

Ф. П. ВАСИЛЬЕВ

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

*Допущено Министерством высшего
и среднего специального образования СССР
в качестве учебного пособия
для студентов вузов, обучающихся
по специальности «Прикладная математика»*



МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1980

22.19

В 19

УДК 519.6

Численные методы решения экстремальных задач.
В а с и л ь е в Ф. П. — М.: Наука. Главная редакция физико-математической литературы, 1980.

Книга написана на основе лекций, которые автор в течение ряда лет читал на факультете вычислительной математики и кибернетики МГУ. В ней изложены наиболее часто используемые на практике методы решения задач минимизации функций и функционалов, а также теоретическое исследование и краткая характеристика вычислительных аспектов этих методов.

Учебное пособие рассчитано на студентов университетов и вузов, обучающихся по специальности «прикладная математика». Книга будет полезна также аспирантам, научным работникам и инженерам.

ОГЛАВЛЕНИЕ

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. Постановка задачи об оптимальных методах	23
§ 6. Оптимальные пассивные методы	28
§ 7. Оптимальный последовательный метод для задачи А	30
§ 8. Оптимальный последовательный метод для задачи Б	33
§ 9. Метод ломаных	37
§ 10. Методы пассивного и последовательного перебора	42
§ 11. Выпуклые функции одной переменной	45
§ 12. Метод касательных	53
§ 13. Метод поиска глобального минимума	61
§ 14. Метод парабол	66
§ 15. Другой метод поиска глобального минимума	69
§ 16. О методе стохастической аппроксимации	73
Глава 2. Предварительные сведения о конечномерных задачах на экстремум	76
§ 1. Постановка задачи минимизации. Теорема Вейерштрасса	76
§ 2. Классический метод	86
§ 3. Вспомогательные предложения	98
Глава 3. Элементы линейного программирования	109
§ 1. Постановка задачи	109
§ 2. Геометрическая интерпретация. Угловые точки	115
§ 3. Симплекс-метод	122
§ 4. Антициклин	136

ОГЛАВЛЕНИЕ

§ 5. Выбор начальной угловой точки	145
§ 6. Об условии разрешимости канонической задачи	155
 Глава 4. Элементы выпуклого анализа	 158
§ 1. Выпуклые множества	158
§ 2. Выпуклые функции	169
§ 3. Сильно выпуклые функции	185
§ 4. Проекция точки на множество	193
§ 5. Отделимость выпуклых множеств	198
§ 6. Субградиент. Субдифференциал	207
§ 7. Равномерно выпуклые функции	212
§ 8. Теорема Куна — Таккера. Двойственная задача	218
§ 9. Правило множителей Лагранжа	239
 Глава 5. Методы минимизации функций многих переменных	 259
§ 1. Градиентный метод	259
§ 2. Метод проекции градиента	273
§ 3. Метод проекции субградиента	281
§ 4. Метод условного градиента	288
§ 5. Метод возможных направлений	296
§ 6. Метод сопряженных направлений	307
§ 7. Метод Ньютона	316
§ 8. Метод с кубической скоростью сходимости	324
§ 9. Метод покоординатного спуска	329
§ 10. Метод поиска глобального минимума	335
§ 11. Метод множителей Лагранжа	343
§ 12. Метод штрафных функций	350
§ 13. Метод барьерных функций	369
§ 14. Метод нагруженных функций	384
§ 15. О методе случайного поиска	399
§ 16. Общие замечания	404
 Глава 6. Принцип максимума Понтрягина	 410
§ 1. Постановка задачи оптимального управления	410
§ 2. Формулировка принципа максимума	424
§ 3. Доказательство принципа максимума для задачи оптимального управления со свободным правым концом	418
§ 4. О методах решения краевой задачи принципа максимума	435
§ 5. Связь между принципом максимума и классическим вариационным исчислением	460

Глава 7. Динамическое программирование	464
§ 1. Схема Беллмана. Проблема синтеза для дискретных систем	474
§ 2. Схема Монсеева	479
§ 3. Проблема синтеза для систем с непрерывным временем	486
§ 4. Достаточные условия оптимальности	496
Основная литература	505
Дополнительная литература	505
Предметный указатель	516

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 наиболее быстрым 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 множество U принято называть экстремальными задачами. 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 управления процессами, описываемыми уравнениями с частными производными, некорректные экстремальные задачи, аппроксимация и устойчивость экстремальных задач, методы минимизации в функциональных пространствах. Этим разделам теории и методов 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 её 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

МЕТОДЫ МИНИМИЗАЦИИ ФУНКЦИЙ ОДНОЙ ПЕРЕМЕННОЙ

Изложение методов решения экстремальных задач начнем с задач минимизации функций одной переменной. Такие задачи возникают при изучении объектов, зависящих от одной скалярной переменной, когда требуется выбрать эту переменную наилучшим в том или ином смысле образом. Кроме того, как увидим в последующих главах, эти задачи входят как составная часть во многие итерационные методы решения задач минимизации функций многих переменных и других более сложных экстремальных задач.

С задачами минимизации функций одной переменной мы впервые сталкиваемся при изучении начальных глав математического анализа уже в средней школе и решаем их методами дифференциального исчисления. Может показаться, что задачи минимизации функций одной переменной относятся к достаточно простым и методы их решения хорошо разработаны и изучены. Однако это не совсем так. Оказалось, что 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 разработке новых методов минимизации функций одной переменной, продолжают появляться на страницах математических книг и журналов (см., например, [9, 26, 58, 61, 67, 76—78, 95, 104, 111, 145, 187, 215, 216, 219, 222, 236, 247]). Мы здесь остановимся на некоторых наиболее известных методах, достаточно хорошо проявивших себя на практике.

§ 1. Постановка задачи

Пусть $R^1 = \{u: -\infty < u < +\infty\}$ — числовая ось, U — некоторое множество из R^1 , $J(u)$ — функция, определенная на множестве U и принимающая во всех точках $u \in U$ конечные значения. 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 $J(u_*) \leq J(u)$ для всех $u \in U$; величину $J(u_*)$ называют *наименьшим* или *минимальным значением* $J(u)$ на U и обозначают $\min_{u \in U} J(u) = J(u_*)$. 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 $U = \{u: 1 \leq u \leq 2\}$ минимальное значение $J(u)$ равно нулю, множество U_* состоит из единственной точки $u_* = 1$. Если $U = \{u: 1/3 \leq u \leq 1\}$, то U_* содержит три точки: $1/3, 1/2, 1$; если $U = \{u: 0 < u \leq 1\}$, то $U_* = \{u: u = 1/n, n = 1, 2, \dots\}$ — счетное множество. В случае $U = \{u: 2 \leq u < +\infty\}$ 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, $v = u/k$ 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 $U = \{u: |u| \leq 1\}$ принимает свое наименьшее значение, равное нулю, во всех точках отрезка $U_* = \{u: 0 \leq u \leq 1\}$. Если $U = \{u: 1 \leq u \leq 2\}$, то U_* содержит одну точку $u_* = 1$; если $U = \{u: 1 < u \leq 2\}$, то $U_* = \emptyset$.

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

Пример 4. Пусть $J(u) = \ln u$, $U = \{u: 0 < u \leq 1\}$. Здесь $U_* = \emptyset$, так как во всех точках из U функция принимает конечные значения, а для последовательности $u_k = 1/k$, $k = 1, 2, \dots$, имеем $\lim_{k \rightarrow \infty} J(u_k) = -\infty$.

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

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

В тех случаях, когда $U_* = \emptyset$, 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) $J_* \leq J(u)$ 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 $J(u)$ неограничена снизу на U , то в качестве нижней грани $J(u)$ на U принимается $J_* = -\infty$. Нижнюю грань $J(u)$ на U обозначают через $\inf_{u \in U} J(u) = J_*$.

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

Если $U_* \neq \emptyset$, то, очевидно, нижняя грань $J(u)$ на U совпадает с наименьшим значением этой функции на U , т. е. $\inf_{u \in U} J(u) = \min_{u \in U} J(u)$. В этом случае говорят, что функция $J(u)$ на U достигает своей нижней грани. Подчеркнем, что $\inf_{u \in U} J(u) = J_*$ всегда существует, а $\min_{u \in U} J(u)$, 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 $\lim_{k \rightarrow \infty} J(u_k) = \inf_{u \in U} J(u) = J_*$.

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 $\lim_{k \rightarrow \infty} \rho(u_k, U) = 0$, где $\rho(u_k, U) = \inf_{u \in U} |u_k - u|$ — расстояние от точки u_k до множества U .

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

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

Теперь можем перейти к формулировке постановки задачи минимизации функции $J(u)$ на множестве U . Здесь будем различать задачи двух типов. К первому типу отнесем задачи, в которых требуется определить величину $J_* = \inf_{u \in U} J(u)$. 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_* пусто и требуется наряду с J_* найти какую-либо точку $u_* \in U_*$.

Заметим, что получить точное решение задачи первого или второго типа удастся лишь в редких случаях. Поэтому на прак-

тике 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 [16]. 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 $\mathbb{R}^1$, 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 [7, 8, 179]. 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. Напомним

Определение 6. Точка $v_* \in U$ называется *точкой локального минимума* функции $J(u)$ на множество U со значением $c = J(v_*)$, если существует число $\alpha > 0$ такое, что $J(v_*) \leq J(u)$ для всех $u \in U \cap \{u: |u - v_*| < \alpha\} = O_\alpha(v_*)$. 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 $u_5 < u \leq u_6$ и $u_8 \leq u \leq u_9$, реализуется нестрогий локальный минимум. Функция из примера 1 в точках $u_k = 1/k$, $k = \pm 1, \pm 2$, имеет строгий локальный минимум на $U = \mathbb{R}^1$, 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 *униmodalной* на отрезке $U = \{u: a \leq u \leq b\} = [a, b]$, если она непрерывна на $[a, b]$, и существуют числа α, β , $a \leq \alpha \leq \beta \leq b$, такие, что 1) $J(u)$ строго монотонно убывает при $a \leq u \leq \alpha$ (если $a < \alpha$); 2) $J(u)$ строго монотонно возрастает при $\beta \leq u \leq b$ (если $\beta < b$); 3) $J(u) = J_* = \inf_{u \in U} J(u)$ при $\alpha \leq u \leq \beta$, так что $U_* = [\alpha, \beta]$. Случай, когда один или два из отрезков $[a, \alpha]$, $[\alpha, \beta]$,

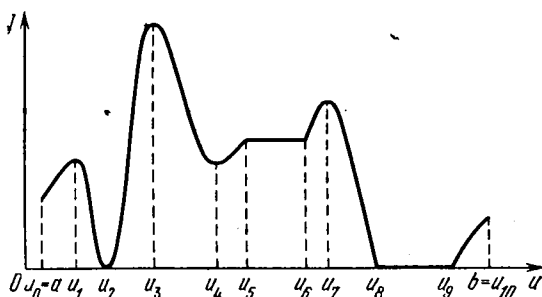


Рис. 1.1.

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

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

Нетрудно видеть, что если функция $J(u)$ униmodalна на $[a, b]$, то она остается униmodalной 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 число B такое, что $J(u) \leq B$ при всех $u \in U$. Функция $J(u)$ не ограничена сверху на U , если существует последовательность $\{u_k\} \in U$, для которой $\lim_{k \rightarrow \infty} J(u_k) = +\infty$. 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) $J(u) \leq J^*$ для всех $u \in U$; 2) для любого числа $\varepsilon > 0$ найдется точка $u_\varepsilon \in U$ такая, 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 $\lim_{k \rightarrow \infty} J(u_k) = J^*$. Если существует точка $u^* \in U$ такая, что $J(u^*) = J^*$, то u^* называется *точкой максимума* $J(u)$ 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 $J^* = \sup_{u \in U} J(u)$.

Заметим, что верхняя грань и максимизирующая последовательность всегда существуют, а максимальное значение может не существовать. Если выполнены условия теоремы 1, то $J^* < +\infty$, $U^* \neq \emptyset$, 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 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 число $\alpha > 0$ такое, что $J(v^*) \geq J(u)$ для всех $u \in U \cap \{u: |u - v^*| < \alpha\} = O_\alpha(v^*)$. 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 $u_5 \leq u < u_6$ 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.

У п р а ж н е н и я. 1. Построить минимизирующую и максимизирующую последовательности для функции $J(u) = \operatorname{arctg} u$ на $U = R^1$. Достигает ли эта функция своих нижних и верхних граней на R^1 ?

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

3. Найти все точки локального экстремума функции $J(u) = ||u^2 - 1| - 1|$ 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 $J(u) = \max_{0 \leq t \leq 1} |t^2 - ut|$ на множествах $U = R^1$ и $U = \{u: 1 \leq u < +\infty\}$.

§ 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. 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 производную. 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 полукрестности граничных точек $u = a$ или $u = b$) подозрительной точки. Для того чтобы подозрительная точка $u \in [a, b]$ была точкой локального минимума, достаточно, чтобы

$\lim_{u \rightarrow v-0} J(u) \geq J(v)$, $\lim_{u \rightarrow v+0} J(u) \geq J(v)$, и при некотором $\alpha > 0$ на множествах $[a, b] \cap \{v - \alpha < u < v + \alpha\} = O_\alpha^+(v)$, $[a, b] \cap \{v - \alpha < u < v\} = O_\alpha^-(v)$ 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 $\lim_{u \rightarrow v-0} J(u) \leq J(v)$, $\lim_{u \rightarrow v+0} J(u) \leq J(v)$ 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 удастся 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 $J'(v)$, $J''(v)$, ..., $J^{(n)}(v)$, причем $J^{(i)}(v) = 0$ при $i = 1, 2, \dots, n-1$, а $J^{(n)}(v) \neq 0$, $n \geq 1$. 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 $v = a$ (или $v = b$) 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 $U = \{u: a \leq u < +\infty\}$, или $U = \{u: -\infty < u \leq b\}$, или $U = \mathbb{R}^1$, то наряду с вышеописанными исследованиями нужно также изучить поведение функции при $u \rightarrow +\infty$ или $u \rightarrow -\infty$.

Классический метод исследования функции на экстремум следует использовать во всех тех случаях, когда достаточно просто удастся 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 затруднительно. Но даже в тех случаях, когда производную все же удастся 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 $\mathbb{R}^1$.

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 отрезке $[-1, 1]$. Исследовать случай, когда $J(u)$ на $[a, b]$ имеет конечное число точек локального экстремума.

4. Пусть функция $J(u)$ определена на отрезке $[a, b]$ 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 $J''(v) \geq 0$. 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 отрезке $[a, b]$ и в точке $v \in [a, b]$ имеет n производных ($n \geq 2$), причем известно, что $J^{(i)}(v) = 0$ при $i = 1, \dots, n-1$ 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 $n \geq 1$? 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 $J'(v)(u - v) \geq 0$ 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 то-

чек u_1 , u_2 вычисляются значения $J(u_1)$, $J(u_2)$ и сравниваются между собой. Если $J(u_1) \leq J(u_2)$, то полагают $a_1 = a$, $b_1 = u_2$; если же $J(u_1) > J(u_2)$, то $a_1 = u_1$, $b_1 = b$. Так как $J(u)$ унимодальна на $[a, b]$, то ясно, что отрезок $[a_1, b_1]$ имеет общую точку со множеством U_* точек минимума $J(u)$ на $[a, b]$ и его длина $b_1 - a_1 = (b - a - \delta)/2 + \delta$.

Пусть отрезок $[a_{k-1}, b_{k-1}]$, имеющий непустое пересечение с U_* , уже известен, и пусть $b_{k-1} - a_{k-1} = (b - a - \delta)/2^{k-1} + \delta > \delta$, $k \geq 2$. 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 $J(u_{2k-1}) \leq J(u_{2k})$, то полагаем $a_k = a_{k-1}$, $b_k = u_{2k}$; если же $J(u_{2k-1}) > J(u_{2k})$, то $a_k = u_{2k-1}$, $b_k = b_{k-1}$. Длина получившегося отрезка $[a_k, b_k]$ равна $b_k - a_k = (b - a - \delta)/2^k + \delta > \delta$, и $[a_k, b_k] \cap U_* \neq \emptyset$.

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 $k > \log_2((b - a - \delta)/(\varepsilon - \delta))$. Так как 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 $J(u_{2k-1}) \leq J(u_{2k})$, и $\bar{u}_n = u_{2k}$ при $J(u_{2k-1}) > J(u_{2k})$, а значение $J(\bar{u}_n)$ может служить приближением для $J_* = \inf_{u \in [a, b]} J(u)$. При та-

ком выборе приближения для U_* будет допущена погрешность $\rho(\bar{u}_n, U_*) \leq \max\{b_k - \bar{u}_n; \bar{u}_n - a_k\} = (b - a - \delta)/2^k$. 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 $\rho(v_n, U_*) \leq (b_k - a_k)/2 = (b - a - \delta)/2^{k+1} + \delta/2$ (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, 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 пре-

дыдущего 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 следующем параграфе.

В заключение отметим, что метод деления отрезка пополам без изменений можно применять для минимизации функций, не являющихся унимодальными. Однако в этом случае нельзя будет гарантировать, что найденное решение явится 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 $u_1 = a + (3 - \sqrt{5})(b - a)/2 = a + 0,381966011 \dots$ $\dots(b - a)$ и $u_2 = a + (\sqrt{5} - 1)(b - a)/2 = a + 0,618033989 \dots$ $\dots(b - a)$, расположенными симметрично относительно середины отрезка, причем $a < u_1 < u_2 < b$, $(b - a)/(b - u_1) = (b - u_1)/(u_1 - a) = (b - a)/(u_2 - a) = (u_2 - a)/(b - u_2) = (\sqrt{5} + 1)/2 = 1,618033989 \dots$

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 $J(u_1) \leq J(u_2)$, то примем $a_2 = a_1$, $b_2 = u_2$, $\bar{u}_2 = u_1$; если же $J(u_1) > J(u_2)$, то $a_2 = u_1$, $b_2 = b_1$, $\bar{u}_2 = u_2$. Так как функция $J(u)$ унимодальна на $[a, b]$, то отрезок $[a_2, b_2]$ имеет хотя бы одну общую точку со множеством U_* .

точек минимума $J(u)$ на $[a, b]$. Кроме того, $b_2 - a_2 = (\sqrt{5} - 1) \times (b - a)/2$, 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 $u_1, u_2, \dots, u_{n-1}$, вычислены значения $J(u_1), J(u_2), \dots, J(u_{n-1})$, найден отрезок $[a_{n-1}, b_{n-1}]$ такой, что $[a_{n-1}, b_{n-1}] \cap U_* \neq \emptyset$, $b_{n-1} - a_{n-1} = ((\sqrt{5} - 1)/2)^{n-2} (b - a)$, и известна точка $\bar{u}_{n-1}$, производящая золотое сечение отрезка $[a_{n-1}, b_{n-1}]$ и такая, что $J(\bar{u}_{n-1}) = \min_{1 \leq i \leq n-1} J(u_i)$, $n \geq 2$. 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 $J(u_n) \leq J(\bar{u}_{n-1})$, 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 $a_n = u_n$, $b_n = b_{n-1}$, $\bar{u}_n = \bar{u}_{n-1}$. Новый отрезок $[a_n, b_n]$ таков, что $[a_n, b_n] \cap U_* \neq \emptyset$, $b_n - a_n = ((\sqrt{5} - 1)/2)^{n-1} (b - a)$, точка $\bar{u}_n$ производит золотое сечение $[a_n, b_n]$ и $J(\bar{u}_n) = \min\{J(u_n), J(\bar{u}_{n-1})\} = \min_{1 \leq i \leq n} J(u_i)$.

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 $J_* = \inf_{u \in [a, b]} J(u)$, а точка $\bar{u}_n$ служит приближением

для множества U_* с погрешностью $\rho(\bar{u}_n, U_*) \leq \max\{b_n - \bar{u}_n, \bar{u}_n - a_n\} = \frac{1}{2}(\sqrt{5} - 1)(b_n - a_n) = \left(\frac{\sqrt{5} - 1}{2}\right)^n (b - a) = A_n$. Напомним, 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 $\rho(\bar{u}_n, U_*) \leq 2^{-n/2} (b - a - \delta) < 2^{-n/2} (b - a) = B_n$. Отсюда имеем $A_n/B_n = (2\sqrt{2}/(\sqrt{5} + 1))^n \approx (0,87 \dots)^n$ — 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

$\Delta_n = b_n - a_n = ((\sqrt{5} - 1)/2)^{n-1} (b - a)$, $n = 1, 2, \dots$. Нетрудно проверить, что Δ_n является решением конечно-разностного уравнения $\Delta_{n-2} = \Delta_{n-1} + \Delta_n$ или

$$\Delta_n = \Delta_{n-2} - \Delta_{n-1}, \quad n = 3, 4, \dots \quad (1)$$

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

Как известно [2], [40], линейно независимые частные решения этого уравнения имеют вид τ_1^n и τ_2^n , $n = 1, 2, \dots$, где $\tau_1 = (\sqrt{5} - 1)/2$, $\tau_2 = -(\sqrt{5} + 1)/2$ — корни характеристического уравнения $\tau^2 + \tau - 1 = 0$, а любое решение (1) представимо в виде

$$\Delta_n = A\tau_1^n + B\tau_2^n, \quad n = 1, 2, \dots, \quad (2)$$

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

$$\begin{cases} A\tau_1 + B\tau_2 = \Delta_1, \\ A\tau_1^2 + B\tau_2^2 = \Delta_2. \end{cases} \quad (3)$$

При $\Delta_1 = b - a$, $\Delta_2 = b - u_1$ 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 $\tilde{\Delta}_n = \tilde{A}\tau_1^n + \tilde{B}\tau_2^n$, $n = 1, 2, \dots$. Так как $0 < \tau_1 = 0,6 \dots < 1$, $|\tau_2| = 1,6 \dots > 1$, 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 $\tilde{u}_n$, $u_{n+1} = a_n + b_n - \tilde{u}_n$ 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 , $n \geq 2$. А именно, на каждом отрезке $[a_n, b_n]$, содержащем точку $\tilde{u}_n$ с предыдущего шага, при выборе следующей точки u_{n+1} нужно остерегаться пользоваться формулой $u_{n+1} = a_n + b_n - \tilde{u}_n$, и вместо этого лучше непосредственно произвести золотое сечение отрезка $[a_n, b_n]$ и в качестве u_{n+1} взять ту из точек

$a_n + (3 - \sqrt{5})(b_n - a_n)/2$, $a_n + (\sqrt{5} - 1)(b_n - a_n)/2$, 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.

В § 8 будет показано, что метод золотого сечения и приведенная его модификация являются почти наилучшими на классе унимодальных функций. Разумеется, этот метод допускает применение 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 ($n \geq 2$) и найдены отрезок $[a_{n-1}, b_{n-1}]$, 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 симметричных 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 $J(\bar{u}_{n-1}) \leq J(u_n)$ 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 $\bar{u}_n \neq (a_n + b_n)/2$.

Симметричный метод описан. Из этого описания 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 $[a_n, b_n]$, $n = 1, 2, \dots$, где $\Delta_1 = b - a$, $\Delta_2 = \max\{b - u_1; u_1 - a\}$. Очевидно, $\Delta_{n+1} \geq \Delta_n/2$ при всех $n \geq 1$. 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 накоплению 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

$$\Delta_n/2 < \Delta_{n+1} \leq 2\Delta_n/3, \quad n = 1, 2, \dots, N, \quad (4)$$

при некотором $N > 1$, то Δ_n будут удовлетворять конечно-разностному уравнению (1) при $n = 2, 3, \dots, N$, 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 $[a_n, b_n]$, $2 \leq n \leq N$, 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 $\bar{u}_n$.

Если условие (4) на некоторых шагах нарушается, то закон изменения Δ_n описывается чуть более сложными конечно-разностными уравнениями, чем (1). Например, если (4) не выполнено при некотором $n = k_1 \geq 1$, т. е. вместо (4) имеет место $2\Delta_{k_1}/3 < \Delta_{k_1+1} < \Delta_{k_1}$, то из описания симметричного метода следует, что Δ_n удовлетворяет уравнениям

$$\Delta_{k_1+n} = \Delta_{k_1} - nA_1, \quad n = 1, 2, \dots, k_2; \quad \Delta_{k_1+k_2+1} = A_1, \quad (5)$$

где $A_1 = \Delta_{k_1} - \Delta_{k_1+1}$, а k_2 — тот номер, для которого $A_1 < \Delta_{k_1+k_2} \leq 2A_1$. Если же при $n = k_1$ условие (4) выполнено, то приняв в (5) $k_2 = 1$, приходим к уравнению (1) для $n = k_1 + 2$. Аналогично выписываются уравнения для Δ_n при $n > k_1 + k_2 + 1$. Пользуясь этими уравнениями, можно проследить за поведением погрешностей от шага к шагу и с помощью индукции показать, что погрешности будут расти не медленнее, чем в случае симметричных методов с условием (4) (см. упражнение 8.9). Доказательство этого факта, которое не очень сложно, но довольно громоздко для изложения, а также описание модификации симметричного метода в общем случае мы здесь не приводим, так как ниже будем иметь дело только с такими симметричными методами, которые удовлетворяют условию (4) и чаще встречаются на практике.

§ 5. Постановка задачи об оптимальных методах

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 минимизируемой функции. В связи с этим возникают вопросы, что такое оптимальный метод, 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 задано некоторое множество методов P , класс функций Q и зафиксирована какая-либо постановка задачи минимизации (например, задача первого или второго типа из § 1). 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 $\delta(p) = \sup_{J \in Q} \Delta(J, p)$, выражающую со-

бой погрешность метода p при минимизации самой «плохой» для p функции из Q .

Определение 1. Величину $\delta(p) = \sup_{J \in Q} \Delta(J, p)$ 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 $\delta(p_*) = \inf_{p \in P} \delta(p) = \gamma$, а величину γ —

наилучшей гарантированной точностью методов P на классе Q . Если для некоторого метода $p_e \in P$ выполняется неравенство $\delta(p_e) \leq \gamma + \varepsilon$, то метод p_e 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 [5, 9, 56, 67, 78, 118, 123, 145, 216, 219, 222, 247].

Здесь мы ограничимся рассмотрением оптимальных методов решения задач минимизации второго типа на классе функций Q , состоящем из всех унимодальных функций на отрезке $[a, b]$. Кроме того, будем пользоваться методами минимизации, использующими лишь значения функции, считая также, что число n вычислений значений минимизируемой функции заранее задано. При выборе точек $u_1, u_2, \dots, u_n$, в которых будут вычисляться значения функции, будем выделять два принципиально различных способа: 1) в первом способе точки $u_1, \dots, u_n$ выбираются последовательно, отдельными порциями, причем при выборе каждой очередной порции учитываются результаты предыдущих вычислений; 2) в другом, так называемом пассивном способе, все точки $u_1, \dots, u_n$ выбираются сразу, до начала вычислений, и в дальнейшем уже не меняются. Перейдем к более точным формулировкам.

Определение 2. Пусть $J(u)$ — унимодальная функция на отрезке $[\alpha, \beta]$, и пусть вычислены значения $J(u)$ в точках $u_1, u_2, \dots, u_m \in [\alpha, \beta]$. Скажем, что *тройка точек $a_m, b_m, \bar{u}_m$ локализует минимум функции $J(u)$ на отрезке $[\alpha, \beta]$ по точкам $u_1, \dots, u_m$, если 1) точка $\bar{u}_m$ совпадает с одной из точек $u_1, \dots, u_m$; 2) $J(\bar{u}_m) = \min_{1 \leq i \leq m} J(u_i)$; 3) $a_m < \bar{u}_m < b_m$, причем*

a_m, b_m — ближайшие к $\bar{u}_m$ точки среди точек $u_1, \dots, u_m, \alpha, \beta$.

Определение 3. Пусть $J(u) \in Q$ — класс унимодальных функций на отрезке $[a, b]$. Скажем, что на $[a, b]$ задан *последовательный метод p_n* , если 1) задано правило выбора первой порции точек $u_1, \dots, u_s \in [a, b]$, $1 \leq s \leq n$, по которым определяется тройка точек $a_s, b_s, \bar{u}_s$, локализирующая минимум $J(u)$ на $[a, b]$; 2) при $s < n$ на отрезке $[a_s, b_s]$ по заданному правилу выбирается вторая порция точек $u_{s+1}, \dots, u_k$, $s < k \leq n$, и по точкам $\bar{u}_s, u_{s+1}, \dots, u_k$ определяется тройка $a_k, b_k, \bar{u}_k$, локализирующая минимум $J(u)$

на $[a_s, b_s]$; 3) при $k < n$ на отрезке $[a_k, b_k]$ по известному правилу выбирается следующая порция точек $u_{k+1}, \dots, u_m$, $k < m \leq n$, и по точкам $\bar{u}_k, u_{k+1}, \dots, u_m$ находится тройка $a_m, b_m, \bar{u}_m$, локализирующая минимум $J(u)$ на $[a_k, b_k]$ и т. д. Этот процесс выбора точек отдельными порциями продолжается до тех пор, пока не будет выбрана последняя n -я точка и определена соответствующая локализирующая тройка $a_n, b_n, \bar{u}_n$.

Примерами последовательного метода служат методы деления отрезка пополам, золотого сечения. В результате применения последовательного метода p_n для минимизации функции $J(u) \in Q$ получим вложенные отрезки $[a, b] \supset [a_s, b_s] \supset [a_k, b_k] \supset \dots \supset [a_n, b_n]$ такие, что $[a_n, b_n] \cap U_* \neq \emptyset$, $J(\bar{u}_n) = \min_{1 \leq i \leq n} J(u_i)$, $a_n < \bar{u}_n < b_n$, причем a_n, b_n — ближайшие к $\bar{u}_n$ точки среди $u_1, \dots, u_n, a, b$. Это значит, что полученная методом p_n тройка $a_n, b_n, \bar{u}_n$ локализует минимум $J(u)$ на $[a, b]$.

Определение 4. Пусть $J(u) \in Q$. Скажем, что на отрезке $[a, b]$ задан пассивный метод p_n , если задано правило выбора сразу всех n точек $u_1, \dots, u_n \in [a, b]$, по которым определяется тройка точек $a_n, b_n, \bar{u}_n$, локализирующая минимум функции $J(u)$ на $[a, b]$.

Примером пассивного метода является метод равномерного перебора. В этом методе точки $u_1, \dots, u_n$ выбираются по правилу $u_i = a_1 + ih$, $i = 1, \dots, n$, где $h = \text{const} > 0$ — шаг метода, a_1 — заданная точка такая, что $|a_1 - a| \leq h$ (например, $a_1 = a$ или $a_1 = a + h/2$), и, кроме того, $nh \leq b - a_1 < (n + 1)h$.

Из определений 3, 4 видно, что пассивный метод является частным случаем последовательного метода, когда все n точек выбираются сразу в первой же порции. Поэтому нетрудно понять, что последовательные методы обладают большей гибкостью и гораздо точнее пассивных методов — это обстоятельство будет строго установлено ниже. Однако это не значит, что пассивные методы вовсе не находят применения. Они незаменимы во всех тех случаях, когда можно одновременно провести n вычислений значений минимизируемой функции, но нет возможностей (например, не хватает времени) для того, чтобы проводить вычисления последовательно, отдельными порциями с учетом результатов предыдущих вычислений.

Предположим, что в результате применения какого-либо последовательного или пассивного метода p_n для минимизации функции $J(u) \in Q$ найдена тройка $a_n, b_n, \bar{u}_n$, локализирующая точку минимума $J(u)$ на $[a, b]$. Что теперь считать приближенным решением задачи минимизации второго типа? Так как $J(\bar{u}_n) = \min_{1 \leq i \leq n} J(u_i)$, то естественно принять величину $J(\bar{u}_n)$ за приближенное значение $J_* = \inf_{u \in [a, b]} J(u)$. Что касается выбора точки, являющейся

приближением к множеству $U_* = \{u: u \in [a, b], J(u) = J_*\}$, то здесь напрашиваются, по крайней мере, две интересные возможности: $\bar{u}_n$ или $v_n = (a_n + b_n)/2$. Что лучше? Для ответа на этот вопрос нам придется уточнить постановку задачи минимизации второго типа. А именно, мы будем различать два варианта этой задачи и для краткости назовем их задачами А и Б.

Задача А. С помощью n вычислений значений функции $J(u)$ найти такую точку w_n , которая удалена от U_* на возможно меньшее расстояние, и соблюсти условие, чтобы приближением к J_* служило как раз значение $J(w_n)$.

Задача Б. С помощью n вычислений значений функции $J(u)$ найти такую точку w_n , которая удалена от U_* на возможно меньшее расстояние, но, в отличие от задачи А, не требуя того, чтобы значение функции, принимаемое за приближение к J_* , обязательно было вычислено в точке w_n .

Поскольку количество вычислений значений функции ограничено сверху числом n , то в задаче А точку w_n придется выбирать среди точек $u_1, \dots, u_n$, в которых значение функции уже вычислено, а в задаче Б эта точка может

быть отличной от $u_1, \dots, u_n$. Это небольшое, казалось бы, различие в постановках задач А и Б приводит к тому, что в задаче А лучше всего принять $w_n = \bar{u}_n$ с погрешностью

$$\rho(\bar{u}_n, U_*) \leq \max \{b_n - \bar{u}_n, \bar{u}_n - a_n\}, \quad (1)$$

а в задаче Б $w_n = v_n = (a_n + b_n)/2$ с погрешностью

$$\rho(v_n, U_*) \leq (b_n - a_n)/2; \quad (2)$$

здесь $a_n, b_n, \bar{u}_n$ — локализирующая тройка метода p_n .

Покажем, что оценки (1), (2) точны на классе Q . Начнем с оценки (1). Из определения последовательного и пассивного методов видно, что локализирующая тройка $a_n, b_n, \bar{u}_n$, полученная методом p_n при минимизации функции $J(u) \in Q$, останется без изменений, если тот же метод p_n применить к любой другой функции $g(u) \in Q$, которая совпадает с $J(u)$ вне $[a_n, b_n]$, а на отрезке $[a_n, b_n]$ удовлетворяет условию $g(\bar{u}_n) < \min \{g(a_n); g(b_n)\}$. Поэтому без умаления общности можем считать, что $J(\bar{u}_n) < \min \{J(a_n); J(b_n)\}$. Далее, в зависимости от того, будет ли $\max \{b_n - \bar{u}_n, \bar{u}_n - a_n\} = b_n - \bar{u}_n$ или $\bar{u}_n - a_n$, определим соответственно функции

$$J_1(u) = \begin{cases} J(u) & \text{вне } [\bar{u}_n, b_n], \\ (u - \bar{u}_n)(J(b_n) - J(\bar{u}_n) - C)/(b_n - \bar{u}_n - \varepsilon) + J(\bar{u}_n) & \text{при } \bar{u}_n \leq u \leq b_n - \varepsilon, \\ (u - b_n)C/\varepsilon + J(b_n) & \text{при } b_n - \varepsilon < u \leq b_n; \end{cases}$$

$$J_2(u) = \begin{cases} J(u) & \text{вне } [a_n, \bar{u}_n], \\ (-u + a_n)C/\varepsilon + J(a_n) & \text{при } a_n \leq u \leq a_n + \varepsilon, \\ (u - \bar{u}_n)(J(\bar{u}_n) - J(a_n) + C)/(\bar{u}_n - a_n - \varepsilon) + J(\bar{u}_n) & \text{при } a_n + \varepsilon < u \leq \bar{u}_n, \end{cases}$$

где $0 < \varepsilon < \min \{b_n - \bar{u}_n; \bar{u}_n - a_n\}$, $\max \{J(a_n), J(b_n)\} - J(\bar{u}_n) < C$. Очевидно, функции $J_1(u)$, $J_2(u)$ строго унимодальны на отрезке $[a, b]$ и принимают свое наименьшее значение в точках $b_n - \varepsilon$ и $a_n + \varepsilon$ соответственно. Отсюда ясно, что выбирая ε достаточно малым, погрешность $\rho(\bar{u}_n, U_*)$ решения задачи А для функций $J_1(u)$, $J_2(u)$ методом p_n можно сделать как угодно близкой к $\max \{b_n - \bar{u}_n, \bar{u}_n - a_n\}$. Это значит, что оценка (1) точна на классе Q . Конечно, в задаче А вместо $w_n = \bar{u}_n$ можно было бы еще попробовать взять $w_n = a$ или $w_n = b$, полагая, что значение $J(w_n)$ вычислено, но те же функции $J_1(u)$, $J_2(u)$ показывают, что погрешность $\rho(w_n, U_*)$ в этом случае стала бы неоправданно большой по сравнению с (1). Из сказанного следует, что в задаче А выбор $w_n = \bar{u}_n$ является наилучшим, а величину $\max \{b_n - \bar{u}_n, \bar{u}_n - a_n\} = \Delta_A(J, p_n)$ можно принять за погрешность последовательного или пассивного метода p_n при решении задачи А для функции $J = J(u) \in Q$.

Аналогично можно убедиться, что в задаче Б выбор точки $w_n = v_n = (a_n + b_n)/2$ является наилучшим на классе Q . В самом деле, если $a_n \leq w_n < v_n$ или $v_n < w_n \leq b_n$, то для функций $J_1(u)$ и соответственно $J_2(u)$, введенных выше при обсуждении задачи А, погрешность $\rho(w_n, U_*)$ будет заведомо больше $(b_n - a_n)/2$. Те же функции показывают, что $\rho(v_n, U_*)$ может быть сколь угодно близкой к $(b_n - a_n)/2$, так что оценка (2) точна на классе Q . Поэтому естественно принять величину $(b_n - a_n)/2 = \Delta_B(J, p_n)$ за погрешность последовательного или пассивного метода p_n при решении задачи Б для функции $J = J(u) \in Q$. Заметим, что всегда $\Delta_B(J, p_n) \leq \Delta_A(J, p_n)$, причем равенство возможно только при $\bar{u}_n = (a_n + b_n)/2$.

Таким образом, n -точечные (т. е. использующие вычисление значений функции в n точках) последовательные и пассивные методы решения задач А и Б полностью 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, $\Delta(J, p_n) = \Delta_A(J, p_n)$, то получим определения гарантированной точности, оптимального и ϵ -оптимального последовательного (или пассивного) метода для задачи А, а при $\Delta(J, p_n) = \Delta_B(J, p_n)$ — определения тех же понятий для задачи Б. В последующих параграфах будет выяснено существование оптимальных последовательных или пассивных методов для задач А и Б и изучены возможности их построения.

Заметим, что все рассмотрения выше проводились для фиксированного отрезка $[a, b]$. Возникает вопрос: если оптимальный или ϵ -оптимальный (последовательный или пассивный) метод поиска минимума унимодальных функций на отрезке $[a, b]$ уже построен, то нельзя ли указать какой-либо простой рецепт получения оптимального или ϵ -оптимального метода поиска минимума унимодальных функций на отрезке $[c, d]$, отличном от $[a, b]$? Для ответа на этот вопрос возьмем преобразование отрезка $[a, b]$ в $[c, d]$, задаваемое линейной функцией $v = (u - a)(d - c)/(b - a) + c$, и каждой унимодальной функции $g(v)$ на $[c, d]$ поставим в соответствие унимодальную функцию $J(u) = g((u - a)(d - c)/(b - a) + c)$ на $[a, b]$. Очевидно, тем самым между классами унимодальных функций на отрезках $[a, b]$ и $[c, d]$ установлено взаимно однозначное соответствие. Теперь естественно принять следующее

Определение 5. Пусть на отрезке $[a, b]$ задан какой-либо последовательный или пассивный метод p_n . Применением метода p_n к отрезку $[c, d]$ для поиска минимума унимодальной функции $g(v)$ на $[c, d]$ будем называть выбор точек $v_i = (u_i - a)(d - c)/(b - a) + c \in [c, d]$, $i = 1, \dots, n$, по тем же правилам и в той же последовательности, по которым выбираются точки $u_i \in [a, b]$, $i = 1, \dots, n$, с соответствующими номерами при применении метода p_n для минимизации функции $J(u) = g((u - a)(d - c)/(b - a) + c)$ на $[a, b]$.

После принятия такого определения имеет смысл говорить о последовательных и пассивных методах p_n безотносительно к какому-либо отрезку. В тех случаях, когда желательно подчеркнуть, что поиск минимума проводится именно на отрезке $[a, b]$, вместо $\Delta(J, p_n)$, $\delta(p_n)$, $\gamma = \gamma_n$ (см. определение 1) будем писать $\Delta(J, p_n, b - a)$, $\delta(p_n, b - a)$, $\gamma_n(b - a)$ соответственно, снабжая эти величины индексами А или Б в зависимости от того, рассматривается задача А или Б. Из определений 1—5 следует, что

$$\Delta(J, p_n, \alpha(b - a)) = \alpha \Delta(J, p_n, b - a),$$

$$\delta(p_n, \alpha(b - a)) = \alpha \delta(p_n, b - a), \quad \gamma_n(\alpha(b - a)) = \alpha \gamma_n(b - a)$$

при любых $\alpha > 0$.

В заключение заметим, что задача Б равносильна следующей задаче Б'. Задача Б'. С помощью n -точечного последовательного (или пассивного) метода получить отрезок $[a_n, b_n]$, имеющий возможно меньшую гарантированную на классе Q длину и содержащий хотя бы одну точку минимума.

Равносильность задач Б и Б' следует из того, что $b_n - a_n = = 2\Delta_B(J, p_n)$.

Упражнения. 1. Найти оптимальные и ϵ -оптимальные последовательные и пассивные методы для задач А и Б при $n = 1$ и $n = 2$.

2. Найти гарантированную на классе Q точность последовательного (или пассивного) метода p_n , если в качестве погрешности метода p_n при минимизации функции $J = J(u) \in Q$ принята величина $\Delta(J, p_n) = |J_n - J(\bar{u}_n)|$.

§ 6. Оптимальные пассивные методы

Пусть Q — класс унимодальных функций на отрезке $[a, b]$, P_n — множество n -точечных пассивных методов на $[a, b]$, $n \geq 1$. Пусть $p_n = \{u_1, \dots, u_n\}$ — какой-либо пассивный метод, $u_0 = a \leq u_1 < \dots < u_n \leq b = u_{n+1}$, а u_{k-1}, u_k, u_{k+1} , u_k — локализирующая тройка этого метода для функции $J = J(u) \in Q$. Тогда погрешность метода p_n при решении задачи А, согласно § 5, равна $\Delta_A(J, p_n, b-a) = \max \{u_{k+1} - u_k, u_k - u_{k-1}\}$.

Найдем $\delta_A(p_n, b-a) = \sup_{J \in Q} \Delta_A(J, p_n, b-a)$ — гарантированную точность метода p_n для задачи А на классе Q , рассчитанную на самые «плохие» для p_n функции $J(u) \in Q$. Прежде всего ясно, что $\Delta_A(J, p_n, b-a) \leq \max_{0 \leq i \leq n} (u_{i+1} - u_i)$ для любой функции $J = J(u) \in Q$. Пусть $\max_{0 \leq i \leq n} (u_{i+1} - u_i) = u_{k+1} - u_k$. Возьмем функцию

$$J_\varepsilon = J_\varepsilon(u) = \begin{cases} \varepsilon(-u + u_{k+1} - \varepsilon) & \text{при } u \leq u_{k+1} - \varepsilon, \\ (u - u_{k+1} + \varepsilon)/\varepsilon & \text{при } u > u_{k+1} - \varepsilon, \end{cases} \quad 0 < \varepsilon < u_{k+1} - u_k.$$

Очевидно, $J_\varepsilon(u)$ строго унимодальна, свое наименьшее значение на $[a, b]$ принимает в точке $u_* = u_{k+1} - \varepsilon$, и кроме того, $\Delta_A(J_\varepsilon, p_n, b-a) = u_{k+1} - u_k$. Отсюда следует, что

$$\delta_A(p_n, b-a) = \max_{0 \leq i \leq n} (u_{i+1} - u_i).$$

Определение 1. Величина $\gamma_n^A(b-a) = \inf_{p_n \in P_n} \delta_A(p_n, b-a)$ называется *наилучшей гарантированной точностью* n -точечных пассивных методов для задачи А на классе Q . Пассивный метод p_n^A называется *оптимальным* для задачи А, если $\delta_A(p_n^A, b-a) = \gamma_n^A(b-a)$. Пассивный метод $p_n^{A, \varepsilon}$ называется *ε -оптимальным* для задачи А, если $\delta_A(p_n^{A, \varepsilon}, b-a) \leq \gamma_n^A(b-a) + \varepsilon$.

Как видим, определение 1 является переформулировкой определения 5.1 на случай задачи А пассивных методов и класса унимодальных функций.

Теорема 1. Для задачи А при всех $n \geq 1$ существует и притом единственный оптимальный пассивный метод p_n^A на классе Q ; он имеет вид $p_n^A = \{v_i = i(b-a)/(n+1), i = 1, 2, \dots, n\}$. Наилучшая гарантированная точность пассивных методов P_n на классе Q равна $(b-a)/(n+1) = \gamma_n^A(b-a)$.

Доказательство. Прежде всего, очевидно, $\delta_A(p_n^A, b-a) = (b-a)/(n+1)$. Для любого другого пассивного метода $p_n = \{u_1, u_2, \dots, u_n\}$, $u_0 = a \leq u_1 < \dots < u_n \leq b = u_{n+1}$, всегда найдется отрезок $[u_k, u_{k+1}]$ длины $u_{k+1} - u_k \geq (b-a)/(n+1)$, ибо в противном случае сумма длин всех отрезков $[u_i, u_{i+1}]$, $i = 0, 1, \dots, n$, будет меньше $b-a$. Следовательно, $\delta_A(p_n, b-a) = \max_{0 \leq i \leq n} (u_{i+1} - u_i) \geq (b-a)/(n+1)$ для любого пассивного метода p_n , причем равенство здесь возможно только при $p_n = p_n^A$.

Перейдем к рассмотрению задачи Б. Погрешность пассивного метода $p_n = \{u_1, \dots, u_n\}$, $u_0 = a \leq u_1 < \dots < u_n \leq b = u_{n+1}$, при решении задачи Б, согласно § 5, равна

$$\Delta_B(J, p_n, b-a) = (u_{k+1} - u_{k-1})/2.$$

Как и в случае задачи А, нетрудно убедиться, что гарантированная точность метода p_n на классе Q тогда равна

$$\Delta_B(p_n, b-a) = (\max_{1 \leq i \leq n} (u_{i+1} - u_{i-1}))/2.$$

Заменяя в определении 1 букву А на Б получим определения наилучшей гарантированной точности пассивных методов, оптимального и ϵ -оптимального пассивного методов для задачи Б на классе Q .

При исследовании пассивных методов для задачи Б, оказывается, нужно различать случаи четного и нечетного n .

Теорема 2. Для задачи Б при всех нечетных $n = 2m + 1$, $m \geq 0$, существует бесконечно много оптимальных пассивных методов на классе Q наилучшая гарантированная точность пассивных методов P_{2m+1} на этом классе равна $(b-a)/2(m+1)$.

Доказательство. Возьмем пассивный метод $p_n^B = \{v_1, \dots, v_n\}$, где $v_{2i} = a + i(b-a)/(m+1)$, $i = 1, 2, \dots, m$, а точки v_{2i+1} , $i = 0, 1, \dots, m$, расположены на отрезке $[a, b]$ произвольно, лишь бы $v_{2i-1} < v_{2i} < v_{2i+1}$, $v_{2i+1} - v_{2i-1} \leq (b-a)/(m+1)$. Очевидно, $\delta_B(p_n^B, b-a) = (b-a)/2(m+1)$. С другой стороны, для любого пассивного метода $p_n = \{u_1, \dots, u_n\}$, $u_0 = a \leq u_1 < \dots < u_n \leq b = u_{n+1}$, $n = 2m + 1$, имеем $\delta_B(p_n, b-a) = \frac{1}{2} \max_{1 \leq i \leq n} (u_{i+1} - u_{i-1}) \geq \frac{1}{2} \max \{b - u_{2m}, u_{2m} - u_{2m-2}, \dots, u_4 - u_2,$

$u_2 - a\} \geq \frac{b-a}{2(m+1)}$. Следовательно, методы p_n^B оптимальны и $\gamma_n^B(b-a) = (b-a)/2(m+1)$.

Теорема 3. Для задачи Б при всех четных $n = 2m$, $m \geq 1$, оптимального пассивного метода на классе Q не существует; наилучшая гарантированная точность пассивных методов P_{2m} на этом классе равна $(b-a)/2(m+1)$. В качестве ϵ -оптимального метода можно взять $p_n^{B, \epsilon} = \{v_1, \dots, v_n\}$, где $v_{2i-1} = a + i(b-a)/(m+1) - \epsilon$, $v_{2i} = a + i(b-a)/(m+1) + \epsilon$, $i = 1, \dots, m$, $0 < \epsilon < (b-a)/2(m+1)$.

Доказательство. Покажем, что $\delta_B(p_n, b-a) > (b-a)/2(m+1)$ для любого пассивного метода $p_n = \{u_1, \dots, u_n\}$, $u_0 = a \leq u_1 < \dots < u_n \leq b = u_{n+1}$, $n = 2m$. Обозначим $\bar{u} = a + (b-a)/(m+1)$. Имеются две возможности: либо $u_2 > \bar{u}$, либо $u_2 \leq \bar{u}$. Если $u_2 > \bar{u}$, то $\delta_B(p_n, b-a) = (\max_{1 \leq i \leq n} (u_{i+1} - u_{i-1}))/2 \geq (u_2 - a)/2 > (\bar{u} - a)/2 = (b-a)/2(m+1)$. Если же $u_2 \leq \bar{u}$, то $\max_{1 \leq i \leq n} (u_{i+1} - u_{i-1}) > u_2 - a$. В самом деле, если бы

$\max_{1 \leq i \leq n} (u_{i+1} - u_{i-1}) \leq u_2 - a$, то $u_{2i+1} - u_{2i-1} \leq u_2 - a$, $i = 1, \dots, m$, и кроме того, $u_1 - a < u_2 - a$. Сложив последние неравенства, придем к противоречивому неравенству $b - a < (m+1)(u_2 - a) \leq (m+1)(\bar{u} - a) = b - a$. Таким образом, при $u_2 \leq \bar{u}$ имеем $\delta_B(p_n, b-a) = (\max_{2 \leq i \leq n} (u_{i+1} - u_{i-1}))/2 =$

$= \delta_B(p'_{n-1}, b - u_1)$, где p'_{n-1} — пассивный метод на отрезке $[u_1, b]$, составленный из точек $u_2, u_3, \dots, u_n$ метода p_n . Но $n-1 = 2m-1$ — нечетное число, поэтому в силу теоремы 2 $\delta_B(p'_{n-1}, b - u_1) \geq (b - u_1)/2m$. Тогда $\delta_B(p_n, b-a) \geq \frac{b-u_1}{2m} > \frac{b-\bar{u}}{2m} = \frac{b-a}{2(m+1)}$. Тем самым доказано, что $\delta_B(p_n, b-a) > (b-a)/2(m+1)$ для любого $p_n \in P_{2m}$. С другой стороны,

$\delta_B(p_n^B, \varepsilon, b-a) = (b-a)/2(m+1) + \varepsilon/2$ для всех $\varepsilon, 0 < \varepsilon < (b-a)/2(m+1)$. Следовательно, $\gamma_n^B(b-a) = (b-a)/2(m+1)$.

Из теорем 2, 3 непосредственно вытекает

Следствие 1. Для любого пассивного метода $p_n = \{u_1, \dots, u_n\}$ на отрезке $[a, b]$ существует функция $J(u) \in Q$ такая, что для тройки точек $a_n, b_n, \bar{u}_n$, локализирующей минимум $J(u)$ на $[a, b]$ по точкам $u_1, \dots, u_n$, справедливо неравенство $b_n - a_n \geq (b-a)/(m+1) = 2\gamma_n^B(b-a)$, где $m = n/2$ для четных n , $m = (n-1)/2$ для нечетных n . С другой стороны, для любого $\varepsilon > 0$ найдется пассивный метод p_n^ε такой, что для него $b_n - a_n < (b-a)/(m+1) + \varepsilon$ для всех $J(u) \in Q$.

Иначе говоря, в следствии 1 утверждается, что

$$\inf_{p_n \in P_n} \sup_{J \in Q} (b_n - a_n) = (b-a)/(m+1) = 2\gamma_n^B(b-a)$$

— это и есть решение задачи Б' из § 5 пассивными методами.

Из теорем 2, 3 также следует, что при решении задачи Б с помощью пассивных методов предпочтительнее брать четное число $n = 2m$ точек, поскольку в случае $n = 2m+1$ наилучшая гарантированная точность остается такой же, как при $n = 2m$.

Уп ра ж н е н и е. Сравните оптимальные и ε -оптимальные пассивные методы на классе Q с методом деления отрезка пополам.

§ 7. Оптимальный последовательный метод для задачи А

Пусть по-прежнему Q — класс функций, унимодальных на отрезке $[a, b]$. Через P_n в настоящем параграфе будем обозначать множество всех n -точечных последовательных методов p_n , $n \geq 1$, через $\delta_A(p_n, b-a)$ — гарантированную точность метода p_n для задачи А на классе Q , через $\gamma_n^A(b-a) = \inf_{p_n \in P_n} \delta_A(p_n, b-a)$ — наилучшую гарантированную точность методов P_n на классе Q (определения всех этих понятий, а также оптимального и ε -оптимального методов см. в § 5).

Построим оптимальный на классе Q последовательный метод для задачи А. Оказывается, этот метод связан со знаменитыми числами Фибоначчи, и поэтому мы его будем называть методом Фибоначчи и обозначать через Φ_n . Как известно [67], числа Фибоначчи определяются соотношениями

$$F_{n+2} = F_{n+1} + F_n, \quad n = 1, 2, 3, \dots; \quad F_1 = F_2 = 1.$$

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

$$F_n = [(1 + \sqrt{5})/2]^n - ((1 - \sqrt{5})/2)^n / \sqrt{5}, \quad n = 1, 2, \dots \quad (1)$$

Вот несколько чисел Фибоначчи: $F_3 = 2$, $F_4 = 3$, $F_5 = 5$, $F_6 = 8$, $F_7 = 13$, $F_{10} = 55$, $F_{15} = 610$, $F_{20} = 6765$, $F_{25} = 75\,025$. Из формулы (1) следует, что при больших n $F_n \approx ((1 + \sqrt{5})/2)^n / \sqrt{5}$, так что числа Фибоначчи с увеличением n растут очень быстро.

Метод Фибоначчи Φ_n относится к классу симметричных методов, описанных в § 4, и определяется заданием на отрезке $[a, b]$ точки $u_1 = a + (b-a)F_n/F_{n+2}$ или симметричной ей точки $u_2 = a + b - u_1 = a + (b-a)F_{n+1}/F_{n+2}$. Пользуясь описанием симметричных методов, нетрудно показать с помощью индукции, что метод Φ_n обладает свойством (4.4) и на k -м шаге метода Φ_n будет получена тройка $a_k, b_k, \bar{u}_k$, локализирующая

минимум функции $J = J(u) \in Q$ и такая, что

$$\Delta_k = b_k - a_k = (b - a) F_{n-k+3}/F_{n+2}, \quad 1 \leq k \leq n, \quad a_1 = a, \quad b_1 = b, \quad (2)$$

а точка $\bar{u}_k$, $a_k < \bar{u}_k < b_k$, с вычисленным значением $J(\bar{u}_k) = \min_{1 \leq i \leq k} J(u_i)$ совпадает с одной из точек

$$\begin{aligned} u'_k &= a_k + (b_k - a_k) F_{n-k+1}/F_{n-k+3} = a_k + (b - a) F_{n-k+1}/F_{n+2}, \\ u''_k &= a_k + (b_k - a_k) F_{n-k+2}/F_{n-k+3} = a_k + (b - a) F_{n-k+2}/F_{n+2}, \end{aligned} \quad (3)$$

расположенных на отрезке $[a_k, b_k]$ симметрично относительно его середины. При $k = n$ процесс заканчивается — в этом случае согласно формулам (2), (3) длина отрезка $[a_n, b_n]$ равна $b_n - a_n = 2(b - a)/F_{n+2}$, а точки $u'_n = a_n + (b - a) F_1/F_{n+2}$, $u''_n = a_n + (b - a) F_2/F_{n+2}$ совпадают и делят отрезок $[a_n, b_n]$ пополам. Принимая точку $\bar{u}_n = u'_n = u''_n$ за приближение ко множеству U_* точек минимума функции $J(u)$ на $[a, b]$, допустим погрешность $\rho(\bar{u}_n, U_*) \leq (b - a)/F_{n+2}$ для всех $J(u) \in Q$. Нетрудно видеть, что для функции $J(u) = u$ метод Φ_n дает погрешность, в точности равную $(b - a)/F_{n+2}$. Это значит, что гарантированная точность метода Φ_n на классе Q равна $\delta_A(\Phi_n, b - a) = (b - a)/F_{n+2}$, $n \geq 1$. Метод Фибоначчи Φ_n полностью описан.

Имея в виду прикладные аспекты метода Φ_n , заметим, что число F_n/F_{n+2} вообще говоря, является бесконечной периодической десятичной дробью, поэтому первая точка u_1 будет задаваться на ЭВМ приближенно. Как отмечалось в § 4, погрешность в задании первой точки метода Φ_n приводит к быстрому росту погрешностей на последующих шагах, и при уже не очень больших n результаты будут сильно искажены. Поэтому на практике нужно пользоваться лишь модифицированным методом Φ_n . А именно, на каждом отрезке $[a_k, b_k]$, содержащем точку $\bar{u}_k$ с предыдущего шага, при выборе следующей точки u_{k+1} нужно остерегаться пользоваться формулой $u_{k+1} = a_k + b_k - \bar{u}_k$ и за точку u_{k+1} принимать ту из точек (3), которая наиболее удалена от $\bar{u}_k$ (здесь в (3) под F_{n-k+1}/F_{n-k+3} , F_{n-k+2}/F_{n-k+3} подразумеваются какие-либо подходящие приближения этих чисел).

Отметим одно замечательное свойство метода Φ_n , присущее, впрочем, всем симметричным методам: применение метода Φ_{n-k+1} на отрезке $[a_k, b_k]$, полученном в результате первых k шагов метода Φ_n , равносильно дальнейшему продолжению метода Φ_n на этом отрезке $[a_k, b_k]$. Этот факт вытекает из того, что первые две точки метода Φ_{n-k+1} на $[a_k, b_k]$ совпадают с точками (3) или, что то же самое, с точками $\bar{u}_k$, $u_{k+1} = a_n + b_n - \bar{u}_k$ метода Φ_n .

Теорема 1. Для задачи А метод Φ_n является единственным оптимальным последовательным методом на классе Q . Наилучшая гарантированная точность множества последовательных методов P_n на классе Q равна $\gamma_n^A(b - a) = (b - a)/F_{n+2}$.

Доказательство будем проводить индукцией по n . При $n = 1$ метод Φ_1 прост: берется точка $u_1 = (a + b)/2$, вычисляется $J(u_1)$, и за решение задачи А принимается пара $J(u_1)$, u_1 с гарантированной на классе Q точностью $\delta_A(\Phi_1, b - a) = (b - a)/2 = (b - a)/F_3$. Очевидно, в этом случае все утверждения теоремы верны.

Пусть оптимальность метода Φ_k и единственность оптимального метода из P_k на классе Q доказаны при всех $k = 1, \dots, n - 1$, $n \geq 2$. Покажем, что тогда Φ_n является единственным в P_n оптимальным методом на классе Q . Возьмем произвольный метод $p_n \in P_n$. Согласно определению 5.3 последовательного метода p_n сначала выбирается первая порция точек $u_1, \dots, u_s \in [a, b]$, $1 \leq s \leq n$, и сравнением величин $J(u_1), \dots, J(u_s)$ находится тройка точек $a_s, b_s, \bar{u}_s$, локализирующая минимум $J(u) \in Q$. Не умаляя

общности, можем считать, что $2 \leq s \leq n$. В самом деле, задание одной точки u_1 и вычисление значения $J(u_1)$ ничего не добавляет к известной информации о том, что множество $U_* \in [a, b]$, и поэтому остается переходить к выбору второй порции точек $u_2, \dots, u_s, s \geq 2$, метода p_n , что, конечно, равносильно заданию точек $u_1, \dots, u_s$ с самого начала, в первой же порции. Итак, пусть $2 \leq s \leq n$.

Отдельно рассмотрим случай $s = 2 \leq n$, когда метод p_n начинается с выбора двух точек $u_1, u_2, a \leq u_1 < u_2 \leq b$. Начальные точки метода Φ_n переобозначим через $v_1 = a + (b-a)F_n/F_{n+2}$, $v_2 = a + (b-a)F_{n+1}/F_{n+2}$. Пусть сначала $u_1 < v_1$. Если $J(u_1) > J(u_2)$, то U_* принадлежит отрезку $[u_1, b]$ длины $b - u_1 > b - v_1$, причем для поиска минимума на этом отрезке мы можем вычислить значение функции $J(u)$ еще в $n-1$ точках, включая точку $\bar{u}_2 = u_2$. Если даже точка $\bar{u}_2$ на $[u_1, b]$ расположена так удачно и допускает применения оптимального (в силу индукции) метода Φ_{n-1} на $[u_1, b]$ с участием точки $\bar{u}_2$, то и в этом случае гарантированная точность решения задачи А методом p_n будет больше, чем при применении метода Φ_n на $[a, b]$: $\delta_A(p_n, b-a) \geq \delta_A(\Phi_{n-1}, b-u_1) = (b-u_1)/F_{n+1} > (b-v_1)/F_{n+1} = (b-a)/F_{n+2} = \delta_A(\Phi_n, b-a)$. Таким образом, последовательный метод p_n , начинающийся с выбора двух точек $u_1, u_2, a \leq u_1 < u_2 \leq b$, где $u_1 < v_1$, заведомо неоптимален. Аналогично показывается, что в случае $u_2 > v_2$ метод p_n также неоптимален.

Пусть теперь $u_1 > v_1$. В худшем случае множество U_* может находиться на отрезке $[a, u_1]$ длиной $u_1 - a > v_1 - a = (b-a)F_n/F_{n+2}$, и на поиск минимума на этом отрезке в нашем распоряжении остается еще $n-2$ вычисления значений функции. Но если даже метод p_n таков, что дальнейший поиск на $[a, u_1]$ совпадает с оптимальным методом Φ_{n-2} , то и в этом случае гарантированная точность будет больше, чем при применении метода Φ_n на $[a, b]$: $\delta_A(p_n, b-a) \geq \delta_A(\Phi_{n-2}, u_1-a) = (u_1-a)/F_n > (v_1-a)/F_n = (b-a)/F_{n+2} = \delta_A(\Phi_n, b-a)$. Таким образом, метод p_n , начинающийся с выбора двух точек $u_1, u_2, a \leq u_1 < u_2 \leq b$, где $u_1 > v_1$, заведомо неоптимален. Аналогично доказывается неоптимальность метода p_n в случае $u_2 < v_2$.

Остается рассмотреть случай $u_1 = v_1, u_2 = v_2$, когда первые две точки методов p_n и Φ_n совпадают, и сравнение величин $J(v_1), J(v_2)$ приведет к одной и той же тройке $a_2, b_2, \bar{u}_2$, локализующей минимум $J(u)$ на $[a, b]$. Для поиска минимума на отрезке $[a_2, b_2]$ мы можем произвести вычисление значений функции еще в $n-1$ точках, включая точку $\bar{u}_2$. Продолжением метода Φ_n на отрезке $[a_2, b_2]$ является метод Φ_{n-1} на этом отрезке, являющийся единственным оптимальным методом из P_{n-1} по предположению индукции. Поэтому любое отклонение метода p_n на $[a_2, b_2]$ от Φ_n приведет лишь к увеличению гарантированной точности: $\delta_A(p_n, b-a) > \delta_A(\Phi_{n-1}, b_2-a_2) = \delta_A(\Phi_n, b-a)$ при $p_n \neq \Phi_n$. Таким образом, случай $s = 2 \leq n$ полностью рассмотрен. Вывод такой: среди методов $p_n \in P_n$, начинающихся с выбора двух точек, наилучшим является план Фибоначчи Φ_n . В частности, при $n = 2$ наилучшим в P_2 будет метод Φ_2 .

Наконец, перейдем к рассмотрению методов $p_n \in P_n$, начинающихся с выбора первой порции из s точек $u_1, \dots, u_s \in [a, b]$, $2 < s \leq n$, по которым затем находится тройка $a_s, b_s, \bar{u}_s$, локализующая минимум $J(u)$ на $[a, b]$. Согласно следствию 6.1 при любых, даже самых наилучших действиях на классе Q мы будем иметь $b_s - a_s \geq (b-a)/(m+1)$, где $m = s/2$ при четном s , $m = (s-1)/2$ при нечетном s , $m \geq 1$. В то же самое время, оказывается, первые s шагов метода Φ_n согласно формуле (2) приводят к отрезку длины $\Delta_s(\Phi_n) = (b-a)F_{n-s+3}/F_{n+2}$, меньшей, чем $\frac{b-a}{m+1} \leq b_s - a_s$. Это

вытекает из следующих неравенств:

$$2F_{n+2} > (s+1)F_{n-s+3} \quad \text{при } s = 2m+1, \quad 3 \leq s \leq n, \quad (4)$$

$$2F_{n+2} > (s+2)F_{n-s+3} \quad \text{при } s = 2m, \quad 4 \leq s \leq n \quad (5)$$

Справедливость неравенств (4) при $s = 3$ и (5) при $s = 4$ легко установить с помощью индукции по n . При всех $s \geq 4$, оказывается, верно более сильное неравенство (5), вытекающее из монотонного убывания $(s+2)F_{n-s+3}$ при возрастании s от $s = 4$ до $s = n$: $(s+2)F_{n-s+3} = (s+2)F_{n-s+2} + (s+2)F_{n-s+1} > (s+2)F_{n-s+2} + F_{n-s+2} = (s+3)F_{n-s+2}$. Для продолжения поиска минимума на отрезке $[a_s, b_s]$ в нашем распоряжении имеется возможность вычислить значения функции еще в $n-s+1$ точках этого отрезка, включая точку $\bar{u}_s$. Если даже точка $\bar{u}_s$ на $[a_s, b_s]$ расположена очень удачно и допускает применение оптимального (в силу индукции метода Φ_{n-s+1} на отрезке $[a_s, b_s]$ с участием точки $\bar{u}_s$, мы сможем получить лишь $\delta_A(p_n, b-a) \geq \delta_A(\Phi_{n-s+1}, b_s - a_s) = (b_s - a_s)/F_{n-s+3} > \Delta_s(\Phi_n)/F_{n-s+3} = (b-a)/F_{n+2} = \delta_A(\Phi_n, b-a)$. Случай $2 < s \leq n$ также рассмотрен. Теорема 1 доказана.

Следствие 1. Количество необходимых при решении задачи А вычислений значений функции $J(u) \in Q$, гарантирующих на Q получение точки, которая удалена от множества U_* на расстояние $\leq \varepsilon$, равно числу n , удовлетворяющему неравенствам $(b-a)/F_{n+2} \leq \varepsilon < (b-a)/F_{n+1}$.

Доказательство непосредственно следует из того, что $\delta_A(\Phi_n, b-a) = \gamma_n^A(b-a) = (b-a)/F_{n+2}$.

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

2. Найти такое наименьшее число N , чтобы $|F_n - ((1 + \sqrt{5})/2)^n / \sqrt{5}| < \varepsilon$ при всех $n \geq N$. Рассмотреть случаи $\varepsilon = 10^{-3}$, $\varepsilon = 10^{-4}$.

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

4. Применить метод Φ_5 для решения задачи А для функций $J(u) = u$, $J(u) = |u|$ на отрезке $[-6, 7]$.

§ 8. Оптимальный последовательный метод для задачи Б

Пусть, как и в § 7, Q — класс функций, унимодальных на отрезке $[a, b]$, P_n — множество всех n -точечных последовательных методов p_n , $n \geq 1$. Обозначим $\delta_B(p_n, b-a)$ — гарантированную точность метода p_n для задачи Б на классе Q , $\gamma_n^B(b-a) = \inf_{p_n \in P_n} \delta_B(p_n, b-a)$ — наилучшую гарантированную

точность методов P_n на классе Q ; определение этих понятий, а также оптимального и ε -оптимального методов см. в § 5.

Теорема 1. Для задачи Б наилучшая гарантированная на классе Q точность равна $\gamma_n^B(b-a) = (b-a)/2F_{n+1}$, однако оптимального метода в P_n при $n > 1$ не существует.

Доказательство. При $n = 1$ независимо от выбора точки $u_1 \in [a, b]$, в которой вычисляется значение функции $J(u) \in Q$, в качестве приближения ко множеству U_* можем взять точку $v_1 = (a+b)/2$. При этом будет допущена погрешность $\rho(v_1, U_*) \leq (b-a)/2 = (b-a)/2F_2$. Очевидно, любой другой выбор точки v_1 может привести лишь к увеличению погрешности, так что $\gamma_1^B(b-a) = (b-a)/2F_2$.

Пусть теперь $n \geq 2$. Покажем, что для любого ε , $0 < \varepsilon < (b-a)/F_{n+1}$, можно построить последовательный метод $\Phi_n^{B, \varepsilon}$, для которого

$$(b-a)/2F_{n+1} < \delta_B(\Phi_n^{B, \varepsilon}, b-a) \leq (b-a)/2F_{n+1} + \varepsilon. \quad (1)$$

Строится этот метод так. На отрезке $[a, b]$ сначала реализуется описанный в § 7 метод Фибоначчи $\Phi_{n-1} = \Phi_{n-1}^A$, и находится отрезок $[a_{n-1}, b_{n-1}]$ длиной $b_{n-1} - a_{n-1} = 2(b-a)/F_{n+1}$, содержащий точку из U_* и точку

$\bar{u}_{n-1} = (a_{n-1} + b_{n-1})/2$ с вычисленным значением $J(\bar{u}_{n-1}) = \min_{1 \leq i \leq n-1} J(u_i)$; здесь $a_1 = a$, $b_1 = b$, $n \geq 2$. Затем полагается $u_n = \bar{u}_{n-1} + \varepsilon$, вычисляется значение $J(u_n)$ и определяется точка

$$v_n = \begin{cases} (a_{n-1} + u_n)/2, & \text{если } J(\bar{u}_{n-1}) \leq J(u_n), \\ (\bar{u}_{n-1} + b_{n-1})/2, & \text{если } J(\bar{u}_{n-1}) > J(u_n). \end{cases}$$

Метод $\Phi_n^{B, \varepsilon}$ описан. Принимая точку v_n в качестве приближения к U_* , допустим погрешность $\rho(v_n, U_*) \leq (u_n - a_{n-1})/2 = (b - a)/2F_{n+1} + \varepsilon/2$ для любых функций $J(u) \in Q$. Нетрудно видеть, что для функции $J(u) = u$ метод $\Phi_n^{B, \varepsilon}$ дает погрешность, в точности равную $(b - a)/2F_{n+1} + \varepsilon/2$. Это значит, что $\delta_B(\Phi_n^{B, \varepsilon}, b - a) = (b - a)/2F_{n+1} + \varepsilon/2$. Следовательно, метод $\Phi_n^{B, \varepsilon}$ удовлетворяет условию (1) при всех $n \geq 2$.

Далее, покажем, что

$$\delta_B(p_n, b - a) > (b - a)/2F_{n+1} \text{ при всех } p_n \in P_n, n \geq 2. \quad (2)$$

Нетрудно видеть, что при $n = 2$ неравенство (2) справедливо. Отсюда и из (1) при $n = 2$ следует, что $\gamma_2^B(b - a) = (b - a)/2F_3 < \delta_B(p_2, b - a)$, $p_2 \in P_2$. Сделаем индуктивное предположение: пусть $\gamma_k^B(b - a) = (b - a)/2F_{k+1} < \delta_B(p_k, b - a)$ при всех $p_k \in P_k$, $k = 2, \dots, n - 1$, и докажем, что тогда

$\gamma_n^B(b - a) = (b - a)/2F_{n+1} < \delta_B(p_n, b - a)$ для любого метода $p_n \in P_n$; $n \geq 3$. Возьмем произвольный метод $p_n \in P_n$. Согласно определению 5.3 последовательного метода p_n сначала выбирается первая порция точек $u_1, \dots, u_s \in [a, b]$, $1 \leq s \leq n$, и по ним находится тройка $a_s, b_s, \bar{u}_s$, локализирующая минимум функции $J(u) \in Q$. Как и в задаче А, не умаляя общности, можем считать, что $2 \leq s \leq n$ (см. доказательство теоремы 7.1).

Отдельно рассмотрим случай $s = 2 \leq n$, когда метод p_n начинается с выбора двух точек u_1, u_2 , $a \leq u_1 < u_2 \leq b$. Начальные точки метода Φ_{n-1}^A на $[a, b]$ переобозначим через $v_1 = a + (b - a)F_{n-1}/F_{n+1}$, $v_2 = a + (b - a)F_n/F_{n+1}$. Возможно, что $u_1 \leq v_1$. Тогда в худшем случае $U \subset [u_1, b]$, причем для поиска минимума на $[u_1, b]$ мы можем вычислить значение функции еще в $n - 1$ точках этого отрезка, включая точку $\bar{u}_2 = u_2$. В силу предположения индукции при любом выборе точек $u_3, \dots, u_n$ и в любом местоположении точки $\bar{u}_2$ на $[u_1, b]$ приближение к U_* можно получить с гарантированной точностью, заведомо большей величины $\gamma_{n-1}^B(b - a) = (b - u_1)/2F_n \geq (b - v_1)/2F_n = (b - a)/2F_{n+1}$. Это значит, что неравенство (2) верно для любого метода p_n , начинающегося с выбора двух точек u_1, u_2 , $a \leq u_1 < u_2 \leq b$, где $u_1 \leq v_1$. Аналогичные рассуждения показывают справедливость (2) и при $u_2 \geq v_2$.

Пусть теперь $u_1 > v_1$. В худшем случае множество U может находиться на отрезке $[a, u_1]$, и для установления этого факта мы должны сделать по крайней мере три вычисления значений функции, причем хотя бы одно из них во внутренней точке отрезка $[a, u_1]$. Следовательно, для поиска минимума на $[a, u_1]$ мы можем использовать значения функции самое большее в $n - 2$ точках этого отрезка. Однако по предположению индукции $\gamma_{n-2}^B(u_1 - a) = (u_1 - a)/2F_{n-1} > (v_1 - a)/2F_{n-1} = (b - a)/2F_{n+1}$, так что (2) справедливо и при $u_1 > v_1$. Аналогично доказывается неравенство (2), если $u_2 < v_2$. Случай $s = 2$ рассмотрен.

Перейдем к рассмотрению методов $p_n \in P_n$, начинающихся с выбора первой порции из s точек $u_1, \dots, u_s \in [a, b]$, $2 < s \leq n$, по которым затем

находится тройка $a_s, b_s, \bar{u}_s$, локализирующая минимум $J(u)$ на $[a, b]$. Согласно следствию 6.1 при любых даже самых наилучших действиях на классе Q будем иметь $b_s - a_s \geq (b - a)/(m + 1)$, где $m = s/2 \geq 2$ при четном s , $m = (s - 1)/2 \geq 1$ при нечетном s .

Далее нам понадобятся неравенства

$$\begin{aligned} 2F_{n+1} &> (s + 1) F_{n-s+2} \quad \text{при} \quad s = 2m + 1, \quad 3 \leq s \leq n, \\ 2F_{n+1} &> (s + 2) F_{n-s+2} \quad \text{при} \quad s = 2m, \quad 4 \leq s \leq n, \end{aligned} \quad (3)$$

справедливость которых при $s \leq n - 1$ вытекает из неравенств (7.4), (7.5), а при $s = n$ они легко доказываются с помощью индукции по n .

Если в методе p_n оказалось $s = n$, то $\delta_B(p_n, b - a) \geq (b_n - a_n)/2 \geq (b - a)/2(m + 1) > (b - a)/2F_{n+1}$ в силу неравенства (3) при $s = n$. Поэтому пусть $3 \leq s \leq n - 1$. Однако в этом случае первые s шагов метода Φ_{n-1}^A приводят к отрезку длины $\Delta_s(\Phi_{n-1}^A) = (b - a) F_{n-s+2}/F_{n+1} < (b - a)/(m + 1) \leq b_s - a_s$, что вытекает из неравенств (3). Для получения отрезка $[a_s, b_s]$ нам понадобилось s вычислений значений функции, поэтому для поиска минимума на $[a_s, b_s]$ мы можем вычислить значение функции еще в $n - s + 1$ точках этого отрезка, включая точку $\bar{u}_s$ с известным $J(\bar{u}_s)$. Однако по предположению индукции $\gamma_{n-s+1}^B(b_s - a_s) = (b_s - a_s)/2F_{n-s+2}$,

поэтому $\delta_B(p_n, b - a) \geq \frac{b_s - a_s}{2F_{n-s+2}} > \frac{\Delta_s(\Phi_{n-1}^A)}{2F_{n-s+2}} = \frac{b - a}{2F_{n+1}}$. Неравенство (2) доказано полностью. Из (1), (2) следуют все утверждения теоремы 1.

Неравенство (1) означает, что для задачи Б метод $\Phi_n^{B, \varepsilon}$ является ε -оптимальным на классе Q . Метод $\Phi_n^{B, \varepsilon}$ будем называть методом Фибоначчи для задачи Б. При практическом применении этого метода следует иметь в виду, что первая точка метода Φ_{n-1}^A , являющегося составной частью метода $\Phi_n^{B, \varepsilon}$, может быть задана неточно, и тогда реализация метода $\Phi_n^{B, \varepsilon}$ будет сопровождаться быстрым ростом погрешности. Поэтому на практике вместо Φ_{n-1}^A здесь надо пользоваться указанной в § 7 модификацией этого метода.

Сформулируем два следствия из теоремы 1.

Следствие 1. Количество необходимых при решении задачи Б вычислений значений функции $J(u) \in Q$, гарантирующее на Q получение точки, которая удалена от множества U_* на расстояние, меньшее ε , равно числу n , удовлетворяющему неравенствам $(b - a)/2F_{n+1} < \varepsilon \leq (b - a)/2F_n$.

Следствие 2. Для любого последовательного метода p_n на отрезке $[a, b]$ существует функция $J(u) \in Q$ такая, что для тройки $a_n, b_n, \bar{u}_n$, полученной методом p_n и локализирующей минимум $J(u)$ на $[a, b]$, справедливо неравенство $b_n - a_n > (b - a)/F_{n+1} = 2\gamma_n^B(b - a)$. С другой стороны, существует метод Фибоначчи $\Phi_n^{B, \varepsilon}$, для которого $b_n - a_n = (b - a)/F_{n+1} + \varepsilon$ при всех $J(u) \in Q$, $0 < \varepsilon < (b - a)/F_{n+1}$.

Иначе говоря,

$$\inf_{p_n \in P_n} \sup_{J \in Q} (b_n - a_n) = 2\gamma_n^B(b - a) = (b - a)/F_{n+1}$$

— это и есть решение задачи Б' из § 5 последовательными методами.

Подведем некоторые итоги. В §§ 6—8 проведено исследование пассивных и последовательных методов решения задач А и Б на классе Q функций, унимодальных на отрезке $[a, b]$, построены оптимальные и ε -оптимальные методы. Выяснилось, что в зависимости от рассматриваемых постановок задач,

используемых методов вопрос существования оптимального метода решается по-разному: в одних случаях оптимальный метод существует и единственный, в других — оптимального метода нет, встретился также случай, когда оптимальных методов много. Из доказанных в §§ 6—8 теорем следует, что пассивные методы на классе Q значительно уступают по точности последовательным методам. Кроме того, $\gamma_n^B(b-a) \leq \gamma_n^A(b-a)$, т. е. в задаче Б удастся, вообще говоря, лучше приблизиться ко множеству точек минимума, чем в задаче А, как при использовании пассивных, так и последовательных методов.

Во избежание быстрого роста погрешности из-за неточности в задании первой точки на практике вместо методов Φ_n^A , $\Phi_n^{B, \varepsilon}$ нужно использовать указанные выше их модификации. Следует также заметить, что методы Φ_n^A , $\Phi_n^{B, \varepsilon}$ для своей реализации требуют, чтобы число n вычислений значений минимизируемой функции было задано заранее — выбор первой точки в этих методах 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 знания n . Для сравнения вспомним, что методом золотого сечения за n вычислений значений функции мы получали отрезок $[a_n, b_n]$ длиной $b_n - a_n = ((\sqrt{5} - 1)/2)^{n-1}(b-a) = (2/(\sqrt{5} + 1))^{n-1}(b-a)$, а метод $\Phi_n^{B, \varepsilon}$ согласно следствию 2 дает отрезок, длина которого близка к $(b-a)/F_{n+1} \approx \approx (2/(\sqrt{5} + 1))^{n+1}(b-a)/\sqrt{5}$ для достаточно больших n . Отсюда следует, что метод золотого сечения хуже метода $\Phi_n^{B, \varepsilon}$ при больших n всего в $((\sqrt{5} + 1)/2)^2/\sqrt{5} = 1,1708 \dots$ раз, т. е. на классе унимодальных функций метод золотого сечения близок к оптимальным методам. На практике возможно сочетание метода золотого сечения с методами Фибоначчи: на первом этапе поиска минимума можно применять метод золотого сечения, а затем, задавшись целым числом $n \geq 1$, можно перейти к методам Фибоначчи Φ_n^A или $\Phi_n^{B, \varepsilon}$ и через n шагов закончить поиск. Здесь полезно также заметить, что

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

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

Упражнения. 1. Указать все точки метода $\Phi_n^{B, \varepsilon}$ при $n = 2, 3, 4$.

2. Применить метод $\Phi_5^{B, \varepsilon}$ для решения задачи Б для функций $J(u) = u$, $J(u) = |u|$ на отрезке $[-6, 5]$.

3. Найти наименьшее n , начиная с которого точность метода золотого сечения больше точности методов Фибоначчи в 2 раза; 1,5 раза.

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

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

6. Используя утверждения упражнений 4 и 5, доказать, что метод золотого сечения является единственным симметричным методом, удовлетворяющим условию (4.4) при всех $n = 1, 2, \dots$

7. Пусть дан симметричный метод с начальными отрезками Δ_1 , Δ_2 , пусть $N \geq 2$ — заданное натуральное число. Используя утверждения упражнений 4,

5, указать промежуток изменения отношения Δ_2/Δ_1 , чтобы метод удовлетворял условию (4.4) при всех $n = 1, 2, \dots, N$.

8. Пусть дан некоторый симметричный метод, удовлетворяющий условию (4.4) при $n = 1$. Используя утверждения упражнений 4, 5, указать максимальное число N , при котором условие (4.4) выполняется для всех $n = 2, 3, \dots, N$.

9. Пусть на отрезке $[a, b]$ задан симметричный метод, начинающийся с точки u_1 (см. § 4), пусть $\Delta_n = b_n - a_n$ — длины получаемых этим методом отрезков $[a_n, b_n]$, $n = 1, 2, \dots$; $\Delta_1 = b - a$; $\Delta_2 = \max\{b - u_1; a - u_1\}$. Пусть $\tilde{\Delta}_n = \Delta_n + \delta_n$, где $\tilde{\Delta}_n$ — приближение для Δ_n , δ_n — погрешность, возникающая из-за неточного задания Δ_2 : $\tilde{\Delta}_2 = \Delta_2 + \delta$; Δ_1 известно точно. Доказать, что тогда $|\delta_{k_1 + \dots + k_m + i}| \geq i F_m \delta$, $i = 1, 2, \dots, k_{m+1}$; $m = 1, 2, \dots$, где $k_1 = 1$, а $k_2, k_3, \dots, k_m, k_{m+1}, \dots$ — последовательные номера, для которых выполняется условие (4.4).

§ 9. 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 важным 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 являющихся унимодальными.

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

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

$$|J(u) - J(v)| \leq L|u - v|, \quad u, v \in [a, b]. \quad (1)$$

Постоянную L называют константой 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 две точки $(u, J(u))$, $(v, J(v))$ графика функции, не превышает константы 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 $[a_i, a_{i+1}]$, $i = 1, 2, \dots, m$, где $a_1 = a$, $a_{m+1} = b$, удовлетворяет условию (1) с константой L_i . Тогда $J(u)$ удовлетворяет условию (1) на всем отрезке с константой $L = \max_{1 \leq i \leq m} L_i$.

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

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

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 константой $L = \sup_{u \in [a, b]} |J'(u)|$.

Доказательство. По формуле конечных приращений для любых $u, v \in [a, b]$ имеем $J(u) - J(v) = J'(v + \theta(u - v)) \times (u - v)$, $0 < \theta < 1$. 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 , $a \leq u \leq b$. 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)

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

т. е.

$$g(u, v) = J(v) - L|u - v| \leq J(u) \quad (2)$$

при всех $u \in [a, b]$, 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 [187], 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 $p_0(u_1) =$

$= \min_{u \in [a, b]} p_0(u)$, $u_1 \in [a, b]$; 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 $p_1(u_2) = \min_{u \in [a, b]} p_1(u)$, $u_2 \in [a, b]$ и т. д. (рис. 1.2). Пусть точки $u_0, u_1, \dots, u_n$ ($n \geq 1$) уже известны. Тогда составляется функция

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

и следующая точка u_{n+1} определяется условиями

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

$$u_{n+1} \in [a, b]. \quad (3)$$

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 константой L , что и функция $J(u)$. Ясно также, что

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

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

$$p_n(u) \leq J(u), \quad u \in [a, b], \quad n = 0, 1, \dots \quad (5)$$

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 последователь-

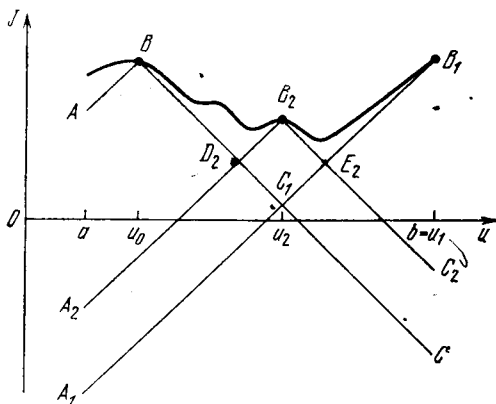


Рис. 1.2. ABC — график $p_0(u) = g(u, u_0)$, A_1B_1 — график $g(u, u_1)$, ABC_1B_1 — график $p_1(u)$, $A_2B_2C_2$ — график $g(u, u_2)$, $ABD_2B_2E_2B_1$ — график $p_2(u)$.

ность $\{u_n\}$, полученная с помощью описанного метода ломаных, такова, что

1) $\lim_{n \rightarrow \infty} J(u_n) = \lim_{n \rightarrow \infty} p_n(u_{n+1}) = J_* = \inf_{u \in [a, b]} J(u)$, причем справедлива оценка

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

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

Доказательство. Возьмем произвольную точку $u_* \in U_*$. С учетом условий (3) и неравенств (4), (5) имеем $p_{n-1}(u_n) = \min_{u \in [a, b]} p_{n-1}(u) \leq p_{n-1}(u_{n+1}) \leq p_n(u_{n+1}) = \min_{u \in [a, b]} p_n(u) \leq p_n(u_*) \leq J(u_*) = J_*$, т. е. последовательность $\{p_n(u_{n+1})\}$ монотонно возрастает и ограничена сверху. Отсюда сразу следует оценка (6) и существование предела $\lim_{n \rightarrow \infty} p_n(u_{n+1}) = p_* \leq J_*$. Покажем, что

$p_* = J_*$. Пусть v_* — какая-либо предельная точка последовательности $\{u_n\}$. Тогда по теореме Больцано — Вейерштрасса существует подпоследовательность $\{u_{n_k}\}$, сходящаяся к v_* , причем можем считать, что $n_1 < n_2 < \dots < n_{k-1} < n_k < \dots$. Заметим, что $J(u_i) = g(u_i, u_i) \leq p_n(u_i) \leq J(u_i)$, т. е. $J(u_i) = p_n(u_i)$ при всех $i = 0, 1, \dots, n$. Тогда $0 \leq p_n(u_i) - \min_{u \in [a, b]} p_n(u) = J(u_i) - p_n(u_{n+1}) = p_n(u_i) - p_n(u_{n+1}) \leq L|u_i - u_{n+1}|$ при любом n и $i = 0, 1, \dots, n$. Принимая здесь $n = n_k - 1$, $i = n_{k-1} \leq n_k - 1$, получим $0 \leq J(u_{n_{k-1}}) - p_{n_k-1}(u_{n_k}) \leq L|u_{n_{k-1}} - u_{n_k}|$, $k \geq 2$. Отсюда при $k \rightarrow \infty$ имеем $J_* \leq J(v_*) = \lim_{k \rightarrow \infty} J(u_{n_{k-1}}) = \lim_{k \rightarrow \infty} p_{n_k-1}(u_{n_k}) = p_* \leq J_*$. т. е. $\lim_{k \rightarrow \infty} J(u_{n_k}) = \lim_{k \rightarrow \infty} p_{n_k-1}(u_{n_k}) = p_* = J_*$. 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 [95] 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 [247].

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 следующем параграфе 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 константы 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

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

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

$$J(v) \geq (J(w) - J(u))(v - u)/(w - u) + J(u) \quad (8)$$

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

Пусть $a = v_0 < v_1 < \dots < v_m = b$; обозначим $L_m = \max_{1 \leq i \leq m} |J(v_i) - J(v_{i-1})| \cdot |v_i - v_{i-1}|^{-1}$. Ясно, что $L_m \leq L$.

Пусть при каждом $m \geq 1$ величина L_{m+1} вычисляется по точкам $a = w_0 < w_1 < \dots < w_{m+1} = b$, полученным добавлением к точкам $v_0, \dots, v_m$ одной новой точки. Тогда согласно (7) имеем $L_m \leq L_{m+1} \leq L$, $m \geq 1$. Это значит, что с возрастанием m величины L_m все лучше и лучше приближают L снизу. Если $\max_{0 \leq i \leq m} |v_i - v_{i-1}| \rightarrow 0$ при $m \rightarrow \infty$, то $\lim_{m \rightarrow \infty} L_m = L$. Приведен-

ные соображения могут помочь в получении оценки для L . При определении L могут быть полезны теоремы 1 и 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 $J(u) = a_n u^n + a_{n-1} u^{n-1} + \dots + a_1 u + a_0$ — многочлен n -й степени на отрезке $[a, b]$, где $0 < a < b$. Обозначим

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

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

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

§ 10. Методы пассивного и последовательного перебора

1. Обозначим через $Q(L)$ класс функций, удовлетворяющих условию Липшица (9.1) на отрезке $[a, b]$, с одной и той же для всех функций этого класса константой $L > 0$. Для функций $J = J(u) \in Q(L)$ будем рассматривать задачу минимизации первого типа, когда ищется величина $J_* = \inf_{u \in [a, b]} J(u)$. Для решения этой задачи будем пользоваться методами p_n , которые

заключаются в выборе точек $u_1, \dots, u_n$, $a \leq u_1 < u_2 < \dots < u_n \leq b$, вычислении значений функции $J(u_1), \dots, J(u_n)$ и определении величины $J(u_k) = \min_{1 \leq i \leq n} J(u_i)$, принимаемой за приближение к J_* . Введем величину

$$\Delta(J, p_n) = \min_{1 \leq i \leq n} J(u_i) - J_*,$$

которую будем называть погрешностью метода p_n при минимизации функции $J = J(u) \in Q(L)$. Очевидно, всегда $\Delta(J, p_n) \geq 0$. Для отдельных функций возможно, что $\Delta(J, p_n) = 0$, но для не-

которых «плохих» для метода p_n функций из $Q(L)$ погрешность $\Delta(J, p_n)$ может оказаться значительной. Величину $\sup_{J \in Q(L)} \Delta(J, p_n) = \delta(p_n)$, выражающую погрешность метода p_n для самой «плохой» для p_n функции из $Q(L)$, назовем гарантированной точностью метода p_n на классе $Q(L)$ (см. определение 5.1).

Возникает вопрос: как выбрать число n и метод $p_n = \{u_1, \dots, u_n\}$, чтобы

$$\sup_{J \in Q(L)} \left(\min_{1 \leq i \leq n} J(u_i) - J_* \right) = \delta(p_n) \leq \varepsilon, \quad (1)$$

где $\varepsilon > 0$ — заданная точность?

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

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

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

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 (9.2) для любого $u \in [u_i - h/2; u_i + h/2]$ имеем $J(u) \geq J(u_i) - L|u - u_i| \geq J(u_i) - Lh/2 \geq \min_{1 \leq i \leq n} J(u_i) - \varepsilon$ при всех $i = 1, \dots, n$. Так как

система отрезков $[u_i - h/2, u_i + h/2]$, $i = 1, 2, \dots, n$, 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 $J(u) \geq \min_{1 \leq i \leq n} J(u_i) - \varepsilon$ для всех $u \in [a, b]$.

Поэтому $J_* \geq \min_{1 \leq i \leq n} J(u_i) - \varepsilon$ 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 $\min_{1 \leq i \leq n} (Lu_i) - La = Lh/2 > \varepsilon$.

3. Метод равномерного перебора (2) относится к пассивным методам (см. определение 5.4), когда точки $u_1, \dots, u_n$ задаются все одновременно до начала вычислений значений функции. На классе $Q(L)$ можно предложить такой же простой, но более эффективный последовательный метод перебора (см. определение 5.3), когда выбор точки u_i при каждом $i > 2$ производится с учетом вычислений значения функции в предыдущих точках $u_1, \dots, u_{i-1}$ и задачу (1) удастся решить, вообще говоря, за

меньшее количество вычислений значений функции, чем методом (2). А именно, следуя работе [104], положим

$$u_1 = a + h/2, \quad u_{i+1} = u_i + h + (J(u_i) - F_i)/L, \quad i = 1, \dots, n-2, \\ u_n = \min \{u_{n-1} + h + (J(u_{n-1}) - F_{n-1})/L; b\}, \quad (3)$$

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

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 (9.2) для всех $u \in [u_i, u_i + h/2 + (J(u_i) - F_i)/L]$ имеем $J(u) \geq J(u_i) - L(h/2 + (J(u_i) - F_i)/L) = F_i - h/2 \geq F_n - \varepsilon$. Аналогично для всех $u \in [u_i - h/2, u_i]$ получим $J(u) \geq J(u_i) - Lh/2 \geq F_n - \varepsilon$. Так как система отрезков $[u_i - h/2, u_i + h/2 + (J(u_i) - F_i)/L]$, $i = 1, \dots, n$, покрывает весь отрезок $[a, b]$, то из предыдущих неравенств следует, что $J(u) \geq F_n - \varepsilon$ при всех $u \in [a, b]$. Тогда $J_* \geq F_n - \varepsilon$ при всех $J = J(u) \in Q(L)$, что равносильно неравенству (1).

В худшем случае, когда, например, $J(u)$ постоянна или монотонно убывает на $[a, b]$ и, следовательно, $F_i = \min_{1 \leq j \leq i} J(u_j) = J(u_i)$, 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 $J(u) = A + L(u-B)$, $A, B = \text{const}$ и, следовательно, $F_i = J(u_i)$, $u_i = u_1 + (2^{i-1} - 1)h$, $i = 1, 2, \dots, n-2$, 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 § 9, но метод (3) выгодно отличается простотой реализации и не требует большой машинной памяти. Недостатком метода (3), как и метода ломаных, является необходимость априорного знания константы L из условия (9.1).

4. 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 $\sup_{u \in [a, b]} J''(u) \leq M$, где M — некоторая фиксированная константа. Обозначим этот класс функций через $R(M)$. Заметим, что если $M \leq 0$, то $J''(u) \leq 0$ и, следовательно, $J'(u)$ монотонно убывает на $[a, b]$. Это значит, что тогда функция $J(u)$ достигает своей нижней грани при $u = a$ или $u = b$. Таким образом, задача минимизации функций из класса $R(M)$ в случае $M \leq 0$ 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 [77]:

$$u_1 = a; \quad u_{i+1} = u_i + \sqrt{2\epsilon/M} + h_i, \quad i = 1, \dots, n-2; \quad u_n = b, \quad (4)$$

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

Теорема 3. Применяя метод (4), задачу минимизации первого типа для любой функции $J(u) \in R(M)$ можно решить с заданной точностью $\epsilon > 0$, т. е.

$$0 \leq \min_{1 \leq i \leq n} J(u_i) - J_* \leq \epsilon. \quad (5)$$

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 (5), очевидно, справедливо. Поэтому пусть $a < u_* < b$. Тогда $J'(u_*) = 0$ и, используя разложение по Тейлору в точке u_* для $J(u) \in R(M)$, получим

$$J(u) = J_* + J''(\xi)(u - u_*)^2/2 \leq J_* + M(u - u_*)^2/2, \quad (6)$$

где $\xi = u_* + \theta(u - u_*)$, $0 < \theta < 1$. Система отрезков $[u_i - \sqrt{2\epsilon/M}, u_i + h_i]$, $i = 1, 2, \dots, n$, покрывает отрезок $[a, b]$, поэтому u_* попадет в один из отрезков этой системы при некотором i . Если $u_i - \sqrt{2\epsilon/M} \leq u_* \leq u_i$, то из (6) при $u = u_i$ имеем $J(u_i) - J_* \leq M/2 \cdot 2\epsilon/M = \epsilon$. Если $u_i \leq u_* \leq u_i + h_i$, то аналогично $J(u_i) - J_* \leq Mh_i^2/2 = \epsilon + J(u_i) - F_i$, или $F_i - J_* \leq \epsilon$. Объединяя оба случая, получаем требуемое неравенство (5).

В худшем случае (например, если $J(u) = M(u - b)^2/2$) может оказаться, что $F_i = J(u_i)$, $i \geq 1$, и тогда метод (4) превратится в метод равномерного перебора с шагом $h = \sqrt{2\epsilon/M}$. Если же $F_i < J(u_i)$ при некоторых i (например, для $J(u) = M(u - a)^2/2$), то методом (4) удается получить неравенство (5), вообще говоря, при меньшем n , чем методом равномерного перебора. Недостатком метода (4) является требование знания константы $M \geq \sup_{u \in [a, b]} J''(u)$.

Упражнения. 1. Пусть одним из методов (2), (3) или (4) найден $\min_{1 \leq i \leq n} J(u_i) = J(u_k)$. Можно ли принять u_k за приближение ко множеству U_* ? Оценить погрешность $\rho(u_k, U_*)$ для метода (2) на классе $Q(L)$; рассмотреть функцию $J(u) = L(u - a) - \epsilon/2$ при $a \leq u \leq a + \epsilon/L$, $J(u) = \epsilon(b - u)/(2(b - a - \epsilon/L))$ при $a + \epsilon/L \leq u \leq b$, где $\epsilon > 0$ — малое число. Оценить $\rho(u_k, U_*)$ для методов (3), (4) на классах $Q(L)$ и $R(M)$ соответственно.

2. Найти оптимальный пассивный и оптимальный последовательный методы на классе функций $Q(L)$. [5], [216].

§ 11. Выпуклые функции одной переменной

Рассмотрим класс функций, для которых существует более эффективный вариант метода ломаных, когда ломаные состоятся 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 $a \leq u \leq b$, называется *выпуклой* на этом отрезке, если

$$J(\alpha u + (1 - \alpha)v) \leq \alpha J(u) + (1 - \alpha)J(v) \quad (1)$$

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

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

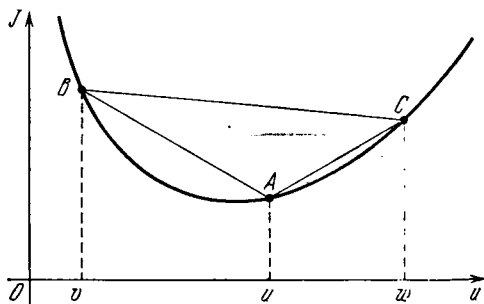


Рис. 1.3

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 $J(u) = u^2$, $J(u) = |u|$, $J(u) = u$.

Наряду с выпуклыми функциями в литературе иногда рассматривают вогнутые функции.

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

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

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

$$(J(u) - J(v))/(u - v) \leq (J(w) - J(v))/(w - v) \leq (J(w) - J(u))/(w - u) \quad (2)$$

при всех $u, v, w, a \leq v < u < w \leq b$.

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 $J(u)$ имеем $J(u) \leq (w - u)J(v)/(w - v) + (1 - (w - u)/(w - v))J(w)$ или

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

Последнее неравенство можно переписать в двойкой форме:

$$(\omega - v)(J(u) - J(v)) \leq (\omega - v)(J(\omega) - J(v))$$

или

$$(\omega - u)(J(\omega) - J(v)) \leq (\omega - v)(J(\omega) - J(u)),$$

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 u отрезка $[a, b]$ непрерывна и имеет конечную правую производную $\lim_{h \rightarrow +0} (J(u+h) - J(u))/h = J'(u+0)$, конечную левую производную $\lim_{\tau \rightarrow +0} (J(u) - J(u-\tau))/\tau = J'(u-0)$, причем $J'(u-0) \leq J'(u+0)$ при всех $u \in \text{int}[a, b] = \{u: a < u < b\}$.*

Доказательство. Из теоремы 1 следует, что

$$(J(u) - J(u-\tau))/\tau \leq (J(u) - J(u-h))/h \leq (J(u+h) - J(u))/h \leq (J(u+\tau) - J(u))/\tau \quad (3)$$

при всех τ, h , лишь бы $0 < h < \tau$ и точки $u, u \pm h, u \pm \tau \in \text{int}[a, b]$ (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 производной $J'(u+0)$. Аналогично доказывается существование левой производной $J'(u-0)$. Из (3) при $h \rightarrow +0$ получаем неравенство $J'(u-0) \leq J'(u+0)$.

Из существования левой и правой производных следует непрерывность $J(u)$ при всех $u \in \text{int}[a, b]$.

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

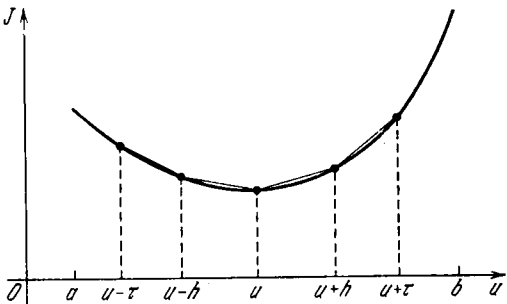


Рис. 1.4.

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

$$J'(a+0)(u-v) \leq J(u) - J(v) \leq J'(b-0)(u-v) \quad (4)$$

при всех u, v , $a \leq v \leq u \leq b$, так что $J(u)$ на $[a, b]$ удовлетворяет условию Липшица (9.1) с константой $L = \max\{|J'(a+0)|, |J'(b-0)|\}$.

Доказательство. Из теоремы 1 имеем

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

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

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

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 $v \rightarrow a+0$ или $u \rightarrow b-0$.

Пример 2 показывает, что конечность $J'(a+0)$, $J'(b-0)$ существенна для выполнения условия Липшица (9.1).

Теорема 4. Пусть функция $J(u)$ выпукла на отрезке $[a, b]$, а $l(v)$ — любая функция, удовлетворяющая неравенствам $J'(v-0) \leq l(v) \leq J'(v+0)$ при $a < v < b$. Тогда $l(v)$ не убывает при $v \in \text{int}[a, b]$, и справедливо неравенство

$$J(u) \geq J(v) + l(v)(u-v), \quad u \in [a, b]. \quad (5)$$

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

$$J(u) \geq J(v) + J'(v)(u-v), \quad u \in [a, b], \quad (6)$$

при любом $v \in [a, b]$. Если неравенство (5) (или (6)) обращаются 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 $J(v + \alpha(u-v)) - J(v) \leq \alpha(J(u) - J(v))$, $0 < \alpha < 1$. Разделив обе части этого неравенства на α и переходя к пределу при $\alpha \rightarrow +0$, получим $J(u) - J(v) \geq J'(v+0)(u-v) \geq l(v)(u-v)$.

при $u > v$ и $J(u) - J(v) \geq J'(v - 0)(u - v) \geq l(v)(u - v)$ при $u < v$. Неравенство (5) доказано. Заметим, что при $a < u < b$ переменные u, v в (5) входят равноправно, поэтому, меняя их ролями, получим $J(v) \geq J(u) + l(u)(v - u)$ при всех $v \in [a, b]$. Сложим последнее неравенство с (5) почленно. Будем иметь $(l(u) - l(v))(u - v) \geq 0$ при всех $u, v \in \text{int}[a, b]$, 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 $u \in \text{int}[a, b]$. Из существования конечных $J'(a + 0)$, $J'(b - 0)$ 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 $J(u) \leq \alpha J(c) + (1 - \alpha)J(v) = \alpha(J(v) + l(v)(c - v)) + (1 - \alpha)J(v) = J(v) + l(v)(u - v)$, $u \in [c, v]$. 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 $J'(v - 0) \leq l(v) \leq J'(v + 0)$, 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 $u \in \text{int}[a, b]$ (если существуют конечные $J'(a + 0)$, $J'(b - 0)$), 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 $\lim_{u \rightarrow a+0} J(u) = J(a)$, $\lim_{u \rightarrow b-0} J(u) = J(b)$. 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

$$J'(u_* + 0) \geq 0, \quad J'(u_* - 0) \leq 0 \quad (7)$$

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

Доказательство. Из условий на функцию $J(u)$ и теореме 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 $J(u_* + h) - J(u_*) \geq 0$ при всех достаточно малых $|h|$, для которых $u_* + h \in [a, b]$. Разделив это неравенство на $h > 0$ и $h < 0$ и устремляя 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 $h \rightarrow +0$ 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 $u_* \in \text{int}[a, b]$ удовлетворяет условию (7). Положим в неравенстве (5) $v = u_*$, $l(v) = 0$, и получим, что $J(u) \geq J(u_*)$ 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 $\lim_{u \rightarrow a+0} J(u) = J(a)$, $\lim_{u \rightarrow b-0} J(u) = J(b)$; пусть U_* — множество точек минимума $J(u)$ на $[a, b]$, и v — некоторая точка, $a < v < b$. Тогда для того чтобы $U_* \cap [a, v] = \emptyset$ (или $U_* \cap [v, b] = \emptyset$), необходимо и достаточно выполнения неравенства $J'(v + 0) < 0$ (или $J'(v - 0) > 0$ соответственно). Для того чтобы $U_* \cap [a, v] \neq \emptyset$ (или $U_* \cap [v, b] \neq \emptyset$), необходимо и достаточно, чтобы $J'(v + 0) \geq 0$ (или $J'(v - 0) \leq 0$ соответственно).

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 $U_* \cap [a, v] = \emptyset$. Если $J'(v - 0) > 0$, то аналогично получаем $J'(u - 0) > 0$ при всех $u \in [v, b]$, так что $U_* \cap [v, b] = \emptyset$.

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

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 $\lim_{u \rightarrow a+0} J(u) = J(a)$, $\lim_{u \rightarrow b-0} J(u) = J(b)$, 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 $J_* = \inf_{u \in [a, b]} J(u) \leq J(\alpha u_* + (1 - \alpha) v_*) \leq \alpha J(u_*) + (1 - \alpha) J(v_*) = J_*$. Это значит, что $J(\alpha u_* + (1 - \alpha) v_*) = J_*$ при всех $\alpha \in [0, 1]$, т. е. $U_* = [u_*, v_*]$.

Далее, так как $U_* \cap [a, v] = \emptyset$ для любого v , $a \leq v < u_*$, то по теореме 6 $J'(v + 0) < 0$ при $a \leq v < u_*$. А тогда $J'(v + 0) \leq (J(v + h) - J(v))/h < 0$ при всех достаточно малых h , т. е. $J(u)$ строго монотонно убывает при $a \leq u \leq u_*$. Аналогично показывается, что при $v_* \leq u \leq b$ функция $J(u)$ строго монотонно возрастает. Теорема 7 доказана.

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 $0 < u \leq 1$, $J(0) = 1$. Функция $J(u)$ выпукла на $[0, 1]$, но множество $U_* = \{u: 0 < u \leq 1\}$ не является отрезком. Здесь $\lim_{u \rightarrow +0} J(u) \neq J(0)$ — 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 $a \leq v < u < w \leq b$. Применяя формулу Лагранжа, имеем

$$(J(u) - J(v))/(u - v) = J'(\xi_1),$$

$$(J(w) - J(u))/(w - u) = J'(\xi_2),$$

где $v < \xi_1 < u < \xi_2 < w$. По условию $J'(\xi_1) \leq J'(\xi_2)$, 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 $J''(u) \geq 0$ на $[a, b]$.

Доказательство. Условие $J''(u) \geq 0$ 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 $r \geq 1$ и $r \leq 0$ 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 $J'(u+0) = \inf_{h>0} (J(u+h) - J(u))/h$, $J'(u-0) = \sup_{h>0} (J(u) - J(u-h))/h$ 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 $J'(v-0) \leq J'(v+0) \leq J'(u-0) \leq J'(u+0)$ 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 $J'(u+0)$ 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 $[a, a_1]$, $[a_1, a_2]$, ..., $[a_{n-1}, a_n]$, $[a_n, b]$, $a < a_1 < \dots < a_n < b$, причем на каждом таком отрезке производная $J'(u)$ суммируема, не убывает и $J'(a_i-0) \leq J'(a_i+0)$, $i = 1, \dots, n$. Доказать, что тогда $J(u)$ выпукла на $[a, b]$.

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

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

8. Если функция $f(t)$ возрастает на отрезке $[a, b]$ и суммируема на этом отрезке, то функция $J(u) = \int_a^u f(t) dt$ 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 $\lim_{u \rightarrow a+0} J(u) = J(a)$, $\lim_{u \rightarrow b-0} J(u) = J(b)$, а U_* — множество точек минимума $J(u)$ на $[a, b]$. Доказать, что $U_* \cap [\alpha, \beta] \neq \emptyset$ (или $U_* \subset \text{int} [\alpha, \beta]$) тогда и только тогда, когда $J'(\alpha - 0) \leq 0$, $J'(\beta + 0) \geq 0$ (соответственно, $J'(\alpha - 0) < 0$, $J'(\beta + 0) > 0$); здесь $a < \alpha < \beta < b$.

11. Доказать, что выпуклая на отрезке $[a, b]$ функция $J(u)$, отличная от константы, 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 $J((u+v)/2) \leq (J(u) + J(v))/2$ 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 [235], стр. 119.

14. Пусть $J(u)$ — выпуклая функция при $u \geq 0$, $J(0) \leq 0$. Доказать, что тогда функция $\varphi(u) = J(u)/u$ монотонно возрастает при $u > 0$. На примере $J(u) = 1 + u^2$ убедиться, что при $J(0) > 0$ это утверждение неверно. Указание: воспользоваться равенством $J(u) = J\left(\frac{u}{u+h}(u+h) + \frac{h}{u+h} \cdot 0\right)$, $h > 0$.

15. Пусть функция $J(u)$ выпукла и дважды дифференцируема при $u \geq 0$, причем $\lim_{u \rightarrow +\infty} (uJ'(u) - J(u)) \leq 0$. Доказать, что тогда $\varphi(u) = J(u)/u$ монотонно убывает при $u > 0$. Указание: вычислить производные функций $\varphi(u)$ и $uJ'(u) - J(u)$.

16. Доказать, что $(a+b)^n \leq 2^{n-1}(a^n + b^n)$ при всех $n \geq 1$, $a \geq 0$, $b \geq 0$. Указание: воспользоваться выпуклостью функции $J(u) = u^n$ при $u \geq 0$, $n \geq 1$.

§ 12. Метод касательных

1. Пусть функция $J(u)$ выпукла и дифференцируема на отрезке $[a, b]$. Согласно теоремам 11.3 и 11.7 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 § 9. 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

$$g(u, v) = J(v) + J'(v)(u - v), \quad a \leq u \leq b.$$

Согласно теореме 11.4

$$g(u, v) \leq J(u) \text{ при всех } u \in [a, b]. \quad (1)$$

В качестве начального приближения возьмем любую точку $u_0 \in$

$\in [a, b]$ (например, $u_0 = a$), составим функцию $p_0(u) = g(u, u_0)$ и определим точку $u_1 \in [a, b]$ из условия $p_0(u_1) = \min_{u \in [a, b]} p_0(u)$ (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 $p_1(u_2) = \min_{u \in [a, b]} p_1(u)$ и т. д.

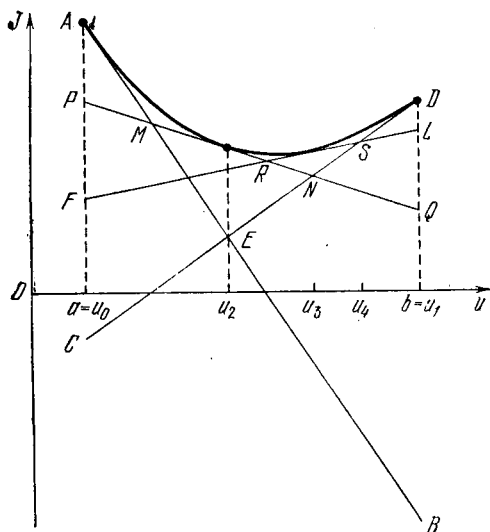


Рис. 1.5. AB — график $g(u, u_0)$, CD — график $g(u, u_1)$, AED — график $p_1(u)$, PQ — график $g(u, u_2)$, $AMND$ — график $p_2(u)$, FL — график $g(u, u_3)$, $AMRSD$ — график $p_3(u)$.

Если точки $u_0, u_1, \dots, u_n$, $n \geq 1$, уже известны, то составляем функцию $p_n(u) = \max \{p_{n-1}(u), g(u, u_n)\} = \max_{0 \leq i \leq n} g(u, u_i)$, и следующую точку u_{n+1} определим из условий

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

$$u_{n+1} \in [a, b].$$

Если при каком-либо $n \geq 0$ окажется $J'(u_n + 0) \geq 0$, $J'(u_n - 0) \leq 0$ (если $a < u_n < b$, то это равносильно условию $J'(u_n) = 0$), то согласно теореме 11.5 $u_n \in U_*$ — в этом случае задача минимизации уже решена и итерации на этом заканчиваются.

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 $u_0, \dots, u_n$ (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 $u_n \notin U_*$, $n = 0, 1, \dots$. Тогда

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

$$0 \leq J(u_{n+1}) - J_* \leq J(u_{n+1}) - p_n(u_{n+1}), \quad n = 1, 2, \dots;$$

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

Доказательство. Так как $J'(a+0)$, $J'(b-0)$ конечны по условию, то в силу теоремы 11.3 $J(u)$ удовлетворяет условию Липшица с константой $L = \max \{|J'(a)|; |J'(b)|\}$. Кроме того, согласно (1) и определению функции $p_n(u)$ имеем

$$p_{n-1}(u) \leq p_n(u) \leq J(u), \quad u \in [a, b], \quad n = 1, 2, \dots \quad (2)$$

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

$$J(u_i) = p_n(u_i), \quad i = 0, 1, \dots, n. \quad (3)$$

Наконец, угловые коэффициенты касательных $g(u, u_i)$ равны $J'(u_i)$, причем $|J'(u_i)| \leq L$. Из теоремы 9.1 тогда следует, что $p_n(u)$ удовлетворяет условию Липшица с константой L . Отсюда с учетом (2), (3) с помощью тех же рассуждений, которые применялись при доказательстве теоремы 9.3, нетрудно убедиться в справедливости всех утверждений доказываемой теоремы. Остается лишь заметить, что из того, что $J(u)$ унимодальна и $u_n \notin U_* = [u_*, v_*]$, $n \geq 0$, равенство $\lim_{n \rightarrow \infty} p(u_n, U_*) = 0$ 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 § 9. 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 $J'(a_1) \geq 0$ или $J'(b_1) \leq 0$, то по теореме 11.5 $a \in U_*$ или $b \in U_*$ — задача решена. Поэтому пусть $J'(a_1) < 0$, $J'(b_1) > 0$, то согласно теореме 11.6 означает $U_* \subset \text{int}[a, b]$. Пусть отрезок $[a_{n-1}, b_{n-1}]$, $n \geq 2$, уже построен, причем $J'(a_{n-1}) < 0$, $J'(b_{n-1}) > 0$, $U_* \subset \text{int}[a_{n-1}, b_{n-1}]$. Обозначим через u_n точку пересечения касательных $g(u, a_{n-1})$, $g(u, b_{n-1})$. 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

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

По построению $J'(a_n) < 0$, $J'(b_n) > 0$, и согласно теореме 11.6 $U_* \subset \text{int}[a, b]$. Индуктивное описание метода закончено. Из геометрических построений нетрудно усмотреть (рис. 1.5), что этот метод совпадает с описанным выше методом касательных, в ко-

тором за начальную точку берется $u_0 = a$; строгое доказательство этого факта приводится ниже 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 $J'(a_n) < 0, J'(b_n) > 0$:

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

3. Поскольку ломаная из отрезков касательных аппроксимирует функцию $J(u)$, вообще говоря, лучше, чем ломаные из § 9, то следует ожидать, что метод касательных для выпуклых функций сходится быстрее метода ломаных из § 9. Исследуем скорость сходимости метода касательных, считая минимизируемую функцию дважды дифференцируемой.

Теорема 2. Пусть функция $J(u)$ дважды непрерывно дифференцируема на $[a, b]$, $\inf_{u \in [a, b]} J''(u) > 0$, 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 $u_n \neq u_*, n = 0, 1, \dots$. Тогда для любого числа $\varepsilon > 0$ существует номер $N = N(\varepsilon)$ такой, что

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

Доказательство. Из теоремы 11.9 следует выпуклость $J(u)$ на $[a, b]$. Кроме того, так как $J'(u)$ строго возрастает, то множество U_* состоит из единственной точки u_* . Тогда из теоремы 1 имеем $\lim_{n \rightarrow \infty} u_n = u_*$. Поскольку $[a_{n+1}, b_{n+1}] \subset [a_n, b_n], n = 1, 2, \dots$, то последовательность $\{a_n\}$ монотонно возрастает, а $\{b_n\}$ — монотонно убывает, причем $a_n < u_* < b_n, n = 1, 2, \dots$. Покажем, что $\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = u_*$. В силу (4)

либо $\{a_n\}$, либо $\{b_n\}$ является подпоследовательностью последовательности $\{u_n\}$ и сходится к u_* . Пусть для определенности $\lim_{n \rightarrow \infty} a_n = u_*$. Допустим, что $\{b_n\}$ не сходится к u_* . Тогда согласно (4) $\{b_n\}$ не может быть подпоследовательностью для $\{u_n\}$, т. е. найдется номер $n_0 \geq 1$ такой, что $b_n = b_{n_0}$ при всех $n \geq n_0$. При $n \rightarrow \infty$ из (5) получим $J(u_*) = J(b_{n_0}) + J'(b_{n_0})(u_* - b_{n_0})$. В силу теоремы 11.4 это возможно только тогда, когда $J(u) \equiv J(b_{n_0}) + J'(b_{n_0})(u - b_{n_0}), u_* \leq u \leq b_{n_0}$. 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 $b_{n+1} - a_{n+1} \leq \max\{u_{n+1} - a_n, b_n - u_{n+1}\} \leq q_n(b_n - a_n)/2$, где $q_n = \max\{J''(\xi_n)/J''(\mu_n); J''(\eta_n)/J''(\mu_n)\}$. Так как последовательности $\{\xi_n\}, \{\eta_n\}, \{\mu_n\}$ вместе с $\{a_n\}, \{b_n\}$ стремятся к u_* , то в силу непрерывности $J''(u)$ и условия $\inf_{u \in [a, b]} J''(u) > 0$ имеем

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

$$|u_n - u_*| \leq (b_n - a_n) \leq q^{n-N}(b_N - a_N), \quad n \geq N.$$

Теорема доказана.

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 $J(u) = u^2$, $-1 \leq u \leq 2$. 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 точная оценка $|u_n - u_*| = |u_n| = 1/2^{n-1}$, в то время как используя методику вывода оценки (5), имеем $|u_n - u_*| \leq b_n - a_n = 3/2^{n-1}$, $n \geq 1$.

Таким образом, из 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 $J'(a_1) \geq 0$ или $J'(b_1) \leq 0$, то по теореме 11.5 $a \in U_*$ или $b \in U_*$ — задача

решена. Поэтому пусть $J'(a_1) < 0$, $J'(b_1) > 0$. Тогда $U_* \subset \subset \text{int}[a_1, b_1]$. Пусть отрезок $[a_{n-1}, b_{n-1}]$, $n \geq 2$, уже построен, причем $J'(a_{n-1}) < 0$, $J'(b_{n-1}) > 0$, так что $U_* \subset \subset \text{int}[a_{n-1}, b_{n-1}]$. 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 $u_n = (a_{n-1} + b_{n-1})/2$.

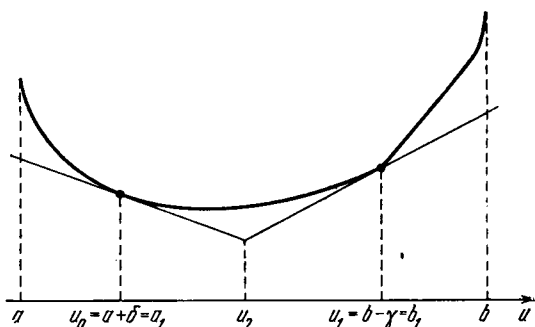


Рис. 1.6.

По построению $J'(a_n) < 0$, $J'(b_n) > 0$, и согласно теореме 11.6 $U_* \subset \subset \text{int}[a_n, b_n]$. Кроме того, ясно, что

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

Описанный метод деления отрезка пополам выгоднее применять для 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 выше мы 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 11.2. Чтобы гарантировать непустоту множества U_* , потребуем еще, чтобы $\lim_{u \rightarrow a+0} J(u) = J(a)$, $\lim_{u \rightarrow b-0} J(u) = J(b)$.

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 11.6 либо $U_* \cap [a, a_1] \neq \emptyset$, либо $U_* \cap [b_1, b] \neq \emptyset$ — в обоих случаях точка из U_* находится с требуемой точностью δ или γ , и задача

решена. Абсциссу точки пересечения касательных $g(u, a_1) = J(a_1) + J'(a_1 + 0)(u - a_1)$, $g(u, b_1) = J(b_1) + J'(b_1 - 0)(u - b_1)$ 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 $p_1(u) = \max \{p_0(u); g(u, u_1)\}$ на $[a_1, b_1]$, причем $a_1 < u_2 < b_1$,

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

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 $u_0, u_1, \dots, u_{n-1}$, $n \geq 2$, 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} совпадает с одной из точек $u_0, \dots, u_{n-1}$; пусть u_n — точка пересечения касательных $g(u, a_{n-1}) = J(a_{n-1}) + J'(a_{n-1} + 0)(u - a_{n-1})$, $g(u, b_{n-1}) = J(b_{n-1}) + J'(b_{n-1} - 0)(u - b_{n-1})$, а функция $p_{n-1}(u) = \max \{p_{n-2}(u), g(u, u_{n-1})\}$ такова, что

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

$p_{n-1}(u)$ на $[a_1, u_n]$ строго убывает, на $[u_n, b_1]$ — строго возрастает — это значит, что u_n — точка минимума $p_{n-1}(u)$ на $[a_1, b_1]$. Вычислим $J'(u_n + 0)$, $J'(u_n - 0)$. Если $J'(u_n + 0) \geq 0$, $J'(u_n - 0) \leq 0$, то по теореме 11.5 $u_n \in U_*$, и задача решена — итерации на этом заканчиваются. Поэтому пусть

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

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 11.6 тогда $U_* \subset \text{int } [a_n, b_n]$. 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

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

где u'_n, u''_n — абсциссы точек пересечения касательных $g(u, a_{n-1})$, $g(u, u_n)$ и $g(u, b_{n-1})$, $g(u, u_n)$ соответственно. Дадим строгое доказательство формулы (9). Из теоремы 11.4 следует, что $g(u, u_i) \leq J(u)$ при всех $u \in [a_1, b_1]$ и $i = 0, 1, \dots$, поэтому $p_{n-1}(u) \leq J(u)$. В частности, $p_{n-1}(u_n) \leq J(u_n)$. 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 11.4 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 11.4 $g(a_{n-1}, u_n) \leq J(a_{n-1})$. 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 значит, что касательные $g(u, a_{n-1})$, $g(u, u_n)$ пересекаются в некоторой точке с абсциссой u'_n , $a_{n-1} < u'_n < u_n$, а касательные

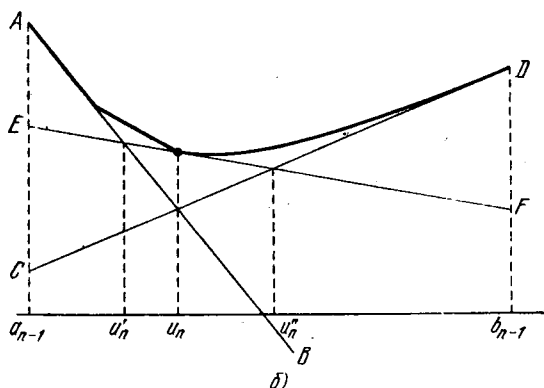
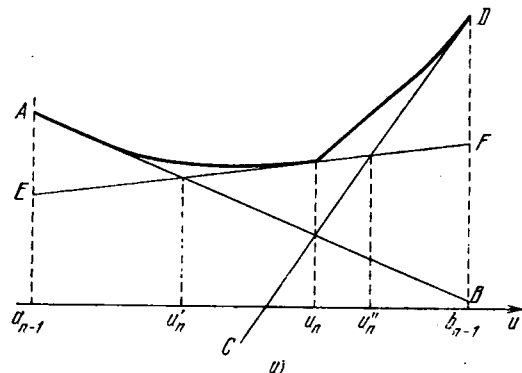


Рис. 1.7. а) $J'(u_n - 0) > 0$, б) $J'(u_n + 0) < 0$; AB — график $g(u, a_{n-1})$, CD — график $g(u, b_{n-1})$, EF — график $g(u, u_n)$.

$g(u, u_n)$, $g(u, b_{n-1})$ — 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 $a_1 \leq u < u'_n$, $g(u, b_{n-1}) > g(u, u_n)$ при $u''_n < u \leq b_1$. Отсюда следует, что $p_{n-1}(u) = \max_{0 \leq i \leq n-1} g(u, u_i) \geq g(u, a_{n-1}) > g(u, u_n)$ при $a_1 \leq u < u'_n$, $p_{n-1}(u) \geq g(u, b_{n-1}) > g(u, u_n)$ при $u''_n < u \leq b_1$. 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 $p_n(u)$ 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 касательных $g(u, a_n), g(u, b_n)$, $a_n < u_{n+1} < b_n$, 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 $[a_n, b_n] \subset [a_{n-1}, b_{n-1}] \subset \dots \subset [a_1, b_1]$, причем согласно теоремам 11.5, 11.6 процесс либо завершится при каком-либо n определением точки $u_n \in U$, либо $U \subset \text{int}[a_n, b_n]$ при всех $n = 1, 2, \dots$. Теорема 1 остается верной и для этого метода. Так как минимум $p_n(u)$ на $[a_1, b_1]$ достигается на $[a_n, b_n]$ в 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 пп. 1, 2, примененными к отрезку $[a + \delta, b - \nu]$ с выбором начальной точки $u_0 = a + \delta$; попутно получено строгое доказательство равносильности методов из пп. 1 и 2.

У п р а ж н е н и е. 1. Применить метод касательных для минимизации функции $J(u) = |u|$, $J(u) = u^2$, $J(u) = |u| + (u - 1)^2$ на отрезках $[-1, 1]$, $[-1, 2]$, $[0, 1]$, $[1, 2]$.

§ 13. Метод поиска глобального минимума

В §§ 9, 10 были рассмотрены методы, пригодные для поиска глобального минимума функций, удовлетворяющих условию Липшица. Эти методы требовали для своей реализации знания константы Липшица минимизируемой функции. Опишем другой метод поиска глобального минимума [215], не требующий априорного знания константы 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 $J(u)$ 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 $u_2 = (u_1 + u_0)/2 - (J(u_1) - J(u_0))/2d_1$. Нетрудно показать (см. ниже общий случай), что $u_0 < u_2 < u_1$. Перенумеруем точки u_0, u_1, u_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 $u_0, u_1, \dots, u_k (k \geq 2)$, причем $u_0 = a < u_1 < \dots < u_{k-1} < u_k = b$. Опишем $(k+1)$ -й шаг метода. Определим величины

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

и

$$d_k = \begin{cases} \mu_k L_k & \text{при } L_k > 0, \\ \nu & \text{при } L_k = 0, \end{cases} \quad (2)$$

где $\nu > 0$, $\mu_k > 1$ — параметры метода, выбираемые вычислителем. Для каждого отрезка $[u_{i-1}, u_i]$ введем следующую числовую характеристику:

$$R_k(i) = d_k(u_i - u_{i-1}) + \frac{(J(u_i) - J(u_{i-1}))^2}{d_k(u_i - u_{i-1})} - 2(J(u_i) + J(u_{i-1})), \quad (3)$$

$i = 1, 2, \dots, k$. После этого перебором значений $R_k(i)$, $i = 1, 2, \dots, k$, определим минимальный номер s , на котором достигается $\max_{1 \leq i \leq k} R_k(i)$, т. е.

$$R_k(s) = \max_{1 \leq i \leq k} R_k(i), \quad R_k(i) < R_k(s) \quad \text{при} \quad i = 1, 2, \dots, s-1. \quad (4)$$

Затем полагаем

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

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

$$\nu > 0; \quad \mu_k \geq \mu > 1, \quad k = 1, 2, \dots; \quad \sup_{k \geq 1} \mu_k < +\infty. \quad (6)$$

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 $k \geq s$ та же точка v_s в зависимости от своего местоположения на $[a, b]$ получает один из номеров i , $0 \leq i \leq k$, 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

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

Так как в силу (1), (2), (6)

$$\left| \frac{J(u_i) - J(u_{i-1})}{d_k(u_i - u_{i-1})} \right| \leq \frac{L_k}{d_k} \leq \frac{1}{\mu} < 1, \quad i = 1, 2, \dots, k, \quad (8)$$

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

$$\begin{aligned} (u_s - u_{s-1})(1 - 1/\mu)/2 &\leq \min \{u_s - v_{k+1}, v_{k+1} - u_{s-1}\} \leq \\ &\leq \max \{u_s - v_{k+1}, v_{k+1} - u_{s-1}\} \leq (u_s - u_{s-1})(1 + 1/\mu)/2, \end{aligned} \quad (9)$$

где $0 < q = (1 + 1/\mu)/2 < 1$, $s = s(k)$, $k = 1, 2, \dots$. Кроме того, из неравенства (9.7) имеем

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

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

$$d_k \geq d > 0, \quad k = 1, 2, \dots \quad (10)$$

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

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

то $L_k \leq L$, и следовательно,

$$d_k \leq \max\{v; L \sup_{k \geq 1} \mu_k\} < \infty, \quad k = 1, 2, \dots \quad (12)$$

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 $\lim_{k \rightarrow \infty} J(v_k) = J(v_*)$ и, кроме того,

$$J(v_*) \leq J(v_i), \quad i = 0, 1, \dots \quad (13)$$

Доказательство. Так как $v_i \neq v_k$ при $i \neq k$ и v_* — предельная точка $\{v_k\}$, то существование последовательности отрезков $[u_{m(k)-1}, u_{m(k)}]$, $k = 1, 2, \dots$, каждый из которых содержит точку v_* и бесконечно много поисковых точек, отличных от v_* , и кроме того, $u_{m(k)-1} \leq u_{m(k+1)-1} < u_{m(k+1)} \leq u_{m(k)}$, $k = 1, 2, \dots$. Введем множество $N(v_*)$, состоящее из всех тех и только тех номеров $k \geq 1$, 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

$$R_k(m(k)) \geq R_k(i), \quad i = 1, 2, \dots, k; \quad k \in N(v_*). \quad (14)$$

Так как 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)

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

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

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

Перепишем формулу (3) в виде

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

$i = 1, 2, \dots, k$. Взяв здесь $i = m(k)$, с учетом соотношений (8), (11), (12), (15) получим

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

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

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

$i = 1, 2, \dots, k$, и примем здесь $i = 1$, то с учетом (14) будем иметь $-4J(u_0) < R_k(1) \leq R_k(m(k))$, $k \in N(v_*)$. Таким образом, $4J(u_i) > -R_k(m(k))$ при всех $i = 0, 1, \dots, k$, $k \in N(v_*)$, т. е. $\min_{0 \leq i \leq k} J(u_i) > -R_k(m(k))$, $k \in N(v_*)$. Так как $\min_{0 \leq i \leq k} J(u_i) = \min_{0 \leq s \leq k} J(v_s) \leq \min_{0 \leq s \leq p} J(v_s)$ при любом $p \leq k$, то из предыдущего неравенства имеем $\min_{0 \leq s \leq p} J(v_s) > -R_k(m(k))$ для всех $p \leq k$, $k \in N(v_*)$. Перейдем здесь к пределу при $k \rightarrow \infty$, $k \in N(v_*)$. С учетом (17) получим $\min_{0 \leq s \leq p} J(v_s) \geq J(v_*)$ при каждом фиксированном $p \geq 1$. Оценка (13) доказана.

Докажем, что $\lim_{k \rightarrow \infty} J(v_k) = J(v_*)$. Пусть c — какая-либо предельная точка последовательности $\{J(v_k)\}$, и пусть $c = \lim_{n \rightarrow \infty} J(v_{k_n})$. Выбирая при необ-

ходимости подпоследовательность, можем считать, что $\{v_{k_n}\}$ сходится к некоторой точке w_* . Тогда из (13) следует, что $J(v_*) \leq J(w_*) = c$. С другой стороны, так как оценка (13) доказана для произвольной предельной точки $\{v_k\}$, то $J(w_*) \leq J(v_i)$, $i = 0, 1, \dots$, и следовательно, $J(w_*) \leq J(v_*)$. Таким образом, $J(w_*) = c = J(v_*)$, т. е. $\{J(v_k)\}$ имеет единственную предельную точку $c = J(v_*)$. Это означает, что $\lim_{k \rightarrow \infty} J(v_k) = J(v_*)$. Теорема 1 доказана.

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 $J(u)$ со значением $c = \lim_{k \rightarrow \infty} J(v_k)$.

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 $[u_{m(k)-1}, u_{m(k)}]$, $k = 1, 2, \dots$, и множество $N(v_*)$, введенные при доказательстве теоремы 1. Если оказалось, что $u_{m(k)-1} < v_* < u_{m(k)}$, $k = 1, 2, \dots$, то с учетом (15) в качестве искомого подпоследовательностей можно взять $\{u_{m(k)-1}\}$ и $\{u_{m(k)}\}$. Остается рассмотреть случаи, когда, начиная с какого-либо s -го шага v_* будет совпадать с одним из концов отрезка $[u_{m(k)-1}, u_{m(k)}]$. Пусть сначала $v_* = u_{m(k)-1}$, $k \geq s$. Тогда подпоследовательность $\{u_{m(k)}\}$ в силу (15) сходится к v_* справа. Рассмотрим подпоследовательность $\{u_{m(k)-2}\}$ при $k \in N(v_*)$. С помощью формулы (16) при $i = m(k) - 1$ и соотношений (14), (17) имеем

$$d_k(v_* - u_{m(k)-2}) \left(1 + \frac{J(v_*) - J(u_{m(k)-2})}{d_k(v_* - u_{m(k)-2})} \right)^2 = R_k(m(k) - 1) + 4J(v_*) \leq \\ \leq R_k(m(k)) + 4J(v_*) \rightarrow 0 \text{ при } k \rightarrow \infty, k \in N(v_*).$$

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

Пусть теперь $v_* = u_{m(k)}$, $k \geq s$. 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

$$d_k(u_{m(k)+1} - v_*) \left(1 - \frac{J(u_{m(k)+1}) - J(v_*)}{d_k(u_{m(k)+1} - v_*)} \right)^2 = R_k(m(k) + 1) + 4J(v_*) \leq \\ \leq R_k(m(k)) + 4J(v_*) \rightarrow 0 \text{ при } k \rightarrow \infty, k \in N(v_*).$$

Отсюда, учитывая (8), (10), имеем $\{u_{m(k)+1}\} \rightarrow v_*$ при $k \rightarrow \infty$, $k \in N(v_*)$. Искомые подпоследовательности для предельных точек v_* , $a < v_* < b$, построены. Заметим, что если $v_* = a$, то в силу (15) подпоследовательность $\{u_{1(k)}\}$ сходится к v_* справа, а если $v_* = b$, то $\{u_{k(k)-1}\}$ сходится к v_* слева.

Перенумеруем точки θ_i , $i = 0, 1, \dots, r$, локального экстремума $J(u)$ на $[a, b]$ в порядке возрастания: $\theta_0 = a < \theta_1 < \dots < \theta_r = b$. 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 вблизи от v_* слева $J(u)$ строго убывает, справа — строго возрастает. Следовательно, v_* совпадает с одной из точек θ_i и представляет собой точку локального минимума $J(u)$ со значением $J(v_*) = \lim_{k \rightarrow \infty} J(v_k)$.

Из теорем 1, 2 и 1.1 непосредственно вытекает

Следствие 1. Пусть функция $J(u)$ на отрезке $[a, b]$ удовлетворяет условию (11), множество $U_* = \{u: u \in [a, b], J(u) = J_* = \inf_{v \in [a, b]} J(v)\}$ 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 $\lim_{k \rightarrow \infty} \rho(v_k, U_*) = 0$.

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, $d_k \geq 2L$. Тогда последовательность $\{v_k\}$, полученная этим методом, такова, что

$$1) \lim_{k \rightarrow \infty} J(v_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(v_k, U_*) = 0;$$

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

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

Возьмем произвольную точку $u_* \in U_*$. Пусть $u_{r(k)-1} \leq u_* \leq u_{r(k)}$, $k = 1, 2, \dots$. Так как $J(u_{r(k)}) + J(u_{r(k)-1}) \leq L(u_{r(k)} - u_{r(k)-1}) + 2J(u_*)$, то из (3) с учетом условия $d_k \geq 2L$ имеем

$$R_k(r(k)) \geq (d_k - 2L)(u_{r(k)} - u_{r(k)-1}) - 4J(u_*) \geq -4J(u_*), \quad (19)$$

$k = 1, 2, \dots$ Далее, пусть v_* — какая-либо предельная точка последовательности $\{v_k\}$. Возьмем отрезки $[u_{m(k)-1}, u_{m(k)}]$, $k = 1, 2, \dots$, и множество $N(v_*)$, введенные при доказательстве теоремы 1 для точки v_* . Так как $J(u_*) \leq J(v_*)$, то из (14) при $i = r(k)$ и из (19) имеем $R_k(m(k)) \geq R_k(r(k)) \geq -4J(u_*) \geq -4J(v_*)$, $k \in N(v_*)$. Отсюда с учетом равенства (17) имеем

$$\lim_{\substack{k \rightarrow \infty \\ k \in N(v_*)}} R_k(m(k)) = \lim_{\substack{k \rightarrow \infty \\ k \in N(v_*)}} R_k(r(k)) = -4J(u_*) = -4J(v_*), \quad (20)$$

т. е. $J(u_*) = J(v_*)$. В силу теоремы 1 тогда $\lim_{k \rightarrow \infty} J(v_k) = J(v_*) = J(u_*)$, а из теоремы 1.1 получим $\lim_{k \rightarrow \infty} \rho(v_k, U_*) = 0$. Наконец, при $\inf_{k \geq k_0} d_k > 2L$ из (19), (20) следует $\lim_{\substack{k \rightarrow \infty \\ k \in N(v_*)}} (u_{r(k)} - u_{r(k)-1}) = 0$, 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 (10.2) на классе функций $Q(L)$, состоящем из функций, удовлетворяющих на $[a, b]$ условию (11) с одной и той же константой L . Согласно теореме 10.1 для обеспечения точности

$$\min_i J(w_i) - J_* \leq \varepsilon \quad (21)$$

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

Теорема 4. Пусть $J(u) \in Q(L)$, а отрезок $[\alpha, \beta] \subseteq \{u: u \in [a, b], J(u) \geq J_* + \gamma\}$, где $\gamma > 0$ — фиксированное число. Тогда на $[\alpha, \beta]$ число поисковых точек метода (1)–(6), реализованного при $d_k = 2L$, $k \geq 1$, по крайней мере в $\gamma/(5\varepsilon)$ раз меньше числа поисковых точек метода равномерного перебора (10.2); здесь число $\varepsilon > 0$ взято из (21), $5\varepsilon < \gamma$.

