольку система (3) при фиксированном λ^* однозначно определяет точку u_*, то необходимо приходим к формуле (7).

Особенно проста задача (6) в том случае, когда в исходной задаче (1), (2) отсутствуют ограничения типа неравенств ($m=0$) и множество U имеет вид

$$U=\{u\in E^n\colon \langle a_i, u\rangle=b^i, \quad i=1, \ldots, s\}. \tag{8}$$

Тогда $\Lambda_0=E^s$, и задача (6) запишется в форме

$$-\psi(\lambda)\to\inf, \quad \lambda\in E^s. \tag{9}$$

Множество Λ^* решений задачи (9) в силу теоремы 4.2.3 совпадает со множеством решений системы

$$-\psi'(\lambda^*)=AC^{-1}A^T\lambda^*+AC^{-1}c+b=0. \tag{10}$$

В общем случае система (10) может иметь более одного решения. Если матрица A невырожденная, т. е. векторы $a_1, \ldots, a_s$ в (8) линейно независимы, то из $C>0$ следует $AC^{-1}A^T>0$, и тогда задача (9) и, следовательно, система (10) будут иметь единственное решение. Таким образом, при $C>0$ для решения задачи (1), (8) достаточно решить две системы линейных алгебраических уравнений (10), (3). Здесь могут быть использованы известные методы линейной алгебры [4, 39, 54, 93, 164]. Поскольку для линейных систем имеется принципиальная возможность получить решение за конечное число арифметических операций (например, методом исключения Гаусса), то такая возможность имеется и для задачи (1), (8) при $C>0$.

2. Следуя [21], покажем, что исходная задача (1), (2) при $C>0$ может быть сведена к решению конечного числа задач вида (1), (8). Здесь важную роль играет понятие особой точки задачи (1), (2).

Определение 1. Точка v называется *особой точкой задачи* (1), (2), если $v\in U$ и v является решением задачи

$$J(u)\to\inf, \quad u\in V=\{u\in E^n\colon \langle a_i, u\rangle=b^i, \ i\in I\cup\{m+1, \ldots, s\}\}, \tag{11}$$

где I — какое-либо подмножество индексов $\{1, \ldots, m\}$ (возможность $I=\varnothing$ не исключается).

Лемма 1. *Пусть в задаче* (1), (2) $C\geqslant 0$, $J_*>-\infty$, $U_*\neq\varnothing$. *Тогда каждое решение u_* задачи* (1), (2) *является особой точкой этой задачи.*

Доказательство. Положим $I(u_*)=\{i\colon 1\leqslant i\leqslant m, \langle a_i, u_*\rangle=b^i\}$ и рассмотрим задачу (11) с $I=I(u_*)$. Заметим, что $u_*\in V$. Так как $\langle a_i, u_*\rangle<b^i$ при $i\notin I(u_*)$ $(1\leqslant i\leqslant m)$ и функция $\langle a_i, u\rangle$ непрерывна, то существует такая окрестность $S(u_*, \varepsilon)=\{u\in E^n\colon$

$|u-u_*|<\varepsilon\}$ $(\varepsilon>0)$ точки u_*, что $\langle a_i, u\rangle<b^i$ при всех $i\notin I(u_*)$ $(1\leqslant i\leqslant m)$ и $u\in S(u_*, \varepsilon)$. Это значит, что $V\cap S(u_*, \varepsilon)\subset U$. Тогда $J(u)\geqslant J(u_*)=J_*$ при всех $u\in V\cap S(u_*, \varepsilon)$. Таким образом, u_* — точка локального минимума выпуклой задачи (11) с $I=I(u_*)$. По теореме 4.2.1 тогда u_* является точкой глобального минимума функции $J(u)$ на множестве V. Следовательно, u_* — решение задачи (11) с $I=I(u_*)$, так что u_* — особая точка задачи (1), (2).

Теорема 1. *Пусть $C>0$, множество* (2) *непусто. Тогда существует конечный метод решения задачи* (1), (2).

Доказательство. Так как множество $\{1, \ldots, m\}$ имеет конечное число подмножеств I, а всякая задача (11) при $C>0$ имеет одно решение (теорема 4.3.1), то и число особых точек задачи (1), (2) конечно. Согласно лемме 1 для отыскания решения задачи (1), (2) достаточно перебрать все ее особые точки и найти ту из них, в которой функция (1) принимает меньшее значение. Так как задача (11) имеет вид (1), (8), то каждая особая точка может быть найдена за конечное число арифметических операций. Таким образом, поиск решения задачи (1), (2) закончится за конечное число шагов.

3. Установленная в теореме 1 принципиальная возможность получения решения задачи (1), (2) за конечное число шагов имеет лишь теоретический интерес. Дело в том, что здесь мы не учли возможную неустойчивость систем (3), (10) по отношению к погрешности задания исходных данных, к погрешности округления при выполнении арифметических операций. Кроме того, полный перебор особых точек задачи (1), (2) на практике требует слишком большого объема вычислений уже при не очень больших значениях n, s. Опишем один из методов упорядоченного перебора особых точек задачи (1), (2), более экономичного по сравнению с полным перебором [21]. При описании этого метода можно выделить три этапа.

На 1-м начальном этапе определяется, будет ли множество (2) непустым, и если $U\neq\varnothing$, то находится какая-либо $v\in U$. Здесь может быть использован, например, симплекс-метод, описанный в главе 3.

2-й этап состоит в переходе от какой-либо точки $v\in U$ к особой точке $w\in U$ со значением $J(w)\leqslant J(v)$. Для построения такой точки w можно воспользоваться следующим итерационным процессом. В качестве начального приближения берется $u_0=v$. Пусть известно k-е приближение $u_k\in U$, $J(u_k)\leqslant J(v)$. Определим вспомогательное приближение $\bar{u}_k$ как решение задачи (11) при

$$I=I(u_k)=\{i\colon\ 1\leqslant i\leqslant m,\ \langle a_i, u_k\rangle=b^i\}.$$

Поскольку $u_k\in V$, то $J(\bar{u}_k)\leqslant J(u_k)\leqslant J(v)$. Поэтому если $\bar{u}_k\in U$,

то в качестве требуемой особой точки можем взять $w=\bar{u}_k$. Допустим, что $\bar{u}_k \notin U$. Тогда $\langle a_j, \bar{u}_k\rangle > b^j$ хотя бы для одного $j \notin I(u_k)$ $(1 \leqslant j \leqslant m)$, так что множество индексов $I_1 = \{j\colon \langle a_j, \bar{u}_k\rangle > b^j,\ j \notin I(u_k),\ 1 \leqslant j \leqslant m\} \neq \varnothing$. Положим

$$u_{k+1} = u_k + \alpha_k(\bar{u}_k - u_k),\quad \alpha_k = \min_{j \in I_1}\left[(b^j - \langle a_j, u_k\rangle)/\langle a_j, \bar{u}_k - u_k\rangle\right]. \tag{12}$$

Из определения I_1 и включения $u_k \in U$ следует, что

$$0 < \frac{b^j - \langle a_j, u_k\rangle}{\langle a_j, \bar{u}_k - u_k\rangle} = \frac{b^j - \langle a_j, u_k\rangle}{(b^j - \langle a_j, u_k\rangle) + (\langle a_j, \bar{u}_k\rangle - b^j)} < 1, \quad j \in I_1,$$

поэтому $0 < \alpha_k < 1$. Покажем, что $u_{k+1} \in U$. Из выпуклости множества V и из $u_k \in V$, $\bar{u}_k \in V$ следует, что $u_{k+1} \in V$, где V взято из (11) при $I = I(u_k)$. Остается доказать, что $\langle a_j, u_{k+1}\rangle \leqslant b^j$ при всех $j \notin I(u_k)$ $(1 \leqslant j \leqslant m)$. Если $j \in I_1$, то с учетом определения (12) величины α_k имеем

$$\langle a_j, u_{k+1}\rangle = \langle a_j, u_k\rangle + \alpha_k\langle a_j, \bar{u}_k - u_k\rangle \leqslant b^j, \quad j \in I_1. \tag{13}$$

Если $j \notin I_1 \cup I(u_k)$ $(1 \leqslant j \leqslant m)$, то $\langle a_j, u_{k+1}\rangle = \alpha_k\langle a_j, \bar{u}_k\rangle + (1-\alpha_k)\langle a_j, u_k\rangle \leqslant b^j$. Таким образом, показано, что $u_{k+1} \in U$. Далее, поскольку $J(\bar{u}_k) \leqslant J(u_k) \leqslant J(v)$ и функция $J(u)$ выпукла, то $J(u_{k+1}) \leqslant \alpha_k J(\bar{u}_k) + (1-\alpha_k)J(u_k) \leqslant J(v)$. Далее, из включения $u_{k+1} \in V$, где V взято из (11) при $I = I(u_k)$, следует, что $I(u_k) \subset I(u_{k+1}) = \{i\colon \langle a_i, u_{k+1}\rangle = b^i,\ 1 \leqslant i \leqslant m\}$. В то же время для тех $j_0 \in I_1$, для которых в (12) реализовывается минимальное значение при определении α_k, неравенство (13) превращается в равенство, так что $j_0 \in I(u_{k+1})$, но $j_0 \notin I(u_k)$. Следовательно, множество $I(u_{k+1})$ содержит по крайней мере на один элемент больше, чем $I(u_k)$. Таким образом, следующее приближение $u_{k+1} \in U$ со значением $J(u_{k+1}) \leqslant J(v)$ построено, причем множество $I(u_{k+1})$ существенно шире $I(u_k)$. Однако множества $I(u_k) \in \{1, \ldots, m\}$ не могут бесконечно расширяться, и поэтому описанный процесс закончится на какой-то k-й итерации, когда $\bar{u}_k \in U$, причем $w = \bar{u}_k$ — особая точка задачи (1), (2) со значением $J(w) \leqslant J(v)$. Поскольку решаемая на каждой итерации задача (11) с $I = I(u_k)$ имеет вид (1), (8) и для ее решения существует конечный метод, то и весь переход от точки v к точке w осуществим за конечное число шагов.

На 3-м этапе выясняется, не будет ли особая точка w, построенная на 2-м этапе, решением задачи (1), (2), и в том случае, если $w \notin U_*$, осуществляется переход к следующей точке $z \in U$, для которой $J(z) < J(w)$. Для этих целей достаточно совершить один шаг несколько модифицированного метода условного градиента, приняв в качестве начальной точку w, полученную на 2-м этапе. А именно, сначала можно решить следующую

задачу линейного программирования

$$\langle J'(w), e\rangle = \langle Cw + c, e\rangle \to \inf,\ e \in \mathscr{E} = \{e = (e^1, \ldots, e^n) \in$$
$$\in E^n\colon \langle a_i, e\rangle \leqslant 0,\ i \in I(w) = \{i\colon 1 \leqslant i \leqslant m,\ \langle a_i, w\rangle = b^i\},$$
$$\langle a_i, e\rangle = 0,\ i = m+1, \ldots, s,\ -1 \leqslant e^j \leqslant 1,\ j = 1, \ldots, n\}. \quad (14)$$

Пусть $e = e_*$ — решение задачи (14), которое может быть получено, например, симплекс-методом. Так как $e = 0 \in \mathscr{E}$, то $\beta = \langle J'(w), e_*\rangle = \min\limits_{e \in \mathscr{E}} \langle J'(w), e\rangle \leqslant \langle J'(w), 0\rangle = 0$. Поэтому имеются две возможности: либо $\beta = 0$, либо $\beta < 0$. Покажем, что в случае $\beta = 0$ точка w — решение задачи (1), (2). С этой целью возьмем произвольную точку $u \in U$, $u \neq w$, и положим $e = t(u - w)$, где $t > 0$ столь мало, что $|e^j| = t|u^j - w^j| \leqslant 1$ $(j = 1, \ldots, n)$. Если $i \in I(w)$, то $\langle a_i, e\rangle = \langle a_i, u - w\rangle t = (\langle a_i, u\rangle - b^i)t \leqslant 0$. Если $m + 1 \leqslant i \leqslant s$, то $\langle a_i, e\rangle = t(\langle a_i, u\rangle - b^i) = 0$. Следовательно, $e = t(u - w) \in \mathscr{E}$. Поэтому $\beta = 0 = \langle J'(w), e_*\rangle \leqslant \langle J'(w), e\rangle$. Пользуясь теоремой 4.2.2 тогда имеем $J(u) - J(w) \geqslant \langle J'(w), u - w\rangle = \langle J'(w), e\rangle t^{-1} \geqslant 0$ при любом $u \in U$. Это значит, что $w \in U_*$, т. е. задача (1), (2) решена.

Рассмотрим вторую возможность: $\beta < 0$. Тогда e_* — возможное направление убывания функции $J(u)$ в точке w. В самом деле, при достаточно малых $\alpha > 0$ с учетом того, что e_* — решение задачи (13), имеем $J(w + \alpha e_*) - J(w) = \langle J'(w), e_*\rangle\alpha + o(\alpha) = \alpha(\beta + o(\alpha)/\alpha) < 0$; если $i \in I(w)$ или $m + 1 \leqslant i \leqslant s$, то $\langle a_i, w + \alpha e_*\rangle = b^i + \alpha\langle a_i, e_*\rangle \leqslant b^i$; если $i \notin I(w)$ $(1 \leqslant i \leqslant m)$, то $\langle a_i, w\rangle < b^i$ и $\langle a_i, w + \alpha e_*\rangle < b^i$. Тогда в качестве искомой точки $z = U$, $J(z) < J(w)$, можно взять $z = w + \alpha_0 e_*$, где $\alpha_0 > 0$ — достаточно малое число, которое может быть найдено за конечное число шагов, например, перебором чисел $\alpha_0 = 2^{-p}$ $(p = 0, 1, \ldots)$. Описание 3-го этапа закончено.

Отправляясь от точки z, полученной на 3-м этапе, можно снова перейти ко 2-му этапу при $v = z$, затем к 3-му этапу и т. д. В результате будет построена последовательность особых точек, на которой функция $J(u)$ строго убывает. Так как при $C > 0$ число особых точек конечно, то на каком-то шаге процесс, состоящий в последовательном применении 2-го и 3-го этапов, закончится нахождением решения задачи (1), (2). Таким образом, описанный метод позволяет за конечное число шагов найти решение задачи (1), (2) при $C > 0$.

Существуют конечные методы решения задачи квадратичного программирования (1), (2) и при $C \geqslant 0$; об этих методах читатель сможет прочесть, например, в [7, 12, 13, 19, 21, 23, 30, 33, 41, 102, 111, 135, 159, 250, 256, 257, 261, 271, 314, 320, 330]. О задачах кубического программирования и, в общем случае, полиноминального программирования, когда минимизируемая функция является многочленом, см. [44, 52].

Упражнения. 1. Уточните описание каждого этапа приведенного выше метода для задачи (1), (2) при $C = I$ — единичная матрица, а также при $U = E^n_+$ или $U = \{u = (u^1, \ldots, u^n) \in E^n\colon \alpha_i \leqslant u^i \leqslant \beta_i,\ i = 1, \ldots, n\}$.

2. Примените описанный выше метод к задачам квадратичного программирования из упражнения 3 к § 4.9.

3. Пусть $U = \{u = (x, y) \in E^2\colon -1 \leqslant x, y \leqslant 1\}$ или $U = \{u = (x, y) \in E^2\colon 0 \leqslant x, y \leqslant 1\}$. Найдите особые точки задачи минимизации функций $J(u) = (x-a)^2 + (y-b)^2$, $J(u) = (ax + by)^2$ на этих множествах при различных a, b.

4. Доказать, что множество точек минимума квадратичной функции $J(u) = |Au - b|^2$ на E^n совпадает со множеством решений системы $A^TAu = A^Tb$ (см. пример 4.2.4).

5. Доказать, что квадратичная функция (1) либо достигает своей нижней грани на множестве (2), либо неограничена снизу.

§ 8. Метод сопряженных направлений

В описанных выше методах, использующих градиент функции, на каждой итерации учитывается информация лишь о линейной части приращения минимизируемой функции в окрестности полученной точки. С помощью этих методов точку минимума квадратичной функции удается найти, вообще говоря, лишь за бесконечное число итераций. Возникает вопрос: нельзя ли придумать метод, использующий лишь градиент функции, который позволяет найти точку минимума квадратичной функции на всем пространстве за конечное число шагов? Если бы такой метод существовал, то можно было бы ожидать, что он сходится к точке минимума гладких функций быстрее градиентного метода, поскольку в окрестности точки минимума гладкая функция достаточно хорошо аппроксимируется квадратичной функцией.

Оказывается, методы с упомянутыми свойствами существуют. Одним из таких методов является метод сопряженных направлений. Опишем один из вариантов этого метода.

1. Сначала рассмотрим квадратичную задачу:

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle \to \inf; \quad u \in U = E^n, \tag{1}$$

где A — симметричная положительная матрица, $b \in E^n$. Тогда, как было показано выше, справедливы формулы

$$J'(u) = Au - b, \quad J''(u) = A,$$

и, кроме того, $J(u)$ сильно выпукла на E^n и достигает своей нижней грани на E^n в единственной точке u_* такой, что

$$J'(u_*) = Au_* - b = 0 \quad \text{или} \quad u_* = A^{-1}b. \tag{2}$$

Возьмем произвольную начальную точку $u_0 \in E^n$ и вычислим $p_0 = J'(u_0)$. Если $J'(u_0) = 0$, то $u_0 = u$ — задача (1) решена. По-

этому пусть $J'(u_0) \neq 0$. Тогда положим

$$u_1 = u_0 - \alpha_0 p_0, \quad \alpha_0 \geqslant 0,$$

где величина α_0 определяется условием

$$f_0(\alpha_0) = \min_{\alpha \geqslant 0} f_0(\alpha), \quad f_0(\alpha) = J(u_0 - \alpha p_0).$$

Таким образом, первая итерация метода сопряженных направлений совпадает с итерацией метода скорейшего спуска. Заметим, что $f_0(\alpha)$ сильно выпукла, поэтому величина α_0 существует и определяется однозначно (см. ниже формулу (14)). Поскольку $f_0'(0) = -\langle J'(u_0), p_0\rangle = -|J'(u_0)|^2 < 0$, то $\alpha_0 > 0$. Следовательно,

$$f_0'(\alpha_0) = 0 = -\langle J'(u_0 - \alpha_0 p_0), p_0\rangle = -\langle J'(u_1), p_0\rangle = \\ = -\langle J'(u_1), J'(u_0)\rangle. \quad (3)$$

Можем считать, что $J'(u_1) \neq 0$, иначе $u_1 = u_*$ и задача (1) решена.

Так как $p_0 \neq 0$, то $Ap_0 \neq 0$ и множество

$$\Gamma_1 = \{u \in E^n\colon \langle Ap_0, u - u_1\rangle = 0\}$$

представляет собой гиперплоскость размерности $n-1$, проходящую через точку u_1. Важно заметить, что искомая точка u_* также принадлежит Γ_1. В самом деле, поскольку матрицы A, A^{-1} симметричны, то с учетом равенств (2), (3) имеем $\langle Ap_0, u_* - u_1\rangle = \langle Ap_0, A^{-1}b - u_1\rangle = \langle A^{-1}Ap_0, b - Au_1\rangle = -\langle p_0, J'(u_1)\rangle = 0$. Поэтому дальнейший поиск точки u_* имеет смысл проводить в гиперплоскости Γ_1. Для этого нужно найти какое-либо направление p_1, параллельное гиперплоскости Γ_1. Можно искать p_1, например, в виде

$$p_1 = J'(u_1) - \beta_0 p_0, \quad \beta_0 = \text{const}.$$

Условие параллельности p_1 гиперплоскости Γ_1 дает равенство $\langle Ap_0, p_1\rangle = \langle Ap_0, J'(u_1) - \beta_0 p_0\rangle = 0$, т. е.

$$\beta_0 = \langle Ap_0, J'(u_1)\rangle / \langle Ap_0, p_0\rangle.$$

Поскольку $J'(u) \neq 0$, то $p_1 \neq 0$. В самом деле, если бы $p_1 = 0$, то $J'(u) = \beta_0 p_0$ и согласно (3) $|J'(u_1)|^2 = \beta_0 \langle p_0, J'(u_1)\rangle = 0$ — противоречие с условием $J'(u_1) \neq 0$. Из $p_1 \neq 0$ следует, что $Ap_1 \neq 0$. Положим

$$u_2 = u_1 - \alpha_1 p_1, \quad \alpha_1 \geqslant 0,$$

где величину α_1 будем определять из условия

$$f_1(\alpha_1) = \min_{\alpha \geqslant 0} f_1(\alpha), \quad f_1(\alpha) = J(u_1 - \alpha p_1).$$

С учетом равенств (3) имеем $f_1'(0) = \langle J'(u_1), -p_1\rangle = \langle J'(u_1),$

$-J'(u_1)+\beta_0 p_0\rangle = -|J'(u_1)|^2 < 0$, поэтому $\alpha_1 > 0$. Тогда

$$f_1'(\alpha_1) = 0 = \langle J'(u_1-\alpha_1 p_1), -p_1\rangle = -\langle J'(u_2), p_1\rangle = 0.$$

Заметим, что

$$J'(u_1) - J'(u_2) = Au_1 - b - Au_2 + b = \alpha_1 A p_1.$$

Тогда в силу (3) и выбора p_1 получаем

$$\langle J'(u_2), J'(u_0)\rangle = \langle J'(u_2), p_0\rangle = \langle J'(u_1) - \alpha_1 A p_1, p_0\rangle = 0.$$

Отсюда следует равенство

$$\langle J'(u_2), J'(u_1)\rangle = \langle J'(u_2), p_1 + \beta_0 p_0\rangle = 0.$$

Таким образом, первые две итерации метода сопряженных направлений для задачи (1) описаны. Показано, что

$$\langle J'(u_1), p_0\rangle = \langle J'(u_1), J'(u_0)\rangle = 0, \quad \langle Ap_0, p_1\rangle = \langle Ap_1, p_0\rangle = 0,$$
$$\langle J'(u_2), p_1\rangle = \langle J'(u_2), p_0\rangle = \langle J'(u_2), J'(u_1)\rangle = \langle J'(u_2), J'(u_0)\rangle = 0.$$

Кроме того, заметим, что векторы Ap_0, Ap_1 линейно независимы. В самом деле, если $\gamma_0 Ap_0 + \gamma_1 Ap_1 = 0$, то, умножая это равенство скалярно сначала на p_0, затем на p_1, получим $\gamma_0 = \gamma_1 = 0$. Можем считать, что $J'(u_2) \neq 0$, иначе $u_2 = u_*$ — задача (1) решена.

Теперь у нас есть основание сделать следующее индуктивное предположение: пусть при некотором $k \geqslant 2$ уже найдены точки $u_0, u_1, \ldots, u_k$, $u_{k+1} = u_i - \alpha_i p_i$ $(i = 0, 1, \ldots, k-1)$, где

$$p_i = J'(u_i) - \beta_i p_{i-1} \neq 0, \quad \beta_i = \langle J'(u_i), Ap_{i-1}\rangle / \langle Ap_{i-1}, p_{i-1}\rangle,$$

а величины $\alpha_i > 0$ определены из условий

$$f_i(\alpha_i) = \min_{\alpha \geqslant 0} f_i(\alpha), \quad f_i(\alpha) = J(u_i - \alpha p_i), \qquad i = 0, 1, \ldots, k-1;$$

пусть

$$\langle Ap_i, p_j\rangle = 0, \quad i \neq j, \quad 0 \leqslant i, \quad j \leqslant k-1, \tag{4}$$

$$\langle J'(u_i), p_j\rangle = 0, \quad 0 \leqslant j < i \leqslant k, \tag{5}$$

$$\langle J'(u_i), J'(u_j)\rangle = 0, \ i \neq j, \ 0 \leqslant i, \ j \leqslant k; \tag{6}$$

и, кроме того, пусть $J'(u_i) \neq 0$ $(i = 0, 1, \ldots, k)$ и система векторов $\{Ap_0, Ap_1, \ldots, Ap_{k-1}\}$ линейно независима.

Тогда множество

$$\Gamma_k = \{u \in E^n\colon \langle Ap_i, u - u_{i+1}\rangle = 0, \quad i = 0, 1, \ldots, k-1\}$$

представляет собой гиперплоскость (аффинное множество — см. пример 4.1.4) размерности $n-k$. Поскольку из (5) следует

$$\langle Ap_i, u_k - u_{i+1}\rangle = \langle p_i, Au_k - Au_{i+1}\rangle = \langle p_i, J'(u_k) - J'(u_{i+1})\rangle = 0$$

для всех $i = 0, 1, \ldots, k-1$, то $u_k \in \Gamma_k$. Замечательно то, что

$u_* \in \Gamma_k$, так как согласно (2), (5)

$$\langle Ap_i, u_* - u_{i+1}\rangle = \langle Ap_i, A^{-1}b - u_{i+1}\rangle = \langle p_i, b - Au_{i+1}\rangle = \\ = \langle p_i, -J'(u_{i+1})\rangle = 0, \quad i = 0, 1, \ldots, k-1.$$

Поэтому дальнейший поиск точки u_* целесообразно продолжать в гиперплоскости Γ_k. Для этого нужно найти направление p_k, параллельное Γ_k, т. е. удовлетворяющее условиям $\langle Ap_i, p_k\rangle = 0$ $(i = 0, 1, \ldots, k-1)$. Будем искать p_k в виде

$$p_k = J'(u_k) - \beta_k p_{k-1}. \tag{7}$$

Заметим, что

$$J'(u_i) - J'(u_{i+1}) = Au_i - Au_{i+1} = \alpha_i Ap_i, \quad i = 0, 1, \ldots, k-1. \tag{8}$$

Из (4), (6), (8) следует

$$\langle Ap_i, p_k\rangle = \langle Ap_i, J'(u_k) - \beta_k p_{k-1}\rangle = \langle Ap_i, J'(u_k)\rangle - \\ - \beta_k \langle Ap_i, p_{k-1}\rangle = \langle J'(u_i) - J'(u_{i+1}), J'(u_k)\rangle \alpha_i^{-1} = 0$$

для всех $i = 0, 1, \ldots, k-2$ при любом выборе β_k в (7). Поэтому для параллельности направления p_k гиперплоскости Γ_k остается удовлетворить равенству $\langle Ap_{k-1}, p_k\rangle = 0$. Отсюда имеем $\langle Ap_{k-1}, J'(u_k) - \beta_k p_{k-1}\rangle = \langle Ap_{k-1}, J'(u_k)\rangle - \beta_k\langle Ap_{k-1}, p_{k-1}\rangle = 0$ или

$$\beta_k = \langle Ap_{k-1}, J'(u_k)\rangle / \langle Ap_{k-1}, p_{k-1}\rangle. \tag{9}$$

Заметим, что $p_k \neq 0$, ибо в противном случае $J'(u_k) = \beta_k p_{k-1}$, и тогда в силу (5) $|J'(u_k)|^2 = \beta_k\langle J'(u_k), p_{k-1}\rangle = 0$, что противоречит индуктивному предположению.

Итак, учитывая выбор направления p_k и равенства (4), имеем

$$\langle Ap_i, p_j\rangle = 0, \quad i \neq j, \quad 0 \leqslant i, \quad j \leqslant k. \tag{10}$$

Следующее $(k+1)$-е приближение будем искать в виде

$$u_{k+1} = u_k - \alpha_k p_k, \quad \alpha_k \geqslant 0, \tag{11}$$

где α_k определяется из условия

$$f_k(\alpha_k) = \min_{\alpha \geqslant 0} f_k(\alpha), \quad f_k(\alpha) = J(u_k - \alpha p_k). \tag{12}$$

Поскольку $f_k(\alpha)$ — сильно выпуклая функция, то величина α_k существует и единственна. С учетом предположений индукции и формулы (7) имеем

$$f_k'(0) = \langle J'(u_k), -p_k\rangle = \langle J'(u_k), -J'(u_k) + \beta_k p_{k-1}\rangle = \\ = -|J'(u_k)|^2 < 0.$$

Это значит, что $\alpha_k > 0$ и $f_k'(\alpha_k) = 0 = \langle J'(u_{k+1}), -p_k\rangle$ или

$$\langle J'(u_{k+1}), p_k\rangle = 0. \tag{13}$$

Отсюда нетрудно получить явное выражение для α_k. В самом деле,

$$0 = \langle J'(u_{k+1}),\ p_k\rangle = \langle Au_{k+1} - b,\ p_k\rangle = \\ = \langle Au_k - \alpha_k Ap_k - b,\ p_k\rangle = \langle J'(u_k),\ p_k\rangle - \alpha_k\langle Ap_k,\ p_k\rangle.$$

Так как $p_k \neq 0$, то $\langle Ap_k,\ p_k\rangle \neq 0$ и из последнего равенства вытекает

$$\alpha_k = \frac{\langle J'(u_k),\ p_k\rangle}{\langle Ap_k,\ p_k\rangle} = \frac{\langle J'(u_k),\ J'(u_k) - \beta_k p_{k-1}\rangle}{\langle Ap_k,\ p_k\rangle} = \frac{|J'(u_k)|^2}{\langle Ap_k,\ p_k\rangle}. \quad (14)$$

Далее, заметим

$$J'(u_k) - J'(u_{k+1}) = Au_k - Au_{k+1} = \alpha_k Ap_k.$$

Отсюда и из (4), (5) имеем

$$\langle J'(u_{k+1}),\ p_i\rangle = \langle J'(u_k) - \alpha_k Ap_k,\ p_i\rangle = 0, \quad i = 0,\ 1,\ \ldots,\ k-1. \quad (15)$$

Собрав все равенства (5), (13), (15), получим

$$\langle J'(u_i),\ p_j\rangle = 0, \quad 0 \leqslant j < i \leqslant k+1. \quad (16)$$

Из предположения индукции и равенств (16) следует

$$\langle J'(u_{k+1}),\ J'(u_i)\rangle = \langle J'(u_{k+1}),\ p_i + \beta_i p_{i-1}\rangle = 0, \quad i = 1,\ \ldots,\ k,$$
$$\langle J'(u_{k+1}),\ J'(u_0)\rangle = \langle J'(u_{k+1}),\ p_0\rangle = 0.$$

Отсюда и из (6) имеем

$$\langle J'(u_i),\ J'(u_j)\rangle = 0, \quad i \neq j, \quad 0 \leqslant i, \quad j \leqslant k+1.$$

Наконец, покажем, что система $\{Ap_0,\ \ldots,\ Ap_k\}$ линейно независима. В самом деле, если $\gamma_0 Ap_0 + \gamma_1 Ap_1 + \ldots + \gamma_k Ap_k = 0$, то умножая это равенство на p_j скалярно, с учетом (10) получим $\gamma_j\langle Ap_j,\ p_j\rangle = 0$ $(j = 0,\ 1,\ \ldots,\ k)$. Так как $p_j \neq 0$, то $\langle Ap_j,\ p_j\rangle > 0$ и последние равенства возможны лишь при $\gamma_j = 0$ $(j = 0, 1, \ldots, k)$.

Тем самым все этапы индукции проведены, следующее $(k+1)$-е приближение u_{k+1} построено. Если $J'(u_{k+1}) = 0$, то $u_{k+1} = u_*$ — решение задачи (1) найдено. Если же $J'(u_{k+1}) \neq 0$, то согласно индукции процесс можно продолжать дальше.

Метод сопряженных направлений для задачи (1), заключающийся в построении последовательности $\{u_k\}$ по правилу (11), где α_k, p_k определяются из (7), (9), (12) (или (14)), $p_0 = J'(u_0)$, описан. Название этого метода объясняет следующее

Определение 1. *Векторы* $p_0,\ p_1,\ \ldots,\ p_k$ называются *сопряженными относительно матрицы* A или *A-ортогональными*, если $\langle Ap_i,\ p_j\rangle = 0$ при всех $i \neq j$, $0 \leqslant i,\ j \leqslant k$.

Нетрудно видеть, что для квадратичной задачи (1) метод сопряженных направлений закончится за конечное число итераций нахождением точки u_*. В самом деле, векторы $J'(u_0),\ J'(u_1),\ \ldots,\ J'(u_k),\ \ldots$, получаемые этим методом, образуют ортогональ-

ную систему: $\langle J'(u_i), J'(u_j)\rangle = 0 \quad (i \neq j)$. Однако в n-мерном пространстве не может быть более n ненулевых взаимно ортогональных векторов. Следовательно, найдется номер k, $0 \leqslant k < n$) такой, что $J'(u_k) = 0$. Тогда $u_k = u_*$ — решение задачи (1).

2. Перейдем к рассмотрению задачи

$$J(u) \to \inf; \quad u \in U \equiv E^n, \tag{17}$$

где функция $J(u) \in C^1(E^n)$, причем в отличие от задачи (1) здесь $J(u)$ не предполагается квадратичной. Так как формула (9) содержит матрицу A, характеризующую квадратичную функцию (1), то описанный выше метод сопряженных направлений (7), (9), (11), (12) не может быть непосредственно применен для решения задачи (17). Поэтому сначала формулу (9) приведем к виду, не содержащему матрицу A. С учетом равенств (6), (8) числитель и знаменатель дроби (9) можно преобразовать так:

$$\langle Ap_{k-1}, J'(u_k)\rangle = \langle J'(u_{k-1}) - J'(u_k), J'(u_k)\rangle \alpha_{k-1}^{-1} = \\ = -|J'(u_k)|^2 \alpha_{k-1}^{-1},$$

$$\langle Ap_{k-1}, p_{k-1}\rangle = \langle J'(u_{k-1}) - J'(u_k), p_{k-1}\rangle \alpha_{k-1}^{-1} = \\ = \langle J'(u_{k-1}), p_{k-1}\rangle \alpha_{k-1}^{-1} = \langle J'(u_{k-1}), J'(u_{k-1}) - \beta_{k-1}p_{k-2}\rangle \alpha_{k-1}^{-1} = \\ = |J'(u_{k-1})|^2 \alpha_{k-1}^{-1}.$$

Тогда формула (9) запишется в виде

$$\beta_k = \frac{\langle J'(u_k), J'(u_{k-1}) - J'(u_k)\rangle}{|J'(u_{k-1})|^2}, \tag{18}$$

где

$$\beta_k = -\frac{|J'(u_k)|^2}{|J'(u_{k-1})|^2}. \tag{19}$$

Кроме того, вспоминая, что для функции (1) $A = J''(u_k)$, формулу (9) можно представить еще и в такой форме:

$$\beta_k = \frac{\langle J''(u_k)\, p_{k-1}, J'(u_k)\rangle}{\langle J''(u_k)\, p_{k-1}, p_{k-1}\rangle}. \tag{20}$$

Для квадратичной функции (1) все три формулы (18)—(20) дают одну и ту же величину β_k. Но если функция $J(u)$ отлична от квадратичной, то из этих формул будут получаться, вообще говоря, различные значения β_k.

В результате, отправляясь от соотношений (7), (11), (12), (18)—(20), придем к следующему описанию метода сопряженных направлений для задачи (17). Пусть u_0 — некоторое началь-

ное приближение. Будем строить последовательность $\{u_k\}$ по правилам

$$u_{k+1} = u_k - \alpha_k p_k, \quad k = 0, 1, \ldots, \tag{21}$$

где

$$p_0 = J'(u_0), \quad p_k = J'(u_k) - \beta_k p_{k-1}, \quad k = 1, 2, \ldots, \tag{22}$$

величина α_k определяется условиями

$$\alpha_k \geqslant 0, \quad f_k(\alpha_k) = \min_{\alpha \geqslant 0} f_k(\alpha), \quad f_k(\alpha) = J(u_k - \alpha p_k), \tag{23}$$

а β_k в (22) вычисляется по одной из формул (18), (19) или (20). Отметим, что в варианте (20)—(23) метода сопряженных направлений требуется, чтобы $J(u) \in C^2(E^n)$, и поэтому на практике он применяется очень редко и лишь в тех случаях, когда матрица $J''(u)$ вычисляется достаточно просто.

Так как в задаче (17) квадратичность функции не предполагается, то нельзя ожидать, что описанный метод сопряженных направлений за конечное число итераций приведет к точке минимума функции $J(u)$ на E^n. Далее, точное определение величины α_k из условий (23) возможно лишь в редких случаях, поэтому реализация каждой итерации метода будет сопровождаться неизбежными погрешностями. Как показывает практика, эти погрешности, накапливаясь, могут привести к тому, что векторы $\{p_k\}$ перестают указывать направление убывания функции, и сходимость метода может нарушиться. Чтобы бороться с этим явлением, метод сопряженных направлений время от времени обновляют, полагая в (22) $\beta_k = 0$. Обозначим множество тех номеров $k \geqslant 1$, при которых принимается $\beta_k = 0$, через I_0. Номера $k \in I_0$ называются *моментами обновления метода*. Если метод используется без обновления, то $I_0 = \varnothing$. На практике часто берут $I_0 = \{n, 2n, 3n, \ldots\}$, где n — размерность рассматриваемого пространства. Возможны и другие правила выбора моментов обновления. Кстати, если $I_0 = \{1, 2, 3, \ldots\}$, то метод (21)—(23) превратится в метод скорейшего спуска.

Если функция $J(u)$ не является квадратичной, то для описанного метода сопряженных направлений равенства (5), (6), вообще говоря, не выполняются. Однако, тем не менее, и в общем случае при любом выборе моментов обновления справедливы равенства

$$\langle J'(u_{k+1}), p_k \rangle = 0, \quad \langle J'(u_k), p_k \rangle = |J'(u_k)|^2, \quad k = 0, 1, \ldots \tag{24}$$

В самом деле, при $k = 0$ имеем $p_0 = J'(u_0)$, поэтому $\langle J'(u_0), p_0 \rangle = |J'(u_0)|^2$. Из условий (23) при $k = 0$ в случае $\alpha_0 > 0$ следует $f_0'(\alpha_0) = \langle J'(u_0 - \alpha_0 p_0), -p_0 \rangle = -\langle J'(u_1), p_0 \rangle = 0$. Если же $\alpha_0 = 0$, то $u_1 = u_0$ и $0 \leqslant f_0'(0) = -\langle J'(u_0), p_0 \rangle = -|J'(u_0)|^2 \leqslant 0$, так что $J'(u_0) = J'(u_1) = 0$, $\langle J'(u_1), p_0 \rangle = 0$. Таким образом, равенства (24) при $k = 0$ верны. Сделаем индуктивное предположение: пусть для некоторого $k \geqslant 1$ имеют место равенства $\langle J'(u_k), p_{k-1} \rangle = 0$, $\langle J'(u_{k-1}), p_{k-1} \rangle = |J'(u_{k-1})|^2$. Тогда из (23) при $\alpha_k > 0$ получим $f_k'(\alpha_k) = -\langle J'(u_{k+1}), p_k \rangle = 0$. Если же $\alpha_k = 0$, то $u_{k+1} = u_k$ и

$0 \leqslant f_k'(0) = -\langle J'(u_k), p_k\rangle = -\langle J'(u_k), J'(u_k) - \beta_k p_{k-1}\rangle = -|J'(u_k)|^2 = -|J'(u_{k+1})|^2 \leqslant 0$, поэтому $J'(u_{k+1}) = 0$ и $\langle J'(u_{k+1}), p_k\rangle = 0$. Наконец,

$$\langle J'(u_k), p_k\rangle = \langle J'(u_k), J'(u_k) - \beta_k p_{k-1}\rangle = |J'(u_k)|^2.$$

Равенства (24) доказаны. Из первого равенства и определения (22) вектора p_k следует

$$|p_k|^2 = |J'(u_k) - \beta_k p_{k-1}|^2 = |J'(u_k)|^2 + \beta_k^2 |p_{k-1}|^2, \qquad k = 1, 2, \ldots \tag{25}$$

3. Пользуясь соотношениями (24), (25), установим сходимость метода сопряженных направлений (21) — (23), (18).

Теорема 1. *Пусть функция $J(u)$ сильно выпукла на E^n, $J(u) \in C^{1,1}(E^n)$. Тогда при любом выборе множества I_0 моментов обновления и любом начальном приближении u_0 последовательность $\{u_k\}$, определяемая условиями* (21) — (23), (18), *сходится к точке u_* минимума функции $J(u)$ на E^n, причем справедливы оценки*

$$0 \leqslant a_k = J(u_k) - J_* \leqslant q^k a_k, \quad |u_k - u_*|^2 \leqslant \frac{2}{\mu} q^k a_0, \qquad k = 0, 1, \ldots, \tag{26}$$

где $q = 1 - \dfrac{\mu^3}{L(\mu^2 + L^2)}$ $(0 < q < 1)$, μ — *постоянная из теоремы* 4.3.3, L — *константа Липшица для градиента $J'(u)$ на E^n.*

Доказательство. Из теоремы 4.3.1 следует существование и единственность точки u_*, в которой $J(u_*) = J_* = \inf\limits_{E^n} J(u)$. Функция $f_k(\alpha) = J(u_k - \alpha p_k)$ при $p_k \neq 0$ также сильно выпукла, и условия (23) однозначно определяют величину $\alpha_k > 0$. Будем считать, что $p_k \neq 0$, $J'(u_k) \neq 0$, $\alpha_k > 0$ при всех $k = 0, 1, \ldots$, ибо в противном случае из (24) при $p_k = 0$ получим $J'(u_k) = 0$ и $u_k = u_*$ — решение задачи (17).

В силу выбора α_k при всех $\alpha \geqslant 0$ имеем $J(u_{k+1}) \leqslant J(u_k - \alpha p_k)$. Отсюда и из леммы 2.3.1 с учетом второго равенства (24) получим

$$J(u_k) - J(u_{k+1}) \geqslant J(u_k) - J(u_k - \alpha p_k) \geqslant \alpha\langle J'(u_k), p_k\rangle - \frac{\alpha^2 L}{2}|p_k|^2 =$$
$$= \alpha|J'(u_k)|^2 - \frac{\alpha^2 L}{2}|p_k|^2, \quad \alpha \geqslant 0, \qquad k = 0, 1, \ldots \tag{27}$$

Докажем неравенство

$$\gamma L|p_k|^2 \leqslant |J'(u_k)|^2, \quad \gamma = \mu^2 L^{-1}(\mu^2 + L^2)^{-1}, \qquad k = 0, 1, \ldots \tag{28}$$

Согласно теореме 4.3.3 $\mu|u_k - u_{k-1}|^2 = \mu\alpha_{k-1}^2|p_{k-1}|^2 \leqslant \langle J'(u_k) - J'(u_{k-1}), u_k - u_{k-1}\rangle = \langle J'(u_k) - J'(u_{k-1}), p_{k-1}\rangle(-\alpha_{k-1})$. Отсюда с учетом равенств (24) имеем

$$\mu\alpha_{k-1}|p_{k-1}|^2 \leqslant \langle J'(u_{k-1}), p_{k-1}\rangle = |J'(u_{k-1})|^2.$$

Тогда из (18) следует

$$|\beta_k| \leqslant |J'(u_k)|L|u_k - u_{k-1}||J'(u_{k-1})|^{-2} \leqslant$$
$$\leqslant |J'(u_k)|L\alpha_{k-1}|p_{k-1}|(\mu\alpha_{k-1}|p_{k-1}|^2)^{-1} = L\mu^{-1}|J'(u_k)||p_{k-1}|^{-1},$$

т. е.

$$|\beta_k|\,|p_{k-1}| \leqslant L\mu^{-1}|J'(u_k)|, \qquad k = 0, 1, \ldots$$

Отсюда и из (25) получим $|p_k|^2 \leqslant |J'(u_k)|^2(1 + L^2\mu^{-2})$, что равносильно неравенству (28).

Теперь нетрудно доказать оценки (26). Из (27) с учетом (28) имеем

$$a_k - a_{k+1} \geqslant \alpha\left(1 - \frac{\alpha}{2\gamma}\right) |J'(u_k)|^2, \quad \alpha \geqslant 0, \quad k = 0, 1, \ldots$$

Следовательно,

$$a_k - a_{k+1} \geqslant \max_{\alpha \geqslant 0}\left(1 - \frac{\alpha}{2\gamma}\right) |J'(u_k)|^2 = \frac{\gamma}{2} |J'(u_k)|^2, \qquad k = 0, 1, \ldots$$

Но $2\mu a_k \leqslant |J'(u_k)|^2$ (см. неравенство (1.18)), поэтому $a_k - a_{k+1} \geqslant \gamma\mu a_k$ или $a_{k+1} \leqslant (1 - \gamma\mu) a_k = q a_k$ $(k = 0, 1, \ldots)$. Отсюда следует первая из оценок (26). Вторая оценка (26) вытекает из первой оценки и неравенства (4.3.2). Остается заметить, что $0 < q < 1$, ибо $\mu \leqslant L$. Теорема доказана.

Отметим, что оценки (26) являются довольно грубыми. Более тонкие исследования показывают, что метод сопряженных направлений на самом деле имеет более высокую скорость сходимости, чем это следует из оценок (26). В то же время этот метод не намного сложнее метода скорейшего спуска. Недостатком метода сопряженных направлений является его чувствительность к погрешностям при определении величины α_k из условия (23) — недостаточно точное определение α_k может привести к ухудшению сходимости метода.

4. В методе (7), (9), (11), (12) направления $p_0, p_1, \ldots, p_k$ строятся с помощью процесса A-ортогонализации последовательно вычисляемых градиентов $J(u_0), J'(u_1), \ldots, J'(u_k)$, и поэтому этот метод для задачи (1) и полученный на его основе метод (21) — (23) для задачи (17) в литературе часто называют *методом сопряженных градиентов*. В общем случае в методе сопряженных направлений могут быть использованы и другие способы построения векторов p_k, отличные от (22). А именно, пусть направления $p_0, p_1, \ldots, p_k$, удовлетворяющие условиям (10), уже известны и с их помощью последовательно построены точки $u_1, \ldots, u_{k+1}$ по формулам (21), (23). Следующий вектор p_{k+1} будем определять из условий $\langle p_{k+1}, Ap_i\rangle = 0$ $(i = 0, 1, \ldots, k)$. В случае квадратичных функций (1) формула (8) остается справедливой при любом выборе векторов $p_0, p_1, \ldots, p_k$ в (21), (23), поэтому условие ортогональности вектора p_{k+1} к векторам $Ap_0, Ap_1, \ldots, Ap_k$ здесь приводит к равенствам

$$\langle p_{k+1}, q_i\rangle = 0, \quad q_i = J'(u_i) - J'(u_{i+1}), \qquad i = 0, 1, \ldots, k. \tag{29}$$

Условия (29) имеют смысл и для неквадратичных функций, и ими пользуются для определения p_{k+1} в общем случае. Обычно вектор p_{k+1} ищут в виде [11, 41, 46, 48, 111, 250, 330]

$$p_{k+1} = H_{k+1} J'(u_{k+1}), \quad H_{k+1} = H_k + \Delta H_k, \tag{30}$$

где матрица ΔH_k определяется из условий (29). Нетрудно видеть, что перечисленные условия (29), (30) матрицу ΔH_k определяют неоднозначно и в зависимости от того, как распорядиться этим произволом, можно получить различные варианты метода сопряженных направлений. Если на каком-либо шаге $H_k = 0$, то метод (21), (23), (29), (30) обновляют, полагая $A_k = I$ — единичная матрица. Приведем один из вариантов этого метода, в котором матрицы H_k определяются по правилу

$$H_{k+1} = H_k - \frac{(Hq_k)(H_k q_k)^T}{\langle Hq_k, q_k\rangle}, \quad q_k = J'(u_{k+1}) - J'(u_k), \quad H_0 = I.$$

В [19] предлагается и исследуется метод сопряженных направлений, позволяющий за конечное число итераций найти точку минимума квадратичной функции (1) на множестве, задаваемом линейными ограничениями типа равенств и неравенств.

Исследование сходимости различных вариантов метода сопряженных направлений, более тонкие оценки скорости сходимости читатель может найти в [11, 19].

Упражнения. 1. Показать, что точка u_k, полученная методом сопряженных направлений для квадратичной функции (1) при $I_0 = \varnothing$, есть точка минимума этой функции на гиперплоскости, проходящей через точку u_0 и натянутой на векторы $J'(u_0), J'(u_1), \ldots, J'(u_{k-1})$.

2. Описать метод сопряженных направлений для функции $J(u) = |Au - b|^2$, $u \in E^n$, где A — матрица порядка $m \times n$, $b \in E^m$.

§ 9. Метод Ньютона

До сих пор мы рассматривали методы первого порядка — так называются методы минимизации, использующие лишь первые производные минимизируемой функции. В этих методах для определения направления убывания функции используется лишь линейная часть разложения функции в ряд Тейлора. Если минимизируемая функция дважды непрерывно дифференцируема и производные $J'(u)$, $J''(u)$ вычисляются достаточно просто, то возможно применение методов минимизации второго порядка, которые используют квадратичную часть разложения этой функции в ряд Тейлора. Поскольку квадратичная часть разложения аппроксимирует функцию гораздо точнее, чем линейная, то естественно ожидать, что методы второго порядка сходятся быстрее, чем методы первого порядка.

Ниже будут описаны два метода второго порядка: в этом параграфе будет рассмотрен метод Ньютона, имеющий квадратичную скорость сходимости на классе сильно выпуклых функций, а в следующем параграфе — метод с кубической скоростью сходимости на этом же классе. Здесь мы пользуемся следующей терминологией, принятой в литературе: говорят, что последовательность $\{u_k\}$ сходится к точке u_* с линейной скоростью или со скоростью геометрической прогрессии (со знаменателем q), если, начиная с некоторого номера, выполняется неравенство $|u_{k+1} - u_*| \leqslant q|u_k - u_*|$ $(0 < q < 1)$; при выполнении неравенства $|u_{k+1} - u_*| \leqslant q_k|u_k - u_*|$, где $\{q_k\} \to 0$, говорят о сверхлинейной скорости сходимости последовательности $\{u_k\}$ к u_*, а если здесь $q_k = C|u_k - u_*|^{s-1}$, т. е. $|u_{k+1} - u_*| \leqslant C|u_k - u_*|^s$, то говорят о скорости сходимости, порядка s (при $s = 2$ получим квадратичную скорость сходимости, при $s = 3$ — кубическую). Для некоторых методов выше была установлена линейная скорость сходимости на классе сильно выпуклых функций; в тех случаях, когда $|u_k - u_*| = O(1/k)$, скорость сходимости ниже линейной; для метода сопряженных направлений можно показать сверхлинейную скорость сходимости [19].

1. Опишем метод Ньютона для задачи

$$J(u) \to \inf; \qquad u \in U, \tag{1}$$

где $J(u) \in C^2(U)$, U — выпуклое замкнутое множество из E^n (например, $U = E^n$). Пусть $u_0 \in U$ — некоторое начальное приближение. Если известно k-е приближение u_k, то приращение

функции $J(u) \in C^2(U)$ в точке u_k можно представить в виде

$$J(u) - J(u_k) = \langle J'(u_k), u - u_k \rangle + \\ + \frac{1}{2} \langle J''(u_k)(u - u_k), u - u_k \rangle + o(|u - u_k|^2).$$

Возьмем квадратичную часть этого приращения

$$J_k(u) \equiv \langle J'(u_k), u - u_k \rangle + \frac{1}{2} \langle J''(u_k)(u - u_k), u - u_k \rangle \qquad (2)$$

и определим вспомогательное приближение $\bar{u}_k$ из условий

$$\bar{u}_k \in U, \quad J_k(\bar{u}_k) = \inf_U J_k(u). \qquad (3)$$

Следующее $(k+1)$-е приближение будем искать в виде

$$u_{k+1} = u_k + \alpha_k(\bar{u}_k - u_k), \qquad 0 \leqslant \alpha_k \leqslant 1. \qquad (4)$$

В зависимости от способа выбора величины α_k в (4) можно получить различные варианты метода (2)—(4), называемого методом Ньютона. Укажем несколько наиболее употребительных способов выбора α_k.

1) В (4) можно принять

$$\alpha_k = 1, \qquad k = 0, 1, \ldots \qquad (5)$$

В этом случае, как следует из (4), $u_{k+1} = \bar{u}_k$ $(k = 0, 1, \ldots)$, т. е. условие (3) сразу определяет следующее $(k+1)$-е приближение. Иначе говоря,

$$u_{k+1} \in U, \quad J_k(u_{k+1}) = \inf_U J_k(u), \quad k = 0, 1, \ldots \qquad (6)$$

В частности, когда $U = E^n$, в точке минимума функции $J_k(u)$ ее производная $J_k'(u)$ обращается в нуль, т. е.

$$J_k'(u_{k+1}) = J'(u_k) + J''(u_k)(u_{k+1} - u_k) = 0. \qquad (7)$$

Это значит, что на каждой итерации метода (2)—(5) или (6) нужно решать линейную алгебраическую систему уравнений (7) относительно неизвестной разности $u_{k+1} - u_k$. Если матрица этой системы $J''(u_k)$ — невырожденная, то из (7) имеем

$$u_{k+1} = u_k - (J''(u_k))^{-1} J'(u_k), \qquad k = 0, 1, \ldots \qquad (8)$$

Широко известный метод Ньютона для решения системы уравнений

$$F(u) = \{F_1(u), \ldots, F_n(u)\} = 0, \qquad u \in E^n,$$

представляет собой итерационный процесс [4, 54]

$$u_{k+1} = u_k - (F'(u_k))^{-1} F(u_k), \qquad k = 0, 1, \ldots, \qquad (9)$$

где $F'(u)$ — матрица, i-я строка которой равна $F'_i(u) = (F_{iu^1}, \ldots, F_{iu^n})$. Сравнение формул (8) и (9) показывает, что метод (8) решения задачи (1) в случае $U = E^n$ представляет собой известный метод Ньютона для решения уравнения $J'(u) = 0$, определяющего стационарные точки функции $J(u)$. Отсюда происходит название метода (2)—(4) и в общем случае.

2) В качестве α_k в (4) можно принять $\alpha_k = \lambda^{i_0}$, где i_0 — минимальный среди $i \geqslant 0$ номер, для которых выполняется неравенство [19]

$$J(u_k) - J(u_k + \lambda^i(\bar{u}_k - u_k)) \geqslant \varepsilon\lambda^i |J_k(\bar{u}_k)|, \tag{10}$$

где λ, ε — параметры метода, $0 < \lambda$; $\varepsilon < 1$.

3) Возможен выбор α_k в (4) из условий [19]

$$0 \leqslant \alpha_k \leqslant 1, \quad f_k(\alpha_k) = \min_{0 \leqslant \alpha \leqslant 1} f_k(\alpha), \quad f_k(\alpha) = J(u_k + \alpha(\bar{u}_k - u_k)). \tag{11}$$

Заметим, что метод (2)—(4) с выбором длины шага α_k по правилам (10), (11) аналогичен соответствующим вариантам метода условного градиента. Для определения $\bar{u}_k$ использовалась линейная часть приращения, а в методе Ньютона — квадратичная часть (2).

Если $J_k(u)$ из (2) сильно выпукла, а $U = E^n$ или U задается линейными ограничениями типа равенств или неравенств, то для определения $\bar{u}_k$ из (3) могут быть использованы методы из § 7, 8. Следует заметить, что задача (3) в общем случае может оказаться весьма сложной и сравнимой по объему требуемой для своего решения вычислительной работы с исходной задачей (1). Метод Ньютона для решения задачи (1) обычно применяют в тех случаях, когда вычисление производных $J'(u)$, $J''(u)$ не представляет особых трудностей и вспомогательная задача (3) решается достаточно просто. Достоинством метода Ньютона является высокая скорость сходимости. Поэтому хотя трудоемкость каждой итерации этого метода, вообще говоря, выше, чем в методах первого порядка, но общий объем вычислительной работы, необходимой для решения задачи (1) с требуемой точностью, при применении метода Ньютона может оказаться меньше, чем при применении других более простых методов.

2. Сначала исследуем сходимость метода Ньютона (2)—(4) с выбором шага α_k из условия (5) при условии $U = E^n$ или, проще говоря, метода (8).

Теорема 1. *Пусть функция $J(u)$ сильно выпукла на E^n, $J(u) \in C^2(E^n)$ и, кроме того,*

$$\|J''(u) - J''(v)\| \leqslant L|u - v|, \quad u, v \in E^n, \quad L = \text{const} > 0. \tag{12}$$

Пусть начальное приближение u_0 выбрано таким, что

$$L|J'(u_0)| \leqslant 2\mu^2 q, \tag{13}$$

где $\mu > 0$ — *постоянная из теоремы* 4.3.4, *а* q — *некоторая константа*, $0 < q < 1$. *Тогда последовательность* $\{u_k\}$, *определяемая условиями* (8), *существует, сходится к точке* u_* ***минимума*** $J(u)$ *на* E^n, *причем справедлива оценка*

$$|u_k - u_*| \leqslant 2\mu L^{-1} q^{2^k}, \qquad k = 0, 1, \ldots \tag{14}$$

Доказательство. Существование и единственность точки u_* установлена в теореме 4.3.1. Согласно теореме 4.3.4

$$\langle J''(u)\xi, \ \xi\rangle \geqslant \mu|\xi|^2, \quad u \in E^n, \quad \xi \in E^n. \tag{15}$$

Отсюда следует, что система уравнений $J''(u)\xi = 0$ имеет единственное решение $\xi = 0$ и, следовательно, матрица $J''(u)$ невырожденная при всех $u \in E^n$. Это значит, что система (7) при каждом $k = 0, 1, \ldots$ имеет, и притом единственное, решение, т. е. последовательность $\{u_k\}$ однозначно определяется условиями (8). Кроме того, полагая в (15) $\xi = (J''(u))^{-1}z$, получим $\mu|(J''(u))^{-1}z|^2 \leqslant \langle z, \ (J''(u))^{-1}z\rangle \leqslant |z|\,|(J''(u))^{-1}z|$ или $|(J''(u))^{-1}z| \leqslant |z|\mu^{-1}$ при всех $z \in E^n$. Это значит, что

$$\|(J''(u))^{-1}\| \leqslant \mu^{-1}, \quad u \in E^n. \tag{16}$$

Введем числовую последовательность $a_k = |J'(u_k)|$ и покажем, что

$$a_k \leqslant 2\mu^2 L^{-1} q^{2^k}, \qquad k = 0, 1, \ldots \tag{17}$$

При $k = 0$ неравенство (17) следует из условия (13). Пусть (17) справедливо при некотором $k \geqslant 0$. Из условия (8) и формулы (2.3.5) имеем

$$J'(u_{k+1}) = J'(u_k) + \int_0^1 J''(u_k + t(u_{k+1} - u_k))(u_{k+1} - u_k)\,dt =$$

$$= \int_0^1 [J''(u_k) - J''(u_k + t(u_{k+1} - u_k))]\,dt\,(J''(u_k))^{-1} J'(u_k).$$

Отсюда и из (8), (12), (16) с помощью предположения индукции получим

$$a_{k+1} \leqslant (L/2)\,|u_{k+1} - u_k|\,\mu^{-1} a_k \leqslant (L/(2\mu^2))\,a_k^2 \leqslant$$

$$\leqslant (L/(2\mu^2))\,(2\mu^2/L)^2\,(q^{2^k})^2 = (2\mu^2/L)\,q^{2^{k+1}}.$$

Неравенства (17) доказаны. Тогда из теоремы 4.3.3 с учетом равенства $J'(u_*) = 0$ имеем $\mu\,|u_k - u_*|^2 \leqslant \langle J'(u_k) - J'(u_*), \ u_k - u_*\rangle \leqslant |J'(u_k)|\,|u_k - u_*|$ или $|u_k - u_*| \leqslant a_k\mu^{-1}$. Отсюда и из неравенства (17) следует оценка (14).

Теорема 1 доказана.

Как видно из оценки (14) и как показывает практика, метод Ньютона (8) сходится очень быстро. Однако у него есть

один существенный недостаток: для его сходимости начальная точка u_0 должна выбираться достаточно близкой к искомой точке u_*. Это требование в теореме 1 выражено условием (13), означающим, что $|u_0 - u_*| \leqslant a_0\mu^{-1} \leqslant (2\mu/L)\,q$. Приведем пример, показывающий, что при отсутствии хорошего начального приближения метод (8) может расходиться.

Пример 1. Пусть

$$J(u) = \begin{cases} -\dfrac{1}{4\delta^3}u^4 + \dfrac{1}{2}\left(1 + \dfrac{3}{\delta}\right)u^2, & |u| \leqslant \delta, \\ \dfrac{u^2}{2} + 2|u| - \dfrac{3}{4}\delta, & |u| > \delta, \end{cases}$$

где $u \in E^1$, а δ — сколь угодно малое фиксированное положительное число, $0 < \delta < 1$. Нетрудно видеть, что $J(u) \in C^2(E^1)$ и, кроме того, $J''(u) \geqslant 1$ при всех $u \in E^1$, так что $J(u)$ сильно выпукла на E^1. Далее, ясно, что $J_* = 0$, $u_* = 0$. В качестве начального приближения возьмем $u_0 = \delta$. Из (8) получим последовательность $u_k = (-1)^k \cdot 2$ $(k = 1, 2, \ldots)$, которая расходится, хотя начальное приближение u_0 отличается от $u_* = 0$ на малое число δ.

Метод (8) часто применяют на завершающем этапе поиска минимума, когда с помощью более грубых, менее трудоемких методов уже найдена некоторая точка, достаточно близкая к точке минимума.

3. Исследуем сходимость метода (2) — (5) без предположения, что $U = E^n$.

Теорема 2. *Пусть U — выпуклое замкнутое множество из E^n, функция $J(u)$ сильно выпукла и принадлежит классу $C^2(U)$ и*

$$|J''(u) - J''(v)| \leqslant L|u - v|, \quad u, v \in U, \quad L = \text{const}. \tag{18}$$

Тогда последовательность $\{u_k\}$ однозначно определяется условиями (6) *при любом выборе начального приближения u_0. Если*

$$q = (L/(2\mu))\,|u_1 - u_0| < 1, \tag{19}$$

то последовательность $\{u_k\}$, определяемая условиями (6), *сходится к точке u_* — решению задачи* (1), *причем справедлива оценка*

$$|u_k - u_*| \leqslant \frac{2\mu}{L}\sum_{m=k}^{\infty} q^{2^m} \leqslant \frac{2\mu}{L}\,q^{2^k}\left(1 - q^{2^k}\right)^{-1}, \qquad k = 0, 1, \ldots; \tag{20}$$

здесь $\mu > 0$ — постоянная из теоремы 4.3.4.

Доказательство. В силу теоремы 4.3.1 функция $J(u)$ ограничена снизу и достигает своей нижней грани на U в единственной точке u_*. Из теоремы 4.3.4 следует

$$\langle J''(u)\xi, \xi\rangle \geqslant \mu|\xi|^2, \quad u \in U, \quad \xi \in L_U, \tag{21}$$

где L_U — подпространство, параллельное аффинной оболочке множества U. Так как $J_k''(u) = J''(u_k)$, то из предыдущего неравенства и теоремы 4.3.4 вытекает сильная выпуклость функции $J_k(u)$ на множестве U при всех $k = 0, 1, \ldots$ Снова обращаясь к теореме 4.3.1, заключаем, что условия (6)

однозначно определяют точку u_{k+1}. Таким образом, существование последовательности $\{u_k\}$ из (6) доказано. Применив теорему 4.2.3 к функции $J_k(u)$ на U, получим

$$\langle J_k'(u_{k+1}),\ u-u_{k+1}\rangle \geqslant 0, \qquad u\in U, \qquad k=0,1,\ldots \tag{22}$$

Так как $J_k'(u)\equiv J'(u_k)+J''(u_k)(u-u_k)$, то неравенство (22) перепишется в виде

$$\langle J'(u_k)+J''(u_k)(u_{k+1}-u_k),\ u-u_{k+1}\rangle \geqslant 0, \qquad u\in U, \qquad k=0,1,\ldots \tag{23}$$

Может случиться, что $u_{k+1}=u_k$. Тогда из (23) имеем $\langle J'(u_k),\ u-u_k\rangle\geqslant 0$ при всех $u\in U$. Согласно теореме 4.2.3 в этом случае $u_k=u_*$ — задача (1) решена. Поэтому можем считать, что $u_k\neq u_{k+1}$ при всех $k=0,1,\ldots$

Положим в (23) $u=u_k$. Получим

$$\langle J'(u_k)+J''(u_k)(u_{k+1}-u_k),\ u_k-u_{k+1}\rangle\geqslant 0.$$

Отсюда и из (21) имеем

$$\mu|u_{k+1}-u_k|^2\leqslant\langle J''(u_k)(u_{k+1}-u_k),\ u_{k+1}-u_k\rangle\leqslant$$
$$\leqslant\langle J'(u_k),\ u_k-u_{k+1}\rangle, \qquad k=0,1,\ldots \tag{24}$$

Оценим правую часть (24) сверху. Для этого в (22) заменим k на $k-1$. Получим $\langle J_{k-1}'(u_k),\ u-u_k\rangle\geqslant 0$, $u\in U$. Полагая здесь $u=u_{k+1}$, имеем

$$\langle J_{k-1}'(u_k),\ u_k-u_{k+1}\rangle\leqslant 0, \qquad k=1,2,\ldots$$

Отсюда, из формулы (2.3.5) и условия (18) следует

$$\langle J'(u_k),\ u_k-u_{k+1}\rangle\leqslant\langle J'(u_k)-J_{k-1}'(u_k),\ u_k-u_{k+1}\rangle=$$
$$=\langle J'(u_k)-J'(u_{k-1})-J''(u_{k-1})(u_k-u_{k-1}),\ u_k-u_{k+1}\rangle=$$
$$=\left\langle\int_0^1[J''(u_{k-1}+t(u_k-u_{k-1}))-J''(u_{k-1})]\,dt\,(u_k-u_{k-1}),\ u_k-u_{k+1}\right\rangle\leqslant$$
$$\leqslant\frac{L}{2}|u_k-u_{k-1}|^2|u_k-u_{k+1}|, \qquad k=1,2,\ldots$$

Подставив полученную оценку в (24), получим

$$|u_{k+1}-u_k|\leqslant(L/(2\mu))|u_k-u_{k-1}|^2, \qquad k=1,2,\ldots \tag{25}$$

Докажем оценку

$$|u_{k+1}-u_k|\leqslant(2\mu/L)\,q^{2^k}, \qquad k=0,1,\ldots \tag{26}$$

При $k=0$ эта оценка следует из условия (19). Сделаем индуктивное предположение: пусть $|u_k-u_{k-1}|\leqslant(2\mu/L)\,q^{2^{k-1}}$ при некотором $k\geqslant 1$. Отсюда и из (25) имеем $|u_{k+1}-u_k|\leqslant(L/(2\mu))\,(2\mu/L)^2\left(q^{2^{k-1}}\right)^2=(2\mu/L)\,q^{2^k}$. Оценка (26) доказана. Из (26) следует

$$|u_k-u_p|\leqslant\sum_{m=k}^{p-1}|u_{m+1}-u_m|\leqslant\sum_{m=k}^{p-1}\frac{2\mu}{L}q^{2^m}\leqslant$$
$$\leqslant\sum_{m=k}^{\infty}\frac{2\mu}{L}q^{2^m}\leqslant\frac{2\mu}{L}q^{2^k}\left(1-q^{2^k}\right)^{-1} \tag{27}$$

для всех p, k, $p > k \geqslant 0$. Так как $0 < q < 1$, то правая часть (27) стремится к нулю при $k \to \infty$. Это значит, что последовательность $\{u_k\}$ фундаментальна и сходится к некоторой точке u_*. В силу замкнутости множества U точка $u_* \in U$. Переходя к пределу при $p \to \infty$, из (27) получим оценку (20). Остается убедиться в том, что u_* — точка минимума $J(u)$ на U. Так как $J(u) \in C^2(U)$, то при $k \to \infty$ из (23) имеем $\langle J'(u_*), u - u_* \rangle \geqslant 0$ при всех $u \in U$. Учитывая выпуклость $J(u)$, отсюда и из теоремы 4.2.3 заключаем, что u_* — решение задачи (1). Теорема 2 доказана.

Из (20) при $k = 0$ имеем $|u_0 - u_*| \leqslant (2\mu/L)\, q\,(1-q)^{-1}$. Это неравенство означает, что метод (6) при $U \neq E^n$, так же как и метод (8), который получен из (6) при $U = E^n$, сходится, вообще говоря, лишь при выборе достаточно хорошего начального приближения.

4. Перейдем к рассмотрению метода (2) — (4) с выбором шага $\alpha_k = \lambda^{i_0}$, где i_0 — минимальный номер, для которого выполняется неравенство (10). Этот вариант метода Ньютона кратко будем называть методом (2) — (4), (10). Покажем, что метод (2) — (4), (10) сходится при любом выборе начального приближения и этим выгодно отличается от метода (2) — (4), (5).

Теорема 3. *Пусть U — замкнутое выпуклое множество из E^n, $J(u) \in C^2(U)$ и*

$$\mu|\xi|^2 \leqslant \langle J''(u)\xi, \xi\rangle \leqslant M|\xi|^2, \quad u \in U, \quad \xi \in L_U, \tag{28}$$

где L_U — подпространство, параллельное аффинной оболочке множества U, а μ, M — постоянные, $0 < \mu \leqslant M$. Тогда последовательность $\{u_k\}$, определяемая методом (2) — (4), (10), при любом начальном приближении $u_0 \in U$ существует и сходится к точке u_* — решению задачи (1). Если, кроме того, $J''(u)$ удовлетворяет условию Липшица (18), то найдется номер k_0 такой, что в (4) $\alpha_k = 1$ при всех $k \geqslant k_0$, и справедлива оценка

$$|u_k - u_*| \leqslant \frac{2\mu}{L} \sum_{m=}^{\infty} q^{2^m} \leqslant \frac{2\mu}{L} q^{2^k} \left(1 - q^{2^k}\right)^{-1}, \quad k \geqslant k_0. \tag{29}$$

Доказательство. Согласно теореме 4.3.4 $J(u)$ сильно выпукла. Тогда из теоремы 4.3.1 следует существование и единственность точки $\bar{u}_k$, удовлетворяющей условиям (3). Согласно теореме 4.2.3 тогда $\langle J_k'(\bar{u}_k), u - \bar{u}_k\rangle \geqslant 0$ или

$$\langle J'(u_k) + J''(u_k)(\bar{u}_k - u_k), u - \bar{u}_k\rangle \geqslant 0 \quad \text{при всех} \quad u \in U. \tag{30}$$

Если оказалось, что $\bar{u}_k = u_k$, то из (30) имеем $\langle J'(u_k), u - u_k\rangle \geqslant 0$, $u \in U$. В силу теоремы 4.2.3 и выпуклости $J(u)$ отсюда следует $\bar{u}_k = u_k = u_*$ — задача (1) решена. Поэтому можем считать, что $\bar{u}_k \neq u_k$. Тогда $J_k(\bar{u}_k) < J_k(u_k) = 0$. Покажем, что тогда существует хотя бы один номер $i \geqslant 0$, для которого выполняется условие (10). С этой целью возьмем произвольное число α $(0 \leqslant \alpha \leqslant 1)$, и положим $u_\alpha = u_k + \alpha(\bar{u}_k - u_k)$. Отсюда и из выпуклости $J_k(u)$ следует

$$J_k(u_\alpha) \leqslant \alpha J_k(\bar{u}_k) + (1-\alpha) J_k(u_k) = \alpha J_k(\bar{u}_k) < 0.$$

Тогда из формулы

$$J(u_\alpha) - J(u_k) = J_k(u_\alpha) + (\alpha^2/2)\langle (J''(u_k + \theta\alpha(\bar{u}_k - u_k)) - J''(u_k))(\bar{u}_k - u_k), \bar{u}_k - u_k\rangle, \quad 0 \leqslant \alpha \leqslant 1, \tag{31}$$

с учетом условий (28) получим

$$J(u_\alpha) - J(u_k) \leqslant J_k(u_\alpha) + (\alpha^2/2)(M - \mu)|\bar{u}_k - u_k|^2 \leqslant \alpha J_k(\bar{u}_k) + (\alpha^2/2) M |\bar{u}_k - u_k|^2, \quad 0 \leqslant \alpha \leqslant 1. \tag{32}$$

Так как $\bar{u}_k$ — точка минимума сильно выпуклой функции $J_k(u)$ на U, то согласно теореме 4.3.1

$$|u_k - \bar{u}_k|^2 \leqslant (2/\mu)\,[J_k(u_k) - J_k(\bar{u}_k)] = (2/\mu)\,|J_k(\bar{u}_k)|. \tag{33}$$

Подставив эту оценку в (32), получим

$$J(u_k) - J(u_k) \leqslant -\alpha|J_k(\bar{u}_k)| + \alpha^2(M/\mu)\,|J_k(\bar{u}_k)|, \quad 0 \leqslant \alpha \leqslant 1.$$

Возьмем произвольное α, удовлетворяющее условиям

$$0 < \varepsilon_0 = \lambda(1-\varepsilon)\mu/M \leqslant \alpha \leqslant (1-\varepsilon)\mu/M < 1. \tag{34}$$

Отсюда и из предыдущего неравенства будем иметь

$$J(u_k) - J(u_k + \alpha(\bar{u}_k - u_k)) \geqslant \alpha(1 - \alpha(M/\mu))\,|J_k(\bar{u}_k)| \geqslant \varepsilon\alpha|J_k(\bar{u}_k)| \tag{35}$$

при всех α, удовлетворяющих условиям (34). Возьмем такой номер $m \geqslant 1$, для которого $\lambda^m \leqslant (1-\varepsilon)\mu/M < \lambda^{m-1}$. Отсюда следует, что

$$0 < \varepsilon_0 = \lambda(1-\varepsilon)\mu/M < \lambda^m \leqslant (1-\varepsilon)\mu/M. \tag{36}$$

Таким образом, $\alpha = \lambda^m$ удовлетворяет условиям (34) и, следовательно, при $\alpha = \lambda^m$ будет справедливо неравенство (35). Это значит, при $i = m$ выполняется условие (10). Тогда найдется наименьший номер $i = i_0$ $(0 \leqslant i_0 \leqslant m)$, удовлетворяющий неравенству (10). Приняв в (4) $\alpha_k = \lambda^{i_0}$, получим следующее приближение u_{k+1}.

Тем самым показано, что последовательность $\{u_k\}$ из метода (2) — (4), (10) при любом начальном приближении существует. Из (10) при $i = i_0$ имеем

$$J(u_k) - J(u_{k+1}) \geqslant \varepsilon\alpha_k|J_k(\bar{u}_k)|, \qquad k = 0, 1, \ldots$$

Учитывая, что согласно (36) $\alpha_k = \lambda^{i_0} \geqslant \lambda^m > \varepsilon_0$, отсюда получим

$$J(u_k) - J(u_{k+1}) \geqslant \varepsilon\varepsilon_0|J_k(\bar{u}_k)|, \quad k = 0, 1, \ldots \tag{37}$$

Таким образом, $J(u_k) \geqslant J(u_{k+1}) \geqslant J_*$ $(k = 0, 1, \ldots)$. Тогда существует $\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$ и $\lim\limits_{k\to\infty} (J(u_k) - J(u_{k+1})) = 0$. Из (37) теперь имеем $\lim\limits_{k\to\infty} J_k(\bar{u}_k) = 0$, а из (33) следует

$$\lim_{k\to\infty} |u_k - \bar{u}_k| = 0. \tag{38}$$

Далее заметим, что согласно (37) $\{u_k\} \in M(u_0) = \{u\colon u \in U,\ J(u) \leqslant J(u_0)\}$. Для сильно выпуклых непрерывных функций множество $M(u_0)$ выпукло, замкнуто и ограничено. Тогда последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть v_* — произвольная предельная точка $\{u_k\}$ и пусть $\{u_{k_m}\} \to v_*$. С учетом (38) и условием $J(u) \in C^2(U)$ из (30) при $k = k_m \to \infty$ получим $\langle J'(v_*), u - v_*\rangle \geqslant 0$ для всех $u \in U$. Согласно теореме 4.2.3 тогда $v_* = u_*$ — точка минимума $J(u)$ на U. Следовательно, $\lim\limits_{k\to\infty} J(u_k) = \lim\limits_{m\to\infty} J(u_{k_m}) = J(u_*) = J_*$, т. е. $\{u_k\}$ — минимизирующая последовательность. Отсюда и из теоремы 4.3.1 следует $\{u_k\} \to u_*$.

Пусть теперь выполнено условие (18). В силу (38) существует номер k_0 такой, что $(L/\mu)\,|\bar{u}_k - u_k| \leqslant 1 - \varepsilon$ при всех $k \geqslant k_0$. Из (31) с учетом условия (18) и оценки (33) тогда имеем

$$J(u_\alpha) - J(u_k) \leqslant \alpha J_k(\bar{u}_k) + (\alpha^3/2)L|\bar{u}_k - u_k|^3 \leqslant$$

$$\leqslant -\alpha|J_k(\bar{u}_k)| + \alpha^2(L/\mu)\,|J_k(\bar{u}_k)|\ |\bar{u}_k - u_k|,$$

т. е.

$$J(u_k) - J(u_\alpha) \geqslant |J_k(\bar{u}_k)|\,\alpha(1 - \alpha(L/\mu)\,|\bar{u}_k - u_k|) \geqslant \alpha\varepsilon\,|J_k(\bar{u}_k)|$$

при всех α, $\varepsilon_0 \leqslant \alpha \leqslant 1$, $k \geqslant k_0$. Это означает, что условие (10) выполнено при $i = i_0 = 0$, и, следовательно, $\alpha_k = \lambda^0 = 1$, $u_{k+1} = \bar{u}_k$ при каждом $k \geqslant k_0$. Таким образом, начиная с номера $k = k_0$, метод (2) — (4), (10) превращается в метод (2) — (5) с начальным приближением u_{k_0}, удовлетворяющим условию

$$q = (L/(2\mu))\left|u_{k_0+1} - u_{k_0}\right| = (L/(2\mu))\left|\bar{u}_{k_0} - u_{k_0}\right| \leqslant (1 - \varepsilon)/2 < 1.$$

Отсюда и из теоремы 2 следует оценка (29), что и требовалось.

Таким образом, метод (2) — (4), (10) не намного сложнее метода (2) — (5), по скорости сходимости не уступает ему и в то же время не столь чувствителен к выбору начального приближения, как метод (2) — (5). При наличии эффективных методов минимизации квадратичной функции $J_k(u)$ на множестве U метод (2) — (4), (10) можно с успехом применять для минимизации достаточно гладких функций.

Другие теоремы о сходимости описанных выше вариантов метода Ньютона читатель может найти в [19].

5. Для задачи безусловной минимизации, когда в (1) $U = E^n$, метод Ньютона является частным случаем *квазиньютоновских методов*

$$u_{k+1} = u_k - \alpha_k A_k J'(u_k), \qquad \alpha_k > 0, \qquad k = 0, 1, \ldots, \tag{39}$$

в которых матрица A_k выбирается из условия

$$\lim_{k\to\infty} \left\| A_k - (J''(u_k))^{-1} \right\| = 0. \tag{40}$$

Взяв в (39) $A_k = (J''(u_k))^{-1}$, $\alpha_k = 1$, приходим к методу (8), который согласно теореме 1 имеет квадратичную скорость сходимости. Оказывается, и методы (39), в которых матрица A_k выбирается близкой к $(J''(u_k))^{-1}$ в соответствии с (40), обладают высокой скоростью сходимости. Другим достоинством методов (39) является возможность определения матриц A_k из достаточно простых рекуррентных соотношений, использующих информацию с предыдущей итерации, обходясь без вычисления и обращения матрицы $J''(u_k)$. Примером квазиньютоновского метода является *метод Давидона — Флетчера — Пауэлла*, в котором матрицы A_k определяются соотношениями

$$A_{k+1} = A_k + \frac{r_k r_k^T}{\langle r_k, q_k\rangle} - \frac{(A_k q_k)(A_k q_k)^T}{\langle A_k q_k, q_k\rangle}, \qquad k = 0, 1, \ldots; \qquad A_0 = I, \tag{41}$$

где $q_k = J'(u_{k+1}) - J'(u_k)$, $r_k = u_{k+1} - u_k$, а величина α_k находится из условия

$$f_k(\alpha_k) = \min_{\alpha\geqslant 0} f_k(\alpha), \qquad f_k(\alpha) = J(u_k - \alpha A_k J'(u_k)). \tag{42}$$

Отметим, что векторы $p_k = A_k J'(u_k)$ удовлетворяют равенствам (8.29), так что метод (39), (41), (42) одновременно является методом сопряженных направлений.

К методам (39) можно прийти, исходя из других соображений. А именно, если B — положительная симметричная матрица, то в $\mathbf{R}^n$ наряду с обычным скалярным произведением $\langle u, v\rangle = \sum_{i=1}^{n} u^i v^i$ можно ввести другое скалярное произведение $\langle u, v\rangle_1 = \langle Bu, v\rangle$. Из представления $J(u + h) - J(u) = \langle BB^{-1}J'(u), h\rangle + o(|h|) = \langle B^{-1}J'(u), h\rangle_1 + o(\langle Bh, h\rangle^{1/2})$

следует, что градиентом функции $J(u)$ в новой метрике является вектор $B^{-1}J'(u)$. Отсюда вытекает, что k-й шаг метода (39) представляет собой шаг градиентного метода в пространстве со скалярным произведением, порожденным матрицей $B = A_k^{-1} > 0$. Поэтому метод (39) часто называют *методом переменной метрики*.

С квазиньютоновскими методами, методами переменной метрики читатель подробнее познакомится в [19, 48, 111, 134, 250, 307, 314, 330, 336].

§ 10. Метод Стеффенсена

1. Описанный выше метод Ньютона на каждой итерации требует вычисления матрицы вторых производных. Отсюда ясно, что в тех случаях, когда вычисление матрицы $J''(u)$ требует значительного объема вычислений, трудоемкость каждой итерации метода Ньютона может стать чрезмерной. Поэтому возникает вопрос о возможности построения методов минимизации, которые по скорости сходимости не уступали бы методу Ньютона, но для своей реализации не требовали вычисления матрицы вторых производных. Одним из таких методов является метод Стеффенсена, представляющий собой разностный аналог метода Ньютона (9.8), в котором матрица вторых производных заменяется разностным отношением первых производных градиента по специально выбранным узловым точкам. Поначалу метод Стеффенсена разрабатывался для решения нелинейных уравнений [239], затем он был обобщен на случай операторных уравнений [301]. Применяя этот метод к решению системы уравнений $J'(u) = \{J_{u^1}(u), \ldots, J_{u^n}(u)\} = 0$, получим следующий итерационный метод решения задачи минимизации

$$J(u) \to \inf, \quad u \in U = E^n. \tag{1}$$

Если приближение u_k $(k \geqslant 0)$ уже известно, то следующее приближение u_{k+1} определяется так:

$$u_{k+1} = u_k - (J'(u_k,\ u_k - \beta_k J'(u_k)))^{-1} J'(u_k), \qquad k = 0,\ 1,\ \ldots, \tag{2}$$

где β_k — числовой параметр, $J'(u,\ v) = \{J_{ij}(u,\ v)\}$ — матрица разделенных разностей первых производных, определяемая по правилу

$$J_{ij}(u, v) = \\ = \begin{cases} \dfrac{J_{u^i}(v^1, \ldots, v^{j-1}, u^j, u^{j+1}, \ldots, u^n) - J_{u^i}(v^1, \ldots, v^{j-1}, v^j, u^{j+1}, \ldots, u^n)}{u^j - v^j}, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u^j \neq v^j, \\ J_{u^i u^j}(v^1, \ldots, v^{j-1}, v^j, u^{j+1}, \ldots, u^n), \quad u^j = v^j; \end{cases} \tag{3}$$

здесь $J_{ij}(u,\ v)$ — элемент i-й строки j-го столбца матрицы $J'(u,\ v)$,

а $J_{u^i}(u)$, $J_{u^iu^j}(u)$, как и выше, обозначают первые и соответственно вторые производные по переменным u^i, u^j функции $J(u)$ $(i, j=1, \ldots, n)$; предполагается, что $J(u)\in C^2(E^n)$. Тогда из (3) следует, что

$$\lim_{v\to u}\|J'(u,v)-J''(u)\|=0, \quad J'(u,u)=J''(u) \qquad \forall u\in E^n.$$

Это значит, что при $\beta_k=0$ $(k=0, 1, \ldots)$, метод (2) превращается в метод Ньютона (9.8).

Как видно из (2), на каждом шаге метода Стеффенсена нужно решать систему линейных алгебраических уравнений

$$J'(u_k, u_k-\beta_kJ'(u_k))\,y=-J'(u_k), \quad y=u_{k+1}-u_k. \tag{4}$$

Здесь подразумевается, что матрица $J'(u_k, u_k-\beta_kJ'(u_k))$ невырожденная. Если $\beta_k\neq 0$, $J_{u^j}(u_k)\neq 0$ при всех $j=1, \ldots, n$, то согласно формуле (3) для определения элементов матрицы $J'(u_k, u_k-\beta_kJ'(u_k))$ достаточно знать первые производные $J_{u^i}(u)$ в точках $u=u_k$ и $u=\left(u_k^1-\beta_kJ_{u^1}(u_k), \ldots, u_k^j-\beta_kJ_{u^j}(u_k), u_k^{j+1}, \ldots \ldots, u_k^n\right)$ $(j=1, \ldots, n)$. Если же $J_{u^j}(u_k)=0$ при некотором j, то в силу (3) для определения j-го столбца матрицы $J'(u_k, u_k-\beta_kJ'(u_k))$ придется вычислять вторые производные $J_{u^iu^j}(u)$ в точке $u=\left(u_k^1-\beta_kJ_{u^1}(u_k), \ldots, u_k^{j-1}-\beta_kJ_{u^{j-1}}(u_k), \quad u_k^j, \ldots, u_k^n\right)$. Если при некотором $k\geqslant 0$ оказалось $J'(u_k)=0$, то процесс (2) или (4) прекращается: в этом случае u_k — стационарная точка, и для выяснения того, будет ли $u_k\in U_*$, при необходимости нужно провести дополнительное исследование. Поэтому будем считать, что $J'(u_k)\neq 0$. А тогда для определения матрицы $J'(u_k, u_k-\beta_kJ'(u_k))$ потребуется вычислить заведомо не все вторые производные $J_{u^iu^j}(u)$, и в этом смысле метод (2) имеет преимущество перед методом Ньютона. С другой стороны, оказывается, метод (2) на классе сильно выпуклых функций сходится с такой же скоростью, как и метод Ньютона (9.8). А именно, справедлива

Теорема 1. *Пусть функция* $J(u)\in C^2(E^n)$,

$$\mu|u-v|^2\leqslant\langle J'(u)-J'(v), u-v\rangle \qquad \forall u, v\in E^n, \quad \mu>0, \tag{5}$$

$$\|J'(u,v)-J'(v,w)\|\leqslant K(|u-v|+|v-w|) \quad \forall u, v, w\in S_0, \ K>0, \tag{6}$$

где $S_0=\left\{u\in E^n: |u-u_0|\leqslant R=\max\left\{\frac{2}{\mu}+\beta; \frac{4}{\mu}\right\}|J'(u_0)|\right\}$, $|\beta_k|\leqslant \leqslant\beta$ $(k=0, 1, \ldots)$; *начальное приближение* u_0 *таково, что*

$$q=K\left(\frac{2}{\mu}+\beta\right)\frac{2}{\mu}|J'(u_0)|<1. \tag{7}$$

Тогда последовательность $\{u_k\}$, *определяемая методом* (2), *сходится к решению* u_* *задачи* (1), *причем справедлива оценка*

$$|u_k - u_*| \leqslant \frac{1}{\mu q} |J'(u_0)| q^{2^k}, \qquad k = 0, 1, \ldots \tag{8}$$

Доказательство. Из условия (5) и теоремы 4.3.3 вытекает сильная выпуклость функции $J(u)$. Отсюда и из теоремы 4.3.1 следует существование и единственность точки u_*. По индукции докажем

$$u_k \in S_0, \ a_k = |J'(u_k)| \leqslant q^{2^k - 1} |J'(u_0)|, \qquad k = 0, 1, \ldots \tag{9}$$

При $k = 0$ действительно $u_0 \in S_0$, $a_0 = q^{2^0 - 1} |J'(u_0)|$. Пусть при некотором $k \geqslant 0$ точка u_k уже найдена и справедливы соотношения (9). Поскольку $q < 1$, то из (9) следует, что $a_k \leqslant a_0$. Тогда из (5) при $u = u_k$, $v = u_0$ имеем $\mu |u_k - u_0|^2 \leqslant 2a_0 |u_k - u_0|$ или

$$|u_k - u_0| \leqslant \frac{2}{\mu} a_0. \tag{10}$$

Обозначим $x_k = u_k - \beta_k J'(u_k)$. Заметим, что в силу (10) $|x_k - u_0| \leqslant |u_k - u_0| + |\beta_k| |J'(u_k)| \leqslant \frac{2}{\mu} a_0 + \beta a_k \leqslant \left(\frac{2}{\mu} + \beta\right) a_0 \leqslant R$, так что $x_k \in S_0$. Далее, с учетом условия (7) имеем

$$|u_k - x_k| \leqslant \beta |J'(u_k)| = \beta a_k \leqslant \beta a_0 \leqslant \beta \frac{\mu}{2K(2/\mu + \beta)} \leqslant \frac{\mu}{2K}. \tag{11}$$

Отсюда и из (6) следует

$$\|J'(u_k, x_k) - J''(u_k)\| = \|J'(u_k, x_k) - J'(u_k, u_k)\| \leqslant K |u_k - x_k| \leqslant \mu/2.$$

Тогда с помощью теорем 4.3.3, 4.3.4 получаем

$$\langle J'(u_k, x_k)\xi, \xi\rangle = \langle J''(u_k)\xi, \xi\rangle + \langle (J'(u_k, x_k) - J''(u_k))\xi, \xi\rangle \geqslant$$
$$\geqslant \mu |\xi|^2 - (\mu/2)|\xi|^2 = (\mu/2)|\xi|^2 \qquad \forall \xi \in E^n. \tag{12}$$

Из (12) вытекает, что система уравнений $J'(u_k, x_k)\xi = 0$ имеет единственное решение $\xi = 0$, так что матрица $J'(u_k, x_k)$ невырожденная, и точка u_{k+1}, определяемая условиями (2) или (4), существует. Полагая в (12) $\xi = (J'(u_k, x_k))^{-1} z$, $z \in E^n$, получим $(\mu/2) |(J'(u_k, x_k))^{-1} z|^2 \leqslant \langle z, (J'(u_k, x_k))^{-1} z\rangle \leqslant |z| |(J'(u_k, x_k))^{-1} z|$ или $|(J'(u_k, x_k))^{-1} z| \leqslant (2/\mu)|z|$, $z \in E^n$. Это значит, что

$$\|(J'(u_k, x_k))^{-1}\| \leqslant 2/\mu. \tag{13}$$

Отсюда и из (2), (10) имеем

$$|u_{k+1} - u_0| \leqslant |u_{k+1} - u_k| + |u_k - u_0| \leqslant (2/\mu) a_k +$$
$$+ (2/\mu) a_0 \leqslant (4/\mu) a_0 \leqslant R,$$

т. е. $u_{k+1} \in S_0$. Далее, из (3) следует

$$\sum_{j=1}^{n} J_{ij}(u, v)(u^j - v^j) = \sum_{j=1}^{n} \left(J_{u^i}(v^1, \ldots, v^{j-1}, u^j, \ldots, u^n) - \right.$$
$$\left. - J_{u^i}(v^1, \ldots, v^{j-1}, v^j, u^{j+1}, \ldots, u^n)\right) = J_{u^i}(u) - J_{u^i}(v), \quad i = 1, \ldots, n,$$

т. е.

$$J'(u, v)(u - v) = J'(u) - J'(v) \qquad \forall u, v \in E^n. \tag{14}$$

Полагая в (14) $u = u_{k+1}$, $v = u_k$ и пользуясь (2), имеем

$$J'(u_{k+1}) = J'(u_k) + J'(u_{k+1}, u_k)(u_{k+1} - u_k) =$$
$$= (J'(u_k, x_k) - J'(u_{k+1}, u_k))(J'(u_k, x_k))^{-1} J'(u_k).$$

Отсюда с помощью (2), (6), (11), (13) и предположения индукции (9) получим

$$a_{k+1} = |J'(u_{k+1})| \leqslant K(|u_{k+1} - u_k| + |u_k - x_k|)\frac{2}{\mu} a_k \leqslant$$
$$\leqslant K\left(\frac{2}{\mu} a_k + \beta a_k\right)\frac{2}{\mu} a_k = K\left(\frac{2}{\mu} + \beta\right)\frac{2}{\mu} a_k^2 \leqslant$$
$$\leqslant K\left(\frac{2}{\mu} + \beta\right)\frac{2}{\mu}\left(q^{2^k - 1} a_0\right)^2 = q^{2^{k+1} - 1} a_0.$$

Рассуждения по индукции закончены, существование последовательности $\{u_k\}$, определяемой методом (2), и соотношения (9) доказаны. Наконец, из (5) при $u = u_k$, $v = u_*$ с учетом равенства $J'(u_*) = 0$ имеем

$$\mu |u_k - u_*|^2 \leqslant \langle J'(u_k) - J'(u_*), u_k - u_* \rangle \leqslant a_k |u_k - u_*|$$

или $|u_k - u_*| \leqslant a_k/\mu$. Отсюда и из (9) следует оценка (8). Теорема доказана.

Недостатком метода (2), как видно из (7), является требование достаточной близости начальной точки u_0 к искомой точке u_*.

2. Как мы убедились, описанные выше методы Ньютона, Стеффенсена на классе сильно выпуклых функций имеют квадратичную скорость сходимости. На основе этих методов, немного усложнив их, можно получить методы, имеющие более высокий порядок сходимости. Примером такого метода является следующий [316]

$$w_k = u_k - (J'(u_k, x_k))^{-1} J'(u_k), \qquad u_{k+1} = w_k - (J'(u_k, x_k))^{-1} J'(w_k), \tag{15}$$

где $x_k = u_k - \beta_k J'(u_k)$, β_k — числовой параметр, $k = 0, 1, \ldots$ На каждой итерации метода (15) последовательно решаются две системы линейных алгебраических уравнений вида (4) с одной и той же матрицей, но с разными правыми частями. Таким

образом, метод (15) не намного сложнее метода (2). Можно показать [83, 316], что на классе сильно выпуклых функций при выборе хорошего начального приближения метод (15) имеет кубическую скорость сходимости: $|u_k - u_*| \leqslant Cq^{3^k}$ $(k=0, 1, \ldots)$. Если в (15) считать β_k $(k=0, 1, \ldots)$, то $x_k = u_k$, $J'(u_k, x_k) = J''(u_k)$ и в результате приходим к модифицированному методу Ньютона, в котором матрица вторых производных обновляется через два шага и который также имеет кубическую скорость сходимости.

Об итерационных методах высокого порядка сходимости для решения задачи минимизации (1), для решения систем уравнений см., например, [4, 8, 19, 20, 45, 54, 73, 76, 111, 209, 238, 239, 301, 330].

§ 11. Метод покоординатного спуска

В предыдущих параграфах мы рассмотрели методы, которые для своей реализации требуют вычисления первых или вторых производных минимизируемой функции. Однако в практических задачах нередко встречаются случаи, когда минимизируемая функция либо не обладает нужной гладкостью, либо является гладкой, но вычисление ее производных с нужной точностью требует слишком большого объема работ, много машинного времени. В таких случаях желательно иметь методы минимизации, которые требуют лишь вычисления значения функции. Одним из таких методов является метод покоординатного спуска [4, 11, 153, 171, 326].

1. Сначала опишем этот метод для задачи

$$J(u) \to \inf; \quad u \in U = E^n. \tag{1}$$

Обозначим $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ — единичный координатный (базис) вектор, у которого i-я координата равна 1, остальные равны нулю $i = 1, \ldots, n$. Пусть u_0 — некоторое начальное приближение, а α_0 — некоторое положительное число, являющееся параметром метода. Допустим, что нам уже известны точка $u_k \in E^n$ и число $\alpha_k > 0$ при каком-либо $k \geqslant 0$. Положим

$$p_k = e_{i_k}, \quad i_k = k - n\left[\frac{k}{n}\right] + 1, \tag{2}$$

где $\left[\frac{k}{n}\right]$ означает целую часть числа k/n. Условие (2) обеспечивает циклический перебор координатных векторов $e_1, e_2, \ldots, e_n$, т. е.

$$p_0 = e_1, \ldots, p_{n-1} = e_n, \quad p_n = e_1, \ldots, p_{2n-1} = e_n,$$
$$p_{2n} = e_1, \ldots$$

Вычислим значение функции $J(u)$ в точке $u = u_k + \alpha_k p_k$ и про-

верим неравенство

$$J(u_k+\alpha_k p_k)<J(u_k). \tag{3}$$

Если (3) выполняется, то примем

$$u_{k+1}=u_k+\alpha_k p_k, \quad \alpha_{k+1}=\alpha_k. \tag{4}$$

В том случае, если (3) не выполняется, то вычисляем значение функции $J(u)$ в точке $u=u_k-\alpha_k p_k$ и проверяем неравенство

$$J(u_k-\alpha_k p_k)<J(u_k). \tag{5}$$

В случае выполнения (5) положим

$$u_{k+1}=u_k-\alpha_k p_k, \quad \alpha_{k+1}=\alpha_k. \tag{6}$$

Назовем $(k+1)$-ю итерацию удачной, если справедливо хотя бы одно из неравенств (3) или (5). Если $(k+1)$-я итерация неудачная, т. е. не выполняются оба неравенства (3) и (5), то полагаем

$$u_{k+1}=u_k, \quad \alpha_{k+1}=\begin{cases}\lambda\alpha_k, & i_k=n,\ u_k=u_{k-n+1},\\ \alpha_k, & i_k\neq n \text{ или } u_k\neq u_{k-n+1},\\ & \text{или } 0\leqslant k\leqslant n-1;\end{cases} \tag{7}$$

здесь λ $(0<\lambda<1)$ — фиксированное число, являющееся параметром метода. Условия (7) означают, что если за один цикл из n итераций при переборе направлений всех координатных осей $e_1, \ldots, e_n$ с шагом α_k реализовалась хотя бы одна удачная итерация, то длина шага α_k не дробится и сохраняется на протяжении по крайней мере следующего цикла из n итераций. Если же среди последних n итераций не оказалось ни одной удачной итерации, то шаг α_k дробится. Таким образом, если на итерации с номером $k=k_m$ произошло дробление α_k, то

$$J(u_{k_m}+\alpha_{k_m}e_i)\geqslant J(u_{k_m}), \quad J(u_{k_m}-\alpha_{k_m}e_i)\geqslant J(u_{k_m}) \tag{8}$$

при всех $i=1, 2, \ldots, n$. Метод покоординатного спуска для задачи (1) описан. Справедлива

Теорема 1. *Пусть функция $J(u)$ выпукла на E^n и принадлежит классу $C^1(E^n)$, а начальное приближение u_0 таково, что множество $M(u_0)=\{u\in E^n\colon J(u)\leqslant J(u_0)\}$ ограничено. Тогда последовательность $\{u_k\}$, получаемая описанным методом* (2)—(7), *минимизирует функцию $J(u)$ на E^n и сходится ко множеству U_*.*

Доказательство. Согласно теореме 2.1.2 $J_*>-\infty$, $U_*\neq\varnothing$. Из описания метода (2)—(7) следует, что $J(u_{k+1})\leqslant J(u_k)$ $(k=0, 1, \ldots)$, так что $\{u_k\}\in M(u_0)$ и существует $\lim\limits_{k\to\infty} J(u_k)\geqslant J_*$. Покажем, что найдется бесконечно много номеров $k_1, \ldots, k_m, \ldots$ итераций, на которых шаг α_k дробится, и поэтому

$\lim\limits_{k\to\infty} \alpha_k = 0$. Допустим противное: пусть процесс дробления конечен, т. е. $\alpha_k = \alpha > 0$ при всех $k \geqslant N$. Обозначим $M_\alpha = \{u\colon u = u_N + \alpha r e_i \in M(u_0),\ u = 1, \ldots, n,\ r = 0, \pm 1, \pm 2, \ldots\}$ — сетку (решетку) с шагом α. Из описания метода покоординатного спуска при $\alpha_k = \alpha$, $k \geqslant N$, следует, что начиная с номера N все последующие циклы из n итераций будут содержать хотя бы одну удачную итерацию, и на каждой удачной итерации будет происходить переход от одной точки сетки M_α к другой соседней точке этой сетки. По определению удачной итерации переход от точки к точке сопровождается строгим уменьшением значения функции $J(u)$, поэтому каждая точка сетки M_α будет просматриваться не более одного раза. Но множество $M(u_0)$ по условию ограничено, и поэтому сетка M_α состоит из конечного числа точек. Следовательно, процесс перебора точек этой сетки закончится через конечное число итераций определением точки u_{k_m} $(k_m > N)$, для которой выполняются неравенства (8) при всех $i = 1, \ldots, n$. А тогда вопреки допущению придется дробить число $\alpha_k = \alpha$. Полученное противоречие показывает, что процесс дробления α_k бесконечен и $\lim\limits_{k\to\infty} \alpha_k = 0$.

Пусть $k_1 < k_2 < \ldots < k_m < \ldots$ — номера тех итераций, на которых длина шага α_k дробится и выполняются неравенства (8). Так как последовательность $\{u_k\}$ принадлежит ограниченному множеству $M(u_0)$, то из $\{u_{k_m}\}$ можно выбрать сходящуюся подпоследовательность. Без умаления общности можем считать, что сама подпоследовательность $\{u_{k_m}\}$ сходится к некоторой точке u_*. С помощью формулы конечных приращений из (8) имеем

$$\langle J'(u_{k_m} + \theta_m \alpha_{k_m} e_i), e_i\rangle \alpha_{k_m} \geqslant 0,$$

$$\langle J'(u_{k_m} - \bar\theta_m \alpha_{k_m} e_i), e_i\rangle (-\alpha_{k_m}) \geqslant 0,$$

откуда

$$J_{u^i}(u_{k_m} + \theta_m \alpha_{k_m} e_i) \geqslant 0, \quad J_{u^i}\left(u_{k_m} - \bar\theta_m \alpha_{k_m} e_i\right) \leqslant 0,$$

$0 \leqslant \theta_m$, $\bar\theta_m \leqslant 1$ при всех $i = 1, \ldots, n$ и $m = 1, 2, \ldots$ Пользуясь тем, что $J(u) \in C^1(E^n)$ и $\lim\limits_{m\to\infty} \alpha_{k_m} = 0$, отсюда получим $J_{u^i}(u_*) = 0$ $(i = 1, \ldots, n)$, т.е. $J'(u_*) = 0$. В силу выпуклости $J(u)$ тогда $u_* \in U_*$. Следовательно, $\lim\limits_{k\to\infty} J(u_k) = \lim\limits_{m\to\infty} J(u_{k_m}) = J(u_*) = J_*$. Таким образом, последовательность $\{u_k\}$ является минимизирующей. Отсюда и из теоремы 2.1.2 следует, что $\rho(u_k, U_*) \to 0$ при $k \to \infty$. Теорема доказана.