Доказательство. Пусть $u_* \in U_*$, $u_{r(k)-1} \leq u_* \leq u_{r(k)}$, $k = 1, 2, \dots$. Тогда из (19) следует, что $R_k(r(k)) \geq -4J(u_*) = -4J_*$. Далее, для любого отрезка $[u_{i-1}, u_i] \subseteq [\alpha, \beta]$, $i = i(k)$ справедливо неравенство $J(u_i) + J(u_{i-1}) \geq \geq 2J_* + 2\gamma$. 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 $R_k(r(k)) - R_k(i(k)) \geq 4\gamma - 5L(u_i - u_{i-1})/2 > 0$, 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 $u_i - u_{i-1} \geq 8\gamma/(5L)$. 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 $d_k = 2L$, $\mu_k = 2L/L_k \geq 2 = \mu$ оценивается снизу величиной $(u_i - u_{i-1})/4 \geq \geq 2\gamma/(5L)$. Это значит, что независимо от номера шага расстояние между точками $u_{i-1}, u_i \in [\alpha, \beta]$ удовлетворяет условию $u_i - u_{i-1} \geq 2\gamma/(5L)$, 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 $(\beta - \alpha)/h \geq (\beta - \alpha)L/(2\varepsilon)$. Сравнивая полученные числа, убеждаемся в справедливости теоремы 4.

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 $v = 1$, $\mu_k = 2$ 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 [215], что метод (1)–(6) позволяет определить глобальный минимум с достаточно высокой надежностью. Замечено, что при увеличении значений параметров v , 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 $[-1, 1]$, $[-1, 2]$, $[1, 2]$.

§ 14. 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 $\Delta^-(v) = J(v - \tau) - J(v) \geq 0$, $\Delta^+(v) = J(v + h) - J(v) \geq 0$, $\Delta^+ + \Delta^- > 0$.

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 [2], [39]:

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

Так как $p = \Delta^+/h + \Delta^-/\tau > 0$, то $L_2(u)$ выпукла на $\mathbf{R}^1$ и достигает своей нижней грани на $\mathbf{R}^1$ в точке

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

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

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

Перейдем к описанию одного из вариантов метода парабол для минимизации функции $J(u)$, унимодальной на отрезке $[a, b]$. Пусть $h > 0$ — некоторый начальный шаг, $2h \leq b - a$, а $u_0 \in [a, b]$ — начальная точка. Поиск минимума начинаем с вычисления значений $J(u_0), J(u_0 + h)$. Если окажется, что $J(u_0 + h) \leq J(u_0)$, то продолжаем вычислять значения $J(u_0 + h \cdot 2^{i-1})$, $i = 2, 3, \dots$ (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 $J(u_0 - h \cdot 2^{i-1})$, $i = 2, 3, \dots$ (рис. 1.9). Перед 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 сле-

дующей точке u_{i+1} и т. д. Этот процесс закончится отысканием такого номера $n \geq 1$, 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 $\mathbb{R}^1$; согласно (3) $w \in [u_{n-2}, u_n]$ (рис. 1.8, $n=4$);

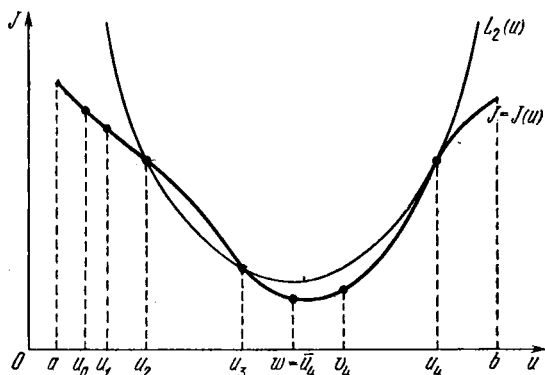


Рис. 1.8.

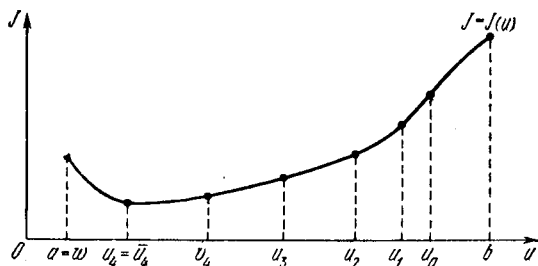


Рис. 1.9.

2) либо ни одна тройка u_{i-2} , u_{i-1} , u_i при $i = 2, 3, \dots, n$ 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} (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 , $u_1, \dots, u_n$ находим точку $\bar{u}_n$ такую, что

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

и за приближение к $J_* = \inf_{u \in [a, b]} J(u)$ берем значение $J(\bar{u}_n)$, а за приближение ко 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.

Метод парабол описан. 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) положительен, хотя в отличие от определения 1 одно из условий $\Delta^+ \geq 0$ или $\Delta^- \geq 0$ может быть и нарушено. В этом случае парабола (1) также выпукла на $\mathbf{R}^1$ 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 $w \notin [a, b]$.

§ 15. 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 $J(u)$ 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 v_0 существует новая точка u_1 более глубокого локального минимума со значением $J(u_1) < J(v_0) < J(u_0)$. Для поиска u_1 можно использовать один из локальных методов минимизации, например, ме-

тод 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 $J(u_1) \leq J(v_0)$, 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 $J(u)$ 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 локального минимума со значением $J(u_2) \leq J(v_1) < J(u_1) < J(u_0)$ 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 $J(u)$ 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 $v_0, v_1, \dots$, т. е. точек, в которых функция принимает меньшее значение, чем в найденной точке локального минимума.

2. Опишем один такой способ, следуя работе [61]. Будем предполагать, что функция $J(u)$ непрерывно дифференцируема на отрезке $[a, b]$ и имеет на $[a, b]$ 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

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

где $g_i(u_i) > 0$ при $a \leq u_i < b$ 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_i 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 $a \leq u_i < b$ и $(-1)^{n_i} J^{(n_i)}(u_i) > 0$ при $u_i = b$ тогда является необходимым для того, чтобы 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 $J(u)$. Тогда $J(u) > J(u_0)$ для всех $u \in O_\alpha(u_0) = \{u: u \in [a, b], 0 < |u - u_0| < \alpha\}$, если число $\alpha > 0$ достаточно мало. Введем функцию

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

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

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 $m = 1, 2, \dots$ 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 $J(u)$ на $[a, b]$ тогда и только тогда, когда $\varphi_m(u, u_0) \geq 0$ при всех $u \in [a, b]$, $u \neq u_0$, $m = 1, 2, \dots$, 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 $\varphi_m(v_0, u_0) < 0$, $m \geq 1$. Очевидно также, что при любом $m = 1, 2, \dots$ 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, если w_0 является точкой локального минимума $J(u)$ со значением $J(w_0) = J(u_0)$.

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

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

Отсюда с учетом представления (1) имеем

$$\varphi'_m(u, u_0)(u - u_0)^{2m-n_0+1} = (u - u_0)g'_0(u) - (2m - n_0)g_0(u). \quad (4)$$

Если $a \leq u_0 < b$ и α достаточно мало, то $g_0(u) \geq \text{const} > 0$ 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 имеем, 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 $a \leq u_0 < b$ рассмотрен. Если же $u_0 = b$, то из непрерывности $g_0(u)$ и условия $(-1)^{n_0} g_0(b) > 0$ следует существование такого $\alpha > 0$, что $(-1)^{n_0} g_0(u) \geq \text{const} > 0$ 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 $v \in \text{int}[a, b]$ 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 $m \geq m_0$. 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 $m \geq m_0$ имеем $\varphi'_m(v, u_0) \times (v - u_0)^{2m+1} < 0$, 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 $m \geq m_0$ 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 $\varphi_m(w_0, u_0) \geq 0$, причем $\varphi_m(u, u_0) \geq 0$ 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 $\varphi_m(u, u_0) \geq 0$, $u \in [u_0 + \alpha, w]$, установленное для какого-то значения m , останется верным при всех $m = 1, 2, \dots$, и свидетельствует об отсутствии на $[u_0 + \alpha, w_0]$ искомой точки v_0 .

Рассмотрим случай $\varphi_m(w_0, u_0) = 0$. Согласно свойству 1) тогда функция $J(u)$ также будет достигать локального минимума в точке w_0 со значением $J(w_0) = J(u_0)$. Так как локальные минимумы $J(u)$ строгие, то можем выбрать $\alpha > 0$ столь малым, чтобы $J(u) > J(w_0)$ и, следовательно, $\varphi_m(u, u_0) > 0$ при $w_0 < u \leq w_0 + \alpha$. 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 $\varphi_m(u, u_0) \geq 0$ при всех u , $u_0 < u \leq b$. 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 $\varphi_m(u, u_0) \geq 0$ при $u \in [a, b]$, $u \neq u_0$, т. е. согласно свойству 1) точка u_0 есть точка глобального минимума $J(u)$ 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 такую точку и реализовать изложенную в начале параграфа схему поиска глобального минимума $J(u)$ 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 $m = 1, 2, \dots, M$. Численные эксперименты показывают, что удачный выбор M и точности определения точек локального минимума рассматриваемых функций обеспечивают достаточно высокую эффективность этого метода [61].

8. 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 $a_{2n} > 0$. Так как $\lim_{|u| \rightarrow +\infty} p_{2n}(u) = +\infty$ и $p_{2n}(u)$ — непрерывная функция, то множество

$M(v) = \{u: u \in R^1, p_{2n}(u) \leq p_{2n}(v)\}$ непусто, ограничено и замкнуто при любом фиксированном v . Кроме того, очевидно $\inf_{u \in R^1} p_{2n}(u) = \inf_{u \in M(v)} p_{2n}(u) = J_*$. 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 R^1 непусто. Более того, так как производная $p'_{2n}(u)$, являющаяся многочленом степени $2n-1$, может иметь не более $2n-1$ действительных корней, то $p_{2n}(u)$ на R^1 может иметь не более $n-1$ точек локального максимума и не более n точек локального минимума.

Поиск глобального минимума функции $p_{2n}(u)$ на R^1 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

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

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 $a_{2n} > 0$. Так как $p_{2n}(u) - p_{2n}(u_0) = (u - u_0)^2 p_{2n-2}(u, u_0)$ при всех $u \in R^1$, то ясно, что точка u_0 будет точкой глобального минимума $p_{2n}(u)$ на R^1 тогда и только тогда, когда $\inf_{u \in R^1} p_{2n-2}(u, u_0) \geq 0$. Это означает, что исходная задача минимизации

на R^1 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 R^1 многочленов меньших степеней $2n-4, 2n-6, \dots, 2$ с одним и тем же старшим коэффициентом $a_{2n} > 0$. Остается заметить, что для многочлена 2-й степени $p_2(u) = b_2 u^2 + b_1 u + b_0$, $b_2 = a_{2n} > 0$, точка минимума на R^1 находится по известной формуле $u_* = -b_1/(2b_2)$.

Если при поиске минимума $p_{2n-2}(u, u_0)$ на R^1 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 локальным методом находим следующую точку u_1 локального минимума $p_{2n}(u)$ со значением $p_{2n}(u_1) \leq p_{2n}(v_0) < p_{2n}(u_0)$ 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 $\inf_{u \in R^1} p_{2n-2}(u, u_i) \geq 0$.

§ 16. 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, 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 $|J(u) - z(u)| \leq \varepsilon$. 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 $u \in \mathbb{R}^1$, причем результаты измерений $z(u)$ не содержат систематических ошибок. Тогда для поиска минимума функции $J(u)$ на $\mathbb{R}^1$ может быть использован следующий итерационный метод:

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

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

$$a_n > 0, \quad c_n > 0, \quad n = 1, 2, \dots, \quad \lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} c_n = 0, \quad (2)$$

$$\sum_{n=1}^{\infty} a_n = +\infty, \quad \sum_{n=1}^{\infty} (a_n/c_n)^2 < \infty.$$

Например, здесь можно взять $a_n = 1/n$, $c_n = 1/n^{1/4}$, $n = 1, 2, \dots$

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:

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

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 $\mathbb{R}^1$. Различные варианты метода стохастической аппроксимации, строгое обоснование этого метода, различные приложения можно найти в [53, 174, 222].

ПРЕДВАРИТЕЛЬНЫЕ СВЕДЕНИЯ О КОНЕЧНОМЕРНЫХ ЗАДАЧАХ НА ЭКСТРЕМУМ

В этой главе собраны основные факты о задачах на экстремум функций конечного числа переменных, обычно излагаемые в учебниках по математическому анализу [7, 8, 179], 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 векторов-столбцов

$$u = \begin{pmatrix} u^1 \\ \vdots \\ u^n \end{pmatrix}, \quad v = \begin{pmatrix} v^1 \\ \vdots \\ v^n \end{pmatrix}, \quad w = \begin{pmatrix} w^1 \\ \vdots \\ w^n \end{pmatrix}$$

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

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

вектор-столбец

$$0 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

называется нулевым. Вектор-строку, полученную транспонированием вектора-столбца u , обозначим через $u^T = (u^1, \dots, u^n)$. 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

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

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

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

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 формулировке постановки задачи минимизации. Пусть U — некоторое 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 $u = (u^1, u^2, \dots, u^n)$ 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

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

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

$$J(u) \rightarrow \inf; \quad u \in U.$$

Как и в § 1.1, будем различать задачи минимизации двух ти-

пов. 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 ко множеству U_* или даже принадлежит U_* , — здесь естественно требовать, чтобы $J_* > -\infty$, $U_* \neq \emptyset$.

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

$$u_k \in U, \quad k = 1, 2, \dots, \quad \lim_{k \rightarrow \infty} J(u_k) = J_*$$

(при $U_* \neq \emptyset$ возможно, например, $u_k = u_* \in U_*$, $k = 1, 2, \dots$). 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 ко множеству U_* , т. е. $\rho(u_k, U_*) = \inf_{u_* \in U_*} |u_k - u_*| \rightarrow 0$ при $k \rightarrow \infty$, 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 $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$ имеет

место не всегда. Поэтому в задачах второго типа построение минимизирующих последовательностей, сходящихся к U_* , в общем случае требует привлечения специальных методов [16].

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 .

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 $\{u_k\} = (u_1, u_2, \dots)$ — некоторая последовательность, $\{u_k\} \in E^n$, т. е. $u_k \in E^n$, $k = 1, 2, \dots$. 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 $M \geq 0$ такая, что $|u_k| \leq M$ для всех $k = 1, 2, \dots$.

Множество U из E^n называется *ограниченным*, если существует постоянная $R \geq 0$ такая, что $|u| \leq R$ для всех $u \in U$. Множество $O(v, \varepsilon) = \{u: u \in E^n, |u - v| < \varepsilon\}$, 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 $U \subseteq E^n$, 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 $k = 1, 2, \dots$ 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 $\{x_k\} = (x_1, x_2, \dots)$ называется ограниченной снизу [сверху], если существует число A такое, что $x_k \geq A$ [$x_k \leq A$] при всех $k = 1, 2, \dots$. Если $\{x_k\}$ неограничена снизу [сверху], то существует подпоследовательность $\{x_{k_m}\}$ такая, что $\lim_{m \rightarrow \infty} x_{k_m} = -\infty$ [$\lim_{m \rightarrow \infty} x_{k_m} = +\infty$].

Определение 2. Число a называется *нижним* [*верхним*] пределом ограниченной снизу [сверху] числовой последовательности $\{x_k\}$ и обозначается через $\lim_{k \rightarrow \infty} x_k = a$ [$\overline{\lim}_{k \rightarrow \infty} x_k = a$], 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 $x_k \geq a - \varepsilon$ [$x_k \leq a + \varepsilon$] для всех $k \geq N$; 2) для любого номера m найдется номер $k_m > m$ такой, что $x_{k_m} \leq a + \varepsilon$ [$x_{k_m} \geq a - \varepsilon$]. В том случае, когда $\{x_k\}$ неограничена снизу [сверху], то по определению принимают $\lim_{k \rightarrow \infty} x_k = -\infty$ [$\lim_{k \rightarrow \infty} x_k = +\infty$]; если $\lim_{k \rightarrow \infty} x_k = -\infty$, то полагают $\lim_{k \rightarrow \infty} x_k = -\infty$; если $\lim_{k \rightarrow \infty} x_k = +\infty$, то $\lim_{k \rightarrow \infty} x_k = +\infty$.

Например, если $x_k = (-1)^k$, $k = 1, 2, \dots$, то $\lim_{k \rightarrow \infty} x_k = -1$, $\overline{\lim}_{k \rightarrow \infty} x_k = +1$; если $x_k = (-1)^k k$, $k = 1, 2, \dots$, то $\lim_{k \rightarrow \infty} x_k = -\infty$, $\overline{\lim}_{k \rightarrow \infty} x_k = +\infty$; если $x_k = [1 + (-1)^k] k$, $k = 1, 2, \dots$, то $\lim_{k \rightarrow \infty} x_k = 0$, $\overline{\lim}_{k \rightarrow \infty} x_k = +\infty$; если $x_k = k^{-1}$, $k = 1, 2, \dots$, то $\lim_{k \rightarrow \infty} x_k = \overline{\lim}_{k \rightarrow \infty} x_k = 0$.

Для того чтобы числовая последовательность $\{x_k\}$ имела предел, необходимо и достаточно, чтобы $\lim_{k \rightarrow \infty} x_k = \overline{\lim}_{k \rightarrow \infty} x_k = a$; тогда $\lim_{k \rightarrow \infty} x_k = a$.

Определение 3. Пусть функция $J(u)$ определена на множестве $U \subseteq E^n$. Говорят, что функция $J(u)$ *полу непрерывна снизу [сверху]* в точке $u \in U$, если для любой последовательности $\{u_k\} \in U$, сходящейся к точке u , имеет место соотношение $\lim_{k \rightarrow \infty} J(u_k) \geq J(u)$ [$\lim_{k \rightarrow \infty} J(u_k) \leq J(u)$]. Функцию $J(u)$ называют

полу непрерывной снизу [сверху] на множестве U , если она полу непрерывна снизу [сверху] в каждой точке этого множества.

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

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

Пример 2. Пусть $U = \{u: u \in E^1, -1 \leq u \leq 1\}$; $J(u) = u$ при $0 < u \leq 1$; $J(u) = 1 - u$, $-1 \leq u < 0$; $J(0) = a$. Нетрудно видеть, что при $a \leq 0$ эта функция полу непрерывна снизу на U ; при $a \geq 1$ — полу непрерывна сверху на U , а при $0 < a < 1$ в точке $u = 0$ она не будет полу непрерывной ни снизу, ни сверху.

Пример 3. Пусть $u = (x, y) \in E^2$; $J(u) = x^2 + y^2$ при $x > 0$, $y \geq 0$; $J(u) = 0$ при $x \leq 0$, $y \geq 0$; $J(u) = 1$ при $x \geq 0$, $y < 0$; $J(u) = -1$ при $x < 0$, $y < 0$. Нетрудно показать, что эта функция на множестве $U_1 = \{(x, y): x > 0, y \geq 0\}$ непрерывна; на $U_2 = \{(x, y): y \geq 0\}$ полу непрерывна снизу; на $U_3 = \{(x, y):$

$x \leq 0$ полунепрерывна сверху; на $U_4 = \{(x, y): x \geq 0\}$ в некоторых точках полунепрерывна снизу, в некоторых — сверху.

Установим связь между свойством полунепрерывности снизу функции и замкнутостью множеств

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

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 $M(c) \neq \emptyset$. 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 $J(u_k) \leq c$, $k = 1, 2, \dots$, с учетом полунепрерывности снизу $J(u)$ в точке w имеем $J(w) \leq \liminf_{k \rightarrow \infty} J(u_k) \leq c$, т. е. $w \in M(c)$. Замкнутость $M(c)$ доказана. В частности, множество $U_* = \{u: u \in U, J(u) \leq J_* = \inf_U J(u)\}$ 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 $\lim_{k \rightarrow \infty} J(u_k) = a = \lim_{m \rightarrow \infty} J(u_{k_m})$. Тогда $J(u_{k_m}) \leq a + \varepsilon$, 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. $J(u) \leq a + \varepsilon$. В силу произвола $\varepsilon > 0$ отсюда имеем $J(u) \leq a = \lim_{k \rightarrow \infty} J(u_k)$.

Лемма 1 доказана.

Установим одно интересное свойство расстояния от точки до множества.

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

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

Доказательство. Прежде всего, из $\rho(u, w) = |u - w| \geq 0$ и $\rho(u, U) \leq |u - w|$, $w \in U$, следует, что функция $\rho(u, U)$ неотрицательна и конечна во всех точках $u \in U$. Возьмем произвольное число $\varepsilon > 0$. По определению нижней грани (см. определение 1.1.3) для любых $u, v \in E^n$ найдутся точки $u_\varepsilon, v_\varepsilon \in U$ такие, что

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

Поскольку $\rho(u, U) \leq \rho(u, v_\varepsilon)$, то с помощью неравенства треугольника $\rho(u, v_\varepsilon) \leq \rho(u, v) + \rho(v, v_\varepsilon)$ имеем $\rho(u, U) - \rho(v, U) \leq \rho(u, v_\varepsilon) - \rho(v, v_\varepsilon) + \varepsilon \leq \rho(u, v) + \varepsilon$. Аналогично получается неравенство $\rho(u, U) - \rho(v, U) \geq \rho(u, u_\varepsilon) - \varepsilon - \rho(v, u_\varepsilon) \geq -\rho(u, v) - \varepsilon$. Объединяя два последних неравенства, имеем $|\rho(u, U) - \rho(v, U)| \leq \rho(u, v) + \varepsilon$. Отсюда при $\varepsilon \rightarrow +0$ получим требуемое неравенство. Лемма 2 доказана.

4. Перейдем к формулировке теоремы Вейерштрасса.

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

Доказательство. Возьмем произвольную минимизирующую последовательность $\{u_k\}$: $u_k \in U$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} J(u_k) = J_*$. 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

$$J_* \leq J(u_*) \leq \lim_{m \rightarrow \infty} J(u_{k_m}) = \lim_{k \rightarrow \infty} J(u_k) = J_*,$$

т. е. $J(u_*) = J_*$. Отсюда следует, что $J_* > -\infty$, $U_* \neq \emptyset$. 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 $J(v_k) = J_*$, $k = 1, 2, \dots$. По вышедшему тогда $v_* \in U_*$. Компактность U_* установлена.

Покажем, что любая минимизирующая последовательность $\{u_k\}$ сходится к U_* . Так как $\rho(u_k, U_*) = \inf_{u \in U_*} \rho(u_k, u) \geq 0$, $k = 1, 2, \dots$, то ясно, что $\lim_{k \rightarrow \infty} \rho(u_k, U_*) \geq 0$. Пусть $\overline{\lim}_{k \rightarrow \infty} \rho(u_k, U_*) = \lim_{m \rightarrow \infty} \rho(u_{k_m}, U_*) = a \leq +\infty$. 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 $\lim_{m \rightarrow \infty} \rho(u_{k_m}, U_*) = \rho(u_*, U_*) = a$. Однако по доказанному $u_* \in U_*$. Тогда $a = \rho(u_*, U_*) = 0$. Это значит, что $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = \overline{\lim}_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. Следовательно, предел $\lim_{k \rightarrow \infty} \rho(u_k, U_*)$ 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 $U = \{u: u^1 \geq 0, \dots, u^n \geq 0\}$ — положительный октант, 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

$$M(v) = \{u: u \in U, J(u) \leq J(v)\}$$

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 $u \in U \setminus M(v)$ и $J(u) \leq J(v)$ при всех $u \in M(v)$. Это значит, что на $U \setminus M(v)$ 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 минимизирующей последовательности $\{u_k\}$ найдется номер k_0 такой, что $J(u_k) < J(v)$ для всех $k \geq k_0$, т. е. $u_k \in M(v)$ при $k \geq k_0$. 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 $\{u_k\} \in U$, $\lim_{k \rightarrow \infty} |u_k| = +\infty$ (если такие u_k существуют) имеет место соотношение

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

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 неограничено, т. е. существует хотя бы одна последовательность $\{v_k\} \in U$ такая, что $\lim_{k \rightarrow \infty} |v_k| = +\infty$. Тогда согласно условию теоремы $\lim_{k \rightarrow \infty} J(v_k) = +\infty$. 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 $M(v) = \{u \in U, J(u) \leq J(v)\}$. Покажем, что $M(v)$ ограничено. Допустим противное: пусть существует последовательность $\{w_k\} \in M(v)$ и такая, что $\lim_{k \rightarrow \infty} |w_k| = +\infty$. Тогда

$\lim_{k \rightarrow \infty} J(w_k) = +\infty$, что противоречит неравенству $J(w_k) \leq J(v) < +\infty$, вытекающему из включения $w_k \in M(v)$, $k = 1, 2, \dots$. 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 $\rho(u, U) = \inf_{w \in U} |u - w| = |u - v(u)|$, т. е. $v(u)$ — ближайшая к u точка из U .

Доказательство. Пусть u — произвольная точка из E^n . Рассмотрим функцию $g(w) = |u - w|$ переменной $w \in E^n$. Ясно, что $g(w)$ непрерывна на E^n . Кроме того, $g(w) \geq |w| - |u|$, так что $\lim_{|w| \rightarrow +\infty} g(w) = +\infty$. 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) $U = \{u = (x, y) \in E^2, x \geq 0, y \geq 0, x + 2y \leq 1\}$; $J(u) = x + y$;

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 $M_\alpha = \{u: u \in U, J(u) < J_* + \alpha\}$ ограничено.

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

а) $\lim_{n \rightarrow \infty} ca_n = c \lim_{n \rightarrow \infty} a_n$, $\overline{\lim}_{n \rightarrow \infty} ca_n = c \overline{\lim}_{n \rightarrow \infty} a_n$, $\lim_{n \rightarrow \infty} (-ca_n) = -c \lim_{n \rightarrow \infty} a_n$,
 $\overline{\lim}_{n \rightarrow \infty} (-ca_n) = -c \lim_{n \rightarrow \infty} a_n$ для любых $c = \text{const} > 0$;

б) если $a_n \leq b_n$, $n = 1, 2, \dots$, то $\lim_{n \rightarrow \infty} a_n \leq \lim_{n \rightarrow \infty} b_n$, $\overline{\lim}_{n \rightarrow \infty} a_n \leq \overline{\lim}_{n \rightarrow \infty} b_n$;

в) $\lim_{n \rightarrow \infty} a_n + \overline{\lim}_{n \rightarrow \infty} b_n \leq \overline{\lim}_{n \rightarrow \infty} (a_n + b_n) \leq \overline{\lim}_{n \rightarrow \infty} a_n + \overline{\lim}_{n \rightarrow \infty} b_n$. Рассмотреть пример: $a_n = (-1)^n$, $b_n = (-1)^n$ или $b_n = (-1)^{n+1}$ и убедиться, что здесь возможны строгие неравенства;

г) $\lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n \leq \lim_{n \rightarrow \infty} (a_n + b_n) \leq \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n$. Привести примеры последовательностей, когда здесь возможны строгие неравенства;

д) если существует $\lim_{n \rightarrow \infty} \overline{a_n}$, то $\overline{\lim_{n \rightarrow \infty} (a_n + b_n)} = \lim_{n \rightarrow \infty} a_n + \overline{\lim_{n \rightarrow \infty} b_n}$,
 $\lim_{n \rightarrow \infty} (a_n + b_n) = \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n$;

е) если $a_n \geq 0$, $b_n \geq 0$, $n = 1, 2, \dots$, то $\lim_{n \rightarrow \infty} a_n \cdot \lim_{n \rightarrow \infty} b_n \leq \lim_{n \rightarrow \infty} a_n b_n \leq \lim_{n \rightarrow \infty} a_n \cdot \overline{\lim_{n \rightarrow \infty} b_n}$; $\lim_{n \rightarrow \infty} a_n \cdot \overline{\lim_{n \rightarrow \infty} b_n} \leq \overline{\lim_{n \rightarrow \infty} a_n b_n} \leq \overline{\lim_{n \rightarrow \infty} a_n} \cdot \overline{\lim_{n \rightarrow \infty} b_n}$. Привести примеры последовательностей, когда здесь возможны строгие неравенства;

ж) если $a_n \geq 0$, $b_n \geq 0$, $n = 1, 2, \dots$, и существует $\lim_{n \rightarrow \infty} a_n$, то $\overline{\lim_{n \rightarrow \infty} a_n b_n} = \lim_{n \rightarrow \infty} a_n \cdot \overline{\lim_{n \rightarrow \infty} b_n}$, $\lim_{n \rightarrow \infty} (a_n b_n) = \lim_{n \rightarrow \infty} a_n \cdot \lim_{n \rightarrow \infty} b_n$.

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 $E^1 = \mathbb{R}$?

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 $\lim_{u \rightarrow v} J(u) = A$ [$\overline{\lim_{u \rightarrow v} J(u)} = A$], если

1) для любой последовательности $\{u_k\} \in U$, сходящейся к v , имеет место неравенство $\lim_{k \rightarrow \infty} J(u_k) \geq A$ [$\lim_{k \rightarrow \infty} J(u_k) \leq A$];

2) существует последовательность $\{u_k\} \in U$, сходящаяся к v и такая, что $\lim_{k \rightarrow \infty} J(u_k) = A$.

Доказать, что функция $J(u)$ полунепрерывна снизу [сверху] в точке v , если $\lim_{u \rightarrow v} J(u) \geq J(v)$ [$\lim_{u \rightarrow v} J(u) \leq J(v)$].

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 $\lim_{u \rightarrow 0} J(u)$, $\overline{\lim_{u \rightarrow 0} J(u)}$. Что изменится, если $U = \{u: |u| = n^{-1}, n = 1, 2, \dots\}$?

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 математи-

ческому анализу. 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 $O_\varepsilon(u) = \{v: v \in E^n, |v - u| < \varepsilon\}$ 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

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

где $o(h, u)$ — величина, бесконечно малая более высокого порядка, чем $|h|$, т. е. $\lim_{\alpha \rightarrow 0} o(h, u) |h|^{-1} = 0$. 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

$$J'(u) = (J'_{u^1}(u), \dots, J'_{u^n}(u)), \quad (2)$$

где

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

— частная производная функции $J(u)$ в точке u по переменной u^i ; $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ — 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 дифференцируемости функции нам понадобится понятие квадратичной формы. Напоминаем, что квадратичной формой называют функцию $\sum_{i,j=1}^n a_{ij} u^i u^j$ переменных $u = (u^1, \dots, u^n) \in E^n$, которая однозначно определяется заданием симметричной числовой матрицы

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \{a_{ij}\}$$

порядка $n \times n$, называемой матрицей квадратичной формы. Обозначая через Au вектор-столбец с координатами $(Au)^i = \sum_{j=1}^n a_{ij} u^j$, $i = 1, 2, \dots, n$, 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

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

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

Квадратичную форму $d^2J(u) = \langle J''(u)h, h \rangle$ переменной $h = (h^1, \dots, h^n) \in E^n$ 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

$$J''(u) = \begin{bmatrix} \frac{\partial^2 J(u)}{(\partial u^1)^2} & \frac{\partial^2 J(u)}{\partial u^1 \partial u^2} & \dots & \frac{\partial^2 J(u)}{\partial u^1 \partial u^n} \\ \frac{\partial^2 J(u)}{\partial u^2 \partial u^1} & \frac{\partial^2 J(u)}{(\partial u^2)^2} & \dots & \frac{\partial^2 J(u)}{\partial u^2 \partial u^n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 J(u)}{\partial u^n \partial u^1} & \frac{\partial^2 J(u)}{\partial u^n \partial u^2} & \dots & \frac{\partial^2 J(u)}{(\partial u^n)^2} \end{bmatrix} = \{J''_{u^i u^j}(u)\}, \quad (4)$$

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

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

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 $\langle Au, u \rangle = \sum_{i, \overline{1, n}}^n a_{ii} u^i u^i$ будет положительно определенной тогда и только тогда, когда все угловые миноры матрицы A , т. е. определители

$$\Delta_i = \det \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1i} \\ \cdot & \cdot & \cdot & \cdot \\ a_{i1} & a_{i2} & \dots & a_{ii} \end{pmatrix}, \quad i = \overline{1, n},$$

положительны; форма $\langle Au, u \rangle$ отрицательно определена тогда и только тогда, когда $(-1)^i \Delta_i > 0$, $i = \overline{1, n}$, т. е. знаки угловых миноров $\Delta_1, \Delta_2, \dots, \Delta_n$ 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 бы привлечь старшие производные и связанные с ними формы более высокого порядка:

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

где суммирование проводится по всем целым $r_1, r_2, \dots, r_n$, таким, что $0 \leq r_i \leq m$, $i = \overline{1, n}$, $r_1 + r_2 + \dots + r_n = m \geq 3$. Однако на практике исследование характера стационарных точек с помощью форм $d^m J(v)$, $m \geq 3$, почти не применяется из-за его громоздкости.

В тех случаях, когда указанным выше путем удастся 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 $u_i = (u_i^1, \dots, u_i^n)$, $i = \overline{1, m}$, и требуется найти точку $u \in E^n$, сум-

ма-квадратов расстояний от которой до этих данных точек была минимальной. Иначе говоря, требуется минимизировать функцию $J(u) = \sum_{i=1}^m |u - u_i|^2$ на E^n .

Так как градиент $J'(u) = 2 \sum_{i=1}^m (u - u_i)$, 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. Мат-

рица вторых производных $J''(u) \equiv 2mE$ при всех $u \in E^n$, где E — единичная матрица порядка $n \times n$, т. е. матрица, у которой элементы $a_{ii} = 1$, $a_{ij} = 0$ при $i \neq j$, $i, j = \overline{1, n}$. 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 $\lim_{|u| \rightarrow \infty} J(u) = +\infty$. Тогда по тео-

реме 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 $u = (u^1, \dots, u^n)$ 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 $u = (u^1, \dots, u^n)$ 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 $u =$

$$= (u^1, \dots, u^n) \text{ удовлетворяют ограничениям} \\ g_i(u) = 0, \dots, g_s(u) = 0; \quad (6)$$

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

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

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

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

выразив из (6) первые p переменных через остальные, то рассматриваемую задачу на условный экстремум можно свести к задаче на безусловный экстремум функции $\varphi(u^{p+1}, \dots, u^n) = J(a_1(u^{p+1}, \dots, u^n), \dots, a_p(u^{p+1}, \dots, u^n), u^{p+1}, \dots, u^n)$ переменных $(u^{p+1}, \dots, u^n) \in E^{n-p}$, 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

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

переменных $(u^1, \dots, u^n, \lambda_0, \lambda_1, \dots, \lambda_s) = (u, \bar{\lambda}) \in E^{n+s+1}$. Оказывается, если u_* — точка локального минимума или максимума функции $J(u)$ на множестве U (см. определения 1.1.6, 1.1.10), то необходимо существуют переменные $(\lambda_0^*, \lambda_1^*, \dots, \lambda_s^*) = \bar{\lambda}^* \neq 0$, называемые множителями Лагранжа, такие, что

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

т. е. векторы $J'(u)$, $g'_1(u)$, $\dots$, $g'_s(u)$ в точке u_* линейно зависимы. Докажем это. В самом деле, если бы эти векторы были линейно независимы, то якобиан системы функций $J(u) - J(u_*) - t, g_1(u), \dots, g_s(u)$ по переменным $u^1, \dots, u^n$ в точке $u = u_*, t = 0$ был бы отличен от нуля ([7], стр. 537). Тогда по

теореме о неявных функциях ([7], стр. 538) система

$$J(u) - J(u_*) - t = 0, \quad g_1(u) = 0, \dots, g_s(u) = 0$$

имеет решение при всех t , $|t| \leq t_0$, где t_0 — достаточно малое число; точнее, существует вектор-функция $u = u(t) = (u^1(t), \dots, u^n(t))$, которая определена и дифференцируема при всех t , $|t| \leq t_0$ и такая, что

$$u(0) = u_*, \quad J(u(t)) = J(u_*) + t, \quad g_i(u(t)) = 0, \quad i = \overline{1, s}.$$

Это значит, что $u(t) \in U$, $|t| \leq t_0$. Отсюда же имеем

$$J(u(t)) = J(u_*) + t > J(u_*) > J(u_*) - t = J(u(-t))$$

при всех t , $0 < t \leq t_0$, что противоречит тому, что u_* — точка локального минимума или максимума. Следовательно, условия (9) необходимы для того, чтобы u_* была точкой экстремума $J(u)$ на U . Иначе говоря, подозрительными на экстремум могут быть лишь те точки u , для которых существуют множители $\bar{\lambda} = (\lambda_0, \lambda_1, \dots, \lambda_s) \neq 0$, такие, что точка $(u, \bar{\lambda}) \in E^{n+s+1}$ удовлетворяет системе $n + s$ уравнений

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

и, кроме того, известно, что $\bar{\lambda} = (\lambda_0, \dots, \lambda_s) \neq 0$. 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 $(\lambda_0, \dots, \lambda_s)$ из (10) определяются с точностью постоянного множителя. Поэтому множители $\bar{\lambda}$ можно подчинить какому-либо дополнительному условию нормировки, например,

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

Если система (10) имеет решения $(v, \bar{\lambda})$ такие, что $\lambda_0 \neq 0$, то задачу минимизации $J(u)$ при условиях (6) называют невырожденной в точке v . В невырожденной задаче условие нормировки (11) можно заменить более простым условием $\lambda_0 = 1$. Нетрудно видеть, что для невырожденности задачи в точке v достаточно, чтобы векторы $g'_1(v), \dots, g'_s(v)$ были линейно независимы, т. е. чтобы равенство $\alpha_1 g'_1(v) + \dots + \alpha_s g'_s(v) = 0$ было возможно только при $\alpha_1 = \alpha_2 = \dots = \alpha_s = 0$.

Условия (10) с условием нормировки (11) (или $\lambda_0 = 1$) представляют собой систему $n + s + 1$ уравнений с $n + s + 1$ неизвестными $(u, \bar{\lambda}) = (u^1, \dots, u^n, \lambda_0, \dots, \lambda_s)$. Решив ее, мы найдем

точки $v \in U$, подозрительные на экстремум, и соответствующие им множители Лагранжа $\bar{\lambda} = \bar{\lambda}_v = (\lambda_0, \dots, \lambda_s) \neq 0$. 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 . А именно, если 1) функция $\mathcal{L}(u, \bar{\lambda})$ в окрестности точки v дважды дифференцируема по u и непрерывны все частные производные второго порядка в самой точке v ; 2) $(v, \bar{\lambda}_v)$ удовлетворяют системе (10), (11); 3) в точке v задача невырождена, т. е. $\lambda_0 > 0$; 4) квадратичная форма

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

для всех $h \neq 0$, для которых

$$\langle g'_i(v), h \rangle = 0, \quad i = \overline{1, s}, \quad (13)$$

то в этой точке v достигается локальный минимум [максимум] функции $J(u)$ при ограничениях (6).

Пример 2. Пусть требуется на n -мерной единичной сфере $U = \{u: u \in E^n, |u|^2 = \langle u, u \rangle = 1\}$ найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \dots, u_m \in E^n$ 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 $\mathcal{L}(u, \bar{\lambda}) = \lambda_0 J(u) + \lambda \langle u, u \rangle - 1$. Система (10) примет вид

$$\mathcal{L}_u(u, \bar{\lambda}) = \lambda_0 \cdot 2m(u - u_0) + 2\lambda u = 0, \quad \langle u, u \rangle = 1, \quad (14)$$

где $u_0 = \frac{1}{m} \sum_{i=1}^m u_i$ (см. пример 1). Очевидно, при $\lambda_0 = 0$ система

(14) не имеет решения с $(\lambda_0, \lambda) \neq 0$, поэтому можем принять $\lambda_0 = 1$. Тогда из (14) 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 $J(u)$ при условии $|u| = 1$. Так как 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 (14). Как выяснилось,

при $u_0 \neq 0$ система (14) имеет всего два решения v_1 и v_2 . Следовательно, $v_1 = u_0/|u_0|$ — точка глобального минимума, $v_2 = -u_0/|u_0|$ — глобального максимума при $u_0 \neq 0$. Таким образом, искомая точка $u_* = u_0/|u_0|$ при $u_0 \neq 0$.

Рассмотрим случай $u_0 = 0$. Тогда решением системы (14) 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 (14) 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: $\mathcal{L}(u, \tilde{\lambda}) = \lambda_0 x + \lambda(x^3 - y^2)$. Система (10), (11) имеет вид

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

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 вырождена в точке $v = 0$. Она просто решается, если из равенства $x^3 - y^2 = 0$ выразить $x = y^{2/3}$ и заметить, что тогда $J(u) = y^{2/3}$, $-\infty < y < \infty$. Ясно, что $v = (0, 0)$ — точка глобального минимума $J(u)$ на U .

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

$$h_1(u) \leq 0, \dots, h_m(u) \leq 0, \quad (15)$$

называемые ограничениями типа неравенств. Будем предполагать, что функции $J(u)$, $g_i(u)$, $h_j(u)$, $i = \overline{1, s}$; $j = \overline{1, m}$, определены и дифференцируемы во всех точках $u \in E^n$. Оказывается, если ввести новые вспомогательные переменные $w = (w^1, \dots, w^s, w^{s+1}, \dots, w^m)$, связанные с исходными переменными $u = (u^1, \dots, u^n)$

соотношениями

$$(w^1)^2 + h_1(u) = 0, \dots, (w^m)^2 + h_m(u) = 0, \quad (16)$$

то задача поиска экстремумов функции $J(u)$ при ограничениях (6), (15) сводится к равносильной задаче поиска экстремумов той же функции в пространстве переменных $\bar{u} = (u^1, \dots, u^n, w^1, \dots, w^m) = (u, w)$ при ограничениях типа равенств (6), (16). Равносильность этих двух задач понимается в следующем смысле: если u_* — точка локального минимума [максимума] функции $J(u)$ при ограничениях (6), (15), то $\bar{u}_* = (u_*, w_*)$, где $w_* = (w_*^1, \dots, w_*^m)$, $w_*^j = (-h_j(u_*))^{1/2}$, $j = 1, m$, будет точкой локального минимума [максимума] $J(u)$ при ограничениях (6), (16), и наоборот, если $\bar{u}_* = (u_*, w_*)$ — точка локального минимума [максимума] $J(u)$ при ограничениях (6), (16), то u_* — точка локального минимума [максимума] $J(u)$ при ограничениях (6), (15). Таким образом, для поиска экстремумов функции $J(u)$ при ограничениях (6), (16) можно ввести функцию Лагранжа

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

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 $w^1, \dots, w^m$, получить точки экстремума функции $J(u)$ при ограничениях (6), (15).

Пример 4. Пусть требуется в n -мерном единичном шаре $U = \{u: u \in E^n, |u|^2 = \langle u, u \rangle \leq 1\}$ найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \dots, u_m \in E^n$ была бы минимальной, т.е. требуется минимизировать функцию

$$J(u) = \sum_{i=1}^m |u - u_i|^2 \text{ при ограничениях } \langle u, u \rangle \leq 1.$$

Из примера 1 следует, что глобальный минимум $J(u)$ на всем пространстве E^n достигается в точке $u_* = u_0 = \frac{1}{m} \sum_{i=1}^m u_i$.

Поэтому если $|u_0| \leq 1$, то эта точка u_0 будет решением рассматриваемой задачи. Остается рассмотреть случай $|u_0| > 1$. Введем новую переменную w соотношением

$$w^2 + \langle u, u \rangle - 1 = 0 \quad (17)$$

и рассмотрим задачу минимизации $J(u)$ в пространстве переменных $\bar{u} = (u, w) = (u^1, \dots, u^n, w) \in E^{n+1}$ при ограничениях

(17). Составим функцию Лагранжа $\mathcal{L}(\bar{u}, \bar{\lambda}) = \lambda_0 \sum_{i=1}^m |u - u_i|^2 + \lambda(\omega^2 + |u|^2 - 1)$. Система (10) будет иметь вид

$$\begin{aligned}\mathcal{L}_u &= 2\lambda_0 m(u - u_0) + 2\lambda u = 0, \\ \mathcal{L}_\omega &= 2\lambda \omega = 0, \quad \omega^2 + |u|^2 - 1 = 0,\end{aligned}$$

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

$$(m + \lambda)u = mu_0, \quad \lambda\omega = 0, \quad \omega^2 + |u|^2 = 1, \quad |u_0| > 1. \quad (18)$$

Заметим, что здесь случай $\lambda = 0$ невозможен, так как тогда $u = u_0$ и $|u|^2 = |u_0|^2 = 1 - \omega^2 \leq 1$, что противоречит условию $|u_0| > 1$. Если же $\lambda \neq 0$, то из (18) 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 (17) может достигаться лишь в точках $\bar{v}_1$ и $\bar{v}_2$. Матрица вторых производных функции Лагранжа в найденных точках $(\bar{v}_1, \lambda_1)$ и $(\bar{v}_2, \lambda_2)$ соответственно имеет вид

$$\begin{pmatrix} m|u_0|E & 0 \\ 0 & m(|u_0| - 1) \end{pmatrix} \text{ и } \begin{pmatrix} -m|u_0|E & 0 \\ 0 & -m(|u_0| + 1) \end{pmatrix}.$$

Отсюда ясно, что при $|u_0| > 1$ в точке $\bar{v}_1 = (u_0/|u_0|, 0)$ достигается локальный минимум, а в точке $\bar{v}_2 = (-u_0/|u_0|, 0)$ — локальный максимум при ограничениях (17). Исключив переменную ω , 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 $J(u)$ на единичном шаре. Таким образом, искомой точкой будет $u_* = u_0$ при $|u_0| \leq 1$ 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), (15); 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 удастся 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.

У п р а ж н е н и я. 1. Найти экстремумы функций:

а) $J(u) = (x + y - 1)e^{-(x^2 - xy + y^2)}$, 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) $U = \{u = (x, y): x \geq 0, y \geq 0, x + y \leq 2\pi\}$.

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) $U = \{u = (x, y): 0 \leq x \leq 1, 0 \leq y \leq 1\}$;

б) $U = \{u = (x, y): x \geq 0, y \geq 0, ax + by = 1\}$, где a, b — неотрицательные числа;

в) $U = \{u = (x, y): x - y^2 \geq 0, x^2 + y^2 \leq 1\}$;

г) $U = \{u = (x, y): x^4 - 4x^3 + 4x^2 + 1 \leq y \leq 1\}$.

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 $\Delta = \Delta(u_1, \dots, u_n)$ — определитель матрицы $(u_1, \dots, u_n)$, столбцами которой являются векторы-столбцы u_i с координатами $(u_i^1, \dots, u_i^n)$, $i = \overline{1, n}$. Найти наибольшее и наименьшее значение величины определителя Δ при условии, что $|u_i| = a_i$, где a_i — заданные числа, $i = \overline{1, n}$. Доказать неравенство Адамара $|\Delta| \leq |u_1| \cdot |u_2| \cdot \dots \cdot |u_n|$. Дать геометрическую интерпретацию задачи при $n = 2$ и $n = 3$ (ср. с упражнением 5) ([7], стр. 554—557).

8. Найти наименьшее и наибольшее значение квадратичной формы

$$J(u) = \langle Au, u \rangle = \sum_{i,j=1}^n a_{ij} u_i u_j \quad \text{при условии } u \in U = \{u: u \in E^n, \langle u, u \rangle = 1\},$$

где A — симметрическая матрица. Показать, что $J_* = \min_U J(u)$ и $J^* = \max_U J(u)$

представляют собой наименьшее и соответственно наибольшее собственное число матрицы A ([128], стр. 209).

9. В множестве $U = \{u: u \in E^n, \langle c, u \rangle = 1\}$ или $U = \{u: u \in E^n, \langle c, u \rangle \leq 1\}$, где c — заданный вектор из E^n , найти точку, сумма квадратов расстояний от которой до m данных точек $u_1, \dots, u_m \in E^n$ 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 ми-

нимум 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 u , то $J(u)$ дифференцируема в точке u . На примере функции $J(u) = |xy|^{1/2}$, $u = (x, y)$, $u = (0, 0)$ показать, что одно лишь существование частных производных в точке u еще не гарантирует ее дифференцируемость в этой точке.

13. Доказать, что если u — точка локального минимума [максимума] дважды дифференцируемой функции $J(u)$ на всем пространстве E^n , то необходимо $\langle J''(u)h, h \rangle \geq 0$ [≤ 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 $\bar{\lambda}^* = (\lambda_0^*, \dots, \lambda_s^*)$ — соответствующие множители Лагранжа, определенные из системы (10), причем $\lambda_0^* = 1$, то $\langle \mathcal{L}_{uu}(u_*, \bar{\lambda}^*)h, h \rangle \geq 0$ [≤ 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 $U = \{u = (x, y) \in E^2: y - x^{3/2} = 0\}$ и $U = \{u: x \geq 0, y - x^{3/2} = 0\}$.

17. Пусть $J(u) = x$, $U = \{u = (x, y) \in E^2, g_1(u) = x^2 - y = 0, g_2(u) = x^2 + y = 0, g_3(u) = x = 0\}$. 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, $|J'(u+h) - J'(u)| \rightarrow 0$ при $|h| \rightarrow 0$ 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 $\|J''(u+h) - J''(u)\| \rightarrow 0$ при $|h| \rightarrow 0$ для всех $u, u+h \in U$. Множество дважды гладких на U функций обозначают через $C^2(U)$.

В определении 2 $\|A\| = \sup_{|e| \leq 1} |Ae|$ — 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 $C^2(U)$, 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 , $0 \leq t \leq 1$. 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

$$f'(t) = \langle J'(u + th), h \rangle, \quad f''(t) = \langle J''(u + th)h, h \rangle, \quad 0 \leq t \leq 1. \quad (1)$$

В самом деле, если, например, $J(u) \in C^2(U)$, то заменив в формуле (2.3) u на $u + th$, h на Δth , получим

$$\begin{aligned} f(t + \Delta t) - f(t) &= \\ &= \Delta t \langle J'(u + th), h \rangle + \frac{1}{2} (\Delta t)^2 \langle J''(u + th)h, h \rangle + o(|\Delta t|^2). \end{aligned}$$

Такое разложение означает, что $f(t) \in C^2[0, 1]$, и указывает на справедливость формул (1).

Для функций одной переменной имеют место формулы

$$f(t) - f(0) = f'(\theta_1 t) t = \int_0^t f'(\tau) d\tau = f'(0) t + \frac{1}{2} f''(\theta_2 t) t^2, \quad -$$

$$f'(t) - f'(0) = f''(\theta_3 t) t, \quad 0 \leq \theta_1, \theta_2, \theta_3 \leq 1.$$

Полагая в этих формулах $t = 1$ и пользуясь равенствами (1), получим различные формулы для конечных приращений функции многих переменных:

$$J(u + h) - J(u) = \langle J'(u + \theta_1 h), h \rangle = \int_0^1 \langle J'(u + th), h \rangle dt, \quad (2)$$

$$J(u + h) - J(u) = \langle J'(u), h \rangle + \frac{1}{2} \langle J''(u + \theta_2 h)h, h \rangle, \quad (3)$$

$$\langle J'(u + h) - J'(u), h \rangle = \langle J''(u + \theta_3 h)h, h \rangle, \quad (4)$$

где $0 \leq \theta_1, \theta_2, \theta_3 \leq 1$. Далее, так как

$$\frac{d}{dt} \langle J'(u + th), h \rangle = \langle J''(u + th)h, h \rangle, \quad 0 \leq t \leq 1,$$

то интегрируя это равенство по t на отрезке $[0, 1]$, получим

$$J'(u+h) - J'(u) = \int_0^1 J''(u+th) h dt = \left(\int_0^1 J''(u+th) dt \right) h. \quad (5)$$

Подчеркнем еще раз, что в формулах (1)–(5) подразумевается, что точки u , $u+h$ принадлежат множеству U вместе с отрезком $u+th$, $0 \leq t \leq 1$. В частности, эти формулы верны на любых выпуклых множествах — множествах, которые содержат вместе с любыми двумя своими точками u и v и отрезок $[u, v] = \{u_\alpha = \alpha u + (1-\alpha)v, 0 \leq \alpha \leq 1\}$, 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 функция $J(u) \in C^1(U)$. Скажем, что градиент $J'(u)$ этой функции удовлетворяет условию Липшица на множестве U с константой $L \geq 0$, если

$$|J'(u) - J'(v)| \leq L |u - v|, \quad u, v \in U. \quad (6)$$

Класс таких функций будем обозначать через $C^{1,1}(U)$.

Лемма 1. Пусть U — выпуклое множество, $J(u) \in C^{1,1}(U)$. Тогда

$$|J(u) - J(v) - \langle J'(v), u - v \rangle| \leq L |u - v|^2/2 \quad (7)$$

при всех $u, v \in U$.

Доказательство. С помощью формулы (2) имеем

$$J(u) - J(v) - \langle J'(v), u - v \rangle = \int_0^1 \langle J'(v + t(u - v)) - J'(v), u - v \rangle dt.$$

Воспользуемся известным неравенством Коши — Буняковского

$$|\langle a, b \rangle| \leq \|a\| \|b\|, \quad a, b \in E^n.$$

С учетом условия (6) получим

$$\begin{aligned} |J(u) - J(v) - \langle J'(v), u - v \rangle| &\leq \\ &\leq \int_0^1 |J'(v + t(u - v)) - J'(v)| \|u - v\| dt \leq \\ &\leq \int_0^1 L \|u - v\|^2 t dt = L \|u - v\|^2/2. \end{aligned}$$

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

$$a_{k+1} \leq a_k + \delta_k, \quad \delta_k \geq 0, \quad k=0, 1, \dots, \quad \sum_{k=0}^{\infty} \delta_k < \infty. \quad (8)$$

Тогда существует $\lim_{k \rightarrow \infty} a_k < +\infty$. Если $\{a_k\}$ ограничена еще и снизу, то $\lim_{k \rightarrow \infty} a_k$ конечен.

Заметим, что если $\delta_k = 0$, $k=0, 1, \dots$, то последовательность $\{a_k\}$ не возрастает, и лемма 1 превращается в хорошо известное утверждение о пределе монотонной последовательности.

Доказательство. Суммируя первое неравенство (8), имеем

$$a_m \leq a_k + \sum_{i=k}^{m-1} \delta_i \leq a_k + \sum_{i=k}^{\infty} \delta_i \quad (9)$$

при всех $m > k \geq 0$. Пусть $\lim_{k \rightarrow \infty} a_k = \lim_{n \rightarrow \infty} a_{k_n}$, $k_n < k_{n+1}$, $n=0, 1, \dots$; $\lim_{n \rightarrow \infty} k_n = \infty$. Положим в (9) $k = k_n$. Получим

$$a_m \leq a_{k_n} + \sum_{i=k_n}^{\infty} \delta_i, \quad m > k_n. \text{ Следовательно, } \overline{\lim}_{m \rightarrow \infty} a_m \leq a_{k_n} + \sum_{i=k_n}^{\infty} \delta_i$$

для всех $n=1, 2, \dots$. Отсюда при $n \rightarrow \infty$ имеем $\overline{\lim}_{m \rightarrow \infty} a_m \leq$

$$\leq \lim_{n \rightarrow \infty} a_{k_n} = \lim_{m \rightarrow \infty} a_m. \text{ Однако всегда } \lim_{m \rightarrow \infty} a_m \leq \overline{\lim}_{m \rightarrow \infty} a_m. \text{ Поэтому}$$

$$\overline{\lim}_{m \rightarrow \infty} a_m = \lim_{m \rightarrow \infty} a_m. \text{ Отсюда следует существование предела } \{a_k\}.$$

Далее, при $k=0$ из (9) следует ограниченность $\{a_k\}$ сверху. Поэтому, если $\{a_k\}$ ограничена еще и снизу, то $\lim_{k \rightarrow \infty} a_k$ конечен.

Лемма 3. Пусть числовая последовательность $\{b_k\}$ такова, что

$$b_{k+1} \geq b_k - \delta_k, \quad \delta_k \geq 0, \quad k=0, 1, \dots, \quad \sum_{k=0}^{\infty} \delta_k < \infty.$$

Тогда существует $\lim_{k \rightarrow \infty} b_k > -\infty$. Если $\{b_k\}$ ограничена еще и сверху, то $\lim_{k \rightarrow \infty} b_k$ конечен.

Эта лемма сводится к лемме 2, если принять $b_k = -a_k$, $k=0, 1, \dots$

Лемма 4. Пусть числовая последовательность $\{a_k\}$ такова, что

$$a_k \geq 0, \quad k = 0, 1, \dots; \quad a_k - a_{k+1} \geq Aa_k^2, \quad k \geq k_0 \geq 0, \quad (10)$$

$$A = \text{const} > 0.$$

Тогда $a_k = O(k^{-1})$, $k = 1, 2, \dots$, т. е. найдется постоянная $B > 0$ такая, что

$$0 \leq a_k \leq Bk^{-1}, \quad k = 1, 2, \dots \quad (11)$$

Доказательство. Если $a_m = 0$ при некотором $m \geq k_0$, то из (10) следует, что $a_k = 0$ при всех $k \geq m$, и оценка (11) становится тривиальной — в (11) достаточно взять $B = m \cdot \max_{1 \leq i \leq m} a_i$. Поэтому пусть $a_n > 0$ при всех $n \geq k_0$. Тогда из (10) имеем

$$\frac{1}{a_{n+1}} - \frac{1}{a_n} = \frac{a_n - a_{n+1}}{a_n a_{n+1}} \geq \frac{a_n}{a_{n+1}} A \geq A > 0, \quad n \geq k_0.$$

Суммируя эти неравенства по n от k_0 до некоторого $k - 1 \geq k_0$, получим $a_k^{-1} - a_{k_0}^{-1} \geq A(k - k_0)$ или $a_k \leq A^{-1}(k - k_0)^{-1}$, $k > k_0$. Но $(k - k_0)^{-1} \leq (k_0 + 1)k^{-1}$ при $k > k_0$, поэтому $0 < a_k < (k_0 + 1)A^{-1}k^{-1}$. Если $1 \leq k \leq k_0$, то $0 \leq a_k = ka_k k^{-1} \leq k_0 (\max_{1 \leq k \leq k_0} a_k) k^{-1}$. Остается в (11) принять $B = \max\{(k_0 + 1)A^{-1}; k_0 \max_{1 \leq k \leq k_0} a_k\}$.

Лемма 5. Пусть числовая последовательность $\{a_k\}$ удовлетворяет условиям

$$a_k \geq 0, \quad k \in N = \{1, 2, \dots\}; \quad (12)$$

$$a_{k+1} \leq a_k - \frac{a_k^2}{A} + \frac{A}{k^{2\rho}}, \quad k \in I_0; \quad (13)$$

$$a_k \leq Bk^{-2\rho}, \quad k \in I_1; \quad (14)$$

$$a_{k+1} \leq a_k + Ck^{-2\rho}, \quad k \in I_1, \quad k + 1 \in I_0, \quad (15)$$

где A, B, C, ρ — положительные постоянные, $\rho \leq 1$, а множества индексов I_0, I_1 таковы, что $I_0 \cup I_1 = N$, $I_0 \cap I_1 = \emptyset$ (случай $I_0 = \emptyset$ или $I_1 = \emptyset$ не исключаются). Тогда существует константа $D > 0$ такая, что

$$0 \leq a_k \leq Dk^{-\rho}, \quad k = 1, 2, \dots \quad (16)$$

Доказательство. Можем считать, что $A \geq B + C$, 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

$$4 \leq k_0^p < (k_0 + 1)^p \leq 6. \quad (17)$$

Убедимся в том, что такое число существует. Для этого перепишем (17) в равносильном виде: $4^e \leq k_0 \leq 6^e - 1$, где $e = p^{-1} \geq 1$. Существование такого числа k_0 будет доказано, если покажем, что длина отрезка $[4^e, 6^e - 1]$ при любом $e \geq 1$ не меньше 1, т. е. $6^e - 4^e - 1 \geq 1$ или $g(e) = 6^e - 4^e \geq 2$ при всех $e \geq 1$. Но $g'(e) = 6^e \ln 6 - 4^e \ln 4 \geq \ln 6 (6^e - 4^e) > 0$, $e \geq 1$, так что $g(e)$ строго монотонно возрастает при $e \geq 1$. Следовательно, $g(e) \geq g(1) = 2$ для всех $e \geq 1$. Таким образом, при каждом p , $0 < p \leq 1$ число k_0 , удовлетворяющее условиям (17), существует.

Покажем, что

$$a_{k_0+1} \leq 2A(k_0 + 1)^{-p}. \quad (18)$$

Может случиться, что $k_0 \in I_0$. Тогда воспользуемся неравенством (13). Заметим, что функция $f_k(a) = a - a^2 A^{-1} + A k^{-2p}$ достигает своего максимума на числовой оси при $a = A/2$, и поэтому $f_k(a) \leq f_k(A/2) = A/4 + A k^{-2p}$ для всех $a \geq 0$, $k = 1, 2, \dots$. Тогда из (13) с учетом неравенств (17) имеем

$$\begin{aligned} a_{k_0+1} &\leq f_k(a_{k_0}) \leq A/4 + A k_0^{-2p} \leq A/4 + A/16 \leq \\ &\leq (5A/16) 6(k_0 + 1)^{-p} < 2A(k_0 + 1)^{-p}. \end{aligned}$$

Если же $k_0 \in I_1$, то возможно и $k_0 + 1 \in I_1$. Тогда из (14) и (17) следует, что

$$a_{k_0+1} \leq B(k_0 + 1)^{-2p} \leq B(k_0 + 1)^{-p}/4 < 2A(k_0 + 1)^{-p}.$$

Если $k_0 \in I_1$, но $k_0 + 1 \in I_0$, то из (14), (15), (17) получим

$$a_{k_0+1} \leq B k_0^{-2p} + C k_0^{-2p} \leq A k_0^{-2p} < 2A(k_0 + 1)^{-p}.$$

Оценка (18) доказана. Далее сделаем индуктивное предположение: пусть при некотором $k \geq k_0 + 1$ верна оценка

$$a_k \leq 2A k^{-p}.$$

Возможно, что $k \in I_0$. Тогда с учетом (17) имеем $a_k \leq 2A k^{-p} \leq 2A k_0^{-p} \leq A/2$. Поскольку $f_k(a)$ монотонно возрастает на отрезке $[0, A/2]$, то из (13) следует $a_{k+1} \leq f_k(a_k) \leq f_k(2A k^{-p}) = 2A k^{-p} - 3A k^{-2p} < 2A(k^{-p} - k^{-2p})$. Но при $0 < p \leq 1$ справедливо соотношение

$$\begin{aligned} k^{-p} - k^{-2p} &< (k^p + 1)^{-1} < (k^p + k^{p-1})^{-1} = k^{-p+1} (k + 1)^{-1} < \\ &< (k + 1)^{-p}, \quad (19) \end{aligned}$$

поэтому $a_{k+1} < 2A(k+1)^{-p}$. Если же $k \in I_1$ и $k+1 \in I_1$, то из (14) и (17) получим

$$a_{k+1} \leq B(k+1)^{-2p} \leq B(k+1)^{-p}(k_0+1)^{-p} < 2A(k+1)^{-p}.$$

Если $k \in I_1$, но $k+1 \in I_0$, то из (14), (15), (17), (19) имеем

$$\begin{aligned} a_{k+1} &\leq (B+C)k^{-2p} \leq Ak^{-2p} < A(k_0^0-1)k^{-2p} < \\ &< A(k^0-1)k^{-2p} = A(k^{-p}-k^{-2p}) < 2A(k+1)^{-p}. \end{aligned}$$

Тем самым показано, что $a_k \leq 2Ak^{-p}$ при всех $k \geq k_0+1$.

Если $1 \leq k \leq k_0$, то $a_k = k^0 a_k k^{-p} \leq \max_{1 \leq k \leq k_0} a_k \cdot k_0 k^{-p}$. Остается в (16) принять $D = \max\{2A; k_0 \cdot \max_{1 \leq k \leq k_0} a_k\}$. Лемма 5 доказана.

Лемма 6. Пусть числовая последовательность $\{\omega_k\}$ такова, что

$$0 \leq \omega_{k+1} \leq (1-s_k)\omega_k + d_k, \quad k=1, 2, \dots; \quad \omega_1 \geq 0, \quad (20)$$

где

$$0 < s_k \leq 1, \quad d_k \geq 0, \quad k=1, 2, \dots; \quad \sum_{k=1}^{\infty} s_k = +\infty, \quad (21)$$

$$\lim_{k \rightarrow \infty} d_k/s_k = 0.$$

Тогда $\lim_{k \rightarrow \infty} \omega_k = 0$.

Доказательство. Поскольку $1-x \leq e^{-x}$ при $0 \leq x \leq 1$, то $1-s_k \leq e^{-s_k}$. Из неравенства (20) тогда имеем $0 \leq \omega_{k+1} \leq \omega_k e^{-s_k} + d_k$, $k=1, 2, \dots$. Отсюда с помощью индукции трудно получить, что

$$0 \leq \omega_{k+1} \leq \left(\omega_1 + \sum_{i=1}^k d_i e^{\sum_{j=1}^i s_j} \right) e^{-\sum_{j=1}^k s_j}, \quad k=1, 2, \dots \quad (22)$$

Далее воспользуемся известной теоремой Штольца ([7], стр. 88), представляющей собой разностный аналог правила Лопиталя и гласящей, что если последовательность $\{y_k\}$ монотонно возрастает, предел $\lim_{k \rightarrow \infty} (x_k - x_{k-1})/(y_k - y_{k-1})$ существует,

$\lim_{k \rightarrow \infty} y_k = +\infty$, то существует и предел $\lim_{k \rightarrow \infty} x_k/y_k$, причем

$$\lim_{k \rightarrow \infty} x_k/y_k = \lim_{k \rightarrow \infty} (x_k - x_{k-1})/(y_k - y_{k-1}).$$

Положим $y_k = e^{\sum_{j=1}^k s_j}$, $x_k = \omega_1 + \sum_{i=1}^k d_i e^{\sum_{j=1}^i s_j}$, $k=1, 2, \dots$. Из условий (21) следует, что $\{y_k\}$ монотонно возрастает и стремится

к бесконечности. Кроме того,

$$\begin{aligned}\lim_{k \rightarrow \infty} \frac{x_k - x_{k-1}}{y_k - y_{k-1}} &= \lim_{k \rightarrow \infty} d_k e^{\sum_{l=1}^k s_l} / \left(e^{\sum_{l=1}^k s_l} - e^{\sum_{l=1}^{k-1} s_l} \right) = \\ &= \lim_{k \rightarrow \infty} \frac{d_k}{1 - e^{-s_k}} = \lim_{k \rightarrow \infty} \left(\frac{d_k}{s_k} \right) \frac{s_k}{1 - e^{-s_k}} = 0,\end{aligned}$$

так как функция $x/(1 - e^{-x})$ ограничена на множестве $0 < x \leq 1$. По теореме Штольца с учетом неравенства (22) получим

$$\lim_{k \rightarrow \infty} \omega_k = \lim_{k \rightarrow \infty} x_k / y_k = \lim_{k \rightarrow \infty} (x_k - x_{k-1}) / (y_k - y_{k-1}) = 0.$$

Лемма 6 доказана.

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

$$0 \leq \omega_{k+1} \leq (1 - s_k) \omega_k + d_k, \quad k = 1, 2, \dots; \quad \omega_1 \geq 0, \quad (23)$$

где

$$0 \leq s_k \leq 1, \quad d_k \geq 0, \quad k = 1, 2, \dots; \quad \sup_{k \geq 1} d_k / s_k = c < \infty. \quad (24)$$

Тогда

$$0 \leq \omega_k \leq \omega_1 + c, \quad k = 1, 2, \dots \quad (25)$$

Доказательство легко проводится по индукции. При $k=1$ оценка (25) очевидна. Если (25) верно для некоторого $k \geq 1$, то из (23), (24) следует $0 \leq \omega_{k+1} \leq (1 - s_k)(\omega_1 + c) + d_k \leq (1 - s_k)\omega_1 + (1 - s_k)c + cs_k \leq \omega_1 + c$, что и требовалось.

Покажем, как может быть применена лемма 7 для оценки конкретных последовательностей.

Лемма 8. Пусть числовая последовательность $\{a_k\}$ такова, что

$$\begin{aligned}0 \leq a_{k+1} &\leq (1 - 1/k) a_k + c_1 / k^2, \quad k = 1, 2, \dots; \\ c_1 &= \text{const} > 0, \quad a_1 \geq 0.\end{aligned} \quad (26)$$

Тогда справедлива оценка

$$0 \leq a_k \leq c_2 \ln(k+1)/k, \quad k = 1, 2, \dots; \quad c_2 = \text{const} > 0. \quad (27)$$

Доказательство. Сделаем замену $\omega_k = a_k k (\ln(k+1))^{-1}$ и, пользуясь леммой 7, докажем ограниченность $\{\omega_k\}$. Из (26)

имеем

$$0 \leq \omega_{k+1} \leq \left(1 - \frac{1}{k}\right) \frac{k+1}{k} \frac{\ln(k+1)}{\ln(k+2)} \omega_k + c_1 \frac{k+1}{k^2 \ln(k+2)},$$

$$k = 1, 2, \dots$$

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 $\lim_{k \rightarrow \infty} d_k/s_k = c_1$, так что $\sup_{k \geq 1} d_k/s_k = c_3 < \infty$. Из леммы 7 тогда имеем $0 \leq \omega_k \leq \omega_1 + c_3$, $k = 1, 2, \dots$, что равносильно оценке (27) с $c_2 = c_3 + a_1/\ln 2$.

Лемма 9. Пусть числовая последовательность $\{a_k\}$ такова, что

$$0 \leq a_{k+1} \leq \left(1 - 1/k^\beta\right) a_k + c_1/k^{2\beta}, \quad k = 1, 2, \dots, \quad (28)$$

$$c_1 = \text{const} > 0, \quad 0 < \beta < 1, \quad a_1 \geq 0.$$

Тогда

$$0 \leq a_k \leq \left(a_1 + \frac{c_1}{1-\beta}\right) \frac{1}{k^\beta}, \quad k = 1, 2, \dots \quad (29)$$

Доказательство. Сделаем замену $\omega_k = k^\beta a_k$. Тогда из (28) имеем

$$0 \leq \omega_{k+1} \leq \left(1 - \frac{1}{k^\beta}\right) \cdot \frac{(k+1)^\beta}{k^\beta} \omega_k + c_1 \frac{(k+1)^\beta}{k^{2\beta}}.$$

Это значит, что $\{\omega_k\}$ удовлетворяет условиям (23) при $s_k = 1 - \left(1 - \frac{1}{k^\beta}\right) \left(1 + \frac{1}{k}\right)^\beta < 1$, $d_k = c_1 \left(1 + \frac{1}{k}\right)^\beta \frac{1}{k^\beta}$, $k = 1, 2, \dots$. Так как $(1 + 1/k)^{-\beta} \geq 1 - \beta/k$, $k = 1, 2, \dots$, то

$$s_k = \left(1 + \frac{1}{k}\right)^\beta \left[\left(1 + \frac{1}{k}\right)^{-\beta} - 1 + \frac{1}{k^\beta}\right] \geq$$

$$\geq \left(1 + \frac{1}{k}\right)^\beta \left[1 - \frac{\beta}{k} - 1 + \frac{1}{k^\beta}\right] =$$

$$= \left(1 + \frac{1}{k}\right)^\beta \cdot \frac{1}{k^\beta} \left(1 - \frac{\beta}{k^{1-\beta}}\right) \geq d_k \cdot \frac{1}{c_1} (1 - \beta) > 0, \quad k = 1, 2, \dots;$$

отсюда же следует, что $d_k/s_k \geq c_1/(1-\beta)$, $k = 1, 2, \dots$. По лемме 7 тогда $0 \leq \omega_k \leq \omega_1 + c_1/(1-\beta)$, что равносильно оценке (29).

Лемма 10. Пусть $\{z_k\}$, $\{w_k\}$ — некоторые последовательности из евклидова пространства E^n , z_* — точка из E^n , такие, что

$$a|w_{k+1} - z_k|^2 + |w_{k+1} - z_*|^2 \leq |z_k - z_*|^2, \quad (30)$$

$$|z_{k+1} - w_{k+1}| \leq b\delta_k, \quad \delta_k \geq 0, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \delta_k < \infty, \quad (31)$$

где a, b — положительные постоянные. Тогда существует конечный предел

$$\lim_{k \rightarrow \infty} |z_k - z_*| = \lim_{k \rightarrow \infty} |w_k - z_*| \quad (32)$$

и справедливо равенство

$$\lim_{k \rightarrow \infty} |w_{k+1} - z_k| = 0. \quad (33)$$

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 $|w_{k+1} - z_*| \leq |z_k - z_*|$. Тогда с помощью (31) имеем

$$|z_{k+1} - z_*| \leq |z_{k+1} - w_{k+1}| + |w_{k+1} - z_*| \leq b\delta_k + |z_k - z_*| \quad (34)$$

или

$$|z_{k+1} - z_*| \leq |z_k - z_*| + b\delta_k, \quad k = 0, 1, 2, \dots$$

Отсюда и из леммы 2 при $a_k = |z_k - z_*|$ вытекает существование конечного предела $\lim_{k \rightarrow \infty} |z_k - z_*|$. Из (34) тогда следует

(32), что в свою очередь гарантирует выполнение равенства (33). Наконец, пусть в (30) z_* — предельная точка z_k , пусть $\{z_{k_m}\} \rightarrow z_*$. Тогда $\lim_{k \rightarrow \infty} |z_k - z_*| = \lim_{m \rightarrow \infty} |z_{k_m} - z_*| = 0$, т. е. $\{z_k\} \rightarrow z_*$.

Отсюда и из (32) получим, что $\{w_k\} \rightarrow z_*$.

Лемма 11. Пусть неотрицательное число z таково, что

$$0 \leq z^p \leq bz + d, \quad (35)$$

где b, d — неотрицательные числа, $p > 1$. Тогда

$$0 \leq z \leq (b^q + qd)^{1/p}, \quad (36)$$

где q определяется равенством $p^{-1} + q^{-1} = 1$.

Доказательство. Если $d = 0$, то из (35) имеем $0 \leq z^{p-1} \leq b$ или $0 \leq z \leq b^{1/(p-1)} = b^{q/p}$, что совпадает с оценкой (36) при $d = 0$. Поэтому можем считать, что $d > 0$. Рассмотрим функцию

$$\varphi(x) = x^p - bx - d, \quad x \geq 0.$$

Нетрудно проверить, что эта функция имеет единственную точку минимума $x_* = (b/p)^{1/(p-1)} \geq 0$, $\varphi(x_*) \leq \varphi(0) = -d < 0$, $\lim_{x \rightarrow +\infty} \varphi(x) = +\infty$, строго монотонно убывает при $0 \leq x \leq x_*$ (если $x_* > 0$), строго монотонно возрастает при $x \geq x_*$. Отсюда следует, что уравнение $\varphi(x) = 0$ имеет единственное решение $x = \gamma$, так что $\varphi(x) < 0$ при $0 \leq x < \gamma$ и $\varphi(x) > 0$ при $x > \gamma$. Согласно (35) $\varphi(z) \leq 0$, поэтому справедлива оценка $0 \leq z \leq \gamma$. Однако получить явное выражение для γ в общем случае не удастся, поэтому в приложениях удобнее оценка (36). Для доказательства оценки (36) достаточно установить, что $\gamma \leq a = (b^q + qd)^{1/p}$. Пользуясь известным неравенством [10]

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q},$$

справедливым для всех действительных чисел $a, b, p > 1, q > 1, p^{-1} + q^{-1} = 1$, имеем $\varphi(a) = a^p - ab - d \geq a^p - a^p p^{-1} - b^q q^{-1} - d = a^{pq-1} - (b^q + qd) q^{-1} = 0$. Следовательно, $a \geq \gamma$ и $0 \leq z \leq a = (b^q + qd)^{1/p}$, что и требовалось.

§ 1. Постановка задачи

$$J(u) = c_1 u^1 + c_2 u^2 + \dots + c_n u^n \quad (1)$$
$$u^k \geq 0, \quad k \in I, \quad (2)$$

$$\begin{array}{rcl} a_{11}u^1 + \dots + a_{1n}u^n & \leq & b^1, \\ \vdots & & \vdots \\ a_{m1}u^1 + \dots + a_{mn}u^n & \leq & b^m \\ a_{m+1,1}u^1 + \dots + a_{m+1,n}u^n & = & b^{m+1} \\ \vdots & & \vdots \\ a_{s1}u^1 + \dots + a_{sn}u^n & = & b^s, \end{array} \quad (3)$$

$$J(u) = \langle c, u \rangle \rightarrow \inf; u \in U = \{u \in E^n: u^k \geq 0, k \in I; \langle a_i, u \rangle \leq b^i, \\ i = \overline{1, m}; \langle a_i, u \rangle = b^i, i = \overline{m+1, s}\}. \quad (4)$$

Если для каких-либо двух векторов $x = (x^1, \dots, x^p)$, $y = (y^1, \dots, y^p)$ справедливы неравенства $x^i \geq y^i$ при всех $i = 1, p$, то будем кратко писать: $x \geq y$. Тогда, например, неравенство $x \geq 0$ означает, что $x^i \geq 0$ для всех $i = 1, p$. Используя принятые обозначения, задачу (1)–(3) или (4) можно записать и в таком виде:

$$J(u) = \langle c, u \rangle \rightarrow \inf; \\ u \in U = \{u \in E^n: u^k \geq 0, \quad k \in I, \quad Au \leq b, \quad \bar{A}u = \bar{b}\}, \quad (5)$$

где

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}, \quad \bar{A} = \begin{pmatrix} a_{m+1,1} & \dots & a_{m+1,n} \\ \dots & \dots & \dots \\ a_{s1} & \dots & a_{sn} \end{pmatrix},$$

$$b = \begin{bmatrix} b^1 \\ \vdots \\ b^m \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} b^{m+1} \\ \vdots \\ b^s \end{bmatrix}.$$

Точку $u_* \in U$ назовем точкой минимума функции $\langle c, u \rangle$ на множестве U или, короче, решением задачи (4) или (5), если $\langle c, u_* \rangle = \inf_u \langle c, u \rangle$.

2. Приведем примеры прикладных задач, приводящих к задачам линейного программирования.

Задача оптимального планирования производства. Пусть на некотором предприятии изготавливаются 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 $u^1, u^2, \dots, u^n$ каждого вида продукции должно изготовить предприятие, чтобы обеспечить себе максимальную прибыль.

Если предприятие наметит себе план производства $u = (u^1, \dots, u^n)$, то оно израсходует $a_{i1}u^1 + \dots + a_{in}u^n$ единиц сырья i -го вида и получит $c_1u^1 + \dots + c_nu^n$ единиц прибыли. Ясно также, что все величины $u^i, i = 1, n$, неотрицательны. Поэтому мы приходим к следующей задаче линейного программирования: максимизировать функцию $J(u) = c_1u^1 + \dots + c_nu^n$ при ограничениях $u^1 \geq 0, \dots, u^n \geq 0, a_{i1}u^1 + \dots + a_{in}u^n \leq b^i, i = 1, s$. Поскольку задача максимизации $J(u)$ равносильна задаче минимизации $(-J(u))$, то с учетом введенных выше обозначений сформулированную задачу линейного программиро-

вания можно кратко записать в виде

$$(-c, u) \rightarrow \inf; \quad u \in U = \{u \in E^n; u \geq 0, \quad Au \leq b\}. \quad (6)$$

Ясно, что задача (6) является частным случаем задачи (5).

Задача об оптимальном использовании посевной площади. Пусть под посев p культур отведено r земельных участков площадью соответственно в $b_1, \dots, b_r$ 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

$$u_{1j} + u_{2j} + \dots + u_{pj} = b_j, \quad j = \overline{1, r}. \quad (7)$$

Ожидаемый средний урожай i -й культуры со всех участков равен $a_{i1}u_{i1} + a_{i2}u_{i2} + \dots + a_{in}u_{in}$ центнеров. Поскольку согласно плану должно быть произведено не менее d_i центнеров i -й культуры, то

$$a_{i1}u_{i1} + a_{i2}u_{i2} + \dots + a_{in}u_{in} \geq d_i, \quad i = \overline{1, p}. \quad (8)$$

Ожидаемая прибыль за урожай i -й культуры равна $c_i(a_{i1}u_{i1} + \dots + a_{in}u_{in})$, а за урожай всех культур —

$$\sum_{i=1}^p c_i(a_{i1}u_{i1} + a_{i2}u_{i2} + \dots + a_{in}u_{in}) = J(u). \quad (9)$$

Таким образом, приходим к задаче максимизации функции (9) (или минимизации функции $(-J(u))$) при условиях (7), (8) и естественных ограничениях

$$u_{ij} \geq 0, \quad i = \overline{1, p}, \quad j = \overline{1, r}.$$

Если умножить соотношения (8) на (-1) и переменные $\{u_{ij}\}$ переобозначить через $u^1, u^2, \dots, u^n$, 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

$$u_{i1} + u_{i2} + \dots + u_{ir} = a_i, \quad i = \overline{1, r} \quad (10)$$

тонн песка, j -му потребителю доставлено

$$u_{1j} + u_{2j} + \dots + u_{rj} = b_j, \quad j = \overline{1, p} \quad (11)$$

тонн песка, а стоимость перевозок будет равна

$$J(u) = \sum_{i=1}^p \sum_{j=1}^r c_{ij} u_{ij}. \quad (12)$$

Естественно требовать, чтобы

$$u_{ij} \geq 0, \quad i = \overline{1, r}, \quad j = \overline{1, p}. \quad (13)$$

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 [83] (стр. 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 попала 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

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au = b\}, \quad (14)$$

и основную задачу

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au \leq b\}, \quad (15)$$

где c, b — заданные векторы, $c \in E^n$, $c \neq 0$, $b \in E^m$, A — матрица порядка $m \times n$, $A \neq 0$. Несмотря на внешнее различие задач (14) и (15) (в одной из них ограничения $Au = b$, в другой — $Au \leq b$), эти задачи, оказывается, в определенном смысле эквивалентны. В самом деле, если ограничения типа равенств $Au = b$ заменить на равносильную систему двух неравенств: $Au \leq b$, $Au \geq b$, то каноническую задачу (14) можно записать в виде основной задачи

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au \leq b, \quad -Au \leq -b\}. \quad (16)$$

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 $v = (v^1, \dots, v^m)$ посредством соотношений $v = b - Au$, $v \geq 0$, и в пространстве E^{n+m} переменных $z = (u, v) = (u^1, \dots, u^n, v^1, \dots, v^m)$ рассмотрим каноническую задачу

$$\langle d, z \rangle \rightarrow \inf; \quad z \in Z = \{z = (u, v): z \in E^{n+m}, z \geq 0, Cz = Au + v = b\}, \quad (17)$$

где $d = (c, 0) \in E^{n+m}$, $C = (A, E)$, E — единичная матрица порядка $m \times m$. Нетрудно видеть, что если u_* — решение задачи (15), т. е. $u_* \in U$, $\langle c, u_* \rangle = \inf_U \langle c, u \rangle$, то $z_* = (u_*, v_*)$, $v_* = b - Au_*$ — решение задачи (17), т. е. $z_* \in Z$, $\langle d, z_* \rangle = \inf_Z \langle d, z \rangle$, 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

$$v = b - Au, \quad \omega^i = \max\{0; u^i\}, \quad \bar{\omega}^i = \max\{0; -u^i\}, \quad i \notin I, \quad (18)$$

и в пространстве переменных $z = (v, u^i (i \in I), \omega^i, \bar{\omega}^i (i \notin I))$ рассмотрим задачу

$$J(z) = \sum_{i \in I} c_i u^i + \sum_{i \notin I} c_i (\omega^i - \bar{\omega}^i) \rightarrow \inf;$$

$$z \in Z = \{z: v \geq 0, u^i \geq 0 (i \in I), \omega^i \geq 0, \bar{\omega}^i \geq 0 (i \notin I),$$

$$\sum_{i \in I} A_i u^i + \sum_{i \notin I} A_i (\omega^i - \bar{\omega}^i) + v = b, \sum_{i \in I} \bar{A}_i u^i + \sum_{i \notin I} \bar{A}_i (\omega^i - \bar{\omega}^i) = \bar{b}\}, \quad (19)$$

где A_i — i -й столбец матрицы A , $\bar{A}_i$ — i -й столбец $\bar{A}$. 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 $u^i = \omega^i - \bar{\omega}^i$ ($i \notin I$), остается лишь заметить, что точка $z_* = (v_*, u_*^i (i \in I), \omega_*^i, \bar{\omega}_*^i (i \notin I))$ будет решением задачи (19) тогда и только тог-

да, когда точка u_* с координатами $u_*^i (i \in I)$, $u_*^i = w_*^i - \bar{w}_*^i (i \notin I)$ 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 применяются, 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:

$$c_1x + c_2y \rightarrow \inf;$$

$$u \in U = \{u = (x, y): x \geq 0, y \geq 0, a_{i1}x + a_{i2}y \leq b^i, i = 1, m\}. \quad (1)$$

Введем множества: $U_0 = \{u = (x, y): x \geq 0, y \geq 0\}$ — положительный квадрант плоскости (x, y) , $U_i = \{u = (x, y): a_{i1}x + a_{i2}y \leq b^i\}$ — полуплоскость, образуемая прямой $a_{i1}x + a_{i2}y = b^i$, $i = \overline{1, m}$. Ясно, что множество U является пересечением множеств $U_0, U_1, \dots, U_m$. 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, составленная из отрезков каких-либо координатных осей и прямых $a_{i1}x + a_{i2}y = b^i$, $i = \overline{1, m}$. 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

$$c_1x + c_2y = \alpha \quad (2)$$

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 $+\infty$ прямая (2), смещаясь параллельно самой себе, «зачертит» («замечет») всю плоскость. При этом вектор c указывает направле-

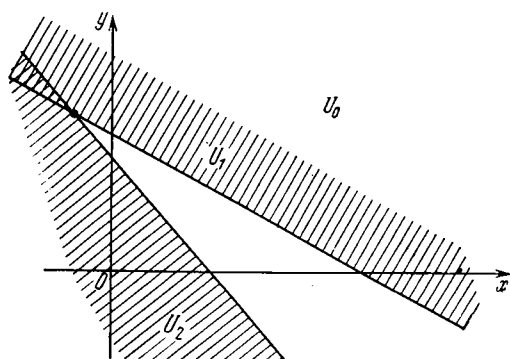


Рис. 3.1.

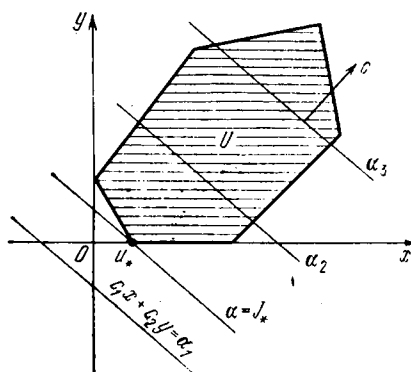


Рис. 3.2.

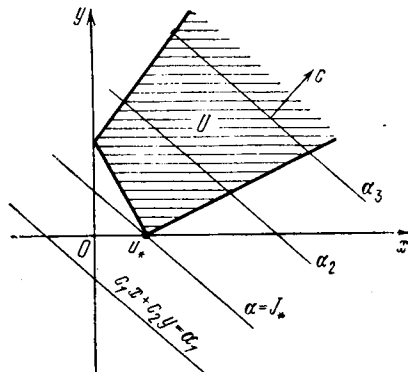


Рис. 3.3.

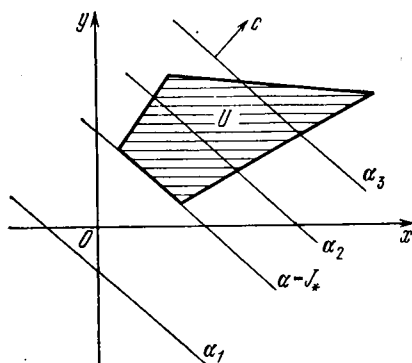


Рис. 3.4.

ние, в котором следует смещать прямую (2), чтобы увеличивать значение функции $J(u) = \langle c, u \rangle$. Если U — многоугольник (рис. 3.2), то при изменении α от $-\infty$ до $+\infty$ 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 $\langle c, u_* \rangle = J_* = \inf_U \langle c, u \rangle$, то есть u_* — решение задачи (1). Возможен случай, когда прямая (2) при первом касании с многоугольником U будет иметь

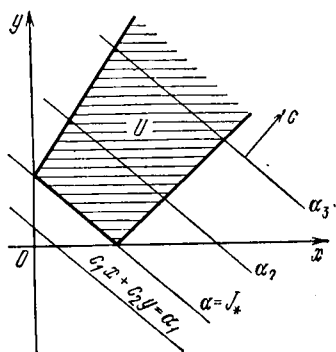


Рис. 3.5.

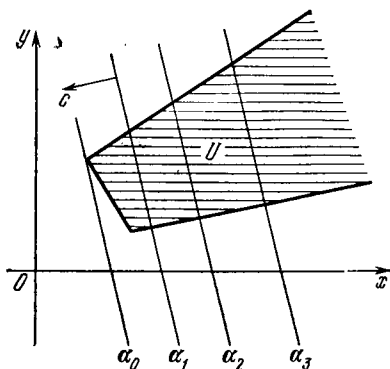


Рис. 3.6.

не одну общую точку с U , а целую сторону (рис. 3.4) — это может случиться, если U имеет сторону, перпендикулярную вектору c .

Если многоугольное множество U неограничено, то наряду со случаями, когда при первом касании прямая (2) будет иметь с U одну общую точку u_* (рис. 3.3) или сторону (рис. 3.5), возможна ситуация, когда прямая (2) при всех α , $-\infty < \alpha \leq \alpha_0 \leq +\infty$, имеет общую точку с U (рис. 3.6) — тогда $\inf_U \langle c, u \rangle = -\infty$ (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 (рис. 3.1, 3.6), может иметь лишь одно решение (рис. 3.2, 3.3), может иметь бесконечно много решений (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 $n = 2$ и $n = 3$.

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 неслучайно: 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

$$U = \{u \in E^n: u \geq 0, Au = b\}, \quad (3)$$

где A — матрица порядка $m \times n$, $A \neq 0$, 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:

$$A_1 u^1 + A_2 u^2 + \dots + A_n u^n = b. \quad (4)$$

Теорема 1. Пусть множество U определено условиями (3), $A \neq 0$, $r = \text{rang } A$ — ранг матрицы A . Для того чтобы точка $v = (v^1, v^2, \dots, v^n) \in U$ была угловой точкой множества U , необходимо и достаточно, чтобы существовали номера $j_1, \dots, j_r$, $1 \leq j_l \leq n$, $l = \overline{1, r}$, такие, что

$$A_{j_1} v^{j_1} + A_{j_2} v^{j_2} + \dots + A_{j_r} v^{j_r} = b; v^l = 0 \text{ при } j \neq j_l, l = \overline{1, r}, \quad (5)$$

причем столбцы $A_{j_1}, A_{j_2}, \dots, A_{j_r}$ линейно независимы.

Доказательство. Необходимость. Пусть v — угловая точка множества U . Если $v = 0$, то из условия $0 \in U$

следует $b = 0$. Так как $A \neq 0$, то $r = \text{rang } A \geq 1$ и существуют линейно независимые столбцы $A_{j_1}, \dots, A_{j_r}$. Отсюда имеем $A_{j_1} \cdot 0 + A_{j_2} \cdot 0 + \dots + A_{j_r} \cdot 0 = 0$. Для случая $v = 0$ соотношения (5) доказаны.

Пусть теперь $v \neq 0$ и пусть $v^{j_1}, \dots, v^{j_k}$ — все положительные координаты точки v . Отсюда и из условия $Av = b$ с учетом представления (4) имеем

$$A_{j_1} v^{j_1} + A_{j_2} v^{j_2} + \dots + A_{j_k} v^{j_k} = b; \quad v^l = 0 \text{ при } l \neq j_1, \dots, j_k. \quad (6)$$

Покажем, что столбцы $A_{j_1}, A_{j_2}, \dots, A_{j_k}$ линейно независимы. Пусть при некоторых $\alpha_1, \dots, \alpha_k$ имеет место равенство

$$\alpha_1 A_{j_1} + \alpha_2 A_{j_2} + \dots + \alpha_k A_{j_k} = 0. \quad (7)$$

Возьмем точку $v_+ = (v_+^1, \dots, v_+^n)$ с координатами $v_+^p = v^p + \varepsilon \alpha_p$, $v_+^l = 0$ при $l \neq j_1, \dots, j_k$ и точку $v_- = (v_-^1, \dots, v_-^n)$ с координатами $v_-^p = v^p - \varepsilon \alpha_p$, $v_-^l = 0$ при $l \neq j_1, \dots, j_k$. Так как $v^p > 0$, $p = j_1, \dots, j_k$, то при достаточно малых $\varepsilon > 0$ будем иметь $v_+ \geq 0$, $v_- \geq 0$. Кроме того, умножая (7) на $+\varepsilon$ или $-\varepsilon$ и складывая с (6), придем к равенствам $Av_+ = b$, $Av_- = b$. Таким образом, $v_+ \in U$, $v_- \in U$. 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 $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$. Таким образом, равенство (7) возможно только при $\alpha_1 = \dots = \alpha_k = 0$. Линейная независимость столбцов $A_{j_1}, \dots, A_{j_k}$ доказана. Отсюда следует, что $k \leq r$.

Если $k = r$, то соотношения (6) равносильны (5). Если $k < r$, то добавим к столбцам $A_{j_1}, \dots, A_{j_k}$ новые столбцы $A_{j_{k+1}}, \dots, A_{j_s}$ матрицы A так, чтобы система $A_{j_1}, \dots, A_{j_k}, A_{j_{k+1}}, \dots, A_{j_s}$ была линейно независимой, а при добавлении любого другого столбца A_l эта система становилась линейно зависимой. Тогда система $A_{j_1}, \dots, A_{j_s}$ образует некоторый базис линейной оболочки векторов $A_1, \dots, A_n$. Размерность линейной оболочки векторов $A_1, \dots, A_n$ равна рангу матрицы A , так что $s = r = \text{rang } A$. Добавив к первому равенству (6) столбцы $A_{j_{k+1}}, \dots, A_{j_r}$, умноженные соответственно на $v^{j_{k+1}} = 0, \dots, v^{j_r} = 0$, из (6) получим соотношения (5). Необходимость доказана.

Достаточность. Пусть некоторая точка $v = (v^1, \dots, v^n)$ удовлетворяет условиям (5), где $A_{j_1}, \dots, A_{j_r}$ — линейно независимы, $r = \text{rang } A$. Пусть $v = \alpha v_1 + (1 - \alpha)v_2$ при некоторых $v_1, v_2 \in U$, $0 < \alpha < 1$. Покажем, что такое представление воз-

можно только при $v_1 = v_2 = v$. Сразу же заметим, что если $v^j = 0$, то из этого представления с учетом неравенств $0 < \alpha < 1$, $v_1^j \geq 0$, $v_2^j \geq 0$ получим $0 \leq \alpha v_1^j + (1 - \alpha) v_2^j = v^j = 0$, 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 $v^{j_1}, \dots, v^{j_r}$. Произведя при необходимости перенумерацию переменных, можем считать, что $v^{j_1} > 0, \dots, v^{j_k} > 0, v^{j_{k+1}} = 0, \dots, v^{j_r} = 0$ (случаи $k = 0$ или $k = r$ здесь не исключаются). Тогда (4) можно переписать в виде $A_{j_1} v^{j_1} + \dots + A_{j_k} v^{j_k} = b$. Кроме того, учитывая, что по доказанному $v_1^j = v_2^j = 0$ при всех $j \neq j_p, p = \overline{1, k}$, равенства $Av_i = b$ также можно записать в виде $A_{j_1} v_i^{j_1} + \dots + A_{j_k} v_i^{j_k} = b, i = 1, 2$. Вспомним, что векторы $A_{j_1}, \dots, A_{j_k}$ линейно независимы. Поэтому вектор b может линейно выражаться через $A_{j_1}, \dots, A_{j_k}$ единственным способом. Это значит, что $v^{j_p} = v_1^{j_p} = v_2^{j_p}$ для $p = \overline{1, k}$. Тем самым установлено, что $v = v_1 = v_2$. Следовательно, v — угловая точка множества U . Теорема 1 доказана.

Определение 2. Систему векторов $A_{j_1}, \dots, A_{j_r}$, входящих в первое из равенств (5), называют *базисом угловой точки* v , а соответствующие им переменные $v^{j_1}, \dots, v^{j_r}$ — *базисными координатами* угловой точки v . Если все базисные координаты угловой точки положительны, то такую угловую точку называют *невырожденной*. Если же среди базисных координат $v^{j_1}, \dots, v^{j_r}$ хотя бы одна равна нулю, то такая угловая точка называется *вырожденной*. При фиксированном базисе $A_{j_1}, \dots, A_{j_r}$ переменные $u^{j_1}, \dots, u^{j_r}$ 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 координаты угловой точки. Если угловая точка вырожденная, то она может обладать несколькими базисами. В самом деле, если $v^{j_1} > 0, \dots, v^{j_k} > 0, k < r = \text{rang } A$, а остальные координаты v^j угловой точки v равны нулю, то, как видно из доказательства теоремы 1, в базис такой точки обязательно войдут столбцы $A_{j_1}, \dots, A_{j_k}$, а остальные базисные столбцы $A_{j_{k+1}}, \dots, A_{j_r}$, входящие в представление (5), могут быть выбраны, вообще говоря, различными способами.

Пример 1. Пусть $U = \{u = (u^1, u^2, u^3, u^4) \in E^4; u^j \geq 0; j = \overline{1, 4}; u^1 + u^2 + 3u^3 + u^4 = 3, u^1 - u^2 + u^3 + 2u^4 = 1\}$. Обозначим

$$A_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad b = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

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 $u_4 \notin U$.

Так как 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 $U \neq \emptyset$ и $\inf_u \langle c, u \rangle >$

$> -\infty$, 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 бы, например, рассмотреть C_n^r систем вида (5) и проверить, будут ли выбранные столбцы $A_{i_1}, \dots, A_{i_r}$ 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 *симплекс — 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 : $U = \left\{ u = (u^1, \dots, u^n): u^i \geq 0, \sum_{i=1}^n u^i = 1 \right\}$; затем

метод был обобщен на случай более общих множеств U , но первоначальное название за ним так и сохранилось; в литературе

чения из (2) соотношений (7) можно воспользоваться известными методами [2, 39, 64] (например, схемой Жордана). О том, как определить угловую точку и номера ее базисных переменных и как привести систему (2) к виду (7), будет рассказано ниже 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:

$$J(u) = \left\langle \bar{c}, \bar{v} - \sum_{i=r+1}^n B^{-1} A_i u^i \right\rangle + \sum_{i=r+1}^n c_i u^i = \\ = \langle \bar{c}, \bar{v} \rangle - \sum_{i=r+1}^n (\langle \bar{c}, B^{-1} A_i \rangle - c_i) u^i.$$

Поскольку $\langle \bar{c}, \bar{v} \rangle = \langle c, v \rangle = J(v)$, то

$$J(u) = J(v) - \sum_{i=r+1}^n \Delta_i u^i, \quad (8)$$

где

$$\Delta_i = \langle \bar{c}, B^{-1} A_i \rangle - c_i = \sum_{s=1}^r c_s \gamma_{si} - c_i. \quad (9)$$

Кстати, заметим, что величины Δ_i имеют смысл и при $i = \overline{1, r}$; тогда $\Delta_i = \langle \bar{c}, e_i \rangle - c_i = 0$, так как по определению обратной матрицы $B^{-1} A_i = e_i$, $i = \overline{1, r}$, 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 $\gamma_{sk} = (B^{-1} A_k)^s$, $\Delta_i = \langle \bar{c}, B^{-1} A_i \rangle - c_i$ пол-

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	$J(v)$

ностью 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, $u \geq 0$. 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 переменные сразу, то вряд ли сможем проследить и за изменением функции $J(u)$ и за соблюдением ограничений $u \geq 0$. Поэтому мы попробуем изменить лишь одну из небазисных переменных, скажем, переменную u^k , $r+1 \leq k \leq n$, а остальные небазисные переменные положим равными нулю. Тогда из соотношений (7) получим

$$\begin{aligned} u^1 &= v^1 - \gamma_{1k} u^k, \quad \dots, \quad u^i = v^i - \gamma_{ik} u^k, \quad \dots, \quad u^s = v^s - \gamma_{sk} u^k, \quad \dots \\ \dots, \quad u^r &= v^r - \gamma_{rk} u^k, \quad u^{r+1} = \dots = u^{k-1} = u^{k+1} = \dots = u^n = 0, \end{aligned} \quad (10)$$

или короче

$$\bar{u} = \bar{v} - (B^{-1}A_k) u^k, \quad u^j = 0, \quad j = \overline{r+1, n}, \quad j \neq k, \quad (11)$$

а функция (8) запишется в виде

$$J(u) = J(v) - \Delta_k u^k. \quad (12)$$

Наша ближайшая задача: выбрать номер k , $r+1 \leq k \leq n$, и величину $u^k \geq 0$ так, чтобы новая точка w , у которой k -я координата равна u^k , а остальные координаты определяются равенствами (10), удовлетворяла требованиям $Aw = b$, $w \geq 0$, $J(w) \leq J(v)$ (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 требованиям: $w \geq 0$, $J(w) \leq J(v)$, и указать правило выбора нужного номера k и нужной величины $u^k \geq 0$. А именно, рассмотрим следующие три взаимоисключающих случая I—III.

Случай I. Справедливы неравенства

$$\Delta_i = \langle \bar{c}, B^{-1}A_i \rangle - c_i \leq 0 \quad \text{при всех } i = \overline{r+1, n}, \quad (13)$$

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 специальным способом изменений переменных (10) ни при каких k , $r+1 \leq k \leq n$, 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 $u \in U = \{u: u \geq 0, Au = b\}$ с учетом представления (4)

$$\text{и неравенств (13) имеем } J(u) = \langle c, u \rangle = \sum_{i=1}^r c_i u^i + \sum_{i=r+1}^n c_i u^i \geq \\ \geq \langle \bar{c}, \bar{u} \rangle + \sum_{i=r+1}^n \langle \bar{c}, B^{-1} A_i \rangle u^i = \left\langle \bar{c}, \bar{u} + \sum_{i=r+1}^n B^{-1} A_i u^i \right\rangle = \langle \bar{c}, \bar{v} \rangle = \\ = J(v). \text{ Таким образом, } J(u) \geq J(v) \text{ при всех } u \in U, \text{ т. е. } v - \\ \text{решение задачи (1).}$$

Случай II. Существует номер k , $r+1 \leq k \leq n$, такой, что

$$\Delta_k > 0 \text{ и } \gamma_{ik} \leq 0 \text{ при всех } i = \overline{1, r}, \text{ т. е. } B^{-1} A_k \leq 0. \quad (14)$$

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 $u^k \geq 0$ точка u , определяемая формулой (10) (или (11)), будет иметь неотрицательные координаты и, следовательно, будет принадлежать множеству U . Тогда, как видно из (12), $J(u) = J(v) - \Delta_k u^k \rightarrow -\infty$ при $u^k \rightarrow +\infty$. Это значит, что $\inf_U \langle c, u \rangle = -\infty$, т. е. задача (1) не имеет решения.

Случай III. Существуют номера k , i , $r+1 \leq k \leq n$, $1 \leq i \leq r$, такие, что

$$\Delta_k > 0, \quad \gamma_{ik} > 0. \quad (15)$$

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

$$I_k = \{i: 1 \leq i \leq r, \gamma_{ik} > 0\}$$

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 $u \geq 0$. Как видно из формул (10) при $\gamma_{ik} \leq 0$, т. е. $i \notin I_k$, будем иметь $u^i = v^i - \gamma_{ik} u^k \geq v^i \geq 0$ при любом выборе $u^k \geq 0$. Если же $\gamma_{ik} > 0$, т. е. $i \in I_k$, то при слишком больших значениях u^k , а именно, при $u^k > \min_{i \in I_k} v^i / \gamma_{ik}$ ве-

личина $u^i = v^i - \gamma_{ik} u^k$ станет отрицательной хотя бы для одного номера $i \in I_k$. Таким образом, для обеспечения условия $u \geq 0$ для точек, определяемых формулами (10), нужно u^k взять так, чтобы $0 \leq u^k \leq \min_{i \in I_k} v^i / \gamma_{ik}$. Пусть

$$\min_{i \in I_k} \frac{v^i}{\gamma_{ik}} = \frac{v^s}{\gamma_{sk}}, \quad s \in I_k. \quad (16)$$

Так как 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 $w = (w^1, \dots, w^n)$ с координатами

$$\begin{aligned} w^1 &= v^1 - \gamma_{1k} \frac{v^s}{\gamma_{sk}}, \dots, w^i = v^i - \gamma_{ik} \frac{v^s}{\gamma_{sk}}, \dots, w^{s-1} = v^{s-1} - \gamma_{s-1,k} \frac{v^s}{\gamma_{sk}}, \\ w^s &= v^s - \gamma_{sk} \cdot \frac{v^s}{\gamma_{sk}} = 0, w^{s+1} = v^{s+1} - \gamma_{s+1,k} \frac{v^s}{\gamma_{sk}}, \dots, w^r = v^r - \gamma_{rk} \frac{v^s}{\gamma_{sk}}, \\ w^{r+1} &= 0, \dots, w^{k-1} = 0, w^k = \frac{v^s}{\gamma_{sk}}, w^{k+1} = 0, \dots, w^n = 0 \quad (17) \end{aligned}$$

и с соответствующим значением функции

$$J(w) = \langle c, w \rangle = J(v) - \Delta_k v^s / \gamma_{sk} \leq J(v). \quad (18)$$

В силу формул (7), (10) и условия (16) ясно, что $w \in U$. Покажем, что w — угловая точка множества U с базисом

$$A_1, \dots, A_{s-1}, A_{s+1}, \dots, A_r, A_k, \quad (19)$$

получающимся из базиса точки v заменой столбца A_s на A_k . Учитывая, что $w^s = w^{r+1} = \dots = w^{k-1} = w^{k+1} = \dots = w^n = 0$, условие $Aw = b$ можно записать в виде $A_1 w^1 + \dots + A_{s-1} w^{s-1} + A_{s+1} w^{s+1} + \dots + A_r w^r + A_k w^k = b$. Согласно теореме 2.1 остается показать, что система векторов (19) линейно независима. Пусть для некоторых чисел $\alpha_1, \dots, \alpha_{s-1}, \alpha_{s+1}, \dots, \alpha_r, \alpha_k$ оказалось, что

$$\alpha_1 A_1 + \dots + \alpha_{s-1} A_{s-1} + \alpha_{s+1} A_{s+1} + \dots + \alpha_r A_r + \alpha_k A_k = 0. \quad (20)$$

Поскольку $A_k = B B^{-1} A_k = \sum_{i=1}^r A_i (B^{-1} A_k)^i = \sum_{i=1}^r \gamma_{ik} A_i$, то из (20) следует

$$\sum_{i=1, i \neq s}^r \alpha_i A_i + \alpha_k \sum_{i=1}^r \gamma_{ik} A_i = \sum_{i=1, i \neq s}^r (\alpha_i + \alpha_k \gamma_{ik}) A_i + \alpha_k \gamma_{sk} A_s = 0.$$

Но система $A_1, \dots, A_s, \dots, A_r$ является базисом точки v и, следовательно, линейно независима. Тогда последнее равенство возможно лишь при $\alpha_i + \alpha_k \gamma_{ik} = 0, i = 1, r, i \neq s; \alpha_k \gamma_{sk} = 0$. Но

$\gamma_{sk} > 0$, как разрешающий элемент, поэтому $\alpha_k = 0$. А тогда все остальные $\alpha_i = 0$, $i = 1, r$, $i \neq s$. Таким образом, равенство (20) возможно лишь при $\alpha_1 = \dots = \alpha_{s-1} = \alpha_{s+1} = \dots = \alpha_r = \alpha_k = 0$, т. е. система (19) линейно независима. Тем самым показано, что ω — угловая точка множества U , причем система (19) является ее базисом, а $u^1, \dots, u^{s-1}, u^{s+1}, \dots, u^r, u^k$ — ее базисными переменными.

Пользуясь соотношениями (7), нетрудно выразить новые базисные переменные $u^1, \dots, u^{s-1}, u^{s+1}, \dots, u^r, u^k$ через остальные, небазисные переменные. Для этого из s -го уравнения системы (7) нужно выразить переменную u^k :

$$u^k = \frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n'} \frac{\gamma_{sj}}{\gamma_{sk}} u^j \quad (21)$$

(здесь и ниже $\sum'$ означает, что суммирование ведется по всем $j = r+1, \dots, n$, исключая $j = k$) и затем подставить ее во все остальные уравнения этой системы. С учетом равенств (17) получим

$$\begin{aligned} u^i &= v^i - \sum_{j=r+1}^n \gamma_{ij} u^j = \\ &= v^i - \sum_{j=r+1}^{n'} \gamma_{ij} u^j - \gamma_{ik} \left(\frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n'} \frac{\gamma_{sj}}{\gamma_{sk}} u^j \right) = \\ &= w^i - \left(-\frac{\gamma_{ik}}{\gamma_{sk}} \right) u^s - \sum_{j=r+1}^{n'} \left(\gamma_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}} \gamma_{sj} \right) u^j \quad (22) \end{aligned}$$

для всех $i = 1, \dots, s-1, s+1, \dots, r$. Аналогично, подставляя выражение (21) для u^k в (8), функцию $J(u)$ также можно выразить через новые небазисные переменные:

$$\begin{aligned} J(u) &= J(v) - \sum_{j=r+1}^{n'} \Delta_j u^j - \Delta_k \left(\frac{v^s}{\gamma_{sk}} - \frac{1}{\gamma_{sk}} u^s - \sum_{j=r+1}^{n'} \frac{\gamma_{sj}}{\gamma_{sk}} u^j \right) = \\ &= J(w) - \left(-\frac{\Delta_k}{\gamma_{sk}} \right) u^s - \sum_{j=r+1}^{n'} \left(\Delta_j - \Delta_k \frac{\gamma_{sj}}{\gamma_{sk}} \right) u^j; \quad (23) \end{aligned}$$

здесь мы учли равенство (18). Таким образом, базисные переменные угловой точки ω и значение функции $J(u)$ выражаются

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^{f_1}	...	$u^{f_{i-1}}$	...	u^k	...	u^{f_s}	...	u^j	...	u^{j_r}	...	u^n	Свободные члены
u^{f_1}	γ_{11}	...	1	...	0	...	γ_{1k}	...	0	...	γ_{1j}	...	0	...	γ_{1n}	v^{f_1}
u^{f_i}	γ_{i1}	...	0	...	1	...	γ_{ik}	...	0	...	γ_{ij}	...	0	...	γ_{in}	v^{f_i}
u^{f_s}	γ_{s1}	...	0	...	0	...	γ_{sk}	...	1	...	γ_{sj}	...	0	...	γ_{sn}	v^{f_s}
u^{f_r}	γ_{r1}	...	0	...	0	...	γ_{rk}	...	0	...	γ_{rj}	...	1	...	γ_{rn}	v^{f_r}
Функция	Δ_1	...	0	...	0	...	Δ_k	...	0	...	Δ_j	...	0	...	Δ_n	$J(v)$

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 $1, 2, \dots, r$. Нетрудно видеть, что если базисные переменные точки v имеют, скажем, номера $j_1, j_2, \dots, j_r$, $1 \leq j_1 < j_2 < \dots < j_r \leq n$, то можно выписать формулы

$$u^{j_i} = v^{j_i} - \sum_{k \notin I_v} \gamma_{ik} u^k, \quad i = \overline{1, r}; \quad J(u) = J(v) - \sum_{k \notin I_v} \Delta_k u^k,$$

выражающие базисные переменные и функцию $J(u)$ через небазисные переменные и аналогичные формулам (5)—(9); здесь

$$I_v = \{j_1, j_2, \dots, j_r\}, \quad \gamma_{ik} = (B^{-1}A_k)^i, \quad i = \overline{1, r}, \quad k \notin I_v,$$

$$B = (A_{j_1} A_{j_2} \dots A_{j_r}), \quad v^{j_i} = (B^{-1}b)^i, \quad i = \overline{1, r};$$

$$\Delta_k = \sum_{i=1}^r c_{j_i} \gamma_{ik} - c_k, \quad k \notin I_v; \quad \Delta_k = 0, \quad i \in I_v;$$

табл. 3 представляет собой симплекс-таблицу точки v . Как и выше, здесь нужно рассмотреть три случая.

С л у ч а й I.

$$\Delta_k \leq 0 \quad \text{при всех} \quad k \notin I_v; \quad (13')$$

С л у ч а й II. Существует номер $k \notin I_v$ такой, что

$$\Delta_k > 0, \quad B^{-1}A_k \leq 0; \quad (14')$$

С л у ч а й III. Существуют номера $k \notin I_v$, $j_i \in I_v$ такие, что

$$\Delta_k > 0, \quad (B^{-1}A_k)^i = \gamma_{ik} > 0. \quad (15')$$

Как и выше, нетрудно убедиться, что в случае I точка v является решением задачи (1), в случае II — $\inf_U \langle c, u \rangle = -\infty$ и задача (1) не имеет решения, а в случае III можно выбрать разрешающий элемент γ_{sk} из условий

$$\Delta_k > 0, \quad \min_{i \in I_{vk}} v^{j_i} / \gamma_{ik} = v^{j_s} / \gamma_{sk}, \quad I_{vk} = \{i: j_i \in I_v, \gamma_{ik} > 0\}, \quad (16')$$

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 значение

функции $J(w) \leq J(v)$. Возникает вопрос, когда возможно $J(w) = J(v)$ 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 $v^s \geq 0$ из-за $v \in U$, то из соотношений (18) нетрудно усмотреть, что $J(w) = J(v)$ тогда и только 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 положительными. 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: $w^s = w^{r+1} = \dots = w^{k-1} = w^{k+1} = \dots = w^n = 0$. 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: $w^i = \gamma_{ik}(v^i/\gamma_{ik} - v^s/\gamma_{sk}) > 0$ или $v^i/\gamma_{ik} > v^s/\gamma_{sk}$, 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 $\inf_U \langle c, u \rangle = -\infty$, — задача (1) не имеет реше-

ния, 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 — $\inf_U \langle c, u \rangle = -\infty$ — 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 формулы (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 $A_1, \dots, A_r$ 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. 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 $A_1, A_2, \dots, A_r$. К симплекс-таблице точки v_0 добавим еще r столбцов $e_1, e_2, \dots, e_r$ 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. формулы (25), а также (4.11)). Пусть уже сделано несколько шагов симплекс-метода с расширенной симплекс-таблицей и найдена очередная угловая точка v_1 , пусть в результате

Таблица 4

Базисные переменные	u^1	...	u^k	...	u^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	$I(v)$					

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	$I(v)$						

преобразований в дополнительных столбцах вместо первоначальных $e_1, \dots, e_r$ появились столбцы $\underline{d}_1, d_2, \dots, d_r$, где d_j имеет координаты $d_{1j}, d_{2j}, \dots, d_{rj}$, $j = 1, r$. Пусть табл. 5 представляет собой расширенную симплекс-таблицу точки v_l — в ней для простоты изложения предполагается, что базисом v_l являются столбцы $A_1, \dots, A_r$ (как уже отмечалось, этого всегда можно добиться перенумерацией переменных). Пусть $\Delta_k > 0$, $I_k = \{i: 1 \leq i \leq r, \gamma_{ik} > 0\} \neq \emptyset$.

Образует множество $I_{k1} = \left\{ s: s \in I_k, \min_{i \in I_k} v^i / \gamma_{ik} = v^s / \gamma_{sk} \right\}$.

В невырожденной задаче, как было выше замечено, множество I_{k1} всегда состоит из единственного номера s . В вырожденных задачах это не всегда так, и множество I_{k1} может состоять из двух и более номеров. В последнем случае образуем множество $I_{k2} = \left\{ s: s \in I_{k1}, \min_{i \in I_{k1}} d_{i1} / \gamma_{ik} = d_{s1} / \gamma_{sk} \right\}$. Если уже определено множество I_{km} , $m \geq 2$, и оно содержит не менее двух номеров, то образуем следующее множество:

$$I_{k, m+1} = \left\{ s: s \in I_{km}, \min_{i \in I_{km}} d_{im} / \gamma_{ik} = d_{sm} / \gamma_{sk} \right\}.$$

Ниже будет показано, что в конце концов мы получим множество I_{kl} , $1 \leq l \leq r+1$, состоящее из единственного номера s , причем все предыдущие множества I_{ki} , $i = 1, l-1$, будут содержать не менее двух номеров. Выведем переменную 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 заикливания, 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 следующем параграфе.

Следует заметить, что хотя среди прикладных задач линейного программирования вырожденные задачи встречаются довольно часто, но заикливание бывает крайне редко. Кроме того, использование антициклина на каждом шаге симплекс-метода приводит к заметному увеличению машинного времени ЭВМ, потребного 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 заикливание, принимают описанный выше или какой-нибудь другой антициклин, а после того, как заикливание будет преодолено, снова возвращаются к упрощенному правилу

выбора разрешающего элемента. Любопытно отметить, что для 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 зацикливание, принципиально важно для полного обоснования симплекс-метода для канонической задачи

$$J(u) = \langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au = b\}. \quad (1)$$

Покажем, что описанное в конце § 3 правило выбора разрешающего элемента в самом деле является антициклином. С этой целью рассмотрим так называемую возмущенную каноническую задачу

$$J(u) = \langle c, u \rangle \rightarrow \inf; \quad u \in U_\varepsilon = \left\{ u \in E^n: u \geq 0, \quad Au = b(\varepsilon) = b + \sum_{i=1}^r \varepsilon^i R_i \right\}, \quad (2)$$

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 предыдущем параграфе, предполагаем, что A — матрица порядка $r \times n$, $r = \text{rang } A < n$. 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 , $0 \leq \varepsilon \leq \varepsilon_0$, $\varepsilon_0 > 0$. Тогда найдется число ε_1 , $0 < \varepsilon_1 \leq \varepsilon_0$, такое, что 1) при всех ε , $0 < \varepsilon < \varepsilon_1$, задача (2) невырожденная; 2) если $v(\gamma) = (v^1(\gamma), \dots, v^n(\gamma))$ — какая-либо угловая точка множества U_γ , $0 < \gamma < \varepsilon_1$, с базисом $A_{j_1}, \dots, A_{j_r}$, то условия

$$A_{j_1} v^{j_1} + A_{j_2} v^{j_2} + \dots + A_{j_r} v^{j_r} = b(\varepsilon), \quad v^i = 0, \quad i \neq j_l, \quad l = \overline{1, r}, \quad (3)$$

при всех ε , $0 \leq \varepsilon < \varepsilon_1$, однозначно определяют угловую точку $v(\varepsilon) = (v^1(\varepsilon), \dots, v^n(\varepsilon))$ 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 $A_{j_1}, \dots, A_{j_r}$.

Доказательство. Возьмем произвольную линейно независимую систему столбцов $A_{j_1}, A_{j_2}, \dots, A_{j_r}$, где $r = \text{rang } A$. Обозначим $B = (A_{j_1}, A_{j_2}, \dots, A_{j_r})$ — это квадратная невырожденная матрица. Тогда система $Au = b(\varepsilon)$ определяет единственную точку $\bar{u} = u(\varepsilon) = (u^1(\varepsilon), \dots, u^n(\varepsilon))$ с координатами $u^i(\varepsilon) = 0$ при $i \neq j_l$, $l = \overline{1, r}$, а остальные координаты $\bar{u}(\varepsilon) = (u^{j_1}(\varepsilon), \dots, u^{j_r}(\varepsilon))$ определяются равенствами $u^{j_l}(\varepsilon) = (B^{-1}b)^l + \sum_{i=1}^r \varepsilon^i (B^{-1}R_i)^l$, $l = \overline{1, r}$, и представляют собой многочлены переменной ε степени не выше r ; последние равенства можно записать в

торной форме

$$\bar{u}(e) = B^{-1}b(e) = B^{-1}b + \sum_{i=1}^r e^i B^{-1}R_i.$$

Так как $\det R \neq 0$, $\det B \neq 0$, 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 $(B^{-1}R_1)^l$, $(B^{-1}R_2)^l$, ..., $(B^{-1}R_r)^l$ этой матрицы служат коэффициентами многочлена $u^{jl}(e)$, то ясно, что ни один из многочленов $u^{jl}(e)$, $l = \overline{1, r}$, не является тождественным нулю. Тогда каждый многочлен $u^{jl}(e)$ имеет не более r положительных корней или же таковых вовсе не имеет. Пусть η — наименьший из всех положительных корней всевозможных многочленов $u^{jl}(e)$, построенных по всевозможным линейно независимым наборам $A_{j_1}, A_{j_2}, \dots, A_{j_r}$ столбцов матрицы A . Так как 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 e , $0 < e < e_1 = \min\{\eta; 0\}$, и произвольную угловую точку $v(e)$ множества U_e с базисом $A_{j_1}, \dots, A_{j_r}$. Тогда базисные координаты $\bar{v}(e) = (v^{j_1}(e), \dots, v^{j_r}(e))$ точки $v(e)$, определяемые условиями $A_{j_1}v^{j_1}(e) + \dots + A_{j_r}v^{j_r}(e) = B\bar{v}(e) = b(e)$, представимы в виде

$$\bar{v}(e) = B^{-1}b(e) = B^{-1}b + \sum_{i=1}^r e^i (B^{-1}R_i)$$

и, следовательно, принадлежат рассмотренному выше множеству многочленов, а небазисные координаты, как им и предполагается, равны нулю. Из того, что $v(e) \in U_e$, следует, что $\bar{v}(e) \geq 0$. Но тогда в силу выбора e из $0 < e < e_1$ имеем $v(e) > 0$, т. е. $v(e)$ — невырожденная угловая точка множества U_e . Отсюда следует невырожденность задачи (2) при всех e , $0 < e < e_1$. Утверждение 1) леммы доказано.

Пусть $v(\gamma)$ — какая-либо угловая точка множества U_γ , $0 < \gamma < e_1$, с базисом $A_{j_1}, \dots, A_{j_r}$, т. е.

$$A_{j_1}v^{j_1}(\gamma) + \dots + A_{j_r}v^{j_r}(\gamma) = B\bar{v}(\gamma) = b(\gamma), \quad v^i(\gamma) = 0, \quad i \neq j_l, \quad l = \overline{1, r},$$

где $B = (A_{j_1}A_{j_2} \dots A_{j_r})$, $\bar{v}(\gamma) = (v^{j_1}(\gamma), \dots, v^{j_r}(\gamma))$. По доказанному выше $\bar{v}(\gamma) > 0$. Покажем, что точка $v(e)$, определяемая условиями (3), является угловой точкой множества U_e при каждом e , $0 \leq e < e_1$. Для этого прежде всего заметим, что условия (3) означают, что $Av(e) = b(e)$ при $0 \leq e < e_1$. Перепишем условия (3) в виде

$$\bar{v}(e) = B^{-1}b + \sum_{i=1}^r e^i (B^{-1}R_i), \quad v^i(e) = 0, \quad i \neq j_l, \quad l = \overline{1, r}. \quad (4)$$

Отсюда и из определения e_1 видно, что ни одна из координат вектора $\bar{v}(e)$ не может обратиться в нуль ни при каком e , $0 < e < e_1$. Но нам уже известно, что $\bar{v}(\gamma) > 0$. С учетом непрерывности $\bar{v}(e)$ тогда заключаем, что $\bar{v}(e) > 0$ при всех e , $0 < e < e_1$. Отсюда и из (4) при $e \rightarrow +0$ имеем $\lim_{e \rightarrow +0} v(e) = B^{-1}b = \bar{v}(0) = \bar{v} \geq 0$. Таким образом, показано, что точка $v(e)$, определяемая условиями (3), при всех e , $0 \leq e < e_1$, имеет неотрица-

тельные координаты. Следовательно, $v(\varepsilon) \in U_\varepsilon$, $0 \leq \varepsilon < \varepsilon_1$. Из (3) с помощью теоремы 2.1 заключаем, что $v(\varepsilon)$ — угловая точка множества U_ε с базисом $A_{j_1}, \dots, A_{j_r}$ при каждом ε , $0 \leq \varepsilon < \varepsilon_1$. Лемма 1 доказана.

2. Предполагая, что условия леммы 1 выполнены, рассмотрим подробнее один шаг симплекс-метода для задачи (2) при фиксированном ε , $0 < \varepsilon < \varepsilon_1$. Пусть $v(\varepsilon) = (v^1(\varepsilon), \dots, v^n(\varepsilon))$ — какая-либо угловая точка множества U_ε . Без умаления общности можем считать, что столбцы $A_1, A_2, \dots, A_r$ являются базисом, а координаты $(v^1(\varepsilon), \dots, v^r(\varepsilon)) = \bar{v}(\varepsilon)$ — базисными координатами этой точки. Согласно лемме 1 задача (2) невырожденная при $0 < \varepsilon < \varepsilon_1$, поэтому

$$\bar{v}(\varepsilon) > 0, \quad v^{r+1}(\varepsilon) = \dots = v^n(\varepsilon) = 0.$$

Обозначим $B = (A_1 A_2 \dots A_r)$. Поскольку $Av(\varepsilon) = B\bar{v}(\varepsilon) = b(\varepsilon)$, то $\bar{v}(\varepsilon) = B^{-1}b(\varepsilon) > 0$. Умножая уравнение $Au = B\bar{u} + A_{r+1}u^{r+1} + \dots + A_n u^n = b(\varepsilon)$ на B^{-1} слева, получим следующее соотношение между базисными переменными $\bar{u} = (u^1, \dots, u^r)$ и небазисными переменными $u^{r+1}, \dots, u^n$ точки $v(\varepsilon)$:

$$\bar{u} + \sum_{k=r+1}^n B^{-1}A_k u^k = B^{-1}b(\varepsilon) = \bar{v}(\varepsilon) = B^{-1}b + \sum_{j=1}^r \varepsilon^j (B^{-1}R_j). \quad (5)$$

Обозначим

$$B^{-1}A_k = \begin{bmatrix} \gamma_{1k} \\ \gamma_{2k} \\ \dots \\ \gamma_{rk} \end{bmatrix}, \quad d_j = B^{-1}R_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \dots \\ d_{rj} \end{bmatrix}, \quad \bar{v} = B^{-1}b = \begin{bmatrix} v^1 \\ v^2 \\ \dots \\ v^r \end{bmatrix}, \quad \bar{c} = \begin{bmatrix} c_1 \\ c_2 \\ \dots \\ c_r \end{bmatrix}.$$

Тогда соотношение (5) можно переписать в следующей покоординатной форме:

$$\begin{aligned} u^1 &+ \gamma_{1r+1}u^{r+1} + \dots + \gamma_{1k}u^k + \dots + \gamma_{1n}u^n = \\ &= v^1 + d_{11}\varepsilon + d_{12}\varepsilon^2 + \dots + d_{1r}\varepsilon^r, \\ u^s &+ \gamma_{sr+1}u^{r+1} + \dots + \gamma_{sk}u^k + \dots + \gamma_{sn}u^n = \\ &= v^s + d_{s1}\varepsilon + d_{s2}\varepsilon^2 + \dots + d_{sr}\varepsilon^r, \\ u^r &+ \gamma_{rr+1}u^{r+1} + \dots + \gamma_{rk}u^k + \dots + \gamma_{rn}u^n = \\ &= v^r + d_{r1}\varepsilon + d_{r2}\varepsilon^2 + \dots + d_{rr}\varepsilon^r, \end{aligned} \quad (6)$$

а для базисных координат $\bar{v}(\varepsilon) = B^{-1}b + \sum_{j=1}^r \varepsilon^j (B^{-1}R_j)$ получим

$$v^i(\varepsilon) = v^i + \sum_{j=1}^r d_{ij}\varepsilon^j, \quad i = \overline{1, r}. \quad (7)$$

Далее, по аналогии с (3.8), с помощью равенств (6) или (7) нетрудно выразить функцию $J(u) = \langle c, u \rangle$ через базисные переменные:

$$J(u) = J(v(\varepsilon)) - \sum_{i=r+1}^n \Delta_i u^i, \quad \Delta_i = \langle \bar{c}, B^{-1}A_i \rangle - c_i. \quad (8)$$

Учитывая, что $\Delta_i = \langle \bar{c}, B^{-1}A_i \rangle - c_i = 0$ при $i = \overline{1, r}$, на основе представлений (6), (8) выпишем симплекс-таблицу для угловой точки $v(\varepsilon)$, выделив в

ней для векторов $d_1, d_2, \dots, d_r$ дополнительные столбцы. В результате получим табл. 6.

Заметим, что точка $v = (v^1, \dots, v^n) = v(0)$ согласно лемме 1 является угловой для множества U с тем же базисом $A_1, \dots, A_r$, что и у точки $v(e)$. Ясно также, что выражения (3.5), (3.8) базисных переменных $\bar{u} = (u^1, \dots, u^r)$ угловой точки v и функции $J(u)$ через небазисные переменные $u^{r+1}, \dots, u^n$ получаются соответственно из (6), (8) при $e = 0$. Это значит что выписанная ранее табл. 5 является симплекс-таблицей точки v , расширенной дополнительными столбцами $d_1, \dots, d_r$; кстати, теперь выясняется, что элементы $d_{1i}, d_{2i}, \dots, d_{ri}$ i -й строки табл. 5 представляют собой коэффициенты многочленов (7).

Опираясь на симплекс-таблицу 6, сделаем один шаг симплекс-метода из угловой точки $v(e)$ для задачи (2) и посмотрим, чему это будет соответствовать в задаче (1). Как и в § 3, рассмотрим три случая.

Случай I. В нижней строчке симплекс-таблицы 6 величины $\Delta_i \leq 0$ при всех $i = \overline{r+1, n}$. Тогда, как и в § 3, нетрудно показать, что $v(e)$ — решение задачи (2), т. е. $\langle c, v(e) \rangle = \inf_{U_e} \langle c, u \rangle > -\infty$. Теперь заметим, что в

нижней строке симплекс-таблицы 5 угловой точки $v = v(0)$ находятся те же величины $\Delta_i \leq 0, i = \overline{r+1, n}$. Следовательно, $v = v(0)$ — решение задачи (1), т. е. $\langle c, v \rangle = \inf_U \langle c, u \rangle > -\infty$.

Случай II. Существует номер $k, r+1 \leq k \leq n$, такой, что $\Delta_k > 0$ и $\gamma_{ik} \leq 0$ при всех $i = \overline{1, r}$. Тогда, как и в § 3, нетрудно показать, что $\inf_U \langle c, u \rangle = -\infty$. Поскольку в симплекс-таблице 5 угловой точки $v = v(0)$ в столбце переменной u^k находятся те же величины $\Delta_k > 0, \gamma_{ik} \leq 0, i = \overline{1, r}$, то и в задаче (1) будем иметь $\inf_U \langle c, u \rangle = -\infty$. Таким образом, в случае II обе задачи (1) и (2) не имеют решения.

Случай III. Существуют номера $k, i, r+1 \leq k \leq n, 1 \leq i \leq r$, такие, что $\Delta_k > 0, \gamma_{ik} > 0$. Тогда множество $I_k = \{i: 1 \leq i \leq r, \gamma_{ik} > 0\}$ непусто и из условия

$$\min_{i \in I_k} \frac{v^i(e)}{\gamma_{ik}} = \frac{v^s(e)}{\gamma_{sk}}, \quad s \in I_k, \quad (9)$$

где $v^i(e)$ задано выражением (7), определяем номер s . Так как 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 $w(e) = (w^1(e), \dots, w^n(e))$ с координатами

$$\begin{aligned} w^i(e) &= v^i(e) - \frac{\gamma_{ik}}{\gamma_{sk}} v^s(e) = v^i + \sum_{j=1}^r d_{ij} e^j - \frac{\gamma_{ik}}{\gamma_{sk}} \left(v^s + \sum_{j=1}^r d_{sj} e^j \right) = \\ &= v^i - \frac{\gamma_{ik}}{\gamma_{sk}} v^s + \sum_{j=1}^r \left(d_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}} d_{sj} \right) e^j > 0, \quad i = 1, \dots, s-1, s+1, \dots, r; \end{aligned} \quad (10)$$

$$w^k(e) = \frac{v^s(e)}{\gamma_{sk}} = \frac{v^s}{\gamma_{sk}} + \sum_{j=1}^r \frac{d_{sj}}{\gamma_{sk}} e^j;$$

$$w^i(e) = 0, \quad i = s, r+1, \dots, k-1, k+1, \dots, n.$$

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	$J(v, \theta)$						

Таким образом, в симплекс-таблице угловой точки $w(e)$ в столбце свободных членов будут находиться числа

$$w^1 = v^1 - \gamma_{1k} \frac{v^s}{\gamma_{sk}}, \dots, w^i = v^i - \gamma_{ik} \frac{v^s}{\gamma_{sk}}, \dots, w^{s-1} = v^{s-1} - \gamma_{s-1,k} \frac{v^s}{\gamma_{sk}}, \\ w^{s+1} = v^{s+1} - \gamma_{s+1,k} \frac{v^s}{\gamma_{sk}}, \dots, w^r = v^r - \gamma_{rk} \frac{v^s}{\gamma_{sk}}, w^k = \frac{v^s}{\gamma_{sk}},$$

а в j -м дополнительном столбце d'_j , $j = \overline{1, r}$, появятся величины

$$d'_{ij} = d_{ij} - \frac{\gamma_{ik}}{\gamma_{sk}} d_{sj}, \quad i = 1, \dots, s-1, s+1, \dots, r; \quad d'_{kj} = \frac{d_{sj}}{\gamma_{sk}}. \quad (11)$$

Это значит, что дополнительные столбцы d_j симплекс-таблицы в задаче (2) преобразовываются по тем же правилам, по которым преобразовываются столбцы, соответствующие небазисным переменным, остающимися небазисными и на следующем шаге симплекс-метода (ср. с формулами (3.25)). Далее, из (8), (10) следует, что

$$J(w(e)) = J(v(e)) - \Delta_k \frac{v^s(e)}{\gamma_{sk}} < J(v(e)).$$

Наконец, исходя из равенств (6) и пользуясь формулами, аналогичными формулам (3.21)—(3.23), нетрудно показать, что остальные элементы γ_{ij} , Δ_j симплекс-таблицы 6 при переходе к следующей угловой точке $w(e)$ преобразовываются 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 $v(e)$ к точке $w(e)$. 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 $i \in I_k$, $\gamma_{ik} \leq 0$ при $i \notin I_k$, что и в симплекс-таблице 6, и 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 $w(e)$ при $e = 0$. Отсюда следует весьма важный вывод: если в возмущенной задаче (2) с помощью симплекс-метода совершается переход от угловой точки $v(e)$ к другой угловой точке $w(e)$ 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 $w = w(0)$. Рассмотренные случаи III закончены.

3. Теперь уже нетрудно дать обоснование антициклина, изложенного выше в конце § 3. Пусть $v_1 = (v_1^1, \dots, v_1^r)$ — некоторая угловая точка множества U с базисом $A_{j_1}, \dots, A_{j_r}$. Обозначим $B = (A_{j_1}, \dots, A_{j_r})$, $\bar{v}_1 = (v_1^1, \dots, v_1^r)$. Согласно теореме 2.1 тогда

$$Av_1 = A_{j_1} v_1^{j_1} + \dots + A_{j_r} v_1^{j_r} = B \bar{v}_1 = b; \quad v_1^l = 0, \quad l \neq j_l, \quad l = \overline{1, r},$$

так что $\bar{v}_1 = B^{-1}b \geq 0$. Рассмотрим следующую возмущенную задачу:

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U_\varepsilon = \left\{ u: u \geq 0, Au = b(\varepsilon) = b + \sum_{i=1}^r \varepsilon^i A_{j_i} \right\}, \quad (12)$$

получающуюся из задачи (2) при $R = (A_{j_1} A_{j_2} \dots A_{j_r}) = B$. Возьмем точку $v_1(\varepsilon) = (v_1^1(\varepsilon), \dots, v_1^n(\varepsilon))$ с координатами

$$v_1^j(\varepsilon) = v_1^j + \varepsilon^i; \quad v_1^j(\varepsilon) = 0, \quad j \neq j_i; \quad i = \overline{1, r}.$$

Напомним, что здесь и в (12) ε^i — i -я степень числа $\varepsilon > 0$. Так как $v_1 \geq 0$, то ясно, что $v_1(\varepsilon) \geq 0$, причем $\bar{v}_1(\varepsilon) = (v_1^1(\varepsilon), \dots, v_1^r(\varepsilon)) > 0$ при всех $\varepsilon > 0$. Кроме того,

$$Av_1(\varepsilon) = \sum_{i=1}^r A_{j_i} v_1^i(\varepsilon) = \sum_{i=1}^r A_{j_i} (v_1^i + \varepsilon^i) = b + \sum_{i=1}^r A_{j_i} \varepsilon^i = b(\varepsilon).$$

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 $A_{j_1}, \dots, A_{j_r}$, что и у точки 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 $v_1(\varepsilon), v_2(\varepsilon), \dots, v_m(\varepsilon)$ множества U_ε со значениями функции $J(v_1(\varepsilon)) > J(v_2(\varepsilon)) > \dots > J(v_m(\varepsilon))$, 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 $v_1(\varepsilon), \dots, v_m(\varepsilon)$, то получим последовательность угловых точек $v_1 = v_1(0), v_2 = v_2(0), \dots, v_m = v_m(0)$, 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 обе 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 заикливания 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 заикливания 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 эле-

мент γ_{sk} . Важно заметить, что коэффициенты многочленов $v^i(e)$, 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. таблицы 4, 5. В частности, для начальной точки v_1 с базисом $A_{j_1}, \dots, A_{j_r}$ в силу выбора матрицы $R = (A_{j_1} \dots A_{j_r}) = B$ 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 преобразовываются по формулам (11), аналогичным формулам (3.25).

Таким образом, коэффициенты многочленов

$$\frac{v_i(e)}{\gamma_{ik}} \equiv f_i(e) = c_{i0} + c_{i1}e + c_{i2}e^2 + \dots + c_{ir}e^r$$

на каждом шаге симплекс-метода могут быть определены без явного привлечения возмущенной задачи по формулам: $c_{i0} = v^i/\gamma_{ik}$, $c_{ij} = d_{ij}/\gamma_{ik}$, $j = \overline{1, r}$. Заметим, что все эти многочлены различны. В самом деле, если бы $f_i(e) \equiv f_m(e)$ при $i \neq m$, то коэффициенты $\{c_{ij}\}$ и $\{c_{mj}\}$ были бы равны: $c_{ij} = c_{mj}$ или $d_{ij} = d_{mj}\gamma_{ik}/\gamma_{mk}$, $j = \overline{1, r}$. Это означало бы, что пропорциональные i -я и m -я строки в матрице $D = (d_1 \dots d_r)$ из дополнительных столбцов расширенной симплекс-таблицы. Однако это невозможно, так как на каждом шаге матрица D невырожденная — она является произведением невырожденных матриц вида $(A_{j_1} \dots A_{j_r})^{-1}$ на R .