Заметим, что хотя метод (2)—(7) для своей реализации не требует знания градиента минимизируемой функции, однако в условии теоремы 1 содержится требование гладкости этой функции. Оказывается, если функция $J(u)$ не является гладкой, то

метод покоординатного спуска может не сходиться ко множеству решений задачи (1). Об этом говорит следующий

Пример 1. Пусть

$$J(u)=x^2+y^2-2(x-y)+2|x-y|, \quad u=(x, y)\in E^2.$$

Нетрудно проверить, что $J(u)$ сильно выпукла на E^2 и, следовательно, ограничена снизу и достигает своей нижней грани на E^2 в единственной точке. Возьмем в качестве начального приближения точку $u_0=(0, 0)$. Тогда имеем $J(u_0+\alpha e_1)=$ $=J(\alpha e_1)=\alpha^2-2\alpha+2|\alpha|\geqslant 0=J(0)$, $J(u_0+\alpha e_2)=J(\alpha e_2)=\alpha^2-$ $-2\alpha+2|\alpha|\geqslant 0=J(0)$ при всех действительных α. Отсюда следует, что все итерации метода (2)—(7) при начальной точке $u_0=(0, 0)$ и любом выборе начального параметра $\alpha=\alpha_0>0$ будут неудачными, т. е. $u_k=u_0$ при всех $k=0, 1, \ldots$ Однако в точке $u_0=(0, 0)$ функция $J(u)$ не достигает своей нижней грани на E^2: например, в точке $v=(1, 1)$ имеем $J(v)=-2<$ $<J(u_0)=0$.

2. Описанный выше метод покоординатного спуска нетрудно модифицировать применительно к задаче минимизации функции на параллелепипеде:

$$J(u)\to\inf; \quad u\in U=\{(u^1, \ldots, u^n): a_i\leqslant u^i\leqslant b_i, \ i=1, \ldots, n\}, \tag{9}$$

где a_i, b_i — заданные числа, $a_i<b_i$ $(i=1, \ldots, n)$. А именно, пусть k-е приближение $u_k\in U$ и число $\alpha_k>0$ при некотором $k\geqslant 0$ уже найдены. Выберем вектор $p_k=e_{i_k}$ согласно формуле (2), составим точку $u_k+\alpha_k p_k$ и проверим условия

$$u_k+\alpha_k p_k\in U, \quad J(u_k+\alpha_k p_k)<J(u_k). \tag{10}$$

Если оба условия (10) выполняются, то следующее приближение u_{k+1}, α_{k+1} определяем по формулам (4). Если же хотя бы одно условие (10) не выполняется, то составляем точку u_k- $-\alpha_k p_k$ и проверяем условия

$$u_k-\alpha_k p_k\in U, \quad J(u_k-\alpha_k p_k)<J(u_k). \tag{11}$$

В случае выполнения обоих условий (11) следующее приближение определяем по формулам (6), а если хотя бы одно из условий (11) не выполняется, то следующее приближение находится из неравенств (7).

Теорема 2. *Пусть функция $J(u)$ выпукла на U и $J(u)\in$ $\in C^1(U)$. Тогда при любом выборе начальных $u_0\in U$ и $\alpha_0>0$ последовательность $\{u_k\}$, получаемая методом* (10), (4), (11), (6), (7), *минимизирует функцию $J(u)$ на U и сходится ко множеству решений задачи* (9).

Доказательство. Так как U — параллелепипед, то множество $M(u_0)=\{u: u\in U, J(u)\leqslant J(u_0)\}$ ограничено. Так как $J(u_{k+1})\leqslant J(u_k)$ $(k=0, 1, \ldots)$, то $\{u_k\}\in U$ и существует

$\lim\limits_{k\to\infty} J(u_k) \geqslant J_*$. Так же, как в теореме 1, доказывается существование бесконечного числа номеров $k_1 < \ldots < k_m < \ldots$ итераций, на которых длина шага α_k дробится, и поэтому $\lim\limits_{k\to\infty} \alpha_k = 0$. В силу ограниченности $M(u_0)$ из $\{u_{k_m}\}$ можно выбрать сходящуюся подпоследовательность. Не умаляя общности, можем считать, что $\{u_{k_m}\} \to u_* = (u_*^1, \ldots, u_*^n)$. При каждом $i = 1, \ldots, n$ возможны следующие три случая.

1) $a_i < u_*^i < b_i$. Так как $\lim\limits_{k\to\infty} \alpha_k = 0$, то найдется номер N такой, что $u_{k_m} + \alpha_{k_m} e_i \in U$ и $u_{k_m} - \alpha_{k_m} e_i \in U$ при всех $m \geqslant N$. Поскольку α_k при $k = k_m$ дробится, то

$$J(u_{k_m} + \alpha_{k_m} e_i) \geqslant J(u_{k_m}), \quad J(u_{k_m} - \alpha_{k_m} e_i) \geqslant J(u_{k_m})$$

для всех $m \geqslant N$. Отсюда, как и в теореме 1, получаем $J_{u^i}(u_*) = 0$, так что

$$J_{u^i}(u_*)(u^i - u_*^i) = 0, \quad a_i \leqslant u^i \leqslant b_i.$$

2) $u_*^i = a_i$. Тогда $u_{k_m} + \alpha_{k_m} e_i \in U$ и $J(u_{k_m} + \alpha_{k_m} e_i) \geqslant J(u_{k_m})$ при всех $m \geqslant N$. Следовательно, $\langle J'(u_{k_m} + \theta_m \alpha_{k_m} e_i) \rangle \alpha_{k_m} \geqslant 0$ или $J_{u^i}(u_{k_m} + \theta_m \alpha_{k_m} e_i) \geqslant 0$ для каждого $m \geqslant N$. Отсюда при $m \to \infty$ получим $J_{u^i}(u_*) \geqslant 0$ или $J_{u^i}(u_*)(u^i - a_i) = J_{u^i}(u_*)(u^i - u_*^i) \geqslant 0 \quad (a_i \leqslant u^i \leqslant b_i)$.

3) $u_*^i = b_i$. Тогда $u_{k_m} - \alpha_{k_m} e_i \in U$ и $J(u_{k_m} - \alpha_{k_m} e_i) \geqslant J(u_{k_m})$ при всех $m \geqslant N$. Поэтому $\langle J(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i), e_i \rangle (-\alpha_{k_m}) \geqslant 0$ или $J_{u^i}\left(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i\right) \leqslant 0 \quad (m \geqslant N)$. Отсюда при $m \to \infty$ получим $J_{u^i}(u_*) \leqslant 0$, следовательно,

$$J_{u^i}(u_*)(u^i - b_i) = J_{u^i}(u_*)(u^i - u_*^i) \geqslant 0, \quad a_i \leqslant u^i \leqslant b_i.$$

Объединяя все три рассмотренных случая, заключаем, что

$$J_{u^i}(u_*)(u^i - u_*^i) \geqslant 0, \quad a_i \leqslant u^i \leqslant b_i, \quad i = 1, \ldots, n.$$

Суммируя эти неравенства по всем $i = 1, \ldots, n$, получим

$$\langle J'(u_*), u - u_* \rangle \geqslant 0 \text{ для всех } u \in U.$$

Согласно теореме 4.2.3 тогда $u_* \in U_*$. Следовательно, $\lim\limits_{k\to\infty} J(u_k) = \lim\limits_{m\to\infty} J(u_{k_m}) = J(u_*) = J_*$, т. е. $\{u_k\}$ — минимизирующая последовательность. Отсюда и из теоремы 2.1.2 следует, что $\lim\limits_{k\to\infty} \rho(u_k, U_*) = 0$. Теорема 2 доказана.

3. Существуют и другие варианты метода покоординатного спуска. Можно, например, строить последовательность $\{u_k\}$ по

правилу

$$u_{k+1} = u_k + \alpha_k p_k, \tag{12}$$

где p_k определяется согласно (2), а α_k — условиями

$$\alpha_k \geqslant 0, \quad f_k(\alpha_k) = \min_{-\infty<\alpha<+\infty} f_k(\alpha), \quad f_k(\alpha) = J(u_k + \alpha p_k). \tag{13}$$

Метод (12), (13) имеет смысл применять в том случае, когда величина α_k из (13) находится в явном виде. Так будет, если функция $J(u)$ — квадратичная, т. е.

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle, \quad u \in E^n, \tag{14}$$

где A — симметричная положительно определенная матрица $b \in E^n$. Нетрудно убедиться, что для функции (14) метод (12), (13) приводит к хорошо известному методу Зейделя из линейной алгебры [4].

Хотя и скорость сходимости метода покоординатного спуска, вообще говоря, невысокая, благодаря простоте каждой итерации, скромным требованиям к гладкости минимизируемой функции этот метод довольно широко применяется на практике.

Существуют и другие методы минимизации, использующие лишь значения функции и не требующие для своей реализации вычисления производных. Например, используя вместо производных их разностные аппроксимации, можно построить модификации рассматривавшихся в предыдущих параграфах методов, требующие вычисления лишь значений функции в подходящим образом выбранных точках (ср. с § 9, 10).

Другой подход для минимизации негладких функций, основанный лишь на вычислении значений функции, дает метод случайного поиска, который будет рассмотрен ниже в § 17. Метод поиска глобального минимума, излагаемый в следующем параграфе, также относится к методам, не требующим вычисления производных минимизируемой функции.

Упражнения. 1. Нарисуйте линии уровня $J(u) = C = \text{const}$ функции из примера 1 и поясните причину расходимости метода покоординатного спуска для этой функции при выборе $u_0 = (0, 0)$.

2. Опишите метод покоординатного спуска и докажите его сходимость для случая, когда в задаче (9) $a_i = -\infty$ или $b_j = \infty$ для каких-либо i, j, $1 \leqslant i, j \leqslant n$.

3. Докажите сходимость метода (12), (13) для функции (14) [4].

§ 12. Метод поиска глобального минимума

1. Заметим, что если задача минимизации

$$J(u) \to \inf; \quad u \in U \tag{1}$$

является *многоэкстремальной* — так называются задачи, в которых имеется хотя бы одна точка локального минимума, отличная от точки глобального минимума, то с помощью описанных

выше методов удается найти, вообще говоря, лишь приближение к какой-либо точке локального минимума. Поэтому упомянутые методы часто называют локальными методами.

Для решения многоэкстремальной задачи (1) локальные методы обычно используются по следующей схеме: на множестве U задают некоторую сетку точек и, выбирая в качестве начальных приближений точки этой сетки, с помощью того или иного локального метода находят локальные минимумы функции, а затем, сравнивая полученные результаты, определяют ее глобальный минимум. Однако ясно, что такой подход к решению многоэкстремальных задач весьма трудоемок и не всегда приводит к цели. Поэтому представляют большой интерес методы поиска глобального минимума в многоэкстремальных задачах.

Если относительно свойств функции $J(u)$ ничего неизвестно, то вряд ли можно предложить какой-либо подход к решению задачи (1), кроме вычисления значений функции $J(u)$ во всех точках $u \in U$. Понятно, что такой подход практически нереализуем. И как показывает наш предыдущий опыт, для построения эффективных численных методов решения задачи (1) на функцию, а также на множество приходится накладывать те или иные ограничения.

Ниже будут изложены методы поиска глобального минимума в задаче (1) для случая, когда множество U является параллелепипедом, т. е.

$$U = \{u = (u^1, u^2, \ldots, u^n) : a_i \leqslant u^i \leqslant b_i, \ i = 1, \ldots, n\}, \qquad (2)$$

a_i, b_i — заданные числа, $a_i < b_i$ $(i = 1, \ldots, n)$, а функция $J(u)$ удовлетворяет условию Липшица

$$|J(u) - J(v)| \leqslant L|u - v| \quad \forall u, v \in U, \quad L = \text{const} \geqslant 0. \qquad (3)$$

Эти методы являются обобщением методов покрытий, рассмотренных в § 1.7 для минимизации функции одной переменной на отрезке. Как и в § 1.7, через $Q(L)$ обозначим класс функций, удовлетворяющих условию (3) с одной и той же для всех функций этого класса константой $L \geqslant 0$.

Пусть p_m — какой-либо метод, представляющий собой правило выбора m точек $u_1, \ldots, u_m$ из U, в которых вычисляются значения функции $J(u_1), \ldots, J(u_m)$ и затем определяется величина $\min\limits_{1 \leqslant i \leqslant m} J(u_i)$, принимаемая за приближенное значение $J_* = \inf\limits_U J(u)$. Зададимся вопросом: как выбрать число m и метод $p_m = \{u_1, \ldots, u_m\}$ так, чтобы

$$\min_{1 \leqslant i \leqslant n} J(u_i) \leqslant J_* + \varepsilon \qquad \forall J(u) \in Q(L), \qquad (4)$$

где $\varepsilon > 0$ — заданная точность? Поставленную задачу будем кратко называть задачей (1) — (4).

2. Можно предложить следующее правило выбора точек $u_1, \ldots, u_m$: пусть эти точки из U таковы, что объединение шаров

$$S(u_i, R) = \{u \in E^n\colon\ |u - u_i| \leqslant R\}, \quad i = 1, \ldots, m,$$

с центрами в точках u_i и радиуса $R = \varepsilon/L$ покрывает множество U, т. е. $U \subset \bigcup\limits_{i=1}^{m} S(u_i, R)$. Оказывается, при таком выборе точек $u_1, \ldots, u_m$, приняв величину $\min\limits_{1 \leqslant i \leqslant m} J(u_i)$ за приближение к J_*, мы решим задачу (1)—(4). В самом деле, возьмем любую точку $u \in U$. Так как шары $S(u_i, R)$ $(i = 1, \ldots, m)$, покрывают множество U, то точка u принадлежит одному из шаров $S(u_i, R)$, т. е. $|u - u_i| \leqslant R = \varepsilon/L$. Отсюда и из условия (3) следует $J(u) \geqslant J(u_i) - L|u - u_i| \geqslant \min\limits_{1 \leqslant i \leqslant m} J(u_i) - \varepsilon$ или $J(u) \geqslant \min\limits_{1 \leqslant i \leqslant m} J(u_i) - \varepsilon$ для всех $u \in U$. Переходя в левой части этого неравенства к нижней грани по $u \in U$, будем иметь $0 \leqslant \min\limits_{1 \leqslant i \leqslant m} J(u_i) - J_* \leqslant \varepsilon$ для каждой функции $J(u) \in Q(L)$. Это значит, что для описанного метода p_m выполняется неравенство (4), что и требовалось.

Однако покрывать множество U шарами не очень-то удобно. Гораздо проще и удобнее покрывать множество (2) параллелепипедами или кубами. Заметим, что шар $S(u_i, R)$ содержит в себе n-мерный куб

$$V_i = \{u = (u^1, \ldots, u^n)\colon\ u_i^j - h \leqslant u^j \leqslant u_i^j + h, \quad j = 1, \ldots, n\},$$

$$h = R/\sqrt{n} = \varepsilon/(L\sqrt{n})$$

с длиной ребра $2h$ и с центром в точке $u_i = (u_i^1, \ldots, u_i^n)$. Пользуясь этим обстоятельством, нетрудно покрыть параллелепипед (2) кубами следующим образом. Возьмем точки

$$u_{i_1, \ldots, i_n} = \left(u_{i_1}^1, \ldots, u_{i_j}^j, \ldots, u_{i_n}^n\right), \quad i_j = 1, \ldots, m_j, \quad j = 1, \ldots, n, \tag{5}$$

где j-е координаты $u_1^j, \ldots, u_{m_j}^j$ образованы по правилу

$$u_i^j = a_j + h(2i - 1), \quad i = 1, \ldots, m_j - 1,$$

$$u_{m_j}^j = \min\{b_j;\ a_j + h(2m_j - 1)\},$$

а номер m_j определяется условием

$$u_{m_j - 1}^j < b_j - h \leqslant a_j + h(2m_j - 1) = u_{m_j - 1}^j + 2h.$$

Тогда кубы $V_{i_1, \ldots, i_n}$ с длиной ребра $2h = 2R/\sqrt{n}$ и с центром в точке $u_{i_1, \ldots, i_n}$ $(i = 1, \ldots, m_j,\ j = 1, \ldots, n)$ покроют параллелепипед (2). Так как куб $V_{i_1, \ldots, i_n}$ принадлежит шару $S(u_{i_1, \ldots, i_n}, R)$,

то объединение шаров $S(u_{i_1,\ldots,i_n}, R)$ $(i_j = 1, \ldots, m_j,\ j = 1, \ldots, n)$ покроет параллелепипед (2). Тогда, как было показано выше, метод p_m, заключающийся в выборе точек $\{u_1, \ldots, u_m\}$, $m = m_1 m_2 \ldots$ по правилу (5), решает задачу (1)—(4) на классе функций $Q(L)$.

3. Рассмотренный выше метод (5) относится к так называемым пассивным методам — в нем все точки $u_1, \ldots, u_m$ задаются одновременно до начала вычислений значений функции. Между тем, если точки $u_1, \ldots, u_m$ выбирать последовательно и при выборе очередной точки u_i как-то учитывать результаты вычислений в предыдущих точках $u_1, \ldots, u_{i-1}$, то шансы найти величину J_* с той же точностью $\varepsilon > 0$ за меньшее число вычислений, чем в описанном пассивном методе (5), имеются. Следуя [139], опишем один такой последовательный метод, одномерный вариант которого был изложен в § 1.7 (см. метод (1.7.3)).

Для простоты и наглядности этот метод сначала изложим для случая $n = 2$, когда

$$U = \{u = (u^1, u^2):\ a_1 \leqslant u^1 \leqslant b_1,\ a_2 \leqslant u^2 \leqslant b_2\}$$

— прямоугольник на плоскости E^2. Введем обозначения

$$\begin{aligned} &F_i = \min\{F_0, J(u_1), \ldots, J(u_i)\}, \\ &R = \varepsilon/L, \quad h = R/\sqrt{2} = \varepsilon/(L\sqrt{2}), \\ &R_i = R + (J(u_i) - F_i)/L, \quad h_i = R_i/\sqrt{2}, \quad i = 1, 2, \ldots, \end{aligned} \tag{6}$$

где точки $u_1, u_2, \ldots$ будут выбираться последовательно по правилам, указанным ниже; в качестве F_0 пока можем взять $F_0 = J(u_1)$, а вопрос о других, более удобных способах выбора F_0 будет обсужден ниже.

Сначала последовательно определим точки

$$u_1, \ldots, u_{m_1}, \quad u_i = (v_i^1, v_1^2), \quad i = 1, \ldots, m_1,$$

где

$$\begin{aligned} &v_1^1 = a_1 + h, \quad v_{i+1}^1 = v_i^1 + h + h_i, \quad i = 1, \ldots, m_1 - 2; \\ &v_{m_1}^1 = \min\left\{b_1;\ v_{m_1-1}^1 + h + h_{m_1-1}\right\}, \quad v_1^2 = a_2 + h, \end{aligned} \tag{7}$$

номер m_1 находится из условий

$$v_{m_1-1}^1 < b_1 - h \leqslant v_{m_1-1}^1 + h + h_{m_1-1}.$$

Затем положим $d_{m_1} = \min\{h_1, \ldots, h_{m_1}\}$ и введем прямоугольник

$$\Pi_{m_1} = \left\{u = (u^1, u^2):\ a_1 \leqslant u^1 \leqslant b_1,\quad a_2 \leqslant u^2 \leqslant v_1^2 + d_{m_1}\right\}.$$

Так как система отрезков $\left[v_i^1 - h, v_i^1 + h_i\right]$ $(i = 1, \ldots, m_1)$ покрывает отрезок $[a_1, b_1]$, то прямоугольник Π_{m_1} будет принадлежать

объединению прямоугольников

$$\Pi_{m_1,i} = \Big\{u = (u^1, u^2)\colon\ v_i^1 - h \leqslant u^1 \leqslant v_i^1 + h_i,$$
$$v_1^2 - h = a_2 \leqslant u^2 \leqslant v_1^2 + d_{m_1}\Big\}, \qquad i = 1, \ldots, m_1.$$

Так как $h \leqslant d_{m_1} \leqslant h_i$ $(i = 1, \ldots, m_1)$, то для любой точки $u = (u^1, u^2) \in \Pi_{m_1,i}$ имеем $|u^1 - v_i^1| \leqslant h_i = R_i/\sqrt{2}$, $|u^2 - v_1^2| \leqslant R_i/\sqrt{2}$, так что $|u - u_i| = (|u^1 - v_i^1|^2 + |u^2 - v_1^2|^2)^{1/2} \leqslant R_i$. Это значит, что $\Pi_{m_1,i}$ принадлежит шару $S(u_i, R_i)$ $(i = 1, \ldots, m_1)$. Отсюда следует, что прямоугольник Π_{m_1} покрывается системой шаров $S(u_i, R_i)$ $(i = 1, \ldots, m_1)$. Возьмем произвольную точку $u \in \Pi_{m_1} \cap U$. Тогда найдется шар $S(u_i, R_i)$, содержащий эту точку, т. е. $|u - u_i| \leqslant R_i$. Отсюда, учитывая условие (3) и обозначения (6), получим

$$J(u) \geqslant J(u_i) - L|u - u_i| \geqslant J(u_i) - LR_i = F_i - \varepsilon \geqslant F_{m_1} - \varepsilon$$

или

$$J(u) \geqslant F_{m_1} - \varepsilon \quad \text{при всех} \quad u \in \Pi_{m_1} \cap U. \tag{8}$$

Если $U \subset \Pi_{m_1}$, то отсюда получим $0 \leqslant F_m - J_* \leqslant \varepsilon$ — задача (1)—(4) решена. Допустим, что $U \not\subset \Pi_{m_1}$. Тогда последовательно введем точки

$$u_{m_1+1}, \ldots, u_{m_2}, \quad u_{m_1+i} = \left(v_{m_1+i}^1, v_2^2\right), \quad i = 1, \ldots, m_2 - m_1,$$

где

$$v_{m_1+1}^1 = a_1 + h,\ v_{m_1+i+1}^1 = v_{m_1+i}^1 + h + h_{m_1+i}, \quad i = 1, \ldots, m_2 - m_1 - 2,$$
$$v_{m_2}^1 = \min\left\{b_1;\ v_{m_2-1}^1 + h + h_{m_2-1}\right\}, \tag{9}$$
$$v_2^2 = v_1^2 + h + d_{m_1},$$

а номер m_2 определяется условиями

$$v_{m_2-1}^1 < b - h \leqslant v_{m_2-1}^1 + h + h_{m_2-1}.$$

Положим $d_{m_2} = \min\{h_{m_1+1}, \ldots, h_{m_2}\}$ и введем прямоугольник

$$\Pi_{m_2} = \Big\{u = (u^1, u^2)\colon\ a_1 \leqslant u^1 \leqslant b_1,\ v_2^2 - h \leqslant u^2 \leqslant v_2^2 + d_{m_2}\Big\},$$

принадлежащий объединению прямоугольников

$$\Pi_{m_2,i} = \Big\{u = (u^1, u^2)\colon\ v_{m_1+i}^1 - h \leqslant u^1 \leqslant v_{m_1+i}^1 + h_{m_1+i},$$
$$v_2^2 - h \leqslant u^2 \leqslant v_2^2 + d_{m_2}\Big\}, \qquad i = 1, \ldots, m_2 - m_1.$$

Как и выше, показываем, что $\Pi_{m_2,i} \subset S(u_{m_1+i}, R_{m_1+i})$ $(i = 1, \ldots, m_2 - m_1)$, и, кроме того,

$$J(u) \geqslant F_{m_2} - \varepsilon \quad \text{при всех} \quad u \in \Pi_{m_2} \cap U.$$

Поскольку $F_{m_1} \geqslant F_{m_2}$, то объединяя последнюю оценку для $J(u)$ с оценкой (8), имеем

$$J(u) \geqslant F_{m_2} - \varepsilon \quad \text{при всех} \quad u \in Q_{m_2} \cap U, \tag{10}$$

где $Q_{m_2} = Q_{m_1} \cup \Pi_{m_2}$, $Q_{m_1} = \Pi_{m_1}$.

Если $U \subset Q_{m_2}$, то из (10) следует $0 \leqslant F_{m_2} - J_* \leqslant \varepsilon$ — задача (1)—(4) решена. Если же $U \not\subset Q_{m_2}$, то процесс продолжаем дальше.

Пусть точки $u_1, \ldots, u_{m_s}$ и величины $F_1, \ldots, F_{m_s}$, $d_{m_1}, \ldots, d_{m_s}$ уже найдены, прямоугольник $Q_{m_s} = Q_{m_{s-1}} \cup \Pi_{m_s}$ рассмотрен и показано, что

$$J(u) \geqslant F_{m_s} - \varepsilon, \quad u \in Q_{m_s} \cap U. \tag{11}$$

Если $U \not\subset Q_{m_s}$, то рассматриваем точки

$$u_{m_s+1}, \ldots, u_{m_{s+1}}, \quad u_{m_s+i} = \left(v^1_{m_s+i}, v^2_{s+1}\right),$$
$$i = 1, \ldots, m_{s+1} - m_s,$$

где координаты $v^1_{m_s+i}$ вычисляются по формулам, аналогичным (7), (9), $v^2_{s+1} = v^2_s + h + d_{m_s}$. Затем полагаем $d_{m_{s+1}} = \min\{h_{m_s+1}, \ldots, h_{m_{s+1}}\}$, вводим прямоугольник

$$\Pi_{m_{s+1}} = \left\{u = (u^1, u^2)\colon\ a_1 \leqslant u^1 \leqslant b_1,\ v^2_{s+1} - h \leqslant u^2 \leqslant v^2_{s+1} + d_{m_{s+1}}\right\}.$$

Аналогично тому, как это делалось выше, устанавливаем, что этот прямоугольник покрывается шарами $S(u_{m_s+i}, R_{m_s+i})$ $(i = 1, \ldots, m_{s+1} - m_s)$ и отсюда получаем неравенство

$$J(u) \geqslant F_{m_s+1} - \varepsilon, \quad u \in Q_{m_{s+1}} \cap U, \quad Q_{m_s+1} = Q_{m_s} \cup \Pi_{m_{s+1}}$$

и т. д. Нетрудно видеть, что этот процесс закончится за конечное число шагов при таком s, для которого $v^2_{s-1} < b_2 - h \leqslant v^2_{s-1} + h + d_{m_s}$, $v^2_s = \min\left\{b_2;\ v^2_{s-1} + h + d_{m_s}\right\}$, $U \subset Q_{m_s}$, неравенство (11) выполняется при всех $u \in U$. Из (11) следует $0 \leqslant F_{m_s} - J_* = \min\{J(u_1), \ldots, J(u_{m_s})\} - J_* \leqslant \varepsilon$ для любой функции $J = J(u) \in Q(L)$. Последовательный метод для решения задачи (1)—(4) для случая $n = 2$ описан.

Кратко опишем такой метод в общем случае, когда $n \geqslant 2$. Аналогично (6) введем обозначения

$$F_i = \min\{F_0;\ J(u_1),\ \ldots,\ J(u_i)\},\quad R = \varepsilon/L,\qquad h = R/\sqrt{n} = \varepsilon/(L\sqrt{n}),\tag{12}$$

$$R_i = R + (J(u_1) - F_i)/L,\ h_i = R_i/\sqrt{n},\quad i = 1,\ 2,\ \ldots,$$

где точки $u_1, \ldots, u_i, \ldots$ выбираются последовательно по следующим правилам. Сначала определяем точки

$$u_1, \ldots, u_{m_1},\quad u_i = (v_i^1, \ldots, v_i^n),\qquad i = 1, \ldots, m_1,$$

где номер m_1 и координаты v_i^1 $(i = 1, \ldots, m_1)$, вычисляются по формулам (7), причем величины h, h_i берутся из (12), а $v_i^j = a_j + h$ $(j = 2, \ldots, n)$. Полагаем $d_{m_1}^1 = \min\{h_1, \ldots, h_{m_1}\}$ и, как и выше, доказываем, что

$$J(u) \geqslant F_{m_1} - \varepsilon,\quad u \in Q_{m_1} \cap U,$$

где $Q_{m_1} = \Pi_{m_1} = \Big\{u = (u^1, \ldots, u^n)\colon\ a_1 \leqslant u^1 \leqslant b_1,\quad a_j \leqslant u^j \leqslant a_j + h + d_{m_1}^1,\quad j = 2, \ldots, n\Big\}$. Если параллелепипед Q_{m_1} не покрывает параллелепипед (2), то далее рассматриваем точки

$$u_{m_1+1}, \ldots, u_{m_2},\quad u_{m_1+i} = \Big(v_{m_1+i}^1, \ldots, v_{m_1+i}^n\Big),\qquad i = 1, \ldots, m_2 - m_1,$$

где номер m_2 и координаты $v_{m_1+i}^1$ $(i = 1, \ldots, m_2 - m_1)$, определяются формулами (9), $v_{m_1+i}^2 = v_{m_1}^2 + h + d_{m_1}^1$, $v_{m_1+i}^j = a_j + h$ $(j = 3, \ldots, n;\quad i = 1, \ldots, m_2 - m_1)$. Далее, полагаем $d_{m_2}^1 = \min\{h_{m_1+1}, \ldots, h_{m_2}\}$, $d_{m_2}^2 = \min\{h_1, \ldots, h_{m_2}\}$, доказываем, что

$$J(u) \geqslant F_{m_2} - \varepsilon,\quad u \in Q_{m_2} \cap U,$$

где $Q_{m_2} = Q_{m_1} \cup \Pi_{m_2}$, $\Pi_{m_2} = \Big\{u\colon\ a_1 \leqslant u^1 \leqslant b_1,\quad v_{m_2}^2 - h \leqslant u^2 \leqslant v_{m_2}^2 + d_{m_2}^1,\quad a_j \leqslant u^j \leqslant a_j + h + d_{m_2}^2,\quad j = 3, \ldots, n\Big\}$ и т. д. Продолжая этот процесс дальше, найдем точки

$$u_{m_{s-1}}, \ldots, u_{m_s},\quad u_{m_{s-1}+i} = \Big(v_{m_{s-1}+i}^1, \ldots, v_{m_{s-1}+i}^n\Big),$$

$$i = 1, \ldots, m_s - m_{s-1},$$

где номер m_s и координаты $v_{m_{s-1}+i}^1$ определяются по формулам, аналогичным (7), (9), $v_{m_{s-1}+i}^2 = \min\Big\{b_2;\ v_{m_{s-1}}^2 + h + d_{m_s-1}^1\Big\}$, $v_{m_{s-1}}^2 < b_2 - h \leqslant v_{m_{s-1}}^2 + h + d_{m_{s-1}}^1$; $v_{m_{s-i+1}}^j = a_j + h$ $(j = 3, \ldots, n)$

устанавливаем неравенство

$$J(u) \geqslant F_{m_s} - \varepsilon, \quad u \in Q_{m_s} \cap U,$$

где $Q_{m_s} = \left\{u\colon\ a_1 \leqslant u^1 \leqslant b_1,\ a_2 \leqslant u^2 \leqslant b_2,\ a_j \leqslant u^j \leqslant a_j + h + d^2_{m_s}\right\}$, $d^1_{m_s} = \min\{h_{m_{s-1}+1}, \ldots, h_{m_s}\}$, $d^2_{m_s} = \min\{h_1, \ldots, h_{m_s}\}$. После этого начинаем менять координату u^3: сначала берем $v^3_{m_s+i} = v^3_{m_s} + h + d^2_{m_s}$ и при $v^j_{m_s+i} = a_j + h$ $(j = 4, \ldots, n)$ осуществляем перебор координат u^1, u^2 по описанной выше схеме до тех пор, пока при некотором m_k не получим оценки $J(u) \geqslant F_{m_k} - \varepsilon$, $u \in Q_{m_k} \cap U$, $Q_{m_k} = \left\{u\colon\ a_1 \leqslant u^1 \leqslant b_1,\ a_2 \leqslant u^2 \leqslant b_2,\ v^3_{m_k} - h \leqslant u^3 \leqslant v^3_{m_k} + d^2_{m_k},\ a_j \leqslant u^j \leqslant a_j + h + d^3_{m_k},\ j = 4, \ldots, n\right\}$, где $d^3_{m_k} = \min\{h_1, \ldots, h_{m_k}\}$, а $d^2_{m_k}$ — минимальное среди чисел $\{h_i\}$, полученных на последнем цикле перебора координат u^1, u^2, соответствующем значению $v^3_{m_s+i} = v^3_{m_s} + h + d^2_{m_s}$; затем берем $v^3_{m_k+1} = v^3_{m_k} + h + d^2_{m_k}$, заново перебираем координаты u^1, u^2 и т. д. Такое изменение координаты u^3 закончится на некоторой итерации m_p установлением оценки $J(u) \geqslant F_{m_p} - \varepsilon$, $u \in Q_{m_p} \cap U$, где $Q_{m_p} = \left\{u\colon\ a_j \leqslant u^j \leqslant b_j,\ j = 1, 2, 3;\ a_j \leqslant u^j \leqslant a_j + d^3_{m_p},\ j = 4, \ldots, n\right\}$, $d^3_{m_p} = \min\{h_1, \ldots, h_{m_p}\}$.

Далее, по такой же схеме изменяем координату u^4 по закону $u^4_{i+1} = u^4_i + h + d^3_{m_q}$, затем — координату u^5 и т. д., продолжая этот процесс до тех пор, пока получившийся параллелепипед Q_{m_r} не покроет исходный параллелепипед U и не будет получена оценка $J(u) \geqslant F_{m_r} - \varepsilon$, $u \in U$. Из этой оценки будет следовать, что $0 \leqslant F_{m_r} - J_* \leqslant \varepsilon$ для любой фиксированной функции $J = J(u) \in Q(L)$. Так как на каждой итерации хотя бы одна координата увеличивается на величину, не меньшую, чем $h = \varepsilon/(L\sqrt{n})$, то описанным методом за конечное число вычислений значений функции $J(u) \in Q(L)$ будет определена величина J_* с требуемой точностью $\varepsilon > 0$.

Так же, как в случае функции одной переменной (см. § 1.7), нетрудно привести примеры самых «плохих» функций из класса $Q(L)$, для которых описанный метод последовательного перебора превратится в пассивный перебор (5), и самых «хороших» функций из $Q(L)$, для которых этот метод дает существенный выигрыш по сравнению с пассивным перебором (5).

Численные эксперименты показывают [139], что перебор существенно сокращается, если в (6) или (12) принять $F_0 = J(u_0)$, где $J(u_0)$ близко к J_*. Поэтому сначала полезно провести пред-

варительное исследование функции $J(u)$ на грубой сетке и получить грубую оценку для J_*, которую можно принять за F_0. Для определения F_0 возможно использование какого-либо простого локального метода минимизации; для этих целей можно также использовать и сам описанный метод последовательного перебора с грубым значением $\varepsilon > 0$. Возможно сочетание описанного метода с каким-либо локальным методом, когда время от времени последовательный перебор прерывается и для уточнения значения F_k из последней найденной точки производится спуск с помощью выбранного локального метода.

Недостатком описанного метода является требование знания константы L из условия (3). Если величина L неизвестна, то можно попытаться взять некоторое начальное значение $L = L_0$ и применять метод с $L = L_0$, $2L_0$, $4L_0$ и т. д. до тех пор, пока полученное приближение значения J_* не будет отличаться от предыдущего приближения не более чем на ε. Для уточнения постоянной L можно также использовать результаты проведенных вычислений значений функции на предыдущих итерациях.

4. Для класса $Q(L)$ можно предложить и другой вариант метода покрытий [231], одномерный вариант которого также был рассмотрен в § 1.7, п. 4. Предположим для простоты, что в (2) $b_1 - a_1 = \ldots = b_n - a_n = d$, т. е. U — куб с ребрами, параллельными осям координат. Найдем наименьшее целое число m из условий $d/2^{m+1} \leqslant \varepsilon/(L\sqrt{n})$ $(m \geqslant 0)$. Через точки $c_{ik} = (c_{ik1}, \ldots, c_{ikn})$, где $c_{ikj} = a_j + i(b_j - a_j)/2^k$ $(i = 0, 1, \ldots, 2^k,\ 0 \leqslant k \leqslant m)$, проведем всевозможные прямые, параллельные осям координат, и разобьем куб U на систему кубов, которые будем называть кубами k-го уровня. Тем самым, исходный куб будет иметь нулевой уровень; кубы $k+1$-го уровня получаются из кубов k-го уровня делением их ребер пополам и проведением через середины ребер прямых, параллельных осям координат. Далее, пользуясь той же схемой, которая описана в § 1.7 (слово «отрезок» теперь нужно заменить словом «куб»), можно организовать перебор кубов различных уровней, последовательно вычислять значение функции $J(u)$ в центрах u_i этих кубов и определять величину $F_s = \min\limits_{1 \leqslant i \leqslant s} J(u_i)$, причем неравенство (1.7.5) здесь следует заменить на

$$F_s \leqslant J(u_s) - \sqrt{n}\, L h_s + \varepsilon,$$

где h_s — длина ребра рассматриваемого на s-м шаге куба с центром в точке u_s. В результате для любой функции $J(u) \in Q(L)$, произведя не более 2^{mn} вычислений ее значений, получим решение задачи (1)—(4).

В общем случае, когда множество (2) не является кубом, для построения покрытия множества U можно воспользоваться параллелепипедами или комбинациями кубов и параллелепипе-

дов. В [231] рассмотрены варианты метода покрытий для более сложных множеств, задаваемых неравенствами $g_i(u) \leqslant 0$ $(i = 1, \ldots, m)$.

Метод последовательного перебора для других классов функций описан в [139]; об оптимальных последовательных методах на различных классах функций см. в [21, 106, 228, 282, 298].

§ 13. Метод модифицированных функций Лагранжа

1. Рассмотрим задачу

$$J(u) \to \inf; \quad u \in U = \{u \in E^n: \ u \in U_0,$$
$$g_i(u) \leqslant 0, \quad i = 1, \ldots, m, \quad g_i(u) = 0, \quad i = m+1, \ldots, s\}, \tag{1}$$

где $J(u)$, $g_1(u)$, ..., $g_s(u)$ — заданные функции на множестве U_0. Пусть $J_* > -\infty, U_* \neq \varnothing$. Выше (см. теоремы 4.9.2, 4.9.4, 4.9.5) при определенных условиях выпуклости и регулярности задачи (1) было установлено, что для любой точки $u_* \in U_*$ найдутся множители Лагранжа $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*)$:

$$\lambda^* \in \Lambda_0 = \{\lambda \in E^s: \ \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\}$$

такие, что точка (u_*, λ^*) образует седловую точку функции Лагранжа

$$L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i g_i(u), \quad u \in U_0, \quad \lambda \in \Lambda_0, \tag{2}$$

т. е.

$$L(u_*, \lambda) \leqslant L(u_*, \lambda^*) = J_* \leqslant L(u, \lambda^*), \quad u \in U_0, \quad \lambda \in \Lambda_0. \tag{3}$$

Была также доказана справедливость обратного утверждения: если $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ является седловой точкой функции (2), то $u_* \in U_*$ (теорема 4.9.1).

Основываясь на этих фактах, можно предложить различные методы решения задачи (1), сводящиеся к поиску седловой точки функции Лагранжа. Например, здесь естественным образом напрашивается итерационный процесс, представляющий собой метод проекции градиента по каждой из переменных u и λ (спуск по переменной u и подъем — по λ):

$$u_{k+1} = P_{U_0}(u_k - \alpha_k L_u(u_k, \lambda_k)), \tag{4}$$

$$\lambda_{k+1} = P_{\Lambda_0}(\lambda_k + \alpha_k L_\lambda(u_k, \lambda_k)) = P_{\Lambda_0}(\lambda_k + \alpha_k g(u_k)), \tag{5}$$

$$k = 0, 1, \ldots, \quad \text{где} \quad L_u(u, \lambda) = (L_{u^1}(u, \lambda), \ldots, L_{u^n}(u, \lambda)),$$

$$L_\lambda(u, \lambda) = (L_{\lambda_1}(u, \lambda), \ldots, L_{\lambda_s}(u, \lambda)) = (g_1(u), \ldots, g_s(u)) = g(u);$$

длину шага α_k в (4), (5) можно выбирать из тех же соображений, как это делалось выше в § 2. Заметим, что проекция любой точки $\lambda \in E^s$ на множество Λ_0 вычисляется просто по

формулам $P_{\Lambda_0}(\lambda) = (\mu_1, \ldots, \mu_s)$, где

$$\mu_i = \lambda_i^* = \max\{\lambda_i, 0\}, \quad i = 1, \ldots, m, \quad \mu_i = \lambda_i, \quad i = m+1, \ldots, s.$$

Вместо (4) возможно использование других итерационных процессов, таких, как метод Ньютона и др. В тех случаях, когда задача минимизации функции $L(u, \lambda)$ по переменной $u \in U_0$ при каждом фиксированном $\lambda \in \Lambda_0$ решается достаточно просто, можно предложить следующий итерационный метод:

$$L(u_{k+1}, \lambda) = \inf_{u \in U_0} L(u, \lambda_k), \quad \lambda_{k+1} = P_{\Lambda_0}(\lambda_k + \alpha_k g(u_{k+1})),$$

$$k = 0, 1, \ldots$$

Однако, как оказалось, сходимость перечисленных методов удается доказать лишь при довольно жестких ограничениях на данные задачи (1).

Приведем простейший пример выпуклой задачи, когда метод (4), (5) не сходится к седловой точке функции Лагранжа.

Пример 1. Пусть $J(u) \equiv 0$, $U = \{u \in E^1\colon g(u) \equiv u = 0\}$. Тогда $J_* = 0$, $U_* = \{0\}$. Функция Лагранжа $L(u, \lambda) = J(u) + \lambda g(u) \equiv \lambda(u)$ на $U_0 \times \Lambda_0 = E^1 \times E^1$ имеет седловую точку $(0, 0)$, так как $L(0, \lambda) = 0 \cdot \lambda \equiv 0 = L(0, 0) = L(u, 0) \equiv 0 \cdot u$. Процесс (4), (5) здесь имеет вид

$$u_{k+1} = u_k - \alpha_k \lambda_k, \quad \lambda_{k+1} = \lambda_k + \alpha_k u_k, \quad k = 0, 1, \ldots$$

Поскольку $u_{k+1}^2 + \lambda_{k+1}^2 = (u_k^2 + \lambda_k^2)(1 + \alpha_k^2) \geqslant u_k^2 + \lambda_k^2$ $(k = 0, 1, \ldots)$, то ясно, что при любых $(u_0, \lambda_0) \neq 0$ и любом выборе длины шага $\alpha_k \geqslant 0$ этот процесс расходится.

Анализ перечисленных методов показывает, что причина их расходимости заключается в том, что функция Лагранжа (2) по переменной λ не очень хорошо «устроена» и допускает, в частности, случаи, когда в (4.9.38) имеют место строгие включения (см. пример 4.9.7). Чтобы преодолеть возникающие здесь трудности, можно попытаться видоизменить функцию Лагранжа, строить так называемые модифицированные функции Лагранжа, которые имеют то же множество седловых точек, что и функция (2), и которые обладают лучшими свойствами, чем функция (2). Такие функции, оказывается, существуют и могут быть использованы для поиска седловой точки функции (2) и для решения задачи (1). Следуя [29], мы рассмотрим один из возможных здесь подходов.

2. Будем рассматривать задачу

$$J(u) \to \inf, \quad u \in U = \{u \in E^n\colon u \in U_0,\ g(u) \leqslant 0\}, \tag{6}$$

где $J(u)$, $g(u) = (g_1(u), \ldots, g_m(u))$ — заданные функции из $C^1(U_0)$. Как и в гл. 3, векторное неравенство $g = (g_1, \ldots, g_m) \geqslant 0$ $[g \leqslant 0]$ здесь и ниже означает, что $g_i \geqslant 0$ $[g_i \leqslant 0]$ при всех

$i=1, \ldots, m$, а неравенство $a \geqslant b$ для $a, b \in E^m$ эквивалентно неравенству $a-b \geqslant 0$.

Наряду с классической функцией Лагранжа задачи (6)

$$L(u, \lambda)=J(u)+\langle g(u), \lambda\rangle, \quad u \in U_0, \ \lambda \in \Lambda_0=\{\lambda \in E^m: \ \lambda \geqslant 0\} \tag{7}$$

еще рассмотрим следующую модифицированную функцию Лагранжа:

$$M(u, \lambda)=J(u)+\frac{1}{2A}\left[(\lambda+Ag(u))^+\right]^2-\frac{1}{2A}|\lambda|^2 \tag{8}$$

переменных $u \in U_0$, $\lambda \in \Lambda_0$, где A — произвольная фиксированная положительная константа; в (8) принято обозначение

$$a^+=P_{E_+^m}(a)=(a_1^+, \ldots, a_m^+), \quad a_i^+=\max\{a_i; 0\}, \quad i=1, \ldots, m, \tag{9}$$

— проекция точки $a \in E^m$ на положительный октант $E_+^m = \{a \in E^m: \ a \geqslant 0\}$.

Нетрудно видеть, что функция $\varphi(z)=(\max\{z; 0\})^2=(z^+)^2$ одной переменной непрерывно дифференцируема на всей числовой оси E^1, причем

$$\varphi'(z)=2\max\{z; 0\}=2z^+.$$

Отсюда следует, что при $J(u)$, $g(u) \in C^1(U_0)$ функция (8) непрерывно дифференцируя по u и λ, причем

$$\begin{aligned} \frac{\partial M}{\partial u}&=M_u(u, \lambda)=J'(u)+(g'(u))^T(\lambda+Ag(u))^+, \\ \frac{\partial M}{\partial \lambda}&=M_\lambda(u, \lambda)=\frac{1}{A}\left[(\lambda+Ag(u))^+-\lambda\right], \quad u \in U_0, \ \lambda \in E^m, \end{aligned} \tag{10}$$

где $g'(u)$ — матрица порядка $m \times n$, у которой в i-й строке, j-м столбце $g_{iu^j}(u)=\dfrac{\partial g_i(u)}{\partial u^j}$ $(i=1, \ldots, m, \ j=1, \ldots, n)$, а матрица $(g'(u))^T$ получена транспонированием $g'(u)$. Далее, пользуясь теоремами 4.2.7, 4.2.8 и следствиями из них, нетрудно показать, что если U_0 — выпуклое множество и функции $J(u)$, $g_i(u)$ выпуклы на U_0, то функция $M(u, \lambda)$ выпукла по переменной u на множестве U_0 при любом фиксированном $\lambda \in E^m$. Отметим также, хотя это ниже явно не будет использовано, что $M(u, \lambda)$ является вогнутой по переменной λ на множестве Λ_0 при любом фиксированном $u \in U_0$ — в этом проще всего убедиться, доказав неравенство $\langle M_\lambda(u, \lambda)-M_\lambda(u, \mu), \lambda-\mu\rangle \leqslant 0$ для всех λ, $\mu \in \Lambda_0$ и затем обратившись к теореме 4.2.4.

Перейдем к описанию метода решения задачи (6), использующего функцию $M(u, \lambda)$. В качестве начального приближения возьмем любые точки $u_0 \in U_0$, $\lambda_0 \in \Lambda_0$. Пусть k-е приближение

$u_k \in U_0$, $\lambda_k \in \Lambda_0$ уже известно. Составим функцию

$$\Phi_k(u) = \frac{1}{2}|u - u_k|^2 + \alpha M(u, \lambda_k), \quad u \in U_0, \tag{11}$$

где α — некоторое положительное число, являющееся параметром метода. Предположим, что существует точка v_k, удовлетворяющая условиям

$$v_k \in U_0, \quad \Phi_k(v_k) = \inf_{U_0} \Phi_k(u). \tag{12}$$

В качестве следующего $(k+1)$-го приближения возьмем точку u_{k+1} такую, что

$$\begin{gathered} u_{k+1} \in U_0, \quad \Phi_k(u_{k+1}) \leqslant \inf_{U_0} \Phi_k(u) + \delta_k^2/2, \\ |g(u_{k+1}) - g(v_k)| \leqslant \delta_k, \end{gathered} \tag{13}$$

где $\delta_k \geqslant 0$, $\lim\limits_{k\to\infty} \delta_k = 0$. В частности, если точка v_k из (12) известна, то можно принять $u_{k+1} = v_k$; в общем случае для определения u_{k+1} из условий (13) нужно решать задачу (12) с помощью какого-либо сходящегося метода минимизации. Дальнейшее изложение не зависит от того, каким методом решается задача (12), поэтому здесь мы можем ограничиться предположением, что имеется какой-либо достаточно эффективный метод решения задачи (12), позволяющий за конечное число итераций найти точку u_{k+1}, которая удовлетворяет условиям (13). После определения u_{k+1} точка λ_{k+1} находится по формуле

$$\lambda_{k+1} = (\lambda_k + Ag(u_{k+1}))^+. \tag{14}$$

Правила получения $(k+1)$-го приближения $u_{k+1} \in U_0$, $\lambda_{k+1} \in \Lambda_0$ изложены.

Описанный метод кратко будем называть методом (13), (14). Для исследования сходимости метода (13), (14) нам понадобятся некоторые свойства функции a^+, определенной равенствами (9). Из теоремы 4.4.2 следует, что

$$|a^+ - b^+| \leqslant |a - b| \qquad \forall a, b \in E^m. \tag{15}$$

Далее, система соотношений

$$g \leqslant 0, \quad \lambda \geqslant 0, \quad \lambda_i g_i = 0, \qquad i = 1, \ldots, m, \tag{16}$$

эквивалентна равенству

$$\lambda = (\lambda + Ag)^+ \tag{17}$$

при любых постоянных $A > 0$. В самом деле, если выполняются соотношения (16), то либо $g_i = 0$, $\lambda_i \geqslant 0$, либо $\lambda_i = 0$, $g_i \leqslant 0$. В каждом из этих случаев, очевидно, равенство (17) верно. Таким образом, из (16) следует (17). Докажем обратное. Пусть имеет место равенство (17). Распишем это равенство в координатной форме

$$\lambda_i = (\lambda_i + Ag_i)^+ = \max\{\lambda_i + Ag_i;\ 0\}, \qquad i = 1, \ldots, m. \tag{17'}$$

Отсюда ясно, что $\lambda_i \geqslant 0$ при всех $i = 1, \ldots, m$, т. е. $\lambda_i \geqslant 0$. Если $\lambda_i = 0$, то $\lambda_i g_i = 0$ и, кроме того, из (17′) получим $0 = (0 + Ag_i)^+ = \lambda_i + Ag_i$,

т. е. $g_i \leqslant 0$. Если же $\lambda_i > 0$, то из (17′) следует $0 < \lambda_i = (\lambda_i + Ag_i)^+ = \lambda_i + Ag_i$, что возможно лишь при $g_i = 0$ и $\lambda_i g_i = 0$. Эквивалентность (16) и (17) доказана.

Далее, пользуясь определением (9) функции a^+, нетрудно получить, что

$$\langle a^+, a\rangle = \langle a^+, a^+\rangle, \quad \langle a^+, b\rangle \leqslant \langle a^+, b^+\rangle \qquad \forall a, b \in E^n.$$

Отсюда имеем

$$\langle a^+ - b^+, a - b\rangle = \langle a^+, a\rangle + \langle b^+, b\rangle - \langle a^+, b\rangle - \langle b^+, a\rangle \geqslant$$
$$\geqslant \langle a^+, a^+\rangle + \langle b^+, b^+\rangle - \langle a^+, b^+\rangle - \langle b^+, a^+\rangle = \langle a^+ - b^+, a^+ - b^+\rangle,$$

т. е.

$$\langle a^+ - b^+, a - b\rangle \geqslant \langle a^+ - b^+, a^+ - b^+\rangle. \tag{18}$$

Теорема 1. *Пусть U_0 — выпуклое замкнутое множество из E^n (например, $U_0 = E^n$), функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ выпуклы на U_0 и принадлежат классу $C^1(U_0)$, $J_* > -\infty$, $U_* \neq \varnothing$, функция Лагранжа (7) имеет хотя бы одну седловую точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ в смысле неравенств (3). Пусть, кроме того, последовательность $\{\delta_k\}$ из (13) неотрицательна и $\sum_{k=0}^{\infty} \delta_k < \infty$. Тогда последовательность $\{(u_k, \lambda_k)\}$, удовлетворяющая условиям (13), (14), при любом выборе начальных $(u_0, \lambda_0) \in U_0 \times \Lambda_0$ и любых фиксированных параметрах $\alpha > 0$, $A > 0$ существует и сходится к некоторой седловой точке функции Лагранжа (7).*

Доказательство. При сделанных предположениях функция $M(u, \lambda)$ выпукла по переменной $u \in U_0$ при всех $\lambda \in \Lambda_0$, $A > 0$, поэтому при любых $u_k \in U_0$, $\lambda_k \in \Lambda_0$, $\alpha > 0$, $A > 0$ функция $\Phi_k(u)$, определяется формулой (11), сильно выпукла на U_0 с константой сильной выпуклости $\varkappa = 1/2$. Отсюда и из теоремы 4.3.1 следует, что точка v_k, удовлетворяющая условиям (12), существует и определяется однозначно. Тогда существует и точка u_{k+1}, удовлетворяющая условиям (13): например, в (13) можно взять $u_{k+1} = v_k$. Здесь важно заметить, что многие из описанных выше методов минимизации для задачи (12) сходятся и при любом $\delta_k > 0$ позволяют получить точку u_{k+1} из (13) за конечное число итераций. Таким образом, при выполнении условий теоремы последовательность $\{(u_k, \lambda_k)\}$ существует и имеются достаточно эффективные способы реализации каждой итерации метода (13), (14).

Наряду с точкой λ_{k+1}, определяемой по формуле (14), введем еще точку

$$\mu_k = (\lambda_k + Ag(v_k))^+, \qquad k = 0, 1, \ldots \tag{19}$$

Покажем, что для любой седловой точки (u_*, λ^*) функции Лагранжа (7) справедливо неравенство

$$|u_k - u_*|^2 + \frac{\alpha}{A}|\lambda_k - \lambda^*|^2 \geqslant |v_k - u_*|^2 + \frac{\alpha}{A}|\mu_k - \lambda^*|^2 +$$
$$+ |v_k - u_k|^2 + \frac{\alpha}{A}|\mu_k - \lambda_k|^2, \qquad k = 0, 1, \ldots \tag{20}$$

Согласно лемме 4.9.1 существование седловой точки (u_*, λ^*) в задаче (6) эквивалентно соотношениям

$$L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \qquad u \in U_0, \tag{21}$$

$$g(u_*) \leqslant 0, \quad \lambda^* \geqslant 0, \quad \lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, m. \tag{22}$$

В силу эквивалентности соотношений (16), (17) условия (22) можно переписать в следующей равносильной форме:

$$\lambda^* = (\lambda^* + Ag(u_*))^+. \tag{23}$$

Кроме того, функция $L(u, \lambda^*)$ выпукла на U_0 и принадлежит $C^1(U_0)$, а тогда согласно теореме 4.2.3 неравенство (21) эквивалентно условию

$$\langle L_u(u_*, \lambda^*), u - u_* \rangle = \langle J'(u_*) + (g'(u_*))^T \lambda^*, u - u_* \rangle \geqslant 0$$

при всех $u \in U_0$. Отсюда с учетом равенства (23) имеем

$$\langle J'(u_*) + (g'(u_*))^T (\lambda^* + Ag(u_*))^+, u - u_* \rangle \geqslant 0, \quad u \in U_0. \tag{24}$$

Далее, из условия (12) и теоремы 4.2.3 следует

$$\langle \Phi_k'(v_k), u - v_k \rangle \geqslant 0, \quad u \in U_0.$$

Отсюда с учетом формулы (10) получим

$$\langle v_k - u_k + \alpha J'(v_k) + \alpha (g'(v_k))^T (\lambda_k + Ag(v_k))^+, u - v_k \rangle \geqslant 0, \quad u \in U_0. \tag{25}$$

Примем в (24) $u = v_k$, умножим это неравенство на $\alpha > 0$ и сложим с неравенством (25) при $u = u_*$. Получим

$$\langle v_k - u_k + \alpha (J'(v_k) - J'(u_*)) + \alpha (g'(v_k))^T (\lambda_k + Ag(v_k))^+ - \\ - \alpha (g'(u_*))^T (\lambda^* + Ag(u_*))^+, u_* - v_k \rangle \geqslant 0, \quad k = 0, 1, \ldots$$

Отсюда имеем

$$\langle v_k - u_k, u_* - v_k \rangle \geqslant \alpha \langle J'(v_k) - J'(u_*), v_k - u_* \rangle + \\ + \alpha \langle (\lambda_k + Ag(v_k))^+, g'(v_k)(v_k - u_*) \rangle - \\ - \alpha \langle (\lambda^* + Ag(u_*))^+, g'(u_*)(v_k - u_*) \rangle, \quad k = 0, 1, \ldots \tag{26}$$

Так как функции $J(u)$, $g_i(u)$ выпуклы, то согласно теореме 4.2.4

$$\langle J'(v_k) - J'(u_*), v_k - u_* \rangle \geqslant 0,$$
$$g'(v_k)(v_k - u_*) \geqslant g(v_k) - g(u_*) \geqslant g'(u_*)(v_k - u_*).$$

Отсюда и из (26) следует

$$\langle v_k - u_k, u_* - v_k \rangle \geqslant \alpha \langle (\lambda_k + Ag(v_k))^+ - (\lambda^* + Ag(u_*))^+, g(v_k) - g(u_*) \rangle = \\ = \frac{\alpha}{A} \langle (\lambda_k + Ag(v_k))^+ - (\lambda^* + Ag(u_*))^+, [(\lambda_k + Ag(v_k)) - \lambda_k] - [(\lambda^* + Ag(u_*)) - \lambda^*] \rangle.$$

К правой части этой оценки применим неравенство (18). С учетом формулы (19), определяющей точку μ_k, и равенства (23) получим

$$\langle v_k - u_k, u_* - v_k \rangle \geqslant \\ \geqslant \frac{\alpha}{A} \langle (\lambda_k + Ag(v_k))^+ - (\lambda^* + Ag(u_*))^+, [(\lambda_k + Ag(v_k))^+ - \lambda_k] - \\ - [(\lambda^* + Ag(u_*))^+ - \lambda^*] \rangle = \frac{\alpha}{A} \langle \mu_k - \lambda^*, \mu_k - \lambda_k \rangle,$$

т. е.

$$\langle v_k - u_k, u_* - v_k \rangle + \frac{\alpha}{A} \langle \mu_k - \lambda_k, \lambda^* - \mu_k \rangle \geqslant 0, \quad k = 0, 1, \ldots \tag{27}$$

Справедливы тождества

$$|u_k - u_*|^2 = |(u_k - v_k) + (v_k - u_*)|^2 = |v_k - u_k|^2 + |u_* - v_k|^2 + 2\langle v_k - u_k, u_* - v_k \rangle,$$
$$|\lambda_k - \lambda^*|^2 = |\mu_k - \lambda_k|^2 + |\lambda^* - \mu_k|^2 + 2\langle \mu_k - \lambda_k, \lambda^* - \mu_k \rangle.$$

Умножим второе из этих тождеств на α/A и сложим с первым. Отсюда с учетом оценки (27) получим обещанное неравенство (20).

Далее, покажем, что

$$|u_{k+1}-v_k|\leqslant\delta_k,\qquad |\lambda_{k+1}-\mu_k|\leqslant 2\delta_k,\qquad k=0,\ 1,\ \ldots \tag{28}$$

Поскольку функция $\Phi_k(u)$ сильно выпукла на U_0, то с помощью теоремы 4.3.1 и первого неравенства (13) получим

$$|u_{k+1}-v_k|^2/2\leqslant\Phi_k(u_{k+1})-\Phi_k^*(v_k)\leqslant\delta_k^2/2,$$

что равносильно первой оценке (28). Из формул (14), (19), определяющих точки λ_{k+1}, μ_{k+1}, неравенства (15) и условий (13) следует

$$|\lambda_{k+1}-\mu_k|\leqslant A|g(u_{k+1})-g(v_k)|\leqslant A\delta_k.$$

Оценки (28) доказаны.

В $(n+m)$-мерном линейном пространстве $\mathbf{R}^{n+m}$ переменных $z=(u,\ \lambda)=(u^1,\ \ldots,\ u^n,\ \lambda_1,\ \ldots,\ \lambda_m)$ введем скалярное произведение $\langle\langle z_1,\ z_2\rangle\rangle=\langle u_1,\ u_2\rangle+(\alpha/A)\langle\lambda^1,\ \lambda^2\rangle$ и соответствующую ему норму

$$\|z\|=(|u|^2+(\alpha/A)|\lambda|^2)^{1/2}. \tag{29}$$

Тогда, обозначив $z_k=(u_k,\ \lambda_k)$, $w_k=(v_k,\ \mu_k)$, $z^*=(u_*,\ \lambda^*)$, неравенства (20) и (28) можем записать в виде

$$\|z_k-z^*\|^2\geqslant\|w_k-z^*\|^2+\|w_k-z_k\|^2,\qquad k=0,\ 1,\ \ldots, \tag{30}$$

$$\|z_{k+1}-w_k\|\leqslant(A\alpha+1)\delta_k,\qquad k=0,\ 1,\ \ldots \tag{31}$$

Напомним, что по условию $\sum\limits_{k=0}^{\infty}\delta_k<\infty$. Таким образом, последовательности $\{z_k\}$, $\{w_k\}$, $\{\delta_k\}$ удовлетворяют условиям леммы 2.3.10. Для полной строгости, конечно, нужно заметить, что в неравенствах (2.3.30), (2.3.31) использована евклидова норма линейного пространства $\mathbf{R}^{n+m}$, а в только что полученных неравенствах (30), (31) — норма (29). Тем не менее, рассуждая так же, как при доказательстве леммы 2.3.10, нетрудно показать, что существует конечный предел $\lim\limits_{k\to\infty}\|z_k-z^*\|$ и, кроме того,

$$\lim_{k\to\infty}\|w_k-z_k\|=0. \tag{32}$$

Заметим, что

$$\min\{1;\ \alpha/A\}|z|^2\leqslant\|z\|^2\leqslant\max\{1;\ \alpha/A\}|z|^2,$$

т. е. нормы $|z|$ и $\|z\|$ эквивалентны. Отсюда и из существования конечного предела $\lim\limits_{k\to\infty}\|z_k-z^*\|$ следует, что последовательность $\{z_k=(u_k,\ \lambda_k)\}\in U_0\times\lambda_0$ ограничена в E^{n+m} и из нее можно выбрать подпоследовательность $\{z_{k_r}=(u_{k_r},\ \lambda_{k_r})\}$, которая сходится в E^{n+m} к некоторой точке $c^*=(a_*,\ b^*)$, причем $a_*\in U_0$, $b^*\in\Lambda_0$ в силу замкнутости U_0 и Λ_0. Покажем, что $c^*=(a_*,\ b^*)$ — седловая точка функции Лагранжа (7). Из $\{z_{k_r}\}\to c^*$ и (31), (32) следует, что $\{w_{k_r}\}\to c^*$, $\{z_{k_r+1}\}\to c^*$. Тогда из (14) при $k=k_r\to\infty$ получим

$$b^*=(b^*+Ag(a_*))^+. \tag{33}$$

В силу эквивалентности соотношений (16) и (17) из (33) следует

$$g(a_*)\leqslant 0,\qquad b^*\geqslant 0,\qquad b_i^*g_i(a_*)=0,\qquad i=1,\ \ldots,\ m. \tag{34}$$

Далее, переходя в (25) к пределу при $k = k_r \to \infty$, будем иметь

$$\left\langle J'(a_*) + (g'(a_*))^T (b^* + Ag(a_*))^+,\ u - a_* \right\rangle \geqslant 0, \quad u \in U_0,$$

или с учетом (33)

$$\left\langle J'(a_*) + (g'(a_*))^T b^*,\ u - a_* \right\rangle = \langle L_u(a_*, b^*),\ u - a_* \rangle \geqslant 0$$

при всех $u \in U_0$. В силу выпуклости $L(u, b^*)$ последнее неравенство эквивалентно неравенству

$$L(a_*, b^*) \leqslant L(u, b^*), \quad u \in U_0. \tag{35}$$

Из соотношений (34), (35) и леммы 4.9.1 следует, что $c^* = (a_*, b^*)$ — седловая точка функции $L(u, \lambda)$ в смысле неравенств (3), а тогда согласно теореме 4.9.1 a_* — решение задачи (6).

Заметим, что неравенство (30) верно для любой седловой точки, в частности, оно верно и для найденной точки $c^* = (a_*, b^*)$. Поэтому существует конечный предел $\lim\limits_{k\to\infty} \| z_k - c^* \|$, причем в силу определения точки c^* имеем $\lim\limits_{k\to\infty} \| z_k - c^* \| = \lim\limits_{r\to\infty} \| z_{k_r} - c^* \| = 0$. Это значит, что вся последовательность $\{z_k = (u_k, \lambda_k)\}$ сходится в точке $c^* = (a_*, b^*)$, и, в частности, $\{u_k\}$ сходится к a_* — решению задачи (6). Теорема 1 доказана.

Другие методы поиска седловой точки функции Лагранжа, другие методы решения задачи (1) или (6), основанные на связи между двойственными задачами (см. теорему 4.9.6), а также библиографию по таким методам читатель найдет в [8, 29, 111, 116, 152, 330].

§ 14. Метод штрафных функций

1. Метод штрафных функций является одним из наиболее простых и широко применяемых методов решения задач минимизаци. Основная идея метода заключается в сведении исходной задачи

$$J(u) \to \inf; \quad u \in U \tag{1}$$

к последовательности задач минимизации

$$\Phi_k(u) \to \inf; \quad u \in U_0, \quad k = 0, 1, \ldots, \tag{2}$$

где $\Phi_k(u)$ — некоторая вспомогательная функция, а множество U_0 содержит U. При этом функция $\Phi_k(u)$ подбирается так, чтобы она с ростом номера k мало отличалась от исходной функции $J(u)$ на множестве U и быстро возрастала на множестве $U_0 \backslash U$. Можно ожидать, что быстрый рост функции $\Phi_k(u)$ вне U приведет к тому, что при больших k нижняя грань этой функции на U_0 будет достигаться в точках, близких ко множеству U, и решение задачи (2) будет приближаться к решению задачи (1). Кроме того, как увидим ниже, имеется достаточно широкий произвол в выборе функций $\Phi_k(u)$ и множества U_0 для задач (2), и можно надеяться на то, что задачи (2) удастся составить более простыми по сравнению с задачей (1) и допускающими применение несложных методов минимизации.

Определение 1. Последовательность функций $\{P_k(u), k=0, 1, \ldots\}$, определенных и неотрицательных на множестве U_0, содержащем множество U, называют *штрафом* или *штрафной функцией* множества U на множестве U_0, если

$$\lim_{k\to\infty} P_k(u) = \begin{cases} 0, & u \in U, \\ \infty, & u \in U_0 \setminus U. \end{cases}$$

Из этого определения видно, что при больших номерах k за нарушение условия $u \in U$ приходится «платить» большой штраф, в то время как при $u \in U$ штрафная функция представляет собой бесконечно малую величину при $k \to \infty$.

Для любого множества $U \subset E^n$ можно указать сколько угодно различных штрафных функций. Например, если $\{A_k\}$ — какая-либо положительная последовательность, $\lim\limits_{k\to\infty} A_k = \infty$, то можно взять

$$P_k(u) = A_k \rho(u, U), \quad u \in E^n = U_0, \quad k = 0, 1, \ldots$$

(здесь U предполагается замкнутым) или

$$P_k(u) = \begin{cases} 0, & u \in U, \\ A_k |u - \bar{u}|, & u \notin U, \end{cases} \quad k = 0, 1, \ldots;$$

здесь $\rho(u, U) = \inf\limits_{v \in U} |u - v|$ — расстояние от точки до множества U, а $\bar{u}$ — какая-либо точка из U. Другие примеры штрафных функций будут приведены ниже.

Допустим, что некоторое множество U_0, содержащее U, а также штрафная функция $\{P_k(u)\}$ множества U на U_0 уже выбраны. Предполагая, что функция $J(u)$ определена на U_0, введем функции

$$\Phi_k(u) = J(u) + P_k(u), \quad u \in U_0, \quad k = 0, 1, \ldots \tag{3}$$

и рассмотрим последовательность задач (2) с функциями (3). Будем считать, что

$$\Phi_{k*} = \inf_{U_0} \Phi_k(u) > -\infty, \quad k = 0, 1, \ldots \tag{4}$$

Если здесь при каждом $k = 0, 1, \ldots$ нижняя грань достигается, то условия

$$\Phi_k(u_k) = \Phi_{k*}, \quad u_k \in U_0, \tag{5}$$

определяют последовательность $\{u_k\}$. Однако точно определить u_k из (5) удается лишь в редких случаях. Кроме того, нижняя грань в (4) при некоторых или даже всех $k = 0, 1, \ldots$ может и не достигаться. Поэтому будем считать, что при каждом $k = 0, 1, \ldots$ с помощью какого-либо метода минимизации найдена

точка u_k, определяемая условиями

$$u_k \in U_0, \quad \Phi_k(u_k) \leqslant \Phi_{k*} + \varepsilon_k, \tag{6}$$

где $\{\varepsilon_k\}$ — некоторая заданная последовательность, $\varepsilon_k > 0$, $k = 0, 1, \ldots, \lim\limits_{k\to\infty} \varepsilon_k = 0$ (если u_k удовлетворяет условиям (5), то в (6) допускается возможность $\varepsilon_k = 0$). Отметим, что, вообще говоря, $u_k \notin U$. Метод штрафных функций описан.

Подчеркнем, что дальнейшее изложение не зависит от того, каким конкретным методом будет найдена точка u_k из (6). Поэтому мы здесь можем ограничиться предположением, что имеется достаточно эффективный метод определения такой точки.

2. Перейдем к исследованию сходимости метода штрафных функций. Так как

$$\lim_{k\to\infty} P_k(u) = \infty \quad \text{при} \quad u \in U_0 \setminus U,$$

то можно ожидать, что для широкого класса задача (1) последовательность $\{u_k\}$, определяемая условиями (6), будет приближаться ко множеству U и будут справедливы равенства

$$\lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} \rho(u_k, U_*) = 0.$$

Мы здесь ограничимся рассмотрением задачи (1) для случая, когда множество U имеет вид

$$U = \{u \in E^n\colon\ u \in U_0,\ g_i(u) \leqslant 0, \quad i = 1, \ldots, m; \quad g_i(u) = 0, \quad i = m+1, \ldots, s\}, \tag{7}$$

где U_0 — заданное множество из E^n (например, $U_0 = E^n$), функции $J(u)$, $g_i(u)$ $(i = 1, \ldots, s)$, определены на U_0. В качестве штрафной функции множества (7) возьмем

$$\begin{gathered} P_k(u) = A_k P(u), \\ P(u) = \sum_{i=1}^{m} (\max\{g_i(u); 0\})^p + \sum_{i=m+1}^{s} |g_i(u)|^p, \quad u \in U_0, \end{gathered} \tag{8}$$

где $A_k > 0$ $(k = 0, 1, \ldots)$, $\lim\limits_{k\to\infty} A_k = \infty$, а $p \geqslant 1$ — фиксированное число.

Очевидно, если функции $g_i(u)$ r раз непрерывно дифференцируемы на множестве U_0, то при любом $p > r$ функция (8) также будет r раз непрерывно дифференцируема на U_0. Если в (8) $p = 1$, то из непрерывности $g_i(u)$ $(i = 1, \ldots, s)$ следует непрерывность $P_k(u)$ на U_0, но гладкости $P_k(u)$ в этом случае ожидать не приходится. Полезно также заметить, что если U_0 — выпуклое множество функции $g_i(u)$ при $i = 1, \ldots, m$ выпуклы на U_0, $g_i(u) = \langle a_i, u\rangle - b^i$ — линейные функции при $i = m+1, \ldots, s$, то функция (8) выпукла на U_0 — это вытекает из следствия к теореме 4.2.8.

Если для краткости ввести обозначения

$$g_i^+(u) = \begin{cases} \max\{g_i(u); 0\} & i = 1, \ldots, m, \\ |g_i(u)|, & i = m+1, \ldots, s, \end{cases} \tag{9}$$

то функцию (8) можно записать в виде

$$P_k(u) = A_k P(u), \quad P(u) = \sum_{i=1}^{s} (g_i^+(u))^p, \quad u \in U_0.$$

Функцию $P(u)$ мы иногда также будем называть *штрафной функцией* множества (7), подразумевая при этом, что после умножения на $A_k > 0$, $\lim\limits_{k\to\infty} A_k = \infty$, она превратится в штрафную функцию в смысле определения 1. Величины A_k из (8) будем называть *штрафными коэффициентами*.

Заметим, что существуют и другие штрафные функции множества (7). Например, вместо (8) можно взять

$$P_k(u) = \sum_{i=1}^{s} A_{ki} (g_i^+(u))^{p_i}, \quad u \in U_0, \quad k = 0, 1, \ldots,$$

где $p_i \geqslant 1$, $A_{ki} > 0$, $\lim\limits_{k\to\infty} A_{ki} = \infty$ $(i = 1, \ldots, s)$; здесь каждое ограничение из (7) имеет свой штрафной коэффициент. Весьма широкий класс штрафных функций множества (7) дает следующая конструкция:

$$P_k(u) = \sum_{i=1}^{s} A_{ki} \varphi_i (g_i^+(u)), \quad u \in U_0, \quad k = 0, 1, \ldots,$$

где $\varphi_i(g)$ — произвольная функция, определенная при $g \geqslant 0$ такая, что $\varphi_i(0) = 0$, $\varphi_i(g) > 0$ при $g > 0$, $i = 1, \ldots, s$. При необходимости можно выбрать функции $\varphi_i(g)$ так, чтобы штрафная функция $P_k(u)$ обладала различными полезными свойствами, такими, как, например, непрерывность, гладкость, выпуклость, простота вычисления значений функции и нужных производных и т. п. Возможны и другие конструкции штрафных функций множества (7). Приведем еще два конкретных примера штрафной функции

$$P_k(u) = \left(1 + \sum_{i=1}^{s} (g_i^+(u))^{p_i}\right)^{A_k} - 1, \quad p_i \geqslant 1,$$

$$P_k(u) = A_k^{-1} \left(\sum_{i=1}^{m} \exp\{A_k g_i(u)\} + \sum_{i=m+1}^{s} \exp\{A_k g_i^2(u)\}\right), \quad u \in U_0,$$

где $A_k > 0$ $(k = 0, 1, \ldots)$, $\lim\limits_{k\to\infty} A_k = \infty$.

Прежде чем переходить к строгим формулировкам теорем сходимости метода штрафных функций, рассмотрим несколько конкретных примеров.

Пример 1. Пусть требуется решить задачу

$$J(u) = x^2 + xy + y^2 \to \inf,$$
$$u \in U = \{u = (x, y) \in E^2:\ x + y - 2 = 0\}.$$

В качестве штрафной функции возьмем $P_k(u) = k(x+y-2)^2$ и положим

$$\Phi_k(u) = x^2 + xy + y^2 + k(x+y-2)^2, \quad u \in U_0 = E^2; \quad k = 1, 2, \ldots$$

Функция $\Phi_k(u)$ при каждом фиксированном $k = 1, 2, \ldots$ сильно выпукла на E^2 и достигает своей нижней грани на E^2 в точке $u_k = (x_k, y_k)$, которая определяется уравнениями

$$\frac{\partial \Phi_k(u_k)}{\partial x} = 2x_k + y_k + 2k(x_k + y_k - 2) = 0,$$

$$\frac{\partial \Phi(u_k)}{\partial y} = x_k + 2y_k + 2k(x_k + y_k - 2) = 0.$$

Отсюда получаем

$$u_k = \left(\frac{4k}{3+4k}, \frac{4k}{3+4k}\right), \quad \Phi_k(u_k) = \frac{12k}{4k+3} = \inf_{E^2} \Phi_k(u).$$

При $k \to \infty$ будем иметь $u_k \to u_* = (1,1)$, $\Phi_k(u_k) \to 3$. Нетрудно видеть, что u_* — решение исходной задачи. В самом деле, $J'(u_*) = (3; 3)$, $\langle J'(u_*), u - u_* \rangle = 3(x-1) + 3(y-1) = 0$ для всех $u \in U$. В силу выпуклости множества U и функции $J(u)$ согласно теореме 4.2.3 тогда u_* — точка минимума $J(u)$ на U, причем $J(u_*) = J_* = 3 = \lim\limits_{k\to\infty} \Phi_k(u_k)$. Таким образом, в рассмотренном примере метод штрафных функций сходится.

Пример 2. Пусть

$$J(u) = e^{-u} \to \inf; \quad u \in U = \{u \in E^1:\ g(u) = ue^{-u} = 0\}.$$

Здесь $U = \{0\} = U_*$, $J_* = 1$. Возьмем штрафную функцию $P_k(u) = kg^2(u) = ku^2e^{-2u}$ и положим $\Phi_k(u) = e^{-u} + ku^2e^{-2u}$, $u \in U_0 = E^1$. Так как $\Phi_k(u) > 0$ при всех $u \in E^1$, $\lim\limits_{u\to\infty} \Phi_k(u) = 0$, то $\Phi_{k*} = \inf\limits_{E^1} \Phi_k(u) = 0$. В качестве точки u_k, удовлетворяющей условиям (6) при $\varepsilon_k = e^{-k} + k^3e^{-2k}$, здесь можно взять $u_k = k$ ($k = 1, 2, \ldots$). Получим $\lim\limits_{k\to\infty} J(u_k) = 0 < J_* = 1$, $\lim\limits_{k\to\infty} \rho(u_k, U_*) = \infty$. Таким образом, выясняется, что метод штрафных функций не всегда сходится.

Перейдем к изложению достаточных условий сходимости метода штрафных функций для задачи (1), (7). Для определенности все формулировки и доказательства теорем проведем для

штрафной функции (8), хотя некоторые из нижеследующих утверждений будут справедливы и для более широкого класса штрафных функций.

Теорема 1. *Пусть функции* $J(u)$, $g_i(u)$ $(i=1, \ldots, s)$, *определены на множестве* U_0, *а последовательность* $\{u_k\}$ *определена условиями* (3), (6), (8). *Тогда*

$$\overline{\lim_{k\to\infty}} J(u_k) \leqslant \overline{\lim_{k\to\infty}} \Phi_k(u_k) = \overline{\lim_{k\to\infty}} \Phi_{k*} \leqslant J_*. \tag{10}$$

Если, кроме того, $J_{**} = \inf\limits_{U_0} J(u) > -\infty$, *то*

$$P(u_k) = \sum_{i=1}^{s} (g_i^+(u_k))^p = O(A_k^{-1}), \quad k = 0, 1, \ldots, \tag{11}$$

$$\overline{\lim_{k\to\infty}} g_i(u_k) \leqslant 0, \quad i = 1, \ldots, m; \quad \lim_{k\to\infty} g_i(u_k) = 0, \quad i = m+1, \ldots, s. \tag{12}$$

Доказательство. Так как $P(u) \geqslant 0$, то из (3), (6), (8) имеем

$$J(u_k) \leqslant J(u_k) + A_k P(u_k) = \Phi_k(u_k) \leqslant \Phi_{k*} + \varepsilon_k \leqslant$$
$$\leqslant \Phi_k(u) + \varepsilon_k = J(u) + A_k P(u) + \varepsilon_k \quad \forall u \in U_0, \quad k = 0, 1, \ldots$$

Отсюда, переходя к нижней грани по $u \in U$ и учитывая, что $P(u) = 0$, $u \in U$, получим

$$J(u_k) \leqslant \Phi_k(u_k) \leqslant \Phi_{k*} + \varepsilon_k \leqslant J_* + \varepsilon_k, \quad k = 0, 1, \ldots \tag{13}$$

При $k \to \infty$ из (13) вытекает (10).

Пусть теперь $J_{**} > -\infty$. Так как $J_* \geqslant J_{**}$, то $J_* > -\infty$. С учетом (13) имеем

$$0 \leqslant A_k P(u_k) = \Phi_k(u_k) - J(u_k) \leqslant J_* + \varepsilon_k - J_{**}, \quad k = 0, 1, \ldots,$$

или

$$0 \leqslant P(u_k) \leqslant \left(J_* + \sup_{k\geqslant 0} \varepsilon_k - J_{**}\right) A_k^{-1}, \quad k = 0, 1, \ldots$$

Оценка (11) доказана. Из нее следует, что $\lim\limits_{k\to\infty} P(u_k) = 0$ или $\lim\limits_{k\to\infty} g_i^+(u_k) = 0$ $(i = 1, \ldots, s)$. Вспоминая определение (9) для $g_i^+(u)$, отсюда получим соотношения (12).

Пример 2 показывает, что в общем случае неравенства в (10) могут быть строгими. Приведем достаточные условия, когда $\lim\limits_{k\to\infty} J(u_k) = J_*$.

Теорема 2. *Пусть* U_0 — *замкнутое множество из* E^n, *функции* $J(u)$, $g_1(u)$, $\ldots$, $g_m(u)$, $|g_{m+1}(u)|$, $\ldots$, $|g_s(u)|$ *полунепрерывны снизу на* U_0, $J_{**} = \inf\limits_{U_0} J(u) > -\infty$. *Пусть последовательность*

$\{u_k\}$, определяемая условиями (3), (6), (8), *имеет хотя бы одну предельную точку. Тогда все предельные точки $\{u_k\}$ принадлежат множеству U_* точек минимума задачи* (1), (7). *Если, кроме того, множество*

$$U_\delta = \{u\colon\ u \in U_0,\ g_i^+(u) \leqslant \delta,\ i = 1, \ldots, s\} \tag{14}$$

ограничено хотя бы при одном значении $\delta > 0$, то

$$\lim_{k\to\infty} \Phi_k(u_k) = \lim_{k\to\infty} \Phi_{k_*} = \lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} \rho(u_k, U_*) = 0. \tag{15}$$

Доказательство. При сделанных предположениях для последовательности $\{u_k\}$ соотношения (10)—(12) сохраняют силу. Пусть v_* — какая-либо предельная точка последовательности $\{u_k\}$, пусть $\{u_{k_r}\} \to v_*$. Заметим, что $v_* \in U_0$ в силу замкнутости U_0. Тогда с учетом полунепрерывности снизу указанных в условии теоремы функций из соотношений (12) получим

$$g_i(v_*) \leqslant \varliminf_{r\to\infty} g_i(u_{k_r}) \leqslant \varlimsup_{k\to\infty} g_i(u_k) \leqslant 0, \qquad i = 1, \ldots, m,$$

$$|g_i(v_*)| \leqslant \varliminf_{r\to\infty} |g_i(u_{k_r})| = \lim_{k\to\infty} |g_i(u_k)| = 0, \qquad i = m+1, \ldots, s.$$

Следовательно, $v_* \in U$. Тогда с учетом (10) имеем $J_* \leqslant J(v_*) \leqslant \varliminf_{r\to\infty} J(u_{k_r}) \leqslant \varlimsup_{k\to\infty} J(u_k) \leqslant J_*$, т. е. $\lim_{r\to\infty} J(u_{k_r}) = J(v_*) = J_*$ или $u_* \in U_*$.

Наконец, пусть множество (14) ограничено при некоторых $\delta > 0$. Из соотношений (12) следует, что $\{u_k\} \in U_\delta$ для всех $k \geqslant k_0$. Это означает, что $\{u_k\}$ имеет хотя бы одну предельную точку. Тогда, как было выше показано, все предельные точки $\{u_k\}$ принадлежат U_*. Следовательно, $\lim_{k\to\infty} \rho(u_k, U_*) = 0$. Из тех же рассуждений и неравенств (10) вытекают остальные равенства (15). Теорема 2 доказана.

Для иллюстрации теоремы 2 рассмотрим

Пример 3. Пусть

$$J(u) = e^{-u} \to \inf; \quad u \in U = \{u \in E^1\colon\ g(u) = u = 0\}.$$

Здесь $J_* = 1$, $U_* = \{0\}$. Функции $J(u)$, $g(u)$ непрерывны на замкнутом множестве $U_0 = E^1$, $J_{**} = \inf_{E^1} e^{-u} = 0$, множество $U_\delta = \{u \in E^1\colon\ |u| \leqslant \delta\}$ ограничено при любом $\delta > 0$. Таким образом, все условия теоремы 2 выполнены. Возьмем штрафную функцию $P(u) = (g(u))^2 = u^2$ и положим

$$\Phi_k(u) = e^{-u} + ku^2, \quad u \in E^1, \qquad k = 1, 2, \ldots$$

Нетрудно видеть, что $\Phi_k(u)$ сильно выпукла на E^1, поэтому $\Phi_{k_*} = \inf_{E^1} \Phi_k(u) > -\infty$. Пусть $\{\varepsilon_k\}$ — произвольная неотрицатель-

ная последовательность, стремящаяся к нулю. Определим точку u_k из условия $\Phi_k(u_k) \leqslant \Phi_{k*} + \varepsilon_k (k = 1, 2, \ldots)$. Для получаемой таким образом последовательности $\{u_k\}$ согласно теореме 2 имеют место равенства (15).

3. Нетрудно видеть, что рассмотренные в примерах 2 и 3 задачи по существу одинаковые: минимизируется одна и та же функция e^{-u} на одном и том же множестве $U = \{0\}$, и отличие лишь в том, что в примере 2 множество U задается ограничениями $g(u) = ue^{-u} = 0$, а в примере 3 — $g(u) = u = 0$. Тем не менее, в примере 2 метод штрафных функций расходится, в примере 3 — сходится.

Отсюда заключаем, что для сходимости метода штрафной функции важное значение имеет способ задания множества U: ограничения, задающие множество U, и штрафные функции этого множества должны быть как-то согласованы с минимизируемой функцией $J(u)$.

Определение 2. Скажем, что задача (1), (7) имеет *согласованную постановку* на множестве U_0, если для любой последовательности $\{u_k\} \in U_0$, для которой

$$\lim_{k\to\infty} g_i^+(u_k) = 0, \qquad i = 1, \ldots, s, \tag{16}$$

имеет место соотношение

$$\varliminf_{k\to\infty} J(u_k) \geqslant J_* = \inf_U J(u). \tag{17}$$

Отметим, что в примере 3 задача имеет согласованную постановку на E^1, а в примере 2 такой согласованности нет.

Теорема 3. *Пусть* $\Phi_k(u) = J(u) + A_k P(u)$, *где* $P(u)$ *определена формулой* (8), *пусть* $\Phi_{k*} = \inf\limits_{U_0} \Phi_k(u)$ $(k = 0, 1, \ldots)$. *Тогда для того, чтобы*

$$\lim_{k\to\infty} \Phi_{k*} = J_*, \tag{18}$$

необходимо, чтобы задача (1), (7) *имела согласованную постановку на множестве* U_0. *Если* $J_{**} = \inf\limits_{U_0} J(u) > -\infty$, *то согласованной постановки задачи* (1), (7) *на* U_0 *достаточно для справедливости равенства* (18).

Доказательство. Необходимость. Пусть имеет место равенство (18). Возьмем произвольную последовательность $\{u_r\} \in U_0$, удовлетворяющую условиям (16). Тогда $\lim\limits_{r\to\infty} P(u_r) = 0$. Справедливы неравенства $\Phi_{k*} \leqslant \Phi_k(u_r) \leqslant J(u_r) + A_k P(u_r)$, $r = 1, 2, \ldots$ Отсюда при $r \to \infty$ получим $\Phi_{k*} \leqslant \varliminf\limits_{r\to\infty} J(u_r)$ при всех $k = 0, 1, \ldots$ Переходя здесь к пределу при $k \to \infty$, с учетом (18) будем иметь $\varliminf\limits_{r\to\infty} J(u_r) \geqslant \lim\limits_{k\to\infty} \Phi_{k*} = J_*$, что и требовалось.

Достаточность. Пусть $J_{**} > -\infty$, задача (1), (7) имеет согласованную постановку на множестве U_0. Поскольку $\Phi_k(u) \geqslant J(u)$ при всех $u \in U_0$, то $\Phi_{k*} \geqslant J_{**} > -\infty$, и имеет смысл говорить о последовательностях, удовлетворяющих условиям (6). Возьмем одну из таких последовательностей $\{u_k\}$. Согласно теореме 1 тогда справедливы соотношения (10)—(12). Заметим, что (12) равносильно (16), откуда следует (17). Из (13), (17) получим $\lim\limits_{k\to\infty} J(u_k) = \lim\limits_{k\to\infty} \Phi_{k*} = J_*$. Теорема 3 доказана.

Класс задач (1), (7), имеющих согласованную постановку на U_0, указан в теореме 2. Другой такой класс задач выделяется в следующей лемме.

Лемма 1. *Пусть функция Лагранжа* $L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i g_i(u)$, $u \in U_0$, $\lambda \in \Lambda_0 = \{\lambda = (\lambda_1, \ldots, \lambda_s) \in E^s\colon \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\}$ *имеет седловую точку на* $U_0 \times \Lambda_0$. *Тогда задача* (1), (7) *имеет согласованную постановку на* U_0.

Доказательство. Пусть $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка функции $L(u, \lambda)$, т. е.

$$L(u_*, \lambda) \leqslant L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \qquad \forall u \in U_0, \quad \lambda \in \Lambda_0. \tag{19}$$

Согласно теореме 4.9.1 тогда $u_* \in U_*$, $J(u_*) = J_* = L(u_*, \lambda^*)$. Из определения (9) функции $g_i^+(u)$ с учетом условия $\lambda^* \in \Lambda_0$ имеем

$$\lambda_i^* g_i(u) \leqslant |\lambda_i^*| g_i^+(u) \qquad \forall u \in U_0, \qquad i = 1, \ldots, s.$$

Отсюда и из (19) получим

$$J_* \leqslant J(u) + \sum_{i=1}^{s} |\lambda_i^*| g_i^+(u) \qquad \forall u \in U_0. \tag{20}$$

Возьмем любую последовательность $\{u_k\} \in U_0$, удовлетворяющую условиям (16). Тогда из (20) при $u = u_k$ получим $\varliminf_{k\to\infty} J(u_k) \geqslant J_*$, т. е. задача (1), (7) имеет согласованную постановку на U_0.

Из теорем 1, 3, леммы 1 следует

Теорема 4. *Пусть функция Лагранжа задачи* (1), (7) *имеет седловую точку и* $J_{**} = \inf_{U_0} J(u) > -\infty$; *пусть последовательность* $\{u_k\}$ *определена условиями* (3), (6), (8). *Тогда* $\lim_{k\to\infty} J(u_*) = \lim_{k\to\infty} \Phi_k(u_k) = \lim_{k\to\infty} \Phi_{k*} = J_*$ *и справедливы соотношения* (11), (12).

4. Покажем, что теорема 4 сохраняет силу и без требования $J_{**} > -\infty$. Более того, для задач, у которых функция Лагранжа имеет седловую точку и даже для несколько более общего класса задач (1), (7), можно получить оценку скорости сходимости метода штрафных функций.

Определение 3. Скажем, что задача (1), (7) имеет *сильно согласованную постановку*, если найдутся такие числа $c_1 \geqslant 0, \ldots, c_s \geqslant 0$, $\nu > 0$, *что*

$$J_* \leqslant J(u) + \sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^{\nu} \qquad \forall u \in U_0. \tag{21}$$

Как видно из неравенства (20), задачи (1), (7), функция Лагранжа которых обладает седловой точкой, имеют сильно согласованную постановку, причем в (21) можно взять $c_i = |\lambda_i^*|$, $\nu = 1$. Другой важный класс задач с сильно согласованной постановкой будет приведен ниже в лемме 5.

Теорема 5. *Пусть задача* (1), (7) *имеет сильно согласованную постановку в смысле определения* 3, $J_* > -\infty$, *последовательность* $\{u_k\}$ *определена условиями* (3), (6), (8), *где* $p > \nu$. *Тогда*

$$0 \leqslant \left(g_i^+(u_k)\right)^p \leqslant P(u_k) \leqslant \rho_k, \quad k = 0, 1, \ldots, \tag{22}$$

$$-|c|(\rho_k)^{\nu/p} \leqslant J(u_k) - J_* \leqslant \varepsilon_k, \quad k = 0, 1, \ldots, \tag{23}$$

$$-BA_k^{-\nu/(p-\nu)} \leqslant \Phi_k(u_k) - J_* \leqslant \varepsilon_k, \quad -BA_k^{-\nu/(p-\nu)} \leqslant \Phi_{k*} - J_* \leqslant \varepsilon_k,$$
$$k = 0, 1, \ldots, \tag{24}$$

где

$$\rho_k = \left(\frac{|c|}{A_k}\right)^{p/(p-\nu)} + \frac{p}{p-\nu}\frac{\varepsilon_k}{A_k}, \quad k = 0, 1, \ldots,$$

$$|c| = \left(\sum_{i=1}^{s} |c_i|^{p/(p-\nu)}\right)^{(p-\nu)/p}, \quad B = (p-\nu)\,\nu^{\nu/(p-\nu)} p^{-p/(p-\nu)} |c|^{p/(p-\nu)}.$$

Если, кроме того, U_0 замкнутое множество, функции $J(u)$, $g_i^+(u)$ полунепрерывны снизу на U_0, $\{A_k\} \to \infty$, $\{\varepsilon_k\} \to 0$, и v_* — предельная точка последовательности $\{u_k\}$, то $v_* \in U_*$.

Доказательство. Прежде всего покажем, что $\Phi_{k*} > -\infty$. Из (21) следует

$$\Phi_k(u) - J_* = J(u) - J_* + A_k P(u) \geqslant -\sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^\nu + A_k \sum_{i=1}^{s} \left(g_i^+(u)\right)^p \geqslant$$

$$\geqslant \sum_{i=1}^{s} \min_{z \geqslant 0} \left(-c_i z^\nu + A_k z^p\right) \qquad \forall u \in U_0. \tag{25}$$

Нетрудно видеть, что функция $\varphi(z) = -c_i z^\nu + A_k z^p$, где $p > \nu$, достигает своей нижней грани при $z \geqslant 0$ в точке $z_* = \left(\frac{\nu c_i}{p A_k}\right)^{1/(p-\nu)}$, причем $\varphi(z_*) = \min_{z \geqslant 0} \varphi(z) = -\nu^{\nu/(p-\nu)} p^{-p/(p-\nu)} c_i^{p/(p-\nu)} (p-\nu) A_k^{-\nu/(p-\nu)}$. Осюда и из (25) следует, что

$$\Phi_k(u) - J_* \geqslant -B A_k^{-\nu/(p-\nu)} \qquad \forall u \in U_0. \tag{26}$$

Переходя к нижней грани по $u \in U_0$, из (26) имеем

$$\Phi_{k*} - J_* \geqslant -B A_k^{-\nu/(p-\nu)}. \tag{27}$$

Отсюда вытекает, что $\Phi_{k*} > -\infty$. Это значит, что при $\varepsilon_k > 0$ точка u_k, удовлетворяющая условиям (6), существует по определению нижней грани (при $\varepsilon_k = 0$ существование такой точки предполагается).

Далее, из (13), (21) имеем

$$J(u_k) + A_k P(u_k) \leqslant J_* + \varepsilon_k \leqslant J(u_k) + \sum_{i=1}^{s} c_i \left(g_i^+(u_k)\right)^\nu + \varepsilon_k, \tag{28}$$

так что

$$0 \leqslant A_k^{-} P(u_k) \leqslant \sum_{i=1}^{s} c_i \left(g_i^+(u_k)\right)^\nu + \varepsilon_k, \qquad k = 0, 1, \ldots \tag{29}$$

Пользуясь неравенством Гельдера

$$\left|\sum_{i=1}^{s} a_i b_i\right| \leqslant \left(\sum_{i=1}^{s} a_i^m\right)^{1/m} \left(\sum_{i=1}^{s} b_i^r\right)^{1/r}, \quad \frac{1}{m} + \frac{1}{r} = 1,$$

при $b_i = c_i$, $a_i = \left(g_i^+(u_k)\right)^\nu$, $m = p/\nu$, $r = p/(p-\nu)$, получаем

$$0 \leqslant \sum_{i=1}^{s} c_i \left(g_i^+(u_k)\right)^\nu \leqslant |c| \left(P(u_k)\right)^{\nu/p}. \tag{30}$$

Отсюда и из (29) следует $0 \leqslant A_k P(u_k) \leqslant |c| (P(u_k))^{\nu/p} + \varepsilon_k$ или $0 \leqslant z^{p/\nu} \leqslant |c| A_k^{-\nu/p} z + \varepsilon_k$ $(k = 0, 1, \ldots)$, где $z = (A_k P(u_k))^{\nu/p}$. С помощью леммы 2.3.11 тогда получаем

$$0 \leqslant (A_k P(u_k))^{\nu/p} \leqslant \left(\left(|c| A_k^{-\nu/p} \right)^{p/(p-\nu)} + \frac{p}{p-\nu} \varepsilon_k \right)^{\nu/p},$$

что равносильно оценке (22). Далее, из (28) с учетом (30) имеем

$$-|c| (P(u_k))^{\nu/p} \leqslant J(u_k) - J_* \leqslant \varepsilon_k.$$

Отсюда и из уже доказанной оценки (22) следует оценка (23). Далее, левые неравенства (24) получаются из (26) при $u = u_k$ и из (27), правые неравенства (24) вытекают из (13). Наконец, утверждение о том, предельная точка v_* последовательности $\{u_k\}$ принадлежит U_*, вытекает из оценок (22)—(24) и доказывается также, как аналогичное утверждение в теореме 2.

Теорема 6. *Пусть задача* (1), (7) *имеет сильно согласованную постановку в смысле определения* 3, $J_* > -\infty$, *последовательность* $\{u_k\}$ *определена условиями* (3), (6), (8), *где* $p = \nu$, $A_k > |c| = \max\limits_{1 \leqslant i \leqslant s} |c_i|$. *Тогда*

$$0 \leqslant \left(g_i^+(u_k)\right)^p \leqslant P(u_k) \leqslant \frac{\varepsilon_k}{A_k - |c|}, \tag{31}$$

$$-|c| \frac{\varepsilon_k}{A_k - |c|} \leqslant J(u_k) - J_* \leqslant \varepsilon_k, \tag{32}$$

$$0 \leqslant \Phi_k(u_k) - J_* \leqslant \varepsilon_k, \quad \Phi_{k*} = J_*. \tag{33}$$

Если $U_* \neq \varnothing$, *то*

$$U_* = U_{k*} = \{u \in U_0 \colon \Phi_k(u) = \Phi_{k*}\}. \tag{34}$$

Если, кроме того, U_0 — *замкнутое множество, функции* $J(u)$, $g_i^+(u)$ *полунепрерывны снизу на* U_0, $\{\varepsilon_k\} \to 0$, *и* v_* — *предельная точка* $\{u_k\}$, *то* $v_* \in U_*$.

Доказательство. Из (25) при $p = \nu$, $A_k > |c|$ имеем $\Phi_k(u) - J_* \geqslant 0$, $u \in U_0$, так что $\Phi_{k*} \geqslant J_* > -\infty$, и последовательность $\{u_k\}$, удовлетворяющая условиям (6), при $\varepsilon_k > 0$ существует. Из (29) при $p = \nu$, $A_k > |c|$ сразу получаем оценку (31). Из нее и из (28) при $p = \nu$ следует оценка (32). Из (13) и (25) при $p = \nu$, $A_k > |c|$ с учетом того, что $\min\limits_{z \geqslant 0} \left(-c_i z^\nu + A_k z^\nu\right) = 0$, приходим к соотношениям (33). Докажем равенство (34). Возьмем произвольную точку $u_* \in U_*$. Тогда $P(u_*) = 0$ и $\Phi_k(u_*) = J(u_*) = J_* = \Phi_{k*}$, так что $u_* \in U_{k*}$. Следовательно, $U_* \subset U_{k*}$. Пусть теперь $u_{k*} \in U_{k*}$, т. е. $\Phi_k(u_{k*}) = \Phi_{k*}$. Это значит, что условие (6) при $u_k = u_{k*}$ выполняется с $\varepsilon_k = 0$. Тогда из оценки (31) при $\varepsilon_k = 0$ получаем $P(u_{k*}) = 0$, т. е. $u_{k*} \in U$. Отсюда, из (33) и из того, что $J(u_{k*}) = \Phi_k(u_{k*}) = \Phi_{k*} = J_*$ следует, что $u_{k*} \in U_*$. Следовательно, $U_{k*} \subset U_*$. Равенство (34) доказано. Последнее утверждение теоремы вытекает из оценок (31)—(33) и доказывается так же, как аналогичное утверждение в теореме 2.

Из теоремы 6 следует, что случай $p = \nu$ интересен тем, что при точной реализации метода штрафных функций (3), (6), (8) решение исходной задачи (1), (7) может быть получено при конечных значениях штрафного коэффициента A_k.

Рассмотрим примеры, которые показывают, что оценки, полученные в теоремах 5, 6, не могут быть существенно улучшены на классе задач (1), (7), имеющих сильно согласованную постановку.

Пример 4. Рассмотрим задачу

$$J(u) = -u \to \inf,\ u \in U = \{u \in E^1\colon g(u) = u \leqslant 0\}.$$

Здесь $J_* = 0$, $U_* = \{0\}$. Функция Лагранжа $L(u, \lambda) = -u + \lambda u$, $u \in U_0 = E^1$, $\lambda \in \Lambda_0 = \{\lambda \in E^1\colon \lambda \geqslant 0\}$ имеет седловую точку ($u_* = 0$, $\lambda^* = 1$), так что согласно (20) неравенство (21) выполнено при $\nu = 1$, $c_i = 1$, $s = 1$. Возьмем штрафную функцию $P_k(u) = A_k(g^+(u))^p = A_k(\max\{u; 0\})^p$, $p \geqslant 1$, $A_k > 1$, $\{A_k\} \to \infty$. Тогда функция (3) будет иметь вид

$$\Phi_k(u) = \begin{cases} -u + A_k u^p, & u \geqslant 0, \\ -u, & u < 0. \end{cases}$$

Нетрудно показать, что

$$\Phi_{k*} = \inf_{E^1} \Phi_k(u) = \begin{cases} 0, & p = \nu = 1, \\ -\dfrac{p-1}{p}(pA_k)^{-1/(p-1)}, & p > 1, \end{cases}$$

причем нижняя грань достигается в точке $u_{k*} = 0$ при $p = 1$ и $u_{k*} = (pA_k)^{-1/(p-1)}$ при $p > 1$, $k = 0, 1, \ldots$ Последовательность $\{u_{k*}\}$ удовлетворяет условиям (5) или (6) с $\varepsilon_k = 0$, причем

$$P(u_{k*}) = \begin{cases} 0, & p = 1, \\ (pA_k)^{-p/(p-1)}, & p > 1, \end{cases} \qquad J(u_{k*}) - J_* = \begin{cases} 0, & p = 1, \\ -(pA_k)^{-1/(p-1)}, & p > 1. \end{cases}$$

Сравнение этих точных равенств с оценками теорем 5, 6 при $\varepsilon_k = 0$ показывает, что в случае $p = 1$ оценки (31), (32) точны, а в случае $p > 1$ оценки (22)—(24) точны по порядку и отличаются от точных оценок лишь константами при степени A_k. Если $\varepsilon_k > 0$, то при $p = 1$ в качестве точки u_k, удовлетворяющей условиям (6) и наиболее удаленной от U_*, здесь можно взять $u_k = \varepsilon_k/(A_k - 1)$ $(k = 0, 1, \ldots)$. Тогда $P(u_k) = u_k = \varepsilon_k(A_k - |c|)^{-1}$, $J(u_k) - J_* = -u_k = -|c|\,\varepsilon_k(A_k - |c|)^{-1}$, $\Phi_k(u_k) - J_* = (A_k - 1)u_k = \varepsilon_k$ $(k = 0, 1, \ldots)$, что совпадает с оценками (31)—(33). Если $\varepsilon_k > 0$, $p = 2$, то точка $u_k = (1/(2A_k)) + (\varepsilon_k/A_k)^{1/2}$ удовлетворяет условиям (6), причем $A_k P(u_k) = A_k u_k^2 = (1/(4A_k)) + \varepsilon_k + (\varepsilon/A_k)^{1/2}$, $J(u_k) - J_* = -u_k$, $\Phi_k(u_k) - J_* = \varepsilon_k - (1/(4A_k))$, что также свидетельствует о том, что оценки (22)—(24) на классе задач с сильно согласованной постановкой не являются грубыми.