Таким образом, приходим к следующей задаче: задано конечное число различных многочленов $f_i(e) = v^i(e)/\gamma_{ik}$, $i \in I_k$, и требуется найти $\min_{i \in I_k} f_i(e)$ при достаточно малом, но заранее неизвестном $e > 0$. Здесь нам будет полезна

Л е м м а 2. Пусть дано множество многочленов

$$f_i(e) = c_{i0} + c_{i1}e + c_{i2}e^2 + \dots + c_{ir}e^r, \quad 0 < e < e_1, \quad i \in S_0,$$

степени не выше r , среди которых нет совпадающих; S_0 — конечное множество номеров. Тогда существует число e_2 , $0 < e_2 \leq e_1$ такое, что $\min_{i \in S_0} f_i(e)$ будет достигаться на единственном многочлене, номер которого не зависит от e , $0 < e < e_2$. Для определения номера этого многочлена нужно рекуррентным образом строить множества $S_0, S_1 = \{s: s \in S_0, c_{s0} = \min_{i \in S_0} c_{i0}\}, \dots, S_{m+1} = \{s: s \in S_m, c_{sm} = \min_{i \in S_m} c_{im}\}, \dots$ до тех пор, пока не будет обнаружено множество S_l , $0 \leq l \leq r+1$, 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 $f_s(e)$ с номерами $s \in S_1$ имеют один и тот же младший коэффициент $c_s = \min_{i \in S_0} c_{i0}$. Возможно, что $S_1 = S_0$. Тогда положим $e_{01} = e_1$ и заметим, что $\min_{i \in S_0} f_i(e) = \min_{i \in S_1} f_i(e)$ при всех e , $0 < e < e_{01}$. Если $S_1 \neq S_0$, т. е. $S_0 \setminus S_1 \neq \emptyset$, то положим

$$e_{01} = \min \left\{ 1; e_1; \frac{1}{2c} \left(\min_{i \in S_0 \setminus S_1} c_{i0} - c_{s0} \right), \quad s \in S_1 \right\} > 0,$$

где $c = \max_{i \in S_0} \sum_{j=1}^r |c_{ij}|$. Тогда для всех $s \in S_1$, $p \in S_0 \setminus S_1$, $0 < \varepsilon < \varepsilon_{01}$, имеем $f_s(\varepsilon) \leq c_{s0} + c\varepsilon < c_{p0} - c\varepsilon \leq f_p(\varepsilon)$, т. е. $f_s(\varepsilon) < f_p(\varepsilon)$. Это значит, что $\min_{i \in S_1} f_i(\varepsilon)$ при всех ε , $0 < \varepsilon < \varepsilon_{01}$, может достигаться лишь на многочленах с номерами $s \in S_1$, т. е. $\min_{i \in S_0} f_i(\varepsilon) = \min_{i \in S_1} f_i(\varepsilon) < f_p(\varepsilon)$ при всех ε , $0 < \varepsilon < \varepsilon_{01}$, и всех $p \notin S_1$. В частности, если S_1 состоит из единственного 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 $S_0 \supseteq S_1 \supseteq \dots \supseteq S_m$, $1 \leq m \leq r$, найдено число $\varepsilon_{0m} > 0$ и показано, что 1) $\min_{i \in S_0} f_i(\varepsilon) = \min_{i \in S_m} f_i(\varepsilon) < f_p(\varepsilon)$ 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 $\varepsilon^0, \varepsilon^1, \dots, \varepsilon^{m-1}$; 3) S_m состоит более чем из одного номера. Рассмотрим множество $S_{m+1} = \{s: s \in S_m, c_{sm} = \min_{i \in S_m} c_{im}\}$. Отсюда с учетом индуктивного предполо-

жения 2) заключаем, что все многочлены $f_s(\varepsilon)$ с номерами $s \in S_{m+1}$ имеют одинаковые коэффициенты при степенях $\varepsilon^0, \varepsilon^1, \dots, \varepsilon^m$. Возможно, что $S_{m+1} = S_m$. Тогда положим $\varepsilon_{0, m+1} = \varepsilon_{0m} > 0$ и с учетом индуктивного предположения 1) заметим, что $\min_{i \in S_0} f_i(\varepsilon) = \min_{i \in S_{m+1}} f_i(\varepsilon) < f_p(\varepsilon)$ 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 $f_s(\varepsilon) \leq c_{s0} + c_{s1}\varepsilon + \dots + c_{s, m-1}\varepsilon^{m-1} + \varepsilon^m(c_{sm} + c\varepsilon) < c_{s0} + c_{s1}\varepsilon + \dots + c_{s, m-1}\varepsilon^{m-1} + \varepsilon^m(c_{pm} - c\varepsilon) = c_{p0} + c_{p1}\varepsilon + \dots + c_{p, m-1}\varepsilon^{m-1} + c_{0m}\varepsilon^m - \varepsilon^{m+1}c \leq f_p(\varepsilon)$, т. е. $f_s(\varepsilon) < f_p(\varepsilon)$. Это значит, что $\min_{i \in S_m} f_i(\varepsilon)$ при всех ε , $0 < \varepsilon < \varepsilon_{0, m+1}$, может

достигаться лишь на многочленах с номерами $s \in S_{m+1}$, т. е. $\min_{i \in S_0} f_i(\varepsilon) = \min_{i \in S_{m+1}} f_i(\varepsilon) < f_p(\varepsilon)$ 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 $S_{r+1} = \{s: s \in S_r, c_{sr} = \min_{i \in S_r} c_{ir}\}$, зная, что $\min_{i \in S_0} f_i(\varepsilon) = \min_{i \in S_{r+1}} f_i(\varepsilon) < f_p(\varepsilon)$,

$p \notin S_{r+1}$, $0 < \varepsilon < \varepsilon_{0, r+1}$, и что все многочлены $f_s(\varepsilon)$ при $s \in S_{r+1}$ имеют одинаковые коэффициенты при степенях $\varepsilon^0, \varepsilon^1, \dots, \varepsilon^r$. 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 $\varepsilon_2 = \varepsilon_{0, r+1}$. Лемма 2 доказана.

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 $\min_{i \in I_k} v^i(\varepsilon)/\gamma_{ik}$, то мы получим именно тот самый антициклон, который был

описан в конце § 3. Этот антициклон в литературе часто называют *лексикографическим правилом выбора разрешающего элемента*. Поясним это название.

Определение 1. Вектор $a = (a^1, \dots, a^m)$ называют *лексикографически положительным* и пишут $a > 0$, 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 $a > b$, если $a - b > 0$.

Другими словами, $a > b$ 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: $a > b$, $b > a$ или $a = b$. Ясно, что если $a > b$, $b > c$, то $a > c$. 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 $S_0 \supseteq S_1 \supseteq \dots \supseteq S_m \supseteq \dots \supseteq S_l$ проводится лексикографическое упорядочивание векторов $\{c_i = (c_{i0}, c_{i1}, \dots, c_{ir}), i \in S_0\}$: $c_p > c_i$ 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. $c_p > c_s$ 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

$$J(u) = \langle c, u \rangle \rightarrow \inf, u \in U = \{u \in E^n: u \geq 0, Au = b\} \quad (1)$$

мы предполагали, что нам уже известна некоторая угловая точка множества U и что система $Au = b$ уже записана в виде (3.5), где $r = \text{rang } A$. Возникают вопросы: как узнать, не пусто ли множество

$$U = \{u \in E^n: u \geq 0, Au = b\}, \quad (2)$$

где A — матрица порядка $m \times n$, 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 $b \geq 0$, так как если $b^i < 0$ при некотором i , $1 \leq i \leq m$, то соответствующее i -е уравнение $(Au)^i = b^i$ из (2) можно умножить на (-1) . Наряду с основными переменными $u = (u^1, \dots, u^n)$ введем вспомогательные (искусственные) переменные $w = (u^{n+1}, u^{n+2}, \dots, u^{n+m})$ и в пространстве E^{n+m} переменных $z = (u^1, \dots, u^n, \dots, u^{n+m}) = (u, w)$ рассмотрим следующую каноническую задачу линейного программирования:

$$J_1(z) = u^{n+1} + u^{n+2} + \dots + u^{n+m} \rightarrow \inf,$$

$$z \in Z = \left\{ z: z = \begin{pmatrix} u \\ w \end{pmatrix} \in E^{n+m}, z \geq 0, Cz \equiv Au + w = b \right\}. \quad (3)$$

Систему $Cz \equiv Au + w = b$ перепишем в покоординатной форме

$$\begin{cases} a_{11}u^1 + \dots + a_{1n}u^n + u^{n+1} &= b^1, \\ a_{21}u^1 + \dots + a_{2n}u^n + u^{n+2} &= b^2, \\ \dots &\dots \\ a_{m1}u^1 + \dots + a_{mn}u^n + u^{n+m} &= b^m. \end{cases} \quad (4)$$

При $u^{n+1} = \dots = u^{n+m} = 0$ система (4), очевидно, равносильна системе $Au = b$. Множество Z непусто: оно содержит, например, точку $z_0 = (0, b) \geq 0$. Более того, нетрудно видеть, что точка z_0 является угловой точкой множества Z с базисом из последних m столбцов $e_1, e_2, \dots, e_m$ матрицы $C = (A, E)$, где E — единичная матрица порядка $m \times m$, $m = \text{rang } C \geq \text{rang } A = r$. Важно также заметить, что из системы (4) легко выразить базисные переменные $w = (u^{n+1}, \dots, u^{n+m})$ угловой точки z_0 через небазисные:

$$w = b - Au = b - A_1u^1 - \dots - A_nu^n.$$

Таким образом, задача (3) уже записана в форме, удобной для применения симплекс-метода с начальной угловой точкой z_0 .

Так как $J_1(z) \geq 0$ при всех $z \in Z$, то случай $\inf_Z J_1(z) = -\infty$ здесь невозможен. Поэтому, взяв в качестве начальной точку z_0 , с помощью симплекс-метода, описанного выше в § 3, за конечное число шагов найдем угловую точку $z_* = (v_1, w_1)$ множества Z , являющуюся решением задачи (3): $v^1 = (v_1^1, \dots, v_1^n)$, $w_1 = (v_1^{n+1}, \dots, v_1^{n+m})$, $\inf_Z J_1(z) = J_1(z_*) = v_1^{n+1} + \dots + v_1^{n+m} \geq 0$.

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, имели бы $J(z) = 0$, что противоречит тому, что $J(z) \geq J(z_*) > 0$ при всех $z \in Z$. Таким образом, при $J_1(z_*) > 0$ множество U пусто и задача (1) не имеет смысла.

Пусть теперь $J_1(z_*) = v_1^{n+1} + \dots + v_1^{n+m} = 0$. Тогда $w_1 = (v_1^{n+1}, \dots, v_1^{n+m}) = 0$, $z_* = (v_1, 0)$. 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 $z_* \geq 0$ следует $v_1 \geq 0$, а из $Cz_* = b$ имеем $Av_1 = b$. Это значит, что $v_1 \in U$. Рассмотрим представление

$$v_1 = \alpha u_1 + (1 - \alpha) u_2, \quad 0 < \alpha < 1, \quad u_1 \in U, \quad u_2 \in U, \quad (5)$$

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 $\text{rang } A$, 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. Так как $m = \text{rang } C \geq \text{rang } A$, то в базис точки z_* могут входить не только столбцы матрицы A , но и столбцы матрицы E (кстати, при $m > \text{rang } A$ в базис обязательно войдут столбцы E). Не умаляя общности, можем считать, что базисом точки z_* являются столбцы $A_1, \dots, A_r, e_1, \dots, e_{m-r}$ матрицы $C = (A, E)$, соответствующие основным переменным $u^1, \dots, u^r$ и вспомогательным переменным $u^{n+1}, \dots, u^{n+m-r}$ — именно эта ситуация отражена в симплекс-таблице 7. Подчеркнем еще раз, что из $z_* = (v_1, 0)$ следует, что $v_1^{n+1} = \dots = v_1^{n+m} = 0$ — эти равенства также нашли отражение в столбце свободных членов табл. 7. Нижняя строка симплекс-таблицы 7, содержащая величины $\Delta_i \geq 0$, $J_1(z_*) = 0$, 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 $u^{n+1}, \dots, u^{n+m}$, т. е. $r = m \leq n$ и столбцы $A_1, \dots, A_r$ составляют базис точки $z_* = (v_1, 0)$. Согласно правилу составления симплекс-таблиц тогда j -й строке табл. 7 будет соответствовать уравнение

$$u^j + \gamma_{j,r+1} u^{r+1} + \dots + \gamma_{j,n} u^n + \gamma_{j,n+1} u^{n+1} + \dots + \gamma_{j,n+m} u^{n+m} = v_1^j, \quad (6)$$

$$j = \overline{1, r=m}.$$

Пользуясь этими уравнениями, исключим вспомогательные переменные $u^{n+1}, \dots, u^{n+m}$ 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 $B = (A_1 \dots A_r)$. Поэтому, положив в уравнении (6)

$u^{n+1} = \dots = u^{n+m} = 0$, придем к системе

$$u^j + \gamma_{j, r+1} u^{r+1} + \dots + \gamma_{jn} u^n = v_j^i, \quad j = \overline{1, r+m}, \quad (7)$$

которая также может быть получена из $Au = b$ умножением слева на B^{-1} и потому равносильна системе $Au = b$. Это значит, что угловая точка v_i имеет своим базисом столбцы $A_1, \dots, A_r, r = \text{rang } A = m$, 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_i нужно вычеркнуть из табл. 7 все столбцы для $u^{n+1}, \dots, u^{n+m}$ и добавить нижнюю строку из

$$\Delta_1 = \dots = \Delta_r = 0, \quad \Delta_i = \sum_{s=1}^r c_s \gamma_{si} - c_i, \quad i = \overline{r+1, n}; \quad J(u_*).$$

Случай II. В табл. 7 имеются строки, соответствующие вспомогательным переменным, т. е. $r < m$, и среди коэффициентов $\gamma_{n+p, j}$, $p = \overline{1, m-r}$, $j = \overline{r+1, n}$, имеются положительные. Пусть, например, $\gamma_{n+i, k} > 0$, $1 \leq i \leq m-r$, $r+1 \leq k \leq n$. Тогда

$$I_k = \{j: 1 \leq j \leq r \text{ или } n+1 \leq j \leq n+m-r, \gamma_{jk} > 0\} \neq \emptyset.$$

Поскольку $v_j^i / \gamma_{jk} \geq 0$ для всех $j \in I_k$, а $v_1^{n+i} / \gamma_{n+i, k} = 0$, то

$$\min_{l \in I_k} v_l^i / \gamma_{lk} = v_1^{n+i} / \gamma_{n+i, k} = 0.$$

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	$\gamma_{n+m-r, r+1}$	...	$\gamma_{n+m-r, k}$	...	$\gamma_{n+m-r, n}$

Это значит, что любой коэффициент $\gamma_{n+i, k} > 0$, $1 \leq i \leq m - r$, $r + 1 \leq k \leq n$, 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 $A_1, \dots, A_r, e_1, \dots, e_{m-r}$ заменится новым базисом $A_1, \dots, A_r, e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_{m-r}, A_k$. Если и в новой симплекс-таблице, соответствующей новому базису точки z_* , на пересечении столбцов $u^{r+1}, \dots, u^{k-1}, u^{k+1}, \dots, u^n$ 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 $e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_{m-r}$ 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 $e_1, \dots, e_{m-r}$ — это будет означать, что после упомянутых операций мы пришли к симплекс-таблице, соответствующей рассмотренному выше первому случаю при $m = \text{rang } A$. Может, однако, получиться и так (при $\text{rang } A < m$ 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 коэф-

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	$\gamma_{1, n+m-r+1}$	...	$\gamma_{1, n+m}$	$v_1^1 \geq 0$
...	...	...	...	...	...	...	...	...
0	...	0	...	0	$\gamma_{j, n+m-r+1}$	...	$\gamma_{j, n+m}$	$v_j^1 \geq 0$
...	...	...	...	...	...	...	...	...
0	...	0	...	0	$\gamma_{r, n+m-r+1}$	...	$\gamma_{r, n+m}$	$v_r^1 \geq 0$
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	$\gamma_{n+m-r, n+m-r+1}$	...	$\gamma_{n+m-r, n+m}$	$v_1^{n+m} = 0$

фициента, и тогда описанный процесс вывода столбцов $\{e_i\}$ из базиса оборвется. Это будет означать, что реализовался третий случай, к рассмотрению которого мы сейчас переходим.

Случай III. В табл. 7 имеются строки, соответствующие вспомогательным переменным, т. е. $r < m$, но $\gamma_{n+p, j} \leq 0$ при всех $p = \overline{1, m-r}$, $j = \overline{r+1, n}$. Согласно правилу составления симплекс-таблиц строке переменной u^j , $j = \overline{1, r}$, в табл. 7 соответствует уравнение

$$u^j + \gamma_{j, r+1} u^{r+1} + \dots + \gamma_{jn} u^n + \dots + \gamma_{j, n+m-r+1} u^{n+m-r+1} + \dots + \gamma_{j, n+m} u^{n+m} = v_j^1, \quad j = \overline{1, r}, \quad (8)$$

а строке переменной u^{n+i} — уравнение

$$\gamma_{n+i, r+1} u^{r+1} + \dots + \gamma_{n+i, n} u^n + u^{n+i} + \gamma_{n+i, n+m-r+1} u^{n+m-r+1} + \dots + \gamma_{n+i, n+m} u^{n+m} = v_i^{n+i} = 0, \quad i = \overline{1, m-r}. \quad (9)$$

Заметим, что система (8), (9) может быть получена из $Au + w = b$ умножением слева на матрицу, обратную матрице $(A_1, \dots, A_r, e_1, \dots, e_{m-r})$, где $A_1, \dots, A_r, e_1, \dots, e_{m-r}$ — базис точки z_* , поэтому системы (4) и (8), (9) равносильны. С другой стороны, напоминаем, что при $u^{n+1} = \dots = u^{n+m} = 0$ система (4) превращается в систему, равносильную $Au = b$. Поэтому, полагая в (8), (9) $u^{n+1} = \dots = u^{n+m} = 0$, получим систему

$$u^j + \gamma_{j, r+1} u^{r+1} + \dots + \gamma_{jk} u^k + \dots + \gamma_{jn} u^n = v_j^1, \quad j = \overline{1, r}, \quad (10)$$

$$\gamma_{n+i, r+1} u^{r+1} + \dots + \gamma_{n+i, k} u^k + \dots + \gamma_{n+i, n} u^n = 0, \quad i = \overline{1, m-r}, \quad (11)$$

которая также равносильна системе $Au = b$.

Заметим, что в рассматриваемом случае в (11) все коэффициенты $\gamma_{n+i, j} \leq 0$, $j = \overline{r+1, n}$, $i = \overline{1, m-r}$. Возможно, что в некотором из уравнений (11), например, в $n+i$ -м уравнении коэффициенты $\gamma_{n+i, j} = 0$ при всех $j = \overline{r+1, n}$. 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) коэффициенты $\gamma_{n+i, s_1} < 0, \dots, \gamma_{n+i, s_q} < 0$, а остальные коэффициенты равны нулю. Тогда

это уравнение запишется в виде

$$\gamma_{n+i, s_1} u^{s_1} + \gamma_{n+i, s_2} u^{s_2} + \dots + \gamma_{n+i, s_q} u^{s_q} = 0.$$

Поскольку $u^{s_j} \geq 0$ для всех $u \in U$, а $\gamma_{n+i, s_j} < 0$, $j = \overline{1, q}$, то последнее уравнение будет удовлетворяться только при $u^{s_1} = 0, \dots, u^{s_q} = 0$.

Координату u^k , $r+1 \leq k \leq n$, назовем *отмеченной*, если $\gamma_{n+i, k} < 0$ хотя бы для одного номера i , $1 \leq i \leq m-r$. 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 однозначно 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_i , 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

$$\sum_{i=1}^r c_i u^i + \sum_{i \in I} c_i u^i \quad (12)$$

при ограничениях

$$u^1 \geq 0, \dots, u^r \geq 0; \quad u^i \geq 0, \quad i \in I, \quad (13)$$

$$u^j + \sum_{k \in I} \gamma_{jk} u^k = v_j^1, \quad j = \overline{1, r}, \quad (14)$$

где I — множество номеров тех из координат $u^{r+1}, \dots, u^n$, которые не являются отмеченными. Задача (12)–(14) уже приведена к виду, удобному для использования симплекс-метода: исходная угловая точка $(v_1^1, \dots, v_r^1, v_i^1 = 0, i \in I)$ известна, r — ранг матрицы ограничений (14), базисные переменные $u^1, \dots, u^r$ легко выражаются из (14) через небазисные.

Решая задачу (12)–(14) симплекс-методом, можно найти ее решение $u_1^1, \dots, u_r^1, u_i^1, i \in I$. 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 $r = n \leq m$, пожалуй, следует ее выделить особо. Из симплекс-таблицы 7 при $r = n$, полагая $u^{n+1} = \dots = u^{n+m} = 0$, получим $u^1 = v_1^1, \dots, u^n = v_n^1$. Это значит, что

при $r = n$ множество U состоит из одной точки $u = v_1$ и задача (1) становится малосодержательной.

Из проведенного 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 таблице 7 столбцы для вспомогательных переменных $u^{n+m-r+1}, \dots, u^{n+m}$ 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 вышеизложенного приведем один пример.

Пример 1. Минимизировать функцию

$$J(u) = u^1 + 3u^2 + 2u^3 + u^4 - 3u^5 \quad (15)$$

на множестве U , определяемом условиями

$$u^1 \geq 0, \quad u^2 \geq 0, \quad \dots, \quad u^5 \geq 0; \quad (16)$$

$$\begin{cases} u^1 + u^2 - 4u^3 + u^4 - 3u^5 = 3, \\ u^1 - 4u^3 + 2u^4 - 5u^5 = 6, \\ -2u^1 - 5u^2 + 8u^3 + u^4 = 3. \end{cases} \quad (17)$$

Для определения начальной угловой точки множества U введем вспомогательные переменные u^6, u^7, u^8 и в пространстве E^8 переменных $z = (u^1, u^2, \dots, u^8)$ рассмотрим задачу: минимизировать функцию

$$J_1(z) = u^6 + u^7 + u^8 \quad (18)$$

на множестве Z , определяемом условиями

$$u^1 \geq 0; \quad u^2 \geq 0, \dots, u^8 \geq 0; \quad (19)$$

$$\begin{cases} u^1 + u^2 - 4u^3 + u^4 - 3u^5 + u^6 = 3, \\ u^1 - 4u^3 + 2u^4 - 5u^5 + u^7 = 6, \\ -2u^1 - 5u^2 + 8u^3 + u^4 + u^8 = 3. \end{cases} \quad (20)$$

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

Т а б л и ц а 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

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 $u^1, u^2, \dots, u^5$ точки z_0 . В табл. 8 в качестве разрешающего элемента возьмем величину 2 из столбца u^4 и совершим симплекс-преобразование (см. формулы (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

Т а б л и ц а 9

Базисные переменные	u^1	u^2	u^3	u^5	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

этой таблицы видно, что в полученной угловой точке $z_1 = (0, 0, 0, 3, 0, 0, 0, 0)$ множества Z значение функции $J_1(z_1) = 0$. Так как $J_1(z) \geq 0$ при всех $z \in Z$, а $J_1(z_1) = 0$, то ясно, что $\inf J_1(z) = J_1(z_1) = 0$, 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

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^8	0	0	0	0

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 ,

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

u^4 соответствуют системе уравнений

$$u^1 = -2u^2 + 4u^3 + u^5, \quad u^4 = u^2 + 2u^5 + 3, \quad (21)$$

которая эквивалентна системе (17). В нижней строке табл. 11 представлены коэффициенты, полученные подстановкой выра-

жений (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) 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. условие (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. Приведем еще две такие теоремы.

Теорема 1. *Для того чтобы каноническая задача (1.14) была разрешима, т. е. существовала точка $u_* \in U$ такая, что $\langle c, u_* \rangle = \inf_U \langle c, u \rangle > -\infty$, 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 $U \neq \emptyset$, по теореме 5.1 следует существование угловой точки множества U . Принимая эту точку за начальную, будем решать задачу (1.14) с помощью симплекс-метода. Так как по условию $\inf_U J(u) > -\infty$, 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 доказана.

Теорема 2. *Если задача (1.14) разрешима, то среди ее решений найдется хотя бы одна угловая точка множества U .*

Доказательство. По условию теоремы $U \neq \emptyset$ и существует точка $u_* \in U$ такая, что $\langle c, u_* \rangle = \inf_U \langle c, u \rangle > -\infty$. 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). Теорема 2 доказана.

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 выше 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.

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 [16]. К задачам линейного программирования мы будем возвращаться еще в § 4.8.

Упражнения. 1. Минимизировать функцию $u^1 - u^2 - u^3 - u^4 + 2u^5$ при условиях $u^i \geq 0$, $i = \overline{1, 5}$; $u^1 + 3u^2 + u^3 + u^4 - 2u^5 = 10$; $2u^1 + 6u^2 + u^3 + 3u^4 - 4u^5 = 20$; $3u^1 + 10u^2 + u^3 + 6u^4 - 7u^5 = 30$.

2. Максимизировать функцию $u^1 - 4u^2 + 3u^3 + 10u^4$ при условиях $u^i \geq 0$, $i = \overline{1, 4}$; $u^1 + u^2 - u^3 + u^4 = 0$, $u^1 + 14u^2 + 10u^3 - 10u^4 = 11$.

3. Минимизировать функцию $u^1 + 2u^2 - 2u^3 + 5u^4$ при условиях $u^i \geq 0$, $i = \overline{1, 4}$; $u^1 + 2u^2 - u^3 - u^4 = 1$, $-u^1 + 2u^2 + 3u^3 + u^4 = 2$, $u^1 + 5u^2 + u^3 - u^4 = 5$.

4. Минимизировать функцию $u^1 + u^2 + u^3$ при условиях $u^1 \geq 0$, $u^2 \geq 0$, $u^3 \geq 0$; $u^1 + u^2 = 1$, $u^1 - u^2 = 1$, $3u^1 - u^2 = 3$.

5. Минимизировать функцию $u^1 + 2u^2 + 3u^3 + 4u^4$ при условиях $u^i \geq 0$, $i = \overline{1, 4}$; $u^1 + u^2 + u^3 + u^4 \geq 1$.

6. Минимизировать функцию $u^1 + u^2 + u^3 + u^4$ при условиях $u^i \geq 0$, $i = \overline{1, 5}$; $u^1 - u^2 \geq 0$, $u^1 + u^2 - u^3 + u^4 - u^5 \geq 1$. Рассмотреть эту же задачу при дополнительном ограничении $1 \leq u^1 \leq 2$.

7. Минимизировать функцию $u^3 - u^4 + u^5 - u^6$ при условиях $u^1 + u^3 - 2u^4 - 3u^5 + 4u^6 = 0$, $u^2 + 4u^3 - 3u^4 - 2u^5 + u^6 = 0$, $u^3 + u^4 + u^5 + u^6 + u^7 = 1$, $u^i \geq 0$, $i = \overline{1, 7}$ [255]. Показать, что в этой задаче можно получить заикливание, если с помощью симплекс-метода организовать перебор базисов в следующем порядке: $(A_3, A_2, A_7) \rightarrow (A_3, A_4, A_7) \rightarrow (A_5, A_4, A_7) \rightarrow$

$\rightarrow (A_5, A_6, A_7) \rightarrow (A_1, A_6, A_7) \rightarrow (A_1, A_2, A_7) \rightarrow (A_3, A_2, A_7)$. Применить для решения этой задачи симплекс-метод с антициклином, изложенный в конце § 3.

8. Обобщить симплекс-метод на случай задачи:

$$\langle c, u \rangle \rightarrow \inf; u \in U = \{u: u \in E^n, Au = b, \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\},$$

где α_i, β_i — заданные величины, $\alpha_i \leq \beta_i, i = \overline{1, n}$ (возможно, некоторые из $\alpha_i = -\infty$ и некоторые из $\beta_i = +\infty$).

9. Пусть $U = \{u: u \in E^n, \langle a_i, u \rangle \leq b^i, i = \overline{1, m}\}, m \geq n$. 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 $\langle a_i, v \rangle \leq b^i$, 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 ($i = \overline{1, m}$) 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).

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, 4, 9, 11, 13, 38, 47, 48, 56, 62, 102, 107, 113, 114, 160, 175, 197, 207, 220, 256].

§ 1. Выпуклые множества

1. Начнем с рассмотрения конкретных примеров выпуклых множеств.

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 $\alpha, 0 \leq \alpha \leq 1$. Иначе говоря, множество U выпукло, если отрезок $[v, u] = \{u_\alpha: u_\alpha = v + \alpha(u - v), 0 \leq \alpha \leq 1\}$, соединяющий любые две точки u, v из U , целиком лежит в U (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

$$S(u_0, R) = \{u: u \in E^n, |u - u_0| \leq R\}$$

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 неравенством треугольника, имеем $|\alpha u + (1 - \alpha)v - u_0| = |\alpha(u - u_0) + (1 - \alpha)(v - u_0)| \leq \alpha|u - u_0| + (1 - \alpha)|v - u_0| \leq \alpha R + (1 - \alpha)R = R$, 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

$$\Gamma = \Gamma(c, \gamma) = \{u: u \in E^n, \langle c, u \rangle = \gamma\},$$

где $c = (c_1, \dots, c_n) \neq 0$ — вектор из E^n , γ — действительное число. Это множество всегда непусто: если, например, $c_i \neq 0$, то точка u_0 с координатами $u^i = \gamma/c_i, u^j = 0, j = 1, n, j \neq i$, удов-

летворяет 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

$$\Gamma = \Gamma(c, u_0) = \{u: u \in E^n, \langle c, u - u_0 \rangle = 0\}.$$

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 $\Gamma = \{u: \langle c, u \rangle = \gamma\}$ — некоторая гиперплоскость. Тогда множества

$$\Gamma^+ = \{u: \langle c, u \rangle > \gamma\}, \quad \Gamma^- = \{u: \langle c, u \rangle < \gamma\}$$

называются открытыми полупространствами, а множества

$$\overline{\Gamma^+} = \{u: \langle c, u \rangle \geq \gamma\}, \quad \overline{\Gamma^-} = \{u: \langle c, u \rangle \leq \gamma\}$$

— замкнутыми полупространствами. Нетрудно видеть, что Γ^+ , Γ^- , $\overline{\Gamma^+}$, $\overline{\Gamma^-}$ — выпуклые множества. Например, если $u, v \in \overline{\Gamma^+}$, то $\langle c, \alpha u + (1 - \alpha)v \rangle = \alpha \langle c, u \rangle + (1 - \alpha) \langle c, v \rangle \geq \alpha \gamma + (1 - \alpha) \gamma = \gamma$ для всех $\alpha \in [0, 1]$.

Пример 4. Множество

$$M = \{u: u \in E^n, Au = b\} = \\ = \{u: u \in E^n, \langle a_i, u \rangle = b^i, i = \overline{1, m}\},$$

где A — заданная матрица порядка $m \times n$, a_i — i -я строка матрицы A ; $b = (b^1, \dots, b^m)$ — заданный вектор из E^m , называется *аффинным множеством* (или *линейным многообразием*). Если $u, v \in M$, то $A(\alpha u + (1 - \alpha)v) = \alpha Au + (1 - \alpha)Av = b$ при всех $\alpha \in [0, 1]$. Это значит, что всякое аффинное множество выпукло.

Согласно теореме Кронекера — Капелли аффинное множество непусто тогда и только тогда, когда матрица A и расширенная матрица $B = (A, b)$ имеют один и тот же ранг. Пустое множество $\emptyset$ (случай $\text{rang } A < \text{rang } B$, например, $A = 0, b \neq 0$) и все пространство E^n (случай $A = 0, b = 0$) представляют собой крайние и тривиальные примеры аффинных множеств. Рассмотрим случай $A \neq 0, \text{rang } A = \text{rang } B = r$. Тогда аффинное множество M состоит из тех и только тех точек, которые

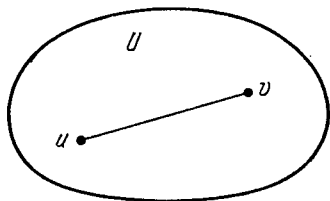


Рис. 4.1.

представимы в виде

$$u = u_0 + \sum_{i=1}^{n-r} t_i u_i,$$

где u_0 — какое-либо частное решение неоднородной системы линейных алгебраических уравнений $Au = b$, $u_1, u_2, \dots, u_{n-r}$ — линейно независимые (фундаментальные) решения однородной системы $Au = 0$, $t_1, \dots, t_{n-r}$ — действительные числа. Если обозначить

$$L_A = \left\{ u: u \in E^n, u = \sum_{i=1}^{n-r} t_i u_i, t_1, \dots, t_{n-r} \text{ — действительные числа} \right\}$$

— подпространство, образуемое всеми решениями однородной системы $Au = 0$, то аффинное множество представимо в виде

$$M = u_0 + L_0 = \{u: u = u_0 + v, v \in L_A\}.$$

Это означает, что всякое нетривиальное аффинное множество получается из некоторого линейного подпространства L_A параллельным сдвигом на некоторый вектор u_0 . Размерность аффинного множества M по определению принимается равной размерности подпространства L_A , параллельного M . Аффинное множество размерности $n-r$ часто обозначают через Γ_{n-r} и называют гиперплоскостью размерности $n-r$. В частности, если $r = n-1 = \text{rang } A = \text{rang } B$ и, следовательно, система $Au = 0$ имеет одно линейно независимое решение $u_1 = d$, то аффинное множество $M = \{u: Au = b\} = \{u: u = u_0 + td, -\infty < t < +\infty\}$, имеющее размерность 1, называют прямой линией в E^n с направляющим вектором $d \neq 0$. Множество $\Gamma_1^+ = \{u: u = u_0 + td, t \geq 0\}$ называют лучом, выходящим из точки u_0 в направлении вектора $d \neq 0$. Очевидно, прямая и луч в E^n являются выпуклыми множествами. Предлагаем читателю убедиться в том, что аффинное множество размерности $n-1$ представляет собой гиперплоскость.

Определение 2. Пересечение всех аффинных множеств, содержащих множество U из E^n , называется *аффинной оболочкой* множества U и обозначается через $\text{aff } U$.

Так как пересечение любого числа аффинных множеств, очевидно, само является аффинным множеством, то $\text{aff } U$ представляет собой минимальное аффинное множество, содержащее U .

Определение 3. *Размерностью произвольного множества U из E^n называется размерность его аффинной оболочки; размерность множества U обозначают $\dim U$.*

Согласно этому определению, отрезок $[u, v] = \{u_\alpha = v + \alpha(u-v), 0 \leq \alpha \leq 1\}$, соединяющий две точки $u, v \in E^n, u \neq v$, имеет размерность 1, так как его аффинной оболочкой является прямая $\Gamma = \{w: w = v + t(u-v), -\infty < t < +\infty\}$.

Пример 5. Множество

$$U = \{u: u \in E^n, Au \leq b\},$$

где A — заданная матрица порядка $m \times n$, b — заданный вектор из E^m , выпукло. Напоминаем, что неравенство $Au \leq b$ означает, что $\langle a_i, u \rangle = (Au)^i \leq b^i$ при всех $i = \overline{1, m}$, a_i — i -я строка матрицы A . Тогда для любых $u, v \in U$ имеем $A(\alpha u + (1 - \alpha)v) = \alpha Au + (1 - \alpha)Av \leq b$ при всех $\alpha \in [0, 1]$.

Пример 6. Множество

$$U = \{u = (u^1, \dots, u^n): \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\},$$

где α_i, β_i — заданные величины, $\alpha_i < \beta_i$ (возможно, что некоторые из $\alpha_i = -\infty$ и некоторые из $\beta_i = +\infty$), выпукло и имеет размерность n . В частности, неотрицательный ортант пространства E^n — это множество $E_+^n = \{u: u \in E^n, u \geq 0\}$ — выпукло, причем $\dim E_+^n = n$. Если в определении множества U величины α_i, β_i конечны при всех $i = \overline{1, n}$, то это множество называют n -мерным параллелепипедом.

Пример 7. Множество

$$U = \{u = (u^1, \dots, u^n): u^i \geq 0, i \in I; Au \leq b, \bar{A}u = \bar{b}\},$$

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 $U_0 = \{u: u^i \geq 0, i \in I\}$, $U_1 = \{u: Au \leq b\}$, $U_2 = \{u: \bar{A}u = \bar{b}\}$, каждое из которых выпукло. Аналогично проверяется выпуклость множеств из основной и канонической задач линейного программирования (3.1.14), (3.1.15). Кстати, в § 3.2 при геометрической интерпретации основной задачи в двумерном случае мы по существу уже пользовались выпуклостью многоугольного множества (3.2.1).

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. Напомним

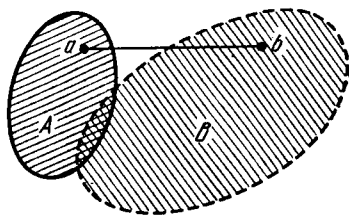


Рис. 4.2.

Определение 4. Суммой множеств $A_1, A_2, \dots, A_m$ называется множество $A = A_1 + A_2 + \dots + A_m = \sum_{i=1}^m A_i$, состоя-

шее из тех и только тех точек a , которые представимы в виде

$$a = \sum_{i=1}^m a_i, \quad a_i \in A_i, \quad i = \overline{1, m}. \quad \text{Разностью двух множеств } A \text{ и } B$$

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 $b = \lambda a$, $a \in A$.

Теорема 1. Если $A_1, A_2, \dots, A_m, A, B$ — выпуклые множества, то множества $C = A_1 + A_2 + \dots + A_m$, $C = A - B$, $C = \lambda A$ выпуклы.

Доказательство проведем, 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 $\alpha a_1 + (1 - \alpha) a_2 \in A$, $\alpha b_1 + (1 - \alpha) b_2 \in B$ при всех $\alpha \in [0, 1]$. Тогда $c_\alpha = \alpha c_1 + (1 - \alpha) c_2 = \alpha(a_1 - b_1) + (1 - \alpha)(a_2 - b_2) = \alpha a_1 - \alpha b_1 + (1 - \alpha)a_2 - (1 - \alpha)b_2 = \alpha a_1 + (1 - \alpha)a_2 - \alpha b_1 - (1 - \alpha)b_2 \in A - B$, так что $c_\alpha \in C$ при всех $\alpha \in [0, 1]$. Выпуклость $C = A - B$ доказана. Аналогично доказывается выпуклость множеств $C = A_1 + A_2 + \dots + A_m$ и $C = \lambda A$.

Определение 5. Замыканием множества U называется множество, являющееся объединением множества U и всех его предельных точек. Замыкание множества U будем обозначать через $\bar{U}$.

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 $\text{int } U$. 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 $E^n \setminus U$ — тогда точка v называется граничной точкой множества U . Совокупность всех граничных точек множества U будем обозначать через $\text{Gr } U$.

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 -окре-

ственность этой точки, не содержащая ни одной точки множества U , отличной от v .

Таким образом, замыкание $\bar{U}$ 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 изолированные точки.

Шар из примера 1 замкнут и его замыкание состоит из внутренних точек $\text{int } S(u_0, R) = \{u: |u - u_0| < R\}$ и граничных точек $\text{Gr } S(u_0, R) = \{u: |u - u_0| = R\}$. Множество $U = \{u = (x, y) \in E^2: 0 \leq x \leq 1, 0 \leq y \leq 1, x + y < 1\}$ выпукло, но не замкнуто — точки прямой $x + y = 1$ при $0 \leq x; y \leq 1$ являются предельными граничными для U , но не принадлежат U ; замыкание $\bar{U} = \{u = (x, y): 0 \leq x; y \leq 1, x + y \leq 1\}$. Множество $U = \{u = (x, y) \in E^2: -1 < x < 1, y = 0\}$ выпукло и не имеет внутренних точек; его замыкание $\bar{U} = \{u = (x, y): -1 \leq x \leq 1, y = 0\}$.

Теорема 2. Если A — выпуклое множество, то его замыкание $\bar{A}$ тоже выпукло.

Доказательство. Пусть a и b — произвольные точки множества $\bar{A}$. Так как выпуклое множество не имеет изолированных точек, то точки a и b будут предельными для A . Тогда существуют последовательности $\{a_k\}$ и $\{b_k\} \in A$, сходящиеся к a и b соответственно. Возьмем произвольное $\alpha \in [0, 1]$. В силу выпуклости A тогда $c_k = \alpha a_k + (1 - \alpha)b_k \in A$. Отсюда при $k \rightarrow \infty$ получим $\lim_{k \rightarrow \infty} c_k = c_\alpha = \alpha a + (1 - \alpha)b$. Таким образом,

точка c_α является предельной для A и, следовательно, принадлежит $\bar{A}$ при любом $\alpha \in [0, 1]$. Теорема доказана.

Теорема 3. Пусть U — выпуклое множество и $\text{int } U \neq \emptyset$. Пусть $u_0 \in \text{int } U$, $v \in \bar{U}$. Тогда $v_\alpha = v + \alpha(u_0 - v) \in \text{int } U$ при всех α , $0 < \alpha \leq 1$.

Доказательство. Так как точка $u_0 \in \text{int } U$, то найдется ее δ -окрестность $O(u_0, \delta) = \{u: |u - u_0| < \delta\}$, целиком принадлежащая U . Сначала рассмотрим случай, когда $v \in U$. Возьмем произвольное α , $0 < \alpha < 1$. Покажем, что окрестность $O(v_\alpha, \alpha\delta) = \{u: |u - v_\alpha| < \alpha\delta\}$ 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 .

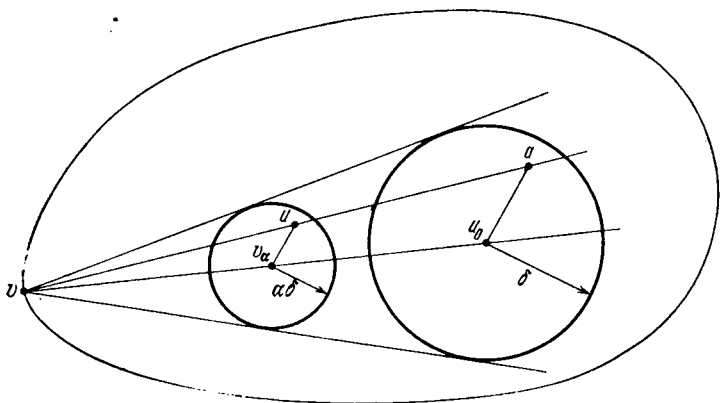


Рис. 4.3.

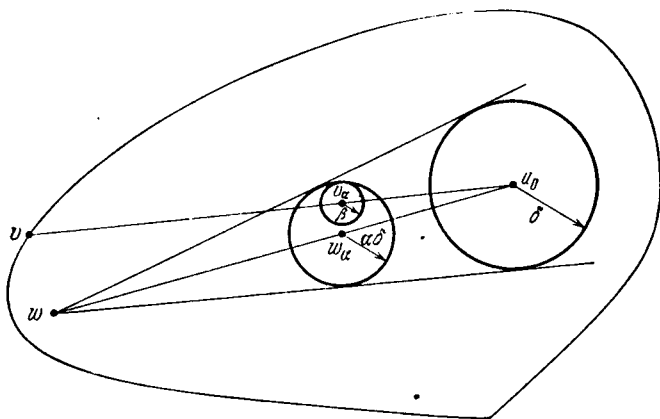


Рис 4.4.

Пусть теперь $v \in \partial \setminus U$. Так как 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 при-

надлежит U . В самом деле, если $u \in O(v_\alpha, \beta)$, то $|u - w_\alpha| \leq |u - v_\alpha| + |v_\alpha - w_\alpha| < \beta + |v_\alpha - w_\alpha| = \delta\alpha$, т. е. $O(v_\alpha, \beta) \in U$. Теорема 3 доказана.

Теорема 4. Если U — выпуклое множество, то $\text{int } U$ тоже выпукло.

Доказательство. Пусть u, v — произвольные точки из $\text{int } U$. В теореме 3 было показано, что $u_\alpha = \alpha u + (1 - \alpha)v \in \text{int } U$ при всех α , $0 < \alpha < 1$. Это и означает выпуклость $\text{int } U$.

3. Выпуклое множество $U = \{u = (x, y, z) \in E^3: x^2 + y^2 \leq 1, z = 0\}$, представляющее собой единичный круг в плоскости $\Gamma = \{u = (x, y, z) \in E^3: z = 0\}$, 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 (то есть не «признавать» точки E^3 , лежащие вне Γ), то U — единичный круг — конечно же, имеет внутренние точки. Приводимая ниже теорема 5 показывает, что это не случайно. Для формулировки теоремы 5 нам понадобится

Определение 6. Точка $v \in U$ называется *относительно внутренней точкой* множества U , если существует ε -окрестность $O(v, \varepsilon) = \{u: u \in E^n, |u - v| < \varepsilon\}$ точки v , такая, что пересечение $O(v, \varepsilon) \cap \text{aff } U$ целиком принадлежит U . Множество всех относительно внутренних точек множества U обозначается через $\text{ri } U$ (иногда обозначают $\text{relint } U$).

Например, если $U = \{u = (x, y, z) \in E^3: x^2 + y^2 \leq 1, z = 0\}$, то $\text{ri } U = \{u = (x, y, z) \in E^3: x^2 + y^2 < 1, z = 0\}$.

Если множество $U \in E^n$ имеет размерность n , т. е. $\dim \text{aff } U = n$, то понятия внутренней и относительно внутренней точки для множества U совпадают и $\text{ri } U = \text{int } U$. Нетрудно указать множества U (например, множество, состоящее из двух различных точек E^n), у которых $\text{ri } U = \emptyset$. Однако для выпуклых множеств верна

Теорема 5. Если U — непустое выпуклое множество из E^n , то $\text{ri } U \neq \emptyset$.

Доказательство этой теоремы будет приведено в § 4.5.

Заметим, что теоремы 3, 4 допускают обобщение на случай относительно внутренних точек. А именно, верна

Теорема 6. Если U — выпуклое множество, точка $u_0 \in \text{ri } U$, $v \in U$, то $v_\alpha = v + \alpha(u_0 - v) \in \text{ri } U$ при всех α , $0 < \alpha \leq 1$, и множество $\text{ri } U$ выпукло.

Для доказательства этой теоремы достаточно повторить доказательства теорем 3, 4, проводя все построения в $\text{aff } U$.

4. 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 $u_1, u_2, \dots, u_m$, если существуют числа $\alpha_1 \geq 0, \alpha_2 \geq 0, \dots, \alpha_m \geq 0, \alpha_1 + \alpha_2 + \dots + \alpha_m = 1$, такие, что $u = \alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_m u_m$.

Теорема 7. 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 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 $\alpha_1 u_1 + \dots + \alpha_m u_m$ произвольных m точек из U . Можем считать, что $\alpha_i > 0, i = \overline{1, m}$. Так как $\alpha_1 + \alpha_2 + \dots + \alpha_m = 1$, то $0 < \alpha_i <$

$< 1, i = \overline{1, m}$. Следовательно, точка $v = \sum_{i=1}^{m-1} \alpha_i (1 - \alpha_m)^{-1} u_i$ является выпуклой комбинацией точек $u_1, \dots, u_{m-1}$ и, по предположению индукции, принадлежит U . Однако тогда $u = (1 - \alpha_m) \sum_{i=1}^{m-1} \alpha_i (1 - \alpha_m)^{-1} u_i + \alpha_m u_m =$

$= (1 - \alpha_m) v + \alpha_m u_m \in U$ 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. Теорема доказана.

Определение 8. Пересечение всех выпуклых множеств, содержащих множество U , называется *выпуклой оболочкой* множества U и обозначается через $co U$.

Ясно, что $co U$, как пересечение выпуклых множеств, является выпуклым множеством. Кроме того, $co U$ содержится в любом выпуклом множестве, содержащем U . Так что $co U$ — это минимальное выпуклое множество, содержащее U .

Теорема 8. 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 $co U = W$. Так как $U \subseteq co U$ и $co U$ — выпуклое множество, то по теореме 7 $co U$ содержит все выпуклые комбинации точек из $co U$ и, в частности, точек из U . Следовательно, $W \subseteq co U$.

Покажем, что W — выпуклое множество. В самом деле, пусть $u, v \in W$, т. е. $u = \alpha_1 u_1 + \dots + \alpha_m u_m, u_i \in U, \alpha_i \geq 0, i = \overline{1, m}, \alpha_1 + \dots + \alpha_m = 1, v = \beta_1 v_1 + \dots + \beta_p v_p, v_i \in U, \beta_i \geq 0, i = \overline{1, p}, \beta_1 + \dots + \beta_p = 1$. Тогда

$$u_\alpha = \alpha u + (1 - \alpha) v = \sum_{i=1}^m \alpha \alpha_i u_i + \sum_{j=1}^p (1 - \alpha) \beta_j v_j, \text{ где } \alpha \alpha_i \geq 0, (1 - \alpha) \beta_j \geq$$

$$\geq 0, \sum_{i=1}^m \alpha \alpha_i + \sum_{j=1}^p (1 - \alpha) \beta_j = 1 \text{ для каждого } \alpha \in [0, 1]. \text{ Следовательно,}$$

u_α является выпуклой комбинацией точек $u_1, \dots, u_m, v_1, \dots, v_p \in U$ и принадлежит W при всех $\alpha \in [0, 1]$. Таким образом, W — выпуклое множество, содержащее U . Но $co U$ по своему определению принадлежит всем выпуклым множествам, содержащим U , и поэтому $co U \subseteq W$. Сравнивая с ранее доказанным включением $W \subseteq co U$, заключаем, что $co U = W$, что и требовалось.

Заметим, что выпуклая оболочка двух точек на плоскости представляет собой отрезок, выпуклая оболочка трех точек, не лежащих на одной пря-

мой, — треугольник. В общем случае выпуклая оболочка конечного числа точек на плоскости образует выпуклый многоугольник, а в пространстве — выпуклый многогранник.

Определение 9. Выпуклая оболочка множества точек $u_0, u_1, \dots, u_m$ из E^n , таких, что система векторов $\{u_i - u_0, i = \overline{1, m}\}$ линейно независима, называется m -мерным симплексом, натянутым на эти точки, и обозначается через $S_m = S_m(u_0, u_1, \dots, u_m)$. Точки $u_0, u_1, \dots, u_m$ называются вершинами симплекса.

В случае $m = 0, 1, 2, 3$ симплекс представляет собой соответственно точку, отрезок, треугольник, тетраэдр.

Согласно теореме 8 симплекс S_m представим в виде

$$S_m = \left\{ u: u = \sum_{i=0}^m \alpha_i u_i, \alpha_i \geq 0, i = \overline{0, m}, \sum_{i=0}^m \alpha_i = 1 \right\}.$$

Предлагаем читателю самостоятельно убедиться в том, что размерность симплекса S_m равна m и в смысле определения 3.

Согласно теореме 8 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 со U , оказывается, достаточно ограничиться рассмотрением выпуклых комбинаций не более чем $n+1$ точек из U . Точнее, верна

Теорема 9. Пусть U — произвольное непустое множество из E^n . Тогда любая точка $u \in \text{co}U$ представима в виде выпуклой комбинации не более чем $n+1$ точек из U .

Доказательство. Согласно теореме 8 любая точка $u \in \text{co}U$ представляема в виде $u = \alpha_1 u_1 + \dots + \alpha_m u_m$, где $u_i \in U, \alpha_i \geq 0, i = \overline{1, m}, \alpha_1 + \dots + \alpha_m = 1$. 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 $\bar{u}_i = (u_i, 1), i = \overline{1, m}$. Так как $m > n+1$, то эти векторы линейно зависимы, т. е. существуют числа $\gamma_1, \dots, \gamma_m$, не все равные нулю и такие, что $\gamma_1 \bar{u}_1 + \gamma_2 \bar{u}_2 + \dots + \gamma_m \bar{u}_m = 0$. Это равенство эквивалентно следующим двум равенствам:

$$\gamma_1 u_1 + \gamma_2 u_2 + \dots + \gamma_m u_m = 0, \quad \gamma_1 + \gamma_2 + \dots + \gamma_m = 0.$$

Тогда точка u представима другими выпуклыми комбинациями тех же точек

$$u_1, u_2, \dots, u_m: \sum_{i=1}^m (\alpha_i - t\gamma_i) u_i = \sum_{i=1}^m \alpha_i u_i - t \sum_{i=1}^m \gamma_i u_i = u. \text{ В самом деле,}$$

здесь $\alpha_i > 0$ и, следовательно, $\alpha_i - t\gamma_i \geq 0, i = \overline{1, m}$, при всех достаточно

малых t , и кроме того, $\sum_{i=1}^m (\alpha_i - t\gamma_i) = \sum_{i=1}^m \alpha_i - t \sum_{i=1}^m \gamma_i = 1$. Поскольку не

все γ_i равны нулю, но $\gamma_1 + \dots + \gamma_m = 0$, то среди $\{\gamma_i\}$ найдутся положительные. Пусть $\alpha_s \gamma_s^{-1} = \min_{\gamma_i > 0} \alpha_i \gamma_i^{-1}$. 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 u удалось представить в виде выпуклой комбинации меньшего числа точек $u_1, \dots, u_{s-1}, u_{s+1}, \dots, u_m$. Ясно, что последовательно применяя описанный прием далее, число точек, участвующих в выпуклой комбинации, можно уменьшить до $n+1$. Теорема доказана.

Теорема 10. Если U — замкнутое ограниченное множество из E^n , то со U замкнуто и ограничено.

Доказательство. По условию существует число $R > 0$ такое, что $|u| \leq R$ 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

со $U \subseteq S(0, R)$, так что $|u| \leq R$ для всех $u \in \text{co } U$. Ограниченность $\text{co } U$ доказана.

Докажем замкнутость $\text{co } U$. Пусть u — предельная точка $\text{co } U$, пусть $\{u_k\} \in \text{co } U$ и $\lim_{k \rightarrow \infty} u_k = u$. Согласно теореме 9 существуют точки $u_{ki} \in U$,

числа $\alpha_{ki} \geq 0$, $i = 1, n+1$, $\alpha_{k1} + \dots + \alpha_{k, n+1} = 1$, такие, что $u_k = \alpha_{k1}u_{k1} + \dots + \alpha_{k, n+1}u_{k, n+1}$. Заметим, что $|u_{ki}| \leq R$, $0 \leq \alpha_{ki} \leq 1$ при всех $i = 1, n+1$ и всех $k = 1, 2, \dots$. Пользуясь теоремой Больцано — Вейерштрасса, сначала из $\{u_{k1}\}$, $\{\alpha_{k1}\}$ выберем подпоследовательности $\{u_{k_11}\}$, $\{\alpha_{k_11}\}$,

сходящиеся к некоторым u_1 , α_1 соответственно; затем из $\{u_{k_12}\}$, $\{\alpha_{k_12}\}$ — подпоследовательности $\{u_{k_22}\} \rightarrow u_2$, $\{\alpha_{k_22}\} \rightarrow \alpha_2$ и т. д., наконец,

$\{u_{k_{n+1}n+1}\} \rightarrow u_{n+1}$, $\{\alpha_{k_{n+1}n+1}\} \rightarrow \alpha_{n+1}$. Тогда из $u_{k_{n+1}} = \sum_{i=1}^{n+1} \alpha_{k_{n+1}i} u_{k_{n+1}i}$,

$u_{k_{n+1}i} \in U$, $\alpha_{k_{n+1}i} \geq 0$, $i = 1, n+1$, $\sum_{i=1}^{n+1} \alpha_{k_{n+1}i} = 1$ предельным переходом

при $k_{n+1} \rightarrow \infty$ получим $u = \sum_{i=1}^{n+1} \alpha_i u_i$, $\alpha_i \geq 0$, $\sum_{i=1}^{n+1} \alpha_i = 1$, где $u_i \in U$, $i =$

$1, n+1$, в силу замкнутости U . Следовательно, по теореме 8 $u \in \text{co } U$, что и требовалось.

Заметим, что требование ограниченности множества U в теореме 10 существенно. Например, множество $U = \{u = (x, y) \in E^2: x \geq 0, y = \sqrt{x}\}$ замкнуто, но $\text{co } U = \{u = (x, y): x \geq 0, 0 < y \leq \sqrt{x}\}$ незамкнуто.

Некоторые другие свойства выпуклых множеств будут рассмотрены в §§ 4.4, 4.5.

Упражнения. 1. Доказать, что множество U будет аффинным тогда и только тогда, если из $u, v \in U$ следует, что $u_\alpha = \alpha u + (1 - \alpha)v \in U$ для всех действительных чисел α .

2. Доказать, что аффинная оболочка множества U состоит из точек вида $\alpha_1 u_1 + \dots + \alpha_m u_m$ при всевозможных $u_1, u_2, \dots, u_m \in U$, $\alpha_1 + \dots + \alpha_m = 1$, α_i — действительные числа, $i = 1, m$, и только из них.

3. Доказать, что если U — выпуклое множество, то $\text{int } U = \text{int } \bar{U}$, т. е. выпуклое множество и его замыкание обладают одним и тем же запасом внутренних точек. Показать, что для невыпуклых множеств это утверждение, вообще говоря, неверно (рассмотреть круг с выколотым центром).

4. Доказать, что если $A \subseteq B$, то $\bar{A} \subseteq \bar{B}$, $\text{int } A \subseteq \text{int } B$, но, вообще говоря, не будучи включения $\text{ri } A \subseteq \text{ri } B$ даже для выпуклых A и B (рассмотреть пример: B — куб в E^3 , A — одна из его граней).

5. Если U — выпуклое множество из E^n , то $\text{aff } A = \text{aff } \bar{A} = \text{aff } (\text{ri } A)$. Доказать.

6. Доказать, что размерность любого множества U совпадает с максимальной размерностью симплексов, содержащихся в U .

7. Если A — выпуклое множество из E^n , то $\bar{\text{ri } A} = \bar{A}$, $\text{ri } \bar{A} = \text{ri } A$. Доказать.

8. Если A, B — выпуклые множества из E^n , то $\bar{A} + \bar{B} \subseteq \overline{A + B}$, $\text{ri } A + \text{ri } B = \text{ri } (A + B)$, $\text{ri } (\lambda A) = \lambda \text{ri } A$ для любых действительных чисел λ . Доказать.

9. Если U — открытое множество, то $\text{co } U$ открыто. Доказать.

10. Доказать, что $\text{co } (A + B) = \text{co } A + \text{co } B$.

11. Доказать, что $\text{co } \bar{U} = \overline{\text{co } U}$, где $\text{co } U$ — пересечение всех выпуклых замкнутых множеств, содержащих U .

12. Доказать, что вершины $u_0, u_1, \dots, u_m$ m -мерного симплекса $S_m = S_m(u_0, \dots, u_m)$ являются его угловыми точками (см. определения 9 и 3.2.1).

§ 2. Выпуклые функции

1. В первой главе 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

$$J(\alpha u + (1 - \alpha)v) \leq \alpha J(u) + (1 - \alpha)J(v) \quad (1)$$

при всех $u, v \in U$, всех α , $0 \leq \alpha \leq 1$. Если в (1) 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 V , 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 1.7 было показано, что выпуклое множество U содержит выпуклые комбинации $\sum_{i=1}^m \alpha_i u_i$, $\alpha_i \geq 0$, $i = \overline{1, m}$;

$\sum_{i=1}^m \alpha_i = 1$, любых своих точек $u_1, u_2, \dots, u_m$ при любых $m = 2, 3, \dots$. Пользуясь индукцией по той же схеме, какая была использована при доказательстве теоремы 1.7, нетрудно показать, что для любой выпуклой функции $J(u)$ на выпуклом множестве имеет место неравенство

$$J\left(\sum_{i=1}^m \alpha_i u_i\right) \leq \sum_{i=1}^m \alpha_i J(u_i) \quad (2)$$

для любых $m = 1, 2, \dots$, любых $u_i \in U$, $\alpha_i \geq 0$, $i = \overline{1, m}$, $\sum_{i=1}^m \alpha_i = 1$.

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

$$U_* = \{u: u \in U, J(u) = J_* = \inf_U J(u)\}$$

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 $O(u_*, \varepsilon) = \{u: |u - u_*| < \varepsilon\}$ точки u_* такая, что $J(u_*) \leq J(v)$ 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 $J(u_*) \leq J(u_* + \alpha(u - u_*)) \leq J(u_*) + \alpha(J(u) - J(u_*))$ или $0 \leq \alpha(J(u) - J(u_*))$. Сокращая на $\alpha > 0$, отсюда получим $J(u) \geq J(u_*)$ при любом $u \in U$. Следовательно, $u_* \in U_*$.

Пусть теперь $u, v \in U_*$, т. е. $J(u) = J(v) = J_*$, $u, v \in U$. Тогда

$$J_* \leq J(\alpha u + (1 - \alpha)v) \leq \alpha J(u) + (1 - \alpha)J(v) = J_*, \quad (3)$$

т. е. $J(\alpha u + (1 - \alpha)v) = J_*$ при всех α , $0 \leq \alpha \leq 1$. Следовательно, $\alpha u + (1 - \alpha)v \in U_*$, $0 \leq \alpha \leq 1$. 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. Теорема 1 доказана.

3. 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.

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

$$J(u) \geq J(v) + \langle J'(v), u - v \rangle \quad (4)$$

при всех $u, v \in U$.

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U . Перепишем неравенство (1) в виде

$$J(v + \alpha(u - v)) - J(v) \leq \alpha[J(u) - J(v)], \quad 0 \leq \alpha \leq 1, \quad u, v \in U.$$

Применяя к левой части формулу конечных приращений (2.3.2), имеем

$$\alpha \langle J'(v + \theta \alpha(u - v)), u - v \rangle \leq \alpha[J(u) - J(v)], \quad 0 \leq \theta \leq 1.$$

Деля обе части этого неравенства на $\alpha > 0$ и устремляя $\alpha \rightarrow +0$, 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 , $0 \leq \alpha \leq 1$. Положим $u_\alpha = \alpha u + (1 - \alpha)v$. Из (4) получим

$$J(u) - J(u_\alpha) \geq \langle J'(u_\alpha), u - u_\alpha \rangle,$$

$$J(v) - J(u_\alpha) \geq \langle J'(u_\alpha), v - u_\alpha \rangle.$$

Умножим первое из этих неравенств на α , второе — на $1 - \alpha$ и сложим. Получим $\alpha J(u) + (1 - \alpha)J(v) - J(u_\alpha) \geq \langle J'(u_\alpha), \alpha u - \alpha u_\alpha + (1 - \alpha)v - (1 - \alpha)u_\alpha \rangle = 0$, что равносильно неравенству (1). Теорема 2 доказана.

Неравенство (4) имеет простой геометрический смысл. Как известно [7, 8, 179], 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 1.11.4).

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, функция $J(u) \in C^1(U)$, пусть U_* — множество точек минимума $J(u)$ на U . Тогда в любой точке $u_* \in U_*$ необходимо выполняется неравенство

$$\langle J'(u_*), u - u_* \rangle \geq 0 \quad (5)$$

при всех $u \in U$, а в случае $u_* \in \text{int } U$ 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 $0 \leq J(u_* + \alpha(u - u_*)) -$

$$-J(u_*) = \alpha \langle J'(u_*), u - u_* \rangle + o(\alpha) \text{ или}$$

$$0 \leq \langle J'(u_*), u - u_* \rangle + o(\alpha)/\alpha, \quad 0 < \alpha < 1.$$

Отсюда при $\alpha \rightarrow +0$ получим условие (5). Если $u_* \in \text{int } U$, то для любого $e \in E^n$ найдется $\varepsilon_0 > 0$, такое, что $u = u_* + \varepsilon e \in U$ при всех ε , $|\varepsilon| \leq \varepsilon_0$. Полагая в (5) $u = u_* + \varepsilon e$, получим $\varepsilon \langle J'(u_*), e \rangle \geq 0$ при всех ε , $|\varepsilon| \leq \varepsilon_0$, 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 $J(u) = u^2$, $U = \{u \in E^1: 1 \leq u \leq 2\}$, 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 $U = \{u \in E^1: 0 \leq u \leq 2\}$, 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 $J(u) - J(u_*) \geq \langle J'(u_*), u - u_* \rangle \geq 0$ или $J(u) \geq J(u_*)$ при всех $u \in U$, т. е. $u_* \in U_*$. Теорема 3 доказана.

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 достаточно, чтобы

$$\langle J'(u) - J'(v), u - v \rangle \geq 0 \quad (6)$$

при всех $u, v \in U$.

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

$$J(v) \geq J(u) + \langle J'(u), v - u \rangle.$$

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

$$\begin{aligned} \alpha J(u) + (1 - \alpha) J(v) - J(\alpha u + (1 - \alpha)v) &= \\ &= \alpha [J(u) - J(\alpha u + (1 - \alpha)v)] + \\ &+ (1 - \alpha) [J(v) - J(\alpha u + (1 - \alpha)v)] = \end{aligned}$$

$$\begin{aligned}
&= \alpha \int_0^1 \langle J'(\alpha u + (1-\alpha)v + t(u - \alpha u - (1-\alpha)v)), u - \alpha u - \\
&- (1-\alpha)v \rangle dt + (1-\alpha) \int_0^1 \langle J'(\alpha u + (1-\alpha)v + t(v - \alpha u - \\
&- (1-\alpha)v)), v - \alpha u - (1-\alpha)v \rangle dt = \\
&= \alpha(1-\alpha) \int_0^1 \langle J'(\alpha u + (1-\alpha)v + t(1-\alpha)(u-v)) - \\
&- J'(\alpha u + (1-\alpha)v + t\alpha(v-u)), u-v \rangle dt,
\end{aligned}$$

или

$$\begin{aligned}
&\alpha J(u) + (1-\alpha)J(v) - J(\alpha u + (1-\alpha)v) = \\
&= \alpha(1-\alpha) \int_0^1 \langle J'(z_1) - J'(z_2), z_1 - z_2 \rangle \frac{1}{t} dt, \quad (7)
\end{aligned}$$

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 $\langle J'(z_1) - J'(z_2), z_1 - z_2 \rangle \geq 0$ при всех t , $0 < t \leq 1$. 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 . Теорема доказана.

Заметим, что для функций одной переменной неравенство (6) равносильно неубыванию производной $J'(u)$. Это значит, что доказанная теорема 4 является естественным обобщением теоремы 1.11.8 на случай гладких функций многих переменных.

Следующий критерий выпуклости обобщает теорему 1.11.9.

Теорема 5. Пусть U — выпуклое множество из E^n , $\text{int } U \neq \emptyset$, пусть функция $J(u) \in C^2(U)$. Тогда для выпуклости $J(u)$ на U необходимо и достаточно, чтобы

$$\langle J''(u)\xi, \xi \rangle \geq 0 \quad (8)$$

при всех $\xi = (\xi^1, \dots, \xi^n) \in E^n$ и всех $u \in U$.

Доказательство. Необходимость. Пусть $J(u)$ выпукла на U . Пусть сначала $u \in \text{int } U$. Тогда для любого $\xi \in E^n$ найдется число $\varepsilon_0 > 0$ такое, что $u + \varepsilon\xi \in U$ при всех ε , $|\varepsilon| \leq \varepsilon_0$. Поскольку для выпуклой гладкой функции выполняется неравенство (6), то из него с учетом формулы (2.3.4) имеем

$$\langle J'(u + \varepsilon\xi) - J'(u), \xi \rangle \varepsilon = \langle J''(u + \theta\varepsilon\xi)\xi, \xi \rangle \varepsilon^2 \geq 0,$$

$0 \leq \theta \leq 1$, или $\langle J''(u + \theta\varepsilon\xi)\xi, \xi \rangle \geq 0$ для всех ε , $|\varepsilon| \leq \varepsilon_0$. Отсюда при $\varepsilon \rightarrow 0$ получим условие (8) для всех $u \in \text{int } U$. Если

u — граничная точка U , то существует последовательность внутренних точек $\{u_m\}$ множества U , сходящаяся к u . По доказанному, $\langle J''(u_m)\xi, \xi \rangle \geq 0$ для всех $\xi \in E^n$, $m = 1, 2, \dots$. Пользуясь непрерывностью $J''(u)$, отсюда при $m \rightarrow \infty$ получим неравенство (8) и для граничных точек U .

Достаточность. Пусть $J(u) \in C^2(U)$ и удовлетворяет условию (8). Тогда, пользуясь формулой (2.3.4), из (8) имеем

$$\langle J'(u) - J'(v), u - v \rangle = \langle J''(v + \theta(u - v))(u - v), u - v \rangle \geq 0$$

при всех $u, v \in U$. Таким образом, для функции $J(u)$ выполняется условие (6). Из теоремы 4 следует выпуклость $J(u)$ на U . Теорема 5 доказана.

Замечание 1. При доказательстве достаточности условие $\text{int } U \neq \emptyset$ не использовалось. И наоборот, как показывает следующий пример, условие $\text{int } U \neq \emptyset$ при доказательстве необходимости было существенным.

Пример 1. Пусть $J(u) = x^2 - y^2$, $U = \{u = (x, y) \in E^2: y = 0\}$. Ясно, что $J(u)$ выпукла на U . Но условие $\langle J''(u)\xi, \xi \rangle = 2(\xi^1)^2 - 2(\xi^2)^2 \geq 0$ не выполняется, например, для $\xi = (0, 1)$. Здесь $\text{int } U = \emptyset$.

Однако, если требовать, чтобы условие (8) выполнялось лишь для тех ξ , которые принадлежат подпространству, параллельному $\text{aff } U$ (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 $u \in \text{ri } U$, а для точек $u \in U \setminus \text{ri } U$ воспользоваться последовательностью $\{u_m\} \in \text{ri } U$, сходящейся к u . Доказательство достаточности остается без изменений, так как вектор $\xi = u - v$ при любых $u, v \in U$ принадлежит подпространству, параллельному $\text{aff } U$.

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 . Имеется следующий простой алгебраический критерий неотрицательности квадратичной формы [75]: для того, чтобы

квадратичная форма $\langle A\xi, \xi \rangle = \sum_{i,j=1}^n a_{ij} \xi^i \xi^j \geq 0$ при всех $\xi = (\xi^1, \dots, \xi^n)$, необходимо и достаточно, чтобы все главные миноры матрицы $A = \{a_{ij}\}$ были неотрицательны. Напоминаем, что главными минорами матрицы $A = \{a_{ij}\}$ называются всевоз-

можные определители

$$\Delta_{i_1 i_2 \dots i_k} = \det \begin{pmatrix} a_{i_1 i_1} & a_{i_1 i_2} & \dots & a_{i_1 i_k} \\ a_{i_2 i_1} & a_{i_2 i_2} & \dots & a_{i_2 i_k} \\ \dots & \dots & \dots & \dots \\ a_{i_k i_1} & a_{i_k i_2} & \dots & a_{i_k i_k} \end{pmatrix},$$

где $1 \leq i_1 < i_2 < \dots < i_k \leq n$, $k = \overline{1, n}$. Симметричную матрицу A называют *неотрицательно определенной*, если она является матрицей неотрицательно определенной квадратичной формы, и обозначают $A \geq 0$.

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

$$J''(u) = \begin{pmatrix} 2 & 2a & 0 \\ 2a & 2b & 0 \\ 0 & 0 & 2c \end{pmatrix}.$$

Условие неотрицательности всех главных миноров этой матрицы дает искомые условия на a, b, c :

$$b - a^2 \geq 0, \quad c \geq 0.$$

Пример 3. Пусть

$$J(u) = \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle, \quad u \in E^n, \quad (9)$$

где A — симметричная неотрицательно определенная матрица порядка $n \times n$, $b \in E^n$. В частности, если $A = E$ — единичная матрица, $b = 0$, то $J(u) = \langle u, u \rangle = |u|^2$.

Приращение функции (9) нетрудно записать в виде

$$J(u+h) - J(u) = \langle Au - b, h \rangle + \frac{1}{2} \langle Ah, h \rangle \quad (10)$$

при любых $u, h \in E^n$. Из (10) имеем

$$J'(u) = Au - b, \quad J''(u) = A.$$

По условию $A \geq 0$. Отсюда и из теоремы 3 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 $A \geq 0$

лежит в основе ряда численных методов линейной алгебры [2, 39, 64, 129].

Пример 4. Пусть

$$J(u) = \|Au - b\|^2, \quad u \in E^n, \quad (11)$$

где A — матрица порядка $m \times n$, $b \in E^m$. Покажем, что такая функция выпукла на E^n . Для этого вычислим ее производные.

Пользуясь следующей просто проверяемой формулой

$$\langle Ax, y \rangle = \langle x, A^T y \rangle, \quad x \in E^n, \quad y \in E^m,$$

где A^T — матрица, полученная транспонированием матрицы A , нетрудно представить приращение функции (11) в виде

$$J(u+h) - J(u) = 2\langle A^T(Au - b), h \rangle + \frac{1}{2}\langle 2A^T A h, h \rangle$$

при всех $u, h \in E^n$. Отсюда имеем

$$J'(u) = 2A^T(Au - b), \quad J''(u) = 2A^T A.$$

Но $\langle J''(u)\xi, \xi \rangle = 2\langle A^T A \xi, \xi \rangle = 2\langle A\xi, A\xi \rangle = 2\|A\xi\|^2 \geq 0$ при всех $\xi \in E^n$. В силу теоремы 5 функция (11) выпукла.

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

$$A^T A u = A^T b.$$

6. Посмотрим, как влияют на выпуклость сложение, умножение на число и некоторые другие операции над выпуклыми функциями. Легко доказывается

Теорема 6. Если функции $J_i(u)$, $i = \overline{1, m}$, выпуклы на выпуклом множестве, то функция

$$J(u) = \alpha_1 J_1(u) + \alpha_2 J_2(u) + \dots + \alpha_m J_m(u)$$

выпукла на этом множестве при любых $\alpha_i \geq 0$, $i = \overline{1, m}$.

Теорема 7. Пусть $J_i(u)$, $i \in I$, — произвольное семейство функций, выпуклых на выпуклом множестве U , пусть

$$J(u) = \sup_{i \in I} J_i(u), \quad u \in U.$$

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 $J_i(u_\alpha) - \varepsilon \leq J_i(u_\alpha)$. С учетом выпуклости $J_i(u)$ тогда

имеем

$$J(u_\alpha) \leq J_i(u_\alpha) + \varepsilon \leq \alpha J_i(u) + (1 - \alpha) J_i(v) + \varepsilon \leq \\ \leq \alpha J(u) + (1 - \alpha) J(v) + \varepsilon$$

или $J(u_\alpha) \leq \alpha J(u) + (1 - \alpha) J(v) + \varepsilon$ при любом $\varepsilon > 0$. Отсюда при $\varepsilon \rightarrow +0$ 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 промежутке $a \leq t \leq b$ (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

$$J(\alpha u + (1 - \alpha)v) = \\ = \varphi(g(\alpha u + (1 - \alpha)v)) \leq \varphi(\alpha g(u) + (1 - \alpha)g(v)) \leq \\ \leq \alpha \varphi(g(u)) + (1 - \alpha)\varphi(g(v)) = \alpha J(u) + (1 - \alpha)J(v),$$

что и требовалось.

Иногда удобнее пользоваться другим вариантом этой теоремы: если функция $\varphi(t)$ выпукла и не возрастает на промежутке $a \leq t \leq b$, 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 $p \geq 1$ и $p \leq 0$.

Следствие 2. Если $g(u)$ выпукла на выпуклом множестве U , то функция

$$J(u) = (\max \{0; g(u)\})^p = (g^+(u))^p$$

выпукла на U при всех $p \geq 1$.

Следствие 3. Если $g(u)$ выпукла на выпуклом множестве U , причем $g(u) \leq 0$ при всех $u \in U$, то функции

$$J(u) = -1/g(u), \quad J(u) = (\max \{-\ln(-g(u)); 0\})^p, \quad p \geq 1,$$

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.

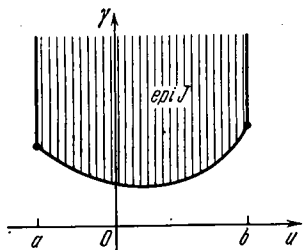


Рис. 4.5. (Надграфик.)

Определение 2. Надграфиком (или эпиграфом) всякой функции $J(u)$, определенной на множестве $U \subseteq E^n$, называется множество (рис. 4.5)

$$\text{epi } J = \{(u, \gamma) \in E^{n+1}: u \in U, \gamma \geq J(u)\}.$$

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.

Доказательство. Необходимость. Пусть функция $J(u)$ выпукла на выпуклом множестве U . Возьмем две произвольные точки $z_1 = (u_1, \gamma_1)$, $z_2 = (u_2, \gamma_2) \in \text{epi } J$, и составим их выпуклую комбинацию $z_\alpha = \alpha z_1 + (1 - \alpha) z_2 = (\alpha u_1 + (1 - \alpha) u_2, \alpha \gamma_1 + (1 - \alpha) \gamma_2)$, $0 \leq \alpha \leq 1$. Из выпуклости U следует, что $u_\alpha = \alpha u_1 + (1 - \alpha) u_2 \in U$. Из выпуклости функции $J(u)$, учитывая, что $z_1, z_2 \in \text{epi } J$, имеем $J(u_\alpha) \leq \alpha J(u_1) + (1 - \alpha) J(u_2) \leq \alpha \gamma_1 + (1 - \alpha) \gamma_2$. Следовательно, $z_\alpha \in \text{epi } J$ при всех $\alpha \in [0, 1]$. Выпуклость $\text{epi } J$ доказана.

Достаточность. Пусть $\text{epi } J$ — выпуклое множество. Возьмем произвольные $u_1, u_2 \in U$ и $\alpha \in [0, 1]$. Тогда $z_1 = (u_1, J(u_1))$, $z_2 = (u_2, J(u_2)) \in \text{epi } J$. В силу выпуклости $\text{epi } J$ точка $z_\alpha = \alpha z_1 + (1 - \alpha) z_2 \in \text{epi } J$. Это значит, что $\alpha J(u_1) + (1 - \alpha) J(u_2) \leq J(\alpha u_1 + (1 - \alpha) u_2)$. Выпуклость $J(u)$ доказана.

Теорема 10. Пусть U — выпуклое множество, а функция $J(u)$ выпукла на U . Тогда множество $M(c) = \{u: u \in U, J(u) \leq c\}$ 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 $J(\alpha u + (1 - \alpha) v) \leq \alpha J(u) + (1 - \alpha) J(v) \leq c$, 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 $M(c) =$

$= \{u: u \in E^1, u^3 \leq c\}$ выпукло при любом c , а функция $J(u) = u^3$ невыпукла на E^1 (см. ниже упражнение 10).

Теорема 11. Пусть U_0 — выпуклое множество, функции $g_i(u)$, $i = \overline{1, m}$, выпуклы на U_0 , а функции $g_i(u) = \langle a_i, u \rangle - b_i$, $i = \overline{m+1, s}$, где a_i — заданные векторы из E^n , b_i — заданные числа, $i = \overline{m+1, s}$. Тогда выпукло множество

$$U = \{u: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}; g_i(u) = 0, i = \overline{m+1, s}\}. \quad (12)$$

Доказательство. В силу теоремы 10 множество $U_i = \{u: u \in U_0, g_i(u) \leq 0\}$ выпукло при всех $i = \overline{1, m}$. Выпукло также множество $M = \{u \in E^n: \langle a_i, u \rangle - b_i = 0, i = \overline{m+1, s}\}$ — см. пример 1.4. Тогда множество (12), являющееся пересечением выпуклых множеств $U_1, \dots, U_m, M$, само будет выпуклым.

8. Рассмотренное в теореме 3 условие оптимальности сформулировано для непрерывно-дифференцируемых функций. Однако аналогичное условие может быть получено 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

$$\frac{dJ(u)}{de} = \lim_{t \rightarrow +0} \frac{J(u + te) - J(u)}{t}. \quad (13)$$

Заметим, что для определения производной по направлению в точке u нужно, чтобы $u + te$ принадлежало области определения $J(u)$ при $0 \leq t \leq t_0$ 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 , $0 \leq t \leq t_0$. Иначе говоря, достаточно малое перемещение из точки u по возможному направлению не выводит за пределы множества U .

Очевидно, если $u \in \text{int } U$, то любое направление $e \neq 0$ является возможным в этой точке. В граничных точках множества возможное направление может и не существовать.

Пример 5. Пусть $U = \{u = (x, y) \in E^2: x \geq 0, x^2 \leq y \leq 2x^2\}$. 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 выпуклого множества U имеется хотя бы одно возможное направление, причем направление $e \neq 0$ будет возможным в точке u тогда и только тогда, если существуют точка $v \in U$, $v \neq u$, и число $\gamma > 0$ такие, что $e = \gamma(v - u)$.

Таким образом, если функция $J(u)$ 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 $0 \leq t \leq t_0$, $t_0 > 0$, и $\frac{dJ(u)}{de} = f'(0) =$ 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

$$\frac{dJ(u)}{de} = \langle J'(u), e \rangle, \quad |e| = 1 \quad (14)$$

(ср. с формулой (2.3.1) при $t \rightarrow +0$). 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

$$J(u) = J(x, y) = \begin{cases} \frac{x^2 y}{x^4 + y^2}, & \text{если } u = (x, y) \neq 0, \\ 0, & \text{если } u = (0, 0) = 0. \end{cases}$$

Возьмем произвольное направление $e = (\cos \alpha, \sin \alpha)$. Тогда

$$\frac{J(0 + te) - J(0)}{t} = \frac{1}{t} J(t \cos \alpha, t \sin \alpha) = \frac{\cos^2 \alpha \sin \alpha}{t^2 \cos^4 \alpha + \sin^2 \alpha}, \quad t > 0.$$

Отсюда имеем

$$\frac{dJ(0)}{de} = \begin{cases} \frac{\cos^2 \alpha}{\sin \alpha} & \text{при } \sin \alpha \neq 0, \\ 0 & \text{при } \sin \alpha = 0. \end{cases}$$

Однако эта функция разрывна в точке $u = 0$. В самом деле, устремим точку $u = (x, y)$ к нулю по параболе $y = x^2$. Тогда $J(x, x^2) = \frac{1}{2} \not\rightarrow J(0, 0) = 0$.

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 J — 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

$$\frac{dJ(u_*)}{de} \geq 0 \quad (15)$$

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 $J(u_* + te) \geq J(u_*)$ или $(J(u_* + te) - J(u_*))/t \geq 0$ при всех достаточно малых $t > 0$. Отсюда при $t \rightarrow +0$ 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 , $0 \leq t \leq t_0 = |u - u_*|$, $t_0 > 0$. Из условия (15) тогда имеем $f'(t) \geq 0$, где $f(t) = J(u_* + te)$. Ниже в теореме 13 будет показано, что $g(t)$ выпукла на $[0, t_0]$. Из неравенства (1.11.5) тогда следует, что $f(t) - f(0) \geq f'(0)t$ или $f(t) \geq f(0)$ при всех $t \in [0, t_0]$. В частности, при $t = t_0 = |u - u_*|$ отсюда имеем $J(u) \geq J(u_*)$, 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 $\frac{dJ(u_*)}{de} = \langle J'(u_*), u - u_* \rangle / |u - u_*|$, 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 выпуклой на отрезке $a \leq t \leq b$, где $a = \inf \{t: u + te \in U\}$, $b = \sup \{t: u + te \in U\}$ (ясно, что $a \leq 0 < b$; если $u + ae \notin U$ или $u + be \notin U$, то функцию $g(t)$ не следует рассматривать при $t = a$ или $t = b$ соответственно).

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 $f(t) = J(u + te)$, $a \leq t \leq b$. Пусть t_1, t_2 — произвольные точки из $[a, b]$ и $\alpha \in [0, 1]$. Тогда $f(\alpha t_1 + (1 - \alpha)t_2) = J(\alpha(u + t_1e) + (1 - \alpha)(u + t_2e)) \leq \alpha J(u + t_1e) + (1 - \alpha)J(u + t_2e) = \alpha f(t_1) + (1 - \alpha)f(t_2)$, 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 $u, v \in U$ и положим $e = v - u$ — это возможное направление в точке u , так как $u + t(v - u) \in U$ при $0 \leq t \leq 1$. Тогда из выпуклости $f(t) = J(u + te)$ получим $J(\alpha v + (1 - \alpha)u) = f(\alpha) = f(\alpha \cdot 1 + (1 - \alpha) \cdot 0) \leq \alpha f(1) + (1 - \alpha)f(0) = \alpha J(v) + (1 - \alpha)J(u)$ при всех $\alpha \in [0, 1]$. Теорема доказана.

Теорема 14. Пусть U — выпуклое множество, функция $J(u)$ выпукла на U . Тогда в любой точке $u \in \text{ri } U$ функция $J(u)$ имеет производные по всем направлениям e , параллельным $\text{aff } U$. В частности, если $\text{int } U \neq \emptyset$, то в точке $u \in \text{int } U$ существуют производные $J(u)$ по всем направлениям $e \in E^n$, $|e| = 1$.

Доказательство. Пусть $\text{aff } U = \{u \in E^n: Au = b\}$, где A — некоторая матрица порядка $m \times n$, $a, b \in E^m$ (см. пример 1.4 и определение 1.2). По определению, направление e параллельно $\text{aff } U$, если e принадлежит подпространству $L_A = \{e \in E^n: Ae = 0\}$. Отсюда ясно, что если $e \in L_A$, то $-e \in L_A$. Зафиксируем какое-либо направление $e \in L_A$, $|e| = 1$, и точку $u \in \text{ri } U$. Согласно определению 1.6 существует ε -окрестность $O(u, e) = \{v \in E^n: |v - u| < \varepsilon\}$ точки u такая, что пересечение $O(u, e) \cap \text{aff } U$ целиком принадлежит U . Учитывая, что $-e$ также принадлежит L_A , можем сказать, что $u + te \in U$ для всех t , $|t| \leq t_0$, $0 < t_0 < \varepsilon$. 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 1.11.2 существует

$$f'(+0) = \lim_{t \rightarrow +0} \frac{(f(t) - f(0))}{t} = \lim_{t \rightarrow +0} \frac{(J(u + te) - J(u))}{t} = \frac{dJ(u)}{de}.$$

Теорема 14 доказана.

Если $u \in U \setminus \text{ri } U$, то в такой точке у выпуклой функции производные по возможным направлениям могут и не существовать — об этом свидетельствует пример 1.11.2.

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 $\text{int } U \neq \emptyset$. 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 $u \in \text{int } U$ 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 $h = (h^1, \dots, h^n)$, $|h| > \delta/n$; здесь $e_i = (0, \dots, 0, 1, 0, \dots, 0)$, $i = \overline{1, n}$, — базис в E^n . Так как 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 , $i = \overline{1, n}$. Поэтому можно взять число δ столь малым, чтобы $|J(u + nh^i e_i) - J(u)| < \varepsilon$ при всех h , $|h| < \delta/n$, $i = \overline{1, n}$. Тогда пользуясь неравенством (2), будем иметь

$$J(u + h) - J(u) =$$

$$= J\left(\frac{1}{n} \sum_{i=1}^n (u + nh^i e_i)\right) - J(u) \leq \frac{1}{n} \sum_{i=1}^n (J(u + nh^i e_i) - J(u)) < \varepsilon \quad (16)$$

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 $J(u) = J((u + h)/2 + (u - h)/2) \leq (J(u + h) + J(u - h))/2$, поэтому $J(u) - J(u + h) \leq J(u - h) - J(u) < \varepsilon$. Отсюда и из (16) следует $|J(u + h) - J(u)| < \varepsilon$ при всех h , $|h| < \delta/n$. Теорема доказана.

Заметим, что если $\text{int } U = \emptyset$, то рассматривая лишь точки из $\text{aff } U$, аналогично можно доказать непрерывность выпуклой на U функции во всех точках $u \in \text{ri } U$. В качестве базиса $\{e_i\}$, участвующего в доказательстве, в этом случае нужно взять базис подпространства, параллельного $\text{aff } U$. В точках $u \in U \setminus \text{ri } U$ выпуклая функция может терпеть разрыв — об этом говорит пример 1.11.1.

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

$$|J'(u) - J'(v)| \leq L |u - v|, \quad u, v \in U, \quad L = \text{const} \geq 0 \quad (17)$$

Для таких функций имеют место неравенства

$$0 \leq \langle J'(u) - J'(v), u - v \rangle \leq L |u - v|^2 \quad (18)$$

при всех $u, v \in U$. 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 [85], полностью характеризующего класс выпуклых функций из $C^{1,1}(U)$ с данной константой $L \geq 0$.

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 константой L , необходимо и достаточно, чтобы

$$|J'(u) - J'(v)|^2 \leq L \langle J'(u) - J'(v), u - v \rangle \quad (19)$$

при всех $u, v \in U$.

Доказательство. Достаточность доказывается просто. А именно, если выполняется неравенство (19), то из него, во-первых, следует, что $\langle J'(u) - J'(v), u - v \rangle \geq 0$, $u, v \in U$, и выпуклость $J(u)$ гарантируется теоремой 4. и, во-вторых, применяя к правой части (19) неравенство Коши — Буняковского и деля на $|J'(u) - J'(v)|$, получим условие (17). Из выпуклости $J(u)$ и условия (17) имеем неравенства (18). Таким образом, из (19) следует (18).

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 $\text{int } U \neq \emptyset$ и $J(u) \in C^2(U)$. Тогда

$$0 \leq \langle J''(u) \xi, \xi \rangle \leq L |\xi|^2, \quad u \in U, \quad (20)$$

при всех $\xi \in E^n$. В самом деле, из неравенств (18) с помощью формулы (2.3.4) в случае $u \in \text{int } U$ имеем

$$0 \leq \langle J'(u + e\xi) - J'(u), e\xi \rangle = e^2 \langle J''(u + \theta e\xi) \xi, \xi \rangle \leq L |\xi|^2 e^2$$

или $0 \leq \langle J''(u + \theta e\xi) \xi, \xi \rangle \leq L |\xi|^2$, $0 \leq \theta \leq 1$, для всех e , $|e| \leq e_0$, $e_0 > 0$. Отсюда при $e \rightarrow +0$ получим (20) для точек $u \in \text{int } U$. Если $u \in \text{Gr } U$, то оценка (20) доказывается с помощью предельного перехода от внутренних точек так же, как это делалось при доказательстве теоремы 5.

Далее, пользуясь формулой (2.3.5), имеем

$$J'(u + h) - J'(u) = Ah, \quad A = \int_0^1 J''(u + th) dt, \quad h = v - u. \quad (21)$$

Разумеется, матрица A зависит от u, v , но эту зависимость мы для краткости не будем явно указывать. Согласно (20) $0 \leq \langle J''(u + th) \xi, \xi \rangle \leq L |\xi|^2$, $0 \leq t \leq 1$, откуда, интегрируя по t , получим

$$0 \leq \langle A \xi, \xi \rangle \leq L |\xi|^2, \quad \xi \in E^n. \quad (22)$$

Таким образом, симметричная матрица A неотрицательно определена. Тогда существует симметричная неотрицательно определенная матрица $A^{1/2}$ такая, что $(A^{1/2})^2 = A$ [63, 128]. Пользуясь оценкой (22) при $\xi = A^{1/2}h$, с помощью формул (21) имеем

$$\begin{aligned} |J'(u) - J'(v)|^2 &= \langle Ah, Ah \rangle = \langle AA^{1/2}h, A^{1/2}h \rangle \leq L |A^{1/2}h|^2 = \\ &= L \langle Ah, h \rangle = L \langle J'(u) - J'(v), u - v \rangle. \end{aligned}$$

Неравенство (19) доказано при дополнительных предположениях $\text{int } U \neq \emptyset$ и $J(u) \in C^2(U)$.

Наметим схему доказательства для случая, когда $\text{int } U \neq \emptyset$, 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 , такие, что $U = \bigcup_{k \geq 1} U_k$, $U_k \subset U_{k+1}$, $k = 1, 2, \dots$, и для всех $k \geq 1$ и всех

$m \geq k$ функция $J_m(u)$ выпукла на U_k , $J_m(u) \in C^2(U_k)$, $|J'_m(u) - J'_m(v)| \leq L |u - v|$ для любых $u, v \in U_k$, $\lim_{m \rightarrow \infty} J_m(u) = J(u)$, $\lim_{m \rightarrow \infty} J'_m(u) = J'(u)$ при всех $u \in U_k$. В силу доказанного тогда $|J'_m(u) - J'_m(v)|^2 \leq L \langle J'_m(u) - J'_m(v), u - v \rangle$, при всех $u, v \in U_k$ и всех $m \geq k$. Отсюда при $m \rightarrow \infty$ получим неравенство (19) на множестве U_k . Далее, при $k \rightarrow \infty$ убеждаемся в справедливости (19) для всех $u, v \in \text{int } U$. 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 $\lim \delta_k = 0$, $\delta_k > \delta_{k+1} > 0$, $k = 1, 2, \dots$. В качестве $J_k(u)$ могут быть взяты средние функции Стеклова — Соболева [179], например,

$$J_k(u) = \int_{E^n} J(w) \omega_k(w - u) dw, \quad \omega_k(u) = \delta_k^{-n} \kappa^{-1} \omega(|u| \delta_k^{-1}),$$

где $\omega(r) = \exp(-1/(1-r^2))$ при $|r| < 1$, $\omega(r) = 0$ при $|r| \geq 1$, $\kappa = \int_{-1}^1 \omega(r) dr$ и функция $J(u)$ вне U доопределена тождественным нулем. Если $\text{int } U = \emptyset$, то аналогичные построения надо провести в $\text{ri } U$.

11. Наконец, остановимся на одном замечательном свойстве выпуклых множеств, задаваемых ограничениями $g(u) \leq c$, 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

$$M(c) = \{u: u \in U_0, g(u) \leq c\}$$

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 очевидна. Докажем достаточность. Пусть множество $M(a) \neq \emptyset$ и ограничено. Так как $M(c) \subset M(a)$ при всех $c \leq a$, то $M(c)$ ограничено при всех $c \leq a$ (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 $t \geq 0$ и всех $u \in M(c)$ (такое направление, задаваемое вектором e , принято называть рецессивным направлением неограниченного выпуклого множества). Так как множество $M(c)$ неограничено по предположению, то существует последовательность $\{u_k\} \in M(c)$ такая, что $|u_k| \rightarrow \infty$ при $k \rightarrow \infty$. Возьмем какую-либо точку $\bar{u} \in M(c)$ и составим вектор $e_k = (u_k - \bar{u}) |u_k - \bar{u}|^{-1}$, $k = 1, 2, \dots$. По теореме Больцано—Вейерштрасса из последовательности $\{e_k\}$ можно выбрать подпоследовательность $\{e_{k_m}\}$, сходящуюся к некоторому вектору e , $|e| = 1$. Возьмем произвольное $t > 0$. Так как $0 < t/|u_k - \bar{u}| < 1$ при всех $k \geq k_0$, то в силу выпуклости $M(c)$ имеем

$$\bar{u} + te_k = \frac{t}{|u_k - \bar{u}|} u_k + \left(1 - \frac{t}{|u_k - \bar{u}|}\right) \bar{u} \in M(c), \quad k \geq k_0.$$

Отсюда при $k \rightarrow \infty$ получим $\bar{u} + te \in M(c)$, поскольку $M(c)$ замкнуто. В силу произвольности $t > 0$ заключаем, что $\bar{u} + te \in M(c)$ при всех $t \geq 0$.

Теперь возьмем любую точку $u \in M(c)$ и покажем, что $u + te \in M(c)$ при каждом $t \geq 0$. По доказанному $\bar{u} + \mu e \in M(c)$, $\mu \geq 0$. В силу выпуклости $M(c)$ тогда

$$\frac{t}{\mu} (\bar{u} + \mu e) + \left(1 - \frac{t}{\mu}\right) u = u + te + t(\bar{u} - u)/\mu \in M(c)$$

при всех $\mu > t$. Отсюда при $\mu \rightarrow \infty$ с учетом замкнутости $M(c)$ получим $u + te \in M(c)$ при каждом $t \geq 0$.

Зафиксируем какую-либо точку $v \in M(a) \subset M(c)$. В силу построения вектора e тогда $v + te \in M(c)$, $t \geq 0$. По условию $M(a)$ ограничено, поэтому луч $\{v + te, t \geq 0\}$ пересекает границу выпуклого множества $M(a)$ в некоторой точке или, точнее говоря, найдется число $t_0 = \sup \{t: v + te \in M(a)\}$ такое, что $v + te \notin M(a)$ при всех $t > t_0$. Это значит, что $v + te \in U_0$, но $c \geq g(v + te) > a \geq g(v)$ для всех $t > t_0$. Зафиксируем какое-либо $t > t_0$. Тогда, пользуясь представлением

$$v + te = \lambda \left(v + \frac{t}{\lambda} e\right) + (1 - \lambda) v, \quad 0 < \lambda < 1$$

и выпуклостью функции $g(u)$, имеем $g(v + te) \leq \lambda g(v + (t/\lambda)e) + (1 - \lambda)g(v)$ или $g\left(v + \frac{t}{\lambda}e\right) \geq (g(v + te) - g(v))/\lambda + g(v)$, $0 < \lambda < 1$.

Так как $g(v + te) > g(v)$, то при $\lambda \rightarrow +0$ отсюда получим $g(v + (t/\lambda)e) \rightarrow +\infty$. 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 $g(v + (t/\lambda)e) \leq c$ при всех λ , $0 < \lambda < 1$. Полученное противоречие доказывает теорему.

Некоторые другие свойства выпуклых функций и множеств будут рассматриваться 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. Если функция $J(u)$ выпукла, то будет ли выпуклой функция $|J(u)|$?

4. 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 $J_1(u)$, $J_2(u)$ потребовать неотрицательность? Или монотонность?

5. Пусть функции $J_k(u)$ выпуклы на выпуклом множестве U при всех $k = 1, 2, \dots$, и пусть существует предел $\lim_{k \rightarrow \infty} J_k(u) = J(u)$ или сходится ряд

$\sum_{k=1}^{\infty} J_k(u) = J(u)$ при всех $u \in U$. Доказать, что тогда $J(u)$ выпукла на U .

6. Пусть U — выпуклое множество, функция $J(u)$ выпукла на U , $J(u) \in C^1(U)$ и $\inf_U J(u) = J_* > -\infty$. Доказать, что для того, чтобы некоторая последовательность $\{u_k\} \in U$ была минимизирующей, т. е. $\lim_{k \rightarrow \infty} J(u_k) = J_*$, необходимо и достаточно выполнения условия $\lim_{k \rightarrow \infty} \langle J'(u_k), u - u_k \rangle \geq 0$ при всех $u \in U$ (ср. с теоремой 3).

7. Выпуклая функция, отличная от постоянной, может достигать своей верхней грани на выпуклом множестве лишь в его граничных точках. Доказать.

8. Для того чтобы функция $\rho(u, U) = \inf_{v \in U} |u - v|$ была выпуклой на E^n , необходимо и достаточно, чтобы замыкание множества U было выпуклым. Доказать.

9. Множество $U_c = \{u \in E^n: g_i(u) \leq c, i = \overline{1, m}\}$, где $g_i(u)$ — выпуклые функции на E^n , $i = \overline{1, m}$, будет ограничено при любых c тогда и только тогда, когда U_c ограничено хотя бы при одном значении $c = c_0$. Доказать.

10. Функция $J(u)$, определенная на выпуклом множестве U , называется квазивыпуклой на U , если $J(\alpha u + (1 - \alpha)v) \leq \max\{J(u); J(v)\}$ при всех $u, v \in U$ и $\alpha \in [0, 1]$. Доказать, что $J(u)$ квазивыпукла на U тогда и только тогда, когда множество $M(v) = \{u: u \in U, J(u) \leq J(v)\}$ выпукло при всех $u \in U$ (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 $J(u) = 1/u$, $U = \{u \in E^1: u \geq 1\}$, то $J_* = \inf J(u) = 0$, 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 $\kappa > 0$ такая, что

$$J(\alpha u + (1 - \alpha)v) \leq \alpha J(u) + (1 - \alpha)J(v) - \alpha(1 - \alpha)\kappa|u - v|^2 \quad (1)$$

при всех $u, v \in U$ и всех α , $0 \leq \alpha \leq 1$. Постоянную κ называют *константой 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 функции на всем пространстве E^n может служить функция

$$J(u) = \langle u, u \rangle = |u|^2, \quad u \in E^n.$$

Для этой функции неравенство (1) превращается в тождественное равенство с константой $\kappa = 1$:

$$|\alpha u + (1 - \alpha)v|^2 = \alpha|u|^2 + (1 - \alpha)|v|^2 - \alpha(1 - \alpha)|u - v|^2$$

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 $u \geq 1$ выпукла, но не будет сильно выпуклой при $u \geq 1$.

Нетрудно видеть, что сумма сильно выпуклой и выпуклой функций на выпуклом множестве U будет сильно выпукла на U с той же константой κ . Если $J(u)$ сильно выпукла на U с константой κ , то $g(u) = cJ(u)$ при любом $c = \text{const} > 0$ будет сильно выпуклой на U с константой $c\kappa$.

Теорема 1. Пусть множество U выпукло и замкнуто, а функция $J(u)$ сильно выпукла и полунепрерывна снизу на U . Тогда

1) множество Лебега

$$M(v) = \{u: u \in U, J(u) \leq J(v)\}$$

выпукло замкнуто и ограничено при всех $v \in U$;

2) $J_* = \inf_U J(u) > -\infty$, множество $U_* = \{u: u \in U; J(u) = J_*\}$ непусто и, более того, состоит из единственной точки u_* ;

3) имеет место неравенство

$$\kappa|u - u_*|^2 \leq J(u) - J(u_*), \quad u \in U; \quad (2)$$

4) любая минимизирующая последовательность $\{u_k\}$: $u_k \in U$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} J(u_k) = J_*$, сходится к точке u_* .

Сформулированная теорема является обобщением теоремы Вейерштрасса — теоремы 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 U ограничено, замкнуто, т. е. U компактно, то все утверждения теоремы 1, кроме неравенства (2), следуют из теорем 2.1.1 и 2.10. Поэтому пусть U — неограниченное множество. Возьмем произвольную точку $v \in U$ и рассмотрим шар

$$S = S(v, 1) = \{u: u \in U, |u - v| \leq 1\}.$$

Из теоремы 2.1.1 следует, что $\inf_S J(u) = J_S^* > -\infty$, так что

$$J(u) \geq J_S^* = J(v) - \nu, \quad \nu = J(v) - J_S^* \geq 0 \quad (3)$$

при всех $u \in S$. Возьмем произвольную точку $u \in U \setminus S$, т. е. $|u - v| > 1$. Тогда

$$0 < \alpha_0 = 1/|u - v| < 1. \quad (4)$$

При $\alpha = \alpha_0$ из (1) получаем

$$\alpha_0 J(u) \geq J(v + \alpha_0(u - v)) - (1 - \alpha_0)J(v) + \alpha_0(1 - \alpha_0)\kappa|u - v|^2. \quad (5)$$

В силу (4) $\alpha_0|u - v| = 1$, поэтому $v + \alpha_0(u - v) \in S$. Согласно (3) тогда $J(v + \alpha_0(u - v)) \geq J(v) - \nu$. Учитывая эту оценку, из (5) имеем

$$\alpha_0 J(u) \geq \alpha_0 J(v) - \nu + \alpha_0(1 - \alpha_0)\kappa|u - v|^2.$$

Отсюда, сокращая на $\alpha_0 > 0$ и вспоминая определение (4) величины α_0 , получим

$$\begin{aligned} J(u) &\geq J(v) + (1 - \alpha_0)\kappa|u - v|^2 - \nu/\alpha_0 = \\ &= J(v) + \kappa|u - v|^2 - (|u - v|\sqrt{\kappa}) \cdot (\sqrt{\kappa} + \nu/\sqrt{\kappa}). \end{aligned}$$

Применяя к последнему слагаемому неравенство $ab \leq (a^2 + b^2)/2$, будем иметь

$$J(u) \geq J(v) + \kappa|u - v|^2/2 - (\sqrt{\kappa} + \nu/\sqrt{\kappa})^2/2 \quad (6)$$

для всех $u \in U \setminus S$. Нетрудно видеть, что неравенство (6) справедливо и при $u \in S$. В самом деле, если $u \in S$, т. е. $|u - v| \leq 1$, то $\nu < (\sqrt{\kappa} + \nu/\sqrt{\kappa})^2/2 - \kappa|u - v|^2/2$. Отсюда и из (3) следует справедливость (6) и для всех $u \in S$.

Таким образом, неравенство (6) имеет место для всех $u \in U$. Для любых $u \in M(v)$ из (6) следует

$$\kappa|u - v|^2/2 - (\sqrt{\kappa} + \nu/\sqrt{\kappa})^2/2 \leq J(u) - J(v) \leq 0$$

или

$$|u - v| \leq 1 + \nu/\kappa, \quad u \in M(v).$$

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 $J_* > -\infty$, $U_* \neq \emptyset$. Так как сильно выпуклая функция строго выпукла, то в силу теоремы 2.1 множество U_* состоит из единственной точки u_* .

Докажем неравенство (2). Учитывая, что $J(u_*) \leq J(u)$ при всех $u \in U$, из (1) имеем $0 \leq J(\alpha u + (1-\alpha)u_*) - J(u_*) \leq \alpha(J(u) - J(u_*)) - \alpha(1-\alpha)\kappa|u - u_*|^2$ или $\alpha(1-\alpha)\kappa|u - u_*|^2 \leq \alpha(J(u) - J(u_*))$ при всех $\alpha \in [0, 1]$ и $u \in U$. Деля на $\alpha > 0$ и устремляя $\alpha \rightarrow +0$, откуда получим неравенство (2).

Наконец, пусть $\{u_k\} \in U$, $\lim_{k \rightarrow \infty} J(u_k) = J_*$. Полагая в (2) $u = u_k$, получим $\kappa|u_k - u_*|^2 \leq J(u_k) - J_*$, $k = 1, 2, \dots$. Отсюда при $k \rightarrow \infty$ следует, что $|u_k - u_*| \rightarrow 0$. Теорема 1 доказана.

З а м е ч а н и е 1. При выполнении условий теоремы 1 утверждение о $J_* > -\infty$, $U_* \neq \emptyset$ остается верным для любого замкнутого (необязательно выпуклого) подмножества $W \subseteq U$ — это следует из замкнутости и ограниченности $M(v) = \{u: u \in W, J(u) \leq J(v)\}$ и теоремы 2.1.2.

2. Укажем критерии сильной выпуклости для гладких функций, аналогичные теоремам 2.2, 2.4, 2.5.

Теорема 2. Пусть U — выпуклое множество, функция $J(u) \in C^1(U)$. Тогда для того чтобы $J(u)$ была сильно выпуклой на U , необходимо и достаточно существования постоянной $\kappa > 0$ такой, что

$$J(u) - J(v) \geq \langle J'(v), u - v \rangle + \kappa|u - v|^2, \quad u, v \in U. \quad (7)$$

Доказательство. Необходимость. Пусть $J(u)$ сильно выпукла на U . Перепишем неравенство (1) в виде

$$J(v + \alpha(u - v)) - J(v) \leq \alpha[J(u) - J(v)] - \alpha(1-\alpha)\kappa|u - v|^2,$$

С помощью формулы (2.3.2) конечных приращений отсюда имеем $\alpha \langle J'(v + \theta\alpha(u - v)), u - v \rangle \leq \alpha[J(u) - J(v)] - \alpha(1-\alpha)\kappa|u - v|^2$, $0 \leq \theta \leq 1$. Деля это неравенство на $\alpha > 0$ и устремляя $\alpha \rightarrow +0$, получим (7).

Д о с т а т о ч н о с т ь. Пусть при некотором $\kappa > 0$ неравенство (7) выполнено при всех $u, v \in U$. В частности, для точки $u_\alpha = \alpha u + (1-\alpha)v$, $0 \leq \alpha \leq 1$, тогда имеем

$$J(u) - J(u_\alpha) \geq \langle J'(u_\alpha), u - u_\alpha \rangle + \kappa|u - u_\alpha|^2,$$

$$J(v) - J(u_\alpha) \geq \langle J'(u_\alpha), v - u_\alpha \rangle + \kappa|v - u_\alpha|^2.$$

Умножим первое из этих неравенств на α , второе — на $1 - \alpha$ и сложим. Получим неравенство (1), что и требовалось.

Теорема 3. Пусть U — выпуклое множество, функция $J(u) \in C^1(U)$. Тогда для сильной выпуклости $J(u)$ на U необхо-

димо и достаточно существования постоянной $\mu > 0$ такой, что

$$\langle J'(u) - J'(v), u - v \rangle \geq \mu |u - v|^2, \quad u, v \in U. \quad (8)$$

Доказательство. Необходимость. Пусть $J(u)$ сильно выпукла на U . Тогда для любых $u, v \in U$ имеет место неравенство (7). Поменяв в (7) переменные u и v ролями, получим

$$J(v) - J(u) \geq \langle J'(u), v - u \rangle + \kappa |u - v|^2.$$

Сложив это неравенство с (7), придем к условию (8) с $\mu = 2\kappa$.

Достаточность. Пусть при некотором $\mu > 0$ выполняется неравенство (8) при всех $u, v \in U$. Тогда, пользуясь формулой (2.7) и обозначениями к ней, получим

$$\begin{aligned} \alpha J(u) + (1 - \alpha) J(v) - J(\alpha u + (1 - \alpha)v) = \\ = \alpha(1 - \alpha) \int_0^1 \langle J'(z_1) - J'(z_2), z_1 - z_2 \rangle dt/t \geq \\ \geq \alpha(1 - \alpha) \int_0^1 \mu |z_1 - z_2|^2 dt/t = \alpha(1 - \alpha) \mu |u - v|^2 \int_0^1 t dt = \\ = \alpha(1 - \alpha) (\mu/2) |u - v|^2, \quad u, v \in U, \quad \alpha \in [0, 1]. \end{aligned}$$

Доказано неравенство (1) с константой $\kappa = \mu/2 > 0$, что и требовалось.

Теорема 4. Пусть U — выпуклое множество, $J(u) \in C^2(U)$. Тогда для сильной выпуклости $J(u)$ на U необходимо и достаточно существования константы $\mu > 0$ такой, что

$$\langle J''(u)\xi, \xi \rangle \geq \mu |\xi|^2 \quad (9)$$

при всех $u \in U$ и всех $\xi = (\xi^1, \xi^2, \dots, \xi^n)$, принадлежащих подпространству, параллельному аффинной оболочке множества U (если $\text{int } U \neq \emptyset$, то (9) выполняется при всех $\xi \in E^n$).

Доказательство. Необходимость. Пусть $J(u)$ сильно выпукла на U . Пусть $\text{aff } U = \{u \in E^n: Au = b\}$, где A — некоторая матрица порядка $m \times n$, а $b \in E^m$ (см. пример 1.4 и определение 1.2). Тогда подпространство L_A , параллельное $\text{aff } U$, имеет вид $L_A = \{\xi \in E^n: A\xi = 0\}$. Возьмем произвольные $\xi \in L_A$, $u \in \text{ri } U$. По определению относительно внутренней точки множества (см. определение 1.6) найдется число $\varepsilon_0 > 0$ такое, что $u + \varepsilon\xi \in U$ при всех ε , $|\varepsilon| \leq \varepsilon_0$. Поскольку для гладкой сильно выпуклой функции справедливо неравенство (8), то из него с учетом формулы (2.3.4) имеем

$$\langle J'(u + \varepsilon\xi) - J'(u), \xi \rangle \varepsilon = \langle J''(u + \theta\varepsilon\xi)\xi, \xi \rangle \varepsilon^2 \geq \mu \varepsilon^2 |\xi|^2,$$

$0 \leq \theta \leq 1$, или $\langle J''(u + \theta \varepsilon \xi) \xi, \xi \rangle \geq \mu |\xi|^2$ для всех ε , $|\varepsilon| \leq \varepsilon_0$. Отсюда при $\varepsilon \rightarrow 0$ получим условие (9) для всех $u \in \text{ri } U$. Если $u \in U \setminus \text{ri } U$, то существует последовательность $\{u_m\} \in \text{ri } U$, сходящаяся к точке u . По доказанному $\langle J''(u_m) \xi, \xi \rangle \geq \mu |\xi|^2$ при всех $\xi \in L_A$. Пользуясь непрерывностью $J''(u)$, отсюда при $m \rightarrow \infty$ получим неравенство (9) и для точек $u \in U \setminus \text{ri } U$.

Достаточность. Пусть $J(u) \in C^2(U)$ и удовлетворяет условию (9). Возьмем произвольные точки $u, v \in U$. Тогда $\xi = u - v \in L_A$. Пользуясь формулой (2.3.4), из (9) при $\xi = u - v$ получим

$$\langle J'(u) - J'(v), u - v \rangle = \langle J''(v + \theta(u - v))(u - v), u - v \rangle \geq \mu |u - v|^2$$

при всех $u, v \in U$. Таким образом, функция $J(u)$ на U удовлетворяет условию (8). Из теоремы 3 следует сильная выпуклость $J(u)$ на U . Теорема 4 доказана (см. замечание 1 к теореме 2.5 и пример 2.1).

Заметим, что для функций одной переменной неравенство (9) имеет вид: $J''(u) \geq \mu > 0$ при всех $u \in U$. Поэтому нетрудно понять, что функция $J(u) = u^p$ при любом $p > 1$ будет сильно выпукла на множестве $U_\varepsilon = \{u \in E^1: u \geq \varepsilon\}$ при всех $\varepsilon > 0$; если здесь $\varepsilon \leq 0$, то такая функция сильно выпукла лишь при $p = 2$. Функция $J(u) = \sin u$ сильно выпукла на $U_\varepsilon = \{-\pi + \varepsilon \leq u \leq -\varepsilon\}$ при всех ε , $0 < \varepsilon < \pi/2$, но не будет сильно выпуклой на $[-\pi, 0]$.

В случае функций многих переменных при $\text{int } U \neq \emptyset$ из условия (9) следует положительная определенность квадратичной формы $\langle J''(u) \xi, \xi \rangle$ на E^n при всех $u \in U$. Как уже упоминалось в § 2.2, для положительной определенности квадратичной

формы $\sum_{i,j=1}^n a_{ij} \xi_i \xi_j = \langle A \xi, \xi \rangle$ необходимо и достаточно, чтобы все угловые миноры

$$\Delta_i = \det \begin{pmatrix} a_{11} & \dots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \dots & a_{ii} \end{pmatrix}, \quad i = \overline{1, n},$$

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 $A^T A$ — невырожденная.

3. Рассмотрим сильно выпуклые функции $J(u)$ из класса $C^{1,1}(U)$, т. е. гладкие сильно выпуклые функции, градиент которых удовлетворяет условию

$$|J'(u) - J'(v)| \leq L|u - v|, \quad u, v \in U. \quad (10)$$

Полезно установить связь между константами κ , μ , L из (1), (7) — (10). Из условия (8) с помощью неравенства Коши — Буняковского и условия (10) имеем $\mu \leq L$. При доказательстве теорем 3, 4 было показано, что $\mu = 2\kappa$. Поэтому $2\kappa \leq L$.

Из определения 1 видно, что если неравенство (1) имеет место при некотором κ , то оно будет иметь место и для всех меньших положительных значений κ . Можно поставить вопрос об определении самой большой, точной константы κ в (1). Очевидно, такой константой в (1) будет

$$\kappa_0 = \inf_{0 < \alpha < 1} \inf_{u, v \in U} \frac{\alpha J(u) + (1 - \alpha) J(v) - J(\alpha u + (1 - \alpha)v)}{\alpha(1 - \alpha)|u - v|^2}.$$

Аналогично, самые большие константы в (8) и (9) соответственно имеют вид

$$\mu_0 = \inf_{u, v \in U} \frac{\langle J'(u) - J'(v), u - v \rangle}{|u - v|^2}, \quad \mu_1 = \inf_{u \in U} \inf_{\xi \in L_A, \xi \neq 0} \frac{\langle J''(u)\xi, \xi \rangle}{|\xi|^2}.$$

Из доказательства теорем 3, 4 следует, что $\mu_0 = \mu_1 = 2\kappa_0$, причем для функций из $C^{1,1}(U)$ все эти константы $\leq L$. Заметим, что для функции $J(u) = |u|^2$ на E^n имеем $\mu_0 = \mu_1 = 2\kappa_0 = L = 2$.

4. Продолжим рассмотрение сильно выпуклых функций $J(u) \in C^{1,1}(U)$. Для таких функций из (8) и (10) имеем неравенства

$$\mu|u - v|^2 \leq \langle J'(u) - J'(v), u - v \rangle \leq L|u - v|^2, \quad u, v \in U. \quad (11)$$

Оказывается, как в теореме 2.16, два неравенства (11) можно записать в виде одного равносильного (11) неравенства [21], полностью характеризующего класс сильно выпуклых функций из $C^{1,1}(U)$ с данными константами L , μ , $L \geq \mu > 0$.

Теорема 5. Пусть U — выпуклое множество из E^n и $J(u) \in C^1(U)$. Тогда для того чтобы $J(u)$ была сильно выпуклой с константой $\kappa = \mu/2 > 0$ и удовлетворяла условию (10) с константой $L > 0$, необходимо и достаточно, чтобы

$$|J'(u) - J'(v)|^2 + L\mu|u - v|^2 \leq (L + \mu)\langle J'(u) - J'(v), u - v \rangle \quad (12)$$

при всех $u, v \in U$.

Доказательство. Необходимость. Пусть $J(u) \in C^{1,1}(U)$ и сильно выпукла на U . Тогда справедливы неравенства (11). 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 (11) следует

$$\langle g'(u) - g'(v), u - v \rangle = \langle J'(u) - J'(v), u - v \rangle - \mu |u - v|^2 \geq 0$$

при всех $u, v \in U$, а из правого неравенства (11) получим

$$\langle g'(u) - g'(v), u - v \rangle \leq (L - \mu) |u - v|^2, \quad u, v \in U.$$

Объединяя оба полученных неравенства, имеем

$$0 \leq \langle g'(u) - g'(v), u - v \rangle \leq (L - \mu) |u - v|^2, \quad u, v \in U.$$

Таким образом, функция $g(u)$ удовлетворяет неравенствам вида (2.18). Согласно теореме 2.16 эти два неравенства равносильны одному неравенству

$$|g'(u) - g'(v)|^2 \leq (L - \mu) \langle g'(u) - g'(v), u - v \rangle, \quad u, v \in U.$$

Подставляя сюда $g'(u) = J'(u) - \mu u$, после несложных тождественных преобразований получим неравенство (12).

Достаточность. Пусть некоторая функция $J(u) \in C^1(U)$ и удовлетворяет неравенству (12). Покажем, что тогда $J(u)$ сильно выпукла с константой $\kappa = \mu/2$ и удовлетворяет условию (10) с константой L , где μ, L взяты из (12). С помощью неравенства Коши — Буняковского из (12) имеем

$$|J'(u) - J'(v)|^2 + L\mu |u - v|^2 \leq (L + \mu) |J'(u) - J'(v)| \cdot |u - v|.$$

Приняв $x = |J'(u) - J'(v)|$, последнее неравенство можно переписать в виде $x^2 - (L + \mu) |u - v|x + L\mu |u - v|^2 \leq 0$. Квадратный трехчлен в левой части этого неравенства имеет корни $x_1 = \mu |u - v|$, $x_2 = L |u - v|$. Поэтому $x_1 \leq x \leq x_2$, т. е.

$$\mu |u - v| \leq |J'(u) - J'(v)| \leq L |u - v|, \quad u, v \in U. \quad (13)$$

Тогда $\langle J'(u) - J'(v), u - v \rangle \leq L |u - v|^2$ — правое неравенство (11) получено. Используя левое неравенство (13), из (12) имеем $\mu^2 |u - v|^2 + L\mu |u - v|^2 \leq (L + \mu) \langle J'(u) - J'(v), u - v \rangle$. Поделив обе части этого неравенства на $L + \mu > 0$, приходим к левому неравенству (11). Таким образом, из (12) получили неравенства (13) и (11). Левое неравенство (11) согласно теореме 3 означает сильную выпуклость $J(u)$ с константой $\kappa = \mu/2$, а правое неравенство (11) (или (13)) дает условие (10). Теорема 5 доказана.

У п р а ж н е н и я. 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 , $0 \leq t \leq 1$, является сильно выпуклой на $[0, 1]$ с одной и той же для всех $u, v \in U$ константой сильной выпуклости.

5. Для того чтобы симметричная матрица A порядка $n \times n$ была положительно определенной, необходимо и достаточно существования постоянных L, μ , $0 < \mu \leq L$, таких, что

$$|Ae|^2 + L\mu |e|^2 \leq (L + \mu) \langle Ae, e \rangle, \quad e \in E^n.$$

Доказать. Убедиться, что в приведенном неравенстве в качестве μ можно взять минимальное собственное число матрицы A , в качестве L — максимальное собственное число.

6. Пусть U — выпуклое множество, $J(u) \in C^1(U)$. Показать, что для того, чтобы функция $J(u)$ была сильно выпуклой и удовлетворяла условию (10), необходимо и достаточно выполнения неравенств (13) при каких-нибудь постоянных $L, \mu, 0 < \mu \leq L$.

§ 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 $P_U(u) = w$.

Так как $\rho(u, U) = \inf_{v \in U} |u - v|$ — расстояние от точки u до множества U , то из определения 1 следует, что

$$\rho(u, U) = |u - P_U(u)| \leq |u - v|, \quad v \in U, \quad u \in E^n.$$

Если $u \in U$, то, очевидно, всегда $P_U(u) = u$. Однако проекция на множество существует не всегда. Например, если $U = \{u \in E^n: |u| < 1\}$ — 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). Однако, как показывает следующая

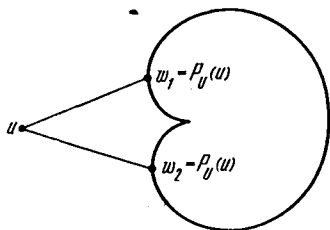


Рис. 4.6.

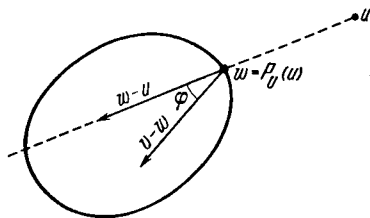


Рис. 4.7.

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 (рис. 4.7)

$$\langle w - u, v - w \rangle \geq 0 \quad \text{при всех } v \in U. \quad (1)$$

При этом если U — аффинное множество (см. пример 1.4), то вместо (1) можно писать

$$\langle w - u, v - w \rangle = 0 \quad \text{при всех } v \in U. \quad (2)$$

Доказательство. Рассмотрим функцию $g(v) = |u - v|^2$ переменной $v \in E^n$ при произвольном фиксированном $u \in U$. Так как $g(v)$ 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 $|v - u|^2 \geq |w - u|^2$ или $|v - u| \geq |w - u|$ при всех $v \in U$, причем равенство здесь возможно только при $v = w$. Остается принять $P_U(u) = w$.

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 $\langle g'(w), v - w \rangle = 2\langle w - u, v - w \rangle \geq 0$ при всех $v \in U$, что равносильно неравенству (1).

Наконец, пусть $U = \{u \in E^n: Au = b\}$ — аффинное множество. Так как это множество выпукло и замкнуто, то неравенство (1) сохраняет силу и здесь. Аффинное множество обладает следующим замечательным свойством: если две точки $v, v_0 \in U$, $v \neq v_0$, то $2v_0 - v$ также принадлежит U , что проверяется непосредственно. Поэтому если здесь взять $v_0 = P_U(u) = w \in U$, то $2w - v \in U$ при любом выборе $v \in U$. Подставим в (1) вместо v точку $2w - v$. Получим $\langle w - u, 2w - v - w \rangle = \langle w - u, w - v \rangle \geq 0$ при всех $v \in U$. Сравнивая полученное неравенство с (1), приходим к равенству (2). Теорема 1 доказана.

Покажем, что оператор проектирования на выпуклое множество обладает сжимающим свойством.

Теорема 2. Если U — выпуклое замкнутое множество из E^n , то

$$|P_U(u) - P_U(v)| \leq |u - v| \quad \text{при всех } u, v \in E^n. \quad (3)$$

Доказательство. Из неравенства (1) имеем

$$\langle P_U(u) - u, P_U(v) - P_U(u) \rangle \geq 0.$$

Поменяв ролями точки u и v в последнем неравенстве, получим

$$\langle P_U(v) - v, P_U(u) - P_U(v) \rangle \geq 0.$$

Сложим эти два неравенства. Будем иметь

$$\langle P_U(u) - u - P_U(v) + v, P_U(v) - P_U(u) \rangle \geq 0.$$

Отсюда следует

$$|P_U(u) - P_U(v)|^2 \leq \langle P_U(u) - P_U(v), u - v \rangle \leq \|P_U(u) - P_U(v)\| \|u - v\|$$

при всех $u, v \in U$. Разделив на $|P_U(u) - P_U(v)| \neq 0$, получим требуемое неравенство (3). Если $|P_U(u) - P_U(v)| = 0$, то (3) очевидно.

2. Приведем примеры множеств, проекция на которые может быть выписана явно.

Пример 1. Пусть $U = S(u_0, R) = \{u \in E^n: |u - u_0| \leq R\}$ — 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

$$\langle w - u, v - w \rangle = (R/|u - u_0| - 1) \langle u - u_0, v - u_0 \rangle - R|u - u_0| \geq 0,$$

так как $|u - u_0| > R$, а $\langle u - u_0, v - u_0 \rangle \leq |u - u_0| \cdot |v - u_0| \leq |u - u_0|R$ в силу неравенства Коши — Буняковского для всех $v \in U$.

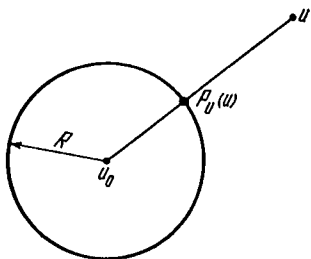


Рис. 4.8.

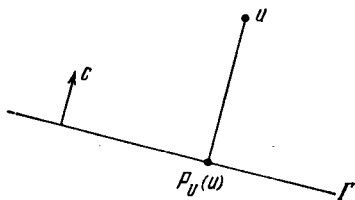


Рис. 4.9.

Пример 2. Пусть $U = \Gamma = \{u \in E^n: \langle c, u \rangle = \gamma\}$ — гиперплоскость; здесь $c \in E^n$, $c \neq 0$, $\gamma = \text{const}$. Пользуясь геометрическими соображениями (рис. 4.9), проекцию точки $u \notin U$ на U будем искать в виде $w = u + \alpha c$. Определяя число α из условия $w \in U$, имеем

$$w = u + (\gamma - \langle c, u \rangle) c / |c|^2.$$

Так как $\langle w - u, v - w \rangle = (\gamma - \langle c, u \rangle) / |c|^2 \cdot \langle c, v - w \rangle = 0$ при всех $v \in U$, то согласно теореме 1 найденная точка w представляет собой проекцию точки u на U .

Пример 3. Пусть $U = \{u \in E^n: \langle a_i, u \rangle = b^i, i = \overline{1, m}\}$ — аффинное множество; здесь $a_i \in E^n$, $b^i = \text{const}$, $i = \overline{1, m}$. Можем считать, что векторы $a_1, \dots, a_m$ 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

$$w = u + \sum_{j=1}^m \alpha_j a_j. \quad (4)$$

Из требования $w \in U$ имеем систему линейных алгебраических уравнений

$$\sum_{j=1}^m \alpha_j \langle a_i, a_j \rangle = b^i - \langle a_i, u \rangle, \quad i = \overline{1, m}, \quad (5)$$

для определения коэффициентов $\alpha_1, \dots, \alpha_m$. Определителем этой системы является определитель Грама [39, 63, 128], который для линейно независимых $a_1, \dots, a_m$ будет отличным от нуля. Поэтому искомые $\alpha_1, \dots, \alpha_m$ 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 , $i = \overline{1, m}$, то точку (4) можно записать в виде

$$w = u - A^T (A A^T)^{-1} (A u - b).$$

Предлагаем читателю провести проверку того, что такая точка $w \in U$, т. е. $A w = b$, и выполняется условие (1).

Пример 4. Пусть $U = \{u \in E^n: \langle c, u \rangle \leq \gamma\}$ — 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 $\langle w - u, v - w \rangle = (\gamma - \langle c, u \rangle) |c|^{-2} (\langle c, v \rangle - \gamma) \geq 0$ при всех $v \in U$. Следовательно, точка w — искомая проекция.

Пример 5. Пусть $U = \{u = (u^1, \dots, u^n) \in E^n: \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\}$ — n -мерный параллелепипед, где α_i, β_i , $\alpha_i < \beta_i$ — заданные числа, $i = \overline{1, n}$. Пусть $u \notin U$. Положим $w = (w^1, \dots, w^n)$, где

$$w^i = \begin{cases} \alpha_i, & \text{если } u^i < \alpha_i, \\ \beta_i, & \text{если } u^i > \beta_i, \\ u^i, & \text{если } \alpha_i \leq u^i \leq \beta_i, \quad i = \overline{1, n}. \end{cases}$$

Тогда $(w^i - u^i)(v^i - w^i) \geq 0$ для всех v^i , $\alpha_i \leq v^i \leq \beta_i$, $i = \overline{1, n}$. Отсюда, суммируя по i от 1 до n , получим $\langle w - u, v - w \rangle \geq 0$ для всех $v \in U$. Следовательно, построенная точка w является проекцией точки u на множество U .

Пример 6. Пусть $U = E_+^n = \{u = (u^1, \dots, u^n): u^i \geq 0, i = \overline{1, n}\}$ — неотрицательный октант пространства E^n . Легко проверить, что проекцией точки u на U является точка $u^+ = ((u^1)^+, (u^2)^+, \dots, (u^n)^+)$, где $(u^i)^+ = \max\{0; u^i\}$, $i = \overline{1, n}$.

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 $J(u)$ на U . Если $u_* \in U_*$, то необходимо выполняется равенство

$$u_* = P_U(u_* - \alpha J'(u_*)) \text{ при всех } \alpha > 0. \quad (6)$$

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 $\langle u_* - (u_* - \alpha J'(u_*)), v - u_* \rangle \geq 0$, $v \in U$, откуда имеем $\alpha \langle J'(u_*), v - u_* \rangle \geq 0$, $v \in U$. Так как $\alpha > 0$, то отсюда получим неравенство $\langle J'(u_*), v - u_* \rangle \geq 0$ 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

$$Au = P_U(u - \alpha J'(u)), \quad \alpha = \text{const} > 0,$$

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 i для определения точки u_* могут быть использованы свойства сжимающих отображений [8, 10].

Упражнения. 1. Найти проекцию точки $u \in E^n$ на множество

$$U = \{u \in E^n: \langle a_1, u \rangle \leq b^1, \langle a_2, u \rangle \leq b^2\}$$

2. Найти проекцию точки $u \in E^n$ на множество

$$U = \{u = (u^1, \dots, u^n): \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\}$$

(здесь $\alpha_i \leq \beta_i$, причем возможно, что $\alpha_i = \beta_i$, или $\alpha_i = -\infty$, или $\beta_i = +\infty$ при некоторых i, j, k).

3. Выяснить геометрический смысл равенства (2).

4. Будут ли верными неравенство (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 $P_U(u) = v$ для всех $v \in U$.

6. Доказать, что для любого замкнутого множества U имеет место неравенство $\|u - P_U(u) - (v - P_U(v))\| \leq \|u - v\|$ для всех $u, v \in U$ (ср. с леммой 2.1.2).

7. Пусть U — выпуклое замкнутое множество из E^n . Доказать, что тогда

$$\begin{aligned} |v - P_U(u)|^2 &\leq \langle v - u, v - P_U(u) \rangle, \\ |v - P_U(u)|^2 + |u - P_U(u)|^2 &\leq |v - u|^2 \end{aligned}$$

при всех $v \in U$.

8. Пусть $J(u) = |Au - b|^2$, где A — матрица порядка $m \times n$, $b \in E^m$ (см. пример 2.4). Доказать, что

$$U_* = \{u \in E^n: J(u) = \inf_{E^n} J(u) = J_*\} \neq \emptyset.$$

Указание: взять проекцию точки b на множество $U = \{v \in E^m: v = Au, u \in E^n\}$ и показать, что $J(u) = |Au - P_U(b)|^2 + |b - P_U(b)|^2$, $J_* = |b - P_U(b)|^2$; замкнутость U см. в лемме 8.2.

§ 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 множество 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 $\langle c, a \rangle \geq \gamma$ при всех $a \in A$ и $\langle c, b \rangle \leq \gamma$ при всех $b \in B$, или, иначе говоря, выполняются неравенства

$$\sup_{b \in B} \langle c, b \rangle \leq \gamma \leq \inf_{a \in A} \langle c, a \rangle.$$

Если $\sup_{b \in B} \langle c, b \rangle < \inf_{a \in A} \langle c, a \rangle$, 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.

Заметим, что в определение 1 множества A и B входят несколько несимметрично. Однако симметрию здесь нетрудно восстановить: нужно лишь взять вектор $(-c)$ вместо c , константу $(-\gamma)$ 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 $|c| = 1$.

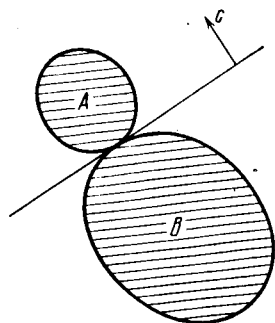


Рис. 4.10.

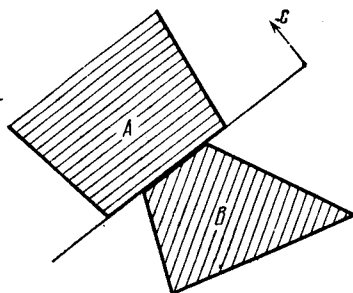


Рис. 4.11.

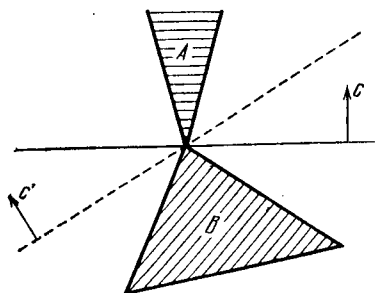


Рис. 4.12.

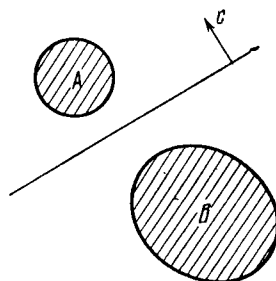


Рис. 4.13.

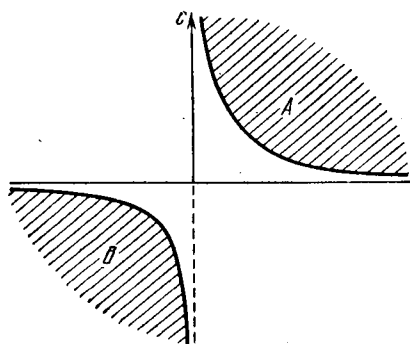


Рис. 4.14.

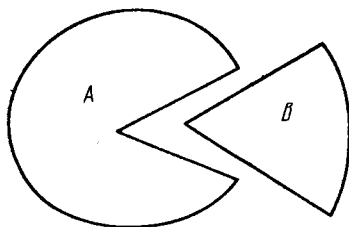


Рис. 4.15.

На рис. 4.10—4.14 изображены случаи, когда два множества отделимы, на рис. 4.13 — сильно отделимы, на рис. 4.14 — строго отделимы. Однако ясно, что не всякие два множества можно отделить гиперплоскостью, — см. рис. 4.15. Ниже приводятся несколько теорем об отделимости выпуклых множеств.

Теорема 1. Пусть X — выпуклое множество из E^n . Тогда для любой точки $y \notin \text{int } X$ существует гиперплоскость $\langle c, u \rangle = \gamma$, отделяющая множество X и точку y . Если точка y не принадлежит $\bar{X}$ — замыканию X , то множество X сильно отделимо от точки y .

Доказательство. Напомним, что если X — выпуклое множество, то выпуклы также $\bar{X}$, $\text{int } X$ (см. теоремы 1.2 и 1.4). Пусть сначала $y \notin \bar{X}$. Согласно теореме 4.1 тогда существует проекция $z = P(y)$ точки y на множество $\bar{X}$, причем $\langle z - y, x - z \rangle \geq 0$ для всех $x \in \bar{X}$. Положим $c = z - y$. С учетом предыдущего неравенства будем иметь $\langle c, x - y \rangle = \langle z - y, x - y \rangle = \langle z - y, x - z \rangle + \langle z - y, z - y \rangle \geq |c|^2 > 0$ или $\langle c, x \rangle \geq \langle c, y \rangle + |c|^2 > \langle c, y \rangle$, $x \in \bar{X}$. Это значит, что гиперплоскость $\langle c, u \rangle = \langle c, y \rangle = \gamma$ сильно отделяет точку y от множества $\bar{X}$ и, тем более, множества X .

Теперь пусть $y \in \text{Gr } X$ — граница X . По определению граничной точки найдется последовательность $\{y_k\} \notin \bar{X}$, сходящаяся к y . По доказанному для каждой точки $y_k \notin \bar{X}$ существует гиперплоскость $\langle c_k, u \rangle = \langle c_k, y_k \rangle$, $|c_k| = 1$, разделяющая множество X и точку y_k , т.е. $\langle c_k, x \rangle > \langle c_k, y_k \rangle$ для всех $x \in X$. По теореме Больцано — Вейерштрасса из последовательности $\{c_k\}$, $|c_k| = 1$, можно выбрать подпоследовательность c_{k_m} , сходящуюся к некоторому c , $|c| = 1$. Так как $\langle c_{k_m}, x \rangle > \langle c_{k_m}, y_{k_m} \rangle$ при

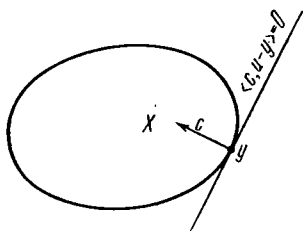


Рис. 4.16.

всех $x \in X$, то при $m \rightarrow \infty$ отсюда получим $\langle c, x \rangle \geq \langle c, y \rangle$ для всех $x \in X$. Искомая гиперплоскость, отделяющая X и y , построена.

Определение 2. Гиперплоскость $\langle c, u \rangle = \gamma$ называют *опорной* ко множеству X , если $\langle c, x \rangle \geq \gamma$ при всех $x \in X$ и $\langle c, y \rangle = \gamma$ для некоторой точки $y \in \bar{X}$. Вектор c , являющийся нормальным вектором опорной к X гиперплоскости, проходящей через точку

$y \in X$, называют *опорным вектором* множества X в точке y (рис. 4.16).

В теореме 1 доказано, что через каждую граничную точку выпуклого множества из E^n можно провести хотя бы одну опорную гиперплоскость.

Теорема 2. Пусть A и B — два выпуклых множества из E^n , пусть A и B не имеют общих точек. Тогда существует гиперплоскость $\langle c, u \rangle = \gamma$, отделяющая множества A и B , а также их замыкания $\bar{A}$ и $\bar{B}$. При этом если $\bar{A}$ и $\bar{B}$ имеют общую граничную точку y , то $\gamma = \langle c, y \rangle$.

Доказательство. Возьмем множество $X = A - B = \{x \in E^n: x = a - b, a \in A, b \in B\}$. Согласно теореме 1.1 множество X выпукло. Так как A и B не имеют общих точек, то $0 \notin X$. В силу теоремы 1 тогда существует гиперплоскость $\langle c, u \rangle = \langle c, 0 \rangle = 0$, отделяющая множество X и точку $y = 0$, т. е. $\langle c, x \rangle \geq 0$ при всех $x \in X$. Это значит, что $\langle c, a - b \rangle \geq 0$ или $\langle c, a \rangle \geq \langle c, b \rangle$ при всех $a \in A$ и $b \in B$. Отсюда $\inf_{a \in A} \langle c, a \rangle \geq \sup_{b \in B} \langle c, b \rangle$, и гиперплоскость $\langle c, u \rangle = \gamma$, где γ — произвольное число, удовлетворяющее неравенству $\inf_{a \in A} \langle c, a \rangle \geq \gamma \geq \sup_{b \in B} \langle c, b \rangle$, отделяет множества A и B .

Нетрудно показать, что та же гиперплоскость $\langle c, u \rangle = \gamma$ отделяет и замыкания $\bar{A}$ и $\bar{B}$ этих множеств. В самом деле, пусть $a \in \bar{A}$, $b \in \bar{B}$. Тогда существуют последовательности $\{a_k\} \in A$, $\{b_k\} \in B$, $a_k \rightarrow a$, $b_k \rightarrow b$ при $k \rightarrow \infty$. По доказанному $\langle c, a_k \rangle \geq \gamma \geq \langle c, b_k \rangle$, $k = 1, 2, \dots$. Отсюда при $k \rightarrow \infty$ получим $\langle c, a \rangle \geq \gamma \geq \langle c, b \rangle$ для всех $a \in \bar{A}$, $b \in \bar{B}$. Если $y \in \bar{A} \cap \bar{B}$, то из последнего неравенства имеем $\gamma = \langle c, y \rangle$. Теорема доказана.

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. Как и в предыдущей теореме, введем множество $X = A - B$. Покажем, что оно замкнуто. Пусть x — некоторая предельная точка множества X , пусть последовательность $\{x_k\} \in X$ и сходится к x . Так как $x_k \in X$, то найдутся $a_k \in A$, $b_k \in B$ такие, что $x_k = a_k - b_k$, $k = 1, 2, \dots$. 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 $a \in A$. Тогда $b_{k_n} = a_{k_n} - x_{k_n} \rightarrow b = a - x$ при $k_n \rightarrow \infty$, 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 $0 \notin X = \bar{X}$. По теореме 1 тогда существует гиперплос-

кость $\langle c, u \rangle = 0$ такая, что $\langle c, x \rangle \geq |c|^2 > 0$ для всех $x \in X$. Отсюда имеем $\langle c, a - b \rangle \geq |c|^2$ или $\langle c, a \rangle \geq \langle c, b \rangle + |c|^2$ для всех $a \in A, b \in B$. Следовательно, $\inf_{a \in A} \langle c, a \rangle \geq \sup_{b \in B} \langle c, b \rangle + |c|^2$.

Любая гиперплоскость $\langle c, u \rangle = \gamma$, где $\sup_{b \in B} \langle c, b \rangle < \gamma < \inf_{a \in A} \langle c, a \rangle$, 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 инструментов исследования свойств выпуклых функций и множеств. Многие приложения этих теорем будут даны в последующих параграфах. Здесь же мы воспользуемся ими для доказательства теоремы о непустоте относительной внутренней выпуклого множества, сформулированной ранее в § 1.

Теорема 1.5. Если U — непустое выпуклое множество из E^n , то $\text{ri } U \neq \emptyset$.

Доказательство. Можем считать, что точка $0 \in U$, так как в противном случае вместо множества U мы рассмотрели бы множество $U - \{u\} = \{w \in E^n: w = u - v, u \in U\}$, где v — какая-либо точка из U . Тогда $0 \in \text{aff } U$, так что $\text{aff } U$ является подпространством в E^n . Пусть множество U имеет размерность $\dim U = m$, $1 \leq m \leq n$ (в случае $m = 0$, когда U состоит из единственной точки, утверждение теоремы принимается по определению). Тогда $\text{aff } U = E^m$. Дальнейшие рассуждения можем проводить в E^m . Поскольку $\dim U = m$, то найдутся точки $u_0, u_1, \dots, u_m \in U$ такие, что $u_1 - u_0, u_2 - u_0, \dots, u_m - u_0$ линейно независимы и, следовательно, образуют базис в E^m . Натянем на эти точки симплекс

$$S_m = \left\{ u \in E^n: u = \sum_{i=0}^m \alpha_i u_i, \alpha_i \geq 0, i = \overline{0, m}; \sum_{i=1}^m \alpha_i = 1 \right\}.$$

Согласно теореме 1.7 $S_m \subseteq U$, а по теореме 1.8 S_m выпукло.

Покажем, что любая точка $w = \sum_{i=0}^m \alpha_i u_i$ при $\alpha_i > 0, i = \overline{0, m}, \sum_{i=0}^m \alpha_i = 1$ является внутренней точкой S_m относительно пространства E^m . Отсюда и из включения $S_m \subseteq U$ будет следовать, что $w \in \text{ri } U$, т. е. $\text{ri } U \neq \emptyset$. Допустим противное: пусть при некоторых $\alpha_i > 0, \sum_{i=0}^m \alpha_i = 1$, точка $w = \sum_{i=0}^m \alpha_i u_i$ оказалась граничной для S_m . По теореме 1 тогда существует вектор $c \in E^m, c \neq 0$, такой, что $\langle c, u \rangle \geq \langle c, w \rangle$ при всех $u \in S_m$. В частности, $\langle c, u_i \rangle \geq \langle c, w \rangle, i = \overline{0, m}$. Кроме того, в силу теоремы 1.7 $w_i = (\alpha_0 u_0 + \dots + \alpha_{i-1} u_{i-1} + \alpha_{i+1} u_{i+1} + \dots + \alpha_m u_m)(\alpha_0 + \dots + \alpha_{i-1} + \alpha_{i+1} + \dots + \alpha_m)^{-1} \in S_m$. Поэтому $\langle c, w \rangle \leq \langle c, w_i \rangle = \langle c, w \rangle - \alpha_i \langle c, u_i \rangle (\alpha_0 + \dots + \alpha_{i-1} + \alpha_{i+1} + \dots + \alpha_m)^{-1}$ или $\alpha_i \langle c, u_i \rangle \leq \alpha_i \langle c, w \rangle$, или $\langle c, u_i \rangle \leq \langle c, w \rangle, i = \overline{0, m}$. Сравнивая с неравенствами $\langle c, w \rangle \leq \langle c, u_i \rangle, i = \overline{0, m}$, заключаем, что $\langle c, u_0 \rangle = \langle c, u_1 \rangle = \dots = \langle c, u_m \rangle = \langle c, w \rangle$. Отсюда $\langle c, u_i - u_0 \rangle = 0, i = \overline{1, m}$. Однако $\{u_i - u_0, i = \overline{1, m}\}$ — базис в E^m , поэтому последние равенства возможны только при $c = 0$. В то же время по построению $c \neq 0$. Полученное противоречие доказывает теорему.

3. Теорему 2 можно истолковать как необходимое условие пустоты пересечения двух выпуклых множеств A и B : если $A \cap B = \emptyset$, A и B выпуклы,

то необходимо существуют вектор $c \in E^n$, $c \neq 0$, и число γ такие, что $\langle c, a \rangle \geq \gamma$ при всех $a \in A$ и $\langle c, b \rangle \leq \gamma$ 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:

$$\langle c_1, u \rangle \geq \gamma_1 \quad \text{при всех } u \in A,$$

$$\langle c_2, u \rangle \geq \gamma_2 \quad \text{при всех } u \in B,$$

$$c_1 + c_2 = 0, \quad \gamma_1 + \gamma_2 = 0,$$

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 [38].

Теорема 4. Пусть непустые множества $A_0, A_1, \dots, A_m$ из E^n выпуклы и $A_0 \cap A_1 \cap \dots \cap A_m = \emptyset$. Тогда необходимо существуют векторы $c_0, c_1, \dots, c_m \in E^n$, не все равные нулю, и числа $\gamma_0, \gamma_1, \dots, \gamma_m$ такие, что

$$\langle c_i, u \rangle \geq \gamma_i \quad \text{для всех } u \in A_i, \quad i = \overline{0, m}, \quad (1)$$

$$c_0 + c_1 + \dots + c_m = 0, \quad (2)$$

$$\gamma_0 + \gamma_1 + \dots + \gamma_m = 0. \quad (3)$$

Для доказательства этой теоремы нам понадобится прямое (декартово) произведение конечного числа множеств, а также прямое произведение евклидовых пространств. Напомним соответствующие определения.

Определение 3. Пусть $A_0, A_1, \dots, A_m$ — какие-либо множества. Множество A , состоящее из всевозможных упорядоченных наборов (точек) $a = (a_0, \dots, a_m)$, где $a_i \in A_i$, $i = \overline{0, m}$, называется прямым произведением множеств $A_0, A_1, \dots, A_m$ и обозначается через $A_0 \times A_1 \times \dots \times A_m = A$.

Пусть $R^{n_0}, \dots, R^{n_m}$ — вещественные линейные пространства. Положим $R = R^{n_0} \times R^{n_1} \times \dots \times R^{n_m}$. Для элементов (точек) $a = (a_0, \dots, a_m)$, $b = (b_0, \dots, b_m) \in R$ определим сумму $a + b = (a_0 + b_0, \dots, a_m + b_m)$ и произведение на вещественное число $\alpha a = (\alpha a_0, \dots, \alpha a_m)$, где под $a_i + b_i$ и αa_i понимаются соответствующие операции в R^{n_i} , $i = \overline{0, m}$. В результате получим вещественное линейное пространство R , называемое прямым произведением линейных пространств $R^{n_0}, \dots, R^{n_m}$. Если $R^{n_i} = E^{n_i}$ — евклидовы пространства размерности n_i , $i = \overline{0, m}$, то в прямом произведении $E = E^{n_0} \times \dots \times E^{n_m}$ также можно ввести скалярное произведение $\langle a, b \rangle = \langle a_0, b_0 \rangle + \langle a_1, b_1 \rangle + \dots + \langle a_m, b_m \rangle$ и норму $|a| = \langle a, a \rangle^{1/2} = (|a_0|^2 + \dots + |a_m|^2)^{1/2}$, где $\langle a_i, b_i \rangle$ и $|a_i|$ — скалярное произведение и соответственно норма в E^{n_i} . Полученное евклидово пространство E называют прямым произведением евклидовых пространств $E^{n_0}, \dots, E^{n_m}$; размерность E равна $n_0 + n_1 + \dots + n_m$. Например, само евклидово пространство E^n является прямым произведением одномерных евклидовых пространств: $E^n = E^1 \times \dots \times E^1$. Параллелепипед $\{u = (u^1, \dots, u^n): \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\}$ представляет собой прямое произведение отрезков $[\alpha_i, \beta_i]$, $i = \overline{1, n}$.

Прямое произведение выпуклых множеств, очевидно, само выпукло.

Доказательство теоремы 4. Пусть $E = E^n \times E^n \times \dots \times E^n$ — прямое произведение $m+1$ экземпляров n -мерных евклидовых пространств. Тогда $A = A_0 \times A_1 \times \dots \times A_m \in E$. Введем в E «диагональное» множество $B = \{b = (b_0, b_1, \dots, b_m): b_0 = b_1 = \dots = b_m, b_i \in E^n, i = \overline{0, m}\}$. Нетрудно

видеть, что пересечение $\bigcap_{i=0}^m A_i$ пусто тогда и только тогда, когда $A \cap B$ пусто. Далее, A и B — выпуклые множества. По теореме 2 множества A и B отделены, т. е. существует $c = (c_0, c_1, \dots, c_m) \in E$, не все c_i равны нулю и $\langle c, a \rangle \geq \langle c, b \rangle$ при всех $a \in A$ и $b \in B$, или

$$\sum_{i=0}^m \langle c_i, a_i \rangle \geq \sum_{i=0}^m \langle c_i, b_i \rangle = \left\langle \sum_{i=0}^m c_i, b \right\rangle, \quad a_i \in A_i, \quad b \in E^n. \quad (4)$$

Отсюда следует ограниченность линейной функции $\left\langle \sum_{i=0}^m c_i, b \right\rangle$ на всем пространстве E^n , что возможно только при

$$\sum_{i=0}^m c_i = c_0 + \dots + c_m = 0.$$

Тогда неравенство (4) принимает вид

$$\sum_{i=0}^m \langle c_i, a_i \rangle \geq 0 \quad \text{при всех } a_i \in A_i, \quad i = \overline{0, m} \quad (5)$$

Если в этом неравенстве зафиксируем какие-либо $a_i = \bar{a}_i \in A_i$ при всех номерах $i = \overline{0, m}$, кроме номера $i = k$, то получим $\langle c_k, a_k \rangle \geq - \sum_{i \neq k} \langle c_i, \bar{a}_i \rangle = \text{const}$ для всех $a_k \in A_k$. Следовательно,

$$\langle c_k, u \rangle \geq \gamma_k = \inf_{a \in A_k} \langle c_k, a \rangle > -\infty \quad \text{для всех } u \in A_k, \quad k = \overline{1, m}. \quad (6)$$

Положим

$$\gamma_0 = -(\gamma_1 + \gamma_2 + \dots + \gamma_m) \quad (7)$$

Тогда, переходя в (5) к нижней грани по всем $a_i \in A_i$, $i = \overline{1, m}$, получим

$$\langle c_0, a_0 \rangle + \sum_{i=1}^m \gamma_i = \langle c_0, a_0 \rangle - \gamma_0 \geq 0 \quad \text{для каждого } a_0 \in A_0 \text{ или}$$

$$\langle c_0, u \rangle \geq \gamma_0 \quad \text{при всех } u \in A_0.$$

Все соотношения (1) — (3) получены. Теорема 4 доказана.

4. Переформулируем теорему 4 для случая, когда $A_0, A_1, \dots, A_m$ 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 K — выпуклое множество, то K называют *выпуклым конусом*, если K замкнуто — *замкнутым конусом*, если K открыто — *открытым конусом*.

Рассмотрим множество

$$K^* = \{c \in E^n: \langle c, u \rangle \geq 0 \text{ для всех } u \in K\}. \quad (8)$$

Это множество всегда непусто, так как $0 \in K^*$. Далее, если $c \in K^*$, то для λc при любом $\lambda > 0$ имеем $\langle \lambda c, u \rangle = \lambda \langle c, u \rangle \geq 0$ для всех $u \in K$, т. е. $\lambda c \in K^*$. Следовательно, K^* — конус.

Определение 5. Конус K^* , определенный посредством (8), называется *двойственным (сопряженным) конусом* к конусу K (рис. 4.17).

Например, если $K = \{u \in E^n: \langle a, u \rangle = 0\}$ — гиперплоскость, то $K^* = \{c \in E^n: c = \lambda a, \lambda = \text{const}\}$; если $K = \{u \in E^n: \langle a, u \rangle \leq 0\}$ — замкнутое полупространство или $K = \{u \in E^n: \langle a, u \rangle < 0\}$ — открытое полупространство, то $K^* = \{c \in E^n: c = -\lambda a, \lambda = \text{const} \geq 0\}$; если $K = E^n$, то $K^* = \{0\}$; если $K = \{0\}$, то $K^* = E^n$, если $K = \{u \in E^n: u \geq 0\}$, то $K^* = \{c \in E^n: c \geq 0\}$.

С помощью двойственных конусов удобно переформулировать теорему 4 для случая, когда множества $A_0, \dots, A_m$ являются конусами.

Теорема 5. Пусть $K_0, K_1, \dots, K_m$ — непустые выпуклые конусы из E^n (с вершиной в нуле), пусть $K_0 \cap K_1 \cap \dots \cap K_m = \emptyset$. Тогда необходимо существуют векторы $c_0, c_1, \dots, c_m$, не все равные нулю, $c_i \in K_i^*, i = 0, m$, и такие, что

$$c_0 + c_1 + \dots + c_m = 0. \quad (9)$$

Доказательство. Согласно теореме 4 существуют векторы $c_0, c_1, \dots, c_m$, не все равные нулю, и числа $\gamma_0, \gamma_1, \dots, \gamma_m$, удовлетворяющие условиям (1)–(3). Воспользуемся тем, что рассматриваемые множества $K_0, \dots, K_m$ являются именно конусами, и покажем, что тогда $\gamma_0 = \gamma_1 = \dots = \gamma_m = 0$. В самом деле, если $\langle c_i, u \rangle \geq \gamma_i$ при всех $u \in K_i$, то $\langle c_i, \lambda u \rangle \geq \gamma_i$ или $\langle c_i, u \rangle \geq \gamma_i / \lambda$ для любых $\lambda > 0$ и $u \in K_i$. Отсюда при $\lambda \rightarrow +\infty$ получим $\langle c_i, u \rangle \geq 0$ при всех $u \in K_i$, т. е. $c_i \in K_i^*, i = 0, m$.

Кроме того, если $u \in K_i$, то, взяв в неравенстве $\langle c_i, u \rangle \geq 0$ 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 $\inf_{u \in K_i} \langle c_i, u \rangle = 0$. Согласно (6) это озна-

чает, что все величины $\gamma_i, i = 1, m$, участвующие в неравенствах (1), равны нулю. Из (7) тогда имеем $\gamma_0 = 0$. Таким образом, если в теореме 4 множества $A_0, \dots, A_m$ являются выпуклыми конусами, то условие (3) выполняется тривиально, так как все $\gamma_i = 0, i = 0, m$, условия (1) означают, что $c_i \in K_i^*, i = 0, m$, а из (2) следует (9). Теорема 5 доказана.

При некоторых дополнительных ограничениях на конусы $K_0, K_1, \dots, K_m$ теорема 5 обратима. А именно, верна

Теорема 6. Пусть $K_0, K_1, \dots, K_m$ — непустые выпуклые конусы из E^n (с вершиной в нуле), пусть все эти конусы, кроме, быть может, одного, открыты. Тогда для того чтобы $K_0 \cap K_1 \cap \dots \cap K_m = \emptyset$, необходимо и достаточно, чтобы существовали векторы $c_0, c_1, \dots, c_m$, не все равные нулю, $c_i \in K_i^*, i = 0, m$, и такие, что

$$c_0 + c_1 + \dots + c_m = 0. \quad (9)$$

Доказательство. Необходимость доказана в теореме 5. Достаточность докажем, рассуждая от противного. Пусть для некоторых $c_i \in K_i^*, i = 0, m$, справедливо равенство (9), не все c_i равны нулю, но пусть тем не менее конусы $K_0, K_1, \dots, K_m$ пересекаются, т. е. существует точка $w \in \bigcap_{i=0}^m K_i$. Из (9) и условия $c_i \in K_i^*, i = 0, m$, имеем $\langle c_0 + c_1 + \dots + c_m, w \rangle =$

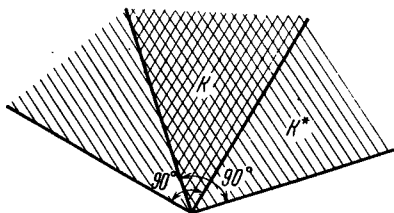


Рис. 4.17.

$= \langle c_0, w \rangle + \dots + \langle c_m, w \rangle = 0$, $\langle c_0, w \rangle \geq 0, \dots, \langle c_m, w \rangle \geq 0$. Отсюда следует $\langle c_i, w \rangle = 0$, $i = 0, \dots, m$. Поскольку не все c_i , $i = 0, \dots, m$, равны нулю, то из (9) вытекает существование по крайней мере двух векторов c_i , c_j , $i \neq j$, отличных от нуля. По условию все конусы $K_0, \dots, K_m$, кроме, быть может, одного, открыты. Поэтому можем считать, что $c_i \neq 0$, $c_i \in K_i^*$, K_i — открытый конус. Тогда точка w принадлежит K_i вместе с некоторой окрестностью $O(w, 2\varepsilon) = \{u \in E^n: |u - w| < 2\varepsilon\}$. В частности, $u = w - \varepsilon c_i / |c_i| \in O(w, 2\varepsilon) \in K_i$. Но $c_i \in K_i^*$, поэтому должно быть $\langle c_i, u \rangle = \langle c_i, w \rangle - \varepsilon |c_i| = -\varepsilon |c_i| \geq 0$, что, однако, невозможно, так как $c_i \neq 0$, $\varepsilon > 0$. Полученное противоречие показывает, что $\bigcap_{i=0}^m K_i = \emptyset$. Теорема 6 доказана.

Теорему 6 называют теоремой Дубовицкого — Милютина [102]. Теоремы 5, 6 играют важную роль в теории экстремальных задач — об этом см. ниже, § 9. Более тонкая теорема, обобщающая теорему 6, доказана в [48].

Упражнения. 1. Пусть A и B — выпуклые множества, не имеющие общих внутренних точек. Можно ли утверждать, что A и B отделимы? Рассмотреть пример:

$$A = \{u = (x, y): y = 0, |x| \leq 1\}, \quad B = \{u = (x, y): x = 0, |y| \leq 1\} \quad \text{в } E^2.$$

2. Пусть X — выпуклое множество, $\text{int } X \neq \emptyset$, $y \in \text{Гр } X$. Показать, что всякая опорная гиперплоскость $\langle c, u \rangle = \langle c, y \rangle$ ко множеству X строго отделяет $\text{int } X$ от y .

3. Пусть X — непустое выпуклое множество, $A = \text{г} A$; пусть M — непустое аффинное множество в E^n , не пересекающееся с A . Доказать, что тогда существует гиперплоскость Γ , содержащая M и строго отделяющая M от A .

4. Пусть A и B — два выпуклых множества из E^n . Доказать, что для существования гиперплоскости Γ , разделяющей A и B , и такой, что хотя бы одно из множеств не содержится целиком в Γ , необходимо и достаточно, чтобы $\text{г} A \cap \text{г} B = \emptyset$.

5. Пусть $\rho(A, B) = \inf_{a \in A} \inf_{b \in B} |a - b|$ — расстояние между двумя множествами A и B . Доказать, что два непустых выпуклых множества $A, B \in E^n$ сильно отделимы тогда и только тогда, когда $\rho(A, B) > 0$.

6. Доказать, что всякое выпуклое замкнутое ограниченное множество A из E^n имеет хотя бы одну угловую точку (см. определение 3.2.1) и является выпуклой оболочкой своих угловых точек.

7. Доказать, что всякое непустое выпуклое замкнутое множество из E^n , отличное от E^n , представимо в виде пересечения открытых (или замкнутых) полупространств.

8. Если $A_1, \dots, A_m$ — выпуклые множества из E^n , причем $\bigcap_{i=1}^m \text{г} A_i \neq \emptyset$, то

$$\overline{\bigcap_{i=1}^m A_i} = \bigcap_{i=1}^m \bar{A}_i; \quad \text{г} \left(\bigcap_{i=1}^m A_i \right) = \bigcap_{i=1}^m (\text{г} A_i); \quad \text{aff} \left(\bigcap_{i=1}^m A_i \right) = \bigcap_{i=1}^m (\text{aff } A_i).$$

Доказать.

9. Пусть K — произвольный конус из E^n . Доказать, что конус K^* замкнут и выпуклый.

10. Доказать, что если K — замкнутый и выпуклый конус, то $(K^*)^* = K$.

11. Доказать, что $(K_1 + \dots + K_m)^* = K_1^* \cap K_2^* \cap \dots \cap K_m^*$, где $K_1, \dots, K_m$ — конусы из E^n .

12. Пусть $K_1, \dots, K_m$ — выпуклые конусы, пусть их пересечение имеет внутреннюю точку. Доказать, что тогда $\left(\bigcap_{i=1}^m K_i\right)^* = K_1^* + K_2^* + \dots + K_m^*$.

13. Множество $A^* = \{c \in E^n: \langle c, u \rangle \leq 1 \text{ для всех } u \in A\}$ называется полярной множества A . Найти полярные множеств A , если $A = \{0\}$; $A = [a, b] \subset \mathbb{C} E^1$; $A = \{u \in E^n: u = te, 0 < t < \infty, e \neq 0\}$; $A = \{u \in E^n: \langle c, u \rangle \leq \gamma\}$; A — шар; A — конус с вершиной в нуле. Выяснить связь между полярной конуса и двойственным конусом.

§ 6. Субградиент. Субдифференциал

1. Для выпуклых дифференцируемых функций на выпуклом множестве выше было доказано неравенство (см. теорему 2.2)

$$J(u) \geq J(v) + \langle J'(v), u - v \rangle, \quad u \in U. \quad (1)$$

К сожалению, выпуклая функция может не быть дифференцируемой даже во 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

$$J(u) \geq J(v) + \langle c(v), u - v \rangle \text{ при всех } u \in U. \quad (2)$$

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).

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 $v \in \text{int } U$, $c(v) \in \partial J(v)$. Так как $J(u) \in C^1(U)$, то $J(u) = J(v) + \langle J'(v), u - v \rangle + o(|u - v|)$, $u \in U$. Отсюда и из (2) следует, что $\langle J'(v) - c(v), u - v \rangle \geq o(|u - v|)$, $u \in U$. Поскольку $v \in \text{int } U$, то $u = v - \varepsilon(J'(v) - c(v)) \in U$ при всех ε , $0 < \varepsilon < \varepsilon_0$. Подставив эту точку в предыдущее неравенство, получим $-\varepsilon |J'(v) - c(v)|^2 \geq o(\varepsilon)$, $0 < \varepsilon < \varepsilon_0$. Деля на $\varepsilon > 0$ и устремляя $\varepsilon \rightarrow +0$, отсюда будем иметь $-|J'(v) - c(v)|^2 \geq 0$, т. е. $c(v) = J'(v)$.

Тем самым показано, что для гладкой выпуклой функции $\partial J(v) = \{J'(v)\}$ при всех $v \in \text{int } U$. Существуют функции, которые недифференцируемы в точке, но тем не менее субдифференциал в этой точке неуст.

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 $|g(u)| - |g(v)| = |g(u)| \geq 0 =$

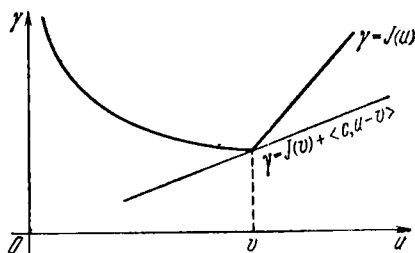


Рис. 4.18.

$= \langle 0, u - v \rangle$ 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 неравенство

$$J(u) - J(0) = |u| \geq \langle c, u - 0 \rangle = \langle c, u \rangle, \quad u \in E^n,$$

для всех c , $|c| \leq 1$. Это значит, что $\partial(|u|)|_{u=0} = \partial J(0) = \{c \in E^n: |c| \leq 1\}$ — 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 $U = \{u \in E^1: u \geq 0\}$ 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 из E^n (например, возможно, $U = E^n$). 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 $\partial J(u) \neq \emptyset$ 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

$$J(u) - J(u_\alpha) \geq \langle c, u - u_\alpha \rangle, \quad J(v) - J(u_\alpha) \geq \langle c, v - u_\alpha \rangle.$$

Умножим первое из этих неравенств на α , второе — на $1 - \alpha$ и сложим. Получим $\alpha J(u) + (1 - \alpha)J(v) - J(u_\alpha) \geq \langle c, u_\alpha - u_\alpha \rangle = 0$ 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. Пусть v — произвольная точка из U . Покажем, что $\partial J(v) \neq \emptyset$. Возьмем некоторый единичный вектор e . Так как U — открытое множество, то $v + te \in U$ при всех t , $0 \leq t < t_0$, $t_0 > 0$. По теореме 2.14 существует производная $\frac{dJ(v)}{de}$ по направлению e .

В пространстве E^{n+1} переменных (u, γ) введем два множества

$$A = \{(u, \gamma) \in E^{n+1}: u \in U, \gamma > J(u)\},$$

$$B = \{(u, \gamma) \in E^{n+1}: u = v + te, \gamma = J(v) + t \frac{dJ(v)}{de}, 0 \leq t < t_0\}.$$

Нетрудно показать, что множество A выпукло — это делается так же, как доказывалась выпуклость надграфика выпуклой функции в теореме 2.9 (кстати, в данном случае $A = \text{int}(\text{epi } J)$). 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 , $0 \leq t < t_0$, — тогда заведомо $(u, \gamma) \notin B$; 2) при некотором t , $0 \leq t < t_0$, оказалось, что $u = v + te$. Тогда с учетом неравенства (1.11.5) и $\gamma > J(u) = J(v + te)$ имеем $\gamma - J(v) > J(v + te) - J(v) \geq t \frac{dJ(v)}{de}$, т. е. $\gamma > J(v) + t \frac{dJ(v)}{de}$, и снова $(u, \gamma) \notin B$.

Итак, A и B выпуклы, $A \cap B = \emptyset$. По теореме 5.2 тогда существует гиперплоскость с нормальным вектором $(d, v) \neq 0$, отделяющая A и B , т. е.

$$\langle d, u \rangle + v\gamma \geq \langle d, v + te \rangle + v \left(J(v) + t \frac{dJ(v)}{de} \right) \quad (3)$$

при всех $\gamma > J(u)$, $u \in U$, $0 \leq t < t_0$. В частности, при $u = v$, $t = 0$ из (3) имеем $v(\gamma - J(v)) \geq 0$ для всех $\gamma > J(v)$. Отсюда следует, что $v \geq 0$.

Допустим, что $v = 0$. Тогда из (3) имеем $\langle d, u \rangle \geq \langle d, v + te \rangle$ для всех $u \in U$, $0 \leq t < t_0$. Положим здесь $u = v + ed$, $t = 0$ — так можно делать, ибо $v + ed \in U$ при всех e , $0 < |e| < \varepsilon_0$, в силу открытости U . Получим $\langle d, v + ed \rangle \geq \langle d, v \rangle$ или $e|d|^2 \geq 0$ при всех e , $0 < |e| \leq \varepsilon_0$, что возможно только при $d = 0$. Однако $\langle d, v \rangle \neq 0$ по построению. Полученное противоречие показывает, что $v = 0$ не может быть.

Итак, $v > 0$. Поделим (3) на $v > 0$. Обозначая $c = -d/v$ и устремляя $\gamma \rightarrow J(u) + 0$, из (3) тогда получим

$$J(u) - J(v) + t \langle c, e \rangle \geq \langle c, u - v \rangle + t \frac{dJ(v)}{de} \quad (4)$$

при всех $u \in U$ и всех t , $0 \leq t < t_0$. Полагая здесь $t = 0$, будем иметь

$$J(u) - J(v) \geq \langle c, u - v \rangle, \quad u \in U.$$

Это означает, что $c \in \partial J(v)$, т. е. $\partial J(v) \neq \emptyset$. Теорема 1 доказана.

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 $\partial J(v)$ 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.

$$J(u) - J(v) \geq \langle c_1, u - v \rangle, \quad J(u) - J(v) \geq \langle c_2, u - v \rangle, \quad u \in U.$$

Возьмем $\alpha \in [0, 1]$. Умножая первое неравенство на α , второе — на $1 - \alpha$ и складывая, получим $J(u) - J(v) \geq \langle \alpha c_1 + (1 - \alpha)c_2, u - v \rangle$ 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 $\partial J(v)$ доказана.

Пусть c — предельная точка множества $\partial J(v)$, пусть $\{c_k\} \in \partial J(v)$ и $c_k \rightarrow c$ при $k \rightarrow \infty$. Из $c_k \in \partial J(v)$ следует, что $J(u) - J(v) \geq \langle c_k, u - v \rangle$, $u \in U$. При $k \rightarrow \infty$ отсюда получим $c \in \partial J(v)$. Замкнутость $\partial J(v)$ доказана.

Покажем ограниченность $\partial J(v)$. Возьмем любой вектор $c \in \partial J(v)$. Так как U — открытое множество, то шар $S(v, \varepsilon) = \{u \in E^n: |u - v| \leq \varepsilon\} \subset U$ при достаточно малом $\varepsilon > 0$. Далее, в силу теоремы 2.15 $J(u)$ непрерывна на U , поэтому $\sup_{S(v, \varepsilon)} J(u) = J^*(S) < \infty$. Положим в неравенстве (2) $u =$

$= v + \varepsilon c / |c| \in S(v, \varepsilon)$. Получим $|c| \leq (J(v + \varepsilon c / |c|) - J(v)) / \varepsilon < (J^*(S) - J(v)) / \varepsilon < \infty$ при всех $c \in \partial J(v)$. Теорема 2 доказана.

3. 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

$$\frac{dJ(v)}{de} = \max_{c \in \partial J(v)} \langle c, e \rangle. \quad (5)$$

Доказательство. Существование производной $\frac{dJ(v)}{de}$ установлено в теореме 2.14. Докажем формулу (5). Из (2) имеем $(J(v + te) - J(v)) / t \geq \langle c, e \rangle$ при всех $c \in \partial J(v)$ и всех достаточно малых $t > 0$. Отсюда при $t \rightarrow +0$ получим $\frac{dJ(v)}{de} \geq \langle c, e \rangle$ для любого $c \in \partial J(v)$, так что

$\frac{dJ(v)}{de} \geq \sup_{c \in \partial J(v)} \langle c, e \rangle$. С другой стороны, при доказательстве теоремы 1

был построен специальный субградиент c , для которого выполняется неравенство (4). Полагая в (4) $u = v$, будем иметь $\frac{\partial J(v)}{\partial e} \leq \langle c, e \rangle$, $c \in \partial J(v)$. Сравнивая это неравенство с предыдущим, приходим к формуле (5). Попутно показали, что максимум в правой части (5) достигается, именно, на том субградиенте, который был построен в теореме (1). Теорема 3 доказана.

Формула (5) обобщает известную формулу $\frac{dJ(v)}{de} = \langle J'(v), e \rangle$ для гладких функций.

4. С помощью субдифференциала можно сформулировать следующий простой (и не очень-то содержательный) критерий оптимальности, не требуя даже выпуклости рассматриваемых множеств и функций.

Теорема 4. Для того чтобы функция $J(u)$ достигала своей нижней грани на множестве U в точке $u_* \in U$, необходимо и достаточно, чтобы $0 \in \partial J(u_*)$.

Доказательство. **Необходимость.** Если u_* — точка минимума $J(u)$ на U , то $J(u) \geq J(u_*) = J(u_*) + \langle 0, u - u_* \rangle$ при всех $u \in U$. Это значит, что $0 \in \partial J(u_*)$.

Достаточность. Если $0 \in \partial J(u_*)$ в некоторой точке $u_* \in U$, то $J(u) \geq J(u_*) + \langle 0, u - u_* \rangle = J(u_*)$ при всех $u \in U$, т. е. u_* — точка минимума $J(u)$ на U . Теорема 4 доказана.

5. 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) 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 $g(u) = J(\lambda u)$, $\lambda > 0$, то $\partial g(u) = \lambda \partial J(\lambda u)$.

4) Если $J_i(u)$ — выпуклые функции на E^n , $i = \overline{1, m}$, то

$$\partial \left(\sum_{i=1}^m J_i(u) \right) = \sum_{i=1}^m \partial J_i(u), \quad u \in E^n. \quad (6)$$

Заметим, что если выпуклая функция $J_i(u)$ определена на выпуклом множестве U_i , необязательно совпадающем с E^n , то $J(u) = \sum_{i=1}^m J_i(u)$ определена на

$U = \bigcap_{i=1}^m U_i$, и здесь вместо (6) в общем случае можно гарантировать лишь

включение $\sum_{i=1}^m \partial J_i(u) \subset \partial \left(\sum_{i=1}^m J_i(u) \right)$, $u \in U$, а для получения равенства достаточно потребовать, например, чтобы функции $J_i(u)$ были непрерывны хотя бы в одной точке $v \in \bigcap_{i=1}^m U_i$. Если $c_i \in \partial J_i(u)$, $i = \overline{1, m}$, то в качестве представителя из $\partial J(u)$ всегда можно брать $c = c_1 + \dots + c_m$.

5) Если $J_1(u), \dots, J_m(u)$ — выпуклые функции на E^n , то функция

$$J(u) = \max_{1 \leq i \leq m} J_i(u), \quad u \in E^n,$$

также выпукла на E^n (см. теорему 2.7), причем

$$\begin{aligned} \partial J(u) &= \text{co} \left(\bigcup_{i \in I(u)} \partial J_i(u) \right) = \\ &= \left\{ c: c = \sum_{i \in I(u)} \alpha_i c_i, \quad c_i \in \partial J_i(u), \quad \alpha_i \geq 0, \quad i \in I(u), \quad \sum_{i \in I(u)} \alpha_i = 1 \right\}. \end{aligned} \quad (7)$$

где $I(u) = \{i: 1 \leq i \leq m, J_i(u) = J(u)\}$. В частности, в качестве представителя из $\partial J(u)$ можно взять $c = c_i \in \partial J_i(u)$ при некотором $i \in I(u)$.

6) Если C — выпуклое замкнутое ограниченное множество из E^n , то функция

$$J(u) = \max_{c \in C} \langle c, u \rangle, \quad u \in E^n,$$

выпукла на E^n (см. теорему 2.7), причем

$$\partial J(u) = \{c: c \in C, \langle c, u \rangle = J(u)\}.$$

7) Если функция $J(u)$ выпукла на E^m , а A — матрица порядка $m \times n$, то функция

$$g(u) = J(Au), \quad u \in E^n,$$

выпукла на E^n , причем

$$\partial g(u) = A^T \partial J(v) |_{v=Au}.$$

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, [1, 89, 99, 109, 113, 114, 197, 207].

В заключение приведем еще один пример:

Пример 3. Пусть функция $g(u)$ выпукла на E^n , пусть

$$J(u) = \max \{0; g(u)\}, \quad u \in E^n. \quad (8)$$

Если $g(u) > 0$, то ясно, что $\partial J(u) = \partial g(u)$; если $g(u) < 0$, то $\partial J(u) = \{0\}$. Рассмотрим случай $g(u) = 0$. Тогда $J(u) = 0$. Согласно (7) тогда

$$\partial J(u) = \{c: \alpha c + (1 - \alpha) \cdot 0, \quad c \in \partial g(u), \quad 0 \in \partial(0), \quad 0 \leq \alpha \leq 1\} = \alpha \partial g(u),$$

$0 \leq \alpha \leq 1$. В качестве представителя из $\partial J(u)$ при $g(u) = 0$ можно, например, взять $c = 0$ или некоторый вектор $c \in \partial g(u)$.

Кстати, заметим, что если даже отказаться от выпуклости функции $g(u)$, то в точке v , где $g(v) = 0$, субдифференциал функции (8) всегда будет непустым, так как $0 \in \partial J(u)$. Это и неудивительно, если вспомнить теорему 4 и заметить, что $\inf_{E^n} J(u) = J(v) = 0$.

E^n

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 $J_1(u), \dots, J_m(u)$, $u \in E^n$, непрерывно дифференцируемы в некоторой окрестности точки v . Доказать, что тогда функции $J(u) = \max_{1 \leq i \leq m} J_i(u)$ в точке v имеет производные по любому направлению

e , $|e| = 1$, причем

$$\frac{dJ(v)}{de} = \max_{i \in I(v)} \langle J'_i(v), e \rangle, \quad I(v) = \{i: 1 \leq i \leq m, J_i(v) = J(v)\}.$$

Установить связь между этой формулой и формулой (5) для случая, когда функции $J_1(u), \dots, J_m(u)$ выпуклы на E^n .

3. Пусть A — замкнутое ограниченное множество из E^m , а функция $g(u, a)$ определена на $E^n \times A$, полунепрерывна сверху по a при каждом $u \in E^n$, выпукла и непрерывна по u при каждом $a \in A$. Тогда функция

$$J(u) = \sup_{a \in A} g(u, a), \quad u \in E^n,$$

выпукла на E^n и ее субдифференциал имеет вид

$$\partial J(u) = \overline{\text{co}} \left(\bigcup_{a \in A(u)} \partial g(u, a) \right), \quad A_0(u) = \{a: a \in A, g(u, a) = J(u)\}$$

(здесь $\overline{\text{co}} B$ — замыкание выпуклой оболочки B). Доказать (ср. с (7)).

4. Найти субдифференциалы функций $J(u) = \max_{|t| \leq 1} |t^2 + xt + y|$,

$$J(u) = \max_{|t| \leq 1} |xt^2 + yt|, \quad J(u) = \max_{0 \leq t \leq 1} |x + ty|,$$

$$u = (x, y) \in E^2.$$

5. 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 $\frac{\partial g(u, a)}{\partial u}$. Доказать, что тогда функция $J(u) = \max_{a \in A} g(u, a)$ во всех

точках $v \in E^n$ имеет производную по любому направлению e , $|e| = 1$, причем

$$\frac{dJ(v)}{de} = \max_{a \in A_0(v)} \left\langle \frac{\partial g(v, a)}{\partial u}, e \right\rangle, \quad A_0(v) = \{a: a \in A, g(v, a) = J(v)\}.$$

Установить связь между этой формулой и формулой (5) для случая, когда $g(u, a)$ выпукла по u при каждом $a \in A$, с учетом результата из упражнения 3.

§ 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. Нельзя ли 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 , $0 \leq t \leq \text{diam } U = \sup_{u, v \in U} |u - v|$, $\delta(0) = 0$, $\delta(t_0) > 0$ при некотором t_0 , $0 < t_0 < \text{diam } U$ и такая, что

$$J(\alpha u + (1 - \alpha)v) \leq \alpha J(u) + (1 - \alpha)J(v) - \alpha(1 - \alpha)\delta(|u - v|) \quad (1)$$

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 , $0 < t < \text{diam } U$, то такую функцию называют *строго равномерно выпуклой* на U [62].

Очевидно, всякая сильно выпуклая функция является строго равномерно выпуклой с модулем $\delta(t) = \kappa t^2$. 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 $\delta(t) \leq \mu(t)$, $0 \leq t \leq \text{diam } U$, неотрицательная, не тождественно 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, возможно, $U = E^n$), а функция $J(u)$ равномерно выпукла и непрерывна снизу на U . Тогда

1) множество Лебега $M(v) = \{u: u \in U, J(u) \leq J(v)\}$ выпукло, замкнуто и ограничено при всех $v \in U$;

2) $J_* = \inf_U J(u) > -\infty$, $U_* = \{u: u \in U, J(u) = J_*\} \neq \emptyset$;

3) имеет место неравенство

$$\delta(|u - u_*|) \leq J(u) - J(u_*) \quad (2)$$

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 $\{u_k\}$: $\{u_k\} \in U$, $\lim_{k \rightarrow \infty} J(u_k) = J_*$ 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

$$\mu(ct) \geq c^2 \mu(t) \quad (3)$$

для всех $c \geq 1$, $t \geq 0$, $0 \leq ct \leq \text{diam } U$.

Доказательство. Сначала рассмотрим случай $1 \leq c < 2$. 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

$$\mu(ct) \leq \frac{\alpha J(u_1) + (1-\alpha)J(u_2) - J(u_\alpha)}{\alpha(1-\alpha)} \leq \mu(ct) + \varepsilon,$$

где $u_\alpha = \alpha u_1 + (1-\alpha)u_2$. Отсюда имеем

$$\alpha J(u_1) + (1-\alpha)J(u_2) - J(u_\alpha) \leq \alpha(1-\alpha)\mu(ct) + \alpha(1-\alpha)\varepsilon. \quad (4)$$

Можем считать, что $0 < \alpha \leq 1/2$, так как в противном случае в (4) точки u_1 и u_2 можно поменять ролями. Тогда с учетом $1 \leq c < 2$ можем сказать, что $0 < \alpha \leq \alpha c < 1$. Кроме того, $1/2 < 1/c \leq 1$, 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

$$J(u_3) \leq (1/c)J(u_1) + (1-1/c)J(u_2) - (1/c)(1-1/c)\mu(ct),$$

$$J(u_\alpha) \leq \alpha c J(u_3) + (1-\alpha c)J(u_2) - \alpha c(1-\alpha c)\mu(t).$$

Умножим первое из этих неравенств на αc и сложим со вторым. Учитывая неравенство (4), получим

$$\begin{aligned} \alpha c(1/c)(1-1/c)\mu(ct) + \alpha c(1-\alpha c)\mu(t) &\leq \\ &\leq \alpha J(u_1) + (\alpha c - \alpha)J(u_2) + (1-\alpha c)J(u_2) - J(u_\alpha) = \\ &= \alpha J(u_1) + (1-\alpha)J(u_2) - J(u_\alpha) \leq \alpha(1-\alpha)\mu(ct) + \alpha(1-\alpha)\varepsilon \end{aligned}$$

или

$$\alpha(1-1/c)\mu(ct) + \alpha c(1-\alpha c)\mu(t) \leq \alpha(1-\alpha)\mu(ct) + \alpha(1-\alpha)\varepsilon.$$

Поскольку здесь $\varepsilon > 0$ — произвольное число, то можем ε устремить к $+0$. Будем иметь

$$\alpha c(1-\alpha c)\mu(t) \leq (\alpha/c)(1-\alpha c)\mu(ct) \quad \text{или} \quad c^2\mu(t) \leq \mu(ct).$$

Неравенство (3) для случая $1 \leq c < 2$ доказано. Пусть теперь $c \geq 1$ — произвольное число, $0 \leq ct \leq \text{diam } U$. Так как $\lim_{n \rightarrow \infty} \sqrt[n]{c} = 1$, то найдется b ,

$1 \leq b < 2$, такое, что $c = b^n$ при некотором натуральном $n \geq 1$. Учитывая, что по доказанному $\mu(bt) \geq b^2\mu(t)$, получим $\mu(ct) = \mu(b^n t) = \mu(b \cdot b^{n-1}t) \geq \geq b^2\mu(b^{n-1}t) \geq b^4\mu(b^{n-2}t) \geq \dots \geq b^{2n}\mu(t) = c^2\mu(t)$. Лемма 1 доказана.

Л е м м а 2. Пусть $\mu(t)$ — точный модуль выпуклости равномерно выпуклой функции $J(u)$ на выпуклом множестве U . Тогда

- 1) $\mu(t) = O(t^2)$ при $t \rightarrow +0$;
- 2) $\mu(t) \equiv 0$ при $0 \leq t < \tau = \inf \{t: \mu(t) > 0\}$, $\mu(t)$ строго монотонна при $\tau < t \leq \text{diam } U$;
- 3) если $\text{diam } U = +\infty$, то $\lim_{t \rightarrow \infty} \mu(t) = +\infty$.

Доказательство. Из определения 1 следует, что $\mu(t_0) > 0$ при некотором t_0 , $0 < t_0 < \text{diam } U$. Поэтому $0 \leq \tau < \text{diam } U$. Если $\tau > 0$, то $\mu(t) \equiv 0$ при $0 \leq t < \tau$ по определению τ . Пусть $\tau < t < a < \text{diam } U$. Тогда с помощью неравенства (3) имеем $\mu(a) = \mu((a/t)t) \geq (a/t)^2\mu(t) > \mu(t) > 0$, т. е. $\mu(t)$ строго монотонна при $\tau < t \leq \text{diam } U$.

Далее, если $\tau > 0$, то условие $\mu(t) = O(t^2)$ при $t \rightarrow +0$ выполняется тривиально. Поэтому пусть $\tau = 0$. Тогда, фиксируя какое-либо t_0 , $0 < t_0 \leq \text{diam } U$, для всех $0 < t < t_0$, имеем $\mu(t_0) = \mu((t_0/t)t) \geq (t_0/t)^2\mu(t)$ или $\mu(t) \leq \mu(t_0)t^2/t_0^2 = \text{const} \cdot t^2$. Это и означает, что $\mu(t) = O(t^2)$ при $t \rightarrow +0$.

Наконец, пусть $\text{diam } U = +\infty$. Тогда $\mu(t)$ определена при всех $t \geq 0$. Пусть $t \geq t_0 > \tau$. Тогда $\mu(t) = \mu((t/t_0)t_0) \geq (t/t_0)^2 \mu(t_0) = \text{const} \cdot t^2$. Это значит, что $\mu(t) \rightarrow \infty$ при $t \rightarrow \infty$ со скоростью не медленнее, чем t^2 . Лемма 2 доказана.

Заметим, что из неравенства $0 \leq \delta(t) \leq \mu(t)$, справедливого для любого модуля выпуклости равномерно выпуклой функции, и из леммы 2 следует, что условие $\delta(t) = O(t^2)$ при $t \rightarrow +0$ является необходимым для того, чтобы некоторая функция $\delta(t)$ могла служить модулем выпуклости для какой-либо равномерно выпуклой функции.

Доказательство теоремы 1. Если U ограничено, замкнуто, т. е. U компактно, то утверждения 1), 2) теоремы следуют из теорем 2.1.1 и 2.10. Остается рассмотреть случай, когда U — неограниченное множество. Тогда $\text{diam } U = +\infty$ и точный модуль выпуклости $\mu(t)$ функции $J(u)$ будет определен при всех $t \geq 0$. Пусть $t_0 > 0$ и $\mu(t_0) > 0$. Возьмем произвольную точку $v \in U$ и рассмотрим шар

$$S = S(v, t_0) = \{u: u \in U, |u - v| \leq t_0\}.$$

Из теоремы 2.1.1 следует, что $\inf_S J(u) = J_S^* > -\infty$, так что

$$J(u) \geq J_S^* = J(v) - v, \quad v = J(v) - J_S^* \geq 0 \quad (5)$$

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

$$0 < \alpha_0 = (\mu(t_0)/\mu(|u - v|))^{1/2} < 1. \quad (6)$$

При $\alpha = \alpha_0$ из (1) получаем

$$\alpha_0 J(u) \geq J(v + \alpha_0(u - v)) - (1 - \alpha_0)J(v) + \alpha_0(1 - \alpha_0)\mu(|u - v|). \quad (7)$$

Из (6) и леммы 1 следует $\mu(t_0) = \alpha_0^2 \mu(|u - v|) = \alpha_0^2 \mu((1/\alpha_0) \cdot \alpha_0 |u - v|) \geq \mu(\alpha_0 |u - v|)$ или $\mu(t_0) \geq \mu(\alpha_0 |u - v|)$. В силу монотонности $\mu(t)$ это означает, что $\alpha_0 |u - v| \leq t_0$. Тогда $v + \alpha_0(u - v) \in S$ и согласно (5) $J(v + \alpha_0(u - v)) \geq J(v) - v$. Учитывая эту оценку, из (7) имеем

$$\alpha_0 J(u) \geq \alpha_0 J(v) - v + \alpha_0(1 - \alpha_0)\mu(|u - v|).$$

Отсюда, сокращая на $\alpha_0 > 0$ и вспоминая определение 6) величины α_0 , получим

$$J(u) \geq J(v) + (1 - \alpha_0)\mu(|u - v|) - v/\alpha_0 = \\ = J(v) + \mu(|u - v|) - \sqrt{\mu(|u - v|)}(\sqrt{\mu(t_0)} + v/\sqrt{\mu(t_0)}).$$

Применяя к последнему слагаемому неравенство $ab \leq (a^2 + b^2)/2$, будем иметь

$$J(u) \geq J(v) + \mu(|u - v|)/2 - (\sqrt{\mu(t_0)} + v/\sqrt{\mu(t_0)})^2/2 \quad (8)$$

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 $\mu(|u - v|) \leq \mu(t_0)$, а тогда $v < (\sqrt{\mu(t_0)} + v/\sqrt{\mu(t_0)})^2/2 - \mu(|u - v|)/2$. Отсюда и из (5) следует справедливость (8) и для $u \in S$.

Для всех $u \in M(v)$ из (8) имеем $\mu(|u - v|)/2 - (\sqrt{\mu(t_0)} + v/\sqrt{\mu(t_0)})^2/2 \leq J(u) - J(v) \leq 0$, т. е. $\mu(|u - v|) \leq (\sqrt{\mu(t_0)} + v/\sqrt{\mu(t_0)})^2$ при любом $u \in M(v)$. Поскольку $\mu(t) \rightarrow +\infty$ при $t \rightarrow +\infty$ 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 $J_* > -\infty$, $U_* \neq \emptyset$.

Докажем неравенство (2). Так как любой модуль выпуклости $\delta(t) \leq \mu(t)$, то неравенство (2) достаточно доказать для $\mu(t)$. Возьмем любую точку $u_* \in U$. Тогда $0 \leq J(\alpha u + (1-\alpha)u_*) - J(u_*) \leq \alpha(J(u) - J(u_*)) - \alpha(1-\alpha)\mu(|u - u_*|)$ или $\alpha(1-\alpha)\mu(|u - u_*|) \leq \alpha(J(u) - J(u_*))$, $0 < \alpha < 1$, $u \in U$. Деля на $\alpha > 0$ и устремляя $\alpha \rightarrow +0$, откуда получим 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 $k \rightarrow \infty$, получим $\delta(|u_k - u_*|) \rightarrow 0$. Это возможно только при $|u_k - u_*| \rightarrow 0$, так как $J(u)$ строго равномерно выпукла. Теорема 1 доказана.

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

$$J(u) \geq J(v) + \langle c(v), u - v \rangle + \delta(|u - v|). \quad (9)$$

$$\langle c(u) - c(v), u - v \rangle \geq 2\delta(|u - v|) \quad (10)$$

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 субдифференциалы $\partial J(u) \neq \emptyset$ 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

$$\begin{aligned} \alpha \langle c(v), u - v \rangle + J(v) &\leq J(\alpha u + (1-\alpha)v) \leq \\ &\leq \alpha J(u) + (1-\alpha)J(v) - \alpha(1-\alpha)\delta(|u - v|) \end{aligned}$$

или $(1-\alpha)\delta(|u - v|) + \langle c(v), u - v \rangle \leq J(u) - J(v)$. Отсюда при $\alpha \rightarrow +0$ получим неравенство (9). Поменяем в (9) переменные u и v ролями; будем иметь

$$J(v) \geq J(u) + \langle c(u), v - u \rangle + \delta(|u - v|), \quad c(u) \in \partial J(u).$$

Складывая это неравенство с (9), приходим к (10). Теорема доказана.

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 $\xi(t)$, $0 \leq t \leq \text{diam } U$, $\xi(t) = O(t^2)$ при $t \rightarrow +0$, $\xi(t) \neq 0$, выполняется неравенство

$$\langle J'(u) - J'(v), u - v \rangle \geq \xi(|u - v|) \quad (11)$$

при всех $u, v \in U$. Тогда функция $J(u)$ равномерно выпукла на U с модулем

$$\text{выпуклости } \delta(t) = \int_0^1 (\xi(\tau t)/\tau) d\tau.$$

Доказательство. Из формулы (2.7) и условия (11) имеем

$$\begin{aligned} \alpha J(u) + (1-\alpha)J(v) - J(\alpha u + (1-\alpha)v) &= \\ &= \alpha(1-\alpha) \int_0^1 \langle J'(z_1) - J'(z_2), z_1 - z_2 \rangle d\tau/\tau \geq \alpha(1-\alpha) \int_0^1 \xi(|z_1 - z_2|) d\tau/\tau = \\ &= \alpha(1-\alpha) \int_0^1 \xi(\tau|u - v|) d\tau/\tau = \alpha(1-\alpha)\delta(|u - v|), \quad u, v \in U, \quad \alpha \in [0, 1], \end{aligned}$$

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 $p \geq 2$.

Доказательство. Покажем, что

$$\langle J'(u) - J'(v), u - v \rangle \geq \frac{p}{2} \min \{1; 2^{3-p}\} |u - v|^p, \quad u, v \in E^n. \quad (12)$$

Здесь $J'(u) = p|u|^{p-2}u$. Тогда

$$\begin{aligned} \langle J'(u) - J'(v), u - v \rangle &= \langle p|u|^{p-2}u - p|v|^{p-2}v, u - v \rangle = \\ &= p[|u|^p + |v|^p - \langle u, v \rangle (|u|^{p-2} + |v|^{p-2})] = \\ &= p \left[|u|^p + |v|^p - \frac{|u|^2 + |v|^2 - |u - v|^2}{2} (|u|^{p-2} + |v|^{p-2}) \right] = \\ &= \frac{p}{2} [(|u|^{p-2} - |v|^{p-2})(|u|^2 - |v|^2) + |u - v|^2 (|u|^{p-2} + |v|^{p-2})] \geq \\ &\geq \frac{p}{2} |u - v|^2 (|u|^{p-2} + |v|^{p-2}), \quad u, v \in E^n. \quad (13) \end{aligned}$$

Покажем, что

$$|u|^{p-2} + |v|^{p-2} \geq |u - v|^{p-2} \min \{1; 2^{3-p}\}, \quad u, v \in E^n. \quad (14)$$

Рассмотрим функцию $\varphi(x) = (x^\alpha + 1)/(x + 1)^\alpha$ при $x \geq 1$, $\alpha > 0$. Имеем $\varphi'(x) = \alpha(x^{\alpha-1} - 1)(x + 1)^{-\alpha-1}$. Если $\alpha \geq 1$, то $\varphi'(x) \geq 0$, и $\varphi(x) \geq \varphi(1) = 2^{1-\alpha}$ для всех $x \geq 1$. Если $0 < \alpha < 1$, то $\varphi'(x) < 0$, и $\varphi(x) \geq \varphi(+\infty) = 1$ при всех $x \geq 1$. Следовательно, $\varphi(x) \geq A_\alpha = \min \{1; 2^{1-\alpha}\}$, $x \geq 1$, или

$$A_\alpha (x + 1)^\alpha \leq x^\alpha + 1, \quad x \geq 1, \alpha > 0. \quad (15)$$

Далее имеем $|u - v|^{p-2} \leq (|u| + |v|)^{p-2}$. Без ограничения общности можем считать $|u| \geq |v|$. Тогда с помощью неравенства (15) получим

$$|u - v|^{p-2} \leq |v|^{p-2} (|u|/|v| + 1)^{p-2} \leq A_{p-2}^{-1} ((|u|/|v|)^{p-2} + 1) |v|^{p-2},$$

что равносильно (14). Из (13) и (14) следует неравенство (12). С помощью теоремы 3 отсюда заключаем, что функция $J(u) = |u|^p$ при $p \geq 2$ равномерно выпукла на E^n с модулем $\delta(t) = t^p \min \{1; 2^{3-p}\}/2$. Теорема доказана. Более тонкие оценки показывают, что функция $J(u) = |u|^p$ при $p \geq 2$ имеет точный модуль выпуклости $\mu(t) = t^p/2^{p-2}$, $t \geq 0$.

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 $x \geq 0$, 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 $x \geq 1$ с некоторым модулем выпуклости $\delta(t)$, то согласно теореме 2 необходимо выполнялось бы неравенство (10). В данном случае неравенство (10) имеет вид

$$2\delta(t) \leq t^p [(x + t)^{p-1} - x^{p-1}], \quad x \geq 0, \quad t \geq 0.$$

С помощью формулы конечных приращений отсюда имеем

$$2\delta(t) \leq t^2 p(p-1)(x + \theta t)^{p-2} \leq t^2 p(p-1)x^{p-2}, \quad x \geq 0, \quad t \geq 0.$$

Зафиксируем здесь произвольное $t > 0$, а x устремим к $+\infty$. 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 [62].

§ 8. 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:

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

$$U = \{u \in E^n: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}; g_i(u) = 0, i = \overline{m+1, s}\}, \quad (2)$$

где U_0 — заданное выпуклое множество из E^n , функции $J(u)$, $g_1(u), \dots, g_m(u)$ определены и выпуклы на U_0 , а $g_i(u) = \langle a_i, u \rangle - b^i$ при $i = \overline{m+1, s}$ — линейные функции, b^i — заданные числа, a_i — заданные векторы из E^n . Здесь не исключаются возможности, когда отсутствуют либо ограничения $g_i(u) \leq 0$ типа неравенств ($m = 0$), либо ограничения $g_i(u) = 0$ типа равенств ($s = m$), либо оба этих вида ограничений ($m = s = 0, U = U_0$). Разумеется, и множество U_0 может задаваться ограничениями типа неравенств или равенств. При выделении множества U_0 в (2) руководствуются обычно тем, чтобы U_0 имело простую структуру, чтобы легко (без трудоемкой вычислительной работы) можно было проверить включение $u \in U_0$, чтобы легко было проектировать точку на U_0 и т.п. Обычно множество U_0 представляет собой параллелепипед

$$U_0 = \{u = (u^1, \dots, u^n): \alpha_i \leq u^i \leq \beta_i, i = \overline{1, n}\},$$

где $\alpha_i, \beta_i, \alpha_i < \beta_i$ — заданные числа, или, как мы видели в задачах линейного программирования, положительный октант

$$U_0 = \{u = (u^1, \dots, u^n): u^i \geq 0, i = \overline{1, n}\}$$

или

$$U_0 = \{u = (u^1, \dots, u^n): u^i \geq 0, i \in I\},$$

где I — заданное множество номеров из $\{1, 2, \dots, n\}$. Конечно, не исключается возможность, когда в (2) $U_0 = E^n$. При сделанных предположениях по теореме 2.11 множество (2) выпукло.

Важное место в теории выпуклого программирования занимает теорема Куна — Таккера, названная так в честь открывших ее американских математиков Куна и Таккера. Эта теорема дает необходимое и достаточное условие оптимальности в задаче (1), (2), т. е. условие принадлежности той или иной точки множеству

$$U_* = \{u: u \in U, J(u) = \inf_{v \in U} J(v) = J_*\},$$

и представляет собой обобщение правила множителей Лагранжа (см. § 2.2) на случай задачи (1), (2). Для формулировки теоремы Куна — Таккера введем функцию Лагранжа

$$L(u, \lambda) = J(u) + \sum_{i=1}^s \lambda_i g_i(u) \quad (3)$$

задачи (1), (2), где $u \in U_0$, а переменные $\lambda = (\lambda_1, \dots, \lambda_s)$, называемые множителями Лагранжа, принадлежат множеству

$$\Lambda_0 = \{\lambda = (\lambda_1, \dots, \lambda_s) \in E^s: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\}. \quad (4)$$

Определение 1. Точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ называют *седловой точкой* функции Лагранжа (3), если

$$L(u_*, \lambda) \leq L(u_*, \lambda^*) \leq L(u, \lambda^*) \quad \text{при всех } u \in U_0, \lambda \in \Lambda_0. \quad (5)$$

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:

$$L(u_*, \lambda^*) \leq L(u, \lambda^*) \quad \text{при всех } u \in U_0, \quad (6)$$

$$\lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, s}; \quad u_* \in U. \quad (7)$$

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:

$$J(u_*) + \sum_{i=1}^s \lambda_i g_i(u_*) \leq J(u_*) + \sum_{i=1}^s \lambda_i^* g_i(u_*), \quad \lambda \in \Lambda_0. \quad (8)$$

Отсюда имеем

$$\sum_{i=1}^s (\lambda_i^* - \lambda_i) g_i(u_*) \geq 0, \quad \lambda \in \Lambda_0. \quad (9)$$

Покажем, что $u_* \in U$. Возьмем точку $\lambda = (\lambda_1, \dots, \lambda_s)$, где $\lambda_j = \lambda_j^* + 1$ при некотором j , $1 \leq j \leq m$, и $\lambda_i = \lambda_i^*$ при всех остальных $i = \overline{1, s}$, $i \neq j$. 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 $(-1)g_j(u_*) \geq 0$, т. е. $g_j(u_*) \leq 0$ при всех $j = \overline{1, m}$. Далее, пусть $\lambda = (\lambda_1, \dots, \lambda_s)$ с координатами $\lambda_j = \lambda_j^* + g_j(u_*)$ при некотором j , $m+1 \leq j \leq s$, и $\lambda_i = \lambda_i^*$ при всех $i = \overline{1, s}$, $i \neq j$. Ясно, что $\lambda \in \Lambda_0$. Поэтому из (9) имеем $-|g_j(u_*)|^2 \geq 0$, т. е. $g_j(u_*) = 0$ при всех $j = \overline{m+1, s}$. Таким образом, доказано, что $u_* \in U$.

Из того, что $g_i(u_*) = 0$ при $i = \overline{m+1, s}$, следует, что $\lambda_i^* g_i(u_*) = 0$, $i = \overline{m+1, s}$. Остается получить равенства (7) при $i = \overline{1, m}$. Возьмем точку $\lambda = (\lambda_1, \dots, \lambda_s)$ с координатами $\lambda_j = 0$ при некотором j , $1 \leq j \leq m$, и $\lambda_i = \lambda_i^*$ при всех остальных $i = \overline{1, s}$, $i \neq j$. Такая точка принадлежит Λ_0 , поэтому из (9) получим $0 \leq \lambda_j^* g_j(u_*)$. Но $\lambda_j^* \geq 0$, $g_j(u_*) \leq 0$ при $j = \overline{1, m}$, поэтому последнее неравенство возможно лишь при $\lambda_j^* g_j(u_*) = 0$, $j = \overline{1, m}$. 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. $g_i(u_*) \leq 0$, $i = \overline{1, m}$, $g_i(u_*) = 0$, $i = \overline{m+1, s}$. Тогда

$$(\lambda_i^* - \lambda_i) g_i(u_*) = 0 \quad (10)$$

при всех $i = \overline{m+1, s}$ и всех тех i , $1 \leq i \leq m$, для которых $g_i(u_*) = 0$. Если $g_i(u_*) < 0$ при некотором i , $1 \leq i \leq m$, то из равенства (7) следует, что $\lambda_i^* = 0$. Поэтому $(\lambda_i^* - \lambda_i) g_i(u_*) = -\lambda_i g_i(u_*) \geq 0$ для всех $\lambda_i \geq 0$, $1 \leq i \leq m$, для которых $g_i(u_*) < 0$. Складывая полученные неравенства с (10), будем иметь $\sum_{i=1}^s (\lambda_i^* - \lambda_i) g_i(u_*) \geq 0$ для всех $\lambda \in \Lambda_0$. Отсюда

$\sum_{i=1}^s \lambda_i g_i(u_*) \leq \sum_{i=1}^s \lambda_i^* g_i(u_*)$ при всех $\lambda \in \Lambda_0$. Добавляя к обеим частям этого неравенства $J(u_*)$, приходим к неравенству (8), представляющему собой левое неравенство (5). Лемма 1 доказана.

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 $u_* \in U$, $J_* = L(u_*, \lambda^*) = J(u_*)$, т. е. u_* является решением задачи (1), (2).

Доказательство. Из условия (7) леммы 1 имеем $u_* \in U$ и $L(u_*, \lambda^*) = J(u_*)$. Тогда неравенство (6) перепишется в виде

$$J(u_*) \leq L(u, \lambda^*) = J(u) + \sum_{i=1}^s \lambda_i^* g_i(u), \quad u \in \dot{U}_0. \quad (11)$$

В частности, (11) верно и для всех $u \in U$. Но $\sum_{i=1}^s \lambda_i^* g_i(u) \leq 0$ при $u \in U$, так как тогда $\overline{g_i(u)} \leq 0$ и $\lambda_i^* \geq 0$ при $i = \overline{1, m}$, так что $\lambda_i^* g_i(u) \leq 0$, $i = \overline{1, m}$, и $g_i(u) = 0$ при $i = \overline{m+1, s}$, так что $\lambda_i g_i(u) = 0$, $i = \overline{m+1, s}$. Поэтому из (11) следует, что $J(u_*) \leq L(u, \lambda^*) \leq J(u)$ при всех $u \in U$, т. е. $u_* \in U_*$. Теорема доказана.

Заметим, что теорема 1, как и лемма 1, доказаны без каких-либо ограничений на функции $J(u)$, $g_i(u)$, $i = \overline{1, s}$, 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, отрицательный: если в задаче (1), (2) $U_* = \emptyset$, 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 $U_* \neq \emptyset$ в общем случае нельзя ожидать, что функция Лагранжа будет иметь седловую точку.

Пример 1. Пусть $J(u) = -u$, $U_0 = \{u \in E^1: u \geq 0\}$, $g(u) = u^2$, $U = \{u: u \in U_0, g(u) \leq 0\}$. 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

$$L(u, \lambda) = -u + \lambda u^2, \quad u \geq 0, \quad \lambda \geq 0,$$

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. Теоремы, в которых устанавливается существование седловой точки функции Лагранжа задачи (1), (2), в литературе обычно называют теоремой Куна — Таккера. Ниже будут приведены несколько вариантов теоремы Куна — Таккера. Начнем с рассмотрения случая, когда в (2) ограничения типа равенств отсутствуют ($m = s$), т. е. множество U имеет вид

$$U = \{u: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}\}. \quad (12)$$

Определение 2. Ограничение $g_i(u) \leq 0$ из (12) называют *регулярным* на множестве U_0 , если существует точка $u_i \in U_0$ такая, что

$$g_i(u_i) < 0. \quad (13)$$

Множество (12) называют *регулярным*, если в (12) все ограничения $g_i(u) \leq 0$, $i = \overline{1, m}$, регулярны на U_0 .

Если U_0 — выпуклое множество, функции $g_i(u)$ выпуклы на U_0 , $i = \overline{1, m}$, а множество (12) регулярно, то существует точка $\bar{u} \in U_0$, для которой

$$g_i(\bar{u}) < 0 \text{ при всех } i = \overline{1, m}. \quad (14)$$

В самом деле, положим $\bar{u} = \sum_{i=1}^m \alpha_i u_i$, где $\alpha_i \geq 0$, $\sum_{i=1}^m \alpha_i = 1$, точки u_i взяты из (13), $i = \overline{1, m}$. Тогда в силу неравенства (2.2) имеем

$$g_i(\bar{u}) \leq \sum_{i=1}^m \alpha_i g_i(u_i) < 0, \quad i = \overline{1, m},$$

т. е. условие (14) выполнено. Условие (14) часто называют *условием Слейтера*.

Теорема 2. Пусть множество U_0 выпукло, функции $J(u)$, $g_i(u)$, $i = \overline{1, m}$, выпуклы на U_0 , а множество (12) регулярно. Пусть множество U_* точек минимума функции $J(u)$ на множестве (12) непусто. Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*) \in \Lambda_0 = \{\lambda \in E^m: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\}$, такие, что пара (u_*, λ^*) образует седловую точку функции Лагранжа на множестве $U_0 \times \Lambda_0$.

Доказательство. В пространстве E^{m+1} переменных $a = (a_0, a_1, \dots, a_m)$ введем два множества

$$\begin{aligned} A = \{a = (a_0, a_1, \dots, a_m) \in E^{m+1}: a_0 \geq \\ \geq J(u), a_1 \geq g_1(u), \dots, a_m \geq g_m(u), u \in U_0\}, \\ B = \{b = (b_0, b_1, \dots, b_m) \in E^{m+1}: b_0 < J_*, b_1 < 0, \dots, b_m < 0\}. \end{aligned}$$

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 $a_0 \geq J(u)$, $a_1 \geq g_1(u)$, $\dots$, $a_m \geq g_m(u)$. Возможно, что $u \in U$. Тогда $a_0 \geq J(u) \geq J_*$ и заведомо $a \notin B$. Если же $u \in U_0 \setminus U$, то найдется номер i , $1 \leq i \leq m$, такой, что $g_i(u) > 0$. Тогда $a_i \geq g_i(u) > 0$ и снова $a \notin B$. Итак, $A \cap B = \emptyset$.

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 $a_0 \geq J(u)$, $c_0 \geq J(v)$, $a_i \geq g_i(u)$, $c_i \geq g_i(v)$, $i = \overline{1, m}$. 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:

$$J(u_\alpha) \leq \alpha J(u) + (1 - \alpha)J(v) \leq \alpha a_0 + (1 - \alpha)c_0, \quad g_i(u_\alpha) \leq \alpha g_i(u) + (1 - \alpha)g_i(v) \leq \alpha a_i + (1 - \alpha)c_i, \quad i = \overline{1, m}.$$

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 $c = (\lambda_0^*, \lambda_1^*, \dots, \lambda_m^*) \neq 0$, отделяющая A и B , а также A и $\bar{B} = \{b = (b_0, \dots, b_m) \in E^{m+1} : b_0 \leq J_*, b_1 \leq 0, \dots, b_m \leq 0\}$. Это значит, что

$$\langle c, b \rangle = \sum_{i=0}^m \lambda_i^* b_i \leq \gamma \leq \langle c, a \rangle = \sum_{i=0}^m \lambda_i^* a_i \quad (15)$$

при всех $a \in A$, $b \in \bar{B}$. Заметим, что точка $y = (J_*, 0, \dots, 0) \in A \cap \bar{B}$. В самом деле, возьмем какую-либо точку $u_* \in U_*$. Тогда $J(u_*) = J_*$, $g_i(u_*) \leq 0$, $i = \overline{1, m}$, что означает $y \in A$. Включение $y \in \bar{B}$ 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

$$\lambda_0^* b_0 + \sum_{i=1}^m \lambda_i^* b_i \leq \lambda_0^* J_* \leq \lambda_0^* a_0 + \sum_{i=1}^m \lambda_i^* a_i, \quad a \in A, \quad b \in \bar{B}. \quad (16)$$

Возьмем точку $b = (J_* - 1, 0, \dots, 0) \in \bar{B}$. Из левого неравенства (16) получим $\lambda_0^* (J_* - 1) \leq \lambda_0^* J_*$, откуда $\lambda_0^* \geq 0$. Далее, беря $b = (J_*, 0, \dots, 0, -1, 0, \dots, 0)$, из левого неравенства (16) имеем $\lambda_0^* J_* - \lambda_1^* \leq \lambda_0^* J_*$, т. е. $\lambda_1^* \geq 0$, $i = \overline{1, m}$. Таким образом, показано, что $\lambda^* = (\lambda_1^*, \dots, \lambda_m^*) \geq 0$, $\lambda_0^* \geq 0$.

Далее, возьмем произвольную точку $u_* \in U_*$. Тогда $a = (J(u_*) = J_*, 0, \dots, 0, g_1(u_*), 0, \dots, 0) \in A \cap \bar{B}$. Подставляя эту точку в левое и правое неравенства (16), имеем $\lambda_0^* J_* + \lambda_i^* g_i(u_*) \leq \lambda_0^* J_* \leq \lambda_0^* J_* + \lambda_i^* g_i(u_*)$, откуда $\lambda_i^* g_i(u_*) \leq 0 \leq \lambda_i^* g_i(u_*)$ или $\lambda_i^* g_i(u_*) = 0$, $i = \overline{1, m}$. Равенства (7) доказаны.

Покажем, что $\lambda_0^* > 0$. В самом деле, в (16) подставим $a = (J(\bar{u}), g_1(\bar{u}), \dots, g_m(\bar{u})) \in A$, где $\bar{u}$ взято из (14). Получим

$$\lambda_0^* J_* \leq \lambda_0^* J(\bar{u}) + \sum_{i=1}^m \lambda_i^* g_i(\bar{u}).$$

Допустим, что $\lambda_0^* = 0$. Поскольку не все λ_i^* , $i = \overline{0, m}$, равны нулю, то $\lambda^* \neq 0$. Тогда из предыдущего неравенства при $\lambda_0^* = 0$ с учетом условия (14) имеем $0 \leq \sum_{i=1}^m \lambda_i^* g_i(\bar{u}) < 0$. 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 $a = (J(u), g_1(u), \dots, g_m(u)) \in A$. Подставим эту точку в правое неравенство (16). С учетом того, что $\lambda_0^* = 1$, получим $J_* \leq J(u) + \sum_{i=1}^m \lambda_i^* g_i(u) = L(u, \lambda^*)$, $u \in U_0$. 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_*, λ^*) — седловая точка. Теорема 2 доказана.

3. Приведенный выше пример 1 показывает, что без дополнительного требования регулярности множества теорема 2, вообще говоря, неверна. Однако если все функции $g_i(u)$, задающие множество (2), линейны, 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

$$K = \{e \in E^n: \langle c_i, e \rangle \leq 0, \quad i = \overline{1, m}; \quad \langle c_i, e \rangle < 0, \\ i = \overline{m+1, p}; \quad \langle c_i, e \rangle = 0, \quad i = \overline{p+1, s}\}, \quad (17)$$

где $c_1, \dots, c_s$ — некоторые векторы из E^n . Тогда для того, чтобы некоторый вектор $c \in E^n$ удовлетворял неравенству

$$\langle c, e \rangle \geq 0 \quad \text{при всех } e \in K, \quad (18)$$

необходимо и достаточно, чтобы существовали числа $\lambda_1, \dots, \lambda_s$, причем $\lambda_1 \geq 0, \dots, \lambda_p \geq 0$, такие, что

$$c = -\lambda_1 c_1 - \lambda_2 c_2 - \dots - \lambda_s c_s. \quad (19)$$

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

$$K^* = \left\{ c \in E^n: c = - \sum_{i=1}^s \lambda_i c_i, \quad \lambda_1 \geq 0, \dots, \lambda_p \geq 0 \right\}. \quad (19)$$

Заметим, что в (17) не исключаются случаи, когда какие-либо виды ограничений (ограничения типа нестрогого или стро-

гого 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 следующая

Л е м м а 2. Пусть $a_1, \dots, a_p$ — некоторое конечное число векторов из E^n . Тогда множество

$$Q = \left\{ a \in E^n: a = \sum_{i=1}^p \alpha_i a_i, \alpha_i \geq 0, i = \overline{1, p} \right\}$$

является выпуклым замкнутым конусом.

Доказательство. Если $a \in Q$, то $\lambda a = \sum_{i=1}^p \lambda \alpha_i a_i$, $\lambda \alpha_i \geq 0$, $i = \overline{1, p}$, при любом $\lambda > 0$. Это значит, что Q — конус. Далее,

пусть a, b — две произвольные точки из Q , т. е. $a = \sum_{i=1}^p \alpha_i a_i$,

$b = \sum_{i=1}^p \mu_i a_i$, $\alpha_i \geq 0$, $\mu_i \geq 0$, $i = \overline{1, p}$. Для всех $\alpha \in [0, 1]$ тогда

$$\alpha a + (1-\alpha)b = \sum_{i=1}^p [\alpha \alpha_i + (1-\alpha)\mu_i] a_i, \alpha \alpha_i + (1-\alpha)\mu_i \geq 0, i = \overline{1, p},$$

так что Q — выпуклый конус. Кстати, тогда из $a, b \in Q$ следует, что $a + b \in Q$. 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 $a/\alpha, b/(1-\alpha) \in Q$.

Замкнутость Q докажем индукцией по числу образующих векторов $a_1, a_2, \dots, a_p$. При $p = 1$ $Q = \{a \in E^n: a = \alpha a_1, \alpha \geq 0\}$ — полупрямая (луч) — замкнутое множество. Пусть известно, что конус с любыми $p - 1$ образующими замкнут. Покажем, что тогда конус Q с p образующими $a_1, a_2, \dots, a_p$ также замкнут. Рассмотрим два случая.

1) Конус Q наряду с векторами $a_1, a_2, \dots, a_p$ содержит также и векторы $-a_1, -a_2, \dots, -a_p$. Тогда наряду с точками

$$a = \sum_{i=1}^p \alpha_i a_i, \alpha_i \geq 0, i = \overline{1, p}, \text{ выпуклый конус } Q \text{ содержит все}$$

$$\text{точки } b = \sum_{i=1}^p \beta_i (-a_i), \beta_i \geq 0, i = \overline{1, p}, \text{ а также точки } a + b.$$

Покажем, что в рассматриваемом случае Q является подпространством (линейной оболочкой), натянутым на векторы $a_1, \dots$

$\dots, a_p$. В самом деле, пусть $d = \sum_{i=1}^p \gamma_i a_i$, где $\gamma_i, i = \overline{1, p}$, — произвольные числа. Положим $\alpha_i = \max \{0; \gamma_i\}$, $\beta_i = \max \{0; -\gamma_i\}$.

Тогда $\gamma_i = \alpha_i - \beta_i$, $i = \overline{1, p}$. Поэтому $d = \sum_{i=1}^p \gamma_i a_i = \sum_{i=1}^p \alpha_i a_i +$

$+\sum_{i=1}^p \beta_i (-a_i) \in Q$ при всех действительных γ_i , $i = \overline{1, p}$. Таким образом, Q — подпространство, натянутое на векторы $a_1, a_2, \dots, a_p$, размерность которого не превышает числа $\min \{p; n\}$. Но в конечномерном пространстве E^n любое подпространство замкнуто [63]. Следовательно, Q замкнут.

2) Хотя бы один из векторов $(-a_1), \dots, (-a_p)$ не принадлежит Q . Пусть для определенности $(-a_p) \notin Q$. Обозначим $Q_{p-1} = \left\{ b \in E^n: b = \sum_{i=1}^{p-1} \alpha_i a_i, \alpha_i \geq 0, i = \overline{1, p-1} \right\}$ — это конус, порожденный векторами $a_1, \dots, a_{p-1}$. По предположению индукции конус Q_{p-1} замкнут. Далее, из представления $a = \sum_{i=1}^{p-1} \alpha_i a_i + \alpha a_p$, $\alpha_i \geq 0, i = \overline{1, p-1}, \alpha \geq 0$, следует, что

$$Q = Q_{p-1} + \alpha a_p, \quad \alpha \geq 0. \quad (20)$$

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 $\mu_m \geq 0$ таких, что $d_m = b_m + \mu_m a_p$, $m = 1, 2, \dots$. Покажем, что $\{\mu_m\}$ ограничена сверху. Допустим, что $\{\mu_m\}$ неограничена сверху. Тогда существует подпоследовательность $\{\mu_{m_k}\} \rightarrow +\infty$. Поскольку $\{d_m\}$, как сходящаяся последовательность, ограничена, то $d_{m_k}/\mu_{m_k} \rightarrow 0$ при $k \rightarrow \infty$. А тогда $\{b_{m_k}/\mu_{m_k} = (d_{m_k}/\mu_{m_k}) - a_p\}$ имеет предел при $k \rightarrow \infty$, 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 $0 \leq \mu_m \leq \text{const}$, $m = 1, 2, \dots$. Тогда существует подпоследовательность $\{\mu_{m_k}\}$, сходящаяся к некоторому числу $\mu \geq 0$. Из равенств $b_{m_k} = d_{m_k} - \mu_{m_k} a_p$, $k = 1, 2, \dots$, при $k \rightarrow \infty$ следует, что предел $b = \lim_{k \rightarrow \infty} b_{m_k}$ 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 $b \in Q_{p-1}$, $\mu \geq 0$. Из (20) тогда следует, что $d \in Q$. Замкнутость Q доказана.

Доказательство теоремы 3. Введем множество

$$Q = \left\{ c \in E^n: c = -\sum_{i=1}^s \lambda_i c_i, \lambda_i \geq 0, \dots, \lambda_p \geq 0 \right\}. \quad (21)$$

Заметим, что Q — конус, порожденный векторами $-c_1, \dots, -c_p, -c_{p+1}, \dots, -c_s, c_{p+1}, \dots, c_s$. В самом деле, с одной стороны,

все точки $c = \sum_{i=1}^p \lambda_i (-c_i) + \sum_{i=p+1}^s \alpha_i (-c_i) + \sum_{i=p+1}^s \beta_i c_i$, $\lambda_i \geq 0$, $i = \overline{1, p}$; $\alpha_i \geq 0$, $\beta_i \geq 0$, $i = \overline{p+1, s}$, принадлежат Q . С другой стороны, любая точка $c = -\lambda_1 c_1 - \dots - \lambda_s c_s$ представима в виде предыдущего равенства при $\alpha_i = \max \{0; \lambda_i\}$, $\beta_i = \max \{0; -\lambda_i\}$, $i = \overline{p+1, s}$, так как $\lambda_i = \alpha_i - \beta_i$. В силу леммы 2 тогда множество (21) является выпуклым замкнутым множеством.

Для доказательства теоремы 3 достаточно установить, что

$K^* = Q$. Возьмем произвольный вектор $c \in Q$, т. е. $c = -\sum_{i=1}^s \lambda_i c_i$, $\lambda_i \geq 0$, ..., $\lambda_p \geq 0$. Тогда для любого $e \in K$ имеем $\langle c, e \rangle = -\sum_{i=1}^s \lambda_i \langle c_i, e \rangle \geq 0$. 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 c . 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 $\langle d, c \rangle > \langle c, y \rangle$ для всех $c \in Q$. Согласно (21) это значит, что

$$\left\langle d, -\sum_{j=1}^s \lambda_j c_j \right\rangle = -\sum_{j=1}^s \lambda_j \langle c_j, d \rangle > \langle d, y \rangle \quad (22)$$

при всех $\lambda_1, \dots, \lambda_s$, лишь бы $\lambda_i \geq 0$, ..., $\lambda_p \geq 0$. Покажем, что тогда $d \in K$ — замыкание K . Зафиксируем некоторый номер примем $\lambda_i = 0$ при всех $j \neq i$, $\lambda_i = t \langle c_i, d \rangle$, $t > 0$. Получим $-\lambda_i \langle c_i, d \rangle > \langle d, y \rangle$ для любого $\lambda_i \geq 0$. Отсюда, деля на $\lambda_i > 0$ и устремляя $\lambda_i \rightarrow +\infty$, имеем $-\langle c_i, d \rangle \geq 0$ или $\langle c_i, d \rangle \leq 0$, $i = \overline{1, p}$. Далее, зафиксируем номер i , $p+1 \leq i \leq s$, 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 $t \rightarrow +\infty$, будем иметь $-|\langle c_i, d \rangle|^2 \geq 0$ или $\langle c_i, d \rangle = 0$, $i = \overline{p+1, s}$. Таким образом, показано, что $d \in K$.

Теперь вспомним, что $y \in K^*$. Это значит, что $\langle y, e \rangle \geq 0$ для всех $e \in K$. Отсюда с помощью предельного перехода нетрудно получить, что $\langle y, e \rangle \geq 0$ для всех $e \in K$. В частности, для $d \in K$ имеем $\langle y, d \rangle \geq 0$. Но, с другой стороны, если в (22) принять $\lambda_i = 0$, $i = \overline{1, s}$, 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

$$U = \{u = (u^1, u^2, \dots, u^n) \in E^n: u^j \geq 0, j \in I; g_i(u) = \langle a_i, u \rangle - b^i \leq 0, i = \overline{1, m}; g_i(u) = \langle a_i, u \rangle - b^i = 0, i = \overline{m+1, s}\}, \quad (23)$$

где I — заданное множество индексов из $\{1, 2, \dots, n\}$ (возможно, что $I = \emptyset$ или $I = \{1, 2, \dots, n\}$); $a_i \in E^n$ — заданные векторы, b^i — заданные числа.

Теорема 4. Пусть функция $J(u)$ выпукла на $U_0 = \{u = (u^1, \dots, u^n): u^j \geq 0, j \in I\}$ и $J(u) \in C^1(U_0)$. Пусть множество U_* точек минимума $J(u)$ на множестве (23) непусто. Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_s^*) \in \Lambda_0 = \{\lambda = (\lambda_1, \dots, \lambda_s) \in E^s: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\}$ такие, что пара (u_*, λ^*) образует седловую точку функции Лагранжа на множестве $U_0 \times \Lambda_0$.

Из этой теоремы следует, что в любой задаче линейного программирования, имеющей решение, функция Лагранжа всегда имеет седловую точку.

Доказательство. Заметим, что согласно теореме 2.11 множество (23) выпукло. Возьмем любую точку $u_* \in U_*$. Рассмотрим возможные направления из точки u_* . Согласно определению 2.3 направление $e = (e^1, \dots, e^n) \neq 0$ называется возможным в точке u_* , если существует число $t_0 > 0$ такое, что $u = u_* + te = (u_*^1 + te^1, \dots, u_*^n + te^n) \in U$ или

$$u_*^j + te^j \geq 0, j \in I; \langle a_i, u_* + te \rangle \leq b^i, i = \overline{1, m}; \quad (24)$$

$$\langle a_i, u_* + te \rangle = b^i, i = \overline{m+1, s},$$

при всех $t, 0 \leq t \leq t_0$. Введем множества индексов

$$I_1 = \{j: j \in I, u_*^j = 0\}, \quad I_2 = \{i: 1 \leq i \leq m, \langle a_i, u_* \rangle = b^i\}.$$

Пусть e — произвольное возможное направление в точке u_* . Тогда из (24) с учетом $u_* \in U_* \subseteq U$ имеем

$$e^j \geq 0, j \in I_1; \langle a_i, e \rangle \leq 0, i \in I_2; \langle a_i, e \rangle = 0, i = \overline{m+1, s}.$$

Это значит, что всякое возможное направление в точке u_* принадлежит конусу

$$K = \{e = (e^1, \dots, e^n) \in E^n: e^j = \langle e_j, e \rangle \geq 0, j \in I_1, \langle a_i, e \rangle \leq 0, i \in I_2; \langle a_i, e \rangle = 0, i = \overline{m+1, s}\}, \quad (25)$$

где $e_j = (0, \dots, 0, 1, 0, \dots, 0)$ — j -я строка единичной матрицы.

Верно и обратное: если $e \in K$, то e — возможное направление в точке u_* . В самом деле, если $j \in I_1^*$, то $u_j^* = 0$ и $u_j^* + te^j \geq 0$ при всех $e \in K$ и $t \geq 0$, а если $j \in I \setminus I_1^*$, т. е. $u_j^* > 0$, то $u_j^* + te^j \geq 0$ при $0 \leq t \leq t_0$ для достаточно малого $t_0 > 0$. Далее, если $i \in I_2^*$, то $\langle a_i, u_* + te \rangle = b^i + t \langle a_i, e \rangle \leq b^i$ для всех $e \in K$ и $t \geq 0$, а если $i \in \{1, 2, \dots, m\} \setminus I_2^*$, то $\langle a_i, u_* \rangle < b^i$ и можно добиться выполнения $\langle a_i, u_* + te \rangle \leq b^i$, $0 \leq t \leq t_0$, при достаточно малом $t_0 > 0$. Наконец, если $e \in K$ и $m+1 \leq i \leq s$, то $\langle a_i, u_* + te \rangle = b^i$ при всех t .

Таким образом, показано, что в точке $u_* \in U_*$ множество возможных направлений представляет собой конус K , определяемый условиями (25).

Согласно теореме 2.3 для того, чтобы $u_* \in U_*$, необходимо и достаточно выполнения неравенства

$$\langle J'(u_*), u - u_* \rangle \geq 0 \quad \text{при всех } u \in U. \quad (26)$$

Возьмем любое $e \in K$. Тогда $u = u_* + te \in U$, $0 \leq t \leq t_0$, $t_0 > 0$. Подставим такую точку u в (26). Получим $\langle J'(u_*), e \rangle t \geq 0$ или $\langle J'(u_*), e \rangle \geq 0$ при всех $e \in K$. Это значит, что $J'(u_*) \in K^*$. По теореме 3 тогда найдутся числа $\mu_j^* \geq 0$, $j \in I_1^*$; $\lambda_i^* \geq 0$, $i \in I_2^*$; $\lambda_{m+1}^*, \dots, \lambda_s^*$ такие, что

$$J'(u_*) = - \sum_{j \in I_1^*} \mu_j^* (-e_j) - \sum_{i \in I_2^*} \lambda_i^* a_i - \sum_{i=m+1}^s \lambda_i^* a_i. \quad (27)$$

Если доопределим $\lambda_i^* = 0$ при $i \in \{1, 2, \dots, m\} \setminus I_2^*$, то получим точку $\lambda^* = (\lambda_1^*, \dots, \lambda_s^*) \in \Lambda_0$. Отсюда, учитывая определение множества I_2^* и условие $u_* \in U_* \subseteq U$, имеем

$$\lambda_i^* (\langle a_i, u_* \rangle - b^i) = \lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, s}, \quad (28)$$

а равенство (27) можем переписать в виде

$$J'(u_*) + \sum_{i=1}^s \lambda_i^* a_i = \sum_{j \in I_1^*} \mu_j^* e_j. \quad (29)$$

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 $J(u) - J(u_*) \geq \langle J'(u_*), u - u_* \rangle$, $u \in U$ (см. теорему 2.2), определение множества I_1^* , условие $\mu_j^* \geq 0$,

$j \in I_1^*$ и равенство (29), для каждого $u \in U_0$ получим

$$\begin{aligned} L(u, \lambda^*) - L(u_*, \lambda^*) &= J(u) - J(u_*) + \sum_{i=1}^s \lambda_i^* \langle a_i, u - u_* \rangle \geqslant \\ &\geqslant \left\langle J'(u_*) + \sum_{i=1}^s \lambda_i^* a_i, u - u_* \right\rangle = \sum_{j \in I_1^*} \mu_j^* \langle e_j, u - u_* \rangle = \\ &= \sum_{j \in I_1^*} \mu_j^* (u^j - u_*^j) = \sum_{j \in I_1^*} \mu_j^* u^j \geqslant 0 \end{aligned}$$

или

$$L(u_*, \lambda^*) \leqslant L(u, \lambda^*) \quad \text{при всех } u \in U_0.$$

Отсюда и из (28) с помощью леммы 1 заключаем, что (u_*, λ^*) — седловая точка функции Лагранжа. Теорема 4 доказана.

5. Наконец, приведем следующий более общий вариант теоремы Куна — Таккера.

Теорема 5. Пусть $\overline{U_0}$ — выпуклое множество из E^n , функции $J(u)$, $g_i(u)$, $i = \overline{1, m}$, выпуклы на U_0 ; $g_i(u) = \langle a_i, u \rangle - b^i$, $i = \overline{m+1, s}$ — линейные функции. Пусть множество U_* точек минимума функции $J(u)$ на множестве

$$U = \{u \in E^n: u \in U_0, \quad g_i(u) \leqslant 0, \quad i = \overline{1, m};$$

$$g_i(u) = \langle a_i, u \rangle - b^i \leqslant 0, \quad i = \overline{m+1, p};$$

$$g_i(u) = \langle a_i, u \rangle - b^i = 0, \quad i = \overline{p+1, s}\}$$

непусто. Кроме того, пусть существует точка $\bar{u} \in \text{ri } U_0 \cap U$ такая, что $g_i(\bar{u}) < 0$, $i = \overline{1, m}$. Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_s^*) \in \in \Lambda_0 = \{\lambda \in E^s: \lambda_i \geqslant 0, \dots, \lambda_p \geqslant 0\}$ 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 опирается на более тонкие теоремы отделимости, чем теоремы 5.1—5.3, и здесь не приводится (см., например, [207], § 28).

Пример 1 показывает, что нарушение условия $\text{ri } U_0 \cap U \neq \emptyset$ может привести к отсутствию седловой точки. С другой стороны, как показывает следующий пример, это условие не является необходимым для существования седловой точки в задачах выпуклого программирования.

Пример 2. Пусть $U_0 = \{u \in E^1: u \geqslant 0\}$, $J(u) = u$, $g(u) = u^2$, $U = \{u: u \in U_0, g(u) \leqslant 0\}$. Здесь U_0 выпукло, функции $J(u)$, $g(u)$ выпуклы на U_0 . Множество U состоит из единствен-

ной точки $u=0$, так что $J_*=0$, $U_*=\{0\}$. Функция Лагранжа $L(u, \lambda)=u+\lambda u^2$, $u \geq 0$, $\lambda \geq 0$, имеет седловую точку ($u_*=0$, $\lambda^*=0$), хотя $U_0 \cap U = \emptyset$. Этот же пример показывает, что множители Лагранжа, вообще говоря, определяются неоднозначно — здесь точки ($u_*=0$, λ^*) при любом $\lambda^* \geq 0$ 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.

Пример 3. Пусть $U_0 = \{u \in E^1: u \leq 1\}$, $J(u) = u^3$, $g(u) = -u^3 - 1$, $U = \{u: u \in U_0, g(u) \leq 0\}$. Здесь U_0 выпукло, но функции $J(u)$, $g(u)$ не являются выпуклыми на U_0 . Множество U представляет собой отрезок $-1 \leq u \leq 1$, 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 $U_0 \times \Lambda_0$, $\Lambda_0 = \{\lambda \in E^1: \lambda \geq 0\}$.

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

$$\chi(u) = \sup_{\lambda \in \Lambda_0} L(u, \lambda), \quad u \in U_0. \quad (30)$$

Заметим, что если $u \in U$, то $\sum_{i=1}^s \lambda_i g_i(u) \leq 0$ 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 $1 \leq i \leq m$ и $g_i(u) > 0$, либо $m+1 \leq i \leq s$ 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

$$\chi(u) = \begin{cases} J(u) & \text{при всех } u \in U, \\ +\infty & \text{при всех } u \in U_0 \setminus U. \end{cases}$$

Отсюда ясно, что $\inf_{U_0} \chi(u) = \inf_U J(u) = J_*$, и задачу (1), (2) можно переписать в равносильном виде

$$\chi(u) \rightarrow \inf; \quad u \in U_0. \quad (31)$$

Как и выше, в задаче (1), (2) будем предполагать, что $J_* > -\infty$, $U_* \neq \emptyset$. Тогда задача (31) будет иметь то же множество решений U_* с тем же минимальным значением J_* , т. е.

$$\inf_U \chi(u) = J_*, \quad U_* = \{u: u \in U_0, \chi(u) = J_*\}. \quad (32)$$

Наряду с функцией (30) введем функцию

$$\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda), \quad \lambda \in \Lambda_0, \quad (33)$$

и рассмотрим задачу

$$\psi(\lambda) \rightarrow \sup; \quad \lambda \in \Lambda_0. \quad (34)$$

Задачу (34) принято называть *двойственной задачей* к задаче (31) (или к исходной, основной задаче (1), (2)), а множители Лагранжа — переменные $\lambda = (\lambda_1, \dots, \lambda_s)$ — называют двойственными переменными в отличие от исходных, основных переменных $u = (u^1, \dots, u^n)$. Обозначим

$$\sup_{\lambda} \psi(\lambda) = \psi^*, \quad \Lambda^* = \{\lambda \in \Lambda_0: \psi(\lambda) = \psi^*\}. \quad (35)$$

Оказывается, задачи (31)' и (34) тесно связаны между собой. Прежде всего, всегда выполняются неравенства

$$\psi(\lambda) \leq \psi^* \leq J_* \leq \chi(u), \quad u \in U_0, \quad \lambda \in \Lambda_0. \quad (36)$$

В самом деле, $\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) \leq L(u, \lambda)$ при всех $\lambda \in \Lambda_0$ и $u \in U_0$. Поэтому $\psi^* = \sup_{\lambda \in \Lambda_0} \psi(\lambda) \leq \sup_{\lambda \in \Lambda_0} L(u, \lambda) = \chi(u)$ для любого $u \in U_0$. Переходя к нижней грани по $u \in U_0$ в этом неравенстве, получим $\psi^* \leq J_*$, откуда следуют неравенства (36).

Интересно выяснить, когда $\psi^* = J_*$, и обе задачи (31) и (34) имеют решение, т. е.

$$U_* \neq \emptyset, \quad \Lambda^* \neq \emptyset, \quad J_* = \psi^*. \quad (37)$$

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

$$\psi^* = \psi(\lambda^*) = \inf_{u \in U_0} L(u, \lambda^*) \leq L(u_*, \lambda^*) = \sup_{\lambda \in \Lambda_0} L(u_*, \lambda) = \chi(u_*) = J_*.$$

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 $L(u_*, \lambda) \leq L(u_*, \lambda^*)$, $\lambda \in \Lambda_0$. Отсюда имеем

$$\sup_{\lambda \in \Lambda_0} L(u_*, \lambda) = \chi(u_*) = L(u_*, \lambda^*).$$

Кроме того, $L(u_*, \lambda^*) \leq L(u, \lambda^*)$, $u \in U_0$, так что

$$L(u_*, \lambda^*) = \inf_{u \in U_0} L(u, \lambda^*) = \psi(\lambda^*),$$

откуда и из неравенств (36) следует

$$L(u_*, \lambda^*) = \psi(\lambda^*) \leq \psi^* \leq J_* \leq \chi(u_*) = L(u_*, \lambda^*),$$

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 $U_* \times \Lambda^*$. Теорема 6 доказана.

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 $(a_*, b^*) \in U_0 \times \Lambda_0$ — две 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

$$L(u_*, b^*) = L(a_*, \lambda^*) = L(u_*, \lambda^*) = L(a_*, b^*) = J_* = \psi^*.$$

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 следствия 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 4. Рассмотрим задачу из примера 1. Было показано, что $J_* = 0$, $U_* = \{0\}$. Поскольку $L(u, \lambda) = -u + \lambda u^2$, $u \geq 0$, $\lambda \geq 0$, то $\psi(\lambda) = \inf_{u \geq 0} L(u, \lambda) = -1/(4\lambda)$ при $\lambda > 0$ и $\psi(0) = -\infty$. Следовательно, $\sup_{\lambda \geq 0} \psi(\lambda) = \psi^* = 0 = J_*$, но $\Lambda^* = \emptyset$.

Однако при отсутствии седловой точки возможно строгое неравенство $\psi^* < J_*$ даже в том случае, когда $U_* \neq \emptyset$, $\Lambda^* \neq \emptyset$.

Пример 5. Пусть $J(u) = e^{-u}$, $U_0 = E^1$, $g(u) = ue^{-u}$, $U = \{u: u \in U_0, g(u) = 0\}$. Здесь множество U состоит из единственной точки $u = 0$, так что $J_* = J(0) = 1$, $U_* = \{0\}$. Поскольку $L(u, \lambda) = e^{-u} + \lambda ue^{-u} = e^{-u}(1 + \lambda u)$, $u \in E^1$, $\lambda \in \Lambda_0 = E^1$, то

$$\psi(\lambda) = \inf_{u \in E^1} L(u, \lambda) = \begin{cases} 0 & \text{при } \lambda = 0, \\ -\infty & \text{при } \lambda > 0, \\ \lambda e^{-1+1/\lambda} & \text{при } \lambda < 0, \end{cases}$$

$$\chi(u) = \sup_{\lambda \in E^1} L(u, \lambda) = \begin{cases} 1 & \text{при } u = 0, \\ +\infty & \text{при } u \neq 0. \end{cases}$$

Отсюда $\psi^* = \sup_{E^1} \psi(\lambda) = 0 = \psi(0)$, $\Lambda^* = \{0\}$; $J_* = \inf_{E^1} \chi(u) = 1 = \chi(0)$, $U_* = \{0\}$. 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

$$U(\lambda^*) = \{u: u \in U_0, L(u, \lambda^*) = L(u_*, \lambda^*)\},$$

$$\Lambda(u_*) = \{\lambda: \lambda \in \Lambda_0, L(u_*, \lambda) = L(u_*, \lambda^*)\},$$

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

$$U_* \subseteq U(\lambda^*), \quad \Lambda^* \subseteq \Lambda(u_*). \quad (38)$$

Пример 6. 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, функ-

ция $L(u, \lambda)$ линейна по λ , поэтому согласно теореме 2.7 функция $-\psi(\lambda) = \sup_{u \in U_0} (-L(u, \lambda))$ выпукла на выпуклом множестве Λ_0 . Тогда задача (34), записанная в виде

$$-\psi(\lambda) \rightarrow \inf; \quad \lambda \in \Lambda_0,$$

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 удастся выписать в явном виде.

Рассмотрим каноническую задачу линейного программирования (см. задачу (3.1.14)):

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au - b = 0\}, \quad (39)$$

где A — матрица порядка $s \times n$, $b \in E^s$, $c \in E^n$. Здесь $U_0 = \{u \in E^n: u \geq 0\}$, $\Lambda_0 = E^s$; 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

$$\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) = \begin{cases} -\langle b, \lambda \rangle & \text{при } c + A^T \lambda \geq 0, \\ -\infty & \text{при } (c + A^T \lambda)^i < 0; \lambda \in \Lambda_0 = E^s. \end{cases}$$

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 $c + A^T \lambda \geq 0$. Поэтому двойственную задачу (34) для задачи (39) можно сформулировать так: $-\langle b, \lambda \rangle \rightarrow \sup$, или

$$\langle b, \lambda \rangle \rightarrow \inf; \quad \lambda \in \Lambda = \{\lambda \in E^s: c + A^T \lambda \geq 0\}. \quad (40)$$

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 $u = (u^1, \dots, u^n)$ и переписав ограничения $c + A^T \lambda \geq 0$ в стандартном виде: $-c - A^T \lambda \leq 0$. 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 $\lambda \in \Lambda_0 = E^s$; $u \in U_0 = \{u \in E^n: u \geq 0\}$. Тогда

$$\psi_1(u) = \inf_{\lambda \in \Lambda} L_1(\lambda, u) = \begin{cases} -\langle c, u \rangle & \text{при } b - Au = 0, \\ -\infty & \text{при } b - Au \neq 0, \quad u \in U_0. \end{cases}$$

Ясно, что $\sup_{u \in U_0} \psi_1(u)$ имеет смысл искать на множестве лишь тех $u \in U_0$, для которых $b - Au = 0$. В результате придем к задаче: $-\langle c, u \rangle \rightarrow \sup$, или

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \in U_0, \quad b - Au = 0\},$$

совпадающей с исходной канонической задачей (39).

Перейдем к рассмотрению основной задачи линейного программирования (см. задачу (3.1.15)):

$$\langle c, u \rangle \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, \quad Au - b \leq 0\}, \quad (41)$$

где A — матрица порядка $m \times n$, $b \in E^m$, $c \in E^n$. Здесь $U_0 = \{u: u \geq 0\}$, $\Lambda_0 = \{\lambda \in E^m: \lambda \geq 0\}$; 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

$$\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) = \begin{cases} -\langle b, \lambda \rangle & \text{при } c + A^T \lambda \geq 0, \\ -\infty & \text{при } (c + A^T \lambda)^i < 0, \quad \lambda \in \Lambda_0 \end{cases}$$

Двойственная к (41) задача запишется в виде:

$$\langle b, \lambda \rangle \rightarrow \inf, \quad \lambda \in \Lambda = \{\lambda \in E^m: \lambda \geq 0, \quad c + A^T \lambda \geq 0\}. \quad (42)$$

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)):

$$\begin{aligned} &\langle c, u \rangle \rightarrow \inf; \\ &u \in U = \{u = (u^1, \dots, u^n): u^j \geq 0, \quad j \in I; \quad Au - b \leq 0, \quad \bar{A}u - \bar{b} = 0\}, \end{aligned} \quad (43)$$

где I — заданное множество номеров из $\{1; 2, \dots, n\}$, 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 $U_0 = \{u = (u^1, \dots, u^n): u^j \geq 0, \quad j \in I\}$. 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

$$\lambda \in \Lambda_0 = \{\lambda = (\mu, \bar{\mu}) \in E^m \times E^s: \mu \geq 0\}.$$

Функция Лагранжа имеет вид

$$\begin{aligned} L(u, \lambda) &= \langle c, u \rangle + \langle \mu, Au - b \rangle + \langle \bar{\mu}, \bar{A}u - \bar{b} \rangle = \\ &= \langle c + A^T \mu + \bar{A}^T \bar{\mu}, u \rangle - \langle b, \mu \rangle - \langle \bar{b}, \bar{\mu} \rangle, \quad u \in U_0, \quad \lambda \in \Lambda_0. \end{aligned}$$

Отсюда $\psi(\lambda) = \inf_{u \in U_0} L(u, \lambda) = -\langle b, \mu \rangle - \langle \bar{b}, \bar{\mu} \rangle$ при $(c + A^T \mu + \bar{A}^T \bar{\mu})_i \geq 0, i \in I$, и $(c + A^T \mu + \bar{A}^T \bar{\mu})_i = 0, i \notin I$, и $\psi(\lambda) = -\infty$ при остальных $\lambda = (\mu, \bar{\mu}) \in \Lambda_0$. Тогда двойственная к (43) задача запишется в виде

$$\langle b, \mu \rangle + \langle \bar{b}, \bar{\mu} \rangle \rightarrow \inf; \quad \lambda = (\mu, \bar{\mu}) \in \Lambda = \{\lambda \mu, \bar{\mu}\} \in E^m \times E^s: \mu \geq 0, \\ (c + A^T \mu + \bar{A}^T \bar{\mu})_i \geq 0, i \in I; (c + A^T \mu + \bar{A}^T \bar{\mu})_i = 0, i \notin I\}. \quad (44)$$

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. $J_* > -\infty, U_* \neq \emptyset$, 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*

$$\langle c, u_* \rangle = -\langle b, \mu^* \rangle - \langle \bar{b}, \bar{\mu}^* \rangle. \quad (45)$$

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*

$$u_*^j \geq 0, j \in I; Au_* \leq b, \bar{A}u_* = \bar{b}, \\ \mu^* \geq 0, (c + A^T \mu^* + \bar{A}^T \bar{\mu}^*)_i \geq 0, i \in I; \\ (c + A^T \mu^* + \bar{A}^T \bar{\mu}^*)_i = 0, i \notin I; \\ \mu_i^* (Au_* - b)_i = 0, i = \overline{1, m}; u_*^i (c + A^T \mu^* + \bar{A}^T \bar{\mu}^*)_i = 0, i = \overline{1, s}.$$

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 4, 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 $u = (u^1, \dots, u^n)$:

$$\begin{aligned} L_1(\lambda, u) &= \langle b, \mu \rangle + \langle \bar{b}, \bar{\mu} \rangle + \\ &+ \sum_{i \in I} u^i (-c - A^T \mu - \bar{A}^T \bar{\mu})_i + \sum_{i \notin I} u^i (-c - A^T \mu - \bar{A}^T \bar{\mu})_i = \\ &= \langle b, \mu \rangle + \langle \bar{b}, \bar{\mu} \rangle - \langle u, c + A^T \mu + \bar{A}^T \bar{\mu} \rangle, \end{aligned}$$

$$\lambda \in \Lambda_0 = \{\lambda = (\mu, \bar{\mu}) \in E^m \times E^s: \mu \geq 0\},$$

$$u \in U_0 = \{u \in E^n: u^j \geq 0, j \in I\}.$$

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.

$$L_1(\lambda^*, u) \leq L_1(\lambda^*, u_*) \leq L_1(\lambda, u_*), \quad u \in U_0, \quad \lambda \in \Lambda_0.$$

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, то равенство (45) вытекает из теоремы 6 и следствия 1 из нее. Теорема 9 доказана.

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 7. Пусть $J(u) = |u|^2 \rightarrow \inf; u \in U = \{u \in E^n: g(u) = |u|^2 - 1 \leq 0\}$. Здесь $U_0 = E^n; J_* = 0, u_* = 0; \Lambda_0 = \{\lambda \in E^1: \lambda \geq 0\}$. Функция Лагранжа $L(u, \lambda) = |u|^2 + \lambda(|u|^2 - 1) = (1 + \lambda)|u|^2 - \lambda$. Отсюда $\psi(\lambda) = \inf_{u \in E^n} L(u, \lambda) = -\lambda$ при

$\lambda + 1 \geq 0$ и $\psi(\lambda) = -\infty$ при $\lambda + 1 < 0$. Следовательно, двойственная задача имеет вид: $\psi(\lambda) = -\lambda \rightarrow \sup; \lambda \in \Lambda = \{\lambda: \lambda \in \Lambda_0, \lambda + 1 \geq 0\}$. Эта задача линейного программирования, поэтому двойственная к ней задача не может совпасть с исходной зада-

чей. Заметим, что в этой задаче ($u_* = 0$, $\lambda^* = 0$) — седловая точка.

Читателей, желающих подробнее ознакомиться с красивой и богатой результатами теорией двойственности, с различными ее приложениями, отсылаем к [1, 107, 197, 207, 220]. 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. Отсюда, в частности, происходят известные в линейном программировании двойственный симплекс-метод, метод одновременного решения прямой и двойственной задач и т. д. [9, 11, 24, 50, 72, 172, 208, 255]. В выпуклом программировании методы, основанные на идеях двойственности, в последние годы также получили значительное развитие [21, 85, 249]. На некоторых из этих методов мы остановимся ниже.

§ 9. Правило множителей Лагранжа

1. Продолжим рассмотрение задачи

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

$$U = \{u \in E^n: u \in U_0, \quad g_i(u) \leq 0, \quad i = \overline{1, m}; \\ g_i(u) = 0, \quad i = \overline{m+1, s}\}, \quad (2)$$

предполагая, что

$$J_* = \inf_U J(u) > -\infty, \quad U_* = \{u: u \in U, \quad J(u) = J_*\} \neq \emptyset.$$

В предыдущем параграфе для исследования этой задачи была введена функция Лагранжа

$$L(u, \lambda) = J(u) + \sum_{i=1}^s \lambda_i g_i(u)$$

переменных $u \in U_0$, $\lambda \in \Lambda_0 = \{\lambda = (\lambda^1, \dots, \lambda^s) \in E^s: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\}$. Для выпуклых задач при дополнительном требовании регулярности множества U (или считая все $g_i(u)$, $i = \overline{1, s}$, линейными функциями) было показано, что для того, чтобы $u_* \in U_*$, необходимо и достаточно существования множителей Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_s^*)$ таких, что

$$\lambda_1^* \geq 0, \quad \lambda_2^* \geq 0, \quad \dots, \quad \lambda_m^* \geq 0, \quad (3)$$

$$L(u_*, \lambda^*) \leq L(u, \lambda^*) \quad \text{при всех } u \in U_0, \quad (4)$$

$$\lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, s}; \quad u_* \in U \quad (5)$$

(см. теоремы 8.1, 8.2, 8.4, 8.5).

Если функции $J(u)$, $g_i(u) \in C^1(U_0)$, то согласно теореме 2.3 из условия (4) следует неравенство

$$\langle L_u(u_*, \lambda^*), u - u_* \rangle \geq 0 \quad \text{при всех } u \in U_0, \quad (6)$$

которое в случае $U_0 = E^n$ превращается в равенство

$$L_u(u_*, \lambda^*) = \frac{\partial L(u_*, \lambda^*)}{\partial u} = J'(u_*) + \sum_{i=1}^s \lambda_i^* g'_i(u_*) = 0. \quad (7)$$

Отсюда видно, что условия (3)—(5) являются обобщением правила множителей Лагранжа в невырожденной задаче (см. § 2.2). А если в (2) $m = s = 0$, т. е. ограничения $g_i(u) \leq 0$ и $g_i(u) = 0$ отсутствуют и $U = U_0$, то $L(u, \lambda) \equiv J(u)$ и условие (6) превратится в условие оптимальности из теоремы 2.3, а в случае $U = U_0 = E^n$ из (7) имеем хорошо знакомое нам равенство $J'(u_*) = 0$.

Таким образом, условия (3)—(5) содержат в себе и обобщают известные нам из § 2.2 и теоремы 2.3 условия оптимальности и являются не только необходимыми, но и достаточными для оптимальности во всех задачах выпуклого программирования с регулярным множеством (2) (или для линейных функций $g_i(u)$, $i = 1, s$).

Однако, к сожалению условия (3)—(5) получены лишь для сравнительно узкого класса задач. Как было показано в примере 8.1, этот класс содержит не все гладкие задачи выпуклого программирования. Кроме того, вытекающее из теорем 8.4, 8.5 правило множителей Лагранжа (7) для задачи

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in E^n: g_1(u) = 0, \dots, g_s(u) = 0\}$$

охватывает лишь случай линейных функций $g_i(u)$, в то время как для такой задачи правило множителей Лагранжа в § 2.2 было установлено для любых гладких функций $g_i(u)$, $i = 1, s$, без требования их линейности и даже выпуклости. Правда, в отличие от теорем 8.4, 8.5, в § 2.2 мы работали с более общей функцией Лагранжа

$$\mathcal{L}(u, \bar{\lambda}) = \lambda_0 J(u) + \sum_{i=1}^s \lambda_i g_i(u), \quad \bar{\lambda} = (\lambda_0, \lambda_1, \dots, \lambda_s), \quad (8)$$

и полученное там правило множителей Лагранжа выражало собой лишь необходимое условие оптимальности и, вообще говоря, не являлось достаточным.

2. Возникает вопрос: нельзя ли с помощью обобщенной функции Лагранжа (8) сформулировать условия оптимальности пусть в более слабой форме, чем это сделано в теоремах 8.2 и 8.4, 8.5,

но зато для более широкого класса задач (1), (2), не требуя, например, выпуклости задачи? Следующая теорема показывает, что это возможно.

Теорема 1. Пусть функции $J(u)$, $g_i(u)$, $i = \overline{1, s}$, определены на E^n и принадлежат классу $C^1(E^n)$, пусть U_0 — выпуклое множество из E^n , причем $\text{int } U_0 \neq \emptyset$. Пусть множество U_* точек минимума функции $J(u)$ на множестве (2) непусто. Тогда для каждой точки $u_* \in U_*$ необходимо существуют множители Лагранжа $\bar{\lambda}^* = (\lambda_0^*, \lambda_1^*, \dots, \lambda_s^*)$ такие, что

$$\bar{\lambda}^* \neq 0, \quad \lambda_0^* \geq 0, \quad \lambda_1^* \geq 0, \quad \dots, \quad \lambda_m^* \geq 0, \quad (9)$$

$$\langle \mathcal{L}_u(u_*, \bar{\lambda}^*), u - u_* \rangle = \left\langle \lambda_0^* J'(u_*) + \sum_{i=1}^s \lambda_i^* g'_i(u_*), u - u_* \right\rangle \geq 0 \quad (10)$$

при всех $u \in U_0$, u

$$\lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, s}. \quad (11)$$

Если $u_* \in \text{inf } U_0$, то из (10) следует

$$\mathcal{L}_u(u_*, \lambda^*) = \lambda_0^* J'(u_*) + \sum_{i=1}^s \lambda_i^* g'_i(u_*) = 0. \quad (12)$$

В теореме 1 не исключаются случаи, когда $U_0 = E^n$, или отсутствуют ограничения типа неравенств $g_i(u) \leq 0$ ($m = 0$) или типа равенств $g_i(u) = 0$ ($m = s$) или ограничения обоих указанных типов ($m = s = 0$).

Заметим, что если какие-либо $\bar{\lambda}^* = (\lambda_0^*, \dots, \lambda_s^*)$ удовлетворяют условиям (9)–(11), то величины $\mu \bar{\lambda}^* = (\mu \lambda_0^*, \dots, \mu \lambda_s^*)$ при любом $\mu > 0$ также удовлетворяют условиям (9)–(11). Это значит, что множители Лагранжа в теореме 1 определяются с точностью до постоянного положительного множителя. Поэтому при необходимости множители Лагранжа $\bar{\lambda}^*$ можно подчинить какому-либо дополнительному условию нормировки; например, условия (9) можно заменить на условия

$$|\bar{\lambda}^*|^2 = (\lambda_0^*)^2 + \dots + (\lambda_s^*)^2 = \alpha, \quad \lambda_0^* \geq 0, \quad \dots, \quad \lambda_m^* \geq 0, \quad (13)$$

где α — некоторое положительное число (в частности, можно взять $\alpha = 1$).

Если множители $\bar{\lambda}^*$, удовлетворяющие условиям (9)–(11), таковы, что $\lambda_0^* > 0$, то задачу (1), (2) называют невырожденной в точке u_* . В невырожденной задаче (1), (2) условия (9) или (13) можно заменить более простым условием

$$\lambda_0^* = 1, \quad \lambda_1^* \geq 0, \quad \dots, \quad \lambda_m^* \geq 0, \quad (13')$$

и тогда условия (9)—(11) превратятся в условия (3), (6), (5) соответственно, а равенство (12) перейдет в (7). В частности, если $m = s = 0$, $U = U_0$, что в задаче (1), (2) во введении множителей $\lambda_1, \dots, \lambda_s$ нет необходимости, и условие (9) или условие (13) можно записать в виде $\bar{\lambda}^* = \lambda_0^* = 1$, условия (11) исчезают, а равенство (10) превратится в условие $\langle J'(u_*), u - u_* \rangle \geq 0$, $u \in U$, известное нам из теоремы 2.3. Другие примеры невырожденных задач дают теоремы 8.2 и 8.4, 8.5.

Если функция $\mathcal{L}(u, \bar{\lambda}^*)$ переменной u выпукла на U_0 (например, так будет, если $J(u)$, $g_1(u)$, $\dots$, $g_m(u)$ выпуклы на U_0 , а $g_i(u) = \langle a_i, u \rangle - b^i$, $i = m+1, s$, — линейные функции), то в силу теоремы 2.3 условие (10) можно переписать в эквивалентном виде

$$\mathcal{L}(u_*, \bar{\lambda}^*) \leq \mathcal{L}(u, \bar{\lambda}^*) \text{ при всех } u \in U_0. \quad (14)$$

В невырожденной задаче из (14) при $\lambda_0^* = 1$ получим условие (4). Это значит, что в невырожденных выпуклых задачах теорема 1 обобщает теорему 8.2, и в этом случае условия (9)—(11) становятся не только необходимыми, но и достаточными для оптимальности (см. теорему 8.1 и лемму 8.1).

Невырожденность задачи (1), (2) в точке u_* гарантируется и в том случае, когда $U_0 = E^n$ и градиенты $\{g'_i(u_*), i \in I\}$, где $I = \{i: 1 \leq i \leq s, g_i(u_*) = 0\}$, линейно независимы. В самом деле, из (11) следует $\lambda_i^* = 0$ при $g_i(u_*) < 0$, и условие (12) переписывается в виде

$$\lambda_0^* J'(u_*) + \sum_{i \in I} \lambda_i^* g'_i(u_*) = 0.$$

Если бы здесь $\lambda_0^* = 0$, то в силу линейной независимости $\{g'_i(u_*), i \in I\}$ получили бы $\lambda_i^* = 0$, $i \in I$, а тогда $\lambda_i^* = 0$ при всех $i = 0, s$, что противоречит условию $\bar{\lambda}^* \neq 0$.

Задача из примера 8.1, оказывается, является вырожденной. В самом деле, пусть $J(u) = -u \rightarrow \inf$; $u \in U = \{u \in E^1: u \in U_0, g(u) \leq 0\}$, где $U_0 = \{u \in E^1: u \geq 0\}$, $g(u) = u^2$. Здесь $U = \{0\}$, поэтому $J(0) = J_* = 0$, $u_* = 0$. Так как $g(u_*) = g(0) = 0$, то условие (11) выполняется при любом выборе множителя Лагранжа. Из (10) имеем $(\lambda_0^* J'(u_*) + \lambda^* g'(u_*))(u - u_*) = (-\lambda_0^*)u \geq 0$ при всех $u \geq 0$. Это возможно только при $\lambda_0^* \leq 0$. С другой стороны, в (9) требуется $\lambda_0^* \geq 0$. Следовательно, $\lambda_0^* = 0$, т. е. рассматриваемая задача вырожденная. В качестве множителей Лагранжа здесь можно взять $\bar{\lambda}^* = (0, \lambda^*)$ при любом $\lambda^* > 0$.

Для иллюстрации теоремы 1 приведем еще несколько примеров.

Пример 1. Пусть $J(u) = J(x, y) = x + \cos y \rightarrow \inf$; $u \in U = \{u = (x, y) \in E^2, g(u) = -x \leq 0\}$. Здесь $U_0 = E^2$,

$\mathcal{L}(u, \bar{\lambda}) = \lambda_0(x + \cos y) + \lambda(-x)$, $\mathcal{L}_u = (\mathcal{L}_x, \mathcal{L}_y) = (\lambda_0 - \lambda, -\lambda_0 \sin y)$. Для определения подозрительных на минимум точек $u = (x, y)$ и соответствующих им множителей $\bar{\lambda} = (\lambda_0, \lambda)$ из (9)–(11) и условия $u \in U$ имеем систему

$$\begin{cases} \lambda_0 \geq 0, \quad \lambda \geq 0, \quad \bar{\lambda} = (\lambda_0, \lambda) \neq 0, \quad \lambda(-x) = 0, \quad -x \leq 0, \\ \mathcal{L}_x = \lambda_0 - \lambda = 0, \quad \mathcal{L}_y = -\lambda_0 \sin y = 0. \end{cases}$$

Отсюда следует, что $\lambda_0 = \lambda > 0$. Учитывая условия нормировки вида (13), можем считать $\lambda_0 = \lambda = 1$. Тогда из предыдущей системы получаем точки $u = (0, \pi k)$, $k = 0, \pm 1, \pm 2, \dots$, 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 $(0, \pi(2m+1))$, $m = 0, \pm 1, \dots$, будут точками минимума, так как $J(0, \pi(2m+1)) = -1 \leq x + \cos y$ при любых $x \geq 0$ и любых y . В точках $(0, 2\pi m)$, $m = 0, \pm 1, \dots$, 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 $J(u) = x \rightarrow \inf$; $u \in U = \{u = (x, y) \in E^2: g_1(u) = -x \leq 0, g_2(u) = x^2 - y \leq 0, g_3(u) = y - 2x^2 \leq 0\}$. Здесь $U_0 = E^2$, $\mathcal{L}(u, \bar{\lambda}) = \lambda_0 x + \lambda_1(-x) + \lambda_2(x^2 - y) + \lambda_3(y - 2x^2)$, $\mathcal{L}_u = (\mathcal{L}_x, \mathcal{L}_y) = (\lambda_0 - \lambda_1 + 2\lambda_2 x - 4\lambda_3 x; -\lambda_2 + \lambda_3)$. 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)–(11) и условия $u \in U$ имеем систему

$$\begin{cases} \lambda_0 \geq 0, \quad \lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_3 \geq 0, \quad \bar{\lambda} \neq 0, \quad -x \leq 0, \\ \phantom{\lambda_0 \geq 0, \quad \lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_3 \geq 0, \quad \bar{\lambda} \neq 0, \quad -x \leq 0,} x^2 - y \leq 0, \quad y - 2x^2 \leq 0, \\ \lambda_1(-x) = 0, \quad \lambda_2(x^2 - y) = 0, \quad \lambda_3(y - 2x^2) = 0, \\ \phantom{\lambda_0 \geq 0, \quad \lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_3 \geq 0, \quad \bar{\lambda} \neq 0, \quad -x \leq 0,} \lambda_0 - \lambda_1 + 2x(\lambda_2 - 2\lambda_3) = 0, \quad -\lambda_2 + \lambda_3 = 0. \end{cases}$$

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 $\lambda_0 = \lambda_1 \geq 0$, $\lambda_2 = \lambda_3 \geq 0$. 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 (13).

Предлагаем читателям рассмотреть примеры 2.2.2—2.2.4 и упражнения 2.2.15—2.2.17, пользуясь теоремой 1.

3. После проведенного обсуждения теоремы 1 приступим к ее доказательству. Доказательство этой теоремы, как и аналогичных теорем из предыдущего параграфа, опирается на теоремы отделимости выпуклых множеств. В данном случае будет использована теорема 5.6 — теорема Дубовицкого — Милутина об условиях пустоты пересечения выпуклых конусов.

Сначала введем понятия направления убывания функции в точке, внутреннего и касательного направления множества в точке и покажем, что множества упомянутых направлений являются конусами.

Определение 1. Вектор $e \in E^n$, $e \neq 0$, называется *направлением убывания функции* $J(u)$ в точке u_* , если существуют числа $\varepsilon_0 > 0$, $\delta > 0$ такие, что при всех $\bar{e} \in O(e, \delta) = \{\bar{e} \in E^n: |\bar{e} - e| < \delta\}$ и всех ε , $0 < \varepsilon < \varepsilon_0$, имеет место неравенство

$$J(u_* + \varepsilon \bar{e}) < J(u_*).$$

Множество всех направлений убывания функции $J(u)$ в точке u_* будем обозначать через $K_y(u_*) = K_y$.

Если u_* — точка глобального минимума функции $J(u)$ на E^n , то множество K_y пусто. Однако K_y может быть пустым и в других случаях. Например, если $u = (x, y) \in E^2$, $J(u) = 0$ при всех $u = (x, y)$, $y \neq x^2$ и $J(u) = -x^2$ при $y = x^2$, то в точке $u_* = (0, 0)$ множество $K_y = \emptyset$, хотя $u_* = (0, 0)$ не является точкой минимума $J(u)$ на E^2 .

Лемма 1. Если $K_y \neq \emptyset$, то K_y является открытым конусом.

Доказательство. Если $e \in K_y$, т. е. $J(u_* + \varepsilon \bar{e}) < J(u_*)$ при всех $\bar{e}$, ε : $|\bar{e} - e| < \delta$, $0 < \varepsilon < \varepsilon_0$, то для λe , $\lambda > 0$, можем указать числа $\delta \lambda > 0$ и $\varepsilon_0/\lambda > 0$ такие, что из $|h - \lambda e| < \delta \lambda$, $0 < t < \varepsilon_0/\lambda$, будет следовать $J(u_* + th) < J(u_*)$. Это значит, что если $e \in K_y$, то $\lambda e \in K_y$ при всех $\lambda > 0$, т. е. K_y — конус. Далее, если $e \in K_y$, то векторы $\bar{e}$ из определения 1 также являются направлениями убывания в точке u_* . В самом деле, положим $\delta_1 = \delta - |\bar{e} - e| > 0$, $\varepsilon_1 = \varepsilon_0$. Тогда для всех h , ε : $|h - \bar{e}| < \delta_1$, $0 < \varepsilon < \varepsilon_1$, имеем $|h - e| \leq |h - \bar{e}| + |\bar{e} - e| < \delta$ и, стало быть, $J(u_* + \varepsilon h) < J(u_*)$. Это значит, что любое направление убывания e принадлежит K_y вместе со своей окрестностью $O(e, \delta)$, т. е. K_y — открытое множество. Лемма 1 доказана.

Определение 2. Вектор $e \in E^n$, $e \neq 0$, называется *внутренним направлением* множества U в точке $u_* \in U$, если существуют числа $\varepsilon_0 > 0$, $\delta > 0$ такие, что при всех $\bar{e} \in O(e, \delta) = \{\bar{e} \in E^n: |\bar{e} - e| < \delta\}$ и всех ε , $0 < \varepsilon < \varepsilon_0$, имеет место включение

$$u_* + \varepsilon \bar{e} \in U.$$

Рис. 4.19.

Множество всех внутренних направлений множества U в точке u_* будем обозначать через $K_b(u_*) = K_b$.

Может случиться, что $K_b = \emptyset$. Об этом говорит следующий

Пример 3. Пусть $U = \{u = (x, y) \in E^2: x \geq 0, 0 \leq y \leq x^2\}$ (рис. 4.19). В точке $u_* = (0, 0)$ это множество не имеет внутренних направлений.

Лемма 2. Если $K_b \neq \emptyset$, то необходимо $\text{int } U \neq \emptyset$ и K_b — открытый конус. Если $u_* \in \text{int } U$, то $K_b(u_*) = E^n$.

Доказательство. Пусть $e \in K_b$, пусть числа e_0, δ взяты из определения 2. Возьмем произвольное число $\lambda > 0$. Тогда для любых $h, |h - \lambda e| < \delta \lambda$, и любых $t, 0 < t < e_0/\lambda$, имеем $u_* + th = u_* + t\lambda(h/\lambda) \in U$. Это значит, что $\lambda e \in K_b$, т. е. K_b — конус. Далее, если $e \in K_b$, то вектор $\bar{e}$ из определения 2 также принадлежит K_b . В самом деле, чтобы гарантировать условие $\bar{e} \neq 0$, число δ из определения 2 можем считать столь малым, что $0 < \delta < |e|$. Далее, для любых $h, |h - \bar{e}| < \delta - |\bar{e} - e|$, имеем $|h - e| \leq |h - \bar{e}| + |\bar{e} - e| < \delta$, поэтому $u_* + eh \in U$ при всех $e, 0 < e < e_0$. Это означает, что $\bar{e} \in K_b$, т. е. $O(e, \delta) \in K_b$ — открытый конус.

Покажем, что если $K_b \neq \emptyset$, то $\text{int } U \neq \emptyset$. Пусть $e \in K_b$, а числа e_0, δ взяты из определения 2. Убедимся в том, что точка $v = u_* + ee \in \text{int } U$ при всех $e, 0 < e < e_0$. Рассмотрим окрестность этой точки $O(v, e\delta) = \{u \in E^n: |u - v| < e\delta\}$. Возьмем произвольную точку $u \in O(v, e\delta)$ и положим $\bar{e} = e + (u - v)/e$. Имеем $|\bar{e} - e| = |u - v|/e < e\delta/e = \delta$. Из того, что $e \in K_b$, следует $u = v + (u - v) = u_* + ee + (u - v) = u_* + e\bar{e} \in U$. Это значит, что $O(v, e\delta) \subset U$, т. е. $v = u_* + ee \in \text{int } U$.

Теперь пусть $u_* \in \text{int } U$. Возьмем произвольный вектор $e \in E^n, e \neq 0$. Из того, что $u_* \in \text{int } U$, следует, что $O(u_*, \delta) = \{u \in E^n: |u - u_*| < \delta\} \subset U$ при достаточно малом $\delta > 0$. Пусть $|\bar{e} - e| < \delta$ и $e_0 = \delta(\delta + |e|)^{-1}$. Тогда $|u_* + e\bar{e} - u_*| = e|\bar{e} - e| \leq e(|\bar{e} - e| + |e|) < e(\delta + |e|) < \delta$ для всех $e, 0 < e < e_0$, т. е. $u_* + e\bar{e} \in O(u_*, \delta) \subset U, 0 < e < e_0$. Это значит, что $e \in K_b$. В силу произвольности $e \in E^n$ отсюда заключаем, что $K_b = E^n$. Лемма 2 доказана.

Попутно мы показали, что если $e \in K_b(u_*)$, то $u_* + ee \in \text{int } U$ при всех $e, 0 < e < e_0$, где e_0 — достаточно малое положительное число. Это обстоятельство нашло отражение во введенном в определении 2 термине «внутреннее направление». Заметим, что из того, что $\text{int } U \neq \emptyset$, не следует, что $K_b(u_*) \neq \emptyset$ при всех $u_* \in U$. В примере 3 множество U имеет непустую внутренность, но в точке $u_* = (0, 0)$ конус K_b пуст.

Всякое внутреннее направление в точке u_* , очевидно, является возможным направлением в этой точке (см. определение 2.3). Отличие внутреннего направления от возможного в том, что всякое внутреннее направление имеет окрестность, также состоящую из внутренних направлений, а возможное направление такой окрестности может не иметь. В примере 3 $e_1 = (1, 0)$ — возможное направление в точке $u_* = (0, 0)$, но оно не является внутренним (рис. 4.19).

Обращаем внимание читателя на то, что в литературе вместо введенного в определении 2 термина «внутреннее направление» как синоним часто употребляется термин «возможное направление». Однако такое употребление термина «возможное направление» может привести к путанице, так как в литературе не менее широко используется определение 2.3 понятия возможного направления. Кроме того, мне кажется, принятая в определениях 2 и 2.3 терминология лучше отражает суть вводимых в них понятий.

Определение 3. Вектор $e \in E^n$ называется *касательным направлением* к множеству U в точке $u_* \in U$, если существуют число $e_0 > 0$ и функция $r(e) = (r_1(e), \dots, r_n(e))$, $0 \leq e < e_0$, такие, что $r(0) = 0$, $\lim_{e \rightarrow +0} r(e)/e = 0$ и $u_* + ee + r(e) \in U$ при всех $e, 0 \leq e < e_0$. Множество всех касательных направлений множества U в точке u_* будем обозначать через $K_k(u_*) = K_k$.

Заметим, что всегда $K_k \neq \emptyset$, так как при $e = 0, r(e) = 0, e \geq 0$, имеем $u_* + e \cdot 0 + 0 \in U$ при всех $e \geq 0$, т. е. $e = 0 \in K_k$.

Лемма 3. Множество K_k — замкнутый конус.

Доказательство. Пусть $e \in K_k$, т. е. $u_* + ee + r(e) \in U$ при всех $0 < e < e_0$, где $\lim_{e \rightarrow +0} r(e)/e = 0$. Тогда для λe при любом $\lambda > 0$ можем указать число $e_1 = e_0/\lambda$ и функцию $r_1(e) = r(\lambda e)$ такие, что $u_* + t(\lambda e) + r_1(t) =$

$= u_* + (t\lambda)e + r(t\lambda) \in U$ для всех t , $0 < t < \varepsilon_1 = \varepsilon_0/\lambda$, причем $\lim_{\varepsilon \rightarrow +0} r_1(\varepsilon)/\varepsilon = 0$.

Таким образом, K_k — конус.

Докажем замкнутость K_k . Пусть $\{e_m\} \in K_k$, $|e_m - e| \rightarrow 0$ при $m \rightarrow \infty$. Тогда найдутся числа $\varepsilon_m > 0$, функции $r_m(\varepsilon)$, $0 \leq \varepsilon < \varepsilon_m$, $r_m(0) = 0$, $\lim_{\varepsilon \rightarrow +0} r_m(\varepsilon)/\varepsilon = 0$ такие, что $u_m(\varepsilon) = u_* + \varepsilon e_m + r_m(\varepsilon) \in U$ при всех ε , $0 \leq \varepsilon < \varepsilon_m$. Можем считать, что все функции $r_m(\varepsilon)$ определены при всех $\varepsilon \geq 0$, так как всегда можно положить $r_m(\varepsilon) = -\varepsilon e_m$ при $\varepsilon \geq \varepsilon_m > 0$, $m = 1, 2, \dots$. Для любого натурального числа i можно указать номер m_i и число $t_i > 0$ такие, что

$$|e_{m_i} - e| + |r_{m_i}(t)/t| < 1/i$$

для всех t , $0 < t \leq t_i$, причем можем считать, что $t_i > t_{i+1}$, $i = 1, 2, \dots$, $\lim t_i = 0$. Положим $r(t) = t(e_{m_i} - e) + r_{m_i}(t)$, $t_{i+1} < t \leq t_i$, $i = 1, 2, \dots$, $r(0) = 0$. Тогда

$$u(t) = u_* + te + r(t) = u_* + te + t(e_{m_i} - e) + r_{m_i}(t) = u_* + te_{m_i} + r_{m_i}(t) \in U$$

при t , $t_{i+1} < t \leq t_i$, для каждого $i = 1, 2, \dots$. Это значит, что $u(t) = u_* + te + r(t) \in U$ при всех t , $0 \leq t < t_i$. Кроме того, $|r(t)/t| \leq |e_{m_i} - e| + |r_{m_i}(t)/t| < 1/i$ при $t_{i+1} < t \leq t_i$, $i = 1, 2, \dots$, так что $\lim_{t \rightarrow +0} r(t)/t = 0$.

Тем самым показано, что вектор e , являющийся предельной точкой конуса K_k , принадлежит этому конусу, т. е. конус K_k замкнут. Лемма 3 доказана.

Заметим, что всякое внутреннее направление (см. определение 2) и возможное направление (см. определение 2.3) в точке являются касательными направлениями в этой точке с функцией $r(\varepsilon) \equiv 0$, $\varepsilon \geq 0$. Следующий пример показывает, что обратное неверно.

Пример 4. Пусть $U = \{u = (x, y) \in E^2; x \geq 0, x^2 \leq y \leq 2x^2\}$ (рис. 4.20). В точке $u_* = (0, 0)$ это множество имеет касательное направление, не являющееся ни внутренним, ни возможным.

4. Используя введенные выше понятия конусов направления убывания, внутреннего и касательного направлений, докажем одно вспомогательное необходимое условие оптимальности для задачи (1), (2). С этой целью переформулируем задачу (1), (2) в несколько ином виде. Введем множества

$$U_i = \{u \in E^n; g_i(u) \leq 0\}, \quad i = \overline{1, m}; \quad (15)$$

$$U_{m+1} = \{u \in E^n; g_{m+1}(u) = 0, \dots, g_s(u) = 0\}. \quad (16)$$

Тогда множество (2) представится в виде $U = \bigcap_{i=0}^{m+1} U_i$, а задачу (1), (2) можно будет кратко записать так:

$$J(u) \rightarrow \inf; \quad u \in U = \bigcap_{i=0}^{m+1} U_i. \quad (17)$$

Лемма 4. Пусть $u_* \in U$ — точка минимума функции $J(u)$ на множестве U , пусть K_y — конус направлений убывания функции $J(u)$ в точке u_* ,

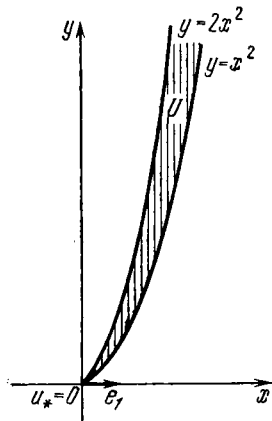


Рис. 4.20.