Этот же пример показывает, что в теореме 6 требования $A_k > |c|$ не может быть опущено. В самом деле, если $A_k < 1 = |c|$, $p = 1$, то $\Phi_{k*} = -\infty$; если же $A_k = 1 = |c|$, $p = 1$, то $\Phi_k(u) \equiv 0$, $\Phi_{k*} = 0$, $U_{k*} = E^1$ и нарушено равенство (34).

Пример 5. Рассмотрим задачу

$$J(u) = -u \to \inf, \quad u \in U = \{u \in E^1\colon g(u) = u^2 \leqslant 0\}.$$

Здесь $J_* = 0$, $U = U_* = \{0\}$. Функция Лагранжа $L(u, \lambda) = -u + \lambda u^2$, $u \in E^1$, $\lambda \in E^1_+$, седловой точки не имеет, но тем не менее задача имеет сильно согласованную постановку. В самом деле, справедливо неравенство $J_* = 0 \leqslant -u + |u| = -u + (g(u))^{1/2}$ при всех $u \in E^1$, так что неравенство (21) выполняется при $c = 1$, $\nu = 1/2$. Возьмем штрафную функцию $P(u) = (\max\{u^2; 0\})^p = (u^2)^p$. Если $p > 1/2 = \nu$, то функция $\Phi_k(u) = -u + A_k u^{2p}$, $A_k > 0$, $\{A_k\} \to \infty$, достигает нижней грани на $U_0 = E^1$ при $u_{k*} =$

$= (2pA_k)^{-1/(2p-1)}$ $(k = 0, 1, \ldots)$, причем $P(u_{k*}) = (2pA_k)^{-2p/(2p-1)}$, $J(u_{k*}) - J_* = -u_{k*}$, $\Phi_{k*} - J_* = -(2p-1)\left((2p)^{2p}A_k\right)^{-1/(2p-1)}$ $(k = 0, 1, \ldots)$. Как видим, эти оценки лишь константами при степенях A_k отличаются от оценок (22)—(24). Интересно заметить, что с увеличением p оценки ухудшаются. Если $p = \nu = 1/2$, $A_k > 1 = |c|$, то $\Phi_k(u) = -u + A_k|u|$, $\Phi_{k*} = 0$. Точка $u_k = \varepsilon_k(A_k - 1)^{-1}$ удовлетворяет условиям (6) и наиболее удалена от $U_* = \{0\}$. Тогда $P(u_k) = |u_k| = \varepsilon_k(A_k - 1)^{-1}$, $J(u_k) - J_* = -u_k$, $\Phi_k(u_k) - J_* = \varepsilon_k$, что совпадает с оценками (31)—(33).

5. Выполнение соотношений (12) или (16), как показывает пример 2, еще не гарантирует сходимость последовательности $\{u_k\}$ из (6) ко множеству U. Для такой сходимости множество должно удовлетворять некоторым дополнительным условиям.

Определение 4. Скажем, что множество (7) задано *корректными ограничениями* на U_0, если всякая последовательность $\{u_k\} \in U_0$, удовлетворяющая условиям (16), сходится ко множеству U.

Примеры 2, 3 показывают, что одно и то же множество может быть задано как корректными, так и некорректными ограничениями. Как следует из доказательства теоремы 2, ограничения из (7) будут корректными на U_0, если функции $g_i^+(u)$ $(i = 1, \ldots, s)$, полунепрерывны снизу на замкнутом множестве U_0, а множество $U(\delta)$, определяемое согласно (14), ограничено при некотором $\delta > 0$. Корректными будут также ограничения, для которых удается доказать неравенство

$$\rho(u, U) \leqslant h\left(g_1^+(u), \ldots, g_s^+(u)\right) \qquad \forall u \in U_0, \tag{35}$$

где функция $h(t) = h(t_1, \ldots, t_s) > 0$ при всех $t \in E_+^s$, $t \neq 0$, $h(0) = 0$, $\lim\limits_{t \to 0} h(t) = 0$. Приведем важные классы множеств (7), задаваемых корректными ограничениями, для которых неравенство (35) имеет вид

$$\rho(u, U) \leqslant M\left(\max_{1 \leqslant i \leqslant s} g_i^+(u)\right)^{\gamma} \qquad \forall u \in U_0; \quad M > 0, \quad \gamma > 0. \tag{36}$$

Лемма 2. *Пусть U_0 — выпуклое замкнутое множество, функции $g_1(u), \ldots, g_m(u)$ выпуклы и непрерывны на U_0, пусть существует такая точка $\bar{u} \in U_0$, что $g_1(\bar{u}) < 0, \ldots, g_m(\bar{u}) < 0$; пусть множество $U = \{u \in U_0: g_1(u) \leqslant 0, \ldots, g_m(u) \leqslant 0\}$ ограничено. Тогда неравенство* (36) *выполняется с $\gamma = 1$, $M = \operatorname{diam} U \left(\min\limits_{1 \leqslant i \leqslant m} |g_i(\bar{u})|\right)^{-1}$, $\operatorname{diam} U = \sup\limits_{u, v \in U} |u - v|$.*

Доказательство. Введем функцию $g(u) = \max\limits_{1 \leqslant i \leqslant m} g_i(u)$. В силу теоремы 4.2.7 функция $g(u)$ выпукла на U_0. Возьмем произвольную точку $u \in U_0 \setminus U$. Тогда $g(u) > 0$. Функция $f(t) = g(u + t(\bar{u} - u))$ переменной t непрерывна на отрезке $[0, 1]$, $f(0) = g(u) > 0$, $f(1) = g(\bar{u}) < 0$. Следовательно, существует точка $t_0 \in (0, 1)$, такая, что $f(t_0) = 0$. Положим $v = u + t_0(\bar{u} - u)$; тогда $t_0 = |v - u|\,|\bar{u} - u|^{-1}$, $1 - t_0 = |v - \bar{u}|\,|\bar{u} - u|^{-1}$. Пользуясь выпуклостью функции $g(u)$, имеем $g(v) = f(t_0) = 0 \leqslant t_0 g(\bar{u}) + (1 - t_0) g(u)$ или $-t_0 g(\bar{u}) \leqslant (1 - t_0) g(u)$ или $|v - u|\,|g(\bar{u})| \leqslant |v - \bar{u}|\, g^+(u)$. Отсюда с учетом $\bar{u}, v \in U$ получаем, что $\rho(u, U) \leqslant |u - v| \leqslant g^+(u) \times |v - \bar{u}|\,(g(\bar{u}))^{-1} = M g^+(u)$, что и требовалось.

Лемма 3. *Пусть $U = \{u \in E^n: g_i(u) = \langle a_i, u\rangle - b^i \leqslant 0,\ i = 1, \ldots, m\} \neq \varnothing$, где $a_i \in E^n$, $b^i \in \mathbf{R}$ $(i = 1, \ldots, m)$. Тогда ограничения, задающие множество U, корректны на E^n, и неравенство* (36) *выполняется с $\gamma = 1$, $U_0 = E^n$.*

Доказательство. Возьмем произвольную точку $u \notin U$. Так как U — выпуклое замкнутое множество, то согласно теореме 4.4.1 однозначно определяется проекция $w = \mathscr{P}_U(u)$ точки u на U. К задаче определения проекции: $g(v) = |v - u| \to \inf$, $u \in U$, применимы теорема 4.9.4 и лем-

ма 4.9.2, которые гарантируют существование таких чисел $\lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0$, что

$$g'(w) + \sum_{i=1}^{m} \lambda_i g_i'(w) = \frac{w-u}{|w-u|} + \sum_{i=1}^{s} \lambda_i a_i = 0, \quad \lambda_i \left(\langle a_i, w\rangle - b^i\right) = 0, \quad i = 1, \ldots, m.$$

Отсюда, учитывая, что $|w-u| = \rho(u, U)$, имеем

$$u - w = \rho(u, U) \sum_{i \in I(u)} \lambda_i a_i, \quad I(u) = \{i\colon 1 \leqslant i \leqslant m,\ \lambda_i > 0,\ \langle a_i, w\rangle - b^i = 0\}. \tag{37}$$

Можно считать, что система векторов $\{a_i,\ i \in I(u)\}$ линейно независима. В самом деле, если существуют числа γ_i $(i \in I(u))$, не все равные нулю, $\sum_{i \in I(u)} \gamma_i a_i = 0$, то $u - w = \rho(u, U) \sum_{i \in I(u)} (\lambda_i - t\gamma_i) a_i$, где $\nu_i = \lambda_i - t\gamma_i \geqslant 0$ $(i \in I(u))$ при всех t, $0 < t < t_0$, t_0 — достаточно малое число. Можно считать, что среди γ_i $(i \in I(u))$, есть положительные числа, иначе изменим знаки всех γ_i $(i \in I(u))$. Положим $t = \alpha_s/\gamma_s = \min\limits_{\gamma_i > 0, i \in I(u)} \alpha_i/\gamma_i$.

Тогда $\nu_i = \lambda_i - t\gamma_i \geqslant 0$ $(i \in I(u))$, причем по крайней мере одно число $\nu_s = \lambda_s - t\gamma_s = 0$. Таким образом, заменив в (37) λ_i на ν_i и исключив из $I(u)$ те номера, для которых $\nu_i = 0$, снова придем к равенству вида (37) с меньшим числом слагаемых. Последовательно применяя этот прием далее, за конечное число шагов придем к представлению (37), в котором система $\{a_i,\ i \in I(u)\}$ линейно независима. Из (37) следует

$$\max_{1 \leqslant i \leqslant m} g_i^+(u) \geqslant \max_{i \in I(u)} g_i^+(u) \geqslant \max_{i \in I(u)} g_i(u) = \max_{i \in I(u)} \left(\langle a_i, u\rangle - b^i\right) =$$
$$= \max_{i \in I(u)} \langle a_i, u - w\rangle = \rho(u, U) \max_{i \in I(u)} \left\langle a_i, \sum_{j \in I(u)} \lambda_j a_j \right\rangle. \tag{38}$$

Покажем, что величину $\max\limits_{i \in I(u)} \left\langle a_i, \sum\limits_{j \in I(u)} \lambda_j a_j \right\rangle$, где система $\{a_i,\ i \in I(u)\}$ линейно независима, можно оценить снизу положительной величиной, не зависящей от u. С этой целью возьмем любое множество индексов $I \subset \{1, \ldots, m\}$, таких, что векторы $\{a_i,\ i \in I\}$ линейно независимы, и введем множество

$$\Lambda_I = \left\{(\lambda_i,\ i \in I)\colon \lambda_i \geqslant 0,\ \left|\sum_{i \in I} \lambda_i a_i\right| = 1\right\}. \tag{39}$$

Заметим, что Λ_I — замкнутое ограниченное множество. В самом деле, если $\lambda^k = \left(\lambda_i^k,\ i \in I\right) \in \Lambda_I$, $\lambda^k \to \lambda$, то предельным переходом в (39) легко убедиться, что $\lambda \in \Lambda_I$. Следовательно, Λ_I замкнуто. Покажем ограниченность Λ_I. Допустим противное: пусть найдутся $\lambda^k \in \Lambda_I$ $(k = 1, 2, \ldots)$, $|\lambda^k| \to \infty$. Тогда последовательность $\mu^k = \lambda^k/|\lambda^k|$ $(k = 1, 2, \ldots)$ ограничена: $|\mu^k| = 1$. Выбирая при необходимости подпоследовательность, можем считать, что $\{\mu^k\} \to \mu$, $|\mu| = 1$. Поскольку $\left|\sum\limits_{i \in I} \lambda_i^k a_i\right| = 1$, то $\left|\sum\limits_{i \in I} \mu_i^k a_i\right| = 1/|\lambda^k| \to 0 = \sum\limits_{i \in I} \mu_i^0 a_i$, где $\mu = \left(\mu_i^0,\ i \in I\right) \neq 0$. Однако это противоречит линейной независимости $\{a_i,\ i \in I\}$. Следовательно, Λ_I замкнуто.

На множестве Λ_I рассмотрим функцию $d(\lambda, I) = \max\limits_{i \in I} \left\langle a_i, \sum\limits_{j \in I} \lambda_j a_j \right\rangle$. Убедимся в том, что $d(\lambda, I) > 0$ при всех $\lambda \in \Lambda_I$. В самом деле, если су-

ществует $\lambda^0 = (\lambda_j^0,\ j \in I) \in \Lambda_I$, что $d(\lambda^0, I) \leqslant 0$, то $\left\langle a_i, \sum_{j \in I} \lambda_j^0 a_j \right\rangle \leqslant 0$ при всех $i \in I$. Умножим эти неравенства на λ_i^0 $(i \in I)$ и сложим; получим равенство $\left| \sum_{i \in I} \lambda_i^0 a_i \right| = 0$, противоречащее определению Λ_I. Таким образом, $d(\lambda, I) > 0$ при всех $\lambda \in \Lambda_I$. Функция $d(\lambda, I)$ непрерывна по λ и на компактном множестве Λ_I достигает своей нижней грани в некоторой точке $\lambda_* \in \Lambda_I$, причем $d_*(I) = \inf_{\lambda \in \Lambda_I} d(\lambda, I) = d(\lambda^*, I) > 0$. Поскольку множество $\{I\}$ различных подмножеств I множества $\{1, \ldots, m\}$, для которых векторы $\{a_i,\ i \in I\}$ линейно независимы, конечно, то $d_* = \inf_{\{I\}} d_*(I) > 0$. Отсюда и из (37), (38) имеем $\max_{1 \leqslant i \leqslant m} g_i^+(u) \geqslant \rho(u, U)\, d(\lambda, I(u)) \geqslant \rho(u, U)\, d_*$, или $\rho(u, U) \leqslant (1/d_*) \max_{1 \leqslant i \leqslant s} g_i^+(u)$ $(u \in E^n)$. Таким образом, неравенство (36) справедливо с $\gamma = 1$, $M = 1/d_*$. Лемма 3 доказана.

Лемма 4. *Непустое множество*

$$U = \{u \in E^n\colon\ g_i(u) = \langle a_i, u\rangle - b^i \leqslant 0,\quad i = 1, \ldots, m;$$
$$g_i(u) = \langle a_i, u\rangle - b^i = 0,\quad u = m+1, \ldots, s\}, \tag{40}$$

где $a_i \in E^n$, $b^i \in R$, задается корректными ограничениями на E^n и неравенство (36) выполняется с $\gamma = 1$, $U_0 = E^n$.

Доказательство. Каждое ограничение $g_i(u) = \langle a_i, u\rangle - b^i = 0$ заменим равносильными ограничениями $h_{1i}(u) = g_i(u) \leqslant 0$, $h_{2i}(u) = -g_i(u) \leqslant 0$ и воспользуемся леммой 3. Получим

$$\rho(u, U) \leqslant M \cdot \max\{g_1^+(u), \ldots, g_m^+(u);\ h_{1m+1}^+(u), \ldots, h_{1s}^+(u),$$
$$h_{2m+1}^+(u), \ldots, h_{2s}^+(u)\}.$$

Отсюда и из

$$h_{1i}^+(u) = \max\{g_i(u);\ 0\} \leqslant |g_i(u)| = g_i^+(u),$$
$$h_{2i}^+(u) = \max\{-g_i(u);\ 0\} \leqslant |g_i(u)| = g_i^+(u), \qquad i = m+1, \ldots, s,$$

приходим к неравенству (36) с $\gamma = 1$. Лемма 4 доказана.

Другие классы множеств (7), заданных корректными ограничениями, читатель найдет в [21].

6. В лемме 1 был выделен класс задач (1), (7), имеющих сильно согласованную постановку (см. неравенства (20), (21)). Следуя [21], приведем еще один содержательный класс таких задач.

Лемма 5. *Пусть функция $J(u)$ на множестве U_0 удовлетворяет условию Гельдера*

$$|J(u) - J(v)| \leqslant L\,|u - v|^\alpha \qquad \forall u, v \in U_0, \qquad L > 0, \qquad 0 < \alpha \leqslant 1; \tag{41}$$

ограничения, задающие множество (7), *корректны на U_0 и удовлетворяют неравенству* (36). *Тогда задача* (1), (7) *имеет сильно согласованную постановку, причем неравенство* (21) *выполняется при* $c_1 = \ldots = c_s = LM^\alpha$, $= \alpha\gamma$.

Доказательство. Возьмем произвольную точку $u \in U_0$. По определению $\rho(u, U) = \inf_{v \in U} |u - v|$ для любого $\varepsilon > 0$ найдется такая точка

$u_\varepsilon \in U$, что $|u - u_\varepsilon| \leqslant \rho(u, U) + \varepsilon$. Тогда с учетом условий (36), (41) имеем

$$LM^\alpha \sum_{i=1}^{s} \left(g_i^+(u)\right)^{\alpha\gamma} + J(u) - J_* \geqslant LM^\alpha \left(\max_{1 \leqslant i \leqslant s} g_i^+(u)\right)^{\alpha\gamma} + J(u) - J(u_\varepsilon) \geqslant$$
$$\geqslant L(\rho(u, U))^\alpha - L|u - u_\varepsilon|^\alpha \geqslant L(\rho(u, U))^\alpha - L(\rho(u, U) + \varepsilon)^\alpha.$$

Пользуясь произволом $\varepsilon > 0$, отсюда при $\varepsilon \to 0$ получим

$$LM^\alpha \sum_{i=1}^{s} \left(g_i^+(u)\right)^{\alpha\gamma} + J(u) - J_* \geqslant 0 \qquad \forall u \in U_0.$$

Заметим, что в общем случае из выполнения условий леммы 5 не следует существование седловой точки функции Лагранжа задачи (1), (7) и, наоборот, существование седловой точки не гарантирует выполнение условий леммы 5. Это означает, что выделенные в леммах 1, 5 два класса задач (1), (7), имеющих сильно согласованную постановку, взаимно дополняют друг друга. Отметим также, что этими двумя классами не исчерпываются задачи (1), (7) с сильно согласованной постановкой. Поясним это на примере.

Пример 6. Рассмотрим задачу

$$J(u) = -u^\alpha \to \inf, \quad u \in U = \{u \geqslant 0\colon g(u) = u^\beta \leqslant 0\}, \tag{42}$$

где $\alpha > 0$, $\beta > 0$, $U_0 = \{u \in E^1\colon u \geqslant 0\} = E^1_+$. Ясно, что $U = U_* = \{0\}$ $J_* = 0$.

Далее, имеем

$$J_* = 0 = -u^\alpha + (u^\beta)^{\alpha/\beta} = J(u) + (g^+(u))^{\alpha/\beta} \qquad \forall u \in U_0,$$

так что неравенство (21) выполняется при $s = m = 1$, $c_i = 1$, $\nu = \alpha/\beta$. Следовательно, задача (42) имеет сильно согласованную постановку и к ней применимы теоремы 5, 6. Отметим, что здесь $\rho(u, U) = |u - 0| = |u| = (g^+(u))^{1/\beta}$, $u \in U_0$, т. е. условие (36) выполняется с $M = 1$, $\gamma = 1/\beta$. Далее, при $0 < \alpha \leqslant 1$ функция $J(u) = -u^\alpha$ удовлетворяет условию Гельдера: $|u^\alpha - v^\alpha| \leqslant |u - v|^\alpha$ $(u, v \in U_0)$, так что в этом случае применима лемма 5. При $\alpha > 1$ условие Гельдера на $U_0 = E^1_+$ не выполняется и лемма 5 неприменима. Далее, функция Лагранжа $L(u, \lambda) = -u^\alpha + \lambda u^\beta$ $(u \geqslant 0, \lambda \geqslant 0)$ задачи (42) при $\alpha = \beta$ имеет седловую точку $(u_* = 0, \lambda^* = 1)$. Кстати, седловая точка здесь не единственная: любая точка $(0, \lambda^*)$, $\lambda^* \geqslant 1$, также является седловой. Заметим также, что функция $J(u) = -u^\alpha$, $u \geqslant 0$, выпукла лишь при $0 < \alpha \leqslant 1$, а $g(u) = u^\beta$, $u \geqslant 0$, выпукла лишь при $\beta \geqslant 1$. Если $\alpha \neq \beta$, то функция Лагранжа не имеет седловой точки. Таким образом, при $\alpha > 1$, $\beta > 0$, $\alpha \neq \beta$ задача (42) не охватывается леммами 1, 5.

7. Покажем, что метод штрафных функций может быть использован для поиска седловой точки функции Лагранжа задачи (1), (7).

Теорема 7. *Пусть U_0 — выпуклое замкнутое множество из E^n; функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ выпуклы на U_0 и принадлежат классу $C^1(U_0)$; $g_i(u) = \langle a_i, u\rangle - b^i$ $(i = m+1, \ldots, s)$; пусть $J_* > -\infty$, $U_* \neq \varnothing$ и функция Лагранжа задачи* (1), (7) *имеет седловую точку* $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ *в смысле неравенства* (19). *Кроме того, пусть функция* $\Phi_k(u) = J(u) + A_k \sum_{i=1}^{s} \left(g_i^+(u)\right)^p$, $u \in U_0$, $p > 1$, *при каждом* $k = 0, 1, \ldots$ *достигает своей нижней грани на U_0, в некоторой точке $u_k \in U_0$. Тогда если последовательность $\{u_k\}$ имеет хотя бы одну предельную точку, то и последовательность $\{\lambda^k\}$, где* $\lambda^k = (\lambda_1^k, \ldots, \lambda_s^k)$, $\lambda_i^k = pA_k \left|g_i^+(u_k)\right|^{p-1}$ $(i = 1, \ldots, m)$; $\lambda_i^k = pA_k \left|g_i^+(u_k)\right|^{p-1} \operatorname{sign} g_i(u_k)$ $(i = m+1, \ldots, s)$ *также будет иметь пре-*

дельную точку, причем любая предельная точка последовательности $\{(u_k, \lambda^k)\}$ *будет седловой точкой функции Лагранжа.*

Доказательство. С учетом леммы 1 из оценки (22) при $\varepsilon_k = 0$, $\nu = 1$, $c_i = |\lambda_i^*|$ получим $|g_i^+(u_k)| \leqslant (|\lambda^*|/A_k)^{1/(p-1)}$. Тогда $A_k |g_i^+(u_k)|^{p-1} \leqslant |\lambda^*|$ при всех $i = 1, \ldots, s$, $k = 0, 1, \ldots$ Отсюда следует, что $|\lambda_i^k| \leqslant p|\lambda^*|$ $(i = 1, \ldots, s,\ k = 0, 1, \ldots)$, т. е. последовательность $\{\lambda^k\}$ ограничена. Пусть v_* — какая-либо предельная точка последовательности $\{u_k\}$, пусть $\{u_{k_r}\} \to v_*$. Согласно теореме 5 $v_* \in U_*$. Из $\{\lambda^{k_r}\}$ выделим подпоследовательность, сходящуюся к некоторой точке μ^*. Можем считать, что сама последовательность $\{\lambda^{k_r}\}$ сходится к μ^*. Так как $\lambda_i^k \geqslant 0$ при $i = 1, \ldots, m$, то и $\mu_i^* \geqslant 0$ $(i = 1, \ldots, m)$, так что $\mu^* \in \Lambda_0$. Далее, так как $\Phi_k(u)$ выпукла и $\Phi_k(u) \in C^1(U_0)$, то согласно теореме 4.2.3 в точке u_k имеем

$$\langle \Phi_k'(u_k),\ u - u_k \rangle \geqslant 0 \quad \forall u \in U_0. \tag{43}$$

Поскольку

$$\Phi_k'(u_k) = J'(u_k) + \sum_{i=1}^{m} pA_k |g_i^+(u_k)|^{p-1} g_i'(u_k) +$$

$$+ \sum_{i=m+1}^{s} pA_k |g_i^+(u_k)|^{p-1} \operatorname{sign} g_i(u_k)\, g_i'(u_k) = J'(u_k) + \sum_{i=1}^{s} \lambda_i^k g_i'(u_k),$$

то

$$\lim_{r\to\infty} \Phi_{k_r}'(u_{k_r}) = J'(v_*) + \sum_{i=1}^{s} \mu_i^* g_i'(u_*) = L_u(v_*, \mu^*).$$

Поэтому из (43) при $k = k_r \to \infty$ получим $\langle L_u(v_*, \mu^*),\ u - v_* \rangle \geqslant 0$, $u \in U_0$. Так как $\lambda^* \in \Lambda_0$, то при выполнении условий теоремы функция $L(u, \lambda^*)$ выпукла на U_0, и последнее неравенство равносильно такому:

$$L(v_*, \mu^*) \leqslant L(u, \mu^*) \qquad \forall u \in U_0. \tag{44}$$

Далее, если $g_i(v_*) < 0$ при некотором i $(1 \leqslant i \leqslant m)$, то из $\lim\limits_{r\to\infty} g_i(u_{k_r}) = g_i(v_*) < 0$ следует, что $g_i(u_{k_r}) < 0$ или $g_i^+(u_{k_r}) = 0$ для всех $r \geqslant r_0$. А тогда $\lambda_i^{k_r} = 0$ при всех $r \geqslant r_0$ и $\lim\limits_{r\to\infty} \lambda_i^{k_r} = \mu_i^* = 0$. Таким образом, $\mu_i^* = 0$ для всех номеров i $(1 \leqslant i \leqslant m)$, для которых $g_i(v_*) < 0$, а для остальных номеров i $(1 \leqslant i \leqslant s)$, мы имеем $g_i(v_*) = 0$. Следовательно, $\mu_i^* g_i(v_*) = 0$ $(i = 1, \ldots, s)$. Отсюда и из (44) с помощью леммы 4.9.2 получим, что (v_*, μ^*) — седловая точка функции Лагранжа. Теорема доказана.

8. Метод штрафных функций может быть использован и для получения условий оптимальности в задаче (1), (7). В частности, с помощью этого метода можно получить другое довольно простое доказательство правила множителей Лагранжа, правда, при несколько больших требованиях, чем в § 4.8.

Теорема 8. *Пусть в задаче* (1), (7) U_0 — *выпуклое замкнутое множество, функции* $J(u)$, $g_1(u), \ldots, g_s(u) \in C^1(U_0)$, $J_* > -\infty$, $U_* \neq \varnothing$. *Если* $u_* \in U_*$, *то существуют числа* $\lambda_0^* \geqslant 0, \ldots, \lambda_m^* \geqslant 0, \lambda_{m+1}^*, \ldots, \lambda_s^*$, *не все равные нулю и такие, что*

$$\langle \lambda_0^* J'(u_*) + \lambda_1^* g_1'(u_*) + \ldots + \lambda_s^* g_s'(u_*),\ u - u_* \rangle \geqslant 0 \qquad \forall u \in U_0, \tag{45}$$

$$\lambda_i^* g_i(u_*) = 0, \qquad i = 1, \ldots, s. \tag{46}$$

Доказательство. Введем новую функцию $g_0(u)=J(u)+|u-u_*|^2$, множество $W_0=U_0\cap S(u_*)$, где $S(u_*)=\{u\in E^n\colon |u-u_*|\leqslant 1\}$, и рассмотрим вспомогательную задачу минимизации

$$g_0(u)\to\inf,\quad u\in W=\{u\in W_0\colon g_1(u)\leqslant 0,\ \ldots,\ g_m(u)\leqslant 0,\ g_{m+1}(u)=0,\ \ldots,\ g_s(u)=0\}. \tag{47}$$

Так как $g_0(u)>J(u)\geqslant J_*$ при всех $u\in W$, $u\neq u_*$, причем $g_0(u_*)=J_*$, $u_*\in W$, то ясно, что u_* — единственное решение задачи (47). Применим к задаче (47) метод штрафных функций. Введем функцию $\Phi_k(u)=g_0(u)+A_k\sum_{i=1}^{s}\left(g_i^+(u)\right)^p$, $u\in W_0$, $p>1$. Так как W_0 — компактное множество, функции $g_0(u)$, $\Phi_k(u)$ непрерывны на W_0, то $g_{0*}=\inf_{W_0} g_0(u)>-\infty$, $\Phi_{k*}=\inf_{W_0}\Phi_k(u)>-\infty$, и существует точка $u_k\in W_0$, для которой $\Phi_k(u_k)=\Phi_{k*}$. Далее, множество $W(\delta)=\{u\in W_0\colon g_i^+(u)\leqslant\delta,\ i=1,\ldots,s\}$ ограничено при всех $\delta>0$, так как W_0 ограничено. По теореме 2 тогда $\lim_{k\to\infty}|u_k-u_*|=0$, $\lim_{k\to\infty} g_0(u_k)=\lim_{k\to\infty} J(u_k)=J_*$. Применяя теорему 4.2.3 к задаче: $\Phi_k(u)\to\inf$, $u\in W_0$, имеем

$$\langle\Phi_k'(u_k),\ u-u_k\rangle\geqslant 0\qquad\forall u\in W_0. \tag{48}$$

Покажем, что это неравенство на самом деле верно для всех $u\in U_0$, если номер k достаточно большой. А именно, выберем номер k_0 таким, чтобы $|u_k-u_*|<1/2$ при всех $k\geqslant k_0$. Возьмем произвольную точку $v\in U_0$. Зафиксируем число α $(0<\alpha\leqslant 1)$, столь малым, чтобы $\alpha|v-u_*|<1/2$. Тогда точка $v_\alpha=u_k+\alpha(v-u_k)\in U_0$ и, кроме того, $|v_\alpha-u_*|=|(1-\alpha)(u_k-u_*)+\alpha(v-u_*)|\leqslant(1-\alpha)|u_k-u_*|+\alpha|v-u_*|\leqslant 1$. Следовательно, $v_\alpha\in W_0$, и в (48) можем принять $u=v_\alpha$. Получим $\langle\Phi_k'(u_k),\ \alpha(v-u_k)\rangle\geqslant 0$ или $\langle\Phi_k'(u_k),\ v-u_k\rangle\geqslant 0$. Таким образом, показано, что при каждом $k\geqslant k_0$ неравенство (48) выполняется при всех $u\in U_0$. Подставим в (48) явное выражение для производной $\Phi_k'(u_k)$; получим

$$\left\langle J'(u_k)+2(u_k-u_*)+\sum_{i=1}^{s}\mu_{ik}g_i'(u_k),\ u-u_k\right\rangle\geqslant 0\qquad\forall u\in U_0,\quad k\geqslant k_0, \tag{49}$$

где $\mu_{ik}=pA_k\left|g_i^+(u_k)\right|^{p-1}\geqslant 0$ $(i=1,\ldots,m)$, $\mu_{ik}=pA_k\left|g_i^+(u_k)\right|^{p-1}\times\operatorname{sign} g_i(u_k)$ $(i=m+1,\ldots,s)$. Разделим неравенство (49) на $\left(1+\sum_{i=1}^{s}\mu_{ik}^2\right)^{1/2}>1$; будем иметь

$$\left\langle\lambda_{0k}J'(u_k)+2\lambda_{0k}(u_k-u_*)+\sum_{i=1}^{s}\lambda_{ik}g_i'(u_k),\ u-u_k\right\rangle\geqslant 0\quad\forall u\in U_0,\ k\geqslant k_0, \tag{50}$$

где $\lambda_{0k}=\left(1+\sum_{i=1}^{s}\mu_{ik}^2\right)^{-1/2}>0$, $\lambda_{ik}=\mu_{ik}\left(1+\sum_{i=1}^{s}\mu_{ik}^2\right)^{-1/2}$ $(i=1,\ldots,s,$

$\lambda_{1k} \geqslant 0, \ldots, \lambda_{mk} \geqslant 0, \ k \geqslant k_0$). Ясно, что $\sum_{i=0}^{s} \lambda_{ik}^2 = 1$, так что последовательность $\{\bar{\lambda}^k = (\lambda_{0k}, \ldots, \lambda_{sk})\}$ ограничена.

Пользуясь теоремой Больцано — Вейерштрасса и выбирая при необходимости сходящуюся подпоследовательность, можем считать, что $\{\bar{\lambda}^k\} \to \bar{\lambda}^* = (\lambda_0^*, \ldots, \lambda_s^*)$, причем $\lambda_i^* \geqslant 0$ $(i = 0, 1, \ldots, m)$, $|\bar{\lambda}_i^*| = 1$, так что не все числа $\lambda_0^*, \lambda_1^*, \ldots, \lambda_s^*$ равны нулю. Так как $J'(u)$, $g_i'(u)$ непрерывны, $\{u_k\} \to u_*$, то из (50) при $k \to \infty$ придем к неравенству (45). Далее, если $g_i(u_*) = 0$, то равенства (46) для таких i, очевидно, выполняются. Если же $g_i(u_*) < 0$ при некотором i, $1 \leqslant i \leqslant m$, то $g_i(u_k) < 0$ при всех $k \geqslant k_1$, а тогда $\mu_{ik} = 0$, $\lambda_{ik} = 0$, $k \geqslant k_1$, и, следовательно, $\lambda_i^* = 0$. Равенства (46) также доказаны.

Замечание 1. Предположим, что наряду с условиями теоремы 8 для задачи (1), (7) справедливо неравенство (21) с параметром $\nu \geqslant 1$. Тогда

$$g_{0*} = \inf_W g_0(u) = J_* \leqslant J(u) + \sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^\nu \leqslant g_0(u) + \sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^\nu,$$
$$\forall u \in U_0,$$

так что задача (47) также имеет сильно согласованную постановку с теми же параметрами c_i, ν. Пользуясь оценкой (22) при $\varepsilon_k = 0$, $p > \nu \geqslant 1$, получим

$$0 \leqslant \mu_{ik} = pA_k \left(g_i^+(u_k)\right)^{p-1} \leqslant p\,|c|^{(p-1)/(p-\nu)} A_k^{-(\nu-1)/(p-\nu)} \leqslant$$
$$\leqslant p\,|c|^{(p-1)/(p-\nu)} < \infty, \qquad i = 1, \ldots, s, \quad k = 0, 1, \ldots \tag{51}$$

Выбирая при необходимости подпоследовательность, можем считать, что $\{\mu_{ik}\} \to \lambda_i^*$ $(i = 1, \ldots, s)$. Отсюда и из (49) при $k \to \infty$ получим неравенство (45) с $\lambda_0^* = 1$. Таким образом, гладкие задачи (1), (7) с сильно согласованной постановкой с параметром $\nu \geqslant 1$, являются регулярными. Если $\nu > 1$, то, как видно из (51), $\lim_{k\to\infty} \mu_{ik} = \lambda_i^* = 0$ $(i = 1, \ldots, s)$, и в качестве множителей Лагранжа в (45), (46) можно взять $\lambda_0^* = 1$, $\lambda_1^* = \ldots = \lambda_s^* = 0$. Это значит, что необходимые условия оптимальности в задаче (1), (7) совпадают с необходимым условием в задаче: $J(u) \to \inf$, $u \in U_0$, и ограничения $g_i(u) \leqslant 0$, $g_j(u) = 0$ в (1), (7) не играют существенной роли. Отсюда следует, что класс гладких задач, удовлетворяющих неравенству (21) с параметром $\nu > 1$, хотя и не пуст (см. пример 6), но не является слишком содержательным.

9. Отдельно остановимся на случае $\nu = 1$. Оказывается, класс выпуклых задач (1), (7), которые удовлетворяют неравенству (21) с параметром $\nu = 1$, является подмножеством задач, у которых функция Лагранжа имеет седловую точку.

Теорема 9. *Пусть W — открытое выпуклое множество из E^n, функции $J(u)$, $g_1(u), \ldots, g_m(u)$ выпуклы на W, $g_i(u) = \langle a_i, u\rangle - b^i$ $(i = m+1, \ldots, s)$, U_0 — выпуклое подмножество из W, пусть в задаче* (1), (7) *$J_* > -\infty$, $U_* \neq \varnothing$. Тогда для того, чтобы в задаче* (1), (7) *функция Лагранжа имела седловую точку, необходимо и достаточно, чтобы неравенство* (21) *выполнялось с показателем $\nu = 1$.*

Доказательство. Необходимость доказана в лемме 1. Докажем достаточность. Пусть выпуклая задача (1), (7) удовлетворяет неравенству (21) с $\nu = 1$, пусть $u_* \in U_*$. Согласно теореме 4.6.1 субдифференциалы $\partial J(u)$, $\partial g_i(u)$ непусты при всех $u \in W$. Штрафная функция $\Phi(u, A) = J(u) + A\sum_{i=1}^{s} g_i^+(u)$ $(A > 0)$ выпукла на W. Пользуясь правилами 7), 9) субдифференцирования из § 4.6 и представлением

$$g_i^+(u) = |g_i(u)| = \max\{\langle a_i, u\rangle - b^i; 0\} + \max\{-\langle a_i, u\rangle + b^i; 0\},$$
$$i = m+1, \ldots, s,$$

имеем

$$\partial\Phi(u, A) = \partial J(u) + \sum_{i=1}^{m} A\mu_i \partial g_i(u) + \sum_{i=m+1}^{s} A\mu_i a_i, \tag{52}$$

где $0 \leqslant \mu_i \leqslant 1$ при $i = 1, \ldots, m$, причем $\mu_i = 0$ при $g_i(u) < 0$, $\mu_i = 1$ при $g_i(u) > 0$; $-1 \leqslant \mu_i \leqslant 1$ при $i = m+1, \ldots, s$, причем $\mu_i = \operatorname{sign}(\langle a_i, u\rangle - b^i)$ при $\langle a_i, u\rangle - b^i \neq 0$. Применяя теорему 6 при $p = \nu = 1$, $A > |c|$, заключаем, что $\Phi(u_*, A) = \Phi_* = \inf\limits_{U_0} \Phi(u, A)$. Тогда по теореме 4.6.4 найдется такой субградиент $c_* \in \partial\Phi(u_*, A)$, что

$$\langle c_*, u - u_*\rangle \geqslant 0 \qquad \forall u \in U_0. \tag{53}$$

Согласно (52) для c_* справедливо представление $c_* = c_0^* + \sum_{i=1}^{s} A\mu_i^* c_i^*$, где $c_0^* \in \partial J(u)$, $c_i^* \in \partial g_i(u_*)$, $0 \leqslant \mu_i^* \leqslant 1$ при $i = 1, \ldots, m$, причем $\mu_i^* = 0$ при $g_i(u_*) < 0$; $c_i^* = a_i$, $-1 \leqslant \mu_i^* \leqslant 1$ при $i = m+1, \ldots, s$. Положим $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*)$: $\lambda_i^* = A\mu_i^*$ $(i = 1, \ldots, s)$, где A фиксировано из условия $A > |c|$. Далее, заметим, что при сделанных предположениях функция $L(u, \lambda) = J(u) + \sum_{i=1}^{s} \lambda_i g_i(u)$ выпукла по переменной $u \in W$ при каждом $\lambda \in \Lambda_0 = \{\lambda = (\lambda_1, \ldots, \lambda_s): \lambda_1 \geqslant 0, \ldots, \lambda_m \geqslant 0\}$, и, следовательно, $\partial L(u, \lambda) \neq \varnothing$ при всех $u \in W$, $\lambda \in \Lambda_0$. Нетрудно видеть, что $\lambda^* \in \Lambda_0$, $c_* = c_0^* + \sum_{i=1}^{s} \lambda_i^* c_i^* \in \partial L(u_*, \lambda^*)$. Тогда из (53) и теоремы 4.6.4 получаем неравенство $L(u_*, \lambda^*) \leqslant L(u, \lambda^*)$ для всех $u \in U_0$. Кроме того, из определения λ_i^* следует, что $\lambda_i^* g_i(u_*) = 0$ $(i = 1, \ldots, s)$. Согласно лемме 4.9.1 (u_*, λ^*) — седловая точка функции Лагранжа задачи (1), (7).

Доказанная теорема 9 дополняет теоремы Куна — Таккера из § 4.9. В частности, опираясь на лемму 5 при $\alpha = \gamma = 1$ и теорему 9, можно получить существование седловой точки для некоторых классов выпуклых задач (1), (7) с нерегулярным множеством в смысле определения 4.9.2 (см. упражнение 9).

10. Рассмотренный выше метод штрафных функций дает простую и универсальную схему решения задач минимизации на множествах, не совпадающих со всем пространством, и часто применяется на практике. Поскольку имеется достаточно богатый выбор штрафных функций, то при составлении функции $\Phi_k(u)$ можно постараться обеспечить нужную гладкость этой функции, выпуклость, подумать об удобствах вычисления значений функции и требуемых ее производных и т. п. Кроме того, имеется определенная свобода в выборе множества U_0 для задачи (2): в задании мно-

жества (7) всегда можно отнести ко множеству U_0 наиболее простые ограничения (например, U_0 может быть шаром или параллелепипедом в E^n, совпадать с полупространством или со всем пространством E^n и т. д.), а остальные ограничения оформить в виде $g_i(u) \leqslant 0$ или $g_j(u) = 0$ и учесть их с помощью штрафной функции. Поэтому можно надеяться на то, что вспомогательные задачи (2), (3) удастся сформировать более простыми, более удобными для применения известных и несложных методов минимизации, чем исходная задача (1).

Следует заметить, что хотя сама схема метода штрафных функций довольно проста, но при практическом использовании этого метода для решения конкретных задач минимизации могут встретиться серьезные трудности. Дело в том, что для получения хорошего приближения решения задачи (1) номер k в (2), (3) (или штрафной коэффициент A_k в (8)) приходится брать достаточно большим. А с увеличением номера k свойства функции $\Phi_k(u) = J(u) + P_k(u)$ $(u \in U_0)$, оказывается, во многих случаях начинают ухудшаться: эта функция может стать более овражной, некоторые координаты градиента $\Phi_k'(u)$ могут быть слишком большими, могут появиться дополнительные локальные минимумы и т. п. Это все может привести к тому, что при больших k методы минимизации, используемые для решения задачи (2), будут плохо сходиться и определение точки u_k, удовлетворяющей условиям (6), с возрастанием k может потребовать все большего и большего объема вычислительной работы.

Поэтому при практическом применении метода штрафных функций вспомогательные задачи (2) обычно решают лишь для таких номеров k (возможно больших), для которых удается обеспечить достаточно быстрое убывание функции $J(u)$ и достаточную близость получаемых точек ко множеству U при небольшом объеме вычислительной работы. Если полученное на этом пути приближение к решению задачи (1) недостаточно хорошее, то привлекают более тонкие и, вообще говоря, более трудоемкие методы минимизации, стараясь при этом получше использовать ту информацию, которая получена с помощью метода штрафных функций.

Заметим, что если выполнены условия теоремы 6 и в качестве штрафной функции множества (7) берется функция (8) при $p = \nu$, то нет необходимости неограниченно увеличивать штрафной коэффициент A_k, и в этом случае упомянутый недостаток метода штрафных функций, вообще говоря, не будет проявляться. Правда, штрафная функция (8) при $p = \nu$ не всегда будет обладать достаточной гладкостью, но появившиеся в последнее время методы минимизации, не требующие гладкости минимизируемой функции (см., например, § 3.11, 12, 17), позволяют надеяться, что численное решение задачи (2) в рассматриваемом случае не будет слишком трудным.

Отметим, что при описании и исследовании метода штрафных функций выше мы предполагали, что функция $J(u)$ и множество (7) известны точно. Если же указанные исходные данные известны лишь приближенно, то метод штрафных функций полезно регуляризовать [6, 22, 84, 86, 87, 177, 329, 343].

Различные прикладные и теоретические аспекты метода штрафных функций исследованы в [6, 8, 11, 12, 19, 21, 109, 111, 122, 129, 144, 148, 159, 173, 174, 177, 240, 306, 307, 314, 330, 338].

Упражнения. 1. Применить метод штрафных функций к задачам

а) $J(u) = x^2 + y^2 \to \inf$; $u \in U = \{u = (x, y) \in E^2\colon g(u) = -x - y + 1 \leqslant 0\}$ или $u \in U = \{u = (x, y) \in E^2\colon g(u) = -x - y + 1 = 0\}$;

б) $J(u) = xy \to \inf$; $u \in U = \{u = (x, y) \in E^2\colon x^2 + y^2 \leqslant 25\}$ или $u \in U = \{u = (x, y) \in E^2\colon x^2 + y^2 = 25\}$;

в) $J(u) = x^2 + y^2 + z^2 \to \inf$; $u \in U = \{u = (x, y, z) \in E^2\colon x + y + z + 1 \leqslant 0\}$.

2. Применить метод штрафных функций к задачам из примеров 2.2.2 и 2.2.4.

3. Пусть $J(u) = e^{-2u}$, а множество $U = \{u \in E^1\colon 0 \leqslant u \leqslant 1\}$ задано либо ограничениями $g_1(u) = -u \leqslant 0$, $g_2(u) = u - 1 \leqslant 0$, либо $g(u) = |u| + |u-1| - 1 = 0$, либо $g(u) = e^{-u}(|u| + |u-1| - 1) = 0$. Выяснить, в каких случаях задача $J(u) \to \inf$, $u \in U$, имеет согласованную или сильно согласованную постановку на E^1.

4. Пусть $\{P_k(u)\}$ — штрафная функция некоторого множества U. Пусть функция $\varphi(t)$ определена при $t \geqslant 0$, $\varphi(0) = 0$, причем $\varphi(t) \to 0$ при $t \to 0$, $\varphi(t) \to \infty$ при $t \to \infty$. Показать, что тогда $\{\varphi(P_k(u))\}$ является штрафной функцией множества U.

5. Применить метод (3), (6), (8) к задаче $J(u) = u^2 - u \to \inf$ и $u \in U = \{u \in E^1\colon g(u) = u \leqslant 0\}$, взяв в качестве штрафной функции $P(u) = (\max\{0; u\})^2$. Получить точную оценку погрешности, сравнить ее с оценками из теоремы 5.

6. Пусть множество U задано либо ограничениями $g_1(u) = u - 1 \leqslant 0$, $g_2(u) = -u - 1 \leqslant 0$, либо $g(u) = e^{-u^2}(u^2 - 1) \leqslant 0$, либо $g(u) = u^2 - 1 \leqslant 0$. Выяснить, какие из этих ограничений являются корректными на E^1 или $U_0 = \{u \in E^1\colon -1 \leqslant u \leqslant 1\}$.

7. Пусть $U = \{u \in E^n\colon g(u) \leqslant 0\}$, где $g(u)$ — непрерывная функция на E^n. Доказать, что для того, чтобы множество U было ограниченным и ограничение $g(u) \leqslant 0$ было корректным на E^n, необходимо и достаточно, чтобы множество $U_\delta = \{u \in E^n\colon g(u) \leqslant \delta\}$ было ограниченным хотя бы при одном $\delta > 0$.

8. Пусть U_0 — выпуклое замкнутое множество из E^n, функция $g(u)$ выпукла и полунепрерывна снизу на U_0, и пусть множество $M(C) = \{u \in U_0\colon g(u) \leqslant C\}$ непусто и ограничено при некотором C. Доказать, что тогда ограничение $g(u) \leqslant 0$ корректно на U_0 (см. теорему 4.2.17).

9. Рассмотреть задачу: $J(u) = u \to \inf$, $u \in U = \{u \in E^1\colon g(u) = u^2 + \varepsilon|u| \leqslant 0\}$, $\varepsilon > 0$. Доказать, что здесь выполняется неравенство (36) с $M = 1/\varepsilon$, $\gamma = 1$. Пользуясь леммой 5 и теоремой 9, установить существование седловой точки Лагранжа. Выполняются ли здесь условия теорем 4.9.2, 4.9.5?

10. Применить метод штрафных функций к задаче (42), получив оценки скорости сходимости метода и сравнить их с оценками из теорем 5, 6.

11. Применить метод штрафных функций к задаче: $J(u) = x^2 + (1 - xy)^2 \to \inf$, $u \in U = \{u = (x, y) \in E^2\colon g(u) = x - a = 0\}$, исследовать его сходимость при различных значениях параметра a.

12. Доказать, что множество $U = \{u \in E^n\colon g_i(u) = \langle a_i, u\rangle - b^i = 0, i = 1, \ldots, s\}$, где $a_1, \ldots, a_s$ — линейно независимые векторы из E^n, $b^i \in \mathbf{R}$, является корректным на E^n и неравенство (36) выполняется с $\gamma = 1$, $M = s\|A^T(AA^T)^{-1}\|$, A — матрица размера $s \times n$, строками которой являются векторы $a_1, \ldots, a_s$. Указание: воспользоваться результатами примера 4.4.3.

13. Пусть задача (44) удовлетворяет условиям теоремы 4.9.2, причем $\bar{u} \notin U_*$. Доказать, что тогда $U_* = \{u \in U_0\colon \Phi(u) = \Phi_*\}$, где

$$\Phi(u) = \ln \frac{1}{J(\bar{u}) - J(u)} + \sum_{i=1}^{m} \max\left\{0; \frac{g_i(u)}{g_i(\bar{u})}\right\}, \qquad \Phi_* = \inf_{U_0} \Phi(u).$$

§ 15. Метод барьерных функций

1. Идеи метода штрафных функций могут быть использованы для построения методов решения задачи

$$J(u) \to \inf; \quad u \in U, \tag{1}$$

позволяющих получить такую минимизирующую последовательность $\{u_k\} \in U$, каждый член которой будет лежать вне некото-

рого заданного «запрещенного» подмножества $\gamma \subset U$. В качестве «запрещенного» множества γ может служить, например, граница $\operatorname{Гр} U$ множества U или какая-либо часть границы. Дело в том, что при применении того или иного метода решения задачи (1) при $U \neq E^n$ может случиться, что каждое получаемое приближение u_k будет принадлежать $\operatorname{Гр} U$. Однако если структура границы множества слишком сложна, то реализация такого метода может потребовать большого объема вычислительной работы и, кроме того, сходимость метода может оказаться очень медленной. В таких случаях можно попробовать как-то построить «барьер» вблизи всей границы $\gamma = \operatorname{Гр} U$ или какой-либо ее части γ (или какого-либо другого заданного подмножества $\gamma \subset U$), который исключал бы возможность попадания очередного приближения u_k на γ.

Определение 1. Пусть γ — некоторое подмножество множества U. Функцию $B(u)$ назовем *барьером* или *барьерной функцией* подмножества γ, если $B(u)$ определена, конечна и неотрицательна во всех точках $u \in U \backslash \gamma$, причем $\lim\limits_{r \to \infty} B(v_r) = \infty$ для всех последовательностей $\{u_r\} \in U \backslash \gamma$, которые сходятся к какой-либо точке $v \in \gamma$.

Заметим, что в определении 1 подразумевается, что $U \backslash \gamma \neq \neq \varnothing$. Это значит, что если $\gamma = \operatorname{Гр} U$, то $\operatorname{int} U = U \backslash \gamma \neq \varnothing$. Заметим также, что в точках $u \in \gamma$ барьерная функция $B(u)$ не определена (можно принять $B(u) = \infty$, $u \in \gamma$).

Пользуясь теми же конструкциями, которые использовались при построении штрафных функций, нетрудно выписать барьерные функции для множеств γ, задаваемых ограничениями типа равенств или неравенств. Например, если $\gamma = \{u \in E^n\colon u \in U, g(u) = 0\}$, где $g(u)$ непрерывна на U, $U \backslash \gamma \neq \varnothing$, то в качестве барьерной функции здесь можно взять $B(u) = |g(u)|^{-1}$, или $B(u) = |g(u)|^{-2}$, или $B(u) = \max\{-\ln|g(u)|;\ 0\}$. Если же $\gamma = \{u \in E^n\colon u \in U, g(u) \leqslant 0\}$, где $U \backslash \gamma \neq \varnothing$, $g(u)$ непрерывна на U, то можно принять $B(u) = (g(u))^{-p}$ $(p > 0)$, или $B(u) = |\ln g(u)|$, $u \in U \backslash \gamma$ и т. п.

Перейдем к описанию метода барьерных функций для решения задачи (1), предполагая, что подмножество $\gamma \subseteq U$ и некоторая его барьерная функция уже заданы. Введем функции

$$F_k(u) = J(u) + a_k B(u), \quad u \in U \backslash \gamma, \quad k = 1, 2, \ldots, \tag{2}$$

где $\{a_k\}$ — положительная последовательность, сходящаяся к нулю. Величины $\{a_k\}$ из (2) называются барьерными коэффициентами. Рассмотрим последовательность задач

$$F_k(u) \to \inf; \quad u \in U \backslash \gamma, \quad k = 1, 2, \ldots \tag{3}$$

Обозначим $F_{k*} = \inf\limits_{U \backslash \gamma} F_k(u)$ $(k = 1, 2, \ldots)$. Будем предполагать,

что в исходной задаче (1) $J_* = \inf\limits_U J(u) > -\infty$. Так как $F_k(u) \geqslant J(u)$ при всех $u \in U \setminus \gamma$, то $F_{k*} \geqslant J_* > -\infty$. Тогда условия

$$u_k \in U \setminus \gamma, \quad F_k(u_k) \leqslant F_{k*} + \varepsilon_k, \quad k = 1, 2, \ldots \tag{4}$$

определяют последовательность $\{u_k\}$, где $\varepsilon_k > 0$, $\lim\limits_{k\to\infty} \varepsilon_k = 0$; если окажется, что $F_k(u_k) = F_{k*}$, то в (4) допускается $\varepsilon_k = 0$.

Поскольку, как обычно, мы подразумеваем, что функция $J(u)$ конечна во всех точках $u \in U$, то согласно определению 1 для любой последовательности $\{v_r\} \in U \setminus \gamma$, $\{v_r\} \to v \in \gamma$ справедливо равенство $\lim\limits_{r\to\infty} F_k(v_r) = \infty$ при каждом фиксированном $k = 1, 2, \ldots$ Таким образом, функция $F_k(u)$ неограниченно возрастает вблизи γ. Поэтому следует ожидать, что при фиксированном k функция $F_k(u)$ вблизи γ не может принимать значения, близкие к F_{k*} и точка u_k, определяемая условиями (4), не будет расположена на слишком близком расстоянии от γ. В то же время благодаря тому, что барьерные коэффициенты $\{a_k\} \to 0$, не исключается возможность того, что с увеличением номера k точки u_k, постепенно «преодолевая барьер», будут приближаться к γ.

Для приближенного решения задачи (3) при фиксированном k и определения точки u_k, удовлетворяющей условиям (4), могут быть использованы различные методы минимизации. В частности, если $\gamma = \operatorname{Гр} U$ и $U \setminus \gamma = \operatorname{int} U \neq \varnothing$, то для решения задачи (3) может быть применен, например, градиентный метод (см. § 1):

$$u_{k,r+1} = u_{kr} - \alpha_r F'_k(u_{kr}), \quad u_{k0} = u_{k-1}, \quad r = 0, 1, \ldots$$

Поскольку $u_{kr} \in \operatorname{int} U$, то при достаточно малых $\alpha_r > 0$ и точка u_{kr+1} также будет принадлежать $\operatorname{int} U$, и мы избавлены от неудобств, связанных с учетом границы U,— нужно лишь на каждой итерации следить за соблюдением включения $u_k \in \operatorname{int} U$, а при его нарушении уменьшать длину шага α_r. Правда, для этого величину α_{r_1}, быть может, придется брать слишком малой, и сходимость градиентного метода, возможно, замедлится, но это уже будет «платой» за выполнение условия $u_k \in \operatorname{int} U$.

Дальнейшее изложение не зависит от того, с помощью какого конкретного метода минимизации будет найдена точка u_k, удовлетворяющая условиям (4). Поэтому мы здесь можем ограничиться предположением, что имеется достаточно удобный метод определения точки u_k из (4).

Метод барьерных функций описан. Отметим, что в литературе этот метод иногда называют *методом внутренних штрафов* (или методом внутренней точки), а метод штрафных функций из § 14 — *методом внешних штрафов* (или методом внешней

точки) [307]. Для иллюстрации метода барьерных функций приведем пример.

Пример 1. Пусть требуется решить задачу

$$J(u) = -u \to \inf; \quad u \in U = \{u \in E^1\colon\ g(u) = u \leqslant 0\}.$$

Очевидно, здесь $J_* = 0$, $U_* = \{0\}$. Границей множества U является $\gamma = \operatorname{Гр} U = \{u \in E^1\colon\ g(u) = u = 0\} = \{0\}$, а $U \setminus \gamma = \{u \in E^1\colon\ g(u) = u < 0\} = \operatorname{int} U$. В качестве барьерной функции для γ возьмем $B(u) = -1/u$ $(u < 0)$. Пусть $a_k = k^{-1}$ $(k = 1, 2, \ldots)$. Тогда функция (2) будет иметь вид $F_k(u) = -u - (ku)^{-1}$ $(u < 0)$. Нетрудно видеть, что здесь $F_{k*} = \inf\limits_{u<0} F_k(u) = 2/\sqrt{k}$ и точка $u_k = -1/\sqrt{k}$ удовлетворяет условиям (4) при $\varepsilon_k = 0$ $(k = 1, 2, \ldots)$. Ясно также, что $\lim\limits_{k\to\infty} F_{k*} = \lim\limits_{k\to\infty} J(u_k) = 0 = J_*$, $\lim\limits_{k\to\infty} u_k = 0 = u_*$.

В качестве барьерной функции здесь можно также взять и $B(u) = |\ln(-u)|$. В этом случае $F_k(u) = -u + |\ln(-u)|k^{-1}$ $(u < 0$, $k = 1, 2, \ldots)$, $F_{k*} = (1 + \ln k)k^{-1}$, а точка $u_k = -k^{-1}$ удовлетворяет условиям (4) при $\varepsilon_k = 0$. И здесь $\{J(u_k)\} \to J_* = 0$, $\{u_k\} \to u_* = 0$.

Перейдем к исследованию сходимости метода барьерных функций.

Теорема 1. *Пусть γ — некоторое подмножество из U, $U \setminus \gamma \neq \varnothing$, и*

$$J_* = J_{**}, \quad \text{где} \quad J_* = \inf_U J(u), \quad J_{**} = \inf_{U\setminus\gamma} J(u) > -\infty. \tag{5}$$

Пусть $B(u)$ — какая-либо барьерная функция подмножества γ, а последовательность $\{u_k\}$ определена условиями (4). Тогда

$$\lim_{k\to\infty} F_{k*} = \lim_{k\to\infty} F_k(u_k) = \lim_{k\to\infty} J(u_k) = J_*, \quad \lim_{k\to\infty} a_k B(u_k) = 0. \tag{6}$$

Кроме того, если множество U ограничено и замкнуто, а $J(u)$ полунепрерывна снизу на U, то $\{u_k\}$ сходится к U_*.

Доказательство. Из определения J_{**}, F_{k*}, неотрицательности барьерной функции и условий (4) следует

$$-\infty < J_{**} \leqslant J(u_k) \leqslant F_k(u_k) \leqslant F_{k*} + \varepsilon_k \leqslant F_k(u) + \varepsilon_k = \\ = J(u) + a_k B(u) + \varepsilon_k \tag{7}$$

при всех $u \in U \setminus \gamma$ $(k = 1, 2, \ldots)$. Так как $B(u)$ конечна в любой точке $u \in U \setminus \gamma$, $\{a_k\} \to 0$, то из (7) при $k \to \infty$ получим

$$J_{**} < \varliminf_{k\to\infty} F_{k*} \leqslant \varlimsup_{k\to\infty} F_{k*} \leqslant J(u), \quad u \in U \setminus \gamma.$$

Переходя в этих неравенствах к нижней грани по $u \in U \setminus \gamma$, будем иметь $J_{**} \leqslant \varliminf\limits_{k\to\infty} F_{k*} \leqslant \varlimsup\limits_{k\to\infty} F_{k*} \leqslant J_{**}$, т. е. $\lim\limits_{k\to\infty} F_{k*} = J_{**}$.

Отсюда и из (7) вытекает $\lim\limits_{k\to\infty} F_k(u_k) = \lim\limits_{k\to\infty} J(u_k) = J_{**}$. Так как $J_* = J_{**}$, то первые соотношения (6) доказаны. А тогда из $0 \leqslant a_k B(u_k) = F_k(u_k) - J(u_k) \to 0$ при $k \to \infty$ получим и второе из соотношений (6). Последнее утверждение о сходимости минимизирующей последовательности $\{u_k\}$ к U_* следует из теоремы 2.1.1.

Полезно заметить, что при доказательстве теоремы 1 были использованы не все свойства барьерных функций: соотношение $\lim\limits_{r\to\infty} B(v_r) = \infty$, где $\{v_r\} \in U\backslash\gamma$, $\{v_r\} \to v \in \gamma$, нам не понадобилось. Поэтому теорему 1 и ряд доказываемых ниже теорем можно использовать не только как теоремы о сходимости метода барьерных функций, но и как утверждения, выражающие собой достаточные условия устойчивости нижней грани относительно возмущений (погрешностей) минимизируемой функции и некоторых типов возмущений множества, на котором ищется минимум.

2. Рассмотрим возможности построения барьерных функций для задачи (1) в случае, когда

$$U = \{u \in U_0,\ g_i(u) \leqslant 0, \quad i = 1, \ldots, m\}; \tag{8}$$

здесь U_0 — заданное множество из E^n, функции $g_1(u), \ldots, g_m(u)$ определены и полунепрерывны снизу на U_0. Положим

$$\gamma = \{u \in U\colon\ g_i(u) = 0 \text{ хотя бы для одного } i, \quad 1 \leqslant i \leqslant m\}. \tag{9}$$

Будем предполагать, что множество

$$U(-0) = \{u \in U_0\colon\ g_i(u) < 0, \quad i = 1, \ldots, m\} \tag{10}$$

непусто. Тогда $U\backslash\gamma = U(-0) \neq \varnothing$. Довольно широкий класс барьерных функций для множества (9) дает следующая конструкция:

$$B(u) = \sum_{i=1}^{m} \varphi_i\big(-g_i(u)\big), \quad u \in U(-0), \tag{11}$$

где $\varphi_i(t)$ — неотрицательная функция переменной $t > 0$ такая, что $\lim\limits_{t\to 0} \varphi_i(t) = \infty$ при всех $i = 1, \ldots, m$. В самом деле, возьмем произвольную последовательность $\{v_r\} \in U\backslash\gamma$, сходящуюся к некоторой точке $v \in \gamma$. Согласно (9) тогда найдется номер j $(1 \leqslant j \leqslant m)$, для которого $g_j(v) = 0$. Так как $g_j(u)$ полунепрерывна снизу, а $g_j(v_r) < 0$ $(r = 1, 2, \ldots)$, то $0 = g_j(v) \leqslant \varliminf\limits_{r\to\infty} g_j(v_r) \leqslant \varlimsup\limits_{r\to\infty} g_j(v_r) \leqslant 0$, т. е. $\lim\limits_{r\to\infty} g_j(v_r) = 0$. Это значит, что $B(v_r) \geqslant \varphi_j(-g_j(v_r)) \to \infty$ при $r \to \infty$, так что функция (11) является барьерной для множества (9).

При необходимости в (11) функции $\varphi_i(t)$ нетрудно выбрать так, чтобы барьерная функция $B(u)$ обладала различными полезными свойствами, такими, как непрерывность, гладкость, вы-

пуклость, простота вычисления значения функции и нужных ее производных и т. п., если, конечно, исходные данные в задаче (1), (8) обладают такими свойствами. Например, взяв в (11) $\varphi_i(t)=1/t$ или $\varphi_i(t)=(\max\{-\ln t;\ 0\})^p$ $(p\geqslant 1)$, получим соответственно

$$B(u)=-\sum_{i=1}^{m}\frac{1}{g_i(u)},$$
$$B(u)=\sum_{i=1}^{m}(\max\{-\ln(-g_i(u));\ 0\})^p,\quad u\in U(-0). \tag{12}$$

Если U_0 выпукло, функции $g_i(u)$ $(i=1,\ \ldots,\ m)$, выпуклы на U_0, то множество $U(-0)=U\setminus\gamma$ выпукло и функции (12) также будут выпуклыми на $U(-0)$,— это следует из следствий к теореме 4.2.8. Далее, функции (12) будут обладать той же гладкостью, какою обладают функции $g_i(u)$ $(i=1,\ \ldots,\ m)$ — у второй функции (12) для этого нужно взять параметр p достаточно большим.

Может сложиться впечатление, что если функции $g_1(u),\ \ldots,\ g_m(u)$ непрерывны на U_0, то множество γ, определяемое условиями (9), будет состоять только лишь из граничных точек множества (8). Однако это не всегда так — множество γ может содержать и внутренние точки U.

Пример 2. Пусть $g(u)=|u|-1$ при $|u|\leqslant 1$, $g(u)=0$ при $1<|u|<2$, $g(u)=|u|-2$ при $|u|\geqslant 2$. Тогда множество $U=\{u\in E^1=U_0:\ g(u)\leqslant 0\}$ представляет собой отрезок $-2\leqslant u\leqslant 2$ на числовой оси, а $\operatorname{int}U=\{u\in E^1:\ -2<u<2\}$. В то же время множество $U(-0)=\{u\in E^1:\ g(u)<0\}=\{u:\ -1<u<1\}\subset\operatorname{int}U$, но $U(-0)\neq\operatorname{int}U$, а $\gamma=\{u\in U:\ g(u)=0\}=\{u:\ 1\leqslant|u|\leqslant 2\}$ наряду с граничными точками $u=2$ и $u=-2$ содержит и внутренние точки множества U.

Таким образом, для множества (8) не всегда выполняется равенство

$$\operatorname{Гр}U=\operatorname{Гр}U_0\cup\gamma, \tag{13}$$

где γ определяется условиями (9), а функции (11), (12), являющиеся барьерными функциями для подмножества γ, могут и не быть таковыми хотя бы для части границы U.

3. Отдельно остановимся на условии (5), которое было существенно использовано в теореме 1 при доказательстве сходимости метода барьерных функций.

Нетрудно привести примеры задач (1), (8), в которых функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ непрерывны, множество U замкнуто и ограничено, но условие (5) не имеет места. Например, если $J(u)=u$, а множество U взято из примера 2, то $J_{**}=-1>J_*=-2$.

Однако даже выполнение условия (13), при котором функции (11), (12) будут барьерными функциями γ — части границы U, еще не гарантирует справедливость равенства (5).

Пример 3. Пусть $J(u) = e^{-x}$, $U = \{u = (x, y) \in E^2 = U_0$: $g(u) = (x^2 + y^2 - 1)(y - 1)^2 \leqslant 0\}$. Тогда $\gamma = \Gamma p\, U = \{u \in U$: $g(u) = 0\} = \{u \in E^2$: $x^2 + y^2 = 1$ или $y = 1\}$, $U_0 = E^2$, $\Gamma p\, U_0 = \varnothing$, так что условие (13) выполнено. Далее, здесь $U \setminus \gamma = U(-0) = \{u \in E^2$: $g(u) < 0\} = \{u$: $x^2 + y^2 < 1\} = \operatorname{int} U$, поэтому $J_{**} = \inf\limits_{U \setminus \gamma} J(u) = e^{-1} > 0$. В то же время $J_* = \lim\limits_{k \to \infty} J(u_k) = 0$, где $u_k = (k, 1) \in U$ $(k = 1, 2, \ldots)$.

Заметим, что в этом примере $\overline{U(-0)} = \{u$: $x^2 + y^2 \leqslant 1\} \subset U$, но $\overline{U(-0)} \neq U$. (Напоминаем, что через $\overline{Z}$ мы условились обозначать замыкание множества Z.)

Приведем две теоремы, дающие достаточные условия для выполнения равенства (5). Имея в виду дальнейшие применения, утверждения сформулируем для множества

$$U(C) = \{u \in E^n\colon u \in U_0,\ g_i(u) \leqslant C,\quad i = 1, \ldots, m\}, \tag{14}$$

где C — некоторая постоянная. Обозначим

$$\begin{gathered} J_*(C) = \inf_{U(C)} J(u), \qquad J_*(C - 0) = \lim_{\varepsilon \to +0} J_*(C - \varepsilon), \\ J_*(C + 0) = \lim_{\varepsilon \to +0} J_*(C + \varepsilon). \end{gathered} \tag{15}$$

Теорема 2. *Пусть для некоторых C, $\varepsilon_0 > 0$ множество $U(C - \varepsilon_0)$ непусто и*

$$\overline{U(C - 0)} = U(C), \tag{16}$$

где

$$U(C - 0) = \{u \in U_0\colon g_i(u) < C,\quad i = 1, \ldots, m\}.$$

Пусть, кроме того, функция $J(u)$ полунепрерывна сверху на множестве $U(C)$. Тогда

$$J_*(C - 0) = J_*(C) = \inf_{U(C-0)} J(u). \tag{17}$$

Доказательство. Прежде всего, заметим, что

$$U(c - \varepsilon) \subseteq U(C - \delta) \subseteq U(C - 0) \subseteq U(C)$$

при всех $0 < \delta < \varepsilon \leqslant \varepsilon_0$, поэтому

$$J_*(C) \leqslant \inf_{U(C-0)} J(u) \leqslant J_*(C - \delta) \leqslant J_*(C - \varepsilon).$$

Это значит, что функция $J_*(C)$ переменной C не возрастает и существует предел

$$\lim_{\varepsilon \to +0} J_*(C - \varepsilon) = J_*(C - 0) \geqslant \inf_{U(C-0)} J(u) \geqslant J_*(C). \tag{18}$$

Возьмем произвольную точку $u \in U(C)$. В силу условия (16) найдется последовательность $\{u_k\} \in U(C - 0)$, сходящаяся к точке u. Это значит, что $u_k \in U_0$, $g_i(u_k) \leqslant C - \varepsilon_{ik} < C$ $(\varepsilon_{ik} > 0,\ k = 1, 2, \ldots)$, где $\lim\limits_{k \to \infty} \varepsilon_{ik} = 0$ $(i = 1, \ldots, m)$. Таким образом, $u_k \in U(C - \varepsilon_k)$, где $\varepsilon_k = \min\limits_{1 \leqslant i \leqslant m} \varepsilon_{ik} > 0$, $\{\varepsilon_k\} \to 0$, и $J_*(C - \varepsilon_k) \leqslant J(u_k)$ $(k = 1, 2, \ldots)$. Отсюда при $k \to \infty$, учиты-

вая полунепрерывность сверху функции $J(u)$, получим $\lim\limits_{\varepsilon\to+0} J_*(C-\varepsilon) = \lim\limits_{k\to\infty} J_*(C-\varepsilon_k) \leqslant J(u)$. В силу произвольности $u \in U(C)$ тогда $J_*(C-0) \leqslant J_*(C)$. Сравнивая это неравенство с (18), приходим к равенству (17). Теорема 2 доказана.

Таким образом, если условия теоремы 2 выполнены при $C = 0$, то методом барьерных функций (2), (4), (8)—(11) для задачи (1), (8) можно получить последовательность $\{u_k\}$, обладающую свойствами (6).

Аналогичное утверждение справедливо для выпуклых задач (1), (8).

Теорема 3. *Пусть U_0 — выпуклое множество из E^n, функции $J(u)$, $g_1(u), \ldots, g_m(u)$ выпуклы на U_0. Тогда равенства* (17) *справедливы при всех $C > C_* = \max\limits_{1\leqslant i\leqslant m} \inf\limits_{U_0} g_i(u)$.*

Доказательство. Как было установлено в теореме 2, функция $J_*(C)$ переменной C не возрастает. Возьмем произвольные $C, \varepsilon > 0$, $C > C - \varepsilon > C_*$. Пусть $u \in U(C)$, $v \in U(C-\varepsilon)$. В силу выпуклости U_0 тогда $u_\alpha = \alpha v + (1-\alpha)u \in U_0$ при всех α $(0 \leqslant \alpha \leqslant 1)$. Кроме того, из выпуклости $g_i(u)$ имеем $g_i(u_\alpha) \leqslant \alpha g_i(v) + (1-\alpha) g_i(u) \leqslant \alpha(C-\varepsilon) + (1-\alpha)C = C - \alpha\varepsilon$ $(0 < \alpha \leqslant 1)$. Это значит, что $u_\alpha \in U(C-\alpha\varepsilon)$. Тогда с учетом выпуклости функции $J(u)$ получим $J_*(C) \leqslant \inf\limits_{U(C-0)} J(u) \leqslant J_*(C-\alpha\varepsilon) \leqslant J(u_\alpha) \leqslant \alpha J(v) + (1-\alpha) J(u)$ для всех $u \in U(C)$, $v \in U(C-\varepsilon)$. Следовательно, $J_*(C) \leqslant \inf\limits_{U(C-0)} J(u) \leqslant J_*(C-\alpha\varepsilon) \leqslant \alpha J_*(C-\varepsilon) + (1-\alpha) J_*(C) \leqslant J_*(C) + \alpha[J_*(C-\varepsilon_0) - J_*(C)]$ для всех α $(0 < \alpha \leqslant 1,\ 0 < \varepsilon \leqslant \varepsilon_0 < C - C_*)$. Отсюда при $\alpha \to +0$ с учетом монотонности $J_*(C)$ получаем равенства (17), что и требовалось.

4. Пусть множество U задается условиями

$$U = \{u \in E^n\colon\ u \in U_0,\ g_i(u) \leqslant 0,\ i = 1, \ldots, m;\ g_i(u) = 0,\ i = m+1, \ldots, s\}. \tag{19}$$

Если это множество не имеет внутренних точек, то реализация ряда методов минимизации (например, методов из § 3—5, 11 и др.) на U может стать затруднительной или даже невозможной. В то же время при применении методов § 13, 14 к задаче (1), (19) могут получиться такие последовательности $\{u_k\}$, которые не принадлежат множеству U и нарушают какие-либо из ограничений $g_i(u) \leqslant 0$, $g_j(u) = 0$ на недопустимо большую величину. В таких случаях может оказаться целесообразным использование метода барьерных функций.

Заметим, что этот метод выше изложен для задачи (1), (8) в предположении, что множество $U(-0)$, определяемое условиями (10), непусто. Однако такое предположение для множества (19) при $m < s$ не имеет смысла. Поэтому описанный выше метод барьерных функций к задаче (1), (19) непосредственно неприменим и требует модификации, обобщения. Опишем один из возможных здесь подходов [178].

Введем последовательность расширенных множеств

$$V_k = \{u \in U_0\colon\ g_i(u) \leqslant \theta_k,\ i = 1, \ldots, m;\ |g_i(u)| \leqslant \theta_k,\ i = m+1, \ldots, s\}, \tag{20}$$

где $\theta_k > 0$ $(k = 1, 2, \ldots)$, $\lim\limits_{k\to\infty} \theta_k = 0$. Так как $U \subset V_k$ $(k = 1, 2, \ldots)$, то из $U \neq \varnothing$ следует $V_k \neq \varnothing$ $(k = 1, 2, \ldots)$. Предполагая, что функция $J(u)$ определена на множестве $\bigcup\limits_{k=1}^{\infty} V_k$, рассмотрим последовательность задач

$$J(u) \to \inf;\quad u \in V_k,\quad k = 1, 2, \ldots \tag{21}$$

Для решения задач (21) могут быть использованы различные методы минимизации. Мы здесь остановимся лишь на методе барьерных функций. Обозначим

$\gamma_k = \{u \in V_k$ и выполняется хотя бы одно из равенств

$$g_i(u) = \theta_k,\ i = 1, \ldots, m;\ g_j(u) = \theta_k,\ g_j(u) = -\theta_k,\ j = m+1, \ldots, s\}. \quad (22)$$

Поскольку $U \subset V_k$, $U \cap \gamma_k = \varnothing$, то $U \subset V_k \setminus \gamma_k \neq \varnothing$ $(k = 1, 2, \ldots)$. В качестве барьерной функции $B_k(u)$ подмножества γ_k возьмем

$$B_k(u) = \sum_{i=1}^{s} \varphi_i(\theta_k - g_i(u)) + \sum_{i=m+1}^{s} \varphi_i(\theta_k + g_i(u)), \quad u \in V_k \setminus \gamma_k, \quad (23)$$

где функция $\varphi_i(t)$ определена, конечна, неотрицательна и не возрастает при $t > 0$, $\lim\limits_{t \to +0} \varphi_i(t) = \infty$ $(i = 1, \ldots, s)$. Например, в качестве $\varphi_i(t)$ можно взять $\varphi_i(t) = t^{-1}$, $\varphi(t) = (\max\{-\ln t; 0\})^p$ $(p \geqslant 1)$.

Далее, составим функцию

$$F_k(u) = J(u) + a_k B_k(u), \quad u \in V_k \setminus \gamma_k, \quad (24)$$

где $\{a_k\}$ — барьерные коэффициенты: $a_k > 0$ $(k = 1, 2, \ldots)$, $\{a_k\} \to 0$. В отличие от рассмотренного выше варианта метода барьерных функций, здесь мы будем требовать, чтобы барьерные коэффициенты $\{a_k\}$ и параметры $\{\theta_k\}$ стремились к нулю согласованно в следующем смысле:

$$\lim_{k \to \infty} a_k \varphi_i(\theta_k) = 0, \quad i = 1, \ldots, s. \quad (25)$$

Предположим, что $J_{k*} = \inf\limits_{V_k} J(u) > -\infty$ $(k = 1, 2, \ldots)$. Так как $B_k(u) \geqslant 0$, $a_k > 0$, то $F_k(u) \geqslant J(u)$ при всех $u \in V_k \setminus \gamma_k$, и поэтому $F_{k*} = \inf\limits_{V_k \setminus \gamma_k} F_k(u) \geqslant J_{k*} > -\infty$ $(k = 1, 2, \ldots)$. С помощью какого-либо метода минимизации определим точку u_k, удовлетворяющую условиям

$$u_k \in V_k \setminus \gamma_k, \quad F_{k*} \leqslant F_k(u_k) \leqslant F_{k*} + \varepsilon_k, \quad k = 1, 2, \ldots, \quad (26)$$

где $\{\varepsilon_k\}$ — некоторая положительная последовательность, сходящаяся к нулю; если $F_k(u_k) = F_{k*}$, то в (26) допускается $\varepsilon_k = 0$. Метод барьерных функций для задачи (1), (19) описан.

Теорема 4. *Пусть функции $F_k(u)$, $B_k(u)$, множества V_k, γ_k определены соотношениями* (20), (22)—(24), *выполняются равенства* (25) *и, кроме того,*

$$\lim_{k \to \infty} J_{k*} = J_* > -\infty, \quad J_{k*} = \inf_{V_k} J(u), \quad J_* = \inf_U J(u). \quad (27)$$

Тогда для последовательности $\{u_k\}$, определяемой условиями (26), *справедливы соотношения*

$$\lim_{k \to \infty} F_{k*} = \lim_{k \to \infty} F_k(u_k) = \lim_{k \to \infty} J(u_k) = J_*, \quad \lim_{k \to \infty} a_k B_k(u_k) = 0. \quad (28)$$

Если, кроме того, множество

$$U(\delta) = \{u \in U_0:\ g_i(u) \leqslant \delta,\ i = 1, \ldots, m;\ |g_i(u)| \leqslant \delta,\ i = m+1, \ldots, s\} \quad (29)$$

компактно при некотором $\delta > 0$, множество U_0 замкнуто, а функции $g_1(u), \ldots, g_m(u)$, $|g_{m+1}(u)|, \ldots, |g_s(u)|$ полунепрерывны снизу на $U(\delta)$, то $\{u_k\} \to U_*$ — множество решений задачи (1), (19).

Доказательство. Из определения J_{k*}, F_{k*}, неотрицательности $B_k(u)$ и условий (26) имеем

$$-\infty < J_{k*} \leqslant J(u_k) \leqslant F_{k*} + \varepsilon_k \leqslant$$
$$\leqslant F_k(u) + \varepsilon_k = J(u) + a_k B_k(u) + \varepsilon_k, \quad u \in V_k \setminus \gamma_k, \quad k = 1, 2, \ldots \tag{30}$$

Так как функции $\varphi_i(t)$ из (23) не возрастают при $t > 0$, то $\varphi_i(\theta_k - g_i(u)) \leqslant \leqslant \varphi_i(\theta_k)$ $(i = 1, \ldots, m)$; $\varphi_i(\theta_k \pm g_i(u)) = \varphi_i(\theta_k)$ $(i = m+1, \ldots, s)$ для всех $u \in U$. Поэтому в силу условия (25)

$$0 \leqslant a_k B_k(u) \leqslant 2a_k \sum_{i=1}^{s} \varphi_i(\theta_k) \to 0, \quad k \to \infty, \quad u \in U. \tag{31}$$

Тогда при $k \to \infty$ из (30) с учетом условия (27) получим

$$J_* \leqslant \varliminf_{k\to\infty} F_{k*} \leqslant \varlimsup_{k\to\infty} F_{k*} \leqslant J(u) \quad \text{при всех} \quad u \in U.$$

Переходя к нижней грани по $u \in U$, отсюда имеем $\lim\limits_{k\to\infty} F_{k*} = J_*$. Тогда из (30) следует $\lim\limits_{k\to\infty} F_k(u_k) = \lim\limits_{k\to\infty} J(u_k) = J_*$. Наконец, $0 \leqslant a_k B_k(u_k) = = F_k(u_k) - J(u_k) \to 0$ при $k \to \infty$. Равенства (28) доказаны.

Пусть теперь выполнены все условия теоремы. Так как $\{\theta_k\} \to 0$, то $V_k \subset U(\delta)$ при всех $k \geqslant k_0$. А тогда $u_k \in U(\delta)$, $k \geqslant k_0$. В силу компактности $U(\delta)$ последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$, пусть подпоследовательность $\{u_{k_r}\} \to u_*$. В силу замкнутости U_0 тогда $u_* \in U_0$. Далее, из полунепрерывности снизу функций $g_1(u), \ldots, g_m(u), |g_{m+1}(u)|, \ldots, |g_s(u)|$ и условия $u_k \in V_k$ следует, что

$$g_i(u_*) \leqslant \varliminf_{r\to\infty} g_i(u_{k_r}) \leqslant \lim_{k\to\infty} \theta_k = 0, \qquad i = 1, \ldots, m,$$

$$|g_i(u_*)| \leqslant \varliminf_{r\to\infty} |g_i(u_{k_r})| \leqslant \lim_{k\to\infty} \theta_k = 0$$

или

$$g_i(u_*) = 0, \qquad i = m+1, \ldots, s.$$

Таким образом, $u_* \in U$. Отсюда с учетом полунепрерывности снизу $J(u)$ на $U(\delta)$ получим $J_* \leqslant J(u_*) \leqslant \varliminf\limits_{r\to\infty} J(u_{k_r}) = \lim\limits_{k\to\infty} J(u_k) = J_*$, т. е. $J(u_*) = = J_*$ или $u_* \in U_*$. Тем самым показано, что любая предельная точка последовательности $\{u_k\}$ принадлежит U_*. Отсюда следует, что $\{u_k\} \to U_*$. Теорема 4 доказана.

При некоторых более жестких ограничениях на данные задачи (1), (19) можно получить оценки погрешности метода (20)—(26).

Теорема 5. *Пусть для задачи* (1), (19) *справедливо неравенство*

$$-\infty < J_* \leqslant J(u) + \sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^{\nu} \quad \forall u \in U_0, \quad c_i \geqslant 0, \quad \nu > 0 \tag{32}$$

(*см. определение* 14.3 *и леммы* 14.1, 14.5). *Тогда последовательность* $\{u_k\}$,

определяемая условиями (20)—(26), *существует и справедливы оценки*

$$-|c|_1\theta_k^\nu \leqslant J(u_k) - J_* \leqslant F_k(u_k) - J_* \leqslant 2a_k \sum_{i=1}^{s} \varphi_i(\theta_k) + \varepsilon_k, \qquad (33)$$

$$0 \leqslant a_k B_k \leqslant 2a_k \sum_{i=1}^{s} \varphi_i(\theta_k) + \varepsilon_k + |c|_1\theta_k^\nu, \qquad k = 1, 2, \ldots, \qquad (34)$$

где $|c|_1 = \sum_{i=1}^{s} |c_i|$. *Если, кроме того, множество* (29) *компактно при некотором* $\delta > 0$, U_0 *замкнуто, а функции* $J(u)$, $g_1(u), \ldots, g_m(u)$, $|g_{m+1}(u)|, \ldots, |g_s(u)|$ *полунепрерывны снизу на* $U(\delta)$, *то* $\{u_k\} \to U_*$.

Доказательство. Из определения (20) множества V_k и условия (32) следует

$$-\infty < J_* \leqslant J(u) + \sum_{i=1}^{s} c_i \left(g_i^+(u)\right)^\nu \leqslant J(u) + |c|_1\theta_k^\nu \leqslant F_k(u) + |c|_1\theta_k^\nu \qquad (35)$$

при всех $u \in V_k \setminus \gamma_k$. Отсюда имеем $F_k(u) \geqslant J_* - |c|_1\theta_k^\nu > -\infty$, $u \in V_k \setminus \gamma_k$, или $F_{k*} \geqslant J_* - |c|_1\theta_k^\nu > -\infty$ $(k = 1, 2, \ldots)$. Таким образом, последовательность $\{u_k\}$, удовлетворяющая условиям (26), существует. Далее из (31) следует

$$0 \leqslant a_k B_k(u_*) \leqslant 2a_k \sum_{i=1}^{s} \varphi_i(\theta_k), \qquad u_* \in U_* \subset U \subset U_k \setminus \gamma_k, \qquad k = 1, 2, \ldots$$

Поэтому с учетом неравенств (26), (35) имеем

$$J_* \leqslant J(u_k) + |c|_1\theta_k^\nu \leqslant F_k(u_k) + |c|_1\theta_k^\nu \leqslant F_{k*} + \varepsilon_k + |c|_1\theta_k^\nu \leqslant$$

$$\leqslant F_k(u_*) + \varepsilon_k + |c|_1\theta_k^\nu \leqslant J_* + 2a_k \sum_{i=1}^{s} \varphi_i(\theta_k) + \varepsilon_k + |c|_1\theta_k^\nu, \qquad k = 1, 2, \ldots$$

Отсюда получаем оценку (33).

Далее, из соотношений $0 \leqslant a_k B_k(u_k) = (F_k(u_k) - J_*) - (J(u_k) - J_*)$ и уже доказанной оценки (33) вытекает оценка (34). Последнее утверждение доказывается так же, как аналогичное утверждение теоремы 4.

5. Отдельно остановимся на условии (27), которое существенно использовалось при доказательстве равенств (28). Нетрудно привести примеры задач (1), (19), когда это условие не выполняется.

Пример 4. Пусть $J(u) = e^{-u}$, $U = \{u \in E^1 = U_0\colon g(u) = (u^2 - 1) \times (1 + u^4)^{-1} \leqslant 0\}$. Ясно, что $U = \{u \in E^1\colon |u| \leqslant 1\}$, $J_* = \inf_U J(u) = e^{-1}$. Возьмем $V_k = \{u \in E^1\colon g(u) \leqslant \theta_k = 1/k^2\}$. Так как $u_r = r \in V_k$ при $r \geqslant k$, то $\lim_{r\to\infty} J(u_r) = 0 = J_{k*}$ $(k = 1, 2, \ldots)$. Таким образом, здесь $\lim_{k\to\infty} J_{k*} = 0 < e^{-1} = J_*$ — условие (27) не выполняется. Заметим, что в рассмотренном примере множество $U(\delta) = \{u \in E^1\colon g(u) \leqslant \delta\}$ не является компактным ни при каком $\delta > 0$.

Приведем две теоремы, дающие достаточные условия для выполнения условия (27).

Теорема 6. *Пусть множество* U_0 *замкнуто, функции* $J(u)$, $g_1(u), \ldots, g_m(u)$, $|g_{m+1}(u)|, \ldots, |g_s(u)|$ *определены и полунепрерывны снизу на* U_0.

Кроме того, пусть множество

$$U(C) = \{u \in E^n\colon\ u \in U_0,\ g_i^+(u) \leqslant C,\ i = 1, \ldots, s\}$$

непусто, а множество $U(C + \varepsilon_0)$ *ограничено и замкнуто при некотором* $\varepsilon_0 > 0$. *Тогда* (*см. обозначения* (15))

$$\lim_{\varepsilon \to +0} J_*(C + \varepsilon) = J_*(C + 0) = J_*(C). \tag{36}$$

Доказательство. Так как $U(C) \subseteq U(C + \delta) \subseteq U(C + \varepsilon)$ при любых $0 < \delta < \varepsilon \leqslant \varepsilon_0$, то $J_*(C + \varepsilon) \leqslant J_*(C + \delta) \leqslant J_*(C)$. Таким образом, функция $J_*(C)$ переменной C не возрастает и существует предел $\lim\limits_{\varepsilon \to +0} J_*(C + \varepsilon) = J_*(C + 0) \leqslant J_*(C)$. Возьмем произвольную последовательность $\{\varepsilon_k\}$, $0 < \varepsilon_k \leqslant \varepsilon_0$, сходящуюся к нулю. При сделанных предположениях множества $U(C + \varepsilon_k) \subseteq U(C + \varepsilon_0)$ при каждом $k = 1, 2, \ldots$ ограничены и замкнуты. Согласно теореме 2.1.1 тогда существует точка $w_k \in U(C + \varepsilon_k)$ такая, что $J(w_k) = J_*(C + \varepsilon_k)$ $(k = 1, 2, \ldots)$. Поскольку $U(C + \varepsilon_0)$ — компактное множество и $w_k \in U(C + \varepsilon_k) \subseteq U(C + \varepsilon_0)$, то последовательность $\{w_k\}$ имеет хотя бы одну предельную точку. Пусть w_* — какая-либо предельная точка $\{w_k\}$. Не умаляя общности, можем считать, что сама последовательность $\{w_k\} \to w_*$. По построению $w_k \in U(C + \varepsilon_k)$, т. е. $w_k \in U_0$, $g_i^+(w_k) \leqslant C + \varepsilon_k$ $(i = 1, \ldots, s)$. Используя замкнутость множества U_0, полунепрерывность рассматриваемых функций, отсюда при $k \to \infty$ получаем $w_* \in U(C)$. А тогда $J_*(C) \leqslant J(w_*) \leqslant \varliminf\limits_{k \to \infty} J(w_k) = \lim\limits_{k \to \infty} J_*(C + \varepsilon_k) = J_*(C + 0)$. Сравнивая с ранее установленным неравенством $J_*(C + 0) \leqslant J_*(C)$, получаем равенство (36).

Нетрудно видеть, что при $C = 0$ из (36) вытекает условие (27).

6. Метод барьерных функций на практике иногда используют в сочетании с методом штрафных функций. Предположим, что множество U задается в виде

$$U = \{u \in U_0\colon\ g_i(u) \leqslant 0,\ i = 1, \ldots, m;\ g_i(u) = 0,\ i = m + 1, \ldots, s;$$
$$h_j(u) \leqslant 0,\ j = 1, \ldots, l;\ h_j(u) = 0,\ j = l + 1, \ldots, r\},$$

где U_0 — заданное множество из E^n, функции $g_i(u)$, $h_j(u)$, а также минимизирующая функция $J(u)$ определены на U_0. Ограничения, задаваемые функциями $g_i(u)$, будем учитывать с помощью штрафных функций

$$P(u) = \sum_{i=1}^{m} (\max\{g_i(u); 0\})^p + \sum_{i=m+1}^{s} |g_i(u)|^p, \quad u \in U_0, \quad p \geqslant 1.$$

Введем множества

$$W_k = \{u \in U_0\colon\ h_j(u) \leqslant \theta_k,\ j = 1, \ldots, l;\ |h_j(u)| \leqslant \theta_k,\ j = l + 1, \ldots, r\},$$

$\gamma_k = \{u \in W_k$ и выполняется хотя бы одно из равенств

$$h_j(u) = \theta_k,\ j = 1, \ldots, r; \quad h_j(u) = -\theta_k,\ j = l + 1, \ldots, r\},$$

$k = 1, 2, \ldots$ В качестве барьерной функции подмножества γ_k возьмем

$$B_k(u) = \sum_{j=1}^{r} \varphi_j(\theta_k - h_j(u)) + \sum_{j=l+1}^{r} \varphi_j(\theta_k + h_j(u)), \qquad u \in W_k \setminus \gamma_k,$$

где функция $\varphi_j(t)$ определена, неотрицательна и не возрастает при $t > 0$, $\lim\limits_{t \to +0} \varphi_j(t) = \infty$ $(j = 1, \ldots, r)$. Рассмотрим последовательность задач

$$F_k(u) = J(u) + A_k P(u) + a_k B_k(u) \to \inf; \quad u \in W_k \setminus \gamma_k,$$

считая, что $\{A_k^{-1}\}$, $\{a_k\}$, $\{\theta_k\}$ — положительные последовательности, сходящиеся к нулю. Пусть $F_{k*} = \inf\limits_{W_k \setminus \gamma_k} F_k(u) > -\infty$ $(k = 1, 2, \ldots)$. Определим точку u_k из условий

$$u_k \in W_k \setminus \gamma_k, \quad F_k(u_k) \leqslant F_{k*} + \varepsilon_k, \tag{37}$$

где $\varepsilon_k > 0$ $(k = 1, 2, \ldots)$, $\lim\limits_{k\to\infty} \varepsilon_k = 0$ (если $F_k(u_k) = F_{k*}$, то в (37) допускается $\varepsilon_k = 0$). Предлагаем самостоятельно сформулировать и доказать для метода (37) теоремы, объединяющие теоремы 4 и 14.2, 5 и 14.5, 14.6.

Различные аспекты метода барьерных функций исследованы в [8, 111, 159, 178, 307, 330, 338].

Упражнения. 1. Применить метод барьерных функций к задачам:

а) $J(u) = x + y \to \inf$; $u \in U = \{u = (x, y) \in E^2: g_1(u) = x^2 - y \leqslant 0, g_2(u) = -x \leqslant 0\}$;

б) $J(u) = y \to \inf$; $u \in U = \{u = (x, y) \in E^2: g(u) = \sin x + x - y \leqslant 0\}$;

в) $J(u) = (u-1)^3 \to \inf$; $u \in U = \{u \in E^1: g(u) = -u - 1 \leqslant 0\}$;

г) к задачам из упражнения 14.1;

д) к задачам из примеров 2.2.2 и 2.2.4, считая, что множество γ совпадает с границей множества U.

2. Сформулировать и доказать аналог теоремы 14.7 для описанных выше вариантов метода барьерных функций.

§ 16. Метод нагруженных функций

1. Будем рассматривать задачу

$$J(u) \to \inf; \quad U = \{u \in E^n: u \in U_0,$$
$$g_i(u) \leqslant 0,\ i = 1, \ldots, m;\ g_i(u) = 0,\ i = m+1, \ldots, s\}. \tag{1}$$

В методе нагруженных функций исходная задача (1) сводится к задачам минимизации некоторых вспомогательных функций на множестве U_0 и к поиску минимального решения (корня) некоторого уравнения. Но, в отличие от метода штрафных функций, в нем нет неограниченно возрастающих коэффициентов, аналогичных штрафным коэффициентам, и кроме того, метод нагруженных функций применим к более широкому классу задач, чем метод модифицированных функций Лагранжа. Введем семейство функций

$$\Phi(u, t) = L|\max\{J(u) - t;\ 0\}|^{p_0} + MP(u), \quad u \in U_0, \tag{2}$$

зависящее от скалярного параметра t, $-\infty < t < \infty$, где

$$P(u) = \sum_{i=1}^{m} (\max\{g_i(u); 0\})^{p_i} + \sum_{i=m+1}^{s} |g_i(u)|^{p_i}, \tag{3}$$

величины $p_i \geqslant 1$, $i = 1, \ldots, s$, $L > 0$, $M > 0$ фиксированы и являются параметрами метода. Положим

$$\rho(t) = \inf_{u \in U_0} \Phi(u, t). \tag{4}$$

Поскольку $\Phi(u, t) \geqslant 0$ при всех t и $u \in U_0$, то $\rho(t) \geqslant 0$ при любом t. Предположим, что в задаче (1) $J_* > -\infty$, $U_* \neq \varnothing$. Возьмем произвольную точку $u_* \in U_*$. Тогда $P(u_*) = 0$ и $\Phi(u_*, J_*) = 0$. Следовательно, $\rho(J_*) = 0$, т. е. J_* является корнем уравнения

$$\rho(t) = 0. \tag{5}$$

С другой стороны, $\Phi(u, t) > 0$ при всех $u \in U_0$ и всех $t < J_*$, и поэтому можно ожидать, что для широкого класса задач будет выполняться неравенство $\rho(t) > 0$ при всех $t < J_*$. Если это в самом деле так, то задача поиска J_* сведется к поиску минимального корня уравнения (5). Такое сведение задачи минимизации привлекательно тем, что для поиска минимального корня уравнения (5) с одной неизвестной могут быть использованы такие широко известные методы решения уравнений, как методы деления отрезка пополам, простой итерации и т. п. [4, 39, 54]. Основная идея метода нагруженных функций описана.

Заметим, что в этом методе могут быть использованы и другие конструкции функции $\Phi(u, t)$, отличные от (2). Например, можно принять

$$\Phi(u, t) = L|J(u) - t|^{p_0} + MP(u), \quad u \in U_0, \quad -\infty < t < \infty, \tag{6}$$

где функция $P(u)$ взята из (3), $L > 0$, $M > 0$, $p_i \geqslant 1$. Повторив предыдущие рассуждения для функции $\rho(t)$, определяемой из условий (4), (6), можно показать, что здесь также $\rho(J_*) = 0$, и высказать гипотезу о том, что для широкого класса задач (1) число J_*, по-видимому, будет минимальным корнем уравнения (5)

2. Прежде чем переходить к формулировке условий, при которых высказанная гипотеза в самом деле будет справедлива, рассмотрим примеры. Во всех примерах ограничимся рассмотрением функций $\Phi(u, t)$ из (2) и (6) при $L = M = p_0 = \ldots = p_s = 1$.

Пример 1. Пусть $J(u) = -u$; $U = \{u \in E^1\colon g(u) = u \leqslant 0\}$. Здесь $J_* = 0$, $u_* = 0$. Функция (2) здесь имеет вид

$$\Phi(u, t) = \max\{-u - t;\ 0\} + \max\{u;\ 0\}, \quad u \in U_0 = E^1.$$

Если $t \geqslant 0$, то $\Phi(0, t) = 0 = \inf\limits_{E^1} \Phi(u, t) = \rho(t)$. Если же $t < 0$, то $\Phi(u, t) = u$ при $u \geqslant -t$, $\Phi(u, t) = -t$ при $0 \leqslant u \leqslant -t$; $\Phi(u, t) = -u - t$ при $u \leqslant 0$ (нарисуйте график функции $\Phi(u, t)$ при различных t). Поэтому $\inf\limits_{E^1} \Phi(u, t) = \rho(t) = -t$ при $t < 0$. Таким образом, $\rho(t) = \max\{-t;\ 0\}$. Очевидно, минимальный корень уравнения (5) здесь совпадает с $J_* = 0$.

Функция (6) будет иметь вид

$$\Phi(u, t) = |-u - t| + \max\{u;\ 0\}, \quad u \in E^1.$$

Если $t \geqslant 0$, то взяв $u = -t$, получим $\Phi(-t;\ t) = 0 = \rho(t)$. Если же $t < 0$, то $\Phi(u,\ t) = 2u + t$ при $u \geqslant -t$; $\Phi(u,\ t) = -t$ при $0 \leqslant u \leqslant -t$; $\Phi(u,\ t) = -u - t$ при $u \leqslant 0$, и следовательно, $\rho(t) = -t$ при $t < 0$. В рассматриваемой задаче функции $\rho(t)$, построенные на основе функций (2) и (6), совпали.

Пример 2. Пусть $J(u) = u$, $U = \{u \in E^1:\ g(u) = u^2 - 1 \leqslant 0\}$. Ясно, что здесь $U = \{u \in E^1:\ -1 \leqslant u \leqslant 1\}$, $J_* = -1$, $u_* = -1$. Если согласно (2) принять

$$\Phi(u,\ t) = \max\{u - t;\ 0\} + \max\{u^2 - 1;\ 0\}, \quad u \in U_0 = E^1,$$

то нетрудно показать, что $\rho(t) = \inf\limits_{E^1} \Phi(u, t) = \max\{-t - 1;\ 0\}$.

Если же за основу взять функцию (6), то

$$\Phi(u, t) = |u - t| + \max\{u^2 - 1;\ 0\}, \quad u \in E^1,$$

$$\rho(t) = \inf_{E^1} \Phi(u, t) = \max\{|t| - 1;\ 0\}.$$

В рассматриваемой задаче функции $\rho(t)$, построенные с помощью функций (2) и (6), оказались разными, но минимальный корень уравнения (5) в обоих случаях совпадает с $J_* = -1$.

Пример 3. Пусть $J(u) = u$, $U = \{u \in E^1:\ g(u) = u^2 = 0\}$. Тогда $U = \{0\}$, $J_* = 0$, $u_* = 0$. Для функции (2) $\Phi(u,\ t) = \max\{u - t;\ 0\} + u^2$, $u \in U_0 = E^1$ получим

$$\rho(t) = \begin{cases} 0, & t \geqslant 0, \\ t^2, & -1/2 \leqslant t < 0, \\ -t - 1/4, & t < -1/2. \end{cases}$$

Если взять функцию (6), то $\Phi(u,\ t) = |u - t| + u^2$, $u \in E^1$, и

$$\rho(t) = \begin{cases} t^2, & |t| \leqslant 1/2, \\ |t| - 1/4, & |t| > 1/2. \end{cases}$$

Здесь также минимальный корень уравнения (5) совпадает с $J_* = 0$.

Однако нетрудно привести примеры задач (1), когда минимальный корень уравнения (5) строго меньше J_*.

Пример 4. Пусть $J(u) = u$, $U = \{u \in E^1:\ g(u) = (u^2 - 1) \times (u^4 + 1)^{-1} \leqslant 0\}$. Здесь $U = \{u \in E^1:\ -1 \leqslant u \leqslant 1\}$ и, очевидно, $J_* = -1$, $u_* = -1$. Если согласно (2) примем

$$\Phi(u,\ t) = \max\{u - t;\ 0\} + \max\{g(u);\ 0\}, \quad u \in U_0 = E^1,$$

то при $t \geqslant -1$ получим $\Phi(-1,\ t) = 0 = \inf\limits_{E^1} \Phi(u, t) = \rho(t)$.

При $t < -1$, взяв $u_k = -k \leqslant t$, также будем иметь $\lim\limits_{k \to \infty} \Phi(-k,\ t) = \lim\limits_{k \to \infty} g(-k) = 0 = \inf\limits_{E^1} \Phi(u, t) = \rho(t)$. Таким образом, в рассматриваемом случае $\rho(t) \equiv 0$ при всех t. Если в качестве мини-

мального решения уравнения (5) здесь взять $t_* = -\infty$, то получим $t_* < J_* = -1$.

Рассмотрим функцию (6)

$$\Phi(u, t) = |u - t| + \max\{g(u); 0\}, \quad u \in U_0 = E^1.$$

Если $|t| \leqslant 1$, то при $u = t$ получим $\Phi(t, t) = 0 = \rho(t)$. Пусть $|t| > 1$. Введем множества $A_1 = \left\{u \in E^1\colon |u - t| \leqslant \frac{|t|-1}{2}\right\}$, $A_2 = \left\{u \in E^2\colon |u - t| > \frac{|t|-1}{2}\right\}$. Так как $A_1 \cup A_2 = E^1$, $A_1 \cap A_2 = \varnothing$, то $\rho(t) = \inf\limits_{E^1} \Phi(u, t) = \min\left\{\inf\limits_{A_1} \Phi(u, t); \inf\limits_{A_2} \Phi(u, t)\right\} \geqslant \min\left\{\min\limits_{A_1} g(u); (|t| - 1)/2\right\} > 0$ при всех t, $|t| > 1$. На первый взгляд создается впечатление, что здесь $J_* = -1$ — минимальный корень уравнения (5). Однако $0 < \rho(t) \leqslant \Phi(t, t) = g(t)$ при $|t| > 1$ и $\lim\limits_{t \to -\infty} \rho(t) = \lim\limits_{t \to -\infty} g(t) = 0$. Поэтому есть основания считать, что минимальный корень уравнения (5) и в этом случае равен $t_* = -\infty < J_*$.

Любопытно сравнить задачи из примеров 2 и 4. В них функции $J(u)$ и множества U совпадают. Но эти задачи отличаются способом задания множества U. Это различие приводит к тому, что в примере 2 минимальный корень t_* уравнения (5) совпадает с J_*, а в примере 4 $t_* < J_*$. Отсюда можно сделать вывод: для выполнения равенства $t_* = J_*$, лежащего в основе метода нагруженных функций, исходные данные задачи (1) должны удовлетворять некоторым дополнительным условиям, они должны быть как-то согласованы. Для формулировки этих условий нам прежде всего нужно уточнить, что понимать под минимальным корнем уравнения (5).

Определение 1. Число t_* назовем *минимальным корнем* уравнения (5), если $\rho(t_*) = 0$, $\rho(t) > 0$ при всех $t < t_*$ и $\varliminf\limits_{t \to -\infty} \rho(t) > 0$. Если же $\varliminf\limits_{t \to -\infty} \rho(t) = 0$, то примем $t_* = -\infty$. Если $\rho(t) > 0$ при всех $t > 0$ и $\varliminf\limits_{t \to -\infty} \rho(t) > 0$, то по определению положим $t_* = \infty$.

Чтобы показать, что все указанные в определении 1 возможности в самом деле могут реализоваться, рассмотрим еще несколько примеров.

Пример 5. Пусть

$$J(u) = \begin{cases} u, & u > -2, \\ -k^2, & -(k+1) < u \leqslant -k, \quad k = 2, 3, \ldots; \end{cases}$$
$$g(u) = \begin{cases} u^2 - 1, & |u| \leqslant 2, \\ 6/|u|, & |u| > 2, \end{cases} \tag{7}$$

$U=\{u\in E^1=U_0:\ g(u)\leqslant 0\}$. Тогда $J_*=-1$, $u_*=-1$. Рассмотрим функцию (6)

$$\Phi(u,\ t)=|J(u)-t|+\max\{g(u);\ 0\},\quad u\in E^1.$$

Покажем, что $\rho(-k^2-k)\geqslant k$ $(k=2,\ 3,\ \ldots)$. В самом деле, если $u\geqslant -2$, то $\Phi(u,\ -k^2-k)\geqslant |u+k^2+k|=k^2+k+u\geqslant k^2\geqslant k$ при всех $k=2,\ 3,\ \ldots$ Если $-(i+1)<u\leqslant -i$ $(2\leqslant i\leqslant k)$, то $\Phi(u,\ -k^2-k)\geqslant |-i^2+k^2+k|=k^2-i^2+k\geqslant k$, а если $-(i+1)<u\leqslant -i$, $i\geqslant k+1$, то $\Phi(u,\ -k^2-k)\geqslant |-i^2+k^2+k|=i^2-(k+1)^2+k+1\geqslant k+1>k$. Таким образом, $\Phi(u,\ -k^2-k)\geqslant k$ для всех $u\in E^1$, поэтому $\rho(-k^2-k)\geqslant k$ $(k=2,\ 3,\ \ldots)$. Следовательно, $\overline{\lim\limits_{t\to-\infty}}\ \rho(t)=\lim\limits_{k\to\infty}\rho(-k^2-k)=\infty$. С другой стороны, $0\leqslant\rho(-k^2)\leqslant\Phi(-k,\ -k^2)=g(-k)=6k^{-1}$ $(k=2,\ 3,\ \ldots)$, так что $\underline{\lim\limits_{t\to-\infty}}\ \rho(t)=\lim\limits_{k\to\infty}\rho(-k^2)=0$. Согласно определению 1 тогда $t_*=-\infty<J_*=-1$.

Остановимся также на функции (2)

$$\Phi(u,\ t)=\max\{J(u)-t;\ 0\}+\max\{g(u);\ 0\},\quad u\in E^1.$$

Нетрудно видеть, что если $t\geqslant -1$, то $\Phi(-1,\ t)=0=\rho(t)$. Если же $t<-1$, то при $k>\sqrt{-t}$ получим $\Phi(-k,\ t)=\max\{-k^2-t;\ 0\}+g(-k)=g(-k)\to 0=\rho(t)$. Таким образом, здесь $\rho(t)\equiv 0$ при всех t и согласно определению 1 $t^*=-\infty$.

Пример 6. Переопределим функцию $J(u)$ из (7) в точках $u=k^{-1}$ так: $J(k^{-1})=-k^2$ $(k=1,\ 2,\ \ldots)$. Функцию $g(u)$ и множество U оставим такими же, как в примере 5. Повторив прежние рассуждения, нетрудно убедиться, что минимальный корень уравнению (5) здесь будет равным $t_*=J_*$ как при использовании функции (2), так и функции (6). В отличие от примера 5, здесь $J_*=-\infty$, поэтому справедливо $t_*=J_*$.

Любопытно посмотреть, что будет, если множество U из (1) пусто, но U_0 непусто. В этом случае задача (1), конечно, перестает быть содержательной, но тем не менее функции $\Phi(u,\ t)$, $\rho(t)$ из (2), (4), (6) будут иметь смысл.

Пример 7. Пусть $J(u)\equiv 1$, $U=\{u\in E^1:\ g(u)=e^{-u^2}\leqslant 0\}$. Здесь $U_0=E^1$, $U=\varnothing$. Согласно формуле (2) $\Phi(u,t)=\max\{1-t;0\}+e^{-u^2}$, $u\in E^1$, поэтому $\rho(t)=\max\{1-t;\ 0\}$ и $t_*=1$. Если же воспользуемся функцией (6) $\Phi(u,t)=|1-t|+e^{-u^2}$, то $\rho(t)=|1-t|$ и $t_*=1$.

Если здесь взять $g(u)=e^{-u^2}+1$, то получим $\rho(t)=\max\{1-t;\ 0\}+1$ для функции (2) и $\rho(t)=|1-t|+1$ для функции (6), так что минимальный корень уравнения (5) согласно определению 1 будет равен $t_*=\infty$.

Пример 8. Пусть $J(u)=u$, $U=\{u\in E^1\colon g(u)=e^{-u^2}\leqslant 0\}$. Здесь $J_0=E^1$, $U=\varnothing$. Согласно (2) $\Phi(u,t)=\max\{u-t;0\}+e^{-u}$. Так как при $u=-k\leqslant t$ $\Phi(-k,t)=e^{-k^2}\to 0$ при $k\to\infty$, то $\rho(t)\equiv 0$ при всех t и $t_*=-\infty$. В случае функции (6) $\Phi(u,\ t)=|u-t|+e^{-u^2}=\min\left\{\inf\limits_{|u-t|\leqslant 1}\Phi(u,t);\ \inf\limits_{|u-t|>1}\Phi(u,t)\right\}\geqslant$ $\geqslant\min\left\{\inf\limits_{|u-t|\leqslant 1}e^{-u^2};1\right\}=c(t)>0$ при всех $u\in E^1$, поэтому $\rho(t)>$ >0 при всех t. Но $0<\rho(t)<\Phi(t,\ t)\to 0$ при $t\to\infty$ или $t\to-\infty$, так что $\lim\limits_{t\to\infty}\rho(t)=\lim\limits_{t\to-\infty}\rho(t)=0$ и $t_*=-\infty$.

Если же здесь взять $g(u)=e^{-u^2}+1$, то $\Phi(u,\ t)\geqslant 1$, $u\in E^1$ и $\rho(t)\geqslant 1$ при всех t, и поэтому $t_*=\infty$.

3. Примеры 7, 8 подсказывают, что для того, чтобы единообразно охватить возможность, когда в задаче (1) $U=\varnothing$, целесообразно принять

$$J_*=\begin{cases}\inf\limits_U J(u), & U\neq\varnothing,\\ \infty, & U=\varnothing,\quad U_0\neq\varnothing.\end{cases}\tag{8}$$

Тогда справедлива следующая

Теорема 1. *Пусть функция* $\rho(t)$ *определена формулой* (4), *где функции* $\Phi(u,\ t)$ *взяты из* (2) *или* (6). *Пусть* t_* — *минимальный корень уравнения* (5) *в смысле определения* 1, *а величина* J_* *определена согласно* (8). *Тогда* $t_*\leqslant J_*$.

Доказательство. Если $U=\varnothing$, то $J_*=\infty$ и утверждение теоремы тривиально. Поэтому пусть $U\neq\varnothing$. Так как мы условились рассматривать функции, принимающие лишь конечные значения в области своего определения, то $J_*<\infty$. По определению J_*, существует последовательность $\{u_k\}\in U$ такая, что $\lim\limits_{k\to\infty}J(u_k)=J_*\geqslant-\infty$. Если $J_*>-\infty$, то $\lim\limits_{k\to\infty}\Phi(u_k,J_*)=$ $=0=\rho(J_*)$ и поэтому $t_*\leqslant J_*$. Если же $J_*=-\infty$, то, взяв $t_k=J(u_k)$, получим $\rho(t_k)=\Phi(u_k,\ t_k)=0$ $(k=1,\ 2,\ \ldots)$. Поскольку $\{t_k\}\to-\infty$, то отсюда следует $\varliminf\limits_{t\to-\infty}\rho(t)=\lim\limits_{k\to\infty}\rho(t_k)=0$, так что $t_*=J_*=-\infty$. Теорема доказана.

Рассмотренные выше примеры показывают, что для выполнения равенства $t_*=J_*$ важное значение имеет способ задания множества U: ограничения, задающие множество U, должны быть как-то согласованы с минимизируемой функцией $J(u)$. Напоминаем, что в § 14 было введено понятие согласованной постановки задачи (1) на U_0 (см. определение 14.2), означающее, что для любой последовательности $\{u_k\}\in U_0$, которая удовлетворяет условиям

$$\lim_{k\to\infty}g_i^+(u_k)=0,\qquad i=1,\ldots,s,\tag{9}$$

имеет место соотношение

$$\varliminf_{k\to\infty} J(u_k) \geqslant J_*. \tag{10}$$

Распространим это понятие на случай, когда $U_0 \neq \varnothing$, $U = \varnothing$ и согласно (8) $J_* = \infty$. Здесь следует различать две возможности: $\inf\limits_{U_0} P(u) = 0$ и $\inf\limits_{U_0} P(u) > 0$. Если $\inf\limits_{U_0} P(u) = 0$, то существует хотя бы одна последовательность $\{u_k\} \in U_0$, удовлетворяющая условиям (9),— в этом случае скажем, что задача (1) имеет согласованную постановку на U_0, если для любой последовательности $\{u_k\} \in U_0$, для которой справедливы соотношения (9), имеет место равенство $\lim\limits_{k\to\infty} J(u_k) = \infty = J_*$. Кстати, это же равенство получается и из (10) при $J_* = \infty$. Наконец, если $U_0 \neq \varnothing$, $U = \varnothing$, $\inf\limits_{U_0} P(u) > 0$, то по определению будем считать, что задача (1) имеет согласованную постановку на множестве U_0.

Оказывается, введенное понятие согласованной постановки задачи (1) играет важную роль при выяснении того, будет ли $t_* = J_*$ или $t_* < J_*$.

Теорема 2. *Пусть функция* $\rho(t)$ *определена формулой* (4), *где функция* $\Phi(u, t)$ *взята из* (2) *или* (6), *пусть* t_* — *минимальный корень уравнения* (5), *а величина* J_* *определена формулой* (7). *Тогда для выполнения равенства* $t_* = J_*$ *необходимо и достаточно, чтобы задача* (1) *имела согласованную постановку на множестве* U_0.

Доказательство. Необходимость. Пусть $t_* = J_*$. Если $J_* = -\infty$, то постановка задачи (1) согласована, так как $\varliminf\limits_{k\to\infty} J(u_k) \geqslant -\infty = J_*$ для любой последовательности $\{u_k\} \in U_0$. Поэтому пусть $J_* > -\infty$. Возьмем произвольную последовательность $\{u_k\} \in U_0$, удовлетворяющую условиям (9). Согласно определению (3) функции $P(u)$ тогда $\lim\limits_{k\to\infty} P(u_k) = 0$. Отсюда и из неравенств $\Phi(u_k, t) \geqslant \rho(t) > 0$, справедливых для всех $t < t_* = J_*$ и $k = 1, 2, \ldots$ при $k \to \infty$ получим

$$\varliminf_{k\to\infty} \max\{J(u_k) - t; 0\} \geqslant \rho(t) > 0, \quad t < t_*, \tag{11}$$

в случае использования функции (2) и

$$\varliminf_{k\to\infty} |J(u_k) - t| \geqslant \rho(t) > 0, \quad t < t_*, \tag{12}$$

в случае использования функции (6).

Покажем, что из (11), (12) следует неравенство $\varliminf\limits_{k\to\infty} J(u_k) \geqslant t_* = J_*$. В самом деле, при выполнении (11) для каждого

$t < t_*$ найдется номер $k_0 = k_0(t)$ такой, что $\max\{J(u_k) - t;\ 0\} \geqslant \rho(t)/2 > 0$ или $J(u_k) - t \geqslant \rho(t)/2 > 0$ для всех $k \geqslant k_0$. Тогда $\varliminf\limits_{k\to\infty} J(u_k) \geqslant t$ при любом $t < t_*$. Устремляя $t \to t_* - 0$, отсюда получим неравенство $\varliminf\limits_{k\to\infty} J(u_k) \geqslant t_* = J_*$.

Рассмотрим случай (12). Пусть $\varliminf\limits_{k\to\infty} J(u_k) = \lim\limits_{r\to\infty} J(u_{k_r}) = a$. Имеются две возможности: либо $a \geqslant t_*$, либо $a < t_*$. Если $a \geqslant t_*$, то требуемое неравенство $\varliminf\limits_{k\to\infty} J(u_k) \geqslant t_* = J_*$ установлено. Остается рассмотреть возможность $a < t_*$. В этом случае величина a не может быть конечной.

Допустим противное: пусть $-\infty < a < t_*$. Тогда при $t = a$ получим

$$\varliminf_{k\to\infty} |J(u_k) - a| = \left|\lim_{r\to\infty} J(u_{k_r}) - a\right| = 0,$$

что противоречит условию (12). Таким образом, если $a < t_*$, то $a = -\infty$, т. е. $\varliminf\limits_{k\to\infty} J(u_k) = \lim\limits_{r\to\infty} J(u_{k_r}) = -\infty$. Тогда, взяв $t_r = J(u_{k_r})$ $(r = 1, 2, \ldots)$, получим $0 < \rho(t_r) \leqslant \Phi(u_{k_r}, J(u_{k_r})) = MP(u_{k_r}) \to 0$ при $r \to \infty$. Это значит, что $\varliminf\limits_{t\to-\infty} \rho(t) = \lim\limits_{r\to\infty} \rho(t_r) = 0$ и согласно определению 1 $t_* = -\infty$. Но по условию $t_* = J_*$, поэтому $\varliminf\limits_{k\to\infty} J(u_k) = J_* = t_* = -\infty$, дело свелось к ранее рассмотренному случаю.

Тем самым установлено, что для любой последовательности $\{u_k\} \in U_0$, удовлетворяющей условиям (9), справедливо неравенство (10). Наконец, если такой последовательности $\{u_k\}$ не существует, т. е. $\inf\limits_{U_0} P(u) > 0$, то задача (1) имеет согласованную постановку по определению. Необходимость доказана.

Достаточность. Пусть задача (1) имеет согласованную постановку на U_0. Покажем, что тогда $t_* = J_*$. Сначала рассмотрим случай, когда $t_* = -\infty$. Это значит, что $\varliminf\limits_{t\to-\infty} \rho(t) = \lim\limits_{k\to\infty} \rho(t_k) = 0$, где $\{t_k\} \to -\infty$. По определению $\rho(t_k)$ согласно формуле (4) следует существование точки $u_k \in U_0$ такой, что $\rho(t_k) \leqslant \Phi(u_k, t_k) \leqslant \rho(t_k) + 1/k$ $(k = 1, 2, \ldots)$. Отсюда при $k \to \infty$ имеем $\lim\limits_{k\to\infty} \Phi(u_k, t_k) = 0$. Это означает, что $\lim\limits_{k\to\infty} P(u_k) = 0$ и $\lim\limits_{k\to\infty} \max\{J(u_k) - t_k;\ 0\} = 0$ в случае использования функции (2) и $\lim\limits_{k\to\infty} |J(u_k) - t_k| = 0$ — в случае функции (6). Но по построению $\{t_k\} \to -\infty$, поэтому последние два равенства возможны

только при $\lim\limits_{k\to\infty} J(u_k) = -\infty = t_*$. С другой стороны, из $\{P(u_k)\} \to 0$ и формулы (3) следует выполнение условий (9). В силу (10) тогда $\varliminf\limits_{k\to\infty} J(u_k) \geqslant J_*$. Следовательно, $J_* = t_* = -\infty$.

Пусть теперь $t_* > -\infty$. В силу теоремы 1 тогда $J_* \geqslant t_* > -\infty$. Возьмем произвольное $t < J_*$. По определению $\rho(t)$ существует последовательность $\{u_k\} \in U_0$ такая, что $\lim\limits_{k\to\infty} \Phi(u_k, t) = \rho(t)$. Может случиться, что $\varliminf\limits_{k\to\infty} P(u_k) = d > 0$. Тогда из $\Phi(u_k, t) \geqslant MP(u_k)$ при $k \to \infty$ следует, что $\rho(t) \geqslant M \varliminf\limits_{k\to\infty} P(u_k) = Md > 0$. Если же $\varliminf\limits_{k\to\infty} P(u_k) = 0 = \lim\limits_{r\to\infty} P(u_{k_r})$, то $\lim\limits_{r\to\infty} g_i^+(u_{k_r}) = 0$ $(i = 1, \ldots, s)$. В силу (10) отсюда имеет $\varliminf\limits_{r\to\infty} J(u_{k_r}) \geqslant J_* > t$. А тогда $\rho(t) = \lim\limits_{r\to\infty} \Phi(u_{k_r}, t) \geqslant L(J_* - t)^{p_0} > 0$ как в случае использования функции (2), так и функции (6). Тем самым показано, что $\rho(t) > 0$ при всех $t < J_*$. Кроме того, в рассматриваемом случае $t_* > -\infty$ по определению 1 $\varliminf\limits_{t\to-\infty} \rho(t) > 0$. Следовательно, $t_* \geqslant J_*$, что в силу теоремы 1 возможно только при $t_* = J_*$. Теорема доказана.

В § 14 были приведены достаточные условия, гарантирующие согласованную постановку задачи (1) на U_0 (см. теорему 14.2, леммы 14.1, 14.5).

4. Подробнее остановимся на частном случае функции (2), когда

$$\Phi(u, t) = L \max\{J(u) - t;\ 0\} + MP(u), \quad u \in U_0, \tag{13}$$

где $L > 0$, $M > 0$, а функция $P(u)$ взята из (3) при некоторых $p_i \geqslant 1$ $(i = 1, \ldots, s)$. Оказывается, функция (13) и соответствующая ей функция $\rho(t)$ обладают рядом полезных свойств, облегчающих поиск минимального корня уравнения (5).

Теорема 3. *Функции $\Phi(u, t)$, $\rho(t)$, определяемые формулами* (13), (14), *монотонно убывают (вообще говоря, не строго) при возрастании t и удовлетворяют неравенствам*

$$|\Phi(u, t) - \Phi(u, \tau)| \leqslant L|t - \tau|, \tag{14}$$

$$|\rho(t) - \rho(\tau)| \leqslant L|t - \tau| \tag{15}$$

при всех $u \in U_0$ и любых t, τ. Если $J_{**} = \inf\limits_{U_0} J(u) > -\infty$, то

$$\Phi(u, t) = -Lt + LJ(u) + MP(u), \quad \rho(t) = -Lt + \inf_{U_0}(LJ(u) + MP(u)) \tag{16}$$

при всех $t \leqslant J_{**}$ — линейные функции по t.

Доказательство. Простым перебором возможных значений функции $\max\{a;\ b\}$ легко доказываются неравенства

$$\max\{J(u) - t;\ 0\} \geqslant \max\{J(u) - \tau;\ 0\}, \quad t \leqslant \tau, \quad u \in U_0,$$

$$|\max\{J(u) - t;\ 0\} - \max\{J(u) - \tau;\ 0\}| \leqslant |t - \tau|, \quad u \in U_0.$$

Отсюда следует невозрастание функции $\Phi(u, t)$ по переменной t и неравенство (14). Далее, для любых $t \leqslant \tau$ имеем $\Phi(u, t) \geqslant \Phi(u, \tau) \geqslant \rho(\tau)$ или $\Phi(u, t) \geqslant \rho(\tau)$ при каждом $u \in U_0$. Отсюда, переходя к нижней грани по $u \in U_0$, получим $\rho(t) \geqslant \rho(\tau)$ при всех $t \leqslant \tau$.

Докажем неравенство (15). Зафиксируем произвольные t, τ. По определению нижней грани при каждом $\varepsilon > 0$ существуют точки u_t, $u_\tau \in U_0$ такие, что

$$\rho(t) \leqslant \Phi(u_t, t) \leqslant \rho(t) + \varepsilon, \quad \rho(\tau) \leqslant \Phi(u_\tau, \tau) \leqslant \rho(\tau) + \varepsilon.$$

Тогда, учитывая уже доказанное неравенство (14), имеем $\rho(t) - \rho(\tau) \leqslant \leqslant \Phi(u_\tau, t) - \Phi(u_\tau, \tau) + \varepsilon \leqslant L|t - \tau| + \varepsilon$, $\rho(t) - \rho(\tau) \geqslant \Phi(u_t, t) - \varepsilon - \Phi(u, \tau) \geqslant -L|t - \tau| - \varepsilon$, т. е. $|\rho(t) - \rho(\tau)| \leqslant L|t - \tau| + \varepsilon$ при любом $\varepsilon > 0$. Отсюда при $\varepsilon \to +0$ получим неравенство (15). Формулы (16) следуют из того, что $J(u) - t \geqslant J_{**} - t \geqslant 0$ при всех $u \in U_0$. Теорема 3 доказана.

Если задача (1) имеет согласованную постановку на U_0 и $J_* > -\infty$, то опираясь на теорему 3 можно предположить следующий итерационный метод определения J_*. Сначала выберем t_0 так, чтобы $\rho(t_0) > 0$ (например, если $J_{**} = \inf\limits_{U_0} J(u) > -\infty$, то можно взять любую точку $t_0 \leqslant J_{**}$). Следующие приближения определим по формулам

$$t_{k+1} = t_k + \rho(t_k)/L, \quad k = 0, 1, \ldots \tag{17}$$

Теорема 4. *Пусть функция $\rho(t) \geqslant 0$ при всех t, $-\infty < t < +\infty$, удовлетворяет условию* (15), *пусть t_* — минимальный корень уравнения* (5) *в смысле определения* 1, $t_* > -\infty$. *Тогда при любом выборе начального приближения t_0, $-\infty < t_0 < t$, последовательность $\{t_k\}$, определяемая условиями* (17), *сходится к t_*.*

Доказательство. Так как $\rho(t) \geqslant 0$, то из (17) следует, что последовательность $\{t_k\}$ монотонно возрастает, и поэтому существует $\lim\limits_{k\to\infty} t_k = a \leqslant \infty$. Покажем, что $a = t_*$. По условию $t_0 < t_*$. Допустим, что при некотором $k \geqslant 0$ оказалось $t_k < t_*$. Тогда $\rho(t) > 0$ при всех $t \leqslant t_k$. Возьмем произвольное t, $t_k \leqslant t < t_{k+1}$. С учетом условий (15), (17) имеем

$$\rho(t) = \rho(t_k) + [\rho(t) - \rho(t_k)] \geqslant \rho(t_k) - L(t - t_k) > \\ > \rho(t_k) - L(t_{k+1} - t_k) = 0, \quad t_k \leqslant t < t_{k+1}.$$

Это значит, что $\rho(t) > 0$ при всех $t < t_{k+1}$, т. е. $t_{k+1} \leqslant t_*$. Может случиться, что $\rho(t_{k+1}) = 0$. Тогда $t_{k+1} = t_*$ — в этом случае итерации (17) заканчивается. Если $\rho(t_{k+1}) > 0$, то $t_{k+1} < t_*$ и итерации продолжаются дальше.

Таким образом, имеются две возможности. Либо процесс (17) закончится тем, что $\rho(t_0) > 0, \ldots, \rho(t_{k-1}) > 0$, $\rho(t_k) = 0$ — тогда $t_k = t_* = a$, утверждение теоремы верно. Либо $\rho(t_k) > 0$, $t_k < t_*$, $\rho(t) > 0$ при $t < t_k$ для всех $k = 0, 1, \ldots$ — в этом случае $\lim\limits_{k\to\infty} t_k = a \leqslant t$ и $\rho(t) > 0$ при всех $t < a$. Покажем, что $a = t_*$. Если последовательность $\{t_k\}$ неограничена сверху, то $a = \infty = t_*$. Если же $t_k < a < \infty$, $k = 0, 1, \ldots$, то, учитывая непрерывность функции $\rho(t)$, из (17) при $k \to \infty$ получим $a = a + \rho(a)/L$ или $\rho(a) = 0$. Это значит, что $t_* = a$ и при $a < \infty$. Теорема доказана.

Заметим, что на каждом шаге метода (17) нужно вычислить одно значение функции $\rho(t)$, и для этого в свою очередь нужно решить задачу минимизации

$$\Phi(u, t) \to \inf; \quad u \in U_0. \tag{18}$$

Поскольку функция (13), вообще говоря, не является гладкой, то это обстоятельство может вызвать некоторые трудности при решении задачи (18). Однако имеющиеся методы решения негладких задач минимизации (см., например, § 3, 11, 12, 17) позволяют надеяться на то, что вычисление приближенного значения $\rho(t)$ не окажется слишком трудным.

При изложении метода (17) предполагалось, что величины $\rho(t_k)$ известны точно. Однако задача (18) на практике, как правило, будет решаться приближенно, и точное значение $\rho(t_k)$ удастся вычислить лишь в редких случаях. Поэтому желательно обобщить итерационный процесс (17) на случай, когда значения функции известны неточно. Опишем одно из возможных таких обобщений [82].

Предположим, что вместо точных значений функции $\rho(t)$ известны лишь некоторые приближения $\rho_\nu(t)$ $(\nu = 1, 2, \ldots)$, удовлетворяющие условиям

$$\rho_\nu(t) \geqslant 0, \quad |\rho_\nu(t) - \rho(t)| \leqslant \gamma_\nu, \quad \nu = 1, 2, \ldots; \quad \lim_{\nu\to\infty} \gamma_\nu = 0. \tag{19}$$

Пусть t_0 — начальное приближение, $t_0 < t_*$. Пусть $(\nu-1)$-е приближение $t_{\nu-1}$ при некотором $\nu \geqslant 1$ уже известно. Для определения следующего приближения t_ν рассмотрим итерационный процесс

$$t_{\nu k+1} = t_{\nu k} + \rho_\nu(t_{\nu k})/L, \qquad k = 0, 1, \ldots; \quad t_{\nu 0} = t_0, \tag{20}$$

аналогичный процессу (17). Поскольку функция $\rho_\nu(t)$ может не обращаться в нуль ни в одной точке даже в том случае, когда уравнение (5) имеет конечный минимальный корень (так будет, например, если $\rho_\nu(t) = \rho(t) + \gamma_\nu \geqslant \gamma_\nu > 0$), то процесс (20) следует прекращать не по критерию $\rho_\nu(t_{\nu k}) = 0$, как было выше в (17), а по условию вида $\rho_\nu(t_{\nu k}) < \theta_\nu$, где величина $\theta_\nu > 0$ стремится к нулю при $\nu \to \infty$ и как-то согласована с погрешностью γ_ν. Предположим, что такая последовательность $\{\theta_\nu\}$ уже задана (условия согласования $\{\theta_\nu\}$ и $\{\gamma_\nu\}$ будут обсуждаться ниже). Тогда имеются две возможности:

1) либо найдется номер $k = k_\nu \geqslant 0$ такой, что

$$\rho_\nu(t_{\nu k}) > \theta_\nu, \quad k = 0, \ldots, k_\nu - 1, \quad \rho_\nu(t_{\nu k_\nu}) \leqslant \theta_\nu; \tag{21}$$

в этом случае процесс (20) заканчивается и полагаем

$$t_\nu = t_{\nu k_\nu}; \tag{22}$$

2) либо

$$\rho_\nu(t_{\nu k}) > \theta_\nu \quad \text{при всех} \quad k = 0, 1, \ldots \tag{23}$$

Тогда, как будет показано ниже, при выполнении условий (15), (19) и согласованном изменении величин θ_ν и γ_ν будет справедливо равенство

$$t_* = \infty. \tag{24}$$

Метод поиска минимального корня уравнения (5) при условиях (15), (19) описан

Теорема 5. *Пусть функция $\rho(t)$ неотрицательная при всех t, не возрастает, удовлетворяет условию* (15), *а $t_* > -\infty$ — минимальный корень уравнения* (5) *в смысле определения* 1. *Кроме того, пусть функция $\{\rho_\nu(t)\}$ удовлетворяет условиям* (19) *и*

$$\theta_\nu \geqslant \gamma_\nu, \qquad \nu = 1, 2, \ldots \tag{25}$$

Тогда последовательность $\{t_\nu\}$, определяемая методом (20)—(24), *сходится к t_* при любом выборе начального приближения t_0, $-\infty < t_0 < t_*$. При этом, если $t_* < \infty$, то итерации* (20) *при каждом $\nu \geqslant 1$ будут заканчи-*

ваться за конечное число шагов выполнением условий (21); *случай* (23) *возможен лишь при* $t_* = \infty$.

Доказательство. Сначала рассмотрим случай $t_* < \infty$. Тогда при каждом фиксированном $\nu \geqslant 1$ имеются две возможности:

1) $t_{\nu k} \leqslant t_* < \infty$ при всех $k = 0, 1, \ldots$ В силу монотонности $\{t_{\nu k}\}$ тогда существует $\lim\limits_{k\to\infty} t_{\nu k} \leqslant t_* < \infty$. Переходя в (20) к пределу при $k \to \infty$, получим $\lim\limits_{k\to\infty} \rho_\nu(t_{\nu k}) = 0$. Это значит, что за конечное число итераций процесс (20) закончится выполнением условий (21).

2) Найдется номер $l \geqslant 0$ такой, что $t_{\nu l} \leqslant t_* < t_{\nu l+1}$. Тогда с учетом соотношений (15), (19), (20) получим

$$t_{\nu,l+1} = t_{\nu l} + \rho_\nu(t_{\nu l})/L = t_{\nu l} + [\rho_\nu(t_{\nu l}) - \rho(t_{\nu l})]/L + \\ + [\rho(t_{\nu l}) - \rho(t_*)]/L \leqslant t_{\nu l} + \gamma_\nu/L + t_* - t_{\nu l} = t_* + \gamma_\nu/L. \quad (26)$$

Далее, в силу монотонности $\rho(t)$ имеем $\rho(t) \equiv 0$ при $t \geqslant t_*$, поэтому $\rho(t_{\nu l+1}) = 0$. Отсюда и из условий (19), (25) следует

$$\rho_\nu(t_{\nu l+1}) = \rho_\nu(t_{\nu l+1}) - \rho(t_{\nu l+1}) \leqslant \gamma_\nu \leqslant \theta_\nu. \quad (27)$$

Это значит, что условия (21) выполнятся при некотором $k_\nu \leqslant l + 1$.

Объединяя обе рассмотренные возможности, заключаем, что при $t_* < \infty$ процесс (20) при каждом $\nu \geqslant 1$ заканчивается за конечное число шагов k_ν выполнением условий (21), причем в силу (26)

$$t_\nu = t_{\nu k_\nu} \leqslant t_* + \gamma_\nu L^{-1}, \qquad \nu = 1, 2, \ldots \quad (28)$$

Покажем, что $\lim\limits_{\nu\to\infty} t_\nu = t_*$. Из (28) имеем $\overline{\lim\limits_{\nu\to\infty}}\, t_\nu \leqslant t_*$. Пусть $\underline{\lim\limits_{\nu\to\infty}}\, t_\nu = \lim\limits_{r\to\infty} t_{\nu_r} = a$. Тогда с учетом условий (15), (19), (21) получим

$$0 \leqslant \rho(a) \leqslant \left[\rho(a) - \rho(t_{\nu_r})\right] + \left[\rho(t_{\nu_r})\right] - \rho_{\nu_r}(t_{\nu_r})\right] + \rho_{\nu_r}(t_{\nu_r}) \leqslant \\ \leqslant L\left|a - t_{\nu_r}\right| + \gamma_{\nu_r} + \theta_{\nu_r} \to 0$$

при $r \to \infty$, т. е. $\rho(a) = 0$. Но t_* — минимальный корень уравнения (5), поэтому $a \geqslant t_*$. Следовательно, $\underline{\lim\limits_{\nu\to\infty}}\, t_\nu = \lim\limits_{r\to\infty} t_{\nu_r} > t_* \geqslant \overline{\lim\limits_{\nu\to\infty}}\, t_\nu$. Это значит, что $\lim\limits_{\nu\to\infty} t_\nu = t_*$, Случай $t_* < \infty$ полностью рассмотрен.

Пусть теперь $t_* = \infty$ и пусть процесс (20) при каждом $\nu \geqslant 1$ заканчивается выполнением условий (21). Покажем, что тогда $\{t_\nu\} \to \infty$. Возьмем произвольное число $T > 0$. Согласно определению 1, если $t_* = \infty$, то $\underline{\lim\limits_{t\to-\infty}}\, \rho(t) > 0$ и $\rho(t) > 0$ при всех t. Отсюда и из непрерывности функции $\rho(t)$ следует, что $\inf\limits_{t\leqslant T} \rho(t) = \rho_T > 0$. Так как $\{\theta_\nu\} \to 0$, $\{\gamma_\nu\} \to 0$, то найдется номер $\nu_0 = \nu_0(T)$ такой, что $\theta_\nu + \gamma_\nu < \rho_T$ при всех $\nu \geqslant \nu_0$. Тогда $\rho_\nu(t) \geqslant \rho(t) - \gamma_\nu \geqslant \rho_T - \gamma_\nu > \theta_\nu$ для всех $t \leqslant T$ и $\nu \geqslant \nu_0$. Это значит, что условия (21) не могут выполняться при $t_{\nu k} \leqslant T$, если $\nu \geqslant \nu_0$. Тогда согласно (21), (22) $t_\nu > T$ для всех $\nu \geqslant \nu_0 = \nu_0(T)$, что означает выполнение равенства $\lim\limits_{\nu\to\infty} t_\nu = \infty$.

Остается рассмотреть случай, когда при некотором $\nu \geqslant 1$ выполняются условия (23). Выше было установлено, что при $t_* < \infty$ процесс (20) при всех $\nu \geqslant 1$ закончится выполнением условий (21). Следовательно, если при каком-либо $\nu \geqslant 1$ реализуются условия (23), то $t_* = \infty$. Теорема 5 доказана.

Заметим, что условие (25) в теореме 5 существенно: его нарушение может привести к тому, что метод (20) — (24) не будет сходиться.

Пример 9. Пусть $J(u) = u \to \inf$; $u \in U = U_0 = \{u \in E^1\colon u \geqslant 0\}$ — это частный случай задачи (1), когда $g_i(u) \equiv 0$ (или $s = 0$). Тогда $\Phi(u, t) = \max\{u - t, 0\}$, $u \geqslant 0$ и $\rho(t) = \inf\limits_{u \geqslant 0} \Phi(u, t) = \max\{-t; 0\}$ (ср. с примером 1). Здесь $t_* = J_* = 0$, $u_* = 0$. Если $\rho_\nu(t) = \max\{-t; 0\} + \gamma_\nu$ и $\theta_\nu < \gamma_\nu$, то $\rho_\nu(t) \geqslant \gamma_\nu > \theta_\nu$ при всех t. Поэтому в методе (20) — (24) реализуется случай (23), и искомый минимальный корень $t_* = 0$ уравнения (5) в рассматриваемом случае не будет найден. Причина этого явления — нарушение условия (25).

Заметим, также, что на практике вместо полуоси $t_0 \leqslant t < \infty$ часто приходится работать на каком-то отрезке $t_0 \leqslant t \leqslant T$, где величина T ограничена, например, разрядной сеткой ЭВМ. В этом случае метод (20) — (24) требует модификации. А именно, итерационный процесс (20) при каждом $\nu \geqslant 1$ здесь будет заканчиваться определением номера $k_\nu \geqslant 0$ такого, что будет выполнено одно из двух следующих условий:

$$\rho_\nu(t_{\nu k}) > \theta_\nu, \quad k = 0, \ldots, k_\nu - 1; \quad \rho_\nu(t_{\nu k_\nu}) \leqslant \theta_\nu, \quad t_{\nu k_\nu} \leqslant T, \tag{29}$$

или

$$\rho_\nu(t_{\nu k}) > \theta_\nu, \quad k = 0, \ldots, k_\nu - 1; \quad t_{\nu k_\nu - 1} \leqslant T < t_{\nu k_\nu}. \tag{30}$$

В качестве ν-го приближения t_ν будем брать

$$t_\nu = \min\{t_{\nu k_\nu}\colon T\}, \quad \nu = 1, 2, \ldots \tag{31}$$

Если выполнены все условия теоремы 5, то, немного видоизменив доказательство этой теоремы, нетрудно установить, что при $t_0 < t_* < T$ для достаточно больших номеров $\nu > \nu_0$ процесс (20) будет заканчиваться выполнением условий (29) и оценки (28), а последовательность $\{t_\nu\}$, определяемая методом (20), (29) — (31), сходится к числу $\min\{t_*; T\}$. Таким образом, метод (20), (29) — (31) позволяет определить, принадлежит ли t_* отрезку $[t_0, T]$, и в случае $t_* \in [t_0, T]$ позволяет найти t_* с нужной точностью.

Для определения t_* при условиях (19) может быть также использован метод деления отрезка пополам.

5. Кратко остановимся на случае, когда функция $\rho(t)$ из (4) определяется с помощью функции

$$\Phi(u, t) = L|J(u) - t| + MP(u), \quad u \in U_0, \tag{32}$$

где $L > 0$, $M > 0$, функция $P(u)$ взята из (3) при некоторых $p_i \geqslant 1$ $(i = 1, \ldots, s)$.

Поскольку $\|J(u) - t| - |J(u) - \tau\| \leqslant |t - \tau|$, то, рассуждая так же, как при доказательстве теоремы 3, убеждаемся, что функции $\Phi(u, t)$, $\rho(t)$ из (4), (32) удовлетворяют условиям (14), (15). Это значит, что для поиска минимального корня уравнения (5) и в этом случае может быть применен описанный выше метод (20)—(24) или его модификация (20), (29)—(31). Только условие (25) здесь нужно заменить условием

$$\theta_\nu \geqslant 2\gamma_\nu, \quad \nu = 1, 2, \ldots \tag{33}$$

Такая замена связана с тем, что функция (32), в отличие от (13), а также соответствующая ей функция $\rho(t)$, вообще говоря, не будут монотонными (см. примеры 1—8). Справедлива

Теорема 6. *Пусть функция $\rho(t)$ неотрицательна при всех t, удовлетворяет условию* (15), *а $t_* > -\infty$ — минимальный корень уравнения* (5) *в смысле определения* 1. *Кроме того, пусть функции $\{\rho_\nu(t)\}$ удовлетворяют условиям* (19), *а последовательности $\{\theta_\nu\}$, $\{\gamma_\nu\}$ — условию* (33). *Тогда*

последовательность $\{t_\nu\}$, *определяемая методом* (20) — (24), *сходится к* t_* *при любом выборе* t_0 $(-\infty < t_0 < t_*)$. *При этом, если* $t_* < \infty$, *то итерации* (20) *при каждом* $\nu \geqslant 1$ *будут заканчиваться за конечное число шагов выполнением условий* (21); *случай* (23) *возможен лишь при* $t_* = \infty$.

Доказательство проводится дословно так же, как доказательство теоремы 5. Нужно лишь неравенство (27), полученное в предположении монотонности $\rho(t)$ и условия (25), заменить следующим неравенством, вытекающим из условий (15), (19), (26), (33):

$$\rho_\nu(t_{\nu,l+1}) = [\rho_\nu(t_{\nu,l+1}) - \rho(t_{\nu,l+1})] + [\rho(t_{\nu,l+1}) - \rho(t_*)] \leqslant$$
$$\leqslant \gamma_\nu + L(t_{\nu,l+1} - t_*) \leqslant 2\gamma_\nu \leqslant \theta_\nu.$$

Нетрудно также показать, что при выполнении условий теоремы 6 последовательность $\{t_\nu\}$, полученная методом (20), (29) — (30), сходится к $\min\{t_*; T\}$.

Приведем пример, показывающий, что условие (33) в общем случае не может быть ослаблено.

Пример 10. Пусть $J(u) \equiv 0$, U — произвольное непустое множество вида (1). Тогда $\Phi(u, t) = |t| + MP(u)$, $u \in U_0$ и $\rho(t) = |t|$; $t_* = 0 = J_*$. Пусть $\rho_\nu(t) = |t| + \gamma_\nu$ $(\nu = 1, 2, \ldots)$. Предположим, что условие (33) нарушено, т. е. $\theta_\nu < 2\gamma_\nu$. Возьмем начальное приближение $t_0 = t_{\nu 0} < \min\{\gamma_\nu - \theta_\nu; 0\}$. Тогда $\rho_\nu(t_{\nu 0}) = |t_{\nu 0}| + \gamma_\nu = -t_{\nu 0} + \nu_\nu > \theta_\nu$. Далее, $t_{\nu 1} = t_{\nu 0} + \rho_\nu(t_{\nu 0}) = \gamma_\nu > 0$, и снова $\rho_\nu(t_{\nu 1}) = \rho_\nu(\gamma_\nu) = 2\gamma_\nu > \theta_\nu$. Отсюда с учетом монотонности $\rho_\nu(t)$ при $t \geqslant 0$ имеем $\rho_\nu(t) \geqslant \rho_\nu(t_{\nu 1}) > \theta_\nu$ для всех $t \geqslant t_{\nu 1}$. Это значит, что $\rho_\nu(t_{\nu k}) > \theta_\nu$ для всех $k = 0, 1, \ldots$ — реализовался случай (23), и искомый минимальный корень $t_* = 0$ уравнения (5) здесь не будет найден.

Полезно заметить, что при описании методов (17), (20) — (24) и (20), (29) — (31), а также при формулировке и доказательстве теорем 4—6 никак не использовался тот факт, что функция $\rho(t)$ получена из (4) и как-то связана с задачей (1), с методом нагруженных функций. Это значит, что описанные методы могут быть использованы для поиска минимального корня уравнения (5) для любой неотрицательной функции $\rho(t)$, удовлетворяющей условию (15) и приближенно заданной посредством условий (19).

По поводу метода нагруженных функций см., например, [8, 21, 82, 85, 257].

Упражнения. 1. Найти минимальное решение уравнения (5), где функции $\Phi(u, t)$, $\rho(t)$ определяются из (2), (3), (6), (13) или (32), и проверить условие $t_* = J_*$ для задачи: $J(u) \to \inf$; $u \in U = \{u \in E^1: u \in U_0, g(u) \leqslant 0\}$, где

а) $J(u) = \operatorname{arctg} u$, $g(u) = u^2 - 4$, $g(u) = (u^2 - 4)(u^4 + 1)^{-1}$, $g(u) = u$, $g(u) = u(u^2 + 1)^{-1}$, $g(u) = u^2$, $U_0 = \{u \in E^1: u \geqslant 1\}$, $U_0 = \{u \in E^1: u \geqslant 0\}$, $U_0 = E^1$;

б) $J(u) = u \sin u$, $g(u) = u^2 - 1$; $g(u) = (u^2 - 1)(u^4 + 1)^{-1}$, $g(u) = (u^2 - 1)e^{-u^2}$; $U_0 = \{u \in E^1: u \geqslant 0\}$;

в) $J(u)$ — произвольная функция, а $U = U_0 = E^1$;

г) $J(u) = 1$ при $u \leqslant 1$, $J(u) = u^{-1}$ при $u > 1$; $g(u) = u - 1$, $g(u) = (u - 2)(u^2 + 1)^{-1}$, $g(u) = e^{-u^2}$, $U_0 = \{u \in E^1: u \geqslant 0\}$ или $U_0 = E^1$;

д) $J(u) \equiv 1$, $g(u) = u$, $g(u) = u^2 + 1$, $g(u) = e^{-u^2}(u^2 + 1)$, $g(u) = u^2 e^{-u^2}$; $U_0 = E^1$, $U_0 = \{u \in E^1: u \geqslant 0\}$.

2. Найти функцию $\rho(t)$ для задачи $J(u) \to \inf$; $u \in U = \{u \in E^1 = U_0: g(u) = |u| - 1 \leqslant 0\}$, беря за основу функции $\Phi(u, t)$ из (13) и (32), сравнить результаты с примером 2.

3. Показать, что если функция $\rho(t)$ построена с помощью функций (2) или (6) при $p_0 > 1$, то условия (14), (15), вообще говоря, не будут иметь места (ни с какой константой L). Указание: рассмотреть задачу (1) при $J(u) \equiv 1$, $U = U_0$.

4. Указать такой способ задания множества U из примера 5, чтобы задача имела согласованную постановку на $U_0 = E^1$. Рассмотреть возможности $g(u) = |u| - 1$, $g(u) = u^2 - 1$, $g(u) = e^{-u^2}(u^2 - 1)$ (воспользуйтесь теоремой 2).

5. Выяснить геометрический смысл методов (17), (20)—(24) и (20), (29) — (31), а также геометрический смысл условий (25), (33).

6. Проверить, будут ли функции $\Phi(u, t)$ из (2), (6), (13), (32), а также соответствующая функция $\rho(t)$ из (4) выпуклы, если исходная задача (1) выпукла.

7. Пусть в задаче (1) $J_* > -\infty$, $U_* \neq \varnothing$, и эта задача имеет согласованную постановку на множестве U_0. Пусть последовательность $\{t_k\}$ построена методом (17), а последовательность $\{u_k\}$ определена условиями: $u_k \in U_0$, $\rho(t_k) = \Phi(u_k, t_k)$ $(k = 0, 1, \ldots)$. Можно ли ожидать, что $\{u_k\} \to U_*$? Приведите примеры.

8. Пусть функция Лагранжа задачи (1) имеет седловую точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$. Показать, что та же точка (u_*, λ^*) является седловой точкой функции Лагранжа для задачи

$$\max\{J(u) - t;\ 0\} \to \inf; \qquad u \in U$$

при всех $t \leqslant J_*$. Выяснить связь между множествами точек минимума последней задачи и задачи (1).

9. Для задачи (1) ввести функцию

$$\Phi_1(u, t) = L \max\{J(u);\ t\} + MP(u), \qquad u \in U_0,$$

где $P(u)$ взята из (3), $L > 0$, $M > 0$, и положить $\rho_1(t) = \inf\limits_{u \in U} \Phi_1(u, t)$. Показать, что $J_* = \rho_1(J_*)$. Можно ли утверждать, что J_* будет минимальным корнем уравнения $\rho_1(t) - t = 0$? Пользуясь равенством $\max\{J(u), t\} = \max\{J(u) - t;\ 0\} + t$, установить связь между функциями $\Phi_1(u, t)$, $\rho_1(t)$ и функциями $\Phi(u, t)$, $\rho(t)$ из (2), (13).

10. Для задачи $J(u) \to \inf$; $u \in U = \{u \in U_0,\ g_1(u) \leqslant 0, \ldots, g_m(u) \leqslant 0\}$ ввести функцию $G(w, u) = \max\{J(w) - J(u);\ g_1(w), \ldots, g_m(w)\}$ или $G(w, u) = (J(w) - J(u))g_1(w) \ldots g_m(w)$ переменных u, $w \in U_0$ и рассмотреть итерационный процесс $G(u_{k+1}, u_k) = \inf\limits_{w \in U_0} G(w, u_k)$; исследовать его сходимость (метод центров, [159]).

§ 17. О методе случайного поиска

Наряду с описанными выше методами минимизации функций переменных существует большая группа методов поиска минимума, объединенных под названием метода случайного поиска. Метод случайного поиска, в отличие от ранее рассмотренных методов, характеризуется намеренным введением элемента случайности в алгоритм поиска. Многие варианты метода случайного поиска сводятся к построению последовательности $\{u_k\}$ по правилу:

$$u_{k+1} = u_k + \alpha_k \xi, \qquad k = 0, 1, \ldots, \tag{1}$$

где α_k — некоторая положительная величина $\xi = (\xi^1, \ldots, \xi^n)$ — какая-либо реализация n-мерной случайной величины ξ с известным законом распределения. Например, координаты ξ^i случайного вектора ξ могут представлять независимые случайные ве-

личины, распределенные равномерно на отрезке $[-1, 1]$. Как видим, метод случайного поиска минимума функции n переменных предполагает наличие датчика (или генератора) случайных чисел, обращаясь к которому, в любой нужный момент можно получить какую-либо реализацию n-мерного случайного вектора ξ с заданным законом распределения. Такие датчики, оформленные в виде стандартных программ, имеются в библиотеках подпрограмм на ЭВМ.

1. Приведем несколько вариантов метода случайного поиска минимума функции $J(u)$ на множестве $U \subseteq E^n$, предполагая, что k-е приближение $u_k \in U$ $(k \geqslant 0)$ уже известно.

а) *Алгоритм с возвратом при неудачном шаге.* Смысл этого алгоритма заключается в следующем. С помощью датчика случайного вектора получают некоторую его реализацию ξ и в пространстве E^n определяют точку $v_k = u_k + \alpha\xi$, $\alpha = \text{const} > 0$. Если $v_k \in U$ и $J(v_k) < J(u_k)$, то сделанный шаг считается удачным, и в этом случае полагается $u_{k+1} = v_k$. Если $v_k \in U$, но $J(v_k) \geqslant J(u_k)$, или же $v_k \notin U$, то сделанный шаг считается неудачным и полагается $u_{k+1} = u_k$.

Если окажется, что $u_k = u_{k+1} = \ldots = u_{k+N}$ для достаточно больших N, то точка u_k может быть принята в качестве приближения искомой точки минимума.

б) *Алгоритм наилучшей пробы.* Берутся какие-либо s реализаций $\xi_1, \ldots, \xi_s$ случайного вектора ξ и вычисляются значения функции $J(u)$ в тех точках $u = u_k + \alpha\xi_i$ $(i = 1, \ldots, s)$, которые принадлежат множеству U. Затем полагается $u_{k+1} = u_k + \alpha\xi_{i_0}$, где индекс i_0 определяется условием

$$J(u_k + \alpha\xi_{i_0}) = \min_{\substack{u_k + \alpha\xi_i \in U \\ 1 \leqslant i \leqslant s}} J(u_k + \alpha\xi_i).$$

Величины $s > 1$ и $\alpha = \text{const} > 0$ являются параметрами алгоритма.

в) *Алгоритм статистического градиента.* Берутся какие-либо s реализаций $\xi_1, \ldots, \xi_s$ случайного вектора ξ и вычисляются разности $\Delta J_{ki} = J(u_k + \gamma\xi_i) - J(u_k)$ для всех $u_k + \gamma\xi_i \in U$. Затем полагают $p_k = \frac{1}{\gamma} \sum_i \xi_i \Delta J_{ki}$, где сумма берется по всем тем i, $1 \leqslant i \leqslant s$, для которых $u_k + \gamma\xi_i \in U$. Если $u_k + \alpha p_k \in U$, то принимается $u_{k+1} = u_k + \alpha p_k$. Если же $u_k + \alpha p_k \notin U$, то повторяют описанный процесс с новым набором из s реализаций случайного вектора ξ. Величины $s > 1$, $\alpha > 0$, $\gamma > 0$ являются параметрами алгоритма. Вектор p_k называют статистическим градиентом. Если $U \equiv E^n$, $s = n$, и векторы ξ_i являются неслучайными и совпадают с соответствующими единичными векторами $e_i = (0, \ldots, 0, 1, \ldots, 0)$ $(i = 1, \ldots, n)$, то описанный алгоритм, как нетрудно видеть, превращается в разностный аналог градиентного метода.

2. В описанных вариантах а)—в) метода случайного поиска предполагается, что закон распределения случайного вектора ξ не зависит от номера итераций. Такой поиск называют *случайным поиском без обучения.* Алгоритмы случайного поиска без обучения не обладают «способностью» анализировать результаты предыдущих итераций и выделять направления, более перспективные в смысле убывания минимизируемой функции, и сходятся, вообще говоря, медленно.

Между тем ясно, что от метода случайного поиска можно ожидать большей эффективности, если на каждой итерации учитывать накопленный опыт поиска минимума на предыдущих итерациях и перестраивать вероятностные свойства поиска так, чтобы направления ξ, более перспективные в смысле убывания функции, становились более вероятными. Иначе говоря, желательно иметь алгоритмы случайного поиска, которые обладают способностью к самообучению и самоусовершенствованию в процессе поиска минимума в зависимости от конкретных особенностей минимизируемой функции. Такой поиск называют *случайным поиском с обучением.* Обучение алгоритма осуществляют посредством целенаправленного изменения закона распределения случайного вектора ξ в зависимости от номера итерации и результатов предыдущих итераций таким образом, чтобы «хорошие» направления, по которым функция убывает, стали более вероятными, а другие направления — менее вероятными. Таким образом, на различных этапах метода случайного поиска с обучением приходится иметь дело с реализациями случайных векторов ξ с различными законами распределения. Имея в виду это обстоятельство, итерационный процесс (1) удобнее записать в виде

$$u_{k+1} = u_k + \alpha_k \xi_k, \qquad k = 0, 1, \ldots, \tag{2}$$

подчеркнув зависимость случайного вектора ξ от k.

В начале поиска закон распределения случайного вектора $\xi = \xi_0$ выбирают с учетом имеющейся априорной информации о минимизируемой функции. Если такая информация отсутствует, то поиск обычно начинают со случайного вектора $\xi_0 = (\xi_0^1, \ldots, \xi_0^n)$, компоненты ξ_0^i $(i = 1, \ldots, n)$ которого представляют собой независимые случайные величины, распределенные равномерно на отрезке $[-1, 1]$.

Для обучения алгоритма в процессе поиска часто берут семейство случайных векторов $\xi = \xi(w)$, зависящих от параметров $w = (w^1, \ldots, w^n)$, и при переходе от k-й итерации к $(k+1)$-й итерации имеющиеся значения параметров w_k заменяют новыми значениями w_{k+1} с учетом результатов предыдущего поиска.

Приведем два варианта метода случайного поиска с обучением для минимизации функции $J(u)$ на всем пространстве.

а) *Алгоритм покоординатного обучения.* Пусть имеется семейство случайных векторов $\xi = \xi(w) = (\xi^1, \ldots, \xi^n)$, каждая координата ξ^i которых принимает два значения: $\xi^i = 1$ с вероятностью p^i и $\xi^i = -1$ с вероятностью $1 - p^i$, где вероятности p^i зависят от параметра w^i следующим образом:

$$p^i = \begin{cases} 0, & w^i < -1, \\ \frac{1}{2}(1 + w^i), & |w^i| \leqslant 1, \\ 1, & w^i > 1, \quad i = 1, \ldots, n. \end{cases} \tag{3}$$

Пусть начальное приближение u_0 уже выбрано. Тогда для определения следующего приближения u_1 в формуле (2) при $k = 0$ берется какая-либо реализация случайного вектора $\xi_0 = \xi(0)$, соответствующего значению параметров $w = w_0 = (0, 0, \ldots, 0)$. Приближение u_2 определяется по формуле (2) при $k = 1$ с помощью случайного вектора $\xi_1 = \xi(0)$. Пусть известны приближения $u_0, u_1, \ldots, u_k$ и значения параметров $w_{k-1} = (w^1_{k-1}, \ldots, w^n_{k-1})$ при некотором $k \geqslant 1$. Тогда полагаем

$$w^i_k = \beta w^i_{k-1} - \delta \operatorname{sign}\left[(J(u_{k-1}) - J(u_{k-2}))(u^i_{k-1} - u^i_{k-2})\right], \quad i = 1, \ldots, n, \quad k = 2, 3, \ldots, \tag{4}$$

где величина $\beta \geqslant 0$ называется *параметром забывания*, $\delta \geqslant 0$ — *параметром интенсивности обучения*, $\beta + \delta > 0$. При определении следующего приближения u_{k+1} в формуле (2) берем какую-либо реализацию случайного вектора $\xi_k = \xi(w_k)$, $w_k = (w^1_k, \ldots, w^n_k)$.

Из (3), (4) видно, что если переход от точки u_{k-2} к u_{k-1} привел к уменьшению значения функции, то вероятность выбора направления $u_{k-1} - u_{k-2}$ на следующем шаге увеличивается. И наоборот, если при переходе от u_{k-2} к u_{k-1} значение функции увеличилось, то вероятность выбора направления $u_{k-1} - u_{k-2}$ на последующем шаге уменьшается. Таким образом, формулы (4) осуществляют обучение алгоритма. Величина $\delta \geqslant 0$ в (4) регулирует скорость обучения: чем больше $\delta > 0$, тем быстрее обучается алгоритм; при $\delta = 0$, как видно, обучения нет. Величина $\beta \geqslant 0$ в формулах (4) регулирует влияние предыдущих значений параметров на обучение алгоритма; при $\beta = 0$ алгоритм «забывает» предыдущие значения w_{k-1}. Для устранения возможного чрезмерного детерминирования алгоритма и сохранения способности алгоритма к достаточно быстрому обучению на параметры w^i_k накладываются ограничения $|w^i_k| \leqslant c_i$, и при нарушении этих ограничений w^i_k заменяются ближайшим из чисел c_i и $-c_i$ $(i = 1, \ldots, n)$. Величины β, δ, c_i являются параметрами алгоритма.

Вместо формул (4), посредством которых производится обучение алгоритма, часто пользуются другими формулами

$$w_k^i = \beta w_{k-1}^i - \delta\left(J(u_{k-1}) - J(u_{k-2})\right)\left(u_{k-1}^i - u_{k-2}^i\right), \quad i = 1, \ldots, n, \quad k = 2, 3, \ldots \tag{5}$$

Описанный алгоритм покоординатного обучения имеет тот недостаток, что поиск и обучение происходят лишь по одному из 2^m направлений $\xi = (\xi^1, \ldots, \xi^n)$, где либо $\xi^i = 1$, либо $\xi^i = -1$. Отсутствие «промежуточных» направлений делает покоординатное обучение немобильным в областях с медленно изменяющимися направлениями спуска. От этого недостатка свободен следующий алгоритм.

б) *Алгоритм непрерывного самообучения.* Пусть имеется семейство случайных векторов $\xi = \xi(w) = \dfrac{\eta + w}{|\eta + w|}$, где $w = (w^1, \ldots, w^n)$ — параметры обучения, $\eta = (\eta^1, \ldots, \eta^n)$ — случайный вектор, координаты η^i которого представляют собой независимые случайные величины, распределенные равномерно на отрезке $[-1, 1]$. Поиск начинается с рассмотрения случайных векторов $\xi_0 = \xi(0)$, $\xi_1 = \xi(0)$, реализации которых используются при определении приближений u_0, u_1 по формулам (2). Обучение алгоритма при $k \geqslant 2$ производится так же, как в алгоритме покоординатного обучения, с помощью формул (4) или (5). При больших значениях $|w_k|$ влияние случайной величины η уменьшается, и направление $\xi_k = \xi(w_k)$ становится более детерминированным и близким к направлению w_k. Во избежание излишней детерминированности метода на параметры $w_k = (w_k^1, \ldots, w_k^n)$ накладываются ограничения $|w_k| \leqslant c = \text{const}$, и при нарушении этих ограничений w_k заменяется на $\dfrac{w_k}{|w_k|} c$.

Приведенные алгоритмы случайного поиска с обучением показывают, что процесс обучения в ходе поиска сопровождается уменьшением фактора случайности и увеличением степени детерминированности алгоритма поиска минимума, направляя поиск преимущественно по направлению убывания функции. В то же время наличие случайного фактора в выборе направления дает возможность алгоритму «переучитываться», если свойства функции в районе поиска изменились или предыдущее обучение было неточным. Случайный поиск с обучением в некотором смысле занимает промежуточное положение между случайным поиском без обучения и детерминированными методами поиска минимума из предыдущих параграфов. Разумеется, и в методах предыдущих параграфов можно обнаружить в том или ином виде элементы самообучения алгоритма, однако наличие случайного фактора в алгоритме делает метод случайного поиска более гибким.

3. Весьма усложняет решение задачи минимизации функций многих переменных наличие помех, когда на значения функции $J(u)$ в каждой точке u накладываются случайные ошибки. В этих случаях можно использовать метод стохастической аппроксимации. Один из вариантов этого метода был описан в § 1.13 применительно к задачам минимизации функций одной переменной. Для функций n переменных эта процедура приводит к построению последовательности $\{u_k\}$ по закону

$$u_{k+1}^i = u_k^i - a_k \frac{z(u_k + c_k e_i) - z(u_k - c_k e_i)}{c_k}, \qquad i = 1, \ldots, n, \qquad k = 0, 1, \ldots,$$

где $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, $z(u)$ — наблюдаемые в эксперименте значения функции $J(u)$ в точке u, последовательности $\{a_k\}$, $\{c_k\}$ удовлетворяют условиям (1.13.2).

Заметим, что задача минимизации функций при наличии случайных ошибок относится к задачам стохастического программирования. Более подробно о стохастическом программировании, о теоретических и вычислительных аспектах методов случайного поиска, стохастической аппроксимации см., например, в [11, 49, 113, 123, 127, 147, 148, 151, 154, 157, 158, 173, 180, 214, 235, 259, 279, 306, 339].

§ 18. Общие замечания

Выше были рассмотрены лишь немногие из известных в настоящее время методов минимизации. В гл. 7 мы обсудим еще один метод решения задач минимизации специального вида — метод динамического программирования. Заметим, что по сей день интенсивно продолжается разработка все новых и новых методов решения экстремальных задач, о чем свидетельствует неуклонно растущее количество публикаций в научной печати по этой тематике.

1. Возникают естественные вопросы: чем руководствоваться при выборе метода для решения той или иной конкретной экстремальной задачи, какой же метод является наилучшим? Иногда считают, что тот метод лучше, у которого выше скорость сходимости на некотором фиксированном классе задач. Однако при таком способе оценки методов не принимается во внимание такое важное качество, как трудоемкость каждой отдельно взятой итерации метода. Нередко бывает, что при решении конкретной задачи выгоднее применять метод, который сходится не очень быстро и для получения решения с нужной точностью требует довольно большого числа итераций, но тем не менее из-за того, что каждая итерация метода осуществляется просто, суммарный объем вычислений и, следовательно, общее машинное время для получения решения оказывается меньшим, чем при применении

другого быстросходящегося метода, каждая итерация которого весьма трудоемка. Таким образом, при характеристике метода минимизации важным является не столько скорость его сходимости, сколько общий объем вычислений, общее машинное время, необходимое для получения решения с нужной точностью.

При практическом использовании методов значительная часть времени, отведенного на расчеты, часто затрачивается на вычисление значений минимизируемой функции или ее производных. Поэтому в тех случаях, когда вычисление значений функции намного проще вычисления ее производных, естественно, выгоднее пользоваться теми методами, которые для своей реализации требуют лишь вычисления значений функции. Конечно, возможны и такие ситуации, когда имеются простые аналитические выражения для производных минимизируемой функции,— в таких случаях, возможно, выгоднее применять методы, использующие градиент или производные более высокого порядка.

Важными характеристиками метода минимизации являются также область сходимости метода, устойчивость метода к погрешностям, объем памяти ЭВМ, необходимой для реализации метода, удобство программирования, широта класса задач, к которым применим метод, и т. п.

Большое количество разнообразных и отчасти противоречивых характеристик методов, недостаточная разработанность методики оценки упомянутых характеристик затрудняют сравнение методов друг с другом. Иногда для сравнения методов минимизации задают некоторой набор тестовых задач (набор таких задач см., например, в [314]) и лучшим признают тот метод, с помощью которого удается решить указанные тестовые задачи с нужной точностью за меньшее число итераций, меньшее число вычислений значений функции или ее производных, или за меньшее машинное время. Несомненно, такие «соревнования» методов полезны, хотя и на их основе нельзя делать окончательные выводы о преимуществах того или иного метода. Здесь следует также заметить, что один и тот же метод, примененный для минимизации одной и той же функции, может привести к различным результатам в зависимости от того, на каком алгоритмическом языке составлена программа, каково качество транслятора (квалификация программиста), на какой ЭВМ решается задача и т. д.

Конечно, хотелось бы иметь метод, наилучший во всех отношениях. Однако такого универсального метода пока нет, и вряд ли такой метод существует. Поэтому для эффективного решения конкретной задачи минимизации, по-видимому, нужно разумно сочетать различные методы с учетом всевозможной априорной информации о решаемой задаче (гладкость исходных данных, выпуклость, физические или какие-либо иные соображения об области возможного расположения точки минимума и т. д.),

имеющихся вычислительных средств, ресурсов машинного времени и т. п. В тех случаях, когда нет никакой априорной информации о задаче, которую нужно решить, по-видимому, сначала полезно попробовать применить не очень точные, но простые методы минимизации (например, метод перебора значений функций на сетке с небольшим числом узловых точек, метод покоординатного спуска, метод случайного поиска), а затем на основе накопленной информации при необходимости перейти к более точным методам.

2. Успешное решение различных классов прикладных экстремальных задач невозможно без пакета минимизации, состоящего из библиотеки подпрограмм, охватывающей достаточно много методов минимизации, а также управляющих и вспомогательных программ. Пакеты минимизации могут быть использованы в автоматизированном или диалоговом режиме.

При работе с пакетом в диалоговом режиме математик — вычислитель, получая сведения о текущих результатах, оперативно вмешивается в процесс минимизации, осуществляет переход от одного метода к другому, изменяет параметры методов, параметры программ. Диалоговый режим работы с пакетом минимизации позволяет лучшим образом использовать опыт и интуицию математика — вычислителя и предъявляет высокие требования к его профессиональным знаниям в области методов решения экстремальных задач.

В тех случаях, когда пользователь, т. е. специалист, проводящий расчеты, не является компетентным в области методов решения экстремальных задач, желательно иметь пакеты минимизации, работающие в автоматическом режиме. Для работы в этом режиме пакет должен содержать управляющую программу, обеспечивающую автоматический выбор наиболее подходящей последовательности используемых методов, их параметров в зависимости от конкретной решаемой задачи.

Принцип построения пакетов минимизации, примеры таких пакетов описаны в [8, 15]. Заметим, что создание эффективно действующих и достаточно универсальных пакетов минимизации, которые могут быть использованы в различных режимах, представляет собой важную и большую научно-техническую задачу.

3. Следует обратить внимание читателя на то, что первоначальная постановка прикладных задач минимизации зачастую бывает достаточно грубой, упрощенной и предполагает, что в процессе решения задача будет уточняться. Это значит, что первоначальный вариант задачи не всегда имеет смысл решать слишком точно. Иногда гораздо выгоднее с помощью простых методов, с небольшой затратой машинного времени получить грубые предварительные результаты и затем проанализировать их вместе с экспертами, с заказчиком. Уже при таком упрощен-

ном анализе может выясниться, что некоторые параметры и ограничения, ранее казавшиеся несущественными и поэтому не учтенные в первоначальной постановке задачи, должны быть включены в нее, и наоборот, часть прежних параметров и ограничений могут оказаться несущественными и без ущерба для существа задачи могут быть опущены. Заметим, что процесс уточнения постановки задачи весьма удобно проводить с помощью пакета минимизации в диалоговом режиме.

Иногда стремятся учесть многие детали задачи и создать слишком подробную математическую модель исследуемого процесса, а затем пытаются найти наилучшие, оптимальные значения всех параметров процесса. Однако такой подход может привести к задаче минимизации с очень большим числом переменных и численное решение такой задачи может встретить непреодолимые трудности. Но даже в тех случаях, когда удается найти оптимальные значения параметров, их практическое использование может оказаться невозможным из-за того, что заказчик, будучи не в состоянии охватить полученную информацию, может не понять разумность выработанных на ее основе рекомендаций и может от них вообще отказаться. Поэтому на первых этапах исследования прикладных задач минимизации желательно пользоваться простыми моделями, учитывающими основные, определяющие параметры.

4. К сожалению, последнему совету удается следовать не всегда. Например, математические модели социально-экономических процессов, достаточно адекватно отражающих основные закономерности, как правило, чрезвычайно сложны, содержат большое число переменных и приводят к так называемым задачам минимизации большой размерности. Численное решение таких задач обычными методами становится невозможным даже при использовании самых мощных современных ЭВМ. Некоторые классы задач большой размерности допускают разбиение на ряд слабо связанных между собой подзадач, имеющих сравнительно небольшие размерности, решая которые, иногда удается получить приближенное решение исходной задачи. Следует заметить, что задачи минимизации большой размерности к настоящему времени изучены слабо. Некоторые методы решения таких задач см., например, в [111, 117, 194, 203, 242, 302, 320, 330].

5. В ряде методов минимизации, описанных выше, предполагалось, что начальная точка u_0, принадлежащая множеству U, известна. Для некоторых множеств, таких, как, например, параллелепипед, шар, гиперплоскость, указать такую точку u_0 совсем нетрудно. Однако не следует думать, что определение точки u_0 из любого множества U всегда просто. Например, если

$$U = \{u \in E^n:\ g_i(u) = 0, \quad i = 1, \ldots, s\}, \tag{1}$$

то для определения точки $u_0 \in U$ нужно решать систему урав-

нений (вообще говоря, нелинейных). Чтобы найти какую-либо точку множества

$$U = \{u \in E^n:\ g_i(u) \leqslant 0, \quad i = 1, \ldots, m\}, \tag{2}$$

придется решать систему неравенств. Определение решения систем линейных или нелинейных уравнений и неравенств представляет собой весьма серьезную задачу, которой посвящена обширная литература; см., например, [4, 8, 13, 20, 22, 39, 42, 45, 54, 73, 76, 93, 106, 111, 112, 117, 119, 162, 200, 209, 221, 237—239, 277, 282, 296, 298, 301, 321, 324, 335].

Полезно заметить, что задачу нахождения какой-либо точки u_0, принадлежащей множеству (1) или (2), можно переформулировать в виде задачи минимизации. А именно, в случае множества (1) введем функцию

$$P(u) = \sum_{i=1} g_i^2(u), \quad u \in E^n,$$

а в случае множества (2) — функцию

$$P(u) = \sum_{i=1}^{m} (\max\{g_i(u); 0\})^p, \quad u \in E^n, \quad p > 0,$$

и рассмотрим задачу минимизации

$$P(u) \to \inf; \quad u \in E^n.$$

Для решения этой задачи могут быть использованы любые подходящие методы минимизации. Если $U \neq \varnothing$, то условие $u_0 \in U$ равносильно условию $P(u_0) = 0 = \inf\limits_{E^n} P(u) = P_*$. Если $P_* > 0$, то $U = \varnothing$. Если же $P_* = 0$, но нижняя грань $P(u)$ на E^n не достигается, то также $U = \varnothing$.

Если

$$U = \{u \in E^n:\ u \in U_0,\ g_i(u) \leqslant 0,\ i = 1, \ldots, m,\ g_i(u) = 0,\ i = m+1, \ldots, s\},$$

то для определения какой-либо точки $u_0 \in U$ можно рассмотреть задачу минимизации

$$P(u) = \sum_{i=1}^{m} (\max\{g_i(u); 0\})^p + \sum_{i=m+1}^{s} |g_i(u)|^p \to \inf; \quad u \in U_0,\ p > 0.$$

Здесь предполагается, что множество U_0 имеет столь простую структуру, что нахождение точки $u_0 \in U_0$ не вызывает трудностей.

Задачу отыскания точки, принадлежащей множеству (2), можно свести к несколько иной задаче минимизации

$$\sigma \to \inf; \quad z = (u, \sigma) \in G = \{z = (u, \sigma) \in E^{n+1}:\ g_i(u) \leqslant \sigma, \quad i = 1, \ldots, m\}. \tag{3}$$

Понятно, что если $U \neq \varnothing$, то $\inf\limits_G \sigma \leqslant 0$. Определение начальной точки $z_0 = (u_0, \sigma_0) \in G$ не вызывает трудностей: достаточно взять какую-либо точку $u_0 \in E^n$ и положить $\sigma_0 = \max\limits_{1 \leqslant i \leqslant m} g_i(u_0)$. Для решения задачи (3) может быть использован, например, метод возможных направлений. Итерационный процесс можно прекратить, как только обнаружится точка $z = (u, \sigma) \in G$, для которой $\sigma \leqslant 0$. В том случае, когда множество (2) регулярно, т. е. существует точка $v \in E^n$, для которой $g_i(v) < 0$ $(i = 1, \ldots, m)$, то $\inf\limits_G \sigma < 0$, и ясно, что любой сходящийся метод минимизации в задаче (3) позволит за конечное число итераций найти точку множества (2).

6. Интересно проанализировать доказательства теорем сходимости градиентного метода, методов проекции градиента, возможных направлений, условного градиента и т. д. Такой анализ показывает, что проводимые рассуждения опираются на предположения одного и того же типа, содержат много общих моментов, техника получения оценок скорости сходимости имеет общие черты. Возникает вопрос, нельзя ли создать общую методику исследования сходимости если и не всех, то хотя бы некоторых достаточно широких семейств методов минимизации? Оказывается, это возможно. К настоящему времени сделаны весьма удачные и интересные попытки создания такой методики, позволяющей единообразно исследовать сходимость широких классов методов минимизации, получать оценку скорости сходимости. К сожалению, мы здесь не имеем возможности останавливаться на этих увлекательных вопросах и отсылаем читателя к работам [8, 11, 49, 140, 159, 214, 235, 248, 250].

Глава 6

ПРИНЦИП МАКСИМУМА ПОНТРЯГИНА

В этой главе рассматриваются задачи оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений. Этот класс экстремальных задач существенно отличается от рассмотренных: если в задачах минимизации функции конечного числа переменных искомая точка минимума являлась точкой n-мерного пространства, то в задачах оптимального управления искомая точка минимума, вообще говоря, представляет собой функцию, принадлежащую некоторому бесконечномерному функциональному пространству. Такие задачи имеют многочисленные приложения в механике космического полета, в вопросах управления электроприводами, химическими или ядерными реакторами, виброзащиты и т. д.

Эффективным средством исследования задач оптимального управления является принцип максимума Понтрягина [17], представляющий собой необходимое условие оптимальности в таких задачах. Принцип максимума, открытый коллективом советских математиков во главе с академиком Л. С. Понтрягиным, представляет собой одно из крупных достижений современной математики и является краеугольным камнем современной математической теории оптимального управления. Принцип максимума Понтрягина существенно обобщает и развивает основные результаты классического вариационного исчисления, созданного Эйлером, Лагранжем и другими выдающимися математиками прошлого. Появление принципа максимума стимулировало последующее бурное развитие теории экстремальных задач и методов их решения.

Из обширной литературы, посвященной различным аспектам современной теории оптимального управления и управляемых систем, их приложений, упомянем [1, 2, 5—9, 14, 17, 22, 28, 31, 32, 34, 36—38, 43, 50, 51, 59—70, 75, 77, 80, 81, 97—101, 104, 117, 118, 120, 121, 124, 125, 136—138, 142, 149, 161, 166, 167, 172, 174, 175, 181—189, 192, 193, 199, 204—206, 210—212, 217, 218—220, 222, 223, 226, 232, 234, 236, 243, 244, 246, 248, 249, 252—254, 260, 263, 268, 271, 273, 275, 280, 281, 285, 286, 289, 290, 293, 294, 300, 303—306, 308—311, 315, 317, 319, 325—328, 339, 341, 342].

§ 1. Постановка задачи оптимального управления

1. Сначала приведем несколько конкретных задач оптимального управления.

Пример 1. Движение плоского маятника, подвешенного к точке опоры при помощи жесткого невесомого стержня (рис. 6.1), как известно, описывается уравнением

$$I\ddot{\theta} + b\dot{\theta} + mgl\sin\theta = M(\tau),$$

где l — длина жесткого стержня маятника, m — масса, сосредоточенная в конце стержня, $I = ml^2$ — момент инерции, g — гравитационная постоянная (ускорение силы тяжести), $b \geqslant 0$ — коэффициент демпфирования, τ — время, $M(\tau)$ — внешний управляющий момент, $\theta = \theta(\tau)$ — угол отклонения стержня от точки устойчивого равновесия. Если сделать замену переменной $t = \tau\sqrt{mgl/I}$, то это уравнение можно привести к виду

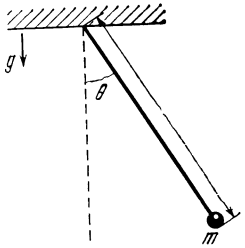

Рис. 6.1

$$\ddot{\varphi} + \beta\dot{\varphi} + \sin\varphi = u(t), \tag{1}$$

где

$$\varphi = \varphi(t) = \theta(t\sqrt{I/(mgl)}), \quad \beta = b/\sqrt{Imgl},$$
$$u(t) = M(t\sqrt{I/(mgl)})/(mgl).$$

Обозначим $x^1(t) = \varphi(t)$ (угол отклонения маятника), $x^2(t) = \dot{\varphi}(t)$ (скорость маятника). Тогда уравнение (1) запишется в виде системы двух уравнений первого порядка:

$$\dot{x}^1(t) = x^2(t), \quad \dot{x}^2(t) = -\beta x^2(t) - \sin x^1(t) + u(t). \tag{2}$$

Пусть в начальный момент $t = 0$ маятник отклонился на угол $x^1(0) = x_0^1$ и имеет начальную скорость $x^2(0) = x_0^2$. Будем также считать, что функция $u(t)$ — управляющий момент (выбор которого может влиять на движение маятника) — удовлетворяет ограничению

$$|u(t)| \leqslant \gamma, \quad \gamma = \text{const} > 0, \quad t \geqslant 0. \tag{3}$$

Здесь возможны следующие постановки задач оптимального управления: выбрать управление $u(t)$, удовлетворяющее условиям (3) так, чтобы:

1) за минимальное время T остановить маятник в одной из точек устойчивого равновесия, т. е. добиться выполнения условий

$$x^1(T) = 2\pi k, \quad x^2(T) = 0 \tag{4}$$

при некотором k ($k = 0, \pm 1, \ldots$) (задача быстродействия);

2) за минимальное время T добиться выполнения условия

$$(x^1(T))^2 + (x^2(T))^2 \leqslant \varepsilon,$$

где $\varepsilon > 0$ — заданное число;

3) к заданному моменту T величина $(x^1(T))^2 + (x^2(T))^2$, или $\int_0^T (x^1(t))^2\,dt$, или $\int_0^T ((x^1(t))^2 + (x^2(t))^2)\,dt$, или $\max\limits_{0 \leqslant t \leqslant T} |x^1(t)|$, или $\max\limits_{0 \leqslant t \leqslant T} \max\{|x^1(t)|, |x^2(t)|\}$ принимала минимально возможное значение, или

4) в заданный момент T выполнялось равенство $x^2(T)=0$, а величина $x^1(T)$ была максимально возможной (задача о накоплении возмущений), или

5) к заданному моменту T добиться выполнения условий (4) и минимизировать величину $\int_0^T u^2(t)\,dt$ (выполнение условия (3) здесь необязательно).

Если колебание маятника ограничено какими-либо упорами, то в перечисленных задачах нужно еще требовать выполнения условия вида

$$|x^1(t)| \leqslant \mu, \quad \mu = \text{const} > 0.$$

На управление $u(t)$ вместо условия (3) (или наряду с условиями (3)) могут накладываться ограничения вида

$$\int_0^T u^2(t)\,dt \leqslant R,$$

где $R = \text{const} > 0$.

При изучении малых колебаний маятника часто полагают $\sin\varphi \approx \varphi$, и тогда уравнение (1) и эквивалентная ему система становятся линейными и будут иметь вид

$$\ddot{\varphi} + \beta\dot{\varphi} + \varphi = u(t)$$

и соответственно

$$\dot{x}^1(t) = x^2(t), \quad \dot{x}^2(t) = -\beta x^2(t) - x^1(t) + u(t).$$

Пример 2. Как известно [121, с. 129], движение центра масс космического аппарата и расход массы описывается системой дифференциальных уравнений

$$\dot{r} = v, \quad \dot{v} = gp/G + F, \quad \dot{G} = -gq, \quad 0 \leqslant t \leqslant T, \tag{5}$$

где t — время, $r = r(t) = (r_1(t),\ r_2(t),\ r_3(t))$ — радиус-вектор центра масс аппарата, $v = v(t) = (v_1(t),\ v_2(t),\ v_3(t))$ — скорость центра масс, $G = G(t)$ — текущий вес аппарата, g — коэффициент пропорциональности между массой и весом, $p = p(t) = (p_1(t),\ p_2(t),\ p_3(t))$ — вектор тяги двигателя, $q = q(t)$ — расход рабочего вещества, $F = F(r,\ t) = (F_1,\ F_2,\ F_3)$ — вектор ускорения от гравитационных сил.

В каждый момент времени t движение космического аппарата характеризуется величинами $r(t)$, $v(t)$, $G(t)$, называемыми фазовыми координатами. Пусть в начальный момент $t=0$ фазовые координаты аппарата известны:

$$r(0) = r_0, \quad v(0) = v_0, \quad G(0) = G_0. \tag{6}$$

Величины $q = q(t)$, $p = p(t)$ являются управлением — задавая их по-разному, можно получить различные фазовые траектории

(решения) задачи (5), (6). Конструктивные возможности аппарата, ограниченность ресурсов рабочего вещества накладывают на управление $q(t)$, $p(t)$ ограничения, например, вида

$$p_{\min} \leqslant |p(t)| \leqslant p_{\max}, \quad q_{\min} \leqslant q(t) \leqslant q_{\max}, \quad 0 \leqslant t \leqslant T,$$

или $\int\limits_0^T q^2(t)\,dt \leqslant R$, $\quad R = \text{const} > 0$. Кроме того, на фазовые траектории задачи (5), (6) могут накладываться некоторые ограничения, вытекающие, например, из условий того, чтобы вес аппарата был не меньше определенной величины или траектория полета проходила вне определенных областей космического пространства (областей повышенной радиации) и др.

Здесь возникают задачи выбора управлений $q(t)$, $p(t)$ так, чтобы управления и соответствующие им траектории задачи (5), (6) удовлетворяли всем наложенным ограничениям, и кроме того, достигалась та или иная цель. Например, здесь возможны следующие задачи:

1) попасть в заданную точку или область космического пространства за минимальное время;

2) к заданному моменту времени попасть в заданную область пространства с заданной скоростью (совершить мягкую посадку, например) и с максимальным весом аппарата или с минимальной затратой энергии;

3) достичь определенной скорости за минимальное время и т. п.

Большое число прикладных задач оптимального управления, связанных с механикой полета летательных аппаратов в космосе и атмосфере, с работой электроприводов, химических и ядерных реакторов, с вопросами виброзащиты и амортизации, с математической экономикой и т. д., читатель найдет в [9, 14, 28, 35, 40, 43, 68, 69, 75, 118, 121, 124, 125, 188, 189, 192, 199, 202, 207, 208, 217, 218, 243, 257, 260, 261, 263, 290, 300, 303—305, 309, 310, 322, 323, 325—328].

2. Приведенные в примерах 1, 2 задачи являются частным случаем более общей задачи оптимального управления, к формулировке которой мы переходим. Пусть движение некоторого управляемого объекта (течение управляемого процесса, изменение управляемой системы) описывается обыкновенными дифференциальными уравнениями

$$\dot{x}^i = f^i(x^1, x^2, \ldots, x^n, u^1, u^2, \ldots, u^r), \qquad i = 1, \ldots, n,$$

которые в векторной форме можно записать в виде

$$\dot{x} = f(x, u, t), \tag{7}$$

где t — время, $x = (x^1, x^2, \ldots, x^n)$ — величины, характеризующие движение объекта в зависимости от времени и называемые *фа-*

зовыми координатами объекта, $u=(u^1, u^2, \ldots, u^r)$ — *параметры управления* («положение рулей» объекта), выбором которых можно влиять на движение объекта, $f=(f^1, f^2, \ldots, f^n)$; функции $f^i(x, u, t)$ $(i=1, \ldots, n)$, описывающие внутреннее устройство объекта и учитывающие различные внешние факторы, предполагаются известными.

Для того, чтобы фазовые координаты объекта (процесса, системы) (7) были определены в виде функций времени $x=x(t)$ на некотором отрезке $t_0 \leqslant t \leqslant T$, необходимо в начальный момент времени t_0 задать начальное условие $x(t_0)=x_0$ и параметры управления $u=(u^1, u^2, \ldots, u^r)$ как функции времени $u=u(t)$ при $t \in [t_0, T]$. Тогда фазовые координаты $x=x(t)$ будут определяться как решение следующей задачи Коши:

$$\dot{x}(t)=f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T, \tag{8}$$

$$x(t_0)=x_0. \tag{9}$$

Нетрудно видеть, что функции $u=u(t)$, называемые *управлениями*, должны удовлетворять определенным требованиям непрерывности, гладкости, так как, с одной стороны, при слишком «плохих» (слишком «разрывных») $u(t)$ задача (8), (9) не будет иметь смысла, с другой стороны, слишком «плохая» функция $u(t)$ не будет иметь физического смысла управления. В большинстве прикладных задач в качестве управлений $u=u(t)$ могут быть взяты кусочно-непрерывные функции. Напоминаем, что функция $u(t)$ называется *кусочно-непрерывной* на отрезке $[t_0, T]$, если $u(t)$ непрерывна во всех точках $t \in [t_0, T]$, за исключением, быть может, лишь конечного числа точек $\tau_1, \ldots, \tau_p \in [t_0, T]$, в которых функция $u(t)$ может терпеть разрывы типа скачка, т. е. существуют конечные пределы

$$\lim_{t \to \tau_i - 0} u(t) = u(\tau_i - 0), \quad \lim_{t \to \tau_i + 0} u(t) = u(\tau_i + 0),$$

но, вообще говоря, $u(\tau_i - 0) \neq u(\tau_i + 0)$ $(i=1, \ldots, p)$. В тех прикладных задачах, в которых разрывные управления технически нереализуемы, рассматриваются лишь непрерывные управления $u(t)$. Встречаются также задачи, в которых, кроме непрерывности, от $u(t)$ требуется существование кусочно-непрерывной производной $\dot{u}(t)$ — такие управления называют *кусочно-гладкими*.

В теоретических исследованиях упомянутые классы кусочно-непрерывных, кусочно-гладких управлений часто бывают слишком узкими, и вместо них приходится рассматривать более широкие классы управлений, такие, как, например, пространство $L_p^r[t_0, T]$ при некотором p $(1 \leqslant p \leqslant \infty)$ [179]. Через $L_p^r[t_0, T]$ при $1 \leqslant p < \infty$ будем обозначать пространство измеримых вектор-функций $u(t)=(u^1(t), \ldots, u^r(t))$ $(t_0 \leqslant t \leqslant T)$, для которых функция $|u(t)|^p$ суммируема на $[t_0, T]$ в смысле Лебега, и,

следовательно, имеет смысл норма

$$\|u\|_{L_p} = \left(\int_{t_0}^{T} |u(t)|^p\,dt\right)^{1/p}, \quad 1 \leqslant p < \infty.$$

Если $p = \infty$, то под $L_\infty^r[t_0, T]$ понимается пространство ограниченных измеримых вектор-функций $u(t) = (u^1(t), \ldots, u^r(t))$ $(t_0 \leqslant t \leqslant T)$, с нормой

$$\|u\|_{L_\infty} = \operatorname{ess}\sup_{t_0 \leqslant t \leqslant T} |u(t)| = \inf_{v(t)} \sup_{t_0 \leqslant t \leqslant T} |v(t)|,$$

где $v(t)$ пробегает множество всех измеримых функций, совпадающих с $u(t)$ почти всюду на отрезке $[t_0, T]$. Если читатель недостаточно знаком с интегралом Лебега и пространствами $L_p^r[t_0, T]$ [179], то всюду в этой главе он может считать, что все рассматриваемые управления $u(t)$ суть кусочно-непрерывные функции.

Далее, как видно из примеров 1, 2, значения управлений не могут быть совершенно произвольными и подчиняются некоторым ограничениям. Такие ограничения можно описать условием

$$u(t) \in V(t), \quad t_0 \leqslant t \leqslant T, \tag{10}$$

где $V(t)$ — заданное множество из E^r при каждом $t \in [t_0, T]$. Например, в случае ограничений (3) $V(t) = \{u\colon u \in E^1, |u| \leqslant \gamma\}$ при всех t. Для кусочно-непрерывных управлений выполнение условий (10) требуется для всех $t \in [t_0, T]$, а для измеримых управлений — почти всюду на $[t_0, T]$.

Кроме ограничений (10), возможны также и ограничения вида

$$\|u\|_{L_p}^p = \int_{t_0}^{T} |u(t)|^p\,dt \leqslant R^p, \quad R = \text{const} > 0, \tag{11}$$

при некотором p $(1 \leqslant p < \infty)$.

Таким образом, постановка задачи оптимального управления прежде всего предполагает, что выбран некоторый класс функций — управлений (например, кусочно-непрерывные, кусочно-гладкие или функции из $L_p^r[t_0, T]$ $(1 \leqslant p \leqslant \infty)$) и указаны налагаемые на них ограничения (например, ограничения вида (10) или (11)).

Заметим, что в учебной литературе символом $z(t) = (z^1(t), \ldots, z^m(t))$ часто обозначают как значение функции в точке t, так и саму функцию, которая представляет собой отображение области определения функции в пространстве E^m, ставящее в соответствие каждой точке t из области определения некоторую точку из E^m. Отдавая дань традициям, мы будем продолжать пользоваться этим не вполне определенным символом

в тех случаях, когда из контекста нетрудно понять, идет ли речь о функции в целом или о ее значении в конкретной точке. В тех случаях, когда обозначение $z(t)$ может привести к недоразумениям, за значением функции в точке t будем сохранять обозначение $z(t)$, а саму функцию будем обозначать через $z(\cdot)$ или просто z. В свете этих обозначений подчеркнем, что ограничения (10) являются ограничениями на значения функции, и поэтому было бы бессмысленно вместо (10) писать $u(\cdot) \in V(t)$. Ограничение (11), наоборот, накладывается на всю функцию $u(\cdot)$ в целом и не является ограничением на значения функции — функция $u(\cdot)$, удовлетворяющая этому ограничению, в отдельных точках или промежутках малой длины может принимать произвольные значения. Поэтому (11) можно записать в виде $\|u(\cdot)\|_{L_p} \leqslant R$, а обозначение $\|u(t)\|_{L_p} \leqslant R$, иногда встречающееся в литературе, не вполне удачное.

3. Итак, пусть заданы точка $x_0 \in E^n$ и некоторое кусочно-непрерывное управление $u = u(\cdot) = u(t)$ $(t_0 \leqslant t \leqslant T)$, или управление $u(\cdot) \in L_p^r[t_0, T]$ при некотором $p \geqslant 1$. Рассмотрим задачу Коши (8), (9). Сразу же возникает вопрос: что понимать под решением этой задачи? Если функции $u(t)$, $f(x, u, t)$ непрерывны, то, как обычно принято в учебниках по дифференциальным уравнениям [172, 251, 295], под решением задачи (8), (9) можно понимать функцию $x = x(\cdot) = x(t)$, $t_0 \leqslant t \leqslant T$, которая непрерывно дифференцируема на отрезке $[t_0, T]$ и удовлетворяет условиям (8), (9). Однако для случая кусочно-непрерывных или измеримых управлений, как видно из (8), требовать существование непрерывно дифференцируемого решения задачи (8), (9), вообще говоря, не имеет смысла. Поэтому мы будем пользоваться следующим более общим определением решения задачи (8), (9).

Определение 1. Непрерывную функцию $x = x(\cdot) = x(t)$, $t_0 \leqslant t \leqslant T$, удовлетворяющую равенству

$$x(t) = \int_{t_0}^{t} f(x(\tau), u(\tau), \tau)\, d\tau + x_0, \quad t_0 \leqslant t \leqslant T, \tag{12}$$

будем называть *решением* или *траекторией задачи* (8), (9), соответствующей начальному условию x_0 и управлению $u = u(\cdot)$, и будем обозначать через $x = x(\cdot, u, x_0) = x(\cdot, u(\cdot), x_0)$ или $x = x(t, u, x_0)$ $(t_0 \leqslant t \leqslant T)$. Начальную точку $x(t_0, u, x_0)$ будем называть *левым концом траектории* $x(\cdot, u, x_0)$, t_0 — *начальным моментом*, $x(T, u, x_0)$ — *правым концом траектории*, T — *конечным моментом*.

В тех случаях, когда ясно, какому именно управлению $u(\cdot)$ или начальному условию x_0 соответствует траектория, в обозначении $x(\cdot, u, x_0)$ букву u или x_0 будем опускать и просто писать $x(\cdot, u)$ или $x(\cdot, x_0)$ или $x = x(\cdot) = x(t)$ $(t_0 \leqslant t \leqslant T)$.

Классические теоремы существования и единственности решения задачи Коши

$$\dot{x} = g(x, t), \quad x(t_0) = x_0,$$

в учебниках по дифференциальным уравнениям обычно доказываются при требовании непрерывности $g(x, t)$ и $g_x(x, t)$ по совокупности переменных в некоторой области, содержащей точку (x_0, t_0) (условие непрерывности $g_x(x, t)$ часто заменяется условием Липшица $g(x, t)$ по переменной x) [172, 251, 295]. Однако если $u(t) \in L_p^1[t_0, T]$ $(p \geqslant 1)$, то непрерывности $g(x, t) \equiv f(x, u(t), t)$ по переменной t ожидать не приходится и классические теоремы существования и единственности решения здесь становятся недостаточными. Тем не менее, используя ту же технику доказательства упомянутых классических теорем, можно получить существование и единственность решения задачи (8), (9) и для кусочно-непрерывных управлений $u(t)$ или $u(t) \in L_p^r[t_0, T]$ $(p \geqslant 1)$. Мы здесь ограничимся доказательством следующей теоремы.

Теорема 1. *Пусть функция $f(x, u, t)$ определена и непрерывна по совокупности переменных при всех $(x, u, t) \in E^n \times E^r \times [t_0, T]$ и пусть*

$$|f(x, u, t) - f(y, u, t)| \leqslant L(t)|x - y| \tag{13}$$

при всех (x, u, t), $(y, u, t) \in E^n \times E^r \times [t_0, T]$, где $L(t)$ — неотрицательная функция, принадлежащая $L_1(t_0, T]$. Тогда для любого ограниченного измеримого управления $u(t)$ (т. е. $u(t) \in L_\infty^r[t_0, T]$) и начального условия x_0 задача (8), (9) *имеет, и притом единственное, решение $x = x(t)$, определенное на всем отрезке $[t_0, T]$. Это решение имеет производную $\dot{x}(t)$ почти всюду на $[t_0, T]$, $\dot{x}(t) \in L_\infty^n[t_0, T]$ и удовлетворяет уравнению* (8) *при почти всех $t \in [t_0, T]$.*

Доказательство. Пространство непрерывных вектор-функций $x(t) = (x^1(t), \ldots, x^n(t))$ $(t_0 \leqslant t \leqslant T)$ с нормой

$$\|x\|_C = \max_{t_0 \leqslant t \leqslant T} |x(t)|$$

обозначим через $C^n[t_0, T]$. Как известно [179], $C^n[t_0, T]$ — полное нормированное пространство. Зафиксируем какие-либо точки x_0 и ограниченное измеримое управление $u = u(t)$, $t_0 \leqslant t \leqslant T$. Можно показать [2], что тогда для любой функции $x(t) \in C^n[t_0, T]$ функция $f(x(t), u(t), t)$ будет ограниченной измеримой функцией переменной t на отрезке $[t_0, T]$. Определим отображение A:

$$z(t) = Ax = \int_{t_0}^{t} f(x(\tau), u(\tau), \tau)\, d\tau + x_0, \quad t_0 \leqslant t \leqslant T, \tag{14}$$

действующее из $C^n[t_0, T]$ в $C^n[t_0, T]$. Значение функции $z(\cdot) = Ax(\cdot)$ в точке t будем обозначать через $Ax(\cdot)(t)$. Покажем, что отображение A^m — m-я степень отображения A — при достаточно большом m будет сжимающим [179]. Для этого с помощью индукции докажем, что для любых $x(\cdot)$, $y(\cdot) \in C^n[t_0, T]$

$$|A^m x(\cdot)(t) - A^m y(\cdot)(t)| \leqslant \frac{1}{m!} \max_{t_0 \leqslant \tau \leqslant t} |x(\tau) - y(\tau)| \left(\int_{t_0}^{t} L(\tau)\, d\tau \right)^m \tag{15}$$

при всех t, $t_0 \leqslant t \leqslant T$ $(m = 1, 2, \ldots)$. Из (14) с учетом условия (13) имеем

$$|Ax(\cdot)(t) - Ay(\cdot)(t)| = \left| \int_{t_0}^{t} [f(x(\tau), u(\tau), \tau) - f(y(\tau), u(\tau), \tau)]\, d\tau \right| \leqslant$$

$$\leqslant \int_{t_0}^{t} L(\tau) |x(\tau) - y(\tau)|\, d\tau \leqslant \max_{t_0 \leqslant \tau \leqslant t} |x(\tau) - y(\tau)| \int_{t_0}^{t} L(\tau)\, d\tau. \tag{16}$$

Оценка (15) при $m = 1$ доказана. Пусть оценка (15) верна для некоторого $m \geqslant 1$. Тогда с помощью неравенства (16) получим

$$|A^{m+1} x(\cdot)(t) - A^{m+1} y(\cdot)(t)| = |A(A^m x(\cdot))(t) - A(A^m y(\cdot))(t)| \leqslant$$

$$\leqslant \int_{t_0}^{t} L(\tau) |A^m x(\cdot)(\tau) - A^m y(\cdot)(\tau)|\, d\tau \leqslant$$

$$\leqslant \int_{t_0}^{t} L(\tau) \frac{1}{m!} \max_{t_0 \leqslant \xi \leqslant \tau} |x(\xi) - y(\xi)| \left(\int_{t_0}^{\tau} L(\xi)\, d\xi \right)^m d\tau \leqslant$$

$$\leqslant \frac{1}{m!} \max_{t_0 \leqslant \xi \leqslant t} |x(\xi) - y(\xi)| \int_{t_0}^{t} L(\tau) \left(\int_{t_0}^{\tau} L(\xi)\, d\xi \right)^m d\tau =$$

$$= \frac{1}{(m+1)!} \max_{t_0 \leqslant \tau \leqslant t} |x(\tau) - y(\tau)| \int_{t_0}^{t} \frac{d}{d\tau} \left(\int_{t_0}^{\tau} L(\xi)\, d\xi \right)^{m+1} d\tau =$$

$$= \frac{1}{(m+1)!} \max_{t_0 \leqslant \tau \leqslant t} |x(\tau) - y(\tau)| \left(\int_{t_0}^{t} L(\tau)\, d\tau \right)^{m+1}$$

для любых $t \in [t_0, T]$. Оценка (15) доказана. Из этой оценки

следует, что

$$\|A^m x(\cdot) - A^m y(\cdot)\|_C \leqslant \frac{1}{m!}\left(\int_{t_0}^{T} L(\tau)\,d\tau\right)^m \|x(\cdot) - y(\cdot)\|_C.$$

Так как $\lim\limits_{m\to\infty} \frac{1}{m!}\left(\int_{t_0}^{T} L(\tau)\,d\tau\right)^m = 0$, то при достаточно большом m будем иметь $\frac{1}{m!}\left(\int_{t_0}^{T} L(\tau)\,d\tau\right)^m < 1$. Таким образом, отображение A^m является сжимающим. Из принципа сжимающих отображений ([179, с. 82]) следует существование единственной функции $x(\cdot) \in C^n[t_0, T]$, для которой $x(\cdot) = Ax(\cdot)$, что равносильно выполнению равенства (12).

Из свойств интеграла Лебега с переменным верхним пределом (см. [179, с. 344]) и из (12) следует, что получившаяся функция $x(t)$, $t_0 \leqslant t \leqslant T$, абсолютно непрерывна, ее производная $\dot{x}(t) \in L_\infty^n[t_0, T]$, и уравнение (8) удовлетворяется почти всюду на $[t_0, T]$. Теорема 1 доказана.

Замечание 1. Вместо условия (13) можно потребовать непрерывность $\frac{\partial f}{\partial x} = \left\{\frac{\partial f^i(x, u, t)}{\partial x^j}\right\}$ $(i, j = 1, \ldots, n)$ при $(x, u, t) \in E^n \times E^r \times [t_0, T]$, однако в этом случае существование решения задачи (8), (9) можно гарантировать, вообще говоря, лишь на отрезке $[t_0, t_0 + \alpha]$, где α — достаточно малое число.

Замечание 2. Если управление $u = u(t) \in L_p^r[t_0, T]$ $(1 \leqslant p < \infty)$, то теорема 1 и ее доказательство останутся в силе, если, например, дополнительно потребовать

$$|f(x, u, t)| \leqslant C_0(|x| + |u|^p) + C_1(t) \tag{17}$$

для всех $(x, u, t) \in E^n \times E^r \times [t_0, T]$, где $C_0 = \text{const} \geqslant 0$, $C_1(t) \geqslant 0$, $C_1(t) \in L_1[t_0, T)$. Условие (17) нужно для обеспечения включения $f(x, u(t), t) \in L_1[t_0, T]$ для любых $x(\cdot) \in C^n[t_0, T]$, $u(\cdot) \in L_p^r[t_0, T]$, чтобы отображение (14) имело смысл.

Более тонкие теоремы существования и единственности решения задачи (8), (9) для управлений $u(t) \in L_p^r[t_0, T]$ $(1 \leqslant p \leqslant \infty)$ можно найти в [2, 77, 104, 166, 199].

Остановимся еще на случае линейной системы, когда вместо (8) имеет место

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad t_0 \leqslant t \leqslant T, \tag{18}$$

где $A(t) = \{a_{ij}(t)\}$, $B(t) = \{b_{ij}(t)\}$ — матрицы порядков $n \times n$ и $n \times r$ соответственно, $f(t) = (f^1(t), \ldots, f^n(t))$ $(t_0 \leqslant t \leqslant T)$.