K_i — конус внутренних направлений множества U_i в точке u_* , $i = \overline{0, m}$; K_{m+1} — конус касательных направлений множества U_{m+1} в точке u_* . Тогда необходимо

$$K_y \cap K_0 \cap K_1 \cap \dots \cap K_{m+1} = \emptyset. \quad (18)$$

Доказательство. Допустим, что равенство (18) не имеет места. Тогда существует вектор e , принадлежащий каждому из конусов $K_y, K_0, \dots, K_{m+1}$, причем $e \neq 0$. Так как $e \in K_y$, то найдутся числа $e_y > 0$, $\delta_y > 0$ такие, что $J(u_* + e\bar{e}) < J(u_*)$ для всех $\bar{e}$, e при $|\bar{e} - e| < \delta_y$, $0 < e < e_y$. Далее, из того, что $e \in K_i$, $i = \overline{0, m}$, следует существование чисел $e_i > 0$, $\delta_i > 0$ таких, что для всех $\bar{e}$, e при $|\bar{e} - e| < \delta_i$, $0 < e < e_i$, справедливо включение $u_* + e\bar{e} \in U_i$, $i = \overline{0, m}$. Положим $\delta = \min \{\delta_y, \delta_0, \dots, \delta_m\}$, $\gamma = \min \{e_y, e_0, \dots, e_m\}$. Тогда

$$u_* + e\bar{e} \in \bigcap_{i=0}^m U_i, \quad J(u_* + e\bar{e}) < J(u_*)$$

для всех $\bar{e}$, e таких, что $|\bar{e} - e| < \delta$, $0 < e < \gamma$. Наконец, вспомним, что $e \in K_{m+1}$. Это значит, что существуют число $e_{m+1} > 0$ и функция $r(e)$, $0 < e < e_{m+1}$, такие, что $r(0) = 0$, $\lim_{e \rightarrow +0} r(e)/e = 0$ и $u(e) = u_* + ee + r(e) \in U_{m+1}$ при всех e , $0 < e < e_{m+1}$. Положим $\bar{e} = e + r(e)/e$. Можем считать e_{m+1} столь малым, что $e_{m+1} \leq \gamma$ и $|\bar{e} - e| = |r(e)/e| < \delta$. Тогда точка $u_* + e\bar{e} = u_* + ee + r(e)$ принадлежит U при всех e , $0 < e < e_{m+1}$, и, кроме того, $J(u_* + e\bar{e}) < J(u_*)$. Получили противоречие с тем, что u_* — решение задачи (17). Лемма 4 доказана.

Заметим, что лемма 4 справедлива при самых общих предположениях относительно функции $J(u)$ и множеств U_i , $i = \overline{0, m+1}$. В частности, при доказательстве этой леммы мы нигде даже не воспользовались тем, что множества U_i , $i = \overline{1, m+1}$, заданы именно в виде (15), (16). Понятно, что при таких широких предположениях относительно исходных данных задачи (17) условие (18) малоинформативно, неконструктивно и особой ценности не имеет. Однако, если задача (17) такова, что конусы $K_y, K_0, \dots, K_{m+1}$ непусты и выпуклы, то пользуясь теоремой 5.6 об отделимости, условие (18) может быть записано в достаточно конструктивной форме. Именно на этом пути ниже будет получено доказательство теоремы 1.

5. Изучим структуру конусов $K_y, K_0, \dots, K_{m+1}$ и двойственных к ним конусов $K_y^*, K_0^*, \dots, K_{m+1}^*$ в предположении, что выполнены условия теоремы 1. Заметим, что теорема 1 требует отдельного доказательства лишь для случая, когда $J'(u_*) \neq 0$, $g'_i(u_*) \neq 0$ при всех $i = \overline{1, m}$, и векторы $\{g'_i(u_*), i = \overline{m+1, s}\}$ линейно независимы. В самом деле, если $J'(u_*) = 0$, то в качестве множителей Лагранжа, удовлетворяющих условию теоремы 1, можно взять $\lambda_0^* = 1$, $\lambda_1^* = \dots = \lambda_s^* = 0$. Если же $g'_i(u_*) = 0$ при некотором i , $1 \leq i \leq m$, то можно принять $\lambda_i^* = 1$, $\lambda_j^* = 0$ при всех $j = \overline{0, s}$, $j \neq i$. Наконец, если $g'_{m+1}(u_*)$, $\dots$, $g'_s(u_*)$ линейно зависимы, то существуют числа $\alpha_{m+1}, \dots, \alpha_s$, не все равные нулю и такие, что

$$\alpha_{m+1}g'_{m+1}(u_*) + \dots + \alpha_s g'_s(u_*) = 0,$$

и в теореме 1 остается принять $\lambda_0^* = \lambda_1^* = \dots = \lambda_m^* = 0$, $\lambda_i^* = \alpha_i$, $i = \overline{m+1, s}$.

Таким образом, нам предстоит изучить структуру следующих конусов: конуса направлений убывания K_y функции $J(u)$ в точке u_* при условии

$J'(u_*) \neq 0$, конуса внутренних направлений K_0 выпуклого множества U_0 в точке u_* при условии $\text{int } U_0 \neq \emptyset$, конуса внутренних направлений K_i множества (15) в точке u_* при условии $g'_i(u_*) \neq 0$, $i = \overline{1, m}$, конуса касательных направлений K_{m+1} множества (16) в точке u_* в предположении, что $\{g'_i(u_*), i = \overline{m+1, s}\}$ линейно независимы, а также выяснить структуру конусов, двойственных к перечисленным конусам $K_y, K_0, \dots, K_{m+1}$.

Начнем с изучения конуса направлений убывания K_y и двойственного к нему конуса K_y^* .

Лемма 5. Пусть $J(u) \in C^1(E^n)$, $J'(u_*) \neq 0$. Тогда конус направлений убывания функции $J(u)$ в точке u_* является открытым выпуклым конусом и имеет вид

$$K_y = \{e \in E^n: \langle J'(u_*), e \rangle < 0\}, \quad (19)$$

а двойственным к нему является конус

$$K_y^* = \{c \in E^n: c = -\lambda J'(u_*), \lambda \geq 0\}. \quad (20)$$

Доказательство. Обозначим

$$Q = \{e \in E^n: \langle J'(u_*), e \rangle < 0\}.$$

Для получения равенства (19) достаточно установить, что $K_y = Q$. Возьмем произвольную точку $e \in K_y$. Согласно определению 1 тогда существуют числа $\delta > 0$, $\varepsilon_0 > 0$ такие, что $J(u_* + e\bar{e}) < J(u_*)$ для всех $\bar{e}, e, |\bar{e} - e| < \delta$, $0 < e < \varepsilon_0$. В частности, для $\bar{e} = e$ с помощью формулы конечных приращений (2.3.2) имеем $J(u_* + e\bar{e}) - J(u_*) = \langle J'(u_* + \theta e\bar{e}), e \rangle e < 0$ для всех $e, 0 < e < \varepsilon_0$. Отсюда, деля на $e > 0$ и устремляя $e \rightarrow 0$, получим $\langle J'(u_*), e \rangle \leq 0$ для всех $e \in K_y$. Покажем, что здесь в действительности будет строгое неравенство $\langle J'(u_*), e \rangle < 0$. В лемме 1 было установлено, что направления $\bar{e}, |\bar{e} - e| < \delta$, также принадлежат K_y и, следовательно, для них тоже $\langle J'(u_*), \bar{e} \rangle \leq 0$. Возьмем число $\gamma > 0$ столь малым, чтобы $\bar{e} = e + \gamma(e + J'(u_*))$ удовлетворяло неравенству $|\bar{e} - e| < \delta$. Отсюда

$$e = \frac{1}{1+\gamma} \bar{e} - \frac{\gamma}{1+\gamma} J'(u_*), \quad \gamma > 0.$$

Учитывая, что $\langle J'(u_*), \bar{e} \rangle \leq 0$, получим $\langle J'(u_*), e \rangle = (\langle J'(u_*), \bar{e} \rangle - \gamma |J'(u_*)|^2) / (1+\gamma) \leq -\gamma(1+\gamma)^{-1} |J'(u_*)|^2 < 0$. Таким образом, показано, что $\langle J'(u_*), e \rangle < 0$ для всех $e \in K_y$. Это значит, что $K_y \subseteq Q$.

Теперь установим, что $Q \subseteq K_y$. Возьмем произвольную точку $e \in Q$, т. е. $\langle J'(u_*), e \rangle = -\alpha < 0$. Заметим, что

$$\begin{aligned} J(u_* + e\bar{e}) - J(u_*) &= \langle J'(u_* + \theta e\bar{e}), \bar{e} \rangle e = \\ &= e [\langle J'(u_*) e \rangle + \langle J'(u_* + \theta e\bar{e}) - J'(u_*), e \rangle + \\ &+ \langle J'(u_* + \theta e\bar{e}), \bar{e} - e \rangle], \quad 0 < \theta < 1, \quad \bar{e} \in E^n, \quad e > 0. \end{aligned} \quad (21)$$

Зафиксируем некоторое число $r > 0$. В силу непрерывности $J'(u)$ найдутся числа $A > 0$, $0 < \mu \leq r$ такие, что

$$\sup_{|h| \leq r} |J'(u_* + h)| \leq A, \quad |J'(u_* + h) - J'(u_*)| < \alpha/(3|e|) \quad (22)$$

при всех $h, |h| < \mu$. Положим $\delta = \alpha/(3A)$, $\varepsilon_0 = \mu(|e| + \delta)^{-1}$. Тогда для всех $\bar{e}, e, |\bar{e} - e| < \delta$, $0 < e < \varepsilon_0$, имеем

$$|e\bar{e}| \leq e(|e| + |\bar{e} - e|) < e(|e| + \delta) < \mu \leq r.$$

Поэтому согласно (22) $|J'(u_* + \theta e \bar{e}), \bar{e} - e| \leq A|\bar{e} - e| < A\delta = \alpha/3$, $|J'(u_* + \theta e \bar{e}) - J'(u_*), e| < \alpha/(3|e|) \cdot |e| = \alpha/3$ при любых $\bar{e}, e, |\bar{e} - e| < \delta$, $0 < e < e_0$. Отсюда и из (21) получаем

$$J(u_* + e\bar{e}) - J(u_*) < e(-\alpha + \alpha/3 + \alpha/3) = -\alpha e/3 < 0$$

для всех $\bar{e}$, $|\bar{e} - e| < \delta$ и всех e , $0 < e < e_0$. Это значит, что $e \in K_y$. Тем самым включение $Q \subseteq K_y$ установлено. Сравнивая его с ранее доказанным включением $K_y \subseteq Q$, приходим к равенству (19). Выпуклость конуса (19) очевидна (см. пример 1.3); его открытость доказана в лемме 1.

Представление (20) для двойственного конуса следует из теоремы 8.3 — теоремы Фаркаша.

Перейдем к рассмотрению конусов внутренних направлений.

Лемма 6. Пусть $g(u) \in C^1(E^n)$, $g(u_*) \leq 0$, $g'(u_*) \neq 0$. Тогда конус K_b внутренних направлений множества

$$U = \{u \in E^n: g(u) \leq 0\} \quad (23)$$

в точке u_* является открытым выпуклым конусом, причем конус K_b и двойственный ему конус K_b^* описываются следующим образом:

а) если $g(u_*) < 0$, то

$$K_b = E^n, \quad K_b^* = \{0\}; \quad (24)$$

б) если $g(u_*) = 0$, то

$$K_b = \{e \in E^n: \langle g'(u_*), e \rangle < 0\}, \quad (25)$$

$$K_b^* = \{c \in E^n: c = -\lambda g'(u_*), \lambda \geq 0\}. \quad (26)$$

Доказательство. Пусть сначала $g(u_*) < 0$. В силу непрерывности $g(u)$ найдется окрестность $O(u_*, \delta) = \{u \in E^n: |u - u_*| < \delta\}$, $\delta > 0$, такая, что $g(u) < 0$ при всех $u \in O(u_*, \delta)$. Это значит, что $u_* \in \text{int } U$. Тогда, как было показано в лемме 2, $K_b = E^n$. Поэтому $K_b^* = \{0\}$. Равенства (24) получены.

Пусть теперь $g(u_*) = 0$. По условию $g'(u_*) \neq 0$. Обозначим $Q = \{e \in E^n: \langle g'(u_*), e \rangle < 0\}$. Возьмем $e \in K_b$. Тогда существуют числа $e_0 > 0$, $\delta > 0$ такие, что $u_* + e\bar{e} \in U$ или $g(u_* + e\bar{e}) \leq 0$ при всех $\bar{e}$, $|\bar{e} - e| < \delta$, и всех e , $0 < e < e_0$. Отсюда, учитывая, что $g(u_*) = 0$, имеем $g(u_* + e\bar{e}) - g(u_*) = \langle g'(u_* + \theta e \bar{e}), e \rangle e \leq 0$. Деля на $e > 0$ и устремляя e к $+0$, получим $\langle g'(u_*), e \rangle \leq 0$ для всех $e \in K_b$. В действительности здесь $\langle g'(u_*), e \rangle < 0$ для любых $e \in K_b$ — это доказывается так же, как аналогичное утверждение в лемме 5. Таким образом, $K_b \subseteq Q$. Обратное включение $Q \subseteq K_b$, из которого будет следовать равенство (25), доказывается с помощью соотношений вида (21), (22), написанных для функции $g(u)$ с учетом равенства $g(u_*) = 0$, — это делается так же, как при доказательстве леммы 5. Интересно заметить, что при этом попутно устанавливается регулярность множества (23) (см. определение 8.2). Выпуклость конуса (25) очевидна; его открытость доказана в лемме 2. Из (25) и теоремы 8.3 следует равенство (26).

Лемма 7. Пусть U — выпуклое множество из E^n , $\text{int } U \neq \emptyset$, $u_* \in U$. Тогда конус K_b внутренних направлений множества U в точке u_* является открытым выпуклым конусом, причем конус K_b и двойственный ему конус K_b^* описываются следующим образом:

а) если $u_* \in \text{int } U$, то $K_b = E^n$, $K_b^* = \{0\}$;

б) если u_* — граничная точка U , то

$$K_b = \{e \in E^n: e = \lambda(u - u_*), \quad u \in \text{int } U, \quad \lambda > 0\}, \quad (27)$$

$$K_b^* = \{c \in E^n: \langle c, u \rangle \geq \langle c, u_* \rangle \text{ при всех } u \in U\}. \quad (28)$$

Доказательство. Случай $u_* \in \text{int } U$ рассмотрен в лемме 2. Поэтому пусть $u_* \in \text{Гр } U$. Обозначим

$$Q = \{e \in E^n: e = \lambda(u - u_*), \quad u \in \text{int } U, \quad \lambda > 0\}.$$

Равенство (27) будет доказано, если покажем, что $K_b = Q$. Пусть $e \in K_b$. В лемме 2 было установлено, что $u = u_* + ee \in \text{int } U$ при всех $e, 0 < e < e_0$, где e_0 взято из определения 2. Отсюда получаем представление $e = (u - u_*)/e = \lambda(u - u_*)$, $u \in \text{int } U$, $\lambda = 1/e > 0$. Итак, если $e \in K_b$, то $e \in Q$, т. е. $K_b \subseteq Q$.

Покажем справедливость обратного включения $Q \subseteq K_b$. Пусть $e \in Q$, т. е. $e = \lambda(u - u_*)$, $u \in \text{int } U$, $\lambda > 0$. Из того, что $u \in \text{int } U$, следует существование числа $\delta > 0$ такого, что $O(u, \delta) = \{v \in E^n: |v - u| < \delta\} \subset U$. Возьмем произвольный вектор $\bar{e}$, $|\bar{e} - e| < \delta\lambda$. Тогда $|u_* + \bar{e}/\lambda - u| = |\bar{e} - e|/\lambda < \delta$, т. е. $u_* + \bar{e}/\lambda \in O(u, \delta) \subset U$. В силу выпуклости U отрезок $[u_*, u_* + \bar{e}/\lambda]$ также будет принадлежать U . Это значит, что $\alpha(u_* + \bar{e}/\lambda) + (1 - \alpha)u_* = u_* + (\alpha/\lambda)\bar{e} \in U$, $0 \leq \alpha \leq 1$, или $u_* + e\bar{e} \in U$ при всех $\bar{e}$, $|\bar{e} - e| < \delta\lambda$, и всех $e, 0 < e < 1/\lambda = e_0$. Иначе говоря, $e \in K_b$. Тем самым показано, что $Q \subseteq K_b$. Равенство (27) установлено. Открытость конуса K_b доказана в лемме 2. Покажем выпуклость K_b . Пусть $e_1, e_2 \in K_b$, т. е. $e_1 = \lambda_1(u_1 - u_*)$, $e_2 = \lambda_2(u_2 - u_*)$, где $u_1, u_2 \in \text{int } U$, $\lambda_1 > 0$, $\lambda_2 > 0$. Пусть $0 \leq \alpha \leq 1$. Тогда $\lambda_\alpha = \alpha\lambda_1 + (1 - \alpha)\lambda_2 > 0$. Положим $\beta = \alpha\lambda_1/\lambda_\alpha$; тогда $1 - \beta = (1 - \alpha)\lambda_2/\lambda_\alpha$. Ясно, что $0 \leq \beta \leq 1$. Поскольку $\text{int } U$ — выпуклое множество (см. теорему 1.4), то $u_\beta = \beta u_1 + (1 - \beta)u_2 \in \text{int } U$. Для вектора $e_\alpha = \alpha e_1 + (1 - \alpha)e_2$ имеем представление $e_\alpha = \alpha\lambda_1(u_1 - u_*) + (1 - \alpha)\lambda_2(u_2 - u_*) = \lambda_\alpha[\beta(u_1 - u_*) + (1 - \beta)(u_2 - u_*)] = \lambda_\alpha(u_\beta - u_*)$, $u_\beta \in \text{int } U$, $\lambda_\alpha > 0$. Следовательно, $e_\alpha \in K_b$ при всех $\alpha, 0 \leq \alpha \leq 1$. Выпуклость конуса K_b установлена.

Докажем равенство (28). Обозначим $M = \{c \in E^n: \langle c, u - u_* \rangle \geq 0 \text{ при всех } u \in U\}$. Возьмем любой вектор $c \in M$. Пусть $e \in K_b$. Согласно равенству (27) тогда найдутся $u \in \text{int } U$ и $\lambda > 0$ такие, что $e = \lambda(u - u_*)$. Тогда, учитывая $c \in M$, получаем $\langle c, e \rangle = \langle c, \lambda(u - u_*) \rangle = \lambda \langle c, u - u_* \rangle \geq 0$, т. е. $\langle c, e \rangle \geq 0$ для всех $e \in K_b$. Это значит, что $c \in K_b^*$, следовательно, $M \subseteq K_b^*$.

Покажем, что $K_b^* \subseteq M$. Возьмем любой вектор $c \in K_b^*$. Тогда $\langle c, e \rangle \geq 0$ для всех $e \in K_b$. Но согласно формуле (27) любой вектор $e \in K_b$ представим в виде $e = \lambda(u - u_*)$, $u \in \text{int } U$, $\lambda > 0$. Следовательно, $\langle c, e \rangle = \langle c, \lambda(u - u_*) \rangle \geq 0$ или $\langle c, u - u_* \rangle \geq 0$ для всех $u \in \text{int } U$. Если $u \in U \setminus \text{int } U$, то найдутся $u_m \in \text{int } U$, $m = 1, 2, \dots$, $\lim_{m \rightarrow \infty} u_m = u$ (согласно теореме 1.3

можно, например, взять $u_m = u + (\bar{u} - u)/m$, где $\bar{u}$ — какая-либо точка из $\text{int } U$). По доказанному $\langle c, u_m - u_* \rangle \geq 0$, $m = 1, 2, \dots$. Отсюда при $m \rightarrow \infty$ получаем $\langle c, u - u_* \rangle \geq 0$ и для $u \in U \setminus \text{int } U$. Таким образом, $\langle c, u - u_* \rangle \geq 0$ при всех $u \in U$. Это значит, что $c \in M$. Тем самым показано, что $K_b^* \subseteq M$.

Тогда $K_b^* = M$, т. е. равенство (28) установлено. Лемма 7 доказана.

Формула (28) показывает, что двойственный конус к конусу (27) при $u_* \in \text{Гр } U$ состоит из всех опорных векторов ко множеству U в точке u_* (см. определение 5.2) и нулевого вектора.

Наконец, выясним структуру конуса касательных направлений для множеств вида (16).

Лемма 8. Пусть

$$U = \{u \in E^n: g_i(u) = 0, \quad i = \overline{1, s}\}, \quad (29)$$

где $g_i(u) \in C^1(E^n)$. Пусть в точке $u_* \in U$ векторы $g'_1(u_*)$, ..., $g'_s(u_*)$ линейно независимы (u , следовательно, $s \leq n$). Тогда конус K_k касательных направлений множества (29) в точке u_* является замкнутым выпуклым конусом, причем конус K_k и двойственный ему конус K_k^* описываются следующим образом:

$$K_k = \{e \in E^n: \langle g'_i(u_*), e \rangle = 0, \quad i = \overline{1, s}\}, \quad (30)$$

$$K_k^* = \left\{ c \in E^n: c = - \sum_{i=1}^s \lambda_i g'_i(u_*), \quad \lambda = (\lambda_1, \dots, \lambda_s) \in E^s \right\}. \quad (31)$$

Доказательство. Сначала рассмотрим случай $s = n$. Тогда u_* — изолированная точка множества U [182, стр. 125], и касательным направлением в этой точке может быть лишь $e = 0$, т. е. $K_k = \{0\}$, что вполне согласуется с формулой (30). В этом случае $K_k^* = E^n$. Поскольку $g'_1(u_*)$, ..., $g'_n(u_*)$ — базис в E^n , то для любого $c \in K_k^* = E^n$ имеет место представление $c = - \sum_{i=1}^n \lambda_i g'_i(u_*)$. Таким образом, при $s = n$ формулы (30), (31) получены.

Пусть теперь $s < n$. Обозначим

$$Q = \{e \in E^n: \langle g'_i(u_*), e \rangle = 0, \quad i = \overline{1, s}\}.$$

Для получения равенства (30) достаточно установить, что $K_k = Q$. Возьмем произвольный вектор $e \in K_k$. По определению касательного направления существуют число $\varepsilon_0 > 0$ и функция $r(\varepsilon)$, $r(0) = 0$, $\lim_{\varepsilon \rightarrow +0} r(\varepsilon)/\varepsilon = 0$, такие, что $u_* + \varepsilon e + r(\varepsilon) \in U$ или $g_i(u_* + \varepsilon e + r(\varepsilon)) = 0, i = \overline{1, s}$, при всех $\varepsilon, 0 \leq \varepsilon < \varepsilon_0$. Отсюда с учетом того, что $g_i(u_*) = 0, i = \overline{1, s}$, получим $g_i(u_* + \varepsilon e + r(\varepsilon)) - g_i(u_*) = \langle g'_i(u_*), \varepsilon e + r(\varepsilon) \rangle + o(\varepsilon, u_*) = 0, i = \overline{1, s}, 0 < \varepsilon < \varepsilon_0$. Деля эти равенства на $\varepsilon > 0$ и устремляя $\varepsilon \rightarrow +0$, будем иметь $\langle g'_i(u_*), e \rangle = 0, i = \overline{1, s}$, т. е. $e \in Q$. Это значит, что $K_k \subseteq Q$.

Покажем, что $Q \subseteq K_k$. Пусть $e = (e^1, \dots, e^n) \in Q, e \neq 0$, т. е. $\langle g'_i(u_*), e \rangle = 0, i = \overline{1, s}$. По условию система $\{g'_i(u_*), i = \overline{1, s}\}$ линейно независима. Тогда ранг матрицы $g'(u_*) = \left\{ \frac{\partial g_i(u_*)}{\partial u^j}, i = \overline{1, s}, j = \overline{1, n} \right\}$, где $s < n$, равен s и существует минор порядка s , отличный от нуля. Без умаления общности можем считать, что этот минор составлен из первых s столбцов матрицы $g'(u_*)$. По теореме о неявных функциях [7, 8, 179] тогда в некоторой достаточно малой окрестности $O(u_*, \nu) = \{u \in E^n: |u - u_*| < \nu\}$ точки u_* существуют дифференцируемые функции

$$u^p = \varphi_p(u^{s+1}, \dots, u^n), \quad p = \overline{1, s},$$

удовлетворяющие условиям $u_*^p = \varphi_p(\bar{u}_*)$, $p = \overline{1, s}$, где $\bar{u}_* = (u_*^{s+1}, \dots, u_*^n)$, и системе уравнений $g_i(u^1, \dots, u^n) = g_i(u) = 0, i = \overline{1, s}$, причем производ-

ные $\frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j}$, $p = \overline{1, s}$, $j = \overline{s+1, n}$, однозначно определяются из линейной системы уравнений

$$\sum_{p=1}^s \frac{\partial g_i(u_*)}{\partial u^p} \frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j} + \frac{\partial g_i(u_*)}{\partial u^j} = 0, \quad i = \overline{1, s}; \quad j = \overline{s+1, n}.$$

Умножим j -е уравнение на e^j и просуммируем получающиеся равенства по j от $s+1$ до n . Будем иметь

$$\sum_{p=1}^s \frac{\partial g_i(u_*)}{\partial u^p} \left(\sum_{j=s+1}^n \frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j} e^j \right) = - \sum_{j=s+1}^n \frac{\partial g_i(u_*)}{\partial u^j} e^j, \quad i = \overline{1, s}. \quad (32)$$

С другой стороны, по координатно расписывая условия $\langle g'_i(u_*), e \rangle = 0$, $i = \overline{1, s}$, получим

$$\sum_{p=1}^s \frac{\partial g_i(u_*)}{\partial u^p} e^p = - \sum_{j=s+1}^n \frac{\partial g_i(u_*)}{\partial u^j} e^j, \quad i = \overline{1, s}. \quad (33)$$

Сравнивая (32) и (33), заключаем, что система линейных уравнений

$$\sum_{p=1}^s \frac{\partial g_i(u_*)}{\partial u^p} z^p = - \sum_{j=s+1}^n \frac{\partial g_i(u_*)}{\partial u^j} e^j, \quad i = \overline{1, s},$$

имеет два решения: $z^p = e^p$, $p = \overline{1, s}$ и $z^p = \sum_{j=s+1}^n \frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j} e^j$, $p = \overline{1, s}$. Но

определителем этой системы является минор, составленный из первых s столбцов матрицы $g'(u_*)$ и отличный от нуля, так что система имеет единственное решение. Это значит, что

$$e^p = \sum_{j=s+1}^n \frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j} e^j, \quad p = \overline{1, s}. \quad (34)$$

Положим $u(\varepsilon) = (u^1(\varepsilon), \dots, u^n(\varepsilon))$, где

$$u^p(\varepsilon) = \begin{cases} \varphi_p(u_*^{s+1} + \varepsilon e^{s+1}, \dots, u_*^n + \varepsilon e^n), & p = \overline{1, s}, \\ u_*^p + \varepsilon e^p, & p = \overline{s+1, n}; \quad 0 \leq \varepsilon < \varepsilon_0 = \nu / \|e\|. \end{cases}$$

Из определения функций $u^p = \varphi_p(u_*^{s+1}, \dots, u_*^n)$, $p = \overline{1, s}$, следует, что $g_i(u(\varepsilon)) = 0$, $i = \overline{1, s}$, т. е. $u(\varepsilon) \in U$, $0 \leq \varepsilon < \varepsilon_0$. Далее, с учетом равенств (34) и $u_*^p = \varphi_p(\bar{u}_*)$, $p = \overline{1, s}$, имеем

$$\begin{aligned} u^p(\varepsilon) &= \varphi_p(\bar{u}_*) + \varepsilon \sum_{j=s+1}^n \frac{\partial \varphi_p(\bar{u}_*)}{\partial u^j} e^j + r_p(\varepsilon) = \\ &= u_*^p + \varepsilon e^p + r_p(\varepsilon), \quad p = \overline{1, s}; \quad \lim_{\varepsilon \rightarrow 0} r_p(\varepsilon)/\varepsilon = 0. \end{aligned} \quad (35)$$

Следовательно, $u(e) = u_* + \frac{e}{\varepsilon} + r(e) \in U$, $0 \leq \varepsilon < \varepsilon_0$, где $r(e) = (r_1(e), \dots, r_n(e))$, $r_p(e)$ при $p = 1, s$ определяется из (35), а $r_p(e) \equiv 0$, $p = s+1, n$, так что $r(0) = 0$, $\lim_{\varepsilon \rightarrow +0} r(e)/\varepsilon = 0$. Это означает, что e — касательное направление.

Таким образом, если $e \in Q$, то $e \in K_k$, т. е. $Q = K_k$. Равенство (30) установлено. Выпуклость и замкнутость конуса (30) очевидны.

Представление (31) для двойственного конуса следует из теоремы 8.3. Лемма 8 доказана.

Доказательство теоремы 1. Выше было замечено, что теорема 1 требует отдельного доказательства лишь в случае, когда $J'(u_*) \neq 0$, $g'_i(u_*) \neq 0$ при всех $i = 1, m$ и векторы $\{g'_i(u_*), i = m+1, s\}$ линейно независимы. Тогда согласно лемме 5 конус направлений убывания функции $J(u)$ в точке u_* имеет вид

$$K_y = \{e \in E^n: \langle J'(u_*), e \rangle < 0\},$$

а двойственный ему конус —

$$K_y^* = \{c \in E^n: c = -\lambda_0 J'(u_*), \lambda_0 \geq 0\}.$$

Из леммы 6 следует, что для множества $U_i = \{u \in E^n: g_i(u) \leq 0\}$ конус внутренних направлений K_i в точке u_* и двойственный ему конус K_i^* описываются следующим образом:

$$K_i = E^n, \quad K_i^* = \{0\} \quad \text{при} \quad g_i(u_*) < 0,$$

а если $g_i(u_*) = 0$, то

$$K_i = \{e \in E^n: \langle g'_i(u_*), e \rangle < 0\},$$

$$K_i^* = \{c \in E^n: c = -\lambda_i g'_i(u_*), \lambda_i \geq 0\}, \quad i = \overline{1, m}.$$

Согласно лемме 7 для выпуклого множества U_0 , $\text{int } U_0 \neq \emptyset$, конус внутренних направлений K_0 в точке u_* и двойственный ему конус K_0^* имеют вид

$$K_0 = E^n, \quad K_0^* = \{0\} \quad \text{при} \quad u_* \in \text{int } U_0, \quad (36)$$

а если u_* — граничная точка U_0 , то

$$K_0 = \{e \in E^n: e = \lambda(u - u_*), \quad u \in \text{int } U_0, \quad \lambda > 0\}, \quad (37)$$

$$K_0^* = \{c \in E^n: \langle c, u - u_* \rangle \geq 0 \quad \text{при всех} \quad u \in U_0\}. \quad (38)$$

Наконец, согласно лемме 8 для множества $U_{m+1} = \{u \in E^n: g_i(u) = 0, i = m+1, s\}$ конус касательных направлений K_{m+1} в точке u_* и двойственный ему конус K_{m+1}^* представляются в виде

$$K_{m+1} = \{e \in E^n: \langle g'_i(u_*), e \rangle = 0, \quad i = m+1, s\},$$

$$K_{m+1}^* = \left\{ c \in E^n: c = - \sum_{i=m+1}^s \lambda_i g'_i(u_*) \right\}.$$

Важно заметить, что конусы $K_y, K_0, \dots, K_m$ выпуклы и открыты, а конус K_{m+1} — выпуклый и замкнутый, причем все эти конусы непустые.

Поскольку u_* — точка минимума функции $J(u)$ на множестве $U = \bigcap_{i=0}^{m+1} U_i$,

то по лемме 4 конусы $K_y, K_0, \dots, K_{m+1}$ необходимо имеют пустое пересечение. А тогда по теореме 5.6 найдутся векторы $c_y, c_0, c_1, \dots, c_{m+1}$, не все равные нулю и такие, что

$$c_y \in K_y^*, \quad c_i \in K_i^*, \quad i = \overline{0, m+1}; \quad c_y + c_0 + c_1 + \dots + c_{m+1} = 0. \quad (39)$$

Учитывая приведенное выше описание двойственных конусов, можем сказать, что

$$c_y = -\lambda_0^* J'(u_*), \quad c_i = -\lambda_i^* g_i'(u_*), \quad c_{m+1} = -\sum_{i=m+1}^s \lambda_i^* g_i'(u_*),$$

где $\lambda_0^* \geq 0, \lambda_1^* \geq 0, \dots, \lambda_m^* \geq 0$, так что равенство из (39) запишется в виде

$$c_0 = -c_y - c_1 - \dots - c_{m+1} = \lambda_0^* J'(u_*) + \sum_{i=1}^s \lambda_i^* g_i'(u_*), \quad (40)$$

Но $c_0 \in K_0^*$, поэтому $\langle c_0, u - u_* \rangle \geq 0$ для всех $u \in U_0$. Отсюда и из (40) получим условие (10), которое в случае $u_* \in \text{int } U_0$ превращается в равенство (12). Условие, что не все $c_y, c_0, \dots, c_{m+1}$ равны нулю, и равенство (39) гарантируют, что не все $\lambda_0^*, \lambda_1^*, \dots, \lambda_s^*$ равны нулю. Таким образом, соотношения (9) также получены.

Далее, если для некоторого $i, 1 \leq i \leq m$, оказалось, что $g_i(u_*) < 0$, то $K_i^* = \{0\}$, так что $c_i = -\lambda_i^* g_i'(u_*) = 0$. Отсюда с учетом $g_i'(u_*) \neq 0$ имеем $\lambda_i^* = 0$, поэтому $\lambda_i^* g_i'(u_*) = 0$. Если же $g_i(u_*) = 0$ при некотором $i, 1 \leq i \leq m$, то тем более $\lambda_i^* g_i'(u_*) = 0$. Наконец, условие $u_* \in U$ предполагает, что $g_i(u_*) = 0, i = \overline{m+1, s}$, поэтому $\lambda_i^* g_i'(u_*) = 0$ и для $i = \overline{m+1, s}$. Равенства (11) получены. Теорема 1 доказана.

З а м е ч а н и е 1. В теореме 1 по существу получены необходимые условия локального минимума функции $J(u)$ на множестве U . В самом деле, пусть u_* — точка локального минимума, т. е. $J(u_*) \leq J(u)$ при всех $u \in U \cap O(u_*, \alpha) = N$, где $O(u_*, \alpha) = \{u \in E^n: |u - u_*| < \alpha\}$, $\alpha > 0$ — окрестность точки u_* . Это значит, что u_* — точка глобального минимума $J(u)$ на множестве $N = \bigcap_{i=0}^{m+1} N_i$, где $N_i = U_i \cap O(u_*, \alpha)$, $i = \overline{0, m+1}$, множества U_i ,

$i = \overline{1, m+1}$, определены посредством (15), (16), а $N_0 = U_0$. Поэтому лемма 4, в которой вместо U_i будут использоваться множества $N_i, i = \overline{0, m+1}$, сохранит силу. Далее, в леммах 5, 6, 8 при построении конусов $K_y, K_i, i = \overline{1, m+1}$, по существу использовались свойства функций $J(u), g_i(u), i = \overline{1, s}$, лишь в некоторой сколь угодно малой окрестности точки u_* , поэтому вид конусов K_y, K_y^*, K_i и K_i^* для множества $N_i, i = \overline{1, m+1}$, сохранит свой прежний вид. Поэтому, проводя прежние рассуждения на множестве

$N = \bigcap_{i=0}^{m+1} N_i$, мы получим в качестве необходимых условий локального минимума те же условия (9)–(11). Отсюда же, кстати, следует, что теорема 1 остается справедливой, если непрерывную дифференцируемость функций $J(u), g_i(u), i = \overline{1, s}$, потребовать не на всем пространстве E^n , а лишь в некоторой

окрестности $O(u_*, \alpha)$ точки u_* . Наконец, если вместо выпуклости множества U_0 и условия $\text{int } U_0 \neq \emptyset$ потребовать выпуклость множества $N_0 = U_0 \cap O(u_*, \alpha)$ и условие $\text{int } N_0 \neq \emptyset$, то также получим условия (9)–(11), но неравенство (10) будет выполняться лишь при всех $u \in N_0$, но, вообще говоря, не для всех $u \in U_0$.

6. Отдельно остановимся на следующей задаче выпуклого программирования:

$$\begin{aligned} J(u) \rightarrow \inf; \quad u \in U, \\ U = \{u \in E^n: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}; g_i(u) = \langle a_i, u \rangle - b^i = 0, \\ i = \overline{m+1, s}\}. \end{aligned} \quad (41)$$

Теорема 2. Пусть U_0 — выпуклое множество из E^n , $\text{int } U_0 \neq \emptyset$, пусть функции $J(u)$, $g_i(u)$, $i = \overline{1, m}$, определены и выпуклы на открытом выпуклом множестве W , содержащем U_0 (например, $W = E^n$). Пусть множество U_* точек минимума функции $J(u)$ на множестве (41) непусто. Тогда для каждой точки $u_* \in U_*$ необходимо найдутся множители Лагранжа $\bar{\lambda}^* = (\lambda_0^*, \lambda_1^*, \dots, \lambda_s^*)$ такие, что

$$\bar{\lambda}^* \neq 0, \quad \lambda_0^* \geq 0, \dots, \lambda_m^* \geq 0, \quad (42)$$

$$\mathcal{L}(u_*, \bar{\lambda}^*) \leq \mathcal{L}(u, \bar{\lambda}^*) \quad \text{при всех } u \in U_0, \quad (43)$$

$$\lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, s}, \quad (44)$$

где функция Лагранжа $\mathcal{L}(u, \bar{\lambda})$ определяется посредством формулы (8).

В теореме 2 не исключаются случаи, когда $U_0 = E^n$, или в (41) отсутствуют ограничения типа неравенства или равенств ($m = 0$, или $m = s$, или $m = s = 0$). В отличие от теоремы 1, гладкости функций $J(u)$, $g_i(u)$, $i = \overline{1, m}$, здесь мы не требуем и вместо градиента этих функций будем пользоваться их субдифференциалом. Поскольку указанные функции выпуклы на открытом выпуклом множестве, содержащем U_0 , то в силу теоремы 6.1 субдифференциалы $\partial J(u)$, $\partial g_i(u)$, $i = \overline{1, m}$, непусты при всех $u \in U_0$, в частности, в точке $u_* \in U_*$.

Доказательство теоремы 2. Достаточно установить существование таких $\bar{\lambda}^* = (\lambda_0^*, \dots, \lambda_s^*)$, удовлетворяющих условиям (42), (44), и таких $a_0 \in \partial J(u_*)$, $a_i \in \partial g_i(u_*)$, $i = \overline{1, m}$, для которых

$$\langle \lambda_0^* a_0 + \lambda_1^* a_1 + \dots + \lambda_s^* a_s, u - u_* \rangle \geq 0 \quad \text{при всех } u \in U_0 \quad (45)$$

(здесь $a_{m+1}, \dots, a_s$ взяты из определения множества (41)). Дело в том, что неравенство (43), оказывается, является следствием неравенства (45). В самом деле, при таких $\bar{\lambda}^*$ из выпуклости $J(u)$, $g_i(u)$, $i = \overline{1, m}$, линейности $g_i(u)$, $i = \overline{m+1, s}$, следует выпуклость функции $\mathcal{L}(u, \bar{\lambda}^*)$ переменной u на

множестве W . Кроме того, нетрудно проверить, что вектор $a = \sum_{i=0}^s \lambda_i^* a_i$,

взятый из (45), является субградиентом функции $\mathcal{L}(u, \bar{\lambda}^*)$ в точке u_* . Это значит, что $\mathcal{L}(u, \bar{\lambda}^*) - \mathcal{L}(u_*, \bar{\lambda}^*) \geq \langle a, u - u_* \rangle$ при всех $u \in W$. Отсюда и из (45) тогда будем иметь неравенство (43), и теорема 2 будет доказана.

Если $0 \in \partial J(u_*)$, то для удовлетворения условий (42), (44), (45) достаточно взять $\lambda_0^* = 1$, $\lambda_1^* = \dots = \lambda_s^* = 0$, $a_0 = 0 \in \partial J(u_*)$, любые $a_i \in \partial g_i(u_*)$,

$i = \overline{1, m}$. Если $0 \in \partial g_i(u_*)$ при некотором i , $1 \leq i \leq m$, то условия (42), (44), (45) будут выполнены при $\lambda_i^* = 1$, $\lambda_j^* = 0$, $j = \overline{0, s}$, $j \neq i$, $a_i = 0 \in \partial g_i(u_*)$, любых $a_0 \in \partial J(u_*)$, $a_j \in \partial g_j(u_*)$, $j = \overline{1, m}$, $j \neq i$. Далее, если векторы $a_{m+1}, \dots, a_s$ линейно зависимы, т. е. $\alpha_{m+1}a_{m+1} + \dots + \alpha_s a_s = 0$, где не все числа $\alpha_{m+1}, \dots, \alpha_s$ равны нулю, то в (42), (44), (45) можно взять $\lambda_0^* = \dots = \lambda_m^* = 0$, $\lambda_i^* = \alpha_i$, $i = m+1, s$, и любые $a_0 \in \partial J(u_*)$, $a_i \in \partial g_i(u_*)$, $i = \overline{1, m}$.

Остается рассмотреть случай, когда $0 \notin \partial J(u_*)$, $0 \notin \partial g_i(u_*)$ при всех $i = \overline{1, m}$, а векторы $a_{m+1}, \dots, a_s$ линейно независимы. При этих условиях изучим структуру конусов направлений убывания функции $J(u)$ в точке u_* , внутренних направлений множеств (15) и касательных направлений множества (16), а также двойственных к ним конусов.

Лемма 9. Пусть W — открытое выпуклое множество из E^n , функция $J(u)$ выпукла на W , пусть точка $u_* \in W$, $0 \notin \partial J(u_*)$. Тогда конус K_y направлений убывания функции $J(u)$ в точке u_* и двойственный ему конус K_y^* будут иметь вид

$$K_y = \{e \in E^n: e = \mu(u - u_*), \mu > 0, u \in W, J(u) < J(u_*)\}, \quad (46)$$

$$K_y^* = \{c \in E^n: c = -\lambda a, \lambda \geq 0, a \in \partial J(u_*)\}. \quad (47)$$

Доказательство. Поскольку $0 \notin \partial J(u_*)$, то из теоремы 6.4 следует существование точек $u \in W$, для которых $J(u) < J(u_*)$. Это значит, что множество

$$Q = \{e \in E^n: e = \mu(u - u_*), \mu > 0, u \in W, J(u) < J(u_*)\}$$

непусто. Равенство (46) будет установлено, если покажем $K_y = Q$.

Возьмем любой вектор $e \in K_y$. Тогда существуют числа $e_0 > 0$, $\delta > 0$ такие, что $u_* + e\bar{e} \in W$ и $J(u_* + e\bar{e}) < J(u_*)$ для всех $\bar{e}$, e , $|\bar{e} - e| < \delta$, $0 < e < e_0$ (обращаем внимание читателя на то, что мы здесь уточнили определение направлений убывания — определение 1 было дано для случая $W = E^n$). В частности, при $\bar{e} = e$ отсюда имеем $u = u_* + ee \in W$, $J(u_* + ee) < J(u_*)$ при всех e , $0 < e < e_0$, т. е. $e = (u - u_*)/e = \mu(u - u_*)$, $\mu = 1/e > 0$, $u \in W$, $J(u) < J(u_*)$. Это значит, что $e \in Q$. Показано, что $K_y \subseteq Q$.

Пусть теперь $e \in Q$, т. е. $e = \mu(u - u_*)$ при некоторых $\mu > 0$, $u \in W$, $J(u) < J(u_*)$. В силу непрерывности функции $J(u)$ (см. теорему 2.15) и открытости множества W найдется число $\delta > 0$ такое, что $J(v) < J(u_*)$ для всех $v \in O(u, \delta) = \{v \in E^n: |v - u| < \delta\} \subset W$. Возьмем произвольный вектор $\bar{e}$, $|\bar{e} - e| < \delta\mu$. Тогда $|u_* + \bar{e}/\mu - u| = |\bar{e} - e|/\mu < \delta$, т. е. $u_* + \bar{e}/\mu \in O(u, \delta)$, и следовательно, $J(u_* + \bar{e}/\mu) < J(u_*)$. С учетом выпуклости $J(u)$ и множества W отсюда имеем

$$\begin{aligned} J(u_* + (\alpha/\mu)\bar{e}) &= J(\alpha(u_* + \bar{e}/\mu) + (1 - \alpha)u_*) \leq \\ &\leq \alpha J(u_* + \bar{e}/\mu) + (1 - \alpha)J(u_*) < J(u_*), \quad u_* + (\alpha/\mu)\bar{e} \in W \end{aligned}$$

при всех α , $0 < \alpha < 1$. Это значит, что $u_* + e\bar{e} \in W$, $J(u_* + e\bar{e}) < J(u_*)$ для всех e , $0 < e < 1/\mu = e_0$, и всех $\bar{e}$, $|\bar{e} - e| < \delta\mu$, т. е. $e \in K_y$. Тем самым показано, что $Q \subseteq K_y$. Отсюда и из ранее установленного включения $K_y \subseteq Q$ следует равенство (46).

Остается доказать равенство (47). Обозначим

$$M = \{c \in E^n: c = -\lambda a, \lambda \geq 0, a \in \partial J(u_*)\}.$$

Требуется показать, что $K_y^* = M$. Возьмем произвольный вектор $c \in M$, т. е. $c = -\lambda a$, $\lambda \geq 0$, $a \in \partial J(u_*)$. Это значит, что $J(u) - J(u_*) \geq \langle a, u - u_* \rangle$ при всех $u \in W$. Отсюда для всех $\mu > 0$ и всех $u \in W$, для которых $J(u) < J(u_*)$, имеем $0 > \langle a, u - u_* \rangle$ или $\langle -\lambda a, \mu(u - u_*) \rangle \geq 0$. Это значит, что $\langle c, e \rangle \geq 0$ при каждом $e \in K_y$, т. е. $c \in K_y^*$. Таким образом, показано, что $M \subseteq K_y^*$.

Пусть теперь $c \in K_y^*$. Покажем, что $c \in M$. Если $c = 0$, то это ясно. Поэтому пусть $c \neq 0$. На плоскости E^2 построим два множества

$$A = \{a = (a^1, a^2): a^1 \geq \langle c, u - u_* \rangle, a^2 \geq J(u) - J(u_*), u \in W\},$$

$$B = \{b = (b^1, b^2): b^1 < 0, b^2 < 0\}.$$

Эти множества не пересекаются. В самом деле, пусть $a = (a^1, a^2) \in A$. Это значит, что существует точка $u \in W$ такая, что $a^1 \geq \langle c, u - u_* \rangle$, $a^2 \geq J(u) - J(u_*)$. Возможно, что $J(u) \geq J(u_*)$. Тогда $a^2 \geq 0$ и заведомо $a \notin B$. Если же $J(u) < J(u_*)$, то согласно равенству (46) вектор $e = u - u_*$ принадлежит K_y . Учитывая, что $c \in K_y^*$, имеем $a^1 \geq \langle c, u - u_* \rangle = \langle c, e \rangle \geq 0$, т. е. снова $a \notin B$.

Итак, $A \cap B = \emptyset$. Кроме того, множества A и B выпуклы. По теореме 5.2 существует вектор $v = (v_1, v_2) \neq 0$ такой, что

$$v_1 b^1 + v_2 b^2 \leq v_1 a^1 + v_2 a^2 \text{ для всех } a \in A, b \in B.$$

В частности, при $a^1 = \langle c, u - u_* \rangle$, $a^2 = J(u) - J(u_*)$ отсюда имеем

$$v_1 b^1 + v_2 b^2 \leq v_1 \langle c, u - u_* \rangle + v_2 (J(u) - J(u_*))$$

для всех $u \in W$, $b^1 \leq 0$, $b^2 \leq 0$. Это возможно только при

$$v_1 \geq 0, v_2 \geq 0, \quad 0 \leq v_1 \langle c, u - u_* \rangle + v_2 (J(u) - J(u_*)), \quad u \in W.$$

Поскольку $c \neq 0$, $u_* \in \text{int } W = W$, то $v_2 > 0$. Поэтому, деля это неравенство на $v_2 > 0$, получим $J(u) - J(u_*) \geq \langle a, u - u_* \rangle$ при всех $u \in W$, где $a = (-v_1/v_2)c$. Следовательно, $a \in \partial J(u_*)$. Но $0 \notin \partial J(u_*)$, поэтому $a \neq 0$. Тогда $v_1 > 0$ и $c = (-v_2/v_1)a = -\lambda a$, $\lambda > 0$, $a \in \partial J(u_*)$, т. е. $c \in M$. Установлено, что $K_y^* \subseteq M$. Это значит, что равенство (47) — верное. Лемма 9 доказана.

Лемма 10. Пусть W — открытое выпуклое множество из E^n , функция $g_i(u)$ выпукла на W ; пусть $u_* \in W$, $g_i(u_*) \leq 0$, $0 \notin \partial g_i(u_*)$. Тогда конус K_i внутренних направлений множества

$$U_i = \{u \in E^n: u \in W, g_i(u) \leq 0\}$$

в точке u_* и двойственный ему конус K_i^* имеют вид:

а) если $g_i(u_*) < 0$, то $K_i = E^n$, $K_i^* = \{0\}$;

б) если $g_i(u_*) = 0$, то

$$K_i = \{e \in E^n: e = \mu(u - u_*), \mu > 0, u \in W, g_i(u) < 0\},$$

$$K_i^* = \{c \in E^n: c = -\lambda a, \lambda \geq 0, a \in \partial g_i(u_*)\}.$$

Доказательство полностью аналогично доказательству леммы 9.

Лемма 11. Пусть векторы $a_{m+1}, a_{m+2}, \dots, a_s$ линейно независимы и $\langle a_i, u_* \rangle = b^i$, $i = m+1, s$. Тогда конус K_{m+1} касательных направлений множества

$$U_{m+1} = \{u \in E^n: \langle a_i, u \rangle - b^i = 0, i = m+1, s\}$$

в точке u_* и двойственный ему конус K_{m+1}^* имеют вид

$$K_{m+1} = \{e \in E^n: \langle a_i, e \rangle = 0, \quad i = \overline{m+1, s}\},$$

$$K_{m+1}^* = \left\{ c \in E^n: c = - \sum_{i=m+1}^s \lambda_i a_i \right\}.$$

Эта лемма является частным случаем леммы 8.

Завершим доказательство теоремы 2. По условию точка u_* является точкой минимума функции $J(u)$ на множестве $U = \bigcap_{i=0}^{m+1} U_i$. Согласно лемме 4 тогда конусы $K_y, K_0, K_1, \dots, K_{m+1}$, где K_0 описывается формулой (37), а $K_y, K_1, \dots, K_{m+1}$ взяты из лемм 9—11, имеют пустое пересечение. Поскольку конусы $K_y, K_0, \dots, K_m$ выпуклы и открыты, а K_{m+1} выпуклый замкнутый конус, то по теореме 5.6 найдутся векторы $c_y \in K_y^*, c_i \in K_i^*, i = \overline{0, m+1}$, не все равные нулю и такие, что

$$c_y + c_0 + c_1 + \dots + c_{m+1} = 0. \quad (48)$$

Учитывая приведенное в леммах 9—11 описание конусов $K_y^*, K_1^*, \dots, K_{m+1}^*$ можем сказать, что

$$c_y = -\lambda_0^* a_0, \quad a_0 \in \partial J(u_*); \quad c_i = -\lambda_i^* a_i, \quad a_i \in \partial g_i(u_*), \quad i = \overline{1, m};$$

$$\lambda_0^* \geq 0, \quad \lambda_1^* \geq 0, \dots, \lambda_m^* \geq 0; \quad c_{m+1} = - \sum_{i=m+1}^s \lambda_i^* a_i.$$

Отсюда и из (48) следует

$$c_0 = -\lambda_0^* a_0 - \lambda_1^* a_1 - \dots - \lambda_s^* a_s.$$

Но $c_0 \in K_0^*$ и согласно формуле (38) $\langle c_0, u - u_* \rangle \geq 0$ при всех $u \in U_0$, что равносильно неравенству (45). Из того, что не все $c_0, c_1, \dots, c_{m+1}$ равны нулю, и из (48) следует, что $\tilde{\lambda}^* = (\lambda_0^*, \lambda_1^*, \dots, \lambda_s^*) \neq 0$. Наконец, если $g_i(u_*) = 0$, то равенство (44) выполняется. Если же $g_i(u_*) < 0$ при некотором $i, 1 \leq i \leq m$, то $K_i^* = \{0\}$ и $c_i = -\lambda_i^* a_i = 0$. Поскольку $0 \notin \partial g_i(u_*)$, то $a_i \neq 0, \lambda_i^* = 0$, так что (44) имеет место и в этом случае. Соотношения (42), (44), (45) и, следовательно, теорема 2 доказаны.

Теоремы 8.2, 8.4, 8.5 дают условия, при которых задача минимизации функции $J(u)$ на множестве (41) — невырожденная. Для невырожденных задач в (42)—(44) можно взять $\lambda_0^* = 1$, и тогда условия (42)—(44) становятся не только необходимыми, но и достаточными для оптимальности (см. лемму 8.1 и теорему 8.1). Кстати, анализируя доказательства теорем 8.4 и только что полученной теоремы 2, нетрудно понять, что в теореме 8.4 вместо гладкости и выпуклости функции $J(u)$ на множестве U_0 достаточно было потребовать выпуклости функции $J(u)$ на каком-либо открытом выпуклом множестве W , содержащем U_0 .

Различные варианты теорем 1, 2, их далеко идущие обобщения, приложения см. в [47, 48, 102, 107, 113, 160, 197, 220, 256].

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, им посвящена обширная литература, из которой мы здесь можем упомянуть лишь весьма незначительную ее часть [2, 5, 9, 11, 13, 15, 16, 18, 20, 21, 23, 29, 30, 33, 35, 38, 41, 44, 45, 51—53, 55—57, 59, 65, 76, 81, 82, 84—86, 89, 90, 93, 94, 96—101, 104—110, 112—119, 121—125, 127, 129—135, 137, 138, 140—142, 144, 154, 155, 157, 161, 164—166, 169, 172, 174, 175, 180, 182, 187—189, 191, 192, 196, 199—202, 208, 213—218, 222—224, 228, 229, 236, 238, 240, 242, 249—251, 253, 254]. Эта область экстремальных задач продолжает бурно развиваться, число работ, посвященных задачам минимизации функций конечного числа переменных и опубликованных в научной печати, быстро растет.

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.

§ 1. Градиентный метод

1. Будем рассматривать задачу

$$J(u) \rightarrow \inf; \quad u \in U \equiv E^n, \quad (1)$$

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

$$J(u+h) - J(u) = \langle J'(u), h \rangle + o(h; u), \quad (2)$$

где $\lim_{|h| \rightarrow 0} o(h; u) |h|^{-1} \neq 0$. 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

$$-|J'(u)| \cdot |h| \leq \langle J'(u), h \rangle \leq |J'(u)| \cdot |h|,$$

причем если $J'(u) \neq 0$, то правое неравенство превращается в равенство только при $h = \alpha J'(u)$, а левое неравенство — только при $h = -\alpha J'(u)$, где $\alpha = \text{const} \geq 0$. Отсюда ясно, что при $J'(u) \neq 0$ направление наивысшего возрастания функции $J(u)$ в точке u совпадает с направлением градиента $J'(u)$, а направление наивысшего 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

$$u_{k+1} = u_k - \alpha_k J'(u_k), \quad \alpha_k > 0, \quad k = 0, 1, 2, \dots \quad (3)$$

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

$$J(u_{k+1}) - J(u_k) = \alpha_k [-|J'(u_k)|^2 + o(\alpha_k) \alpha_k^{-1}] < 0$$

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 $\{u \in E^n: u = u_k - \alpha J'(u_k), \alpha \geq 0\}$, направленном по антиградиенту, введем функцию одной переменной

$$f_k(\alpha) = J(u_k - \alpha J'(u_k)), \quad \alpha \geq 0,$$

и определим α_k из условий

$$f_k(\alpha_k) = \inf_{\alpha \geq 0} f_k(\alpha) = f_{k*}, \quad \alpha_k > 0. \quad (4)$$

Метод (3), (4) принято называть методом скорейшего спуска. При $J'(u_k) \neq 0$ согласно формуле (2.3.1) $g'_k(0) = -|J'(u_k)|^2 < 0$, 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

$$J(u) = \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle, \quad (5)$$

где A — симметричная положительно определенная матрица порядка 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:

$$u_{k+1} = u_k - \alpha_k (Au_k - b), \quad k = 0, 1, \dots$$

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

$$f_k(\alpha) = J(u_k) - \alpha |J'(u_k)|^2 + (\alpha^2/2) \langle AJ'(u_k), J'(u_k) \rangle, \quad \alpha \geq 0.$$

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 $\alpha \geq 0$. 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

$$f_{k*} \leq f_k(\alpha_k) \leq f_{k*} + \delta_k, \quad \delta_k \geq 0, \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty, \quad (6)$$

или из условий [9]

$$f_{k*} \leq f_k(\alpha_k) \leq (1 - \lambda_k) f_k(0) + \lambda_k f_{k*}, \quad 0 < \bar{\lambda} \leq \lambda_k \leq 1. \quad (7)$$

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 быстреего 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 $k \geq 0$ 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

$$|J'(u) - J'(v)| \leq L|u - v|, \quad u, v \in E^n,$$

причем константа L известна, то в (3) в качестве α_k может быть взято любое число, удовлетворяющее условиям

$$0 < \varepsilon_0 \leq \alpha_k \leq 2/(L + 2\varepsilon), \quad (8)$$

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 [15]

$$J(u_k) - J(u_k - \alpha_k J'(u_k)) \geq \varepsilon \alpha_k |J'(u_k)|^2, \quad \varepsilon > 0. \quad (9)$$

Для удовлетворения условия (9) сначала обычно берут некоторое число $\alpha_k = \alpha > 0$ (одно и то же на всех итерациях; напри-

мер, $\alpha_k = 1$), а затем при необходимости дробят его, т.е. изменяют по закону $\alpha_k = \lambda^i \alpha$, $i = 0, 1, \dots$, $0 < \lambda < 1$, до тех пор, пока впервые не выполнится условие (9).

5) Возможно априорное задание величин α_k из условий

$$\alpha_k > 0, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \quad (10)$$

Например, в качестве α_k можно взять $\alpha_k = c(k+1)^{-\alpha}$, где $c = \text{const} > 0$, а число α таково, что $1/2 < \alpha \leq 1$. В частности, если $\alpha = 1$, $c = 1$, то получим $\alpha_k = (k+1)^{-1}$, $k = 0, 1, \dots$. 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 $J_* = \inf_{E^n} J(u) > -\infty$, то 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 света. Здесь часто используются следующие критерии:

$$|u_k - u_{k+1}| \leq \varepsilon, \quad \text{или} \quad |J(u_k) - J(u_{k+1})| \leq \delta, \quad \text{или} \quad |J'(u_k)| \leq \gamma,$$

где $\varepsilon, \delta, \gamma$ — 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 ко множеству точек минимума

$$U_* = \{u \in E^n, \quad J(u) = J_* = \inf_{E^n} J(u)\},$$

или иначе говоря, выполняются ли соотношения

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0? \quad (11)$$

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 $J_* = \inf_{E^n} J(u) > -\infty$, $J(u) \in C^{1,1}(E^n)$.

Тогда последовательность $\{u_k\}$, полученная методом (3), (6) при произвольном начальном приближении u_0 , такова, что $\lim_{k \rightarrow \infty} J'(u_k) = 0$. Если при этом множество $M_\delta(u_0) = \{u \in E^n$:

$J(u) \leq J(u_0) + \delta\}$, где δ взято из (6), ограничено, то $\lim_{k \rightarrow \infty} \rho(u_k, S_*) = 0$, где $S_* = \{u \in M_\delta(u_0) : J'(u) = 0\}$ — множество стационарных точек функции $J(u)$ на $M_\delta(u_0)$.

Доказательство. Если при некотором $k \geq 0$ окажется, что $J'(u_k) = 0$, то из (3), (6) формально получаем $u_k = u_{k+1} = \dots$ и утверждения теоремы становятся тривиальными. Поэтому будем считать, что $J'(u_k) \neq 0$ при всех $k = 0, 1, \dots$. Так как $J(u_{k+1}) = f_k(\alpha_k) \leq \inf_{\alpha \geq 0} f_k(\alpha) + \delta_k \leq J(u_k - \alpha J'(u_k)) + \delta_k$ при всех $\alpha \geq 0$, то из неравенства (2.3.7) при $v = u_k$, $u = u_k - \alpha J'(u_k)$ имеем

$$\begin{aligned} J(u_k) - J(u_{k+1}) &\geq J(u_k) - J(u_k - \alpha J'(u_k)) - \delta_k \geq \\ &\geq \alpha |J'(u_k)|^2 - L\alpha^2 |J'(u_k)|^2 / 2 - \delta_k \geq \alpha (1 - \alpha L/2) |J'(u_k)|^2 - \delta_k \end{aligned}$$

при всех $\alpha \geq 0$ и $k = 0, 1, \dots$. Следовательно,

$$\begin{aligned} J(u_k) - J(u_{k+1}) &\geq \max_{\alpha \geq 0} \alpha (1 - \alpha L/2) |J'(u_k)|^2 - \delta_k = \\ &= (1/2L) |J'(u_k)|^2 - \delta_k, \quad k = 0, 1, \dots \end{aligned} \quad (12)$$

Отсюда получаем

$$J(u_{k+1}) \leq J(u_k) + \delta_k, \quad k = 0, 1, \dots \quad (13)$$

Так как $J(u_k) \geq J_* > -\infty$, $k = 0, 1, \dots$, то из леммы 2.3.2 и (13) следует существование предела $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Тогда

$\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0$ и из (12) будем иметь $\lim_{k \rightarrow \infty} J'(u_k) = 0$.

Наконец, пусть множество $M_\delta(u_0)$ ограничено. Суммируя неравенства (13) по k от 0 до $m-1$, получим

$$J(u_m) \leq J(u_0) + \sum_{k=0}^{m-1} \delta_k \leq J(u_0) + \delta, \quad m = 1, 2, \dots,$$

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 $\{u_{k_m}\} \rightarrow u_*$. Пользуясь непрерывностью $J'(u)$, отсюда имеем $\lim_{m \rightarrow \infty} J'(u_{k_m}) = J'(u_*) = 0$, т. е. $u_* \in S_*$. Так как расстояние $\rho(u, S_*)$ непрерывно (см. лемму 2.1.2), то $\lim_{m \rightarrow \infty} \rho(u_{k_m}, S_*) = \rho(u_*, S_*) = 0$. Отсюда следует, что числовая последовательность $\{\rho(u_k, S_*)\}$ имеет единственную предельную точку, равную нулю, т. е. $\lim_{k \rightarrow \infty} \rho(u_k, S_*) = 0$. 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

$$0 \leq J(u_k) - J_* \leq c_0 k^{-1}, \quad c_0 = \text{const} > 0. \quad (14)$$

Доказательство. Из ограниченности $M_\delta(u_0)$, непрерывности $J(u)$ согласно теореме 2.1.2 имеем $J_* > -\infty$, $U_* \neq \emptyset$, $U_* \subset M_\delta(u_0)$. Тогда для любой точки $u_* \in U_*$ с помощью неравенства (4.2.4) получаем

$$0 \leq J(u_k) - J_* = J(u_k) - J(u_*) \leq \langle J'(u_k), u_k - u_* \rangle \leq |J'(u_k)| \cdot |u_k - u_*| \leq d |J'(u_k)|, \quad k = 0, 1, \dots, \quad (15)$$

где $d \geq \text{diam } M_\delta(u_0) = \sup_{u, v \in M_\delta(u_0)} |u - v|$ — диаметр множества $M_\delta(u_0)$. В теореме 1 было доказано, что $\lim_{k \rightarrow \infty} J'(u_k) = 0$. Отсюда и из (15) следует, что $\lim_{k \rightarrow \infty} J(u_k) = J_*$. 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 $a_k - a_{k+1} = J(u_k) - J(u_{k+1}) \geq (1/2L) |J'(u_k)|^2 - \delta_k \geq a_k^2 / (2Ld^2) - \delta_k$. По условию $\delta_k = O(k^{-2})$, т. е. $0 \leq \delta_k \leq c_1 k^{-2}$, $k = 1, 2, \dots$, $c_1 = \text{const} > 0$. Полагая $A = \max\{c_1; 2Ld^2\}$, получим

$$a_{k+1} \leq a_k - a_k^2/A + Ak^{-2}, \quad k = 1, 2, \dots$$

Отсюда и из леммы 2.3.5 при $I_0 = \{1, 2, \dots\}$, $I_1 = \emptyset$ следует оценка (14). Если $\delta_k = 0$, $k = 0, 1, \dots$, 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 $\delta_k = 0$, $k = 0, 1, \dots$, то верна более сильная, чем (14), оценка

$$0 \leq J(u_k) - J_* \leq (J(u_0) - J_*) q^k, \quad (16)$$

$$|u_k - u_*|^2 \leq (2/\mu)(J(u_0) - J_*) q^k, \quad k = 0, 1, \dots, \quad (17)$$

где u_* — точка минимума $J(u)$ на E^n , $q = 1 - \mu/L$, $0 \leq q < 1$, 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

$$\begin{aligned} a_k = J(u_k) - J(u_*) &\leq \langle J'(u_k), u_k - u_* \rangle - \kappa |u_k - u_*|^2 \leq \\ &\leq |J'(u_k)| \cdot |u_k - u_*| - \kappa |u_k - u_*|^2 \leq \sup_{z \geq 0} (|J'(u_k)|z - \kappa z^2) = \\ &= |J'(u_k)|^2 / 4\kappa, \end{aligned}$$

т. е.

$$a_k = J(u_k) - J(u_*) \leq |J'(u_k)|^2 / 4\kappa, \quad k = 0, 1, \dots \quad (18)$$

Подставив неравенство (18) в правую часть (12) при $\delta_k = 0$, получим

$$a_k - a_{k+1} \geq (4\kappa/2L) a_k = (2\kappa/L) a_k, \quad k = 0, 1, \dots$$

В § 4.3 было установлено, что $2\kappa = \mu \leq L$. Поэтому $0 \leq q = 1 - (\mu/L) < 1$, и предыдущее неравенство можно переписать в виде $0 \leq a_{k+1} \leq a_k(1 - \mu/L) = qa_k$. Отсюда имеем $a_k \leq qa_{k-1} \leq q^2 a_{k-2} \leq \dots \leq q^k a_0$, что равносильно оценке (16). Наконец, из неравенства (4.3.2) следует

$$\kappa |u_k - u_*|^2 \leq J(u_k) - J(u_*) = a_k, \quad k = 0, 1, \dots$$

Отсюда и из (16) получаем оценку (17). Теорема 3 доказана.

Метод скорейшего спуска имеет простой геометрический смысл: оказывается, точка u_{k+1} , определяемая условиями (3), (4), лежит на луче $L_k = \{u: u = u_k - \alpha J'(u_k), \alpha \geq 0\}$ в точке его касания линии уровня (при $n \geq 3$ — поверхности уровня) $\Gamma_{k+1} = \{u \in E^n: J(u) = J(u_{k+1})\}$, а сам луч L_k перпендикулярен линии уровня $\Gamma_k = \{u \in E^n: J(u) = J(u_k)\}$ — см. рис. 5.1 и 5.2. В самом деле, пусть $u = u(t)$, $a \leq t \leq b$, — некоторое параметрическое уравнение линии уровня Γ_k , т. е. $J(u(t)) = J(u_k) = \text{const}$, $a \leq t \leq b$, причем $u(t_0) = u_k$. Тогда $\frac{d}{dt} J(u(t)) = \langle J'(u(t)), \dot{u}(t) \rangle = 0$, $a \leq t \leq b$. 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 $f'_k(\alpha_k) = -\langle J'(u_k - \alpha_k J'(u_k)), J'(u_k) \rangle = -\langle J'(u_{k+1}), J'(u_k) \rangle = 0$. 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

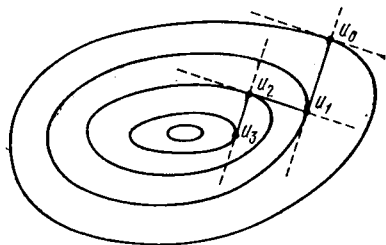


Рис. 5.1.

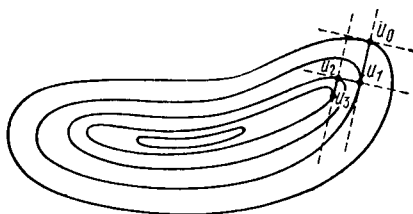


Рис. 5.2.

(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 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 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 другом «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 $\bar{u}_0, \bar{u}_1$, из которых производят спуск с помощью какого-либо варианта градиентного метода и получают две точки u_0, u_1 на «дне оврага». Затем полагают

$$\bar{u}_2 = u_1 - (u_1 - u_0) |u_1 - u_0|^{-1} h \operatorname{sign} (J(u_1) - J(u_0)),$$

где h — положительная постоянная, называемая овражным шагом. Из точки $\bar{u}_2$, которая, вообще говоря, находится на «склоне

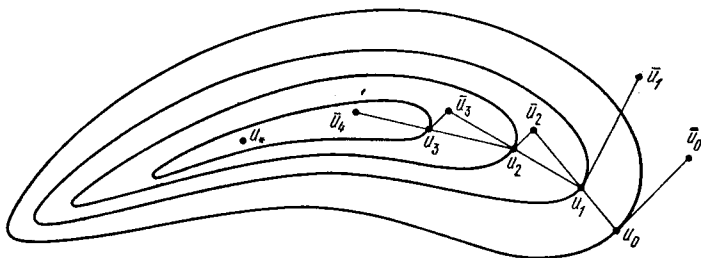


Рис. 5.3.

оврага», производят спуск с помощью градиентного метода и определяют следующую точку u_2 на «дне оврага».

Если уже известны точки $u_0, u_1, \dots, u_k$ ($k \geq 2$), то из точки $\bar{u}_{k+1} = u_k - (u_k - u_{k-1}) |u_k - u_{k-1}|^{-1} h \operatorname{sign} [J(u_k) - J(u_{k-1})]$ совершают спуск с помощью градиентного метода и находят следующую точку u_{k+1} на «дне оврага» (см. рис. 5.3; спуск из точки $\bar{u}_k$ в точку u_k , состоящий, быть может, из нескольких итерационных шагов градиентного метода, на рис. 5.3 условно изображен отрезком прямой, соединяющей точки $\bar{u}_k, u_k, k=0, 1, \dots$).

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 $\bar{u}_k$ могут слишком удаляться от «дна оврага» и спуск из точки $\bar{u}_k$ в точку u_k может потребовать большого объема вычислений. Кроме того, при больших h на крутых поворотах может произойти выброс точки $\bar{u}_k$ 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, реа-

гирующей 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 $\bar{u}_k$ от «дна оврага» и тем самым сократить объем вычислений, требуемый на градиентный спуск из точки $\bar{u}_k$ в точку u_k , $k = 0, 1, 2, \dots$. 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 [214] предлагается следующий способ выбора овражного шага:

$$h_{k+1} = h_k \cdot c^{\cos \alpha_k - \cos \alpha_{k-1}}, \quad k = 2, 3, \dots, \quad (19)$$

где α_k — угол между векторами $\bar{u}_k - u_{k-1}$, $u_k - u_{k-1}$, определяемый условием

$$\cos \alpha_k = \langle \bar{u}_k - u_{k-1}, u_k - u_{k-1} \rangle / |\bar{u}_k - u_{k-1}|^{-1} |u_k - u_{k-1}|^{-1},$$

а постоянная $c > 1$ является параметром алгоритма. Точка $\bar{u}_{k+1}$ тогда определяется так:

$$\bar{u}_{k+1} = u_k - \frac{u_k - u_{k-1}}{|u_k - u_{k-1}|} \times \\ \times h_{k+1} \operatorname{sign} [J(u_k) - J(u_{k-1})].$$

Разность $\cos \alpha_k - \cos \alpha_{k-1}$ в равенстве (19) 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.

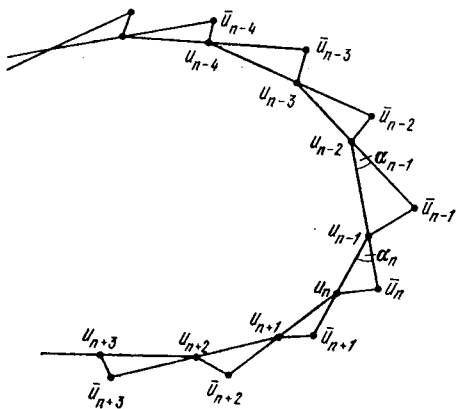


Рис. 5.4.

рис. 5.4). Тогда в силу (19) $h_{k+1} < h_k$, т. е. овражный шаг уменьшается, приспособляясь к повороту «дна оврага», что в свою очередь приводит к уменьшению выбросов точки $\bar{u}_{k+1}$ 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.

Параметр c в равенстве (19) 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 [214]. Выражение (19) для овражного шага удобнее преобразовать так:

$$h_{k+1} = h_k c^{\cos \alpha_k - \cos \alpha_{k-1}} = h_{k-1} c^{\cos \alpha_k - \cos \alpha_{k-2}} = \dots = h_2 c^{\cos \alpha_k - \cos \alpha_1},$$

откуда имеем

$$h_{k+1} = A c^{\cos \alpha_k}, \quad A = h_2 c^{-\cos \alpha_1} = \text{const} > 0, \quad k = 2, 3, \dots$$

Другой способ ускорения сходимости градиентного метода заключается в выборе подходящей замены переменных $u = g(\xi) = (g_1(\xi), \dots, g_n(\xi))$ с тем, чтобы поверхности уровня функции $J(g(\xi)) = G(\xi)$ в пространстве переменных $\xi = (\xi^1, \dots, \xi^n)$ были близки к сферам. Заметим, что $G'(\xi) = (g'(\xi))^T J'(g(\xi))$, где $g'(\xi) = \{g_{i\xi^j}(\xi)\}$ — матрица, i -я строка которой представляет собой $g'_i(\xi) = (g_{i\xi^1}(\xi), \dots, g_{i\xi^n}(\xi))$, а $(g'(\xi))^T$ — матрица, полученная из $g'(\xi)$ транспонированием. В пространстве переменных ξ градиентный метод выглядит так:

$$\xi_{k+1} = \xi_k - \beta_k (g'(\xi_k))^T J'(g(\xi_k)), \quad \beta_k > 0, \quad k = 0, 1, \dots$$

В пространстве исходных переменных $u = (u^1, \dots, u^n)$ этот подход можно трактовать как итерационный процесс вида

$$u_{k+1} = u_k - \alpha_k A_k J'(u_k), \quad \alpha_k > 0, \quad k = 0, 1, \dots,$$

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 $A_k = (J''(u_k))^{-1}$, $k = 0, 1, \dots$. О методах минимизации овражных функций, о различных приемах ускорения сходимости итерационных методов см. [2, 15, 39, 182, 201, 236, 253].

4. 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

$$(u(t + \Delta t) - u(t))/\Delta t = -\beta(t) J'(u(t)), \quad t \geq 0; \quad u(0) = u_0.$$

Отсюда, формально переходя к пределу при $\Delta t \rightarrow +0$, приходим к следующей задаче Коши:

$$\dot{u}(t) = -\beta(t) J'(u(t)), \quad t \geq 0; \quad u(0) = u_0. \quad (20)$$

Задача (20) представляет собой непрерывный аналог градиентного метода, а исходный процесс (3) является методом ломаных Эйлера для решения задачи (20). Понятно, что задачу Коши (20) можно решать и другими численными методами, которые, возможно, будут сходиться быстрее метода Эйлера и лучше приспособлены для минимизации овражных функций [2, 40, 201].

Определение 1. Траекторию (решение) $u(t)$ задачи (20) будем называть *минимизирующей*, если $u(t)$ определена при всех $t \geq 0$ и $\lim_{t \rightarrow +\infty} J(u(t)) = J_*$.

Ограничимся следующей теоремой о сходимости метода (20).

Теорема 4. Пусть функция $J(u)$ сильно выпукла на E^n и $J(u) \in C^{1,1}(E^n)$, функция $\beta(t)$ определена, непрерывна и $\beta(t) \geq \beta_0 > 0$ при всех $t \geq 0$. Тогда траектория задачи (20) при любом выборе начальной точки u_0 является минимизирующей и сходится к точке минимума u_* с оценкой

$$|u(t) - u_*| \leq |u_0 - u_*| e^{-\mu\beta_0 t}, \quad t \geq 0, \quad (21)$$

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 (20) определена при всех $t \geq 0$ для любой начальной точки u_0 . Положим

$$V(t) = |u(t) - u_*|^2/2, \quad t \geq 0. \quad (22)$$

Тогда с учетом условий (20) и теоремы 4.3.3 имеем

$$\begin{aligned} \dot{V}(t) &= \langle u(t) - u_*, \dot{u}(t) \rangle = \\ &= -\beta(t) \langle J'(u(t)) - J'(u_*), u(t) - u_* \rangle \leq -\mu\beta_0 |u(t) - u_*|^2 = \\ &= -2\mu\beta_0 V(t), \quad t \geq 0; \quad V(0) = |u_0 - u_*|^2/2. \end{aligned}$$

Отсюда следует $\frac{d}{dt}(V(t)e^{+2\mu\beta_0 t}) \leq 0, t \geq 0$. Интегрируя это неравенство, получим,

$$0 \leq V(t) \leq V(0)e^{-2\mu\beta_0 t} = |u_0 - u_*|^2 e^{-2\mu\beta_0 t}/2,$$

что равносильно оценке (21). В силу непрерывности функции $J(u)$ тогда $\lim_{t \rightarrow \infty} J(u(t)) = J_*$, что и требовалось.

Пользуясь терминологией, принятой в теории устойчивости обыкновенных дифференциальных уравнений [6, 137, 193], мож-

но сказать, что в теореме 4 доказана асимптотическая устойчивость системы (20) относительно точки равновесия u_* этой системы. Для доказательства этого факта был использован второй метод Ляпунова, в качестве функции Ляпунова была взята функция (22). 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 (22) 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 [35].

Существуют и другие дифференциальные уравнения, траектории которых являются минимизирующими. Например, так называемый метод тяжелого шарика [2] заключается в рассмотрении системы дифференциальных уравнений

$$\ddot{u}(t) + \gamma \dot{u}(t) + J'(u(t)) = 0, \quad t \geq 0, \quad \gamma = \text{const} > 0.$$

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 (20).

5. 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, усиливается 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

$$J(u) \rightarrow \inf; \quad u \in U \subseteq E^n, \quad (1)$$

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

$$u_{k+1} = P_U(u_k - \alpha_k J'(u_k)), \quad k = 0, 1, \dots, \quad (2)$$

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 $u_k = P_U(u_k - \alpha_k J'(u_k))$ (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 $f_k(\alpha) = J(P_U(u_k - \alpha J'(u_k)))$, $\alpha \geq 0$, и определим α_k из условий

$$f_k(\alpha_k) = \inf_{\alpha \geq 0} f_k(\alpha) = f_{k*}, \quad \alpha_k > 0. \quad (3)$$

Очевидно, при $U = E^n$ метод (2), (3) превратится в метод скорейшего спуска. Поскольку величину α_k из условий (3) удастся 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

$$0 < \varepsilon_0 \leq \alpha_k \leq 2/(L + 2\varepsilon), \quad (4)$$

где $\varepsilon_0, \varepsilon$ — положительные числа, являющиеся параметрами метода.

4) Возможен выбор α_k из условия

$$J(u_k) - J(P_U(u_k - \alpha_k J'(u_k))) \geq \varepsilon \|u_k - P_U(u_k - \alpha_k J'(u_k))\|^2, \quad (5)$$

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

$$\alpha_k > 0, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < +\infty, \quad (6)$$

например, $\alpha_k = (k+1)^{-1}$, $k = 0, 1, \dots$. 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

$$\begin{aligned} u_{k+1} &= u_k + \beta_k (P_U(u_k - \alpha_k J'(u_k)) - u_k) = \\ &= \beta_k P_U(u_k - \alpha_k J'(u_k)) + (1 - \beta_k) u_k, \quad 0 < \beta_k \leq 1, \quad \alpha_k > 0, \quad (2') \end{aligned}$$

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 . Однако задача отыскания проекции некоторой точки u на множество U сама, в свою очередь, является задачей минимизации функции $g(v) = |v - u|$ или $g(v) = |v - u|^2$ на этом множестве и далеко не всегда просто решается. Поэтому методом проекции градиента обычно пользуются лишь в тех случаях, когда проекция точки на множество легко определяет-

ся. 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.

Полезно заметить, что задача определения проекции (2) точки $u_k - \alpha_k J'(u_k)$ на выпуклое замкнутое множество U равносильна следующей задаче минимизации квадратичной функции:

$$\Phi_k(u) = |u - u_k|^2/2 + \alpha_k \langle J'(u_k), u - u_k \rangle \rightarrow \inf; \quad u \in U. \quad (7)$$

В самом деле, функция $\Phi_k(u)$ сильно выпукла и по теореме 4.3.1 достигает своей нижней грани на U в единственной точке $u_{k+1} \in U$. В силу теоремы 4.2.3 для этого необходимо и достаточно, чтобы $\langle \Phi'_k(u_{k+1}), u - u_{k+1} \rangle \geq 0$ при всех $u \in U$. Отсюда, учитывая, что

$$\Phi'_k(u) = u - u_k + \alpha_k J'(u_k),$$

имеем

$$\langle u_{k+1} - [u_k - \alpha_k J'(u_k)], u - u_{k+1} \rangle \geq 0 \quad (8)$$

для любого $u \in U$. Согласно теореме 4.4.1 неравенство (8) означает, что точка u_{k+1} является проекцией точки $u_k - \alpha_k J'(u_k)$ на множество U . Равносильность задачи (7) и равенства (2) установлена.

2. Остановимся на вопросах сходимости метода (2), (4).

Теорема 1. Пусть U — выпуклое замкнутое множество из E^n , функция $J(u) \in C^{1,1}(U)$, $\inf_U J(u) = J_* > -\infty$. Тогда для последовательности $\{u_k\}$, полученной методом (2), (4) при любом начальном приближении u_0 , имеет место соотношение $\lim_{k \rightarrow \infty} |u_{k+1} - u_k| = 0$. Если при этом множество $M(u_0) = \{u: u \in U, J(u) \leq J(u_0)\}$ ограничено, то $\lim_{k \rightarrow \infty} \rho(u_k, S_*) = 0$, где $S_* = \{u: u \in M(u_0), \langle J'(u), v - u \rangle \geq 0 \text{ при всех } v \in U\}$ — 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

$$J(u_k) - J(u_{k+1}) \geq \langle J'(u_k), u_k - u_{k+1} \rangle - (L/2) |u_k - u_{k+1}|^2, \quad k=0, 1, \dots \quad (9)$$

Перепишем неравенство (8) в виде

$$\langle J'(u_k), u - u_{k+1} \rangle \geq \langle u_k - u_{k+1}, u - u_{k+1} \rangle / \alpha_k, \quad u \in U, \quad k=0, 1, \dots \quad (10)$$

Отсюда при $u = u_k$ с учетом условия (4) получим

$$\langle J'(u_k), u_k - u_{k+1} \rangle \geq |u_k - u_{k+1}|^2 / \alpha_k \geq (L/2 + \varepsilon) |u_k - u_{k+1}|^2.$$

Подставим эту оценку в (9):

$$J(u_k) - J(u_{k+1}) \geq \varepsilon |u_k - u_{k+1}|^2, \quad k = 0, 1, \dots \quad (11)$$

Так как $J(u_k) \geq J_* > -\infty$ и последовательность $\{J(u_k)\}$ — убывающая, то существует конечный предел $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$ и, следовательно, $\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0$. Отсюда и из (11) сразу получим $\lim_{k \rightarrow \infty} |u_k - u_{k+1}| = 0$.

Пусть теперь множество $M(u_0)$ ограничено. Так как согласно (11) $J(u_{k+1}) \leq J(u_k) \leq \dots \leq J(u_0)$, 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 $\{u_{k_m}\} \rightarrow u_*$. По доказанному $\lim_{k \rightarrow \infty} |u_{k+1} - u_k| = 0$, поэтому $\{u_{k_m+1}\} \rightarrow u_*$. Переходя в (10) к пределу при $k = k_m \rightarrow \infty$, с учетом условий (4) и непрерывности $J'(u)$ получим $\langle J'(u_*), u - u_* \rangle \geq 0$ при любом $u \in U$, т. е. $u_* \in S_*$. По лемме 2.1.2 расстояние $\rho(u, S_*)$ непрерывно по u , поэтому $\lim_{m \rightarrow \infty} \rho(u_{k_m}, S_*) = \rho(u_*, S_*) = 0$. Отсюда следует, что $\{\rho(u_k, S_*)\}$ имеет единственную предельную точку, равную нулю, т. е. $\lim_{k \rightarrow \infty} \rho(u_k, S_*) = 0$. 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

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0, \quad (12)$$

причем справедлива оценка

$$0 \leq J(u_k) - J_* \leq C_0 k^{-1}, \quad C_0 = \text{const} > 0, \quad k = 1, 2, \dots \quad (13)$$

Доказательство. Из ограниченности $M(u_0)$, непрерывности $J(u)$ согласно теореме 2.1.2 следует $J_* > -\infty$, $U_* = \{u: u \in U, J(u) = J_*\} \neq \emptyset$, $U_* \subset M(u_0)$. Возьмем произвольную точку $u_* \in U_*$. Из неравенства (4.2.4) тогда имеем

$$\begin{aligned} 0 \leq \alpha_k &= J(u_k) - J(u_*) \leq \langle J'(u_k), u_k - u_* \rangle = \\ &= \langle J'(u_k), u_k - u_{k+1} \rangle - \langle J'(u_k), u_* - u_{k+1} \rangle, \quad k = 0, 1, \dots \end{aligned}$$

Пользуясь неравенством (10) при $u = u_*$ и условием (4) выбора α_k , отсюда получим

$$0 \leq \alpha_k \leq \langle J'(u_k), u_k - u_{k+1} \rangle - \langle u_k - u_{k+1}, u_* - u_{k+1} \rangle / \alpha_k \leq \\ \leq |u_k - u_{k+1}| (\sup_{M(u_k)} |J'(u)| + D/\varepsilon_0) = C_1 |u_k - u_{k+1}|, \quad k = 0, 1, \dots \quad (14)$$

Здесь мы учли ограниченность множества $M(u_0)$, поэтому $D = \sup_{u, v \in M(u_0)} |u - v| < \infty$, и кроме того, $|J'(u)| \leq |J'(u) - J'(u_0)| + |J'(u_0)| \leq L|u - u_0| + |J'(u_0)| \leq LD + |J'(u_0)|$ при любом $u \in M(u_0)$, так что $\sup_{M(u_k)} |J'(u)| < \infty$. Из (11), (14)

следует $\alpha_k - \alpha_{k+1} \geq \varepsilon C_1^{-2} \alpha_k^2 = A \alpha_k^2$, $k = 0, 1, \dots$. 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 , $\mu \leq L$, взяты из (2.3.6), (4.3.8). Тогда последовательность $\{u_k\}$, получаемая из (2) при $\alpha_k = \alpha$, $k = 0, 1, \dots$, сходится к точке минимума u_* , причем справедлива оценка

$$|u_k - u_*| \leq |u_0 - u_*| (q(\alpha))^k, \quad k = 0, 1, \dots, \quad (15)$$

где $q(\alpha) = (1 - 2\mu\alpha + \alpha^2 L^2)^{1/2}$, $0 < q(\alpha) < 1$.

Доказательство. Введем отображение

$$Au = P_U(u - \alpha J'(u)),$$

действующее из U в U . Покажем его сжимаемость при $0 < \alpha < 2\mu L^{-2}$. С помощью теоремы 4.4.2 имеем

$$|Au - Av|^2 = |P_U(u - \alpha J'(u)) - P_U(v - \alpha J'(v))|^2 \leq \\ \leq |u - \alpha J'(u) - v + \alpha J'(v)|^2 = |u - v|^2 + \alpha^2 |J'(u) - J'(v)|^2 - \\ - 2\alpha \langle J'(u) - J'(v), u - v \rangle \leq |u - v|^2 (1 + \alpha^2 L^2 - 2\mu\alpha) = \\ = q^2(\alpha) |u - v|^2.$$

т. е.

$$|Au - Av| \leq q(\alpha) |u - v|, \quad u, v \in U. \quad (16)$$

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 ([10], стр. 74).

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 [10], что такая точка u_* существует, единственна и $\lim_{k \rightarrow \infty} |u_k - u_*| = 0$. Из (16) следует, что

$$|u_k - u_m| \leq (q(\alpha))^k |u_0 - u_{m-k}| \quad \text{при всех } m \geq k.$$

Отсюда при $m \rightarrow \infty$ получим оценку (15). Так как $u_* = P_U(u_* - \alpha J'(u_*))$, 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 [21], рассмотрим сходимость метода (2), (4), не требуя, в отличие от теорем 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 , $J(u) \in C^{1,1}(U)$, $J_* > -\infty$, $U_* \neq \emptyset$. Пусть вместо точного значения градиента $J'(u)$ и проекции $P_U(u) \equiv P(u)$ известны их приближения $J'_k(u)$ и соответственно $P_k(u)$ с погрешностью

$$|J'(u) - J'_k(u)| \leq \delta_k, \quad u \in U; \quad |P(u) - P_k(u)| \leq C_0 \delta_k, \quad (17)$$

$$u \in E^n, \quad C_0 = \text{const} > 0, \quad \delta_k \geq 0, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty.$$

Наконец, пусть последовательность $\{u_k\}$ определяется условиями

$$u_{k+1} = P_k(u_k - \alpha_k J'_k(u_k)), \quad u_0 \in U, \quad k = 0, 1, \dots, \quad (18)$$

где α_k выбирается так:

$$0 < \varepsilon_0 \leq \alpha_k \leq 2(1 - \varepsilon)/L, \quad k = 0, 1, \dots, \quad 0 < \varepsilon < 1. \quad (19)$$

Тогда $\{u_k\}$ сходится к некоторой точке $v_* \in U_*$.

Доказательство. Наряду с $\{u_k\}$ введем вспомогательную последовательность $\{v_k\}$, определяемую следующим образом:

$$v_{k+1} = P(u_k - \alpha_k J'(u_k)), \quad k = 0, 1, \dots; \quad v_0 = u_0. \quad (20)$$

Отсюда и из (18) с помощью теоремы 4.4.2 и условий (17) получаем

$$\begin{aligned} |u_{k+1} - v_{k+1}| &\leq |P_k(u_k - \alpha_k J'_k(u_k)) - P(u_k - \alpha_k J'(u_k))| + \\ &\quad + |P(u_k - \alpha_k J'_k(u_k)) - P(u_k - \alpha_k J'(u_k))| \leq \\ &\leq C_0 \delta_k + \alpha_k |J'_k(u_k) - J'(u_k)| \leq C_0 \delta_k + (2(1 - \varepsilon)/L) \delta_k = C_1 \delta_k, \quad (21) \\ &k = 0, 1, \dots \end{aligned}$$

Возьмем произвольную точку $u_* \in U_*$. Согласно теореме 4.2.3 тогда

$$\langle J'(u_*), u - u_* \rangle \geq 0, \quad u \in U. \quad (22)$$

Так же, как неравенство (8), из (20) получаем

$$\langle v_{k+1} - u_k + \alpha_k J'(u_k), u - v_{k+1} \rangle \geq 0, \quad u \in U. \quad (23)$$

Положим в (23) $u = u_*$, а (22) умножим на $\alpha_k > 0$ и примем $u = v_{k+1}$. Сложим получившиеся неравенства

$$\langle v_{k+1} - u_k, u_* - v_{k+1} \rangle + \alpha_k \langle J'(u_k) - J'(u_*), u_* - v_{k+1} \rangle \geq 0, \quad (24)$$

$$k = 0, 1, \dots$$

Преобразуем каждое слагаемое в правой части (24). Прежде всего имеем

$$2 \langle v_{k+1} - u_k, u_* - v_{k+1} \rangle = |u_k - u_*|^2 - |u_k - v_{k+1}|^2 - |v_{k+1} - u_*|^2. \quad (25)$$

Далее, воспользуемся неравенством (4.2.19):

$$|J'(u_k) - J'(u_*)|^2 \leq L \langle J'(u_k) - J'(u_*), u_k - u_* \rangle,$$

из которого следует

$$\begin{aligned} \langle J'(u_k) - J'(u_*), u_* - v_{k+1} \rangle &= \\ &= \langle J'(u_k) - J'(u_*), u_k - v_{k+1} \rangle - \langle J'(u_k) - J'(u_*), u_k - u_* \rangle \leq \\ &\leq \langle J'(u_k) - J'(u_*), u_k - v_{k+1} \rangle - |J'(u_k) - J'(u_*)|^2 L^{-1} = \\ &= (L/4) |u_k - v_{k+1}|^2 - \{L^{-1/2} [J'(u_k) - J'(u_*)] - (L^{1/2}/2)(u_k - v_{k+1})\}^2 \leq \\ &\leq (L/4) |u_k - v_{k+1}|^2, \quad k = 0, 1, \dots \quad (26) \end{aligned}$$

Подставив (25), (26) в (24), получим

$$|u_k - u_*|^2 - |v_{k+1} - u_*|^2 - (1 - \alpha_k L/2) |u_k - v_{k+1}|^2 \geq 0.$$

Отсюда, учитывая условие (19), имеем

$$|u_k - u_*|^2 \geq |v_{k+1} - u_*|^2 + \varepsilon |u_k - v_{k+1}|^2, \quad k = 0, 1, \dots \quad (27)$$

Далее, воспользуемся леммой 2.3.10 при $z_k = u_k$, $z_* = u_*$, $w_k = v_k$. Тогда из (21), (27) получим

$$\lim_{k \rightarrow \infty} |u_k - u_*| = \lim_{k \rightarrow \infty} |v_k - u_*| < \infty, \quad \lim_{k \rightarrow \infty} |v_{k+1} - u_k| = 0. \quad (28)$$

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 (21) и (28) имеем $\lim_{m \rightarrow \infty} v_{k_m+1} = \lim_{m \rightarrow \infty} u_{k_m+1} = v_*$. Согласно (23)

$$\begin{aligned} \langle J'(u_k), u - v_{k+1} \rangle &\geq - \langle v_{k+1} - u_k, u - v_{k+1} \rangle \alpha_k^{-1} \geq \\ &\geq - |v_{k+1} - u_k| \cdot |u - v_{k+1}| \varepsilon_0^{-1}. \end{aligned}$$

Отсюда при $k = k_m \rightarrow \infty$ будем иметь $\langle J'(v_*), u - v_* \rangle \geq 0$ 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.

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 неутраченной 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 $\alpha_k = \alpha$, $k = 0, 1, \dots$, $0 < \alpha < 2/L$. Тогда

$$|u_k - u_*| \leq (q(\alpha))^k |u_0 - u_*|, \quad k = 0, 1, \dots, \quad (29)$$

где

$$q(\alpha) = \begin{cases} 1 - \mu\alpha & \text{при } 0 < \alpha < 2(L + \mu)^{-1}, \\ L\alpha - 1 & \text{при } 2(L + \mu)^{-1} \leq \alpha < 2L^{-1}, \end{cases}$$

$0 < q(\alpha) < 1$, 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 $\lim_{k \rightarrow \infty} |u_k - u_*| = 0$.

Здесь мы предполагаем, что в (17) $\delta_k = 0$, $k = 0, 1, \dots$, поэтому из (18), (20), (21) следует $u_{k+1} = u_{k+1}$, $k = 0, 1, \dots$. Учитывая последнее равенство и условие $\alpha_k = \alpha$, подставим (25) в (24). Получим

$$\begin{aligned} |u_{k+1} - u_*|^2 &\leq |u_k - u_*|^2 - |u_{k+1} - u_k|^2 + 2\alpha \langle J'(u_k) - J'(u_*), u_* - u_{k+1} \rangle = \\ &= |u_k - u_*|^2 - |u_k - u_{k+1} - \alpha(J'(u_k) - J'(u_*))|^2 + \\ &+ \alpha^2 |J'(u_k) - J'(u_*)|^2 - 2\alpha \langle J'(u_k) - J'(u_*), u_k - u_* \rangle, \quad k = 0, 1, \dots \end{aligned} \quad (30)$$

Вспомним неравенства (4.3.12), (4.3.13), из которых имеем

$$\begin{aligned} |J'(u_k) - J'(u_*)|^2 + L\mu |u_k - u_*|^2 &\leq \\ &\leq (L + \mu) \langle J'(u_k) - J'(u_*), u_k - u_* \rangle, \quad k = 0, 1, \dots, \end{aligned} \quad (31)$$

$$\mu |u_k - u_*| \leq |J'(u_k) - J'(u_*)| \leq L |u_k - u_*|, \quad k = 0, 1, \dots \quad (32)$$

Из (30), (31) следует

$$\begin{aligned} |u_{k+1} - u_*|^2 &\leq |u_k - u_*|^2 + \alpha^2 |J'(u_k) - J'(u_*)|^2 - \\ &- 2\alpha(L + \mu)^{-1} |J'(u_k) - J'(u_*)|^2 - 2\alpha L\mu(L + \mu)^{-1} |u_k - u_*|^2 = \\ &= \alpha[\alpha - 2(L + \mu)^{-1}] |J'(u_k) - J'(u_*)|^2 + \\ &+ [1 - 2\alpha L\mu(L + \mu)^{-1}] |u_k - u_*|^2, \quad k = 0, 1, \dots \end{aligned} \quad (33)$$

Рассмотрим два случая: 1) если $0 < \alpha < 2(L + \mu)^{-1} \leq \mu^{-1}$, то из (33) и левого неравенства (32) имеем $|u_{k+1} - u_*|^2 \leq |u_k - u_*|^2 [\alpha(\alpha - 2(L + \mu)^{-1})\mu^2 + 1 - 2\alpha L\mu(L + \mu)^{-1}] = (1 - \alpha\mu)^2 |u_k - u_*|^2$;

2) если $L^{-1} \leq 2(L + \mu)^{-1} \leq \alpha < 2L^{-1}$, то из (33) и правого неравенства (32) получим $|u_{k+1} - u_*|^2 \leq |u_k - u_*|^2 [\alpha(\alpha - 2(L + \mu)^{-1})L^2 + 1 - 2\alpha L\mu(L + \mu)^{-1}] = (L\alpha - 1)^2 |u_k - u_*|^2$. Объединяя оба случая, имеем $|u_{k+1} - u_*| \leq q(\alpha) |u_k - u_*|$, $k = 0, 1, \dots$, 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 $q(\alpha_*) = (L - \mu)(L + \mu)^{-1}$. Теорема 5 доказана.

Приведем пример, показывающий, что оценка (29) неулучшаема на классе сильно выпуклых функций из $C^{1,1}(U)$.

Пример 1. Пусть $u = (x, y) \in U = E^2$, $J(u) = (Lx^2 + \mu y^2)/2$, $0 < \mu \leq L$. Ясно, что эта функция сильно выпукла с константой $\kappa = \mu/2$, 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

$$x_{k+1} = x_k - \alpha L x_k = (1 - \alpha L) x_k, \quad y_{k+1} = y_k - \alpha \mu y_k = (1 - \alpha \mu) y_k, \\ k = 0, 1, \dots$$

Положим здесь $\alpha = 2(L + \mu)^{-1}$, $q = (L - \mu)(L + \mu)^{-1}$. Тогда $x_{k+1} = -q x_k$, $y_{k+1} = q y_k$. 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

$$J(u) = (x - 1)^2 + (y + 1)^2, \quad -$$

$$u \in U = E^2 =$$

$$= \{u = (x, y) \in E^2 : x \geq 0, y \geq 0\}.$$

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

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

можно предложить следующий итерационный метод:

$$u_{k+1} = P_U(u_k - \alpha_k c_k), \quad \alpha_k > 0, \quad c_k \in \partial J(u_k), \quad k = 0, 1, \dots, \quad (2)$$

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).

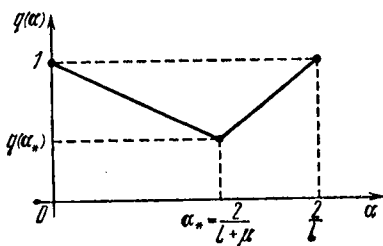


Рис. 5.5.

В самом деле, при $u_k = P_U(u_k - \alpha_k c_k)$ согласно теореме 4.4.1 $\langle u_k - (u_k - \alpha_k c_k), u - u_k \rangle = \langle c_k, u - u_k \rangle \alpha_k \geq 0$ или $\langle c_k, u - u_k \rangle \geq 0$ при всех $u \in U$. А тогда по определению субградиента

$$J(u) - J(u_k) \geq \langle c_k, u - u_k \rangle \geq 0, \quad u \in U,$$

т. е. $J(u_k) = \inf_U J(u) = J_*$.

В том случае, когда функция $J(u)$ дифференцируема во всех точках $u \in U$, то 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

$$\alpha_k > 0, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty. \quad (3)$$

Как уже отмечалось в § 1, в качестве α_k можно взять $\alpha_k = C(k+1)^{-\alpha}$, где $C = \text{const} > 0$, $1/2 < \alpha \leq 1$; например, $\alpha_k = (k+1)^{-1}$, $k = 0, 1, \dots$

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,

$$\sup_{u \in U} \sup_{c \in \partial J(u)} |c| = A < +\infty. \quad (4)$$

Тогда последовательность $\{u_k\}$, определяемая условиями (2), (3), такова, что

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0. \quad (5)$$

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 $M(C) = \{u \in U: J(u) \leq C\}$ ограничено при любом $C \geq J_*$. 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

$$\begin{aligned} \rho^2(u_{k+1}, U_*) &= |u_{k+1} - P_{U_*}(u_{k+1})|^2 \leq |u_{k+1} - P_{U_*}(u_k)|^2 = \\ &= |P_U(u_k - \alpha_k c_k) - P_U(P_{U_*}(u_k))|^2 \leq |u_k - \alpha_k c_k - P_{U_*}(u_k)|^2 = \\ &= \rho^2(u_k, U_*) + \alpha_k^2 |c_k|^2 - 2\alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle \end{aligned}$$

или

$$2\alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle \leq \alpha_k^2 |c_k|^2 + \rho^2(u_k, U_*) - \rho^2(u_{k+1}, U_*), \quad (6)$$

$$k = 0, 1, \dots$$

Суммируя неравенства (6) по k от 0 до некоторого $m \geq 1$, с учетом условий (3), (4) получим

$$\begin{aligned} \sum_{k=0}^m 2\alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle &\leq \\ &\leq A^2 \sum_{k=0}^m \alpha_k^2 + \rho^2(u_0, U_*) - \rho^2(u_{m+1}, U_*) \leq \\ &\leq A^2 \sum_{k=0}^{\infty} \alpha_k^2 + \rho^2(u_0, U_*) = B < \infty, \quad m = 1, 2, \dots \end{aligned} \quad (7)$$

Далее, по определению субградиента имеем

$$0 \leq J(u_k) - J_* = J(u_k) - J(P_{U_*}(u_k)) \leq \langle c_k, u_k - P_{U_*}(u_k) \rangle, \quad (8)$$

$$k = 0, 1, \dots$$

Из (7), (8) следует, что ряд

$$\sum_{k=0}^{\infty} \alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle$$

сходится. Но согласно (3) $\sum_{k=0}^{\infty} \alpha_k = +\infty$. Поэтому сходимость предыдущего ряда возможна лишь при $\lim_{k \rightarrow \infty} \langle c_k, u_k - P_{U_*}(u_k) \rangle = 0$. Это значит, что существуют номера $k_1 < k_2 < \dots < k_m < \dots$ такие, что

$$\lim_{m \rightarrow \infty} \langle c_{k_m}, u_{k_m} - P_{U_*}(u_{k_m}) \rangle = 0. \quad (9)$$

Тогда из (8) при $k = k_m \rightarrow \infty$ получим $\lim_{m \rightarrow \infty} J(u_{k_m}) = J_*$. Кроме того, из (8), (9) следует, что $J(u_{k_m}) \leq J_* + \sup_m \langle c_{k_m}, u_{k_m} - P_{U_*}(u_{k_m}) \rangle = C < \infty$, 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 $\lim_{m \rightarrow \infty} \rho(u_{k_m}, U_*) = 0$.

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 $\{u_k\}$. С этой целью введем множества номеров $I_0 = \{k: k \geq 1, \rho(u_{k-1}, U_*) < \rho(u_k, U_*)\}$ и $I_1 = \{0, 1, 2, \dots\} \setminus I_0$. Сначала рассмотрим случай, когда I_0 содержит бесконечно много номеров. Пусть $I_0 = \{k_1, k_2, \dots, k_m, \dots\}$, $k_m < k_{m+1}$, $m = 1, 2, \dots$. Тогда из (6) имеем $2\alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle \leq \alpha_k^2 A^2$, или $0 \leq \langle c_k, u_k - P_{U_*}(u_k) \rangle \leq \alpha_k (A^2/2) \rightarrow 0$ при $k \rightarrow \infty$, $k \in I_0$. Это значит, что подпоследовательность $\{u_{k_m}, k_m \in I_0\}$ удовлетворяет условию (9). По доказанному тогда $J(u_{k_m}) \rightarrow J_*$, $\rho(u_{k_m}, U_*) \rightarrow 0$ при $m \rightarrow \infty$. Поэтому для любого $\varepsilon > 0$ найдется номер N такой, что $\rho(u_{k_m}, U_*) < \varepsilon$ для всех $m \geq N$. Если $k \in I_0$ при всех $k > k_N$, то равенства (5) уже доказаны. Пусть среди $k > k_N$ существует номер $k \in I_1$. Возьмем произвольный номер $k \in I_1$, $k > k_N$. Тогда найдутся $k_m, k_{m+1} \in I_0$ такие, что $k_N \leq k_m < k < k_{m+1}$. По

определению I_1 будем иметь

$$\rho(u_k, U_*) \leq \rho(u_{k-1}, U_*) \leq \dots \leq \rho(u_{k_m}, U_*) < \varepsilon,$$

или $\rho(u_k, U_*) < \varepsilon$ для всех $k \in I_1$, $k > k_N$.

Тем самым установлено, что $\rho(u_k, U_*) < \varepsilon$ для всех $k > k_N$ независимо от того, будет ли $k \in I_0$ или $k \in I_1$. Это значит, что $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$.

Покажем, что тогда $\lim_{k \rightarrow \infty} J(u_k) = J_*$. По условию множество U_* ограничено. Тогда последовательность $\{u_k\}$, сходящаяся к U_* , также ограничена. Возьмем любую предельную точку u_* этой последовательности. Пусть $\{u_{k_r}\} \rightarrow u_*$. Так как U_* — замкнутое множество и $\lim_{r \rightarrow \infty} \rho(u_{k_r}, U_*) = \rho(u_*, U_*) = 0$, то $u_* \in U_*$. А тогда $\lim_{r \rightarrow \infty} J(u_{k_r}) = J(u_*) = J_*$. Это означает, что числовая последовательность $\{J(u_k)\}$ имеет единственную предельную точку, равную J_* , т. е. $\lim_{k \rightarrow \infty} J(u_k) = J_*$.

Случай, когда множество I_0 бесконечно, полностью рассмотрен. Теперь пусть I_0 конечно, пусть $k_N = \max_{I_0} k$ (если $I_0 = \emptyset$, то $k_N = 0$). Тогда $k \in I_1$ при всех $k \geq k_N$. По определению I_1 имеем

$$\rho(u_k, U_*) \geq \rho(u_{k+1}, U_*) \text{ для всех } k \geq k_N.$$

Следовательно, существует $\lim_{k \rightarrow \infty} \rho(u_k, U_*)$. Выше было показано существование подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (9), и было установлено, что тогда $\lim_{m \rightarrow \infty} \rho(u_{k_m}, U_*) = 0$. Отсюда следует, что $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. Тогда, как и в предыдущем случае, получаем $\lim_{k \rightarrow \infty} J(u_k) = J_*$.

Теорема 1 доказана.

З а м е ч а н и е 1. В условии теоремы 1 предполагается выполнение условия (4). В том случае, когда U ограничено, то, оказывается, условие (4) всегда выполняется. Покажем это. Сначала установим, что $\inf_{v \in U} \rho(v, \text{Гр } W) > 0$,

где $\text{Гр } W$ — граница множества W . От противного: пусть существуют $v_k \in U$, $k = 1, 2, \dots$, такие, что $\rho(v_k, \text{Гр } W) \rightarrow 0$ при $k \rightarrow \infty$. Так как U — компактное множество, то можем считать, что $\{v_k\} \rightarrow v \in U$. Из непрерывности $\rho(v, \text{Гр } W)$ (см. лемму 2.1.2) тогда $\lim_{k \rightarrow \infty} \rho(v_k, \text{Гр } W) = \rho(v, \text{Гр } W) = 0$, т. е. $v \in \text{Гр } W$.

С другой стороны, поскольку $v \in U \subset W$ и W — открытое множество, то существует число $\delta > 0$ такое, что $O(v, \delta) = \{u \in E^n: |u - v| < \delta\} \subset W$, т. е. $\rho(v, \text{Гр } W) \geq \delta$. Полученное противоречие показывает, что $\inf_{v \in U} \rho(v, \text{Гр } W) = 2\varepsilon > 0$. Это значит, что замкнутое ограниченное множе-

ство $U_\varepsilon = \{u \in E^n: |u - v| \leq \varepsilon, v \in U\}$, представляющее собой ε -раздutie множества U , принадлежит W . В силу непрерывности $J(u)$ на U_ε тогда $\sup_{U_\varepsilon} J(u) = J_\varepsilon^* < \infty$. Возьмем произвольные $v \in U$ и $c \in \partial J(v)$. В неравен-

стве $J(u) - J(v) \geq \langle c, u - v \rangle$, $u \in W$, определяющем субградиент $c \in \partial J(v)$ (см. также теорему 4.6.1), положим $u = v + \varepsilon c / |c| \in U_\varepsilon$. Получим $c \leq (J(v + \varepsilon c / |c|) - J(v)) \varepsilon^{-1} \leq 2J_\varepsilon^* \varepsilon^{-1} = A < \infty$ для всех $c \in \partial J(v)$ и всех $v \in U$, что равносильно условию (4).

2. Описанный выше метод проекции субградиента после некоторой модификации можно использовать для решения следующей задачи выпуклого программирования:

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in E^n: u \in U_0, \quad g_i(u) \leq 0, \quad i = \overline{1, m}\}. \quad (10)$$

Заметим, что система неравенств $g_i(u) \leq 0, \quad i = \overline{1, m}$, равносильна одному неравенству $g(u) \leq 0$, где $g(u) = \max_{1 \leq i \leq m} g_i(u), \quad u \in U$. Кроме того, из выпуклости функций $g_i(u), \quad i = \overline{1, m}$, на U_0 следует выпуклость $g(u)$ на U_0 (см. теорему 4.2.7). Поэтому задачу (10) можно переформулировать в виде эквивалентной задачи

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in U_0, \quad g(u) \leq 0\}, \quad (11)$$

также являющейся задачей выпуклого программирования.

Предположим, что субдифференциалы $\partial J(u), \quad \partial g(u)$ не пусты при всех $u \in U_0$. Следуя [251], рассмотрим метод

$$u_{k+1} = P_{U_0}(u_k - \alpha_k c_k), \quad k = 0, 1, \dots; \quad u_0 \in U_0, \quad (12)$$

где $\{\alpha_k\}$ выбирается из условий

$$\alpha_k > 0, \quad k = 0, 1, \dots, \quad \sum_{k=0}^{\infty} \alpha_k^{1+\gamma} = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \quad 0 < \gamma < 1, \quad (13)$$

а субградиенты $\{c_k\}$ таковы, что

$$c_k \in \partial J(u_k) \quad \text{при} \quad g(u_k) \leq \alpha_k^\gamma \quad \text{и} \quad c_k \in \partial g(u_k) \quad \text{при} \quad g(u_k) > \alpha_k^\gamma. \quad (14)$$

Таким образом, метод (12) — (14) работает так: если ограничение $g(u) \leq 0$ при $u = u_k$ не нарушено или нарушено ненамного, то минимизируем функцию $J(u)$, а если нарушение этого ограничения велико, то минимизируем функцию $g(u)$. Если функции $J(u), \quad g(u)$ 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 $\alpha = 3/5, \quad \gamma = 1/2, \quad C = 1$ получим $\alpha_k = (k+1)^{-3/5}, \quad k = 0, 1, \dots$

Теорема 2. Пусть U_0 — выпуклое замкнутое множество из E^n , функции $J(u), \quad g(u)$ определены и выпуклы на некотором открытом выпуклом множестве W , содержащем U_0 (например, $W = E^n$). Пусть $J_* = \inf_U J(u) > -\infty$, 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 $J(u), \quad g(u)$ непрерывны на U_0 , субдифференциалы $\partial J(u), \quad \partial g(u)$ 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

$$M(C_1, C_2) = \{u \in U_0: J(u) \leq C_1, \quad g(u) \leq C_2\}$$

ограничено при всех $C_1 > \inf_{U_0} J(u)$, $C_2 > \inf_U g(u)$. В самом деле, $M(J_*, 0) = U_*$ ограничено по условию. Тогда по теореме 4.2.17 множество $M(C_1, 0) = \{u: u \in U_0, g(u) \leq 0, J(u) \leq C_1\}$ ограничено при всех $C_1 > \inf_{U_0} J(u)$. Теперь, фиксируя любое C_1 , по той же теореме 4.2.17 получаем ограниченность $M(C_1, C_2)$ при каждом $C_2 > \inf_{U_0} g(u)$.

Нетрудно видеть, что неравенства (6), (7) сохраняют силу и для метода (12) — (14). Из (7) имеем

$$\sum_{k=0}^r \alpha_k^{1+\gamma} \alpha_k^{-\gamma} \langle c_k, u_k - P_{U_*}(u_k) \rangle \leq B < \infty, \quad r = 1, 2, \dots \quad (15)$$

Отсюда следует существование номеров $k_1 < k_2 < \dots < k_m < \dots$ таких, что

$$\langle c_{k_m}, u_{k_m} - P_{U_*}(u_{k_m}) \rangle \leq \alpha_{k_m}^{\gamma}, \quad m = 1, 2, \dots \quad (16)$$

В самом деле, допустим, что (16) не имеет места. Тогда $\langle c_k, u_k - P_{U_*}(u_k) \rangle > \alpha_k^{\gamma}$ при всех $k = 0, 1, \dots$. Отсюда и из (15) имеем

$$\sum_{k=0}^r \alpha_k^{1+\gamma} \leq \sum_{k=0}^r \alpha_k \langle c_k, u_k - P_{U_*}(u_k) \rangle \leq B < \infty, \quad r = 1, 2, \dots,$$

что противоречит расходимости ряда $\sum_{k=0}^{\infty} \alpha_k^{1+\gamma}$.

Тем самым показано существование подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (16). Докажем, что

$$\lim_{m \rightarrow \infty} J(u_{k_m}) = J_*, \quad \lim_{m \rightarrow \infty} \rho(u_{k_m}, U_*) = 0. \quad (17)$$

Сначала убедимся в том, что

$$c_{k_m} \in \partial J(u_{k_m}), \quad m = 1, 2, \dots \quad (18)$$

Для этого достаточно показать, что $g(u_{k_m}) \leq \alpha_{k_m}^{\gamma}$, $m = 1, 2, \dots$, и вспомнить условия (14). Допустим, что $g(u_{k_m}) > \alpha_{k_m}^{\gamma}$ при некотором $m \geq 1$. Учитывая, что тогда $c_{k_m} \in \partial g(u_{k_m})$ и, кроме того, $P_{U_*}(u_{k_m}) \in U_* \subset U$, т. е. $g(P_{U_*}(u_{k_m})) \leq 0$, из (16) имеем

$$\alpha_{k_m}^{\gamma} < g(u_{k_m}) \leq g(u_{k_m}) - g(P_{U_*}(u_{k_m})) \leq \langle c_{k_m}, u_{k_m} - P_{U_*}(u_{k_m}) \rangle \leq \alpha_{k_m}^{\gamma}.$$

Получили противоречивое неравенство. Включения (18) доказаны, и попутно установлено, что

$$g(u_{k_m}) \leq \alpha_{k_m}^{\gamma}, \quad m = 1, 2, \dots \quad (19)$$

Множество номеров $\{k_m\}$, удовлетворяющих условиям (16), представим в виде объединения непересекающихся множеств $I_2 = \{k_m: J(u_{k_m}) \geq J_*\}$ и $I_3 = \{k_m: J(u_{k_m}) < J_*\}$.

Сначала рассмотрим случай, когда множество I_2 бесконечно. Из (16), (18) имеем

$$0 \leq J(u_{k_m}) - J_* = J(u_{k_m}) - J(P_{U_*}(u_{k_m})) \leq \langle c_{k_m}, u_{k_m} - P_{U_*}(u_{k_m}) \rangle \leq \alpha_{k_m}^y, \quad k_m \in I_2.$$

Отсюда следует, что $J(u_{k_m}) \rightarrow J_*$ при $m \rightarrow \infty$, $k_m \in I_2$. Тогда $J(u_{k_m}) \leq C_1 < \infty$, и кроме того, согласно (13), (19) $g(u_{k_m}) \leq \alpha_{k_m}^y \leq \sup_{k \geq 0} \alpha_k^y = C_2 < \infty$, т. е. $\{u_{k_m}\} \in M(C_1, C_2)$, $k_m \in I_2$. Так как $M(C_1, C_2)$ ограничено, то $\{u_{k_m}, k_m \in I_2\}$ имеет хотя бы одну предельную точку. Не умаляя общности, можем считать, что $\{u_{k_m}\} \rightarrow u_*$, $k_m \in I_2$. Из замкнутости U_0 , неравенства (19) и непрерывности $g(u)$ следует, что $u_* \in U$. Но $J(u_{k_m}) \rightarrow J(u_*) = J_*$ при $k_m \rightarrow \infty$, $k_m \in I_2$, так что $u_* \in U_*$. Отсюда и из непрерывности $\rho(u, U_*)$ имеем $\rho(u_{k_m}, U_*) \rightarrow \rho(u_*, U_*) = 0$ при $k_m \rightarrow \infty$, $k_m \in I_2$.

Теперь рассмотрим случай, когда множество I_3 бесконечно. Если $k_m \in I_3$, то $J(u_{k_m}) < J_*$, а $g(u_{k_m}) \leq \alpha_{k_m}^y \leq \sup_{k \geq 0} \alpha_k^y = C_2$ в силу (19). Это значит, что $\{u_{k_m}\} \in M(J_*, C_2)$, $k_m \in I_3$. Поскольку множество $M(J_*, C_2)$ ограничено, то $\{u_{k_m}, k_m \in I_3\}$ имеет предельную точку. Не умаляя общности, можем считать, что $\{u_{k_m}\} \rightarrow u_*$, $k_m \in I_3$. Из (19) и замкнутости U_0 следует, что $u_* \in U$. Поэтому $J(u_*) \geq J_*$. С другой стороны, $J(u_{k_m}) < J_*$, $k_m \in I_3$, так что $\lim_{k_m \in I_3, m \rightarrow \infty} J(u_{k_m}) = J(u_*) \leq J_*$. Следовательно, $J(u_*) = J_*$, т. е. $u_* \in U_*$. Отсюда и из непрерывности $\rho(u, U_*)$ получаем $\rho(u_{k_m}, U_*) \rightarrow \rho(u_*, U_*) = 0$ при $k_m \rightarrow \infty$, $k_m \in I_3$.

Объединяя оба рассмотренных случая, заключаем, что для подпоследовательности $\{u_{k_m}\}$, удовлетворяющей условию (16), справедливы равенства (17). Далее остается ввести множества I_0, I_1 , как это было сделано при доказательстве теоремы 1, и провести прежние рассуждения, используя вместо (9) неравенство (16).

На этом закончим рассмотрение методов минимизации негладких выпуклых функций. Отметим, что негладкие и невыпуклые задачи в последние годы интенсивно исследуются, продолжается разработка различных методов их решения [65, 89, 97—99, 112, 180, 228, 250, 253].

У п р а ж н е н и я. 1. Рассмотреть возможность применения метода проекции субградиента к задачам из упражнений 4.6.1 и 4.6.4.

2. Описать метод (12)—(14) применительно к задаче

$$J(u) = |x + y| + |x - y| \rightarrow \inf, \quad u \in U = \{u = (x, y) \in E^2: u \in U_0, \\ g(u) = u^2 - 1 \leq 0\}, \quad U_0 = \{u = (x, y): x \geq 0, y \geq 0\}.$$

3. Проверить условия теоремы 2 для задачи

$$J(u) = |\langle c, u \rangle| \rightarrow \inf; \quad u \in U = \{u \in E^n: u \geq 0, g(u) = |\langle a, u \rangle| - 1 \leq 0\},$$

где $a, c \in E^n$. Описать метод (12)—(14) применительно к этой задаче.

4. 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

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

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 $u_k \in U$, $k \geq 0$, 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

$$J_k(u) = \langle J'(u_k), u - u_k \rangle, \quad (2)$$

и определим вспомогательное приближение $\bar{u}_k$ из условий

$$\bar{u}_k \in U, \quad \inf_U J_k(u) = J_k(\bar{u}_k) = \langle J'(u_k), \bar{u}_k - u_k \rangle. \quad (3)$$

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

$$U = \{u \in E^n: u \geq 0, \quad \langle a_i, u \rangle \leq b^i, \\ i = \overline{1, m}; \quad \langle a_i, u \rangle = b^i, \quad i = \overline{m+1, s}\},$$

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

$$U = \{u = (u^1, \dots, u^n): \alpha_i \leq u^i \leq \beta_i, \quad i = \overline{1, n}\}$$

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 $\bar{u}_k = (\bar{u}_k^1, \dots, \bar{u}_k^n)$, где

$$\bar{u}_k^i = \begin{cases} \alpha_i, & \text{если } J_{u^i}(u_k) > 0, \\ \beta_i, & \text{если } J_{u^i}(u_k) < 0; \end{cases}$$

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

$$U = \{u \in E^n: |u - v_0| \leq r\}$$

— шар радиуса r с центром в точке v_0 , то с помощью неравенства Коши — Буняковского $\langle J'(u_k), u \rangle = \langle J'(u_k), u - v_0 \rangle + \langle J'(u_k), v_0 \rangle \geq -|J'(u_k)|r + \langle J'(u_k), v_0 \rangle$, $u \in U$, получаем, что

$$\bar{u}_k = v_0 - rJ'(u_k)|J'(u_k)|^{-1}.$$

Разумеется, так просто получить вспомогательное приближение $\bar{u}_k$ удастся далеко не всегда, и вместо точного решения задачи (3) часто приходится довольствоваться определением какого-либо приближенного решения. А именно, будем предполагать, что $\bar{u}_k$ определяется из следующих условий:

$$\bar{u}_k \in U, \quad J_k(\bar{u}_k) \leq \min_U J_k(u) + \varepsilon_k, \quad \varepsilon_k \geq 0, \quad \lim_{k \rightarrow \infty} \varepsilon_k = 0. \quad (4)$$

Допустим, что точка $\bar{u}_k$, удовлетворяющая условиям (4) (или (3)), уже найдена. Тогда следующее $(k+1)$ -е приближение будем искать в виде

$$u_{k+1} = u_k + \alpha_k(\bar{u}_k - u_k), \quad 0 \leq \alpha_k \leq 1. \quad (5)$$

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

$$J_k(\bar{u}_k) = J_k(u_k) = 0 = \min_U J_k(u) \quad \text{или} \quad \langle J'(u_k), u - u_k \rangle \geq 0$$

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 $-\varepsilon_k \leq \min_U J_k(u) \leq J_k(\bar{u}_k) = J_k(u_k) = 0$, 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 получать различные варианты описанного метода, часто имеваемого в литературе методом условного градиента. Укажем некоторые наиболее употребительные способы выбора α_k в (5).

1) Величина α_k может выбираться из условий

$$0 \leq \alpha_k \leq 1, \quad f_k(\alpha_k) = \min_{0 \leq \alpha \leq 1} f_k(\alpha) = f_{k*}, \quad f_k(\alpha) = J(u_k + \alpha(\bar{u}_k - u_k)). \quad (6)$$

Для некоторых классов задач удастся 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 , $b \in E^n$. Тогда $J'(u_k) = Au_k - b$. Пользуясь формулой (4.2.10), имеем

$$f_k(\alpha) = J(u_k) + \alpha \langle J'(u_k), \bar{u}_k - u_k \rangle + (\alpha^2/2) \langle A(\bar{u}_k - u_k), \bar{u}_k - u_k \rangle. \quad (7)$$

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

$$\alpha_k^* = - \langle J'(u_k), \bar{u}_k - u_k \rangle / (\langle A(\bar{u}_k - u_k), \bar{u}_k - u_k \rangle)^{-1}.$$

Рассматривая возможные случаи $\alpha_k^* < 0$, $0 \leq \alpha_k^* \leq 1$, $\alpha_k^* > 1$, из условий (6) тогда получаем

$$\alpha_k = \begin{cases} 0, & \text{если } \alpha_k^* < 0; \\ \alpha_k^*, & \text{если } 0 \leq \alpha_k^* \leq 1; \\ 1, & \text{если } \alpha_k^* > 1. \end{cases} \quad (8)$$

Кстати, если точка $\bar{u}_k$ в (7) найдена из условий (3), то $J_k(\bar{u}_k) \leq J_k(u_k) = 0$ и, следовательно, $\alpha_k^* \geq 0$ — 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

$$0 \leq \alpha_k \leq 1; \quad f_k(\alpha_k) \leq f_{k*} + \delta_k, \quad \delta_k \geq 0, \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty \quad (9)$$

или [9]

$$0 \leq \alpha_k \leq 1, \quad f_k(\alpha_k) \leq (1 - \lambda_k) f_k(0) + \lambda_k f_{k*}, \quad 0 < \bar{\lambda} \leq \lambda_k \leq 1.$$

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

$$\alpha_k = \begin{cases} \min \{1; \rho_k |J'_k(\bar{u}_k)| |\bar{u}_k - u_k|^{-2}\}, & \text{если } J'_k(\bar{u}_k) \leq 0, \\ 0, & \text{если } J'_k(\bar{u}_k) > 0, \end{cases} \quad (10)$$

где

$$0 < \varepsilon_0 \leq \rho_k \leq 2(1 - \varepsilon)/L, \quad (11)$$

$\varepsilon_0, \varepsilon$ — параметры метода, $0 < \varepsilon < 1$.

3) Другой способ выбора α_k : при $J'_k(\bar{u}_k) > 0$ полагают $\alpha_k = 0$, а если $J'_k(\bar{u}_k) \leq 0$, то $\alpha_k = \lambda^i$, где i_0 — минимальный номер среди номеров $i \geq 0$, удовлетворяющих условию [15]

$$J(u_k) - J(\bar{u}_k + \lambda^i (\bar{u}_k - u_k)) \geq \lambda^i \varepsilon |J'_k(\bar{u}_k)|,$$

где λ, ε — параметры метода, $0 < \lambda; \varepsilon < 1$.

4) Величины α_k в (5) можно априорно задавать из условий

$$0 < \alpha_k \leq 1, \quad \lim_{k \rightarrow \infty} \alpha_k = 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad (12)$$

например, $\alpha_k = (k+1)^{-1}$, $k = 0, 1, \dots$ (предложен М. Ячимо-вичем). Такой выбор α_k очень прост для реализации на ЭВМ, но, вообще говоря, не гарантирует выполнение условия монотонности: $J(u_{k+1}) < J(u_k)$.

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.

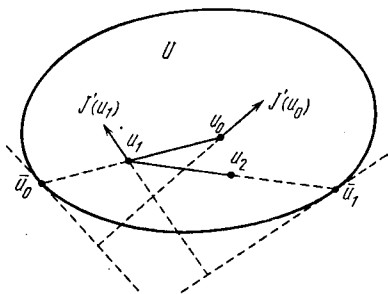


Рис. 5.6

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

$$\lim_{k \rightarrow \infty} \langle J'(u_k), \bar{u}_k - u_k \rangle = 0, \quad \lim_{k \rightarrow \infty} \rho(u_k, S_*) = 0, \quad (13)$$

где $S_* = \{u: u \in U, \langle J'(u), v - u \rangle \text{ при всех } v \in U\}$ — множество стационарных точек функции $J(u)$ на U .

Если, кроме перечисленных условий, $J(u)$ выпукла на U и

$$\varepsilon_k + \delta_k \leq C_0 k^{-2\rho}, \quad C_0 = \text{const} > 0, \quad 1/2 < \rho \leq 1, \quad (14)$$

то

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0, \quad (15)$$

и справедлива оценка

$$0 \leq J(u_k) - J_* \leq C_1 k^{-\rho}, \quad k = 1, 2, \dots; \quad C_1 = \text{const} \geq 0. \quad (16)$$

Наконец, если, кроме того, $J(u)$ сильно выпукла на U , то

$$\|u_k - u_*\|^2 \leq (C_1/\chi) k^{-\rho}, \quad k = 1, 2, \dots \quad (17)$$

Доказательство. При сделанных предположениях $J_* > -\infty$, $U_* \neq \emptyset$. Так как множество U ограничено, то $\sup_{u, v \in U} \|u - v\| \leq d < \infty$. Из условия (9) следует $J(u_{k+1}) = f_k(\alpha_k) \leq f_{k*} + \delta_k \leq J(u_k + \alpha(\bar{u}_k - u_k)) + \delta_k$ при всех α , $0 \leq \alpha \leq 1$. Поэтому, пользуясь неравенством (2.3.7), имеем

$$\begin{aligned} J(u_k) - J(u_{k+1}) + \delta_k &\geq J(u_k) - J(u_k + \alpha(\bar{u}_k - u_k)) \geq \\ &\geq -\alpha \langle J'(u_k), \bar{u}_k - u_k \rangle - \alpha^2 L \|\bar{u}_k - u_k\|^2 / 2 \geq \\ &\geq -\alpha J_k(\bar{u}_k) - \alpha^2 L d^2 / 2, \quad 0 \leq \alpha \leq 1, \quad k = 0, 1, \dots \end{aligned} \quad (18)$$

Множество $N = \{0, 1, 2, \dots\}$ разобьем на два множества $N^+ = \{k: k \in N, \langle J'(u_k), \bar{u}_k - u_k \rangle > 0\}$ и $N^- = N \setminus N^+$. Так как $\inf_U J_k(u) \leq J_k(u_k) = 0$, то из (4) получаем $0 \leq J_k(\bar{u}_k) \leq \varepsilon_k$ при всех $k \in N^+$. Поэтому если N^+ — бесконечное множество, то $J_k(\bar{u}_k) \rightarrow 0$ при $k \rightarrow \infty$, $k \in N^+$.

Теперь пусть $k \in N^-$. Тогда из (18) имеем

$$0 \leq -J_k(\bar{u}_k) \leq (J(u_k) - J(u_{k+1}) + \delta_k) / \alpha + \alpha L d^2 / 2 \quad (19)$$

при всех α , $0 < \alpha < 1$, $k \in N^-$. Далее, из (9) следует, что $J(u_{k+1}) \leq J(u_k) + \delta_k$, $k = 0, 1, \dots$. Так как $J(u_k) \geq J_* > -\infty$, $k = 0, 1, \dots$, то из леммы 2.3.2 вытекает существование конечного предела $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Следовательно, $\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0$.

Если N^- — бесконечное множество, то при $k \rightarrow \infty$, $k \in N^-$, из (19) имеем

$$0 \leq \lim_{k \rightarrow \infty} |J_k(\bar{u}_k)| \leq \overline{\lim}_{k \rightarrow \infty} |J_k(\bar{u}_k)| \leq \alpha L d^2 / 2$$

при всех α , $0 < \alpha < 1$. Устремляя $\alpha \rightarrow +0$, отсюда получим $J_k(\bar{u}_k) \rightarrow 0$ при $k \rightarrow \infty$, $k \in N^-$. Объединяя оба случая $k \in N^+$ и $k \in N^-$, приходим к первому равенству (13). Так как U ограни-

чено и $\{u_k\} \in U$, то последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$, пусть $\{u_{k_m}\} \rightarrow u_*$. Согласно (4) $J_k(\bar{u}_k) - \varepsilon_k \leq \inf_U J_k(u) \leq \langle J'(u_k), u - u_k \rangle$ при всех $u \in U$ и $k = 0, 1, \dots$

Отсюда при $k = k_m \rightarrow \infty$ с учетом первого равенства (13) получим, что $\langle J'(u_*), u - u_* \rangle \geq 0$ 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

$$0 \leq a_k = J(u_k) - J(u_*) \leq \langle J'(u_k), u_k - u_* \rangle = -J_k(u_*) \leq -\min_U J_k(u) \leq -J_k(\bar{u}_k) + \varepsilon_k, \quad k = 0, 1, \dots \quad (20)$$

Отсюда и из первого равенства (13) следует $\lim_{k \rightarrow \infty} J(u_k) = J_*$, т.е. $\{u_k\}$ — минимизирующая последовательность. Из теоремы 2.1.1 тогда получаем $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. 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 $N = \{0, 1, 2, \dots\}$ разобьем на два множества $I_0 = \{k: k \in N, a_k \geq \varepsilon_k\}$, $I_1 = \{k: k \in N, 0 \leq a_k < \varepsilon_k\}$. Из оценки

$$0 \leq a_k - \varepsilon_k \leq -J_k(\bar{u}_k), \quad k \in I_0, \quad (21)$$

являющейся следствием неравенства (20), следует, что $I_0 \subseteq N^-$. Поэтому (18) можно переписать в виде

$$a_k - a_{k+1} \geq \alpha |J_k(\bar{u}_k)| - \alpha^2 L d^2 / 2 - \delta_k, \quad 0 \leq \alpha \leq 1, \quad k \in I_0. \quad (22)$$

Так как в силу (13) $\{|J_k(\bar{u}_k)|\}$ ограничена, то взяв при необходимости d еще большим, можем сделать $0 \leq \bar{a}_k = |J_k(\bar{u}_k)| \times \times d^{-2} L^{-1} \leq 1$ при всех $k = 0, 1, \dots$. Принимая в (22) $\alpha = \bar{a}_k$, получим

$$a_k - a_{k+1} \geq 1/(2Ld^2) |J_k(\bar{u}_k)|^2 - \delta_k, \quad k \in I_0.$$

Отсюда и из (21) с учетом условия (14) имеем

$$a_{k+1} \leq a_k - (a_k^2 - \varepsilon_k^2)/(2Ld^2) + \delta_k \leq a_k - a_k^2/(2Ld^2) + + (\sup_{k \geq 0} a_k) L^{-1} d^{-2} \varepsilon_k + \delta_k \leq a_k - a_k^2/A + Ak^{-2p}, \quad k \in I_0, \quad (23)$$

где $A = \max \{2Ld^2; (\sup_{k \geq 0} a_k) L^{-1} d^{-2} C_0; C_0\}$.

Если $k \in I_1$, то $0 \leq a_k < \varepsilon_k \leq C_0 k^{-2p}$. Кроме того, из (18) при $\alpha \rightarrow +0$ получим $a_k - a_{k+1} + \delta_k \geq 0$ или $a_{k+1} \leq a_k + \delta_k \leq$

$\leq a_k + C_0 k^{-2\rho}$ для всех $k = 0, 1, \dots$. 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 $e_k \leq C_0 k^{-2\rho}$, $C_0 = \text{const} > 0$, $0 < \rho \leq 1$, верна оценка (16). Для сильно выпуклой функции справедлива оценка (17).

Доказательство. Так же, как неравенство (18), нетрудно показать, что

$$J(u_k) - J(u_{k+1}) \geq -\alpha_k J_k(\bar{u}_k) - \alpha_k^2 L |\bar{u}_k - u_k|^2 / 2, \quad k = 0, 1, \dots \quad (24)$$

В соответствии с формулой (10), определяющей величину α_k , рассмотрим три возможных случая:

1) Если $J_k(\bar{u}_k) \leq 0$, $\alpha_k = 1 \leq \rho_k |J_k(\bar{u}_k)| |\bar{u}_k - u_k|^{-2}$, то из (24) с учетом (11) имеем

$$J(u_k) - J(u_{k+1}) \geq |J_k(\bar{u}_k)| - L \rho_k |J_k(\bar{u}_k)| / 2 \geq \varepsilon |J_k(\bar{u}_k)|. \quad (25)$$

2) Если $J_k(\bar{u}_k) \leq 0$, $\alpha_k = \rho_k |J_k(\bar{u}_k)| |\bar{u}_k - u_k|^{-2} < 1$, то из (24) с учетом (11) получаем

$$\begin{aligned} J(u_k) - J(u_{k+1}) &\geq \rho_k |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2} - L \rho_k^2 |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2} / 2 = \\ &= |J_k(\bar{u}_k)|^2 |\bar{u}_k - u_k|^{-2} \rho_k (1 - L \rho_k / 2) \geq \\ &\geq |J_k(\bar{u}_k)|^2 d^{-2} \varepsilon_0 \varepsilon, \quad d \geq \sup_{u, v \in U} |u - v|. \end{aligned} \quad (26)$$

3) Наконец, если $J_k(\bar{u}_k) > 0$, то согласно (10) $\alpha_k = 0$, и из (24) имеем

$$J(u_k) - J(u_{k+1}) \geq 0, \quad (27)$$

а из (4) следует

$$0 < J_k(\bar{u}_k) \leq e_k. \quad (28)$$

Из (25) — (27) вытекает, что последовательность $\{J(u_k)\}$ не возрастает. Так как $J(u_k) \geq J_* > -\infty$, то существует $\lim_{k \rightarrow \infty} J(u_k) = J_*$ и, следовательно, $\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0$. Отсюда и из (25), (26), (28) имеем $0 \leq |J_k(\bar{u}_k)| \leq \max \{e_k; \text{const} \cdot (J(u_k) - J(u_{k+1}))^{1/2}\} \rightarrow 0$ при $k \rightarrow \infty$. 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 $e_k \leq C_0 k^{-2\rho}$, $0 < \rho \leq 1$, докажем оценку (16). Предварительно заметим, что $0 \leq a_k = J(u_k) - J_* \leq \sup_U J(u) - J_* = C_2 < \infty$, поэтому

$$a_k^2 \leq \sup_{k \geq 0} a_k \cdot a_k \leq C_2 a_k, \quad k = 0, 1, \dots \quad (29)$$

Еще раз переберем рассмотренные выше три возможности.

1) Если $J_k(\bar{u}_k) \leq 0$, $\alpha_k = 1 \leq \rho_k |J_k(\bar{u}_k)| \cdot |\bar{u}_k - u_k|^{-2}$, то из (20), (25), (29) имеем $a_k - a_{k+1} \geq \varepsilon a_k - \varepsilon e_k$, или

$$a_{k+1} \leq a_k - \varepsilon a_k + \varepsilon e_k \leq a_k - a_k^2 (\varepsilon / C_2) + \varepsilon C_0 k^{-2\rho}. \quad (30)$$

2) Пусть $J_k(\bar{u}_k) \leq 0$, $\alpha_k = \rho_k |J_k(\bar{u}_k)| \cdot |\bar{u}_k - u_k|^{-2} < 1$. Здесь, в свою очередь, имеются две возможности: $a_k \geq e_k$ или $0 \leq a_k < e_k$. Если $a_k \geq e_k$, то из (20), (26), (29) получим $a_k - a_{k+1} \geq (a_k - e_k)^2 d^{-2} \varepsilon_0 \varepsilon \geq a_k^2 d^{-2} \varepsilon_0 \varepsilon - 2C_2 e_k d^{-2} \varepsilon_0 \varepsilon$ или

$$a_{k+1} \leq a_k - a_k^2 (\varepsilon_0 \varepsilon / d^2) + 2C_2 \varepsilon_0 \varepsilon d^{-2} C_0 k^{-2\rho}. \quad (31)$$

Если же $0 \leq a_k < e_k$, то достаточно воспользоваться более простым следствием (26): $a_{k+1} \leq a_k$. Последние два неравенства можно переписать в виде

$$0 \leq a_k \leq C_0 k^{-2\rho}, \quad a_{k+1} \leq a_k \leq a_k + C_0 k^{-2\rho}. \quad (32)$$

3) Наконец, пусть $J_k(\bar{u}_k) > 0$, $\alpha_k = 0$. Тогда из (20), (27) получим $0 \leq a_k \leq e_k$, $a_{k+1} \leq a_k$, 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 $u_0 \in U_0$ для последовательности $\{u_k\}$, определяемой условиями (4), (5), (12), справедливо равенство (15). Если при этом $\alpha_k = (k+1)^{-1}$, $e_k = C_3 (k+1)^{-1}$, $k = 0, 1, \dots$, то

$$0 \leq J(u_k) - J_* \leq C_4 \ln(k+1)/k, \quad k = 1, 2, \dots, \quad (33)$$

а если $\alpha_k = (k+1)^{-\beta}$, $e_k = C_3 (k+1)^{-\beta}$, $k = 0, 1, \dots$, $0 < \beta < 1$, то

$$0 \leq J(u_k) - J_* \leq C_4 k^{-\beta}, \quad k = 1, 2, \dots; \quad (34)$$

здесь C_3, C_4 — некоторые положительные постоянные.