Теорема 2. *Пусть элементы* $\{a_{ij}(t)\}$, $\{b_{ij}(t)\}$ *матриц* $A(t)$, $B(t)$ *принадлежат* $L_\infty[t_0, T]$, *а* $f(t) \in L_1^n[t_0, T]$. *Тогда для каждого управления* $u = u(t) \in L_p^r[t_0, T]$, *где* p — *какое-либо фиксированное число*, $1 \leqslant p \leqslant \infty$, *и любой точки* $x_0 \in E^n$ *задача Коши для системы* (18) *с начальным условием* $x(t_0) = x_0$ *имеет, и притом единственное, решение* $x = x(t)$ *в смысле определения* 1, *определенное на всем отрезке* $[t_0, T]$. *Это решение имеет производную* $\dot{x}(t)$ *почти всюду на* $[t_0, T]$, $x(t) \in L_1^n[t_0, T]$ *и удовлетворяет уравнению* (18) *почти всюду на* $[t_0, T]$. *Если кроме перечисленных условий, еще имеет место включение* $f(t) \in L_p^n[t_0, T]$, *то* $\dot{x}(t) \in L_p^n[t_0, T]$.

Доказательство. Нетрудно видеть, что правая часть уравнения (18) удовлетворяет условию (13) с $L(t) = \|A(t)\| \in L_\infty[t_0, T]$. Кроме того, $g(t) \equiv A(t)x(t) + B(t)u(t) + f(t) \in L_1^n[t_0, T]$ для любых $x(t) \in C^n[t_0, T]$, $u(t) \in L_p^r[t_0, T]$. Дальнейшее доказательство проводится так же, как доказательство теоремы 1.

4. Вернемся к постановке задачи оптимального управления. Как видно из примеров 1, 2, не только на управляющие параметры объекта, но и на его фазовые координаты могут накладываться некоторые дополнительные ограничения, которые не вытекают из свойств системы (8) и ограничений на управления. Такие ограничения можно описать условием

$$x(t) = x(t, u(\cdot), x_0) \in G(t), \quad t_0 \leqslant t \leqslant T, \tag{19}$$

где $G(t)$ — некоторое заданное множество из E^n при каждом $t \in [t_0, T]$. Ограничения (19) часто называют *фазовыми ограничениями*.

Далее, начальный и конечный моменты времени t_0 и T, $t_0 \leqslant T$, характеризующие продолжительность движения объекта, могут зависеть от управления (например, в задачах быстродействия) и не всегда могут быть заданы заранее. В таких случаях обычно указывают ограничения

$$t_0 \in \Theta_0, \quad T \in \Theta_1, \tag{20}$$

где Θ_0, Θ_1 — заданные множества на числовой оси $\mathbf{R} = \{t\colon -\infty < t < +\infty\}$ (не исключается возможность, что $\Theta_0 = \mathbf{R}$ или $\Theta_1 = \mathbf{R}$).

Наконец, остановимся на условиях, которым должны удовлетворять левый и правый концы траектории. Собственно говоря, из включений (19) при $t = t_0$ и $t = T$ уже следуют условия $x(t_0) \in G(t_0)$, $x(T) \in G(T)$, и в некоторых случаях нет необходимости как-то еще иначе выделять ограничения на концы траектории. Однако могут возникнуть ситуации (например, при $G(t) = E^n$ $(t_0 < t < T)$), когда такие ограничения удобнее выделить и рассматривать самостоятельно. В таких случаях будем считать, что в E^n при каждом $t_0 \in \Theta_0$ задано множество $S_0(t_0)$

и при каждом $T \in \Theta_1$ — множество $S_1(T)$, и условия на концах траекторий будем записывать в виде

$$x(t_0) \in S_0(t_0), \quad t_0 \in \Theta_0; \quad x(T) \in S_1(T), \quad T \in \Theta_1. \tag{21}$$

В задачах оптимального управления принята следующая классификация условий (20), (21). Если множество Θ_0 состоит из единственной точки t_0, то *начальный момент называют закрепленным*; если Θ_1 состоит из одной точки T, то *конечный момент называют закрепленным*. Если множество $S_0(t_0)$ [или $S_1(T)$] состоит из одной точки и не зависит от t_0, т. е. $S_0(t_0) = \{x_0\}$, $t_0 \in \Theta_0$ [или соответственно $S_1(T) = \{x_1\}$, $T \in \Theta_1$], то говорят, что *левый* [*правый*] *конец траектории закреплен*. Если $S_0(t_0) \equiv E^n$, $t_0 \in \Theta_0$ [или $S_1(T) \equiv E^n$, $T \in \Theta_1$], то *левый* [*правый*] *конец траектории называют свободным*. В остальных случаях *левый* [*соответственно правый*] *конец траектории называют подвижным*. Примером того, как могут задаваться множества $S_0(t_0)$, является

$$S_0(t_0) = \{y\colon\ y \in G(t_0), \quad h_i(y, t_0) \leqslant 0, \quad i = 1, \ldots, m_0,$$
$$h_i(y, t_0) = 0, \quad i = m_0 + 1, \ldots, s_0\}, \tag{22}$$

где функции $h_i(y, t)$ $(i = 1, \ldots, s_0)$, определены при $y \in G(t)$, $t \in \Theta_0$. Аналогично, примером множества $S_1(T)$ служит

$$S_1(T) = \{x\colon\ x \in G(T),\ g_i(x, T) \leqslant 0, \quad i = 1, \ldots, m_1,$$
$$g_i(x, T) = 0, \quad i = m_1 + 1, \ldots, s_1\}, \tag{23}$$

где функции $g_i(x, t)$ $(i = 1, \ldots, s_1)$, определены при $x \in G(t)$, $t \in \Theta_1$.

В приложениях нередко возникают также задачи, в которых левый и правый концы траектории должны выбираться согласованно, в зависимости друг от друга. Это требование можно записать в виде

$$(x(t_0), x(T)) \in S(t_0, T), \quad t_0 \in \Theta_0, \quad T \in \Theta_1, \tag{24}$$

где $S(t_0, T)$ при каждом $(t_0, T) \in \Theta_0 \times \Theta_1$ представляет собой заданное множество из $E^n \times E^n$. Примером такого множества является

$$S(t_0, T) = \{(x, y) \in E^n \times E^n\colon\ g_i(x, y, t_0, T) \leqslant 0, \quad i = 1, \ldots, m,$$
$$g_i(x, y, t_0, T) = 0, \quad i = m + 1, \ldots, s\}, \tag{25}$$

где $g_i(x, y, t, T)$ — заданные функции переменных $(x, y, t, T) \in E^n \times E^n \times \Theta_0 \times \Theta_1$. Понятно, что множества (21) являются частным случаем множества (24), когда $S(t_0, T) = S_0(t_0) \times S_1(T)$; множества (22), (23) — частный случай (25).

5. Теперь перейдем к непосредственной формулировке задачи оптимального управления. Пусть заданы множества Θ_0, Θ_1 на числовой оси $\mathbf{R}$, $\inf \Theta_0 < \sup \Theta_1$; $V(t) \subseteq E^r$, $G(t) \subset E^n$ при всех t,

$\inf \Theta_0 \leqslant t \leqslant \sup \Theta_1$; $S(t_0, T)$, $t_0 \in \Theta_0$, $T \in \Theta_1$. Пусть движение фазовой точки $x = (x^1, \ldots, x^n)$ описывается системой обыкновенных дифференциальных уравнений (7), где функция $f(x, u, t)$ определена при $x \in G(t)$, $u \in V(t)$, $t \in [t_0, T]$.

Набор $(t_0, T, x_0, u(\cdot), x(\cdot))$ назовем *допустимым*, если $t_0 \in \Theta_0$, $T \in \Theta_1$, $t_0 \leqslant T$, управление $u = u(\cdot) = (u^1(t), \ldots, u^r(t))$ определено и кусочно-непрерывно на отрезке $[t_0, T]$ и удовлетворяет ограничению (10) на этом отрезке, а $x = x(\cdot) = x(\cdot, u(\cdot), x_0)$ — траектория задачи (8), (9) (см. определение 1), которая определена на отрезке $[t_0, T]$ и удовлетворяет фазовому ограничению (19), $(x(t_0) = x_0, x(T)) \in S(t_0, T)$. Будем предполагать, что множество допустимых наборов непусто.

Пусть на множестве допустимых наборов задана функция (или, как часто говорят, *целевая функция* или *функционал*)

$$J(x_0, u(\cdot), x(\cdot), t_0, T) = \\ = \int_{t_0}^{T} f^0(x(t), u(t), t)\, dt + g_0(x_0, x(T), t_0, T), \tag{26}$$

где $f^0(x, u, t)$, $g_0(x, y, t, T)$ — заданные функции при $x \in G(t)$, $u \in V(t)$, $\inf \Theta_0 \leqslant t \leqslant \sup \Theta_1$, $T \in \Theta_1$.

Задача оптимального управления заключается в том, чтобы минимизировать или максимизировать функцию (26) *на множестве допустимых наборов.* Мы ограничимся рассмотрением лишь задач минимизации, так как задача минимизации J всегда может быть сведена к эквивалентной задаче минимизации $(-J)$.

Обозначим

$$J_* = \inf J(x_0), u(\cdot), x(\cdot), t_0, T),$$

где нижняя грань берется по всем допустимым наборам. Допустимый набор $(x_{0*}, u_*(\cdot), x_*(\cdot), t_{0*}, T_*)$ назовем *решением задачи оптимального управления*, $u_*(\cdot)$ — *оптимальным управлением*, $x_*(\cdot)$ — *оптимальной траекторией*, если $J(x_{0*}, u_*(\cdot), x_*(\cdot), t_{0*}, T_*) = J_*$.

Сформулированную задачу оптимального управления можно записать в следующем кратком виде:

$$J(x_0, u(\cdot), x(\cdot), t_0, T) = \\ = \int_{t_0}^{T} f^0(x(t), u(t), t)\, dt + g_0(x_0, x(T), t_0, T) \to \inf, \tag{27}$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T, \tag{28}$$

$$x(t) \in G(t), \quad t_0 \leqslant t \leqslant T, \tag{29}$$

$$x(t_0) = x_0, \quad x(T) \in S(t_0, T), \quad t_0 \in \Theta_0, \quad T \in \Theta_1, \tag{30}$$

$$u(t) \in V(t), \quad t_0 \leqslant t \leqslant T, \tag{31}$$

подразумевая (если не оговорено другое), что управление $u=u(\cdot)$ — кусочно-непрерывно на отрезке $[t_0, T]$.

Если $f^0\equiv 1$, $g_0\equiv 0$, то $J(t_0, T, x_0, u(\cdot), x(\cdot))\equiv T-t_0$ — в этом случае задачу (27)—(31) называют *задачей быстродействия*. Если $f^i(x, u, t)\equiv f^i(x, u)$ $(i=0, 1, \ldots, n)$, $g_0(x, y, t, T)\equiv g_0(x, y)$, множества $S(t_0, T)$, $V(t)$, $G(t)$ не зависят от времени, то задачу (27)—(31) называют *автономной* или *стационарной*.

Если начальный момент закреплен, т. е. $\Theta_0=\{t_0\}$, то в формулировке задачи (27)—(31) включение $t_0\in\Theta_0$ опускают, вместо $J(x_0, u(\cdot), x(\cdot), t_0, T)$ пишут короче: $J(x_0, u(\cdot), x(\cdot), T)$ или $J(x_0, u(\cdot), T)$, вместо $S(t_0, T)$ пишут $S(T)$. Аналогично поступают, если закреплены конечный момент T или один из концов траектории. В том случае, когда $G(t)\equiv E^n$ при всех t или $S(t_0, T)=S_0(t_0)\times S_1(T)$, где $S_0(t)\equiv E^n$, или $S_1(t)\equiv E^n$, или $V(t)\equiv E^r$, то соответствующие из условий (29), (30), (31) в постановке задачи (27)—(31) также явно не указывают. Если $S(t_0, T)=S_0(t_0)\times S_1(T)$, где $S_0(t_0)\equiv G(t_0)$ или $S_1(T)\equiv G(T)$, то включение $x(t_0)\in S_0(t_0)$ или соответственно $x(T)\in S_1(T)$ будут учтены в условии (29), поэтому в (30) эти включения можно опустить.

В приложениях встречаются задачи оптимального управления более общего вида, чем задача (27)—(31). Возможны ситуации, когда наряду с ограничениями на управления и фазовые координаты, записанными, так сказать, в разделенном виде (29), (31), имеются более сложные ограничения вида

$$(u(t), x(t))\in W(t), \quad t_0\leqslant t\leqslant T, \tag{32}$$

где $W(t)$ — заданные множества из $E^r\times E^n$. Ограничения (29), (31), (32) накладываются на значения функций $u(t)$, $x(t)$ в каждой точке t, поэтому их можно назвать *точечными ограничениями*. Наряду с точечными ограничениями возможны также ограничения вида

$$\begin{aligned} &J_i(x_0, u(\cdot), x(\cdot), t_0, T)\leqslant 0, \quad && i=1, \ldots, m_2,\\ &J_i(x_0, u(\cdot), x(\cdot), t_0, T)=0, \quad && i=m+1, \ldots, s_2, \end{aligned} \tag{33}$$

где

$$J_i(x_0, u(\cdot), x(\cdot), t_0, T)=\int_{t_0}^{T} f_i^0(x(t), u(t), t)\,dt+g_i^0(x_0, x(T), t_0, T),$$

$f_i^0(x, u, t)$, $g_i^0(x, y, t, T)$ $(i=1, \ldots, s_2)$ — заданные функции. Ограничения (33) накладываются на функции $u(\cdot)$, $x(\cdot)=x(\cdot, u(x), x_0)$ в целом, поэтому их, в отличие от точечных, можно назвать *интегральными ограничениями*. Кстати, заметим, что ограничения вида (11) относятся к интегральным.

В теории оптимального управления рассматриваются также задачи, учитывающие запаздывание информации, задачи с пара-

метрами, с дискретным временем, с более общим видом целевой функции, задачи для интегро-дифференциальных уравнений, для уравнений с частными производными, для стохастических уравнений и др. С некоторыми из этих задач мы познакомимся ниже.

Важнейшим обобщением задач оптимального управления являются дифференциальные игры, описывающие конфликтно-управляемые системы, управляемые системы в условиях неопределенности. При исследовании управляемых систем наряду с задачами оптимизации описанного выше типа рассматриваются также и другие важные проблемы, такие, как управляемость, наблюдаемость, инвариантность, чувствительность, устойчивость, стабилизация, идентификация, фильтрация и т. д. У нас здесь нет возможности хотя бы бегло остановиться на перечисленных аспектах теории управляемых систем, и по этим вопросам мы отсылаем читателя к литературе, упомянутой во введении к настоящей главе.

§ 2. Формулировка принципа максимума. Примеры

1. Начнем с рассмотрения задачи оптимального управления с закрепленным временем. А именно, пусть требуется минимизировать функцию

$$J(x_0, u(\cdot), x(\cdot)) = \int_{t_0}^{T} f^0(x(t), u(t), t)\,dt + g^0(x_0, x(T)) \tag{1}$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T; \quad x(t_0) = x_0, \tag{2}$$

$$\begin{gathered} g^i(x_0, x(T)) \leqslant 0, \quad i = 1, \ldots, m, \\ g^i(x_0, x(T)) = 0, \quad i = m+1, \ldots, s, \end{gathered} \tag{3}$$

$$u = u(t) \in V, \quad t_0 \leqslant t \leqslant T, \tag{4}$$

где моменты t_0, T предполагаются заданными, $u = (u^1, \ldots, u^r)$, $x = (x^1, \ldots, x^n)$, $f = (f^1, \ldots, f^n)$; управление $u = u(\cdot)$ является кусочно-непрерывной функцией на отрезке $[t_0, T]$; $f^i(x, u, t)$, $g^i(x, y)$ — заданные функции. Подчеркнем, что в задаче (1)—(4) множество $V \subset E^r$ не зависит от времени и фазовые ограничения при $t_0 < t < T$ отсутствуют. Очевидно, задача (1)—(4) является частным случаем задачи (1.27)—(1.31). В (3) не исключаются возможности, когда отсутствуют ограничения типа неравенств ($m = 0$), типа равенств ($s = m \geqslant 1$) или все ограничения (3) ($s = m = 0$). Предполагается, что функции $f^j(x, u, t)$, $g^j(x, y)$ имеют частные производные $\partial f^j/\partial x^i = f^j_{x^i}$, $\partial g^j/\partial x^i = g^j_{x^i}$, $\partial g^j/\partial y^i = g$ ($i = 1, \ldots, n$). Обозначим $f^j_x = \left(f^j_{x^1}, \ldots, f^j_{x^n}\right)$, $g^j_x = \left(g^j_{x^1}, \ldots, g^j_{x^n}\right)$, $g^j_y = \left(g^j_{y^1}, \ldots, g^j_{y^n}\right)$.

Для формулировки принципа максимума введем функцию

$$H(x, u, t, \psi, a_0) = \\ = -a_0 f^0(x, u, t) + \psi_1 f^1(x, u, t) + \ldots + \psi_n f^n(x, u, t) = \\ = -a_0 f^0(x, u, t) + \langle \psi, f(x, u, t) \rangle, \quad (5)$$

называемую функцией Гамильтона — Понтрягина; здесь $\psi = (\psi_1, \ldots, \psi_n)$, a_0 — вспомогательные переменные, определяемые ниже.

Пусть $u = u(t)$ — кусочно-непрерывное управление на отрезке $[t_0, T]$, $x(t) = x(t, u, x_0)$ — решение задачи (2), соответствующее этому управлению $u = u(\cdot)$ и начальному условию x_0 и определенное на всем отрезке $[t_0, T]$. Паре $(u(t), x(t))$ $(t_0 \leqslant t \leqslant T)$ поставим в соответствие следующую систему линейных дифференциальных уравнений относительно переменных $\psi = \psi(t) = (\psi_1(t), \ldots, \psi_n(t))$:

$$\dot{\psi}_i(t) = -\left.\frac{\partial H(x, u, t, \psi(t), a_0)}{\partial x^i}\right|_{u=u(t),\, x=x(t)} = \\ = a_0 f^0_{x^i}(x(t), u(t), t) - \sum_{j=1}^{n} \psi_j(t) f^j_{x^i}(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T, \quad (6)$$

называемую сопряженной системой. Систему (6) можно записать в векторной форме

$$\dot{\psi}(t) = -H_x(x(t), u(t), t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T, \quad (7)$$

где $H_x = (H_{x^1}, \ldots, H_{x^n})$. Подчеркнем, что в (6), (7) и всюду ниже запись вида $H_{x^i}(x(t), u(t), t, \psi(t), a_0)$, $g^j_{x^i}(x_0, x(T))$, $g^j_{y^i}(x_0, x(T))$, как это обычно принято, означает, что сначала вычисляется соответствующая частная производная функций $H(x, u, t, \psi, a_0)$, $g^j(x, y)$ и затем вместо аргументов подставляются их конкретные значения.

Если система (2) линейна относительно x, u, т. е.

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad t_0 \leqslant t \leqslant T$$

(см. обозначения в (1.18)), то $H(x, u, t, \psi, a_0) = -a_0 f^0(x, u, t) + \langle \psi, A(t)x + B(t)u + f(t) \rangle$ и сопряженная система (7) может быть записана в виде

$$\dot{\psi}(t) = a_0 f^0_x(x(t), u(t), t) - (A(t))^T \psi(t), \quad t_0 \leqslant t \leqslant T,$$

где $(A(t))^T$ — матрица, полученная транспонированием матрицы $A(t)$.

Из теоремы 1.2 следует, что если зафиксировать постоянную a_0, момент времени t_1, $t_0 \leqslant t_1 \leqslant T$, и точку $\psi_0 \in E^n$, то линейная

система (7) будет иметь и притом единственное решение $\psi(t) = \psi(t, u, x_0, t_1, \psi_0)$, удовлетворяющее условию $\psi(t_1) = \psi_0$ и определенное на всем отрезке $[t_0, T]$.

Теперь можем перейти к формулировке теоремы, выражающей необходимое условие оптимальности — принцип максимума для задачи (1) — (4).

Теорема 1. *Пусть функции* $f^j(x, u, t)$ $(j = 0, 1, \ldots, n)$, $g^j(x, y)$ $(j = 0, 1, \ldots, s)$ *имеют частные производные* $f^j_{x^i}$, $g^j_{x^i}$, $g^j_{y^i}$ $(i = 1, \ldots, n)$ *и непрерывны вместе с этими производными по совокупности своих аргументов при* $x \in E^n$, $y \in E^n$, $u \in V$, $t \in [t_0, T]$. *Пусть* $(x_0, u(t), x(t))$ $(t_0 \leqslant t \leqslant T)$ — *решение задачи* (1), (4). *Тогда необходимо существуют числа* $a_0, a_1, \ldots, a_s$ *и вектор-функция* $\psi(t) = (\psi_1(t), \ldots, \psi_n(t))$, $t_0 \leqslant t \leqslant T$, *такие, что*

1) $a = (a_0, a_1, \ldots, a_s) \neq 0$, $a_0 \geqslant 0$, $a_1 \geqslant 0, \ldots, a_m \geqslant 0$; (8)

2) $\psi(t)$ *является решением сопряженной системы* (6), *соответствующей рассматриваемому решению* $(x_0, u(\cdot), x(\cdot))$;

3) *для всех* $t \in [t_0, T]$, *являющихся точками непрерывности оптимального управления* $u(\cdot)$, *функция* $H(x(t), u, t, \psi(t), a_0)$ *переменной* $u = (u^1, \ldots, u^r)$ *достигает своей верхней грани на множестве* V *при* $u = u(t)$, *т. е.*

$$\sup_{u \in V} H(x(t), u, t, \psi(t), a_0) = H(x(t), u(t), t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T; \tag{9}$$

4) *выполнены условия*

$$\begin{aligned} \psi_i(t_0) &= \sum_{j=0}^{s} a_j g^j_{x^i}(x_0, x(T)), \\ \psi_i(T) &= -\sum_{j=0}^{s} a_j g^j_{y^i}(x_0, x(T)), \qquad i = 1, \ldots, n, \end{aligned} \tag{10}$$

$$a_j g^j(x_0, x(T)) = 0, \qquad j = 1, \ldots, m. \tag{11}$$

Условия (10) принято называть *условием трансверсальности*, условие (11) — *условием дополняющей нежесткости*. В том случае, когда управление $u(\cdot)$ является ограниченной измеримой функцией, т. е. $u(\cdot) \in L^r_\infty[t_0, T]$, то формулировка теоремы 1 полностью сохраняется, но равенство (9) (как и включение (4)) будет выполняться, вообще говоря, лишь почти всюду на отрезке $[t_0, T]$.

Центральное место в теореме 1 занимает условие максимума (9): оказывается, если $u(\cdot)$ — оптимальное управление, а $x(\cdot)$ — оптимальная траектория, то непременно найдутся такие числа $a_0, a_1, \ldots, a_s$ и такое решение $\psi(t)$ системы (6), (10), что функция $H(x(t), u, t, \psi(t), a_0)$ переменной u будет достигать своего

максимума на V именно при $u=u(t)$ во всех точках $t\in[t_0, T]$ непрерывности управления $u(\cdot)$. Поэтому теорему 1 и нижеследующую теорему 2, дающие необходимое условие оптимальности, принято называть *принципом максимума.*

2. Однако как практически пользоваться теоремой 1 для поиска решения задачи (1)—(4)? Здесь обычно поступают следующим образом. Рассматривают функцию $H(x, u, t, \psi, a_0)$ как функцию r переменных $u=(u^1, \ldots, u^r)\in V$, считая остальные переменные (x, t, ψ, a_0) параметрами, и при каждом фиксированном наборе (x, t, ψ, a_0) решают задачу максимизации:

$$H(x, u, t, \psi, a_0)\to\sup, \quad u\in V. \tag{12}$$

Отсюда находят функцию

$$u=u(x, t, \psi, a_0)\in V, \tag{13}$$

на которой достигается верхняя грань в задаче (12), т. е.

$$H(x, u(x, t, \psi, a_0), t, \psi, a_0)=\sup_{u\in V} H(x, u, t, \psi, a_0). \tag{14}$$

Если исходная задача (1)—(4) имеет решение, то, как следует из (9), функция (13) определена на непустом множестве.

В ряде случаев функция (13) может быть выписана в явном виде. Например, если

$$f^j(x, u, t)=f_0^j(x, t)+\sum_{i=1}^{r} f_{1i}^j(x, t)\,u^i, \qquad j=0, 1, \ldots, n,$$

$$V=\{u=(u^1, \ldots, u^r)\in E^r\colon \alpha_i\leqslant u^i\leqslant\beta_i, \ i=1, \ldots, r\},$$

где α_i, β_i — заданные числа, то

$$H(x, u, t, \psi, a_0)=-a_0 f_0^0(x, t)+\sum_{j=1}^{r}\psi_j f_0^j(x, t)+\sum_{i=1}^{r}\varphi_i(x, t, \psi, a_0)u^i;$$

здесь для краткости обозначено

$$\varphi_i(x, t, \psi, a_0)=-a_0 f_{1i}^0(x, t)+\sum_{j=1}^{n}\psi_j f_{1i}^j(x, t), \qquad i=1, \ldots, r.$$

Ясно, что решением задачи (12) тогда будет вектор-функция $u(x, t, \psi, a_0)$ с координатами

$$u^i=u^i(x, t, \psi, a_0)=\begin{cases}\beta_i, & \varphi_i(x, t, \psi, a_0)>0,\\ \alpha_i, & \varphi_i(x, t, \psi, a_0)<0,\end{cases} \qquad i=1, \ldots, r.$$

В частности, если $\alpha_i=-1$, $\beta_i=+1$, то $u^i=\operatorname{sign}\varphi_i(x, t, \psi, a_0)$ $(i=1, \ldots, r)$. Полученная формула дает довольно много информации о структуре оптимального управления: i-я координата оптимального управления является ступенчатой функцией со

значениями α_i или β_i, причем точки переключения определяются условием $\varphi_i(x, t, \psi, a_0)=0$. Обратим внимание читателя на возможный особый случай, когда $\varphi_i(x(t), t, \psi(t), a_0)=0$ на каком-либо промежутке $[\alpha, \beta]\subset[t_0, T]$. В этом случае функция H не будет зависеть от u^i и из условия (9) не удастся извлечь никакой полезной информации об i-й координате управления $u(\cdot)$ при $t\in[\alpha, \beta]$; некоторым источником информации об $u^i(\cdot)$ на этом особом участке $[\alpha, \beta]$ здесь может служить само равенство $\varphi_i(x(t), t, \psi(t), a_0)=0$.

Если множество V имеет вид

$$V=\left\{u\in E^r\colon |u|=\left(\sum_{i=1}^{r}(u^i)^2\right)^{1/2}\leqslant R\right\},$$

то, пользуясь известным неравенством Коши — Буняковского, также нетрудно выписать функцию (13) в явном виде:

$$u(x, t, \psi, a_0)=\frac{\varphi(x, t, \psi, a_0)}{|\varphi(x, t, \psi, a_0)|}R, \quad \varphi=(\varphi_1, \ldots, \varphi_r).$$

Ряд задач, в которых удается получить явное выражение для функции (13), приводятся ниже в примерах.

Допустим, что функция (13) нам уже известна. Тогда можем рассмотреть следующую систему из $2n$ дифференциальных уравнений

$$\begin{aligned}&\dot{x}=f(x, u(x, t, \psi, a_0), t),\\&\dot{\psi}=-H_x(x, u(x, t, \psi, a_0), t, \psi, a_0), \qquad t_0\leqslant t\leqslant T,\end{aligned} \tag{15}$$

относительно неизвестных $x(\cdot)$, $\psi(\cdot)$. Как известно [172, 251, 295] общее решение системы (15) зависит от $2n$ произвольных числовых параметров (например, такими параметрами могли бы служить начальные условия $x(t_0)$, $\psi(t_0)$) и для определения этих параметров нам нужно иметь $2n$ условий. Кроме того, параметры $a_0, a_1, \ldots, a_s$, встречающиеся в теореме 1, также неизвестны и для их определения нужно еще $s+1$ условие. Таким образом, для определения $2n+s+1$ неизвестных числовых параметров нам нужно $2n+s+1$ условие. Где их взять? Оказывается, эти условия также могут быть извлечены из теоремы 1. А именно, условия трансверсальности (10) и дополняющей нежесткости (11) нам дают $2n+m$ уравнений; еще $s-m$ уравнений

$$g^j(x_0, x(T))=0, \qquad j=m+1, \ldots, s, \tag{16}$$

вытекают из условий (3). Для получения еще одного уравнения заметим, что функция $H(x, u, t, \psi, a_0)$, определенная соотношением (5), линейна и однородна относительно переменных $\psi_1, \ldots, \psi_n, a_0$, т. е. $H(x, u, t, \alpha\psi, \alpha a_0)=\alpha H(x, u, t, \psi, a_0)$ при любых

α. Отсюда и из условия (14) тогда имеем

$$u(x, t, \alpha\psi, \alpha a_0) \equiv u(x, t, \psi, a_0) \quad \forall \alpha > 0. \tag{17}$$

Из (8), (10), (11), (17) следует, что если некоторый набор $a_0, \ldots, a_s, \psi_1, \ldots, \psi_n$ удовлетворяет условиям теоремы 1, то этим условиям удовлетворяет также набор $\alpha a_0, \ldots, \alpha a_s, \alpha\psi_1, \ldots, \alpha\psi_n$ при любых $\alpha > 0$. Это означает, что теорема 1 определяет величины $a_0, \ldots, a_s, \psi_1, \ldots, \psi_n$ лишь с точностью до положительного множителя, и этим множителем мы можем распорядиться по своему усмотрению. Например, опираясь на первое из условий (8), можно положить

$$|a|^2 = \sum_{i=0}^{s} a_i^2 = 1. \tag{18}$$

В тех задачах, в которых удается показать, что $a_0 > 0$, вместо условия нормировки (18) часто берут $a_0 = 1$.

Таким образом, для определения $2n + s + 1$ параметров — $2n$ параметров общего решения системы (5) и параметров $a_0, a_1, \ldots$ $\ldots, a_s$ — у нас имеется система $2n + s + 1$ уравнений (10), (11), (16), (18). Разумеется, эти уравнения надо решать совместно с неравенствами

$$a_0 \geqslant 0, \ldots, a_m \geqslant 0, \quad g_i(x_0, x(T)) \leqslant 0, \quad i = 1, \ldots, m. \tag{19}$$

Если исходная задача (1)—(4) имеет решение, то согласно теореме 1 система (10), (11), (16), (18), (19) также имеет решение. Попутно заметим, что для тех i $(1 \leqslant i \leqslant m)$, для которых $g_i(x_0, x(T)) < 0$ (неактивные ограничения), из (11) вытекает, что $a_i = 0$, и неопределенными остаются лишь a_i с номерами i, для которых $g_i(x_0, x(T)) = 0$ (активные ограничения). Это означает, что из уравнений (11) существенное значение имеют лишь подсистемы $g_i(x_0, x(T)) = 0$ $(i \in I)$, состоящие из активных ограничений.

Итак, основываясь на теореме 1, от исходной задачи (1)—(4) мы пришли к специальной краевой задаче, состоящей из условия максимума (14), системы дифференциальных уравнений (15) и условий (10), (11), (16), (18), (19). Такую краевую задачу естественно назвать *краевой задачей принципа максимума* для задачи оптимального управления (1)—(4).

Можно ожидать, что имеются лишь отдельные, изолированные функции $(x(t), \psi(t))$ $(t_0 \leqslant t \leqslant T)$, и значения параметров $a_0, a_1, \ldots, a_s$, удовлетворяющие условиям (10), (11), (15), (16), (18), (19). Возьмем один из таких наборов $x(t), \psi(t), a_0, a_1, \ldots$ $\ldots, a_s$, и подставим их в (13); получим функцию

$$u(t) = u(x(t), t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T. \tag{20}$$

Пусть эта функция оказалась кусочно-непрерывной на $[t_0, T]$. Из (13), (14), (20) тогда следует, что полученное таким образом управление $u(t)$ $(t_0 \leqslant t \leqslant T)$ удовлетворяет условию (9)

и, следовательно, согласно теореме 1 может претендовать на роль оптимального управления задачи (1)—(4), а функция $x(t)=x(t,\ u(\cdot),\ x(t_0))$ $(t_0\leqslant t\leqslant T)$ — на роль оптимальной траектории этой задачи. Будет ли найденная пара $(u(t),\ x(t))$ $(t_0\leqslant t\leqslant T)$ в самом деле решением задачи (1)—(4), теорема 1 не гарантирует, так как эта теорема, вообще говоря, дает лишь необходимое условие оптимальности. Более того, ниже на примерах мы увидим, что бывают случаи, когда пара $(u(t),\ x(t))$ удовлетворяет условиям теоремы 1, но не является решением задачи (1)—(4). Однако, если из каких-либо соображений известно, что задача (1)—(4) имеет решение, а краевая задача принципа максимума однозначно определяет функции $(x(t),\ \psi(t))$ и параметры $a_0,\ a_1,\ \ldots,\ a_s$, то управление (20) будет оптимальным. Если информации о существовании решения задачи (1)—(4) нет или краевая задача принципа максимума имеет несколько решений, то для выяснения вопроса об оптимальности получаемых здесь управлений требуется дополнительное и порою весьма сложное исследование.

Заметим также, что в задаче (12) верхняя грань может достигаться в нескольких точках, и тогда функция (13) будет определяться неоднозначно. В этом случае для получения всех подозрительных на оптимальность управлений нужно найти все наборы $(x(t),\ \psi(t),\ a_0,\ a_1,\ \ldots,\ a_s)$ и управления $u(t)$ $(t_0\leqslant t\leqslant T)$, удовлетворяющие условиям (10), (11), (15), (16), (18)—(20), для всех функций $u=u(x,\ t,\ \psi,\ a_0)$ из (13).

Таким образом, схема использования принципа максимума — теоремы 1 — для получения решения задачи (1)—(4) описана. Как видно, принцип максимума дает просто и изящно выписываемые необходимые условия оптимальности для задачи (1)—(4) и сводит ее к специального вида краевой задаче.

3. Посмотрим, как выглядит краевая задача принципа максимума для задач оптимального управления (1)—(4) для некоторых конкретных классов функций $g^i(x,\ y)$, соответствующих различным режимам на левом и правом концах траектории. Начнем с рассмотрения случая, когда концы траектории закреплены:

$$x(t_0)=x_0,\quad x(T)=x_1. \tag{21}$$

Если положить $g^i(x,\ y)=x^i-x_0^i$ $(i=1,\ \ldots,\ n)$, $g^i(x,\ y)=y^i-x_1^i$ $(i=n+1,\ \ldots,\ 2n)$, то условия (21) запишутся в виде (3): $g^i(x_0,\ x(T))=0$ $(i=1,\ \ldots,\ 2n=s,\ m=0)$. Поэтому условия (11) здесь отсутствуют, а условия трансверсальности (10) дадут

$$\begin{gathered}\psi(t_0)=\sum_{j=0}^{2n}a_jg_x^j(x_0,\ x(T))=a_0g_x^0(x_0,\ x(T))+(a_1,\ \ldots,\ a_n),\\ \psi(T)=-\sum_{j=0}^{2n}a_jg_y^j(x_0,\ x(T))=-a_0g_y^0(x_0,\ x(T))-(a_{n+1},\ \ldots,\ a_{2n}).\end{gathered} \tag{22}$$

Оказывается, условие $a \neq 0$ из (8) здесь может быть заменено условием

$$|a_0| + |\psi(t)| \neq 0 \qquad \forall t \in [t_0, T]. \tag{23}$$

В самом деле, если (23) не выполняется, то $a_0 = 0$, $\psi(t) \equiv 0$ $(t_0 \leqslant t \leqslant T)$. А тогда в силу (22) $\psi(t_0) = 0 = (a_1, \ldots, a_n)$, $\psi(T) = 0 = (a_{n+1}, \ldots, a_{2n})$ и, следовательно, $a = (a_0, a_1, \ldots, a_{2n}) = 0$, что противоречит (8). Это значит, что для задачи (1), (2), (4), (21) выполняется условие (23). Тогда условие нормировки (18) можно заменить условием

$$|a_0| + |\psi(t_1)| = 1, \tag{24}$$

где t_1 — какая-либо подходящая точка из отрезка $[t_0, T]$ (часто здесь берут $t_1 = t_0$ или $t_1 = T$). Таким образом, краевая задача принципа максимума для задачи (1), (2), (4), (21) состоит из системы (15), граничных условий (21), (22), неравенства $a_0 \geqslant 0$ и условия нормировки (24). Так как неизвестные параметры $a_1, \ldots, a_{2n}$ входят лишь в условие трансверсальности (22) и не входят в (15), (21), (24), то эти параметры и условия (22) можно исключить из дальнейшего рассмотрения. В итоге, для определения функций $(x(t), \psi(t))$ $(t_0 \leqslant t \leqslant T)$ и параметра $a_0 \geqslant 0$ будем иметь краевую задачу принципа максимума, состоящую из системы $2n$ дифференциальных уравнений (15), $2n$ граничных условий (21) и условия нормировки (24). Конечно, при необходимости исключенные параметры $a_1, \ldots, a_{2n}$ могут быть определены из условий (22) после того, как уже будут найдены $x(t)$, $\psi(t)$, a_0 из (15), (21), (24).

Теперь рассмотрим задачу (1), (2), (4) при условиях, когда левый конец траектории закреплен: $x(t_0) = x_0$, а правый конец свободный. Этот случай граничных режимов соответствует задаче (1)—(4), в которой $g^i(x, y) = x^i - x_0^i$ $(i = 1, \ldots, n = s, m = 0)$. Поэтому условия (11) здесь отсутствуют, а условия трансверсальности (10) запишутся в виде:

$$\begin{aligned} \psi(t_0) &= a_0 g_x^0(x_0, x(T)) + (a_1, \ldots, a_n), \\ \psi(T) &= -a_0 g_y^0(x_0, x(T)). \end{aligned} \tag{25}$$

Покажем, что в рассматриваемом случае также выполняется условие (23) и, более того, можно гарантировать, что $a_0 > 0$. В самом деле, если $a_0 = 0$, то, как видно из (25), $\psi(T) = 0$ и система однородных уравнений (6) будет иметь лишь тривиальное решение $\psi(t) \equiv 0$, а тогда $\psi(t_0) = 0 = (a_1, \ldots, a_n)$. Пришли к противоречию с первым условием (8): $a = (a_0, a_1, \ldots, a_n) \neq 0$. Следовательно, $a_0 > 0$, и можем принять условие нормировки $a_0 = 1$. Таким образом, краевая задача принципа максимума для задачи (1), (2), (4) с закрепленным левым концом и свободным правым концом состоит из системы (15), граничных условий (25),

$x(t_0)=x_0$ и условия нормировки $a_0=1$. Так как параметры $a_1, \ldots, a_n$ входят лишь в первое из условий (25), то это условие и параметры $a_1, \ldots, a_n$ можно исключить из дальнейшего рассмотрения, и в результате краевая задача принципа максимума сведется к системе $2n$ дифференциальных уравнений (15), которая решается при условиях

$$x(t_0)=x_0, \quad \psi(T)=-g_y^0(x_0, x(T)), \quad a_0=1. \tag{26}$$

Аналогичные рассуждения показывают, что если в задаче (1), (2), (4) левый конец траектории свободный, правый конец закреплен, то соответствующая краевая задача принципа максимума представляет собой систему (15), которая должна решаться при условиях

$$x(T)=x_1, \quad \psi(t_0)=g_x^0(x_0, x(T)), \quad a_0=1. \tag{27}$$

Если в задаче (1), (2), (4) оба конца траектории свободны, то краевая задача принципа максимума состоит из системы (15) и условий

$$\psi(t_0)=g_x^0(x_0, x(T)), \quad \psi(T)=-g_y^0(x_0, x(T)), \quad a_0=1. \tag{28}$$

Далее, рассмотрим задачу (1), (2), (4) в случае, когда левый конец траектории закреплен: $x(t_0)=x_0$, а правый конец подвижный и удовлетворяет условиям

$$\begin{gathered} g^i(x(T))\leqslant 0, \quad i=1, \ldots, m_1; \\ g^i(x(T))=0, \quad i=m_1+1, \ldots, s_1. \end{gathered} \tag{29}$$

Здесь мы имеем дело с задачей (1)—(4), в которой функции $g^i(x, y)$ определены так: $g^i(x, y)=g^i(y)$ $(i=1, \ldots, s_1)$; $g^i(x, y)=x^i-x_0^i (i=s_1+1, \ldots, s_1+n=s, m=m_1)$. Условия (10), (11) на правом конце траектории с учетом (8) запишутся в виде

$$\begin{gathered} \psi(T)=-a_0 g_y^0(x_0, x(T))-\sum_{j=1}^{s_1} a_j g_y^j(x(T)), \\ a_j g^j(x(T))=0, \quad a_j\geqslant 0, \quad j=1, \ldots, m_1; \quad a_0\geqslant 0, \end{gathered} \tag{30}$$

на левом конце — в виде:

$$\psi(t_0)=a_0 g_x^0(x_0, x(T))+(a_{s_1+1}, \ldots, a_{s_1+n}). \tag{31}$$

Условие $a=(a_0, a_1, \ldots, a_{s_1+n})\neq 0$ здесь может быть заменено условием $(a_0, a_1, \ldots, a_{s_1})\neq 0$. В самом деле, если бы $(a_0, a_1, \ldots, a_{s_1})=0$, то в силу (30) $\psi(T)=0$, и однородная система (6) будет иметь тривиальное решение $\psi(t)\equiv 0$; тогда $\psi(t_0)=0$ и из (31) будет следовать $(a_{s_1+1}, \ldots, a_{s_1+n})=0$, что противоречит условию $a\neq 0$. Поэтому условие нормировки (18) можно заменить на

$$a_0^2+a_1^2+\ldots+a_{s_1}^2=1, \tag{32}$$

а условие (31) и параметры $a_{s_1+1}, \ldots, a_{s_1+n}$ исключить из рассмотрения. В результате, краевая задача принципа максимума будет состоять из системы (15), начального условия $x(t_0)=x_0$, условий (29), (30), (32).

Аналогично показывается, что в задаче (1), (2), (4), когда левый конец траектории свободный, а на правом конце заданы условия (29), краевая задача принципа максимума будет состоять из системы (15), условий (29), (30) на правом конце, условия нормировки (32) и условия на левом конце

$$\psi(t_0) = a_0 g_x^0(x_0, x(T)). \tag{33}$$

Рассмотрим задачу (1), (2), (4), когда оба конца траектории подвижны, причем условия на правом конце задаются в виде (29), на левом конце пусть

$$\begin{aligned} &h^i(x(t_0)) \leqslant 0, \quad i=1, \ldots, m_0; \\ &h^i(x(t_0)) = 0, \quad i=m_0+1, \ldots, s_0. \end{aligned} \tag{34}$$

Здесь мы имеем дело с задачей (1)—(4), в которой функции $g^i(x, y)$ определены так: $g^i(x, y)=g^i(y)$ $(i=1, \ldots, m_1)$; $g^i(x, y)=h^{i-m_1}(x)$ $(i=m_1+1, \ldots, m_1+m_0=m)$, $g^i(x, y)=g^{i-m_0}(y)$ $(i=m_1+m_0+1, \ldots, m_0+s_1)$; $g^i(x, y)=h^{i-s_1}$ $(i=m_0+s_1+1, \ldots, s_0+s_1=s)$. Условия (10), (11) на правом конце траектории с учетом (8) запишутся в виде (30), а на левом конце получим

$$\begin{aligned} &\psi(t_0) = a_0 g_x^0(x_0, x(T)) + \sum_{j=1}^{s_0} b_j h_x^j(x_0), \\ &b_j h^j(x_0) = 0, \quad b_j \geqslant 0, \quad j=1, \ldots, m_0. \end{aligned} \tag{35}$$

Таким образом, краевая задача принципа максимума, соответствующая задаче (1), (2), (4), (29), (34), состоит из системы (15), условий (29), (30), (34), (35), условия нормировки

$$a_0^2 + a_1^2 + \ldots + a_{s_1}^2 + b_1^2 + \ldots + b_{s_0}^2 = 1.$$

Если в (1), (2), (4) левый конец удовлетворяет условиям (34), а правый конец закрепленный, то систему (15) нужно решать при условиях (34), (35), $x(T)=x_1$, условии нормировки

$$a_0^2 + b_1^2 + \ldots + b_{s_0}^2 = 1. \tag{36}$$

Если в задаче (1), (2), (4) левый конец удовлетворяет условиям (34), а правый конец свободный, то система (15) решается при условиях (34)—(36), $\psi(T) = -a_0 g_y^0(x_0, x(T))$.

4. Сформулируем принцип максимума для задачи оптимального управления, когда начальный или конечный моменты вре-

мени не закреплены. А именно, пусть требуется минимизировать функцию

$$J(x_0, u(\cdot), x(\cdot), t_0, T) = \int_{t_0}^{T} f^0(x(t), u(t), t)\,dt + g^0(x_0, x(T), t_0, T) \tag{37}$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T; \quad x(t_0) = x_0, \tag{38}$$

$$\begin{gathered} g^i(x_0, x(T), t_0, T) \leqslant 0, \quad i = 1, \ldots, m; \\ g^i(x_0, x(T), t_0, T) = 0, \quad i = m+1, \ldots, s, \end{gathered} \tag{39}$$

$$u = u(t) \in V, \quad t_0 \leqslant t \leqslant T, \tag{40}$$

где один из моментов t_0 или T или оба эти момента заранее неизвестны и подлежат определению вместе с управлением $u(t)$ и траекторией $x(t)$ из условия минимума функции (37); $g^i(x, y, t, T)$ $(i = 0, 1, \ldots, n)$ — заданные функции переменных $x \in E^n$, $y \in E^n$, $t \in \mathbf{R}$, $T \in \mathbf{R}$, $t \leqslant T$; остальные обозначения те же, что и в задаче (1)—(4). В (39) не исключаются возможности, когда отсутствуют ограничения типа неравенств $(m = 0)$, типа равенств $(s = m \geqslant 1)$ или все ограничения (39) $(s = m = 0)$.

Теорема 2. *Пусть функции* $f^j(x, u, t)$ $(j = 0, 1, \ldots, n)$; $g^j(x, y, t, T)$ $(j = 0, 1, \ldots, s)$ *имеют частные производные* $f^j_{x^i}$, $g^j_{x^i}$, $g^j_{y^i}$ $(i = 1, \ldots, n)$; g^j_t, g^j_T *и непрерывны вместе с этими производными по совокупности своих аргументов при* $x \in E^n$, $y \in E^n$, $u \in V$, $t \in \mathbf{R}$, $T \in \mathbf{R}$, $t \leqslant T$. *Пусть* $(x_0, u(\cdot), x(\cdot), t_0, T)$ — *решение задачи* (37)—(40). *Тогда необходимо существуют числа* $a_0, a_1, \ldots, a_s$ *и вектор-функция* $\psi(t) = (\psi_1(t), \ldots, \psi_n(t))$ $(t_0 \leqslant t \leqslant T)$, *удовлетворяющие условиям* 1)—3) *теоремы* 1, *условиям трансверсальности*

$$\begin{gathered} \psi(t_0) = \sum_{j=0}^{s} a_j g^j_x(x_0, x(T), t_0, T), \\ \psi(T) = -\sum_{j=0}^{s} a_j g^j_y(x_0, x(T), t_0, T), \end{gathered} \tag{41}$$

$$\max_{u \in V} H(x(t_0), u, t_0, \psi(t_0), a_0) = -\sum_{j=0}^{s} a_j g^j_t(x_0, x(T), t_0, T) \tag{42}$$

(*если* t_0 *закреплено, то условие* (42) *отсутствует*);

$$\max_{u \in V} H(x(T), u, T, \psi(T), a_0) = \sum_{j=0}^{s} a_j g^j_T(x_0, x(T), t_0, T) \tag{43}$$

(*если* T *закреплено, то условие* (43) *отсутствует*) *и условию дополняющей нежесткости*

$$a_j g^j(x_0, x(T), t_0, T) = 0, \quad j = 1, \ldots, m. \tag{44}$$

С помощью теоремы 2 задачу (37)—(40) можно свести к краевой задаче принципа максимума, действуя по той же схеме, что и в задаче (1)—(4). Это снова приведет нас к конечномерной задаче максимизации (12), функции (13) и к системе $2n$ дифференциальных уравнений (15) относительно функций $x(t)$, $\psi(t)$. Для определения $2n$ параметров, от которых зависит общее решение системы (15), и параметров $a_0, a_1, \ldots, a_s$ имеем $2n$ условий (41), m условий (44), $s-m$ условий типа равенств из (39), условие нормировки (18), а наличие неизвестных моментов t_0, T здесь компенсируется появлением дополнительных условий (42), (43); разумеется, поиск упомянутых параметров нужно вести с учетом неравенств из (8), (39). Расшифровка условий трансверсальности (41), (42) для различных режимов на концах траекторий, когда каждый из концов может быть закрепленным, свободным или подвижным, проводится точно так же, как и выше; в частности, условия (41) здесь приведут к тем же условиям (21)—(36). Естественно, в задаче оптимального управления с незакрепленным временем функции g^i, h^i из условий (29), (34) наряду с x, y могут зависеть еще от переменных t_0, T, как например, в (1.22), (1.23).

5. Для иллюстрации теорем 1, 2 рассмотрим конкретные примеры задач оптимального управления.

Пример 1. Пусть требуется минимизировать функцию $J(u)=\int\limits_0^T (u^2(t)+x^2(t))\,dt$ при условиях $\dot{x}(t)=u(t)$, $0\leqslant t\leqslant T$, $x(0)=x(T)=0$.

Здесь момент $T>0$ задан; $V=E^1$. Задача, конечно, несложная: пара $(u(t)\equiv 0,\ x(t)\equiv 0)$ $(0\leqslant t\leqslant T)$, очевидно, является единственным ее решением. Продемонстрируем на этой простой задаче изложенную выше схему использования принципа максимума — теоремы 1. Выпишем функцию Гамильтона — Понтрягина $H(x, u, \psi, a_0)=-a_0(u^2+x^2)+\psi u$ и сопряженную систему $\dot{\psi}=-H_x=2a_0x$. Если $a_0=0$, то функция $H=\psi u$ может достигать своей верхней грани на множестве $V=E^1$ лишь при $\psi=0$. Однако соотношения $a_0=\psi=0$ противоречат условию (23). Следовательно, $a_0>0$. Тогда можем считать, что $a_0=1$. В этом случае функция $H=-u^2-x^2+\psi u$ достигает верхней грани на E^1 при $u=\psi/2$ — вот какой вид имеет функция (13) в рассматриваемой задаче. Тогда краевая задача принципа максимума запишется в виде

$$\dot{x}=\psi/2,\quad \dot{\psi}=2x,\quad 0\leqslant t\leqslant T;\quad x(0)=x(T)=0.$$

Отсюда однозначно определяем $x(t)\equiv\psi(t)\equiv 0$ $(0\leqslant t\leqslant T)$. Тогда $u(t)=\psi(t)/2\equiv 0$ $(0\leqslant t\leqslant T)$ — получили уже известное нам оптимальное управление.

Перейдем к рассмотрению более интересной задачи оптимального управления, которая в зависимости от величины конечного момента T имеет единственное решение или бесконечно много решений, или не имеет решения. Эта задача любопытна также и тем, что даже в том случае, когда она не имеет решения, краевая задача принципа максимума будет иметь одно или даже бесконечно много решений.

Пример 2. Пусть требуется минимизировать функцию $J(u)=\int\limits_0^T (u^2(t)-x^2(t))\,dt$ при условиях $\dot{x}(t)=u(t)$, $0\leqslant t\leqslant T$, $x(0)=x(T)=0$, $T>0$.

Функция Гамильтона — Понтрягина здесь имеет вид $H=-a_0(u^2-x^2)+\psi(u)$, сопряженная система такая: $\dot{\psi}=-H_x=-2a_0x$. Если $a_0=0$, то $H=\psi u$ достигает своей верхней грани на $V=E^1$ лишь при $\psi=0$, что противоречит условию (23). Следовательно, $a_0>0$. Можно считать, что $a_0=1$. Тогда $H=x^2-u^2+\psi u$ и $\sup\limits_{u\in E} H$ достигается при $u=\psi/2$. Краевая задача принципа максимума имеет вид $\dot{x}=\psi/2$, $\dot{\psi}=-2x$, $0\leqslant t\leqslant T$, $x(0)=x(T)=0$. Общее решение этой системы дается формулой $x(t)=C\sin t+D\cos t$, $\psi(t)=2C\cos t-2D\sin t$, где C, D — произвольные постоянные. С учетом условия $x(0)=0$ отсюда имеем $D=0$, и тогда $x(t)=C\sin t$, $\psi(t)=2C\cos t$. Условие $x(T)=0$ приводит к равенству $C\sin T=0$. Возможно, что $T\neq\pi k$ $(k=1,2,\ldots)$; тогда $C=0$ и краевая задача принципа максимума будет иметь единственное решение $x(t)\equiv 0$, $\psi(t)\equiv 0$ $(0\leqslant t\leqslant T)$, а управление, подозрительное на оптимальность, равно $u(t)=\psi(t)/2=0$ $(0\leqslant t\leqslant T)$. Если же $T=\pi k$, k — целое положительное число, то краевая задача принципа максимума имеет бесчисленное множество решений $x(t)=C\sin t$, $\psi(t)=2C\cos t$, зависящих от одного параметра C, и управлений, подозрительных на оптимальность, будет бесконечно много: $u(t)=C\cos t$ $(0\leqslant t\leqslant T)$.

Спрашивается, будут ли найденные управления оптимальными? Оказывается, ответ на этот вопрос зависит от величины T. Рассмотрим случаи $T>\pi$ и $0<T\leqslant\pi$.

1) $T>\pi$. Покажем, что тогда $\inf J(u)=-\infty$. Для этого возьмем последовательность управлений $u_m=u_m(t)=\frac{m\pi}{T}\cos\frac{\pi t}{T}$ и соответствующих им траекторий $x_m=x_m(t)=m\sin\frac{\pi t}{T}$ $(0\leqslant t\leqslant T$, $m=1,2,\ldots)$. Тогда $J(u_m)=\int\limits_0^T\left(u_m^2(t)-x_m^2(t)\right)dt=\frac{1}{2}Tm^2\times\left(\frac{\pi^2}{T^2}-1\right)\to-\infty$ при $m\to\infty$. Следовательно, при $T>\pi$ рассматриваемая задача оптимального управления не имеет реше-

ния. В то же время краевая задача принципа максимума разрешима, причем для $T \neq \pi k$ $(k=1, 2, \ldots)$ она имеет единственное решение, при $T=\pi k$ $(k=1, 2, \ldots)$ — бесконечно много решений.

2) $0<T\leqslant\pi$. Тогда для любых кусочно-непрерывных $v(t)$, для которых существует решение $x(t)$ задача $\dot{x}(t)=v(t)$ $(0\leqslant t\leqslant T)$, $x(0)=x(T)=0$, имеем

$$J(v)=\int_0^T (v^2-x^2)\,dt=\int_0^T (v^2+x^2\operatorname{ctg}^2 t-x^2\sin^{-2} t)\,dt=$$

$$=\int_0^T (v^2+x^2\operatorname{ctg}^2 t-2x\dot{x}\operatorname{ctg} t)\,dt=\int_0^T (v(t)-x(t)\operatorname{ctg} t)^2\,dt\geqslant 0.$$

Заметим, что проделанные преобразования законны, так как все подынтегральные функции, встретившиеся при этих преобразованиях, ограничены и $\lim\limits_{t\to+0} x^2(t)\operatorname{ctg} t=0$, а в случае $T=\pi$ еще и $\lim\limits_{t\to T-0} x^2(t)\operatorname{ctg} t=0$. Итак, $J(v)\geqslant 0$, а на управлениях $u(t)\equiv 0$ при $T<\pi$ и $u(t)=c\cos t$ при $T=\pi$ будем иметь $J(u)=0$. Таким образом, при $T<\pi$ рассматриваемая задача оптимального управления имеет единственное решение, при $T=\pi$ — бесчисленное множество решений, причем все решения найдены с помощью принципа максимума.

Пример 3. Минимизировать функцию $J(u)=\frac{1}{2}\int_0^T (x^2(t)+u^2(t))\,dt$ при условиях $\dot{x}(t)=-ax(t)+u(t)$, $x(0)=x_0$.

Здесь x_0, $a>0$, $T>0$ — заданные постоянные, $V=E^1$, правый конец траектории свободный. Составим функцию Гамильтона — Понтрягина $H=-a_0(x^2+u^2)/2+\psi(-ax+u)$ и выпишем сопряженную систему

$$\dot{\psi}=-H_x=a_0x+a\psi,\quad 0\leqslant t\leqslant T.$$

Из условия (26) трансверсальности для свободного правого конца траектории имеем $\psi(T)=0$, $a_0=1$. Тогда функция $H=-(x^2+u^2)/2+\psi(-ax+u)$ достигает своей верхней грани по u на $V=E^1$ при $u=\psi$, и краевая задача принципа максимума запишется в виде

$$\dot{x}=-ax+\psi,\quad \dot{\psi}=a\psi+x,\quad x(0)=x_0,\quad \psi(T)=0.$$

Отсюда следует, что подозрительным на оптимальность является управление

$$u(t)=\psi(t)=x_0\,\frac{e^{\lambda t}-e^{2\lambda T}e^{-\lambda T}}{(\lambda-a)+(\lambda+a)\,e^{2\lambda T}},\quad 0\leqslant t\leqslant T,\quad \lambda=\sqrt{a^2+1}.$$

Пример 4. Пусть точка движется по оси Ox по закону $\ddot{x}(t)=u(t)$ $(t\geqslant 0)$. Требуется найти кусочно-непрерывное уп-

равление $u(t)$, $|u(t)| \leqslant 1$ $(0 \leqslant t \leqslant T)$, такое, чтобы точка, выйдя из начального положения $x(0)=1$ с нулевой скоростью, пришла в начало координат с нулевой скоростью за минимальное время T.

Положим, что $x^1 = x$, $x^2 = \dot{x}$ — фазовые координаты точки. Тогда задачу можно переформулировать так: быстрейшим образом перевести фазовую точку (x^1, x^2) из состояния $(1, 0)$ в состояние $(0, 0)$, считая, что движение подчиняется уравнениям $\dot{x}^1(t) = x^2(t)$, $\dot{x}^2(t) = u(t)$ $(t \geqslant 0)$. Здесь $V = \{u \in E^1\colon |u| \leqslant 1\}$, $f^0 \equiv 1$, $g^0 \equiv 0$.

Составим функцию $H = -a_0 + \psi_1 x^2 + \psi_2 u$ и выпишем сопряженную систему

$$\dot{\psi}_1 = -H_{x^1} = 0, \quad \dot{\psi}_2 = -H_{x^2} = -\psi_1, \quad t \geqslant 0.$$

Отсюда имеем $\psi_1(t) = C$, $\psi_2(t) = -Ct + D$, где C, D — постоянные. Отметим, что $\psi_2(t) \not\equiv 0$, так как в противном случае $C = D = 0$, а тогда $\psi_1(t) \equiv 0$ и равенство $H|_{t=T} = -a_0 = 0$, вытекающее из (43), приводит к противоречию с условием (23). Из условия $\max\limits_{|u| \leqslant 1} H$ следует $u(t) = \operatorname{sign} \psi_2(t) = \operatorname{sign}(-Ct + D)$ $(t \geqslant 0)$. Таким образом, оптимальное управление (если оно существует) является кусочно-постоянной функцией, принимающей значения $+1$, -1 и имеющей не более одной точки переключения t_1, при переходе через которую $u(t)$ меняет знак. Нетрудно убедиться, что траектория, выходящая из точки $(1, 0)$ и соответствующая управлению $u(t) = +1$ при $t \geqslant 0$, или $u(t) \equiv -1$ при $t \geqslant 0$, или $u(t) \equiv +1$ $(0 \leqslant t < t_1)$, $u(t) = -1$ $(t \geqslant t_1)$, никогда не будет проходить через точку $(0, 0)$. Остается рассмотреть управление $u(t) = -1$ $(0 \leqslant t < t_1)$, $u(t) = +1$ $(t \geqslant t_1)$. Этому управлению соответствует траектория $(x_1(t), x_2(t))$:

$$x^1(t) = \begin{cases} 1 - t^2/2, & 0 \leqslant t \leqslant t_1, \\ t^2/2 - 2t_1 t + t_1^2 + 1, & t \geqslant t_1, \end{cases}$$

$$x^2(t) = \begin{cases} -t, & 0 \leqslant t \leqslant t_1, \\ t - 2t_1, & t \geqslant t_1. \end{cases}$$

Из условия $x^1(T) = x^2(T) = 0$ находим $t_1 = 1$, $T = 2$. Тогда

$$x^1(t) = \begin{cases} 1 - t^2/2, & 0 \leqslant t \leqslant 1, \\ (t-2)^2/2, & 1 \leqslant t \leqslant 2, \end{cases} \quad x^2(t) = \begin{cases} -t, & 0 \leqslant t \leqslant 1, \\ t - 2, & 1 \leqslant t \leqslant 2. \end{cases}$$

В качестве величин a_0, ψ_1, ψ_2, участвующих в формулировке принципа максимума, могут служить $a_0 = 0$, $\psi_1(t) = -1$, $\psi_2(t) = t - 1$ $(0 \leqslant t \leqslant 2)$. Можно показать, что полученные управление и траектория в самом деле являются решением поставленной задачи быстродействия об этом см. пример 7.4.4.

Пример 5. Требуется перевести точку $x = (x^1, x^2)$ из состояния $x_0 = (2, -2)$ на множество $S_1 = \{x \in E^2\colon g^1(x) \equiv x^1 = 0\}$

быстрейшим образом, предполагая, что движение точки подчиняется уравнениям $\dot{x}^1(t)=x^2(t)$, $\dot{x}^2(t)=u(t)$, причем $u(t)\in V=\{u\in E^1\colon |u|\leqslant 1\}$.

Как и в предыдущем примере, здесь $H=-a_0+\psi_1x^2+\psi_2u$, сопряженная система имеет вид: $\dot{\psi}_1=0$, $\dot{\psi}_2=-\psi_1$, откуда следует $\psi_1(t)\equiv C$, $\psi_2(t)=-Ct+D$, C, $D=\text{const}$. Условия трансверсальности (30), (43) здесь дают

$$\psi_1(T)=-a_1,\ \psi_2(T)=0,\ -a_0+\psi_1(T)x^2(T)+\psi_2(T)u(T)=H|_{t=T}=0.$$

Следовательно, $\psi_2(t)=C(T-t)$ $(0\leqslant t\leqslant T)$. Заметим, что здесь $C\neq 0$, так как при $C=0$ получим $\psi_1(t)\equiv\psi_2(t)\equiv 0$ $(0\leqslant t\leqslant T)$, а тогда $a_1=0$, из условия $H|_{t=T}=0$ вытекает $a_0=0$ — противоречие с условием (32). Итак, $C\neq 0$, $\psi(t)=C(T-t)\neq 0$ при $0\leqslant t<T$. Из условия $\sup\limits_{|u|\leqslant 1} H$ тогда имеем

$$u(t)=\operatorname{sign}\psi_2(t)=\operatorname{sign}C,\quad 0\leqslant t\leqslant T.$$

Значит, подозрительными на оптимальность здесь могут быть лишь управления $u(t)\equiv 1$ или $u(t)\equiv -1$ $(0\leqslant t\leqslant T)$. Если $u(t)=1$, то из краевой задачи

$$\dot{x}^1=x^2,\quad \dot{x}^2=1,\quad 0\leqslant t\leqslant T;\quad x^1(0)=2,\quad x^2(0)=-2,\ x^1(T)=0$$

получим $T=2$, $x^1(t)=(t-2)^2/2$, $x^2(t)=t-2$ $(0\leqslant t\leqslant 2)$. Если $u(t)=-1$, то из

$$\dot{x}^1=x^2,\quad \dot{x}^2=-1,\quad 0\leqslant t\leqslant T;\quad x^1(0)=2,\quad x^2(0)=-2;\quad x^1(T)=0$$

будем иметь $T=\sqrt{8}-2$, $x^1(t)=4-(t+2)^2/2$, $x^2(t)=-t-2$, $0\leqslant t\leqslant\sqrt{8}-2$.

Таким образом, краевая задача принципа максимума здесь дает два решения. Однако лишь управление $u(t)=-1$ $(0\leqslant t\leqslant T=\sqrt{8}-2)$ может претендовать на оптимальность, так как $T=2>\sqrt{8}-2$, а управление $u(t)=1$ $(0\leqslant t\leqslant T)$ заведомо неоптимально.

Пример 6. Рассмотрим задачу минимизации функции

$$J(u)=\int_0^1((u^1(t))^2+(u^2(t))^2)\,dt+x^1(1)+x^2(1) \tag{45}$$

при условиях

$$\begin{gathered}\dot{x}^1(t)=u^1(t),\quad \dot{x}^2(t)=u^2(t),\quad 0\leqslant t\leqslant 1,\\ x^1(0)=0,\quad x^2(0)=0,\quad x^1(1)\leqslant 0,\quad (x^2(1))^2-x^1(t)\leqslant 0,\end{gathered} \tag{46}$$

где $u=u(t)=(u^1(t),\ u^2(t))\in L^2_\infty[0,1]$. Эта задача является частным случаем задачи (1), (2), (4), (29) при $t_0=0$, $T=1$, $n=r=2$, $f^0\equiv(u^1)^2+(u^2)^2$, $f(x,\ u,\ t)=u$, $V=E^2$, $x=(x^1,\ x^2)$, $y=(y^1,\ y^2)$, $g^0(x,\ y)=y^1+y^2$, $g^1(x,\ y)=y^1$, $g^2(x,\ y)=-y^1+(y^2)^2$,

$m_1 = s_1 = 2$; левый конец траектории закреплен. Из (46) видно, что правый конец любой допустимой траектории этой задачи удовлетворяет равенствам: $x^1(1) = 0$, $x^2(1) = 0$. Тогда $J(u) = \int_0^1 |u(t)|^2 dt \geqslant 0$ для всех допустимых управлений. Поскольку $u = u(t) \equiv 0$ допустимое управление и $J(0) = 0$, то $J_* = 0$, $u(t) \equiv 0$ — единственное оптимальное управление задачи (45), (46). Функция Гамильтона — Понтрягина $H = -a_0((u^1)^2 + (u^2)^2) + \psi_1 u^1 + \psi_2 u^2$ не зависит от x, поэтому сопряженная система имеет вид $\dot{\psi}_1 = 0$, $\dot{\psi}_2 = 0$ $(0 \leqslant t \leqslant 1)$. Следовательно, $\psi_1(t) \equiv c_1$, $\psi_2(t) \equiv c_2$, c_1, c_2 — постоянные. Условие (30), (32) здесь дают

$$\begin{gathered} -\psi_1(1) = a_0 \cdot 1 + a_1 \cdot 1 + a_2(-1) = a_0 + a_1 - a_2, \\ -\psi_2(1) = a_0 \cdot 1 + a_1 \cdot 0 + a_2 \cdot 2x^2(1) = a_0, \\ a_0^2 + a_1^2 + a_2^2 = 1, \quad a_0 \geqslant 0, \quad a_1 \geqslant 0, \quad a_2 \geqslant 0. \end{gathered} \tag{47}$$

Покажем, что в этой задаче $a_0 = 0$. В самом деле, если $a_0 > 0$, то можем воспользоваться условием нормировки $a_0 = 1$. Тогда функция $H = -|u|^2 + \langle \psi, u \rangle$ достигает своего максимума на $V = E^2$ в точке $u = \psi/2 = c/2$, $c = (c_1, c_2)$. Соответствующая траектория $x(t) = tc/2$ условию $x(1) = 0$ может удовлетворять лишь при $c_1 = c_2 = 0$. Таким образом, $\psi(t) \equiv 0$ $(0 \leqslant t \leqslant 1)$. Из второго условия (47) тогда следует, что $a_0 = 0$, что противоречит равенству $a_0 = 1$. Следовательно, $a_0 = 0$. Но тогда линейная функция $H = \langle \psi, u \rangle$ на E^2 может иметь конечный максимум (который, кстати, должен достигаться на оптимальном управлении $u(t) \equiv 0$) лишь при $\psi = \psi(t) \equiv c = 0$. Из условий (47), учитывая, что $a_0 = 0$, получаем $a_1 = a_2 > 0$. Таким образом, краевая задача принципа максимума здесь дает $a = (a_0 = 0,\ a_1 = a,\ a_2 = a)$, $a > 0$, $\psi(t) \equiv 0$, $0 \leqslant t \leqslant 1$. Как видим, условие (23) в этой задаче не выполняется, функция $H \equiv 0$, и условие максимума (9) не позволяет определить оптимальное управление $u = u(t) \equiv 0$.

Говорят, что оптимальное управление $u(\cdot)$ является *особым* на отрезке $[\alpha, \beta] \subset [t_0, T]$, если $H(x(t), u, t, \psi(t), a_0)$ при $t \in [\alpha, \beta]$ не зависит от u. В этом случае для набора $(x = x(t), t, \psi = \psi(t), a_0)$ при $t \in [\alpha, \beta]$ условие (14) не дает никакой полезной информации об оптимальном управлении, функция (13) становится неопределенной и пользоваться формулой (20) невозможно. В частности, когда нарушается условие (23), т. е. $a_0 = 0$, $\psi(t) \equiv 0$ $(t_0 \leqslant t \leqslant T)$, то имеем дело с одним из типичных случаев появления особого управления. Так случилось в только что рассмотренном примере 6. Разумеется, условие (23) само по себе не исключает возможность появления особого управления, но тем не менее полезно подчеркивать случаи, когда оно выполняется.

К сожалению, условие $a=(a_0, \ldots, a_s)\neq 0$ из (8) само по себе не всегда приводит к условию (23). Например, в задаче (1), (2), (4), (29), как показывает пример 6, в общем случае (23) не выполняется и приходится довольствоваться условием $a=(a_0, \ldots, a_{s_1})\neq 0$ и вытекающим из него условием нормировки (32). Для того чтобы в задаче (1), (2), (4), (29) из $a\neq 0$ следовало условие (23), можно дополнительно потребовать, например, чтобы градиенты $g_y^1(x(T)), \ldots, g_y^{s_1}(x(T))$ были линейно независимы. Тогда либо $a_0\neq 0$, либо $a_0=0$, но согласно (30) $\psi(T)\neq 0$ и однородная система (6) будет иметь нетривиальное решение $\psi(t)$ $(t_0\leqslant t\leqslant T)$. Аналогичные требования, гарантирующие условие (23), можно сформулировать и для задачи (1), (2), (4), (29), (34) и других задач оптимального управления, рассмотренных выше. Об особых управлениях читатель может прочесть, например, в [68, 98, 205].

В следующих примерах покажем, как выписывается краевая задача принципа максимума для некоторых задач оптимального управления движением математического маятника (см. пример 1.1).

Пример 7. Пусть требуется минимизировать функцию

$$J(u)=(x^1(T))^2+(x^2(T))^2 \tag{48}$$

при условиях

$$\begin{gathered}\dot{x}^1(t)=x^2(t), \quad \dot{x}^2(t)=-\sin x^1(t)-\beta x^2(t)+u(t),\\ 0\leqslant t\leqslant T, \quad x(0)=x_0,\end{gathered} \tag{49}$$

$$u(t)\in V=\{u\in E^1:\ |u|\leqslant 1\}, \tag{50}$$

где $x=(x^1, x^2)$ — фазовые координаты, $x_0=(x_0^1, x_0^2)$ — заданная точка, $T>0$ — заданный момент времени. В этой задаче правый конец траектории свободен, $f^0\equiv 0$, $g^0(y)=(y^1)^2+(y^2)^2$. Выпишем функцию Гамильтона — Понтрягина

$$H(x, u, \psi)=\psi_1 x^2+\psi_2(-\sin x^1-\beta x^2+u)$$

и сопряженную систему

$$\dot{\psi}_1=-H_{x^1}=\psi_2\cos x^1, \quad \dot{\psi}_2=-H_{x^2}=-\psi_1+\beta\psi_2, \quad 0\leqslant t\leqslant T. \tag{51}$$

Из условий (26) имеем

$$\begin{aligned}\psi_1(T)&=-g_{y^1}^0(x(T))=-2x^1(T),\\ \psi_2(T)&=-g_{y^2}^0(x(T))=-2x^2(T).\end{aligned} \tag{52}$$

Из условия $\max\limits_{|u|\leqslant 1} H$ следует $u=\operatorname{sign}\psi_2$. Тогда краевая задача

принципа максимума запишется в виде

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -\sin x^1 - \beta x^2 + \operatorname{sign} \psi_2,$$
$$\dot{\psi}_1 = \psi_2 \cos x^1, \quad \dot{\psi}_2 = -\psi_1 + \beta\psi_2, \quad 0 \leqslant t \leqslant T, \tag{53}$$
$$x(0) = x_0, \quad \psi_1(T) = -2x^1(T), \quad \psi_2(T) = -2x^2(T). \tag{54}$$

Если краевая задача (53), (54) имеет решение $(x(t), \psi(t))$ $(0 \leqslant t \leqslant T)$, причем $\psi_2(t)$ обращается в нуль в конечном числе точек, то функция $u(t) = \operatorname{sign} \psi_2(t)$ будет управлением, подозрительным на оптимальность в задаче (48)—(50).

Заметим, что если для некоторого управления $v = v(\cdot)$ из (50) решение $x(\cdot, v)$ задачи Коши (49) таково, что $x(T, v) = 0$, то $J(v) = 0$. Это значит, что $v(\cdot)$ — оптимальное управление в задаче (48)—(50). Любопытно, что это управление является особым — его нельзя получить из принципа максимума. В самом деле, при $x(T, v) = 0$ из (51), (52) следует $\psi(t, v) \equiv 0$ $(0 \leqslant t \leqslant T)$, а тогда $H(x(t, v), u, \psi(t, v)) \equiv 0$ $(0 \leqslant t \leqslant T)$, при всех $u \in V$ и условие $\sup\limits_{|u|<0} H$ не суживает исходное множество управлений, подозрительных на оптимальность.

Пример 8. Пусть требуется минимизировать функцию $J(u) = \int\limits_0^T u^2(t)\,dt$ при условиях (49) и закрепленном правом конце $x(T) = 0$; $T > 0$ — задано.

Здесь $V = E^1$, $H = -a_0 u^2 + \psi_1 x^2 + \psi_2(-\sin x^1 + \beta x^2 + u)$, сопряженная система имеет вид (51). В случае $a_0 = 0$ функция H может достигать своей верхней грани на E^1 лишь при $\psi_2 = 0$. Но если $\psi_2(t) \equiv 0$ $(0 \leqslant t \leqslant T)$, то из второго уравнения (51) получим $\psi_1(t) \equiv 0$, что противоречит условию (23). Таким образом, можем считать $a_0 = 1$. Тогда из условия $\max\limits_{u \in E^1} H$ получим $u = \psi_2/2$. Краевая задача принципа максимума будет иметь вид

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -\sin x^1 - \beta x^2 + \psi_2/2,$$
$$\dot{\psi}_1 = \psi_2 \cos x^1, \quad \dot{\psi}_2 = -\psi_1 + \beta\psi_2, \quad 0 \leqslant t \leqslant T,$$
$$x(0) = x_0, \quad x(T) = 0.$$

Пример 9. Рассмотрим задачу быстрейшего перевода точки $x = (x^1, x^2)$ из состояния $x_0 \neq 0$ в начало координат (0, 0), предполагая, что движение точки подчиняется условиям (49), (50). Эта задача является частным случаем задачи (37)—(40). Если $f^0 \equiv 1$, $g^0 \equiv 0$; функция Гамильтона — Понтрягина имеет вид

$$H = -a_0 + \psi_1 x^2 + \psi_2(-\sin x^1 - \beta x^2 + u). \tag{55}$$

Отсюда ясно, что сопряженная система будет иметь вид (51),

а условие $\max\limits_{|u|\leqslant 1} H$ выделит функцию $u = \operatorname{sign}\psi_2$. Краевая задача принципа максимума в этом случае будет состоять из системы (53), граничных условий $x(0) = x_0$, $x(T) = 0$, условия трансверсальности $H|_{t=T} = -a_0 + \psi_2(T)u(T) = 0$, вытекающего из (43), и условия (23). Отметим, что в этой задаче $\psi_2(t) \not\equiv 0$. В самом деле, если бы $\psi_2(t) \equiv 0$, то из (51) будем иметь $\psi_1(t) \equiv 0$, а из $H|_{t=T} = 0$ получим $a_0 = 0$, что противоречит (23).

Пример 10. Пусть требуется быстрейшим образом перевести точку $x = (x^1, x^2)$ из состояния x_0 в начальный момент $t_0 = 0$ в состояние, удаленное от точки (0, 0) на расстояние, равное $\varepsilon > 0$, предполагая, что движение точки подчиняется условиям (49), (50). Это значит, что правый конец траектории является подвижным и удовлетворяет условиям $|x(T)|^2 = (x^1(T))^2 + (x^2(T))^2 = \varepsilon^2$. Здесь функция H имеет тот же вид (55), сопряженная система — вид (51), условие $\max\limits_{|u|\leqslant 1} H$ дает $u = \operatorname{sign}\psi_2$. Из условия трансверсальности (30), (43) имеем

$$\psi(T) = -2a_1 x(T), \quad H(x(T), u(T), T, \psi(T), a_0) = 0. \tag{56}$$

Оказывается, здесь $a_1 \neq 0$. В самом деле, если $a_1 = 0$, то из (56) будем иметь $\psi(T) = 0$, а из (51) будет следовать $\psi(t) \equiv 0$; тогда из (55), (56) получим $a_0 = 0$, что противоречит условию (32). Итак, $a_1 \neq 0$; тогда условие нормировки (32) можем заменить на равенство $|a_1| = 1$. Таким образом, краевая задача принципа максимума в рассматриваемой задаче состоит из системы (53), граничных условий $x(0) = x_0$, $\psi(T) = \pm 2x(T)$, $|x(T)|^2 = \varepsilon^2$, $H|_{t=T} = 0$, из которых нужно определить функции $x(t)$, $\psi(t)$, параметры a_0, T. Отметим, что здесь $\psi_2(t) \not\equiv 0$. В самом деле, если $\psi_2(t) \equiv 0$, то в силу (51) $\psi_1(t) \equiv 0$, а тогда $x(T) = 0$, что противоречит равенству $|x(T)|^2 = \varepsilon^2 > 0$.

Пример 11. В задаче из примера 10 условие $|x(T)| = \varepsilon^2$ на правом конце траектории заменим неравенством $|x(T)|^2 \leqslant \varepsilon^2$, считая, что $|x(t_0)| > \varepsilon^2$. Тогда функция H, сопряженная система (51), условия (56) останутся без изменений; здесь также выполняется условие дополняющей нежесткости $a_1(|x(T)|^2 - \varepsilon^2) = 0$. Оказывается, и в этой задаче можно показать, что $a_1 \neq 0$. В самом деле, если $a_1 = 0$, то в силу (56) $\psi(T) = 0$, из (51) будет следовать $\psi(t) \equiv 0$, а из $H|_{t=T} = 0$ получаем $a_0 = 0$, что противоречит условию (32). Итак, $a_1 \neq 0$. С учетом $a_1 \geqslant 0$ можем принять $a_1 = 1$. Таким образом, краевая задача принципа максимума будет состоять из системы (53), условий $x(0) = x_0$, $\psi(T) = -2x(T)$, $|x(T)|^2 = \varepsilon^2$, $H|_{t=T} = 0$. Как и в предыдущем примере можно показать, что $\psi_2(t) \not\equiv 0$.

Предлагаем читателю самостоятельно выписать краевую задачу принципа максимума для задач из примеров 10, 11 с заменой условий на правом конце одним из условий $x^i(T) = 0$, $|x^i(T)|^2 \leqslant \varepsilon^2$, $|x^i(T)|^2 = \varepsilon^2$, где $i = 1$ или 2.

6. При формулировке принципа максимума было оговорено, что условие максимума (9) имеет место для всех точек непрерывности кусочно-непрерывного оптимального управления $u(t)$. Возникает вопрос: как истолковать условие (9) в точках разрыва $u(t)$? Нельзя ли определить оптимальное управление $u(t)$ в точках разрыва так, чтобы равенство (9) выполнялось во всех точках отрезка $[t_0, T]$? Представляется естественным доопределить $u(t)$ в точках разрыва предельным значением слева или справа, приняв $u(t) = u(t+0) = \lim\limits_{\tau \to t+0} u(\tau)$ или $u(t) = u(t-0) = \lim\limits_{\tau \to t-0} u(\tau)$ при $t_0 < t < T$, а на концах отрезка $[t_0, T]$ взяв $u(T) = u(T-0)$, $u(t_0) = u(t_0+0)$. Сразу же отметим, что значения кусочно-непрерывного управления $u(\cdot)$ в точках разрыва не влияют на решение уравнения (2) (см. определение 1.1), на значение интеграла в (1) и, следовательно, на задачу (1)—(4). Покажем, что способ доопределения управления $u(t)$ в точках разрыва не оказывает существенного влияния и на условие (9).

Теорема 3. *Пусть в задаче* (1)—(4) *выполнены все условия теоремы* 1, *множество* V *замкнуто, пусть* $u(\cdot)$ — *оптимальное кусочно-непрерывное управление,* $x(\cdot)$ — *оптимальная траектория, функция* $\psi(\cdot)$ *и параметры* $a_0, \ldots, a_s$ *определены согласно теореме* 1. *Тогда функция*

$$H(t) = \sup_{w \in V} H(x(t), w, t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T, \tag{57}$$

непрерывна во всех точках отрезка $[t_0, T]$, *причем верхняя грань в* (57) *достигается как при* $u = u(t+0) \in V$, *так и при* $u = u(t-0)$ *для всех* $t \in [t_0, T]$. *Если в дополнение к условиям теоремы* 1 *функции* $f^i(x, u, t)$ $(i = 0, 1, \ldots, n)$, *имеют частные производные* $f_t^i(x, u, t)$, *непрерывные по совокупности аргументов* $(x, u, t) \in E^n \times E^n \times [t_0, T]$, *то функция* (57) *имеет левую и правую производные во всех точках* $t \in [t_0, T]$, *причем*

$$\dot{H}(t \pm 0) = H_t(x, u, t, \psi, a_0)\,|_{x=x(t),\, u=u(t\pm 0),\, \psi=\psi(t)}, \quad t_0 \leqslant t \leqslant T, \tag{58}$$

где $H_t(x, u, t, \psi, a_0) = -a_0 f_t^0(x, u, t) + \langle \psi, f_t(x, u, t) \rangle$ — *частная производная функции* H *по переменной* t; *производная* $\dot{H}(t)$ *существует и непрерывна во всех точках непрерывности оптимального управления.*

Доказательство. Пусть $\tau_1, \ldots, \tau_M$ — точки разрыва оптимального управления $u(t)$; тогда $A = [t_0, T] \setminus \{\tau_1, \ldots, \tau_M\}$ — множество точек непрерывности $u(t)$. Для краткости положим $H(t, w) = H(x(t), w, t, \psi(t), a_0)$. Согласно (57) $H(t) = \sup\limits_{w \in V} H(t, w)$. Точку из V, в которой достигается верхняя грань

в (57), обозначим через $v(t)$. Заметим, что верхняя грань может достигаться в нескольких точках, поэтому точка $v(t)$ определяется неоднозначно. В частности, согласно (9) можно взять $v(t)=u(t)$ при всех $t\in A$; существование $v(t)$ при $t\notin A$ для простоты изложения будем предполагать. Зафиксируем произвольную точку $t\in[t_0, T]$. Пусть $|\Delta t|>0$ столь мало, что полуинтервалы $[t-|\Delta t|, t)$, $(t, t+|\Delta t|]\in A$ (если $t=t_0$ или $t=T$, то, конечно, принадлежность A требуется лишь для одного из этих полуинтервалов). Тогда $H(t+\Delta t)=H(t+\Delta t, u(t+\Delta t))\geqslant \geqslant H(t+\Delta t, v(t))$, $H(t)=H(t, v(t))\geqslant H(t, u(t+\Delta t))$. Поэтому

$$H(t+\Delta t, v(t))-H(t, v(t))\leqslant H(t+\Delta t)-H(t)\leqslant \\ \leqslant H(t+\Delta t, u(t+\Delta t))-H(t, u(t+\Delta t)). \quad (59)$$

Так как функции $x(t)$, $\psi(t)$, $H(x, w, t, \psi, a_0)$ непрерывны по совокупности своих аргументов, то функция $H(t, w)=H(x(t), w, t, \psi(t), a_0)$ непрерывна по $(t, w)\in[t_0, T]\times V$. Поэтому переходя к пределу при $\Delta t\to+0$ или $\Delta t\to-0$, из (59) получаем $H(t)= =H(t+0)=H(t-0)$, где $H(t\pm 0)=H(t, u(t\pm 0))$. Тем самым, непрерывность функции $H(t)$ на отрезке $[t_0, T]$ установлена. Кроме того, показано, что в (57) верхняя грань достигается как при $v(t)=u(t+0)$, так и при $v(t)=u(t-0)$.

Пусть теперь функции $f^i(x, u, t)$ имеют непрерывно частные производные по (x, t). Тогда, как следует из (5), функция $H(x, u, t, \psi, a_0)$ непрерывно дифференцируема по (x, t, ψ). Далее, как видно из (2), (6), функции $x(t)$, $\psi(t)$ непрерывно дифференцируемы при всех $t\in A$. Тогда сложная функция $H(t, w)=H(x(t), w, t, \psi(t), a_0)$ также непрерывно дифференцируема при всех $t\in A$ и ее производная равна

$$\frac{dH(t, w)}{dt}=\langle H_x(x(t), w, t, \psi(t), a_0), f(x(t), u(t), t)\rangle + \\ +\langle f(x(t), w, t), -H_x(x(t), u(t), t, \psi(t), a_0)\rangle + \\ + H_t(x(t), w, t, \psi(t), a_0), \quad t\in A. \quad (60)$$

Из этой формулы видно, что при сделанных предположениях относительно функций $f^i(x, u, t)$ производная $dH(t, w)/dt$ непрерывна по совокупности $(t, w)\in A\times V$, причем существуют односторонние пределы $\lim\limits_{\Delta t\to\pm 0}\frac{dH(t+\Delta t, w)}{dt}=\frac{dH(t\pm 0, w)}{dt}$, $\lim\limits_{\Delta t\to\pm 0}\frac{dH(t+\Delta t, u(t+\Delta t))}{dt}=\frac{dH(t\pm 0, u(t\pm 0))}{dt}$ при всех $t\in[t_0, T]$, $w\in V$. Из (60) также следует, что

$$\frac{dH(t, u(t))}{dt}=H_t(x(t), u(t), t, \psi(t), a_0), \quad t\in A, \quad (61)$$

откуда

$$\frac{dH(t, u(t \pm 0))}{dt} = H_t(x(t), u(t \pm 0), t, \psi(t), a_0), \quad t \in [t_0, T]. \quad (62)$$

Возьмем в (59) $\Delta t > 0$, $v(t) = u(t+0)$; получим

$$H(t+\Delta t, u(t+0)) - H(t, u(t+0)) \leqslant H(t+\Delta t) - H(t) \leqslant$$
$$\leqslant H(t+\Delta t, u(t+\Delta t)) - H(t, u(t+\Delta t)).$$

По теореме Лагранжа о конечных приращениях отсюда имеем

$$\frac{dH(t+\theta_1\Delta t, u(t+0))}{dt}\Delta t \leqslant H(t+\Delta t) - H(t) \leqslant$$
$$\leqslant \frac{dH(t+\theta_2\Delta t, u(t+\Delta t))}{dt}\Delta t,$$

где $0 < \theta_1, \theta_2 < 1$. Разделив эти неравенства на $\Delta t > 0$ и устремив $\Delta t \to +0$, получим $\dot{H}(t+0) = \frac{dH(t, u(t+0))}{dt}$. Далее, взяв в (59) $\Delta t < 0$, $v(t) = u(t-0)$ рассуждая аналогично, имеем $\dot{H}(t-0) = \frac{dH(t, u(t-0))}{dt}$. Отсюда и из (61), (62) следует формула (58) и непрерывность $\dot{H}(t)$ при $t \in A$. Теорема 3 доказана.

Нетрудно видеть, что теорема 3 остается верной и для задачи (37)—(40).

7. В следующем параграфе будет приведено доказательство теорем 1, 2. Здесь мы хотим привести эвристические соображения, которые помогают понять, откуда появляется сопряженная система, условия максимума, условия трансверсальности, фигурирующие в теоремах 1, 2, и установить связь между принципом максимума Понтрягина и правилом множителей Лагранжа.

Для простоты рассмотрим задачу (1), (2), (4) при дополнительном предположении, что $V = E^r$, $g^0(x, y) = g^0(y)$, условия на правом конце траектории имеют вид

$$g^1(x(T)) = 0, \ldots, g^{s_1}(x(T)) = 0, \quad (63)$$

на левом конце

$$h^1(x(t_0)) = 0, \ldots, h^{s_0}(x(t_0)) = 0. \quad (64)$$

Воспользуемся процедурой исследования задач на условный экстремум (см. § 2.2, 4.8) и введем множители Лагранжа: непрерывную вектор-функцию $\psi(t) = (\psi_1(t), \ldots, \psi_n(t))$ для учета ограничений (2), постоянные $(a_1, \ldots, a_{s_1}) = a$ для учета ограниче-

ний (63), постоянные $(b_1, \ldots, b_{s_0}) = b$ — для учета (64), постоянную a_0 — для функции (1), и составим функцию Лагранжа

$$\mathscr{L}(x(\cdot), u(\cdot), \psi(\cdot), a_0, a, b) = \\ = -a_0\left[\int_{t_0}^{T} f^0(x(t), u(t), t)\,dt + g^0(x(T))\right] + \\ + \int_{t_0}^{T} \langle \psi(t), f(x(t), u(t), t) - \dot{x}(t)\rangle\,dt - \\ - \sum_{j=1}^{s_1} a_j g^j(x(T)) - \sum_{j=1}^{s_0} b_j h^j(x(t_0)).$$

С помощью функции Гамильтона — Понтрягина (5) можно записать функцию Лагранжа в следующем виде:

$$\mathscr{L}(x(\cdot), u(\cdot), \psi(\cdot), a_0, a, b) = \\ = \int_{t_0}^{T} [H(x(t), u(t), t, \psi(t), a_0) - \langle \psi(t), \dot{x}(t)\rangle]\,dt - \\ - \sum_{j=0}^{s_1} a_j g^j(x(T)) - \sum_{j=1}^{s_0} b_j h^j(x(t_0)).$$

Все аргументы функции Лагранжа считаются независимыми переменными. Дадим приращения (вариации) переменным $x(t)$, $u(t)$, т. е. рассмотрим $x(t) + \delta x(t)$, $u(t) + \delta u(t)$ $(t_0 \leqslant t \leqslant T)$, здесь $u(t)$, $\delta u(t)$ — кусочно-непрерывны, а $x(t)$, $\delta x(t)$ — непрерывно дифференцируемы на $[t_0, T]$. Тогда вариация функции Лагранжа, представляющая собой главную линейную часть приращения этой функции, имеет вид

$$\delta\mathscr{L} = \int_{t_0}^{T} [\langle H_x, \delta x\rangle + \langle H_u, \delta u\rangle - \langle \psi, \delta\dot{x}\rangle]\,dt - \\ - \sum_{j=0}^{s_1} a_j \langle g_x^j(x(T)), \delta x(T)\rangle - \sum_{j=1}^{s_0} b_j \langle h_x^j(x(t_0)), \delta x(t_0)\rangle;$$

для краткости аргументы у функций под интегралом опущены. Интегрируя по частям, находим

$$\int_{t_0}^{T} \langle \psi(t), \delta\dot{x}(t)\rangle\,dt = \langle \psi(t), \delta x(t)\rangle\big|_{t=t_0}^{T} - \int_{t_0}^{T} \langle \dot{\psi}(t), \delta x(t)\rangle\,dt.$$

Тогда

$$\delta\mathscr{L} = \int_{t_0}^{T} [\langle H_x + \dot{\psi}, \delta x\rangle + \langle H_u, \delta u\rangle]\, dt -$$

$$- \left\langle \sum_{j=0}^{s_1} a_j g_x^j (x(T)) + \psi(T), \delta x(T) \right\rangle +$$

$$+ \left\langle - \sum_{j=1}^{s_0} b_j h_x^j (x(t_0)) + \psi(t_0), \delta x(t_0) \right\rangle.$$

Пользуясь независимостью вариаций $\delta x(\cdot)$, $\delta u(\cdot)$, из условия стационарности $\delta\mathscr{L} = 0$ имеем

$$\dot{\psi} + H_x(x, u, t, \psi, a_0) = 0, \qquad H_u(x, u, t, \psi, a_0) = 0,$$

$$\psi(T) = - \sum_{j=0}^{s_1} a_j g_x^j (x(T)), \quad \psi(t_0) = \sum_{j=1}^{s_0} b_j h_x^j (x(t_0)).$$

Таким образом, получены почти все основные соотношения теоремы 1, кроме условий (8), (9). Вместо условия (9) мы получили равенство $H_u = 0$, являющееся следствием (9) при $V = E^r$. Подчеркнем, что приведенные здесь рассуждения, конечно, не могут считаться строгими и являются лишь полезными наводящими соображениями при получении необходимых условий оптимальности.

В заключении заметим, что принцип максимума выше был сформулирован для задач оптимального управления, не содержащих фазовые ограничения при $t_0 < t < T$, и в предположении, что область управлений — множество V — не зависит от времени. Принцип максимума может быть сформулирован и для задач с фазовыми ограничениями и с переменной областью управления [31, 77, 137, 166, 206, 306], однако, надо отметить, получающаяся отсюда краевая задача принципа максимума будет иметь более сложный вид. Другой подход к исследованию задач с фазовыми ограничениями и с переменной областью интегрирования будет изложен в главе 7.

Упражнения. 1. С помощью принципа максимума решить задачу быстродействия при условиях $\dot{x}^1 = x^2$, $\dot{x}^2 = u(t)$, $x(0) \in S_0$, $x(T) \in S_1$; $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$, где $S_0 = \{x \in E^2\colon h(x) \equiv |x|^2 = 1\}$ или $S_0 = \{x \in E^2\colon h(x) \equiv |x^2| \leqslant 1\}$, или $S_0 = \{(0, 0)\}$, или $S_0 = \{(1, 0)\}$, а $S_1 = \{x \in E^2, g(x) \equiv x^1 = 0\}$, или $S_1 = \{x \in E^2\colon g(x) \equiv |x|^2 = 4\}$ или $S_1 = \{x \in E^2\colon g(x) \equiv |x|^2 \leqslant 4\}$.

2. Применить принцип максимума к задаче: $J(u) = \int_0^1 |u(t)|^2\, dt \to \inf$;

$\dot{x}^1 = x^2$, $\dot{x}^2 = u^1(t)$, $\dot{x}^3 = x^4$, $\dot{x}^4 = u^2(t) - g$ $(0 \leqslant t \leqslant 1)$; $x(0) = (-1, 0, 0, 0)$. $x(T) = (0, 0, 0, 0)$. Здесь $x = (x^1, x^2, x^3, x^4)$, $u = (u^1, u^2)$; $g = \text{const} \geqslant 0$.

3. Применить принцип максимума к задаче о мягкой посадке космического корабля на Луну с минимальной затратой горючего ([308], с. 36, 44, 54): $J(u) = m(T) \to \sup$; $\dot{h}(t) = v(t)$, $\dot{v}(t) = -g + u(t)/m(t)$, $\dot{m}(t) = -ku(t)$, $u(t) \in V = \{u \in E^1\colon 0 \leqslant u \leqslant \alpha\}$ $(0 \leqslant t \leqslant T)$; $h(0) = h_0 > 0$,

$v(0) = v_0$, $m(0) = m_0 > 0$; $h(T) = 0$, $v(T) = 0$; момент T заранее не задан. Здесь $m(t)$ — масса корабля, $h(t)$ — высота, $v(t)$ — вертикальная скорость корабля над Луной, $u(t)$ — тяга двигателя, g — гравитационное ускорение Луны, $\alpha > 0$, $k > 0$ — постоянные (ср. пример 1.2).

4. Рассмотреть задачи из примеров 7—11 для малых колебаний маятника, считая $\sin x^1 \approx x^1$, $\cos x^1 \approx 1$. Найти решения краевых задач принципа максимума.

5. Рассмотреть задачи из примеров 1, 2 при дополнительном ограничении $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$.

6. Применить принцип максимума к задаче: $J(u) = x^2(T) \to \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant T)$; $x(0) = x_0$; момент $T > 0$ задан. Сколько решений имеет эта задача при $x_0 = 0$? $x_0 = T$?

7. Показать, что в задаче $J(u) = \int\limits_0^1 (x^2(t) - u^2(t))\,dt \to \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant 1)$; $x(0) = 0$, оптимальное управление не существует, $\inf J(u) = -1$ (см. пример 7.4.2). Что дает здесь применение принципа максимума?

8. Найти минимум функции $J(u) = \int\limits_0^1 \sin u(t)\,dt$ при условии $\dot{x} = \cos u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant \pi/2\}$ $(0 \leqslant t \leqslant 1)$; $x(0) = 0$, $x(1) = 1$. Показать, что $u(t) \equiv 0$ $(0 \leqslant t \leqslant 1)$, — оптимальное управление, и убедиться в том, что в принципе максимума здесь надо принять $a_0 = 0$.

9. Применить принцип максимума к задаче: $J(u) = \int\limits_0^T |x(t)|^2\,dt \to \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant T)$; $x(0) = 1$, $x(T) = 0$; момент $T > 0$ задан. Показать, что при $T > 1$ оптимальное управление: $u(t) \equiv -1$ $(0 \leqslant t < 1)$; $u(t) \equiv 0$ $(1 \leqslant t \leqslant T)$ на участке $1 \leqslant t \leqslant T$ является особым, т. е. его нельзя определить из принципа максимума.

10. Применить принцип максимума для $J(u) = \int\limits_0^1 u(t) \cos \frac{\pi x(t)}{2}\,dt \to \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant 1)$; $x(0) = 0$.

11. Показать, что задача: $J(u) = \int\limits_0^T u^2(t)(u(t) - 1)^2\,dt \to \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1\colon 0 \leqslant u \leqslant 1\}$ $(0 \leqslant t \leqslant T)$; $x(0) = 0$, $x(T) = 1$ при $T = 1$ имеет единственное решение, а при $T > 1$ — бесконечно много решений. Изменится ли этот вывод, если $V = E^1$? Если $x(T) = -1$?

12. С помощью принципа максимума исследовать задачу: $J(u) = |x^2(1)|^2 \to \inf$; $\dot{x}^1 = x^2$, $\dot{x}^2 = u(t)$, $u(t) \in V = \{u \in E^1\colon |u| \leqslant 1\}$, $x(0) = (x_0^1, x_0^2)$.

13. Пусть задача оптимального управления автономна (см. § 1), $(u(\cdot), x(\cdot))$ — ее решение, а $\psi(\cdot)$, $a_0, a_1, \ldots, a_s$, определены из принципа максимума (теоремы 1, 1). Показать, что тогда $H(x(t), u(t), \psi(t), \psi_0) \equiv \mathrm{const}$ $(t_0 \leqslant t \leqslant T)$. Указание: воспользоваться теоремой 3.

14. Сформулировать принцип максимума для задачи оптимального управления с параметрами:

$$J(u(\cdot), w) = \int\limits_{t_0}^T f^0(x(t), u(t), t, w)\,dt + g_0(x_0, x(T), t_0, T, w) \to \inf,$$

$$\dot{x}(t) = f(x(t), u(t), t, w), \quad t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0,$$
$$g^i(x_0, x(T), t_0, T, w) \leqslant 0, \quad i = 1, \ldots, m,$$
$$g^i(x_0, x(T), t_0, T, w) = 0, \quad i = m+1, \ldots, s,$$
$$u(t) \in V, \quad t_0 \leqslant t \leqslant T, \quad w \in W,$$

где $w = (w^1, \ldots, w^s)$ — параметры (они не зависят от времени), W — заданное множество из E^s; остальные обозначения см. выше. Рассмотреть случаи $W = E^s$ и $W = \{w\colon q_i(w) = 0, i = 1, \ldots, k\}$, где $q_i(w)$ — заданные гладкие функции на E^s, $i = 1, \ldots, k$. Указание: ввести новые переменные x^{n+i} посредством условий $\dot{x}^{n+i}(t) = 0$ $(t_0 \leqslant t \leqslant T)$, $x^{n+i}(t_0) = w^i$ $(i = 1, \ldots, k)$ в пространстве $(x^1, \ldots, x^{n+k})$ воспользоваться теоремой 1 или 2, а затем исключить переменные x^{n+i}, ψ_{n+i} $(i = 1, \ldots, k)$.

15. Сформулировать принцип максимума для задачи получающейся из задачи (1)—(4) или (37)—(40) добавлением условий (1.33).

Указание: ввести новые переменные x^{n+i} посредством условий $\dot{x}^{n+i}(t) = f_i^0(x(t), u(t), t)$ $(t_0 \leqslant t \leqslant T)$, $x^{n+i}(t_0) = 0$ $(i = 1, \ldots, s_2)$; $x^{n+i}(T) + g_i^0(x_0, x(T), t_0, T) \leqslant 0$ $(i = 1, \ldots, m_2)$; $x^{n+i}(T) + g_i^0(x_0, x(T), t_0, T) = 0$ $(i = m_2 + 1, \ldots, s_2)$ в пространстве $(x^1, \ldots, x^{n+s_2})$, воспользоваться теоремой 1 или 2, а затем исключить переменные x^{n+i}, ψ_{n+i} $(i = 1, \ldots, s_2)$.

§ 3. Доказательство принципа максимума

1. Доказательство теорем 2.1, 2.2 проведем при дополнительном предположении, что вектор-функция $f(x, u, t) = (f^1(x, u, t), \ldots, f^n(x, u, t))$ удовлетворяет условию Липшица по переменным (x, u), т. е.

$$|f(x, u, t) - f(y, v, t)| \leqslant L(|x - y| + |u - v|), \quad L > 0, \tag{1}$$

при всех $(x, u, t), (y, v, t) \in E^n \times E^n \times [t_0, T]$. Покажем, что тогда решение задачи Коши

$$\dot{x} = f(x, u(t), t), \quad t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0, \tag{2}$$

непрерывно зависит от начальной точки x_0 и управления $u = u(t)$. Сначала докажем одно утверждение, которое часто приводится в учебных пособиях по дифференциальным уравнениям и в литературе известно как неравенство Гронуолла.

Лемма 1. *Пусть функции $\varphi(t)$, $b(t)$ неотрицательны и непрерывны на отрезке $t_0 \leqslant t \leqslant T$, $a = \text{const}$. Пусть*

$$\varphi(t) \leqslant a \int_{t_0}^{t} \varphi(\tau)\, d\tau + b(t), \quad t_0 \leqslant t \leqslant T. \tag{3}$$

Тогда

$$0 \leqslant \varphi(t) \leqslant a \int_{t_0}^{t} b(\tau)\, e^{a(t-\tau)} d\tau + b(t), \quad t_0 \leqslant t \leqslant T. \tag{4}$$

В частности, если $b(t) \equiv b = \text{const} \geqslant 0$, то

$$0 \leqslant \varphi(t) \leqslant b e^{a(t-t_0)}, \quad t_0 \leqslant t \leqslant T. \tag{5}$$

Если же

$$\varphi(t) \leqslant a \int_{t}^{T} \varphi(\tau)\, d\tau + b(t), \quad t_0 \leqslant t \leqslant T, \tag{6}$$

то

$$0 \leqslant \varphi(t) \leqslant a \int_t^T b(\tau) e^{a(\tau-t)} d\tau + b(t), \quad t_0 \leqslant t \leqslant T, \tag{7}$$

а при $b(t) \equiv b = \text{const} \geqslant 0$ *будем иметь*

$$0 \leqslant \varphi(t) \leqslant b e^{a(T-t)}, \quad t_0 \leqslant t \leqslant T. \tag{8}$$

Доказательство. Положим $R(t) = a \int_{t_0}^t \varphi(\tau)\, d\tau$. Заметим, что $R(t_0) = 0$, $R(t) \geqslant 0$, $\dot{R}(t) = a\varphi(t)$ $(t_0 \leqslant t \leqslant T)$. С учетом (3) имеем

$$\dot{R}(t) \leqslant aR(t) + ab(t) \quad \text{или} \quad \dot{R}(t) - aR(t) \leqslant ab(t), \quad t_0 \leqslant t \leqslant T.$$

Умножая обе части последнего неравенства на $e^{-a(t-t_0)}$, получим

$$\frac{d}{dt}\left(R(t) e^{-a(t-t_0)}\right) \leqslant ab(t) e^{-a(t-t_0)}, \quad t_0 \leqslant t \leqslant T.$$

Интегрирование этого неравенства от t_0 до t с учетом $R(t_0) = 0$ приводит к оценке

$$R(t) e^{-a(t-t_0)} \leqslant a \int_{t_0}^t b(\tau) e^{-a(\tau-t_0)} d\tau,$$

или

$$R(t) \leqslant a \int_{t_0}^t b(\tau) e^{a(t-\tau)} d\tau, \quad t_0 \leqslant t \leqslant T.$$

Подставив эту оценку в правую часть (3), сразу получим требуемое неравенство (4). Если $b(t) = b = \text{const}$, то непосредственно вычисляя интеграл в правой части (4), придем к оценке (5).

Неравенства (7), (8), вытекающие из (6), доказываются аналогично с помощью вспомогательной функции $R(t) = a \int_t^T \varphi(\tau)\, d\tau$ $(t_0 \leqslant t \leqslant T)$. Лемма 1 доказана.

Теорема 1. *Пусть вектор-функция* $f(x, u, t)$ *непрерывна по совокупности переменных* $(x, u, t) \in E^n \times V \times [t_0, T]$, *удовлетворяет условию* (1). *Тогда*

$$\max_{t_0 \leqslant t \leqslant T} |x(t, v, y_0) - x(t, u, x_0)| \leqslant C_1 |y_0 - x_0| + C_2 \int_{t_0}^T |v(t) - u(t)|\, dt, \tag{9}$$

где $x(t, u, x_0)$ — *решение задачи Коши* (2), *соответствующее управлению* $u = u(t) \in L_\infty^r[t_0, T]$: $u(t) \in V$ $(t_0 \leqslant t \leqslant T)$ *и начальному условию* x_0; $C_1 = e^{L(T-t_0)}$, $C_2 = C_1 L$.

Доказательство. Существование траекторий $x(t, u, x_0)$, $x(t, v, y_0)$, $t_0 \leqslant t \leqslant T$, следует из теоремы 1.1. Обозначим для краткости $\Delta u(t) = v(t) - u(t)$, $\Delta x(t) = x(t, v, y_0) - x(t, u, x_0)$, $\Delta x_0 = y_0 - x_0$. Тогда из (1.12)

следует

$$\Delta x(t) = \int_{t_0}^{t} [f(x(\tau, v, y_0), v(\tau), \tau) - f(x(\tau, u, x_0), u(\tau), \tau)]\, d\tau + \Delta x_0.$$

Отсюда с учетом условия (1) имеем

$$|\Delta x(t)| \leqslant L \int_{t_0}^{t} |\Delta x(\tau)|\, d\tau + L \int_{t_0}^{T} |\Delta u(\tau)|\, d\tau + |\Delta x_0|.$$

Это неравенство запишется в виде (3), если принять $\varphi(t) = |\Delta x(t)|$, $a = L$, $b(t) \equiv b = L \int_{t_0}^{T} |\Delta u(t)|\, dt + |\Delta x_0|$. Отсюда и из леммы 1 следует оценка (9).

При доказательстве принципа максимума для задачи с незакрепленным временем наряду с задачей (2) нам ниже понадобится также задача Коши, записанная в виде

$$\dot{x} = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T, \quad x(T) = x_1. \tag{10}$$

При выполнении условий теоремы 1 для решения $x(t, u, x_1)$ задачи (10) справедлива оценка

$$\max_{t_0 \leqslant t \leqslant T} |x(t, v, y_1) - x(t, u, x_1)| \leqslant C_1 |y_1 - x_1| + C_2 \int_{t_0}^{T} |v(t) - u(t)|\, dt, \tag{11}$$

которая доказывается так же, как (9), но с использованием неравенств (6)—(8); постоянные C_1, C_2 в (11) те же, что и в (9).

Более тонкие теоремы о непрерывной зависимости решения задач (2), (10) от исходных данных, справедливые без дополнительного требования (1) и удобные для использования при доказательстве принципа максимума, читатель найдет, например, в [2, 77, 104].

2. Приступим к доказательству теорем 2.1, 2.2. Приводимое ниже простое и изящное доказательство этих теорем принадлежит А. В. Арутюнову и М. Д. Марданову [206].

Доказательство теоремы 2.1 будет состоять из трех этапов. Вначале исходная задача (2.1)—(2.4) аппроксимируется семейством конечномерных задач. Затем к конечномерной задаче будет применен метод штрафных функций для учета ограничений на концах траекторий и выведено необходимое условие оптимальности для получившейся штрафной задачи. Наконец, будет совершен предельный переход в полученных необходимых условиях и установлена справедливость принципа максимума.

Сначала доказательство проведем в предположении, что решение $(x_{0*}, u_*(t), x_*(t))$ задачи (2.1)—(2.4) единственно. Через $t_1, t_2, \ldots$ обозначим каким-либо образом занумерованные рациональные точки интервала (t_0, T), являющиеся точками непрерывности оптимального управления $u_*(t)$. Выберем во множестве V всюду плотную последовательность точек $v_1, v_2, \ldots$ Зафиксируем произвольный номер $N \geqslant 1$ и для любого натурального числа $l \leqslant N$ выберем такие точки $t_{lp} = t_{lp}(N) \in (t_0, T)$ $(p = 1, \ldots, N+1)$, что

$$\begin{gathered} t_l = t_{l1} < \ldots < t_{lp} < \ldots < t_{lN+1}, \quad t_{lN+1} - t_l \leqslant 1/N, \\ [t_l, t_{lN+1}] \cap [t_k, t_{kN+1}] = \varnothing \qquad \forall\, l, k, l \neq k, \quad 1 \leqslant l, k \leqslant N, \end{gathered} \tag{12}$$

причем оптимальное управление $u(t)$ непрерывно на отрезках $[t_l, t_{lN+1}]$ $(l = 1, \ldots, N)$. Обозначим через ξ квадратную матрицу $\xi = \{\xi_{lp}\}$ размер-

ности N, элементы которой удовлетворяют ограничению

$$0 \leqslant \xi_{lp} \leqslant d_N = \min_{1 \leqslant l, p \leqslant N} (t_{lp+1} - t_{lp}).$$

Для каждой такой матрицы ξ определим управление $u(\cdot, \xi)$ следующим образом:

$$u(t, \xi) = \begin{cases} v_p, & t_{lp} < t \leqslant t_{lp} + \xi_{lp}, \quad 1 \leqslant l, \ p \leqslant N, \\ u_*(t) & \text{в остальных точках отрезка } [t_0, T]. \end{cases} \tag{13}$$

Нетрудно видеть, что управление (13) кусочно-непрерывно, $u(t, \xi) \in V$ при всех $t \in [t_0, T]$. Рассмотрим задачу: минимизировать функцию

$$J_N(x_0, \xi) \equiv J(x_0, u(\cdot, \xi)) = \int_{t_0}^{T} f^0(x(t), u(t, \xi), t)\, dt + g^0(x_0, x(T)) \tag{14}$$

при условиях:

$$\dot{x}(t) = f(x(t), u(t, \xi), t), \quad t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0, \tag{15}$$

$$\begin{aligned} g^i(x_0, x(T)) &\leqslant 0, \quad i = 1, \ldots, m, \\ g^i(x_0, x(T)) &= 0, \quad i = m+1, \ldots, s, \end{aligned} \tag{16}$$

$$\xi = \{\xi_{lp}\}: \quad 0 \leqslant \xi_{lp} \leqslant d_N, \quad l, p = 1, 2, \ldots, N, \tag{17}$$

где управление $u = u(t, \xi)$ определяется согласно (13). Подчеркнем, что при каждом фиксированном $N \geqslant 1$ задача (14)—(17) является конечномерной задачей минимизации в пространстве переменных (x_0, ξ): $x_0 = (x_0^1, \ldots, x_0^n)$, $\xi = \{\xi_{lp}, l, p = 1, \ldots, N\}$.

Нетрудно видеть, что любой набор $(x_0, u(\cdot, \xi))$, допустимый в задаче (14)—(17), является допустимым и в задаче (2.1)—(2.4), причем значения целевых функций в обеих задачах на таком наборе совпадают. Отсюда следует, что нижняя грань J_{N_*} целевой функции задачи (14)—(17) не меньше нижней грани J_* в задаче (2.1)—(2.4): $J_{N_*} \geqslant J_*$. В то же время оптимальный набор $(x_{0*}, u_*(\cdot))$ задачи (2.1)—(2.4) является допустимым в задаче (14)—(17), так как согласно (13) $u(t, 0) = u_*(t)$ $(t_0 \leqslant t \leqslant T)$. Тогда $J_* = J(x_{0*}, u_*(\cdot)) = J(x_{0*}, u(\cdot, \cdot)) = J_N(x_{0*}, 0) \geqslant J_{N_*}$. Следовательно, $J_* = J_{N_*} = J_N(x_{0*}, 0)$. По предположению задача (2.1)—(2.4) имеет единственное решение. Тогда и задача (14)—(17) также будет иметь единственное решение $(x_{0*}, \xi_* = 0)$ при каждом фиксированном $N \geqslant 1$.

Применим к задаче (14)—(17) метод штрафных функций. Обозначим

$$g_{iN}(x_0, \xi) = g_i(x_0, x(T)) = g_i(x_0, x(T, u(\cdot, \xi), x_0), \qquad i = 1, \ldots, s,$$

$$g_i^+(x, y) = \begin{cases} \max\{0;\ g^i(x, y)\}, & i = 1, \ldots, m, \\ g^i, & i = m+1, \ldots, s. \end{cases} \tag{18}$$

Рассмотрим задачу: минимизировать функцию

$$\Phi_k(x_0, \xi) = J_N(x_0, \xi) + A_k \sum_{i=1}^{s} \left(g_{iN}^+(x_0, \xi)\right)^2 =$$

$$= \int_{t_0}^{T} f^0(x(t), u(t, \xi), t)\, dt + g^0(x_0, x(T)) + A_k \sum_{i=1}^{s} \left(g_i^+(x_0, x(T))\right)^2 \tag{19}$$

при условиях

$$\dot{x}(t) = f(x(t), u(t, \xi), t), \ t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0, \tag{20}$$

$$\xi = \{\xi_{lp}\}, \ 0 \leqslant \xi_{lp} \leqslant d_N, \quad l, p = 1, \ldots, N, \tag{21}$$

$$|x_0 - x_{0*}| \leqslant 1, \tag{22}$$

где $A_k > 0$ $(k = 0, 1, \ldots)$; $\{A_k\} \to \infty$. Если ввести множество

$$U_0 = \{(x_0, \xi): |x_0 - x_{0*}| \leqslant 1, \quad 0 \leqslant \xi_{lp} \leqslant d_N, \quad l, p = 1, \ldots, N\},$$

то задача (19)—(22) может быть кратко записана в виде

$$\Phi_k(x_0, \xi) \to \inf, \qquad (x_0, \xi) \in U_0. \tag{23}$$

Покажем, что задача (23) имеет хотя бы одно решение. Сначала убедимся, что функции $J_N(x_0, \xi)$, $g_{iN}(x_0, \xi)$ непрерывны по совокупности (x_0, ξ) на множестве U_0. Пусть (x_0, ξ), $(x_0 + \Delta x_0, \xi + \Delta\xi) \in U_0$. Согласно оценке (9) с учетом (13) имеем

$$\max_{t_0 \leqslant t \leqslant T} |x(t, u(\cdot, \xi + \Delta\xi), x_0 + \Delta x_0) - x(t, u(\cdot, \xi), x_0| \leqslant$$

$$\leqslant C_1 |\Delta x_0| + C_2 \int_{t_0}^{T} |u(t, \xi + \Delta\xi) - u(t, \xi)| \, dt =$$

$$= C_1 |\Delta x_0| + C_2 \sum_{l,p=1}^{N} \int_{t_{lp}+\xi_{lp}}^{t_{lp}+\xi_{lp}+\Delta\xi_{lp}} |v_p - u_*(t)| \, dt \leqslant C_1 |\Delta x_0| + C_{2N} |\Delta\xi|, \tag{24}$$

где

$$C_{2N} = C_2 \sup_{1 \leqslant p \leqslant N} \sup_{t_0 \leqslant t \leqslant T} |v_p - u_*(t)|, \quad |\Delta\xi| = \sum_{l,p=1}^{N} |\Delta\xi_{lp}|.$$

Из непрерывности функций $f^0(x, u, t)$, $g^i(x, y)$ по совокупности своих аргументов и оценки (24) следует непрерывность функций $J_N(x_0, \xi)$, $g_{iN}(x_0, \xi)$, $g_{iN}^+(x_0, \xi)$, $\Phi_k(x_0, \xi)$ на компактном множестве U_0. Согласно теореме 1.1.1 $\Phi_{k*} = \inf\limits_{U_0} \Phi_k(x_0, \xi) > -\infty$ и существует хотя бы одна точка $(x_{0k}, \xi_k) \in U_0$, в которой нижняя грань достигается. Покажем, что последовательность решений $(x_{0k}, \xi_k = \{\xi_{lp}^k\})$ $(k = 0, 1, \ldots)$ решений задач (23) сходится к решению $(x_{0*}, \xi_* = 0)$ задачи (14)—(17). Воспользуемся теоремами 5.14.1, 5.14.2. Проверим выполнение условий этих теорем. Непрерывность функций $J_N(x_0, \xi)$, $g_{iN}(x_0, \xi)$ мы уже установили. Из непрерывности $J_N(x_0, \xi)$ и компактности U_0 следует, что $J_{**} = \inf\limits_{U_0} J_N(x_0, \xi) > -\infty$. Множество $U_\delta = \{(x_0, \xi) \in U_0: g_{iN}^+(x_0, \xi) \leqslant \delta, \ i = 1, \ldots, s\}$ ограничено в силу ограниченности U_0 при любом $\delta > 0$. Все условия теорем 5.14.1, 5.14.2 выполнены. Поэтому решение задачи (23) или (19)—(22) сходится к решению задачи (14)—(17) как по функции, так и по аргументу. Поскольку задача (14)—(17) имеет единственное решение $(x_{0*}, \xi_* = 0)$, то

$$\lim_{k \to \infty} \xi_k = \xi_* = 0, \quad \lim_{k \to \infty} x_{0k} = x_{0*}, \quad \lim_{k \to \infty} \Phi_k(x_{0k}, \xi_k) = J_{N*} = J_*. \tag{25}$$

Из оценки (24) при $\xi = \xi_* = 0$, $x_0 = x_{0*}$, $\Delta\xi = \xi_k$, $\Delta x_0 = x_{0k} - x_{0*}$ получаем

$$\lim_{k \to \infty} \max_{t_0 \leqslant t \leqslant T} |x_k(t) - x_*(t)| = 0, \tag{26}$$

где $x_*(t) = x(t, u_*(\cdot), x_{0*}) = x(t, u(\cdot, 0), x_{0*})$ — оптимальная траектория в задачах (14)—(17), (2.1)—(2.4), $x_k(t) = x(t, u(\cdot, \xi_k), x_{0k})$ — оптимальная траектория в задаче (19)—(22). Кроме того, из теорем 5.14.1, 5.14.2 следует

$$\lim_{k\to\infty} g_i^+(x_{0k}, x_k(T)) = g_i^+(x_{0*}, x_*(T)) = 0, \quad i = 1, \ldots, s. \tag{27}$$

Задача (19)—(22) представляет собой задачу оптимального управления с подвижным левым концом и свободным правым концом траектории и поэтому для приращения функции (19) нетрудно получить формулу, которая понадобится нам при получении необходимых условий оптимальности для задачи (19)—(22). Будем рассматривать задачи (19)—(22) со столь большими номерами k, чтобы

$$|x_{0k} - x_{0*}| < 1, \quad 0 \leqslant \xi_{lp}^k < d_N, \quad l, p = 1, 2, \ldots, N; \tag{28}$$

это возможно в силу равенств (25), из которых вытекает, что неравенства (28) будут выполняться для всех $k \geqslant k_0$, где k_0 — достаточно большое число. Оптимальному решению (x_{0k}, ξ_k) задачи (19)—(22) в силу (28) можно дать такие малые ненулевые приращения $(\Delta x_0, \Delta\xi)$, что

$$|x_{0k} + \Delta x_0 - x_{0*}| < 1, \quad 0 \leqslant \xi_{lp}^k + \Delta\xi_{lp} < d_N,$$
$$\Delta\xi_{lp} \geqslant 0, \qquad l, p = 1, \ldots, N, \ k \geqslant k_0. \tag{29}$$

Тогда оптимальные управления $u_k(t) = u(t, \xi_k)$ и траектория $x_k(t) = x(t, u_k(\cdot), x_{0k})$ задачи (19)—(22) получат приращение $\Delta u(t) = u(t, \xi_k + \Delta\xi_k) - u_k(t)$, $\Delta x(t) = x(t, u_k + \Delta u, x_{0k} + \Delta x_0) - x_k(t)$ $(t_0 \leqslant t \leqslant T)$. Приращение $\Delta x(t)$ удовлетворяет условиям

$$\Delta\dot{x}(t) = \Delta f = f(x_k(t) + \Delta x(t), \quad u_k(t), t) - f(x_k(t), \quad u_k(t), t), \ t_0 \leqslant t \leqslant T,$$
$$\Delta x(t_0) = \Delta x_0, \tag{30}$$

вытекающим из (20). Из (24) для $\Delta x(t)$ следует оценка

$$\max_{t_0 \leqslant t \leqslant T} |\Delta x(t)| \leqslant c_1 |\Delta x_0| + c_{2N} |\Delta\xi|. \tag{31}$$

С учетом оптимальности (x_{0k}, ξ_k) для приращения функции (19) получим

$$0 \leqslant \Delta\Phi_k = \Phi_k(x_{0k} + \Delta x_0, \xi_k + \Delta\xi) - \Phi_k(x_{0k}, \xi_k) =$$
$$= \int_{t_0}^{T} \left[f^0(x_k(t) + \Delta x(t), u_k(t) + \Delta u(t), t) - f^0(x_k(t), u_k(t), t)\right] dt +$$
$$+ g^0(x_{0k} + \Delta x_0, x_k(T) + \Delta x(T)) - g^0(x_{0k}, x_k(T)) +$$
$$+ A_k \sum_{i=1}^{s} \left[\left(g_i^+(x_{0k} + \Delta x_0, x_k(T) + \Delta x(T))\right)^2 - \left(g_i^+(x_0, x_k(T))\right)^2\right] =$$
$$= \int_{t_0}^{T} \Delta f^0 dt + \left\langle g_x^0(x_{0k}, x_k(T)) + \sum_{i=1}^{s} 2A_k g_i^+(x_{0k}, x_k(T))\, g_x^i(x_{0k}, x_k(T)), \Delta x_0 \right\rangle +$$
$$+ \left\langle g_y^0(x_{0k}, x_k(T)) + \sum_{i=1}^{s} 2A_k g_i^+(x_{0k}, x_k(T))\, g_y^i(x_{0k}, x_k(T)), \Delta x(T) \right\rangle + R_{1k}, \tag{32}$$

где

$$\Delta f^0 = f^0(x_k(t) + \Delta x(t),\quad u_k(t) + \Delta u(t), t) - f^0(x_k(t),\ u_k(t), t),$$

$$R_{1k} = \left\langle \left(g_x^0\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) - g_x^0(\ldots)\right) + \sum_{i=1}^{s} 2A_k\left(g_i^+\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) g_x^i\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) - \right.\right.$$

$$\left.\left. - g_i^+(\ldots)\, g_x^i(\ldots)\right), \Delta x_0\right\rangle + \left\langle \left(g_y^0\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) - g_y^0(\ldots)\right) + \right.$$

$$\left. + \sum_{i=1}^{s} 2A_k\left(g_i^+\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) g_y^i\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) - g_i^+(\ldots)\, g_y^i(\ldots)\right), \Delta x(T)\right\rangle; \tag{33}$$

для краткости здесь обозначено $\left(\begin{smallmatrix}\theta\\ \ldots\end{smallmatrix}\right) = (x_{0k} + \theta\Delta x_0,\ x_k(T) + \theta\Delta x(T))\,(0 < \theta < 1)$, $(\ldots) = (x_{0k}, x_k(T))$.

Введем обозначения

$$a_{0k} = \left[1 + \sum_{i=1}^{s}\left(2A_k g_i^+(x_{0k}, x_k(T))\right)^2\right]^{-1/2},\quad a_{ik} = 2A_k g_i^+(x_{0k}, x_k(T))\, a_{0k},$$

$$i = 1, \ldots, s. \tag{34}$$

Отсюда и из (18) следует, что

$$0 < a_{0k} \leqslant 1,\quad a_{1k} \geqslant 0, \ldots,\quad a_{mk} \geqslant 0,\quad \sum_{i=0}^{s} a_{ik}^2 = 1. \tag{35}$$

Для преобразования правой части равенства (32) к более удобному виду, введем функцию Гамильтона — Понтрягина

$$H(x, u, t, \psi, a_0) = -a_0 f^0(x, u, t) + \langle \psi, f(x, u, t)\rangle \tag{36}$$

и сопряженную задачу

$$\dot{\psi}_k(t) = -H_x(x, u, t, \psi, a_0)\,|_{x=x_k(t), u=u_k(t), \psi=\psi_k, a_0=a_{0k}}, \tag{37}$$

$$\psi_k(T) = -\sum_{i=0}^{s} a_{ik} g_y^i(x_{0k}, x_k(T)), \tag{38}$$

где a_{ik} взяты из (34). Напоминаем, что в соответствии с определением 1.1 решением задачи (37), (38) называется непрерывная вектор-функция $\psi_k(t)$, являющаяся непрерывным решением интегрального уравнения

$$\psi_k(t) = \int_t^T H_x(x_k(\tau), u_k(\tau), \tau, \psi_k(\tau), a_{0k})\, d\tau + \psi_k(T). \tag{39}$$

Система (37) линейна относительно ψ_k; существование и единственность решения задачи (37), (38) следует из теоремы 1.2. При сделанных предположениях относительно функций $f^i(x, u, t)$ $(i = 0, 1, \ldots, n)$, как видно из (39), функция $\psi_k(t)$ во всех точках непрерывности управления $u_k(t)$ непрерывно дифференцируема и удовлетворяет уравнению (37).

Умножим равенство (32) на $a_{0k} > 0$ и с учетом обозначений (34), (38) перепишем его в виде

$$0 \leqslant a_{0k}\Delta\Phi_k = \int_{t_0}^{T} a_{0k}\Delta f^0\, dt + \left\langle \sum_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)), \Delta x_0\right\rangle -$$

$$- \langle \psi_k(T), \Delta x(T)\rangle + a_{0k} R_{1k}. \tag{40}$$

Преобразуем третье слагаемое из правой части (40). С учетом соотношений (30), (37), (38) имеем

$$\langle \psi_k(T), \Delta x(T)\rangle = \int_{t_0}^{T} \frac{d}{dt}\langle \psi_k(t), \Delta x(t)\rangle\, dt + \langle \psi_k(t_0), \Delta x_0\rangle =$$

$$= \int_{t_0}^{T} (\langle \psi_k(t), \Delta \dot{x}(t)\rangle + \langle \dot{\psi}_k(t), \Delta x(t)\rangle)\, dt + \langle \psi_k(t_0), \Delta x_0\rangle =$$

$$= \int_{t_0}^{T} \langle \psi_k(t), \Delta f\rangle\, dt - \int_{t_0}^{T} \langle H_x(x_k(t), u_k(t), t, \psi_k(t), a_{0k}), \Delta x(t)\rangle\, dt +$$

$$+ \langle \psi_k(t_0), \Delta x_0\rangle.$$

Подставим полученное выражение в (40); с учетом обозначения (36) будем иметь

$$0 \leqslant a_{0k}\Delta\Phi_k = -\int_{t_0}^{T} [H(x_k(t) + \Delta x(t), u_k(t) + \Delta u(t), t, \psi_k(t), a_{0k}) -$$

$$- H(x_k(t), u_k(t), t, \psi_k(t), a_{0k})]\, dt + \left\langle \sum_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)) - \psi_k(t_0), \Delta x_0 \right\rangle +$$

$$+ \int_{t_0}^{T} \langle H_x(x_k(t), u_k(t), t, \psi_k(t), a_{0k}), \Delta x(t)\rangle\, dt + a_{0k} R_{1k}. \qquad (41)$$

В силу формулы конечных приращений

$$H(x_k + \Delta x, u_k + \Delta u, t, \psi_k, a_{0k}) = H(x_k, u_k + \Delta u, t, \psi_k, a_{0k}) +$$
$$+ \langle H_x(x_k + \theta_1 \Delta x, u_k + \Delta u, t, \psi_k, a_{0k}), \Delta x\rangle, \quad 0 < \theta_1 < 1.$$

Отсюда и из (41) получим

$$0 \leqslant a_{0k}\Delta\Phi_k = -\int_{t_0}^{T} [H(x_k(t), u_k(t) + \Delta u(t), t, \psi_k(t), a_{0k}) -$$

$$- H(x_k(t), u_k(t), t, \psi_k(t), a_{0k})]\, dt + \left\langle \sum_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)) - \right.$$

$$\left. - \psi_k(t_0), \Delta x_0 \right\rangle + R_k, \quad R_k = a_{0k} R_{1k} + R_{2k}, \qquad (42)$$

где

$$R_{2k} = -\int_{t_0}^{T} \langle H_x(x_k(t) + \theta_1 \Delta x(t), u_k(t) + \Delta u(t), t, \psi_k(t), a_{0k}) -$$

$$- H_x(x_k(t), u_k(t), t, \psi_k(t), a_{0k}), \quad \Delta x(t)\rangle\, dt. \qquad (43)$$

Покажем, что

$$R_k = o_k(|\Delta x_0| + |\Delta \xi|), \quad \lim_{t \to 0} o_k(t)/t = 0. \qquad (44)$$

Из непрерывности функций $g^i(x, y)$, $g_i^+(x, y)$, $g_x^i(x, y)$, $g_y^i(x, y)$, оценки (31) и выражения (33) для R_{1k} следует, что $a_{0k}R_{1k} = o_k(|\Delta x_0| + |\Delta\xi|)$. Далее, перепишем выражение (43) для R_{2k} в виде $R_{2k} = R_{3k} + R_{4k}$, где

$$
\begin{aligned}
R_{3k} = -\int_{t_0}^{T} \langle H_x(x_k(t) + \theta_1\Delta x(t),\ u(t, \xi_k + \Delta\xi),\ t,\ \psi_k(t),\ a_{0k}) - \\
- H_x(x_k(t),\ u(t, \xi_k + \Delta\xi),\ t,\ \psi_k(t),\ a_{0k}),\ \Delta x(t)\rangle\, dt, \\
R_{4k} = -\int_{t_0}^{T} \langle H_x(x_k(t),\ u(t, \xi_k + \Delta\xi),\ t,\ \psi_k(t),\ a_{0k}) - \\
- H_x(x_k(t),\ u(t, \xi_k),\ t,\ \psi_k(t),\ a_{0k}),\ \Delta x(t)\rangle\, dt.
\end{aligned}
\tag{45}
$$

Напомним, что управления $u(t, \xi_k)$, $u(t, \xi_k + \Delta\xi)$ определены согласно (13). Отсюда и из неравенств (21), (22), (28), (29), (31) следует, что аргументы x, u, t, ψ функции H_x в формулах (45) принадлежат ограниченному замкнутому множеству $Q_{kN} = \{(x, u, t, \psi)\colon |x| \leqslant \max\limits_{t_0 \leqslant t \leqslant T} |x_k(t)| + c_1(|x_{0*}| + 1) + c_{2N}N^2 d_N,\quad |u| \leqslant \sup\limits_{t_0 \leqslant t \leqslant T} |u_*(t)| + \max\limits_{1 \leqslant p \leqslant N} |v_p|,\ t_0 \leqslant t \leqslant T,\quad |\psi| \leqslant \max\limits_{t_0 \leqslant t \leqslant T} |\psi_k(t)|$. Непрерывная функция $H_x(x, u, t, \psi, a_{0k})$ переменных (x, u, t, ψ) на компактном множестве Q_{kN} будет равномерно непрерывна на этом множестве. Отсюда и из оценки (31) следует, что $R_{3k} = o_k(|\Delta x_0| + |\Delta\xi|)$. Далее, с учетом определения (13) управлений $u(t, \xi_k)$, $u(t, \xi_k + \Delta\xi)$ имеем

$$
\begin{aligned}
|R_{4k}| = \Bigg| \sum_{l,p=1}^{N} \int_{t_{lp}+\xi_{lp}^k}^{t_{lp}+\xi_{lp}^k+\Delta\xi_{lp}} \langle H_x(x_k(t),\ v_p,\ t,\ \psi_k(t),\ a_{0k}) - \\
- H_x(x_k(t),\ u_*(t),\ t,\ \psi_k(t),\ a_{0k}),\ \Delta x(t)\rangle\, dt \Bigg| \leqslant \\
\leqslant 2|\Delta\xi| \sup_{Q_{kN}} |H_x(x, u, t, \psi, a_{0k})|\, |\Delta x(t)| = o_k(|\Delta x_0| + |\Delta\xi|).
\end{aligned}
$$

Суммируя полученные оценки для R_{1k}, R_{3k}, R_{4k}, придем к оценке (44).

Итак, нужная формула (42) для приращения функции (19) с оценкой остаточного члена (44) получена. Положим $\Delta\xi = 0$, $\Delta x_0 = -\varepsilon\left(\sum\limits_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)) - \psi_k(t_0)\right)$, где число $\varepsilon > 0$ столь мало, что выполняется первое из неравенств (29); тогда $\Delta u(t) \equiv 0$ и из (42), (44) получим $0 \leqslant a_{0k}\Delta\Phi_k = -\varepsilon \left|\sum\limits_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)) - \psi_k(t_0)\right|^2 + o_k(\varepsilon)$ или

$$
\left|\sum_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)) - \psi_k(t_0)\right|^2 \leqslant \frac{o_k(\varepsilon)}{\varepsilon}.
$$

Отсюда при $\varepsilon \to 0$ имеем

$$
\psi_k(t_0) = \sum_{i=0}^{s} a_{ik} g_x^i(x_{0k}, x_k(T)), \qquad k \geqslant k_0. \tag{46}
$$

Далее, положим в (42) $\Delta x_0 = 0$, матрицу $\Delta\xi = \{\xi_{lp}\}$ возьмем так, чтобы элемент, находящийся на пересечении произвольным образом фиксированных l-й строки и p-го столбца, $1 \leqslant l,\ p \leqslant N$, равнялся $\varepsilon > 0$, а остальные элементы равны нулю, причем ε возьмем столь малым, чтобы выполнялось второе неравенство (29). Тогда согласно (13)

$$\Delta u(t) = u(t, \xi_k + \Delta\xi) - u(t, \xi_k) = \begin{cases} v_p - u_k(t) = v_p - u_*(t) \\ \text{при } t_{lp} + \xi_{lp} < t \leqslant t_{lp} + \xi_{lp}^k + \varepsilon, \\ 0 \text{ в остальных точках из } [t_0, T], \end{cases}$$

и из (42), (44) получим

$$0 \leqslant a_{0k}\Delta\Phi_k = -\int\limits_{t_{lp}+\xi_{lp}^k}^{t_{lp}+\xi_{lp}^k+\varepsilon} [H(x_k(t), v_p, t, \psi_k(t), a_{0k}) - \\ - H(x_k(t), u_*(t), t, \psi_k(t), a_{0k})]\, dt + o_k(\varepsilon). \quad (47)$$

Заметим, что подынтегральная функция $g_k(t) = H(x_k(t), v_p, t, \psi_k(t), a_{0k}) - H(x_k(t), u_*(t), t, \psi_k(t), a_{0k})$ непрерывна на отрезке $\left[t_{lp} + \xi_{lp}^k, t_{lp} + \xi_{lp}^k + \varepsilon\right]$. Применяя теорему о среднем, из (47) имеем $0 \leqslant -g_k\left(t_{lp} + \xi_{lp}^k + \theta_2\varepsilon\right)\varepsilon + o_k(\varepsilon)$ $(0 < \theta_2 < 1)$. Разделим это неравенство на $\varepsilon > 0$ и совершим предельный переход при $\varepsilon \to 0$. Получим $0 \leqslant -g_k\left(t_{lp} + \xi_{lp}^k\right)$ или

$$[H(x_k(t), v_p, t, \psi_k(t), a_{0k}) - H(x_k(t), u_*(t), t, \psi_k(t), a_{0k})]_{t=t_{lp}+\xi_{lp}^k} \leqslant 0 \quad (48)$$

при всех $k \geqslant k_0$, $l = 1, \ldots, N$, $p = 1, \ldots, N$.

Заметим, что соотношения (35), (37), (38), (46), (48), представляющие собой необходимое условие оптимальности в задаче (19) — (22), вполне аналогичны соотношениям (2.7) — (2.10) из теоремы 2.1. Соотношения (35), (37), (38), (46), (48) получены при фиксированном $N \geqslant 1$ и справедливы при каждом $k \geqslant k_0$. Для завершения доказательства теоремы 2.1 остается сначала перейти в этих соотношениях к пределу при $k \to \infty$, считая N фиксированным, затем совершить предельный переход при $N \to \infty$. Поскольку построенные выше последовательности $\{a_k\}$, $\{\psi_k(t)\}$ зависят от N, то их предельные точки, вообще говоря, также будут зависеть от N. В дальнейшем нам будет полезно явно подчеркнуть эту зависимость и поэтому упомянутые последовательности и их предельные точки ниже будем снабжать индексом N.

Согласно (35) последовательность $a_k = (a_{0k}, a_{1k}, \ldots, a_{sk}) = a_k(N)$ $(k = 0, 1, \ldots)$, ограничена и, пользуясь теоремой Больцано — Вейерштрасса, из нее можно выбрать подпоследовательность, сходящуюся к некоторой точке $a = a(N) = (a_0(N), a_1(N), \ldots, a_k(N))$. Без умаления общности дальнейших рассуждений можем считать, что сама последовательность $\{a_k(N)\}$ сходится к $a(N)$. Из (35) следует

$$0 \leqslant a_0(N) \leqslant 1,\ a_1(N) \geqslant 0, \ldots, a_m(N) \geqslant 0, \quad |a(N)|^2 = \sum_{i=0}^{s} a_i^2(N) = 1. \quad (49)$$

Далее, пользуясь непрерывностью $g_y^i(x, y)$ и равенствами (25), (26), из (38) при $k \to \infty$ получим

$$\lim_{k\to\infty} \psi_k(T; N) = \psi(T; N) = -\sum_{i=0}^{s} a_i(N)\, g_y^i(x_{0*}, x_*(T)). \quad (50)$$

Покажем, что последовательность $\{\psi_k(t)\} = \{\psi_k(t, N)\}$ равномерно на $[t_0, T]$ сходится к решению $\psi(t; N)$ системы уравнений

$$\dot{\psi}(t; N) = -H_x(x_*(t), u_*(t), t, \psi(t; N), a_0(N)), \quad t_0 \leqslant t \leqslant T, \tag{51}$$

с начальным условием (50). Как и в задаче (37), (38) решение задачи (51), (50) существует, единственно, является непрерывным решением интегрального уравнения

$$\psi(t; N) = \int_t^T H_x(x_*(\tau), u_*(\tau), \tau, \psi(\tau; N), a_0(N))\, d\tau + \psi(T; N), \tag{52}$$

во всех точках непрерывности оптимального управления $u_*(t)$ непрерывно дифференцируемо и удовлетворяет уравнению (51).

Обозначим $\Delta\psi_k(t) = \psi_k(t; N) - \psi(t; N)$ $(t_0 \leqslant t \leqslant T)$. Из (39), (52) с учетом определения (36) функции H и ее производной H_x имеем

$$\Delta\psi_k(t) = \int_t^T \sum_{i=1}^n \Delta\psi_{ik}(\tau) f_x^i(x_k(\tau), u_k(\tau), \tau)\, d\tau + b_k(t; N) + \Delta\psi_k(T), \tag{53}$$

где

$$\begin{aligned} b_k(t; N) = &-\int_t^T (a_{0k}(N) - a_0(N)) f_x^0(x_k(\tau), u_k(\tau), \tau)\, d\tau - \\ &- a_0(N) \int_t^T \left[f_x^0(x_k(\tau), u_k(\tau), \tau) - f_x^0(x_*(\tau), u_k(\tau), \tau)\right] d\tau - \\ &- a_0(N) \int_t^T \left[f_x^0(x_*(\tau), u_k(\tau), \tau) - f_x^0(x_*(\tau), u_*(\tau), \tau)\right] d\tau + \\ &+ \int_t^T \sum_{i=1}^n \psi_i(\tau; N) \left[f_x^i(x_k(\tau), u_k(\tau), \tau) - f_x^i(x_*(\tau), u_k(\tau), \tau)\right] d\tau + \\ &+ \int_t^T \sum_{i=1}^n \psi_i(\tau; N) \left[f_x^i(x_*(\tau), u_k(\tau), \tau) - f_x^i(x_*(\tau), u_*(\tau), \tau)\right] d\tau. \end{aligned} \tag{54}$$

Покажем, что $\{b_k(t; N)\} \to 0$ при $k \to \infty$ (N фиксировано!) равномерно на отрезке $[t_0, T]$. С этой целью заметим, что в силу (21), (22), (28), оценки (24) при $\xi = 0$, $\Delta\xi = \xi_k$ и определения (13) управления $u_k(t) = u(\cdot, \xi_k)$ аргументы (x, u, t) функций $f_x^i(x, u, t)$, входящих в (53), (54), принадлежат компактному множеству $Q_N = \Big\{(x, u, t)\colon\ |x| \leqslant \max\limits_{t_0 \leqslant t \leqslant T} |x_*(t)| + c_1(|x_{0*}| + 1) + c_{2N} N^2 d_N,\ |u| \leqslant \sup\limits_{t_0 \leqslant t \leqslant T} |u_*(t)| + \sup\limits_{1 \leqslant p \leqslant N} |v_p|,\ t_0 \leqslant t \leqslant T\Big\}$ при всех $k \geqslant k_0$. Непрерывные функции $f_x^i(x, u, t)$ на Q_N будут ограничены и равномерно непрерывны по совокупности аргументов $(x, u, t) \in Q_N$. Следовательно, $\max\limits_{0 \leqslant i \leqslant n} \max\limits_{(x,u,t) \in Q_N} \left|f_x^i(x, u, t)\right| = L_N < \infty$. Отсюда и из $\{a_{0k}(N)\} \to a_0(N)$ получаем, что 1-е слагаемое из правой части (54) стремится к нулю равномерно на $[t_0, T]$. Равномерная сходимость к нулю 2-го

слагаемого из (54) следует из равномерной непрерывности $f^0(x, u, t)$ на Q_N и равномерной сходимости $\{x_k(t)\}$ к $x_*(t)$, вытекающей из оценки (24) при $\xi = 0$, $\Delta\xi = \xi_k$. Для 3-го слагаемого из (54) с учетом определения (13) управлений $u_k(t) = u(t, \xi_k)$, $u_*(t) = u(t; 0)$ имеем

$$\left| a_0(N) \int_t^T \left[f_x^0(x_*(\tau), u_k(\tau), \tau) - f_x^0(x_*(\tau), u_*(\tau), \tau \right] d\tau \right| =$$

$$= a_0(N) \sum_{l,p=1}^{N} \int_{t_{lp}}^{t_{lp}+\xi_{lp}^k} \left| f_x^0(x_*(\tau), v_p, \tau) - f_x^0(x_*(\tau), u_*(t), \tau \right| d\tau \leqslant$$

$$\leqslant a_0(N) \left| \xi_k \right| \cdot 2 \max_{(x,u,t)\in Q_N} \left| f_x^0(x, u, t) \right| \to 0$$

при $k \to \infty$ равномерно на $[t_0, T]$, так как $|\xi_k| = \sum_{l,p=1}^{N} |\xi_{lp}^k| \to 0$ в силу (25). Аналогично доказывается равномерная на $[t_0, T]$ сходимость при $k \to \infty$ 4-го и 5-го слагаемых из (54). Таким образом,

$$\lim_{k\to\infty} \sup_{t_0 \leqslant t \leqslant T} |b_k(t; N)| = 0. \tag{55}$$

Далее, из (53) имеем

$$|\Delta\psi_k(t)| \leqslant \int_t^T L_N \sum_{i=1}^{n} |\Delta\psi_k^i(\tau)| \, d\tau + |b_k(t; N)| + |\Delta\psi_k(T)| \leqslant$$

$$\leqslant L_N \sqrt{n} \int_t^T |\Delta\psi(\tau)| \, d\tau + |b_k(t; N)| + |\Delta\psi_k(T)|, \quad t_0 \leqslant t \leqslant T.$$

Как видно, функция $\varphi(t) = |\Delta\psi_k(t)|$ удовлетворяет неравенству (6) с $a = L_N\sqrt{n}$, $b(t) = |b_k(t; N)| + |\Delta\psi_k(T)|$. Тогда в силу (7) получаем $|\Delta\psi_k(t)| = |\psi_k(t; N) - \psi(t; N)| \leqslant$

$$\leqslant L_N \sqrt{n} \int_t^T (|b_k(\tau; N)| + |\Delta\psi_k(T)|) \exp\{L_N \sqrt{n}(\tau - t)\} \, d\tau +$$

$$+ |b_k(t; N)| + |\Delta\psi_k(T)|, \quad t_0 \leqslant t \leqslant T.$$

Отсюда и из (50), (55) следует

$$\lim_{k\to\infty} \max_{t_0 \leqslant t \leqslant T} |\psi_k(t; N) - \psi(t; N)| = 0. \tag{56}$$

Из (46) с учетом (26), $\{a_k(N)\} \to a(N)$ при $k \to \infty$ тогда имеем

$$\lim_{k\to\infty} \psi_k(t_0; N) = \psi(t_0; N) = \sum_{i=0}^{s} a_i(N) \, g_x^i(x_{0*}, x_*(T)). \tag{57}$$

Перейдем к пределу при $k \to \infty$ в неравенстве (48); с учетом (25), (26), (56), $\{a_{0k}(N)\} \to a_0(N)$ и непрерывности функции H по совокупности своих

аргументов получим

$$[H(x_*(t), v_p, t, \psi(t; N), a_0(N)) - H(x_*(t), u_*(t), t, \psi(t; N), a_0(N))]|_{t=t_{lp}} \leqslant 0 \tag{58}$$

для всех l, $p = 1, \ldots, N$.

Наконец, совершим предельный переход при $N \to \infty$ в соотношениях (49)—(51), (57), (58). Выбирая при необходимости подпоследовательность из $\{a(N)\}$, можем считать, что сама последовательность $\{a(N)\} \to a = (a_0, a_1, \ldots, a_s)$. Тогда из (49) сразу получим

$$0 \leqslant a_0 \leqslant 1, \quad a_1 \geqslant 0, \ \ldots, \ a_m \geqslant 0, \quad |a|^2 = \sum_{i=0}^{s} a_i^2 = 1. \tag{59}$$

Из (50) при $N \to \infty$ имеем

$$\lim_{N\to\infty} \psi(T; N) = \psi(T) = -\sum_{i=0}^{s} a_i g_y^i (x_{0*}, x_*(T)). \tag{60}$$

Покажем, что последовательность $\{\psi(t; N)\}$ равномерно на $[t_0, T]$ сходится к решению $\psi(t)$ системы уравнений

$$\dot{\psi}(t) = -H_x(x_*(t), u_*(t), t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T; \tag{61}$$

с начальным условием (60). Существование и единственность решения задачи (61), (60) следует из теоремы 1.1. Выпишем интегральное уравнение, которому удовлетворяет решение задачи (61), (60):

$$\psi(t) = \int_t^T H_x(x_*(\tau), u_*(\tau), \tau\psi(\tau), a_0)\, d\tau + \psi(T).$$

Отсюда и из (52) для разности $\Delta\psi(t) = \psi(t; N) - \psi(t)$ имеем

$$\Delta\psi(t) = \int_t^T \Big[(-a_0(N) + a_0) f_x^0(x_*(\tau), u_*(\tau), \tau) + \\ + \sum_{i=1}^{n} \Delta\psi_i(\tau) f_x^i(x_*(\tau), u_*(\tau), \tau)\Big] d\tau + \Delta\psi(T), \quad t_0 \leqslant t \leqslant T.$$

Тогда

$$|\Delta\psi(t)| \leqslant \int_t^T L\sqrt{n}\, |\Delta\psi(\tau)|\, d\tau + L(T - t_0)\, |a_0(N) - a_0| + |\Delta\psi(T)|, \\ t_0 \leqslant t \leqslant T,$$

где $L = \max\limits_{0 \leqslant i \leqslant n} \sup\limits_{t_0 \leqslant t \leqslant T} |f_x^i(x_*(\tau), u_*(\tau), \tau|$. Отсюда и из неравенств (6)—(8) следует

$$|\Delta\psi(t)| = |\psi(t; N) - \psi(t)| \leqslant \\ \leqslant \exp\{L\sqrt{n}(T - t_0)\}(L(T - t_0)\,|a_0(N) - a_0| + |\Delta\psi(T)|), \quad t_0 \leqslant t \leqslant T.$$

Тогда в силу (60), $\{a_0(N)\} \to a_0$ будем иметь

$$\lim_{N\to\infty} \max_{t_0 < t < T} |\psi(t; N) - \psi(t)| = 0. \tag{62}$$

Из (57) с учетом (62), $\{a(N)\} \to a$ получаем

$$\lim_{N\to\infty} \psi(t_0; N) = \psi(t_0) = \sum_{i=0}^{s} a_i g_x^i (x_{0*}, x_*(T)). \tag{63}$$

Перейдем к пределу при $N \to \infty$ в неравенстве (58). Поскольку в силу (12) $\{t_{lp}\} \to t_l$ при $N \to \infty$, $\{a_0(N)\} \to a_0$, то из (58) с учетом (62) имеем

$$H(x_*(t_l), v_p, t_l, \psi(t_l), a_0) - H(x_*(t_l), u_*(t_l), t_l, \psi(t_l), a_0) \leqslant 0, \quad l = 1, 2, \ldots$$

В силу плотности последовательности точек $\{v_p\}$ во множестве V отсюда получаем

$$H(x_*(t_l), v, t_l, \psi(t_l), a_0) \leqslant H(x_*(t_l), u_*(t_l), t_l, \psi(t_l), a_0) \qquad \forall v \in V, \quad l = 1, 2, \ldots$$

Последовательность $\{t_l\}$ рациональных точек интервала (t_0, T), являющихся точками непрерывности оптимального управления $u_*(t)$, всюду плотна на отрезке $[t_0, T]$. Поэтому предыдущее неравенство справедливо для всех $t \in [t_0, T]$, в которых $u_*(t)$ непрерывно:

$$H(x_*(t), v, t, \psi(t), a_0) \leqslant H(x_*(t), u_*(t), t, \psi(t), a_0) \qquad \forall v \in V.$$

Поскольку при $v = u_*(t) \in V$ здесь получаем равенство, то

$$\sup_{v \in V} H(x_*(t), v, t, \psi(t), a_0) = H(x_*(t), u_*(t), t, \psi(t), a_0) \tag{64}$$

во всех точках непрерывности $u_*(t)$.

Далее, докажем, что

$$a_i g^i(x_{0*}, x_*(T)) = 0, \qquad i = 1, \ldots, m. \tag{65}$$

Для тех номеров i $(1 \leqslant i \leqslant m)$, для которых $g^i(x_{0*}, x_*(T)) = 0$, равенства (65), конечно, выполняются. Пусть $g^i(x_{0*}, x_*(T)) < 0$. Из (25), (26) тогда следует: $g^i(x_{0k}, x_k(T)) < 0$ при всех $k \geqslant k_0$, где k_0 достаточно большое число. Поэтому из формул (18), (34) получаем $a_{ik} = a_{ik}(N) = 0$ для всех $k \geqslant k_0$. Тогда $\lim\limits_{k \to \infty} a_{ik}(N) = a_i(N) = 0$ для каждого $N \geqslant 1$. Следовательно, $\lim\limits_{N \to \infty} a_i(N) = a_i = 0$ для тех номеров i $(1 \leqslant i \leqslant m)$, для которых $g^i(x_{0*}, x_*(T)) < 0$. Равенства (65) доказаны.

Соотношения (59)—(61), (63)—(65), составляющие основное содержание теоремы 2.1, установлены. Тем самым теорема 2.1 полностью доказана для случая, когда задача (2.1)—(2.4) имеет единственное решение $(x_{0*}, u_*(t), x_*(t))$. Общий случай, когда задача (2.1)—(2.4) имеет более чем одно решение, легко сводится к рассмотренному случаю. А именно, пусть $(x_{0*}, u_*(t), x_*(t))$ одно из решений задачи (2.1)—(2.4). Введем новую фазовую координату x^{n+1} и к системе (2.2) добавим еще одно уравнение

$$\dot{x}^{n+1}(t) = |u(t) - u_*(t)| = f^{n+1}(u(t), t), \qquad t_0 \leqslant t \leqslant T, \qquad x^{n+1}(t_0) = x_0^{n+1}. \tag{66}$$

Введем функцию

$$J_1(x_0, x_0^{n+1}, u(\cdot)) = \int_{t_0}^{T} f^0(x(t), u(t), t)\, dt + g^0(x_0, x(T)) + (x^{n+1}(T))^2 +$$

$$+ |x_0 - x_{0*}|^2 = \int_{t_0}^{T} f^0(x(t), u(t), t)\, dt + g_1^0(x_0, x(T), x^{n+1}(T)), \tag{67}$$

где $g_1^0(x, y, y^{n+1}) = g^0(x, y) + (y^{n+1})^2 + |x - x_{0*}|^2$. Рассмотрим задачу минимизации функции (67) при условиях (2.2)—(2.4), (66). Нетрудно видеть,

что $J_1\left(x_0, x_0^{n+1}, u(\cdot)\right) > J(x_0, u(\cdot)) \geqslant J_*$ для всех допустимых наборов $\left(x_0, x_0^{n+1}, u(\cdot), x(\cdot), x^{n+1}(\cdot)\right)$ задачи (67), (2.2)—(2.4), (66), для которых $x_0 \neq x_{0*}$, $x_0^{n+1} \neq 0$, $u(\cdot) \neq u_*(\cdot)$. В то же время $J_1(x_{0*}, 0, u_*(\cdot)) = J(x_{0*}, u_*(\cdot)) = J_*$. Это значит, что функция (67) при условиях (2.2)—(2.4), (66) достигает своей нижней грани на единственном допустимом наборе $\left(x_{0*}, x_{0*}^{n+1} = 0, u_*(\cdot), x_*(\cdot), x_*^{n+1}(\cdot) \equiv 0\right)$. Следовательно, для задачи (67), (2.2)—(2.4), (66) справедлив принцип максимума. Конечно, для полной строгости надо оговорить, что функция $f^{n+1}(u, t) = |u - u_*(t)|$, находящаяся в правой части уравнения (66), необязательно непрерывна по $t \in [t_0, T]$, и строго говоря, не удовлетворяет условиям теоремы 2.1. Но, тем не менее, благодаря тому, что $f^{n+1}(u, t)$ кусочно-непрерывна по t, не зависит от x, удовлетворяет условию (1), нетрудно последить, что все вышеприведенные рассуждения, приведшие к соотношениям (59)—(61), (63)—(65), сохраняют силу и для задачи (67), (2.2)—(2.4), (66). В этой задаче функция Гамильтона — Понтрягина имеет вид

$$H_1\left(x, x^{n+1}, u, t, \psi, \psi_{n+1}, a_0\right) = -a_0 f^0(x, u, t) + \langle\psi, f(x, u, t)\rangle + \\ + \psi_{n+1}|u - u_*(t)| = H(x, u, t, \psi, a_0) + \psi_{n+1}|u - u_*(t)|.$$

Согласно уже доказанному случаю принципа максимума найдутся числа $a_0, a_1, \ldots, a_s$ и вектор-функция $(\psi(t), \psi_{n+1}(t))$ $(t_0 \leqslant t \leqslant T)$ такие, что

$$a = (a_0, a_1, \ldots, a_s) \neq 0, \quad a_0 \geqslant 0, a_1 \geqslant 0, \ldots, a_m \geqslant 0,$$

$$\dot{\psi}(t) = -H_{1x}\left(x_*(t), x_*^{n+1}(t), u_*(t), t, \psi(t), \psi_{n+1}(t), a_0\right) = \\ = -H_x(x_*(t), u_*(t), t, \psi(t), a_0), \quad t_0 \leqslant t \leqslant T,$$

$$\dot{\psi}_{n+1}(t) = -H_{1x^{n+1}} \equiv 0, \quad t_0 \leqslant t \leqslant T;$$

условие максимума

$$\max_{v \in V} H_1\left(x_*(t), x_*^{n+1}(t), v, t, \psi(t), \psi_{n+1}(t), a_0\right) = \\ = H_1\left(x_*(t), x_*^{n+1}(t), u_*(t), t, \psi(t), \psi_{n+1}(t), a_0\right)$$

выполняется во всех точках непрерывности $u_*(t)$; справедливы условия трансверсальности

$$\psi(t_0) = a_0 g_{1x}^0\left(x_{0*}, x_*(T), x_*^{n+1}(T) = 0\right) + \sum_{j=1}^{s} a_j g_x^j(x_{0*}, x_*(T)) = \\ = \sum_{j=0}^{s} a_j g_x^j(x_{0*}, x_*(T)),$$

$$\psi_{n+1}(t_0) = a_0 g_{1x^{n+1}}^0 + \sum_{j=1}^{s} a_j g_{x^{n+1}}^j = 0,$$

$$\psi(T) = -a_0 g_{1y}^0\left(x_{0*}, x_*(T), x_*^{n+1}(T) = 0\right) - \sum_{j=1}^{s} a_j g_y^j(x_{0*}, x_*(T)) = \\ = -\sum_{j=0}^{s} a_j g_y^j(x_{0*}, x_*(T)),$$

$$\psi_{n+1}(T) = -a_0 g^0_{1y^{n+1}}\left(x_{0*}, x_*(T), x_*^{n+1}(T) = 0\right) - \\ - \sum_{j=1}^{s} a_j g^j_{y^{n+1}}\left(x_{0*}, x_*(T)\right) = 0$$

и условия дополняющей нежесткости

$$a_i g_i\left(x_{0*}, x_*(T)\right) = 0, \qquad i = 1, \ldots, m.$$

Из этих условий следует, что $\psi_{n+1}(t) \equiv 0$. Учитывая это равенство в полученных условиях, снова придем к соотношениям (59)—(61), (63)—(65) и в том случае, когда в задаче (2.1)—(2.4) оптимальное решение не единственно. Теорема 2.1 доказана.

Доказательство теоремы 2.2 также приведем при дополнительном предположении выполнения условия (1), считая, что в (1) $t \in \mathbf{R}$. Пусть $(x_{0*}, u_*(t), x_*(t), t_{0*}, T_*)$ — оптимальное решение задачи (2.37)—(2.40). Если задачу (2.37)—(2.40) рассматривать как задачу с закрепленным временем, взяв в ней $t_0 = t_{0*}$, $T = T_*$, то она превратится в задачу (2.1)—(2.4) с оптимальным решением $(x_{0*}, u_*(t), x_*(t))$, и к ней применима уже доказанная теорема 2.1, из которой следуют соотношения (59), (61), (64) с $t_0 = t_{0*}$, $T = T_*$, а также условия трансверсальности

$$\psi(t_{0*}) = \sum_{j=0}^{s} a_j g^j_x(x_{0*}, x_*(T_*), t_{0*}, T_*), \quad \psi(T_*) = -\sum_{j=0}^{s} a_j g^j_y(x_{0*}, x_*(T_*), t_{0*}, T_*)$$

и дополняющей нежесткости

$$a_j g^j\left(x_{0*}, x_*(T_*), t_{0*}, T_*\right) = 0, \qquad i = 1, \ldots, m.$$

Поэтому в отдельном доказательстве нуждаются лишь условия трансверсальности (2.42), (2.43):

$$\max_{v \in V} H\left(x_*(t_{0*}), v, t_{0*}, \psi(t_{0*}), a_0\right) = -\sum_{j=0}^{s} a_j g^j_t\left(x_{0*}, x_*(T_*), t_{0*}, T_*\right), \tag{68}$$

$$\max_{v \in V} H\left(x_*(T_*), v, T_*, \psi(T_*), a_0\right) = \sum_{j=0}^{s} a_j g^j_T\left(x_{0*}, x_*(T_*), t_{0*}, T_*\right). \tag{69}$$

Начнем с доказательства равенства (69). Сначала рассуждения проведем в предположении, что решение $(x_{0*}, u_*(t), x_*(t), t_{0*}, T_*)$ задачи (2.37)—(2.40) с закрепленным $t_0 = t_{0*}$ единственно. Как и выше, на интервале (t_{0*}, T_*), введем точки $\{t_{lp}\}$ согласно (12), матрицу ξ, перейдем к задаче (14)—(17), зафиксировав в ней $t_0 = t_{0*}$; управление $u(t, \xi)$ на отрезке $[t_{0*}, T_*]$ будем определять согласно (13). Поскольку теперь время T окончания процесса заранее неизвестно и нам придется рассматривать значения $T > T_*$, то нужно как-то определить $u(t, \xi)$ и при $t \geqslant T_*$. Для наших целей будет достаточно принять

$$u(t, \xi) = u_*(T_* - 0, \xi) = u_*(T_* - 0) \qquad \forall t \geqslant T_*. \tag{70}$$

Кроме того, говоря о задаче (14)—(17) и других возникающих ниже задачах, мы всюду дальше будем иметь в виду, что здесь функции g^i являются функциями аргументов (x, y, t, T).

Пусть число $\Delta T_N > 0$ столь мало, что отрезок $[T_* - \Delta T_N, T_*]$ не содержит точек $\{t_{lp}\}$ $(l = 1, \ldots, N, p = 1, \ldots, N+1)$, и управление $u_*(t)$

непрерывно при $T_* - \Delta T_N \leqslant t < T_*$. Тогда, как видно из (13), (70), управление $u(t, \xi)$ будет непрерывно при $t \geqslant T_* - \Delta T_N$, а тогда соответствующая траектория $x(t, u(\cdot, \xi), x_0)$ при $t \geqslant T_* - \Delta T_N$ будет непрерывно дифференцируема. Задачу (14)—(17) будем рассматривать при дополнительном условии

$$T_* - \Delta T_N < T < T_* + \Delta T_N. \tag{71}$$

Получившаяся задача (14)—(17), (71) представляет собой конечномерную задачу минимизации в пространстве переменных (x_0, ξ, T): $x_0 = (x_0^1, \ldots, x_0^n)$, $\xi = \{\xi_{lp}, l, p = 1, \ldots, N\}$. Любой набор $(x_0, u(\cdot, \xi), T)$, допустимый в задаче (14)—(17), (71), является допустимым и в задаче (2.37)—(2.40), причем значения целевых функций в обеих задачах на таком наборе совпадают (напомним, что момент $t_0 = t_{0*}$ у нас пока фиксирован). В то же время оптимальный набор $(x_{0*}, u_*(\cdot) = u(\cdot, 0), T_*)$, задачи (2.37)—(2.40) является допустимым в задаче (14)—(17), (71). Отсюда ясно, что тройка $(x_{0*}, \xi_* = 0, T_*)$ является единственным решением задачи (14)—(17), (71).

Применяя к задаче (14)—(17), (71) метод штрафных функций, придем к задаче (19)—(22), (71) с фиксированным $t_0 = t_{0*}$. Как и выше, пользуясь оценкой (24), нетрудно показать, что функция $\Phi_k(x_0, \xi, T)$ непрерывна на компактном множестве

$$U_0 = \{(x_0, \xi, T)\colon |x_0 - x_{0*}| \leqslant 1;\ 0 \leqslant \xi_{lp} \leqslant d_N,\ l, p = 1, \ldots, N;\ |T - T_*| \leqslant \Delta T_N\}.$$

Отсюда будет следовать, что задача (19)—(22), (71) имеет хотя бы одно решение $(x_{0k}, \xi_k = \{\xi_{lp}^k\}, T_k)$. Пользуясь теоремами 5.14.1, 5.14.2, убеждаемся в справедливости предельных соотношений (25)—(27) с $t_0 = t_{0*}$, $T = T_*$ и, кроме того,

$$\lim_{k\to\infty} T_k = T_*, \quad \lim_{k\to\infty} x_k(T_k) = x_*(T_*), \tag{72}$$

где $x_k(t) = x(t, u(\cdot, \xi_k), x_{0k})$ $(t_0 = t_{0*} \leqslant t \leqslant T_k)$. Поэтому можем считать, что наряду с неравенствами (28) выполняется неравенство $|T_k - T_*| < \Delta T_N$ при всех $k \geqslant k_0$. Оптимальному решению (x_{0k}, ξ_k, T_k) дадим приращение $(\Delta x_0 = 0, \Delta\xi = 0, \Delta T)$, где $\Delta T > 0$ столь мало, что

$$|T_k + \Delta T - T_*| < \Delta T_N. \tag{73}$$

Тогда на отрезке $t_0 = t_{0*} \leqslant t \leqslant \min\{T_k, T_k + \Delta T\}$ траектория не получит приращения, а функция $\Phi_k(x_0, \xi, T)$ получит приращение только за счет того, что одна и та же траектория $x_k(t) = x(t, u(\cdot, \xi_k), x_{0k})$ будет рассматриваться на различных отрезках $[t_{0k}, T_k]$, $[t_{0*}, T_k + \Delta T]$. С учетом оптимальности набора (x_{0k}, ξ_k, T_k), определения (13), (70) управления $u_k(t) = u(t, \xi_k)$ имеем

$$0 \leqslant \Delta\Phi_k = \Phi_k(x_{0k}, \xi_k, T_k + \Delta T) - \Phi_k(x_{0k}, \xi_k, T_k) =$$
$$= \int_{t_{0*}}^{T_k+\Delta T} f^0(x_k(t), u_k(t), t)\,dt - \int_{t_0}^{T_k} f^0(x_k(t), u_k(t), t)\,dt +$$
$$+ g^0(x_{0k}, x_k(T_k + \Delta T), t_{0*}, T_k + \Delta T) - g^0(x_{0k}, x_k(T_k), t_{0*}, T_k) +$$

$$+ A_k \sum_{i=1}^{s} \left[\left(g_i^+ (x_{0k}, x_k(T_k+\Delta T), t_{0*}, T_k+\Delta T) \right)^2 - \left(g_i^+ (x_{0k}, x_k(T_k), t_{0*}, T_k) \right)^2 \right] =$$

$$= \int_{T_k}^{T_k+\Delta T} f^0 (x_k(t), u_*(t), t)\, dt + \Big\langle g_y^0 (x_{0k}, x_k(T_k), t_{0*}, T_k) +$$

$$+ \sum_{i=1}^{s} 2A_k g_i^+ (x_{0k}, x_k(T_k), t_{0*}, T_k)\, g_y^i (x_{0k}, x_k(T_k), t_{0*}, T_k),\ x_k(T_k+\Delta T) -$$

$$- x_k(T_k) \Big\rangle + \Big[g_T^0 (x_{0k}, x_k(T_k), t_{0*}, T_k) +$$

$$+ \sum_{i=1}^{s} 2A_k g_i^+ (x_{0k}, x_k(T_k), t_{0*}, T_k)\, g_T^i (x_{0k}, x_k(T_k), t_{0*}, T_k) \Big] \Delta T + R_{5k},$$

где (74)

$$R_{5k} = \Big\langle g_y^0 \binom{\theta}{\dots} - g_y^0 (\cdots) + \sum_{i=1}^{s} 2A_k \Big(g_i^+ \binom{\theta}{\dots} g_y^i \binom{\theta}{\dots} -$$

$$- g_i^+ (\cdots)\, g_y^i (\cdots) \Big),\ x_k(T_k+\Delta T) - x_k(T_k) \Big\rangle + \Big[g_T^0 \binom{\theta}{\dots} - g_T^0 (\cdots) +$$

$$+ \sum_{i=1}^{s} 2A_k \Big(g_i^+ \binom{\theta}{\dots} g_T^i \binom{\theta}{\dots} - g_i^+ (\cdots)\, g_T^i (\cdots) \Big) \Big] \Delta T, \quad 0 < \theta < 1; \tag{75}$$

здесь по аналогии с (33) для краткости обозначено $\binom{\theta}{\dots} = (x_{0k}, x_k(T_k) + \theta (x_k(T_k+\Delta T) - x_k(T_k)), t_{0*}, T_k + \theta\Delta T)$, $(\cdots) = (x_{0k}, x_k(T_k), t_{0*}, T_k)$. Заметим, что функции $f^i(x_k(t), u_k(t), t)$ $(i = 0, 1, \dots, n)$ непрерывны на отрезке $[T_k, T_k+\Delta T]$ (при $\Delta T < 0$ здесь надо рассматривать отрезок $[T_k+\Delta T, T_k]$) — это следует из выбора ΔT_N, условий (13), (70), (71), (73), непрерывности $u_k(t) = u(t, \xi_k) = u_*(t)$ на этом отрезке, непрерывности функций $f^i(x, u, t)$, $x_k(t)$. Поэтому

$$x_k(T_k+\Delta T) - x_k(T_k) = \int_{T_k}^{T_k+\Delta T} f(x_k(t), u_*(t), t)\, dt =$$

$$= f(x_k(T_k), u_*(T_k), T_k)\, \Delta T + R_{6k},$$

$$\int_{T_k}^{T_k+\Delta T} f^0 (x_k(t), u_*(t), t)\, dt = f^0 (x_k(T_k), u_*(T_k), T_k)\, \Delta T + R_{7k}, \tag{76}$$

где

$$R_{6k} = \int_{T_k}^{T_k+\Delta T} [f(x_k(t), u_*(t), t) - f(x_k(T_k), u_*(T_k), T_k)]\, dt = o(\Delta T),$$

$$R_{7k} = \int_{T_k}^{T_k+\Delta T} \left[f^0 (x_k(t), u_*(t), t) - f^0 (x_k(T_k), u_*(T_k), T_k) \right] \Delta T = o_k(\Delta T),$$

$\lim\limits_{\Delta T\to 0} o_k(\Delta T)/\Delta T = 0$. Отсюда и из непрерывности функций g^i, g_i^+, g_y^i, g_T^i для R_{5k} из (75) также имеем $R_{5k} = o_k(\Delta T)$. Далее, воспользуемся величинами a_{ik} из (34) (разумеется аргументы функций g^i, g_i^+ здесь должны быть дополнены величинами $t_0 = t_{0*}$, $T = T_k$). Умножим (74) на $a_{0k} > 0$; с учетом равенств (76) получим

$$0 \leqslant a_{0k}\Delta\Phi_k = a_{0k}f^0(x_k(T_k), u_*(T_k), T_k)\Delta T +$$
$$+ \left\langle \sum_{i=0}^{s} a_{ik}g_y^i(x_{0k}, x_k(T_k), t_{0*}, T_k), f(x_k(T_k), u_*(T_k), T_k \right\rangle \Delta T +$$
$$+ \sum_{i=0}^{s} a_{ik}g_T^i(x_{0k}, x_k(T_k), t_{0*}, T_k)\Delta T + R_{8k},$$

где

$$R_{8k} = a_{0k}(R_{5k} + R_{7k}) + \left\langle \sum_{i=0}^{s} a_{ik}g_y^i(x_{0k}, x_k(T_k), t_{0*}, T_k), R_{6k} \right\rangle = o_k(\Delta T).$$

Отсюда, пользуясь условием (38) с $T = T_k$ и определением (36) функции H, имеем

$$0 \leqslant a_{0k}\Delta\Phi_k = \Delta T\left[-H(x_k(T_k), u_*(T_k), T_k, \psi_k(T_k), a_{0k}) + \right.$$
$$\left. + \sum_{i=0}^{s} a_{ik}g_T^i(x_{0k}, x_k(T_k), t_{0*}, T_k) + \frac{o_k(\Delta T)}{\Delta T}\right]$$

при всех достаточно малых $|\Delta T|$, причем ΔT могут быть как положительными, так и отрицательными. Это возможно лишь при

$$H(x_k(T_k), u_*(T_k), T_k, \psi_k(T_k), a_{0k}) = \sum_{i=0}^{s} a_{ik}g_T^i(x_{0k}, x_k(T_k), t_{0*}, T_k), \quad k \geqslant k_0.$$

Отсюда, пользуясь (72), рассуждая также, как при доказательстве теоремы 2.1, совершим последовательно предельные переходы сначала при $k \to \infty$, затем при $N \to \infty$ и получим условие (69). Напоминаем, что наши рассуждения проводились в предположении, что задача (2.37)—(2.40) с закрепленным $t_0 = t_{0*}$ имеет единственное решение. Если эта задача имеет неединственное решение, то можно перейти к задаче минимизации функции

$$J_1(x_0, x_0^{n+1}, u(\cdot), T) = J(x_0, u(\cdot), t_{0*}, T) +$$
$$+ (x^{n+1}(T))^2 - (T - T_*)^2 + |x_0 - x_{0*}|^2$$

при условиях (2.38)—(2.40), (66), которая имеет единственное решение $(x_{0*}, u_*(t), x_*(t), T_*)$, применить к ней уже доказанное и, учитывая, что оптимальные $x_*^{n+1}(t) \equiv 0$, $\psi_{n+1}(t) \equiv 0$, получить условие (69) и в этом случае.

Для доказательства условия (68) нужно провести те же рассуждения, поменяв ролями левый и правый концы траектории, зафиксировав в (2.37)—(2.40) $T = T_*$, пользуясь вместо задачи (2) задачей Коши (10) и варьируя t_0 в достаточно малой окрестности t_{0*}. Теорема 2.2 доказана.

§ 4. О методах решения краевой задачи принципа максимума

1. Выше в § 2 было показано, что задача оптимального управления (см. задачи (2.1)—(2.4), (2.37)—(2.40)) с помощью принципа максимума может быть сведена к краевой задаче, состоящей из условия максимума (2.14)

$$\sup_{u\in V} H(x, u, t, \psi, a_0)=H(x, u(x, t, \psi, a_0), t, \psi, a_0), \tag{1}$$

определяющего функцию $u=u(x, t, \psi, a_0)$, системы $2n$ дифференциальных уравнений

$$\dot{x}=f(x, u(x, t, \psi, a_0), t), \quad \dot{\psi}=-H_x(x, u(x, t, \psi, a_0), t, \psi, a_0), \tag{2}$$
$$t_0\leqslant t\leqslant T,$$

граничных условий (см. условие (2.18), ограничения типа равенств из (2.39), условия (2.41)—(2.44)), которые могут быть записаны в следующем виде

$$Q_i(t_0, x(t_0), \psi(t_0), a_0, \ldots, a_s, T, x(T), \psi(T))=0,$$
$$i=1, \ldots, 2n+s+3, \tag{3}$$

и неравенств (см. (2.19), (2.39))

$$a_0\geqslant 0, \quad a_1\geqslant 0, \ldots, a_m\geqslant 0, \quad g^i(x(t_0), x(T), t_0, T)\leqslant 0,$$
$$i=1, \ldots, m. \tag{4}$$

Для численного решения такой краевой задачи могут быть приспособлены известные методы, такие, как метод стрельбы, метод прогонки, различные итерационные методы [4, 13, 20, 39, 54, 209]. Здесь мы кратко остановимся на методе стрельбы, с помощью которого краевая задача (1)—(4) сводится к решению задачи Коши и некоторой задаче минимизации функции конечного числа переменных. Для применения этого метода сначала из набора t_0, $x(t_0)=x_0$, $\psi(t_0)=\psi_0$, $a=(a_0, a_1, \ldots, a_s)$, T, $x(T)=x_1$, $\psi(T)=\psi_1$ выделяют величины, называемые параметрами стрельбы, задав которые можно сформулировать и однозначно решить какую-либо задачу Коши для системы (2). В качестве параметров стрельбы часто берут t_0, x_0, ψ_0, a. Затем задают какие-либо конкретные числовые значения этих параметров и решают задачу Коши для системы (2) с начальными условиями

$$x(t_0)=x_0, \quad \psi(t_0)=\psi_0. \tag{5}$$

При численном решении задачи (2), (5) можно пользоваться известными методами Эйлера, Адамса, Рунге — Кутта и др. [4, 13, 39, 54, 258].

Допустим, что решение этой задачи $x(t;\ t_0,\ x_0,\ \psi_0,\ a_0)$, $\psi(t;\ t_0,\ x_0,\ \psi_0,\ a_0)$, соответствующее взятым значениям параметров стрельбы, уже найдено. Тогда можно вычислить значения функций Q_i из (3), взяв в качестве аргументов этих функций выбранные значения параметров стрельбы и соответствующие им $x(T)=x(T;\ t_0,\ x_0,\ \psi_0,\ a_0)$, $\psi(T)=\psi(T;\ t_0,\ x_0,\ \psi_0,\ a_0)$ при некотором $T>t_0$, и значение функции

$$\varphi(t_0, x_0, \psi_0, a, T) = \sum_{i=1}^{2n+s+3} Q_i^2(t_0, x_0, \psi_0, a, T, x(T), \psi(T)), \qquad (6)$$

называемой функцией невязки системы (3). В случае удачного «выстрела», когда выбор параметров t_0, x_0, ψ_0, a, T обеспечивает выполнение неравенств (4) и функция невязки (6) обратится в нуль, что равносильно условиям (3), краевая задача принципа максимума (1)—(4) будет решена. Однако вряд ли с первого «выстрела» удастся добиться выполнения с нужной точностью неравенств (4) и обратить невязку (6) в нуль — скорее всего придется продолжать «стрельбу» с другими значениями параметров t_0, x_0, ψ_0, a, T, которые дают меньшее значение функции (6) при соблюдении условий (4). Таким образом, для выбора наилучших параметров стрельбы нужно решить задачу минимизации функции невязки (6) при ограничениях (4); здесь возможно использование известных методов минимизации, например, методов из главы 5. Конечно, вместо этой задачи минимизации можно непосредственно решать систему уравнений

$$Q_i(t_0,\ x_0,\ \psi_0,\ a,\ T,\ x(T;\ t_0,\ x_0,\ \psi_0,\ a_0),\ \psi(T;\ t_0,\ x_0,\ \psi_0,\ a_0))=0, \qquad (7)$$
$$i=1,\ \ldots,\ 2n+s+3,$$

относительно $2n+s+3$ неизвестных t_0, x_0, ψ_0, a, T с учетом ограничений (4); здесь могут быть использованы подходящие модификации известных методов решения нелинейных систем уравнений [4, 8, 13, 20, 22, 39, 42, 45, 54, 73, 76, 106, 209, 221, 238, 239, 296, 301]. Подчеркнем, что вычисление значений функции (6) и левых частей системы (7) в каждой отдельно взятой точке $(t_0,\ x_0,\ \psi_0,\ a,\ T)$, вообще говоря, весьма трудоемко — для этого всякий раз нужно решать задачу Коши (2), (5).

Аналогично может быть описан метод стрельбы для случая, когда в качестве параметров стрельбы берутся T, x_1, ψ_1, a и задача Коши для системы (2) решается с начальными условиями $x(T)=x_1$, $\psi(T)=\psi_1$. Разумеется, описанная схема метода стрельбы при применении к конкретным задачам оптимального уравнения каждый раз должна модифицироваться с учетом граничных режимов на концах траекторий (см., например, (2.21)—(2.36))

и других особенностей решаемой задачи; количество параметров стрельбы по возможности нужно стараться уменьшить.

Следует заметить, что краевая задача (1)—(4) принципа максимума имеет ряд специфических особенностей, затрудняющих применение метода стрельбы и других стандартных методов решения краевых задач. Такие методы обычно разрабатываются для краевых задач вида (2), (3) при отсутствии каких-либо дополнительных ограничений на переменные; здесь же задачу (2), (3) приходится решать совместно с дополнительными неравенствами (4). Наличие числовых параметров a_0, a_1, ..., a_s также осложняет краевую задачу (1)—(4). Далее, функция $u = u(x, t, \psi, a_0)$, определяемая из условия (1), вообще говоря, нелинейно зависит от своих аргументов, и поэтому система дифференциальных уравнений (2), как правило, нелинейна относительно (x, ψ) даже в том случае, когда исходная система $\dot{x} = f(x, u, t)$ была линейна относительно (x, u). Кроме того, функция $u = u(x, t, \psi, a_0)$, вообще говоря, не обладает хорошими дифференциальными свойствами и даже может быть разрывной (об этом говорят функции вида $u = \operatorname{sign}\psi$ из примеров 2.7—2.11), что влечет за собой плохие аналитические свойства правых частей системы (2). Краевая задача (1)—(4) существенно осложняется в тех случаях, когда из условия (1) функция $u = u(x, t, \psi, a_0)$ определяется неоднозначно. Указанные обстоятельства весьма затрудняют исследование вопросов существования, единственности, устойчивости решения краевой задачи принципа максимума, сходимости методов ее решения. При численном решении прикладных задач оптимального управления трудности, связанные с плохой сходимостью методов, с их неустойчивостью, обычно преодолеваются на основе учета специфики конкретно решаемой задачи, ее физического смысла и т. п. [14, 38, 68, 80, 217, 326]. Некоторые приемы преодоления возможной неустойчивости в краевых задачах принципа максимума описаны, например, в [14, 81, 204, 217, 326].

2. Для некоторых классов задач оптимального управления разработаны специальные методы решения краевой задачи принципа максимума. Здесь мы кратко остановимся на следующей задаче с закрепленным левым и свободным правым концами траектории:

$$J(u) = \int_{t_0}^{T} f^0(x(t)\ u(t), t)\, dt + g^0(x(T)) \to \inf, \tag{8}$$

$$\dot{x}(t) = f(x(t),\ u(t),\ t), \quad t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0, \tag{9}$$

$$u = u(t) \in V, \quad t_0 \leqslant t \leqslant T, \tag{10}$$

где $u = u(t)$ кусочно-непрерывное управление; обозначения взяты из § 2. Как показано в § 2, краевая задача принципа макси-

мума для задачи (8)—(10) имеет вид (см. (2.15), (2.26)):

$$\dot{x} = f(x, u, t), \quad t_0 \leqslant t \leqslant T; \quad x(t_0) = x_0, \tag{11}$$

$$\dot{\psi} = -H_x(x, u, t, \psi), \quad t_0 \leqslant t \leqslant T; \quad \psi(T) = -g_y^0(x(T)), \tag{12}$$

где $H(x, u, t, \psi) = -f^0(x, u, t) + \langle \psi, f(x, u, t) \rangle$, функция $u = u(x, t, \psi)$ в (11), (12) определяется из условия

$$\sup_{u \in V} H(x, u, t, \psi) = H(x, u(x, t, \psi), t, \psi). \tag{13}$$

Попутно сделаем замечание, касающееся возможностей применения метода стрельбы к краевой задаче (11)—(13). В этой задаче начальное условие x_0, моменты t_0, T известны, поэтому в качестве параметра стрельбы здесь достаточно взять ψ_0; систему (2) нужно решать с начальными условиями $x(t_0) = x_0$, $\psi(t_0) = \psi_0$ при $a_0 = 1$; функция невязки (6) будет иметь вид $\varphi(\psi_0) = |\psi(T; \psi_0) + g_y^0(x(T; \psi_0))|^2$; ограничения (4) здесь отсутствуют. Если же за параметр стрельбы взять $x(T) = x_1$, то систему (2) нужно решать с начальными условиями $x(T) = x_1$, $\psi(T) = -g_y^0(x_1)$ при $a_0 = 1$; функция (6) будет иметь вид $\varphi(x_1) = |x(t_0; x_1) - x_0|^2$.

Перейдем к описанию итерационного метода, специально предназначенного для решения краевой задачи (11)—(13) [326]. Пусть $u_0 = u_0(t) \in V$ $(t_0 \leqslant t \leqslant T)$ — начальное приближение. Допустим, что уже найдено k-е приближение $u_k = u_k(t) \in V$ $(t_0 \leqslant t \leqslant T)$. Тогда, решая задачу Коши (11) при $u = u_k(t)$, находим ее решение $x_k(t) = x(t, u_k)$ $(t_0 \leqslant t \leqslant T)$. Затем, приняв в (12) $u = u_k(t)$, $x = x_k(t)$, из (12) определяем $\psi_k(t) = \psi(t; u_k)$. Наконец, из условия

$$\sup_{u \in V} H(x_k(t), u, t, \psi_k(t)) = H(x_k(t), u_{k+1}(t), t, \psi_k(t)), \quad t_0 \leqslant t \leqslant T, \tag{14}$$

получаем следующее приближение $u_{k+1} = u_{k+1}(t) \in V$ $(t_0 \leqslant t \leqslant T)$ и т. д. Метод описан. Таким образом, на каждом шаге метода нужно решить две задачи Коши — задачи (11) и (12) — и конечномерную задачу максимизации (14) при каждом $t \in [t_0, T]$. Предполагаем, что все получаемые этим методом функции $u_k(t)$ кусочно-непрерывны на отрезке $[t_0, T]$.

Может случиться, что при некотором k имеет место равенство $u_k(t) = u_{k+1}(t)$ $(t_0 \leqslant t \leqslant T)$. Согласно (14) это будет означать, что управление $u_k(t)$ удовлетворяет принципу максимума (теорема 2.1) и, следовательно, является подозрительным на оптимальность. В этом случае итерационный процесс прекращается, а управление $u_k(t)$ при необходимости подвергается дальнейшему исследованию для выяснения того, будет ли оно на самом деле оптимальным.

Рассмотрим случай, когда $u_{k+1}(\tau) \neq u_k(\tau)$ в некоторой точке $\tau \in [t_0, T]$, являющейся точкой непрерывности этих управлений,

и, кроме того,

$$H(x_k(\tau),\ u_{k+1}(\tau),\ \tau,\ \psi_k(\tau))-H(x_k(\tau),\ u_k(\tau),\ \tau,\ \psi_k(\tau))\geqslant 2\alpha>0. \tag{15}$$

Можно ли ожидать, что тогда $J(u_{k+1})<J(u_k)$? Для ответа на этот вопрос полезно обратиться к формуле приращения функции (8):

$$J(u+\Delta u)-J(u)=-\int_{t_0}^{T}[H(x(t;\ u),\ u(t)+\Delta u(t),\ t,\ \psi(t;\ u))-$$
$$-H(x(t;\ u),\ u(t),\ t,\ \psi(t;\ u))]\,dt+R, \tag{16}$$

где $R=R_1+R_2$,

$$R_1=\langle g_y^0(x(T;\ u)+\theta_1\Delta x(T))-g_y^0(x(T;\ u)),\ \Delta x(T)\rangle,$$

$$R_2=-\int_{t_0}^{T}\langle H_x(x(t;\ u)+\theta_2\Delta x(t),\ u(t)+\Delta u(t),\ t,\ \psi(t;\ u))-$$
$$-H_x(x(t;u),\ u(t),\ t,\ \psi(t;\ u)),\ \Delta x(t)\rangle\,dt, \tag{17}$$

$$\Delta x(t)=x(t;\ u+\Delta u)-x(t;\ u),\quad t_0\leqslant t\leqslant T,\quad 0<\theta_1,\quad \theta_2<1,$$

$x(t;\ u)$, $x(t;\ u+\Delta u)$ — решения системы (11) при $u=u(t)$ или $u=u(t)+\Delta u(t)$ соответственно, $\psi(t;\ u)$ — решение задачи (12) при $u=u(t)$. Формула (16), (17) выводится совершенно так же, как аналогичная формула (3.42), (3.33), (3.43) при $g^0(x,\ y)=g^0(y)$, $g^1=\ldots=g^s=0$, $\Delta x_0=0$, $a_{0k}=1$, $a_{1k}=\ldots=a_{sk}=0$, $A_k=1$. Из формулы (16), (17) при $u=u_k(t)$, $\Delta u=u_{k+1}-u_k(t)$ имеем

$$J(u_{k+1})-J(u_k)=$$
$$=-\int_{t_0}^{T}[H(x_k(t),\ u_{k+1}(t),\ t,\ \psi_k(t))-H(x_k(t),\ u_k(t),\ t,\ \psi_k(t))]\,dt+R. \tag{18}$$

Так как функции $x_k(t)$, $\psi_k(t)$, $H(x,\ u,\ t,\ \psi)$ непрерывны по совокупности своих аргументов, а $u_k(t)$, $u_{k+1}(t)$ кусочно-непрерывны на $[t_0,\ T]$, то из (15) следует существование отрезка $[\tau,\ \tau+\varepsilon]\subseteq[t_0,\ T]$, $\varepsilon>0$ (или отрезка $[t-\varepsilon,\ \tau]$ при $\tau=T$), такого, что

$$g(t)=H(x_k(t),\ u_{k+1}(t),\ t,\ \psi_k(t))-H(x_k(t),\ u_k(t),\ t,\ \psi_k(t))\geqslant\alpha>0 \tag{19}$$

при всех t $(\tau\leqslant t\leqslant\tau+\varepsilon)$. Из (14) и (19) заключаем, что первое слагаемое из формулы (18) для $J(u_{k+1})-J(u_k)$ отрицательно. Это значит, что если остаточный член R в (18) достаточно мал, то $J(u_{k+1})<J(u_k)$, т. е. управление u_{k+1} «лучше» управления u_k. Однако может случиться, что $J(u_{k+1})>J(u_k)$. Что тогда делать? В этом случае вместо управления $u_{k+1}(t)$ можно указать

другое «лучшее» управление $v_{k+1}(t)$, для которого $J(v_{k+1}) < J(u_k)$. А именно, возьмем игольчатую вариацию

$$\Delta u_k(t) = \begin{cases} u_{k+1}(t) - u_k(t), & \tau \leqslant t \leqslant \tau + \varepsilon, \\ 0, & t \in [t_0, T] \setminus [\tau, \tau + \varepsilon], \end{cases}$$

и положим $v_{k+1}(t) = u_k(t) + \Delta u_k(t)$ $(t_0 \leqslant t \leqslant T)$. Из формул (16), (17) при $u = u_k$, $\Delta u = \Delta u_k$ и оценки (3.9) при $u = u_k(t)$, $v = u_k(t) + \Delta u_k(t)$, $y_0 = x_0$ с учетом непрерывности производных f_x^0, f_x, g_y^0 имеем

$$J(v_{k+1}) - J(u_k) = -\int_{\tau}^{\tau+\varepsilon} g(t)\,dt + R, \quad R = o(\varepsilon), \quad \lim_{\varepsilon \to 0} \frac{o(\varepsilon)}{\varepsilon} = 0.$$

Отсюда и из (19) следует $J(v_{k+1}) - J(u_k) \leqslant \varepsilon[-\alpha + o(\varepsilon)/\varepsilon] < 0$, если $\varepsilon > 0$ взять достаточно малым. Таким образом, при выполнении условия (15) всегда можно подобрать управление $v_{k+1} = v_{k+1}(t)$ такое, что $J(v_{k+1}) < J(u_k)$.

Описанный итерационный метод может быть применен для решения более сложных задач оптимального управления, когда имеющиеся ограничения на управление и фазовые координаты удается учесть с помощью штрафных функций и свести задачу к задаче вида (8)—(10). Например, если функцию (8) нужно минимизировать при условиях (9), (10) и закрепленном правом конце $x(T) = x_1$ или фазовом ограничении вида $x^1(t) \leqslant 1$ $(t_0 \leqslant t \leqslant T)$ и т. п., то можно ввести штрафную функцию $P_k(u) = A_k|x(T) - x_1|^2$ или $P_k(u) = A_k \int_{t_0}^{T} (\max\{x^1(t) - 1; 0\})^2\,dt$ и перейти к рассмотрению последовательности задач минимизации функции $\Phi_k(u) = J(u) + A_k P_k(u)$ при условиях (9), (10), $\lim_{k\to\infty} A_k = \infty$. Более подробно об описанном итерационном методе, о приемах ускорения его сходимости, о его возможностях см. в [204, 326].

§ 5. Связь между принципом максимума и классическим вариационным исчислением

Основной задачей классического вариационного исчисления, как известно [64, 74, 108, 172, 181], является следующая задача: среди всех непрерывных кривых $x = x(t_0)$ $(t_0 \leqslant t \leqslant T)$, имеющих кусочно-непрерывные производные $\dot{x}(t)$ и удовлетворяющих условиям $x(t_0) \in S_0$, $x(T) \in S_1$, найти такую, которая доставляет функции (функционалу)

$$J = \int_{t_0}^{T} f^0(x(t), \dot{x}(t))\,dt$$

минимальное значение. Здесь $x(t)=(x^1(t), \ldots, x^n(t))$, S_0 и S_1 — заданные множества в E^n. Будем предполагать, что функция $f_x^0(x, u, t)$ непрерывна и имеет непрерывные производные f_x^0, f_u^0, f_t^0, f_{ut}^0, f_{ux}^0, f_{uu}^0 при $(x, u, t)\in E^n\times E^n\times[t_0, \infty)$. Далее, в этом параграфе для простоты мы ограничимся рассмотрением случая закрепного левого конца: $x(t_0)=x_0$, t_0 — задано, а правый конец $x(T)$ либо закреплен: $x(T)=x_1$, T — задано, либо свободный: $S_1\equiv E^n$, T — задано, либо является подвижным и лежит на заданной гладкой кривой

$$S_1=S_1(T)=\{y\in E^n\colon g(y, T)=y-\varphi(T)=0\},$$
$$T\in \mathbf{R}=\{-\infty<t<+\infty\}.$$

Обозначим $\dot{x}(t)=u(t)$ и запишем рассматриваемую задачу в эквивалентном виде как задачу оптимального управления:

$$J(u)=\int_{t_0}^{T} f^0(x(t), u(t), t)\,dt\to\inf,$$
$$\dot{x}(t)=u(t), \quad t_0\leqslant t\leqslant T,$$
$$x(t_0)=x_0, \quad x(T)\in S_1(T).$$

Для исследования этой задачи воспользуемся принципом максимума Понтрягина. Выпишем функцию Гамильтона — Понтрягина

$$H(x, u, t, \psi, a_0)=-a_0f^0(x, u, t)+\langle\psi, u\rangle \tag{1}$$

и сопряженную систему

$$\dot{\psi}=-H_x=a_0f_x^0(x, u, t), \quad a_0\geqslant 0. \tag{2}$$

Для решения $(u(t), x(t))$ $(t_0\leqslant t\leqslant T)$ рассматриваемой задачи должно выполняться необходимое условие

$$H(x(t), u(t), t, \psi(t), a_0)=\sup_{u\in E^n} H(x(t), u, t, \psi(t), a_0), \tag{3}$$
$$t_0\leqslant t\leqslant T,$$

где $\psi(t)$ — решение системы (2) при $u=u(t)$, $x=x(t, u)$ $(t_0\leqslant t\leqslant T)$. Так как в данном случае множество V совпадает со всем пространством E^n, то условие (3) может соблюдаться лишь в стационарной точке, т. е.

$$H_u\equiv -a_0f_u^0(x(t), u(t), t)+\psi(t)=0, \quad t_0\leqslant t\leqslant T. \tag{4}$$

Отсюда ясно, что $a_0\neq 0$, так как при $a_0=0$ из (4) получим $\psi(t)\equiv 0$, что противоречит теоремам 2.1, 2.2. Следовательно, можно считать, что $a_0=1$. Тогда соотношения (1)—(4) пере-

пишутся соответственно в виде

$$H(x, u, t, \psi) = -f^0(x, u, t) + \langle \psi, u \rangle, \tag{5}$$

$$\dot{\psi}(t) = f^0_x(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T, \tag{6}$$

$$H(x(t), u(t), t, \psi(t)) = \sup_{u \in E^n} H(x(t), u, t, \psi(t)), \quad t_0 \leqslant t \leqslant T, \tag{7}$$

$$\psi(t) = f^0_u(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T. \tag{8}$$

Из уравнения (6) имеем: $\psi(t) = \int\limits_{t_0}^{t} f^0_x(x(t), u(t), t)\, dt + \psi(t_0)$.

С учетом (8) отсюда получаем

$$f^0_u(x(t), u(t), t) = \int\limits_{t_0}^{t} f^0_x(x(\tau), u(\tau), \tau)\, d\tau + \psi(t_0), \quad t_0 \leqslant t \leqslant T. \tag{9}$$

Уравнение (9) называется уравнением Эйлера в интегральной форме; здесь $u(t) = \dot{x}(t)$ $(t_0 \leqslant t \leqslant T)$. Если (9) продифференцировать по t, то получим уравнение Эйлера классического вариационного исчисления в дифференциальной форме

$$\frac{d}{dt}\left(f^0_u(x(t), u(t), t)\right) - f^0_x(x(t), u(t), t) = 0, \quad u(t) = \dot{x}(t), \quad t_0 \leqslant t \leqslant T.$$

Далее, необходимым условием достижения функцией $H(x(t), u(t), t, \psi(t))$ максимума при $u = u(t)$ является неположительность следующей квадратичной формы:

$$\sum_{i,j=1}^{n} H_{u^i u^j}(x(t), u(t), t, \psi(t))\, \xi_i \xi_j \leqslant 0$$

при любых

$$\xi = (\xi_1, \xi_2, \ldots, \xi_n), \quad t_0 \leqslant t \leqslant T.$$

Отсюда, учитывая выражение (5) для H, имеем

$$\sum_{i,j=1}^{n} f^0_{u^i u^j}(x(t), u(t), t)\, \xi_i \xi_j \geqslant 0, \quad \xi \in E^n, \quad t_0 \leqslant t \leqslant T. \tag{10}$$

Условие (10) называется необходимым условием Лежандра. В частности, при $n = 1$ отсюда имеем

$$f^0_{uu}(x(t), u(t), t) \geqslant 0, \quad t_0 \leqslant t \leqslant T.$$

Теперь выведем необходимое условие Вейерштрасса. Для этого перепишем условие (7) с учетом (5), (8) в следующем виде:

$$0 \leqslant H(x(t), u(t), t, \psi(t)) - H(x(t), v, t, \psi(t)) =$$

$$= f^0(x(t), v, t) - f^0(x(t), u(t), t) - \langle v - u(t), f^0_u(x(t), u(t), t) \rangle. \tag{11}$$

Это неравенство справедливо при любых $v \in E^n$, $t \in [t_0, T]$, если

$(u(t),\ x(t))$ $(t_0 \leqslant t \leqslant T)$ — решение исходной задачи. Введем в рассмотрение функцию

$$E(t, x, u, v) \equiv f^0(x, v, t) - f^0(x, u, t) - \langle v - u, f_u^0(x, u, t)\rangle, \quad (12)$$

называемую функцией Вейерштрасса. Известно в классическом вариационном исчислении необходимое условие Вейерштрасса

$$E(t, x(t), u(t), v) \geqslant 0, \quad t_0 \leqslant t \leqslant T, \quad v \in E^n,$$

является следствием неравенства (11). Далее, из теорем 2.1 — 2.3 следует непрерывность функций $\psi(t)$ и $H(t) = \sup\limits_{u \in E^n} H(x(t), u, t, \psi(t))$ на отрезке $[t_0, T]$. Поэтому с учетом соотношений (5), (7), (8) имеем

$$\begin{gathered} [f_u^0(x(t), u(t), t]_t = 0, \\ [\langle u(t), f_u^0(x(t), u(t), t)\rangle - f^0(x(t), u(t), t)]_t = 0, \quad t_0 \leqslant t \leqslant T; \end{gathered} \quad (13)$$

здесь принято обозначение: $[z(t)]_t = z(t+0) - z(t-0)$. Поскольку равенства (13) выполнены при всех t $(t_0 \leqslant t \leqslant T)$, то они сохраняют силу, в частности, и в те моменты t, когда функция $x(t)$ может иметь излом, т. е. производная $\dot{x}(t)$ терпит разрыв. Таким образом, если учесть связь $u(t) \equiv \dot{x}(t)$, условия (13) превращаются в известные из классического вариационного исчисления условия Эрдмана — Вейерштрасса в точках излома кривой $x(t)$ $(t_0 \leqslant t \leqslant T)$.

Перейдем к рассмотрению условий на правом конце оптимальной кривой $x(t)$ $(t_0 \leqslant t \leqslant T)$. Если конец $x(T)$ свободен, то в силу условия (2.26) $\psi(T) = 0$. Отсюда с учетом выражения (8) имеем

$$f_u^0(x(T), u(T), T) = 0. \quad (14)$$

Если правый конец $x(T)$ подвижен, точнее, $x(T) \in S_1(T) = \{y \in E^n\colon\ g^j(y, T) \equiv y^j - \varphi_j(T) \equiv 0,\ \ j = 1, \ldots, n\}$, то согласно условиям (2.30), (2.41) существуют постоянные $a_1, \ldots, a_n$ такие, что

$$\psi_i(T) = -\sum_{j=1}^n a_j g_{y^i}^j(x(T), T) = -a_i,$$

$$H(x(T), u(T), T, \psi(T)) = \sum_{j=1}^n a_j g_t^j(x(T), T) =$$

$$= -\sum_{j=1}^n a_j \dot{\varphi}_j(T) = \sum_{j=1}^n \psi_j(T)\dot{\varphi}_j(T) = \langle \psi(T), \dot{\varphi}(T)\rangle.$$

Так как $H(x, u, t, \psi) \equiv \langle \psi, u\rangle - f^0(x, u, t)$ и $\psi(t)$ выражается формулой (8), то последнее неравенство можно переписать так:

$$f^0(x(T), u(T), T) + \langle f_u^0(x(T), u(T), T), \dot{\varphi}(T) - u(T)\rangle = 0. \quad (15)$$

Условия (14), (15) при учете связи $\dot{x}(t) \equiv u(t)$ выражают собой известные в классическом вариационном исчислении условия трансверсальности для свободного и соответственно подвижного правого конца.

Таким образом, в случае $V \equiv E^n$ из принципа максимума следуют все основные необходимые условия, известные в классическом вариационном исчислении [64, 74, 108, 161, 172, 181, 182, 234, 342]. Однако, если V — замкнутое множество и $V \neq E^n$, то соотношение (4), вообще говоря, не выполняется. Более того, имеются примеры, когда и условие Вейерштрасса в этом случае не имеет места ([17, с. 284]). Условие максимума (2.9), являясь естественным обобщением условия Вейерштрасса из классического вариационного исчисления, имеет то существенное преимущество перед условием Вейерштрасса, что оно применимо для любого (в частности, и замкнутого) множества $V \in E^r$ и для более общих задач. Заметим, что именно случай замкнутого множества наиболее интересен в прикладных вопросах, поскольку значения оптимальных управлений чаще всего лежат на границе V.

Глава 7

ДИНАМИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

В этой главе мы остановимся на методе динамического программирования, часто используемом для численного решения задач оптимального управления при наличии фазовых ограничений, конечномерных задач минимизации специального вида. С помощью динамического программирования можно также наметить пути решения проблемы синтеза, сформулировать достаточные условия оптимальности для задач оптимального управления и т. д. Изложение метода динамического программирования начнем с простейшей схемы Беллмана для задачи оптимального управления, затем опишем более совершенную и удобную для практики схему Моисеева, а также обсудим некоторые другие аспекты этого метода [12, 14, 15, 25, 26, 50, 57—60, 65, 68, 81, 101, 124, 161, 169, 188—190, 213, 215, 217, 225, 234, 262, 263, 265, 285, 305, 308, 317, 319, 326, 327].

§ 1. Схема Беллмана. Проблема синтеза для дискретных систем

1. Рассмотрим следующую задачу оптимального управления:

$$J(t_0, x_0, u(\cdot)) = \int_{t_0}^{T} f^0(x(\tau), u(\tau), \tau)\, d\tau + \Phi(x(T)) \to \inf, \qquad (1)$$

$$\dot{x}(\tau) = f(x(\tau),\ u(\tau),\ \tau), \quad t_0 \leqslant \tau \leqslant T, \quad x(t_0) = x_0, \qquad (2)$$

$$x(\tau) \in G(\tau), \quad t_0 \leqslant \tau \leqslant T, \qquad (3)$$

$$u = u(\tau) \in V(\tau), \quad t_0 \leqslant \tau \leqslant T; \quad u(\tau) \text{ — кусочно-непрерывна}; \qquad (4)$$

моменты времени t_0, T будем считать заданными (описание обозначений см. в § 6.1).

Для приближенного решения этой задачи разобьем отрезок $t_0 \leqslant t \leqslant T$ на N частей точками $t_0 < t_1 < \ldots < t_{N-1} < t_N = T$, и, приняв эти точки в качестве узловых, интеграл в (1) заменим квадратурной формулой прямоугольников, уравнения (2) — разностными уравнениями с помощью явной схемы Эйлера [4, 13, 20, 39, 54]. В результате придем к следующей дискретной задаче

оптимального управления

$$I_0(x; [u_i]_0) = \sum_{i=0}^{N-1} F_i^0(x_i, u_i) + \Phi(x_N) \to \inf, \tag{5}$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = 0, 1, \ldots, N-1, \quad x_0 = x \in G_0, \tag{6}$$

$$x_i \in G_i, \quad i = 0, 1, \ldots, N, \tag{7}$$

$$[u_i]_0 = (u_0, u_1, \ldots, u_{N-1}): u_i \in V_i, \quad i = 0, 1, \ldots, N-1, \tag{8}$$

где $F_i^0(x, u) = f^0(x, u, t_i)(t_{i+1} - t_i)$, $F_i(x, u) = x + f(x, u, t_i) \times \times (t_{i+1} - t_i)$, $G_i = G(t_i)$, $V_i = V(t_i)$. Заметим, что задача (5)—(8) имеет также и самостоятельный интерес и возникает при описании управляемых дискретных (импульсных) систем, в которые сигналы управления поступают в дискретные моменты времени, фазовые координаты также меняются дискретно [44, 66, 101, 213, 254, 322, 323].