Доказательство. Заметим, что неравенства (20), (24) не зависят от способа выбора α_k , $0 \leq \alpha_k \leq 1$, в (5), поэтому сохраняют силу и в рассматриваемом случае. Из них имеем $a_k - a_{k+1} \geq \alpha_k (a_k - e_k) - \alpha_k^2 L d^2 / 2$ или

$$a_{k+1} \leq (1 - \alpha_k) a_k + \alpha_k^2 L d^2 / 2 + \alpha_k e_k, \quad k = 0, 1, \dots$$

Отсюда с учетом свойств последовательностей $\{\alpha_k\}$, $\{e_k\}$ из (4), (12) заключаем, что $\{a_k\}$ удовлетворяет условиям леммы 2.3.6. Поэтому $\lim_{k \rightarrow \infty} a_k = 0$ или

$\lim_{k \rightarrow \infty} J(u_k) = J_*$. 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 $u \in U = \{u = (x, y) \in E^2: 0 \leq x \leq 1, -1 \leq y \leq 0\}$, 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 , $0 \leq t \leq t_0$, 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 β — положительное число, $\beta \leq t_0$.

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:

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in E^n: g_i(u) \leq 0, i = \overline{1, m}\}, \quad (1)$$

где функции $J(u)$, $g_i(u)$, $i = \overline{1, m}$, 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 $u_k \in U$, $k \geq 0$. Введем множество номеров

$$I_k = \{i: 1 \leq i \leq m, \quad g_i(u_k) = 0\}.$$

Возможно, что $I_k = \emptyset$, — это будет означать, что $g_i(u_k) < 0$ при всех $i = 1, m$, т. е. $u_k \in \text{int } U$ — такая возможность ниже не исключается. В пространстве переменных

$$z = (e, \sigma) = (e^1, \dots, e^n, \sigma) \in E^{n+1}$$

рассмотрим вспомогательную задачу

$$\sigma \rightarrow \inf, \quad z = (e, \sigma) \in W_k = \{(e, \sigma): \langle J'(u_k), e \rangle \leq \sigma, \\ \langle g'_i(u_k), e \rangle \leq \sigma, \quad i \in I_k; |e^j| \leq 1, \quad j = \overline{1, n}\}. \quad (2)$$

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 $e = (e^1, \dots, e^n)$. Далее, ясно, что точка $z = (e=0, \sigma=0) = (0, 0) \in W_k$, так что $W_k \neq \emptyset$ и $\inf_{W_k} \sigma = \sigma_k \leq 0$ при всех $k = 0, 1, \dots$

Очевидно, множество W_k замкнуто. Наконец, условия $|e^j| \leq 1$, $j = \overline{1, n}$, 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 $\sigma_k = \inf_{W_k} \sigma$. Выше было замечено, что $\sigma_k \leq 0$.

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

$$\langle J'(u_k), e_k \rangle \leq \sigma_k < 0, \quad \langle g'_i(u_k), e_k \rangle \leq \sigma_k < 0, \quad i \in I_k.$$

Отсюда ясно, что $e_k \neq 0$. Кроме того, для любого номера $i \in I_k$ имеем

$$g_i(u_k + \alpha e_k) = g_i(u_k + \alpha e_k) - g_i(u_k) = \langle g'_i(u_k), e_k \rangle \alpha + o(\alpha) \leq \\ \leq \alpha [\sigma_k + o(\alpha)/\alpha] < 0 \quad \text{при всех } \alpha, \quad 0 < \alpha < \alpha_i, \quad \alpha_i > 0.$$

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 $\alpha_0 = \min \{\alpha_1, \dots, \alpha_m\} > 0$. Тогда

$$g_i(u_k + \alpha e_k) < 0, \quad i = \overline{1, m}; \quad 0 < \alpha < \alpha_0,$$

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

$$J(u_k + \alpha e_k) - J(u_k) = \langle J'(u_k), e_k \rangle \alpha + o(\alpha) \leq \\ \leq \alpha [\sigma_k + o(\alpha)/\alpha] < 0 \quad \text{при всех } \alpha, \quad 0 < \alpha < \alpha_0.$$

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:

$$u_{k+1} = u_k + \alpha_k e_k, \quad 0 < \alpha_k \leq \beta_k, \quad (3)$$

где

$$\beta_k = \sup \{ \alpha: u_k + t e_k \in U, \quad 0 \leq t \leq \alpha \} > 0. \quad (4)$$

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:

$$0 < \alpha_k \leq \beta_k, \quad f_k(\alpha_k) = \inf_{0 < \alpha \leq \beta_k} f_k(\alpha) = f_{k*}; \quad f_k(\alpha) = J(u_k + \alpha e_k). \quad (5)$$

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 :

$$0 < \alpha_k \leq \beta_k, \quad f_k(\alpha_k) \leq f_{k*} + \delta_k, \quad \delta_k \geq 0, \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty \quad (6)$$

или

$$J(u_k + \alpha_k e_k) \leq (1 - \lambda_k) J(u_k) + \lambda_k f_{k*}, \quad 0 < \lambda \leq \lambda_k \leq 1.$$

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

$$J(u_k) - J(u_k + \alpha_k e_k) \geq \varepsilon \alpha_k |\sigma_k|, \quad 0 < \alpha_k \leq \beta_k, \quad 0 < \varepsilon < 1/2.$$

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 $\langle J'(u_k), e_k \rangle \leq \sigma$) и в то же время оставалось возможным направлением для множества U в точке u_k (это обеспечивается условиями $\langle g'_i(u_k), e_k \rangle \leq \sigma$, $i \in I_k$), причем чем меньше σ , тем ярче выражены указанные свойства направления e_k . Кстати, если $I_k = \emptyset$, т.е. $u_k \in \text{int } U$, то $e_k = -\alpha J'(u_k)$, $\alpha = (\max_{1 \leq i \leq n} |J'_{ui}(u_k)|)^{-1} > 0$ — 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 4.9.1. Для выпуклой регулярной задачи (1) условие $\sigma_k = 0$ гарантирует, что $u_k \in U_*$. Покажем это.

Теорема 1. Пусть функции $J(u)$, $g_i(u)$, $i = \overline{1, m}$, 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. $J_* > -\infty$, $U_* \neq \emptyset$. Тогда для любой точки $u_* \in U_*$ задача

$$\sigma \rightarrow \inf; \quad z = (e, \sigma) \in W_* = \{(e, \sigma): \langle J'(u_*), e \rangle \leq \sigma, \quad \langle g'_i(u_*), e \rangle \leq \sigma, \\ i \in I_*; \quad |e^j| \leq 1, \quad j = \overline{1, n}\}, \quad (7)$$

где

$$I_* = \{i: 1 \leq i \leq m, \quad g_i(u_*) = 0\},$$

необходимо имеет решение (e_*, σ_*) с $\sigma_* = \min_{W_*} \sigma = 0$. Если, кроме того, $J(u)$, $g_i(u)$ выпуклы на E^n , а множество U регулярно (см. определение 4.8.2), то всякая точка $u_* \in U$, для которой задача (7) определяет величину $\sigma_* = \inf_{W_*} \sigma = 0$, является решением задачи (1).

Доказательство. Необходимость. Пусть $u_* \in U_*$. По теореме 4.9.1 тогда существуют множители Лагранжа $\lambda_0^*, \dots, \lambda_m^*$, неотрицательные и не все равные нулю, такие, что

$$\lambda_0^* J'(u_*) + \sum_{i=1}^m \lambda_i^* g'_i(u_*) = 0, \quad \lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, m}. \quad (8)$$

Если $i \notin I_*$, то из второго равенства (8) следует $\lambda_i^* = 0$, поэтому первое равенство (8) можно переписать в виде

$$\lambda_0^* J'(u_*) + \sum_{i \in I_*} \lambda_i^* g'_i(u_*) = 0. \quad (9)$$

Возьмем любую точку $(e, \sigma) \in W_*$. Тогда $\langle J'(u_*), e \rangle \leq \sigma$, $\langle g'_i(u_*), e \rangle \leq \sigma$, $i \in I_*$. Умножим первое из этих неравенств на $\lambda_0^* \geq 0$, остальные — на соответствующие $\lambda_i^* \geq 0$ и сложим. С учетом равенства (9) получим

$$\langle \lambda_0^* J'(u_*) + \sum_{i \in I_*} \lambda_i^* g'_i(u_*), e \rangle = 0 \leq \sigma (\lambda_0^* + \lambda_1^* + \dots + \lambda_m^*).$$

Так как $\lambda_0^* + \dots + \lambda_m^* > 0$, то отсюда имеем $\sigma \geq 0$ для любой точки $(e, \sigma) \in W_*$. Поскольку $(0, 0) \in W_*$, то $\sigma_* = \inf_{W_*} \sigma = 0$.

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 $\sigma_* = \inf_{W_*} \sigma = 0$. Покажем, что тогда $u_* \in U_*$. С этой целью введем конус

$$K = \{z = (e, \sigma) \in E^{n+1}: \langle J'(u_*), e \rangle + (-1)\sigma \leq 0,$$

$$\langle g'_i(u_*), e \rangle + (-1)\sigma \leq 0, \quad i \in I_*\},$$

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 $\inf_{W_*} \sigma = \sigma_* = 0$ имеем

$$\langle d, z \rangle = \langle 0, e \rangle + 1 \cdot \sigma = \sigma \geq \sigma_* = 0 \quad (10)$$

для всех $z = (e, \sigma) \in K$, для которых $|e'| \leq 1$, $j = \overline{1, n}$. Однако условие $|e'| \leq 1$, $j = \overline{1, n}$, 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 , $1 \leq j \leq n$. Тогда $\|e\| = \max_{1 \leq j \leq n} |e^j| > 1$. Положим

$$\bar{z} = (\bar{e}, \bar{\sigma}), \quad \bar{e} = e/\|e\|, \quad \bar{\sigma} = \sigma/\|e\|.$$

Ясно, что $\bar{z} \in W_*$. Следовательно, $\langle d, \bar{z} \rangle = \langle d, z \rangle / \|e\| = \bar{\sigma} = \sigma / \|e\| \geq 0$, так что $\langle d, z \rangle = \sigma \geq 0$. Тем самым показано, что неравенство (10) верно для всех $z \in K$, т. е. $d \in K^*$.

По теореме 4.8.3 тогда существуют неотрицательные числа $\lambda_0^*, \dots, \lambda_m^*$ такие, что

$$d = (0, 1) = -\lambda_0^* c_0 - \sum_{i \in I_*} \lambda_i^* c_i.$$

Вспомяная определения векторов $c_0, c_i, i \in I_*$, отсюда имеем

$$0 = -\lambda_0^* J'(u_*) - \sum_{i \in I_*} \lambda_i^* g'_i(u_*) = 0, \quad 1 = \lambda_0^* + \sum_{i \in I_*} \lambda_i^*. \quad (11)$$

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

$$\lambda_0^* J'(u_*) + \sum_{i=1}^m \lambda_i^* g'_i(u_*) = 0, \quad \lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, m}, \quad (12)$$

а из второго равенства (11) следует, что не все числа $\lambda_0^*, \lambda_i^*, i \in I_*$, равны нулю.

Покажем, что $\lambda_0^* > 0$. Если $I_* = \emptyset$, то из (11) сразу имеем $\lambda_0^* = 1$. Допустим, что $I_* \neq \emptyset$, но тем не менее $\lambda_0^* = 0$. Тогда среди неотрицательных чисел $\lambda_i^*, i \in I_*$, 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 $i = \overline{1, m}$. Поскольку $I_* \neq \emptyset$, 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 $\alpha, 0 \leq \alpha \leq 1$. 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 $0 > g_i(\bar{u}) = g_i(\bar{u}) - g_i(u_*) \geq \langle g'_i(u_*), \bar{u} - u_* \rangle = \langle g'_i(u_*), e \rangle$.

Поэтому $\sum_{i=1}^m \lambda_i^* \langle g'_i(u_*), e \rangle \leq \lambda_p^* \langle g'_p(u_*), e \rangle < 0$. Но с другой сторо-

ны, 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, приходим 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^* , $i = \overline{1, m}$. Поэтому предыдущее равенство в силу теоремы 4.2.3 равносильно условию $L(u_*, \lambda^*) \geq L(u, \lambda^*)$ при всех $u \in E^n$. Отсюда и из второго равенства (12) с помощью леммы 4.8.1 и теоремы 4.8.1 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 $J(u) = x + \cos y$, $u \in U = \{u = (x, y) \in E^2: g(u) = -x \leq 0\}$ (ср. с примером 4.9.1). Возьмем точку $u_* = (0, 0)$. Тогда $J'(u_*) = (1, 0)$, $g'(u_*) = (-1, 0)$, $W_* = \{(e, \sigma) = (e^1, e^2, \sigma): e^1 \leq \sigma, -e^1 \leq \sigma, |e^1| \leq 1, |e^2| \leq 1\}$. Отсюда $|e^1| \leq \sigma$ при всех $(e, \sigma) \in W_*$. Это значит, что $\inf_{W_*} \sigma = \sigma_* = 0$, 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 . Любопытно 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, e_k) , $u_k \in U$, $e_k > 0$, при каком-то $k \geq 0$ уже известно. Определим множество номеров

$$I_k = \{i: 1 \leq i \leq m, \quad -e_k \leq g_i(u_k) \leq 0\} \quad (13)$$

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 $\sigma_k = \inf_{W_k} \sigma \leq 0$.

Имеются две возможности:

1) $\sigma_k \leq -\varepsilon_k$. В этом случае считаем, что e_k является достаточно хорошим возможным направлением убывания в точке u_k , и полагаем

$$u_{k+1} = u_k + \alpha_k e_k, \quad 0 < \alpha_k \leq \beta_k, \quad e_{k+1} = e_k, \quad (14)$$

где β_k определяется из (4), а выбор α_k может быть осуществлен одним из описанных выше способов.

2) $-\varepsilon_k < \sigma_k \leq 0$. В этом случае считаем, что направление e_k не обладает ясно выраженным свойством возможного направления в точке u_k , полагаем

$$u_{k+1} = u_k, \quad e_{k+1} = e_k/2 \quad (15)$$

и снова переходим к рассмотрению задачи (2) с заменой множества I_k на множество $I_{k+1} = \{i: 1 \leq i \leq m, -\varepsilon_{k+1} = -\varepsilon_k/2 \leq g_i(u_k) \leq 0\}$, 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 $\varepsilon_0, \varepsilon_1, \dots$, 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 e_k , например, $e_{k+1} = 0,9 e_k$.

3. Следуя [9, 44], изучим сходимость метода (2), (4), (6), (13)—(15). Предварительно докажем несколько лемм.

Лемма 1. Пусть $J(u)$, $g_i(u) \in C^{1,1}(U)$, $i = \overline{1, m}$, и I — некоторое фиксированное множество номеров, взятых из $\{1, 2, \dots, m\}$ (возможности $I = \emptyset$ или $I = \overline{1, m}$ не исключаются). Для каждого $u \in U$ положим $\sigma(u) = \min_{G(u)} \sigma$, где $G(u) = \{(e, \sigma) = (e^1, \dots, e^n, \sigma) \in E^{n+1}: \langle J'(u), e \rangle \leq \sigma, \langle g'_i(u), e \rangle \leq \sigma, i \in I; |e^j| \leq 1, j = \overline{1, n}\}$. Тогда

$$|\sigma(u) - \sigma(v)| \leq L \sqrt{n} \|u - v\|, \quad u, v \in U, \quad (16)$$

где L — константа Липшица для градиентов $J'(u)$, $g'_i(u)$, $i = \overline{1, m}$.

Доказательство. Возьмем произвольные точки $u, v \in U$. Пусть $(e, \sigma) \in G(v)$, т. е.

$$\langle J'(v), e \rangle \leq \sigma, \quad \langle g'_i(v), e \rangle \leq \sigma, \quad i \in I; |e^j| \leq 1, \quad j = \overline{1, n}.$$

Тогда

$$\langle J'(u), e \rangle = \langle J'(v), e \rangle + \langle J'(u) - J'(v), e \rangle \leq$$

$$\leq \sigma + L \|u - v\| \|e\| \leq \sigma + L \|u - v\| \sqrt{n}$$

и аналогично

$$\langle g'_i(u), e \rangle \leq \sigma + L |u - v| \sqrt{n}, \quad i \in I.$$

Это значит, что при каждом $(e, \sigma) \in G(v)$ точка $(e, \sigma + L \sqrt{n} |u - v|)$ принадлежит множеству $G(u)$. Тогда $\sigma(u) = \min_{G(u)} \sigma \leq \sigma + L \sqrt{n} |u - v|$ для любых $(e, \sigma) \in G(v)$. Следовательно, $\sigma(u) \leq \sigma(v) + L \sqrt{n} |u - v|$. Поменяв в этих рассуждениях точки u, v ролями, получим $\sigma(v) \leq \sigma(u) + L \sqrt{n} |u - v|$. Из последних двух неравенств следует неравенство (16).

Лемма 2. Пусть

$$J(u), g_1(u), \dots, g_m(u) \in C^{1,1}(U), A_0 = \max_{1 \leq i \leq m} \sup_{u \in U} |g'_i(u)| < \infty,$$

а последовательности $\{u_k\}$, $\{\alpha_k\}$, $\{\beta_k\}$, $\{e_k\}$, $\{\sigma_k\}$ определены условиями (2), (4), (6), (13) — (15). Тогда

$$\beta_k \geq A_1 \min \{e_k, |\sigma_k|\}, \quad k = 0, 1, \dots, \quad (17)$$

где $A_1 = \min \{1/(A_0 \sqrt{n}); 1/(nL)\} > 0$, L — константа Липшица для градиентов $J'(u)$, $g'_1(u)$, $\dots$, $g'_m(u)$.

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 $g_i(u_k) < -e_k$. Тогда $e_k < -g_i(u_k) = g_i(u_k + \beta_k e_k) - g_i(u_k) = \langle g'_i(u_k + \beta_k e_k), \beta_k e_k \rangle \leq A_0 \beta_k |e_k| \leq A_0 \sqrt{n} \beta_k$, т. е. $\beta_k \geq e_k / (A_0 \sqrt{n})$.

Если же оказалось, что $-e_k \leq g_i(u_k) \leq 0$, то $i \in I_k$ и $\langle g'_i(u_k), e_k \rangle \leq \sigma_k \leq 0$. 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 $g_i(u_k + \alpha e_k) \leq 0$ при всех α , $0 < \alpha < \beta_k$. Кроме того, $g_i(u_k + \beta_k e_k) = 0$ по выбору номера i . Тогда $0 \geq g_i(u_k + \alpha e_k) - g_i(u_k + \beta_k e_k) = \langle g'_i(u_k + \beta_k e_k), e_k \rangle (\alpha - \beta_k) + o(|\alpha - \beta_k|)$ или $\langle g'_i(u_k + \beta_k e_k), e_k \rangle \geq o(|\alpha - \beta_k|) \cdot (\beta_k - \alpha)^{-1}$ при всех α , $0 < \alpha < \beta_k$. Отсюда при $\alpha \rightarrow \beta_k - 0$ получим $\langle g'_i(u_k + \beta_k e_k), e_k \rangle \geq 0$. Тогда $|\sigma_k| = -\sigma_k \leq \langle -g'_i(u_k), e_k \rangle \leq \langle g'_i(u_k + \beta_k e_k) - g'_i(u_k), e_k \rangle \leq L \beta_k |e_k|^2 \leq Ln \beta_k$, т. е. $\beta_k \geq |\sigma_k| / (nL)$. Если $\sigma_k = 0$, то последнее неравенство также остается верным, так как согласно (4) всегда $\beta_k \geq 0$. Объединяя обе полученные оценки для β_k , приходим к оценке (17). Лемма 2 доказана.

Лемма 3. Пусть $J(u)$, $g_i(u) \in C^{1,1}(U)$, $i = \overline{1, m}$. Пусть, кроме того, в процессе (2), (4), (6), (13) — (15) на некоторой k -й итерации оказалось $\sigma_k \leq -e_k$. Тогда

$$J(u_k) - J(u_{k+1}) \geq A_2 \min \{\beta_k |\sigma_k|; \sigma_k^2\} - \delta_k, \quad (18)$$

где $A_2 = \min \{1/2; 1/(2nL)\} > 0$.

Доказательство. Из неравенства $\sigma_k \leq -e_k$ и определения e_k , σ_k следует, что $\langle J'(u_k), e_k \rangle \leq \sigma_k \leq -e_k < 0$. Кроме того, e_k является возможным направлением в точке u_k и, следовательно, $\beta_k > 0$. Из (6) и леммы 2.3.1

имеем

$$J(u_k) - J(u_{k+1}) \geq J(u_k) - \inf_{0 \leq \alpha \leq \beta_k} f_k(\alpha) - \delta_k \geq J(u_k) - J(u_k + \alpha e_k) - \delta_k \geq \\ \geq -\alpha \langle J'(u_k), e_k \rangle - \alpha^2 L |e_k|^2 / 2 - \delta_k \geq -\alpha \sigma_k - \alpha^2 nL / 2 - \delta_k \quad (19)$$

при всех α , $0 \leq \alpha \leq \beta_k$. Положим здесь $\alpha = \min \{\beta_k; |\sigma_k|/(nL)\}$.

Может случиться, что $\alpha = \beta_k \leq |\sigma_k|/(nL)$. Тогда из (19) получаем

$$J(u_k) - J(u_{k+1}) \geq \alpha |\sigma_k| - \alpha \cdot \alpha \cdot nL / 2 - \delta_k \geq \beta_k |\sigma_k| - \\ - \beta_k (|\sigma_k|/(nL)) (nL / 2) - \delta_k = \beta_k |\sigma_k| / 2 - \delta_k.$$

Если же $\alpha = |\sigma_k|/(nL) < \beta_k$, то из (19) следует

$$J(u_k) - J(u_{k+1}) \geq \sigma_k^2 / (nL) - (\sigma_k^2 / (nL)^2) (nL / 2) - \delta_k = \sigma_k^2 / (2nL) - \delta_k.$$

Объединяя оба рассмотренных случая, приходим к оценке (18).

Теорема 2. Пусть функции $J(u)$, $g_i(u)$, $i = 1, m$, определены и выпуклы на E^n ; множество U из (1) регулярно; $J(u)$, $g_i(u) \in C^{1,1}(U)$, $A_0 = \max_{1 \leq i \leq m} \sup_U |g'_i(u)| < \infty$. Пусть задача (1) имеет решение, т. е. $J_* > -\infty$, $U_* \neq \emptyset$, и начальная точка $u_0 \in U$ такова, что множество $M_\delta(u_0) = \{u: u \in U, J(u) \leq J(u_0) + \delta\}$ ограничено. Тогда при любом выборе $\varepsilon_0 > 0$ для последовательности $\{u_k\}$, определяемой условиями (2), (4), (6), (13)–(15), справедливы равенства

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0. \quad (20)$$

Доказательство. Сначала установим, что

$$\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0. \quad (21)$$

Если $\sigma_k \leq -\varepsilon_k$, то из (14), (6) имеем $J(u_{k+1}) = f_k(\alpha_k) \leq f_k(0) + \delta_k = J(u_k) + \delta_k$. Если же $-\varepsilon_k < \sigma_k \leq 0$, то из (15) следует $J(u_{k+1}) = J(u_k) \leq J(u_k) + \delta_k$. Таким образом, $J(u_k) \geq J_* > -\infty$ и, кроме того,

$$J(u_{k+1}) \leq J(u_k) + \delta_k, \quad k = 0, 1, \dots; \quad \sum_{k=0}^{\infty} \delta_k = \delta < \infty. \quad (22)$$

Согласно лемме 2.3.2 тогда существует $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Отсюда следует равенство (21).

Далее, покажем, что $\lim_{k \rightarrow \infty} \varepsilon_k = 0$. Согласно (14), (15) последовательность $\{\varepsilon_k\}$ получается дроблением и не возрастает. Допустим, что $\lim_{k \rightarrow \infty} \varepsilon_k = \varepsilon > 0$. Это значит, что в процессе построения $\{u_k\}$ было конечное число дроблений и $\varepsilon_k = \varepsilon > 0$ при всех $k \geq k_0$. Из (14) тогда имеем $\sigma_k \leq -\varepsilon_k = -\varepsilon$, т. е. $|\sigma_k| \geq \varepsilon$, $k \geq k_0$. В этом случае согласно лемме 2 $\beta_k \geq A_1 \varepsilon$, $k \geq k_0$. Поэтому из леммы 3 получим $J(u_k) - J(u_{k+1}) \geq A_2 \min \{A_1 \varepsilon^2; \varepsilon^2\} - \delta_k$, $k \geq k_0$, что противоречит равенству (21). Итак, показано, что $\lim_{k \rightarrow \infty} \varepsilon_k = 0$.

Пусть $k_1 < k_2 < \dots < k_m < \dots$ — номера тех итераций, когда происходит дробление e_k . Согласно (14), (15) тогда $-e_{k_m} \leq \sigma_{k_m} \leq 0$, $m = 1, 2, \dots$. Следовательно, $\lim_{m \rightarrow \infty} \sigma_{k_m} = 0$. Тем самым установлено, что существует хотя бы одна подпоследовательность $\{\sigma_{k_m}\}$, сходящаяся к нулю.

Возьмем произвольную подпоследовательность $\{\sigma_{k_m}\} \rightarrow 0$. Покажем, что тогда любая предельная точка соответствующей подпоследовательности $\{u_{k_m}\}$ принадлежит множеству U_* . Из (22) следует, что $J(u_{k+1}) \leq J(u_0) + \delta$, $k = 0, 1, \dots$, т. е. $\{u_k\} \in M_\delta(u_0)$. По условию множество $M_\delta(u_0)$ ограничено. Поэтому можем считать, что взятая выше подпоследовательность $\{u_{k_m}\}$ сходится к некоторой точке u_* . Далее множество номеров l_k , определяемое согласно (13), представляет собой подмножество конечного числа номеров $\{1, 2, \dots, m\}$, поэтому число различных множеств l_k конечно. Это значит, что среди $\{l_{k_m}, m = 1, 2, \dots\}$ найдется хотя бы одно множество $l_{k_r} = l$, которое повторяется бесконечно много раз. Выбирая при необходимости подпоследовательности, можем, таким образом, считать, что

$$\{\sigma_{k_m}\} \rightarrow 0, \quad \{u_{k_m}\} \rightarrow u_*, \quad l_{k_m} = l, \quad m = 1, 2, \dots$$

Согласно лемме 1 при $G(u_{k_m}) = W_{k_m}$ имеем

$$|\sigma_{k_m} - \sigma(u_*)| = |\sigma(u_{k_m}) - \sigma(u_*)| \leq L \sqrt{n} |u_{k_m} - u_*| \rightarrow 0,$$

т. е. $\lim_{m \rightarrow \infty} \sigma(u_{k_m}) = \lim_{m \rightarrow \infty} \sigma_{k_m} = 0 = \sigma(u_*)$, где $\sigma(u_*) = \inf_{G(u_*)} \sigma$, $G(u_*) = \{(e, \sigma) \in E^{n+1} : \langle J'(u_*), e \rangle \leq \sigma, \langle g'_i(u_*), e \rangle \leq \sigma, i \in I; |e^j| \leq 1, j = \overline{1, n}\}$.

Рассмотрим задачу (7), соответствующую точке $u_* = \lim_{m \rightarrow \infty} u_{k_m}$. Покажем, что $W_* \subseteq G(u_*)$. По определению множества $I = I_{k_m} = \{i: 1 \leq i \leq m, -e_{k_m} \leq g_i(u_{k_m}) \leq 0\}$, $m = 1, 2, \dots$. Отсюда при $m \rightarrow \infty$ 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 тогда, замечая, что одна и та же функция на более широком множестве имеет меньшую нижнюю грань, получаем $\sigma_* = \inf_{W_*} \sigma \geq \inf_{G(u_*)} \sigma = \sigma(u_*) = 0$.

С другой стороны, $(0, 0) \in W_*$, поэтому $\sigma_* \leq 0$. 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 $\lim_{k \rightarrow \infty} J(u_k)$. Теперь можем сказать, чему равен этот предел: $\lim_{k \rightarrow \infty} J(u_k) = \lim_{m \rightarrow \infty} J(u_{k_m}) = J(u_*) = J_*$.

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 $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. Равенства

(20) и, тем самым, теорема 2 доказаны.

4. Для задачи

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in E^n: g_i(u) \leq 0, \quad i = \overline{1, m},$$

$$g_i(u) = \langle a_i, u \rangle - b^i = 0, \quad i = \overline{m+1, s}\}.$$

содержащей линейные ограничения типа равенств, метод возможных направлений описывается так же, как выше, лишь в задаче (2) нужно добавить еще ограничения $\langle a_i, e \rangle = 0$, $i = m+1, s$.

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 $|e| \leq 1$, $j = 1, n$. Возможны и другие условия нормировки [121], например, $|e|^2 = \langle e, e \rangle \leq 1$. Заметим, что при нормировке $|e|^2 \leq 1$ задача (2) уже не будет задачей линейного программирования. О других способах нормировки, о сходимости различных вариантов метода возможных направлений и других аспектах этого метода см., например, в [9, 44, 45, 116, 121, 127].

У п р а ж н е н и я. 1. Сделать несколько итераций метода возможных направлений для задачи минимизации $J(u) = x + y$ на множестве $U = \{u = (x, y): g_1(u) = x^2 - y \leq 0, g_2(u) = y - 1 \leq 0\}$ 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 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 удастся 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:

$$J(u) = \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle \rightarrow \inf; \quad u \in U = E^n, \quad (1)$$

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

$$J'(u_*) = Au_* - b = 0 \quad \text{или} \quad u_* = A^{-1}b. \quad (2)$$

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. Поэтому пусть $J'(u_0) \neq 0$. Тогда положим

$$u_1 = u_0 - \alpha_0 p_0, \quad \alpha_0 \geq 0,$$

где величина α_0 определяется условием

$$f_0(\alpha_0) = \min_{\alpha \geq 0} f_0(\alpha), \quad f_0(\alpha) = J(u_0 - \alpha p_0).$$

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

$$\Gamma_1 = \{u \in E^n: \langle Ap_0, u - u_1 \rangle = 0\}$$

представляет собой гиперплоскость размерности $n-1$, проходящую через точку u_0 . 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 $J'(u_1) \neq 0$, то $p_1 \neq 0$. В самом деле, если бы $p_1 = 0$, то $J'(u_1) = \beta_0 p_0$ 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

$$u_2 = u_1 - \alpha_1 p_1, \quad \alpha_1 \geq 0,$$

где величину α_1 будем определять из условия

$$f_1(\alpha_1) = \min_{\alpha \geq 0} f_1(\alpha), \quad f_1(\alpha) = J(u_1 - \alpha p_1).$$

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

$$J'(u_1) - J'(u_2) = Au_1 - b - Au_2 + b = \alpha_1 Ap_1.$$

Тогда в силу (3) и выбора p_1 получаем

$$\langle J'(u_2), J'(u_0) \rangle = \langle J'(u_2), p_0 \rangle = \langle J'(u_1) - \alpha_1 Ap_1, p_0 \rangle = 0.$$

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 $k \geq 2$ уже найдены точки $u_0, u_1, \dots, u_k, u_{i+1} = u_i - \alpha_i p_i, i = 0, k-1$, 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

$$f_i(\alpha_i) = \min_{\alpha \geq 0} f_i(\alpha), \quad f_i(\alpha) = J(u_i - \alpha p_i), \quad i = 0, k-1;$$

пусть

$$\langle Ap_i, p_j \rangle = 0, \quad i \neq j, \quad 0 \leq i, j \leq k-1, \quad (4)$$

$$\langle J'(u_i), p_j \rangle = 0, \quad 0 \leq j < i \leq k, \quad (5)$$

$$\langle J'(u_i), J'(u_j) \rangle = 0, \quad i \neq j, \quad 0 \leq i, j \leq k; \quad (6)$$

и, кроме того, пусть $J'(u_i) \neq 0, i = 0, k$, и система векторов $\{Ap_0, Ap_1, \dots, Ap_{k-1}\}$ линейно независима.

Тогда множество

$$\Gamma_k = \{u \in E^n: \langle Ap_i, u - u_{i+1} \rangle = 0, \quad i = \overline{0, k-1}\}$$

представляет собой гиперплоскость (аффинное множество — см. пример 4.1.4) размерности $m - k$. 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 $i = \overline{0, k-1}$, то $u_k \in \Gamma_k$. Замечательно то, что $u_* \in \Gamma_k$, так как согласно (2), (5)

$$\begin{aligned} \langle Ap_i, u_* - u_{i+1} \rangle &= \langle Ap_i, A^{-1}b - u_{i+1} \rangle = \langle p_i, b - Au_{i+1} \rangle = \\ &= \langle p_i, -J'(u_{i+1}) \rangle = 0, \quad i = \overline{0, k-1}. \end{aligned}$$

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 $\langle Ap_i, p_k \rangle = 0$, $i = \overline{0, k-1}$. Будем искать p_k в виде

$$p_k = J'(u_k) - \beta_k p_{k-1}. \quad (7)$$

Заметим, что

$$J'(u_i) - J'(u_{i+1}) = Au_i - Au_{i+1} = \alpha_i Ap_i, \quad i = \overline{0, k-1}. \quad (8)$$

Из (4), (6), (8) следует

$$\begin{aligned} \langle Ap_i, p_k \rangle &= \langle Ap_i, J'(u_k) - \beta_k p_{k-1} \rangle = \\ &= \langle Ap_i, J'(u_k) \rangle - \beta_k \langle Ap_i, p_{k-1} \rangle = \langle J'(u_i) - J'(u_{i+1}), J'(u_k) \rangle \alpha_i^{-1} = 0 \end{aligned}$$

для всех $i = \overline{0, k-2}$ 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

$$\beta_k = \langle Ap_{k-1}, J'(u_k) \rangle / \langle Ap_{k-1}, p_{k-1} \rangle. \quad (9)$$

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

$$\langle Ap_i, p_j \rangle = 0, \quad i \neq j, \quad 0 \leq i, j \leq k. \quad (10)$$

Следующее $(k+1)$ -е приближение будем искать в виде

$$u_{k+1} = u_k - \alpha_k p_k, \quad \alpha_k \geq 0, \quad (11)$$

где α_k определяется из условия

$$f_k(\alpha_k) = \min_{\alpha \geq 0} f_k(\alpha), \quad f_k(\alpha) = J(u_k - \alpha p_k). \quad (12)$$

Поскольку $f_k(\alpha)$ — сильно выпуклая функция, то величина α_k существует и единственна. С учетом предположений индукций 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

$$\langle J'(u_{k+1}), p_k \rangle = 0. \quad (13)$$

Отсюда нетрудно получить явное выражение для α_k . В самом деле,

$$\begin{aligned} 0 &= \langle J'(u_{k+1}), p_k \rangle = \langle Au_{k+1} - b, p_k \rangle = \\ &= \langle Au_k - \alpha_k Ap_k - b, p_k \rangle = \langle J'(u_k), p_k \rangle - \alpha_k \langle Ap_k, p_k \rangle. \end{aligned}$$

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

$$\langle J'(u_{k+1}), p_i \rangle = \langle J'(u_k) - \alpha_k Ap_k, p_i \rangle = 0, \quad i = \overline{0, k-1}. \quad (15)$$

Собрав все равенства (5), (13), (15), получим

$$\langle J'(u_i), p_j \rangle = 0, \quad 0 \leq j < i \leq k+1. \quad (16)$$

Из предположения индукции и равенств (16) следует

$$\begin{aligned} \langle J'(u_{k+1}), J'(u_i) \rangle &= \langle J'(u_{k+1}), p_i + \beta_i p_{i-1} \rangle = 0, \quad i = \overline{1, k}, \\ \langle J'(u_{k+1}), J'(u_0) \rangle &= \langle J'(u_{k+1}), p_0 \rangle = 0. \end{aligned}$$

Отсюда и из (6) имеем

$$\langle J'(u_i), J'(u_j) \rangle = 0, \quad i \neq j, \quad 0 \leq i, j \leq k+1.$$

Наконец, покажем, что система $\{Ap_0, \dots, Ap_k\}$ линейно независима. В самом деле, если $\gamma_0 Ap_0 + \gamma_1 Ap_1 + \dots + \gamma_k Ap_k = 0$, то умножая это равенство на p_j скалярно, с учетом (10) получим $\gamma_j \langle Ap_j, p_j \rangle = 0$, $j = \overline{0, k}$. Так как $p_j \neq 0$, то $\langle Ap_j, p_j \rangle > 0$ и последние равенства возможны лишь при $\gamma_j = 0$, $j = \overline{0, k}$.

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 $p_0, p_1, \dots, p_k$ называются *сопряженными относительно матрицы A или A -ортогональными*, если $\langle A p_i, p_j \rangle = 0$ при всех $i \neq j, 0 \leq i, j \leq k$.

Нетрудно видеть, для квадратичной задачи (1) метод сопряженных направлений закончится за конечное число итераций нахождением точки u_* . В самом деле, векторы $J'(u_0), \dots, J'(u_k), \dots$, получаемые этим методом, образуют ортогональную систему: $\langle J'(u_i), J'(u_j) \rangle = 0, i \neq j$. Однако в n -мерном пространстве не может быть более n ненулевых взаимно ортогональных векторов. Следовательно, найдется номер $k, 0 \leq k < n$, такой, что $J'(u_k) = 0$. Тогда $u_k = u_*$ — решение задачи (1).

2. Перейдем к рассмотрению задачи

$$J(u) \rightarrow \inf; \quad u \in U \equiv E^n, \quad (17)$$

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:

$$\begin{aligned} \langle A p_{k-1}, J'(u_k) \rangle &= \langle J'(u_{k-1}) - J'(u_k), J'(u_k) \rangle \alpha_k^{-1} = -|J'(u_k)|^2 \alpha_k^{-1}, \\ \langle A p_{k-1}, p_{k-1} \rangle &= \langle J'(u_{k-1}) - J'(u_k), p_{k-1} \rangle \alpha_k^{-1} = \\ &= \langle J'(u_{k-1}), p_{k-1} \rangle \alpha_k^{-1} = \langle J'(u_{k-1}), J'(u_{k-1}) - \beta_{k-1} p_{k-2} \rangle \alpha_k^{-1} = \\ &= |J'(u_{k-1})|^2 \alpha_k^{-1}. \end{aligned}$$

Тогда формула (9) запишется в виде

$$\beta_k = \frac{\langle J'(u_k), J'(u_{k-1}) - J'(u_k) \rangle}{|J'(u_{k-1})|^2}, \quad (18)$$

или

$$\beta_k = - \frac{|J'(u_k)|^2}{|J'(u_{k-1})|^2}. \quad (19)$$

Кроме того, вспоминая, что для функции (1) $A = J''(u_k)$, формулу (9) можно представить еще и в такой форме:

$$\beta_k = \frac{\langle J''(u_k) p_{k-1}, J'(u_k) \rangle}{\langle J''(u_k) p_{k-1}, p_{k-1} \rangle}. \quad (20)$$

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), приходим к следующему описанию метода сопряженных направлений для задачи (17). Пусть u_0 — некоторое начальное приближение. Будем строить последовательность $\{u_k\}$ по правилам

$$u_{k+1} = u_k - \alpha_k p_k, \quad k = 0, 1, \dots, \quad (21)$$

где

$$p_0 = J'(u_0), \quad p_k = J'(u_k) - \beta_k p_{k-1}, \quad k = 1, 2, \dots, \quad (22)$$

величина α_k определяется условиями

$$\alpha_k \geq 0, \quad f_k(\alpha_k) = \min_{\alpha \geq 0} f_k(\alpha), \quad f_k(\alpha) = J(u_k - \alpha p_k), \quad (23)$$

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 $k \geq 1$, при которых принимается $\beta_k = 0$, через I_0 . Номера $k \in I_0$ называются *моментами обновления метода*. Если метод используется без обновления, то $I_0 = \emptyset$. На практике часто берут $I_0 = \{n, 2n, 3n, \dots\}$, где n — размерность рассматриваемого пространства. Возможны и другие правила выбора моментов обновления. Кстати, если $I_0 = \{1, 2, 3, \dots\}$, 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

$$\langle J'(u_{k+1}), p_k \rangle = 0, \quad \langle J'(u_k), p_k \rangle = |J'(u_k)|^2, \quad k = 0, 1, \dots \quad (24)$$

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 $0 \leq f'_0(0) = -\langle J'(u_0), p_0 \rangle = -|J'(u_0)|^2 \leq 0$, 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 $k \geq 1$ 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 $0 \leq f'_k(0) = -\langle J'(u_k), p_k \rangle = -\langle J'(u_k), J'(u_k) - \beta_k p_{k-1} \rangle = -|J'(u_k)|^2 = -|J'(u_{k+1})|^2 \leq 0$, 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

$$|p_k|^2 = |J'(u_k) - \beta_k p_{k-1}|^2 = |J'(u_k)|^2 + \beta_k^2 |p_{k-1}|^2, \quad k = 1, 2, \dots \quad (25)$$

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

$$0 \leq \alpha_k = J(u_k) - J_* \leq q^k a_k, \quad |u_k - u_*|^2 \leq \frac{2}{\mu} q^k a_0, \quad k = 0, 1, \dots, \quad (26)$$

где $q = 1 - \frac{\mu^3}{L(\mu^2 + L^2)}$, $0 < q < 1$, 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 $J(u_*) = J_* = \inf_{E^n} J(u)$. 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 $k = 0, 1, \dots$, 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 $\alpha \geq 0$ имеем $J(u_{k+1}) \leq J(u_k - \alpha p_k)$. Отсюда и из леммы 2.3.1 с учетом второго равенства (24) получим

$$\begin{aligned} J(u_k) - J(u_{k+1}) &\geq J(u_k) - J(u_k - \alpha p_k) \geq \alpha \langle J'(u_k), p_k \rangle - \frac{\alpha^2 L}{2} |p_k|^2 = \\ &= \alpha |J'(u_k)|^2 - \frac{\alpha^2 L}{2} |p_k|^2, \quad \alpha \geq 0, \quad k = 0, 1, \dots \end{aligned} \quad (27)$$

Докажем неравенство

$$\gamma L |p_k|^2 \leq |J'(u_k)|^2, \quad \gamma = \mu^2 L^{-1} (\mu^2 + L^2)^{-1}, \quad k = 0, 1, \dots \quad (28)$$

Согласно теореме 4.3.3 $\mu |u_k - u_{k-1}|^2 = \mu \alpha_{k-1}^2 |p_{k-1}|^2 \leq \langle J'(u_k) - J'(u_{k-1}), u_k - u_{k-1} \rangle = \langle J'(u_k) - J'(u_{k-1}), p_{k-1} \rangle (-\alpha_{k-1})$. Отсюда с учетом равенств (24)

имеем

$$\mu \alpha_{k-1} |p_{k-1}|^2 \leq \langle J'(u_{k-1}), p_{k-1} \rangle = |J'(u_{k-1})|^2.$$

Тогда из (18) следует

$$\begin{aligned} |\beta_k| &\leq |J'(u_k)| L |u_k - u_{k-1}| \cdot |J'(u_{k-1})|^{-2} \leq \\ &\leq |J'(u_k)| L \alpha_{k-1} |p_{k-1}| (\mu \alpha_{k-1} |p_{k-1}|^2)^{-1} = L \mu^{-1} |J'(u_k)| \cdot |p_{k-1}|^{-1} \end{aligned}$$

т. е.

$$|\beta_k| |p_{k-1}| \leq L \mu^{-1} |J'(u_k)|, \quad k = 0, 1, \dots$$

Отсюда и из (25) получим $|p_k|^2 \leq |J'(u_k)|^2 (1 + L^2 \mu^{-2})$, что равносильно неравенству (28).

Теперь нетрудно доказать оценки (26). Из (27) с учетом (28) имеем

$$a_k - a_{k+1} \geq \alpha \left(1 - \frac{\alpha}{2\gamma}\right) |J'(u_k)|^2, \quad \alpha \geq 0, \quad k = 0, 1, \dots$$

Следовательно,

$$a_k - a_{k+1} \geq \max_{\alpha \geq 0} \alpha \left(1 - \frac{\alpha}{2\gamma}\right) |J'(u_k)|^2 = \frac{\gamma}{2} |J'(u_k)|^2, \quad k = 0, 1, \dots$$

Но $2\mu \alpha_k \leq |J'(u_k)|^2$ (см. неравенство (1.18)), поэтому $a_k - a_{k+1} \geq \gamma \mu \alpha_k$ или $a_{k+1} \leq (1 - \gamma \mu) a_k = q a_k$, $k = 0, 1, \dots$. 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 $\mu \leq L$. Теорема доказана.

Отметим, что оценки (26) являются довольно грубыми. Более тонкие исследования показывают [15], что метод сопряженных направлений на самом деле имеет более высокую скорость сходимости, чем это следует из оценок (26). В то же время этот метод ненамного 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.

В заключение заметим, что для минимизации гладкой функции $J(u)$ на замкнутом выпуклом множестве $U \neq E^n$ возможно использование метода проекции сопряженных направлений по схеме

$$u_{k+1} = P_U(u_k - \alpha_k p_k), \quad k = 0, 1, \dots,$$

где p_k определяется согласно формулам (22), (18) или (19), а величина $\alpha_k > 0$ определяется из условия монотонности $J(u_{k+1}) < J(u_k)$.

В [15] 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 общий подход к построению таких методов, более тонкие оценки скорости сходимости читатель может найти в [9, 15].

Упражнения. 1. Показать, что точка u_k , полученная методом сопряженных направлений для квадратичной функции (1) при $I_0 = \emptyset$, есть точка минимума этой функции на гиперплоскости, проходящей через точку u_0 и натянутой на векторы $J'(u_0)$, $J'(u_1)$, ..., $J'(u_{k-1})$.

2. Описать метод сопряженных направлений для функции $J(u) = |Au - b|^2$, $u \in E^n$, где A — матрица порядка $m \times n$, $b \in E^m$.

§ 7. 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 $|u_{k+1} - u_*| \leq q|u_k - u_*|$, $0 < q < 1$; при выполнении неравенства $|u_{k+1} - u_*| \leq q_k|u_k - u_*|$, где $\{q_k\} \rightarrow 0$, говорят о сверхлинейной скорости сходимости последовательности $\{u_k\}$ к u_* , а если здесь $q_k = C|u_k - u_*|^{s-1}$, т. е. $|u_{k+1} - u_*| \leq C|u_k - u_*|^s$, 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 [15].

1. Опишем метод Ньютона для задачи

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

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

$$\begin{aligned} J(u) - J(u_k) = & \langle J'(u_k), u - u_k \rangle + \frac{1}{2} \langle J''(u_k)(u - u_k), u - u_k \rangle + o(|u - u_k|^2). \end{aligned}$$

Возьмем квадратичную часть этого приращения

$$J_k(u) \equiv \langle J'(u_k), u - u_k \rangle + \frac{1}{2} \langle J''(u_k)(u - u_k), u - u_k \rangle, \quad (2)$$

и определим вспомогательное приближение $\bar{u}_k$ из условий

$$\bar{u}_k \in U, \quad J_k(\bar{u}_k) = \inf_U J_k(u). \quad (3)$$

Следующее $(k+1)$ -е приближение будем искать в виде

$$u_{k+1} = u_k + \alpha_k (\bar{u}_k - u_k), \quad 0 \leq \alpha_k \leq 1. \quad (4)$$

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

$$\alpha_k = 1, \quad k = 0, 1, \dots \quad (5)$$

В этом случае, как следует из (4), $u_{k+1} = \bar{u}_k$, $k = 0, 1, \dots$, т. е. условие (3) сразу определяет следующее $(k+1)$ -е приближение. Иначе говоря

$$u_{k+1} \in U, \quad J_k(u_{k+1}) = \inf_U J_k(u), \quad k = 0, 1, \dots \quad (6)$$

В частности, когда $U = E^n$, в точке минимума функции $J_k(u)$ ее производная $J'_k(u)$ обращается в нуль, т. е.

$$J'_k(u_{k+1}) = J'(u_k) + J''(u_k)(u_{k+1} - u_k) = 0. \quad (7)$$

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

$$u_{k+1} = u_k - (J''(u_k))^{-1} J'(u_k), \quad k = 0, 1, \dots \quad (8)$$

Широко известный метод Ньютона для решения системы уравнений

$$F(u) = \{F_1(u), \dots, F_n(u)\} = 0, \quad u \in E^n,$$

представляет собой итерационный процесс [2, 39]

$$u_{k+1} = u_k - (F'(u_k))^{-1} F(u_k), \quad k = 0, 1, \dots, \quad (9)$$

где $F'(u)$ — матрица, i -я строка которой равна $F'_i(u) = (F_{iu^1}, \dots, F_{iu^n})$. 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 проистихает название метода (2) — (4) и в общем случае.

2) В качестве α_k в (4) можно принять $\alpha_k = \lambda^i$, где i_0 — минимальный среди $i \geq 0$ номер, для которых выполняется неравенство [15]

$$J(u_k) - J(u_k + \lambda^i (\bar{u}_k - u_k)) \geq \varepsilon \lambda^i |J_k(\bar{u}_k)|, \quad (10)$$

где λ, ε — параметры метода, $0 < \lambda; \varepsilon < 1$.

3) Возможен выбор α_k в (4) из условий [15]

$$0 \leq \alpha_k \leq 1, \quad f_k(\alpha_k) = \min_{0 \leq \alpha \leq 1} f_k(\alpha), \quad f_k(\alpha) = J(u_k + \alpha(\bar{u}_k - u_k)). \quad (11)$$

Заметим, что метод (2) — (4) с выбором длины шага α_k по правилам (10), (11) аналогичен соответствующим вариантам метода условного градиента — различие лишь в том, что в методе 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) может быть использован метод сопряженных направлений. 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,

$$\|J''(u) - J''(v)\| \leq L \|u - v\|, \quad u, v \in E^n, \quad L = \text{const} > 0. \quad (12)$$

Пусть начальное приближение u_0 выбрано таким, что

$$L \|J'(u_0)\| \leq 2\mu^2 q, \quad (13)$$

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

$$|u_k - u_*| \leq 2\mu L^{-1} q^{2k}, \quad k = 0, 1, \dots \quad (14)$$

Доказательство. Существование и единственность точки u_* установлена в теореме 4.3.1. Согласно теореме 4.3.4

$$\langle J''(u)\xi, \xi \rangle \geq \mu |\xi|^2, \quad u \in E^n, \quad \xi \in E^n. \quad (15)$$

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 $k = 0, 1, \dots$ 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 $\mu |(J''(u))^{-1}z|^2 \leq \langle z, (J''(u))^{-1}z \rangle \leq |z| |(J''(u))^{-1}z|$ или $|(J''(u))^{-1}z| \leq |z| \mu^{-1}$ при всех $z \in E^n$. Это значит, что

$$\|(J''(u))^{-1}\| \leq \mu^{-1}, \quad u \in E^n. \quad (16)$$

Введем числовую последовательность $a_k = |J'(u_k)|$ и покажем, что

$$a_k \leq 2\mu^2 L^{-1} q^{2k}, \quad k = 0, 1, \dots \quad (17)$$

При $k = 0$ неравенство (17) следует из условия (13). Пусть (17) справедливо при некотором $k \geq 0$. Из условия (8) и формулы (2.3.5) имеем

$$\begin{aligned} J'(u_{k+1}) &= J'(u_k) + \int_0^1 J''(u_k + t(u_{k+1} - u_k))(u_{k+1} - u_k) dt = \\ &= \int_0^1 [J''(u_k) - J''(u_k + t(u_{k+1} - u_k))] dt (J''(u_k))^{-1} J'(u_k). \end{aligned}$$

Отсюда и из (8), (12), (16) с помощью предположения индукции получим

$$\begin{aligned} a_{k+1} &\leq (L/2) |u_{k+1} - u_k| \mu^{-1} a_k \leq (L/2\mu^2) a_k^2 \leq \\ &\leq (L/2\mu^2) (2\mu^2/L)^2 (q^{2k})^2 = (2\mu^2/L) q^{2k+1}. \end{aligned}$$

Неравенства (17) доказаны. Тогда из теоремы 4.3.3 с учетом равенства $J'(u_*) = 0$ имеем $\mu |u_k - u_*|^2 \leq \langle J'(u_k) - J'(u_*), u_k - u_* \rangle \leq |J'(u_k)| |u_k - u_*|$ или $|u_k - u_*| \leq a_k \mu^{-1}$. 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 $|u_0 - u_*| \leq a_0 \mu^{-1} \leq (2\mu/L)q$. Приведем пример, показывающий, что при отсутствии хорошего начального приближения метод (8) может расходиться.

Пример 1. Пусть

$$J(u) = \begin{cases} -\frac{1}{4\delta^3} u^4 + \frac{1}{2} \left(1 + \frac{3}{\delta}\right) u^2, & \text{если } |u| \leq \delta, \\ u^2/2 + 2|u| - (3/4)\delta, & \text{если } |u| > \delta, \end{cases}$$

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, $J''(u) \geq 1$ 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 $u_k = (-1)^k \cdot 2$, $k = 1, 2, \dots$, 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

$$|J''(u) - J''(v)| \leq L|u - v|, \quad u, v \in U, \quad L = \text{const}. \quad (18)$$

Тогда последовательность $\{u_k\}$ однозначно определяется условиями (6) при любом выборе начального приближения u_0 . Если

$$q = (L/2\mu)|u_1 - u_0| < 1, \quad (19)$$

то последовательность $\{u_k\}$, определяемая условиями (6), сходится к точке u_* — решению задачи (1), причем справедлива оценка

$$|u_k - u_*| \leq (2\mu/L) \sum_{m=k}^{\infty} q^{2^m} \leq (2\mu/L) q^{2^k} (1 - q^{2^k})^{-1}, \quad k = 0, 1, \dots; \quad (20)$$

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

$$\langle J''(u) \xi, \xi \rangle \geq \mu |\xi|^2, \quad u \in U, \quad \xi \in L_U, \quad (21)$$

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 $k = 0, 1, \dots$. 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

$$\langle J'_k(u_{k+1}), u - u_{k+1} \rangle \geq 0, \quad u \in U, \quad k = 0, 1, \dots \quad (22)$$

Так как $J'_k(u) \equiv J'(u_k) + J''(u_k)(u - u_k)$, то неравенство (22) перепишется в виде

$$\langle J'(u_k) + J''(u_k)(u_{k+1} - u_k), u - u_{k+1} \rangle \geq 0, \quad u \in U, \quad k = 0, 1, \dots \quad (23)$$

Может случиться, что $u_{k+1} = u_k$. Тогда из (23) имеем $\langle J'(u_k), u - u_k \rangle \geq 0$ 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 $k = 0, 1, \dots$.

Положим в (23) $u = u_k$. Получим

$$\langle J'(u_k) + J''(u_k)(u_{k+1} - u_k), u_k - u_{k+1} \rangle \geq 0.$$

Отсюда и из (21) имеем

$$\begin{aligned} \mu |u_{k+1} - u_k|^2 &\leq \langle J''(u_k)(u_{k+1} - u_k), u_{k+1} - u_k \rangle \leq \\ &\leq \langle J'(u_k), u_k - u_{k+1} \rangle, \quad k = 0, 1, \dots \end{aligned} \quad (24)$$

Оценим правую часть (24) сверху. Для этого в (22) заменим k на $k-1$. Получим $\langle J'_{k-1}(u_k), u - u_k \rangle \geq 0, u \in U$. Полагая здесь $u = u_{k+1}$, имеем

$$\langle J'_{k-1}(u_k), u_k - u_{k+1} \rangle \leq 0, \quad k = 1, 2, \dots$$

Отсюда, из формулы (2.3.5) и условия (18) следует

$$\begin{aligned} \langle J'(u_k), u_k - u_{k+1} \rangle &\leq \langle J'(u_k) - J'_{k-1}(u_k), u_k - u_{k+1} \rangle = \\ &= \langle J'(u_k) - J'(u_{k-1}) - J''(u_{k-1})(u_k - u_{k-1}), u_k - u_{k+1} \rangle = \\ &= \left\langle \int_0^1 [J''(u_{k-1} + t(u_k - u_{k-1})) - J''(u_{k-1})] dt (u_k - u_{k-1}), u_k - u_{k+1} \right\rangle \leq \\ &\leq (L/2) |u_k - u_{k-1}|^2 |u_k - u_{k+1}|, \quad k = 1, 2, \dots \end{aligned}$$

Подставив полученную оценку в (24), получим

$$|u_{k+1} - u_k| \leq (L/2\mu) |u_k - u_{k-1}|^2, \quad k = 1, 2, \dots \quad (25)$$

Докажем оценку

$$|u_{k+1} - u_k| \leq (2\mu/L) q^{2^k}, \quad k = 0, 1, \dots \quad (26)$$

При $k=0$ эта оценка следует из условия (19). Сделаем индуктивное предположение: пусть $|u_k - u_{k-1}| \leq (2\mu/L) q^{2^{k-1}}$ при некотором $k \geq 1$. Отсюда и из (25) имеем $|u_{k+1} - u_k| \leq (L/2\mu) (2\mu/L)^2 (q^{2^{k-1}})^2 = (2\mu/L) q^{2^k}$. Оценка (26)

доказана. Из (26) следует

$$\begin{aligned} |u_k - u_p| &\leq \sum_{m=k}^{p-1} |u_{m+1} - u_m| \leq \sum_{m=k}^{p-1} (2\mu/L) q^{2^m} \leq \\ &\leq \sum_{m=k}^{\infty} (2\mu/L) q^{2^m} \leq (2\mu/L) q^{2^k} (1 - q^{2^k})^{-1} \end{aligned} \quad (27)$$

для всех $p, k, p > k \geq 0$. Так как $0 < q < 1$, то правая часть (27) стремится к нулю при $k \rightarrow \infty$. 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 $p \rightarrow \infty$, 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 $k \rightarrow \infty$ из (23) имеем $\langle J'(u_*), u - u_* \rangle \geq 0$ 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 $|u_0 - u_*| \leq (2\mu/L) q(1 - q)^{-1}$. 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

$$\mu \|\xi\|^2 \leq \langle J''(u) \xi, \xi \rangle \leq M \|\xi\|^2, \quad u \in U, \quad \xi \in L_U, \quad (28)$$

где L_U — подпространство, параллельное аффинной оболочке множества U , а μ, M — постоянные, $0 < \mu \leq M$. 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 $k \geq k_0$, и справедлива оценка

$$|u_k - u_*| \leq (2\mu/L) \sum_{m=k}^{\infty} q^{2^m} \leq (2\mu/L) q^{2^k} (1 - q^{2^k})^{-1}, \quad k \geq k_0. \quad (29)$$

Доказательство. Согласно теореме 4.3.4 $J(u)$ сильно выпукла. Тогда из теоремы 4.3.1 следует существование и единственность точки u_* . Пусть u_0 — произвольная точка из U . Допустим, что точка $u_k \in U$ при некотором $k \geq 0$ уже известна. Поскольку $J'_k(u) = J''(u_k)$, то из условия (28) и теоремы 4.3.4 следует сильная выпуклость квадратичной функции (2), а из теоремы 4.3.1 — существование и единственность точки $\bar{u}_k$, удовлетворяющей условиям (3). Согласно теореме 4.2.3 тогда $\langle J'_k(\bar{u}_k), u - \bar{u}_k \rangle \geq 0$ или

$$\langle J'(u_k) + J''(u_k)(\bar{u}_k - u_k), u - \bar{u}_k \rangle \geq 0 \quad \text{при всех } u \in U. \quad (30)$$

Если оказалось, что $\bar{u}_k = u_k$, то из (30) имеем $\langle J'(u_k), u - u_k \rangle \geq 0, u \in U$. 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 $i \geq 0$, для которого выполняется условие (10). С этой целью возьмем произ-

вольное число α , $0 \leq \alpha \leq 1$, и положим $u_\alpha = u_k + \alpha(\bar{u}_k - u_k)$. Отсюда и из выпуклости $J_k(u)$ следует

$$J_k(u_\alpha) \leq \alpha J_k(\bar{u}_k) + (1 - \alpha) J_k(u_k) = \alpha J_k(\bar{u}_k) < 0.$$

Тогда из формулы

$$J(u_\alpha) - J(u_k) = J_k(u_\alpha) + (\alpha^2/2) \langle J''(u_k + \theta \alpha(\bar{u}_k - u_k)) - J''(u_k) \rangle (\bar{u}_k - u_k), \quad 0 \leq \alpha \leq 1, \quad (31)$$

с учетом условий (28) получим

$$J(u_\alpha) - J(u_k) \leq J_k(u_\alpha) + (\alpha^2/2) (M - \mu) |\bar{u}_k - u_k|^2 \leq \alpha J_k(\bar{u}_k) + (\alpha^2/2) M |\bar{u}_k - u_k|^2, \quad 0 \leq \alpha \leq 1. \quad (32)$$

Так как $\bar{u}_k$ — точка минимума сильно выпуклой функции $J_k(u)$ на U , то согласно теореме 4.3.1

$$|u_k - \bar{u}_k|^2 \leq (2/\mu) [J_k(u_k) - J_k(\bar{u}_k)] = (2/\mu) |J_k(\bar{u}_k)|. \quad (33)$$

Подставив эту оценку в (32), получим

$$J(u_\alpha) - J(u_k) \leq -\alpha |J_k(\bar{u}_k)| + \alpha^2 (M/\mu) |J_k(\bar{u}_k)|, \quad 0 \leq \alpha \leq 1.$$

Возьмем произвольное α , удовлетворяющее условиям

$$0 < \varepsilon_0 = \lambda(1 - \varepsilon) \mu/M \leq \alpha \leq (1 - \varepsilon) \mu/M < 1. \quad (34)$$

Отсюда и из предыдущего неравенства будем иметь

$$J(u_k) - J(u_k + \alpha(\bar{u}_k - u_k)) \geq \alpha(1 - \alpha(M/\mu)) |J_k(\bar{u}_k)| \geq \varepsilon \alpha |J_k(\bar{u}_k)| \quad (35)$$

при всех α , удовлетворяющих условиям (34). Возьмем такой номер $m \geq 1$, для которого $\lambda^m \leq (1 - \varepsilon) \mu/M < \lambda^{m-1}$. Отсюда следует, что

$$0 < \varepsilon_0 = \lambda(1 - \varepsilon) \mu/M < \lambda^m \leq (1 - \varepsilon) \mu/M. \quad (36)$$

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 $i = i_0$, $0 \leq i_0 \leq m$, 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

$$J(u_k) - J(u_{k+1}) \geq \varepsilon \alpha_k |J_k(\bar{u}_k)|, \quad k = 0, 1, \dots$$

Учитывая, что согласно (36) $\alpha_k = \lambda^{i_0} \geq \lambda^m > \varepsilon_0$, отсюда получим

$$J(u_k) - J(u_{k+1}) \geq \varepsilon \varepsilon_0 |J_k(\bar{u}_k)|, \quad k = 0, 1, \dots \quad (37)$$

Таким образом, $J(u_k) \geq J(u_{k+1}) \geq J_*$, $k = 0, 1, \dots$. Тогда существует $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$ и $\lim_{k \rightarrow \infty} (J(u_k) - J(u_{k+1})) = 0$. Из (37) теперь имеем $\lim_{k \rightarrow \infty} J_k(\bar{u}_k) = 0$, а из (33) следует

$$\lim_{k \rightarrow \infty} |u_k - \bar{u}_k| = 0. \quad (38)$$

Далее заметим, что согласно (37) $\{u_k\} \in M(u_0) = \{u: u \in U, J(u) \leq J(u_0)\}$. 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 $\{u_{k_m}\} \rightarrow v_*$. С учетом (38) и условия $J(u) \in C^2(U)$ из (30) при $k = k_m \rightarrow \infty$ получим $\langle J'(v_*), u - v_* \rangle \geq 0$ для всех $u \in U$. Согласно теореме 4.2.3 тогда $v_* = u_*$ — точка минимума $J(u)$ на U . Следовательно, $\lim_{k \rightarrow \infty} J(u_k) = \lim_{m \rightarrow \infty} J(u_{k_m}) = J(u_*) = J_*$, т. е. $\{u_k\}$ — минимизирующая последовательность. Отсюда и из теоремы 4.3.1 следует $\{u_k\} \rightarrow u_*$.

Пусть теперь выполнено условие (18). В силу (38) существует номер k_0 такой, что $(L/\mu) |\bar{u}_k - u_k| \leq 1 - \varepsilon$ при всех $k \geq k_0$. Из (31) с учетом условия (18) и оценки (33) тогда имеем

$$J(u_\alpha) - J(u_k) \leq \alpha J_k(\bar{u}_k) + (\alpha^3/2) L |\bar{u}_k - u_k|^3 \leq \\ \leq -\alpha |J_k(\bar{u}_k)| + \alpha^2 (L/\mu) |J_k(\bar{u}_k)| |\bar{u}_k - u_k|,$$

т. е.

$$J(u_k) - J(u_\alpha) \geq |J_k(\bar{u}_k)| \alpha (1 - \alpha (L/\mu) |\bar{u}_k - u_k|) \geq \alpha \varepsilon |J_k(\bar{u}_k)|$$

при всех α , $\varepsilon_0 \leq \alpha \leq 1$, $k \geq k_0$. 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 $k \geq k_0$. 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

$$q = (L/2\mu) |u_{k_0+1} - u_{k_0}| = (L/2\mu) |\bar{u}_{k_0} - u_{k_0}| \leq (1 - \varepsilon)/2 < 1.$$

Отсюда и из теоремы 2 следует оценка (29), что и требовалось.

Таким образом, метод (2) — (4), (10) ненамного 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 сходимости описанных выше вариантов метода Ньютона читатель может найти в [15].

§ 8. Метод с кубической скоростью сходимости

Описанный выше метод Ньютона на каждой итерации требует вычисления матрицы вторых производных. Однако возможны случаи, когда вычисление матрицы $J''(u)$ требует значительного объема вычислений, из-за чего трудоемкость каждой итерации метода Ньютона может стать непомерно большой. Поэтому возникает вопрос о возможности построения методов минимизации, которые по скорости сходимости были бы близки к методу Ньютона или даже превосходили его, но для своей реализации требовали бы меньшего количества вычислений на каждой итерации.

В настоящее время уже разработан и исследован ряд методов, обладающих указанными свойствами [15, 249]. Следуя [238], здесь мы изложим один из таких методов.

Пусть функция $J(u) \in C^2(E^n)$. Будем рассматривать задачу

$$J(u) \rightarrow \inf; \quad u \in U = E^n. \quad (1)$$

Введем матрицу

$$J'(u, v) = \{J'_{ij}(u, v), \quad i, j = \overline{1, n}\},$$

называемую разделенной разностью для градиента $J'(u)$ в точках u и v и определяемую следующим образом:

$$J'_{ij}(u, v) = \begin{cases} \frac{J_{u^i}^j(v^1, \dots, v^{l-1}, u^l, u^{l+1}, \dots, u^n) - J_{u^i}^j(v^1, \dots, v^{l-1}, v^l, u^{l+1}, \dots, u^n)}{u^l - v^l} \\ \text{при } u^l \neq v^l, \\ J_{u^i u^l}^j(v^1, \dots, v^{l-1}, v^l, u^{l+1}, \dots, u^n) \quad \text{при } u^l = v^l; \end{cases} \quad (2)$$

здесь $J'_{ij}(u, v)$ — элемент i -й строки j -го столбца матрицы $J'(u, v)$, а $J_{u^i}^j(u)$, $J_{u^i u^l}^j(u)$, как обычно, обозначают первые и соответственно вторые производные по переменным u^i , u^l ; $i = \overline{1, n}$; $j = \overline{1, n}$.

Из (2) нетрудно получить, что

$$\sum_{i=1}^n J'_{ij}(u, v)(u^i - v^i) = \sum_{i=1}^n (J_{u^i}^j(v^1, \dots, v^{l-1}, u^l, \dots, u^n) - J_{u^i}^j(v^1, \dots, v^{l-1}, v^l, \dots, u^n)) = J_{u^i}^j(u) - J_{u^i}^j(v), \quad i = \overline{1, n},$$

т. е.

$$J'(u, v)(u - v) = J'(u) - J'(v), \quad u, v \in E^n. \quad (3)$$

Поскольку $J(u) \in C^2(E^n)$, то из (2) следует

$$J'(u, v) = J''(u) + \varepsilon(|u - v|), \quad u, v \in E^n, \quad (4)$$

где $\|\varepsilon(t)\| \rightarrow 0$ при $t \rightarrow +0$.

Предположим, что матрица $J'(u, v)$ — невырожденная при всех $u, v \in E^n$. Рассмотрим следующий двухшаговый итерационный процесс, на первом шаге которого решается система

$$J'(u_k, u_k - \beta_k J'(u_k))(u_k - \bar{u}_k) = J'(u_k) \quad (5)$$

и определяется вспомогательное приближение $\bar{u}_k$, а на втором шаге из системы

$$J'(u_k, u_k - \beta_k J'(u_k))(\bar{u}_k - u_{k+1}) = J'(\bar{u}_k) \quad (6)$$

находится следующее $(k+1)$ -е приближение u_{k+1} ; здесь $\beta_k \neq 0$ — параметр метода, $k = 0, 1, \dots$

Таким образом, на каждой итерации описанного метода (5), (6) решаются две системы линейных алгебраических уравнений, причем обе системы имеют одну и ту же матрицу $J'(u_k, u_k - \beta_k J'(u_k))$ и отличаются лишь правыми частями. Если $J_{u^j}(u_k) \neq 0$ при всех $j = \overline{1, n}$, то согласно формуле (2) для определения матрицы $J'(u_k, u_k - \beta_k J'(u_k))$ достаточно знать первые производные $J_{u^i}^j(u)$ в точках $u = u_k$ и $u = (u_k^1 - \beta_k J_{u^1}^1(u_k), \dots, u_k^j - \beta_k J_{u^j}^j(u_k), u_k^{j+1}, \dots, u_k^n)$, $j = \overline{1, n}$. Если же $J_{u^j}(u_k) = 0$ при некотором j , то в силу (2) для определения j -го столбца матрицы $J'(u_k, u_k - \beta_k J'(u_k))$ придется вычислить вторые производные $J_{u^i u^j}^k(u)$ в точке $u = (u_k^1 - \beta_k J_{u^1}^1(u_k), \dots, u_k^{j-1} - \beta_k J_{u^{j-1}}^{j-1}(u_k), u_k^j, \dots, u_k^n)$. Если при некотором $k \geq 0$ оказалось $J'(u_k) = 0$, то процесс (5), (6) 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 $J'(u_k, u_k - \beta_k J'(u_k))$ потребуются вычислить заведомо не все вторые производные $J_{u^i u^j}^k(u)$, и в этом смысле метод (5), (6) имеет преимущество перед методом Ньютона. Кроме того, оказывается, метод (5), (6) на классе сильно выпуклых функций сходится быстрее метода Ньютона (7.8). А именно, справедлива

Теорема 1. Пусть функция $J(u)$ сильно выпукла на E^n , $J(u) \in C^2(E^n)$ и

$$\|J'(u, w) - J'(w, v)\| \leq L(|u - w| + |w - v|), \quad L = \text{const} > 0, \quad (7)$$

для всех $u, v \in S_0 = \{u \in E^n: |u - u_0| \leq 4\mu^{-1}|J'(u_0)|\}$, где $\mu > 0$ — постоянная из теорем 4.3.3, 4.3.4. Пусть, кроме того,

$$0 < |\beta_k| \leq 2\mu^{-1}, \quad k = 0, 1, \dots; \quad h = 8L\mu^{-2}|J'(u_0)| < 1, \quad q = Ah < 1, \quad (8)$$

где $A = ((5 + 3h)/2)^{1/2}$. Тогда последовательность $\{u_k\}$, определяемая условиями (5), (6), при достаточно малых значениях $\{\beta_k\}$ существует и сходится к точке минимума u_* функции $J(u)$ на E^n , причем справедлива оценка

$$|u_k - u_*| \leq (\mu/8LA)q^{3k}, \quad k = 1, 2, \dots \quad (9)$$

Доказательство. Существование и единственность точки u_* — решения задачи (1) — следует из теоремы 4.3.1, а равенство $J'(u_*) = 0$ — из теоремы 4.2.3. Кроме того, в силу теоремы 4.3.1 множество S_0 выпукло, замкнуто и ограничено. Тогда $\sup_{S_0} |J'(u)| = B < \infty$, и из формулы (4) получим

$$\|J'(u, u - \beta J'(u)) - J''(u)\| = \|\varepsilon(|\beta|B)\| \rightarrow 0$$

при $\beta \rightarrow 0$ равномерно для всех $u \in S_0$. Следовательно, существует число $\bar{\beta}$, $0 < \bar{\beta} \leq 2\mu^{-1}$, такое, что

$$\|J'(u, u - \beta J'(u)) - J''(u)\| \leq \mu/2$$

при всех β , $|\beta| \leq \bar{\beta}$, и всех $u \in S_0$. Отсюда с помощью теоремы 4.3.4 имеем

$$\begin{aligned} \langle J'(u, u - \beta J'(u)) \xi, \xi \rangle &\geq \langle J''(u) \xi, \xi \rangle + \langle [J'(u, u - \beta J'(u)) - J''(u)] \xi, \xi \rangle \geq \\ &\geq \mu |\xi|^2 - (\mu/2) |\xi|^2 = (\mu/2) |\xi|^2 \quad (10) \end{aligned}$$

для всех $\xi \in E^n$, $u \in S_0$, $|\beta| \leq \bar{\beta}$. Из (10) следует, что система уравнений $J'(u, u - \beta J'(u)) \xi = 0$ имеет единственное решение $\xi = 0$, так что матрица $J'(u, u - \beta J'(u))$ — невырожденная при всех $u \in S_0$ и любом выборе $|\beta| \leq \bar{\beta}$. Далее, полагая в (10) $\xi = (J'(u, u - \beta J'(u)))^{-1} z$, получим

$$\begin{aligned} (\mu/2) |(J'(u, u - \beta J'(u)))^{-1} z|^2 &\leq \langle z, (J'(u, u - \beta J'(u)))^{-1} z \rangle \leq \\ &\leq |z| |(J'(u, u - \beta J'(u)))^{-1} z| \end{aligned}$$

или

$$|(J'(u, u - \beta J'(u)))^{-1} z| \leq (2/\mu) |z|$$

при всех $z \in E^n$. Это значит, что

$$\|(J'(u, u - \beta J'(u)))^{-1}\| \leq 2/\mu \quad (11)$$

при всех $u \in S_0$, $|\beta| \leq \bar{\beta}$.

С помощью индукции покажем, что последовательность $\{u_k\}$, определяемая условиями (5), (6), при любом выборе β_k , лишь бы $|\beta_k| \leq \bar{\beta}$, существует, принадлежит шару S_0 и удовлетворяет неравенствам

$$|J'(u_k)| \leq q^{3k-1} |J'(u_0)|, \quad k = 1, 2, \dots \quad (12)$$

Введем обозначения

$$a_k = |J'(u_k)|, \quad w_k = u_k - \beta_k J'(u_k), \quad \Gamma_k = (J'(u_k, w_k))^{-1}. \quad (13)$$

Тогда итерационный процесс (5), (6) можно переписать в виде

$$\bar{u}_k = u_k - \Gamma_k J'(u_k), \quad u_{k+1} = \bar{u}_k - \Gamma_k J'(\bar{u}_k), \quad k = 0, 1, \dots \quad (14)$$

Сделаем предварительные оценки, предполагая, что точки u_k , $\bar{u}_k$, w_k , u_{k+1} из (13), (14) существуют и принадлежат S_0 . Согласно (11) $\|\Gamma_k\| \leq 2/\mu$, $k = 0, 1, \dots$. Отсюда и из (8), (13), (14) имеем

$$|\bar{u}_k - u_k| \leq 2a_k \mu^{-1}, \quad |u_k - w_k| \leq |\beta_k| a_k \leq 2a_k \mu^{-1}, \quad (15)$$

$$|u_{k+1} - \bar{u}_k| \leq |J'(\bar{u}_k)| 2\mu^{-1},$$

$$\begin{aligned} |u_{k+1} - w_k| &\leq |u_k - \bar{u}_k| + |\beta_k| a_k + 2|J'(\bar{u}_k)| \mu^{-1} \leq \\ &\leq 4a_k \mu^{-1} + 2|J'(\bar{u}_k)| \mu^{-1}. \quad (16) \end{aligned}$$

Далее, из свойства (3) разделенных разностей для градиента и (13), (14) следует

$$\begin{aligned} J'(\bar{u}_k) &= J'(u_k) + J(\bar{u}_k, u_k)(\bar{u}_k - u_k) = \\ &= (J'(u_k, w_k) - J'(\bar{u}_k, u_k)) \Gamma_k J'(u_k). \end{aligned}$$

Отсюда с помощью условия (7) и оценок (15) получим

$$|J'(\bar{u}_k)| \leq L(|u_k - w_k| + |\bar{u}_k - u_k|) 2\mu^{-1}a_k \leq 8La_k^2\mu^{-2}. \quad (17)$$

Снова обращаясь к (3), (13), (14), имеем

$$\begin{aligned} J'(u_{k+1}) &= J'(\bar{u}_k) + J'(u_{k+1}, \bar{u}_k)(u_{k+1} - \bar{u}_k) = \\ &= [J'(u_k, w_k) - J'(w_k, u_{k+1}) + J'(w_k, u_{k+1}) - J'(u_{k+1}, \bar{u}_k)] \Gamma_k J'(\bar{u}_k). \end{aligned}$$

Отсюда и из оценок (7), (15)–(17) вытекает

$$\begin{aligned} |J'(u_{k+1})| &= a_{k+1} \leq \\ &\leq L(|u_k - w_k| + 2|w_k - u_{k+1}| + |u_{k+1} - \bar{u}_k|) 2\mu^{-1}|J'(\bar{u}_k)| \leq \\ &\leq L[2a_k\mu^{-1} + 2(4a_k\mu^{-1} + 2|J'(\bar{u}_k)|\mu^{-1}) + 2|J'(\bar{u}_k)|\mu^{-1}] 2\mu^{-1}|J'(\bar{u}_k)| \leq \\ &\leq L(10a_k\mu^{-1} + 6\mu^{-1} \cdot 8La_k^2\mu^{-2}) 2\mu^{-1} \cdot 8La_k^2\mu^{-2} = \\ &= 32L^2\mu^{-4}a_k^3(5 + 24La_k\mu^{-2}), \end{aligned}$$

т. е.

$$a_{k+1} \leq 32L^2\mu^{-4}a_k^3(5 + 24La_k\mu^{-2}). \quad (18)$$

Еще раз подчеркнем, что оценки (15)–(18) получены в предположении, что точки $u_k, \bar{u}_k, w_k, u_{k+1}$ из (13), (14) существуют и принадлежат S_0 .

Теперь приступим к доказательству существования последовательности $\{u_k\}$, включения $\{u_k\} \in S_0$ и оценки (12). Заметим, что $u_0 \in S_0$ по определению S_0 . Отсюда и из (15) следует, что $|u_0 - w_0| \leq 2a_0\mu^{-1}$, т. е. $w_0 \in S_0$. Тогда $\Gamma_0 = (J'(u_0, u_0 - \beta_0 J'(u_0)))^{-1}$ при $|\beta_0| \leq \bar{\beta}$ существует, и точки $\bar{u}_0, u_1$ из (5), (6) при $k=0$ определяются однозначно. Кроме того, из условий (8) и оценок (15), (17) имеем $|\bar{u}_0 - u_0| \leq 2a_0\mu^{-1}$, т. е. $\bar{u}_0 \in S_0$, и $|J'(\bar{u}_0)| \leq 8La_0^2\mu^{-2} = ha_0 \leq a_0$. Тогда

$$|u_1 - u_0| = |\bar{u}_0 - \Gamma_0 J'(\bar{u}_0) - u_0| \leq 4a_0\mu^{-1}, \text{ т. е. } u_1 \in S_0.$$

Итак, $u_0, \bar{u}_0, w_0, u_1 \in S_0$. Поэтому можем пользоваться оценкой (18) при $k=0$. Получим

$$a_1 \leq 8^2L^2\mu^{-4}a_0^2(a_0/2)(5 + 3 \cdot 8L\mu^{-2}a_0) = h^2a_0(5 + 3h)/2 = q^2a_0.$$

Справедливость оценки (12) при $k=1$ установлена. Допустим, что $u_k \in S_0$ и оценка (12) верна при некотором $k \geq 1$. Поскольку $q < 1$, то из (12) следует $a_k \leq a_0$. Отсюда и из неравенства

$$\mu|u - v|^2 \leq \langle J'(u) - J'(v), u - v \rangle, \quad u, v \in E^n, \quad (19)$$

имеем $|u_k - u_0| \leq \mu^{-1} |J'(u_k) - J'(u_0)| \leq 2\mu^{-1}a_0$. Тогда $|w_k - u_0| = |u_k - \beta_k J'(u_k) - u_0| \leq 4\mu^{-1}a_0$, т. е. $w_k \in S_0$. Это значит, что $\Gamma_k = (J'(u_k, u_k - \beta_k J'(u_k)))^{-1}$ при $|\beta_k| \leq \bar{\beta}$ существует и точки $\bar{u}_k, u_{k+1}$ из (5), (6) определяются однозначно. Из (15) следует $|\bar{u}_k - u_0| \leq |\bar{u}_k - u_k| + |u_k - u_0| \leq 2a_k\mu^{-1} + 2a_0\mu^{-1} \leq 4a_0\mu^{-1}$, т. е. $\bar{u}_k \in S_0$. Тогда можем воспользоваться оценкой (17), из которой получим

$$|J'(\bar{u}_k)| \leq 8La_k^2\mu^{-2} \leq 8La_0^2\mu^{-2} = ha_0 \leq a_0.$$

Отсюда и из (19) вытекает уточненная оценка $|\bar{u}_k - u_0| \leq \mu^{-1} |J'(\bar{u}_k) - J'(u_0)| \leq 2a_0\mu^{-1}$. Поэтому с учетом первого неравенства (16) будем иметь $|u_{k+1} - u_0| \leq |u_{k+1} - \bar{u}_k| + |\bar{u}_k - u_0| \leq 2\mu^{-1} |J'(\bar{u}_k)| + 2a_0\mu^{-1} \leq 4a_0\mu^{-1}$, т. е. $u_{k+1} \in S_0$.

Итак, показано, что если $u_k \in S_0$ и оценка (12) верна при $k \geq 1$, то $\bar{u}_k, w_k, u_{k+1}$ существуют и принадлежат S_0 . Тогда можем пользоваться оценкой (18), из которой с помощью предположения индукции получим

$$a_{k+1} \leq 8^2 L^2 \mu^{-4} a_0^2 a_0^{-2} a_k^3 (5 + 3 \cdot 8L\mu^{-2} a_0) / 2 = \\ = q^2 a_0^{-2} a_k^3 \leq q^2 a_0^{-2} (q^{3^k-1} a_0)^3 = a_0 q^{3^{k+1}-1}.$$

Все этапы индукции проведены, существование последовательности $\{u_k\} \in S_0$ и оценка (12) доказаны. Из неравенства (19) и оценки (12) имеем

$$|u_k - u_*| \leq a_k \mu^{-1} \leq q^{3^k-1} a_0 \mu^{-1} = q^{3^k} (\mu/8AL), \quad k = 1, 2, \dots$$

Оценка (9) получена. Теорема доказана.

Таким образом, метод (5), (6) на классе сильно выпуклых функций имеет кубическую скорость сходимости, т. е. по скорости сходимости превосходит метод Ньютона. Недостатком метода (5), (6), как это видно из условий (8), является требование достаточной близости начальной точки u_0 к искомой точке u_* .

§ 9. Метод покоординатного спуска

В предыдущих параграфах мы рассмотрели методы, которые для своей реализации требуют вычисления первых или даже 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 [2, 9, 100, 110, 244].

1. Сначала опишем этот метод для задачи

$$J(u) \rightarrow \inf; \quad u \in U = E^n. \quad (1)$$

Обозначим $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ — единичный координатный (базисный) вектор, у которого i -я координата равна 1, остальные равны нулю, $i = 1, n$. 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 $k \geq 0$. Положим

$$p_k = e_{i_k}, \quad i_k = k - n \left[\frac{k}{n} \right] + 1, \quad (2)$$

где $\left[\frac{k}{n} \right]$ означает целую часть числа k/n . Условие (2) обеспечивает циклический перебор координатных векторов $e_1, e_2, \dots, \dots, e_n$, т. е.

$$p_0 = e_1, \dots, p_{n-1} = e_n, \quad p_n = e_1, \dots, p_{2n-1} = e_n, \\ p_{2n} = e_1, \dots$$

Вычислим значение функции $J(u)$ в точке $u = u_k + \alpha_k p_k$ и проверим неравенство

$$J(u_k + \alpha_k p_k) < J(u_k). \quad (3)$$

Если (3) выполняется, то примем

$$u_{k+1} = u_k + \alpha_k p_k, \quad \alpha_{k+1} = \alpha_k. \quad (4)$$

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

$$J(u_k - \alpha_k p_k) < J(u_k). \quad (5)$$

В случае выполнения (5) положим

$$u_{k+1} = u_k - \alpha_k p_k, \quad \alpha_{k+1} = \alpha_k. \quad (6)$$

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

$$u_{k+1} = u_k, \quad \alpha_{k+1} = \begin{cases} \lambda \alpha_k & \text{при } i_k = n, \quad u_k = u_{k-n+1}, \\ \alpha_k & \text{при } i_k \neq n \text{ или } u_k \neq u_{k-n+1} \\ & \text{или } 0 \leq k \leq n-1; \end{cases} \quad (7)$$

здесь λ , $0 < \lambda < 1$, — фиксированно число, являющееся параметром метода. Условия (7) означают, что если за один цикл из n итераций при переборе направлений всех координатных осей $e_1, \dots, e_n$ с шагом α_k реализовалась хотя бы одна удачная итерация, то длина шага α_k не дробится и сохраняется на про-

тяжении 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

$$J(u_{k_m} + \alpha_{k_m} e_i) \geq J(u_{k_m}), \quad J(u_{k_m} - \alpha_{k_m} e_i) \geq J(u_{k_m}) \quad (8)$$

при всех $i = 1, 2, \dots, n$. 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 $M(u_0) = \{u \in E^n: J(u) \leq J(u_0)\}$ ограничено. Тогда последовательность $\{u_k\}$, получаемая описанным методом (2) — (7), минимизирует функцию $J(u)$ на E^n и сходится к множеству U_* .

Доказательство. Согласно теореме 2.1.2 $J_* > -\infty$, $U_* \neq \emptyset$. Из описания метода (2) — (7) следует, что $J(u_{k+1}) \leq J(u_k)$, $k = 0, 1, \dots$, так что $\{u_k\} \in M(u_0)$ и существует $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Покажем, что найдется бесконечно много номе-

ров $k_1, k_2, \dots, k_m, \dots$ итераций, на которых шаг α_k дробится, и поэтому $\lim_{k \rightarrow \infty} \alpha_k = 0$. Допустим противное: пусть процесс дробления

α_k конечен, т. е. $\alpha_k = \alpha > 0$ при всех $k \geq N$. Обозначим $M_\alpha = \{u: u = u_N + \alpha r e_i \in M(u_0), i = \overline{1, n}, r = 0, \pm 1, \pm 2, \dots\}$ — сетку (решетку) с шагом α . Из описания метода покоординатного спуска при $\alpha_k = \alpha$, $k \geq N$, 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 $i = 1, 2, \dots, n$. А тогда вопреки допущению придется дробить число $\alpha_k = \alpha$. Полученное противоречие показывает, что процесс дробления α_k бесконечен и $\lim_{k \rightarrow \infty} \alpha_k = 0$.

Пусть $k_1 < k_2 < \dots < k_m < \dots$ — 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

$$\begin{aligned} \langle J'(u_{k_m} + \theta_m \alpha_{k_m} e_i), e_i \rangle \alpha_{k_m} &\geq 0, \\ \langle J'(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i), e_i \rangle (-\alpha_{k_m}) &\geq 0, \end{aligned}$$

откуда

$$J_{u^i}(u_{k_m} + \theta_m \alpha_{k_m} e_i) \geq 0, \quad J_{u^i}(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i) \leq 0,$$

$0 \leq \theta_m, \bar{\theta}_m \leq 1$ при всех $i = 1, 2, \dots, n$ и $m = 1, 2, \dots$. Пользуясь тем, что $J(u) \in C^1(E^n)$ и $\lim_{m \rightarrow \infty} \alpha_{k_m} = 0$, откуда получим

$J_{u^i}(u_*) = 0, \quad i = \overline{1, n}$, т.е. $J'(u_*) = 0$. В силу выпуклости $J(u)$ тогда $u_* \in U_*$. Следовательно, $\lim_{k \rightarrow \infty} J(u_k) = \lim_{m \rightarrow \infty} J(u_{k_m}) = J(u_*) = J_*$. Таким образом, последовательность $\{u_k\}$ является минимизирующей. Отсюда и из теоремы 2.1.2 следует, что $\rho(u_k, U_*) \rightarrow 0$ при $k \rightarrow \infty$. Теорема 1 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 $J(u_0 + \alpha e_1) = J(\alpha e_1) = \alpha^2 - 2\alpha + 2|\alpha| \geq 0 = J(0)$, $J(u_0 + \alpha e_2) = J(\alpha e_2) = \alpha^2 - 2\alpha + 2|\alpha| \geq 0 = J(0)$ 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 $k = 0, 1, \dots$. 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:

$$J(u) \rightarrow \inf; \quad u \in U = \{u^1, \dots, u^n: a_i \leq u^i \leq b_i, \quad i = \overline{1, n}\}, \quad (9)$$

где a_i, b_i — заданные числа, $a_i < b_i, i = \overline{1, n}$. А именно, пусть k -е приближение $u_k \in U$ и число $\alpha_k > 0$ при некотором $k \geq 0$

уже найдены. Выберем вектор $p_k = e_{i_k}$ согласно формуле (2), составим точку $u_k + \alpha_k p_k$ и проверим условия

$$u_k + \alpha_k p_k \in U, \quad J(u_k + \alpha_k p_k) < J(u_k). \quad (10)$$

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

$$u_k - \alpha_k p_k \in U, \quad J(u_k - \alpha_k p_k) < J(u_k). \quad (11)$$

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 равенств (7).

Теорема 2. Пусть функция $J(u)$ выпукла на U и $J(u) \in C^1(U)$. 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 к множеству решений задачи (9).

Доказательство. Так как U — параллелепипед, то множество $M(u_0) = \{u: u \in U, J(u) \leq J(u_0)\}$ ограничено. Так как $J(u_{k+1}) \leq J(u_k)$, $k = 0, 1, \dots$, то $\{u_k\} \in U$ и существует $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Так же, как в теореме 1, доказывается существование бесконечного числа номеров $k_1 < k_2 < \dots < k_m < \dots$ итераций, на которых длина шага α_k дробится, и поэтому $\lim_{k \rightarrow \infty} \alpha_k = 0$. В силу ограниченности $M(u_0)$ из $\{u_{k_m}\}$ можно выбрать сходящуюся подпоследовательность. Не умаляя общности, можем считать, что $\{u_{k_m}\} \rightarrow u_* = (u_*^1, \dots, u_*^n)$. При каждом $i = 1, 2, \dots, n$ возможны следующие три случая.

1) $a_i < u_*^i < b_i$. Так как $\lim_{k \rightarrow \infty} \alpha_k = 0$, 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 $m \geq N$. Поскольку α_k при $k = k_m$ дробится, то

$$J(u_{k_m} + \alpha_{k_m} e_i) \geq J(u_{k_m}), \quad J(u_{k_m} - \alpha_{k_m} e_i) \geq J(u_{k_m})$$

для всех $m \geq N$. Отсюда, как и в теореме 1, получаем $J_{u^i}(u_*) = 0$, так что

$$J_{u^i}(u_*)(u^i - u_*^i) = 0, \quad a_i \leq u^i \leq b_i.$$

2) $u_*^i = a_i$. Тогда $u_{k_m} + \alpha_{k_m} e_i \in U$ и $J(u_{k_m} + \alpha_{k_m} e_i) \geq J(u_{k_m})$ при всех $m \geq N$. Следовательно, $\langle J'(u_{k_m} + \theta_m \alpha_{k_m} e_i), e_i \rangle \alpha_{k_m} \geq 0$ или $J_{u^i}(u_{k_m} + \theta_m \alpha_{k_m} e_i) \geq 0$ для каждого $m \geq N$. Отсюда при $m \rightarrow \infty$ получим $J_{u^i}(u_*) \geq 0$ или $J_{u^i}(u_*)(u^i - a_i) = J_{u^i}(u_*)(u^i - u_*^i) \geq 0$, $a_i \leq u^i \leq b_i$.

3) $u_i^t = b_i$. Тогда $u_{k_m} - \alpha_{k_m} e_i \in U$ и $J(u_{k_m} - \alpha_{k_m} e_i) \geq J(u_{k_m})$ при всех $m \geq N$. Поэтому $\langle J'(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i), e_i \rangle (-\alpha_{k_m}) \geq 0$ или $J_{u^i}(u_{k_m} - \bar{\theta}_m \alpha_{k_m} e_i) \leq 0, m \geq N$. Отсюда при $m \rightarrow \infty$ получим $J_{u^i}(u_*) \leq 0$, следовательно,

$$J_{u^i}(u_*)(u^i - b_i) = J_{u^i}(u_*)(u^i - u_i^t) \geq 0, \quad a_i \leq u^i \leq b_i.$$

Объединяя все три рассмотренных случая, заключаем, что

$$J_{u^i}(u_*)(u^i - u_i^t) \geq 0, \quad a_i \leq u^i \leq b_i, \quad i = \overline{1, n}.$$

Суммируя эти неравенства по всем $i = \overline{1, n}$, получим

$$\langle J'(u_*), u - u_* \rangle \geq 0 \text{ для всех } u \in U.$$

Согласно теореме 4.2.3 тогда $u_* \in U_*$. Следовательно, $\lim_{k \rightarrow \infty} J(u_k) = \lim_{m \rightarrow \infty} J(u_{k_m}) = J(u_*) = J_*$, т. е. $\{u_k\}$ — минимизирующая последовательность. Отсюда и из теоремы 2.1.2 следует, что $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. Теорема 2 доказана.

3. Существуют и другие варианты метода покоординатного спуска. Можно, например, строить последовательность $\{u_k\}$ по правилу

$$u_{k+1} = u_k + \alpha_k p_k, \quad (12)$$

где p_k определяется согласно (2), а α_k — условиями

$$\alpha_k \geq 0, \quad f_k(\alpha_k) = \min_{-\infty < \alpha < +\infty} f_k(\alpha), \quad f_k(\alpha) = J(u_k + \alpha p_k). \quad (13)$$

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.

$$J(u) = \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle, \quad u \in E^n, \quad (14)$$

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 [2].

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 моди-

фикации рассматривавшихся в предыдущих параграфах методов, требующие вычисления лишь значений функции в подходящим образом выбранных точках. На этом пути, в частности, могут быть получены и методы высокого порядка сходимости, сравнимые с методом Ньютона, [15].

Другой подход для минимизации негладких функций, основанный лишь на вычислении значений функции, дает метод случайного поиска, который будет рассмотрен ниже в § 15. Метод поиска глобального минимума, излагаемый в следующем параграфе, также относится к методам, не требующим вычисления производных минимизируемой функции.

У п р а ж н е н и я. 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 $i, j, 1 \leq i, j \leq n$.

3. Докажите сходимость метода (12), (13) для функции (14) [2].

§ 10. Метод поиска глобального минимума

1. Заметим, что если задача минимизации

$$J(u) \rightarrow \inf; \quad u \in U,$$

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 удастся 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, кроме, например, непрерывности, 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 ме-

тодов решения задачи (1) на функцию, а также на множество приходится накладывать те или иные ограничения.

Ниже будет изложен метод поиска глобального минимума в задаче (1) для случая, когда множество U является параллелепипедом, т. е.

$$U = \{u = (u^1, u^2, \dots, u^n): a_i \leq u^i \leq b_i, \quad i = \overline{1, n}, \quad (2)$$

a_i, b_i — заданные числа, $a_i < b_i, i = \overline{1, n}$, а функция $J(u)$ удовлетворяет условию Липшица

$$|J(u) - J(v)| \leq L|u - v|, \quad u, v \in U, \quad L = \text{const} \geq 0. \quad (3)$$

Этот метод является обобщением метода (1.10.3), рассмотренного в § 1.10 для минимизации функции одной переменной на отрезке. Как и в § 1.10, через $Q(L)$ обозначим класс функций, удовлетворяющих условию (3) с одной и той же для всех функций этого класса константой $L \geq 0$.

Пусть p_m — какой-либо метод, представляющий собой правило выбора m точек $u_1, u_2, \dots, u_m$ из U , в которых вычисляются значения функции $J(u_1), \dots, J(u_m)$ и затем определяется величина $\min_{1 \leq i \leq m} J(u_i)$, принимаемая за приближенное значение $J_* = \inf_U J(u)$. Как и в § 1.10, введем величины

$$\Delta(J, p_m) = \min_{1 \leq i \leq m} J(u_i) - J_*, \quad \delta(p_m) = \sup_{J \in Q(L)} \Delta(J, p_m),$$

где $\Delta(J, p_m) \geq 0$ характеризует погрешность метода p_m при минимизации конкретной функции $J = J(u)$ из класса $Q(L)$, а $\delta(p_m)$ выражает собой погрешность метода p_m для самой «плохой» для этого метода функции из $Q(L)$ и называется гарантированной точностью метода p_m на классе $Q(L)$.

Зададимся вопросом: как выбрать число m и метод $p_m = \{u_1, \dots, u_m\}$ так, чтобы

$$\delta(p_m) \leq \varepsilon, \quad (4)$$

где $\varepsilon > 0$ — заданная точность? Поставленную задачу будем кратко называть задачей (1) — (4).

2. Можно предложить следующее правило выбора точек $u_1, u_2, \dots, u_m$: пусть эти точки из U таковы, что объединение шаров

$$S(u_i, R) = \{u \in E^n: |u - u_i| \leq R\}, \quad i = \overline{1, m},$$

с центрами в точках u_i и радиуса $R = \varepsilon/L$ покрывает множество U , т. е. $U \subset \bigcup_{i=1}^m S(u_i, R)$. Оказывается, при таком выборе точек $u_1, \dots, u_m$, приняв величину $\min_{1 \leq i \leq m} J(u_i)$ за приближение к J_* , мы решим задачу (1) — (4). В самом деле, возьмем любую точ-

ку $u \in U$. Так как шары $S(u_i, R)$, $i = \overline{1, n}$, покрывают множество U , то точка u принадлежит одному из шаров $S(u_i, R)$, т. е. $|u - u_i| \leq R = \varepsilon/L$. Отсюда и из условия (3) следует

$$J(u) \geq J(u_i) - L|u - u_i| \geq \min_{1 \leq i \leq m} J(u_i) - \varepsilon \text{ или}$$

$$J(u) \geq \min_{1 \leq i \leq m} J(u_i) - \varepsilon$$

для всех $u \in U$. Переходя в левой части этого неравенства к нижней грани по $u \in U$, будем иметь $0 \leq \min_{1 \leq i \leq m} J(u_i) - J_* \leq \varepsilon$

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

$$V_i = \{u = (u^1, \dots, u^n): u_i^j - h \leq u^j \leq u_i^j + h, j = \overline{1, n}\},$$

$$h = R/\sqrt{n} = \varepsilon/(L\sqrt{n}),$$

с длиной ребра $2h$ и с центром в точке $u_i = (u_i^1, \dots, u_i^n)$. Пользуясь этим обстоятельством, нетрудно покрыть параллелепипед (2) кубами следующим образом. Возьмем точки

$$u_{i_1, i_2, \dots, i_n} = (u_{i_1}^1, u_{i_2}^2, \dots, u_{i_j}^j, \dots, u_{i_n}^n), \quad i_j = \overline{1, m_j}, \quad j = \overline{1, n}, \quad (5)$$

где j -е координаты $u_1^j, u_2^j, \dots, u_{m_j}^j$ образованы по правилу

$$u_i^j = a_j + h(2i - 1), \quad i = \overline{1, m_j - 1};$$

$$u_{m_j}^j = \min \{b_j; a_j + h(2m_j - 1)\},$$

а номер m_j определяется условием

$$u_{m_j-1}^j < b_j - h \leq a_j + h(2m_j - 1) = u_{m_j}^j + 2h.$$

Тогда кубы $V_{i_1, i_2, \dots, i_n}$ с длиной ребра $2h = 2R/\sqrt{n}$ и с центром в точке $u_{i_1, i_2, \dots, i_n}$, $i_j = \overline{1, m_j}$, $j = \overline{1, n}$, покроют параллелепипед (2). Так как куб $V_{i_1, i_2, \dots, i_n}$ принадлежит шару $S(u_{i_1, i_2, \dots, i_n}, R)$, то объединение шаров $S(u_{i_1, i_2, \dots, i_n}, R)$, $i_j = \overline{1, m_j}$, $j = \overline{1, n}$, покроем параллелепипед (2). Тогда, как было показано выше, метод p_m , заключающийся в выборе точек $\{u_1, \dots, u_m\}$, $m = m_1 m_2 \dots m_n$, 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 $u_1, \dots, u_m$ задаются одновременно до начала вычислений значений функции. Между

тем, если точки $u_1, \dots, u_m$ выбирать последовательно и при выборе очередной точки u_i как-то учитывать результаты вычислений в предыдущих точках $u_1, \dots, u_{i-1}$, то есть шансы найти величину J_* с той же точностью $\varepsilon > 0$ за меньшее число вычислений, чем в описанном пассивном методе (5). Следуя [104] опишем один такой последовательный метод, одномерный вариант которого был изложен в § 1.10 (см. метод (1.10.3)).

Для простоты и наглядности этот метод сначала изложим для случая $n = 2$, когда

$$U = \{u = (u^1, u^2): a_1 \leq u^1 \leq b_1, a_2 \leq u^2 \leq b_2\}$$

— прямоугольник на плоскости E^2 . Введем обозначения

$$F_i = \min \{F_0, J(u_1), \dots, J(u_i)\}, \quad R = \varepsilon/L, \quad h = R/\sqrt{2} = \varepsilon/(L\sqrt{2}),$$

$$R_i = R + (J(u_i) - F_i)/L, \quad h_i = R_i/\sqrt{2}, \quad i = 1, 2, \dots \quad (6)$$

где точки $u_1, u_2, \dots$ 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

$$u_i, u_2, \dots, u_{m_1}, \quad u_i = (v_i^1, v_i^2), \quad i = \overline{1, m_1},$$

где

$$v_i^1 = a_1 + h, \quad v_{i+1}^1 = v_i^1 + h + h_i, \quad i = \overline{1, m_1 - 2};$$

$$v_{m_1}^1 = \min \{b_1; v_{m_1-1}^1 + h + h_{m_1-1}\}, \quad v_i^2 = a_2 + h, \quad (7)$$

номер m_1 находится из условий

$$v_{m_1-1}^1 < b_1 - h \leq v_{m_1-1}^1 + h + h_{m_1-1}.$$

Затем положим $d_{m_1} = \min \{h_1, h_2, \dots, h_{m_1}\}$ и введем прямоугольник

$$P_{m_1} = \{u = (u^1, u^2): a_1 \leq u^1 \leq b_1, a_2 \leq u^2 \leq v_1^2 + d_{m_1}\}.$$

Так как система отрезков $[v_i^1 - h, v_i^1 + h_i], i = \overline{1, m_1}$, покрывает отрезок $[a_1, b_1]$, то прямоугольник P_{m_1} будет принадлежать объединению прямоугольников

$$P_{m_1, i} = \{u = (u^1, u^2): v_i^1 - h \leq u^1 \leq v_i^1 + h_i,$$

$$v_i^2 - h = a_2 \leq u^2 \leq v_i^2 + d_{m_1}\}, \quad i = \overline{1, m_1}.$$

Так как $h \leq d_{m_1} \leq h_i, i = \overline{1, m_1}$, то для любой точки $u = (u^1, u^2) \in P_{m_1, i}$ имеем $|u^1 - v_i^1| \leq h_i = R_i/\sqrt{2}, |u^2 - v_i^2| \leq R_i/\sqrt{2}$, так что $|u - u_i| = (|u^1 - v_i^1|^2 + |u^2 - v_i^2|^2)^{1/2} \leq R_i$. Это значит, что $P_{m_1, i}$ принадлежит шару $S(u_i, R_i), i = \overline{1, m_1}$. Отсюда следует, что пря-

моугольник Π_{m_1} покрывается системой шаров $S(u_i, R_i)$, $i = 1, m_1$. Возьмем произвольную точку $u \in \Pi_{m_1} \cap U$. Тогда найдется шар $S(u_i, R_i)$, содержащий эту точку, т.е. $|u - u_i| \leq R_i$. Отсюда, учитывая условие (3) и обозначения (6), получим

$$J(u) \geq J(u_i) - L|u - u_i| \geq J(u_i) - LR_i = F_i - \varepsilon \geq F_{m_1} - \varepsilon$$

или

$$J(u) \geq F_{m_1} - \varepsilon \quad \text{при всех } u \in \Pi_{m_1} \cap U. \quad (8)$$

Если $U \subset \Pi_{m_1}$, то отсюда получим $0 \leq F_{m_1} - J_* \leq \varepsilon$ — задача (1) — (4) решена. Допустим, что $U \not\subset \Pi_{m_1}$. Тогда последовательно введем точки

$$u_{m_1+1}, u_{m_1+2}, \dots, u_{m_2}, u_{m_1+i} = (v_{m_1+i}^1, v_2^2), \quad i = \overline{1, m_2 - m_1},$$

где

$$\begin{aligned} v_{m_1+1}^1 &= a_1 + h, \quad v_{m_1+i+1}^1 = v_{m_1+i}^1 + h + h_{m_1+i}, \quad i = \overline{1, m_2 - m_1 - 2}; \\ v_{m_2}^1 &= \min \{b_1; v_{m_2-1}^1 + h + h_{m_2-1}\}, \\ v_2^2 &= v_1^2 + h + d_{m_1}, \end{aligned} \quad (9)$$

а номер m_2 определяется условиями

$$v_{m_2-1}^1 < b - h \leq v_{m_2-1}^1 + h + h_{m_2-1}.$$

Положим $d_{m_2} = \min \{h_{m_1+1}, \dots, h_{m_2}\}$ и введем прямоугольник

$$\Pi_{m_2} = \{u = (u^1, u^2): a_1 \leq u^1 \leq b_1, \quad v_2^2 - h \leq u^2 \leq v_2^2 + d_{m_2}\},$$

принадлежащий объединению прямоугольников

$$\begin{aligned} \Pi_{m_2, i} = \{u = (u^1, u^2): v_{m_1+i}^1 - h \leq u^1 \leq v_{m_1+i}^1 + h_{m_1+i}, \\ v_2^2 - h \leq u^2 \leq v_2^2 + d_{m_2}\}, \quad i = \overline{1, m_2 - m_1}. \end{aligned}$$

Как и выше, показываем, что $\Pi_{m_2, i} \subset S(u_{m_1+i}, R_{m_1+i})$, $i = \overline{1, m_2 - m_1}$, и, кроме того,

$$J(u) \geq F_{m_2} - \varepsilon \quad \text{при всех } u \in \Pi_{m_2} \cap U.$$

Поскольку $F_{m_1} \geq F_{m_2}$, то объединяя последнюю оценку для $J(u)$ с оценкой (8), имеем

$$J(u) \geq F_{m_2} - \varepsilon \quad \text{при всех } u \in Q_{m_2} \cap U, \quad (10)$$

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 $0 \leq F_{m_2} - J_* \leq \varepsilon$ — задача (1) — (4) решена. Если же $U \not\subset Q_{m_2}$, то процесс продолжаем дальше.

Пусть точки $u_1, \dots, u_{m_s}$ и величины $F_1, \dots, F_{m_s}, d_{m_1}, \dots, d_{m_s}$ уже найдены, прямоугольник $Q_{m_s} = Q_{m_{s-1}} \cup \Pi_{m_s}$ рассмотрен и показано, что

$$J(u) \geq F_{m_s} - \varepsilon, \quad u \in Q_{m_s} \cap U. \quad (11)$$

Если $U \not\subset Q_{m_s}$, то рассматриваем точки

$$u_{m_s+1}, \dots, u_{m_{s+1}}, \quad u_{m_s+i} = (v_{m_s+i}^1, v_{s+1}^2), \quad i = \overline{1, m_{s+1} - m_s},$$

где координаты $v_{m_s+i}^1$ вычисляются по формулам, аналогичным (7), (9), $v_{s+1}^2 = v_s^2 + h + d_{m_s}$. Затем полагаем $d_{m_{s+1}} = \min \{h_{m_s+1}, \dots, h_{m_{s+1}}\}$, вводим прямоугольник

$$\Pi_{m_{s+1}} = \{u = (u^1, u^2): a_1 \leq u^1 \leq b_1, v_{s+1}^2 - h \leq u^2 \leq v_{s+1}^2 + d_{m_{s+1}}\}.$$

Аналогично тому, как это делалось выше, устанавливаем, что этот прямоугольник покрывается шарами $S(u_{m_s+i}, R_{m_s+i})$, $i = \overline{1, m_{s+1} - m_s}$, и отсюда получаем неравенство

$$J(u) \geq F_{m_{s+1}} - \varepsilon, \quad u \in Q_{m_{s+1}} \cap U; \quad Q_{m_{s+1}} = Q_{m_s} \cup \Pi_{m_{s+1}},$$

и т. д. Нетрудно видеть, что этот процесс закончится за конечное число шагов при таком s , для которого $v_{s-1}^2 < b_2 - h \leq v_{s-1}^2 + h + d_{m_s}$, $v_s^2 = \min \{b_2; v_{s-1}^2 + h + d_{m_s}\}$, $U \subset Q_{m_s}$, неравенство (11) выполняется при всех $u \in U$. Из (11) следует $0 \leq F_{m_s} - J_* = \min \{J(u_1), \dots, J(u_{m_s})\} - J_* \leq \varepsilon$ 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 $n \geq 2$. Аналогично (6) введем обозначения

$$F_i = \min \{F_0; J(u_1), \dots, J(u_i)\}, \quad R = \varepsilon/L, \quad h = R/\sqrt{n} = \varepsilon/(L\sqrt{n}), \quad (12)$$

$$R_i = R + (J(u_i) - F_i)/L, \quad h_i = R_i/\sqrt{n}, \quad i = 1, 2, \dots,$$

где точки $u_1, u_2, \dots, u_i, \dots$ выбираются последовательно по следующим правилам. Сначала определяем точки

$$u_1, \dots, u_{m_1}, \quad u_i = (v_i^1, v_i^2, \dots, v_i^n), \quad i = \overline{1, m_1},$$

где номер m_1 и координаты v_i^j , $i = \overline{1, m_1}$, вычисляются по формулам (7), причем величины h, h_i берутся из (12), а $v_i^1 = a_1 + h$, $j = 2, n$. Полагаем $d_{m_1}^1 = \min \{h_1, \dots, h_{m_1}\}$ и, как и

выше, доказываем, что

$$J(u) \geq F_{m_1} - \varepsilon, \quad u \in Q_{m_1} \cap U,$$

где $Q_{m_1} = \Pi_{m_1} = \{u = (u^1, \dots, u^n): a_1 \leq u^1 \leq b_1, a_j \leq u^j \leq a_j + h + d_{m_1}^1, j = 2, n\}$. Если параллелепипед Q_{m_1} не покрывает параллелепипед (2), то далее рассматриваем точки

$$u_{m_1+1}, \dots, u_{m_2}, \quad u_{m_1+i} = (v_{m_1+i}^1, \dots, v_{m_1+i}^n), \quad i = \overline{1, m_2 - m_1},$$

где номер m_2 и координаты $v_{m_1+i}^1, i = \overline{1, m_2 - m_1}$, определяются формулами (9), $v_{m_1+i}^2 = v_{m_1}^2 + h + d_{m_1}^1, v_{m_1+i}^j = a_j + h, j = \overline{3, n}; i = \overline{1, m_2 - m_1}$. Далее, полагаем $d_{m_2}^1 = \min \{h_{m_1+1}, \dots, h_{m_2}\}, d_{m_2}^2 = \min \{h_1, \dots, h_{m_2}\}$, доказываем, что

$$J(u) \geq F_{m_2} - \varepsilon, \quad u \in Q_{m_2} \cap U,$$

где $Q_{m_2} = Q_{m_1} \cup \Pi_{m_2}, \Pi_{m_2} = \{u: a_1 \leq u^1 \leq b_1, v_{m_2}^2 - h \leq u^2 \leq v_{m_2}^2 + d_{m_2}^1, a_j \leq u^j \leq a_j + h + d_{m_2}^1, j = \overline{3, n}\}$ и т. д. Продолжая этот процесс дальше, найдем точки

$$u_{m_{s-1}}, \dots, u_{m_s}, \quad u_{m_{s-1}+i} = (v_{m_{s-1}+i}^1, \dots, v_{m_{s-1}+i}^n),$$

$$i = \overline{1, m_s - m_{s-1}},$$

где номер m_s и координаты $v_{m_{s-1}+i}^1$ определяются по формулам, аналогичным (7), (9), $v_{m_{s-1}+i}^2 = \min \{b_2; v_{m_{s-1}}^2 + h + d_{m_{s-1}}^1\}, v_{m_{s-1}}^2 < b_2 - h \leq v_{m_{s-1}}^2 + h + d_{m_{s-1}}^1; v_{m_{s-1}+i}^j = a_j + h, j = \overline{3, n};$ устанавливаем неравенство

$$J(u) \geq F_{m_s} - \varepsilon, \quad u \in Q_{m_s} \cap U,$$

где $Q_{m_s} = \{u: a_1 \leq u^1 \leq b_1, a_2 \leq u^2 \leq b_2, a_j \leq u^j \leq a_j + h + d_{m_s}^2\}, d_{m_s}^1 = \min \{h_{m_{s-1}+1}, \dots, h_{m_s}\}, d_{m_s}^2 = \min \{h_1, \dots, h_{m_s}\}$. После этого начинаем менять координату u^3 : сначала берем $v_{m_s+i}^3 = v_{m_s}^3 + h + d_{m_s}^2$ и при $v_{m_s+i}^j = a_j + h, j = \overline{4, n}$, осуществляем перебор координат u^1, u^2 по описанной выше схеме до тех пор, пока при некотором m_k не получим оценки $J(u) \geq F_{m_k} - \varepsilon, u \in Q_{m_k} \cap U, Q_{m_k} = \{u: a_1 \leq u^1 \leq b_1, a_2 \leq u^2 \leq b_2, v_{m_k}^3 - h \leq u^3 \leq v_{m_k}^3 + d_{m_k}^2, a_j \leq u^j \leq a_j + h + d_{m_k}^3, j = \overline{4, n}\}$, где $d_{m_k}^3 = \min \{h_1, \dots, h_{m_k}\}$, а $d_{m_k}^2$ — минимальное среди чисел $\{h_i\}$, полученных на последнем цикле перебора координат u^1, u^2 , со-

ответствующем значению $v_{m_s+i}^3 = v_{m_s}^3 + h + d_{m_s}^2$; затем берем $v_{m_k+i}^3 = v_{m_k}^3 + h + d_{m_k}^2$, заново перебираем координаты u^1 , u^2 и т. д. Такое изменение координаты u^3 закончится на некоторой итерации m_p установлением оценки $J(u) \geq F_{m_p} - \varepsilon$, $u \in Q_{m_p} \cap U$, где $Q_{m_p} = \{u: a_j \leq u^j \leq b_j, j = 1, 2, 3; a_j \leq u^j \leq a_j + d_{m_p}^3, j = 4, \dots, n\}$, $d_{m_p}^3 = \min \{h_1, \dots, h_{m_p}\}$.

Далее, по такой же схеме изменяем координату u^4 по закону $u_{i+1}^4 = u_i^4 + h + d_{m_q}^3$, 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 $J(u) \geq F_{m_r} - \varepsilon$, $u \in U$. Из этой оценки будет следовать, что $0 \leq F_{m_r} - J_* \leq \varepsilon$ 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. § 1.10), 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 [104], 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 предварительное 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.

Метод последовательного перебора для других классов функций описан в [104]; об оптимальных последовательных методах на различных классах функций см. в [76, 118, 123, 124, 216].

§ 11. Метод множителей Лагранжа

1. Рассмотрим задачу

$$J(u) \rightarrow \inf; \quad u \in U = \{u \in E^n: u \in U_0, \\ g_i(u) \leq 0, \quad i = \overline{1, m}, \quad g_i(u) = 0, \quad i = \overline{m+1, s}\}, \quad (1)$$

где $J(u)$, $g_1(u)$, ..., $g_s(u)$ — заданные функции на множестве U_0 . Пусть $J_* > -\infty$, $U_* \neq \emptyset$. Выше (см. теоремы 4.8.2, 4.8.4, 4.8.5) при определенных условиях выпуклости и регулярности задачи (1) было установлено, что для любой точки $u_* \in U_*$ найдутся множители Лагранжа $\lambda^* = (\lambda_1^*, \dots, \lambda_s^*)$:

$$\lambda^* \in \Lambda_0 = \{\lambda \in E^s: \lambda_i \geq 0, \dots, \lambda_m \geq 0\},$$

такие, что точка (u_*, λ^*) образует седловую точку функции Лагранжа

$$L(u, \lambda) = J(u) + \sum_{i=1}^s \lambda_i g_i(u), \quad u \in U_0, \quad \lambda \in \Lambda_0, \quad (2)$$

т. е.

$$L(u_*, \lambda) \leq L(u_*, \lambda^*) = J_* \leq L(u, \lambda^*), \quad u \in U_0, \quad \lambda \in \Lambda_0. \quad (3)$$

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 4.8.1).

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):

$$u_{k+1} = P_{U_0}(u_k - \alpha_k L_u(u_k, \lambda_k)), \quad (4)$$

$$\lambda_{k+1} = P_{\Lambda_0}(\lambda_k + \alpha_k L_\lambda(u_k, \lambda_k)) = P_{\Lambda_0}(\lambda_k + \alpha_k g(u_k)), \quad (5)$$

$$k = 0, 1, \dots, \quad \text{где} \quad L_u(u, \lambda) = (L_{u^1}(u, \lambda), \dots, L_{u^n}(u, \lambda)),$$

$$L_\lambda(u, \lambda) = (L_{\lambda_1}(u, \lambda), \dots, L_{\lambda_s}(u, \lambda)) = (g_1(u), \dots, g_s(u)) = g(u);$$

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 $P_{\Lambda_0}(\lambda) = (\mu_1, \mu_2, \dots, \mu_s)$, где

$$\mu_i = \lambda_i^+ = \max\{\lambda_i, 0\}, \quad i = \overline{1, m}; \quad \mu_i = \lambda_i, \quad i = \overline{m+1, s}.$$

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:

$$L(u_{k+1}, \lambda) = \inf_{u \in U_0} L(u, \lambda_k), \quad \lambda_{k+1} = P_{\Lambda_0}(\lambda_k + \alpha_k g(u_{k+1})), \\ k = 0, 1, \dots$$

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 $J(u) \equiv 0$, $U = \{u \in E^1: g(u) \equiv u = 0\}$. 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

$$u_{k+1} = u_k - \alpha_k \lambda_k, \quad \lambda_{k+1} = \lambda_k + \alpha_k u_k, \quad k = 0, 1, \dots$$

Поскольку $u_{k+1}^2 + \lambda_{k+1}^2 = (u_k^2 + \lambda_k^2)(1 + \alpha_k^2) \geq u_k^2 + \lambda_k^2$, $k = 0, 1, \dots$, то ясно, что при любых $(u_0, \lambda_0) \neq 0$ и любом выборе длины шага $\alpha_k \geq 0$ 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 (4.8.38) имеют место строгие включения (см. пример 4.8.6). 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 [21], мы рассмотрим один из возможных здесь подходов.

2. Будем рассматривать задачу

$$J(u) \rightarrow \inf, \quad u \in U = \{u \in E^n: u \in U_0, \quad g(u) \leq 0\}, \quad (6)$$

где $J(u)$, $g(u) = (g_1(u), \dots, g_m(u))$ — заданные функции из $C^1(U_0)$. Как и в гл. 3, векторное неравенство $g = (g_1, \dots, g_m) \geq$

≥ 0 [$g \leq 0$] здесь и ниже означает, что $g_i \geq 0$ [$g_i \leq 0$] при всех $i = \overline{1, m}$, а неравенство $a \geq b$ для $a, b \in E^m$ эквивалентно неравенству $a - b \geq 0$.

Наряду с классической функцией Лагранжа задачи (6)

$$L(u, \lambda) = J(u) + \langle g(u), \lambda \rangle, \quad u \in U_0, \quad \lambda \in \Lambda_0 = \{\lambda \in E^m: \lambda \geq 0\} \quad (7)$$

еще рассмотрим следующую модифицированную функцию Лагранжа:

$$M(u, \lambda) = J(u) + \frac{1}{2A} [(\lambda + Ag(u))^+]^2 - \frac{1}{2A} |\lambda|^2 \quad (8)$$

переменных $u \in U_0, \lambda \in \Lambda_0$, где A — произвольная фиксированная положительная константа; в (8) принято обозначение

$$a^+ = P_{E_+^m}(a) = (a_1^+, \dots, a_m^+), \quad a_i^+ = \max\{a_i; 0\}, \quad i = \overline{1, m}, \quad (9)$$

— проекция точки $a \in E^m$ на положительный октант $E_+^m = \{a \in E^m: a \geq 0\}$.

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 дифференцируема по u и λ , причем

$$\begin{aligned} \frac{\partial M}{\partial u} &= M_u(u, \lambda) = J'(u) + (g'(u))^T (\lambda + Ag(u))^+, \\ \frac{\partial M}{\partial \lambda} &= M_\lambda(u, \lambda) = \frac{1}{A} [(\lambda + Ag(u))^+ - \lambda], \quad u \in U_0, \quad \lambda \in E^m, \end{aligned} \quad (10)$$

где $g'(u)$ — матрица порядка $m \times n$, у которой в i -й строке, j -м столбце находится производная $g_{iu,j}(u) = \frac{\partial g_i(u)}{\partial u^j}$, $i = \overline{1, m}$, $j = \overline{1, n}$, 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 $\langle M_\lambda(u, \lambda) - M_\lambda(u, \mu), \lambda - \mu \rangle \leq 0$ для всех $\lambda, \mu \in \Lambda_0$ 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

$$\Phi_k(u) = \frac{1}{2} |u - u_k|^2 + \alpha M(u, \lambda_k), \quad u \in U_0, \quad (11)$$

где α — некоторое положительное число, являющееся параметром метода. Предположим, что существует точка v_{k+1} , удовлетворяющая условиям

$$v_{k+1} \in U_0, \quad \Phi_k(v_{k+1}) = \inf_{U_0} \Phi_k(u). \quad (12)$$

В качестве следующего $(k+1)$ -го приближения возьмем точку u_{k+1} такую, что

$$\begin{aligned} u_{k+1} \in U_0, \quad |\Phi'_k(u_{k+1}) - \Phi'_k(v_{k+1})| &\leq \delta_k, \\ |g(u_{k+1}) - g(v_{k+1})| &\leq \delta_k, \end{aligned} \quad (13)$$

где $\delta_k \geq 0$, $\lim_{k \rightarrow \infty} \delta_k = 0$. В частности, если точка v_{k+1} из (12)

известна, то можно принять $u_{k+1} = v_{k+1}$; 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

$$\lambda_{k+1} = (\lambda_k + Ag(u_{k+1}))^+. \quad (14)$$

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

$$|a^+ - b^+| \leq |a - b|, \quad a, b \in E^m. \quad (15)$$

Далее, система соотношений

$$g \leq 0, \quad \lambda \geq 0, \quad \lambda_i g_i = 0, \quad i = \overline{1, m}, \quad (16)$$

эквивалентна равенству

$$\lambda = (\lambda + Ag)^+ \quad (17)$$

при любых постоянных $A > 0$. В самом деле, если выполняются соотношения (16), то либо $g_i = 0$, $\lambda_i \geq 0$, либо $\lambda_i = 0$, $g_i \leq 0$. 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

$$\lambda_i = (\lambda_i + Ag_i)^+ = \max \{\lambda_i + Ag_i; 0\}, \quad i = \overline{1, m}, \quad (17')$$

Отсюда ясно, что $\lambda_i \geq 0$ при всех $i = \overline{1, m}$, т. е. $\lambda \geq 0$. Если $\lambda_i = 0$, то $\lambda_i g_i = 0$ и, кроме того, из (17') получим $0 = (0 + Ag_i)^+ = Ag_i^+$, т. е. $g_i \leq 0$. 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

$$\langle a^+, a \rangle = \langle a^+, a^+ \rangle, \quad \langle a^+, b \rangle \leq \langle a^+, b^+ \rangle, \quad a, b \in E^m.$$

Отсюда имеем

$$\begin{aligned} \langle a^+ - b^+, a - b \rangle &= \langle a^+, a \rangle + \langle b^+, b \rangle - \langle a^+, b \rangle - \langle b^+, a \rangle \geq \\ &\geq \langle a^+, a^+ \rangle + \langle b^+, b^+ \rangle - \langle a^+, b^+ \rangle - \langle b^+, a^+ \rangle = \langle a^+ - b^+, a^+ - b^+ \rangle, \end{aligned}$$

т. е.

$$\langle a^+ - b^+, a - b \rangle \geq \langle a^+ - b^+, a^+ - b^+ \rangle. \quad (18)$$

Теорема 1. Пусть U_0 — выпуклое замкнутое множество из E^n (например, $U_0 = E^n$), функции $I(u)$, $g_1(u), \dots, g_m(u)$ выпуклы на U_0 и принадлежат классу $C^1(U_0)$, $J_* > -\infty$, $U_* \neq \emptyset$, 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 $\Phi_k(u)$, определяемая формулой (11), сильно выпукла на U_0 с константой сильной выпуклости $\kappa = 1/2$. Отсюда и из теоремы 4.3.1 следует, что точка v_{k+1} , удовлетворяющая условиям (12), существует и определяется однозначно. Тогда существует и точка u_{k+1} , удовлетворяющая условиям (13): например, в (13) можно взять $u_{k+1} = v_{k+1}$. 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

$$\mu_{k+1} = (\lambda_k + Ag(v_{k+1}))^+, \quad k = 0, 1, \dots \quad (19)$$

Покажем, что для любой седловой точки (u_*, λ^*) функции Лагранжа (7) справедливо неравенство

$$\begin{aligned} |u_k - u_*|^2 + \frac{\alpha}{A} |\lambda_k - \lambda^*|^2 &\geq \\ &\geq |v_{k+1} - u_*|^2 + \frac{\alpha}{A} |\mu_{k+1} - \lambda^*|^2 + |v_{k+1} - u_k|^2 + \frac{\alpha}{A} |\mu_{k+1} - \lambda_k|^2, \quad (20) \\ k &= 0, 1, \dots \end{aligned}$$

Согласно лемме 4.8.1 существование седловой точки (u_*, λ^*) в задаче (6) эквивалентно соотношениям

$$L(u_*, \lambda^*) \leq L(u, \lambda^*), \quad u \in U_0, \quad (21)$$

$$g(u_*) \leq 0, \quad \lambda^* \geq 0, \quad \lambda_i^* g_i(u_*) = 0, \quad i = \overline{1, m}. \quad (22)$$

В силу эквивалентности соотношений (16), (17) условия (22) можно переписать в следующей равносильной форме:

$$\lambda^* = (\lambda^* + Ag(u_*))^+. \quad (23)$$

Кроме того, функция $L(u, \lambda^*)$ выпукла на U_0 , принадлежит $C^1(U_0)$, а тогда согласно теореме 4.2.3 неравенство (21) эквивалентно условию

$$\langle L_u(u_*, \lambda^*), u - u_* \rangle = \langle J'(u_*) + (g'(u_*))^T \lambda^*, u - u_* \rangle \geq 0$$

при всех $u \in U_0$. Отсюда с учетом равенства (23) имеем

$$\langle J'(u_*) + (g'(u_*))^T (\lambda^* + Ag(u_*))^+, u - u_* \rangle \geq 0, \quad u \in U_0. \quad (24)$$

Далее, из условия (12) и теоремы 4.2.3 следует

$$\langle \Phi'_k(v_{k+1}), u - v_{k+1} \rangle \geq 0, \quad u \in U_0.$$

Отсюда с учетом формулы (10) получим

$$\langle v_{k+1} - u_k + \alpha J'(v_{k+1}) + \alpha (g'(v_{k+1}))^T (\lambda_k + Ag(v_{k+1}))^+, u - v_{k+1} \rangle \geq 0, \\ u \in U_0. \quad (25)$$

Примем в (24) $u = v_{k+1}$, умножим это неравенство на $\alpha > 0$ и сложим с неравенством (25) при $u = u_*$. Получим

$$\langle v_{k+1} - u_k + \alpha (J'(v_{k+1}) - J'(u_*)) + \alpha (g'(v_{k+1}))^T (\lambda_k + Ag(v_{k+1}))^+ - \\ - \alpha (g'(u_*))^T (\lambda^* + Ag(u_*))^+, u_* - v_{k+1} \rangle \geq 0, \quad k = 0, 1, \dots$$

Отсюда имеем

$$\langle v_{k+1} - u_k, u_* - v_{k+1} \rangle \geq \alpha \langle J'(v_{k+1}) - J'(u_*), v_{k+1} - u_* \rangle + \\ + \alpha \langle (\lambda_k + Ag(v_{k+1}))^+, g'(v_{k+1})(v_{k+1} - u_*) \rangle - \\ - \alpha \langle (\lambda^* + Ag(u_*))^+, g'(u_*)(v_{k+1} - u_*) \rangle, \quad k = 0, 1, \dots \quad (26)$$

Так как функции $J(u)$, $g_i(u)$ выпуклы, то согласно теореме 4.2.4

$$\langle J'(v_{k+1}) - J'(u_*), v_{k+1} - u_* \rangle \geq 0,$$

$$g'(v_{k+1})(v_{k+1} - u_*) \geq g(v_{k+1}) - g(u_*) \geq g'(u_*)(v_{k+1} - u_*).$$

Отсюда и из (26) следует

$$\langle v_{k+1} - u_k, u_* - v_{k+1} \rangle \geq \alpha \langle (\lambda_k + Ag(v_{k+1}))^+ - \\ - (\lambda^* + Ag(u_*))^+, g(v_{k+1}) - g(u_*) \rangle = \\ = \frac{\alpha}{A} \langle (\lambda_k + Ag(v_{k+1}))^+ - (\lambda^* + Ag(u_*))^+, \\ [(\lambda_k + Ag(v_{k+1})) - \lambda_k] - [(\lambda^* + Ag(u_*)) - \lambda^*] \rangle.$$

К правой части этой оценки применим неравенство (18). С учетом формулы (19), определяющей точку μ_{k+1} , и равенства (23) получим

$$\begin{aligned} \langle v_{k+1} - u_k, u_* - v_{k+1} \rangle &\geq \frac{\alpha}{A} \langle (\lambda_k + Ag(v_{k+1}))^+ - (\lambda^* + Ag(u_*))^+ , \\ &[(\lambda_k + Ag(v_{k+1}))^+ - \lambda_k] - [(\lambda^* + Ag(u_*))^+ - \lambda^*] \rangle = \\ &= \frac{\alpha}{A} \langle \mu_{k+1} - \lambda^*, \mu_{k+1} - \lambda_k \rangle. \end{aligned}$$

т. е.

$$\langle v_{k+1} - u_k, u_* - v_{k+1} \rangle + \frac{\alpha}{A} \langle \mu_{k+1} - \lambda_k, \lambda^* - \mu_{k+1} \rangle \geq 0, \quad k = 0, 1, \dots \quad (27)$$

Справедливы тождества

$$\begin{aligned} \|u_k - u_*\|^2 &= \|u_k - v_{k+1}\|^2 + \|v_{k+1} - u_*\|^2 = \\ &= \|v_{k+1} - u_k\|^2 + \|u_* - v_{k+1}\|^2 + 2\langle v_{k+1} - u_k, u_* - v_{k+1} \rangle, \\ \|\lambda_k - \lambda^*\|^2 &= \|\mu_{k+1} - \lambda_k\|^2 + \|\lambda^* - \mu_{k+1}\|^2 + 2\langle \mu_{k+1} - \lambda_k, \lambda^* - \mu_{k+1} \rangle. \end{aligned}$$

Умножим второе из этих тождеств на α/A и сложим с первым. Отсюда с учетом оценки (27) получим обещанное неравенство (20).

Далее, покажем, что

$$\|u_{k+1} - v_{k+1}\| \leq \delta_k, \quad \|\lambda_{k+1} - \mu_{k+1}\| \leq A\delta_k, \quad k = 0, 1, \dots \quad (28)$$

Поскольку функция $\Phi_k(u)$ сильно выпукла на U_0 , то с помощью теоремы 4.3.3 и первого неравенства (13) получим

$$\|u_{k+1} - v_{k+1}\|^2 \leq \langle \Phi'_k(u_{k+1}) - \Phi'_k(v_{k+1}), u_{k+1} - v_{k+1} \rangle \leq \delta_k \|u_{k+1} - v_{k+1}\|.$$

Отсюда после деления на $\|u_{k+1} - v_{k+1}\|$ придем к первой оценке (28). Из формул (14), (19), определяющих точки λ_{k+1} , μ_{k+1} , неравенства (15) и условий (13) следует

$$\|\lambda_{k+1} - \mu_{k+1}\| \leq A \|g(u_{k+1}) - g(v_{k+1})\| \leq A\delta_k.$$

Оценки (28) доказаны.

В $(n+m)$ -мерном линейном пространстве $\mathbb{R}^{n+m}$ переменных $z = (u, \lambda) = (u^1, \dots, u^n, \lambda^1, \dots, \lambda^m)$ 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

$$\|z\| = (\|u\|^2 + (\alpha/A) \|\lambda\|^2)^{1/2}. \quad (29)$$

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

$$\|z_k - z^*\|^2 \geq \|w_{k+1} - z^*\|^2 + \|w_{k+1} - z_k\|^2, \quad k = 0, 1, \dots, \quad (30)$$

$$\|z_{k+1} - w_{k+1}\| \leq (A\alpha + 1) \delta_k, \quad k = 0, 1, \dots \quad (31)$$

Напоминаем, что по условию $\sum_{k=0}^{\infty} \delta_k < \infty$. 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 $\mathbb{R}^{n+m}$, 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 $\lim_{k \rightarrow \infty} \|z_k - z^*\|$ и, кроме того,

$$\lim_{k \rightarrow \infty} \|w_{k+1} - z_k\| = 0. \quad (32)$$

Заметим, что

$$\min \{1; \alpha/A\} \|z\|^2 \leq \|z\|^2 \leq \max \{1; \alpha/A\} \|z\|^2.$$

т. е. нормы $\|z\|$ и $\|z\|$ эквивалентны. Отсюда и из существования конечного предела $\lim_{k \rightarrow \infty} \|z_k - z^*\|$ 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 $\{z_{k_r}\} \rightarrow c^*$ и (31), (32) следует, что $\{w_{k_r+1}\} \rightarrow c^*$, $\{z_{k_r+1}\} \rightarrow c^*$. Тогда из (14) при $k = k_r \rightarrow \infty$ получим

$$b^* = (b^* + Ag(a_*))^+. \quad (33)$$

В силу эквивалентности соотношений (16) и (17) из (33) следует

$$g(a_*) \leq 0, \quad b^* \geq 0, \quad b_i^* g_i(a_*) = 0, \quad i = \overline{1, m}. \quad (34)$$

Далее, переходя в (25) к пределу при $k = k_r \rightarrow \infty$, будем иметь

$$\langle J'(a_*) + (g'(a_*))^T (b^* + Ag(a_*))^+, u - a_* \rangle \geq 0, \quad u \in U_0,$$

или с учетом (33)

$$\langle J'(a_*) + (g'(a_*))^T b^*, u - a_* \rangle = \langle L_u(a_*, b^*), u - a_* \rangle \geq 0$$

при всех $u \in U_0$. В силу выпуклости $L(u, b^*)$ последнее неравенство эквивалентно неравенству

$$L(a_*, b^*) \leq L(u, b^*), \quad u \in U_0. \quad (35)$$

Из соотношений (34), (35) и леммы 4.8.1 следует, что $c^* = (a_*, b^*)$ — седловая точка функции $L(u, \lambda)$ в смысле неравенств (3), а тогда согласно теореме 4.8.1 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 $\lim_{k \rightarrow \infty} \|z_k - c^*\|$, причем в силу определения точки c^* имеем $\lim_{k \rightarrow \infty} \|z_k - c^*\| = \lim_{r \rightarrow \infty} \|z_{k_r} - c^*\| = 0$. 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 4.8.6), а также библиографию по таким методам читатель найдет в [21].

§ 12. Метод штрафных функций

1. Метод штрафных функций является одним из наиболее простых и широко применяемых методов решения задач минимизации. Основная идея метода заключается в сведении исходной задачи

$$J(u) \rightarrow \inf; \quad u \in U \quad (1)$$

к последовательности задач минимизации

$$\Phi_k(u) \rightarrow \inf; \quad u \in U_0, \quad k = 1, 2, \dots, \quad (2)$$

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 $U_0 \setminus U$. 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 к 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 $\{P_k(u)$, $k = 1, 2, \dots\}$, определенных и неотрицательных на множестве U_0 , содержащем множество U , называют *штрафом или штрафной функцией* множества U на множестве U_0 , если

$$\lim_{k \rightarrow \infty} P_k(u) = \begin{cases} 0 & \text{при } u \in U; \\ +\infty & \text{при } u \in U_0 \setminus U. \end{cases}$$

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 $k \rightarrow \infty$.