Если задать какое-либо дискретное управление $[u_i]_0 = (u_0, u_1, \ldots, u_{N-1})$ и начальное условие $x_0 = x \in G_0$, то система (6) однозначно определяет соответствующую дискретную траекторию $[x_i]_0 = [x_i(x; [u_i]_0)]_0 = (x_0, x_1, \ldots, x_N)$. Зафиксируем некоторое $x \in G_0$ и через $\Delta_0(x)$ обозначим множество управлений $[u_i]_0$ таких, что 1) выполнены условия (8); 2) дискретная траектория $[x_i]_0$, соответствующая управлению $[u_i]_0$ и выбранному начальному условию $x_0 = x$, удовлетворяют ограничениям (7). Пару $([u_i]_0, [x_i]_0)$, состоящую из управления и траектории, будем называть допустимой для задачи (5)—(8) или, короче, допустимой парой, если эта пара удовлетворяет всем условиям (6)—(8) или, иначе говоря, $[u_i]_0 \in \Delta_0(x_0)$.

Множество $\Delta_0(x)$ может быть пустым или непустым. Если $\Delta_0(x) = \varnothing$ при всех $x \in G_0$, то условия (6)—(8) несовместны и функция (5) будет определена на пустом множестве. Поэтому, чтобы задача (5)—(8) имела смысл, естественно требовать существование хотя бы одной точки $x \in G_0$, для которой $\Delta_0(x) \neq \varnothing$. Обозначим $X_0 = \{x: x \in G_0, \Delta_0(x) \neq \varnothing\}$. Тогда задача (5)—(8) может быть сформулирована совсем кратко: минимизировать функцию $I_0(x, [u_i]_0)$ при $[u_i]_0 \in \Delta_0(x)$ $(x \in X_0)$. Положим

$$I_0^* = \inf_{x \in X_0} \inf_{[u_i]_0 \in \Delta_0(x)} I_0(x, [u_i]_0).$$

Допустимую пару $([u_i^*]_0, [x_i^*]_0)$ назовем решением задачи (5)—(8), если $I_0(x_0^*, [u_i^*]_0) = I_0^*$, $[u_i^*]_0$ назовем оптимальным управлением, $[x_i^*]_0$ — оптимальной траекторией задачи (5)—(8).

Как видим, задача (5)—(8) является уже известной нам задачей минимизации функции $n + Nr$ переменных $x, u_0, u_1, \ldots, u_{N-1}$ и для ее решения в принципе могут быть использованы методы, описанные в главах 1—3, 5. Однако в практических за-

дачах число $n+Nr$ обычно бывает столь большим, что непосредственное использование методов глав 1—3, 5, вообще говоря, сильно осложняется: вызывает трудности также и то обстоятельство, что множества $\Delta_0(x)$, X_0, на которых минимизируется функция $I_0(x, [u_i]_0)$, заданы неявно. Для преодоления этих трудностей здесь часто пользуются методом динамического программирования, с помощью которого задачу (5)—(8) большого числа переменных удается свести к последовательности конечного числа задач минимизации функций меньшего числа переменных.

2. Для изложения метода динамического программирования нам понадобятся следующие вспомогательные задачи:

$$I_k(x, [u_i]_k) = \sum_{i=k}^{N-1} F_i^0(x_i, u_i) + \Phi(x_N) \to \inf, \tag{9}$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = k, \ldots, N-1, \quad x_k = x, \tag{10}$$

$$x_i \in G_i, \quad i = k, \ldots, N, \tag{11}$$

$$[u_i]_k = (u_k, u_{k+1}, \ldots, u_{N-1}): u_i \in V_i, \quad i = k, \ldots, N-1, \tag{12}$$

где точка x и целое число k фиксированы, $x \in G_k$ $(0 \leqslant k \leqslant N-1)$. При $k=0$ отсюда получим исходную задачу (5)—(8). Через $\Delta_k(x)$ обозначим множество всех управлений $[u_i]_k$, удовлетворяющих условиям (12) и таких, что соответствующая ему траектория $[x_i]_k = (x_k = x, x_{k+1}, \ldots, x_N)$ из (10) удовлетворяет фазовым ограничениям (11). Пару $([u_i]_k, [x_i]_k)$ назовем допустимой парой для задачи (9)—(12), если $[u_i]_k \in \Delta_k(x)$. Допустимую пару $([u_i^*]_k, [x_i^*]_k)$ будем называть решением задачи (9)—(12), если $I_k(x, [u_i^*]_k) = I_k^*(x) = \inf\limits_{\Delta_k(x)} I_k(x, [u_i]_k)$; $[u_i^*]_k$ назовем оптимальным управлением, $[x_i^*]_k$ — оптимальной траекторией задачи (9)—(12).

Нетрудно видеть, что если $X_0 \neq \varnothing$, то $\Delta_k(x) \neq \varnothing$ хотя бы для одного $x \in G_k$. Введем функцию

$$B_k(x) = \inf_{\Delta_k(x)} I_k(x, [u_i]_k), \quad k = 0, 1, \ldots, N-1,$$

называемую *функцией Беллмана* задачи (5)—(8). Область определения функции $B_k(x)$ представляет собой множество $X_k = \{x \in G_k: \Delta_k(x) \neq \varnothing\}$. Покажем, что функция Беллмана задачи (5)—(8) удовлетворяет некоторым рекуррентным соотношениям, называемым *уравнением Беллмана*.

Теорема 1. *Функция Беллмана задачи* (5)—(8) *необходимо является решением уравнения*

$$B_k(x) = \inf_{u \in D_k(x)} \left[F_k^0(x, u) + B_{k+1}(F_k(x, u))\right],$$

$$x \in X_k, \quad k = 0, 1, \ldots, N-1, \tag{13}$$

где

$$B_N(x) = \Phi(x), \quad x \in G_N, \tag{14}$$

$D_k(x)$ — *множество всех тех* $u \in V_k$, *для которых существует хотя бы одно управление* $[u_i]_k \in \Delta_k(x)$ *с компонентой* $u_k = u$. *Верно и обратное: функция* $B_k(x)$, $x \in X_k$ $(k = 0, 1, \ldots, N-1)$, *определяемая условиями* (13), (14), *является функцией Беллмана задачи* (5)—(8).

Доказательство проведем в предположении, что все упоминаемые в теореме 1 нижние грани конечны.

Необходимость. Пусть $B_k(x) = \inf\limits_{\Delta_k(x)} I_k(x, [u_i]_k)$ $(x \in X_k$, $k = 0, 1, \ldots, N-1)$; $B_N(x) = \Phi(x)$, $x \in G_N$. Покажем, что эти функции удовлетворяют уравнению (13). Из определения множеств $\Delta_k(x)$, $D_k(x)$ видно, что множества $D_k(x)$ и $\Delta_k(x)$ оба пусты или непусты одновременно, и поскольку $x_{k+1} = F_k(x, u)$, то для непустоты этих множеств необходимо и достаточно, чтобы $\Delta_{k+1}(F_k(x, u)) \neq \varnothing$. Справедливость соотношения (13) при $k = N-1$ очевидным образом вытекает из условия $B_N(x) \equiv \Phi(x)$ и представления $I_{N-1}(x, [u_i]_{N-1}) \equiv F^0_{N-1}(x, u) + \Phi(F_{N-1}(x, u))$, верного для любого $u \in D_{N-1}(x) \equiv \Delta_{N-1}(x) \equiv \{u\colon u \in V_{N-1},\ x_N = F_{N-1}(x, u) \in G_N,\ x \in G_{N-1}\}$. Докажем (13) при k $(0 \leqslant k < N-1)$. Для этого сначала убедимся в том, что

$$B_k(x) \leqslant \inf_{u \in D_k(x)} \left[F^0_k(x, u) + B_{k+1}(F_k(x, u))\right], \quad x \in X_k. \tag{15}$$

Возьмем произвольное $u \in D_k(x)$ и с этим управлением «выйдем» из точки x в момент k. В момент $k+1$ придем в точку $x_{k+1} = F_k(x, u)$, для которой $\Delta_{k+1}(x_{k+1}) \neq \varnothing$. По определению $B_{k+1}(x_{k+1}) = \inf\limits_{\Delta_{k+1}(x_{k+1})} I_{k+1}(x_{k+1}, [u_i]_{k+1})$ для любого $\varepsilon > 0$ найдется управление $[u_i^\varepsilon]_{k+1} \in \Delta_{k+1}(x_{k+1})$ такое, что $B_{k+1}(x_{k+1}) \leqslant I_{k+1}(x_{k+1}, [u_i^\varepsilon]_{k+1}) \leqslant B_{k+1}(x_{k+1}) + \varepsilon$. Поскольку $[\bar{u}_i]_k = (u, u^\varepsilon_{k+1}, \ldots, u^\varepsilon_{N-1}) \in \Delta_k(x)$, то $B_k(x) \leqslant I_k(x, [\bar{u}_i]_k) = F^0_k(x, u) + I_{k+1}(x_{k+1}, [u_i^\varepsilon]_{k+1}) \leqslant F^0_k(x, u) + B_{k+1}(x_{k+1}) + \varepsilon \equiv F^0_k(x, u) + B_{k+1}(F_k(x, u)) + \varepsilon$. В силу произвольности $u \in D_k(x)$ и величины $\varepsilon > 0$ отсюда следует неравенство (15).

Теперь покажем, что в (15) на самом деле знак неравенства можно заменить знаком равенства. По определению $\inf\limits_{\Delta_k(x)} I_k(x, [u_i]_k) = B_k(x)$ для каждого $\varepsilon > 0$ найдется такое управление $[v_i^\varepsilon]_k \in \Delta_k(x)$, что $B_k(x) \leqslant I_k(x, [v_i^\varepsilon]_k) \leqslant B_k(x) + \varepsilon$. Но $[\bar{v}_i]_{k+1} \equiv (v^\varepsilon_{k+1}, \ldots, v^\varepsilon_{N-1}) \in \Delta_{k+1}(F_k(x, v^\varepsilon_k))$, поэтому

$$F^0_k(x, v^\varepsilon_k) + B_{k+1}(F_k(x, v^2_k)) \leqslant F^0_k(x, v^\varepsilon_k) + I_{k+1}(F_k(x, v^\varepsilon_k), [\bar{v}_i]_{k+1}) = I_k(x, [v_i^\varepsilon]_k) \leqslant B_k(x) + \varepsilon.$$

Так как $v_k^* \in D_k(x)$, то отсюда имеем: $\inf\limits_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1} \times \times (F_k(x, u))] \leqslant B_k(x) + \varepsilon$, или в силу произвольности $\varepsilon > 0$:

$$\inf_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1}(F_k^*(x, u))] \leqslant B_k(x), \quad x \in X_k.$$

Отсюда и из (15) немедленно следует равенство (13).

Достаточность. Пусть функции $B_k(x)$ $(x \in X_k,\ k = 0, 1, \ldots, N-1)$, определены из условий (13), (14). Спрашивается: какое отношение имеют эти функции к задаче (5)—(8)? Покажем, что при каждом $x \in X_k$ величина $B_k(x)$ равна нижней грани функции (9) при условиях (10)—(12). Отметим, что условия (13), (14) однозначно определяют функции $B_k(x)$ $(x \in X_k,\ k = 0, 1, \ldots, N)$. Это легко доказывается с помощью индукции последовательным перебором в (13), (14) номеров $k = N, N-1, \ldots, 0$ с учетом того, что функции $\Phi(x)$, $F_k^0(x, u)$, $F_k(x, u)$ однозначны и нижняя грань функций определяется также однозначно. С другой стороны, как было установлено выше, функции $\inf\limits_{\Delta_k(x)} I_k(x, [u_i]_k)$ $(x \in X_k)$, также являются решением уравнения (13) при условии (14). Из единственности решения системы (13), (14) тогда следует, что $B_k(x) = \inf\limits_{\Delta_k(x)} I_k(x, [u_i]_k)$ $(x \in X_k,\ k = 0, 1, \ldots, N-1)$. Теорема 1 доказана.

3. Пользуясь условиями (13), (14), можно последовательно определить функции $B_k(x)$ и их области определения X_k $(k = N, N-1, \ldots, 1; 0)$. А именно $B_N(x) \equiv \Phi(x)$, $x \in G_N \equiv X_N$ — известны. Если известны $B_{k+1}(x)$ и X_{k+1} $(k \leqslant N-1)$, то для определения $B_k(x)$ нужно решить задачу минимизации функции $\varphi(x, u) \equiv F_k^0(x, u) + B_{k+1}(F_k(x, u))$ переменных $u = (u^1, \ldots, u^r)$ на известном множестве $D_k(x) = \{u\colon u \in V_k,\ F_k(x, u) \in X_{k+1}\}$. Здесь могут быть использованы методы глав 1—3, 5. Очевидно, функция $B_k(x)$ определена в точке x тогда и только тогда, когда $D_k(x) \neq \varnothing$. Таким образом, при определении значений функции $B_k(x)$ одновременно находится и область ее определения $X_k = \{x\colon x \in G_k,\ D_k(x) \neq \varnothing\} \equiv \{x\colon x \in G_k,\ \Delta_k(x) \neq \varnothing\}$. Так как $\Delta_k(x) \neq \varnothing$ хотя бы при одном $x \in G_k$, то $X_k \neq \varnothing$ $(k = N, N-1, \ldots, 1, 0)$.

Предположим, что нам удалось найти функции $B_k(x)$ из условий (13), (14) и, кроме того, пусть также известны функции $u_k(x) \in D_k(x)$ $(x \in X_k,\ k = 0, 1, \ldots, N-1)$, на которых достигается нижняя грань в правой части (13). Тогда, оказывается, решения задач (5)—(8) и (9)—(12) выписываются совсем просто. А именно, оптимальное управление $[u_i^*]_0$ и соответст-

вующая траектория $[x_i^*]_0$ для задачи (5)—(8) определяются следующим образом: сначала из условия

$$\inf_{x \in X_0} B_0(x) = B_0(x_0^*) \tag{16}$$

находят $x_0^* \in X_0$, затем последовательно полагают

$$\begin{gathered} u_0^* = u_0(x_0^*), \quad x_1^* = F_0(x_0^*, u_0^*), \quad u_1^* = u_1(x_1^*), \\ x_2^* = F_1(x_1^*, u_1^*), \ldots, x_N^* = F_{N-1}(x_{N-1}^*, u_{N-1}^*). \end{gathered} \tag{17}$$

Оптимальное управление $[u_i^*]_k$ и траектория $[x_i^*]_k$ для задачи (9)—(12) определяются аналогично:

$$\begin{gathered} x_k^* = x, \quad u_k^* = u_k(x_k^*), \\ x_{k+1}^* = F_k(x_k^*, u_k^*), \ldots, x_N^* = F_{N-1}(x_{N-1}^*, u_{N-1}^*). \end{gathered} \tag{18}$$

Для доказательства этих утверждений введем вспомогательные функции:

$$R_i(x, u) \equiv B_{i+1}(F_i(x, u)) - B_i(x) + F_i^0(x, u), \quad i = 0, 1, \ldots, N-1. \tag{19}$$

Очевидно, уравнения Беллмана (13) тогда можно переписать в эквивалентном виде:

$$\inf_{u \in D_k(x)} R_k(x, u) \equiv R_k(x, u_k(x)) = 0, \quad k = 0, 1, \ldots. N-1. \tag{20}$$

Кроме того, с помощью функций $R_i(x, u)$ значение функции (9) на любом управлении $[u_i]_k \in \Delta_k(x)$ и $x \in X_k$ можно выразить формулой

$$I_k(x, [u_i]_k) = \sum_{i=k}^{N-1} R_i(x_i, u_i) + B_k(x) \tag{21}$$

при всех $k = 0, 1, \ldots, N-1$. В самом деле, учитывая равенство $B_N(x) \equiv \Phi(x)$, из (10), (19) имеем $\sum_{i=k}^{N-1} R_i(x_i, u_i) = \sum_{i=k}^{N-1} [B_{i+1}(x_{i+1}) - B_i(x_i) + F_i^0(x_i, u_i)] = B_N^-(x_N) - B_k(x) + \sum_{i=k}^{N-1} F_i^0(x_i, u_i) = I_k(x, [u_i]_k) - B_k(x)$, что равносильно (21).

Теорема 2. *Пусть найдены функции $B_k(x)$ из* (13), (14) *и их области определения X_k, а также функции $u = u_k(x)$ ($x \in X_k$, $k = 0, 1, \ldots, N-1$), на которых достигается нижняя грань в уравнении* (13) *или* (20), *и пусть x_0^* определена условием* (16). *Тогда оптимальное управление $[u_i^*]_0$ и траектория $[x_i^*]_0$ для задачи* (5)—(8) *определяются соотношениями* (16), (17).

Доказательство. Из определения $u(x)$, $[u_i^*]_0$, $[x_i^*]_0$ и эквивалентности записей уравнения Беллмана (13) и (20) имеем

$$R_i(x_i^*, u_i^*) \equiv R_i(x_i^*, u_i(x_i^*)) = \inf_{u \in D_i(x_i^*)} R_i(x_i^*, u) = 0,$$

$$i = 0, 1, \ldots, N-1. \quad (22)$$

Возьмем произвольные $x \in X_0$; управление $[u_i]_0 \in \Delta_0(x)$ с соответствующей траекторией $[x_i]_0$ из (6). Так как $u_i \in D_i(x_i)$, то из уравнения (20) и определения $u_i(x)$ следует

$$R_i(x_i, u_i) \geqslant \inf_{u \in D_i(x_i)} R_i(x_i, u) = R_i(x_i, u_i(x_i)) = 0,$$

$$i = 0, 1, \ldots, N-1. \quad (23)$$

С помощью формулы (21) при $k = 0$ с учетом соотношений (16), (22), (23) получаем $I_0(x, [u_i]_0) - I_0(x_0^*, [u_i^*]_0) = \sum_{i=0}^{N-1} [R_i(x_i, u_i) - R_i(x^*, u_i^*)] + B_0(x) - B_0(x_0^*) \geqslant 0$ для любых $x \in X_0$ и $[u_i]_0 \in \Delta_0(x)$, что и требовалось.

Теорема 3. *Пусть известны $B_k(x)$ $(x \in X_k)$ из* (13), (14), *а также функции $u_k(x)$, на которых достигается нижняя грань в уравнении* (13) (*или* (20)). *Тогда оптимальное управление $[u_i^*]_k$ и траектория $[x_i^*]_k$ для задачи* (9)—(12) *определяются формулами* (18).

Доказательство. Возьмем произвольное управление $[u_i]_k \in \Delta_k(x)$ и соответствующую траекторию $[x_i]_k$ из (10). Очевидно, соотношения (22), (23) остаются справедливыми и здесь при всех $i = k, \ldots, N-1$. Отсюда с помощью (21) получим

$$I_k(x, [u_i]_k) - I_k(x, [u_i^*]_k) = \sum_{i=k}^{N-1} [R_i(x_i, u_i) - R_i(x_i^*, u_i^*)] \geqslant 0,$$

что и требовалось.

4. В теории оптимального управления и ее приложениях важное место занимает так называемая *проблема синтеза*, заключающаяся в построении функции $u = u_k(x)$, выражающей собой оптимальное управление при условии, что в момент k объект находится в точке x фазового пространства. Такая функция $u_k(x)$ называется синтезирующей.

Теорема 3 показывает, что решение уравнения Беллмана (13) равносильно решению проблемы синтеза для задачи (5)—(8). А именно, функция $u_k(x)$, на которой достигается нижняя грань в (13), является синтезирующей: если в момент k объект находится в точке $x \in X_k$, то дальнейшее оптимальное движение объекта определяется условиями: $x_{i+1} = F_i(x_i, u_i(x_i))$ $(i = k, \ldots, N-1)$, $x_k = x$ (если $x \notin X_k$, то $\Delta_k(x) = \varnothing$,— движение с со-

блюдением условий (10)—(12) невозможно). Достаточные условия существования функции Беллмана и синтезирующей функции для задачи (5)—(8) даются в следующей теореме.

Теорема 4. *Пусть множества* G_k $(k=0, 1, \ldots, N)$ *замкнуты, множества* V_k $(k=0, 1, \ldots, N-1)$ *замкнуты и ограничены, функция* $F_k^0(x, u)$ *полунепрерывна снизу, а функция* $F_k(x, u)$ *непрерывна по совокупности аргументов* (x, u) *при* $x \in G_k$, $u \in V_k$ $(k=0, 1, \ldots, N-1)$, $\Phi(x)$ *полунепрерывна снизу на множестве* G_N. *Тогда*: 1) *множества* X_k $(k=0, 1, \ldots, N)$ *замкнуты, множества* $D_k(x)$ $(k=0, 1, \ldots, N-1)$ *замкнуты и ограничены равномерно по* $x \in X_k$; 2) *нижняя грань в правой части* (13) *достигается хотя бы при одном* $u=u_k(x) \in D_k(x)$; 3) *функция* $B_k(x)$ *полунепрерывна снизу на* X_k $(k=0, 1, \ldots, N)$.

Доказательство. По условию $G_N \equiv X_N$ замкнуто, $\Phi(x) \equiv B_N(x)$ полунепрерывна снизу на X_N. Сделаем индуктивное предположение: пусть X_{k+1} замкнуто, $B_{k+1}(x)$ полунепрерывна снизу на X_{k+1} при некотором k $(0 \leqslant k \leqslant N-1)$. Докажем, что тогда X_k замкнуто и на X_k справедливы все утверждения теоремы. Так как $D_k(x)=\{u\colon u \in V_k,\ F_k(x, u) \in X_{k+1}\} \subseteq V_k$ и V_k ограничено, то $D_k(x)$ ограничено равномерно по $x \in X_k$. Докажем замкнутость $D_k(x)$ при любом фиксированном $x \in X_k$. Пусть $v_m=D_k(x)$ $(m=1, 2, \ldots, v_m \to v)$ при $m \to \infty$. Это значит, что $v_m \in V_k$, $F_k(x, v_m) \in X_{k+1}$ $(m=1, 2, \ldots)$. Из замкнутости V_k, X_{k+1} и непрерывности $F_k(x, v)$ сразу имеем: $v \in V_k$, $\lim\limits_{m\to\infty} F_k(x, v_m)=F_k(x, v) \in X_{k+1}$, т. е. $v \in D_k(x)$. Замкнутость $D_k(x)$ доказана.

Покажем замкнутость $X_k=\{x\colon x \in G_k,\ D_k(x) \neq \varnothing\}$. Пусть $y_m \in X_k$ $(m=1, 2, \ldots)$, $y_m \to y$ при $m \to \infty$. Из замкнутости G_k следует $y \in G_k$. Если мы еще покажем, что $D_k(y) \neq \varnothing$, то это будет означать, что $y \in X_k$, и замкнутость X_k будет доказана. Так как $D_k(y_m) \neq \varnothing$, то существует такое $v_m \in V_k$, что $F_k(y_m, v_m) \in X_{k+1}$ $(m=1, 2, \ldots)$. В силу компактности V_m из последовательности $\{v_m\}$ можно выбрать подпоследовательность $\{v_{m_p}\} \to v \in V_k$ при $p \to \infty$. Поскольку X_{k+1} замкнуто, $F_k(x, u)$ непрерывна, то $\lim\limits_{p\to\infty} F_k(y_{m_p}, v_{m_p})=F_k(y, v) \in X_{k+1}$, т. е. $v \in D_k(y)$. Таким образом, $D_k(y) \neq \varnothing$.

Далее, функция $\varphi(x, u) \equiv F_k^0(x, u)+B_{k+1}(F_k(x, u))$ полунепрерывна снизу по (x, u) при $x \in X_k$, $u \in D_k(x)$ — это следует из непрерывности $F_k(x, u)$ и полунепрерывности снизу $F_k^0(x, u)$, $B_{k+1}(x)$. Поскольку $D_k(x)$ — замкнутое ограниченное множество, то в силу теоремы 2.1.1 $\varphi(x, u)$ при каждом фиксированном $x \in X_k$ достигает своей нижней грани на $D_k(x)$ хотя бы в одной точке $u=u_k(x) \in D_k(x)$. Таким образом, $B_k(x)=\inf\limits_{u\in D_k(x)} \varphi(x, u)=\varphi(x, u_k(x))$, в силу уравнения Беллмана (13).

Остается еще доказать полунепрерывность снизу $B_k(x)$ на X_k. Пусть $x, y_m \in X_k$, $y_m \to x$ при $m \to \infty$, $B_k(y_m) = \varphi(y_m, u_k(y_m))$. Так как $u_k(y_m) \in D_k(y_m) \in V_k$, то в силу компактности V_k последовательность $\{u_k(y_m),\ m = 1, 2, \ldots\}$ имеет хотя бы одну предельную точку $v \in V_k$. Можем считать, что сама последовательность $\{u_k(y_m)\} \to v$ при $m \to \infty$. Поскольку $F_k(x, u)$ непрерывна, X_{k+1} замкнуто, кроме того, $F_k(y_m, v_k(y_m)) \in X_{k+1}$, то $\lim\limits_{m\to\infty} F_k(y_m, v_k(y_m)) = F_k(y, v) \in X_{k+1}$. Это значит, что $v \in D_k(x)$. Тогда

$$\varliminf_{m\to\infty} B_k(y_m) = \varliminf_{m\to\infty} \varphi(y_m, u_k(y_m)) \geqslant \varphi(x, v) \geqslant \inf_{u\in D_k(x)} \varphi(x, u) = B_k(x).$$

Полунепрерывность $B_k(x)$ на X_k доказана, что и требовалось.

5. Нетрудно привести примеры задач типа (5)—(8), когда нижняя грань в (13) или (16) не достигается (см. ниже упражнение 2). В таких задачах, конечно, приходится пользоваться величинами, лишь приближенно реализующими нижнюю грань в (13), (16). Но даже в том случае, когда нижняя грань в (13), (16) достигается, получить точные выражения для функций $B_k(x)$, $u_k(x)$ и точки x_0^* из (13), (16) часто бывает затруднительно. Поэтому на практике часто пользуются соотношениями (16), (17) для приближенных $B_k(x)$, $u_k(x)$ и вместо точных управлений $[u_i^*]_0$ и траектории $[x_i^*]_0$ получают какие-то их приближения. Возникает вопрос, насколько отличается полученное таким образом приближенное решение задачи (5)—(8) от ее точного решения? Приводимая ниже оценка погрешности дает некоторый ответ на этот вопрос.

Пусть $K_i(x)$ — приближенное значение функции Беллмана $B_i(x)$ $(i = 0, 1, \ldots, N)$. По аналогии с (19) введем функцию

$$S_i(x, u) \equiv K_{i+1}(F_i(x, u)) - K_i(x) + F_i^0(x, u), \qquad i = 0, 1, \ldots, N-1, \tag{24}$$

и, кроме того, положим

$$s_N(x) = \Phi(x) - K_N(x), \quad x \in G_N. \tag{25}$$

Возьмем произвольную допустимую пару $[u_i]_0 = (u_0, u_1, \ldots, u_{N-1})$, $[x_i]_0 = (x_0, x_1, \ldots, x_{N-1})$ задачи (5)—(8). Тогда $x_0 \in X_0$, $[u_i]_0 \in \Delta_0(x_0)$, $u_i \in D_i(x_i)$ $(i = 0, 1, \ldots, N-1)$. Учитывая условие (6), из (24) имеем

$$S_i(x_i, u_i) = K_{i+1}(x_{i+1}) - K_i(x_i) + F_i^0(x_i, u_i), \qquad i = 0, 1, \ldots, N-1.$$

Суммируя эти равенства по i от 0 до $N-1$, с помощью (25) получим формулу

$$J_0(x_0, [u_i]_0) = \sum_{i=0}^{N-1} S_i(x_i, u_i) + s_N(x_N) + K_0(x_0). \tag{26}$$

Если $K_i(x)=B_i(x)$, то $s_N(x)\equiv 0$, и формула (26) превратится в знакомую нам формулу (21) при $k=0$.

Предположим, что каким-либо образом (например, из соотношений (16), (17) с приближенными функциями $B_k(x)$, $u_k(x)$), нам удалось найти некоторое управление $[\bar{u}_i]_0$ и соответствующую ему траекторию $[\bar{x}_i]_0$, удовлетворяющую условиям (6)—(8), т. е. $\bar{x}_0\in X_0$, $[\bar{u}_i]_0\in\Delta_0(\bar{x}_0)$, $\bar{u}_i\in D_i(\bar{x}_i)$ $(i=0, 1, \ldots, N-1)$. Согласно (26) тогда

$$I_0(\bar{x}_0, [\bar{u}_i]_0)=\sum_{i=0}^{N-1} S_i(\bar{x}_i, \bar{u}_i)+s_N(\bar{x}_N)+K_0(\bar{x}_0).$$

Отсюда и из (26) имеем

$$I_0(\bar{x}_0, [\bar{u}_i]_0)-I_0(x_0, [u_i]_0)=$$
$$=\sum_{i=0}^{N-1}[S_i(\bar{x}_i, \bar{u}_i)-S_i(x_i, u_i)]+s_N(\bar{x}_N)-s_N(x_N)+K_0(\bar{x}_0)-K_0(x_0) \quad (27)$$

для любых $x_0\in X_0$, $[u_i]_0\in\Delta_0(x_0)$. Учитывая, что $\bar{x}_0$, $x_0\in X_0$, $[\bar{u}_i]_0\in\Delta_0(\bar{x}_0)$, $[u_i]_0\in\Delta_0(x_0)$, $(x_i, u_i)\in X_i\times D_i(x_i)$, перейдем к нижней грани по $(x_0, [u_i]_0)$ сначала в правой части (27), а затем в левой части (27). Получим неравенства

$$0\leqslant I_0(\bar{x}_0, [\bar{u}_i]_0)-I_0^*\leqslant\sum_{i=0}^{N-1}\left[S_i(\bar{x}_i, \bar{u}_i)-\inf_{x\in X_i}\inf_{u\in D_i(x)} S_i(x, u)\right]+$$
$$+s_N(\bar{x}_N)-\inf_{X_N} s_N(x)+K_0(\bar{x}_0)-\inf_{X_0} K_0(x), \quad (28)$$

представляющие собой оценку погрешности, которая будет допущена, если $[\bar{u}_i]_0$, $[\bar{x}_i]$ будут взяты в качестве приближенного решения задачи (5)—(8).

Если $K_i(x)=B_i(x)$ $(i=0, 1, \ldots, N)$, то $S_i(x, u)=R_i(x, u)$, $s_N(x)=0$. Кроме того, из (20) следует, что $\inf\limits_{u\in D_i(x)} R_i(x, u)=0$ для всех $x\in X_i$, так что $\inf\limits_{x\in X_i}\inf\limits_{u\in D_i(x)} R_i(x, u)=0$. Поэтому при $K_i(x)=B_i(x)$ из (28) получим

$$0\leqslant I_0(\bar{x}_0, [\bar{u}_i]_0)-I_0^*\leqslant\sum_{i=0}^{N-1} R_i(\bar{x}_i, \bar{u}_i)+B_0(\bar{x}_0)-\inf_{X_0} B_0(x). \quad (29)$$

Из оценки (29) следует, что если определить $\bar{u}_i(x)\in\Delta_i(x)$, $\bar{x}_0\in X_0$ так, чтобы $R_i(x, \bar{u}_i(x))$ было поближе к $\inf\limits_{u\in D_i(x)} R_i(x, u)=0$, $x\in X_i$, а $B_0(\bar{x}_0)$ — поближе к $\inf\limits_{X_0} B_0(x)$, и затем строить

управление $[\bar{u}_i]$ и траекторию $[\bar{x}_i]$ следующим образом:

$$\bar{u}_0 = \bar{u}_0(\bar{x}_0), \quad \bar{x}_1 = F_0(\bar{x}_0, \bar{u}_0), \quad \bar{u}_1 = \bar{u}_1(x_1), \ldots, x_N = \\ = F_{N-1}(\bar{x}_{N-1}, \bar{u}_{N-1}),$$

то и величина $I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^*$ будет небольшой.

Заметим, что поскольку на практике конструктивное описание множеств X_i, $D_i(x)$ часто отсутствует, то в правой части оценки (28) вместо X_i, $D_i(x)$ часто берут G_i и V_i соответственно — такая замена, очевидно, может привести лишь к увеличению правой части (28). В результате получим достаточно удобную апостериорную оценку погрешности

$$0 \leqslant I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^* \leqslant \sum_{i=0}^{N-1} \left[S_i(\bar{x}_i, \bar{u}_i) - \inf_{x \in G_i} \inf_{u \in V_i} S(x, u) \right] + \\ + s_N(\bar{x}_N) - \inf_{G_N} s_N(x) + K_0(\bar{x}_0) - \inf_{G_0} K_0(x). \quad (30)$$

Конечно, при пользовании оценкой (30) надо помнить, что если правые части оценок (28), (30) отличаются намного, то оценка (30) может оказаться слишком грубой.

6. Оценка (28) полезна также и тем, что она указывает пути получения достаточных условий оптимальности для задачи (5)—(8).

Теорема 5. *Для того чтобы управление* $[\bar{u}_i]_0$ *и траектория* $[\bar{x}_i]_0$, *удовлетворяющие условиям* (6)—(8), *были решением задачи* (5)—(8), *достаточно, чтобы существовала функция* $K_i(x)$ $(i = 0, 1, \ldots, N)$ *такая, что*

$$S_i(\bar{x}_i, \bar{u}_i) = \inf_{x \in X_i} \inf_{u \in D_i(x)} S_i(x, u) = S_{i\min}, \quad i = 0, 1, \ldots, N-1, \quad (31)$$

$$s_N(\bar{x}_N) = \inf_{X_N} s_N(x) = s_{N\min}, \quad K_0(\bar{x}_0) = \inf_{X_0} K_0(x) = K_{0\min}, \quad (32)$$

где функции $S_i(x, u)$, $s_N(x)$ *определяются формулами* (24), (25).

Доказательство следует из того, что при выполнении условий (31), (32) правая часть оценки (28) обращается в нуль, и $I_0(\bar{x}_0, [\bar{u}_i]_0) = I_0^*$.

С помощью оценки (28) нетрудно также получить условия, достаточные для того, чтобы та или иная последовательность допустимых пар управлений и траекторий была минимизирующей для задачи (5)—(8).

Теорема 6. *Пусть последовательность управлений* $[u_{im}]_0$ *и траекторий* $[x_{im}]_0$ $(m = 1, 2, \ldots)$ *удовлетворяет условиям* (6)—(8). *Для того чтобы*

$$\lim_{m \to \infty} I_0(x_{0m}, [u_{im}]_0) = I_0^*, \quad (33)$$

достаточно существования функции $K_i(x)$ $(i = 0, 1, \ldots, N)$

такой, что

$$\lim_{m\to\infty} S_i(\bar{x}_{im}, \bar{u}_{im}) = S_{i\min}, \quad i = 0, 1, \ldots, N-1, \tag{34}$$

$$\lim_{m\to\infty} s_N(x_{Nm}) = s_{N\min}, \quad \lim_{m\to\infty} K_0(x_{0m}) = K_{0\min}, \tag{35}$$

где $S_{i\min}$, $s_{N\min}$, $K_{0\min}$ *определены в* (31), (32).

Доказательство. В оценку (28) вместо $[\bar{u}_i]_0$, $[\bar{x}_i]_0$ поставим соответственно $[u_{im}]_0$, $[x_{im}]_0$ и перейдем к пределу при $m\to\infty$. С учетом условий (34), (35) получим равенство (33).

Утверждения, аналогичные теоремам 5, 6, доказаны в [124, 188, 189] для более общих задач, чем задача (5)—(8).

Всякую функцию $K_i(x)$ $(i=0, 1, \ldots, N)$, удовлетворяющую условиям теоремы 5 (теоремы 6), назовем *функцией Кротова* задачи (5)—(8), соответствующей допустимой паре $[u_i]_0$, $[x_i]_0$ [или последовательности допустимых пар $[u_{im}]_0$, $[x_{im}]_0$, $m=1, 2, \ldots$].

Заметим, что если существует хотя бы одна функция Кротова $K_i(x)$ $(i=0, 1, \ldots, N)$, то функция $K_i(x)+\alpha_i$ $(i=0, 1, \ldots, N)$ при любых α_i также является функцией Кротова. Поэтому без ограничения общности в теоремах 5, 6 можно принять $S_{i\min}=0$ $(i=0, 1, \ldots, N-1)$, $s_{N\min}=0$, ибо в противном случае функцию $K_i(x)$ заменим новой функцией $K_i(x)+\alpha_i$, где

$$\alpha_i = S_{i\min} + S_{i+1,\min} + \ldots + S_{N-1,\min}, \quad i = 1, \ldots, N-1, \quad \alpha_N = s_{N\min}.$$

Таким образом, функция Кротова для допустимой пары $([\bar{u}_i]_0, [\bar{x}_i]_0)$ или последовательности $([u_{im}]_0, [x_{im}]_0)$ $(m=1, 2, \ldots)$ согласно теоремам 5, 6 удовлетворяет условиям

$$S_i(x, u) \equiv K_{i+1}(F_i(x, u)) - K_i(x) + \\ + F_i^0(x, u) \geqslant 0, \quad u \in D_i(x), \quad x \in X_i, \quad i = 0, 1, \ldots, N-1, \tag{36}$$

$$s_N(x) = K_N(x) + \Phi(x) \geqslant 0, \quad x \in X_N = G_N, \\ K_0(x) \geqslant K_{0\min}, \quad x \in X_0, \tag{37}$$

причем неравенства (36), (37) должны обратиться в равенства при $u=\bar{u}_i$, $x=\bar{x}_i$ или при $u=u_{im}$, $x=x_{im}$ в пределе при $m\to\infty$ $(i=0, 1, \ldots, N)$.

Сравнение соотношений (36), (37) с (13), (14), (16), (20) показывает, что функция Беллмана всегда является функцией Кротова, и обратное, вообще говоря, неверно. Заметим также, что с помощью функции Кротова удается установить оптимальность допустимых пар, не решая проблемы синтеза, так как согласно условиям (36), (37) функция $K_i(x)$ выбирается с учетом индивидуальных свойств конкретной допустимой пары управления и траектории (или последовательности пар), подозрительных на оптимальность.

7. Остановимся на одном специальном классе задач минимизации функций большого числа переменных, которые с помощью

метода динамического программирования могут быть сведены к последовательности задач минимизации функций меньшего числа переменных. А именно, пусть требуется минимизировать функцию

$$I_0([u_i]_0) = I_0(u_0, u_1, \ldots, u_{N-1}) = \sum_{i=0}^{N-1} f_i^0(u_i) \tag{38}$$

при условиях

$$u_i \in V_i, \quad i = 0, 1, \ldots, N-1, \tag{39}$$

$$\sum_{i=0}^{N-1} g_i^j(u_i) \leqslant b^j, \quad j = 1, \ldots, p, \qquad \sum_{i=0}^{N-1} g_i^j(u_i) = b^j, \quad j = p+1, \ldots, n, \tag{40}$$

где V_i — заданные множества из E^r; $f_i^0(u)$, $g_i^j(u)$, $u \in V_i$ — заданные функции, b^j — заданные числа.

Задачу (38)—(40), оказывается, нетрудно записать в виде задачи (5)—(8). В самом деле, введем переменные x_i ($i = 0, 1, \ldots, N$) как решение системы

$$x_{k+1} = x_k + g_k(u_k), \quad k = 0, 1, \ldots, N-1, \quad x_0 = 0, \tag{41}$$

где $g_k(u) = (g_k^1(u), \ldots, g_k^n(u))$, $x_k = (x_k^1, \ldots, x_k^n)$ $(k = 0, 1, \ldots, N)$.

Так как из (41) следует, что $x_N = \sum_{k=0}^{N-1} g_k(u_k)$, то ясно, что ограничения (40) равносильны условию

$$x_N \in G_N = \{x\colon\ x \in E^n,\ x^j \leqslant b^j,\ j = 1, \ldots, p;\ x^j = b^j,\ j = p+1, \ldots, n\}. \tag{42}$$

Таким образом, задача (38)—(40) эквивалентна задаче минимизации функции (38) при условиях (39), (41), (42) и является частным случаем задачи (5)—(8) при $F_i^0(x, u) = f_i^0(u)$, $F_i(x, u) = x + g_i(u)$, $\Phi(x) \equiv 0$, $G_i = E^n$ $(i = 1, \ldots, N-1)$; $G_0 = \{0\}$, G_N определено соотношением (42). Это значит, что для исследования задачи (38)—(40) может быть применен метод динамического программирования, изложенный выше. Пользуясь введенными ранее обозначениями, можем переписать уравнение Беллмана (13), (14) применительно к задаче (38)—(40):

$$B_k(x) = \inf_{u \in D_k(x)} \left[f_k^0(u) + B_{k+1}(x + g_i(u))\right], \tag{43}$$

$$x \in X_k, \quad k = 0, 1, \ldots, N-1, \quad B_N(x) = 0.$$

В том случае, когда ограничения (40) и, следовательно, (42) отсутствуют, то $G_N = E^n$ и в (43) можно положить $D_k(x) = V_k$ $(k = 0, 1, \ldots, N-1)$. Уравнениями (43) можно пользоваться для решения задачи (38)—(40), как это показано выше в п. 3.

Предлагаем читателю в качестве упражнения переформулировать теоремы 1—6 применительно к задаче (38)—(40).

Подчеркнем, что в задаче (38)—(40) функция и ограничения имеют весьма специальный вид — это обстоятельство было весьма существенно для применения метода динамического программирования. Этот метод применим и к несколько более общим, чем (38)—(40), задачам — об этом см. подробнее, например, в [101].

8. Изложенный выше метод динамического программирования является достаточно эффективным средством решения задач вида (5)—(8) или (38)—(40) — с его помощью исходная задача сводится к последовательности вспомогательных и, вообще говоря, более простых задач минимизации функций меньшего числа переменных для определения $B_k(x)$, $u_k(x)$ (см. условия (13), (14) или (43)). Если эти вспомогательные задачи решены с достаточно хорошей точностью, то тем самым и в исходной задаче глобальный минимум функции будет найден с высокой точностью. Далее, метод динамического программирования позволяет решить важную в приложениях проблему синтеза. Как показано в п. 6, с помощью этого метода и его обобщений могут быть получены достаточные условия оптимальности для дискретных управляемых систем. Кроме того, этот метод дает значительный выигрыш в объеме вычислений по сравнению с простым перебором всевозможных допустимых управлений и траекторий, поскольку при определении $B_k(x)$, $u_k(x)$ рассматриваются лишь такие управления, которые переводят точку $x \in G_k$ в точку $x_{k+1} = F(x, u) \in G_{k+1}$, а дальнейшее движение из точки x_{k+1} осуществляется по оптимальной траектории, при этом неоптимальные траектории вовсе не рассматриваются. Указанные достоинства метода динамического программирования, простота схемы, применимость к задачам оптимального управления с фазовыми ограничениями делают этот метод весьма привлекательным, и его широко используют при решении задач типа (5)—(8) или (38)—(40). Что касается задачи (1)—(4), с которой мы начали изложение, то можно показать, что при некоторых ограничениях решение дискретной задачи (5)—(8) при $\lim\limits_{N\to\infty} \max\limits_{0\leqslant i\leqslant N-1} (t_{i+1} - t_i) = 0$ будет приближаться в некотором смысле к решению задачи (1)—(4).

Заметим, что поскольку аналитическое выражение для $B_k(x)$, $u_k(x)$ при всех $x \in X_k$ в общем случае найти не удается, то на практике приходится ограничиваться приближенным вычислением $B_k(x)$, $u_k(x)$ в некоторых заранее выбранных узловых точках множества X_k. Однако согласно (13) при вычислении $B_k(x)$ нужно знать значение $B_{k+1}(F_k(x, u))$ при некоторых u, и здесь вполне возможны случаи, когда точка $x_{k+1} = F_k(x, u)$ не будет принадлежать заранее выбранному множеству узловых точек из

X_{k+1} и нужное значение $B_{k+1}(x_{k+1})$ еще не будет вычислено. Если же мы захотим вычислить недостающее значение $B_{k+1}(x_{k+1})$, то здесь могут понадобиться значения ранее вычисленных функций $B_{k+2}(x)$, ..., $B_N(x)$ в новых дополнительных точках, а для этого в свою очередь придется еще более расширить множества узловых точек в X_{k+2}, X_{k+3}, ... и т. д. На практике в таких случаях недостающее значение $B_{k+1}(x)$ получают с помощью интерполяции по значениям $B_{k+1}(x)$ в близлежащих узловых точках, что, вообще говоря, снижает точность. Заметим также, что принятый выше способ аппроксимации задачи (1)—(4) с помощью разностной задачи (5)—(8) довольно груб, поскольку опирается на простейший метод ломаных Эйлера для интегрирования дифференциальных уравнений и квадратурную формулу прямоугольников. В следующем параграфе будет описана схема Моисеева, которая не требует интерполяции и оставляет достаточную свободу при выборе способа аппроксимации задачи (1)—(4).

В заключение отметим, что метод динамического программирования относится к классу методов декомпозиции — так называются методы минимизации, позволяющие задачи большой размерности свести к задачам меньшей размерности; о методах декомпозиции см., например, в [111, 117, 194, 203, 242, 302, 320, 330].

Упражнения. 1. Найти функцию Беллмана для задачи

$$I_0([u_i]_0) = \sum_{i=0}^{N-1} [\langle a_i, x_i\rangle + b_i(u_i)] + \langle c, x_N\rangle \to \inf,$$

$$x_{i+1} = A_i x_i + B_i(u_i), \quad u_i \in V_i, \quad i = 0, 1, \ldots, N-1, \; x_0 = a,$$

где A_i — матрица порядка $n \times n$; $B_i(u) = \left(B_i^1(u), \ldots, B_i^n(u)\right)$, $B_i^j(u)$, $b_i(u)$ — функции переменной $u \in V_i \subseteq E^r$, a_i, c, a — n-мерные векторы, $i = 0, 1 \ldots$ $\ldots, N-1$. Указание: функцию Беллмана искать в виде $B_k(x) = \langle \psi_k, x\rangle$ $(k = 0, 1, \ldots, N)$.

2. Найти функцию Беллмана для задачи: $I_0(u) = \Phi(x_1) \to \inf$: $x_1 = x_0 + u$, $x_i \in G_i$ $(i = 0, 1)$; $u \in V_0$, где $\Phi(x) = (1 + e^{-1/x})^{-1}$ при $x \neq 0$, $\Phi(0) = 1/2$; $G_0 = \{x \in E^1\colon |x| \leqslant 1/2\}$, $G_1 = E^1$; $V_0 = \{u \in E^1\colon |u| \leqslant 1\}$. Показать, что $X_0 = G_0$, $X_1 = E^1$, и убедиться, что для этой задачи нижняя грань в (13), (16) не достигается.

3. Пусть функция $B_k(x)$ $(x \in X_k,\ k = 0, 1, \ldots, N)$ удовлетворяет условиям (13), (14), а функция $u_{km}(x) \in D_k(x)$ и точки $x_{0m} \in X_0$ $(m = 1, 2, \ldots)$ таковы, что $\lim\limits_{m\to\infty} (x, u_{im}(x)) = 0$, $\lim\limits_{m\to\infty} B_0(x_{0m}) = \inf\limits_{X_0} B_0(x)$. Пусть управление $[u_{im}]_0$ и траектория $[x_{im}]_0$ построены по правилу: $u_{0m} = u_{0m}(x_{0m})$, $x_{1m} = F_0(x_{0m}, u_{0m})$, $u_{1m} = u_{1m}(x_{1m})$, ..., $x_{Nm} = F_{N-1}(x_{N-1,m}, u_{N-1,m})$. Тогда последовательность пар $(x_{0m}, [u_{im}]_0)$ — минимизирующая для задачи (5) — (8), т. е. $\lim\limits_{m\to\infty} I_0(x_{0m}, [u_{im}]_0) = I_0^*$. Доказать.

4. Доказать, что последовательность функций $u_{im}(x)$ $(i = 0, 1, \ldots, N-1$, $m = 1, 2, \ldots)$ из упражнения 3 дает приближенное решение проблемы синтеза для задачи (5) — (8), т. е. если в момент k система находится в точке $x_k = x \in X_k$, то движение по закону $x_{i+1,m} = F_i(x_{im}, u_{im}(x_{im}))$ $(i = k, \ldots$

$\ldots, N-1$); $x_{km} = x$ $(m = 1, 2, \ldots)$, при больших m доставляет функции $I_k(x, [u_i]_k)$ значения, близкие к I_k^*.

5. Для задачи (9) — (12) получить оценку погрешности, аналогичную оценке (28).

6. Вывести уравнения Беллмана и доказать теоремы, аналогичные теоремам 1—4 для задачи:

$$I_0(x, [u_i]_0) = \sum_{i=0}^{N} F_i^0(x_i, u_i) + \Phi_0(x_0) + \Phi_1(x_N) \to \inf,$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = 0, 1, \ldots, N-1, \quad x_i \in G_i, \quad u_i \in V_i, \ i = 0, 1, \ldots, N,$$

где $x_i = (x_i^1, \ldots, x_i^n)$, $F_i = (F_i^1, \ldots, F_i^n)$, $u_i = (u_i^1, \ldots, u_i^r)$; множества $G_i \subseteq E^n$, $V_i \subseteq E^r$ $(i = 0, 1, \ldots, N)$ заданы; функции $F_i^j(x, u)$ определены при $x \in G_i$, $u \in V_i$ $(i = 0, 1, \ldots, N, j = 0, 1, \ldots, n)$; функции $\Phi_0(x)$, $\Phi_1(x)$ определены при $x \in G_0$, $x \in G_N$ соответственно; i — дискретное время; момент N считается заданным.

7. Обобщить оценку (28) и теоремы 5, 6 для задачи из упражнения 6.

8. Пусть в задаче (5) — (8) $F_i^0(x, u) \equiv 0$ $(i = 0, 1, \ldots, N-1)$. Показать, что тогда $B_k(x) = \Phi(x)$ для всех $k = 0, 1, \ldots, N$, причем функции $B_k(x)$ при различных k отличаются друг от друга областями своего определения X_k.

9. Применить метод динамического программирования к задаче минимизации функции

$$I_0(u_0, u_1, \ldots, u_N) = \sum_{i=0}^{N-1} f_i(u_i, u_{i+1}) + f_N(u_N)$$

при условиях $u_i \in V_i$ $(i = 0, 1, \ldots, N)$. Указание: ввести функцию

$$B_k(u) = \inf\left(\sum_{i=k}^{N-1} f_i(u_i, u_{i+1}) + f_N(u_N)\right),$$

где нижняя грань берется по всем наборам $(u_k = u, u_{k+1}, \ldots, u_N)$, $u_i \in V_i$ $(i = k, \ldots, N)$ и показать, что $B_k(u) = \inf\limits_{v \in V_{k+1}} [f_k(u, v) + B_{k+1}(v)]$ $(k = 0, 1, \ldots, N-1)$; $B_N(u) = f_N(u)$.

§ 2. Схема Моисеева

1. По-прежнему будем рассматривать задачу (1.1) — (1.4). Для приближенного решения этой задачи, как и раньше, разобьем отрезок $t_0 \leqslant t \leqslant T$ на N частей точками $t_0 < t_1 < \ldots < t_{N-1} < t_N = T$. На множестве $G_i \equiv G(t_i)$ возьмем некоторую дискретную сетку точек $x_{ij} \in G_i$; следуя [14, 217] множество всех точек выбранной сетки будем называть *шкалой состояний* и обозначать через H_i $(i = 0, 1, \ldots, N)$. Шкалы состояний H_i и H_{i+1} будем называть *соседними*. На двух соседних шкалах H_i и H_{i+1} возьмем точки $x \in H_i$ и $y \in H_{i+1}$ и рассмотрим следующую

вспомогательную задачу:

$$J_i(x, y, u(\cdot)) = \int_{t_i}^{t_{i+1}} f^0(x(t), u(t), t)\, dt \to \inf, \tag{1}$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_i \leqslant t \leqslant t_{i+1}; \quad x(t_i) = x, \quad x(t_{i+1}) = y, \tag{2}$$

$$x(t) \in G(t), \quad t_i \leqslant t \leqslant t_{i+1}, \tag{3}$$

$u = u(\cdot)$ кусочно-непрерывна и $u(t) \in V(t)$ при $t_i \leqslant t \leqslant t_{i+1}$. (4)

Следуя [12, 217], задачу (1)—(4) будем называть элементарной операцией, соединяющей точки x и y. Через $\Delta_i(x, y)$ обозначим множество всех управлений $u = u(\cdot)$ из (4), для которых соответствующая траектория $x(\cdot)$ удовлетворяет всем условиям (2), (3). Положим $M_i(x, y) = \inf\limits_{u \in \Delta_i(x,y)} J_i(x, y, u)$; если $\Delta_i(x, y) = \varnothing$, то $M_i(x, y) = \infty$ по определению.

Пусть все точки всех соседних шкал попарно соединены элементарными операциями. Если $\inf\limits_{\Delta_i(x,y)} J_i(x, y, u)$ достигается на некотором управлении $u_i(\cdot) \in \Delta_i(x, y)$ и соответствующей траектории $x_i(\cdot)$, то величина

$$\sum_{i=0}^{N-1} M_i(x_{ij_i}, x_{i+1, j_{i+1}}) + \Phi(x_{Nj_N}) \tag{5}$$

выражает собой значение исходной функции (1.1) на управлении $u = u(t) = u_i(t)$ $(t_i \leqslant t \leqslant t_{i+1}, \ i = 0, 1, \ldots, N-1)$ и соответствующей траектории $x(t, u) = x_i(t)$ $(t_i \leqslant t \leqslant t_{i+1}, \ i = 0, 1, \ldots, N-1)$; $x(t_0) = x_{0j_0}$ при соблюдении всех ограничений (1.2)—(1.4). Поэтому исходную задачу (1.1)—(1.4) естественно аппроксимировать задачей отыскания минимума суммы (5) по всевозможным наборам точек $(x_{0j_0}, x_{1j_1}, \ldots, x_{Nj_N})$, $x_{ij_i} \in H_i$ $(i = 0, 1, \ldots, N)$.

Такая аппроксимация задачи (1.1)—(1.4) имеет смысл, конечно, и в том случае, когда $\inf\limits_{\Delta_i(x,y)} J_i(x, y, u)$ не достигается при каких-либо x, y, i. Очевидно, приведенная аппроксимация задачи (1.1)—(1.4) более гибкая и лучше приспособлена для приближенного решения этой задачи, чем схема из § 1, ибо здесь представляется свобода в выборе способа разрешения элементарной операции (1)—(4), и кроме того, рассмотрение лишь таких траекторий, концы которых лежат в известных точках соседних шкал, избавляет нас от необходимости интерполирования. На способах разрешения элементарной операции мы остановимся ниже.

Описанная аппроксимация задачи (1.1)—(1.4) имеет простой геометрический смысл. А именно, траекторию $x_i(t)$ из (2),

на которой достигается $\inf\limits_{\Delta_i(x,y)} J_i(x, y, u)$, назовем дугой, соединяющей точки x, y, а числа $M_i(x, y)$ — длиной этой дуги, $i = 0, 1, \ldots, N-1$. Дуги, последовательно соединяющие пары точек x_{ij_i}, $x_{i+1,j_{i+1}}$ соседних шкал H_i, H_{i+1} $(i = 0, 1, \ldots, N-1)$ назовем путем, соединяющим шкалы H_0 и H_N и проходящим через точки $x_{0i_0}, x_{1j_1}, \ldots, x_{Nj_N}$, и в качестве его длины примем величину (5). Тогда наша задача сведется к отысканию кратчайшего пути, соединяющего шкалы H_0 и H_N.

2. Обозначим

$$G_k(x) = \inf\left\{\sum_{i=k}^{N-1} M_i(x_{ij_i}, x_{i+1,j_{i+1}}) + \Phi(x_{Nj_N})\right\},$$

где нижняя грань берется по всем наборам точек $(x_{kj_k} = x, x_{k+1,j_{k+1}}, \ldots, x_{Nj_N})$, $x_{ij_i} \in H_i$ $(i = k, \ldots, N)$. Иначе говоря, $C_k(x)$ выражает собой кратчайшее расстояние между фиксированной точкой $x \in H_k$ и шкалой H_N. Покажем, что функции $C_k(x)$ удовлетворяют следующим рекуррентным соотношениям:

$$C_k(x) = \inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\}, \quad k = 0, 1, \ldots, N-1;$$
$$C_N(x) \equiv \Phi(x), \quad (6)$$

аналогичным условиям (1.13), (1.14). Справедливость (6) при $k = N-1$ следует из определения $C_{N-1}(x)$, $C_N(x)$. Докажем (6) при других k $(0 \leqslant k \leqslant N-1)$. Для этого сначала убедимся в том, что

$$C_k(x) \leqslant \inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\}, \quad x \in H_k. \quad (7)$$

Возьмем произвольное $y \in H_{k+1}$. По определению $C_{k+1}(y)$ для любого $\varepsilon > 0$ найдется путь, соединяющий точку y со шкалой H_N, длина которого не превышает $C_{k+1}(y) + \varepsilon$. Если этот путь «удлинить», добавив к нему дугу, соединяющую точки x и y, то получим путь, соединяющий точку $x \in H_k$ со шкалой H_N, длина которого не превышает $M_k(x, y) + C_{k+1}(y) + \varepsilon$. Поэтому заведомо $C_k(x) \leqslant M_k(x, y) + C_{k+1}(y) + \varepsilon$. В силу произвольности точки $y \in H_{k+1}$ и величины ε отсюда имеем неравенство (7).

Теперь покажем, что в (7) на самом деле знак неравенства можно заменить знаком равенства. По определению $C_k(x)$ для любого $\varepsilon > 0$ найдется путь, соединяющий точку $x \in H_k$ со шкалой H_N, длина которого не превосходит $C_k(x) + \varepsilon$. Пусть этот путь проходит через точку $y_\varepsilon \in H_{k+1}$. Ясно, что отрезок этого пути от y_ε до H_N не меньше $C_{k+1}(y_\varepsilon)$, и поэтому весь путь от x до H_N не меньше $M_k(x, y_\varepsilon) + C_{k+1}(y_\varepsilon)$. Следовательно, $M_k(x, y_\varepsilon) + C_{k+1}(y_\varepsilon) \leqslant C_k(x) + \varepsilon$. Так как $y_\varepsilon \in H_{k+1}$, то отсюда име-

ем $\inf\limits_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\} \leqslant C_k(x) + \varepsilon_2$ или в силу произвольности ε: $\inf\limits_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\} \leqslant C_k(x)$. Сравнивая это неравенство с (7), немедленно получаем требуемые соотношения (6).

3. Соотношения (6) могут быть использованы так же, как и уравнение Беллмана (1.13), (1.14). Опишем порядок работы с этими соотношениями в предположении, что каждая из шкал состояний H_i состоит из конечного числа точек p_i $(i = 0, 1, \ldots, N)$. Заметим, что в этом случае в (6) вместо inf можно писать min. Функция $C_N(x) \equiv \Phi(x)$, $x \in H_N$ нам известна. Для вычисления $C_{N-1}(x)$ с помощью элементарных операций соединим попарно все точки шкал H_{N-1} и H_N. Сравнивая $p_{N-1}p_N$ величин $M_{N-1}(x, y) + \Phi(y)$ при всевозможных $x \in H_{N-1}$, $y \in H_N$, найдем $\min\limits_{y \in H_N} [M_{N-1}(x, y) + \Phi(y)] = C_{N-1}(x)$, а также точку $y = y_{N-1}(x) \in H_N$, $x \in H_{N-1}$, на которой этот минимум достигается. Если $C_{k+1}(x)$ и $y = y_{k+1}(x) \in H_{k+2}$, $x \in H_{k+1}$ уже известны, то для вычисления $C_k(x)$, $x \in H_k$, соединим элементарными операциями всевозможные пары точек шкал H_k и H_{k+1}. Перебором $p_k p_{k+1}$ величин $M_k(x, y) + C_{k+1}(y)$ при всех $x \in H_k$, $y \in H_{k+1}$ найдем $\min\limits_{y \in H_{k+1}} [M_k(x, y) + C_{k+1}(y)]$, а также точку $y = y_k(x) \in H_{k+1}$, $x \in H_k$, на которой этот минимум достигается, и т. д. для всех $k = N, N-1, \ldots, 1, 0$. На вычисление всех $C_k(x)$ в $y = y_k(x)$, $x \in H_k$ $(k = 0, 1, \ldots, N-1)$, понадобится перебор $\sum\limits_{i=0}^{N-1} p_i p_{i+1}$ величин. Наконец, находим $x_0^* \in H_0$ из условия $C_0(x_0^*) = \inf\limits_{x \in H_0} C_0(x)$ — для этого нужно перебрать еще p_0 величин, и определяем последовательно точки $x_1^* = y_0(x_0^*)$, $x_2^* = y_1(x_1^*), \ldots, x_N^* = y_{N-1}(x_{N-1}^*)$. Путь, проходящий через найденные точки $x_0^*, x_1^*, \ldots, x_N^*$, будет кратчайшим среди всех путей, соединяющих крайние шкалы H_0, H_N. В самом деле, по определению $y_k(x)$ $(k = 0, 1, \ldots, N-1)$ и $x_k^* \in H_k$ $(k = 0, 1, \ldots, N)$ имеем

$$C_k(x_k^*) = M_k(x_k^*, x_{k+1}^*) + C_{k+1}(x_{k+1}^*), \qquad k = 0, 1, \ldots, N-1. \quad (8)$$

Возьмем произвольный путь, соединяющий шкалы H_0 и H_N и проходящий через точки $x_0, x_1, \ldots, x_N$, $x_i \in H_i$ $(i = 0, 1, \ldots, N)$. Так как $x_{k+1} \in H_{k+1}$, то согласно (6) будем иметь $C_k(x_k) \leqslant M_k(x_k, x_{k+1}) + C_{k+1}(x_{k+1})$. Отсюда и из (8) тогда следует $M_k(x_k^*, x_{k+1}^*) + C_{k+1}(x_{k+1}^*) - C_k(x_k^*) = 0 \leqslant M_k(x_k, x_{k+1}) + C_{k+1}(x_{k+1}) - C_k(x_k)$ $(k = 0, 1, \ldots, N-1)$. Просуммируем это

неравенство по k от нуля до $N-1$. Получим

$$\sum_{k=0}^{N-1} M_k(x_k^*, x_{k+1}^*) + C_N(x_N^*) - C_0(x_0^*) \leqslant$$

$$\leqslant \sum_{k=0}^{N-1} M_k(x_k, x_{k+1}) + C_N(x_N) - C_0(x_0).$$

Но $C_0(x_0^*) = \inf\limits_{x \in H_0} C_0(x) \leqslant C_0(x_0)$, поэтому $\sum\limits_{k=0}^{N-1} M_k(x_k^*, x_{k+1}^*) + \Phi(x_N^*) \leqslant \sum\limits_{k=0}^{N-1} M_k(x_k, x_{k+1}) + \Phi(x_N)$ для любых путей, соединяющих шкалы H_0 и H_N. Таким образом, путь, проходящий через точки $x_0^*, x_1^*, \ldots, x_N^*$ в самом деле кратчайший. Если $u_i^*(t)$ и $x_i^*(t)$ $(t_i \leqslant t \leqslant t_{i+1})$ представляют собой то управление и соответствующую траекторию, на которых приближенно реализуется элементарная операция (1)—(4) при $x = x_i^*$, $y = x_{i+1}^*$, то в качестве приближенного решения исходной задачи (1.1)—(1.4) можно взять управление $u^*(t) = u_i^*(t)$ и траекторию $x^*(t) = x_i^*(t)$ $(t_i \leqslant t \leqslant t_{i+1},\ i = 0, 1, \ldots, N-1)$.

Аналогично доказывается, что путь, проходящий через точки $x_k^* = x \in H_k$, $x_{k+1}^* = y_k(x_k^*) \in H_{k+1}, \ldots, x_N^* = y_{N-1}(x_{N-1}^*) \in H_N$, является кратчайшим между точкой $x \in H_k$ и шкалой H_N. Это означает, что функция $y_k(x)$ дает нам приближенное решение проблемы синтеза для задачи (1.1)—(1.4).

Заметим, что определение кратчайшего пути между шкалами H_0 и H_N по описанной схеме потребовало перебора $\sum\limits_{i=0}^{N-1} p_i p_{i+1} + p_0$ величин, в то время как полный перебор всех путей, как нетрудно видеть, потребовал бы сравнения $p_0 p_1 \ldots p_N$ величин. Таким образом, уже при не слишком больших p_i перебор с помощью соотношений (6) по сравнению с полным перебором дает существенную экономию памяти ЭВМ и машинного времени. Такая экономия достигается за счет того, что при вычислении $C_k(x)$ из (6) рассматриваются лишь пути, проходящие через всевозможные точки $x \in H_k$ и $y \in H_{k+1}$ и соединяющие точки $y \in H_{k+1}$ со шкалой H_N кратчайшим образом, и тем самым все некратчайшие пути, соединяющие $y \in H_{k+1}$ со шкалой H_N, из рассмотрения полностью исключаются.

4. Для получения более точного решения задачи (1.1)—(1.4) необходимо взять более густую сетку точек на шкалах и увеличить число шкал. Однако при этом число перебираемых величин даже при использовании описанной выше схемы перебора катастрофически быстро растет, и уже при небольших размер-

ностях векторов x, u становится невозможным решить задачу о кратчайшем пути за разумное время с помощью самых лучших современных ЭВМ. В этом случае часто используют прием, известный под названием *метода блуждающих трубок* [326]. Суть этого приема заключается в следующем.

Сначала берут небольшое число шкал с небольшим количеством точек на них, и по описанной выше схеме поиска находят кратчайший путь l_1, соединяющий крайние шкалы H_0 и H_N. Затем уменьшают шаг сетки на каждой шкале, путь l_1 окружают некоторой «трубкой» из путей, проходящих вблизи l_1 по точкам новой сетки. Для построения трубки вокруг l_1 на каждой шкале обычно берут небольшие окрестности точки, через которую проходит l_1, и рассматривают пути, проходящие через выбранные точки. С помощью описанной выше схемы перебора находят кратчайший путь в полученной трубке; все пути вне этой трубки в переборе пока не участвуют. Таким образом, получают новый улучшенный путь l_2, длина которого не превышает длину l_1. Далее, сохраняя прежние шкалы и точки на них, окружают путь l_2 новой трубкой и находят следующее приближение l_3 и т.д., продолжая процесс до тех пор, пока трубка не перестает «блуждать» и впервые не получится равенство $l_s = l_{s+1}$. После этого измельчают сетку на каждой шкале, окружают путь l_s новой трубкой и продолжают поиск кратчайшего пути описанным приемом блуждающих трубок. Процесс измельчения сетки на шкалах и поиска кратчайшего пути указанным способом повторяют, пока два кратчайших пути, полученные после двух последовательных измельчений сетки, не совпадают с удовлетворительной точностью. Затем увеличивают число шкал, т. е. сгущают сетку по времени, и повторяют процесс поиска методом блуждающих трубок с постепенным измельчением сетки на шкалах состояний до удовлетворительного совпадения двух последовательных приближений. Попеременно измельчая сетку на шкалах состояния и сетку по времени, поиск с помощью блуждающих трубок продолжают до получения приближенного решения исходной задачи с достаточной точностью. Изменение шагов сеток на шкалах состояния и по времени должно быть согласованным; например, в случае равномерных сеток эти шаги должны удовлетворять соотношению $|\Delta x| = o(\Delta t)$ [217].

Оказывается, метод блуждающих трубок во многих случаях существенно сокращает перебор. В то же время следует заметить, что метод блуждающих трубок позволяет определить, вообще говоря, лишь локально кратчайший путь между крайними шкалами при фиксированной сетке, поскольку на каждом шаге в переборе участвуют лишь пути, попавшие в трубку.

5. Другой подход к поиску кратчайшего пути между шкалами дает *метод локальных вариаций* [14, 217, 326]. Этот метод

предполагает, что какой-то путь l_1, соединяющий шкалы H_0 и H_N, уже известен. Для определения следующего более короткого пути последовательно просматриваются шкалы $H_0, H_1, \ldots$ $\ldots, H_N$. Допустим, что шкалы $H_0, \ldots, H_{i-1}$ уже просмотрены и получен путь $l_{1,i-1}$, соединяющий шкалу H_0 и H_N. Пусть $x_i(l_1)$ — точка пути l_1, лежащая на шкале H_i. Выберем на шкале H_i некоторое количество точек, расположенных близко к точке $x_i(l_1)$, и переберем пути, проходящие через эти выбранные точки шкалы H_i, а в остальном совпадающие с путем $l_{1,i-1}$. Если среди перебираемых путей найдется путь, имеющий меньшую длину, чем путь $l_{1,i-1}$, то его обозначаем через l_{1i} и просмотр шкалы H_i на этом заканчиваем. Если же длины всех перебираемых путей оказались не меньше длины $l_{1,i-1}$, то полагаем $l_{1i} = l_{1,i-1}$. После определения пути l_{1i} переходим к просмотру следующей шкалы H_{i+1} и т. д. Такой перебор всех шкал $H_0, H_1, \ldots$ $\ldots, H_N$ закончится получением пути $l_{1N} = l_2$. Если длина пути l_2 меньше длины l_1, то для пути l_2 повторяют описанный выше просмотр шкал $H_0, H_1, \ldots, H_N$ и находят следующий путь l_3 и т. д. Если же окажется, что $l_2 = l_1$, т. е. путь l_1 улучшить не удалось, то на шкалах берут густую сетку точек, а также при необходимости измельчают сетку по времени, и снова просматривают шкалы и т. д.

Метод локальных вариаций описан. Нетрудно видеть, что этот метод является аналогом метода покоординатного спуска. Поскольку здесь в переборе участвуют гораздо меньше путей, чем в методе, основанном на соотношениях (6), или методе блуждающих трубок, то ясно, что метод локальных вариаций гораздо экономичнее и позволяет увеличить размерность решаемых задач. С другой стороны, нетрудно привести примеры задач, когда этим методом не удается найти даже локально кратчайший путь между шкалами. На рис. 7.1 приведен такой пример — здесь $N = 3$, шкалы H_0, H_1, H_2, H_3 состоят соответственно из точек $\{A\}$, $\{B, C\}$, $\{D, E\}$, $\{F\}$; на дугах, соединяющих точки соседних шкал, указаны их длины. Кратчайшим путем, соединяющим шкалы H_0 и H_3, является путь $ACDF$. Если в качестве начального приближения l_1 взять путь $ABEF$, то улучшить его методом локальных вариаций не удается. Любопытно также заметить, что если за l_1 взять путь $ABDF$, то в зависимости от того, начнем ли просмотр со шкалы H_1 или H_2, придем к кратчайшему пути $ACDF$ или уже рассмотренному пути $ABEF$.

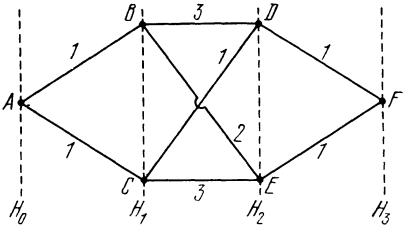

Рис. 7.1

Различные модификации метода локальных вариаций, примеры задач, к которым применялся этот метод, см. в [14, 217, 326].

6. Успех в применении описанных в этом параграфе методов приближенного решения задачи (1.1)—(1.4) во многом зависит от умения строить элементарные операции (1)—(4). По существу элементарная операция представляет собой экстремальную задачу той же трудности, что и исходная задача. Однако малость отрезка $t_i \leqslant t \leqslant t_{i+1}$ позволяет здесь сделать ряд упрощающих предположений. Прежде всего, если шаг сетки по времени достаточно мал, то элементарную операцию строят, исходя из условий (1), (2), (4), полагая, что дуги траекторий или не нарушают ограничений (3), или этим нарушением можно пренебречь. Далее, вместо минимизации функции (1) при условиях (2), (4) часто ограничиваются построением какого-либо допустимого управления и соответствующей траектории, удовлетворяющих условиям (2), (4), и в качестве длины дуги $M_i(x, y)$ тогда берут значение функции (1) на полученных управлении и траектории.

Во многих задачах полезно задаться каким-либо семейством управлений $u(t) \equiv u(t, c_1, \ldots, c_m)$, зависящих от параметров $c_1, \ldots, c_m$ $(m \geqslant n)$. Например, это могут быть алгебраические или тригонометрические многочлены с коэффициентами $c_1, \ldots, c_m$ или кусочно-постоянные функции со значениями $c_1, \ldots, c_m$ и т. п. Значения этих параметров затем можно определить из следующей системы n уравнений с m неизвестными $(m \geqslant n)$:

$$\int_{t_i}^{t_{i+1}} \dot{x}(t)\,dt = \int_{t_i}^{t_{i+1}} f(x(t), u(t, c_1, \ldots, c_m))\,dt = y - x, \tag{9}$$

$$u(t, c_1, \ldots, c_m) \in V(t_i), \quad t_i \leqslant t \leqslant t_{i+1},$$

используя для этого различные методы [4, 20, 39, 54, 209]. Если $m > n$, то параметры $c_1, \ldots, c_m$ отсюда будут определяться, вообще говоря, неоднозначно, и свободные параметры можно использовать для минимизации функции (1). Для упрощения системы (9) дифференциальное уравнение (2) часто заменяют более простыми уравнениями:

$$\dot{x}(t) = f(x, u(t), t) \text{ или } \dot{x}(t) = f\left(\frac{x+y}{2}, u(t), t\right)$$

или другими более точными разностными уравнениями; здесь возможно использование линеаризованной системы:

$$\dot{x}(t) = f(x, u(t), t) + \langle f_x(x, u(t), t), x(t) - x\rangle.$$

При решении задачи (1), (2), (4) часто применяется также

принцип максимума в сочетании с различными упрощающими приемами, описанными выше. Другие способы построения элементарных операций, примеры решенных конкретных задач см. в [14, 217].

§ 3. Проблема синтеза для систем с непрерывным временем

Продолжим исследование задачи (1.1)—(1.4) с непрерывно меняющимся временем. *Проблема синтеза* для задачи (1.1)—(1.4) заключается в построении функции $u = u(x, t)$, называемой синтезирующей функцией этой задачи и представляющей собой значение оптимального управления при условии, что в момент t система (1.2) находится в точке x, т. е. $x(t) = x$. Умение решать проблему синтеза крайне важно в различных прикладных задачах оптимального управления. В самом деле, если известна синтезирующая функция $u(x, t)$, то техническое осуществление оптимального хода процесса может быть произведено по следующей схеме, называемой схемой с обратной связью: с измерительного прибора, замеряющего в каждый момент t фазовое состояние $x(t)$, на ЭВМ или какое-либо другое вычислительное средство подается величина $x(t)$, вычисляется значение управления $u(t) = u(x(t), t)$, после чего найденное значение $u(t)$ оптимального управления передается на исполнительный механизм, непосредственно реализующий требуемое течение управляемого процесса.

1. Проблема синтеза для задачи (1.1)—(1.4) сводится к решению следующей задачи: определить управление $u_*(\tau) = u_*(\tau, x, t)$, доставляющее функции

$$J(t, x, u(\cdot)) = \int_t^T f^0(x(\tau), u(\tau), \tau)\, d\tau + \Phi(x(T)) \tag{1}$$

минимальное значение при условиях

$$\dot{x}(\tau) = f(x(\tau), u(\tau), \tau), \quad t \leqslant \tau \leqslant T; \quad x(t) = x, \tag{2}$$

$$x(\tau) \in G(\tau), \quad t \leqslant \tau \leqslant T, \tag{3}$$

$$u = u(\tau) \in V(\tau), \quad t \leqslant \tau \leqslant T; \quad u(\cdot) \text{ — кусочно-непрерывно}, \tag{4}$$

где x — произвольная точка множества $G(t)$, а t — произвольный фиксированный момент времени, $t_0 \leqslant t \leqslant T$. Заметим, что при $t = t_0$ задача (1)—(4) превращается в исходную задачу (1.1)—(1.4).

Обозначим через $\Delta(x, t)$ множество всех управлений $u(\tau)$ $(t \leqslant \tau \leqslant T)$, удовлетворяющих условиям (4) и $u(\tau) = u(\tau + 0) = \lim\limits_{s \to \tau + 0} u(s)$ $(t \leqslant \tau < T)$, и таких, что соответствующая траектория $x(\tau) = x(\tau, u)$ $(t \leqslant \tau \leqslant T)$, системы (2) определена на всем отрезке $[t, T]$ и удовлетворяет фазовому ограничению (3). Положим $X(t) = \{x\colon x \in G(t), \Delta(x, t) \neq \varnothing\}$ $(t_0 \leqslant t < T)$; $X(T) = G(T)$.

Пару $(u(\tau), x(\tau))$ $(t \leqslant \tau \leqslant T)$ назовем *допустимой парой* задачи (1)—(4), если $x(t) = x \in X(t)$, $u(\cdot) \in \Delta(x, t)$ и $x(\cdot)$ является решением системы (2), соответствующим рассматриваемому управлению $u(\cdot)$. Аналогично, пару $(u(\tau), x(\tau))$ $(t_0 \leqslant \tau \leqslant T)$ будем называть допустимой парой исходной задачи (1.1)—(1.4), если $x(t_0) = x_0 \in X(t_0)$, $u(\cdot) \in \Delta(x_0, t_0)$ и выполнены все соотношения (1.2)—(1.4). Допустимую пару $(u_*(\tau), x_*(\tau))$ $(t \leqslant \tau \leqslant T)$, задачи (1)—(4) будем называть решением этой задачи, если $J(t, x, u_*(\cdot)) = \inf\limits_{\Delta(x,t)} J(t, x, u(\cdot))$; при этом $u_*(\cdot)$ назовем оптимальным управлением, а $x_*(\cdot)$ — оптимальной траекторией задачи (1)—(4).

Зная оптимальное управление $u_*(\tau) = u_*(\tau, x, t)$ задачи (1)—(4) при всех тех (x, t), $x \in G(t)$ $(t_0 \leqslant t < T)$, при которых эта задача имеет ре-

шение, нетрудно получить синтезирующую функцию задачи (1.1)—(1.4): достаточно положить $u(x, t) \equiv u_*(t, x, t)$. Однако получить явное аналитическое выражение для оптимального управления $u_*(\tau, x, t)$ задачи (1)—(4) удается лишь в редких случаях. Поэтому желательно иметь другие подходы к решению проблемы синтеза.

Вспомним, что при решении проблемы синтеза для дискретных систем важную роль играло уравнение Беллмана (1.13), (1.14). Оказывается, аналогичное уравнение может быть получено и для задачи (1)—(4). Введем функцию

$$B(x, t) = \inf_{\Delta(x,t)} J(t, x, u(\cdot)),$$

называемую функцией Беллмана для задачи (1.1)—(1.4). Если задача (1.1)—(1.4) удовлетворяет некоторым ограничениям и функция $B(x, t)$ непрерывно дифференцируема, то можно показать, что функция Беллмана удовлетворяет следующим условиям, называемым уравнением Беллмана задачи (1.1)—(1.4):

$$\inf_{u \in D(x,t)} \left[\langle B_x(x, t), f(x, u, t)\rangle + B_t(x, t) + f^0(x, u, t)\right] = 0,$$
$$x \in X(t), \quad t_0 \leqslant t < T, \tag{5}$$

$$B(x, T) = \Phi(x), \quad x \in X(T) = G(T), \tag{6}$$

где $B_x = (B_{x^1}, \ldots, B_{x^n})$, B_{x^i}, B_t — частные производные функции $B(x, t)$, а $D(x, t)$ — множество всех тех $u \in V(t)$, для которых существует хотя бы одно управление $u(\cdot) \in \Delta(x, t)$ со значением $u(t) = u(t+0) = u$.

Приведем эвристические соображения, из которых следуют соотношения (5), (6) в следующей форме:

$$\inf_{u \in D_k(x)} [B_{k+1}(F_k(x, u))] - \left[B_k(x) + F_k^0(x, u)\right] = 0, \quad x \in X_k,$$
$$k = 0, 1, \ldots, N-1, \tag{7}$$

$$B_N(x) = \Phi(x), \quad x \in X_N = G_N. \tag{8}$$

Вспомним также обозначения, связывающие задачи (1.1)—(1.4) и (1.5)—(1.8) и принятые в (7), (8):

$$F_k^0(x, u) = (t_{k+1} - t_k) f^0(x, u, t_k), \quad F_k(x, u) = x + (t_{k+1} - t_k) f(x, u, t_k).$$

Исключим из этих обозначений и соотношений (7), (8) индекс k, приняв $t_k = t$, $\Delta t = t_{k+1} - t_k$, $t_N = T$, $B_k(x) = B(x, t)$, $B_{k+1}(y) = B(y, t + \Delta t)$, $D_k(x) = D(x, t)$, $X_k = X(t)$. Тогда соотношения (7), (8) могут быть переписаны в следующей безындексной форме:

$$\inf_{u \in D(x,t)} [B(x + \Delta t f(x, u, t), t + \Delta t) - B(x, t) + \Delta t f^0(x, u, t)] = 0,$$
$$x \in X(t), \quad t_0 \leqslant t \leqslant T,$$
$$B(x, T) = \Phi(x), \quad x \in X(T) = G(T).$$

Если теперь поделим первое из этих равенств на $\Delta t > 0$ и совершим формальный предельный переход при $\Delta t \to +0$, то придем к соотношениям (5), (6). Подчеркнем, что приведенные рассуждения никоим образом не претендуют на какую-либо строгость и могут служить лишь наводящими соображениями при получении соотношений (5), (6). Аналогичным образом можно было бы «вывести» эти соотношения, опираясь на уравнения (2.6).

Заметим, что уравнение (5) является дифференциальным уравнением в частных производных первого порядка, левая часть которого осложнена

взятием нижней грани, и вопросы существования и единственности решения задачи (5), (6), свойства ее решения в настоящее время исследованы слабо [327]. Задача (5), (6) здесь нас будет интересовать лишь с точки зрения решения проблемы синтеза.

Под решением задачи (5), (6) мы будем понимать функцию $B(x, t)$, которая определена и непрерывна при всех (x, t), $x \in X(t)$ $(t_0 \leqslant t \leqslant T)$ обладает кусочно-непрерывными частными производными B_x, B_t и удовлетворяет уравнению (5) всюду, где существуют эти производные, удовлетворяет условию (6) и, кроме того, для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)—(4) при всех $x \in X(t)$ $(t_0 \leqslant t < T)$ функция $B(x(\tau), \tau)$ переменной τ имеет кусочно-непрерывную производную (или $B(x(\tau), \tau)$ абсолютно непрерывна) на отрезке $[t, T]$.

Теорема 1. *Пусть $B(x, t)$ — решение задачи* (5), (6) *и, кроме того, пусть нижняя грань в левой части* (5) *достигается на кусочно непрерывной функции $u(x, t) \in D(x, t)$, $x \in X(t)$ $(t_0 \leqslant t \leqslant T)$. Тогда $u(x, t)$ — синтезирующая функция задачи* (1.1)—(1.4).

Доказательство. Возьмем произвольные t $(t_0 \leqslant t < T)$ и $x \in X(t)$. Пусть $x_*(\tau)$ $(t \leqslant \tau \leqslant T)$ является решением задачи Коши

$$\dot{x}(\tau) = f(x(\tau), u(x(\tau), \tau), \tau), \quad t \leqslant \tau < T; \quad x(t) = x,$$

и пусть $x_*(\tau) \in X(\tau)$ при всех $\tau \in [t, T]$. Положим $u_*(\tau) = u(x_*(\tau), \tau)$ $(t \leqslant \tau \leqslant T)$. Ясно, что $u_*(\cdot) \in \Delta(x, t)$ и $(u_*(\cdot), x_*(\cdot))$ — допустимая пара задачи (1)—(4). Для доказательства теоремы достаточно показать, что пара $(u_*(\cdot), x_*(\cdot))$ является решением задачи (1)—(4).

Сначала покажем, что для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)—(4) справедлива формула

$$J(t, x, u(\cdot)) = \int_t^T R(x(\tau), u(\tau), \tau)\, d\tau + B(x, t), \tag{9}$$

где

$$R(x, u, t) = \langle B_x(x, t), f(x, u, t)\rangle + B_t(x, t) + f^0(x, u, t). \tag{10}$$

В самом деле, по условию функция $B(x(\tau), \tau)$ переменной τ непрерывна и имеет кусочно-непрерывную производную. Тогда в силу уравнения (2) имеем

$$\frac{dB(x(\tau), \tau)}{d\tau} = R(x(\tau), u(\tau), \tau) - f^0(x(\tau), u(\tau), \tau)$$

всюду на $[t, T]$, за исключением, быть может, конечного числа точек. Интегрируя это тождество по τ на $[t, T]$, с учетом условия (6) получим

$$\Phi(x(T)) - B(x, t) = \int_t^T R(x(\tau), u(\tau), \tau)\, d\tau - \int_t^T f^0(x(\tau), u(\tau), \tau)\, d\tau,$$

что равносильно (9). Заметим, что формула (9) является аналогом формулы (1.21).

Уравнение (5) с помощью функции (10) можно переписать в виде $\inf\limits_{u \in D(x,\tau)} R(x, u, \tau) = 0$. Отсюда и из определения функции имеем

$$R(x, u(x, \tau), \tau) = 0 = \inf_{u \in D(x,\tau)} R(x, u, \tau) \leqslant R(x, u, \tau) \tag{11}$$

для всех $u \in D(x, \tau)$, $x \in X(\tau)$ $(t \leqslant \tau \leqslant T)$. Если $(u(\cdot), x(\cdot))$ — допустимая пара задачи (1)—(4), то $x(\tau) \in X(\tau)$, $u(\tau) = u(\tau + 0) = u \in D(x(\tau), \tau)$ $(t \leqslant \tau \leqslant T)$. Поэтому из (11) получаем

$$R(x_*(\tau), u(x_*(\tau)), \tau) = R(x_*(\tau), u_*(\tau), \tau) = 0 \leqslant R(x(\tau), u(\tau), \tau), \tag{12}$$

$t \leqslant \tau \leqslant T$, для любой допустимой пары задачи (1)—(4). Отсюда и из формулы (9) с учетом условия $x_*(t) = x(t) = x$ имеем

$$J(t, x, u(\cdot)) - J(t, x, u_*(\cdot)) = \int_t^T R(x(\tau), u(\tau), \tau) d\tau \geqslant 0 \tag{13}$$

для всех допустимых пар $(u(\cdot), x(\cdot))$ задачи (1)—(4).

Из (9), (12), (13) следует, что

$$J(t, x, u(\cdot)) = \inf_{\Delta(x,t)} J(t, x, u(\cdot)) = B(x, t).$$

Тем самым показано, что функция $B(x, t)$, определяемая соотношениями (5), (6), в самом деле является функцией Беллмана задачи (1.1)—(1.4), и функция $u(x, t)$, на которой достигается нижняя грань в левой части (5), является синтезирующей для этой задачи.

С помощью функций $B(x, t)$, $u(x, t)$ нетрудно получить решение и для исходной задачи (1.1)—(1.4). А именно, верна

Теорема 2. *Пусть $B(x, t)$ — решение задачи* (5), (6) *и пусть нижняя грань в левой части* (5) *достигается на кусочно-непрерывной функции $u(x, t)$. Кроме того, пусть точка $x_0^* \in X(t_0)$ определена из условия*

$$B(x_0^*, t_0) = \inf_{x \in X(t_0)} B(x, t_0), \tag{14}$$

а пара $(u_*(\cdot), x_*(\cdot))$, где $x_*(\cdot)$ — решение задачи Коши

$$\dot{x}(\tau) = f(x(\tau), u(x(\tau), \tau), \tau) \quad t_0 \leqslant \tau \leqslant T; \qquad x(t_0) = x_0^*,$$

и $u_*(\tau) = u(x_*(\tau), \tau)$, является допустимой парой задачи (1.1)—(1.4). *Тогда пара $(u_*(\cdot), x_*(\cdot))$ является решением задачи* (1.1)—(1.4), *т. е.*

$$J(t_0, x_0^*, u_*(\cdot)) = \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot)) = J_*.$$

Доказательство. Возьмем произвольную допустимую пару $(u(\tau), x(\tau))$ $(t_0 \leqslant \tau \leqslant T)$, $x(t_0) = x_0 \in X(t_0)$ задачи (1.1)—(1.4). Из формулы (9) и неравенств (12) при $t = t_0$ с учетом условия (14) имеем

$$J(t_0, x_0, u(\cdot)) - J(t_0, x_0^*, u_*(\cdot)) =$$
$$= \int_{t_0}^T R(x(\tau), u(\tau), \tau) d\tau + B(x_0, t_0) - \inf_{x \in X(t_0)} B(x, t_0) \geqslant 0,$$

что и требовалось доказать.

2. Таким образом, согласно теореме 1 для решения рассматриваемой проблемы синтеза достаточно найти решение задачи (5), (6). Возникает вопрос, как же решить задачу (5), (6)? Прежде всего заметим, что удобное для работы конструктивное описание множеств $X(t)$, $D(x, t)$, входящих в формулировку задачи (5), (6), часто отсутствует, и поэтому на практике вместо задачи (5), (6) обычно пользуются следующей более конструктивной задачей:

$$\begin{gathered} \inf_{u \in V(t)} \left[\langle B_x(x, t), f(x, u, t) \rangle + B_t(x, t) + f^0(x, u, t) \right] = 0, \\ x \in G(t), \quad t_0 \leqslant t < T, \\ B(x, T) = \Phi(x), \quad x \in G(T), \end{gathered} \tag{15}$$

получающейся из (5), (6) заменой $D(x, t)$, $X(t)$ на $V(t)$, $G(t)$ соответственно. Конечно, здесь надо помнить, что задача (15) может и не иметь решения, в то время как задача (5), (6) может оказаться разрешимой.

Наиболее удобными и эффективными при решении задачи (5), (6) или задачи (15), по-видимому, являются методы, изложенные выше в § 1, 2,— рекуррентные соотношения (1.13), (1.14) и (2.6) по существу представляют собой некоторую дискретную аппроксимацию задач (5), (6) и (15), а функции $B_k(x)$, $C_k(x)$ являются приближенным значением для $B(x, t_k)$. Существуют и другие методы решения таких задач. Во многих прикладных задачах часто удается получить явное выражение $u = u(x, t, B_x)$ для точки u, в которой достигается нижняя грань

$$\inf_{u \in V(t)} \left[\langle B_x, f(x, u, t) \rangle + f^0(x, u, t) \right]$$

при фиксированных значениях параметров (x, t, B_x). Подставив такое $u = u(x, t, B_x)$ в (15), приходим к следующей задаче Коши:

$$B_t + \left[\langle B_x, f(x, u, t) \rangle + f^0(x, u, t) \right] \Big|_{u=u(x,t,B_x)} = 0,$$
$$x \in G(t), \quad t_0 \leqslant t < T,$$
$$B(x, T) = \Phi(x), \quad x \in G(T),$$

для нелинейного уравнения с частными производными первого порядка. Для численного решения задачи Коши можно воспользоваться известным арсеналом методов — разностными методами, методом характеристик, методом прямых, методом коллокации и т. п. [4, 13, 20, 39, 54].

Иногда удается найти решение $B(x, t)$ задачи (15) в виде многочлена по переменным $x^1, \ldots, x^n$ с неопределенными коэффициентами, зависящими от времени:

$$B(x, t) = \sum_{i_1=0}^{m_1} \sum_{i_2=0}^{m_2} \ldots \sum_{i_n=0}^{m_n} \psi_{i_1 \ldots i_n}(t) (x^1)^{i_1} \ldots (x^n)^{i_n}.$$

Если подставим это выражение для $B(x, t)$ в (15), то для определения коэффициентов $\psi_{i_1 \ldots i_n}(t)$ получим дифференциальное уравнение следующего вида

$$B_t(x, t) = \sum_{i_1=0}^{m_1} \ldots \sum_{i_n=0}^{m_n} \dot{\psi}_{i_1 \ldots i_n}(t) (x^1)^{i_1} \ldots (x^n)^{i_n} =$$
$$= \inf_{u \in V(t)} F\left(\psi_{0 \ldots 0}(t), \ldots, \psi_{m_1 \ldots m_n}(t); x^1 \ldots x^n, u, t\right), \tag{16}$$
$$x \in G(t), \quad t_0 \leqslant t \leqslant T,$$

с начальным условием

$$\sum_{i_1=0}^{m_1} \ldots \sum_{i_n=0}^{m_n} \psi_{i_1 \ldots i_n}(T) (x^1)^{i_1} \ldots (x^n)^{i_n} = \Phi(x), \; x \in G(T). \tag{17}$$

Если $\Phi(x)$, $\inf\limits_{V(t)} F$, в свою очередь, являются многочленами относительно $x^1, \ldots, x^n$, то, приравняв коэффициенты при одинаковых степенях в (16), (17), получим задачу Коши для системы обыкновенных дифференциальных уравнений относительно $\psi_{i_1 \ldots i_n}(t)$, записанной в нормальной форме Коши. Далее, здесь можно использовать различные численные методы решения задачи Коши, такие, как методы Эйлера, Адамса, Рунге — Кутта и т. д. [4, 13, 39, 54, 258, 295].

Если же $\Phi(x)$ или $\inf\limits_{V(t)} F$ не являются многочленами относительно $x^1, \ldots, x^n$, то условия (16), (17) не могут быть, вообще говоря, удовлетворены во всей области $G(t)$ $(t_0 \leqslant t \leqslant T)$ ни при каком выборе $N = (m_1 + 1) \ldots (m_n + 1)$ коэффициентов $\psi_{i_1 \ldots i_n}(t)$. В этом случае в [188] предлагается задать в области $G(t)$ N кривых $\xi_1(t), \ldots, \xi_N(t)$ и рекомендуется определять $\psi_{i_1 \ldots i_n}(t)$ из условия удовлетворения равенств (16), (17) не всюду в $G(t)$, а лишь на этих кривых. Этот подход перекликается с известными методами коллокаций и интегральных соотношений и приводит к задаче Коши для системы обыкновенных дифференциальных уравнений, не разрешенных относительно производной $\dot{\psi}_{i_1 \ldots i_n}(t)$ (отметим, что эти производные в уравнение будут входить линейно). Кривые $\xi_1(t), \ldots, \xi_N(t)$ обычно выбирают так, чтобы они имели достаточно простое аналитическое выражение (например, семейство прямых, параллельных осям координат, семейство парабол и т. п.) и задавали достаточно густую сетку в области $G(t)$ $(t_0 \leqslant t \leqslant T)$.

Для иллюстрации вышесказанного приведем пример.

Пример 1. Пусть требуется минимизировать функцию

$$J(u) = \int_0^T u^2(t)\,dt + \lambda x^2(T), \qquad \lambda = \text{const} > 0$$

при условиях $\dot{x} = u(t)$, $x(0) = x_0$, $u = u(t)$ — кусочно-непрерывная функция; числа T, x_0 заданы. Здесь $G(t) \equiv E^1$, $V(t) \equiv E^1$ $(0 \leqslant t \leqslant T)$.

Задача (15) в рассматриваемом случае имеет вид

$$\inf_{u \in E^1} \left[B_x(x, t)\,u + B_t(x, t) + u^2\right] = 0, \qquad x \in E^1, \qquad 0 \leqslant t \leqslant T, \tag{18}$$

$$B(x, T) = \lambda x^2, \qquad x \in E^1. \tag{19}$$

Нижняя грань в (18) достигается при $u = -B_x/2$, поэтому уравнение (18) перепишется так:

$$B_t(x, t) - B_x^2(x, t)/4 = 0, \qquad x \in E^1, \qquad 0 \leqslant t \leqslant T. \tag{20}$$

Функцию $B(x, t)$ будем искать в виде многочлена

$$B(x, t) \equiv \psi_0(t) + \psi_1(t)x + \psi_2(t)x^2$$

переменной x. Подставим это выражение в (19), (20); получим

$$\dot{\psi}_0 + \dot{\psi}_1 x + \dot{\psi}_2 x^2 - (\psi_1 + 2\psi_2 x)^2/4 = 0, \qquad x \in E^1, \qquad 0 \leqslant t \leqslant T,$$
$$\psi_0(T) + \psi_1(T)x + \psi_2(T)x^2 = \lambda x^2, \qquad x \in E^1.$$

Приравнивая коэффициенты при одинаковых степенях x, придем к следующей задаче Коши:

$$\dot{\psi}_0 - \psi_1^2/4 = 0, \qquad \dot{\psi}_1 - \psi_1\psi_2 = 0, \qquad \dot{\psi}_2 - \psi_2^2 = 0, \qquad 0 \leqslant t \leqslant T,$$
$$\psi_0(T) = 0, \qquad \psi_1(T) = 0, \qquad \psi_2(T) = \lambda.$$

Отсюда находим

$$\psi_0(t) \equiv \psi_1(t) \equiv 0, \qquad \psi_2(t) = \frac{\lambda}{1 - \lambda(t - T)}.$$

Таким образом, функция Беллмана здесь имеет вид

$$B(x, t) = \frac{\lambda x^2}{1 - \lambda(t - T)};$$

синтезирующей является функция

$$u(x, t) = -\frac{B_x}{2} = \frac{\lambda x}{1 - \lambda(t - T)}, \quad x \in E^1, \quad 0 \leqslant t \leqslant T.$$

3. Предположим, что с помощью того или иного метода нам удалось получить некоторое приближенное решение $B(x, t)$ задачи (5), (6) или (15). Если это решение получено разностным методом (например, методами § 1, 2) на какой-то дискретной сетке точек, то доопределим ее (например, интерполяцией или с помощью сплайнов) во всех точках области $G(t)$ $(t_0 \leqslant t \leqslant T)$ до некоторой непрерывной кусочно-гладкой функции $B(x, t)$. Тогда функцию $u = u(x, t)$, на которой реализуется точная или приближенная нижняя грань функции $R(x, u, t)$ из (10) на множестве $D(x, t)$ или $V(t)$, можем принять в качестве приближенного решения проблемы синтеза для задачи (1.1)—(1.4). Это значит, что приближенное решение $(\bar{u}(\cdot), \bar{x}(\cdot))$ задачи (1)—(4) будем определять из условий

$$\begin{gathered} \dot{\bar{x}}(\tau) = f(\bar{x}(\tau), \quad u(\bar{x}(\tau), \tau), \tau), \quad t \leqslant \tau \leqslant T, \ \bar{x}(T) = x, \\ \bar{x}(\tau) \in G(\tau), \quad \bar{u}(\tau) = u(\bar{x}(\tau), \tau), \quad t \leqslant \tau \leqslant T. \end{gathered} \tag{21}$$

Приближенное решение исходной задачи (1.1)—(1.4) находится аналогично: сначала определяем точку $\bar{x}_0$, на которой точно или приближенно реализуется нижняя грань функции $B(x, t_0)$ на множестве $X(t_0)$ или $G(t_0)$, а затем решая задачу (21) при $t = t_0$, $x = \bar{x}_0$, находим траекторию $\bar{x}(\tau)$ и управление $\bar{u}(\tau) = u(\bar{x}(\tau))$ $(t_0 \leqslant \tau \leqslant T)$. Найденную пару $(\bar{u}(\cdot), \bar{x}(\cdot))$ примем за приближенное решение задачи (1.1)—(1.4). Спрашивается, какая при этом будет допущена погрешность? Приводимая ниже оценка погрешности дает некоторый ответ на этот вопрос.

Пусть $K(x, t)$ — какая-либо функция, которая определена и непрерывна при всех $x \in X(t)$ $(t_0 \leqslant t \leqslant T)$, обладает кусочно-непрерывными производными K_x, K_t и такова, что для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)—(4) при всех $x \in X(t)$ $(t_0 \leqslant t < T)$, функция $K(x(\tau), \tau)$ переменной τ — кусочно-гладкая (или абсолютно непрерывная) на $[t, T]$. На практике в качестве функции $K(x, t)$ обычно берут какое-либо приближенное решение $B(x, t)$ задачи (5), (6) или (15).

По аналогии с (10) введем функцию

$$\begin{gathered} S(x, u, t) = \langle K_x(x, t), f(x, u, t)\rangle + K_t(x, t) + f^0(x, u, t), \\ x \in X(t), \quad t_0 \leqslant t \leqslant T, \end{gathered} \tag{22}$$

и, кроме того, положим

$$s(x, T) = \Phi(x) - K(x, T), \quad x \in G(T). \tag{23}$$

Возьмем произвольную допустимую пару $(u(\cdot), x(\cdot))$ задачи (1)—(4). В силу уравнения (2) тогда имеем

$$\frac{dK(x(\tau), \tau)}{d\tau} = S(x(\tau), u(\tau), \tau) - f^0(x(\tau), u(\tau), \tau), \quad t_0 \leqslant \tau \leqslant T.$$

Учитывая непрерывность и кусочно-гладкую функцию $K(x(\tau), \tau)$, проинтегрируем это тождество по τ от t до T. Получим формулу

$$J(t, x, u(\cdot)) = \int_t^T S(x(\tau), u(\tau), \tau)\, d\tau + s(x(T), T) + K(x, t). \tag{24}$$

Если $K(x, t) = B(x, t)$, то $S(x, u, \tau) = R(x, u, \tau)$ $s(x, T) \equiv 0$, и эта формула превратится в выведенную выше формулу (9).

Предположим, что каким-то образом мы получили пару $(\bar{u}(\tau), \bar{x}(\tau))$ $(t \leqslant \tau \leqslant T)$, удовлетворяющую условиям (3), (4) и уравнению (2) с на-

чальным условием $\bar{x}(t) = \bar{x} \in X(t)$. Согласно (24) тогда

$$J(t, \bar{x}, \bar{u}(\cdot)) - J(t, x, u(\cdot)) = \int_t^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau) - S(x(\tau), u(\tau), \tau)]\, d\tau +$$

$$+ [s(\bar{x}(T), T) - s(x(T), T)] + [K(\bar{x}, t) - K(x, t)] \quad (25)$$

для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)—(4). Из (25) уже нетрудно получить требуемые оценки погрешности для задач (1)—(4) и (1.1)—(1.4). Обозначим

$$S_{\min}(\tau) = \inf_{x \in X(\tau)} \inf_{u \in D(x,\tau)} S(x, u, \tau), \quad s_{\min} = \inf_{x \in X(T)} s(x, T),$$

$$K_{0\min} = \inf_{x \in X(t_0)} K(x, t_0). \quad (26)$$

Пусть $(\bar{u}(\cdot), \bar{x}(\cdot))$ — некоторая допустимая пара задачи (1)—(4), которую мы хотим взять в качестве приближенного решения этой задачи. Учитывая, что для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)—(4) имеют место включения $x(t) = x \in X(t)$, $u(\cdot) \in \Delta(x, t)$, $x(\tau) \in X(\tau)$, $u(\tau) \in D(x(\tau), \tau)$ $(t \leqslant \tau \leqslant T)$ из 25, 26 получим требуемую оценку погрешности:

$$0 \leqslant J(t, x, \bar{u}(\cdot)) - \inf_{\Delta(x,t)} J(t, x, u(\cdot)) \leqslant$$

$$\leqslant \int_t^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau] - S_{\min}(\tau)]\, d\tau + [s(\bar{x}(T), T) - s_{\min}]. \quad (27)$$

Если же $(\bar{u}(\tau), \bar{x}(\tau))$ $(t_0 \leqslant \tau \leqslant T)$ — допустимая пара задачи (1.1)—(1.4), $\bar{x}(t_0) = \bar{x}_0$, которая берется за приближенное решение этой задачи, то из (25), (26) имеем такую оценку погрешности:

$$0 \leqslant J(t_0, \bar{x}_0, \bar{u}(\cdot)) - \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot)) \leqslant$$

$$\leqslant \int_{t_0}^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau) - S_{\min}(\tau)]\, d\tau + [s(\bar{x}(T), T) - s_{\min}] +$$

$$+ K(\bar{x}_0, t_0) - K_{0\min}. \quad (28)$$

Если $K(x, t) = B(x, t)$, то $S(x, u, t) = R(x, u, t)$, $s(x, T) = 0$ и, кроме того, из (11) следует, что $\inf\limits_{u \in D(x,t)} R(x, u, t) = 0$ при всех $x \in X(t)$, так что $\inf\limits_{x \in X(t)} \inf\limits_{u \in D(x,t)} R(x, u, t) = 0$. Поэтому при $K(x, t) = B(x, t)$ из (27), (28) соответственно получим

$$0 \leqslant J(t, x, \bar{u}(\cdot)) - \inf_{u(\cdot) \in \Delta(x,t)} J(t, x, u(\cdot)) \leqslant \int_t^T R(\bar{x}(\tau), \bar{u}(\tau), \tau)\, d\tau, \quad (29)$$

$$0 \leqslant J(t_0, \bar{x}_0, \bar{u}(\cdot)) - \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot)) \leqslant$$

$$\leqslant \int_{t_0}^T R(\bar{x}(\tau), \bar{u}(\tau), \tau)\, d\tau + B(\bar{x}_0, t_0) - \inf_{x \in X(t_0)} B(x, t_0). \quad (30)$$

Из оценок (29), (30) следует, что если определить $\bar{u}(x, t) \in \Delta(x, t)$ [$\bar{x}_0 \in X(t_0)$ — для задачи (1.1)—(1.4)] так, чтобы $R(x, \bar{u}(x, t), t)$ было поближе к $\inf\limits_{\Delta(x,t)} R(x, u, t)$ $\left[B(x_0, t_0)\right.$ — поближе к $\left.\inf\limits_{x \in X(t_0)} B(x, t_0)\right]$ и затем найти пару $(\bar{u}(\tau), \bar{x}(\tau))$ из условий

$$\dot{\bar{x}}(\tau) = f(\bar{x}(\tau), \bar{u}(\bar{x}(\tau), \tau), \tau), \quad \bar{x}(\tau) \in G(\tau), \quad t \leqslant \tau \leqslant T,$$
$$\bar{x}(t) = x, \quad \bar{u}(\tau) = \bar{u}(\bar{x}(\tau), \tau)$$

[для задачи (1.1)—(1.4) здесь надо взять $t = t_0$, $\bar{x}(t_0) = \bar{x}_0$], то величина $J(t, x, \bar{u}(\cdot))$ [в случае задачи (1.1)—(1.4) — величина $J(t, \bar{x}_0, \bar{u}(\cdot))$] будет мало отличаться от искомого оптимального значения, а функция $\bar{u}(x, t)$ будет хорошим приближением для синтезирующей функции. Заметим, что оценки (28), (30) являются аналогами оценок (1.28) и (1.29).

Заметим также, что в приложениях могут оказаться удобнее более грубые оценки, получающиеся из (26)—(30) при замене неконструктивно определенных множеств $\Delta(x, t)$, $X(t)$ на множества $V(t)$, $G(t)$ соответственно.

Упражнения. 1. Решить проблему синтеза для задачи минимизации функции $J(x_0, u(\cdot)) = \int\limits_0^T (x^2(t) + u^2(t))\, dt$ при условиях $\dot{x}(t) = -x(t) + u(t)$, $x(0) = x_0$; здесь $G(t) \equiv E^1$, $V(t) \equiv E^1$ при всех $t \in [0, T]$.

2. Решить проблему синтеза для задачи минимизации функции $J(x_0, u(\cdot)) = x^2(1)$ при условиях $\dot{x}(t) = u(t)$ $(0 \leqslant t \leqslant 1)$, $x(0) = x_0$, $u(t) \in V = \{u \in E^1\colon 0 \leqslant u \leqslant 1\}$, $x(t) \in G(t) \equiv E^1$ при $0 \leqslant t \leqslant 1$. Показать, что в этой задаче синтезирующих функций бесконечно много. Убедиться, например, что синтезирующими являются функции

$$u(x, t) = \begin{cases} 0, & x \geqslant 0, \\ 1, & x < 0, \end{cases} \quad \text{или} \quad u(x, t) = \begin{cases} 0, & x > t - 1, \\ 1, & x \leqslant t - 1. \end{cases}$$

3. Решить проблему синтеза для задачи минимизации функций

$$J(x_0, u(\cdot)) = x^2(T) \quad \text{или} \quad J(x_0, u(\cdot)) = \int\limits_0^T x^2(t)\, dt$$

при условиях $\dot{x}(t) = u(t)$ $(0 \leqslant t \leqslant T)$; $x(0) = x_0$, $u(t) \in V(t) = \{u \in E^1\colon -1 \leqslant u \leqslant 1\}$, $x(t) \in G(t) \equiv E^1$ $(0 \leqslant t \leqslant T)$. Будет ли в этих задачах синтезирующая функция единственной?

4. Написать уравнения Беллмана (5), (6) для задачи быстродействия:

$$J(x_0, t_0, u(\cdot)) = T - t_0 \to \inf, \quad \dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T,$$

$x(t_0) = x_0$, $x(T) = x_1$, $u(\cdot)$ — кусочно-непрерывна и принадлежит множеству $V \subseteq E^r$, $t \geqslant t_0$.

5. Показать, что функция Беллмана для задачи из примера 6.2.4 не является непрерывно дифференцируемой.

6. Найти функцию Беллмана для задачи

$$J(t_0, x_0, u(\cdot)) = \int\limits_{t_0}^T [\langle a(t), x(t)\rangle + b(u(t), t)]\, dt + \langle c, x(T)\rangle \to \inf,$$

$$\dot{x}(t) = A(t)x(t) + C(t)u(t) + f(t), \quad t_0 \leqslant t \leqslant T, \quad x(t_0) = x_0, \tag{31}$$

$$u = u(t) \in V(t), \ u(\cdot) \text{ — кусочно-непрерывна}, \tag{32}$$

считая известными моменты времени t_0, T, матрицы $A(t)$, $C(t)$ порядка $n \times n$ и $n \times r$ соответственно, n-мерные вектор-функции $a(t)$, $f(t)$, скалярную функцию $b(u, t)$, n-мерные векторы c, x_0 и множества $V(t) \in E^r$ $(t_0 \leqslant t \leqslant T)$. Указание: пользуясь уравнениями (5), (6) искать функцию $B(x, t)$ в виде $B(x, t) = \langle \psi(t), x \rangle$ — многочлена первой степени относительно переменных $x = (x^1, \ldots, x^n)$.

7. Исследовать уравнения (5), (6) для задачи минимизации функции

$$J(t_0, x_0, u(\cdot)) = \alpha_1 \int_{t_0}^{T} x^2(t)\, dt + \alpha_2 x^2(T), \quad \alpha_1, \alpha_2 = \text{const} \geqslant 0,$$

при условиях (31), (32). Рассмотреть случаи $V(t) = E^r$, $V(t) = \{u \in E^r: |u| \leqslant 1\}$, $V(t) = \{u = (u^1, \ldots, u^r): -1 \leqslant u^i \leqslant 1, i = 1, \ldots, r\}$.

§ 4. Достаточные условия оптимальности

При решении задач оптимального управления часто возникает следующий вопрос: будут ли на самом деле оптимальными те управления и соответствующие им траектории, которые найдены с помощью каких-либо точных или приближенных методов? Такой вопрос, например, естественно возникает, когда управление и траектория найдены из краевой задачи принципа максимума, поскольку принцип максимума выражает собой необходимое условие оптимальности, не являясь, в общем случае, достаточным для оптимальности. Один из подходов, с помощью которого можно получить достаточные условия оптимальности, связан с методом динамического программирования [65, 81, 101, 124, 125, 188, 189, 199, 263]. Этот подход уже был использован в § 1 для получения достаточных условий в дискретных системах. Покажем возможности этого подхода для задач оптимального управления с непрерывно меняющимся временем.

1. Начнем с рассмотрения задачи (1.1)—(1.4) с закрепленным временем. Будем пользоваться обозначениями и некоторыми формулами из § 3. Согласно (3.24) и (3.28) для каждой допустимой пары $(\bar{u}(t), \bar{x}(t))$ $(t_0 \leqslant t \leqslant T)$, $\bar{x}(t_0) = \bar{x}_0$, задачи (1.1)—(1.4) справедливы формула

$$J(t_0, \bar{x}_0, \bar{u}(\cdot)) = \int_{t_0}^{T} S(\bar{x}(t), \bar{u}(t), t)\, dt + s(x(T), T) + K(\bar{x}_0, t_0) \tag{1}$$

и оценка

$$0 \leqslant J(t_0, \bar{x}_0, \bar{u}(\cdot)) - J_* \leqslant \int_{t_0}^{T} [S(\bar{x}(t), \bar{u}(t), t) - S_{\min}(t)]\, dt + \\ + [s(\bar{x}(T), T) - s_{\min}] + [K(\bar{x}_0, t_0) - K_{0\min}], \tag{2}$$

где

$$S(x, u, t) = \langle K_x(x, t), f(x, u, t) \rangle + K_t(x, t) + f^0(x, u, t), \tag{3}$$

$$S(x, T) = \Phi(x) - K(x, T), \tag{4}$$

$$S_{\min}(t) = \inf_{x \in X(t)} \inf_{u \in D(x,t)} S(x, u, t), \quad s_{\min} = \inf_{x \in X(T)} s(x, T), \tag{5}$$

$$K_{0\min} = \inf_{x \in X(t_0)} K(x, t_0), \quad J_* = \inf_{x \in X(t_0)} \operatorname*{int}_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot));$$

остальные обозначения см. в § 3. Напомним, что для справедливости соотношений (1), (2) достаточно было, чтобы функция $K(x, t)$ была определена и непрерывна при всех $x \in G(t)$, $t \in [t_0, T]$, обладала кусочно-непре-

рывными производными K_x, K_t, а функция $K(x(\tau), \tau)$ переменной τ для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1.1)—(1.4) была непрерывной и кусочно-гладкой на отрезке $[t_0, T]$.

Теорема 1. *Для того чтобы допустимая пара* $(\bar{u}(t), \bar{x}(t))$ $(t_0 \leqslant t \leqslant T)$, $\bar{x}(t_0) = \bar{x}_0$, *задачи* (1.1)—(1.4) *была решением этой задачи, достаточно существования функции* $K(x, t)$, *для которой формула* (1) *верна для любой допустимой пары задачи* (1.1)—(1.4) *и*

$$S(\bar{x}(t), \bar{u}(t), t) = S_{\min}(t), \quad t_0 \leqslant t \leqslant T, \tag{6}$$

$$s(x(T), T) = s_{\min}, \quad K(\bar{x}_0, t_0) = K_{0\min}, \tag{7}$$

где $S(x, u, t)$, $s(x, t)$, $S_{\min}(t)$, $s_{\min}$, $K_{0\min}$ *определяются* (3)—(5).

Доказательство этой теоремы следует из того, что при выполнении условий (6), (7) правая часть оценки (2) обращается в нуль и $J(t_0, \bar{x}_0, \bar{u}(\cdot)) = J_*$.

С помощью оценки (2) нетрудно также получить условия, достаточные для того, чтобы та или иная последовательность допустимых управлений и траекторий была минимизирующей для задачи (1.1) — (1.4).

Теорема 2. *Для того чтобы некоторая последовательность* $(u_m(t), x_m(t))$ $(t_0 \leqslant t \leqslant T)$, $x_m(t_0) = x_{0m}$ $(m = 1, 2, \ldots)$ *допустимых пар задачи* (1.1)—(1.4) *была минимизирующей, достаточно существования функции* $K(x, t)$, *для которой формула* (1) *верна для любой допустимой пары задачи* (1.1)—(1.4) *и*

$$\lim_{m\to\infty} \int_{t_0}^{T} S(x_m(t), u_m(t), t)\, dt = \int_{t_0}^{T} S_{\min}(t)\, dt, \tag{8}$$

$$\lim_{m\to\infty} s(x_m(T), T) = s_{\min}, \quad \lim_{m\to\infty} K(x_{0m}, t_0) = K_{0\min}. \tag{9}$$

Доказательство. В оценке (2) вместо $\bar{u}(t)$, $\bar{x}(t)$ подставим $u_m(t)$, $x_m(t)$ и перейдем к пределу при $m \to \infty$. С учетом условий (8), (9) получим $\lim\limits_{m\to\infty} J(t_0, x_{0m}, u_m(\cdot)) = J_*$, что и требовалось.

Чтобы пользоваться приведенными в теоремах 1, 2 достаточными условиями оптимальности, нужно знать функцию $K(x, t)$, с помощью которой строятся функции (3) — (5). Как найти такую функцию $K(x, t)$, каким условиям она удовлетворяет?

Всякую функцию $K(x, t)$, удовлетворяющую условиям теоремы 1 [теоремы 2], назовем функцией Кротова задачи (1.1)—(1.4), соответствующей допустимой паре $(\bar{u}(t), \bar{x}(t))$ $(t_0 \leqslant t \leqslant T)$ [или последовательности $(u_m(t), x_m(t))$ допустимых пар].

Заметим, что если существует какая-либо функция Кротова $K(x, t)$, то функцией Кротова является также функция $\tilde{K}(x, t) = K(x, t) + \alpha(t)$, где $\alpha(t)$ — произвольная непрерывная, кусочно-гладкая (или абсолютно непрерывная) на $[t_0, T]$ функция. В частности, если $\alpha(t) = -\int_T^t S_{\min}(\tau)\, d\tau - s_{\min}$, то функции $\tilde{S}(x, u, t)$, $\tilde{s}(x, T)$, построенные по формулам (3), (4) с заменой K на $\tilde{K} = K + \alpha$, таковы, что $\tilde{S}(x, u, t) = S(x, u, t) - S_{\min}(t)$, $\tilde{s}(x, T) = s(x, T) - s_{\min}$ и, следовательно, $\inf\limits_{x\in X(t)} \inf\limits_{u\in D(x,t)} \tilde{S}(x, u, t) \equiv 0$ $(t_0 \leqslant t \leqslant T)$, $\inf\limits_{x\in X(T)} \tilde{s}(x, T) = 0$, а $\inf\limits_{x\in X(t_0)} \tilde{K}(x, t_0) = K_{0\min} + \alpha(t_0)$ отли-

чается от $K_{0\min}$ на постоянную $\alpha(t_0)$. Поэтому в теоремах 1, 2 без ограничения общности можем принять $S_{\min}(t) \equiv 0$, $s_{\min} = 0$.

С учетом этого замечания заключаем, что функция Кротова, соответствующая допустимой паре $(\bar{u}(t), \bar{x}(t))$ или последовательности $(u_m(t), x_m(t))$ допустимых пар задачи (1.1)—(1.4), согласно теоремам 1, 2 удовлетворяет условиям

$$S(x, u, t) = \langle K_x(x, t), f(x, u, t)\rangle + K_t(x, t) + f^0(x, u, t) \geqslant 0, \tag{10}$$

$$u \in D(x, t), \quad x \in X(t), \quad t_0 \leqslant t \leqslant T;$$

$$s(x, T) = \Phi(x) - K(x, T) \geqslant 0, \quad x \in X(T), \tag{11}$$

$$K(x, t_0) \geqslant K_{0\min}, \quad x \in X(t_0),$$

причем неравенства (10), (11) должны обратиться в равенства при $u = \bar{u}(t)$, $x = \bar{x}(t)$ или при $u = u_m(t)$, $x = x_m(t)$ в пределе при $m \to \infty$.

Задача (10), (11) для определения функции Кротова несколько необычна тем, что, во-первых, здесь мы имеем дело не с дифференциальным уравнением, а с дифференциальным неравенством (10) в частных производных, во-вторых, начальное условие при $t = T$ также задано в виде неравенства и, наконец, задача (10), (11) тесно связана с конкретной допустимой парой $(u(\cdot), x(\cdot))$ или последовательностью $(u_m(\cdot), x_m(\cdot))$ допустимых пар, подозреваемых на оптимальность.

Сравнение соотношений (10), (11) с (3.5), (3.6), (3.14) показавает, что функция Беллмана всегда является функцией Кротова. С другой стороны, функция Кротова определяется из более широких условий (10), (11), и она может существовать даже тогда, когда функция Беллмана не существует.

В тех случаях, когда отсутствует удобное для работы конструктивное описание множеств $D(x, t)$, $X(t)$, функцию Кротова можно попытаться определить из условий, получающихся из (10), (11) при замене $D(x, t)$, $X(t)$ на $V(t)$, $G(t)$ соответственно.

Проиллюстрируем вышесказанное на примерах.

Пример 1. Пусть требуется минимизировать функцию $J(u(\cdot)) = \int_0^1 (x^2(t) - u(t))\,dt$ при условиях $\dot{x}(t) = u(t)$, $x(0) = x(1) = 0$, $u(t) \in V(t) = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant 1)$.

Здесь фазовые ограничения $G(t)$ такие: $G(0) = G(1) = \{0\}$, $G(t) \equiv E^1$ $(0 < t < 1)$. Очевидно, пара $(\bar{u}(t) \equiv 0, \bar{x}(t) \equiv 0)$ является допустимой для этой задачи. Покажем, что она является решением задачи. Для этого возьмем функцию $K(x, t) \equiv x$. Тогда

$$S(x, u, t) = x^2 \geqslant 0 = S(\bar{x}(t), \bar{u}(t), t) = S_{\min}(t),$$

$$s(x, 1) = -x = 0 = s(\bar{x}(1), 1) = \inf_{x \in G(1)} s(x, 1),$$

$$K(x, 0) = x = 0 = K(\bar{x}(0), 0) = \inf_{x \in G(0)} K(x, 0).$$

Кроме того, ясно, что формула (1) здесь будет верна для любой допустимой пары. Согласно теореме 1 тогда пара $(\bar{u}(t) \equiv 0, \bar{x}(t) \equiv 0)$ оптимальна. Предлагаем читателю найти функцию Беллмана и синтезирующую функцию этой задачи.

Пример 2. Пусть требуется минимизировать функцию $J(u(\cdot)) = \int_0^1 (x^2(t) - u^2(t))\,dt$ при условиях $\dot{x}(t) = u(t)$, $x(0) = 0$, $u(t) \in V(t) = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant 1)$.

Здесь фазовые ограничения $G(t)$ такие: $G(0) = \{0\}$, $G(t) \equiv E^1$ $(0 < t \leqslant 1)$. Возьмем последовательность пар функций $(u_m(\cdot), x_m(\cdot))$, где

$$u_m(t) = \begin{cases} 1, & \frac{p}{m} < t \leqslant \frac{p}{m} + \frac{1}{2m}, \\ -1, & \frac{p}{m} + \frac{1}{2m} < t \leqslant \frac{p}{m} + \frac{1}{m}, \end{cases}$$

$$x_m(t) = \begin{cases} t - \frac{p}{m}, & \frac{p}{m} < t \leqslant \frac{p}{m} + \frac{1}{2m}, \\ -t + \frac{p+1}{m}, & \frac{p}{m} + \frac{1}{2m} < t \leqslant \frac{p}{m} + \frac{1}{m}, \end{cases}$$

$$p = 0, 1, \ldots, m-1, \quad m = 1, 2, \ldots$$

Нетрудно проверить, что пара $(u_m(\cdot), x_m(\cdot))$ допустима для рассматриваемой задачи при всех $m = 1, 2, \ldots$ Покажем, что последовательность этих пар является минимизирующей. Для этого возьмем функцию $K(x, t) = t - 1$. Тогда $S(x, u, t) = x^2 + 1 - u^2 \geqslant 0 = \min\limits_{x \in E^1} \min\limits_{|u| \leqslant 1} S(x, u, t) = \lim\limits_{m \to \infty} S(x_m(t), u_m(t), t)$, $s(x, 1) = -K(x, 1) \equiv 0 = \inf\limits_{x \in E^1} K(x, 1) = \lim\limits_{m \to \infty} s(x_m(1), 1)$, $K(x, 0) \equiv -1 = \inf\limits_{x \in G(0)} K(x, 0) = \lim\limits_{m \to \infty} K(x_m(0), 0)$. Согласно теореме 2 последовательность $(u_m(\cdot), x_m(\cdot))$ будет минимизирующей: $\lim\limits_{m \to \infty} J(u_m(\cdot)) = -1 = J_*$.

Заметим, что в этом примере $\inf J(u) = -1$ не достигается ни на какой допустимой паре. В самом деле, если $x^2(t) \equiv 0$, то в силу уравнения $\dot{x} = u(t) \equiv 0$ и $J(0) = 0 > -1$. Если же $x^2(t) \not\equiv 0$, то $J(u) > -\int\limits_0^1 u^2(t)\,dt \geqslant -1$. Таким образом, $J(u(\cdot)) > -1$ для всех допустимых управлений и траекторий. В этой задаче мы имеем дело с так называемым скользящим режимом [37, 77, 104, 124, 125, 188, 189, 304]. Предлагаем читателю найти функцию Беллмана и приближенную синтезирующую функцию этой задачи.

2. Перейдем к рассмотрению следующей задачи оптимального управления с незакрепленным временем: минимизировать функцию

$$J(t_0, T, x_0, u(\cdot)) = \int\limits_{t_0}^{T} f^0(x(t), u(t), t)\,dt + \Phi(x(T), T) \tag{12}$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leqslant t \leqslant T; \quad x(t_0) = x_0, \tag{13}$$

$$x(t) \in G(t), \quad t_0 \leqslant t \leqslant T; \quad x(T) \in S_1(T) \subseteq G(T), \tag{14}$$

$$u = u(t) \in V(t), \quad u(t) \text{ — кусочно-непрерывна}, \quad t_0 \leqslant t \leqslant T,$$

$$u(t) = u(t+0) = \lim_{\tau \to t+0} u(\tau), \quad t_0 \leqslant t < T, \tag{15}$$

где моменты t_0, T, в отличие от задачи (1.1)—(1.4), неизвестные и таковы, что

$$t_0 \in \theta_0, \quad T \in \theta_1; \tag{16}$$

θ_0, θ_1 — некоторые множества на числовой оси $-\infty < t < \infty$; остальные обозначения см. в § 6.1. В частности, если $f^0 \equiv 1$, $\Phi \equiv 0$, то задача (12)—(16) превратится в задачу быстродействия. Всюду ниже будем предполагать, что $t_0 \leqslant T$.

Через $\Delta(x, t, T)$ обозначим множество всех управлений $u(\cdot)$, определенных на отрезке $[t, T]$, удовлетворяющих условиям (15) и таких, что траектория системы $\dot{x}(\tau) = f(x(\tau), u(\tau), \tau)$, $x(t) = x \in G(t)$ также определена на отрезке $[t, T]$, причем $x(\tau) \in G(\tau)$ $(t \leqslant \tau \leqslant T)$, $x(T) \in S_1(T)$. Положим $X(t, T) = \{x\colon x \in G(t),\ \Delta(x, t, T) \neq \varnothing\}$ при $t < T$; $X(T, T) \equiv S_1(T)$. Введем также множество $D(x, t, T)$ всех тех $u \in V(t)$, для которых существует хотя бы одно управление $u(\tau) \in \Delta(x, t, T)$ со значением $u(t) = u(t+0) = u$. Пару $(u(t), x(t))$ назовем допустимой парой задачи (12)—(16), если функции $u(t)$, $x(t)$ определены на каком-либо отрезке $[t_0, T]$, где $t_0 \in \theta_0$, $T \in \theta_1$, и такие, что $x(t_0) = x_0 \in G(t_0)$, $u(\cdot) \in \Delta(x_0, t_0, T)$, $x(\cdot)$ — траектория системы (13). Если $(u(\tau), x(\tau))$ $(t_0 \leqslant \tau \leqslant T)$ является допустимой парой задачи (12)—(16), то $u(\cdot) \in \Delta(x(t), t, T)$, $x(t) \in X(t, T)$, $u(t) \in D(x(t), t, T)$ для всех t $(t_0 \leqslant t < T)$.

Моменты времени $t_0^* \in \theta_0$, $T^* \in \theta_1$ и допустимую пару $(u_*(t), x_*(t))$, определенную на отрезке $\left[t_0^*, T^*\right]$, назовем решением задачи (12)—(16), если

$$J\left(t_0^*, T^*, x_*(t_0), u_*(\cdot)\right) = \inf_{t_0 \in \theta_0} \inf_{T \in \theta_1} \inf_{x_0 \in X(t_0, T)} \inf_{u(\cdot) \in \Delta(x_0, t_0, T)} J(t_0, T, x_0, u(\cdot)) = J_*.$$

Скажем, что последовательности моментов $\{t_{0m}\} \in \theta_0$, $\{T_m\} \in \theta_1$ и допустимых пар $(u_m(t), x_m(t))$, определенных на отрезке $[t_{0m}, T_m]$, являются минимизирующими для задачи (12)—(16), если $\lim\limits_{m \to \infty} J(t_{0m}, T_m, x_m(t_{0m}), u_m(\cdot)) = J_*$.

Для формулировки достаточных условий оптимальности снова воспользуемся функциями $S(x, u, t)$, $s(x, T)$, определяемыми равенствами (3), (4), причем для случая рассматриваемой задачи (12)—(16) в (4) вместо $\Phi(x)$ будем брать $\Phi(x, T)$. Будем считать, что функция $K(x, t)$ такова, что формула (1) остается верной для любых допустимых пар задачи (12)—(16) — для этого достаточно, чтобы функция $K(x, t)$ была определена и непрерывна при всех $x \in G(t)$, $t \in [\inf \theta_0, \sup \theta_1]$, обладала кусочно-непрерывными K_x, K_t, а функция $K(x(t), t)$ переменной t для любой допустимой пары $(u(t), x(t))$ $(t_0 \leqslant t \leqslant T)$ задачи (12)—(16) была кусочно-гладкой на отрезке $[t_0, T]$.

Пусть $(u_*(t), x_*(t))$ и $(u(t), x(t))$ — какие-либо допустимые пары задачи (12)—(16), определенные на отрезках $\left[t_0^*, T^*\right]$ и $[t_0, T]$ соответственно. Тогда из формулы (1) получим

$$J\left(t_0^*, T^*, x_*\left(t_0^*\right), u_*(\cdot)\right) - J\left(t_0, T, x(t_0), u(\cdot)\right) = \int_{t_0^*}^{T^*} S(x_*(t)), u_*(t), t)\, dt - \int_{t_0}^{T} S(x(t), u(t), t)\, dt + [s(x_*(T^*), T^*) - s(x(T), T)] + \left[K\left(x_*\left(t_0^*\right), t_0^*\right) - K\left(x(t_0), t_0\right)\right]. \quad (17)$$

Обозначим

$$S_{\min} = \inf_{t_0 \in \theta_0} \inf_{T \in \theta_1} \int_{t_0}^{T} \inf_{x \in X(t, T)} \inf_{u \in D(x, t, T)} S(x, u, t)\, dt, \quad (18)$$

$$s_{\min} = \inf_{T \in \theta_1} \operatorname{int}_{x \in S_1(T)} s(x, T), \qquad K_{0\,\min} = \inf_{t_0 \in \theta_0} \inf_{T \in \theta_1} \inf_{x \in X(t_0, T)} K(x, t_0).$$

Учитывая, что для любой допустимой пары $(u(t), x(t))$ $(t_0 \leqslant t \leqslant T)$ задачи (12)—(16) имеют место включения $u(t) \in D(x(t), t, T)$, $x(t) \in X(t, T)$ при всех t $(t_0 \leqslant t < T)$ и $u(\cdot) \in \Delta(x(t_0), t_0, T)$, $x(t_0) \in X(t_0, T)$, $x(T) \in X(T, T) = S_1(T)$, $t_0 \in \theta_0$, $T \in \theta_1$, из (17), (18) получим следующие неравенства:

$$0 \leqslant J\left(t_0^*, T^*, x_*\left(t_0^*\right), u_*(\cdot)\right) - J_* \leqslant \int_{t_0^*}^{T^*} S(x_*(t), u_*(t), t)\, dt - S_{\min} +$$

$$+ \left[s(x_*(T^*), T^*) - s_{\min}\right] + \left[K\left(x_*\left(t_0^*\right), t_0^* - K_{0\min}\right]. \quad (19)$$

Неравенства (19) обобщают формулу (2) на случай задач оптимального управления с незакрепленным временем и представляют собой оценку погрешности, которая будет допущена, если допустимую пару $(u_*(t), x_*(t))$ $\left(t_0^* \leqslant t \leqslant T^*\right)$ задачи (12)—(16) возьмем за приближенное решение этой задачи. Оценка (19) станет более конструктивной, если в формулах (18), определяющих величины $S_{\min}$, $s_{\min}$, $K_{0\min}$, множества $D(x, t, T)$, $X(t, T)$ заменим на $V(t)$, $G(t)$ соответственно.

Опираясь на оценку (19), нетрудно сформулировать достаточные условия оптимальности для задачи (12)—(16).

Теорема 3. *Для того чтобы допустимая пара* $(u_*(t), x_*(t))$ $\left(t_0^* \leqslant t \leqslant T^*\right)$ *задачи* (12)—(16) *была решением этой задачи, достаточно существования функции* $K(x, t)$, *для которой формула* (1) *верна для любой допустимой пары задачи* (12)—(16) *и*

$$\int_{t_0^*}^{T^*} S(x_*(t), u_*(t), t)\, dt = S_{\min}, \quad (20)$$

$$s(x_*(T^*), T^*) = s_{\min}, \qquad K\left(x_*\left(t_0^*\right), t_0^*\right) = K_{0\min}. \quad (21)$$

Доказательство. При выполнении условий (20), (21) правая часть оценки (19) обращается в нуль, откуда и следует утверждение теоремы.

Теорема 4. *Для того чтобы некоторая последовательность* $(u_m(t), x_m(t))$ $(t_{0m} \leqslant t \leqslant T_m,\ m = 1, 2, \ldots)$ *допустимых пар задачи* (12)—(16) *была минимизирующей для этой задачи, достаточно существования функции* $K(x, t)$, *для которой формула* (1) *верна для любой допустимой пары задачи* (12)—(16) *и*

$$\lim_{m\to\infty} \int_{t_{0m}}^{T_m} S(x_m(t), u_m(t), t)\, dt = S_{\min}, \quad (22)$$

$$\lim_{m\to\infty} s(x_m(T_m), T_m) = s_{\min}, \qquad \lim_{m\to\infty} K(x_m(t_{0m}), t_{0m}) = K_{0\min}. \quad (23)$$

Доказательство. В оценке (19) вместо $u_*(t)$, $x_*(t)$, t_0^*, T^* подставим соответственно $u_m(t)$, $x_m(t)$, t_{0m}, T_m и перейдем к пределу при $m \to \infty$. С учетом условий (22), (23) получим утверждение теоремы.

Предлагаем читателю самостоятельно выписать условия, аналогичные условиям (10), (11), для определения функции Кротова $K(x, t)$ для задачи (12)—(16).

Пример 3. Требуется наибыстрейшим образом перевести точку $(x, y) \in E^2$ из начала координат (0, 0) в точку (1, 0), предполагая, что

движение точки подчиняется условиям $\dot{x}(t) = -y^2(t) + u^2(t)$, $\dot{y}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1: |u| \leqslant 1\}$ $(0 \leqslant t \leqslant T)$.

Здесь $G(0) = \{(0, 0)\}$, $G(t) \equiv E^2$ при $0 < t \leqslant T$, $S_1(T) = \{(1, 0)\}$, $\theta_0 = \{t_0 = 0\}$, $\theta_1 = \{T: T \geqslant 0\}$, $f^0 \equiv 1$, $\Phi \equiv 0$, $J(T, u(t)) = T$.

Пусть T_m — корень уравнения $(T-1)^3 = 12(T-1) - m^{-2}$, расположенный в пределах $1 < T < 1 + m^{-2/3}$ $(m = 1, 2, \ldots)$. Положим

$$u_m(t) = \begin{cases} 1 \text{ при } \dfrac{p}{m} < t \leqslant \dfrac{p}{m} + \dfrac{1}{2m} \text{ и } 1 \leqslant t \leqslant \dfrac{T_m + 1}{2}, \\ -1 \text{ при } \dfrac{p}{m} + \dfrac{1}{2m} < t \leqslant \dfrac{p}{m} + \dfrac{1}{m} \text{ и } \dfrac{T_m + 1}{2} < t \leqslant T_m, \end{cases}$$

$p = 0, 1, \ldots, m-1$, $m = 1, 2, \ldots$ Через $(x_m(t), y_m(t))$ $(0 \leqslant t \leqslant T_m)$ обозначим траекторию точки, соответствующую управлению $u_m(\cdot)$ и начальным условиям $x_m(0) = y_m(0) = 0$. Нетрудно видеть, что $x_m(T_m) = 1$, $y_m(T_m) = 0$, $(1 - m^{-4/3})t \leqslant x_m(t) \leqslant t$, $0 \leqslant y_m(t) \leqslant m^{-2/3}$ при $0 \leqslant t \leqslant T_m$, $m = 1, 2, \ldots$

Покажем, что $\lim\limits_{m\to\infty} T_m = 1 = T^*$ — оптимальное время. Для этого возьмем функцию $K(x, y, t) = -x$. Тогда

$$S(x, y, u, t) = y^2 - u^2 + 1, \quad \inf_{(x,y)\in E^2} \inf_{|u|\leqslant 1} S(x, y, u, t) \equiv 0$$

при всех $t \geqslant 0$, $S_{\min} = 0$, $\int\limits_0^{T_m} S(x_m(t), y_m(t), u_m(t), t)\,dt = \int\limits_0^{T_m} y_m^2(t)\,dt - \int\limits_0^{T_m} (u_m^2(t) - 1)\,dt = \int\limits_0^{T_m} y_m^2(t)\,dt \to 0 = S_{\min}$ при $m \to \infty$; $s(x, y, T) = -K(x, y, T) = -x = -1$ при $(x, y) \in S_1(T)$, $s(x_m(T_m), y_m(T_m), T_m) = -1 = s_{\min}$, $K(x, y, 0) = 0$ при $(x, y) \in G(0)$, $K(x_m(0), y_m(0), 0) = 0 = K_{0\min}$ $(m = 1, 2, \ldots)$.

Кроме того, очевидно, формула (1) для данной задачи будет справедлива при всех допустимых $(x(\cdot), y(\cdot), u(\cdot))$. Таким образом, для последовательностей $\{T_m\}$, $\{(x_m(t), y_m(t), u_m(t))\}$ $(0 \leqslant t \leqslant T_m)$ все условия теоремы 4 выполнены. Следовательно, $\lim\limits_{m\to\infty} J(T_m, u_m(\cdot)) = \lim\limits_{m\to\infty} T_m = 1$ — оптимальное время. Остается заметить, что в рассмотренной задаче $\inf J(T, u) = 1$ не достигается — здесь, как и в примере 2, мы имеем дело со скользящим режимом.

3. Приведем еще одно достаточное условие оптимальности, касающееся задачи быстродействия.

Теорема 5. *Пусть в задаче* (12)—(16) $f^0 \equiv 1$, $\Phi \equiv 0$; $\theta_0 = \{t_0\}$, *т. е. начальный момент* t_0 *закреплен*; $\theta_1 \equiv \{T: T \geqslant t_0\}$. *Пусть имеется некоторая последовательность* $(u_m(t), x_m(t))$ $(t_0 \leqslant t \leqslant T_m,\ m = 1, 2, \ldots)$ *допустимых пар рассматриваемой задачи быстродействия, причем* $\lim\limits_{m\to\infty} T_m = T^*$. *Тогда для того, чтобы* T^* *было оптимальным временем, достаточно существования функции* $K(x, t)$, *для которой формула* (1) *верна для любой допустимой пары* u, *кроме того,*

$$\lim_{m\to\infty} \int\limits_{t_0}^{T_m} S(x_m(t), u_m(t), t)\,dt < \int\limits_t^T S_{\min}(t, T)\,dt + T^* - T \tag{24}$$

для любого T $(t_0 \leqslant T < T^*)$,

$$\lim_{m\to\infty} s(x_m(T_m), T_m) = s_{\min}, \quad \lim_{m\to\infty} K(x_m(t_0), t_0) = K_{0\min}, \tag{25}$$

где

$$S_{\min}(t, T) = \inf_{x \in X(t,T)} \inf_{u \in D(x,t,T)} S(x, u, t),$$

$$s_{\min} = \inf_{0 \leqslant T \leqslant T^*} \inf_{x \in S_1(T)} s(x, T), \quad K_{0\min} = \inf_{t_0 \leqslant T \leqslant T^*} \inf_{x \in X(t_0, T)} K(x, t_0).$$

Для получения формулировки достаточного условия оптимальности для фиксированной допустимой пары $(u_*(t), x_*(t))$ $(t_0 \leqslant t \leqslant T^*)$ в этой теореме надо принять $T_m = T^*$, $u_m(t) = u_*(t)$, $x_m(t) = x_*(t)$ $(m = 1, 2, \ldots)$ и в (24), (25) всюду опустить знак lim.

Доказательство. Пусть вопреки утверждению теоремы T^* не является оптимальным временем. Тогда существуют момент T $(t_0 \leqslant T < T^*)$ и допустимая пара $(u(t), x(t))$ $(t_0 \leqslant t \leqslant T)$. В формуле (17) вместо t_0^*, T^*, $(u_*(t), x_*(t))$ $(t_0^* \leqslant t \leqslant T^*)$ примем соответственно t_0, T_m, $(u_m(t), x_m(t))$ $(t_0 \leqslant t \leqslant T_m)$ и перейдем к пределу при $m \to \infty$. С учетом условий (24), (25) будем иметь

$$\lim_{m \to \infty} J(t_0, T_m, x_m(t_0), u_m(\cdot)) - J(t_0, T, x(t_0), u(\cdot)) = T^* - T \leqslant$$

$$\leqslant \lim_{m \to \infty} \int_{t_0}^{T_m} S(x_m(t), u_m(t), t)\, dt - \int_{t_0}^{T} S(x(t), u(t), t)\, dt < T^* - T.$$

Полученное противоречивое неравенство доказывает теорему.

Пример 4. Пусть требуется наибыстрейшим образом перевести точку $(x, y) \in E^2$ из положения $(1, 0)$ в начало координат $(0, 0)$, предполагая, что движение точки подчиняется условиям $\dot{x}(t) = y(t)$, $\dot{y}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1 \colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant T)$.

Здесь $G(0) = \{(1, 0)\}$, $G(t) \equiv E^2$ при $0 \leqslant t < T$, $S_1(T) = \{(0, 0)\}$, $\theta_0 = \{t_0 = 0\}$, $\theta_1 = \{T \geqslant 0\}$.

В примере 6.2.4 с помощью принципа максимума была найдена допустимая пара $((x_*(t), y_*(t)), u_*(t)$ $(0 \leqslant t \leqslant T^* = 2)$, где

$$u_*(t) = \begin{cases} -1, & 0 \leqslant t \leqslant 1, \\ 1, & 1 < t \leqslant 2, \end{cases} \quad x_*(t) = \begin{cases} 1 - t^2/2, & 0 \leqslant t \leqslant 1, \\ (t-2)^2/2, & 1 \leqslant t \leqslant 2, \end{cases}$$

$$y_*(t) = \begin{cases} -t, & 0 \leqslant t \leqslant 1, \\ t - 2, & 1 \leqslant t \leqslant 2. \end{cases}$$

Покажем, что эта пара является решением рассматриваемой задачи быстродействия. Возьмем функцию $K(x, y, t) = x - (t-1)y$. Тогда $S(x, y, u, t) = K_x y + K_y u + K_t + 1 = -(t-1)u + 1$, $S_{\min}(t, T) = \inf_{(x,y) \in E^2} \inf_{|u| \leqslant 1} S(y, u, t) = -|t-1| + 1$ при всех T и $T \geqslant 0$, $S(x_*(t), y_*(t), u_*(t), t) = -|t-1| + 1$; $\int_0^{T^*} S(x_*(t), y_*(t), u_*(t), t)\, dt - \int_0^T S_{\min}(t, T)\, dt = \int_T^2 (1 - |t-1|)\, dt < 2 - T = T^* - T$ для всех T $(0 \leqslant T < T^* = 2)$; $s(x, y, T) = -K(x, y, T) = 0$ при $(x, y) \in S_1(T)$, $s(x_*(T^*), y_*(T^*), T^*) = s_{\min}$; $K(x, y, 0) = 1$ при $(x, y) \in G(0) = X(0, T) = \{(1, 0)\}$, $K(x_*(0), y_*(0), 0) = 1 = K_{0\min}$.

Кроме того, очевидно, формула (1) для данной задачи будет справедлива при всех допустимых управлениях и траекториях. В силу теоремы 5 момент $T^* = 2$ и пара $((x_*(\cdot), y_*(\cdot)), u_*(\cdot))$ являются оптимальными. Заметим, что функция Беллмана в этой задаче имеет разрывы первой производной именно на оптимальной траектории [17, 65], в то время как функция Кротова $K(x, t)$ является просто многочленом.

В заключение упомянем, что функции Беллмана, Кротова тесно связаны с функцией Ляпунова, широко используемой в теории устойчивости [9].

Упражнения. 1. Рассмотреть задачу минимизации функции $J(u(\cdot)) = \int_0^T (u^2(t) - x^2(t))\,dt$ при условиях $x(0) = x(T) = 0$. Показать, что пара $(u_*(t) \equiv 0,\ x_*(t) \equiv 0)$ является оптимальной при $0 < T < \pi$. Указание: функцию Кротова искать в виде $K(x, t) = \psi(t)x^2$.

2. С помощью принципа максимума найти подозрительные на оптимальность управления и траектории, а затем доказать их оптимальность для следующей задачи быстродействия: наибыстрейшим образом перевести точку (x_0, y_0) из заданного состояния в начало координат $(0, 0)$, предполагая, что движение точки подчиняется одному из следующих условий:

а) $\dot{x}(t) = y(t)$, $\dot{y}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1\colon |u| \leqslant 1\}$, $0 \leqslant t \leqslant T$;

б) $\dot{x}(t) = y(t)$, $\dot{y}(t) = -x(t) + u(t)$, $u(t) \in V(t) = \{u \in E^1\colon |u| \leqslant 1\}$, $0 \leqslant t \leqslant T$;

в) $\dot{x}(t) = y(t) + u(t)$, $\dot{y}(t) = -x(t) + v(t)$, $(u(t), v(t)) \in V(t) = \{(u, v) \in E^2\colon |u| \leqslant 1, |v| \leqslant 1\}$, $0 \leqslant t \leqslant T$. Указание: функцию Кротова искать в виде $K(x, t) = \psi_1(t)x + \psi_2(t)y$.

3. Перевести точку $(x, y, z) \in E^3$ из начала координат $(0, 0, 0)$ в точку $(a, 0, 0)$ быстрейшим образом, если $\dot{x}(t) = y(t)$, $\dot{y}(t) = z(t)$, $\dot{z}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1\colon |u| \leqslant 1\}$ $(0 \leqslant t \leqslant T)$; $a = \text{const}$. Показать, что оптимальное время $T^* = (32|a|)^{1/3}$. Указание: функцию Кротова искать в виде $K(x, t) = \psi_1(t)x + \psi_2(t)y + \psi_3(t)z$.

4. Рассмотреть задачу минимизации функции

$$J(t_0, T, x_0, u(\cdot)) = \int_{t_0}^{T} f^0(x(t), u(t), t)\,dt + \Phi_0(x(t_0), t_0) + \Phi(x(T), T)$$

при условиях (13)—(16). Для этой задачи сформулировать и доказать теоремы, аналогичные теоремам 1—4.

СПИСОК ЛИТЕРАТУРЫ

ОСНОВНАЯ ЛИТЕРАТУРА

1. Алексеев В. М., Галеев Э. М., Тихомиров В. М. Сборник задач по оптимизации. Теория. Примеры. Задачи.— М.: Наука, 1984.— 288 с.
2. Алексеев В. М., Тихомиров В. М., Фомин С. В. Оптимальное управление.— М.: Наука, 1979.— 432 с.
3. Ашманов С. А. Линейное программирование.— М.: Наука, 1981.— 304 с.
4. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы.— М.: Наука, 1987.— 600 с.
5. Бублик Б. Н., Кириченко Н. Ф. Основы теории управления.— Киев: Вища школа, 1975.— 328 с.
6. Васильев Ф. П. Методы решения экстремальных задач.— М.: Наука, 1981.— 400 с.
7. Габасов Р., Кириллова Ф. М. Методы оптимизации.— Минск: Изд-во БГУ, 1981.— 352 с.
8. Евтушенко Ю. Г. Методы решения экстремальных задач и их применение в системах оптимизации.— М.: Наука, 1982.— 432 с.
9. Зубов В. И. Лекции по теории управления.— М.: Наука, 1975.— 496 с.
10. Ильин В. А., Садовничий В. А., Сендов Бл. Х. Математический анализ. Начальный курс.— М.: Изд-во МГУ, 1985.— 660 с.
11. Карманов В. Г. Математическое программирование.— М.: Наука, 1986.— 288 с.
12. Ляшенко И. Н., Карагодова Е. А., Черникова Н. В., Шор Н. З. Линейное и нелинейное программирование.— Киев: Вища школа, 1975.— 372 с.
13. Марчук Г. И. Методы вычислительной математики.— М.: Наука, 1980.— 536 с.
14. Моисеев Н. Н. Элементы теории оптимальных систем.— М.: Наука, 1975.— 528 с.
15. Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М. Методы оптимизации.—М.: Наука, 1978.— 352 с.
16. Морозов В. В., Сухарев А. Г., Федоров В. В. Исследование операций в задачах и упражнениях.— М.: Высшая школа, 1986.— 287 с.
17. Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая теория оптимальных процессов.— М.: Наука, 1976.— 392 с.
18. Пшеничный Б. Н. Выпуклый анализ и экстремальные задачи.— М.: Наука, 1980.— 320 с.
19. Пшеничный Б. Н., Данилин Ю. М. Численные методы в экстремальных задачах.— М.: Наука, 1975.— 320 с.

20. Самарский А. А., Николаев Е. С. Методы решения сеточных уравнений.— М.: Наука, 1978.— 592 с.
21. Сухарев А. Г., Тимохов А. В., Федоров В. В. Курс методов оптимизации.— М.: Наука, 1986.— 328 с.
22. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач.— М.: Наука, 1986.— 288 с.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

23. Абрамов Л. М., Капустин В. Ф. Математическое программирование.— Л.: Изд-во ЛГУ, 1981.— 328 с.
24. Аваков Е. Р. Условия экстремума для гладких задач с ограничениями типа равенств // Журн. вычислит. матем. и матем. физики.— 1985.— Т. 25, № 5.— С. 680—693.
25. Акулич И. Л. Математическое программирование в примерах и задачах.— М.: Высшая школа, 1986.— 319 с.
26. Алексеев О. Г. Комплексное применение методов дискретной оптимизации.— М.: Наука, 1987.— 248 с.
27. Альбер Я. И., Шильман С. В. Метод обобщенного градиента: сходимость, устойчивость и оценки погрешности // Журн. вычисл. матем. и матем. физики.— 1982.— Т. 22, № 4.— С. 814—823.
28. Анрион Р. Теория второй вариации и ее приложения в оптимальном управлении.— М.: Наука, 1979.— 208 с.
29. Антипин А. С. Методы нелинейного программирования, основанные на прямой и двойственной модификации функции Лагранжа.— М.: Изд-во Всесоюзного научно-исследовательского института системных исследований, 1979.— 74 с.
30. Аоки М. Введение в методы оптимизации.— М.: Наука, 1977.— 344 с.
31. Арутюнов А. В. О необходимых условиях оптимальности в задаче с фазовыми ограничениями // ДАН СССР.— 1985.— Т. 280, № 5.— С. 1033—1037.
32. Арутюнов А. В., Марданов М. Дж. К теории оптимальных процессов с запаздываниями // Дифференциальные уравнения.— 1986.— Т. 22, № 8.— С. 1291—1298.
33. Астафьев Н. Н. Линейные неравенства и выпуклость.— М.: Наука, 1982.— 152 с.
34. Ахиев С. С. О необходимых условиях оптимальности для систем функционально-дифференциальных уравнений // ДАН СССР.— 1979.— Т. 247, № 1.— С. 11—14.
35. Ашманов С. А. Введение в математическую экономику.— М.: Наука, 1984.— 296 с.
36. Ащепков Л. Т. Оптимальное управление линейными системами.— Иркутск: Изд-во ИГУ, 1982.— 116 с.
37. Ащепков Л. Т. Оптимальное управление разрывными системами.— Новосибирск: Наука, 1987.— 226 с.
38. Ащепков Л. Т., Белов Б. И., Булатов В. П., Васильев О. В., Срочко В. А., Тарасенко Н. В. Методы решения задач математического программирования и оптимального управления.— Новосибирск: Наука, 1984.— 234 с.
39. Бабенко К. И. Основы численного анализа.— М.: Наука, 1986.— 744 с.
40. Багриновский К. А., Бусыгин В. П. Математика плановых решений.— М.: Наука, 1980.— 224 с.
41. Базара М., Шетти К. Нелинейное программирование. Теория и алгоритмы.— М.: Мир, 1982.— 584 с.
42. Бакушинский А. Б. Итерационные регуляризующие алгоритмы для нелинейных задач // Журн. вычисл. матем. и матем. физики.— 1987.— Т. 27, № 4.— С. 617—621.

43. Баничук Н. В. Оптимизация форм упругих тел.— М.: Наука, 1980.— 256 с.
44. Банк Б., Белоусов Е. Г., Мандель Р., Черемных Ю. Н., Широнин В. М. Математическая оптимизация: вопросы разрешимости и устойчивости.— М.: Изд-во МГУ, 1986.— 216 с.
45. Бартиш М. Я. Об одном классе методов типа Ньютона // Вестник МГУ. Сер. вычислит. матем. и киберн.— 1987, № 2.— С. 16—20.
46. Батищев Д. И. Методы оптимального проектирования.— М.: Радио и связь, 1984.— 248 с.
47. Батухтин В. Д., Майборода Л. А. Оптимизация разрывных функций.— М.: Наука, 1984.— 208 с.
48. Бейко И. В., Бублик Б. Н., Зинько П. Н. Методы и алгоритмы решения задач оптимизации.— Киев: Вища школа, 1983.— 512 с.
49. Беленький В. З., Волконский В. А., Иванков С. А., Поманский А. Б., Шапиро А. Д. Итеративные методы в теории игр и программировании.— М.: Наука, 1974.— 240 с.
50. Беллман Р. Процессы регулирования с адаптацией.— М.: Наука,— 1964.— 360 с.
51. Белолипецкий А. А., Рябов А. Ю. Асимптотические оценки решений задачи оптимального быстродействия вблизи точек излома изохронной поверхности // Журн. вычисл. матем. и матем. физики.— 1986.— Т. 26, № 4.— С. 521—535.
52. Белоусов Е. Г. Введение в выпуклый анализ и целочисленное программирование.— М.: Изд-во МГУ, 1977.— 196 с.
53. Бердышев В. И. Непрерывность многозначного отображения, связанного с задачей минимизации функционала // Изв. АН СССР. Сер. матем.— 1980.— Т. 44, № 3.— С. 483—509.
54. Березин И. С., Жидков Н. П. Методы вычислений. Том I.— М.: Наука, 1966.— 632 с. Том 2.— М.: Физматгиз, 1962.— 640 с.
55. Березнев В. А. Математические методы планирования производственной программы предприятий легкой промышленности.— М.: Легкая индустрия, 1980.— 144 с.
56. Березнев В. А., Карманов В. Г., Третьяков А. А. О стабилизирующих свойствах градиентного метода // Журн. вычислит. матем. и матем. физики.— 1986.— Т. 26, № 1.— С. 134—137.
57. Береснев В. Л., Гимади Э. Х., Дементьев В. Т. Экстремальные задачи стандартизации.— Новосибирск: Наука, 1978.— 334 с.
58. Бертсекас Д. Условная оптимизация и методы множителей Лагранжа.— М.: Радио и связь, 1987.— 400 с.
59. Бертсекас Д., Шрив С. Стохастическое оптимальное управление: случай дискретного времени.— М.: Наука, 1985.— 280 с.
60. Бистрицкас В. Б. Приближенное решение уравнений динамического программирования // Журн. вычисл. матем. и матем. физики.— 1985.— Т. 25, № 8.— С. 1131—1142.
61. Благодатских В. И. Линейная теория оптимального управления.— М.: Изд-во МГУ, 1978.— 96 с.
62. Благодатских В. И. Принцип максимума для дифференциальных включений // Тр. Мат. ин-та АН СССР.— 1984.— Т. 166.— С. 23—43.
63. Благодатских В. И., Филиппов А. Ф. Дифференциальные включения и оптимальное управление // Тр. Мат. ин-та АН СССР.— 1985.— Т. 169.— С. 194—252.
64. Блисс Г. А. Лекции по вариационному исчислению.— М.: Изд-во иностр. литературы, 1950.— 348 с.
65. Болтянский В. Г. Математические методы оптимального управления.— М.: Наука, 1969.— 408 с.
66. Болтянский В. Г. Оптимальное управление дискретными системами.— М.: Наука, 1973.— 448 с.

67. Болтянский В. Г. Метод шатров в топологических векторных пространствах // ДАН СССР.— 1986.— Т. 289, № 5.— С. 1036—1039.
68. Брайсон А., Хо Ю-Ши. Прикладная теория оптимального управления.— М.: Мир, 1972.— 544 с.
69. Бублик Б. Н., Гаращенко Ф. Г., Кириченко Н. Ф. Структурно-параметрическая оптимизация и устойчивость динамики пучков.— Киев: Наукова думка, 1985.— 304 с.
70. Будак Б. М., Васильев Ф. П. Некоторые вычислительные аспекты задач оптимального управления.— М.: Изд-во МГУ, 1975.— 172 с.
71. Булавский В. А., Звягина Р. А., Яковлева М. А. Численные методы линейного программирования.— М.: Наука, 1977.— 368 с.
72. Булатов В. П. Методы погружения в задачах оптимизации.— Новосибирск: Наука, 1977.— 160 с.
73. Бурдаков О. П. Устойчивые варианты метода секущих для решения систем уравнений // Журн. вычисл. матем. и матем. физики.— 1983.— Т. 23, № 5.— С. 1027—1040.
74. Буслаев В. С. Вариационное исчисление.— Л.: Изд-во ЛГУ, 1980.— 288 с.
75. Бутковский А. Г. Фазовые портреты управляемых динамических систем.— М.: Наука, 1985.— 136 с.
76. Вайникко Г. М., Веретенников А. Ю. Итерационные процедуры в некорректных задачах.— М.: Наука, 1986.— 182 с.
77. Варга Дж. Оптимальное управление дифференциальными и функциональными уравнениями.— М.: Наука, 1977.— 624 с.
78. Васильев Н. С. О численном решении экстремальных задач построения эллипсоидов и параллелепипедов // Журн. вычислит. матем. и матем. физики.— 1987.— Т. 27, № 3.— С. 340—348.
79. Васильев О. В. Методы оптимизации в конечномерных пространствах.— Иркутск: Изд-во Иркутск. ун-та, 1979.— 90 с.
80. Васильев О. В. Методы оптимизации в функциональных пространствах.— Иркутск: Изд-во Иркутск. ун-та, 1979.— 120 с.
81. Васильев Ф. П. Лекции по методам решения экстремальных задач.— М.: Изд-во МГУ, 1974.— 376 с.
82. Васильев Ф. П. О методе нагруженных функционалов // Вестник МГУ. Сер. вычисл. матем. и киберн., 1978.— № 3.— С. 24—32.
83. Васильев Ф. П. Численные методы решения экстремальных задач.— М.: Наука, 1980.— 520 с.
84. Васильев Ф. П. Применение негладких штрафных функций в методе регуляризации неустойчивых задач минимизации // Журн. вычислит. матем. и матем. физики.— 1987.— Т. 27, № 10.— С. 1444—1450.
85. Васильев Ф. П., Константинова Т. В. Об одном обобщении метода нагруженных функционалов // Вестник МГУ. Сер. вычисл. матем. и киберн.— 1983, № 2.— С. 3—8.
86. Васильев Ф. П., Солодкая М. С., Ячимович М. Д. О регуляризованном методе линеаризации при наличии погрешностей в исходных данных // Вестник МГУ. Сер. вычисл. матем. и киберн.— 1985, № 4.— С. 3—8.
87. Васильев Ф. П., Хромова Л. Н., Ячимович М. Д. Итеративная регуляризация одного метода минимизации третьего порядка // Вестник МГУ. Сер. вычисл. матем. и киберн.— 1981, № 1.— С. 31—36.
88. Васин А. А. Модели процессов с несколькими участниками.— М.: Изд-во МГУ, 1983.— 84 с.
89. Васин В. В. Дискретная аппроксимация и устойчивость в экстремальных задачах // Журн. вычисл. матем. и матем. физики.— 1982.— Т. 22, № 4.— С. 824—839.

90. Вилков А. В., Жидков Н. П., Щедрин Б. М. Метод отыскания глобального минимума функции одного переменного // Журн. вычислит. матем. и матем. физики.— 1975.— Т. 15, № 4.— С. 1040—1042.
91. Вилков В. Б. Некоторые свойства функции Лагранжа в задачах математического программирования.— Кибернетика, 1986, № 1.— С. 65—69.
92. Владимиров А. А., Нестеров Ю. Е., Чеканов Ю. Н. О равномерно выпуклых функционалах // Вестник МГУ. Сер. вычисл. матем. и киберн.— 1978, № 3.— С. 12—23.
93. Воеводин В. В. Линейная алгебра.— М.: Наука, 1980.— 400 с.
94. Волошин А. Ф. Метод локализации области оптимума в задачах математического программирования // ДАН СССР.— 1987.— Т. 293, № 3.— С. 549—553.
95. Воробьев Н. Н. Числа Фибоначчи.— М.: Наука, 1978.— 144 с.
96. Воробьев Н. Н. Теория игр. Лекции для экономистов-кибернетиков.— Л.: Изд-во ЛГУ, 1985.— 268 с.
97. Габасов Р. Ф., Кириллова Ф. М. Качественная теория оптимальных процессов.— М.: Наука, 1971.— 508 с.
98. Габасов Р. Ф., Кириллова Ф. М. Особые оптимальные управления.— М.: Наука, 1973.— 256 с.
99. Габасов Р. Ф., Кириллова Ф. М. Оптимизация линейных систем.— Минск: Изд-во БГУ, 1973.— 248 с.
100. Габасов Р. Ф., Кириллова Ф. М. Принцип максимума в теории оптимального управления.— Минск: Наука и техника, 1974.— 272 с.
101. Габасов Р. Ф., Кириллова Ф. М. Основы динамического программирования.— Минск: Изд-во БГУ, 1975.— 264 с.
102. Габасов Р. Ф., Кириллова Ф. М. Методы линейного программирования.— Минск: Изд-во БГУ. Часть 1, 1977.— 176 с., часть 2, 1978.— 240 с., часть 3, 1980.— 368 с.
103. Гавурин М. К., Малоземов В. Н. Экстремальные задачи с линейными ограничениями.— Л.: Изд-во ЛГУ, 1984.— 176 с.
104. Гамкрелидзе Р. В. Основы оптимального управления.— Тбилиси: Изд-во Тбилисского ун-та, 1977.— 254 с.
105. Гантмахер Ф. Р. Теория матриц.— М.: Наука, 1967.— 576 с.
106. Ганшин Г. С. Методы оптимизации и решение уравнений.— М.: Наука, 1987.— 128 с.
107. Гапоненко Ю. Л. Метод последовательной аппроксимации для решения нелинейных экстремальных задач // Известия вузов. Сер. матем.— 1980, № 5.— С. 12—15.
108. Гельфанд И. М., Фомин С. В. Вариационное исчисление.— М.: Физматгиз, 1961.— 228 с.
109. Гермейер Ю. Б. Введение в теорию исследования операций.— М.: Наука, 1971.— 384 с.
110. Гермейер Ю. Б. Игры с непротивоположными интересами.— М.: Наука, 1976.— 328 с.
111. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация.— М.: Мир, 1985.— 509 с.
112. Гилязов С. Ф. Методы решения линейных некорректных задач.— М.: Изд-во МГУ, 1987.— 120 с.
113. Гладков Д. И. Оптимизация систем неградиентным случайным поиском.— М.: Энергоатомиздат, 1984.— 256 с.
114. Гнеденко Б. В. Математика — народному хозяйству.— Новое в жизни, науке, технике. Сер. Математика, кибернетика.— М.: Знание, 1977, № 10.— 64 с.
115. Голиков А. И., Жадан В. Г. Две модификации метода линеаризации в нелинейном программировании // Журн. вычислит. матем. и матем. физики.— 1983.— Т. 23, № 2.— С. 314—325.

116. Гольштейн Е. Г., Третьяков Н. В. Модифицированные функции Лагранжа и их применение // Экономика и матем. методы.— 1983.— Т. 19, № 3.— С. 528—547.
117. Горбунов В. К. Методы редукции неустойчивых вычислительных задач.— Фрунзе: Илим, 1984.— 134 с.
118. Горелик В. А., Кононенко А. Ф. Теоретико-игровые модели принятия решений в эколого-экономических системах.— М.: Радио и связь, 1982.— 145 с.
119. Гребенников А. И. Метод сплайнов и решение некорректных задач теории приближений.— М.: Изд-во МГУ, 1983.— 208 с.
120. Григоренко Н. Л. Дифференциальные игры преследования несколькими объектами.— М.: Изд-во МГУ, 1983.— 79 с.
121. Гродзовский Г. Л., Иванов Ю. Н., Токарев В. В. Механика космического полета с малой тягой.— М.: Наука, 1966.— 680 с.
122. Гроссман К., Каплан А. А. Нелинейное программирование на основе безусловной минимизации.— Новосибирск: Наука, 1981.— 184 с.
123. Гупал А. М. Стохастические методы решения негладких экстремальных задач.— Киев: Наукова думка, 1979.— 152 с.
124. Гурман В. И. Вырожденные задачи оптимального управления.— М.: Наука, 1977.— 304 с.
125. Гурман В. И. Принцип расширения в задачах управления.— М.: Наука, 1985.— 288 с.
126. Давыдов Э. Г. Методы и модели теории антагонистических игр.— М.: Изд-во МГУ, 1978.— 208 с.
127. Дамбраускас А. П. Симплексный поиск.— М.: Энергия, 1979.— 176 с.
128. Данилин Ю. М. Оценка эффективности одного алгоритма отыскания абсолютного минимума // Журн. вычисл. матем. и матем. физики.— 1971.— Т. 11, № 4.— С. 1026—1031.
129. Данилин Ю. М., Ковнир В. Н. Об одной точной штрафной функции для задачи нелинейного программирования.— Кибернетика, 1986, № 5.— С. 43—46.
130. Данциг Дж. Линейное программирование, его применения и обобщения.— М.: Прогресс, 1966.— 600 с.
131. Даффин Р., Питерсон Э., Зенер К. Геометрическое программирование.— М.: Мир, 1972.— 312 с.
132. Демьянов В. Ф., Васильев Л. В. Недифференцируемая оптимизация.— М.: Наука, 1981.— 384 с.
133. Демьянов В. Ф., Малоземов В. Н. Введение в минимакс.— М.: Наука, 1972.— 368 с.
134. Денисов Д. В., Карманов В. Г., Третьяков А. А. Ускоренный метод Ньютона для решения функциональных уравнений // ДАН СССР.— 1985.— Т. 281, № 6.— 1293—1297.
135. Дикин И. И., Зоркальцев В. И. Итеративное решение задач математического программирования.— Новосибирск: Наука, 1980.— 144 с.
136. Дончев А. Системы оптимального управления. Возмущения, приближения и анализ чувствительности.— М.: Мир, 1987.— 156 с.
137. Дубовицкий А. Я., Милютин А. А. Необходимые условия слабого экстремума в общей задаче оптимального управления.— М.: Наука, 1971.— 114 с.
138. Дюркович Е. Численный метод нахождения времени быстродействия с заданной точностью // Журн. вычисл. матем. и матем. физики.— 1983.— Т. 23, № 1.— С. 51—60.
139. Евтушенко Ю. Г. Численный метод поиска глобального экстремума функций (перебор на неравномерной сетке) // Журн. вычисл. матем. и матем. физики.— 1971.— Т. 11, № 6.— С. 1390—1703.

140. Евтушенко Ю. Г., Жадан В. Г. Об одном подходе к систематизации численных методов нелинейного программирования.— Техническая кибернетика, 1983, 4 1.— С. 47—59.
141. Евтушенко Ю. Г., Ратькин В. А. Метод половинных делений для глобальной оптимизации функции многих переменных.— Техническая кибернетика, 1987, № 1.— С. 119—127.
142. Егоров А. И. Оптимальное управление тепловыми и диффузионными процессами.— М.: Наука, 1978.— 464 с.
143. Емеличев В. А., Комлик В. И. Метод построения последовательности планов для решения задач дискретной оптимизации.— М.: Наука, 1981.— 208 с.
144. Еремин И. И., Астафьев Н. Н. Введение в теорию линейного и выпуклого программирования.— М.: Наука, 1976.— 192 с.
145. Еремин И. И., Мазуров В. Д. Нестационарные процессы математического программирования.— М.: Наука, 1979.— 288 с.
146. Еремин И. И., Мазуров В. Д., Астафьев Н. Н. Несобственные задачи линейного и выпуклого программирования.— М.: Наука, 1983.— 336 с.
147. Ермаков С. М., Жиглявский А. А. Математическая теория оптимального эксперимента.— М.: Наука, 1987.— 320 с.
148. Ермольев Ю. М. Методы стохастического программирования.— М.: Наука, 1976.— 240 с.
149. Ермольев Ю. М., Гуленко В. П., Царенко Т. И. Конечно-разностный метод в задачах оптимального управления.— Киев: Наукова думка, 1978.— 164 с.
150. Ермольев Ю. М., Ляшко И. И., Михалевич В. С., Тюптя В. И. Математические методы исследования операций.— Киев: Вища школа, 1979.— 312 с.
151. Ермольев Ю. М., Ястремский А. И. Стохастические модели и методы в экономическом планировании.— М.: Наука, 1979.— 254 с.
152. Жадан В. Г. Об одном классе итеративных методов решения задач выпуклого программирования // Журн. вычисл. матем. и матем. физики.— 1984.— Т. 24, № 5.— С. 665—676.
153. Жданов В. А. О методе покоординатного спуска // Мат. заметки.— 1977.— Т. 22, вып. 1.— С. 137—142.
154. Жиглявский А. А. Математическая теория глобального случайного поиска.— Л.: Изд-во ЛГУ, 1985.— 296 с.
155. Заботин Я. И. Лекции по линейному программированию.— Казань: Изд-во Казанск. ун-та, 1985.— 98 с.
156. Заботин Я. И., Кораблев А. И., Хабибуллин Р. Ф. Условия экстремума функционала при наличии ограничений.— Кибернетика, 1973, № 6. С. 65—70.
157. Заботин Я. И., Крейнин М. И. К сходимости методов отыскания минимакса // Известия вузов. Сер. матем.— 1977.— № 10 (185). С. 56—64.
158. Завриев С. К. Стохастические градиентные методы решения минимаксных задач.— М.: Изд-во МГУ, 1984.— 82 с.
159. Зангвилл У. И. Нелинейное программирование.— М.: Советское радио, 1973.— 312 с.
160. Зорич В. А. Математический анализ.— М.: Наука, ч. I, 1981.— 543 с., ч. II, 1984.— 640 с.
161. Иванов В. А., Фалдин Н. В. Теория оптимальных систем автоматического управления.— М.: Наука, 1981.— 336 с.
162. Иванов В. К., Васин В. В., Танана В. П. Теория линейных некорректных задач и ее приложения.— М.: Наука, 1978.— 208 с.
163. Ижуткин В. С., Кокурин М. Ю. О гибридном методе нелинейного программирования, использующем криволинейный спуск // Известия вузов. Сер. матем.— 1986, № 2.— С. 61—64.

164. Ильин В. А., Позняк Э. Г. Линейная алгебра.— М.: Наука, 1974, 296 с.
165. Ильин В. А., Позняк Э. Г. Основы математического анализа.— М.: Наука, ч. I, 1971.— 600 с., ч. II, 1973.— 448 с.
166. Иоффе А. Д., Тихомиров В. М. Теория экстремальных задач.— М.: Наука, 1974.— 480 с.
167. Казимиров В. И., Плотников В. И., Старобинец И. М. Абстрактная схема метода вариаций и необходимые условия экстремума // Изв. АН СССР. Сер. матем.— 1985.— Т. 49, № 1.— С. 141—159.
168. Калихман И. Л. Сборник задач по математическому программированию.— М.: Высшая школа, 1975.— 270 с.
169. Калихман И. Л., Войтенко М. А. Динамическое программирование в примерах и задачах.— М.: Высшая школа, 1979.— 126 с.
170. Капустин В. Ф. Практические занятия по курсу математического программирования.— Л.: Изд-во ЛГУ, 1976.— 192 с.
171. Карманов В. Г., Третьяков А. А. Оценка скорости сходимости некоторых методов покоординатного спуска // Вестн. МГУ. Сер. 15. Вычисл. матем. и киберн.— 1985, № 2.— С. 41—46.
172. Карташев А. П., Рождественский Б. Л. Обыкновенные дифференциальные уравнения и основы вариационного исчисления.— М.: Наука, 1986.— 287 с.
173. Катковник В. Я. Линейные оценки и стохастические задачи оптимизации.— М.: Наука, 1976.— 488 с.
174. Кирин Н. Е. Методы последовательных оценок в задачах оптимизации управляемых систем.— Л.: Изд-во ЛГУ, 1975.— 160 с.
175. Киселев Ю. Н. Линейная теория быстродействия с возмущениями.— М.: Изд-во МГУ, 1986.— 106 с.
176. Ковалев М. М. Дискретная оптимизация.— Минск: Изд-во БГУ, 1977.— 191 с.
177. Ковач М. Непрерывный аналог итеративной регуляризации градиентного типа // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1979.— № 3.— С. 36—42.
178. Ковач М. О сходимости метода обобщенных барьерных функций // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1981.— № 1.— С. 40—45.
179. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа.— М.: Наука, 1976.— 544 с.
180. Коростелев А. П. Стохастические рекуррентные процедуры (локальные свойства).— М.: Наука, 1984.— 208 с.
181. Коша А. Вариационное исчисление.— М.: Высшая школа, 1983.— 279 с.
182. Краснов М. Л., Макаренко Г. И., Киселев А. И. Вариационное исчисление. Задачи и упражнения.— М.: Наука, 1973.— 192 с.
183. Краснощеков П. С., Петров А. А. Принципы построения моделей.— М.: Изд-во МГУ, 1983.— 264 с.
184. Красовский Н. Н. Теория управления движением.— М.; Наука, 1968.— 476 с.
185. Красовский Н. Н. Игровые задачи о встрече движений.— М.: Наука, 1970.— 420 с.
186. Красовский Н. Н. Управление динамической системой. Задача о минимуме гарантированного результата.— М.: Наука, 1985.— 520 с.
187. Красовский Н. Н., Субботин А. И. Позиционные дифференциальные игры.— М.: Наука, 1974.— 456 с.
188. Кротов В. Ф., Букреев В. З., Гурман В. И. Новые методы вариационного исчисления в динамике полета.— М.: Машиностроение, 1969.— 288 с.
189. Кротов В. Ф., Гурман В. И. Методы и задачи оптимального управления.— М.: Наука, 1973.— 448 с.

190. Кузнецов Ю. Н., Кузубов В. И., Волощенко А. Б. Математическое программирование.— М.: Высшая школа, 1980.— 302 с.
191. Кукушкин Н. С., Морозов В. В. Теория неантагонистических игр.— М.: Изд-во МГУ, 1984.— 104 с.
192. Куржанский А. Б. Управление и наблюдение в условиях неопределенности.— М.: Наука, 1977.— 392 с.
193. Лагунов В. Н. Введение в дифференциальные игры.— Вильнюс: Институт матем. и кибернетики АН Литовской ССР, 1979.— 342 с.
194. Лебедев В. Ю. Декомпозиционный метод решения блочных задач линейного программирования со связывающими переменными // Журн. вычисл. матем. и матем. физики.— 1981.— Т. 21, № 4.— С. 881—886.
195. Левин В. Л. Выпуклый анализ в пространствах измеримых функций и его применение в математике и экономике.— М.: Наука, 1985.— 352 с.
196. Левитин Е. С. К теории возмущений негладких экстремальных задач с ограничениями // ДАН СССР.— 1975.— Т. 224, № 6.— С. 1260—1263.
197. Лейхтвейс К. Выпуклые множества.— М.; Наука, 1985.— 336 с.
198. Леонов А. С. О применении обобщенного принципа невязки для решения некорректных экстремальных задач // ДАН СССР.— 1982.— Т. 262, № 6.— С. 1306—1310.
199. Ли Э. Б., Маркус Л. Основы теории оптимального управления.— М.: Наука, 1972.— 576 с.
200. Лисковец О. А. Вариационные методы решения неустойчивых задач.— Минск: Изд-во Наука и техника, 1981.— 344 с.
201. Лоран П.-Ж. Аппроксимизация и оптимизация.— М.: Мир, 1975.— 496 с.
202. Лотов А. В. Введение в экономико-математическое моделирование.— М.: Наука, 1984.— 392 с.
203. Лэсдон Л. С. Оптимизация больших систем.— М.: Наука, 1975.— 432 с.
204. Любушин А. А., Черноусько Ф. Л. Метод последовательных приближений для решения задач оптимального управления // Изв. АН СССР. Сер. технич. киберн.— 1983.— № 2.— С. 147—159.
205. Мансимов К. Б. Многоточечные необходимые условия оптимальности особых в классическом смысле управлений в системах с запаздыванием // Дифференц. уравнения.— 1985.— Т. 21, № 3.— С. 527—530.
206. Марданов М. Д. Некоторые вопросы математической теории оптимальных процессов в системах с запаздываниями.— Баку: Изд-во Азерб. ун-та, 1987.— 120 с.
207. Марчук Г. И. Математическое моделирование в проблеме окружающей среды.— М.: Наука, 1982.— 320 с.
208. Марчук Г. И. Окружающая среда и проблемы оптимизации // Тр. МИАН СССР. Т. 166.— М.: Наука, 1984.— С. 123—129.
209. Марчук Г. И., Лебедев В. И. Численные методы в теории переноса нейтронов.— М.: Атомиздат, 1981.— 454 с.
210. Матвеев А. С. О необходимых условиях экстремума в задаче оптимального управления с фазовыми ограничениями // Дифференц. управления.— 1987.— Т. 23, № 4.— С. 629—640.
211. Мееров М. В. Исследование и оптимизация многосвязных систем управления.— М.: Наука, 1986.— 236 с.
212. Мезенцев А. В. Сборник задач по теории оптимального управления.— М.: Изд-во МГУ, 1980.— 48 с.
213. Михалевич В. С., Волкович В. Л. Вычислительные методы исследования и проектирования сложных систем.— М.: Наука, 1982.— 286 с.

214. Михалевич В. С., Гупал А. М., Норкин В. И. Методы невыпуклой оптимизации.— М.: Наука, 1987.— 280 с.
215. Михалевич В. С., Кукса А. И. Методы последовательной оптимизации в дискретных сетевых задачах оптимального распределения ресурсов.— М.: Наука, 1983.— 208 с.
216. Михалевич В. С., Трубин В. А., Шор Н. З. Оптимизационные задачи производственно-транспортного планирования: модели, методы, алгоритмы.— М.: Наука, 1986.— 259 с.
217. Моисеев Н. Н. Численные методы в теории оптимальных систем.— М.: Наука, 1971.— 424 с.
218. Моисеев Н. Н. Математические задачи системного анализа.— М.: Наука, 1981.— 488 с.
219. Мордухович Б. Ш. Методы аппроксимаций в задачах оптимизации и управления.— М.: Наука, 1988.— 360 с.
220. Мороз А. И. Курс теории систем.— М.: Высшая школа, 1987.— 304 с.
221. Морозов В. А. Регулярные методы решения некорректно поставленных задач.— М.: Изд-во МГУ, 1974.— 360 с.
222. Морозов С. Ф., Сумин М. И. Об одном классе задач управления динамическими системами с разрывной правой частью.— Кибернетика, 1985, № 3.— С. 59—71.
223. Москаленко А. И. Методы нелинейных отображений в оптимальном управлении.— Новосибирск: Наука, 1983.— 223 с.
224. Муртаф Б. Современное линейное программирование.— М.: Мир, 1984.— 224 с.
225. Мухачева Э. А., Рубинштейн Г. Ш. Математическое программирование.— Новосибирск: Наука, 1977.— 320 с.
226. Наконечный А. Г. Минимальное оценивание функционалов от решений вариационных уравнений в гильбертовых пространствах.— Киев: Изд-во Киевск. ун-та, 1985.— 84 с.
227. Немировский А. С., Нестеров Ю. Е. Оптимальные методы гладкой выпуклой минимизации // Журн. вычисл. матем. и матем. физики.— 1985.— Т. 25, № 3.— С. 356—369.
228. Немировский А. С., Юдин Д. Б. Сложность задач и эффективность методов оптимизации.— М.: Наука, 1979.— 384 с.
229. Нестеров Ю. Е. Об одном классе методов безусловной минимизации выпуклой функции, обладающих высокой скоростью сходимости // Журн. вычисл. матем. и матем. физики.— 1984.— Т. 24, № 7.— С. 1090—1093.
230. Нефедов В. Н. Методы регуляризации многокритериальных задач оптимизации.— М.: Изд-во Московск. авиацион. ин-та, 1984.— 56 с.
231. Нефедов В. Н. Отыскание глобального максимума функции нескольких переменных на множестве, заданном ограничениями типа неравенств // Журн. вычисл. матем. и матем. физики.— 1987.— Т. 27, № 1.— С. 35—51.
232. Никольский С. М. Первый прямой метод Л. С. Понтрягина в дифференциальных играх.— М.: Изд-во МГУ, 1984.— 64 с.
233. Никольский С. М. Курс математического анализа.— М.: Наука, 1973.— Т. 1.— 432 с. Т. 2.— 392 с.
234. Ногин В. Д., Протодьяконов И. О., Евлампиев И. И. Основы теории оптимизации.— М.: Высшая школа, 1986.— 384 с.
235. Нурминский Е. А. Численные методы решения детерминированных и стохастических минимаксных задач.— Киев: Наукова думка, 1979.— 158 с.
236. Орлов М. В. О некоторых численных методах решения линейной задачи быстродействия // Вестник МГУ. Сер. вычисл. матем. и кибернетики.— 1986.— № 4.— С. 41—46.
237. Орловский С. А. Проблемы принятия решений при нечетной исходной информации.— М.: Наука, 1981.— 208 с.

238. Ортега Д., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными.— М.: Мир, 1975.— 560 с.
239. Островский А. М. Решение уравнений и систем уравнений.— М.: Изд-во иностр. литературы, 1963.— 220 с.
240. Панин В. М. Методы конечных штрафов с линейной аппроксимацией ограничений // Кибернетика.— 1984.— Ч. 1, № 2.— С. 44—50.— Ч. 2, № 4.— С. 73—81.
241. Певный А. Б. Об оптимальных стратегиях поиска максимума функции с ограниченной старшей производной // Журн. вычисл. матем. и матем. физики.— 1982.— Т. 22, № 5.— С. 1061—1066.
242. Первозванский А. А., Гайцгори В. Г. Декомпозиция, агрегирование и приближенная оптимизация.— М.: Наука, 1979.— 344 с.
243. Петров Ю. П. Вариационные методы теории оптимального управления.— Ленинград: Энергия, 1977.— 288 с.
244. Петросян Л. А., Томский Г. В. Динамические игры и их приложения.— Л.: Изд-во ЛГУ, 1982.— 252 с.
245. Пионтковский О. В. О минимизации нелинейных функционалов в нормированных пространствах.— Успехи матем. наук, 1974.— Т. 29, № 3.— С. 225—226.
246. Пиявский С. А. Один алгоритм отыскания абсолютного экстремума функции // Журн. вычисл. матем. и матем. физики.— 1972.— Т. 12, № 4.— С. 888—896.
247. Подиновский В. В., Ногин В. Д. Парето-оптимальные решения многокритериальных задач.— М.: Наука, 1982.— 256 с.
248. Полак Э. Численные методы оптимизации. Единый подход.— М.: Мир, 1974.— 376 с.
249. Половинкин Е. С., Смирнов Г. В. О задаче быстродействия для дифференциальных включений // Дифференц. уравнения, 1986.— Т. 22, № 8.— С. 1351—1365.
250. Поляк Б. Т. Введение в оптимизацию.— М.: Наука, 1983.— 384 с.
251. Понтрягин Л. С. Обыкновенные дифференциальные уравнения.— М.: Наука, 1983.— 332 с.
252. Понтрягин Л. С. Математическая теория оптимальных процессов и дифференциальные игры // Тр. МИАН СССР. Т. 169.— М.: Наука, 1985.— С. 119—158.
253. Потапов М. М. Об аппроксимации задач оптимизации с гладкими допустимыми управлениями при наличии ограничений // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1983.— № 4.— С. 3—7.
254. Пропой А. И. Элементы теории оптимальных дискретных процессов.— М.: Наука, 1973.— 256 с.
255. Пшеничный Б. Н. Необходимые условия экстремума.— М.: Наука, 1982.— 144 с.
256. Пшеничный Б. Н. Метод линеаризации.— М.: Наука, 1983.— 136 с.
257. Разумихин Б. С. Физические модели и методы теории равновесия в программировании и экономике.— М.: Наука, 1975.— 304 с.
258. Ракитский Ю. В., Устинов С. М., Черноруцкий И. Г. Численные методы решения жестких систем.— М.: Наука, 1979.— 208 с.
259. Растригин Л. А. Системы экстремального управления.— М.: Наука, 1974.— 632 с.
260. Раушенбах Б. В., Токарь Е. Н. Управление ориентацией космических аппаратов.— М.: Наука, 1974.— 600 с.
261. Рейклейтис Г., Рейвиндран А., Рэгсдел К. Оптимизация в технике. В двух книгах.— М.: Мир, 1986, книга 1 — 350 с., книга 2 — 320 с.

262. Рихтер К. Динамические задачи дискретной оптимизации.— М.: Радио и связь, 1985.— 136 с.
263. Ройтенберг Я. Н. Автоматическое управление.— М.: Наука, 1978.— 552 с.
264. Рокафеллар Р. Выпуклый анализ.— М.: Мир, 1973.— 472 с.
265. Романовский И. В. Алгоритмы решения экстремальных задач.— М.: Наука, 1977.— 352 с.
266. Рубальский Г. Б. Поиск экстремума унимодальной функции одной переменной на неограниченном множестве // Журн. вычисл. матем. и матем. физики.— 1982.— Т. 22, № 1.— С. 10—16.
267. Саати Т. Целочисленные методы оптимизации и связанные с ними экстремальные проблемы.— М.: Мир, 1973.— 302 с.
268. Самсонов С. П. Восстановление выпуклого множества по его опорной функции с заданной точностью // Вестн. МГУ. Сер. вычисл. матем. и кибернетики.— 1983.— № 1.— С. 68—71.
269. Саульев В. К., Самойлова И. И. Приближенные методы безусловной оптимизации функций многих переменных // Сб. работ ВИНИТИ: Матем. анализ. Сер. итоги науки и техники.— М.: ВИНИТИ АН СССР.— 1973.— Т. 11.— С. 91—128.
270. Сачков В. Н. Введение в комбинаторные методы дискретной математики.— М.: Наука, 1982.— 384 с.
271. Сеа Ж. Оптимизация. Теория и алгоритмы.— М.: Мир, 1973.— 244 с.
272. Сергиенко И. В., Лебедева Т. Т., Рощин В. А. Приближенные методы решения дискретных задач оптимизации.— Киев: Наукова думка, 1980.— 274 с.
273. Силин Д. Б. Линейные задачи оптимального быстродействия с разрывными на множестве положительной меры управлениями.— Матем. сборник, 1986.— Т. 129(171), № 2.— С. 264—278.
274. Скарин В. Д. Об одном подходе к анализу несобственных задач линейного программирования // Журн. вычисл. матем. и матем. физики.— 1986.—Т. 26, № 3.— С. 439—448.
275. Срочко В. А. Вычислительные методы оптимального управления.— Иркутск: Изд-во Иркутск. ун-та, 1982.— 110 с.
276. Старосельский Л. А., Шелудько Г. А., Кантор Б. Я. Об одной реализации метода оврагов с адаптацией величины овражного шага по экспоненциальному закону // Журн. вычисл. матем. и матем. физики.— 1968.— Т. 8, № 5.— С. 1161—1167.
277. Старостенко В. И. Устойчивые численные методы в задачах гравиметрии.— Киев: Наукова думка, 1978.— 228 с.
278. Стрекаловский А. С. К проблеме глобального экстремума // ДАН СССР.— 1987.— Т. 292, № 5.— С. 1062—1066.
279. Стронгин Р. Г. Численные методы в многоэкстремальных задачах.— М.: Наука, 1978.— 240 с.
280. Субботин А. И., Ченцов А. Г. Оптимизация гарантии в задачах управления.— М.: Наука, 1981.— 288 с.
281. Сумин М. И. Оптимальное управление системами с приближенно известными исходными данными //Журн. вычислит. матем. и матем. физики.— 1987.— Т. 27, № 2.— С. 163—177.
282. Сухарев А. Г. Оптимальный поиск экстремума.— М.: Изд-во МГУ, 1975.— 100 с.
283. Сухинин М. Ф. Правило множителей Лагранжа в локально выпуклых пространствах // Сибирск. матем. журн.— 1982.— Т. 23, № 4.— С. 153—165.
284. Сухинин М. Ф. О двух вариантах градиентного метода // Журн. вычисл. матем. и матем. физики.— 1984.— Т. 24, № 8.— С. 1265—1267.
285. Сухинин М. Ф. Об одном аналоге уравнения Беллмана // Мат. заметки.— 1985.— Т. 38, № 2.— С. 265—269.

286. Тадумадзе Т. А. Некоторые вопросы качественной теории оптимального управления.— Тбилиси: Изд-во Тбилисского ун-та, 1983.— 128 с.
287. Танаев В. С., Шкурба В. В. Введение в теорию расписаний.— М.: Наука, 1975.— 256 с.
288. Тарасова В. П. Оптимальный поиск экстремума для класса локально унимодальных функций.— Кибернетика, 1984.— № 1.— С. 65—68.
289. Телеснин В. Р. Об одной задаче оптимизации переходных процессов // Тр. Мат. ин-та АН СССР.— М.: Наука, 1984.— Т. 166.— С. 235—244.
290. Тер-Крикоров А. М. Оптимальное управление и математическая экономика.— М.: Наука, 1977.— 216 с.
291. Тетерев А. Г. Анализ сходимости и устойчивости методов одномерной оптимизации // Вестн. МГУ. Серия вычисл. матем. и кибернетики.— 1981.— № 4.— С. 21—27.
292. Тимохов А. В. Математические модели экономического воспроизводства.— М.: Изд-во МГУ, 1982.— 128 с.
293. Тихомиров В. М. Некоторые вопросы теории приближений.— М.: Изд-во МГУ, 1976.— 304 с.
294. Тихомиров В. М. Рассказы о максимумах и минимумах.— М.: Наука, 1986.— 192 с.
295. Тихонов А. Н., Васильева А. Б., Свешников А. Г. Дифференциальные уравнения.— М.: Наука, 1985.— 232 с.
296. Тихонов А. Н., Гончарский А. В., Степанов В. В., Ягола А. Г. Регуляризующие алгоритмы и априорная информация.—М.: Наука, 1983.— 200 с.
297. Тихонов А. Н., Рютин А. А., Агаян Г. М. Об устойчивом методе решения задачи линейного программирования с приближенными данными // Докл. АН СССР.— 1983.— Т. 272, № 5.— С. 1058—1063.
298. Трауб Дж., Вожьняковский Х. Общая теория оптимальных алгоритмов.— М.: Мир, 1983.— 384 с.
299. Третьяков А. А. Необходимые и достаточные условия оптимальности p-го порядка // Журн. вычисл. матем. и матем. физики.— 1984.— Т. 24, № 2.— С. 203—209.
300. Троицкий В. А., Петухов Л. В. Оптимизация формы упругих тел.— М.: Наука, 1982.— 432 с.
301. Ульм С. Ю. Обобщение метода Стеффенсена для решения нелинейных операторных уравнений // Журн. вычисл. матем. и матем. физики, 1964.— Т. 4, № 6.— С. 1093—1097.
302. Ульм С. Ю. Методы декомпозиции для решения задач оптимизации.— Таллин: Изд-во Валгус, 1979.— 132 с.
303. Уонэм М. Линейные многомерные системы управления.— М.: Наука, 1980.— 376 с.
304. Уткин В. И. Скользящие режимы в задачах оптимизации и управления.— М.: Наука, 1981.— 368 с.
305. Федоренко Р. П. Приближенное решение задач оптимального управления.— М.: Наука, 1978.— 488 с.
306. Федоров В. В. Численные методы максимина.— М.: Наука, 1979.— 280 с.
307. Фиакко А., Мак-Кормик Г. Нелинейное программирование. Методы последовательной безусловной минимизации.— М.: Мир, 1972.— 240 с.
308. Флеминг У., Ришел Р. Оптимальное управление детерминированными и стохастическими системами.— М.: Мир, 1978.— 318 с.
309. Формальский А. М. Управляемость и устойчивость систем с ограниченными ресурсами.— М.: Наука, 1974.— 368 с.

310. Фурасов В. Д. Устойчивость движения, оценки и стабилизация.— М.: Наука, 1977.— 248 с.
311. Харатишвили Г. Л., Мачаидзе З. А., Маркозашвили Н. И., Тадумадзе Т. А. Абстрактная вариационная теория и ее применения к оптимальным задачам с запаздываниями.— Тбилиси: Мецниереба, 1973.— 112 с.
312. Харди Г. Г., Литтльвуд Дж. Е., Полиа Г. Неравенства.— М.: Изд-во иностран. лит-ры, 1948.— 456 с.
313. Хачиян Л. Г. Полиномиальные алгоритмы в линейном программировании // Журн. вычисл. матем. и матем. физики. 1986.— Т. 20, № 1.— С. 51—68.
314. Химмельблау Д. Прикладное нелинейное программирование.— М.: Мир, 1975.— 536 с.
315. Хоменюк В. В. Оптимальные системы управления.— М.: Наука, 1977.— 150 с.
316. Хромова Л. Н. Об одном методе минимизации с кубической скоростью сходимости // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1980.— № 3.— С. 52—56.
317. Хрусталев М. М. Необходимые и достаточные условия оптимальности в форме уравнения Беллмана // Докл. АН СССР.— 1978.— Т. 242, № 5.— С. 1023—1026.
318. Ху Т. Целочисленное программирование и потоки в сетях.— М.: Мир, 1974.— 520 с.
319. Цирлин А. М., Балакирев В. С., Дудников Е. Г. Вариационные методы оптимизации управляемых объектов.— М.: Энергия; 1976.— 448 с.
320. Цурков В. И. Декомпозиция в задачах большой размерности.— М.: Наука, 1981.— 352 с.
321. Чарин В. С. Линейные преобразования и выпуклые множества.— Киев: Вища школа, 1978.— 192 с.
322. Черемных Ю. Н. Анализ поведения траекторий динамики народнохозяйственных моделей.— М.: Наука, 1982.— 177 с.
323. Черемных Ю. Н. Математические модели развития народного хозяйства.— М.: Изд-во МГУ, 1986.— 104 с.
324. Черников С. Н. Линейные неравенства.— М.: Наука, 1968.— 488 с.
325. Черноусько Ф. Л., Акуленко Л. Д., Соколов Б. Н. Управление колебаниями.— М.: Наука, 1980.— 384 с.
326. Черноусько Ф. Л., Баничук Н. В. Вариационные задачи механики и управления.— М.: Наука, 1973.— 238 с.
327. Черноусько Ф. Л., Колмановский В. Б. Оптимальное управление при случайных возмущениях.— М.: Наука, 1978.— 352 с.
328. Черноусько Ф. Л., Меликян А. А. Игровые задачи управления и поиска.— М.: Наука, 1978.— 270 с.
329. Чирич Н. Т. О регуляризованном методе линеаризации для минимизации выпуклой функции на многогранном множестве при наличии погрешностей в исходных данных // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1987.— № 2.— С. 20—25.
330. Численные методы условной оптимизации // Сб. работ под ред. Гилл Ф., Мюррэй У.— М.: Мир, 1977.— 292 с.
331. Чичинадзе В. К. Решение невыпуклых нелинейных задач оптимизации.— М.: Наука, 1983.— 256 с.
332. Чуян О. Р. Оптимальный одношаговый алгоритм максимизации дважды дифференцируемых функций // Журн. вычисл. матем. и матем. физики.— 1986.— Т. 26, № 3.— С. 381—397.
333. Швартин С. М. Общая задача устойчивости для некоторых классов задач линейного программирования. // ДАН СССР.— 1985.— Т. 285, № 1.— С. 56—59.

334. Шепилов М. А. О методе обобщенного градиента для экстремальных задач // Журн. вычисл. матем. и матем. физики.— 1976.— Т. 16, № 1.— С. 242—247.
335. Шепилов М. А. Об отыскании корней и глобального экстремума липшицевой функции.— Кибернетика, 1987.— № 2.— С. 71—74.
336. Шор Н. З. Методы минимизации недифференцируемых функций и их приложения.— Киев: Наукова думка, 1979.— 200 с.
337. Экланд И., Темам Р. Выпуклый анализ и вариационные проблемы.— М.: Мир, 1979.— 400 с.
338. Эльстер К.-Х., Рейнгардт Р., Шойбле М., Донат Г. Введение в нелинейное программирование.— М.: Наука, 1985.— 264 с.
339. Юдин Д. Б. Задачи и методы стохастического программирования.— М.: Советское радио, 1979.— 392 с.
340. Юдин Д. Б., Гольштейн Е. Г. Линейное программирование. Теория, методы и приложения.— М.: Наука, 1969.— 424 с.
341. Якубович В. А. К абстрактной теории оптимального управления // Сибирск. матем. журн.— I, 1977, 18, № 3.— С. 685—707; II, 1978.— 19, № 2.— С. 436—460.
342. Янг Л. Лекции по вариационному исчислению и теории оптимального управления.— М.: Мир, 1974.— 488 с.
343. Ячимович М. Итеративная регуляризация одного варианта метода условного градиента // Вестн. МГУ. Сер. вычисл. матем. и киберн.— 1980.— № 4.— С. 13—19.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Антициклин 124

Базис угловой точки 111
Базисные координаты 111
— переменные 111

Вектор опорный 198
— собственно опорный 198
Верхний предел последовательности 71
— — функции 78
Верхняя грань функции 13
Выпуклая комбинация точек 156

Гиперплоскость 149
— опорная 198
— отделяющая 194
— собственно опорная 198
Градиент 79

Двойственные переменные 248

Задача быстродействия 434
— двойственная 248
— классического вариационного исчисления 485
— Коши 425
— минимизации второго типа 11, 70
— — первого типа 11, 70
— многоэкстремальная 347
— на безусловный экстремум 82
— на условный экстремум 82
— оптимального управления 433
— — — автономная 434
— — — с закрепленным временем 432, 435
— — — с закрепленным концом 432, 441, 442
— — — со свободным концом 432, 442, 443
Задача оптимального управления с подвижным концом 432, 443, 444
— — — с фазовыми ограничениями 431
— регулярная 84, 225
— с сильно согласованной постановкой 371
— с согласованной постановкой 370
Замыкание множества 153
Зацикливание 124
Золотое сечение отрезка 19

Квадратичная форма неотрицательная 168
— — отрицательно определенная 80
— — положительно определенная 80
Конус 204
— выпуклый 204
— двойственный (сопряженный) 204
— замкнутый 204
— открытый 204
Координата базисная 111
— отмеченная 141
— фазовая 425
Коэффициент барьерный 385
— штрафной 366
Краевая задача принципа максимума 440
Критерий выпуклости функции 39, 43, 44, 164, 165, 167
— оптимальности 42, 165, 173, 192, 210, 234—247
— сильной выпуклости функции 184, 185

Лексикографически положительный вектор 135
Лексикографическое правило 135
Линейного программирования задача вырожденная 123
— — — каноническая 105
— — — невырожденная 123

Линейного программирования задача общая 101
— — — общая 101
— — — основная 105
Луч 149

Метод барьерных функций 384
— блуждающих трубок 510
— возможных направлений 299
— градиентный 261
— Давидона — Флетчера — Пауэлла 337
— декомпозиции 504
— деления отрезка пополам 17
— золотого сечения 19
— касательных 45
— квазиньютоновский 337
— классический 15, 78
— линеаризации 309
— локальных вариаций 510
— ломаных 28
— модифицированных функций Лагранжа 356
— нагруженных функций 396
— Ньютона 329
— овражный 269
— оптимальный 23
— парабол 59
— пассивный 24, 350
— — оптимальный 25
— переменной метрики 338
— поиска глобального минимума 28, 33, 53, 62, 347
— покоординатного спуска 342
— покрытий 33, 348
— последовательный 24, 350
— — оптимальный 27
— проекции градиента 277
— — субградиента 285
— равномерного перебора 24, 33
— симметричный 21
— скорейшего спуска 262
— случайного поиска 410
— — — без обучения 412
— — — с обучением 412
— сопряженных градиентов 328
— — направлений 320
— Стеффенсена 338
— стохастической аппроксимации 66, 415
— стрельбы 480
— тяжелого шарика 276
— условного градиента 291
— Фибоначчи 26
— штрафных функций 363
Минимальный корень уравнения 399
Множество аффинное 149
— выпуклое 148
— замкнутое 71
Множество компактное 71
— Лебега 73
— многогранное 152
— ограниченное 71
— открытое 153
— регулярное 238
Множитель Лагранжа 83, 224
Модуль выпуклости 218
— — точный 218
Момент времени конечный 427
— — — закрепленный 432
— — начальный 425, 427
— — — закрепленный 432

Надграфик (эпиграф) функции 171
Наибольшее (максимальное) значение функции 14
Наименьшее (минимальное) значение функции 9
Направление возможное 172
— — убывания 299
— рецессивное 177
Неравенство Гронуолла 461
— Йенсена 163
Нижний предел последовательности 71
— — функции 78
Нижняя грань функции 10
Нормальный вектор гиперплоскости 149

Оболочка аффинная 152
— выпуклая 157
Ограничения активные 224
— интегральные 434
— корректные 375
— пассивные 224
— типа неравенств 87
— — равенств 82
— точечные 434
— фазовые 431
Окрестность точки 71
Ортант неотрицательный 152
Отделимость множеств 193
— — сильная 194
— — собственная 194
— — строгая 194
Отображение многозначное 211
— — выпуклозначное 211
— — компактное 211
— — монотонное 211
— — полунепрерывное сверху 211
— — — снизу 211
— субдифференциальное 211
Отрезок локализации минимума 24

Параллелепипед 152
Подпространство несущее 152

Позином 256
Полупространство замкнутое 149
— открытое 149
Поляра 206
Последовательность максимизирующая 13
— минимизирующая 11
— ограниченная 71
Постоянная Липшица 28
— сильной выпуклости 181
Принцип максимума 438
Проблема синтеза 496, 513
Программирование выпуклое 234
— геометрическое 255
— динамическое 490
— квадратичное 314
— линейное 101
— полиномиальное 319
— стохастическое 415
Проекция точки на множество 188
Произведение множества на число 153
Производная по направлению 172
Прямая линия 149
Прямое произведение множеств 200

Размерность множества 151, 152
Разность множеств 153
Разрешающий элемент 118
Расстояние от точки до множества 11

Симплекс 113, 157
Симплекс-метод 112
Скользящий режим 525
Сопряженная система 436
Субградиент 206
Субдифференциал 207
Сумма множеств 153
Схема Беллмана 490
— Моисеева 505
Сходимость последовательности ко множеству 11

Теорема Вейерштрасса 12
— Куна — Таккера 235
— Фаркаша 240
Точка глобального (абсолютного) максимума 13
— — — минимума 12
— локального максимума 14
— — минимума 12
— множества внешняя 154
— — внутренняя 153
— — граничная 154
— — изолированная 154
Точка множества относительно внутренняя 160
— множества предельная 71
— — угловая 109
— — — вырожденная 111
— — — невырожденная 111
—, подозрительная на экстремум 15, 85, 88
— седловая 235
— стационарная 80
— строгого локального максимума 14
— — — минимума 12
— экстремума 14
Точность метода гарантированная 23
— — — наилучшая 23
Траектории левый конец 427
— — — закрепленный 432
— — — подвижный 432
— — — свободный 432
— правый конец 427
— — — закрепленный 432
— — — подвижный 432
— — — свободный 432
Траектория (решение) задачи Коши 427
— оптимальная 433

Управление 425
— оптимальное 433
— особое 451
Уравнение Беллмана 492
— Эйлера 487
Условие Вейерштрасса 487
— дополняющей нежесткости 224, 437
— достаточное оптимальности (максимума, минимума) 15, 80, 85, 165, 173, 192, 210, 237, 500, 522
— Лежандра 487
— необходимое оптимальности (максимума, минимума, экстремума) 15, 80, 83, 165, 173, 192, 210, 224, 239, 244, 246, 379, 437, 445
— Слейтера 238
— трансверсальности 437, 489
— Эрдмана — Вейерштрасса 488

Формула конечных приращений 92
Функция барьерная 385
— Беллмана 492
— Вейерштрасса 488
— вогнутая 163
— выпуклая 162
— Гамильтона — Понтрягина 436
— дважды дифференцируемая 79
— — непрерывно дифференцируемая (дважды гладкая) 91

Функция дифференцируемая 78
— квазивыпуклая 181
— Кротова 501, 523
— кусочно гладкая 425
— — непрерывная 425
— Лагранжа 83, 224
— — модифицированная 358
— — регулярная 235
— Ляпунова 276, 530
— Минковского 180
— непрерывно дифференцируемая (гладкая) 91
— овражная 268
— ограниченная 13
— — сверху 13
— — снизу 10
— опорная 180, 199
— полунепрерывная сверху 72
— — снизу 72
Функция равномерно выпуклая 218
— сильно выпуклая 181
— синтезирующая 496, 513
— строго вогнутая 163
— — выпуклая 162
— — равномерно выпуклая 218
— — унимодальная 13
—, удовлетворяющая условию Гельдера 377
—, — — Липшица 28
— унимодальная 13
— штрафная 364

Шар 148
Шкала состояний 505

Элементарная операция 506

Учебное издание

ВАСИЛЬЕВ Федор Павлович

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ
ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

Заведующий редакцией *Е. Ю. Ходан*
Редактор *И. В. Викторенкова*
Художественный редактор *Г. М. Коровина*
Технический редактор *В. Н. Кондакова*
Корректоры: *О. А. Бутусова, Т. С. Вайсберг*

ИБ № 12659

Сдано в набор 07.12.87. Подписано к печати 02.11.88. Формат 60×90/16. Бумага офсетная. Гарнитура обыкновенная новая. Печать высокая. Усл. печ. л. 34,5. Усл. кр.-отт. 34,5. Уч.-изд. л. 38,64. Тираж 19 500 экз. Заказ № 1219. Цена 1 р. 60 к.

Ордена Трудового Красного Знамени издательство «Наука»
Главная редакция физико-математической литературы
117071 Москва В-71, Ленинский проспект, 15

Четвертая типография издательства «Наука»
630077 г. Новосибирск-77, Станиславского, 25