Для любого множества $U \subset E^n$ можно указать сколько угодно различных штрафных функций. Например, если $\{A_k\}$ — какая-либо положительная последовательность, $\lim_{k \rightarrow \infty} A_k = +\infty$,

то можно взять

$$P_k(u) = A_k \rho(u, U), \quad u \in E^n = U_0, \quad k = 1, 2, \dots$$

(здесь U предполагается замкнутым) или

$$P_k(u) = \begin{cases} 0 & \text{при } u \in U, \\ A_k |u - \bar{u}| & \text{при } u \notin U, \quad k = 1, 2, \dots; \end{cases}$$

здесь $\rho(u, U) = \inf_{v \in U} |u - v|$ — расстояние от точки u 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

$$\Phi_k(u) = J(u) + P_k(u), \quad u \in U_0, \quad k = 1, 2, \dots, \quad (3)$$

и рассмотрим последовательность задач (2) с функциями (3). Будем считать, что

$$\Phi_{k*} = \inf_{U_0} \Phi_k(u) > -\infty, \quad k = 1, 2, \dots \quad (4)$$

Если здесь при каждом $k = 1, 2, \dots$ нижняя грань достигается, то условия

$$\Phi_k(u_k) = \Phi_{k*}, \quad u_k \in U_0, \quad (5)$$

определяют последовательность $\{u_k\}$. Однако точно определить u_k из (5) удастся лишь в редких случаях. Кроме того, нижняя грань в (4) при некоторых или даже всех $k = 1, 2, \dots$ может и не достигаться. Поэтому будем считать, что при каждом $k = 1, 2, \dots$ с помощью какого-либо метода минимизации найдена точка u_k , определяемая условиями

$$u_k \in U_0, \quad \Phi_k(u_k) \leq \Phi_{k*} + \varepsilon_k, \quad (6)$$

где $\{\varepsilon_k\}$ — некоторая заданная последовательность, $\varepsilon_k > 0$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} \varepsilon_k = 0$ (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

$$\lim_{k \rightarrow \infty} P_k(u) = 0 \quad \text{при} \quad u \in U_0 \setminus U,$$

то можно ожидать, что для широкого класса задач (1) последовательность $\{u_k\}$, определяемая условиями (6), будет приближаться к множеству U и будут справедливы равенства

$$\lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0.$$

Мы здесь ограничимся рассмотрением задачи (1) для случая, когда множество U имеет вид

$$U = \{u \in E^n: u \in U_0, \quad g_i(u) \leq 0, \quad i = \overline{1, m}; \quad g_i(u) = 0, \quad i = \overline{m+1, s}\}, \quad (7)$$

где U_0 — заданное множество из E^n (например, $U_0 = E^n$), функции $J(u)$, $g_i(u)$, $i = \overline{1, s}$, определены на U_0 . В качестве

штрафной функции множества (7) возьмем

$$P_k(u) = A_k P(u);$$

$$P(u) = \sum_{i=1}^m (\max \{g_i(u); 0\})^p + \sum_{i=m+1}^s |g_i(u)|^p, \quad u \in U_0, \quad (8)$$

где $A_k > 0$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} A_k = +\infty$, а $p \geq 1$ — 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 $g_i(u)$, $i = \overline{1, s}$, 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 $i = \overline{1, m}$ выпуклы на U_0 , $g_i(u) = \langle a_i, u \rangle - b^i$ — линейные функции при $i = m+1, s$, то функция (8) выпукла на U_0 — это вытекает из следствий к теореме 4.2.8.

Если для краткости ввести обозначения

$$g_i^+(u) = \begin{cases} \max \{g_i(u); 0\} & \text{при } i = \overline{1, m}; \\ |g_i(u)| & i = \overline{m+1, s}, \end{cases} \quad (9)$$

то функцию (8) можно записать в виде

$$P_k(u) = A_k P(u), \quad P(u) = \sum_{i=1}^s (g_i^+(u))^p, \quad u \in U_0. \quad (8)$$

Функцию $P(u)$ мы иногда также будем называть *штрафной функцией* множества (7), подразумевая при этом, что после умножения на $A_k > 0$, $\lim_{k \rightarrow \infty} A_k = +\infty$, 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

$$P_k(u) = \sum_{i=1}^s A_{ki} (g_i^+(u))^{p_i}, \quad u \in U_0, \quad k = 1, 2, \dots,$$

где $p_i \geq 1$, $A_{ki} > 0$, $\lim_{k \rightarrow \infty} A_{ki} = +\infty$, $i = \overline{1, s}$; здесь каждое ограничение из (7) имеет свой штрафной коэффициент. Весьма широкий класс штрафных функций множества (7) дает следующая конструкция:

$$P_k(u) = \sum_{i=1}^s A_{ki} \varphi_i(g_i^+(u)), \quad u \in U_0, \quad k = 1, 2, \dots,$$

где $\varphi_i(g)$ — произвольная функция, определенная при $g \geq 0$, такая, что $\varphi_i(0) = 0$, $\varphi_i(g) > 0$ при $g > 0$, $i = \overline{1, s}$. 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

$$P_k(u) = \left(1 + \sum_{i=1}^s (g_i^+(u))^{p_i} \right)^{A_k} - 1, \quad p_i \geq 1;$$

$$P_k(u) = A_k^{-1} \left(\sum_{i=1}^m e^{A_k g_i(u)} + \sum_{i=m+1}^s e^{A_k g_i^2(u)} \right), \quad u \in U_0,$$

где $A_k > 0$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} A_k = +\infty$.

Прежде чем переходить к строгим формулировкам теорем сходимости метода штрафных функций, рассмотрим несколько конкретных примеров.

Пример 1. Пусть требуется решить задачу

$$J(u) = x^2 + xy + y^2 \rightarrow \inf;$$

$$u \in U = \{u = (x, y) \in E^2: x + y - 2 = 0\}.$$

В качестве штрафной функции возьмем $P_k(u) = k(x + y - 2)^2$ и положим

$$\Phi_k(u) = x^2 + xy + y^2 + k(x + y - 2)^2, \quad u \in U_0 = E^2; \quad k = 1, 2, \dots$$

Функция $\Phi_k(u)$ при каждом фиксированном $k = 1, 2, \dots$ сильно выпукла на E^2 и достигает своей нижней грани на E^2 в точке $u_k = (x_k, y_k)$, которая определяется уравнениями

$$\frac{\partial \Phi_k(u_k)}{\partial x} = 2x_k + y_k + 2k(x_k + y_k - 2) = 0,$$

$$\frac{\partial \Phi_k(u_k)}{\partial y} = x_k + 2y_k + 2k(x_k + y_k - 2) = 0.$$

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 $k \rightarrow \infty$ будем иметь $u_k \rightarrow u_* = (1, 1)$, $\Phi_k(u_k) \rightarrow 3$. 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 $J(u_*) = J_* = 3 = \lim_{k \rightarrow \infty} \Phi_k(u_k)$. Таким образом, в рассмотренном примере метод штрафных функций сходится.

Пример 2. Пусть

$$J(u) = e^{-u} \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) = ue^{-u} = 0\}.$$

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 $u \in E^1$, $\lim_{u \rightarrow +\infty} \Phi_k(u) = 0$, то $\Phi_{k*} = \inf_{E^1} \Phi_k(u) = 0$. В качестве точки u_k , удовлетворяющей условиям (6) при $\epsilon_k = e^{-k} + k^3e^{-2k}$, здесь можно взять $u_k = k$, $k = 1, 2, \dots$. Получим $\lim_{k \rightarrow \infty} J(u_k) = 0 < J_* = 1$, $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = +\infty$. 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 $J(u)$, $g_i(u)$, $i = \overline{1, s}$, определены на множестве U_0 , а последовательность $\{u_k\}$ определена условиями (3), (6), (8). Тогда

$$\overline{\lim}_{k \rightarrow \infty} J(u_k) \leq \overline{\lim}_{k \rightarrow \infty} \Phi_k(u_k) = \overline{\lim}_{k \rightarrow \infty} \Phi_{k*} \leq J_* \quad (10)$$

Если, кроме того, $J_{**} = \inf_U J(u) > -\infty$, то

$$P(u_k) = \sum_{i=1}^s (g_i^+(u_k))^p = O(A_k^{-1}), \quad k = 1, 2, \dots; \quad (11)$$

$$\overline{\lim}_{k \rightarrow \infty} g_i(u_k) \leq 0, \quad i = \overline{1, m}; \quad \lim_{k \rightarrow \infty} g_i(u_k) = 0, \quad i = \overline{m+1, s}. \quad (12)$$

Доказательство. Так как $P(u) \geq 0$, то из (3), (6), (8) имеем

$$\begin{aligned} J(u_k) &\leq J(u_k) + A_k P(u_k) = \Phi_k(u_k) \leq \Phi_{k*} + \epsilon_k \leq \\ &\leq \Phi_k(u) + \epsilon_k = J(u) + A_k P(u) + \epsilon_k \end{aligned} \quad (13)$$

для всех $u \in U_0$ и $k = 1, 2, \dots$. Учитывая, что $P(u) = 0$ при $u \in U$, из (13) при $k \rightarrow \infty$ получим

$$\overline{\lim}_{k \rightarrow \infty} J(u_k) \leq \overline{\lim}_{k \rightarrow \infty} \Phi_k(u_k) = \overline{\lim}_{k \rightarrow \infty} \Phi_{k*} \leq J(u)$$

при всех $u \in U$. Переходя здесь к нижней грани по $u \in U$, приходим к соотношениям (10).

Пусть теперь $J_{**} > -\infty$. Так как $J_* \geq J_{**}$, то $J_* > -\infty$. С учетом условий (6) имеем

$$0 \leq A_k P(u_k) = \Phi_k(u_k) - J(u_k) \leq \Phi_{k*} + e_k - J_{**}, \quad k = 1, 2, \dots$$

Согласно (10) $\lim_{k \rightarrow \infty} \overline{\Phi_{k*}} \leq J_*$, поэтому $\Phi_{k*} \leq C = \text{const}$, $k = 1, 2, \dots$. Тогда

$$0 \leq P(u_k) \leq (C + \sup_{k \geq 0} e_k - J_{**}) A_k^{-1}, \quad k = 1, 2, \dots$$

Оценка (11) доказана. Из нее следует, что $\lim_{k \rightarrow \infty} P(u_k) = 0$ или

$\lim_{k \rightarrow \infty} g_i^+(u_k) = 0$, $i = \overline{1, s}$. 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 $\lim_{k \rightarrow \infty} J(u_k) = J_*$.

Теорема 2. Пусть U_0 — замкнутое множество из E^n , функции $J(u)$, $g_1(u)$, $\dots$, $g_m(u)$, $|g_{m+1}(u)|$, $\dots$, $|g_s(u)|$ непрерывны снизу на U_0 , $J_{**} = \inf_{U_0} J(u) > -\infty$. 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

$$U_\delta = \{u: u \in U_0, \quad g_i(u) \leq \delta, \quad i = \overline{1, m}; \quad |g_i(u)| \leq \delta, \quad i = \overline{m+1, s}\} \quad (14)$$

ограничено хотя бы при одном значении $\delta > 0$, то

$$\lim_{k \rightarrow \infty} \Phi_k(u_k) = \lim_{k \rightarrow \infty} \Phi_{k*} = \lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0. \quad (15)$$

Доказательство. При сделанных предположениях для последовательности $\{u_k\}$ соотношения (10) — (12) сохраняют силу. Пусть u_* — какая-либо предельная точка последовательности $\{u_k\}$, пусть $\{u_k\} \rightarrow u_*$. Заметим, что $u_* \in U_0$ в силу замкнутости U_0 . Тогда с учетом полунепрерывности снизу указанных в условии теоремы функций из соотношений (12) получим

$$g_i(u_*) \leq \lim_{r \rightarrow \infty} g_i(u_{k_r}) \leq \lim_{k \rightarrow \infty} g_i(u_k) \leq 0, \quad i = \overline{1, m};$$

$$|g_i(u_*)| \leq \lim_{r \rightarrow \infty} |g_i(u_{k_r})| = \lim_{k \rightarrow \infty} |g_i(u_k)| = 0, \quad i = \overline{m+1, s}.$$

Следовательно, $u_* \in U$. Тогда с учетом (10) имеем $J_* \leq J(u_*) \leq \lim_{r \rightarrow \infty} J(u_{k_r}) \leq \overline{\lim}_{k \rightarrow \infty} J(u_k) \leq J_*$, т. е. $\lim_{r \rightarrow \infty} J(u_{k_r}) = J(u_*) = J_*$ или $u_* \in U_*$.

Наконец, пусть множество (14) ограничено при некотором $\delta > 0$. Из соотношений (12) следует, что $\{u_k\} \in U_\delta$ для всех $k \geq k_0$. Это означает, что $\{u_k\}$ имеет хотя бы одну предельную точку. Тогда, как было выше показано, все предельные точки $\{u_k\}$ принадлежат U_* , т. е. $\lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0$. Из тех же рассуждений и неравенств (10) вытекают остальные равенства (15). Теорема 2 доказана.

Для иллюстрации теоремы 2 рассмотрим

Пример 3. Пусть

$$J(u) = e^{-u} \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) = u = 0\}.$$

Здесь $J_* = 0$, $U_* = \{0\}$. Функции $J(u)$, $g(u)$ непрерывны на замкнутом множестве $U_0 = E^1$, $J_* = \inf_{E^1} e^{-u} = 0$, множество $U_\delta = \{u \in E^1: |u| \leq \delta\}$ ограничено при любом $\delta > 0$. Таким образом, все условия теоремы 2 выполнены. Возьмем штрафную функцию $P(u) = (g(u))^2 = u^2$ и положим

$$\Phi_k(u) = e^{-u} + ku^2, \quad u \in E^1, \quad k = 1, 2, \dots$$

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 неотрицательная последовательность, стремящаяся к нулю. Определим точку u_k из условия $\Phi_k(u_k) \leq \Phi_{k*} + \varepsilon_k$, $k = 1, 2, \dots$ 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. Ограничения

$$g_i(u) \leq 0, \quad i = \overline{1, m}; \quad g_i(u) = 0, \quad i = \overline{m+1, s}, \quad u \in U_0, \quad (16)$$

назовем *согласованными с функцией $J(u)$ на множестве $u \in U_0$* , если для любой последовательности $\{u_k\} \in U_0$, для которой

$$\overline{\lim}_{k \rightarrow \infty} g_i(u_k) \leq 0; \quad i = \overline{1, m}; \quad \lim_{k \rightarrow \infty} g_i(u_k) = 0, \quad i = \overline{m+1, s}, \quad (17)$$

имеет место соотношение

$$\overline{\lim}_{k \rightarrow \infty} J(u_k) \geq J_* = \inf_U J(u),$$

где множество U имеет вид (7).

Отметим, что в примере 3 ограничение $g(u) = 0$ согласовано с функцией 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 $\Phi_{k*} = \inf_{U_0} \Phi_k(u)$, $k = 1, 2, \dots$

Тогда для того чтобы

$$\lim_{k \rightarrow \infty} \Phi_{k*} = J_*, \quad (18)$$

необходимо, чтобы ограничения (16) были согласованы с функцией $J(u)$ на множестве U_0 . Если $J_{**} = \inf_{U_0} J(u) > -\infty$, то согласованности ограничений с функцией $J(u)$ 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 (17). Тогда $\lim_{r \rightarrow \infty} P(u_r) = 0$. Справедливы неравенства

$$\Phi_{k*} \leq \Phi_k(u_r) \leq J(u_r) + A_k P(u_r), \quad r = 1, 2, \dots$$

Отсюда при $r \rightarrow \infty$ получим $\Phi_{k*} \leq \lim_{r \rightarrow \infty} J(u_r)$ при всех $k = 1, 2, \dots$

Переходя здесь к пределу при $k \rightarrow \infty$, с учетом (18) будем иметь $\lim_{r \rightarrow \infty} J(u_r) \geq \lim_{k \rightarrow \infty} \Phi_{k*} = J_*$, что и требовалось.

Достаточность. Пусть $J_{**} > -\infty$, и ограничения (16) согласованы с функцией $J(u)$ на множестве U_0 . Поскольку $\Phi_k(u) \geq J(u)$ при всех $u \in U_0$, то $\Phi_{k*} \geq J_{**} > -\infty$, и имеет смысл говорить о последовательностях $\{u_k\}$, удовлетворяющих условиям (6). Возьмем одну из таких последовательностей $\{u_k\}$. Согласно теореме 1 тогда справедливы соотношения (10)–(12). По условию ограничения (16) согласованы с $J(u)$ на U_0 . Поэтому из (12) следует $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Отсюда и из (10) с учетом неравенств $\varepsilon_k + \Phi_{k*} \geq \Phi_k(u_k) \geq J(u_k)$, $k = 1, 2, \dots$, получим $\lim_{k \rightarrow \infty} J(u_k) = \lim_{k \rightarrow \infty} \Phi_{k*} = J_*$. Теорема 3 доказана.

Пример ограничений, согласованных с $J(u)$ на U_0 , указан в теореме 2. Другой такой пример дают задачи (1), (7), у которых функция Лагранжа

$$L(u, \lambda) = J(u) + \sum_{k=1}^s \lambda_k g_k(u), \quad u \in U_0, \quad \lambda \in \Lambda_0, \quad (19)$$

$$\Lambda_0 = \{\lambda \in E^s: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\},$$

имеет седловую точку на $U_0 \times \Lambda_0$. В самом деле, пусть $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка функции (19), т. е.

$$L(u_*, \lambda) \leq L(u_*, \lambda^*) = J_* \leq L(u, \lambda^*), \quad u \in U_0, \quad \lambda \in \Lambda_0. \quad (20)$$

Возьмем произвольную последовательность $\{u_k\}$, удовлетворяющую соотношениям (17). Согласно (20) тогда $J_* \leq$

$$\leq L(u_k, \lambda^*) = J(u_k) + \sum_{i=1}^s \lambda_i^* g_i(u_k). \quad \text{Учитывая, что } \lambda^* \in \Lambda_0, \text{ отсюда при } k \rightarrow \infty \text{ получим } \lim_{k \rightarrow \infty} J(u_k) \geq J_*. \text{ Таким образом, если}$$

функция Лагранжа задачи (1), (7) имеет седловую точку, то ограничения согласованы с функцией $J(u)$ на U_0 . Отсюда и из теорем 1, 3 вытекает следующая

Теорема 4. Пусть функция Лагранжа (19) имеет седловую точку и $J_* = \inf_{U_0} J(u) > -\infty$; пусть последовательность $\{u_k\}$ определена условиями (3), (6), (8). Тогда

$$\lim_{k \rightarrow \infty} J(u_k) = \lim_{k \rightarrow \infty} \Phi_k(u_k) = \lim_{k \rightarrow \infty} \Phi_k = J_*.$$

и справедливы соотношения (11), (12).

4. Покажем, что при выполнении условий теоремы 4 можно получить оценку скорости сходимости метода штрафных функций, а также убедимся в том, что утверждения теоремы 4 сохраняют силу и без требования $J_* > -\infty$.

Теорема 5. Пусть функция Лагранжа (19) задачи (1), (7) имеет седловую точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$. Пусть последовательность $\{u_k\}$ определена условиями (3), (6), (8). Тогда при $p = 1$ справедливы оценки

$$0 \leq A_k P(u_k) \leq \frac{e_k}{1 - \|\lambda^*\|_\infty A_k^{-1}}, \quad (21)$$

$$- \frac{e_k \|\lambda^*\|_\infty}{A_k (1 - \|\lambda^*\|_\infty A_k^{-1})} \leq J(u_k) - J_* \leq e_k \quad (22)$$

для всех тех номеров $k \geq 1$, для которых $A_k > \|\lambda^*\|_\infty = \max_{1 \leq i \leq s} |\lambda_i^*|$; если же $p > 1$, то

$$0 \leq A_k P(u_k) \leq (\|\lambda^*\|_q)^q A_k^{1-q} + q e_k, \quad (23)$$

$$- \|\lambda^*\|_q ((\|\lambda^*\|_q)^q A_k^{-q} + q e_k A_k^{-1})^{1/p} \leq J(u_k) - J_* \leq e_k, \quad (24)$$

где $k=1, 2, \dots, p$ и q связаны условием $p^{-1} + q^{-1} = 1$, $|\lambda^*|_q = \left(\sum_{i=1}^s |\lambda_i^*|^q \right)^{1/q}$.

Если, кроме того, U_0 замкнуто, а функции $J(u)$, $g_1(u)$, ..., $g_m(u)$, $|g_{m+1}(u)|$, ..., $|g_s(u)|$ полунепрерывны снизу на U_0 и u_* — предельная точка последовательности $\{u_k\}$, то $u_* \in U_*$.

Доказательство. Если $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ — седловая точка функции Лагранжа, то согласно теореме 4.8.1 $u_* \in U_*$. Из определения (9) функций $g_i^+(u)$ и условия $\lambda^* \in \Lambda_0$ следует, что $\lambda_i^* g_i(u) \leq |\lambda_i^*| g_i^+(u)$ для всех $i = \overline{1, s}$ и $u \in U_0$. Отсюда и из неравенств (20) получим

$$J_* \leq J(u) + \sum_{i=1}^s |\lambda_i^*| g_i^+(u), \quad u \in U_0. \quad (25)$$

Далее нам понадобится неравенство Гельдера [10]

$$\sum_{i=1}^s |\lambda_i^* g_i| \leq \left(\sum_{i=1}^s |\lambda_i^*|^q \right)^{1/q} \left(\sum_{i=1}^s |g_i|^p \right)^{1/p},$$

справедливое для всех действительных чисел λ_i^* , g_i , $p > 1$, $p^{-1} + q^{-1} = 1$. Это неравенство, записанное в виде

$$\sum_{i=1}^s |\lambda_i^* g_i| \leq |\lambda^*|_q \left(\sum_{i=1}^s |g_i|^p \right)^{1/p},$$

останется верным для всех $p \geq 1$, если при $p=1$ условимся принять $q = +\infty$, $|\lambda^*|_\infty = \max_{1 \leq i \leq s} |\lambda_i^*|$. При таком соглашении имеем

$$\sum_{i=1}^s |\lambda_i^*| g_i^+(u) \leq |\lambda^*|_q (P(u))^{1/p}, \quad u \in U_0, \quad p \geq 1.$$

Отсюда и из неравенства (25) следует

$$J_* \leq J(u) + |\lambda^*|_q (P(u))^{1/p}, \quad u \in U_0, \quad p \geq 1. \quad (26)$$

Так как $\lim_{k \rightarrow \infty} A_k = +\infty$, то $A_k \geq |\lambda^*|_\infty$ для всех $k \geq k_0$. Тогда из (26) при $p=1$ получим

$$J_* \leq J(u) + A_k P(u) = \Phi_k(u), \quad u \in U_0, \quad k \geq k_0.$$

Отсюда вытекает неравенство $\Phi_{k_*} = \inf_{U_0} \Phi_k(u) \geq J_* > -\infty$ при любом $k \geq k_0$.

Это значит, что последовательность $\{u_k\}$, удовлетворяющая условиям (6), при $p=1$ существует (без ущерба для общности ниже можем считать $k_0=1$).

Рассмотрим случай $p > 1$. Воспользуемся неравенством [10]

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q},$$

справедливым для всех действительных a , b и $p > 1$, $p^{-1} + q^{-1} = 1$. Положив $a = (pA_k)^{1/p} (P(u))^{1/p}$, $b = (pA_k)^{-1/p} |\lambda^*|_q$, с учетом равенства $q/p = q-1$ получим

$$|\lambda^*|_q (P(u))^{1/p} = ab \leq A_k P(u) + (|\lambda^*|_q)^q q^{-1} p^{1-q} A_k^{1-q}, \quad u \in U_0.$$

Подставим это неравенство в правую часть (26). Будем иметь

$$J_* \leq J(u) + A_k P(u) + B A_k^{1-q} = \Phi_k(u) + B A_k^{1-q}, \quad u \in U_0,$$

при всех $p > 1$, $k = 1, 2, \dots$, где $B = (|\lambda^*|_q)^q q^{-1} p^{1-q}$. Отсюда вытекает неравенство $\Phi_k \geq J_* - B A_k^{1-q} > -\infty$, $k = 1, 2, \dots$. Это значит, что и при каждом $p > 1$ существует последовательность $\{u_k\}$, удовлетворяющая условиям (6).

Перейдем к доказательству оценок (21) — (24). Так как $P(u_*) = 0$ для всех $u_* \in U_*$, то из неравенств (6), (26) и определений $\Phi_k(u)$, Φ_{k*} , $P_k(u) \geq 0$ имеем

$$\begin{aligned} J(u_k) &\leq J(u_k) + A_k P(u_k) = \Phi_k(u_k) \leq \Phi_{k*} + \varepsilon_k \leq \\ &\leq \Phi_k(u_*) + \varepsilon_k = J(u_*) + \varepsilon_k = J_* + \varepsilon_k \leq \\ &\leq J(u_k) + |\lambda^*|_q (P(u_k))^{1/p} + \varepsilon_k, \quad k = 1, 2, \dots \end{aligned} \quad (27)$$

Отсюда следует, что

$$A_k P(u_k) \leq |\lambda^*|_q (P(u_k))^{1/p} + \varepsilon_k, \quad k = 1, 2, \dots \quad (28)$$

Если $p = 1$, то при всех $A_k > |\lambda^*|_\infty$ неравенство (28) равносильно оценке (21). Кроме того, из (26), (27) при $p = 1$ будем иметь

$$-|\lambda^*|_\infty P(u_k) \leq J(u_k) - J_* \leq \varepsilon_k, \quad k = 1, 2, \dots$$

Отсюда и из уже доказанной оценки (21) вытекает оценка (22).

Рассмотрим случай $p > 1$. Если обозначить $z_k = (P(u_k))^{1/p}$, то неравенство (28) можно переписать в виде $0 \leq A_k z_k^p \leq |\lambda^*|_q z_k + \varepsilon_k$ или $0 \leq z_k^p \leq z_k b_k + d_k$, где $b_k = |\lambda^*|_q A_k^{-1}$, $d_k = \varepsilon_k A_k^{-1}$, $k = 1, 2, \dots$. Отсюда с помощью леммы 2.3.11 получим $0 \leq z_k \leq (b_k^q + q d_k)^{1/p}$ или $0 \leq P(u_k) \leq (|\lambda^*|_q)^q A_k^{-q} + q \varepsilon_k A_k^{-1}$, что равносильно оценке (23). Далее, из (26), (27) следует

$$-|\lambda^*|_q (P(u_k))^{1/p} \leq J(u_k) - J_* \leq \varepsilon_k, \quad k = 1, 2, \dots$$

Отсюда и из уже доказанной оценки (23) вытекает оценка (24).

Таким образом, $\lim_{k \rightarrow \infty} J(u_k) = J_*$, $\lim_{k \rightarrow \infty} P(u_k) = 0$. Поэтому, если u_* — предельная точка последовательности $\{u_k\}$, то включение $u_* \in U_*$ доказывается так же, как аналогичное утверждение в теореме 2. Теорема 5 доказана.

Заметим, что если последовательность $\{u_k\}$, определяемая условиями (5), существует, то при выполнении условий теоремы 5 оценки для такой последовательности будут получаться из оценок (21) — (24) при $\varepsilon_k = 0$, $k = 1, 2, \dots$. В частности, если в (24) $\varepsilon_k = 0$, то

$$0 \geq J(u_k) - J_* \geq -|\lambda^*|_q (|\lambda^*|_q)^{q/p} A_k^{-q/p} = -(|\lambda^*|_q)^q A_k^{1-q}, \quad k = 1, 2, \dots$$

Что же будет, если в (21), (22) принять $\varepsilon_k = 0$, $k = 1, 2, \dots$ — этот вопрос мы обсудим ниже в теореме 7.

Рассмотрим пример, который показывает, что оценки (21) — (24) не могут быть существенно улучшены на классе задач (1), (7), у которых функция Лагранжа имеет седловую точку.

Пример 4. Рассмотрим задачу

$$J(u) = -u \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) \equiv u \leq 0\}.$$

Здесь $J_* = 0$, $U_* = \{0\}$. Функция Лагранжа $L(u, \lambda) = -u + \lambda u$, $u \in U_0 = E^1$, $\lambda \in \Lambda_0 = \{\lambda \in E^1: \lambda \geq 0\}$ имеет седловую точку ($u_* = 0$, $\lambda^* = 1$). Возьмем штрафную функцию $P_k(u) = A_k(\max\{u; 0\})^p$, $p \geq 1$, $A_k > 1$, $\lim_{k \rightarrow \infty} A_k = +\infty$.

Тогда функция (3) будет иметь вид

$$\Phi_k(u) = \begin{cases} -u + A_k u^p, & \text{если } u \geq 0, \\ -u, & \text{если } u < 0. \end{cases}$$

Нетрудно показать (нарисуйте график функции $\Phi_k(u)$), что

$$\Phi_{k*} = \inf_{E^1} \Phi_k(u) = \begin{cases} 0 & \text{при } p = 1, \\ -q^{-1} p^{1-q} A_k^{1-q} & \text{при } p > 1, \end{cases}$$

причем нижняя грань достигается в точке $u_{k*} = 0$ при $p = 1$ и $u_{k*} = p^{1-q} A_k^{1-q} > 0$ при $p > 1$, $k = 1, 2, \dots$. Попутно отметим, что $\Phi_{k*} = -\infty$, если $p = 1$ и $A_k < |\lambda^*| = 1$. Указанная последовательность $\{u_{k*}\}$ удовлетворяет условиям (5) или условиям (6) при $\varepsilon_k = 0$, причем

$$A_k P(u_{k*}) = \begin{cases} 0 & \text{при } p = 1, \\ p^{-q} A_k^{1-q} & \text{при } p > 1; \end{cases}$$

$$J(u_{k*}) - J_* = \begin{cases} 0 & \text{при } p = 1, \\ -p^{1-q} A_k^{1-q} & \text{при } p > 1. \end{cases}$$

Сравнение полученных здесь точных оценок с оценками (21) — (24) при $\varepsilon_k = 0$ показывает, что в случае $p = 1$ оценки (21), (22) точны на классе задач (1), (7), у которых функция Лагранжа имеет седловую точку, а в случае $p > 1$ оценки (23), (24) точны по порядку A_k^{1-q} и отличаются лишь константами при A_k^{1-q} .

Если $\varepsilon_k > 0$, то при $p = 1$ в качестве точки u_k , удовлетворяющей условиям (6), в рассматриваемой задаче можно взять $u_k = \varepsilon_k (A_k - 1)^{-1}$, $k = 1, 2, \dots$. Тогда

$$P(u_k) = u_k = \varepsilon_k (A_k - |\lambda^*|)^{-1}, \quad J(u_k) - J_* = -u_k = -\varepsilon_k |\lambda^*| (A_k - |\lambda^*|)^{-1},$$

$k = 1, 2, \dots$, что полностью совпадает с оценками (21), (22).

Если $\varepsilon_k > 0$, $p = 2$, то точка $u_k = (1/2A_k) + (\varepsilon_k/A_k)^{1/2}$ удовлетворяет условиям (6), $k = 1, 2, \dots$. Отсюда имеем

$$A_k P(u_k) = A_k u_k^2 = (1/4A_k) + \varepsilon_k + (\varepsilon_k/A_k)^{1/2},$$

$$J(u_k) - J_* = -u_k, \quad k = 1, 2, \dots,$$

что также свидетельствует о том, что оценки (23), (24) на рассматриваемом классе задач не являются грубыми.

Любопытно, что если функция Лагранжа задачи не имеет седловой точки, то оценки для $A_k P(u_k)$, $J(u_k) - J_*$, вообще говоря, ухудшаются по сравнению с оценками (21) — (24).

Пример 5. Пусть

$$J(u) = -u \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) = u^2 \leq 0\}.$$

Здесь $J_* = 0$, $U_* = \{0\}$, функция Лагранжа $L(u, \lambda) = -u + \lambda u^2$ при $u \in U_0 = E^1$, $\lambda \in \Lambda_0 = \{\lambda \in E^1: \lambda \geq 0\}$ седловой точки не имеет. Возьмем штрафную функцию $P(u) = (\max\{u^2; 0\})^p = u^{2p}$. Нижняя грань функции $\Phi_k(u) = -u + A_k u^{2p}$, $A_k > 0$, $p > 1$, на $U_0 = E^1$ достигается при $u_{k*} =$

$$A_k P(u_{k*}) = (2p)^{-\frac{1}{2p-1}} A_k^{-\frac{1}{2p-1}},$$

$$J(u_k) - J_* = -(2p)^{-\frac{1}{2p-1}} A_k^{-\frac{1}{2p-1}}, \quad k = 1, 2, \dots$$

Так как $q-1 = (p-1)^{-1} > (2p-1)^{-1}$ при всех $p > 1$, то ясно, что полученные точные оценки уступают оценкам (23), (24). Интересно заметить, что с увеличением $p > 1$ оценка ухудшается.

5. Выше мы выяснили, что при определенных условиях последовательность $\{u_k\}$, получаемая методом штрафных функций (3), (6), (8), такова, что

$$\lim_{k \rightarrow \infty} P(u_k) = 0 \quad \text{или} \quad \lim_{k \rightarrow \infty} g_i^+(u_k) = 0, \quad i = \overline{1, s}, \quad (29)$$

что равносильно соотношениям (17). Однако, как показывает пример 2, отсюда, вообще говоря, еще не следует, что последовательность $\{u_k\}$ будет сходиться к множеству U .

Определение 3. Ограничения (16) называются *корректными* на множестве U_0 , если любая последовательность $\{u_k\} \in U_0$, удовлетворяющая условиям (29) (или (17)), сходится к множеству U .

Примеры 2, 3 показывают, что одно и то же множество может быть задано как корректными, так и некорректными ограничениями. Возникает вопрос, какие ограничения являются корректными? Здесь мы ограничимся приведением одного простого достаточного условия корректности ограничений (16).

Теорема 6. Пусть существует функция $\delta(t)$, определенная и неотрицательная при $t \geq 0$, $\delta(0) = 0$, $\lim_{t \rightarrow +0} \delta(t) = 0$, такая, что

$$\rho(u, U) = \inf_{v \in U} |u - v| \leq \delta(P(u)), \quad u \in U_0, \quad (30)$$

где штрафная функция $P(u)$ определена формулой (8) при каком-либо $p \geq 1$. Тогда ограничения (16) корректны на множестве U_0 .

Доказательство. Возьмем произвольную последовательность $\{u_k\}$, удовлетворяющую условиям (29) (или (17)). Полагая в (30) $u = u_k$ и пользуясь свойствами функции $\delta(t)$, получим $\rho(u_k, U) \leq \delta(P(u_k)) \rightarrow 0$ при $k \rightarrow \infty$, что и требовалось.

Заметим, что из неравенства (30) и оценок (21) — (24) следует оценка и для $\rho(u_k, U)$, $k = 1, 2, \dots$

Приведем несколько примеров корректных ограничений.

Пример 6. Пусть

$$U = \{u \in E^n: g_i(u) = \langle a_i, u \rangle - b^i = 0, \quad i = \overline{1, m}\}$$

— аффинное множество, где $a_1, \dots, a_m$ — линейно независимая система век-

торов из E^n ; $b^1, \dots, b^m$ — заданные числа. Проекция произвольной точки $u \in E^n$ на множество U имеет вид

$$P_U(u) = u - A^T (AA^T)^{-1} (Au - b),$$

где A — матрица порядка $m \times n$, строками которой являются векторы $a_1, \dots, a_m$, а $b = (b^1, \dots, b^m)$ (см. пример 4.4.3). Тогда

$$\rho(u, U) = |u - P_U(u)| \leq \|A^T (AA^T)^{-1}\| \cdot |Au - b|, \quad u \in E^n.$$

Это неравенство можно записать в виде $\rho(u, U) \leq \delta(P(u))$, где $\delta(t) = \|A^T (AA^T)^{-1}\| t^{1/2}$, $P(u) = |Au - b|^2 = \sum_{i=1}^m |g_i(u)|^2$ — штрафная функция множества U . Отсюда и из теоремы 6 следует корректность ограничений $g_i(u) = \langle a_i, u \rangle - b^i = 0$, $i = \overline{1, m}$, на E^n .

Пример 7. Пусть

$$U = \{u \in E^n: g(u) = \langle c, u \rangle - \gamma \leq 0\}$$

— замкнутое полупространство, где $c \neq 0$ — заданный вектор из E^n , γ — число. Проекция любой точки $u \in E^n$ на множество U имеет вид

$$P_U(u) = u - \max\{\langle c, u \rangle - \gamma; 0\} c |c|^{-2}$$

(см. пример 4.4.4). Следовательно,

$$\rho(u, U) = |u - P_U(u)| \leq |c| g^+(u), \quad u \in E^n.$$

Это неравенство можно переписать в виде $\rho(u, U) \leq \delta(P(u))$, где $\delta(t) = |c| t^{1/p}$, $P(u) = (g^+(u))^p$, $p \geq 1$, — штрафная функция множества U . Отсюда и из теоремы 6 следует корректность ограничения $g(u) = \langle c, u \rangle - \gamma \leq 0$ на E^n .

Пример 8. Пусть функции $g_1(u), \dots, g_m(u), |g_{m+1}(u)|, \dots, |g_s(u)|$ определены и полунепрерывны снизу на замкнутом множестве U_0 , а множество U_δ , определенное формулами (14), ограничено при некотором $\delta > 0$. Тогда ограничения (16) корректны на U_0 — это следует из доказательства теоремы 2.

6. При описании и исследовании метода штрафных функций (3), (6), (8) всюду выше предполагалось, что штрафные коэффициенты $\{A_k\}$ стремятся к бесконечности. Однако, оказывается, существуют классы задач и классы штрафных функций, когда можно ограничиться рассмотрением достаточно больших, но конечных штрафных коэффициентов.

Теорема 7. Пусть функция Лагранжа (19) задачи (1), (7) имеет седловую точку (u_*, λ^*) на множестве $U_0 \times \Lambda_0$. Пусть

$$\Phi_A(u) = J(u) + AP(u), \quad P(u) = \sum_{i=1}^s g_i^+(u), \quad u \in U_0, \quad (31)$$

где функции $g_i^+(u)$ определены по формулам (9). Тогда при всех $A > |\lambda^*|_\infty = \max_{1 \leq i \leq s} |\lambda_i^*|$ множество $U_{A*} = \{u \in U_0: \Phi_A(u) = \Phi_{A*} = \inf_{u \in U_0} \Phi_A(u)\}$ непусто и совпадает со множеством U_* точек минимума в задаче (1), (7) и, кроме того, $\Phi_{A*} = J_*$.

Доказательство. Согласно теореме 4.8.1, если (u_*, λ^*) — седловая точка функции Лагранжа, то $u_* \in U_*$, так что $U_* \neq \emptyset$. При $A_k > |\lambda^*|_\infty =$

$= \max_{1 \leq i \leq s} |\lambda_i^*|$ из неравенства (25) получим

$$J_* \leq J(u) + AP(u) = \Phi_A(u) \quad \text{при всех } u \in U_0. \quad (32)$$

Следовательно, $J_* \leq \Phi_{A_*}$. Возьмем произвольную точку $u_* \in U_*$. Тогда из неравенства (32) с учетом $P(u_*) = 0$ будем иметь $J_* \leq \Phi_{A_*} \leq \Phi_A(u_*) = J(u_*) + AP(u_*) = J(u_*) = J_*$ или $\Phi_{A_*} = \Phi_{A_*}(u_*) = J_*$. Таким образом, $\Phi_{A_*} = J_*$ и $u_* \in U_{A_*}$. Отсюда следует, что $U_{A_*} \neq \emptyset$ и $U_* \subseteq U_{A_*}$.

Докажем справедливость включения $U_{A_*} \subseteq U_*$. Возьмем произвольную точку $u_A \in U_{A_*}$. Учитывая неравенство (26) при $p = 1$ и уже доказанное равенство $\Phi_{A_*} = J_*$, получим

$$J(u_A) + AP(u_A) = \Phi_A(u_A) = \Phi_{A_*} = J_* \leq J(u_A) + |\lambda^*|_\infty P(u_A). \quad (33)$$

Отсюда имеем $AP(u_A) \leq |\lambda^*|_\infty P(u_A)$ или $(A - |\lambda^*|_\infty)P(u_A) \leq 0$. При $A > |\lambda^*|_\infty$ это возможно лишь при $P(u_A) = 0$, т. е. $u_A \in U$. Тогда из (33) вытекает $J(u_A) = J_*$, т. е. $u_A \in U_*$. Это значит, что $U_{A_*} \subseteq U_*$. С учетом вышесказанного включения $U_* \subseteq U_{A_*}$ отсюда получаем требуемое.

В теореме 7 условие $A > |\lambda^*|_\infty$ заменить на $A = |\lambda^*|_\infty$, вообще говоря, нельзя. В этом нетрудно убедиться, рассмотрев задачу из примера 4.

7. При $p > 1$ утверждение, аналогичное теореме 7, неверно. Точнее, для функции $\Phi_k(u) = J(u) + A_k P(u)$ со штрафной функцией (8) при $p > 1$ нельзя ожидать, что множество $U_{k*} = \{u \in U_0, \Phi_k(u) = \Phi_{k*}\}$ будет совпадать со множеством U_* при каких-либо $A_k > 0$. Чтобы убедиться в этом, достаточно рассмотреть задачу из примера 1: полученная в этом примере точка $u_k \in U_{k*}$, но $u_k \notin U$, и в то же время $u_* \notin U_{k*}$ при всех $k = 1, 2, \dots$. Тем не менее и при $p > 1$ существует интересная связь между множителями Лагранжа, штрафными коэффициентами $\{A_k\}$ и последовательностью $\{u_k\}$, определяемой условиями (3), (6), (8).

Теорема 8. Пусть U_0 — выпуклое замкнутое множество из E^n , функции $J(u)$, $g_1(u), \dots, g_m(u)$ выпуклы на U_0 и принадлежат классу $C^1(U_0)$, $g_i(u) = \langle a_i, u \rangle - b^i$, $i = m+1, s$, — линейные функции, пусть функция Лагранжа (19) задачи (1), (7) имеет седловую точку на $U_0 \times \Lambda_0$ в смысле неравенств (20). Кроме того, пусть функция

$$\Phi_k(u) = J(u) + A_k \sum_{i=1}^s (g_i^+(u))^p, \quad u \in U_0, \quad p > 1,$$

при каждом $k = 1, 2, \dots$ достигает своей нижней грани на множестве U_0 в некоторой точке $u_k \in U_0$, т. е. u_k удовлетворяет условиям (5) или условиям (6) при $\varepsilon_k = 0$. Тогда если последовательность $\{u_k\}$ имеет хотя бы одну предельную точку, то и последовательность $\{\lambda^k\}$, где $\lambda^k = (\lambda_1^k, \dots, \lambda_s^k)$, $\lambda_i^k = p A_k |g_i^+(u_k)|^{p-1}$, $i = \overline{1, m}$; $\lambda_i^k = p A_k |g_i^+(u_k)|^{p-1} \text{sign } g_i(u_k)$, $i = \overline{m+1, s}$, также будет иметь предельную точку, причем любая предельная точка (u_*, λ^*) последовательности $\{(u_k, \lambda^k)\}$ будет седловой точкой функции Лагранжа.

Доказательство. Из оценки (23) при $\varepsilon_k = 0$ получим $|g_i^+(u_k)|^p \leq P(u_k) \leq A_k^{-q} (|\lambda^*|_q)^q$ или $A_k |g_i^+(u_k)|^{p-1} \leq A_k (A_k^{-q} (|\lambda^*|_q)^q)^{(p-1)/p} =$

$= |\lambda^*|_q$ при всех $i = \overline{1, s}$ и $k = 1, 2, \dots$. Отсюда следует, что $|\lambda_i^k| \leq p |\lambda^*|_q < \infty$, $i = \overline{1, s}$, $k = 1, 2, \dots$, т. е. последовательность $\{\lambda^k\}$ ограничена. Пусть u_* — какая-либо предельная точка последовательности $\{u_k\}$, пусть $\{u_{k_r}\} \rightarrow u_*$.

Согласно теореме 5 $u^* \in U_*$. Из $\{\lambda^{k_r}\}$ выделим подпоследовательность, сходящуюся к некоторой точке λ^* . Можем считать, что сама подпоследовательность $\{\lambda^{k_r}\}$ сходится к λ^* . Так как $\lambda_i^k \geq 0$ при всех $i = \overline{1, m}$, то и $\lambda_i^* \geq 0$, $i = \overline{1, m}$, так что $\lambda_* \in \Lambda_0$. Далее, так как $\Phi_k(u)$ выпукла и $\Phi_k(u) \in C^1(U_0)$, то в точке u_k справедливо неравенство

$$\langle \Phi'_k(u_k), u - u_k \rangle \geq 0, \quad u \in U_0. \quad (34)$$

Поскольку

$$\begin{aligned} \Phi'_k(u_k) &= J'(u_k) + \sum_{i=1}^m p A_k |g_i^+(u_k)|^{p-1} g'_i(u_k) + \\ &+ \sum_{i=m+1}^s p A_k |g_i^+(u_k)|^{p-1} \operatorname{sign} g_i(u_k) g'_i(u_k) = J'(u_k) + \sum_{i=1}^s \lambda_i^k g'_i(u_k), \end{aligned}$$

то

$$\lim_{r \rightarrow \infty} \Phi'_{k_r}(u_{k_r}) = J'(u_*) + \sum_{i=1}^s \lambda_i^* g'_i(u_*) = L_u(u_*, \lambda^*).$$

Поэтому из (34) при $k = k_r \rightarrow \infty$ получим $\langle L_u(u_*, \lambda^*), u - u_* \rangle \geq 0$, $u \in U_0$. Так как $\lambda^* \in \Lambda_0$, то при выполнении условий теоремы функция $L(u, \lambda^*)$ выпукла на U_0 , и последнее неравенство равносильно такому:

$$L(u_*, \lambda^*) \leq L(u, \lambda^*), \quad u \in U_0. \quad (35)$$

Далее, если $g_i(u_*) < 0$ при некотором i , $1 \leq i \leq m$, то из $\lim_{r \rightarrow \infty} g_i(u_{k_r}) = g_i(u_*) < 0$ следует $g_i(u_{k_r}) < 0$ или $g_i^+(u_{k_r}) = 0$ для всех $r \geq r_0$. А тогда $\lambda_i^{k_r} = 0$ при всех $r \geq r_0$ и $\lim_{r \rightarrow \infty} \lambda_i^{k_r} = \lambda_i^* = 0$. Таким образом, $\lambda_i^* = 0$ для всех номеров i , $1 \leq i \leq m$, для которых $g_i(u_*) < 0$, а для остальных номеров i , $1 \leq i \leq s$, мы имеем $g_i(u_*) = 0$. Следовательно, $\lambda_i^* g'_i(u_*) = 0$, $i = \overline{1, s}$. Отсюда и из (35) с помощью леммы 4.8.1 заключаем, что (u_*, λ^*) — седловая точка функции Лагранжа. Теорема доказана.

8. 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 за-

дачи (2): в задании множества (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 $g_i(u) \leq 0$ 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 достаточно 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 в качестве штрафной функции множества (7) берется функция (8) при $p = 1$, то согласно теореме 7 в выпуклых регулярных задачах 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 $p = 1$ не будет обладать достаточной гладкостью, но появившиеся в последние годы методы минимизации, не требующие гладко-

сти минимизируемой функции (см. §§ 3, 9, 10, 15), 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 [16, 59].

Различные прикладные и теоретические аспекты метода штрафных функций исследованы в [5, 9, 11, 15, 59, 84, 101, 107, 109, 119, 137, 138, 141, 228, 229, 236, 250] и др.

В заключение заметим, что рассмотренный в § 11 метод множителей Лагранжа идейно примыкает к методу штрафных функций. В самом деле, в методе множителей Лагранжа ограничения типа $g_i(u) \leq 0$ или $g_i(u) = 0$ учитываются в функции Лагранжа или ее модификации. Затем исходная задача (1), (7) сводится к последовательности вспомогательных задач минимизации на множестве U_0 (см., например, задачу (11.12)). Достоинством метода множителей Лагранжа является то, что в нем приближения $\{\lambda_k\}$ для множителей Лагранжа пересчитываются по простому правилу и отсутствуют неограниченно растущие коэффициенты типа штрафных коэффициентов. В то же время метод множителей Лагранжа требует от задачи (1), (7) существования седловой точки функции Лагранжа, а метод штрафных функций может сходиться для более широких классов задач (см., например, теорему 2) и является более универсальным.

У п р а ж н е н и я. 1. Применить метод штрафных функций к задачам
а) $J(u) = x^2 + y^2 \rightarrow \inf; u \in U = \{u = (x, y) \in E^2: g(u) = -x - y + 1 \leq 0\}$
или $u \in U = \{u = (x, y) \in E^2: g(u) = -x - y + 1 = 0\};$

б) $J(u) = xy \rightarrow \inf; u \in U = \{u = (x, y) \in E^2: x^2 + y^2 \leq 25\}$ или $u \in U = \{u = (x, y) \in E^2: x^2 + y^2 = 25\};$

в) $J(u) = x^2 + y^2 + z^2 \rightarrow \inf; u \in U = \{u = (x, y, z) \in E^3: x + y + z + 1 \leq 0\}.$

2. Применить метод штрафных функций к задачам из примеров 2.2.2 и 2.2.4.

3. Пусть $J(u) = e^{-2u}$, а множество $U = \{u \in E^1: 0 \leq u \leq 1\}$ задано либо ограничениями $g_1(u) = -u \leq 0$, $g_2(u) = u - 1 \leq 0$, либо $g(u) = |u| + |u - 1| - 1 = 0$, либо $g(u) = e^{-u}(|u| + |u - 1| - 1) = 0$. Выяснить, в каких случаях ограничения согласованы с функцией $J(u)$ на E^1 .

4. Пусть $\{P_k(u)\}$ — штрафная функция некоторого множества U . Пусть функция $\varphi(t)$ определена при $t \geq 0$, $\varphi(0) = 0$, причем $\varphi(t) \rightarrow 0$ при $t \rightarrow +0$, $\varphi(t) \rightarrow +\infty$ при $t \rightarrow +\infty$. Показать, что тогда $\{\varphi(P_k(u))\}$ является штрафной функцией множества U .

5. Применить метод (3), (6), (8) к задаче $J(u) = u^2 - u \rightarrow \inf; u \in U = \{u \in E^1: g(u) = u \leq 0\}$, 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 $U = \{u \in E^1: -1 \leq u \leq 1\}$ задано либо ограничениями $g_1(u) = u - 1 \leq 0$, $g_2(u) = -u - 1 \leq 0$, либо $g(u) = e^{-u^2}(u^2 - 1) \leq 0$, либо

$g(u) = u^2 - 1 \leq 0$. Выяснить, какие из этих ограничений являются корректными на E^1 .

7. Пусть $U = \{u \in E^n: g(u) \leq 0\}$, где $g(u)$ — непрерывная функция на E^n . Доказать, что для того, чтобы множество U было ограниченным и ограничением $g(u) \leq 0$ было корректным на E^n , необходимо и достаточно, чтобы множество $U_\delta = \{u \in E^n: g(u) \leq \delta\}$ 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 $M(C) = \{u \in U_0: g(u) \leq C\}$ непусто и ограничено при некотором C . Доказать, что тогда ограничение $g(u) \leq 0$ корректно на U_0 (см. теорему 4.2.17 и пример 8).

9. Пусть W — открытое выпуклое множество из E^n , U_0 — выпуклое множество, $U_0 \subset W$, функции $J(u)$, $g_1(u)$, ..., $g_m(u)$ выпуклы на W , а $g_i(u) = \langle a_i, u \rangle - b^i$, $i = \overline{m+1}, s$, — линейные функции. Пусть функция $J(u)$ на множестве (7) ограничена снизу и множество U_* точек минимума $J(u)$ на множестве (7) непусто. Кроме того, пусть при некотором $A > 0$ непусто

множество U_{A*} точек минимума функции $\Phi_A(u) = J(u) + A \sum_{i=1}^m \max\{g_i(u); 0\} +$
 $+ \sum_{i=m+1}^s |g_i(u)|$ на множестве U_0 и $U_{A*} = U_*$. Доказать, что тогда функция Лагранжа (19) имеет седловую точку в смысле неравенств (20).

§ 13. Метод барьерных функций

1. Идеи метода штрафных функций могут быть использованы для построения методов решения задачи

$$J(u) \rightarrow \inf; \quad u \in U, \quad (1)$$

позволяющих получить такую минимизирующую последовательность $\{u_k\} \in U$, каждый член которой будет лежать вне некоторого заданного «запрещенного» подмножества $\gamma \subset U$. В качестве «запрещенного» множества γ может служить, например, граница $\text{Гр } U$ 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 $\text{Гр } U$. 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 $\gamma = \text{Гр } U$ 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 не-

отрицательна во всех точках $u \in U \setminus \gamma$, причем $\lim_{r \rightarrow \infty} B(v_r) = +\infty$ для всех последовательностей $\{v_r\} \in U \setminus \gamma$, которые сходятся к какой-либо точке $v \in \gamma$.

Заметим, что в определении 1 подразумевается, что $U \setminus \gamma \neq \emptyset$. Это значит, что если $\gamma = \text{Гр } U$, то $\text{int } U = U \setminus \gamma \neq \emptyset$. 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 $\gamma = \{u \in E^n: u \in U, g(u) = 0\}$, где $g(u)$ непрерывна на U , $U \setminus \gamma \neq \emptyset$, то в качестве барьерной функции здесь можно взять $B(u) = |g(u)|^{-1}$, или $B(u) = |g(u)|^{-p}$, или $B(u) = \max\{-\ln |g(u)|; 0\}$. Если же $\gamma = \{u \in E^n: u \in U, g(u) \leq 0\}$, где $U \setminus \gamma \neq \emptyset$, $g(u)$ непрерывна на U , то можно принять $B(u) = (\max\{g(u); 0\})^{-p}$, $p > 0$, или $B(u) = |\ln g(u)|$, $u \in U \setminus \gamma$ и т. п.

Перейдем к описанию метода барьерных функций для решения задачи (1), предполагая, что подмножество $\gamma \in U$ и некоторая его барьерная функция уже заданы. Введем функции

$$F_k(u) = J(u) + a_k B(u), \quad u \in U \setminus \gamma, \quad k = 1, 2, \dots, \quad (2)$$

где $\{a_k\}$ — положительная последовательность, сходящаяся к нулю. Величины $\{a_k\}$ из (2) называются *барьерными коэффициентами*. Рассмотрим последовательность задач

$$F_k(u) \rightarrow \inf; \quad u \in U \setminus \gamma, \quad k = 1, 2, \dots \quad (3)$$

Обозначим $F_{k*} = \inf_{U \setminus \gamma} F_k(u)$, $k = 1, 2, \dots$ Будем предполагать,

что в исходной задаче (1) $J_* = \inf_U J(u) > -\infty$. Так как $F_k(u) \geq J(u)$ при всех $u \in U \setminus \gamma$, то $F_{k*} \geq J_* > -\infty$. Тогда условия

$$u_k \in U \setminus \gamma, \quad F_k(u_k) \leq F_{k*} + \varepsilon_k, \quad k = 1, 2, \dots, \quad (4)$$

определяют последовательность $\{u_k\}$, где $\varepsilon_k > 0$, $\lim_{k \rightarrow \infty} \varepsilon_k = 0$;

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 $\{v_r\} \in U \setminus \gamma$, $\{v_r\} \rightarrow v \in \gamma$ справедливо равенство $\lim_{r \rightarrow \infty} F_k(v_r) = +\infty$ при каждом фиксированном

$k = 1, 2, \dots$ 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 $\{a_k\} \rightarrow 0$, 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 $\gamma = \text{Gr } U$ и $U \setminus \gamma = \text{int } U \neq \emptyset$, то для решения задачи (3) может быть применен, например, градиентный метод (см. § 1):

$$u_{kr+1} = u_{kr} - \alpha_r F'_k(u_{kr}), \quad u_{k0} = u_{k-1}, \quad r = 0, 1, \dots$$

Поскольку $u_{kr} \in \text{int } U$, то при достаточно малых $\alpha_r > 0$ и точка u_{kr+1} также будет принадлежать $\text{int } U$, и мы избавлены от неудобств, связанных с учетом границы U , — нужно лишь на каждой итерации следить за соблюдением включения $u_k \in \text{int } U$, а при его нарушении уменьшать длину шага α_r . Правда, для этого величину α_r , быть может, придется брать слишком малой, и сходимость градиентного метода, возможно, замедлится, но это уже будет «платой» за выполнение условия $u_k \in \text{int } U$.

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 § 12 — методом внешних штрафов (или методом внешней точки) [229]. Для иллюстрации метода барьерных функций приведем пример.

Пример 1. Пусть требуется решить задачу

$$J(u) = -u \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) = u \leq 0\}.$$

Очевидно, здесь $J_* = 0$, $U_* = \{0\}$. Границей множества U является $\gamma = \text{Gr } U = \{u \in E^1: g(u) = u = 0\} = \{0\}$, а $U \setminus \gamma = \{u \in E^1: g(u) = u < 0\} = \text{int } U$. В качестве барьерной функции для γ возьмем $B(u) = -1/u$, $u < 0$. Пусть $a_k = k^{-1}$, $k = 1, 2, \dots$. Тогда функция (2) будет иметь вид $F_k(u) = -u - (ku)^{-1}$, $u < 0$. Нетрудно видеть, что здесь $F_{k*} = \inf_{u < 0} F_k(u) =$

$= 2/\sqrt{k}$ и точка $u_k = -1/\sqrt{k}$ удовлетворяет условиям (4) при $\epsilon_k = 0$, $k = 1, 2, \dots$. Ясно также, что $\lim_{k \rightarrow \infty} F_{k*} = \lim_{k \rightarrow \infty} J(u_k) = 0 = J_*$, $\lim_{k \rightarrow \infty} u_k = 0 = u_*$.

В качестве барьерной функции здесь можно также взять и $B(u) = |\ln(-u)|$. В этом случае $F_k(u) = -u + |\ln(-u)|k^{-1}$, $u < 0$, $k = 1, 2, \dots$, $F_{k*} = (1 + \ln k)k^{-1}$, а точка $u_k = -k^{-1}$ удовлетворяет условиям (4) при $\varepsilon_k = 0$. И здесь $\{J(u_k)\} \rightarrow J_* = 0$, $\{u_k\} \rightarrow u_* = 0$.

Перейдем к исследованию сходимости метода барьерных функций.

Теорема 1. Пусть γ — некоторое подмножество из U , $U \setminus \gamma \neq \emptyset$, и

$$J_* = J_{**}, \quad \text{где } J_* = \inf_U J(u), \quad J_{**} = \inf_{U \setminus \gamma} J(u) > -\infty. \quad (5)$$

Пусть $B(u)$ — какая-либо барьерная функция подмножества γ , а последовательность $\{u_k\}$ определена условиями (4). Тогда

$$\lim_{k \rightarrow \infty} F_{k*} = \lim_{k \rightarrow \infty} F_k(u_k) = \lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} a_k B(u_k) = 0. \quad (6)$$

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

$$-\infty < J_{**} \leq J(u_k) \leq F_k(u_k) \leq F_{k*} + \varepsilon_k \leq F_k(u) + \varepsilon_k = J(u) + a_k B(u) + \varepsilon_k \quad (7)$$

при всех $u \in U \setminus \gamma$, $k = 1, 2, \dots$. Так как $B(u)$ конечна в любой точке $u \in U \setminus \gamma$, $\{a_k\} \rightarrow 0$, то из (7) при $k \rightarrow \infty$ получим

$$J_{**} \leq \lim_{k \rightarrow \infty} F_{k*} \leq \overline{\lim}_{k \rightarrow \infty} F_{k*} \leq J(u), \quad u \in U \setminus \gamma.$$

Переходя в этих неравенствах к нижней грани по $u \in U \setminus \gamma$, будем иметь $J_{**} \leq \lim_{k \rightarrow \infty} F_{k*} \leq \overline{\lim}_{k \rightarrow \infty} F_{k*} \leq J_{**}$, т. е. $\lim_{k \rightarrow \infty} F_{k*} = J_{**}$.

Отсюда и из (7) вытекает $\lim_{k \rightarrow \infty} F_k(u_k) = \lim_{k \rightarrow \infty} J(u_k) = J_{**}$. Так как $J_* = J_{**}$, то первые соотношения (6) доказаны. А тогда из $0 \leq a_k B(u_k) = F_k(u_k) - J(u_k) \rightarrow 0$ при $k \rightarrow \infty$ 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 $\lim_{r \rightarrow \infty} B(v_r) = +\infty$, где $\{v_r\} \in U \setminus \gamma$, $\{v_r\} \rightarrow v \in \gamma$, 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

$$U = \{u \in U_0, \quad g_i(u) \leq 0, \quad i = \overline{1, m}\}; \quad (8)$$

здесь U_0 — заданное множество из E^n , функции $g_1(u), \dots, g_m(u)$ определены и полунепрерывны снизу на U_0 . Положим

$$\gamma = \{u \in U: g_i(u) = 0 \text{ хотя бы для одного } i, \quad 1 \leq i \leq m\}. \quad (9)$$

Будем предполагать, что множество

$$U(-0) = \{u \in U_0: g_i(u) < 0, \quad i = \overline{1, m}\} \quad (10)$$

непусто. Тогда $U \setminus \gamma = U(-0) \neq \emptyset$. Довольно широкий класс барьерных функций для множества (9) дает следующая конструкция:

$$B(u) = \sum_{i=1}^m \varphi_i(-g_i(u)), \quad u \in U(-0), \quad (11)$$

где $\varphi_i(t)$ — неотрицательная функция переменной $t > 0$ такая, что $\lim_{t \rightarrow +0} \varphi_i(t) = +\infty$ при всех $i = \overline{1, m}$. В самом деле, возьмем произвольную последовательность $\{v_r\} \in U \setminus \gamma$, сходящуюся к некоторой точке $v \in \gamma$. Согласно (9) тогда найдется номер j , $1 \leq j \leq m$, для которого $g_j(v) = 0$. Так как $g_j(u)$ полунепрерывна снизу, а $g_j(v_r) < 0$, $r = 1, 2, \dots$, то $0 = g_j(v) \leq \lim_{r \rightarrow \infty} g_j(v_r) \leq \lim_{r \rightarrow \infty} \overline{g_j(v_r)} = 0$, т. е. $\lim_{r \rightarrow \infty} g_j(v_r) = 0$. Это значит, что $B(v_r) \geq \varphi_j(-g_j(v_r)) \rightarrow \infty$ при $r \rightarrow \infty$, 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, выпуклость, 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) $\varphi_i(t) \equiv 1/t$ или $\varphi_i(t) = (\max\{-\ln t; 0\})^p$, $p \geq 1$, получим соответственно

$$B(u) = - \sum_{i=1}^m 1/g_i(u) \quad (12)$$

или

$$B(u) = \sum_{i=1}^m (\max\{-\ln(-g_i(u)); 0\})^p, \quad u \in U(-0).$$

Если U_0 выпукло, функции $g_i(u)$, $i = \overline{1, m}$, 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, какую обладают функции $g_i(u)$, $i = \overline{1, m}$, — у второй функции (12) для этого нужно взять параметр p достаточно большим.

Может сложиться впечатление, что если функции $g_1(u), \dots, g_m(u)$ 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 $|u| \leq 1$, $g(u) = 0$ при $1 < |u| < 2$, $g(u) = |u| - 2$ при $|u| \geq 2$. Тогда множество $U = \{u \in E^1 = U_0: g(u) \leq 0\}$ представляет собой отрезок $-2 \leq u \leq 2$ на числовой оси, а $\text{int } U = \{u \in E^1: -2 < u < 2\}$. В то же время множество $U(-0) = \{u \in E^1: g(u) < 0\} = \{u: -1 < u < 1\} \subset \text{int } U$, но $U(-0) \neq \text{int } U$, а $\gamma = \{u \in U: g(u) = 0\} = \{u: 1 \leq |u| \leq 2\}$ наряду с граничными точками $u = 2$ и $u = -2$ содержит и внутренние точки множества U .

Таким образом, для множества (8) не всегда выполняется равенство

$$\text{Гр } U = \text{Гр } U_0 \cup \gamma, \quad (13)$$

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 $J(u)$, $g_1(u), \dots, g_m(u)$ 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 $J(u) = e^{-x}$, $U = \{u = (x, y) \in E^2 = U_0: g(u) = (x^2 + y^2 - 1)(y - 1)^2 \leq 0\}$. Тогда $\gamma = \text{Гр } U = \{u \in U: g(u) = 0\} = \{u \in E^2: x^2 + y^2 = 1 \text{ или } y = 1\}$, $U_0 = E^2$, $\text{Гр } U_0 = \emptyset$, так что условие (13) выполнено. Далее, здесь $U \setminus \gamma = U(-0) = \{u \in E^2: g(u) < 0\} = \{u: x^2 + y^2 < 1\} = \text{int } U$, поэтому $J_{**} = \inf_{U \setminus \gamma} J(u) = e^{-1} > 0$. В то же время $J_* = \lim_{k \rightarrow \infty} J(u_k) = 0$, где $u_k = (k, 1) \in U$, $k = 1, 2, \dots$

Заметим, что в этом примере $\overline{U(-0)} = \{u: x^2 + y^2 \leq 1\} \subset U$, но $\overline{U(-0)} \neq U$. (Напоминаем; что через $\bar{Z}$ мы условились обозначать замыкание множества Z .)

Приведем две теоремы, дающие достаточные условия для выполнения равенства (5). Имая в виду дальнейшие применения, утверждения сформулируем для множества

$$U(C) = \{u \in E^n: u \in U_0, \quad g_i(u) \leq C, \quad i = \overline{1, m}\}, \quad (14)$$

где C — некоторая постоянная. Обозначим

$$J_*(C) = \inf_{U(C)} J(u), \quad J_*(C-0) = \lim_{\varepsilon \rightarrow +0} J_*(C-\varepsilon), \quad J_*(C+0) = \lim_{\varepsilon \rightarrow +0} J_*(C+\varepsilon). \quad (15)$$

Теорема 2. Пусть для некоторых $C, \varepsilon_0 > 0$ множество $U(C-\varepsilon_0)$ непусто и

$$\overline{U(C-0)} = U(C), \quad (16)$$

где

$$U(C-0) = \{u \in U_0: g_i(u) < C, \quad i = \overline{1, m}\}.$$

Пусть, кроме того, функция $J(u)$ полунепрерывна сверху на множестве $U(C)$. Тогда

$$J_*(C-0) = J_*(C) = \inf_{U(C-0)} J(u). \quad (17)$$

Доказательство. Прежде всего, заметим, что

$$U(C-\varepsilon) \subseteq U(C-\delta) \subseteq U(C-0) \subseteq U(C)$$

при всех $0 < \delta < \varepsilon \leq \varepsilon_0$; поэтому

$$J_*(C) \leq \inf_{U(C-0)} J(u) \leq J_*(C-\delta) \leq J_*(C-\varepsilon).$$

Это значит, что функция $J_*(C)$ переменной C не возрастает и существует предел

$$\lim_{\varepsilon \rightarrow +0} J_*(C-\varepsilon) = J_*(C-0) \geq \inf_{U(C-0)} J(u) \geq J_*(C). \quad (18)$$

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 $u_k \in U_0, g_i(u_k) \leq C - \varepsilon_{ik} < C, \varepsilon_{ik} > 0, k = 1, 2, \dots$, где $\lim_{k \rightarrow \infty} \varepsilon_{ik} = 0, i = \overline{1, m}$.

Таким образом, $u_k \in U(C-\varepsilon_k)$, где $\varepsilon_k = \min_{1 \leq i \leq m} \varepsilon_{ik} > 0, \{e_k\} \rightarrow 0$, и $J_*(C-\varepsilon_k) \leq J(u_k), k = 1, 2, \dots$. Отсюда при $k \rightarrow \infty$, учитывая полунепрерывность сверху функции $J(u)$, получим $\lim_{\varepsilon \rightarrow +0} J_*(C-\varepsilon) = \lim_{k \rightarrow \infty} J_*(C-\varepsilon_k) \leq J(u)$. В силу произвольности $u \in U(C)$ тогда $J_*(C-0) \leq J_*(C)$. 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 $J(u)$, $g_1(u), \dots, g_m(u)$ выпуклы на U_0 . Тогда равенства (17) справедливы при всех $C > C_* = \max_{1 \leq i \leq m} \inf_{U_0} g_i(u)$.

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 $\alpha, 0 \leq \alpha \leq 1$. Кроме того, из выпуклости $g_i(u)$ имеем $g_i(u_\alpha) \leq \alpha g_i(v) + (1 - \alpha)g_i(u) \leq \alpha(C - \varepsilon) + (1 - \alpha)C = C - \alpha\varepsilon, 0 < \alpha \leq 1$. Это значит, что $u_\alpha \in U(C - \alpha\varepsilon)$. Тогда с учетом выпуклости функции $J(u)$ получим $J_*(C) \leq \inf_{U(C - \varepsilon)} J(u) \leq J_*(C - \alpha\varepsilon) \leq J(u_\alpha) \leq \alpha J(v) + (1 - \alpha)J(u)$ для всех $u \in U(C), v \in U(C - \varepsilon)$. Следовательно, $J_*(C) \leq \inf_{U(C - \varepsilon)} J(u) \leq J_*(C - \alpha\varepsilon) \leq \alpha J_*(C - \varepsilon) + (1 - \alpha)J_*(C) \leq J_*(C) + \alpha[J_*(C - \varepsilon) - J_*(C)]$ для всех $\alpha, 0 < \alpha \leq 1, 0 < \varepsilon \leq \varepsilon_0 < C - C_*$. Отсюда при $\alpha \rightarrow +0$ с учетом монотонности $J_*(C)$ получаем равенства (17), что и требовалось.

4. Пусть множество U задается условиями

$$U = \{u \in E^n: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}; g_i(u) = 0, i = \overline{m+1, s}\}. \quad (19)$$

Если это множество не имеет внутренних точек, то реализация ряда методов минимизации (например, методов проекции градиента, возможных направлений, условного градиента, покоординатного спуска и др.) на U может стать затруднительной или даже невозможной. В то же время при применении методов множителей Лагранжа или штрафных функций к задаче (1), (19) могут получиться такие последовательности $\{u_k\}$, которые не принадлежат множеству U и нарушают какие-либо из ограничений $g_i(u) \leq 0, g_i(u) = 0$ на недопустимо большую величину. В таких случаях может оказаться целесообразным использование метода барьерных функций.

Заметим, что этот метод выше был 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 [140].

Введем последовательность расширенных множеств

$$V_k = \{u \in U_0: g_i(u) \leq \theta_k, i = \overline{1, m}; |g_i(u)| \leq \theta_k, i = \overline{m+1, s}\}, \quad (20)$$

где $\theta_k > 0, k = 1, 2, \dots, \lim_{k \rightarrow \infty} \theta_k = 0$. Так как $U \subset V_k, k = 1, 2, \dots$, то из $U \neq \emptyset$ следует $V_k \neq \emptyset, k = 1, 2, \dots$. Предполагая, что функция $J(u)$ определена на множестве $\bigcup_{k=1}^{\infty} V_k$, рассмотрим последовательность задач

$$J(u) \rightarrow \inf; u \in V_k, k = 1, 2, \dots \quad (21)$$

Для решения задач (21) могут быть использованы различные методы минимизации. Мы здесь остановимся лишь на методе барьерных функций. Обозначим

$V_k = \{u \in V_k \text{ и выполняется хотя бы одно из равенств}$

$$g_i(u) = \theta_k, i = \overline{1, m}; g_j(u) = \theta_k, g_j(u) = -\theta_k, j = \overline{m+1, s}\}. \quad (22)$$

Поскольку $U \subset V_k$, $U \cap \gamma_k = \emptyset$, то $U \subset V_k \setminus \gamma_k \neq \emptyset$, $k = 1, 2, \dots$. 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 $t > 0$, $\lim_{t \rightarrow +0} \varphi_i(t) = +\infty$, $i = \overline{1, s}$. Например, в качестве $\varphi_i(t)$ можно взять $\varphi_i(t) = t^{-1}$, $\varphi(t) = (\max\{-\ln t; 0\})^p$, $p \geq 1$.

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: $a_k > 0$, $k = 1, 2, \dots$, $\{a_k\} \rightarrow 0$. 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:

$$\lim_{k \rightarrow \infty} a_k \varphi_i(\theta_k) = 0, \quad i = \overline{1, s}. \quad (25)$$

Предположим, что $J_{k*} = \inf_{V_k} J(u) > -\infty$, $k = 1, 2, \dots$. Так как $B_k(u) \geq 0$, $a_k > 0$, то $F_k(u) \geq J(u)$ при всех $u \in V_k \setminus \gamma_k$, и поэтому $F_{k*} = \inf_{V_k \setminus \gamma_k} F_k(u) \geq J_{k*} > -\infty$, $k = 1, 2, \dots$. С помощью какого-либо метода минимизации определим точку u_k , удовлетворяющую условиям

$$u_k \in V_k \setminus \gamma_k, \quad F_{k*} \leq F_k(u_k) \leq F_{k*} + \varepsilon_k, \quad k = 1, 2, \dots, \quad (26)$$

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,

$$\lim_{k \rightarrow \infty} J_{k*} = J_* > -\infty, \quad J_{k*} = \inf_{V_k} J(u), \quad J_* = \inf_U J(u). \quad (27)$$

Тогда для последовательности $\{u_k\}$, определяемой условиями (26), справедливости соотношения

$$\lim_{k \rightarrow \infty} F_{k*} = \lim_{k \rightarrow \infty} F_k(u_k) = \lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} a_k B_k(u_k) = 0. \quad (28)$$

Если, кроме того, множество

$$U(\delta) = \{u \in U_0: g_i(u) \leq \delta, \quad i = \overline{1, m}; \quad |g_i(u)| \leq \delta, \quad i = \overline{m+1, s}\} \quad (29)$$

компактно при некотором $\delta > 0$, множество U_0 замкнуто, а функции $J(u)$, $g_1(u)$, $\dots$, $g_m(u)$, $|g_{m+1}(u)|$, $\dots$, $|g_s(u)|$ полунепрерывны снизу на $U(\delta)$, то $\{u_k\} \rightarrow U_*$ — множество решений задачи (1), (19).

Доказательство. Из определения J_{k*} , F_{k*} , неотрицательности $B_k(u)$ и условий (26) имеем

$$-\infty < J_{k*} \leq J(u_k) \leq F_k(u_k) \leq F_{k*} + \varepsilon_k \leq \\ \leq F_k(u) + \varepsilon_k = J(u) + a_k B_k(u) + \varepsilon_k, \quad u \in V_k \setminus \gamma_k, \quad k = 1, 2, \dots \quad (30)$$

Так как функции $\varphi_i(t)$ из (23) не возрастают при $t > 0$, то $\varphi_i(\theta_k - g_i(u)) \leq \varphi_i(\theta_k)$, $i = \overline{1, m}$; $\varphi_i(\theta_k \pm g_i(u)) = \varphi_i(\theta_k)$, $i = \overline{m+1, s}$, для всех $u \in U$. Поэтому в силу условия (25)

$$0 \leq a_k B_k(u) \leq 2a_k \sum_{i=1}^s \varphi_i(\theta_k) \rightarrow 0 \quad (k \rightarrow \infty), \quad u \in U. \quad (31)$$

Тогда при $k \rightarrow \infty$ из (30) с учетом условия (27) получим

$$J_* \leq \lim_{k \rightarrow \infty} F_{k*} \leq \overline{\lim}_{k \rightarrow \infty} F_{k*} \leq J(u) \quad \text{при всех } u \in U.$$

Переходя к нижней грани по $u \in U$, отсюда имеем $\lim_{k \rightarrow \infty} F_{k*} = J_*$. Тогда из (30) следует $\lim_{k \rightarrow \infty} F_k(u_k) = \lim_{k \rightarrow \infty} J(u_k) = J_*$. Наконец, $0 \leq a_k B_k(u_k) = F_k(u_k) - J(u_k) \rightarrow 0$ при $k \rightarrow \infty$. Равенства (28) доказаны.

Пусть теперь выполнены все условия теоремы. Так как $\{\theta_k\} \rightarrow 0$, то $V_k \subset U(\delta)$ при всех $k \geq k_0$. А тогда $u_k \in U(\delta)$, $k \geq k_0$. В силу компактности $U(\delta)$ последовательность $\{u_k\}$ имеет хотя бы одну предельную точку. Пусть u_* — произвольная предельная точка $\{u_k\}$, пусть подпоследовательность $\{u_{k_r}\} \rightarrow u_*$. В силу замкнутости U_0 тогда $u_* \in U_0$. Далее, из полунепрерывности снизу функций $g_1(u), \dots, g_m(u), |g_{m+1}(u)|, \dots, |g_s(u)|$ и условия $u_k \in V_k$ следует, что

$$g_i(u_*) \leq \lim_{r \rightarrow \infty} g_i(u_{k_r}) \leq \lim_{k \rightarrow \infty} \theta_k = 0, \quad i = \overline{1, m},$$

$$|g_i(u_*)| \leq \lim_{r \rightarrow \infty} |g_i(u_{k_r})| \leq \lim_{k \rightarrow \infty} \theta_k = 0 \quad \text{или} \quad g_i(u_*) = 0, \quad i = \overline{m+1, s}.$$

Таким образом, $u_* \in U$. Отсюда с учетом полунепрерывности снизу $J(u)$ на $U(\delta)$ получим $J_* \leq J(u_*) \leq \lim_{r \rightarrow \infty} J(u_{k_r}) = \lim_{k \rightarrow \infty} J(u_k) = J_*$, т. е. $J(u_*) = J_*$ или $u_* \in U_*$. Тем самым показано, что любая предельная точка последовательности $\{u_k\}$ принадлежит U_* . Отсюда следует, что $\{u_k\} \rightarrow U_*$. Теорема 4 доказана.

При некоторых более жестких ограничениях на данные задачи (1), (19) можно получить оценки погрешности метода (20)–(26).

Теорема 5. Пусть функция Лагранжа $L(u, \lambda) = J(u) + \sum_{i=1}^s \lambda_i g_i(u)$, $u \in U_0$, $\lambda \in \Lambda_0 = \{\lambda = (\lambda_1, \dots, \lambda_s) \in E^s: \lambda_1 \geq 0, \dots, \lambda_m \geq 0\}$, задачи (1), (19) имеет седловую точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$, т. е. $L(u_*, \lambda) \leq L(u_*, \lambda^*) = J_* \leq L(u, \lambda^*)$, $u \in U_0$, $\lambda \in \Lambda_0$. Тогда последовательность $\{u_k\}$, определяемая

условиями (20)–(26), существует и справедливы оценки

$$-|\lambda^*|_1 \theta_k \leq J(u_k) - J_* \leq F_k(u_k) - J_* \leq 2a_k \sum_{i=1}^s \varphi_i(\theta_k) + \varepsilon_k, \quad (32)$$

$$0 \leq a_k B_k(u_k) \leq 2a_k \sum_{i=1}^s \varphi_i(\theta_k) + \varepsilon_k + |\lambda^*|_1 \theta_k, \quad k = 1, 2, \dots, \quad (33)$$

где $|\lambda^*|_1 = \sum_{i=1}^s |\lambda_i^*|$. Если, кроме того, множество (29) компактно при некотором $\delta > 0$, U_0 замкнуто, а функции $J(u)$, $g_1(u)$, ..., $g_m(u)$, $|g_{m+1}(u)|$, ..., $|g_s(u)|$ полунепрерывны снизу на $U(\delta)$, то $\{u_k\} \rightarrow U_*$.

Доказательство. Из существования седловой точки следует, что

$$-\infty < J_* \leq J(u) + \sum_{i=1}^s \lambda_i^* \gamma_i(u) \leq J(u) + |\lambda^*|_1 \theta_k \leq F_k(u) + |\lambda^*|_1 \theta_k \quad (34)$$

при всех $u \in V_k \setminus \gamma_k$. Отсюда имеем $F_k(u) \geq J_* - |\lambda^*|_1 \theta_k > -\infty$, $u \in V_k \setminus \gamma_k$, или $F_{k*} \geq J_* - |\lambda^*|_1 \theta_k > -\infty$, $k = 1, 2, \dots$. Таким образом, последовательность $\{u_k\}$, удовлетворяющая условиям (26), существует. Далее из (31) следует

$$0 \leq a_k B_k(u_*) \leq 2a_k \sum_{i=1}^s \varphi_i(\theta_k), \quad u_* \in U_* \subset U \subset V_k \setminus \gamma_k, \quad k = 1, 2, \dots$$

Поэтому с учетом неравенств (26), (34) имеем

$$J_* \leq J(u_k) + |\lambda^*|_1 \theta_k \leq F_k(u_k) + |\lambda^*|_1 \theta_k \leq F_{k*} + \varepsilon_k + |\lambda^*|_1 \theta_k \leq F_k(u_*) + \varepsilon_k + |\lambda^*|_1 \theta_k \leq J_* + 2a_k \sum_{i=1}^s \varphi_i(\theta_k) + \varepsilon_k + |\lambda^*|_1 \theta_k, \quad k = 1, 2, \dots$$

Отсюда получаем оценку (32).

Далее, из соотношений $0 \leq a_k B_k(u_k) = (F_k(u_k) - J_*) - (J(u_k) - J_*)$ и уже доказанной оценки (32) вытекает оценка (33). 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 $J(u) = e^{-u}$, $U = \{u \in E^1 = U_0: g(u) = (u^2 - 1)(1 + u^4)^{-1} \leq 0\}$. Ясно, что $U = \{u \in E^1: |u| \leq 1\}$, $J_* = \inf U J(u) = e^{-1}$. Возьмем $V_k = \{u \in E^1: g(u) \leq \theta_k = 1/k^2\}$. Так как $u_r = r \in V_k$ при $r \geq k$, то $\lim_{r \rightarrow \infty} J(u_r) = 0 = J_{k*}$, $k = 1, 2, \dots$. Таким образом, здесь $\lim_{k \rightarrow \infty} J_{k*} = 0 < e^{-1} = J_*$ — условие (27) не выполняется. Заметим, что в рассмотренном примере множество $U(\delta) = \{u \in E^1: g(u) \leq \delta\}$ 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 $J(u)$, $g_1(u)$, ..., $g_m(u)$, $|g_{m+1}(u)|$, ..., $|g_s(u)|$ определены и полунепрерывны снизу

на U_0 . Кроме того, пусть множество

$$U(C) = \{u \in E^n: u \in U_0, g_i(u) \leq C, i = \overline{1, m}; |g_i(u)| \leq C, i = \overline{m+1, s}\} \quad (35)$$

непусто, а множество $U(C + \varepsilon_0)$ ограничено и замкнуто при некотором $\varepsilon_0 > 0$. Тогда (см. обозначения (15))

$$\lim_{\varepsilon \rightarrow 0} J_*(C + \varepsilon) = J_*(C + 0) = J_*(C). \quad (36)$$

Доказательство. Так как $U(C) \subseteq U(C + \delta) \subseteq U(C + \varepsilon)$ при любых $0 < \delta < \varepsilon \leq \varepsilon_0$, то $J_*(C + \varepsilon) \leq J_*(C + \delta) \leq J_*(C)$. Таким образом, функция $J_*(C)$ переменной C не возрастает и существует предел $\lim_{\varepsilon \rightarrow +0} J_*(C + \varepsilon) = J_*(C + 0) \leq J_*(C)$. Возьмем произвольную последовательность $\{\varepsilon_k\}$, $0 < \varepsilon_k \leq \varepsilon_0$, сходящуюся к нулю. При сделанных предположениях множества $U(C + \varepsilon_k) \subseteq U(C + \varepsilon_0)$ при каждом $k = 1, 2, \dots$ ограничены и замкнуты. Согласно теореме 2.1.1 тогда существует точка $w_k \in U(C + \varepsilon_k)$ такая, что $J(w_k) = J_*(C + \varepsilon_k)$, $k = 1, 2, \dots$. 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 $\{w_k\} \rightarrow w_*$. По построению $w_k \in U(C + \varepsilon_k)$, т. е. $w_k \in U_0$, $g_i(w_k) \leq C + \varepsilon_k$, $i = \overline{1, m}$; $|g_i(w_k)| \leq C + \varepsilon_k$, $i = \overline{m+1, s}$. Используя замкнутость множества U_0 , полунепрерывность рассматриваемых функций, отсюда при $k \rightarrow \infty$ получаем $w_* \in U(C)$. А тогда $J_*(C) \leq J(w_*) \leq \lim_{k \rightarrow \infty} J(w_k) = \lim_{k \rightarrow \infty} J_*(C + \varepsilon_k) = J_*(C + 0)$. Сравнивая с ранее установленным неравенством $J_*(C + 0) \leq J_*(C)$, получаем равенство (36).

Нетрудно видеть, что при $C = 0$ из (36) вытекает условие (27).

Заметим, что в теореме 5 условие (27) не упоминается. Однако, как показывает следующая теорема, оно по существу присутствует и в теореме 5.

Теорема 7. Пусть в задаче

$$J(u) \rightarrow \inf; \quad u \in U(C), \quad (37)$$

где $U(C)$ определено условиями (35), функция Лагранжа

$$\begin{aligned} L(u, \lambda, \mu, v) = & J(u) + \sum_{i=1}^m \lambda_i (g_i(u) - C) + \sum_{i=m+1}^s \mu_i (g_i(u) - C) + \\ & + \sum_{i=m+1}^s v_i (-g_i(u) - C), \quad u \in U_0, \quad \lambda = (\lambda_1, \dots, \lambda_m) \geq 0, \\ & \mu = (\mu_{m+1}, \dots, \mu_s) \geq 0, \quad v = (v_{m+1}, \dots, v_s) \geq 0, \end{aligned}$$

имеет седловую точку $(u_*, \lambda^*, \mu^*, v^*)$, $u_* \in U_0$, $\lambda^* \geq 0$, $\mu^* \geq 0$, $v^* \geq 0$, т. е.

$$L(u_*, \lambda, \mu, v) \leq L(u_*, \lambda^*, \mu^*, v^*) = J_*(C) \leq L(u, \lambda^*, \mu^*, v^*)$$

при всех $u \in U_0$, $\lambda \geq 0$, $\mu \geq 0$, $v \geq 0$. Тогда имеет место (36).

Доказательство. По определению седловой точки $J_*(C) \leq J(u) +$
 $+ \sum_{i=1}^m \lambda_i^* (g_i(u) - C) + \sum_{i=m+1}^s \mu_i^* (g_i(u) - C) + \sum_{i=m+1}^s v_i^* (-g_i(u) - C)$ при

всех $u \in U_0$. Отсюда для $u \in U(C + \varepsilon) \subset U_0$ имеем

$$J_*(C) \leq J(u) + \varepsilon A, \quad A = \sum_{i=1}^m \lambda_i^* + \sum_{i=m+1}^s (\mu_i^* + \nu_i^*) \geq 0, \quad \varepsilon > 0.$$

Следовательно, $J_*(C) \leq J_*(C + \varepsilon) + \varepsilon A$. В то же время из $U(C) \subseteq U(C + \varepsilon)$ вытекает $J_*(C + \varepsilon) \leq J_*(C)$. Из последних двух неравенств при $\varepsilon \rightarrow +0$ получим $J_*(C + 0) = J_*(C)$, что и требовалось.

Заметим, что при $C = 0$ задача (37) превращается в задачу (1), (19), функция Лагранжа $L(u, \lambda, \mu, \nu)$ задачи (37) — в функцию Лагранжа $L(u, \lambda)$ задачи (1), (19) с $\lambda_i = \mu_i + \nu_i$, $i = \overline{m+1, s}$. Из теоремы 7 тогда следует, что при выполнении условий теоремы 5 будем иметь $J_*(+0) = J_*(0) = J_*$, что равносильно условию (27).

Объединяя теоремы 2 и 6, а также теоремы 3 и 7, можно получить достаточные условия непрерывности функции $J_*(C)$ переменной C , представляющей собой минимальное значение функции $J(u)$ на множестве (14).

6. Метод барьерных функций на практике иногда используют в сочетании с методом штрафных функций. Предположим, что множество U задается в виде

$$U = \{u \in U_0; g_i(u) \leq 0, \quad i = \overline{1, m}; \quad g_i(u) = 0, \quad i = \overline{m+1, s}; \\ h_j(u) \leq 0, \quad j = \overline{1, l}; \quad h_j(u) = 0, \quad j = \overline{l+1, r}\}, \quad (38)$$

где U_0 — заданное множество из E^n , функции $g_i(u)$, $h_j(u)$, а также минимизируемая функция $J(u)$ определены на U_0 . Ограничения, задаваемые функциями $g_i(u)$, будем учитывать с помощью штрафных функций

$$P(u) = \sum_{i=1}^m (\max\{g_i(u); 0\})^p + \sum_{i=m+1}^s |g_i(u)|^p, \quad u \in U_0, \quad p \geq 1.$$

Введем множества

$$W_k = \{u \in U_0; h_j(u) \leq \theta_k, \quad j = \overline{1, l}; \quad |h_j(u)| \leq \theta_k, \quad j = \overline{l+1, r}\},$$

$\gamma_k = \{u \in W_k \text{ и выполняется хотя бы одно из равенств}$

$$h_j(u) = \theta_k, \quad j = \overline{1, r}; \quad h_j(u) = -\theta_k, \quad j = \overline{l+1, r}\},$$

$k = 1, 2, \dots$ В качестве барьерной функции подмножества γ_k возьмем

$$B_k(u) = \sum_{j=1}^r \varphi_j(\theta_k - h_j(u)) + \sum_{j=l+1}^r \varphi_j(\theta_k + h_j(u)), \quad u \in W_k \setminus \gamma_k,$$

где функция $\varphi_j(t)$ определена, неотрицательна и не возрастает при $t > 0$,

$\lim_{t \rightarrow +0} \varphi_j(t) = +\infty$, $j = \overline{1, r}$. Рассмотрим последовательность задач

$$F_k(u) = J(u) + A_k P(u) + a_k B_k(u) \rightarrow \inf; \quad u \in W_k \setminus \gamma_k.$$

считая, что $\{A_k^{-1}\}$, $\{a_k\}$, $\{\theta_k\}$ — положительные последовательности, сходящиеся к нулю. Пусть $F_{k*} = \inf_{W_k \setminus \gamma_k} F_k(u) > -\infty$, $k = 1, 2, \dots$ Определим

точку u_k из условий

$$u_k \in W_k \setminus \gamma_k, \quad F_k(u_k) \leq F_{k*} + \varepsilon_k, \quad (39)$$

где $e_k > 0$, $k = 1, 2, \dots$, $\lim_{k \rightarrow \infty} e_k = 0$ (если $F_k(u_k) = F_{k*}$, то в (39) допускается $e_k = 0$).

Теорема 8. Пусть: 1) множество U_0 замкнуто, функции $J(u)$, $g_1(u), \dots, g_m(u)$, $|g_{m+1}(u)|, \dots, |g_s(u)|$, $h_1(u), \dots, h_l(u)$, $|h_{l+1}(u)|, \dots, |h_r(u)|$ полунепрерывны снизу на U_0 ; 2) множество

$$W(\delta) = \{u \in U_0: h_j(u) \leq \delta, \quad j = \overline{1, l}; \quad |h_j(u)| \leq \delta, \quad j = \overline{l+1, r}\}$$

ограничено для некоторого $\delta > 0$; 3) последовательности $\{A_k^{-1}\}$, $\{a_k\}$, $\{\theta_k\}$,

$\{e_k\}$ положительны и сходятся к нулю, причем $\lim_{k \rightarrow \infty} a_k \sum_{j=1}^r \varphi_j(\theta_k) = 0$. Тогда последовательность $\{u_k\}$, определяемая условиями (39), такова, что

$$\lim_{k \rightarrow \infty} F_{k*} = \lim_{k \rightarrow \infty} F_k(u_k) = \lim_{k \rightarrow \infty} J(u_k) = J_*, \quad \lim_{k \rightarrow \infty} A_k P(u_k) = 0,$$

$$\lim_{k \rightarrow \infty} a_k B_k(u_k) = 0, \quad \lim_{k \rightarrow \infty} \rho(u_k, U_*) = 0.$$

Доказательство. Так как $U \subset W(\delta)$, то U ограничено. Кроме того, U замкнуто. Согласно теореме 2.2.1 тогда $J_* > -\infty$, $U_* \neq \emptyset$, а также $\inf J(u) = J_{**} > -\infty$. Поскольку $\{\theta_k\} \rightarrow 0$, то можем считать, что $\theta_k \leq \delta$ и $W_k \subset W(\delta)$, $k = 1, 2, \dots$. Так как $F_k(u) \geq J(u) \geq J_{**} > -\infty$ при всех $u \in W_k \setminus \gamma_k$, то $F_{k*} > -\infty$, а последовательность $\{u_k\}$, удовлетворяющая условиям (39), существует. Далее, поскольку $\{u_k\} \in W_k \setminus \gamma_k \subset W(\delta)$, то $\{u_k\}$ имеет хотя бы одну предельную точку. Заметим, что $P(u_*) = 0$, $0 \leq a_k B_k(u_*) \leq$

$$\leq 2a_k \sum_{j=1}^r \varphi_j(\theta_k), \quad u_* \in U_*. \text{ Отсюда и из (39) имеем}$$

$$J_{**} \leq J(u_k) \leq J(u_k) + A_k P(u_k) \leq F_k(u_k) \leq F_{k*} + e_k \leq$$

$$\leq F_k(u_*) + e_k \leq J_* + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + e_k, \quad k = 1, 2, \dots$$

Следовательно, $\overline{\lim}_{k \rightarrow \infty} J(u_k) \leq \overline{\lim}_{k \rightarrow \infty} F_k(u_k) = \overline{\lim}_{k \rightarrow \infty} F_{k*} \leq J_*$. Тогда $0 \leq A_k P(u_k) \leq \leq F_k(u_k) - J(u_k) \leq \sup_{k \geq 1} F_k(u_k) - J_{**} < \infty$, и поэтому $\lim_{k \rightarrow \infty} P(u_k) = 0$. Далее, рассуждая так же, как при доказательстве теорем 4 и 12.2, получаем, что все предельные точки последовательности $\{u_k\}$ принадлежат U_* , и убеждаемся в справедливости всех утверждений теоремы.

Теорема 9. Пусть функция Лагранжа задачи (1), (38)

$$L(u, \lambda, \mu) = J(u) + \sum_{i=1}^s \lambda_i g_i(u) + \sum_{j=1}^r \mu_j h_j(u), \quad u \in U_0, \quad (\lambda, \mu) \in \Lambda_0,$$

$\Lambda_0 = \{(\lambda, \mu) = (\lambda_1, \dots, \lambda_s, \mu_1, \dots, \mu_r): \lambda_1 \geq 0, \dots, \lambda_m \geq 0, \mu_1 \geq 0, \dots, \mu_l \geq 0\}$ имеет седловую точку $(u_*, \lambda^*, \mu^*) \in U_0 \times \Lambda_0$, т. е.

$$L(u_*, \lambda, \mu) \leq L(u_*, \lambda^*, \mu^*) = J_* \leq L(u, \lambda^*, \mu^*), \quad u \in U_0, \quad (\lambda, \mu) \in \Lambda_0.$$

Пусть последовательности $\{a_k\}$, $\{A_k^{-1}\}$, $\{\theta_k\}$, $\{e_k\}$ положительны, сходятся к нулю и $\lim_{k \rightarrow \infty} a_k \sum_{j=1}^r \varphi_j(\theta_k) = 0$. Тогда последовательность $\{u_k\}$, определяемая

условиями (39), существует и такова, что

$$-|\mu^*|_1 \theta_k - |\lambda^*|_q (P(u_k))^{1/p} \leq J(u_k) - J_* \leq F_k(u_k) - J_* \leq 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + \varepsilon_k, \quad (40)$$

$$0 \leq a_k B_k(u_k) \leq |\lambda^*|_q (P(u_k))^{1/p} + |\mu^*|_1 \theta_k + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + \varepsilon_k, \quad (41)$$

где $p \geq 1$, $p^{-1} + q^{-1} = 1$ при $p > 1$ и $q = \infty$ при $p = 1$; $|\lambda^*|_q = \left(\sum_{i=1}^s |\lambda_i^*|^q \right)^{1/q}$

при $1 < q < \infty$, $|\lambda^*|_\infty = \max_{1 \leq i \leq s} |\lambda_i^*|$; $|\mu^*|_1 = \sum_{j=1}^r |\mu_j^*|$, а $P(u_k)$ оценивается так:

$$0 \leq A_k P(u_k) \leq \frac{\varepsilon_k + 2a_k \sum_{j=1}^r \varphi_j(\theta_k)}{1 - |\lambda^*|_\infty A_k^{-1}} \quad \text{при } p = 1, \quad A_k > |\lambda^*|_\infty, \quad (42)$$

$$0 \leq A_k P(u_k) \leq |\lambda^*|_q^q A_k^{1-q} + q \left(\varepsilon_k + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) \right) \quad \text{при } p > 1. \quad (43)$$

Если, кроме того, выполнено условие 1) теоремы 8, то любая предельная точка последовательности $\{u_k\}$ является решением задачи (1), (38).

Доказательство аналогично доказательству теорем 5 и 12.5, поэтому мы наметим лишь схему рассуждений. Из существования седловой точки функции Лагранжа следует, что при всех $u \in W_k \setminus \gamma_k$

$$J_* \leq F_k(u) + |\mu^*|_1 \theta_k + |\lambda^*|_q^q q^{-1} p^{q-1} A_k^{1-q}$$

(при $p = 1$ последнее слагаемое отсутствует). Поэтому $F_{k*} > -\infty$, и последовательность $\{u_k\}$, определяемая условиями (39), существует. Далее, рассуждая так же, как при выводе оценки (12.26), устанавливаем

$$J_* \leq J(u_k) + |\lambda^*|_q (P(u_k))^{1/p} + |\mu^*|_1 \theta_k, \quad p \geq 1, \quad k = 1, 2, \dots \quad (44)$$

Тогда, учитывая, что $P(u_*) = 0$, $a_k B_k(u_*) \leq 2a_k \sum_{j=1}^r \varphi_j(\theta_k)$ при всех $u_* \in U_*$ и условия (39), получим

$$\begin{aligned} J(u_k) &\leq J(u_k) + A_k P(u_k) \leq F_k(u_k) \leq F_{k*} + \varepsilon_k \leq F_k(u_*) + \varepsilon_k \leq \\ &\leq J_* + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + \varepsilon_k \leq J(u_k) + |\lambda^*|_q (P(u_k))^{1/p} + \\ &+ |\mu^*|_1 \theta_k + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + \varepsilon_k, \quad k = 1, 2, \dots \end{aligned} \quad (45)$$

Отсюда следует, что

$$A_k P(u_k) \leq |\lambda^*|_q (P(u_k))^{1/p} + |\mu^*|_1 \theta_k + 2a_k \sum_{j=1}^r \varphi_j(\theta_k) + \varepsilon_k, \quad k = 1, 2, \dots \quad (46)$$

Так же, как из неравенства (12.28) были выведены оценки (12.21) и (12.23), из (46) получаем оценки (42), (43). Далее, левое неравенство (40) следует из (44), а правое — из (45). Наконец, из $0 \leq a_k B_k(u_k) \leq (F_k(u_k) - J_*) - (J(u_k) - J_*)$ и оценки (40) имеем оценки (41). Последнее утверждение доказывается так же, как аналогичное утверждение в теоремах 4 и 12.2.

Различные аспекты метода барьерных функций исследованы в [140, 229].

У п р а ж н е н и я. 1. Применить метод барьерных функций к задачам

а) $J(u) = x + y \rightarrow \inf; \quad u \in U = \{u = (x, y) \in E^2: g_1(u) = x^2 - y \leq 0, g_2(u) = -x \leq 0\};$

б) $J(u) = y \rightarrow \inf; \quad u \in U = \{u = (x, y) \in E^2: g(u) = \sin x + x - y \leq 0\};$

в) $J(u) = (u - 1)^3 \rightarrow \inf; \quad u \in U = \{u \in E^1: g(u) = -u - 1 \leq 0\};$

г) к задачам из упражнения 12.1;

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 12.8 для описанных выше вариантов метода барьерных функций.

§ 14. Метод нагруженных функций

1. Будем рассматривать задачу

$$J(u) \rightarrow \inf; \quad U = \{u \in E^n: u \in U_0,$$

$$g_i(u) \leq 0, \quad i = \overline{1, m}; \quad g_i(u) = 0, \quad i = \overline{m+1, s}\}. \quad (1)$$

Для решения таких задач могут быть использованы описанные выше методы множителей Лагранжа, штрафных функций. Однако метод множителей Лагранжа применим к сравнительно узкому классу задач, у которых функция Лагранжа имеет седловую точку. Применение метода штрафных функций к задаче (1), как отмечалось выше, может быть затруднено из-за наличия в этом методе неограниченно возрастающих штрафных коэффициентов.

Перейдем к описанию метода нагруженных функций, в котором, как и в упомянутых методах, 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 множителей Лагранжа. Введем семейство функций

$$\Phi(u, t) = L | \max \{J(u) - t; 0\} |^{p_0} + MP(u), \quad u \in U_0, \quad (2)$$

зависящее от скалярного параметра t , $-\infty < t < +\infty$, где

$$P(u) = \sum_{i=1}^m (\max \{g_i(u); 0\})^{p_i} + \sum_{i=m+1}^s |g_i(u)|^{p_i}, \quad (3)$$

величины $p_i \geq 1$, $i = \overline{1, s}$, $L > 0$, $M > 0$ фиксированы и являются параметрами метода. Положим

$$\rho(t) = \inf_{u \in U_0} \Phi(u, t). \quad (4)$$

Поскольку $\Phi(u, t) \geq 0$ при всех t и $u \in U_0$, то $\rho(t) \geq 0$ при любом t . Предположим, что в задаче (1) $J_* > -\infty$, $U_* \neq \emptyset$. 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

$$\rho(t) = 0. \quad (5)$$

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. [2, 39, 129]. 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

$$\Phi(u, t) = L|J(u) - t|^{p_0} + MP(u), \quad u \in U_0, \quad -\infty < t < \infty, \quad (6)$$

где функция $P(u)$ взята из (3), $L > 0$, $M > 0$, $p_i \geq 1$. 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 $L = M = p_0 = \dots = p_s = 1$.

Пример 1. Пусть $J(u) = -u$; $U = \{u \in E^1: g(u) = u \leq 0\}$. 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 $t \geq 0$, то $\Phi(0, t) = 0 = \inf_{E^1} \Phi(u, t) = \rho(t)$. Если же $t < 0$, то $\Phi(u, t) = u$ при $u \geq -t$, $\Phi(u, t) = -t$ при $0 \leq u \leq -t$; $\Phi(u, t) = -u - t$ при $u \leq 0$ (нарисуйте график функции $\Phi(u, t)$).

при различных t). Поэтому $\inf_{E^1} \Phi(u, t) = \rho(t) = -t$ 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 $t \geq 0$, то, взяв $u = -t$, получим $\Phi(-t; t) = 0 = \rho(t)$. Если же $t < 0$, то $\Phi(u, t) = 2u + t$ при $u \geq -t$; $\Phi(u, t) = -t$ при $0 \leq u \leq -t$; $\Phi(u, t) = -u - t$ при $u \leq 0$, 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 $J(u) = u$, $U = \{u \in E^1: g(u) = u^2 - 1 \leq 0\}$. Ясно, что здесь $U = \{u \in E^1: -1 \leq u \leq 1\}$, $J_* = -1$, $u_* = -1$. 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 $\rho(t) = \inf_{E^1} \Phi(u, t) = \max\{-t - 1; 0\}$.

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

$$\rho(t) = \begin{cases} 0 & \text{при } t \geq 0, \\ t^2 & \text{при } -1/2 \leq t < 0, \\ -t - 1/4 & \text{при } t < -1/2. \end{cases}$$

Если взять функцию (6), то $\Phi(u, t) = |u - t| + u^2$, $u \in E^1$, и

$$\rho(t) = \begin{cases} t^2 & \text{при } |t| \leq 1/2, \\ |t| - 1/4 & \text{при } |t| > 1/2. \end{cases}$$

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 $J(u) = u$, $U = \{u \in E^1: g(u) = (u^2 - 1)(u^4 + 1)^{-1} \leq 0\}$. Здесь $U = \{u \in E^1: -1 \leq u \leq 1\}$

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 $t \geq -1$ получим $\Phi(-1, t) = 0 = \inf_{E^1} \Phi(u, t) = \rho(t)$. При $t < -1$, взяв $u_k = -k \leq t$, также будем иметь $\lim_{k \rightarrow \infty} \Phi(-k, t) = \lim_{k \rightarrow \infty} g(-k) = 0 = \inf_{E^1} \Phi(u, t) = \rho(t)$. Таким образом, в рассматриваемом случае $\rho(t) \equiv 0$ при всех t . Если в качестве минимального 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 $|t| \leq 1$, то при $u = t$ получим $\Phi(t, t) = 0 = \rho(t)$. Пусть $|t| > 1$. Введем множества $A_1 = \left\{ u \in E^1: |u - t| \leq \frac{|t| - 1}{2} \right\}$, $A_2 = \left\{ u \in E^2: |u - t| > \frac{|t| - 1}{2} \right\}$. Так как $A_1 \cup A_2 = E^1$, $A_1 \cap A_2 = \emptyset$, то $\rho(t) = \inf_{E^1} \Phi(u, t) = \min\left\{ \inf_{A_1} \Phi(u, t); \inf_{A_2} \Phi(u, t) \right\} \geq \min\left\{ \min_{A_1} g(u); (|t| - 1)/2 \right\} > 0$ при всех t , $|t| > 1$. На первый взгляд создается впечатление, что здесь $J_* = -1$ — минимальный корень уравнения (5). Однако $0 < \rho(t) \leq \Phi(t, t) = g(t)$ при $|t| > 1$ и $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{t \rightarrow -\infty} g(t) = 0$. 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 $\lim_{t \rightarrow -\infty} \rho(t) > 0$. Если же $\lim_{t \rightarrow -\infty} \rho(t) = 0$, то примем $t_* = -\infty$. Если $\rho(t) > 0$ при всех $t > 0$ и $\lim_{t \rightarrow -\infty} \rho(t) > 0$, 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

$$J(u) = \begin{cases} u, & u > -2, \\ -k^2, & -(k+1) < u \leq -k, \quad k = 2, 3, \dots; \end{cases}$$

$$g(u) = \begin{cases} u^2 - 1, & |u| \leq 2; \\ 6/|u|, & |u| > 2, \end{cases}$$

$U = \{u \in E^1 = U_0: g(u) \leq 0\}$. 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 $\rho(-k^2 - k) \geq k$, $k = 2, 3, \dots$. В самом деле, если $u \geq -2$, то $\Phi(u, -k^2 - k) \geq |u + k^2 + k| = k^2 + k + u \geq k^2 \geq k$ при всех $k = 2, 3, \dots$. Если $-(i+1) < u \leq -i$, $2 \leq i \leq k$, то $\Phi(u, -k^2 - k) \geq |-i^2 + k^2 + k| = k^2 - i^2 + k \geq k$, а если $-(i+1) < u \leq -i$, $i \geq k+1$, то $\Phi(u, -k^2 - k) \geq |-i^2 + k^2 + k| = i^2 - (k+1)^2 + k + 1 \geq k + 1 > k$. Таким образом, $\Phi(u, -k^2 - k) \geq k$ для всех $u \in E^1$, поэтому $\rho(-k^2 - k) \geq k$, $k = 2, 3, \dots$. Следовательно $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{k \rightarrow \infty} \rho(-k^2 - k) = +\infty$.

С другой стороны, $0 \leq \rho(-k^2) \leq \Phi(-k, -k^2) = g(-k) = 6k^{-1}$, $k = 2, 3, \dots$, так что $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{k \rightarrow \infty} \rho(-k^2) = 0$. 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 $t \geq -1$, то $\Phi(-1, t) = 0 = \rho(t)$. Если же $t < -1$, то при $k > \sqrt{-t}$ получим $\Phi(-k, t) = \max\{-k^2 - t; 0\} + g(-k) = g(-k) \rightarrow 0 = \rho(t)$. Таким образом, здесь $\rho(t) \equiv 0$ при всех t и согласно определению 1 $t_* = -\infty$.

Пример 6. Переопределим функцию $J(u)$ из предыдущего примера в точках $u = k^{-1}$ так: $J(k^{-1}) = -k^2$, $k = 1, 2, \dots$. Функцию $g(u)$ и множество U оставим такими же, как в примере 5. Повторив прежние рассуждения, нетрудно убедиться, что минимальный корень уравнения (5) здесь будет равным $t_* = -\infty$ 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 не пусто. 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 $J(u) \equiv 1$, $U = \{u \in E^1: g(u) = e^{-u^2} \leq 0\}$. Здесь $U_0 = E^1$, $U = \emptyset$. 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 $J(u) = u$, $U = \{u \in E^1: g(u) = e^{-u^2} \leq 0\}$. Здесь $J_0 = E^1$, $U = \emptyset$. Согласно (2) $\Phi(u, t) = \max\{u - t; 0\} + e^{-u^2}$. Так как при $u = -k \leq t$ $\Phi(-k, t) = e^{-k^2} \rightarrow 0$ при $k \rightarrow \infty$, то $\rho(t) \equiv 0$ при всех t и $t_* = -\infty$. В случае функции (6) $\Phi(u, t) = |u - t| + e^{-u^2} = \min \left\{ \inf_{|u-t| \leq 1} \Phi(u, t); \inf_{|u-t| > 1} \Phi(u, t) \right\} \geq \min \left\{ \inf_{|u-t| \leq 1} e^{-u^2}; 1 \right\} = c(t) > 0$ при всех $u \in E^1$, поэтому $\rho(t) > 0$ при всех t . Но $0 < \rho(t) < \Phi(t, t) \rightarrow 0$ при $t \rightarrow +\infty$ или $t \rightarrow -\infty$, так что $\lim_{t \rightarrow +\infty} \rho(t) = \lim_{t \rightarrow -\infty} \rho(t) = 0$ и $t_* = -\infty$.

Если же здесь взять $g(u) = e^{-u^2} + 1$, то $\Phi(u, t) \geq 1$, $u \in E^1$, и $\rho(t) \geq 1$ при всех t , и поэтому $t_* = +\infty$.

3. Примеры 7, 8 подсказывают, что для того, чтобы единообразно охватить возможные, когда в задаче (1) $U = \emptyset$, целесообразно принять

$$J_* = \begin{cases} \inf_U J(u), & \text{если } U \neq \emptyset, \\ +\infty, & \text{если } U = \emptyset, \quad U_0 \neq \emptyset. \end{cases} \quad (7)$$

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 (7). Тогда $t_* \leq J_*$.

Доказательство. Если $U = \emptyset$, то $J_* = +\infty$ и утверждение теоремы тривиально. Поэтому пусть $U \neq \emptyset$. Так как мы условились рассматривать функции, принимающие лишь конечные значения в области своего определения, то $J_* \leq +\infty$. По определению J_* существует последовательность $\{u_k\} \subset U$ такая, что $\lim_{k \rightarrow \infty} J(u_k) = J_* \geq -\infty$. Если $J_* > -\infty$, то $\lim_{k \rightarrow \infty} \Phi(u_k, J_*) = 0 = \rho(J_*)$ и поэтому $t_* \leq J_*$. Если же $J_* = -\infty$, то, взяв $t_k = J(u_k)$, получим $\rho(t_k) = \Phi(u_k, t_k) = 0$, $k = 1, 2, \dots$. Поскольку $\{t_k\} \rightarrow -\infty$, то отсюда следует $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{k \rightarrow \infty} \rho(t_k) = 0$, 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 § 12 было введено следующее понятие согласованности исходных данных задачи (1) (см. определение 12.2): ограничения

$$g_i(u) \leq 0, \quad i = \overline{1, m}; \quad g_i(u) = 0, \quad i = \overline{m+1, s}; \quad u \in U_0, \quad (8)$$

согласованы с функцией $J(u)$ на множестве U_0 , если для любой последовательности $\{u_k\} \in U_0$, которая удовлетворяет условиям

$$\lim_{k \rightarrow \infty} g_i(u_k) \leq 0, \quad i = \overline{1, m}; \quad \lim_{k \rightarrow \infty} g_i(u_k) = 0, \quad i = \overline{m+1, s}, \quad (9)$$

имеет место соотношение

$$\lim_{k \rightarrow \infty} J(u_k) \geq J_*. \quad (10)$$

Распространим это понятие на случай, когда $U_0 \neq \emptyset$, $U = \emptyset$ и согласно (7) $J_* = +\infty$. Здесь следует различать две возможности: $\inf_{U_0} P(u) = 0$ и $\inf_{U_0} P(u) > 0$. Если $\inf_{U_0} P(u) = 0$, то существует хотя бы одна последовательность $\{u_k\} \in U_0$, удовлетворяющая условиям (9), — в этом случае скажем, что ограничения (8) согласованы с функцией $J(u)$ на U_0 , если для любой последовательности $\{u_k\} \in U_0$, для которой справедливы соотношения (9), имеет место равенство $\lim_{k \rightarrow \infty} J(u_k) = +\infty = J_*$. Кстати, это же равенство получается и из (10) при $J_* = +\infty$. Наконец, если $U_0 \neq \emptyset$, $U = \emptyset$, $\inf_{U_0} P(u) > 0$, то по определению будем считать, что ограничения (8) согласованы с функцией $J(u)$ на множестве U_0 .

Оказывается, введенное понятие согласованности исходных данных 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 ограничения (8) были согласованы с функцией $J(u)$ на множестве U_0 .

Доказательство. Необходимость. Пусть $t_* = J_*$. Если $J_* = -\infty$, то ограничения (8) корректны, так как $\lim_{k \rightarrow \infty} J(u_k) \geq -\infty = J_*$ 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 $\lim_{k \rightarrow \infty} P(u_k) = 0$. Отсюда и из нера-

венств $\Phi(u_k, t) \geq \rho(t) > 0$, справедливых для всех $t < t_* = J_*$ и $k = 1, 2, \dots$, при $k \rightarrow \infty$ получим

$$\lim_{k \rightarrow \infty} \max \{J(u_k) - t; 0\} \geq \rho(t) > 0, \quad t < t_*, \quad (11)$$

в случае использования функции (2) и

$$\lim_{k \rightarrow \infty} |J(u_k) - t| \geq \rho(t) > 0, \quad t < t_*, \quad (12)$$

в случае использования функции (6).

Покажем, что из (11), (12) следует неравенство $\lim_{k \rightarrow \infty} J(u_k) \geq t_* = J_*$. В самом деле, при выполнении (11) для каждого $t < t_*$ найдется номер $k_0 = k_0(t)$ такой, что $\max \{J(u_k) - t; 0\} \geq \rho(t)/2 > 0$ или $J(u_k) - t \geq \rho(t)/2 > 0$ для всех $k \geq k_0$. Тогда $\lim_{k \rightarrow \infty} J(u_k) \geq t$ при любом $t < t_*$. Устремляя $t \rightarrow t_* - 0$, отсюда получим неравенство $\lim_{k \rightarrow \infty} J(u_k) \geq t_* = J_*$.

Рассмотрим случай (12). Пусть $\lim_{k \rightarrow \infty} J(u_k) = \lim_{r \rightarrow \infty} J(u_{k_r}) = a$. Имеются две возможности: либо $a \geq t_*$, либо $a < t_*$. Если $a \geq t_*$, то требуемое неравенство $\lim_{k \rightarrow \infty} J(u_k) \geq t_* = J_*$ 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

$$\lim_{k \rightarrow \infty} |J(u_k) - a| = \lim_{r \rightarrow \infty} |J(u_{k_r}) - a| = 0,$$

что противоречит условию (12). Таким образом, если $a < t_*$, то $a = -\infty$, т. е. $\lim_{k \rightarrow \infty} J(u_k) = \lim_{r \rightarrow \infty} J(u_{k_r}) = -\infty$. Тогда, взяв $t_r = J(u_{k_r})$, $r = 1, 2, \dots$, получим $0 < \rho(t_r) \leq \Phi(u_{k_r}, J(u_{k_r})) = MP(u_{k_r}) \rightarrow 0$ при $r \rightarrow \infty$. Это значит, что $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{r \rightarrow \infty} \rho(t_r) = 0$ и согласно определению 1 $t_* = -\infty$. Но по условию $t_* = J_*$, поэтому $\lim_{k \rightarrow \infty} J(u_k) = J_* = t_* = -\infty$, дело свелось к ранее рассмотренному случаю.

Тем самым установлено, что для любой последовательности $\{u_k\} \in U_0$, удовлетворяющей условиям (9), справедливо неравенство (10), т. е. ограничения (8) согласованы с $J(u)$ на U_0 . Наконец, если такой последовательности $\{u_k\}$ не существует, т. е. $\inf_{U_0} P(u) > 0$, то ограничения (8) согласованы с $J(u)$ на U_0 по определению. Необходимость доказана.

Достаточность. Пусть ограничения (8) согласованы с функцией $J(u)$ на U_0 . Покажем, что тогда $t_* = J_*$. Сначала рассмотрим случай, когда $t_* = -\infty$. Это значит, что $\lim_{t \rightarrow -\infty} \rho(t) = \lim_{k \rightarrow \infty} \rho(t_k) = 0$, где $\{t_k\} \rightarrow -\infty$. По определению $\rho(t_k)$ согласно формуле (4) следует существование точки $u_k \in U_0$ такой, что $\rho(t_k) \leq \Phi(u_k, t_k) \leq \rho(t_k) + 1/k$, $k = 1, 2, \dots$. Отсюда при $k \rightarrow \infty$ имеем $\lim_{k \rightarrow \infty} \Phi(u_k, t_k) = 0$. Это означает, что $\lim_{k \rightarrow \infty} P(u_k) = 0$ и $\lim_{k \rightarrow \infty} \max \{J(u_k) - t_k; 0\} = 0$ в случае использования функции (2) и $\lim_{k \rightarrow \infty} |J(u_k) - t_k| = 0$ — в случае функции (6). Но по построению $\{t_k\} \rightarrow -\infty$, поэтому последние два равенства возможны только при $\lim_{k \rightarrow \infty} J(u_k) = -\infty = t_*$. С другой стороны, из $\{P(u_k)\} \rightarrow 0$ и формулы (3) следует выполнение условий (9). В силу согласованности ограничений (8) с функцией $J(u)$ на U_0 тогда $\lim_{k \rightarrow \infty} J(u_k) \geq J_*$. Следовательно, $J_* = t_* = -\infty$.

Пусть теперь $t_* > -\infty$. В силу теоремы 1 тогда $J_* \geq t_* > -\infty$. Возьмем произвольное $t < J_*$. По определению $\rho(t)$ существует последовательность $\{u_k\} \in U_0$ такая, что $\lim_{k \rightarrow \infty} \Phi(u_k, t) = \rho(t)$. Может случиться, что $\lim_{k \rightarrow \infty} P(u_k) = d > 0$. Тогда из $\Phi(u_k, t) \geq MP(u_k)$ при $k \rightarrow \infty$ следует, что $\rho(t) \geq M \lim_{k \rightarrow \infty} P(u_k) = Md > 0$. Если же $\lim_{k \rightarrow \infty} P(u_k) = 0 = \lim_{r \rightarrow \infty} P(u_{k_r})$, то $\lim_{r \rightarrow \infty} g_i(u_{k_r}) \leq 0$, $i = \overline{1, m}$, $\lim_{r \rightarrow \infty} g_i(u_{k_r}) = 0$, $i = \overline{m+1, s}$. В силу согласованности ограничений (8) с функцией $J(u)$ на множестве U_0 отсюда имеем $\lim_{r \rightarrow \infty} J(u_{k_r}) \geq J_* > t$. А тогда $\rho(t) = \lim_{r \rightarrow \infty} \Phi(u_{k_r}, t) \geq L(J_* - t)^{p_0} > 0$ 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 $1 \lim_{t \rightarrow -\infty} \rho(t) > 0$. Следовательно, $t_* \geq J_*$, что в силу теоремы 1 возможно только при $t_* = J_*$. Теорема доказана.

В § 12 были приведены следующие два достаточных условия, гарантирующие согласованность ограничений (8) с функцией $J(u)$ на U_0 :

1) функция Лагранжа (12.19) задачи (1) имеет седловую точку в смысле неравенств (12.20); или

2) U_0 — замкнутое множество из E^n , функции $J(u)$, $g_1(u)$, $\dots$, $g_m(u)$, $|g_{m+1}(u)|$, $\dots$, $|g_s(u)|$ полунепрерывны снизу на

U_0 , а множество

$$U(\delta) = \{u \in U_0: g_i(u) \leq \delta, \quad i = \overline{1, m}; \quad |g_i(u)| \leq \delta, \quad i = \overline{m+1, s}\}$$

ограничено для некоторого числа $\delta > 0$.

Заметим, что в примере 1 функция Лагранжа имеет седловую точку, но множество $U(\delta)$ неограничено при всех $\delta > 0$, в примере 3 множество $U(\delta)$ ограничено при любом $\delta > 0$, но функция Лагранжа не имеет седловой точки; в примере 2 выполнены оба приведенных достаточных условия согласованности исходных данных.

4. Подробнее остановимся на частном случае функции (2), когда

$$\Phi(u, t) = L \max \{J(u) - t; 0\} + MP(u), \quad u \in U_0, \quad (13)$$

где $L > 0$, $M > 0$, а функция $P(u)$ взята из (3) при некоторых $p_i \geq 1$, $i = \overline{1, s}$. 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), (4), монотонно убывают (вообще говоря, не строго) при возрастании t и удовлетворяют неравенствам*

$$|\Phi(u, t) - \Phi(u, \tau)| \leq L |t - \tau|, \quad (14)$$

$$|\rho(t) - \rho(\tau)| \leq L |t - \tau| \quad (15)$$

при всех $u \in U_0$ и любых t, τ . Если $J_{**} = \inf_{U_0} J(u) > -\infty$, то

$$\Phi(u, t) = -Lt + LJ(u) + MP(u), \quad \rho(t) = -Lt + \inf_{U_0} (LJ(u) + MP(u)) \quad (16)$$

при всех $t \leq J_{**}$ — линейные функции по t .

Доказательство. Простым перебором возможных значений функции $\max \{a; b\}$ легко доказываются неравенства

$$\max \{J(u) - t; 0\} \geq \max \{J(u) - \tau; 0\}, \quad t \leq \tau, \quad u \in U_0,$$

$$|\max \{J(u) - t; 0\} - \max \{J(u) - \tau; 0\}| \leq |t - \tau|, \quad u \in U_0.$$

Отсюда следуют невозрастание функции $\Phi(u, t)$ по переменной t и неравенство (14). Далее, для любых $t \leq \tau$ имеем $\Phi(u, t) \geq \Phi(u, \tau) \geq \rho(\tau)$ или $\Phi(u, t) \geq \rho(\tau)$ при каждом $u \in U_0$. Отсюда, переходя к нижней грани по $u \in U_0$, получим $\rho(t) \geq \rho(\tau)$ при всех $t \leq \tau$.

Докажем неравенство (15). Зафиксируем произвольные t, τ . По определению нижней грани при каждом $\varepsilon > 0$ существуют точки $u_\tau, u_\tau \in U_0$ такие, что

$$\rho(t) \leq \Phi(u_\tau, t) \leq \rho(t) + \varepsilon, \quad \rho(\tau) \leq \Phi(u_\tau, \tau) \leq \rho(\tau) + \varepsilon.$$

Тогда, учитывая уже доказанное неравенство (14), имеем $\rho(t) - \rho(\tau) \leq \Phi(u_\tau, t) - \Phi(u_\tau, \tau) + \varepsilon \leq L |t - \tau| + \varepsilon$, $\rho(t) - \rho(\tau) \geq \Phi(u_\tau, t) - \varepsilon - \Phi(u_\tau, \tau) \geq -L |t - \tau| - \varepsilon$, т. е. $|\rho(t) - \rho(\tau)| \leq L |t - \tau| + \varepsilon$ при любом $\varepsilon > 0$. Отсюда при $\varepsilon \rightarrow +0$ получим неравенство (15). Формулы (16) следуют из того, что $J(u) - t \geq J_{**} - t \geq 0$ при всех $u \in U_0$. Теорема 3 доказана.

Если ограничения (8) согласованы с функцией $J(u)$ на U_0 и $J_* > -\infty$, то опираясь на теорему 3, можно предложить следующий итерационный метод определения J_* . Сначала выберем t_0 так, чтобы $\rho(t_0) > 0$ (например, если $J_* = \inf_{U_0} J(u) > -\infty$, то можно взять любую точку $t_0 \leq J_*$). Следующие приближения определим по формулам [210]

$$t_{k+1} = t_k + \rho(t_k)/L, \quad k = 0, 1, \dots \quad (17)$$

Теорема 4. Пусть функция $\rho(t) \geq 0$ 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 $\rho(t) \geq 0$, то из (17) следует, что последовательность $\{t_k\}$ монотонно возрастает, и поэтому существует $\lim_{k \rightarrow \infty} t_k = a \leq +\infty$. Покажем, что $a = t_*$. По условию $t_0 < t_*$. Допустим, что при некотором $k \geq 0$ оказалось $t_k < t_*$. Тогда $\rho(t) > 0$ при всех $t \leq t_k$. Возьмем произвольное t , $t_k \leq t < t_{k+1}$. С учетом условий (15), (17) имеем

$$\begin{aligned} \rho(t) &= \rho(t_k) + [\rho(t) - \rho(t_k)] \geq \rho(t_k) - L(t - t_k) > \\ &> \rho(t_k) - L(t_{k+1} - t_k) = 0, \quad t_k \leq t < t_{k+1} \end{aligned}$$

Это значит, что $\rho(t) > 0$ при всех $t < t_{k+1}$, т. е. $t_{k+1} \leq t_*$. Может случиться, что $\rho(t_{k+1}) = 0$. Тогда $t_{k+1} = t_*$ — в этом случае итерации (17) заканчиваются. Если $\rho(t_{k+1}) > 0$, то $t_{k+1} < t_*$ и итерации продолжают дальше.

Таким образом, имеются две возможности. Либо процесс (17) закончится тем, что $\rho(t_0) > 0, \dots, \rho(t_{k-1}) > 0, \rho(t_k) = 0$, — 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 $k = 0, 1, \dots$ — в этом случае $\lim_{k \rightarrow \infty} t_k = a \leq t_*$ и $\rho(t) > 0$ при всех $t < a$.

Покажем, что $a = t_*$. Если последовательность $\{t_k\}$ неограничена сверху, то $a = +\infty = t_*$. Если же $t_k < a < \infty, k = 0, 1, \dots$, то, учитывая непрерывность функции $\rho(t)$, из (17) при $k \rightarrow \infty$ 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 вычислять одно значение функции $\rho(t)$, и для этого в свою очередь нужно решить задачу минимизации

$$\Phi(u, t) \rightarrow \inf; \quad u \in U_0. \quad (18)$$

Поскольку функция (13), вообще говоря, не является гладкой, то это обстоятельство может вызвать некоторые трудности при решении задачи (18). Однако успехи в разработке методов решения негладких задач минимизации (см. §§ 3, 9, 10, 15) 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 [57].

Предположим, что вместо точных значений функции $\rho(t)$ известны лишь некоторые приближения $\rho_v(t), v = 1, 2, \dots$, удовлетворяющие условиям

$$\rho_v(t) \geq 0, \quad |\rho_v(t) - \rho(t)| \leq \gamma_v, \quad v = 1, 2, \dots; \quad \lim_{v \rightarrow \infty} \gamma_v = 0. \quad (19)$$

Пусть t_0 — начальное приближение, $t_0 < t_*$. Пусть $(v-1)$ -е приближение t_{v-1} при некотором $v \geq 1$ уже известно. Для определения следующего приближения t_v рассмотрим итерационный процесс

$$t_{vk+1} = t_{vk} + \rho_v(t_{vk})/L, \quad k=0, 1, \dots; \quad t_{v0} = t_0, \quad (20)$$

аналогичный процессу (17). Поскольку функция $\rho_v(t)$ может не обращаться в нуль ни в одной точке даже в том случае, когда уравнение (5) имеет конечный минимальный корень (так будет, например, если $\rho_v(t) = \rho(t) + \gamma_v \geq \gamma_v > 0$), то процесс (20) следует прекращать не по критерию $\rho_v(t_{vk}) = 0$, как было выше в (17), а по условию вида $\rho_v(t_{vk}) < \theta_v$, где величина $\theta_v > 0$ стремится к нулю при $v \rightarrow \infty$ и как-то согласована с погрешностью γ_v . Предположим, что такая последовательность $\{\theta_v\}$ уже задана (условия согласования $\{\theta_v\}$ и $\{\gamma_v\}$ будут обсуждаться ниже). Тогда имеются две возможности:

1) либо найдется номер $k = k_v \geq 0$ такой, что

$$\rho_v(t_{vk}) > \theta_v, \quad k = \overline{0, k_v - 1}; \quad \rho_v(t_{vk_v}) \leq \theta_v; \quad (21)$$

в этом случае процесс (20) заканчиваем и полагаем

$$t_v = t_{vk_v}; \quad (22)$$

2) либо

$$\rho_v(t_{vk}) > \theta_v \quad \text{при всех } k=0, 1, \dots \quad (23)$$

Тогда, как будет показано ниже, при выполнении условий (15), (19) и согласованном изменении величин θ_v и γ_v , будет справедливо равенство

$$t_* = +\infty. \quad (24)$$

Метод поиска минимального корня уравнения (5) при условиях (15), (19) описан.

Теорема 5. Пусть функция $\rho(t)$ неотрицательна при всех t , не возрастает, удовлетворяет условию (15), а $t_* > -\infty$ — минимальный корень уравнения (5) в смысле определения 1. Кроме того, пусть функция $\{\rho_v(t)\}$ удовлетворяет условиям (19) и

$$\theta_v \geq \gamma_v, \quad v=1, 2, \dots \quad (25)$$

Тогда последовательность $\{t_v\}$, 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 $v \geq 1$ будут заканчиваться за конечное число шагов выполнением условий (21); случай (23) возможен лишь при $t_* = +\infty$.

Доказательство. Сначала рассмотрим случай $t_* < +\infty$. Тогда при каждом фиксированном $v \geq 1$ имеются две возможности:

1) $t_{vk} \leq t_* < +\infty$ при всех $k=0, 1, \dots$. В силу монотонности $\{t_{vk}\}$ тогда существует $\lim_{k \rightarrow \infty} t_{vk} \leq t_* < +\infty$. Переходя в (20) к пределу при $k \rightarrow \infty$, получим $\lim_{k \rightarrow \infty} \rho_v(t_{vk}) = 0$. Это значит, что за конечное число итераций процесс (20) закончится выполнением условий (21).

2) Найдется номер $l \geq 0$ такой, что $t_{vl} \leq t_* < t_{vl+1}$. Тогда с учетом соотношений (15), (19), (20) получим

$$\begin{aligned} t_{vl+1} &= t_{vl} + \rho_v(t_{vl})/L = t_{vl} + [\rho_v(t_{vl}) - \rho(t_{vl})]/L + \\ &+ [\rho(t_{vl}) - \rho(t_*)]/L \leq t_{vl} + \gamma_v/L + t_* - t_{vl} = t_* + \gamma_v/L. \end{aligned} \quad (26)$$

Далее, в силу монотонности $\rho(t)$ имеем $\rho(t) \equiv 0$ при $t \geq t_*$, поэтому $\rho(t_{vI+1}) = 0$. Отсюда и из условий (19), (25) следует

$$\rho_v(t_{vI+1}) = \rho_v(t_{vI+1}) - \rho(t_{vI+1}) \leq \gamma_v \leq \theta_v. \quad (27)$$

Это значит, что условия (21) выполняются при некотором $k_v \leq l + 1$.

Объединяя обе рассмотренные возможности, заключаем, что при $t_* < +\infty$ процесс (20) при каждом $v \geq 1$ заканчивается за конечное число шагов k_v выполнением условий (21), причем в силу (26)

$$t_v = t_{vk_v} \leq t_* + \gamma_v L^{-1}, \quad v = 1, 2, \dots \quad (28)$$

Покажем, что $\lim_{v \rightarrow \infty} t_v = t_*$. Из (28) имеем $\overline{\lim}_{v \rightarrow \infty} t_v \leq t_*$. Пусть $\lim_{v \rightarrow \infty} t_v = \lim_{r \rightarrow \infty} t_{v_r} = a$. Тогда с учетом условий (15), (19), (21) получим

$$\begin{aligned} 0 \leq \rho(a) &\leq [\rho(a) - \rho(t_{v_r})] + [\rho(t_{v_r}) - \rho_{v_r}(t_{v_r})] + \rho_{v_r}(t_{v_r}) \leq \\ &\leq L |a - t_{v_r}| + \gamma_{v_r} + \theta_{v_r} \rightarrow 0 \end{aligned}$$

при $r \rightarrow \infty$, т. е. $\rho(a) = 0$. Но t_* — минимальный корень уравнения (5), поэтому $a \geq t_*$. Следовательно, $\lim_{v \rightarrow \infty} t_v = \lim_{r \rightarrow \infty} t_{v_r} \geq t_* \geq \overline{\lim}_{v \rightarrow \infty} t_v$. Это значит, что $\lim_{v \rightarrow \infty} t_v = t_*$. Случай $t_* < +\infty$ полностью рассмотрен.

Пусть теперь $t_* = +\infty$ и пусть процесс (20) при каждом $v \geq 1$ заканчивается выполнением условий (21). Покажем, что тогда $\{t_v\} \rightarrow +\infty$. Возьмем произвольное число $T > 0$. Согласно определению 1, если $t_* = +\infty$, то $\lim_{t \rightarrow -\infty} \rho(t) > 0$ и $\rho(t) > 0$ при всех t . Отсюда и из непрерывности функции $\rho(t)$ следует, что $\inf_{t \leq T} \rho(t) = \rho_T > 0$. Так как $\{\theta_v\} \rightarrow 0$, $\{\gamma_v\} \rightarrow 0$, то найдется номер $v_0 = v_0(T)$ такой, что $\theta_v + \gamma_v < \rho_T$ при всех $v \geq v_0$. Тогда $\rho_v(t) \geq \rho(t) - \gamma_v \geq \rho_T - \gamma_v > \theta_v$ для всех $t \leq T$ и $v \geq v_0$. Это значит, что условия (21) не могут выполняться при $t_{vk} \leq T$, если $v \geq v_0$. Тогда согласно (21), (22) $t_v > T$ для всех $v \geq v_0 = v_0(T)$, что означает выполнение равенства $\lim_{v \rightarrow \infty} t_v = +\infty$.

Остается рассмотреть случай, когда при некотором $v \geq 1$ выполняются условия (23). Выше было установлено, что при $t_* < +\infty$ процесс (20) при всех $v \geq 1$ закончится выполнением условий (21). Следовательно, если при каком-либо $v \geq 1$ 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 $J(u) = u \rightarrow \inf$; $u \in U = U_0 = \{u \in E_1: u \geq 0\}$ — это частный случай задачи (1), когда $g_i(u) \equiv 0$ (или $s = 0$). Тогда $\Phi(u, t) = \max\{u - t, 0\}$, $u \geq 0$, и $\rho(t) = \inf_{u \geq 0} \Phi(u, t) = \max\{-t, 0\}$ (ср. с примером 1). Здесь $t_* = J_* = 0$, $u_* = 0$. Если $\rho_v(t) = \max\{-t, 0\} + \gamma_v$ и $\theta_v < \gamma_v$, то $\rho_v(t) \geq \gamma_v > \theta_v$ 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 $t_0 \leq t < +\infty$ нам придется работать на каком-то отрезке $t_0 \leq t \leq T$, где величина T ограничена, например, разрядной сеткой ЭВМ. В этом случае метод (20) — (24) требует

модификации. А именно, итерационный процесс (20) при каждом $v \geq 1$ здесь будет заканчиваться определением номера $k_v \geq 0$ такого, что будет выполнено одно из двух следующих условий:

$$\rho_v(t_{vk_v}) > \theta_v, \quad k = \overline{0, k_v - 1}; \quad \rho_v(t_{vk_v}) \leq \theta_v, \quad t_{vk_v} \leq T, \quad (29)$$

или

$$\rho_v(t_{vk_v}) > \theta_v, \quad k = \overline{0, k_v - 1}; \quad t_{vk_v-1} \leq T < t_{vk_v}. \quad (30)$$

В качестве v -го приближения t_v будем брать

$$t_v = \min \{t_{vk_v}; T\}, \quad v = 1, 2, \dots \quad (31)$$

Если выполнены все условия теоремы 5, то, немного видоизменив доказательство этой теоремы, нетрудно установить, что при $t_0 < t_* < T$ для достаточно больших номеров $v > v_0$ процесс (20) будет заканчиваться выполнением условий (29) и оценки (28), а последовательность $\{t_v\}$, 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

$$\Phi(u, t) = L|J(u) - t| + MP(u), \quad u \in U_0, \quad (32)$$

где $L > 0$, $M > 0$, функция $P(u)$ взята из (3) при некоторых $p_i \geq 1$, $i = \overline{1, s}$.

Поскольку $|J(u) - t| - |J(u) - \tau| \leq |t - \tau|$, 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

$$\theta_v \geq 2\gamma_v, \quad v = 1, 2, \dots \quad (33)$$

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 $\{\rho_v(t)\}$ удовлетворяют условиям (19), а последовательности $\{\theta_v\}$, $\{\gamma_v\}$ — условию (33). Тогда последовательность $\{t_v\}$, 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 $v \geq 1$ 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):

$$\begin{aligned} \rho_v(t_{v, l+1}) &= [\rho_v(t_{v, l+1})] - \rho(t_{v, l+1}) + [\rho(t_{v, l+1}) - \rho(t_*)] \leq \\ &\leq \gamma_v + L(t_{v, l+1} - t_*) \leq 2\gamma_v \leq \theta_v. \end{aligned}$$

Нетрудно также показать, что при выполнении условий теоремы 6 последовательность $\{t_v\}$, полученная методом (20), (29) — (31), 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 $\rho_v(t) = |t| + \gamma_v$, $v = 1, 2, \dots$. Предположим, что условие (33) нарушено, т. е. $\theta_v < 2\gamma_v$. Возьмем начальное приближение $t_0 = t_{v0} < \min \{\gamma_v - \theta_v; 0\}$. Тогда $\rho_v(t_{v0}) = |t_{v0}| + \gamma_v = -t_{v0} + \gamma_v > \theta_v$. Далее, $t_{v1} = t_{v0} + \rho_v(t_{v0}) = \gamma_v > 0$ и снова $\rho_v(t_{v1}) = \rho_v(\gamma_v) = 2\gamma_v > \theta_v$. Отсюда с учетом монотонности $\rho_v(t)$ при $t \geq 0$ имеем $\rho_v(t) \geq \rho_v(t_{v1}) > \theta_v$ для всех $t \geq t_{v1}$. Это значит, что $\rho_v(t_{vk}) > \theta_v$ для всех $k = 0, 1, \dots$ — 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, [57, 125, 200].

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: $J(u) \rightarrow \inf$; $u \in U = \{u \in E^1: u \in U_0, g(u) \leq 0\}$, где

а) $J(u) = \operatorname{arctg} u$, $g(u) = u^2 - 4$, $g(u) = (u^2 - 4)(u^4 + 1)^{-1}$, $g(u) = u$, $g(u) = u(u^2 + 1)^{-1}$, $g(u) = u^2$, $U_0 = \{u \in E^1: u \geq 1\}$, $U_0 = \{u \in E^1: u \geq 0\}$, $U_0 = E^1$;

б) $J(u) = u \sin u$, $g(u) = u^2 - 1$; $g(u) = (u^2 - 1)(u^4 + 1)^{-1}$, $g(u) = (u^2 - 1)e^{-u^2}$; $U_0 = \{u \in E^1: u \geq 0\}$;

в) $J(u)$ — произвольная функция, а $U = U_0 = E^1$;

г) $J(u) = 1$ при $u \leq 1$, $J(u) = u^{-1}$ при $u > 1$; $g(u) = u - 1$, $g(u) = (u - 2)(u^2 + 1)^{-1}$, $g(u) = e^{-u^2}$, $U_0 = \{u \in E^1: u \geq 0\}$ или $U_0 = E^1$.

д) $J(u) \equiv 1$, $g(u) = u$, $g(u) = u^2 + 1$, $g(u) = e^{-u^2}(u^2 + 1)$, $g(u) = u^2 e^{-u^2}$; $U_0 = E^1$, $U_0 = \{u \in E^1: u \geq 0\}$.

2. Найти функцию $\rho(t)$ для задачи $J(u) \rightarrow \inf$; $u \in U = \{u \in E^1: U_0: g(u) = |u| - 1 \leq 0\}$, 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 $\rho_0 > 1$, 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 ограничение $g(u) \leq 0$ было согласовано с функцией $J(u)$ 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) выпуклыми, если исходная задача (1) выпукла.

7. Пусть в задаче (1) $J_* > -\infty$, $U_* \neq \emptyset$, ограничения (8) согласованы с функцией $J(u)$ на множестве U_0 . Пусть последовательность $\{t_k\}$ построена методом (17), а последовательность $\{u_k\}$ определена условиями: $u_k \in U_0$,

$\rho(t_k) = \Phi(u_k, t_k)$, $k = 0, 1, \dots$ Можно ли ожидать, что $\{u_k\} \rightarrow U_*$? Приведите примеры.

8. Пусть функция Лагранжа задачи (1) имеет седловую точку $(u_*, \lambda^*) \in U_0 \times \Lambda_0$ (см. соотношения (12.19), (12.20)). Показать, что та же точка (u_*, λ^*) является седловой точкой функции Лагранжа для задачи

$$\max \{J(u) - t; 0\} \rightarrow \inf; \quad u \in U$$

при всех $t \leq J_*$. Выяснить связь между множествами точек минимума последней задачи и задачи (1).

9. Для задачи (1) ввести функцию

$$\Phi_1(u, t) = L \max \{J(u); t\} + MP(u), \quad u \in U_0,$$

где $P(u)$ взята из (3), $L > 0$, $M > 0$, и положить $\rho_1(t) = \inf_{u \in U_*} \Phi_1(u, t)$. 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 $J(u) \rightarrow \inf; u \in U = \{u \in U_0, g_1(u) \leq 0, \dots, g_m(u) \leq 0\}$ ввести функцию $G(w, u) = \max \{J(w) - J(u); g_1(w), \dots, g_m(w)\}$ или $G(w, u) = (J(w) - J(u)) g_1(w) \dots g_m(w)$ переменных $u, w \in U_0$ и рассмотреть итерационный процесс $G(u_{k+1}, u_k) = \inf_{w \in U_0} G(w, u_k)$; исследовать его сходимость (метод центров, [119]).

§ 15. О методе случайного поиска

Наряду с описанными выше методами минимизации функций n 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:

$$u_{k+1} = u_k + \alpha_k \xi, \quad k = 0, 1, 2, \dots, \quad (1)$$

где α_k — некоторая положительная величина, $\xi = (\xi^1, \dots, \xi^n)$ — какая-либо реализация n -мерной случайной величины ξ с известным законом распределения. Например, координаты ξ^i случайного вектора ξ могут представлять независимые случайные величины, 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 $u_k \in U$, $k \geq 0$, 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 $v_k \notin U$, но $J(v_k) \geq J(u_k)$, или же $v_k \notin U$, то сделанный шаг считается неудачным и полагается $u_{k+1} = u_k$.

Если окажется, что $u_k = u_{k+1} = \dots = u_{k+N}$ 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 $\xi_1, \dots, \xi_s$ случайного вектора ξ и вычисляются значения функции $J(u)$ в тех точках $u = u_k + \alpha \xi_i$, $i = \overline{1, s}$, которые принадлежат множеству U . Затем полагается $u_{k+1} = u_k + \alpha \xi_{i_0}$, где индекс i_0 определяется условием

$$J(u_k + \alpha \xi_{i_0}) = \min_{\substack{u_k + \alpha \xi_i \in U \\ 1 \leq i \leq s}} J(u_k + \alpha \xi_i).$$

Величины $s > 1$ и $\alpha = \text{const} > 0$ являются параметрами алгоритма.

в) *Алгоритм статистического градиента.* Берутся какие-либо s реализаций $\xi_1, \dots, \xi_s$ 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 , $1 \leq i \leq s$, 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 $e_i = (0, \dots, 0, 1, \dots, 0)$, $i = 1, 2, \dots, n$, 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 результа-

ты 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

$$u_{k+1} = u_k + \alpha_k \xi_k, \quad k = 0, 1, 2, \dots, \quad (2)$$

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 $\xi_0 = (\xi_0^1, \dots, \xi_0^n)$, компоненты ξ_i^l , $i = \overline{1, n}$, 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 $w = (w^1, \dots, w^n)$, 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 $\xi = \xi(w) = (\xi^1, \dots, \xi^n)$, 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:

$$p^i = \begin{cases} 0 & \text{при } w^i < -1, \\ \frac{1}{2}(1 + w^i) & \text{при } |w^i| \leq 1, \\ 1 & \text{при } w^i > 1, \quad i = \overline{1, n}. \end{cases} \quad (3)$$

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 $w = w_0 = (0, 0, \dots, 0)$. Приближение u_2 определяется по формуле (2) при $k=1$ с помощью случайного вектора $\xi_1 = \xi(0)$. Пусть известны приближения $u_0, u_1, \dots, u_k$ и значения параметров $w_{k-1} = (w_{k-1}^1, \dots, w_{k-1}^n)$ при некотором $k \geq 1$. Тогда полагаем

$$w_k^i = \beta w_{k-1}^i - \delta \operatorname{sign} [(J(u_{k-1}) - J(u_{k-2}))(u_{k-1}^i - u_{k-2}^i)], \quad (4) \\ i = \overline{1, n}, \quad k = 2, 3, \dots,$$

где величина $\beta \geq 0$ называется *параметром забывания*, $\delta \geq 0$ — *параметром интенсивности обучения*, $\beta + \delta > 0$. При определении следующего приближения u_{k+1} в формуле (2) берем какую-либо реализацию случайного вектора $\xi_k = \xi(w_k)$, $w_k = (w_k^1, \dots, w_k^n)$.

Из формул (3), (4) видно, что если переход от точки u_{k-2} к точке 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 $\delta \geq 0$ в (4) регулирует скорость обучения: чем больше $\delta > 0$, тем быстрее обучается алгоритм; при $\delta = 0$, как видно, обучения нет. Величина $\beta \geq 0$ 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_k^i накладываются ограничения $|w_k^i| \leq c_i$, и при нарушении этих ограничений w_k^i заменяются ближайшим из чисел c_i и $-c_i$, $i = \overline{1, n}$. Величины β, δ, c_i являются параметрами алгоритма.

Вместо формул (4), посредством которых производится обучение алгоритма, часто пользуются другими формулами

$$w_k^i = \beta w_{k-1}^i - \delta (J(u_{k-1}) - J(u_{k-2}))(u_{k-1}^i - u_{k-2}^i), \quad (5) \\ i = \overline{1, n}, \quad k = 2, 3, \dots$$

Описанный алгоритм покоординатного обучения имеет тот недостаток, что поиск и обучение происходят лишь по одному из 2^m направлений $\xi = (\xi^1, \dots, \xi^n)$, 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 $\xi = \xi(w) = \frac{\eta + w}{|\eta + w|}$, где $w = (w^1, \dots, w^n)$ — параметры обучения, $\eta = (\eta^1, \dots, \eta^n)$ — 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 $k \geq 2$ 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 $w_k = (w_k^1, \dots, w_k^n)$ накладываются ограничения $|w_k| \leq c = \text{const}$, и при нарушении этих ограничений w_k заменяется на $\frac{w_k}{|w_k|} c$.

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 направлениям убывания функции. В то же время наличие случайного фактора в выборе направления дает возможность алгоритму «переучиваться», 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 § 1.16 применительно к задачам минимизации функций одной переменной. Для функций n переменных эта процедура приводит к построению последовательности $\{u_k\}$ по закону

$$u_{k+1}^i = u_k^i - a_k \frac{z(u_k + c_k e_i) - z(u_k - c_k e_i)}{c_k}, \quad i = \overline{1, n}, \quad k = 0, 1, \dots$$

где $e_i = (0, \dots, 0, 1, 0, \dots, 0)$, $z(u)$ — наблюдаемые в эксперименте значения функции $J(u)$ в точке u , последовательности $\{a_k\}$, $\{c_k\}$ удовлетворяют условиям (1.16.2).

Заметим, что задача минимизации функций при наличии случайных ошибок относится к задачам стохастического программирования — так называется раздел математического программирования, изучающий теорию и методы решения экстремальных задач при неполной информации об исходных данных. Более подробно о стохастическом программировании, о теоретических и вычислительных аспектах методов случайного поиска, стохастической аппроксимации см., например, в [9, 35, 45, 53, 89, 90, 100, 101, 109, 133, 137, 174, 180, 202, 215, 222, 228, 254].

§ 16. 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 оказываются меньшими, чем при применении другого быстро сходящегося метода, каждая итерация которого весьма трудоемка. Таким образом, при характеристике метода минимизации важным является не столько ско-

рость 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 [236]) и лучшим признают тот метод, с помощью которого удастся 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 специалистом 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 [13]. 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 существования 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 удастся 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 [164, 223, 249].

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

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

то для определения точки $u_0 \in U$ нужно решать систему уравнений (вообще говоря, нелинейных). Чтобы найти какую-либо точку множества

$$U = \{u \in E^n: g_i(u) \leq 0, i = \overline{1, m}\}, \quad (2)$$

придется решать систему неравенств. Определение решения систем линейных или нелинейных уравнений и неравенств представляет собой весьма серьезную задачу, которой посвящены отдельные монографии или большие главы в учебниках по численным методам [2, 39, 129, 182].

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

$$P(u) = \sum_{i=1}^s g_i^2(u), \quad u \in E^n,$$

а в случае множества (2) — функцию

$$P(u) = \sum_{i=1}^m (\max \{g_i(u); 0\})^p, \quad u \in E^n; \quad p \geq 1,$$

и рассмотрим задачу минимизации

$$P(u) \rightarrow \inf; \quad u \in E^n.$$

Для решения этой задачи могут быть использованы любые подходящие методы минимизации. Если $U \neq \emptyset$, то условие $u_0 \in U$ равносильно условию $P(u_0) = 0 = \inf_{E^n} P(u) = P_*$. Если $P_* > 0$,

то $U = \emptyset$. Если же $P_* = 0$, но нижняя грань $P(u)$ на E^n не достигается, то также $U = \emptyset$.

Если

$$U = \{u \in E^n: u \in U_0, g_i(u) \leq 0, i = \overline{1, m}; g_i(u) = 0, i = \overline{m+1, s}\},$$

то для определения какой-либо точки $u_0 \in U$ можно рассмотреть задачу минимизации

$$P(u) = \sum_{i=1}^m (\max \{g_i(u); 0\})^p + \sum_{i=m+1}^s |g_i(u)|^p \rightarrow \inf; \quad u \in U_0; \quad p \geq 1.$$

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

$$\sigma \rightarrow \inf; \quad z = (u, \sigma) \in G =$$

$$= \{z = (u, \sigma) \in E^{n+1}: g_i(u) \leq \sigma, i = \overline{1, m}\}. \quad (3)$$

Понятно, что если $U \neq \emptyset$, то $\inf_{\sigma} \sigma \leq 0$. 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 $\sigma_0 = \max_{1 \leq i \leq m} g_i(u_0)$. 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 $\sigma \leq 0$. В том случае, когда множество (2) регулярно, т. е. существует точка $v \in E^n$, для которой $g_i(v) < 0$, $i=1, m$, то $\inf_{\sigma} \sigma < 0$,

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 [9, 35, 119, 134, 135, 180, 189].

ПРИНЦИП МАКСИМУМА ПОНТРЯГИНА

В этой главе будут рассматриваться задачи оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений. Этот класс экстремальных задач существенно отличается от выше-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 бесконечно-мерному функциональному пространству. Такие задачи имеют многочисленные приложения в механике космического полета, в вопросах управления электроприводами, химическими или ядерными реакторами, виброзащиты и т. д.

Эффективнейшим средством исследования задач оптимального управления является принцип максимума Понтрягина [14], 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, 3, 4, 6, 12, 14, 16, 17, 19, 22, 25, 27, 28, 31, 32, 34, 36, 37, 42, 43, 46, 48, 49, 54, 56, 66, 68—71, 73, 74, 79, 80, 87, 88, 91, 92, 102, 103, 120, 124, 126, 136, 138, 139, 143, 146—153, 156, 158, 159, 162, 163, 167—171, 173, 176—178, 181, 183—186, 189, 190, 194, 195, 198, 203, 204, 206, 209, 212, 217, 220, 221, 225—228, 230—234, 237, 241, 244—248, 252, 254, 257, 258].

§ 1. Постановка задачи оптимального управления

1. Сначала приведем несколько примеров 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), $b \geq 0$ — 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

$$\ddot{\varphi} + \beta \dot{\varphi} + \sin \varphi = u(t), \quad (1)$$

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 эквивалентной системы двух уравнений первого порядка

$$\dot{x}^1(t) = x^2(t), \quad \dot{x}^2(t) = -\beta x^2(t) - \sin x^1(t) + u(t). \quad (2)$$

Пусть в начальный момент времени $t = 0$ маятник отклонен 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, выбором которого можно влиять на движение маятника, — удовлетворяет ограничению

$$|u(t)| \leq \gamma, \quad \gamma = \text{const} > 0, \quad t \geq 0. \quad (3)$$

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

$$x^1(T) = 2\pi k, \quad x^2(T) = 0, \quad (4)$$

при некотором $k = 0, \pm 1, \pm 2, \dots$ (задача быстрогодействия), или

2) за минимальное время T добиться выполнения условия

$$(x^1(T))^2 + (x^2(T))^2 \leq \varepsilon,$$

где $\varepsilon > 0$ — заданное число, или

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 $\max_{0 \leq t \leq T} |x^1(t)|$, или

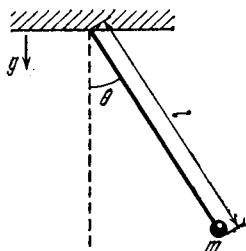


Рис. 6.1.

$\max_{0 \leq t \leq T} \max \{ |x^1(t)|, |x^2(t)| \}$ 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

$$|x^1(t)| \leq \mu, \quad \mu = \text{const} > 0.$$

На управление $u(t)$ вместо условия (3) (или наряду с условиями (3)) могут накладываться ограничения вида

$$\int_0^T u^2(t) dt \leq R,$$

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 ([88], стр. 129), движение центра масс космического аппарата и расход массы описывается системой дифференциальных уравнений

$$\dot{r} = v, \quad \dot{v} = gp/G + F, \quad \dot{G} = -gq, \quad 0 \leq t \leq T, \quad (5)$$

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, $q = q(t)$ — массовый 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:

$$r(0) = r_0, \quad v(0) = v_0, \quad G(0) = G_0. \quad (6)$$

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

$$p_{\min} \leq |p(t)| \leq p_{\max}, \quad q_{\min} \leq q(t) \leq q_{\max}, \quad 0 \leq t \leq T,$$

или $\int_0^T q^2(t) dt \leq R$, $R = \text{const} > 0$. 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, связанные 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 [6, 12, 22, 25, 29, 49, 73, 88, 91, 126, 152, 153, 156, 159, 162, 163, 167, 169, 171, 173, 181, 185, 200, 203, 204, 206, 209, 220, 221, 225, 226, 232, 233, 244—247, 252].

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

$$\dot{x}^i = f^i(x^1, x^2, \dots, x^n, u^1, u^2, \dots, u^r), \quad i = \overline{1, n},$$

которые в векторной форме можно записать в виде

$$\dot{x} = f(x, u, t), \quad (7)$$

где t — время, $x = (x^1, x^2, \dots, x^n)$ — величины, характеризующие движение объекта в зависимости от времени и называемые фазовыми координатами объекта, $u = (u^1, u^2, \dots, u^r)$ — параметры управления («положение рулей» объекта), выбором которых можно влиять на движение объекта, $f = (f^1, f^2, \dots, f^n)$; функции $f^i(x, u, t)$, $i = \overline{1, n}$, 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 $t_0 \leq t \leq T$, необходимо в начальный момент времени t_0 задать начальное условие $x(t_0) = x_0$ и также параметры управления $u = (u^1, u^2, \dots, u^r)$ 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:

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T, \quad (8)$$

$$x(t_0) = x_0. \quad (9)$$

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 $\tau_1, \tau_2, \dots, \tau_p \in [t_0, T]$, в которых функция $u(t)$ может терпеть разрывы типа скачка, т. е. существуют конечные пределы

$$\lim_{t \rightarrow \tau_i - 0} u(t) = u(\tau_i - 0), \quad \lim_{t \rightarrow \tau_i + 0} u(t) = u(\tau_i + 0),$$

но, вообще говоря, $u(\tau_i - 0) \neq u(\tau_i + 0)$, $i = \overline{1, p}$. 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 , $1 \leq p \leq +\infty$ [10]. Через $L_p^r[t_0, T]$ при $1 \leq p < +\infty$ будем обозначать пространство измеримых вектор-функций $u(t) = (u^1(t), \dots, u^r(t))$, $t_0 \leq t \leq T$, для которых функция $|u(t)|^p$ суммируема на $[t_0, T]$ в смысле Лебега, и следовательно, имеет смысл норма

$$\|u\|_{L_p} = \left(\int_{t_0}^T |u(t)|^p dt \right)^{1/p}, \quad 1 \leq p < +\infty.$$

Если $p = +\infty$, то под $L_\infty^r[t_0, T]$ понимается пространство ограниченных измеримых вектор-функций $u(t) = (u^1(t), \dots, u^r(t))$, $t_0 \leq t \leq T$, с нормой

$$\|u\|_{L_\infty} = \text{ess sup}_{t_0 \leq t \leq T} |u(t)| = \inf_{v(t)} \sup_{t_0 \leq t \leq T} |v(t)|,$$

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 $L_p^r[t_0, T]$ [10], 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

$$u(t) \in V(t), \quad t_0 \leq t \leq T, \quad (10)$$

где $V(t)$ — заданное множество из E^r при каждом $t \in [t_0, T]$. Например, в случае ограничений (3) $V(t) = \{u: u \in E^1, |u| \leq \gamma\}$ 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

$$\|u\|_{L_p}^p = \int_{t_0}^T |u(t)|^p dt \leq R^p, \quad R = \text{const} > 0, \quad (11)$$

при некотором p , $1 \leq p < +\infty$.

Таким образом, постановка задачи оптимального управления прежде всего предполагает, что выбран некоторый класс функций — управлений (например, кусочно непрерывные, кусочно

гладкие или функции из $L_p^r[t_0, T]$, $1 \leq p \leq \infty$) и указаны налагаемые на них ограничения (например, ограничения вида (10) или (11)).

Заметим, что в учебной литературе символом $z(t) = (z^1(t), \dots, z^m(t))$ часто обозначают как значение функции в точке t , так и саму функцию, которая представляет собой отображение области определения функции в пространство 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 $\|u(\cdot)\|_{L_p} \leq R$, а обозначение $\|u(t)\|_{L_p} \leq R$, иногда встречающееся в литературе, не вполне удачное.

3. Итак, пусть заданы точка $x_0 \in E^n$ и некоторое кусочно непрерывное управление $u = u(\cdot) = u(t)$, $t_0 \leq t \leq T$, или управление $u(\cdot) \in L_p^r[t_0, T]$ при некотором $p \geq 1$. 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, под решением задачи (8), (9) можно понимать функцию $x = x(\cdot) = x(t)$, $t_0 \leq t \leq T$, 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 существования непрерывно дифференцируемого решения задачи (8), (9), вообще говоря, не имеет смысла. Поэтому мы будем пользоваться следующим более общим определением решения задачи (8), (9).

О п р е д е л е н и е 1. Непрерывную функцию $x = x(\cdot) = x(t)$, $t_0 \leq t \leq T$, удовлетворяющую равенству

$$x(t) = \int_{t_0}^t f(x(\tau), u(\tau), \tau) d\tau + x_0, \quad t_0 \leq t \leq T, \quad (12)$$

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 $x = x(t, u, x_0)$, $t_0 \leq t \leq T$. 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 $x(\cdot, u, x_0)$ буквы u или x_0 будем опускать и просто писать $x(\cdot, u)$ или $x(\cdot, x_0)$ или $x = x(\cdot) = x(t)$, $t_0 \leq t \leq T$.

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) [193]. Однако если $u(t) \in L_p^r[t_0, T]$, $p \geq 1$, 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 $u(t) \in L_p^r[t_0, T]$, $p \geq 1$. 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

$$|f(x, u, t) - f(y, u, t)| \leq L(t)|x - y| \quad (43)$$

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 $x(t) = (x^1(t), \dots, x^n(t))$, $t_0 \leq t \leq T$, с нормой

$$\|x\|_C = \max_{t_0 \leq t \leq T} |x(t)|$$

обозначим через $C^n[t_0, T]$. Как известно [10] $C^n[t_0, T]$ — полное нормированное пространство. Зафиксируем какие-либо точки x_0 и ограниченное измеримое управление $u = u(t)$, $t_0 \leq t \leq T$. Можно показать [1], 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 :

$$z(t) = Ax = \int_{t_0}^t f(x(\tau), u(\tau), \tau) d\tau + x_0, \quad t_0 \leq t \leq T, \quad (14)$$

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 [10]. Для этого с помощью индукции докажем, что для любых $x(\cdot), y(\cdot) \in C^n[t_0, T]$

$$|A^m x(\cdot)(t) - A^m y(\cdot)(t)| \leq \frac{1}{m!} \max_{t_0 \leq \tau \leq t} |x(\tau) - y(\tau)| \left(\int_{t_0}^t L(\tau) d\tau \right)^m \quad (15)$$

при всех t , $t_0 \leq t \leq T$, $m = 1, 2, \dots$. Из (14) с учетом условия (13) имеем

$$\begin{aligned} |Ax(\cdot)(t) - Ay(\cdot)(t)| &= \left| \int_{t_0}^t [f(x(\tau), u(\tau), \tau) - f(y(\tau), u(\tau), \tau)] d\tau \right| \leq \\ &\leq \int_{t_0}^t L(\tau) |x(\tau) - y(\tau)| d\tau \leq \max_{t_0 \leq \tau \leq t} |x(\tau) - y(\tau)| \int_{t_0}^t L(\tau) d\tau. \end{aligned} \quad (16)$$

Оценка (15) при $m = 1$ доказана. Пусть оценка (15) верна для некоторого $m \geq 1$. Тогда с помощью неравенства (16) получим

$$\begin{aligned} |A^{m+1} x(\cdot)(t) - A^{m+1} y(\cdot)(t)| &= |A(A^m x(\cdot))(t) - A(A^m y(\cdot))(t)| \leq \\ &\leq \int_{t_0}^t L(\tau) |A^m x(\cdot)(\tau) - A^m y(\cdot)(\tau)| d\tau \leq \\ &\leq \int_{t_0}^t L(\tau) \frac{1}{m!} \max_{t_0 \leq \xi \leq \tau} |x(\xi) - y(\xi)| \left(\int_{t_0}^{\tau} L(\xi) d\xi \right)^m d\tau \leq \\ &\leq \frac{1}{m!} \max_{t_0 \leq \xi \leq t} |x(\xi) - y(\xi)| \int_{t_0}^t L(\tau) \left(\int_{t_0}^{\tau} L(\xi) d\xi \right)^m d\tau = \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{(m+1)!} \max_{t_0 \leq \tau \leq t} |x(\tau) - y(\tau)| \int_{t_0}^t \frac{d}{d\tau} \left(\int_{t_0}^{\tau} L(\xi) d\xi \right)^{m+1} d\tau = \\
&= \frac{1}{(m+1)!} \max_{t_0 \leq \tau \leq t} |x(\tau) - y(\tau)| \left(\int_{t_0}^t L(\tau) d\tau \right)^{m+1}
\end{aligned}$$

для любых $t \in [t_0, T]$. Оценка (15) доказана. Из этой оценки следует, что

$$\|A^m x(\cdot) - A^m y(\cdot)\|_C \leq \frac{1}{m!} \left(\int_{t_0}^T L(\tau) d\tau \right)^m \|x(\cdot) - y(\cdot)\|_C.$$

Так как $\lim_{m \rightarrow \infty} \frac{1}{m!} \left(\int_{t_0}^T L(\tau) d\tau \right)^m = 0$, то при достаточно большом

m будем иметь $\frac{1}{m!} \left(\int_{t_0}^T L(\tau) d\tau \right)^m < 1$. Таким образом, отобра-

жение A^m является сжимающим. Из принципа сжимающих отображений ([10], стр. 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 ([10], стр. 344) и из (12) следует, что получившаяся функция $x(t)$, $t_0 \leq t \leq T$, 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.

З а м е ч а н и е 1. Вместо условия (13) можно потребовать непрерывность $\frac{\partial j}{\partial x} = \left\{ \frac{\partial j^i(x, u, t)}{\partial x^i} \right\}$, $i, j = \overline{1, n}$, 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.

З а м е ч а н и е 2. Если управление $u = u(t) \in L_p^r[t_0, T]$, $1 \leq p < +\infty$, то теорема 1 и ее доказательство останутся в силе, если, например, дополнительно потребовать

$$|f(x, u, t)| \leq C_0(|x| + |u|^p) + C_1(t) \quad (17)$$

для всех $(x, u, t) \in E^n \times E^r \times [t_0, T]$, где $C_0 = \text{const} \geq 0$, $C_1(t) \geq 0$, $C_1(t) \in L_1[t_0, T]$. Условие (17) нужно для обеспечения включения $f(x(t), u(t), t) \in L_1[t_0, T]$ для любых $x(\cdot) \in$

$\in C^n[t_0, T]$, $u(\cdot) \in L_p^r[t_0, T]$, чтобы отображение (14) имело смысл.

Более тонкие теоремы существования и единственности решения задачи (8), (9) для управлений $u(t) \in L_p^r[t_0, T]$, $1 \leq p \leq +\infty$, можно найти в [1, 54, 74, 163].

Остановимся еще на случае линейной системы, когда вместо (8) имеет место

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad t_0 \leq t \leq T, \quad (18)$$

где $A(t) = \{a_{ij}(t)\}$, $B(t) = \{b_{ij}(t)\}$ — матрицы порядков $n \times n$ и $n \times r$ соответственно, $f(t) = (f^1(t), \dots, f^n(t))$, $t_0 \leq t \leq T$.

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, $1 \leq p \leq +\infty$, 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 $[t_0, T]$, $\dot{x}(t) \in L_1^n[t_0, T]$ 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

$$x(t) = x(t, u(\cdot), x_0) \in G(t), \quad t_0 \leq t \leq T, \quad (19)$$

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 , $t_0 \leq T$, 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

$$t_0 \in \Theta_0, \quad T \in \Theta_I, \quad (20)$$

где Θ_0, Θ_I — заданные множества на числовой оси $\mathbf{R} = \{t: -\infty < t < +\infty\}$ (не исключается возможность, что $\Theta_0 = \mathbf{R}$ или $\Theta_I = \mathbf{R}$).

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, 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 $T \in \Theta_I$ — множество $S_I(T)$, и условия на концах траекторий будем записывать в виде

$$x(t_0) \in S_0(t_0), \quad t_0 \in \Theta_0; \quad x(T) \in S_I(T), \quad T \in \Theta_I. \quad (21)$$

Можем считать, что $S_0(t_0) \subseteq G(t_0)$, $t_0 \in \Theta_0$, так как в противном случае вместо $S_0(t_0)$ можно взять его пересечение с $G(t_0)$; аналогично, можно считать, что $S_I(T) \subseteq G(T)$, $T \in \Theta_I$.

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_I состоит из одной точки T , то *конечный момент называют закрепленным*. Если множество $S_0(t_0)$ [или $S_I(T)$] 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 $S_I(T) = \{x_I\}$, $T \in \Theta_I$], то говорят, что *левый [правый] конец траектории закреплен*. Если $S_0(t_0) = E^n$, $t_0 \in \Theta_0$ [или $S_I(T) = E^n$, $T \in \Theta_I$], 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

$$S_0(t_0) = \{x: x \in G(t_0), \quad h_i(x, t_0) \leq 0, \quad i = \overline{1, m_0}; \\ h_i(x, t_0) = 0, \quad i = \overline{m_0 + 1, s_0}\}, \quad (22)$$

где функции $h_i(x, t)$, $i = \overline{1, s_0}$, определены при $x \in G(t)$, $t \in \Theta_0$. Аналогично, примером множества $S_I(T)$ служит

$$S_I(T) = \{x: x \in G(T), \quad g_i(x, T) \leq 0, \quad i = \overline{1, m_I}; \\ g_i(x, T) = 0, \quad i = \overline{m_I + 1, s_I}\}, \quad (23)$$

где функции $g_i(x, t)$, $i = \overline{1, s_I}$, определены при $x \in G(t)$, $t \in \Theta_I$.

5. Теперь перейдем к непосредственной формулировке задачи оптимального управления. Пусть заданы множества Θ_0 , Θ_I на числовой оси $\mathbb{R}$, $\inf \Theta_0 < \sup \Theta_I$; $V(t) \subseteq E^r$, $G(t) \subseteq E^n$ при всех t , $\inf \Theta_0 \leq t \leq \sup \Theta_I$; $S_0(t_0) \subseteq G(t_0)$, $t_0 \in \Theta_0$; $S_I(T) \subseteq G(T)$, $T \in \Theta_I$. Пусть движение фазовой точки $x = (x^1, \dots, x^n)$ 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 $t_0 \in \Theta_0$, $T \in \Theta_I$, $t_0 \leq T$, $x_0 \in S_0(t_0)$, управление $u = u(\cdot) = (u^1(t), \dots, u^r(t))$ 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), $x(T) \in S_I(T)$. В некоторых случаях вместо кусочно непрерывного управления в допустимых наборах будут рассматриваться управления $u = u(\cdot) \in L_p^r(t_0, T)$, $t_0 \in \Theta_0$, $T \in \Theta_I$, $1 \leq p \leq +\infty$, удовлетворяющие ограничениям (10), (11), — эти случаи будем оговаривать особо.

Будем предполагать, что множество допустимых наборов не пусто.

Пусть на множестве допустимых наборов задана функция (или, как часто говорят, *целевая функция* или *функционал*)

$$J(t_0, T, x_0, u(\cdot), x(\cdot)) = \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T), T), \quad (24)$$

где $f^0(x, u, t)$, $\Phi(x, T)$ — заданные функции при $x \in G(t)$, $u \in V(t)$, $\inf \Theta_0 \leq t \leq \sup \Theta_I$, $T \in \Theta_I$.

Задача оптимального управления заключается в том, чтобы минимизировать или максимизировать функцию (24) на множестве допустимых наборов. Мы ограничимся рассмотрением лишь задач минимизации, так как задача максимизации J всегда может быть сведена к эквивалентной задаче минимизации ($-J$).

Обозначим

$$J_* = \inf J(t_0, T, x_0, u(\cdot), x(\cdot)),$$

где нижняя грань берется по всем допустимым наборам. Допустимый набор $(t_0^*, T^*, x_0^*, u_*(\cdot), x_*(\cdot))$ назовем *решением задачи оптимального управления*, $u_*(\cdot)$ — *оптимальным управлением*, $x_*(\cdot)$ — *оптимальной траекторией*, если $J(t_0^*, T^*, x_0^*, u_*(\cdot), x_*(\cdot)) = J_*$.

Сформулированную задачу оптимального управления мы часто будем записывать в следующем кратком виде:

$$J(t_0, T, x_0, u(\cdot), x(\cdot)) =$$

$$= \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T), T) \rightarrow \inf; \quad (25)$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T; \quad (26)$$

$$x(t) \in G(t), \quad t_0 \leq t \leq T, \quad (27)$$

$$x(t_0) = x_0 \in S_0(t_0), \quad t_0 \in \Theta_0; \quad x(T) \in S_1(T), \quad T \in \Theta_1, \quad (28)$$

$$u(t) \in V(t), \quad t_0 \leq t \leq T, \quad (29)$$

подразумевая (если не оговорено противное), что управление $u = u(\cdot)$ — кусочно непрерывно на отрезке $[t_0, T]$.

Если $f^0 \equiv 1$, $\Phi \equiv 0$, то $J(t_0, T, x_0, u(\cdot), x(\cdot)) \equiv T - t_0$ — в этом случае задачу (25) — (29) называют *задачей быстрогодействия*. Если $f^i(x, u, t) \equiv f^i(x, u)$, $i = \overline{1, n}$, $\Phi(x, T) \equiv \Phi(x)$, множества $S_0(t)$, $S_1(t)$, $V(t)$, $G(t)$ не зависят от времени, то задачу (25) — (29) называют *автономной* или *стационарной*.

Если начальный момент закреплен, т. е. $\Theta_0 = \{t_0\}$, то в формулировке задачи (25) — (29) включение $t_0 \in \Theta_0$ опускают, вместо $J(t_0, T, x_0, u(\cdot), x(\cdot))$ пишут короче: $J(T, x_0, u(\cdot), x(\cdot))$ или $J(T, x_0, u(\cdot))$, вместо $S_0(t_0)$ пишут S_0 . 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 $S_0(t) \equiv E^n$, или $S_1(t) \equiv E^n$, или $V(t) \equiv E^r$, то соответствующие из условий (27), (28), (29) в постановке задачи (25) — (29) также явно не указывают. Если $S_0(t_0) \equiv G(t_0)$ или $S_1(T) \equiv G(T)$, то включения $x(t_0) \in S_0(t_0)$ или соответственно $x(T) \in S_1(T)$ будут учтены в условии (27), поэтому в (28) эти включения можно опустить.

В приложениях встречаются задачи оптимального управления более общего вида, чем задача (25) — (29). Возможны ситуации, когда наряду с ограничениями на управления и фазовые координаты, записанными, так сказать, в разделенном виде (27), (29), имеются более сложные ограничения вида

$$(u(t), x(t)) \in W(t), \quad t_0 \leq t \leq T, \quad (30)$$

где $W(t)$ — заданные множества из $E^r \times E^n$. Ограничения (27), (29), (30) 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

$$J_i(t_0, T, x_0, u(\cdot), x(\cdot)) \leq 0, \quad i = \overline{1, m};$$

$$J_i(t_0, T, x_0, u(\cdot), x(\cdot)) = 0, \quad i = \overline{m+1, s}, \quad (31)$$

где

$$J_i(t_0, T, x_0, u(\cdot), x(\cdot)) = \int_{t_0}^T f_i^0(x(t), u(t), t) dt + \Phi_i(x(T), T),$$

$f_i^0(x, u, t)$, $\Phi_i(x, T)$, $i = \overline{1, s}$, — заданные функции. Ограничения (31) накладываются на функции $u(\cdot)$, $x(\cdot) = x(\cdot, u(\cdot), x_0)$ 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 параметрами, 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

1. Сначала рассмотрим задачу оптимального управления с закрепленными концами и с закрепленным временем. А именно, пусть требуется минимизировать функцию

$$J(u(\cdot)) = \int_{t_0}^T f^0(x(t), u(t), t) dt \quad (1)$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T, \quad (2)$$

$$x(t_0) = x_0, \quad x(T) = x_1, \quad (3)$$

$$u(t) \in V, \quad t_0 \leq t \leq T, \quad (4)$$

где управления $u = u(\cdot)$ предполагаются кусочно непрерывными функциями на $[t_0, T]$: моменты t_0 , T и точки x_0 , x_1 заданы.

Подчеркнем, что в задаче (1) — (4) множество $V \subseteq E^r$ не зависит от времени и фазовые ограничения при $t_0 < t < T$ отсутствуют.

Заметим, что значения управления $u(\cdot)$ в точках разрыва не влияют на решения уравнения (2) (см. определение 1.1), на значение интеграла (1) и, следовательно, на задачу (1) — (4). Поэтому в точках разрыва управления можно доопределять совершенно произвольно, лишь бы не нарушалось условие (4). При формулировке принципа максимума мы будем для определенности считать, что $u(t) = u(t+0) = \lim_{\tau \rightarrow t+0} u(\tau)$ при всех t , $t_0 \leq t < T$, и $u(T) = u(T-0)$.

Будем предполагать, что функции $f^j(x, u, t)$, $j = \overline{0, n}$, имеют частные производные $\frac{\partial f^j}{\partial x^i} = f_{x^i}^j$ по переменным $(x^1, \dots, x^n)$. Обозначим

$$\frac{\partial f}{\partial x} = f_x = \begin{pmatrix} f_{x^1}^1 & \dots & f_{x^1}^n \\ \dots & \dots & \dots \\ f_{x^n}^1 & \dots & f_{x^n}^n \end{pmatrix} = \begin{pmatrix} f_x^1 \\ \dots \\ f_x^n \end{pmatrix}, \quad \frac{\partial f^0}{\partial x} = f_x^0 = (f_{x^1}^0, \dots, f_{x^n}^0).$$

Будем считать, что $f^j(x, u, t)$, $j = \overline{0, n}$, $f_x(x, u, t)$, $f_x^0(x, u, t)$ непрерывны по совокупности аргументов $(x, u, t) \in E^n \times E^r \times [t_0, T]$.

Для формулировки принципа максимума введем вспомогательные переменные $\psi = (\psi_1, \dots, \psi_n) \in E^n$ и постоянную ψ_0 и определим функцию

$$\begin{aligned} H(x, u, t, \psi, \psi_0) &= \\ &= \psi_0 f^0(x, u, t) + \psi_1 f^1(x, u, t) + \dots + \psi_n f^n(x, u, t) = \\ &= \sum_{j=0}^n \psi_j f^j(x, u, t) = \psi_0 f^0(x, u, t) + \langle \psi, f(x, u, t) \rangle, \end{aligned} \quad (5)$$

называемую функцией Гамильтона — Понтрягина.

Пусть $u = u(\cdot)$ — кусочно непрерывное управление, удовлетворяющее условию (4), а $x(\cdot) = x(\cdot, u, x_0)$ — решение уравнения (2), соответствующее этому управлению $u(\cdot)$ и начальному условию x_0 и определенное на всем отрезке $[t_0, T]$. Паре $(u(t), x(t))$, $t_0 \leq t \leq T$, поставим в соответствие следующую систему линейных дифференциальных уравнений относительно переменных $\psi(t) = (\psi_1(t), \dots, \psi_n(t))$:

$$\dot{\psi}_i(t) = - \left. \frac{\partial H(x, u, t, \psi(t), \psi_0)}{\partial x^i} \right|_{\substack{u=u(t) \\ x=x(t)}} = - \sum_{j=0}^n \psi_j(t) \left. \frac{\partial f^j(x, u, t)}{\partial x^i} \right|_{\substack{u=u(t) \\ x=x(t)}}, \quad i = \overline{1, n}, \quad (6)$$

где $\psi_0(t) = \psi_0$ — постоянная. Систему (6) называют сопряженной системой, соответствующей паре $(u(t), x(t, u, x_0))$, $t_0 \leq t \leq T$. Пользуясь введенными выше обозначениями f_x , f_x^0 , систему (6) можно записать в векторной форме:

$$\begin{aligned} \dot{\psi}(t) &= -H_x(x, u, t, \psi(t), \psi_0) \Big|_{\substack{u=u(t) \\ x=x(t)}} = \\ &= -\psi_0 f_x^0(x(t), u(t), t) - (f_x(x(t), u(t), t))^T \psi(t), \quad t_0 \leq t \leq T, \end{aligned} \quad (7)$$

где $(f_x)^T$ — матрица, полученная транспонированием матрицы f_x , а под $\psi(t)$ в (7) надо понимать вектор-столбец с координатами $\psi_j(t)$, $j = \overline{1, n}$.

В частности, если уравнение (2) линейно относительно x, u , т. е.

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad t_0 \leq t \leq T,$$

(см. обозначения в (1.18)), то $H(x, u, t, \psi, \psi_0) = \psi_0 f^0(x, u, t) + \langle \psi, A(t)x + B(t)u + f(t) \rangle$, и сопряженная система (7) будет иметь вид

$$\dot{\psi}(t) = -\psi_0 f_x^0(x(t), u(t), t) - (A(t))^T \psi(t), \quad t_0 \leq t \leq T.$$

Из теоремы 1.2 следует, что если зафиксировать постоянную ψ_0 , момент t_1 , $t_0 \leq t_1 \leq T$, и точку $a \in E^n$, то линейная система (7) будет иметь, и притом единственное, решение $\psi(\cdot) = \psi(\cdot, u, x_0, t_1, a)$, удовлетворяющее условию $\psi(t_1) = a$ и определенное на всем отрезке $[t_0, T]$.

Перейдем к формулировке теоремы, выражающей необходимое условие оптимальности — принцип максимума для задачи (1) — (4).

Теорема 1. Пусть $(u(t), x(t))$, $t_0 \leq t \leq T$, — решение задачи (1) — (4). Тогда необходимо существуют непрерывная вектор-функция $\psi(t)$, $t_0 \leq t \leq T$, и постоянная ψ_0 такие, что

$$1) \quad \psi_0 \leq 0, \quad |\psi_0| + |\psi(t)| \neq 0, \quad t_0 \leq t \leq T; \quad (8)$$

2) $\psi(\cdot)$ является решением сопряженной системы (6), соответствующей рассматриваемому решению $(u(\cdot), x(\cdot))$;

3) при каждом $t \in [t_0, T]$ функция $H(x(t), u, t, \psi(t), \psi_0)$ переменной $u = (u^1, \dots, u^r)$ достигает своей верхней грани на множестве V при $u = u(t)$, т. е.

$$\sup_{u \in V} H(x(t), u, t, \psi(t), \psi_0) = H(x(t), u(t), t, \psi(t), \psi_0), \quad t_0 \leq t \leq T. \quad (9)$$

Теорема 1 сформулирована для случая, когда управление $u(t)$ — кусочно непрерывная функция. Если же управление $u(\cdot)$

является ограниченной измеримой функцией, т. е. $u(\cdot) \in L^\infty [t_0, T]$, то формулировка теоремы 1 полностью сохраняется, но соотношение (9) (как и включение (4)) будет выполняться, вообще говоря, лишь почти всюду на $[t_0, T]$.

Центральное место в теореме 1 занимает условие максимума (9): оказывается, если $u(\cdot)$ — оптимальное управление, а $x(\cdot)$ — оптимальная траектория, то непременно найдутся такая постоянная ψ_0 и такое решение $\psi(t)$ системы (6), что функция $H(x(t), u, t, \psi(t), \psi_0)$ переменной u при всех $t \in [t_0, T]$ будет достигать своего максимума на V именно при $u = u(t)$. Поэтому теорему 1 и нижеследующие аналогичные теоремы, дающие необходимое условие оптимальности в задачах оптимального управления, принято называть принципом максимума. Условие (8) гарантирует, что функция H не обратится в тождественный нуль, и делает условие максимума (9) содержательным.

Однако как практически воспользоваться условием (9)? — ведь функции $x(t)$, $\psi(t)$, $t_0 \leq t \leq T$, и постоянная ψ_0 нам еще неизвестны... Здесь обычно поступают следующим образом. Рассматривают функцию $H(x, u, t, \psi, \psi_0)$ как функцию r переменных $u = (u^1, \dots, u^r) \in V$, считая остальные переменные (x, t, ψ, ψ_0) параметрами, и при каждом фиксированном наборе (x, t, ψ, ψ_0) решают задачу максимизации

$$H(x, u, t, \psi, \psi_0) \rightarrow \sup; \quad u \in V.$$

Отсюда находят функцию

$$u = u(x, t, \psi, \psi_0) \in V, \quad (10)$$

на которой достигается верхняя грань функции $H(x, u, t, \psi, \psi_0)$ по переменной $u \in V$, т. е.

$$H(x, u(x, t, \psi, \psi_0), t, \psi, \psi_0) = \sup_{u \in V} H(x, u, t, \psi, \psi_0). \quad (11)$$

Если исходная задача (1) — (4) имеет решение, то как следует из (9), функция (10) определена на непустом множестве.

В ряде случаев функция (10) может быть выписана в явном виде. Например, если

$$f^j(x, u, t) = f_1^j(x, t) + \sum_{i=1}^r f_{2i}^j(x, t) u^i, \quad j = \overline{0, n},$$

$$V = \{u = (u^1, \dots, u^r): \alpha_i \leq u^i \leq \beta_i, \quad i = \overline{1, r}\},$$

где α_i, β_i — заданные числа, то

$$H(x, u, t, \psi, \psi_0) = \sum_{j=0}^n \psi_j f_1^j(x, t) + \sum_{i=1}^r \left(\sum_{j=0}^n \psi_j f_{2i}^j(x, t) \right) u^i.$$

Ясно, что эта функция достигает своей верхней грани на V в точке $u = u(x, t, \psi, \psi_0) \in V$ с координатами

$$u^i = \begin{cases} \beta_i, & \text{если } \sum_{j=0}^n \psi_j f'_{2i}(x, t) > 0; \\ \alpha_i, & \text{если } \sum_{j=0}^n \psi_j f'_{2i}(x, t) < 0, \quad i = \overline{1, r}. \end{cases}$$

В частности, если $\alpha_i = -1$, $\beta_i = +1$, то $u^i = \text{sign} \left(\sum_{j=0}^n \psi_j f'_{2i}(x, t) \right)$.

Полученная формула дает достаточно много информации о структуре оптимального управления: i -я координата оптимального управления является ступенчатой (кусочно постоянной) функцией со значениями α_i или β_i , причем точки переключения определяются условием $\sum_{j=0}^n \psi_j f'_{2i}(x, t) = 0$. Обратим внимание чи-

тателя на возможный особый случай, когда $\sum_{j=0}^n \psi_j(t) f'_{2i}(x(t), t) = 0$ на каком-либо промежутке $[\alpha, \beta] \subseteq [t_0, T]$. В этом случае функция H не будет зависеть от u и из условия (9) не удастся извлечь никакой полезной информации об управлении $u(\cdot)$ при $t \in [\alpha, \beta]$. Тогда говорят, что *оптимальное управление* $u(\cdot)$ является *особым* на промежутке $[\alpha, \beta]$. О том, что особые управления возможны, говорит приводимый ниже пример 6. Мы здесь случаи особого управления рассматривать не будем, и по этим вопросам посылаем читателя к [49, 69].

Другие задачи, когда удастся получить явное выражение для функции (10), приводятся ниже в примерах.

Допустим, что функция (10) нам уже известна. Рассмотрим следующую систему из $2n$ дифференциальных уравнений

$$\begin{aligned} \dot{x} &= f(x, u(x, t, \psi, \psi_0), t), \\ \dot{\psi} &= -H_x(x, u(x, t, \psi, \psi_0), t, \psi, \psi_0), \quad t_0 \leq t \leq T, \end{aligned} \quad (12)$$

относительно неизвестных функций $x(\cdot)$, $\psi(\cdot)$. Как известно [193], общее решение такой системы зависит от $2n$ произвольных числовых параметров (например, такими параметрами являются начальные условия $x(t_0)$, $\psi(t_0)$) и для определения этих параметров нужно иметь еще $2n$ условий. В рассматриваемой задаче такими условиями являются

$$x(t_0) = x_0, \quad x(T) = x_1. \quad (3)$$

Не следует забывать, что система (12) содержит еще один неизвестный параметр $\psi_0 \leq 0$. Как определить этот параметр,

из каких условий? Заметим, что функция $H(x, u, t, \psi, \psi_0)$, определенная соотношением (5), линейна и однородна относительно переменных $\psi_0, \psi_1, \dots, \psi_n$, т. е. $H(x, u, t, \alpha\psi, \alpha\psi_0) = \alpha H(x, u, t, \psi, \psi_0)$ при любых α . Отсюда и из условия (11) тогда имеем

$$u(x, t, \alpha\psi, \alpha\psi_0) \equiv u(x, t, \psi, \psi_0) \quad (13)$$

для любых $\alpha > 0$. Отсюда следует, что теорема 1 определяет величины $\psi_0, \psi_1, \dots, \psi_n$ лишь с точностью до положительного множителя, и этим множителем мы можем распорядиться по своему усмотрению. На практике с учетом условий (8) чаще всего полагают

$$|\psi_0|^2 + |\psi(\bar{t})|^2 = 1, \quad \psi_0 \leq 0, \quad (14)$$

где $\bar{t}$ — какой-либо момент времени, $t_0 \leq \bar{t} \leq T$, например, $\bar{t} = t_0$ или $\bar{t} = T$. В тех задачах, в которых удастся заранее показать, что $\psi_0 < 0$, вместо условия нормировки (14) часто берут $\psi_0 = -1$.

Таким образом, для определения $2n + 1$ параметров системы (12) — $2n$ параметров общего решения и еще параметра ψ_0 — у нас имеется $2n + 1$ условий (3), (14). Можно ожидать, что имеются лишь отдельные, изолированные функции $(x(t), \psi(t))$, $t_0 \leq t \leq T$, и значения параметра ψ_0 , удовлетворяющие условиям (12), (3), (14). Пусть нам удалось определить из этих условий какие-то $(x(t), \psi(t), \psi_0)$, $t_0 \leq t \leq T$. Подставим их в (10) и получим функцию

$$u(t) = u(x(t), t, \psi(t), \psi_0) \in V, \quad t_0 \leq t \leq T. \quad (15)$$

Пусть эта функция оказалась кусочно непрерывной функцией. Из (10), (11), (15) тогда следует, что полученное таким образом управление $u(t)$, $t_0 \leq t \leq T$, удовлетворяет условию (9) и, следовательно, согласно теореме 1 может претендовать на роль оптимального управления задачи (1) — (4), а функция $x(t) = x(t, u(\cdot), x_0)$, $t_0 \leq t \leq T$, — на роль оптимальной траектории этой задачи. Будет ли найденная пара $(u(t), x(t))$, $t_0 \leq t \leq T$, 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 $(u(t), x(t))$, $t_0 \leq t \leq T$, 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 из условий (12), (3), (14) однозначно найдены $(x(t), \psi(t), \psi_0)$, $t_0 \leq t \leq T$, то управление (15) будет оптимальным. Если информации о существовании решения задачи (1) — (4) нет, или условиям (10) — (12), (3), (14), (15) удовлетворяют несколько управлений, то для выясне-

ния вопроса об оптимальности этих управлений требуется дополнительное и порою весьма сложное исследование.

Заметим также, что верхняя грань в правой части (11) может достигаться в нескольких точках, и тогда функция (10) будет определяться неоднозначно. В этом случае для получения всех подозрительных на оптимальность управлений нужно найти все наборы $(x(t), \psi(t), \psi_0)$ и управления $u(t)$, $t_0 \leq t \leq T$, удовлетворяющие условиям (12), (3), (14), (15) для всех функций $u = u(x, t, \psi, \psi_0)$ из (11).

Таким образом, схема использования принципа максимума — теоремы 1 — для получения решения задачи (1) — (4) описана. Как видим, принцип максимума дает просто и изящно выписываемые необходимые условия оптимальности для задачи (1) — (4) и сводит ее к специального вида краевой задаче, состоящей из условия максимума (11), системы дифференциальных уравнений (12), краевых условий (3) и условия нормировки (14). Такую краевую задачу естественно назвать *краевой задачей принципа максимума* для задачи оптимального управления (1) — (4).

Для иллюстрации вышесказанного рассмотрим примеры.

Пример 1. Пусть требуется минимизировать функцию

$$J(u) = \int_0^T (u^2(t) + x^2(t)) dt \quad \text{при условиях } \dot{x}(t) = u(t), \quad 0 \leq t \leq T, \\ x(0) = x(T) = 0.$$

Здесь момент $T > 0$ задан; $V = E^1$. Задача, конечно, несложная: пара $(u(t) \equiv 0, x(t) \equiv 0)$, $0 \leq t \leq T$, очевидно, является единственным ее решением. Продемонстрируем на этой простой задаче изложенную выше схему использования принципа максимума — теоремы 1. Выпишем функцию Гамильтона — Понтрягина $H(x, u, \psi, \psi_0) = \psi_0(u^2 + x^2) + \psi u$ и сопряженную систему $\dot{\psi} = -H_x = -2\psi_0 x$. Если $\psi_0 = 0$, то функция $H = \psi u$ может достигать своей верхней грани на множестве $V = E^1$ лишь при $\psi = 0$. Однако соотношения $\psi_0 = \psi = 0$ противоречат условию (8). Следовательно, $\psi_0 < 0$. Тогда можем считать, что $\psi_0 = -1$. 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 (10) в рассматриваемой задаче. Тогда краевая задача принципа максимума запишется в виде

$$\dot{x} = \psi/2, \quad \dot{\psi} = 2x, \quad 0 \leq t \leq T; \quad x(0) = x(T) = 0.$$

Отсюда однозначно определяем $x(t) \equiv \psi(t) \equiv 0$, $0 \leq t \leq T$. Тогда $u(t) = \psi(t)/2 \equiv 0$, $0 \leq t \leq T$, — 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 $J(u) = \int_0^T (u^2(t) - x^2(t)) dt$ при условиях $\dot{x}(t) = u(t)$, $0 \leq t \leq T$, $x(0) = x(T) = 0$, $T > 0$.

Функция Гамильтона — Понтрягина здесь имеет вид $H = \psi_0(u^2 - x^2) + \psi u$, сопряженная система такая: $\dot{\psi} = -H_x = 2\psi_0 x$. Если $\psi_0 = 0$, то $H = \psi u$ достигает своей верхней грани лишь при $\psi = 0$, что противоречит условию (8). Следовательно, $\psi_0 < 0$. Можем считать, что $\psi_0 = -1$. Тогда $H = x^2 - u^2 + \psi u$ и $\sup_{u \in E_1} H$ достигается при $u = \psi/2$. Краевая задача принципа максимума имеет вид

$$\dot{x} = \psi/2, \quad \dot{\psi} = -2x, \quad 0 \leq t \leq T, \quad x(0) = x(T) = 0.$$

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 $T \neq \pi k$, $k = 1, 2, \dots$; тогда $C = 0$ и краевая задача принципа максимума будет иметь единственное решение $x(t) \equiv 0$, $\psi(t) \equiv 0$, $0 \leq t \leq T$, а управление, подозрительное на оптимальность, равно $u(t) = \psi(t)/2 = 0$, $0 \leq t \leq T$. 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: $u(t) = C \cos t$, $0 \leq t \leq T$.

Спрашивается, будут ли найденные управления оптимальными? Оказывается, ответ на этот вопрос зависит от величины T . Рассмотрим случаи $T > \pi$ и $0 < T \leq \pi$.

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 $x_m = x_m(t) = m \sin \frac{\pi t}{T}$, $0 \leq t \leq T$, $m = 1, 2, \dots$. Тогда $J(u_m) = \int_0^T (u_m^2(t) - x_m^2(t)) dt = \frac{1}{2} T m^2 \left(\frac{\pi^2}{T^2} - 1 \right) \rightarrow -\infty$ при $m \rightarrow \infty$. Следовательно, при $T > \pi$

рассматриваемая задача оптимального управления не имеет решения. В то же время краевая задача принципа максимума разрешима, причем для $T \neq \pi k$, $k = 1, 2, \dots$, она имеет единственное решение, а при $T = \pi k$, $k = 1, 2, \dots$ — бесконечно много решений.

2) $0 < T \leq \pi$. Тогда для любых кусочно непрерывных $v(t)$, для которых существует решение $x(t)$ задачи $\dot{x}(t) = v(t)$, $0 \leq t \leq T$, $x(0) = x(T) = 0$, имеем

$$\begin{aligned} J(v) &= \int_0^T (v^2 - x^2) dt = \int_0^T (v^2 + x^2 \operatorname{ctg}^2 t - x^2 \sin^{-2} t) dt = \\ &= \int_0^T (v^2 + x^2 \operatorname{ctg}^2 t - 2x\dot{x} \operatorname{ctg} t) dt = \int_0^T (v(t) - x(t) \operatorname{ctg} t)^2 dt \geq 0. \end{aligned}$$

Заметим, что проделанные преобразования законны, так как все подынтегральные функции, встретившиеся при этих преобразованиях, ограничены и $\lim_{t \rightarrow +0} x^2(t) \operatorname{ctg} t = 0$, а в случае $T = \pi$ еще и $\lim_{t \rightarrow T-0} x^2(t) \operatorname{ctg} t = 0$. Итак, $J(v) \geq 0$, 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.

2. Перейдем к формулировке принципа максимума для задачи оптимального управления с граничными условиями более общего вида, чем в задаче (1)–(4), а начальный и конечный моменты времени по-прежнему будем считать закрепленными. А именно, рассмотрим задачу

$$J(x_0, u(\cdot)) = \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T)) \rightarrow \inf; \quad (16)$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T, \quad (17)$$

$$x(t_0) \in S_0, \quad x(T) \in S_T, \quad (18)$$

$$u(t) \in V, \quad t_0 \leq t \leq T, \quad (19)$$

где управления $u(\cdot)$ кусочно непрерывны на $[t_0, T]$, $u(t) = u(t+0)$ при $t_0 \leq t < T$, $u(T) = u(T-0)$. Задача (16)–(19) является частным случаем сформулированной в § 1 более общей задачи (1.25)–(1.29) и получается из нее при $\Theta_0 = \{t_0\}$, $\Theta_1 = \{T\}$, $G(t) \equiv E^n$ и $V(t) \equiv V$ при $t_0 \leq t \leq T$.

Будем предполагать, что правый конец траектории либо свободен, т. е. $S_l = E^n$, либо множество S_l имеет вид

$$S_l = \{x \in E^n: g_j(x) = 0, \quad j = \overline{1, s_l}\}, \quad (20)$$

либо

$$S_l = \{x \in E^n: g_j(x) \leq 0, \quad j = \overline{1, m_l}; \quad g_j(x) = 0, \quad j = \overline{m_l + 1, s_l}\}. \quad (21)$$

В частности, если $g_j(x) = x^j - x_l^j$, $j = \overline{1, n}$, $s_l = n$, то из (20) получим случай закрепленного правого конца: $x(T) = x_l$. Аналогично, на левом конце траектории будем предполагать, что либо $S_0 = E^n$, либо

$$S_0 = \{x \in E^n: h_j(x) = 0, \quad j = \overline{1, s_0}\}, \quad (22)$$

либо

$$S_0 = \{x \in E^n: h_j(x) \leq 0, \quad j = \overline{1, m_0}; \quad h_j(x) = 0, \quad j = \overline{m_0 + 1, s_0}\}. \quad (23)$$

Далее, будем считать, что функции $f^j(x, u, t)$, $j = \overline{0, n}$, $\Phi(x)$, $g_j(x)$, $j = \overline{1, s_l}$, $h_j(x)$, $j = \overline{1, s_0}$, имеют частные производные по переменным $(x^1, \dots, x^n)$ и непрерывны вместе с указанными производными по совокупности своих аргументов при всех $x \in E^n$, $u \in V$, $t \in [t_0, T]$. В дополнение к введенным выше обозначениям f_x , f_x^0 еще положим

$$\frac{\partial \Phi}{\partial x} = \Phi_x = (\Phi_{x^1}, \dots, \Phi_{x^n}), \quad \frac{\partial g_j}{\partial x} = g_{jx} = (g_{jx^1}, \dots, g_{jx^n}),$$

$$\frac{\partial h_j}{\partial x} = h_{jx} = (h_{jx^1}, \dots, h_{jx^n}).$$

Теорема 2. Пусть $(u(t), x(t))$, $t_0 \leq t \leq T$, — решение задачи (16)–(19). Тогда необходимо существуют непрерывная вектор-функция $\psi(t)$, $t_0 \leq t \leq T$, и постоянная ψ_0 такие, что

$$1) \quad \psi_0 \leq 0, \quad |\psi_0| + |\psi(t)| \neq 0, \quad t_0 \leq t \leq T; \quad (8)$$

2) $\psi(\cdot)$ является решением сопряженной системы (6), соответствующей рассматриваемому решению $(u(\cdot), x(\cdot))$;

3) при каждом $t \in [t_0, T]$ функция $H(x(t), u, t, \psi(t), \psi_0)$ переменной $u = (u^1, \dots, u^r)$ достигает своей верхней грани на множестве V при $u = u(t)$, т. е.

$$\sup_{u \in V} H(x(t), u, t, \psi(t), \psi_0) = H(x(t), u(t), t, \psi(t), \psi_0), \quad t_0 \leq t \leq T; \quad (9)$$

4) на левом и правом концах траектории $x(\cdot)$ выполняются условия трансверсальности, которые в случае задачи (16)–(19) означают, что вектор $\psi(T) - \psi_0 \Phi_x(x(T))$ ортогонален к множеству S_l в точке $x(T) \in S_l$, а вектор $\psi(t_0)$ ортогонален к множеству S_0 в точке $x(t_0) \in S_0$.

Если управление $u(\cdot)$ является ограниченной измеримой функцией, то формулировка теоремы 2 сохраняется, только лишь соотношение (9) (как и включение (19)) будет выполняться в смысле почти всюду на $[t_0, T]$.

Как видим, утверждения 1)–3) теорем 1 и 2 совпадают. Поэтому, применяя принцип максимума к задаче (16)–(19) по описанной выше схеме, мы снова придем к условию максимума (11) и к системе $2n$ дифференциальных уравнений (12) относительно функций $(x(\cdot), \psi(\cdot))$, но граничные условия теперь, вообще говоря, будут отличаться от условий (3). Заметим, что равенство (13) сохраняет силу и здесь, а свойство ортогональности векторов $\psi(T) - \psi_0 \Phi_x(x(T))$, $\psi(t_0)$ к множествам S_I и S_0 соответственно также не нарушится от умножения всех величин $\psi_0, \psi_1, \dots, \psi_n$ на одно и то же число $\alpha > 0$. Поэтому и в рассматриваемой задаче можно принять условие нормировки (14) или условие $\psi_0 = -1$, если известно, что $\psi_0 < 0$. Остается указать еще $2n$ условий, из которых можно было бы определить $2n$ числовых параметров, от которых зависит общее решение системы (12). Для получения этих условий расшифруем условие трансверсальности при различных режимах на левом и правом концах траектории.

Начнем с правого конца. Если правый конец свободен, т. е. $S_I = E^n$, то условие ортогональности вектора $\psi(T) - \psi_0 \Phi_x(x(T))$ ко всему пространству E^n означает, что

$$\psi(T) - \psi_0 \Phi_x(x(T)) = 0. \quad (24)$$

Таким образом, в случае свободного правого конца условие трансверсальности дает n граничных условий для системы (12), записанных в векторной форме (24).

Пусть теперь правый конец является подвижным. Сначала рассмотрим случай, когда множество S_I задается в виде (20) и $g_{jx}(x(T)) \neq 0$, $j = 1, s_I$. Тогда гиперплоскость $\langle g_{jx}(x(T)), x - x(T) \rangle = 0$ представляет собой касательную плоскость к поверхности $g_j(x) = 0$ в точке $x(T)$, а аффинное множество $\Gamma = \{x \in E^n: \langle g_{jx}(x(T)), x - x(T) \rangle = 0, j = 1, s_I\}$ — касательную плоскость к множеству S_I в точке $x(T)$. Условие ортогональности вектора a к множеству S_I в точке $x(T)$, по определению, означает ортогональность a к касательной плоскости Γ , т. е. $\langle a, x - x(T) \rangle = 0$ при всех $x \in \Gamma$. Таким образом, из условия трансверсальности имеем: $\langle \psi(T) - \psi_0 \Phi_x(x(T)), x - x(T) \rangle = 0$ для всех x , для которых $\langle g_{ix}(x(T)), x - x(T) \rangle = 0$, $i = 1, s_I$. По теореме 4.8.3 тогда найдутся числа $a_1, a_2, \dots, a_{s_I}$ такие, что

$$\psi(T) - \psi_0 \Phi_x(x(T)) = \sum_{j=1}^{s_I} a_j g_{jx}(x(T)). \quad (25)$$

Сюда же добавим еще условие принадлежности $x(T) \in S_l$, т. е.

$$g_j(x(T)) = 0, \quad j = \overline{1, s_l}. \quad (26)$$

Таким образом, равенства (25), (26) дают всего $n + s_l$ условий, из которых s условий можно употребить на определение дополнительно появившихся параметров $a_1, \dots, a_{s_l}$, а остальные n условий присоединить к системе (12).

В случае, когда множество S_l имеет вид (21) и $g_{jx}(x(T)) \neq 0$, $j = \overline{1, s_l}$, условие трансверсальности означает [70], что существуют числа $a_1, \dots, a_{s_l}$, такие, что

$$\psi(T) - \psi_0 \Phi_x(x(T)) = \sum_{j=1}^{s_l} a_j g_{jx}(x(T)), \quad (27)$$

$$a_j g_j(x(T)) = 0, \quad a_j \geq 0, \quad j = \overline{1, m_l}; \quad g_j(x(T)) = 0, \quad j = \overline{m_l + 1, s_l}. \quad (28)$$

Таким образом, и соотношения (27), (28) дают $n + s_l$ равенств, из которых S_l условий можно употребить на определение параметров $a_1, \dots, a_{s_l}$, а остальные n условий присоединить к системе (12). Кстати, для всех тех j , $1 \leq j \leq m_l$, для которых $g_j(x(T)) < 0$ (неактивные ограничения), из (28) следует, что $a_j = 0$, и неопределенными остаются лишь a_j с номерами j , для которых $g_j(x(T)) = 0$ (активные ограничения).

Заметим, что в случае закрепленного правого конца, когда

$$x(T) = x_l, \quad (29)$$

условие трансверсальности тоже можно истолковать как условие ортогональности вектора $\psi(T) - \psi_0 \Phi_x(x(T))$ к вектору нулевой длины — к точке x_l . Понятно, что это всегда тривиально выполняется. Это значит, что условие трансверсальности в случае закрепленного конца вырождается и не несет в себе никакой информации. К счастью, никакой дополнительной информации к условию (29) здесь и не требуется — из (29) мы имеем n граничных условий, которые присоединяем к системе (12).

Таким образом, из условия трансверсальности на правом конце траектории во всех случаях получаем n граничных условий для системы (12). Для получения еще n граничных условий для этой системы обратимся к условию трансверсальности на левом конце траектории.

В случае свободного левого конца, когда $S_0 = E^n$, имеем

$$\psi(t_0) = 0. \quad (30)$$

Рассмотрим случай подвижного левого конца. Если множество S_0 имеет вид (22) и $h_{jx}(x(t_0)) \neq 0$, $j = \overline{1, s_0}$, то условие трансверсальности означает существование чисел $b_1, \dots, b_{s_0}$,

таких, что

$$\psi(t_0) = - \sum_{j=1}^s b_j h_{jx}(x(t_0)). \quad (31)$$

К условиям (31) надо добавить еще условие $x(t_0) \in S_0$, означающее, что

$$h_j(x(t_0)) = 0, \quad j = \overline{1, s_0}. \quad (32)$$

Если множество S_0 имеет вид (23) и $h_{jx}(x(t_0)) \neq 0$, $j = \overline{1, s_0}$, то условие трансверсальности и условие $x(t_0) \in S_0$ запишутся так:

$$\psi(t_0) = - \sum_{j=1}^s b_j h_{jx}(x(t_0)), \quad (33)$$

$$b_j h_j(x(t_0)) = 0, \quad b_j \geq 0, \quad j = \overline{1, m_0}; \quad h_j(x(t_0)) = 0, \quad j = \overline{m_0 + 1, s_0}. \quad (34)$$

Наконец, в случае закрепленного левого конца, когда

$$x(t_0) = x_0, \quad (35)$$

условие трансверсальности тривиально выполняется и ничего нового к (35) не добавляет.

Как и в случае правого конца, анализируя получившиеся при различных режимах на левом конце соотношения (30) — (35), заключаем, что эти соотношения дают недостающие n граничных условий для системы (12).

Таким образом, принцип максимума — теорема 2 — позволяет свести задачу оптимального управления (16) — (19) к краевой задаче специального вида, состоящей из условия максимума (11), системы дифференциальных уравнений (12), условия нормировки (14) и граничных условий, получаемых из (24) — (35) в зависимости от того, является ли соответствующий конец траектории закрепленным, подвижным или свободным. Эту краевую задачу будем называть краевой задачей принципа максимума для задачи оптимального управления (16) — (19). Вообще говоря, можно ожидать, что из этой краевой задачи определяться лишь отдельные значения параметра ψ_0 и отдельные функции $x(t)$, $\psi(t)$, $t_0 \leq t \leq T$, с помощью которых по формуле (15) можно получить управления $u(t)$, $t_0 \leq t \leq T$, подозрительные на оптимальность.

Заметим, что в литературе под условием трансверсальности на правом конце траектории часто подразумевают лишь условия (24) или (25), или (27) вместе с первыми m_1 соотношениями (28), на левом конце — лишь условия (30) или (31), или (33) вместе с первыми m_0 соотношениями (34), а в случае закрепленного конца об условии трансверсальности вообще не говорят. Приведенная в теореме 2 геометрическая формулировка условий

трансверсальности, конечно, ничего не добавляет к принятой в литературе трактовке этих условий, но позволяет кратко и единообразно охватить сразу все возможные граничные режимы при закрепленных начальных и конечных моментах времени.

3. Рассмотрим задачу оптимального управления, когда начальный и конечный моменты времени, вообще говоря, неизвестны и также подлежат определению. А именно, пусть требуется минимизировать функцию

$$J(t_0, T, x_0, u(\cdot)) = \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T), T) \quad (36)$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T; \quad (37)$$

$$x(t_0) \in S_0(t_0); \quad x(T) \in S_1(T), \quad (38)$$

$$u(t) \in V, \quad t_0 \leq t \leq T, \quad (39)$$

где управления $u(\cdot)$ кусочно непрерывны на $[t_0, T]$, $u(t) = u(t+0)$ при $t_0 \leq t < T$, $u(T) = u(T-0)$. Задача (36)–(39) является частным случаем задачи (1.25)–(1.29) при $\Theta_0 = \Theta_1 = \mathbb{R} = \{t: -\infty < t < +\infty\}$, $G(t) \equiv E^n$ и $V(t) \equiv V$ при $t \in \mathbb{R}$.

Будем предполагать, что правый конец траектории либо свободен, т. е. $S_1(t) \equiv E^n$, $t \in \mathbb{R}$, либо множество $S_1(T)$ имеет вид (1.23):

$$S_1(T) = \{x \in E^n: g_j(x, T) \leq 0, \quad j = \overline{1, m_j}; \\ g_j(x, T) = 0, \quad j = \overline{m_j+1, s_j}\}, \quad T \in \mathbb{R}. \quad (40)$$

Аналогично, на левом конце траектории пусть либо $S_0(t_0) \equiv E^n$, $t_0 \in \mathbb{R}$, либо множество $S_0(t_0)$ имеет вид (1.22):

$$S_0(t_0) = \{x \in E^n: h_j(x, t_0) \leq 0, \\ j = \overline{1, m_0}; \quad h_j(x, t_0) = 0, \quad j = \overline{m_0+1, s_0}\}, \quad t_0 \in \mathbb{R}_0. \quad (41)$$

Случаи $m_1 = 0$ или $s_1 = m_1$, или $m_0 = 0$, или $s_0 = m_0$ в (40), (41) не исключаются.

Далее, будем считать, что функции $f^j(x, u, t)$, $j = \overline{0, n}$, и их частные производные $f_x^j(x, u, t)$ непрерывны по совокупности $(x, u, t) \in E^n \times V \times \mathbb{R}$, а функции $\Phi(x, t)$, $g_j(x, t)$, $j = \overline{1, s_1}$, $h_j(x, t)$, $j = \overline{1, s_0}$, и их частные производные Φ_x , Φ_t , g_{jx} , g_{jt} , h_{jx} , h_{jt} непрерывны по совокупности $(x, t) \in E^n \times \mathbb{R}$ (обозначения f_x^0 , f_x^j , Φ_x , g_{jx} , h_{jx} см. выше).

Теорема 3. Пусть набор $(t_0, T, u(\cdot), x(\cdot))$ является решением задачи (36)–(39). Тогда необходимо существуют непрерывная вектор-функция $\psi(t)$, $t_0 \leq t \leq T$, и параметр ψ_0

такие, что

$$1) \quad \psi_0 \leq 0, \quad |\psi_0| + |\psi(t)| \neq 0, \quad t_0 \leq t \leq T; \quad (8)$$

2) $\psi(\cdot)$ — решение сопряженной системы (6), соответствующей рассматриваемому решению $(u(t), x(t))$ на рассматриваемом отрезке $t_0 \leq t \leq T$;

3) при каждом $t \in [t_0, T]$ функция $H(x(t), u, t, \psi(t), \psi_0)$ непрерывной и достигает своей верхней грани на множестве V при $u = u(t)$, т. е.

$$\sup_{u \in V} H(x(t), u, t, \psi(t), \psi_0) = H(x(t), u(t), t, \psi(t), \psi_0); \quad (9)$$

4) на левом и правом концах траектории $x(\cdot)$ выполняются условия трансверсальности.

Расшифруем условия трансверсальности на концах оптимальной траектории. Начнем с правого конца. Если правый конец свободен, то

$$\psi(T) - \psi_0 \Phi_x(x(T), T) = 0, \quad (24)$$

$$H(x(T), u(T), T, \psi(T), \psi_0) - \psi_0 \Phi_t(x(T), T) = 0. \quad (42)$$

Если правый конец подвижный и множество $S_I(T)$ имеет вид (40), причем $(g_{jx}(x(T), T), g_{jt}(x(T), T)) \neq 0$, $j = \overline{1, s_I}$, то существуют числа $a_1, \dots, a_{s_I}$ такие, что

$$\psi(T) - \psi_0 \Phi_x(x(T), T) = \sum_{j=1}^{s_I} a_j g_{jx}(x(T), T); \quad (27)$$

$$a_j g_{jt}(x(T), T) = 0, \quad a_j \geq 0, \quad j = \overline{1, m_I};$$

$$g_{jt}(x(T), T) = 0, \quad j = \overline{m_I + 1, s_I}. \quad (28)$$

$$H(x(T), u(T), T, \psi(T), \psi_0) - \psi_0 \Phi_t(x(T), T) =$$

$$= - \sum_{j=1}^{s_I} a_j g_{jt}(x(T), T). \quad (43)$$

Если правый конец закреплен, то

$$x(T) = x_I, \quad (29)$$

$$H(x_I, u(T), T, \psi(T), \psi_0) - \psi_0 \Phi_t(x_I, T) = 0. \quad (44)$$

Аналогично, если левый конец свободен, то

$$\psi(t_0) = 0, \quad (30)$$

$$H(x(t_0), u(t_0), t_0, \psi(t_0), \psi_0) = 0, \quad (45)$$

Если левый конец подвижный и множество $S_0(t_0)$ имеет вид (41), причем $(h_{jx}(x(t_0), t_0), h_{jt}(x(t_0), t_0)) \neq 0$, то существуют числа

$b_1, \dots, b_{s_0}$ такие, что

$$\psi(t_0) = - \sum_{j=1}^{s_0} b_j h_{jx}(x(t_0), t_0), \quad (33)$$

$$b_j h_j(x(t_0), t_0) = 0, \quad b_j \geq 0, \quad j = \overline{1, m_0};$$

$$h_j(x(t_0), t_0) = 0, \quad j = \overline{m_0 + 1, s_0}, \quad (34)$$

$$H(x(t_0), u(t_0), t_0, \psi(t_0), \psi_0) = \sum_{j=1}^{s_0} b_j h_{jx}(x(t_0), t_0). \quad (46)$$

Если левый конец закреплен, то

$$x(t_0) = x_0, \quad (35)$$

$$H(x_0, u(t_0), t_0, \psi(t_0), \psi_0) = 0. \quad (47)$$

Нетрудно видеть, что с помощью теоремы 3 задачу (36) — (39) можно свести к такой же краевой задаче принципа максимума, какая была получена для задачи (16) — (19) с помощью теоремы 2, а наличие неизвестных моментов t_0, T здесь «компенсируется» появлением дополнительных условий вида (42) — (47).

З а м е ч а н и е 1. Если в задаче (36) — (39) начальный или конечный момент известен, то теорема 3 сохраняет силу, нужно лишь условие трансверсальности на левом [правом] конце траектории расшифровывать так же, как в теореме 2.

4. Рассмотрим примеры задач оптимального управления, иллюстрирующие теоремы 1—3.

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. Воспользуемся теоремой 2. Составим функцию Гамильтона — Понтрягина $H = \psi_0(x^2 + u^2)/2 + \psi(-ax + u)$ и выпишем сопряженную систему

$$\dot{\psi} = -H_x = -\psi_0 x + a\psi, \quad 0 \leq t \leq T.$$

Из условия трансверсальности для свободного правого конца траектории имеем $\psi(T) = 0$. Если $\psi_0 = 0$, то из полученных условий следует $\psi(t) \equiv 0$, что противоречит требованию (8). Это значит, что $\psi_0 < 0$. Можем считать, что $\psi_0 = -1$. 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

$$u(t) = \psi(t) = x_0 \frac{e^{\lambda t} - e^{2\lambda T} e^{-\lambda t}}{(\lambda - a) + (\lambda + a) e^{2\lambda T}}, \quad 0 \leq t \leq T; \quad \lambda = \sqrt{a^2 + 1}.$$

Пример 4. Пусть точка движется по оси Ox по закону $\ddot{x}(t) = u(t)$, $t \geq 0$. Требуется найти кусочно непрерывное управление $u(t)$, $|u(t)| \leq 1$, $0 \leq t \leq T$, 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: быстрее всего перевести фазовую точку (x^1, x^2) из состояния $(1, 0)$ в состояние $(0, 0)$, считая, что движение подчиняется уравнениям $\dot{x}^1(t) = x^2(t)$, $\dot{x}^2(t) = u(t)$, $t \geq 0$. Здесь $V = \{u \in E^1: |u| \leq 1\}$, $f^0 \equiv 1$, $\Phi \equiv 0$.

Вспользуемся теоремой 3 и замечанием 1 к ней. Составим функцию $H = \psi_0 + \psi_1 x^2 + \psi_2 u$ и выпишем сопряженную систему

$$\dot{\psi}_1 = -H_{x^1} = 0, \quad \dot{\psi}_2 = -H_{x^2} = -\psi_1, \quad t \geq 0.$$

Отсюда имеем $\psi_1(t) = C$, $\psi_2(t) = -Ct + D$, где C, D — постоянные. Из условия $\max_{|u| \leq 1} H$ следует $u(t) = \text{sign } \psi_2(t) = \text{sign}(-Ct + D)$, $t \geq 0$. 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 $t \geq 0$, или $u(t) \equiv -1$ при $t \geq 0$, или $u(t) \equiv +1$, $0 \leq t < t_1$, $u(t) = -1$, $t \geq t_1$, никогда не будет проходить через точку $(0, 0)$. Остается рассмотреть управление $u(t) = -1$, $0 \leq t < t_1$, $u(t) = +1$, $t \geq t_1$. Этому управлению соответствует траектория $(x_1(t), x_2(t))$:

$$x^1(t) = \begin{cases} 1 - t^2/2, & 0 \leq t \leq t_1, \\ t^2/2 - 2t_1 t + t_1^2 + 1, & t \geq t_1, \end{cases}$$

$$x^2(t) = \begin{cases} -t, & 0 \leq t \leq t_1, \\ t - 2t_1, & t \geq t_1. \end{cases}$$

Из условий $x^1(T) = x^2(T) = 0$ находим $t_1 = 1$, $T = 2$. Тогда

$$x^1(t) = \begin{cases} 1 - t^2/2, & 0 \leq t \leq 1, \\ (t - 2)^2/2, & 1 \leq t \leq 2, \end{cases} \quad x^2(t) = \begin{cases} -t, & 0 \leq t \leq 1, \\ t - 2, & 1 \leq t \leq 2. \end{cases}$$

В качестве величин ψ_0, ψ_1, ψ_2 , участвующих в формулировке принципа максимума, могут служить $\psi_0 = 0, \psi_1(t) = -1, \psi_2(t) = t - 1, 0 \leq t \leq 2$. Можно показать, что полученные управление и траектория в самом деле являются решением поставленной задачи быстрогодействия — 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 $S_1 = \{x \in E^2: g(x) \equiv x^1 = 0\}$ быстрее всего образом, предполагая, что движение точки подчиняется уравнениям $\dot{x}^1(t) = x^2(t), \dot{x}^2(t) = u(t)$, причем $u(t) \in V = \{u \in E^1: |u| \leq 1\}$.

Как и в предыдущем примере, здесь $H = \psi_0 + \psi_1 x^2 + \psi_2 u$, 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 (27), (28), (43) здесь дают

$$\psi_2(T) = 0, \quad x^1(T) = 0, \quad \psi_0 + \psi_1(T)x^2(T) + \psi_2(T)u(T) = H|_{t=T} = 0.$$

Следовательно, $\psi_2(t) = C(T - t), 0 \leq t \leq T$. Заметим, что здесь $C \neq 0$, так как при $C = 0$ получим $\psi_1(t) \equiv \psi_2(t) \equiv 0, 0 \leq t \leq T$, а тогда из условия $H|_{t=T} = 0$ вытекает $\psi_0 = 0$ — противоречие с условием (8). Итак, $C \neq 0, \psi(t) = C(T - t) \neq 0$ при $0 \leq t < T$. Из условия $\sup_{|u| \leq 1} H$ тогда имеем

$$u(t) = \text{sign } \psi_2(t) = \text{sign } C, \quad 0 \leq t \leq T.$$

Значит, подозрительными на оптимальность здесь могут быть лишь управления $u(t) \equiv 1$ или $u(t) \equiv -1, 0 \leq t \leq T$. Если $u(t) = 1$, то из краевой задачи

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = 1, \quad 0 \leq t \leq T; \quad x^1(0) = 2, \quad x^2(0) = -2, \quad x^1(T) = 0$$

получим $T = 2, x^1(t) = (t - 2)^2/2, x^2(t) = t - 2, 0 \leq t \leq 2$. Если $u(t) = -1$, то из

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -1, \quad 0 \leq t \leq T; \quad x^1(0) = 2, \quad x^2(0) = -2; \quad x^1(T) = 0$$

будем иметь $T = \sqrt{8} - 2, x^1(t) = 4 - (t + 2)^2/2, x^2(t) = -t - 2, 0 \leq t \leq \sqrt{8} - 2$.

Таким образом, краевая задача принципа максимума здесь дает два решения. Однако лишь управление $u(t) = -1, 0 \leq t \leq T = \sqrt{8} - 2$ может претендовать на оптимальность, так как $T = 2 > \sqrt{8} - 2$ и управление $u(t) = 1, 0 \leq t \leq T$, заведомо неоптимально.

В следующих примерах 6—9 покажем, как выписывается краевая задача принципа максимума для некоторых задач оптимального управления движением математического маятника (см. пример 1.1).

Пример 6. Пусть требуется минимизировать функцию

$$J(u) = (x^1(T))^2 + (x^2(T))^2 \quad (48)$$

при условиях

$$\begin{aligned} \dot{x}^1(t) &= x^2(t), \quad \dot{x}^2(t) = -\sin x^1(t) - \beta x^2(t) + u(t), \\ 0 \leq t \leq T; \quad x(0) &= x_0, \end{aligned} \quad (49)$$

$$u(t) \in V = \{u \in E^1: |u| \leq 1\}, \quad (50)$$

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, $f^0 \equiv 0$, $\Phi(x) = (x^1)^2 + (x^2)^2$. Выпишем функцию Гамильтона — Понтрягина

$$H(x, u, \psi, \psi_0) = \psi_1 x^2 + \psi_2 (-\sin x^1 - \beta x^2 + u)$$

и сопряженную систему

$$\dot{\psi}_1 = -H_{x^1} = \psi_2 \cos x^1, \quad \dot{\psi}_2 = -H_{x^2} = -\psi_1 + \beta \psi_2, \quad 0 \leq t \leq T. \quad (51)$$

Из условия трансверсальности (24) имеем

$$\begin{aligned} \psi_1(T) &= \psi_0 \Phi_{x^1}(x(T)) = 2\psi_0 x^1(T); \\ \psi_2(T) &= \psi_0 \Phi_{x^2}(x(T)) = 2\psi_0 x^2(T). \end{aligned} \quad (52)$$

Сразу заметим, что здесь $\psi_0 < 0$, так как при $\psi_0 = 0$ из (51), (52) получим $\psi_1(t) \equiv \psi_2(t) \equiv 0$, $0 \leq t \leq T$, что противоречит условию (8). Следовательно, можно считать $\psi_0 = -1$. Из условия $\max_{|u| \leq 1} H$ следует $u = \text{sign } \psi_2$. Тогда краевая задача принципа

максимума запишется в виде

$$\begin{aligned} \dot{x}^1 &= x^2, \quad \dot{x}^2 = -\sin x^1 - \beta x^2 + \text{sign } \psi_2, \\ \dot{\psi}_1 &= \psi_2 \cos x^1, \quad \dot{\psi}_2 = -\psi_1 + \beta \psi_2, \quad 0 \leq t \leq T, \\ x^1(0) &= x_0^1, \quad x^2(0) = x_0^2; \quad \psi_1(T) = -2x^1(T), \quad \psi_2(T) = -2x^2(T). \end{aligned} \quad (53)$$

Если эта краевая задача имеет решение $(x(t), \psi(t))$, $0 \leq t \leq T$, причем $\psi_2(t)$ обращается в нуль в конечном числе точек, то функция $u(t) = \text{sign } \psi_2(t)$ 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 $\psi(t, v) \equiv 0$, $0 \leq t \leq T$, а тогда $H(x(t, v), u, \psi(t, v)) \equiv 0$, $0 \leq t \leq T$, при всех $u \in$

$\in V$ и условие $\sup_{|u| \leq 1} H$ не суживает исходное множество управлений, подозрительных на оптимальность.

Пример 7. Рассмотрим задачу быстрого перевода точки $x = (x^1, x^2)$ из состояния $x_0 \neq 0$ в начало координат $(0, 0)$, предполагая, что выполняются условия (49), (50).

Здесь $f^0 \equiv 1$, $\Phi \equiv 0$ и поэтому $H = \psi_0 + \psi_1 x^2 + \psi_2 (-\sin x^1 - \beta x^2 + u)$. Отсюда ясно, что сопряженная система будет иметь вид (53), а условие $\max_{|u| \leq 1} H$ выделит $u = \text{sign } \psi_2$. Краевая за-

дача принципа максимума в этом случае будет состоять из системы (53), граничных условий $x(0) = x_0$, $x(T) = 0$, и еще условия $\psi_0 + \psi_2(T)u(T) = 0$, вытекающего из (44).

Пример 8. Пусть требуется быстрее всего перевести точку $x = (x^1, x^2)$ из состояния x_0 на множество $S_1 = \{x \in E^2: g(x) = |x|^2 = \varepsilon^2\}$, $\varepsilon = \text{const} > 0$, предполагая, что выполняются условия (49), (50).

Функция H и сопряженная система здесь имеют тот же вид, что и в примере 6. Из условия $\max_{|u| \leq 1} H$ следует $u = \text{sign } \psi_2$. Из условий трансверсальности (27), (28), (43) имеем

$$\psi(T) = 2ax(T), \quad |x(T)|^2 = \varepsilon^2, \quad H(x(T), u(T), \psi(T), \psi_0) = 0. \quad (54)$$

Таким образом, краевая задача принципа максимума в этом случае состоит из системы (53), граничных условий (54) и $x(0) = x_0$. Кстати, из (54) нетрудно исключить параметр a : $|\psi(T)| = 2|a||x(T)| = 2|a|\varepsilon$, откуда $a = \pm |\psi(T)|/(2\varepsilon)$. Это значит, что первые два условия (54) можно заменить на условие $\psi(T) = \pm |\psi(T)|x(T)/\varepsilon$.

Если в рассматриваемой задаче множество S_1 заменить на множество $\{x = (x^1, x^2): g(x) = x^1 = 0\}$, то краевая задача принципа максимума остается той же, за исключением первых двух условий (54), которые следует заменить условиями $x^1(T) = 0$, $\psi_2(T) = 0$.

Пример 9. Пусть требуется минимизировать функцию $J(u) = \int_0^T u^2(t) dt$ при условиях (49) и закреплённом правом конце $x(T) = 0$; $T > 0$ — задано.

Здесь $V = E^1$, $H = \psi_0 u^2 + \psi_1 x^2 + \psi_2 (-\sin x^1 + \beta x^2 + u)$, сопряженная система имеет вид (51). В случае $\psi_0 = 0$ функция H может достигать своей верхней грани на E^1 лишь при $\psi_2 = 0$. Но если $\psi_2(t) \equiv 0$, $0 \leq t \leq T$, то из второго уравнения (51) получим $\psi_1(t) \equiv 0$, что противоречит условию (8). Таким образом, можем считать $\psi_0 = -1$. Тогда из условия $\max_{u \in E^1} H$ получим

$u = \psi_2/2$. Краевая задача принципа максимума будет иметь вид

$$\dot{x}^1 = x^2, \quad \dot{x}^2 = -\sin x^1 - \beta x^2 + \psi_2/2;$$

$$\dot{\psi}_1 = \psi_2 \cos x^1, \quad \dot{\psi}_2 = -\psi_1 + \beta \psi_2, \quad 0 \leq t \leq T;$$

$$x(0) = x_0, \quad x(T) = 0.$$

5. В следующем параграфе будет приведено доказательство принципа максимума для задачи оптимального управления, когда начальный и конечный моменты времени и левый конец траектории зафиксированы, а правый конец свободен. Доказательство принципа максимума — теорем 1, 2, 3 — в остальных случаях читатель может найти, например, в [1, 14, 46, 54, 70, 102, 103, 163].

Здесь мы хотим привести эвристические соображения, которые помогают понять, откуда появляются сопряженная система, условие максимума, условия трансверсальности, фигурирующие в принципе максимума, и установить связь между принципом максимума Понтягина и теоремой Куна — Таккера.

Будем рассматривать задачу (16) — (19) при дополнительном предположении, что $V \equiv E^r$, а множества S_I, S_0 имеют вид (20), (22). Следуя известной процедуре исследования задач на условный экстремум (для задач минимизации функций конечного числа переменных эта процедура была изложена в § 2.2; для задач вариационного исчисления — см. [80, 257]), введем множители Лагранжа: непрерывную вектор-функцию $\psi(t) = (\psi_1(t), \dots, \psi_n(t))$ для учета ограничений (17), постоянные $(a_1, \dots, a_{s_I}) = a$ для учета ограничений (20), постоянные $(b_1, \dots, b_{s_0}) = b$ для учета (22), постоянную ψ_0 — для функции (16), и составим функцию Лагранжа

$$\mathcal{L}(x(\cdot), u(\cdot), \psi(\cdot), \psi_0, a, b) =$$

$$= \psi_0 \left[\int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T)) \right] +$$

$$+ \int_{t_0}^T \langle \psi(t), f(x(t), u(t), t) - \dot{x}(t) \rangle dt +$$

$$+ \sum_{j=1}^{s_I} a_j g_j(x(T)) + \sum_{j=1}^{s_0} b_j h_j(x(t_0)).$$

С помощью функции Гамильтона — Понтрягина (5) можно записать функцию Лагранжа в следующем виде:

$$\begin{aligned} \mathcal{L}(x(\cdot), u(\cdot), \psi(\cdot), \psi_0, a, b) = \\ = \int_{t_0}^T [H(x(t), u(t), t, \psi(t), \psi_0) - \langle \psi(t), \dot{x}(t) \rangle] dt + \\ + \psi_0 \Phi(x(T)) + \sum_{j=1}^{s_I} a_j g_j(x(T)) + \sum_{j=1}^{s_0} b_j h_j(x(t_0)). \end{aligned}$$

Все аргументы функции Лагранжа считаются независимыми переменными. Дадим приращения (вариации) всем этим переменным, кроме ψ_0 , т. е. рассмотрим $x(t) + \delta x(t)$, $u(t) + \delta u(t)$, $\psi(t) + \delta \psi(t)$, $t_0 \leq t \leq T$, $a + \delta a$, $b + \delta b$; здесь $u(t)$, $\delta u(t)$ — кусочно непрерывны, а $x(t)$, $\delta x(t)$, $\psi(t)$, $\delta \psi(t)$ — непрерывно дифференцируемы на $[t_0, T]$. Тогда вариация функции Лагранжа, представляющая собой главную линейную часть приращения этой функции, имеет вид

$$\begin{aligned} \delta \mathcal{L} = \int_{t_0}^T [\langle H_x, \delta x \rangle + \langle H_u, \delta u \rangle + \langle H_\psi, \delta \psi \rangle - \langle \psi, \delta \dot{x} \rangle - \langle \dot{x}, \delta \psi \rangle] dt + \\ + \psi_0 \langle \Phi_x(x(T)), \delta x(T) \rangle + \sum_{j=1}^{s_I} a_j \langle g_{jx}(x(T)), \delta x(T) \rangle + \\ + \sum_{j=1}^{s_T} \delta a_j g_j(x(T)) + \sum_{j=1}^{s_0} \delta b_j h_j(x(t_0)) + \sum_{j=1}^{s_0} b_j \langle h_{jx}(x(t_0)), \delta x(t_0) \rangle; \end{aligned}$$

для краткости аргументы у функций под интегралом опущены. Интегрируя по частям, находим

$$\int_{t_0}^T \langle \psi(t), \delta \dot{x}(t) \rangle dt = \langle \psi(t), \delta x(t) \rangle \Big|_{t_0}^T - \int_{t_0}^T \langle \dot{\psi}(t), \delta x(t) \rangle dt.$$

Тогда

$$\begin{aligned} \delta \mathcal{L} = \int_{t_0}^T [\langle H_x + \dot{\psi}, \delta x \rangle + \langle H_u, \delta u \rangle + \langle H_\psi - \dot{x}, \delta \psi \rangle] dt + \\ + \langle \psi_0 \Phi_x(x(T)) + \sum_{j=1}^{s_I} a_j g_{jx}(x(T)) - \psi(T), \delta x(T) \rangle + \sum_{j=1}^{s_I} \delta a_j g_j(x(T)) + \\ + \left\langle \sum_{j=1}^{s_0} b_j h_{jx}(x(t_0)) + \psi(t_0), \delta x(t_0) \right\rangle + \sum_{j=1}^{s_0} \delta b_j h_j(x(t_0)). \end{aligned}$$

Пользуясь независимостью вариаций $\delta x(\cdot)$, $\delta u(\cdot)$, $\delta \psi(\cdot)$, δa , δb , из условия стационарности $\delta \mathcal{L} = 0$ имеем

$$\dot{x} = f(x, u, t), \quad \dot{\psi} + H_x(x, u, t, \psi, \psi_0) = 0;$$

$$H_u(x, u, t, \psi, \psi_0) = 0;$$

$$\psi(T) = \psi_0 \Phi_x(x(T)) + \sum_{j=1}^{s_I} a_j g_{jx}(x(T)), \quad g_j(x(T)) = 0, \quad j = \overline{1, s_I};$$

$$\psi(t_0) = - \sum_{j=1}^{s_0} b_j h_{jx}(x(t_0)), \quad h_j(x(t_0)) = 0, \quad j = \overline{1, s_0}.$$

Таким образом, получены почти все основные соотношения теоремы 2, кроме условий (8), (9). Вместо условия (9) мы получили равенство $H_u = 0$, являющееся следствием (9) при $V = E^r$. О том, как получить условие (9), а также все основные соотношения теоремы 3, пользуясь приемами классического вариационного исчисления, читатель сможет прочесть в [66, 221]. Подчеркнем, что приведенные здесь рассуждения, конечно, не могут считаться строгими и являются лишь полезными наводящими соображениями при получении необходимых условий оптимальности.

В заключение заметим, что принцип максимума выше был сформулирован для задач оптимального управления, не содержащих фазовые ограничения, 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 [54, 102, 103, 227, 228], однако, надо отметить, получающаяся отсюда крайняя задача принципа максимума будет иметь весьма 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 быстрогодействия при условиях $\dot{x}^1 = x^2$, $\dot{x}^2 = u(t)$, $x(0) \in S_0$, $x(T) \in S_I$; $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, где $S_0 = \{x \in E^2: h(x) = |x|^2 = 1\}$ или $S_0 = \{x \in E^2: h(x) = |x^2| \leq 1\}$, или $S_0 = \{(0, 0)\}$, или $S_0 = \{(1, 0)\}$, а $S_I = \{x \in E^2, g(x) = x^1 = 0\}$, или $S_I = \{x \in E^2: g(x) = |x|^2 = 4\}$ или $S_I = \{x \in E^2: g(x) = |x|^2 \leq 4\}$.

2. Применить принцип максимума к задаче: $J(u) = \int_0^1 |u(t)|^2 dt \rightarrow \inf$;

$\dot{x}^1 = x^2$, $\dot{x}^2 = u^1(t)$, $\dot{x}^3 = x^4$, $\dot{x}^4 = u^2(t) - g$, $0 \leq t \leq 1$; $x(0) = (-1, 0, 0, 0)$, $x(T) = (0, 0, 0, 0)$. Здесь $x = (x^1, x^2, x^3, x^4)$, $u = (u^1, u^2)$; $g = \text{const} \geq 0$.

3. Применить принцип максимума к задаче о мягкой посадке космического корабля на Луну с минимальной затратой горючего [231, стр. 36, 44, 54]: $J(u) = m(T) \rightarrow \sup$; $\dot{h}(t) = v(t)$, $\dot{v}(t) = -g + u(t)/m(t)$, $\dot{m}(t) = -ku(t)$, $u(t) \in V = \{u \in E^1: 0 \leq u \leq \alpha\}$, $0 \leq t \leq T$; $h(0) = h_0 > 0$, $v(0) = v_0$, $m(0) = m_0 > 0$; $h(T) = 0$, $v(T) = 0$; 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 6—9 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 $u(t) \in V = \{u \in E^1: |u| \leq 1\}$.

6. Применить принцип максимума к задаче: $J(u) = x^2(T) \rightarrow \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq T$; $x(0) = x_0$; момент $T > 0$ задан. Сколько решений имеет эта задача при $x_0 = 0$? $x_0 = T$?

7. Показать, что в задаче $J(u) = \int_0^1 (x^2(t) - u^2(t)) dt \rightarrow \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq 1$; $x(0) = 0$, оптимальное управление не существует, $\inf J(u) = -1$ (см. пример 7.4.2). Что дает здесь применение принципа максимума?

8. Найти минимум функции $J(u) = \int_0^1 \sin u(t) dt$ при условии $\dot{x} = \cos u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq \pi/2\}$, $0 \leq t \leq 1$; $x(0) = 0$, $x(1) = 1$. Показать, что $u(t) \equiv 0$, $0 \leq t \leq 1$, — оптимальное управление, и убедиться в том, что в принципе максимума здесь надо принять $\psi_0 = 0$.

9. Применить принцип максимума к задаче: $J(u) = \int_0^T |x(t)|^2 dt \rightarrow \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq T$; $x(0) = 1$, $x(T) = 0$; момент $T > 0$ задан. Показать, что при $T > 1$ оптимальное управление: $u(t) \equiv -1$, $0 \leq t < 1$; $u(t) \equiv 0$, $1 \leq t \leq T$, на участке $1 \leq t \leq T$ является особым, т. е. его нельзя определить из принципа максимума.

10. Применить принцип максимума к задаче: $J(u) = \int_0^T u(t) \cos(\pi x(t)/2) \times \times dt \rightarrow \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq 1$; $x(0) = 0$.

11. Показать, что задача: $J(u) = \int_0^1 u^2(t) (u(t) - 1)^2 dt \rightarrow \inf$; $\dot{x} = u(t)$, $u(t) \in V = \{u \in E^1: 0 \leq u \leq 1\}$, $0 \leq t \leq T$; $x(0) = 0$, $x(T) = 1$ 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: $J(u) = |x^2(1)|^2 \rightarrow \inf$; $\dot{x}^1 = x^2$, $\dot{x}^2 = u(t)$, $u(t) \in V = \{u \in E^1: |u| \leq 1\}$, $x(0) = (x_0^1, x_0^2)$.

13. Пусть задача оптимального управления автономна (см. § 1), $(u(\cdot), x(\cdot))$ — ее решение, а $\psi(\cdot)$, ψ_0 определены из принципа максимума (теоремы 1—3). Показать, что тогда $H(x(t), u(t), \psi(t), \psi_0) \equiv \text{const}$, $t_0 \leq t \leq T$ [14, 70].

14. Сформулировать принцип максимума для задачи оптимального управления с параметрами:

$J(u(\cdot), w) = \int_{t_0}^T f^0(x(t), u(t), t, w) dt \rightarrow \inf$; $\dot{x}(t) = f(x(t), u(t), t, w)$,
 $t_0 \leq t \leq T$; $x(t_0) \in S_0(t_0)$, $x(T) \in S_1(T)$; $u(t) \in V$, $t_0 \leq t \leq T$; $w \in W$.

где $w = (w^1, \dots, w^k)$ — параметры (они не зависят от времени), W — заданное множество из E^s ; остальные обозначения см. выше. Рассмотрим случаи $W = E^s$ и $W = \{w: q_i(w) = 0, i = \overline{1, k}\}$, где $q_i(w)$ — заданные гладкие функции на E^s , $i = \overline{1, k}$. Указание: ввести новые переменные x^{n+i} посредством условий $\dot{x}^{n+i}(t) = 0$, $t_0 \leq t \leq T$, $x^{n+i}(t_0) = w^i$, $i = \overline{1, k}$ в пространстве $(x^1, \dots, x^{n+k})$ воспользоваться теоремой 2 или 3, а затем исключить переменные x^{n+i} , ψ_{n+i} , $i = \overline{1, k}$ [14].

15. Сформулировать принцип максимума для изопериметрической задачи получающейся из задачи (16) — (19) или (36) — (39) добавлением условий

$$\int_{t_1}^T G^i(x(t), u(t), t) dt = \alpha_i = \text{const}, \quad i = \overline{1, k}.$$

Указание: ввести новые переменные x^{n+i} посредством условий $\dot{x}^{n+i}(t) = G^i(x(t), u(t), t)$, $t_0 \leq t \leq T$, $x^{n+i}(t_0) = 0$, $x^{n+i}(T) = \alpha_i$, $i = \overline{1, k}$, в пространстве $(x^1, \dots, x^{n+k})$ воспользоваться теоремой 2 или 3, а затем исключить переменные x^{n+i} , ψ_{n+i} , $i = \overline{1, k}$ [14].

§ 3. Доказательство принципа максимума для задачи оптимального управления со свободным правым концом

Рассмотрим задачу

$$J(u) = \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T)) \rightarrow \inf; \quad (1)$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T; \quad x(t_0) = x_0, \quad (2)$$

$$u(t) \in V, \quad t_0 \leq t \leq T, \quad (3)$$

где управление $u = u(\cdot)$ кусочно непрерывно на $[t_0, T]$; $u(t) = u(t+0)$, $t_0 \leq t < T$, $u(T) = u(T-0)$; моменты t_0 , T и точка x_0 заданы, V — заданное множество из E^r ; $x = (x^1, \dots, x^n)$, $f = (f^1, \dots, f^n)$, $u = (u^1, \dots, u^r)$. Будем предполагать, что функции $f^0(x, u, t)$, $f(x, u, t)$, $\Phi(x)$ имеют частные производные

$$f_x^0 = (f_{x^1}^0, \dots, f_{x^n}^0), \quad f_x = \{f_{x^i}^j, i, j = \overline{1, n}\}, \quad \Phi_x = (\Phi_{x^1}, \dots, \Phi_{x^n})$$

по переменным $(x^1, \dots, x^n)$ и непрерывны вместе с этими производными по совокупности своих аргументов при $x \in E^n$, $u \in V$, $t \in [t_0, T]$. При этих условиях сформулируем принцип максимума для задачи (1) — (3).

Теорема 1. Пусть $(u(t), x(t))$, $t_0 \leq t \leq T$, — решение задачи (1) — (3). Тогда

$$\sup_{u \in V} H(x(t), u, t, \psi(t)) = H(x(t), u(t), t, \psi(t)), \quad t_0 \leq t \leq T, \quad (4)$$

где

$$H(x, u, t, \psi) = -f^0(x, u, t) + \langle \psi, f(x, u, t) \rangle, \quad (5)$$

а функция $\psi(t) = \psi(t, u)$, $t_0 \leq t \leq T$, является решением следующей сопряженной задачи:

$$\dot{\psi}(t) = -H_x(x, u, t, \psi(t)) \Big|_{\substack{u=u(t) \\ x=x(t)}} = \\ = f_x^0(x(t), u(t), t) - (f_x(x(t), u(t), t))^T \psi(t), \quad t_0 \leq t \leq T, \quad (6)$$

$$\psi(T) = -\Phi_x(x(T)). \quad (7)$$

Эта теорема является частным случаем теоремы 2.2 и несколько уточняет ее — здесь $\psi_0 = -1$, и поэтому условие (2.8) всегда выполнено, а условие трансверсальности (2.24) обретает вид (7).

Доказательство теоремы 1 мы проведем при дополнительном предположении, считая, что функции f , f_x , f , f_x^0 , Φ_x удовлетворяют условию Липшица по переменным (x, u) , т. е:

$$|f(x + \Delta x, u + h, t) - f(x, u, t)| \leq L(|\Delta x| + |h|), \quad (8)$$

$$\|f_x(x + \Delta x, u + h, t) - f_x(x, u, t)\| \leq L(|\Delta x| + |h|), \quad (9)$$

$$|f_x^0(x + \Delta x, u + h, t) - f_x^0(x, u, t)| \leq L(|\Delta x| + |h|), \quad (10)$$

$$|\Phi_x(x + \Delta x) - \Phi_x(x)| \leq L|\Delta x| \quad (11)$$

при всех (x, u, t) , $(x + \Delta x, u + h, t) \in E^n \times V \times [t_0, T]$; константы Липшица для всех функций обозначены одной буквой L , так как конкретные величины этих констант ниже не понадобятся. Как следует из теоремы 1.1, при выполнении условия (8) задача Коши (2) имеет решение $x(\cdot, u)$, определенное на всем отрезке $[t_0, T]$ для всех кусочно непрерывных управлений, удовлетворяющих условию (3).

Управлению $u = u(\cdot)$, $u(t) \in V$, $t_0 \leq t \leq T$, дадим кусочно непрерывное приращение $h(t)$, $t_0 \leq t \leq T$, такое, что $u + h = u(t) + h(t) \in V$, $t_0 \leq t \leq T$. Тогда траектория $x(t) = x(t, u)$, $t_0 \leq t \leq T$, получит приращение

$$\Delta x(t) = x(t, u + h) - x(t, u), \quad t_0 \leq t \leq T.$$

Из условий (2) следует, что приращение $\Delta x(t)$ удовлетворяет условиям

$$\Delta \dot{x}(t) = f(x(t) + \Delta x(t), u(t) + h(t), t) - f(x(t), u(t), t), \quad (12)$$

$$t_0 \leq t \leq T;$$

$$\Delta x(t_0) = 0.$$

Покажем, что

$$|\Delta x(t)| = |x(t, u + h) - x(t, u)| \leq C_1 \int_{t_0}^T |h(t)| dt, \quad t_0 \leq t \leq T, \quad (13)$$

где $C_1 = Le^{L(T-t_0)} = \text{const}$.

Для этого нам понадобится одно утверждение, которое часто приводится в курсах лекций по обыкновенным дифференциальным уравнениям и в литературе известно как неравенство Гронулла.

Л е м м а 1. Пусть функция $\varphi(t)$, $b(t)$ неотрицательны и непрерывны на отрезке $t_0 \leq t \leq T$, $a = \text{const}$. Пусть

$$\dot{\varphi}(t) \leq a \int_{t_0}^t \varphi(\tau) d\tau + b(t), \quad t_0 \leq t \leq T. \quad (14)$$

Тогда

$$0 \leq \varphi(t) \leq a \int_{t_0}^t b(\tau) e^{a(t-\tau)} d\tau + b(t), \quad t_0 \leq t \leq T. \quad (15)$$

В частности, если $b(t) \equiv b = \text{const} \geq 0$, то

$$0 \leq \varphi(t) \leq b e^{a(t-t_0)}, \quad t_0 \leq t \leq T. \quad (16)$$

Если же

$$\dot{\varphi}(t) \leq a \int_t^T \varphi(\tau) d\tau + b(t), \quad t_0 \leq t \leq T,$$

то

$$0 \leq \varphi(t) \leq a \int_t^T b(\tau) e^{a(\tau-t)} d\tau + b(t), \quad t_0 \leq t \leq T,$$

а при $b(t) \equiv b = \text{const} \geq 0$ будем иметь

$$0 \leq \varphi(t) \leq b e^{a(T-t)}, \quad t_0 \leq t \leq T.$$

Доказательство. Положим $R(t) = a \int_{t_0}^t \varphi(\tau) d\tau$. Заметим, что $R(t_0) = 0$, $R(t) \geq 0$, $\dot{R}(t) = a\varphi(t)$, $t_0 \leq t \leq T$. С учетом (14) имеем $\dot{R}(t) \leq aR(t) + ab(t)$, т. е. $\dot{R}(t) - aR(t) \leq ab(t)$, $t_0 \leq t \leq T$. Умножая обе части последнего неравенства на $e^{-a(t-t_0)}$, получим

$$\frac{d}{dt} (R(t) e^{-a(t-t_0)}) \leq ab(t) e^{-a(t-t_0)}, \quad t_0 \leq t \leq T.$$

Интегрирование этого неравенства от t_0 до t с учетом $R(t_0) = 0$ приводит к оценке

$$R(t) e^{-a(t-t_0)} \leq a \int_{t_0}^t b(\tau) e^{-a(\tau-t_0)} d\tau,$$

т. е.

$$R(t) \leq a \int_{t_0}^t b(\tau) e^{a(t-\tau)} d\tau, \quad t_0 \leq t \leq T.$$

Подставив эту оценку в правую часть (14), сразу получим требуемое неравенство (15). Если $b(t) = b = \text{const}$, то непосредственно вычисляя интеграл в правой части (15), придем к оценке (16).

Второе утверждение леммы доказывается аналогично с помощью вспомогательной функции $R(t) = a \int_t^T \varphi(\tau) d\tau$, $t_0 \leq t \leq T$.

Лемма 1 доказана.

Из соотношений (12) для $\Delta x(t)$ с учетом условия (8) имеем

$$\begin{aligned} |\Delta x(t)| &= \left| \int_{t_0}^t [f(x(\tau) + \Delta x(\tau), u(\tau) + h(\tau), \tau) - f(x(\tau), u(\tau), \tau)] d\tau \right| \leq \\ &\leq L \int_{t_0}^t |\Delta x(\tau)| d\tau + L \int_{t_0}^T |h(\tau)| d\tau. \end{aligned}$$

Отсюда, приняв в (14) $\varphi(t) = |\Delta x(t)|$, $a = L$, $b(T) = L \int_{t_0}^T |h(\tau)| d\tau = \text{const}$, из леммы 1 получим оценку (13).

Докажем, что приращение $\Delta J = J(u + h) - J(u)$ функции (1) представимо в виде

$$\begin{aligned} \Delta J &= - \int_{t_0}^T [H(x(t), u(t) + h(t), t, \psi(t)) - \\ &\quad - H(x(t), u(t), t, \psi(t))] dt + R, \end{aligned} \quad (17)$$

где остаточный член R удовлетворяет неравенству

$$|R| \leq C_2 \left(\int_{t_0}^T |h(t)| dt \right)^2, \quad C_2 = \text{const} \geq 0. \quad (18)$$

Так как $\Phi(x + \Delta x) - \Phi(x) = \langle \Phi_x(x + \theta \Delta x), \Delta x \rangle$, $0 \leq \theta \leq 1$, то из (1) получаем

$$\begin{aligned} \Delta J &= \int_{t_0}^T [f^0(x(t) + \Delta x(t), u(t) + h(t), t) - f^0(x(t), u(t), t)] dt + \\ &\quad + \langle \Phi_x(x(T)), \Delta x(T) \rangle + R_1, \end{aligned} \quad (19)$$

где

$$R_1 = \langle \Phi_x(x(T) + \theta \Delta x(T)) - \Phi_x(x(T)), \Delta x(T) \rangle.$$

Из (11), (13) следует оценка

$$|R_1| \leq L |\Delta x(T)|^2 \leq LC_1^2 \left(\int_{t_0}^T |h(t)| dt \right)^2. \quad (20)$$

Преобразуем второе слагаемое в (19). С учетом соотношений (6), (7), (12) имеем

$$\begin{aligned} \langle \Phi_x(x(T)), \Delta x(T) \rangle &= - \langle \psi(T), \Delta x(T) \rangle = \\ &= - \int_{t_0}^T \frac{d}{dt} \langle \psi(t), \Delta x(t) \rangle dt = - \int_{t_0}^T [\langle \dot{\psi}(t), \Delta x(t) \rangle + \langle \psi(t), \Delta \dot{x}(t) \rangle] dt = \\ &= - \int_{t_0}^T \langle \psi(t), f(x(t) + \Delta x(t), u(t) + h(t), t) - f(x(t), u(t), t) \rangle dt + \\ &\quad + \int_{t_0}^T \langle H_x(x(t), u(t), t, \psi(t)), \Delta x(t) \rangle dt. \end{aligned}$$

Подставим полученное выражение в (19), откуда с помощью функции (5) будем иметь

$$\begin{aligned} \Delta J &= - \int_{t_0}^T [H(x(t) + \Delta x(t), u(t) + h(t), t, \psi(t)) - \\ &\quad - H(x(t), u(t), t, \psi(t))] dt + \\ &\quad + \int_{t_0}^T \langle H_x(x(t), u(t), t, \psi(t)), \Delta x(t) \rangle dt + R_1. \quad (21) \end{aligned}$$

Из формулы конечных приращений следует

$$\begin{aligned} H(x + \Delta x, u + h, t, \psi) &= H(x, u + h, t, \psi) + \\ &\quad + \langle H_x(x + \theta \Delta x, u + h, t, \psi), \Delta x \rangle, \quad 0 \leq \theta \leq 1. \end{aligned}$$

Отсюда и из (21) получим формулу (17), где $R = R_1 + R_2$,

$$R_2 = - \int_{t_0}^T \langle H_x(x + \theta \Delta x, u + h, t, \psi) - H_x(x, u, t, \psi), \Delta x \rangle dt.$$

Так как $H_x(x, u, t, \psi) = -\dot{f}_x^0(x, u, t) + (f_x(x, u, t))^T \psi$, то с учетом условий (9), (10) для R_2 имеем оценку

$$|R_2| \leq L(1 + \max_{t_0 \leq t \leq T} |\psi(t, u)|) \int_{t_0}^T (|\Delta x(t)| + |h(t)|) |\Delta x(t)| dt.$$

Отсюда и из оценки (13) следует

$$|R_2| \leq L(1 + \max_{t_0 \leq t \leq T} |\psi(t, u)|) [C_1^2(T - t_0) + C_1] \left(\int_{t_0}^T |h(t)| dt \right)^2. \quad (22)$$

Так как $R = R_1 + R_2$, то из (20), (22) получаем требуемую оценку (18).

Завершим доказательство теоремы 1. Пусть $u = u(\cdot)$ — оптимальное управление задачи (1) — (3). Тогда

$$\Delta J = J(u + h) - J(u) \geq 0$$

для всех кусочно непрерывных $h = h(\cdot)$, для которых $u(t) + h(t) \in V$, $t_0 \leq t \leq T$. Возьмем произвольную точку $v \in V$, произвольные моменты времени $t, t + \varepsilon \in [t_0, T]$, $\varepsilon > 0$, и рассмотрим приращение $h(\cdot)$ специального вида:

$$h(\tau) = \begin{cases} v - u(\tau), & t \leq \tau \leq t + \varepsilon, \\ 0, & \tau \in [t_0, T] \setminus [t, t + \varepsilon]. \end{cases} \quad (23)$$

Приращения вида (23) принято называть игольчатыми вариациями управления. Этим названием подчеркивается тот факт, что при малых $\varepsilon > 0$ приращение $h(\tau)$ будет отлично от нуля лишь на отрезке малой длины. Нетрудно видеть, что приращение (23) кусочно непрерывно и $u(t) + h(t) \in V$ при всех $t \in [t_0, T]$. Для приращения (23) из (17), (18) имеем

$$\Delta J = - \int_t^{t+\varepsilon} g(\tau) d\tau + R \geq 0, \quad |R| \leq \varepsilon C_2 \int_t^{t+\varepsilon} |h(\tau)|^2 d\tau; \quad (24)$$

здесь

$$g(\tau) = H(x(\tau), v, \tau, \psi(\tau)) - H(x(\tau), u(\tau), \tau, \psi(\tau)), \quad t \leq \tau \leq t + \varepsilon,$$

а оценка для R получена из оценки (18) применением неравенства Коши — Буняковского. Так как функция $u(\cdot)$ кусочно непрерывна, $u(t) = u(t + 0)$, а $x(\tau)$, $\psi(\tau)$, $H(x, u, \tau, \psi)$ — непрерывные функции своих аргументов, то, взяв $\varepsilon > 0$ достаточно малым, можем считать, что функция $g(\tau)$ непрерывна на отрезке

$[t, t + \varepsilon]$. Тогда по теореме о среднем $\int_t^{t+\varepsilon} g(\tau) d\tau = \varepsilon g(t + \theta\varepsilon)$,

$0 \leq \theta \leq 1$. Поэтому из (24) будем иметь

$$0 \leq \Delta J = -\varepsilon g(t + \theta\varepsilon) + R \leq -\varepsilon g(t + \theta\varepsilon) + \varepsilon C_2 \int_t^{t+\varepsilon} |h(\tau)|^2 d\tau.$$

Далее поделив на $\varepsilon > 0$ и устремив $\varepsilon \rightarrow +0$, отсюда получим $\lim_{\varepsilon \rightarrow +0} g(t + \theta\varepsilon) = g(t) \leq 0$, т. е.

$$H(x(t), v, t, \psi(t)) \leq H(x(t), u(t), t, \psi(t)).$$

В силу произвольности $v \in V$ отсюда имеем условие максимума (4) для всех t , $t_0 \leq t < T$. Для доказательства (4) при $t = T$ нужно взять приращение

$$h(\tau) = \begin{cases} v - u(\tau), & T - \varepsilon \leq \tau \leq T, \\ 0, & \tau \in [t_0, T] \setminus [T - \varepsilon, T], \end{cases}$$

и провести аналогичные рассуждения. Теорема 1 доказана.

З а м е ч а н и е 1. До сих пор значения кусочно непрерывного управления в точках разрыва мы определяли условием $u(t) = u(t+0)$ при $t_0 \leq t < T$ и $u(T) = u(T-0)$. Поэтому, если ввести функцию

$$H(t) = \sup_{u \in V} H(x(t), u, t, \psi(t)), \quad t_0 \leq t \leq T,$$

то условие максимума (4) можно записать в виде

$$\begin{aligned} H(t) &= H(x(t), u(t+0), t, \psi(t)), & t_0 \leq t < T; \\ H(T) &= H(x(T), u(T-0), T, \psi(T)). \end{aligned}$$

С другой стороны, если бы мы приняли $u(t) = u(t-0)$ при $t_0 < t \leq T$ и $u(t_0) = u(t_0+0)$, то рассуждая так же, как при доказательстве теоремы 1, получили бы

$$\begin{aligned} H(t) &= H(x(t), u(t-0), t, \psi(t)), & t_0 < t \leq T; \\ H(t_0) &= H(x(t_0), u(t_0+0); t_0, \psi(t_0)). \end{aligned}$$

Отсюда следует, что функция $H(t)$ непрерывна на отрезке $[t_0, T]$ и условие максимума (4) остается справедливым для всех $t \in [t_0, T]$ при любом из двух указанных выше способах доопределения $u(\cdot)$ в точках разрыва.

Оказывается, указанное свойство непрерывности функции $H(t)$ имеет место и для более общих, чем (1)–(3), задач оптимального управления. В самом деле, пусть $(u(t), x(t))$, $t_0 \leq t \leq T$, — оптимальные управления и траектория в задаче (2.36)–(2.39), а $\psi_0, \psi(t)$, $t_0 \leq t \leq T$, определены из теоремы 2.3. Положим

$$H(t) = \sup_{u \in V} H(x(t), u, t, \psi(t), \psi_0), \quad t_0 \leq t \leq T. \quad (25)$$

Поскольку функции $x(t)$, $\psi(t)$, $H(x, u, t, \psi, \psi_0)$ непрерывны по совокупности своих аргументов, а $u(t)$ — кусочно непрерывна

на отрезке $[t_0, T]$, то из условия максимума (2.9)

$$H(t) = H(x(t), u(t), t, \psi(t), \psi_0)$$

следует непрерывность $H(t)$ во всех точках t , в которых непрерывно управление $u(\cdot)$. Остается рассмотреть точки разрыва управления $u(\cdot)$. Пусть τ — одна из таких точек и пусть

$$H(\tau) = H(x(\tau), v, \tau, \psi(\tau), \psi_0), \quad v \in V,$$

здесь v — та точка, в которой достигается верхняя грань из (25) при $t = \tau$, и ясно, что пока v никак не связано со значениями $u(\tau + 0)$ или $u(\tau - 0)$. Из определения (25) функции $H(t)$ следует, что $H(\tau) \geq H(x(\tau), u(t), \tau, \psi(\tau), \psi_0)$, $H(t) \geq H(x(t), v, t, \psi(t), \psi_0)$. Поэтому $H(x(t), v, t, \psi(t), \psi_0) - H(x(\tau), v, \tau, \psi(\tau), \psi_0) \leq H(t) - H(\tau) \leq H(x(t), u(t), t, \psi(t), \psi_0) - H(x(\tau), u(t), \tau, \psi(\tau), \psi_0)$ во всех точках t , в которых $u(\cdot)$ непрерывно. Отсюда, переходя к пределу при $t \rightarrow \tau + 0$ и $t \rightarrow \tau - 0$, получим $H(\tau) = H(\tau + 0) = H(x(\tau), u(\tau + 0), \tau, \psi(\tau), \psi_0) = H(\tau - 0) = H(x(\tau), u(\tau - 0), \tau, \psi(\tau), \psi_0)$. Непрерывность функции $H(t)$ на отрезке $[t_0, T]$ доказана. Кроме того выяснено, что для справедливости условия максимума (2.9) оптимальное управление в точках своего разрыва может быть доопределено предельным значением либо справа, либо слева.

З а м е ч а н и е 2. Оценка (18) остаточного члена в формуле приращения (17) имеет второй порядок малости относительно $\|h\|_{L_1}$ — это было получено с помощью условий (8) — (11). Однако, как было замечено выше, теорема 1 верна и в том случае, если вместо условий (8) — (11) потребовать лишь непрерывность функций $f, f_x, f^0, f^0_x, \Phi, \Phi_x$ по совокупности своих аргументов. В самом деле, при доказательстве теоремы мы пользовались лишь игольчатыми вариациями (23). А для таких вариаций оценка $\max_{t_0 \leq t \leq T} |\Delta x(t)| = O(\varepsilon)$, аналогичная оценке (13), является

следствием теоремы о непрерывной зависимости решения задачи (2) от начального условия и правой части [54, 70, 74, 193].

Остается заметить, что тогда в формуле (17) вместо оценки (18) будет $R \equiv o(\varepsilon)$, и провести рассуждения, аналогичные рассуждениям из приведенного выше доказательства теоремы 1.

§ 4. О методах решения краевой задачи принципа максимума

Выше было показано, что задача оптимального управления (см. задачи (2.1) — (2.4), (2.16) — (2.19), (2.36) — (2.39)) с помощью принципа максимума может быть сведена к краевой задаче, состоящей из системы $2n$ дифференциальных уравнений

$$\dot{x} = f(x, u(x, t, \psi, \psi_0), t), \quad \dot{\psi} = -H_x(x, u(x, t, \psi, \psi_0), t, \psi, \psi_0), \quad (1) \\ t_0 \leq t \leq T,$$

относительно переменных $(x, \psi) = (x^1, \dots, x^n, \psi_1, \dots, \psi_n)$, граничных условий, которые в общем случае могут быть записаны в виде

$$P_i(t_0, x(t_0), \psi(t_0), \psi_0, b_1, \dots, b_{s_0}) = 0, \quad i = \overline{1, n + s_0 + 2} \quad (2)$$

(см. условия (2.14) при $\bar{t} = t_0$, (2.30) — (2.35), (2.45) — (2.47) на левом конце траектории),

$$Q_i(T, x(T), \psi(T), \psi_0, a_1, \dots, a_{s_I}) = 0, \quad i = \overline{1, n + s_I + 1} \quad (3)$$

(см. условия (2.24) — (2.29), (2.42) — (2.44) на правом конце траектории), и условия максимума (2.9)

$$\sup_{u \in V} H(x, u, t, \psi, \psi_0) = H(x, u(x, t, \psi, \psi_0), t, \psi, \psi_0), \quad (4)$$

из которого определяются функции $u = u(x, t, \psi, \psi_0)$, используемые в (1).

1. Для численного решения такой краевой задачи могут быть использованы известные методы, такие, как метод стрельбы, метод прогонки, различные итерационные методы [2, 40, 129, 211].

Кратко остановимся на методе стрельбы, с помощью которого краевая задача (1) — (3) сводится к задаче Коши и задаче минимизации функции конечного числа переменных. Допустим, что из соотношений

$$P_i(t_0, x, \psi, \psi_0, b_1, \dots, b_{s_0}) = 0, \quad i = \overline{1, n + s_0 + 2},$$

нам удалось выразить параметры $\psi_0, t_0, b_1, \dots, b_{s_0}$ и какие-то n координат вектора (x, ψ) через остальные координаты этого вектора. Пусть для определенности

$$t_0 = t_0(x), \quad \psi = \omega(x), \quad \psi_0 = \psi_0(x), \quad b_j = b_j(x), \quad j = \overline{1, s_0}. \quad (5)$$

Зададим каким-либо образом $x_0 = \lambda$, где $\lambda = (\lambda_1, \dots, \lambda_n)$ — параметры стрельбы. Тогда из (5) при $x = \lambda$ однозначно определяются $t_0 = t_0(\lambda)$, $\psi = \omega(\lambda)$, $\psi_0 = \psi_0(\lambda)$, $b_j = b_j(\lambda)$, $j = \overline{1, s_0}$. Сделаем пробный «выстрел» с некоторым параметром λ . А именно, решая задачу Коши для системы (1) с $\psi_0 = \psi_0(\lambda)$ и с начальным условием $x(t_0) = \lambda$, $\psi(t_0) = \omega(\lambda)$, найдем ее решение $x(t, \lambda)$, $\psi(t, \lambda)$, $t \geq t_0$. В силу выбора величин $t_0(\lambda)$, $\omega(\lambda)$, $\psi_0(\lambda)$, $b_j(\lambda)$ это решение удовлетворяет левым граничным условиям (2) при любых λ . Подставим в (3) $\psi_0 = \psi_0(\lambda)$, $x(T, \lambda)$, $\psi(T, \lambda)$ при некотором T и какие-либо значения параметров $a_1, \dots, a_{s_I}$. В случае удачного «выстрела», т. е. удачного выбора параметров $\lambda, T, a_1, \dots, a_{s_I}$, мы получим

$$Q_i(T, x(T, \lambda), \psi(T, \lambda), \psi_0(\lambda), a_1, \dots, a_{s_I}) = 0, \quad i = \overline{1, n + s_I + 1}. \quad (6)$$

Иначе говоря, для $n + s_l + 1$ параметров $\lambda, T, a_1, \dots, a_{s_l}$ имеем систему (6) из $n + s_l + 1$ уравнений. Для решения системы (6) могут быть использованы известные методы [2, 39, 129]. В частности, систему (6) можно записать в виде задачи минимизации $\Phi(T, \lambda, a) =$

$$= \sum_{i=1}^{n+s_l+1} Q_i^2(T, x(T, \lambda), \psi(T, \lambda), \psi_0(\lambda), a_1, \dots, a_{s_l}) \rightarrow \inf, \quad (7)$$

которая может решаться, например, одним из методов, описанных выше, в гл. 5. Подчеркнем, что вычисление значения функции (7) в некоторой точке (T, λ, a) , вообще говоря, весьма трудоемко — для этого нужно решить задачу Коши для системы (1). При численном решении задачи Коши можно пользоваться методами Эйлера, Адамса, Рунге — Кутты и др. [2, 40, 129, 211]. Метод стрельбы описан.

Следует заметить, что краевая задача принципа максимума (1) — (4) имеет ряд специфических особенностей, затрудняющих применение упомянутых стандартных методов решения краевых задач. Дело в том, что функция $u = u(x, t, \psi, \psi_0)$, определяемая из условия (4), вообще говоря, нелинейно зависит от своих аргументов, и поэтому дифференциальные уравнения (1), как правило, нелинейны даже в том случае, когда исходная система $\dot{x} = f(x, u, t)$ была линейна относительно (x, u) . Кроме того, функция $u = u(x, t, \psi, \psi_0)$, вообще говоря, не всюду дифференцируема и даже может быть разрывной (например, возможна функция вида $u = \text{sign } \psi$; см. примеры 2.4—2.8), что влечет за собой плохие аналитические свойства правых частей системы (1). Краевая задача (1) — (4) существенно усложняется в тех случаях, когда из условия (4) функция $u = u(x, t, \psi, \psi_0)$ 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. [12, 49, 171, 183, 244]. Некоторые приемы преодоления возможной неустойчивости в краевых задачах принципа максимума описаны, например, в [12, 56, 171].

2. Остановимся на одном специальном итерационном методе [244] решения краевой задачи принципа максимума, соответствующей задаче оптимального управления со свободным правым концом, которая была рассмотрена в § 3. В этом случае, как это следует из теоремы 3.1, краевая задача принципа максимума

имеет вид

$$\dot{x} = f(x, u, t), \quad t_0 \leq t \leq T; \quad x(t_0) = x_0, \quad (8)$$

$$\dot{\psi} = -H_x(x, u, t, \psi), \quad t_0 \leq t \leq T; \quad \psi(T) = -\Phi_x(x(T)), \quad (9)$$

где $u = u(x, t, \psi)$ определяется из условия

$$\sup_{u \in V} H(x, u, t, \psi) = H(x, u(x, t, \psi), t, \psi). \quad (10)$$

Пусть $u_0 = u_0(t)$, $t_0 \leq t \leq T$, — начальное приближение, $u_0(t) \in V$, $t_0 \leq t \leq T$. Допустим, что уже найдено k -е приближение $u_k = u_k(\cdot)$, $u_k(t) \in V$, $t_0 \leq t \leq T$. Тогда, решая задачу Коши (8) при $u = u_k(\cdot)$, находим $x_k(\cdot) = x(\cdot, u_k)$. Затем, приняв в (9) $u = u_k(\cdot)$, $x = x_k(\cdot)$, из (9) определяем $\psi_k(\cdot) = \psi(\cdot, u_k)$. Наконец, из условия

$$\sup_{u \in V} H(x_k(t), u, t, \psi_k(t)) = H(x_k(t), u_{k+1}(t), t, \psi_k(t)), \quad (11)$$

$$t_0 \leq t \leq T,$$

получаем следующее приближение $u_{k+1} = u_{k+1}(\cdot)$ и т. д. Метод описан. На одном шаге этого метода нужно решить две задачи Коши — задачи (8) и (9) — и конечномерную задачу максимизации (11) при каждом $t \in [t_0, T]$. Будем предполагать, что все получаемые этим методом функции $u_k(\cdot)$ кусочно непрерывны на отрезке $[t_0, T]$.

Может случиться, что при некотором k имеет место равенство $u_k(t) = u_{k+1}(t)$, $t_0 \leq t \leq T$. Согласно (11) это будет означать, что управление $u_k(\cdot)$ удовлетворяет принципу максимума (теорема 3.1) и, следовательно, является подозрительным на оптимальность. В этом случае итерационный процесс прекращается, а управление $u_k(\cdot)$ при необходимости подвергается дальнейшему исследованию для выяснения того, будет ли оно на самом деле оптимальным.

Рассмотрим случай, когда $u_{k+1}(\tau) \neq u_k(\tau)$ при некотором $\tau \in [t_0, T]$ и, кроме того,

$$H(x_k(\tau), u_{k+1}(\tau), \tau, \psi'_k(\tau)) - H(x_k(\tau), u_k(\tau), \tau, \psi_k(\tau)) \geq 2\alpha > 0. \quad (12)$$

Можно ли ожидать, что тогда $J(u_{k+1}) < J(u_k)$? Для ответа на этот вопрос обратимся к формуле приращения (3.17) с оценкой остаточного члена (3.18). Из этой формулы при $u = u_k$, $h = u_{k+1} - u_k$ имеем

$$J(u_{k+1}) - J(u_k) =$$

$$= - \int_{t_0}^T [H(x_k(t), u_{k+1}(t), t, \psi_k(t)) - H(x_k(t), u_k(t), t, \psi_k(t))] dt + R, \quad (13)$$

$$|R| \leq C_2 \left(\int_{t_0}^T |u_{k+1}(t) - u_k(t)| dt \right)^2.$$

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 (12) следует существование отрезка $[\tau, \tau + \varepsilon] \subseteq [t_0, T]$, $\varepsilon > 0$ (или отрезка $[\tau - \varepsilon, \tau]$ при $\tau = T$), такого, что $g(t) = H(x_k(t), u_{k+1}(t), t, \psi_k(t)) - H(x_k(t), u_k(t), t, \psi_k(t)) \geq \alpha > 0$ (14)

при всех t , $\tau \leq t \leq \tau + \varepsilon$. Из (11) и (14) заключаем, что первое слагаемое из формулы (13) для $J(u_{k+1}) - J(u_k)$ отрицательно. Это значит, что если остаточный член R в (13) 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 $u_{k+1}(\cdot)$ можно указать другое «лучшее» управление $v_{k+1}(\cdot)$, для которого $J(v_{k+1}) < J(u_k)$. А именно, возьмем игольчатую вариацию

$$h_k(t) = \begin{cases} u_{k+1}(t) - u_k(t), & \tau \leq t \leq \tau + \varepsilon, \\ 0, & t \in [t_0, T] \setminus [\tau, \tau + \varepsilon], \end{cases}$$

и положим $v_{k+1}(t) = u_k(t) + h_k(t)$, $t_0 \leq t \leq T$. Из формул (3.17), (3.18) при $u = u_k$, $h = h_k$ имеем

$$J(v_{k+1}) - J(u_k) = - \int_{\tau}^{\tau+\varepsilon} g(t) dt + R,$$

$$|R| \leq C_2 \left(\int_{\tau}^{\tau+\varepsilon} |h(t)| dt \right)^2 \leq \varepsilon C_2 \int_{\tau}^{\tau+\varepsilon} |h(t)|^2 dt.$$

Отсюда и из (14) следует $J(v_{k+1}) - J(u_k) \leq \varepsilon \left[-\alpha + C_2 \int_{\tau}^{\tau+\varepsilon} |h(t)|^2 dt \right] < 0$, если $\varepsilon > 0$ взять достаточно малым. Таким образом, при выполнении условия (12) всегда можно подобрать управление $v_{k+1}(\cdot)$ 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 (3.1) — (3.3). Например, если функцию (3.1) нужно минимизировать при условиях (3.2), (3.3) и закрепленном правом конце $x(T) = x_I$ или фазовом ограничении вида $x^1(t) \leq 1$, $t_0 \leq t \leq T$, и т. п., то можно ввести штрафную функцию $P_k(u) = A_k |x(T) - x_I|^2$ 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 (3.2), (3.3), $\lim_{k \rightarrow \infty} A_k = +\infty$. Более подробно об описанном итерационном методе, о приемах ускорения его сходимости, о его возможностях, см. в [244].

§ 5. Связь между принципом максимума и классическим вариационным исчислением

Основной задачей классического вариационного исчисления, как известно [80, 136] является следующая задача: среди всех непрерывных кривых $x = x(t)$, $t_0 \leq t \leq T$, 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)

$$J = \int_{t_0}^T f^0(x(t), \dot{x}(t), t) dt$$

минимальное значение. Здесь $x(t) = (x^1(t), \dots, x^n(t))$, S_0 и S_1 — заданные множества в E^n . Будем предполагать, что функция $f^0(x, u, t)$ 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 закрепленного 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: $S_1 = E^n$, T — задано, либо является подвижным и лежит на заданной гладкой кривой

$$S_1 = S_1(T) = \{x \in E^n: g(x, T) = x - \varphi(T) = 0\},$$

$$T \in \mathbb{R} = \{-\infty < t < +\infty\}.$$

Обозначим $\dot{x}(t) = u(t)$ и запишем рассматриваемую задачу в эквивалентном виде как задачу оптимального управления:

$$J(u) = \int_{t_0}^T f^0(x(t), u(t), t) dt \rightarrow \inf;$$

$$\dot{x}(t) = u(t), \quad t_0 \leq t \leq T;$$

$$x(t_0) = x_0, \quad x(T) \in S_1(T).$$

Для исследования этой задачи воспользуемся принципом максимума Понтрягина. Согласно теореме 2.3 и замечанию 1 к ней мы должны выписать функцию Гамильтона — Понтрягина

$$H(x, u, t, \psi, \psi_0) = \psi_0 f^0(x, u, t) + \langle \psi, u \rangle \quad (1)$$

и сопряженную систему

$$\dot{\psi} = -H_x = -\psi_0 f_x^0(x, u, t), \quad \psi_0 = \text{const} \leq 0. \quad (2)$$

Для решения $(u(t), x(t))$, $t_0 \leq t \leq T$, рассматриваемой задачи должно выполняться необходимое условие

$$H(x(t), u(t), t, \psi(t), \psi_0) = \sup_{u \in E^n} H(x(t), u, t, \psi(t), \psi_0), \quad (3)$$

$$t_0 \leq t \leq T,$$

где $\psi(t)$ — решение системы (2) при $u = u(t)$, $x = x(t, u)$, $t_0 \leq t \leq T$. Так как в данном случае множество V совпадает со всем пространством E^n , то условие (3) может соблюдаться лишь в стационарной точке, т. е.

$$H_u = \psi_0 f_u^0(x(t), u(t), t) + \psi(t) = 0, \quad t_0 \leq t \leq T. \quad (4)$$

Отсюда ясно, что $\psi_0 \neq 0$, так как при $\psi_0 = 0$ из (4) получим $\psi(t) \equiv 0$, что противоречит условию (2.8). Следовательно, можно считать, что $\psi_0 = -1$. Тогда соотношения (1) — (4) переписутся соответственно в виде

$$H(x, u, t, \psi) = -f^0(x, u, t) + \langle \psi, u \rangle, \quad (5)$$

$$\dot{\psi}(t) = f_x^0(x(t), u(t), t), \quad t_0 \leq t \leq T, \quad (6)$$

$$H(x(t), u(t), t, \psi(t)) = \sup_{u \in E^n} H(x(t), u, t, \psi(t)), \quad t_0 \leq t \leq T, \quad (7)$$

$$\psi(t) = f_u^0(x(t), u(t), t), \quad t_0 \leq t \leq T, \quad (8)$$

Из уравнения (6) имеем: $\psi(t) = \int_{t_0}^t f_x^0(x(t), u(t), t) dt + \psi(t_0)$.

С учетом (8) отсюда получаем

$$f_u^0(x(t), u(t), t) = \int_{t_0}^t f_x^0(x(\tau), u(\tau), \tau) d\tau + \psi(t_0), \quad t_0 \leq t \leq T. \quad (9)$$

Уравнение (9) называется уравнением Эйлера в интегральной форме; здесь $u(t) = \dot{x}(t)$, $t_0 \leq t \leq T$. Если (9) продифференцировать по t , то получим уравнение Эйлера классического вариационного исчисления в дифференциальной форме

$$\frac{d}{dt} (f_u^0(x(t), u(t), t)) - f_x^0(x(t), u(t), t) = 0, \quad u(t) = \dot{x}(t), \quad t_0 \leq t \leq T.$$

Далее, необходимым условием достижения функцией $H(x(t), u, t, \psi(t))$ максимума при $u = u(t)$ является неположительность

следующей квадратичной формы:

$$\sum_{i,j=1}^n H_{u^i u^j}(x(t), u(t), t, \psi(t)) \xi_i \xi_j \leq 0$$

при любых

$$\xi = (\xi_1, \xi_2, \dots, \xi_n), \quad t_0 \leq t \leq T.$$

Отсюда, учитывая выражение (5) для H , имеем

$$\sum_{i,j=1}^n f_{u^i u^j}^0(x(t), u(t), t) \xi_i \xi_j \geq 0, \quad \xi \in E^n, \quad t_0 \leq t \leq T. \quad (10)$$

Условие (10) называется необходимым условием Лежандра. В частности, при $n = 1$ отсюда имеем:

$$f_{uu}^0(x(t), u(t), t) \geq 0, \quad t_0 \leq t \leq T.$$

Теперь выведем необходимое условие Вейерштрасса. Для этого перепишем условие (7) с учетом (5), (8) в следующем виде:

$$\begin{aligned} 0 &\leq H(x(t), u(t), t, \psi(t)) - H(x(t), v, t, \psi(t)) = \\ &= f^0(x(t), v, t) - f^0(x(t), u(t), t) - \langle v - u(t), f_u^0(x(t), u(t), t) \rangle. \end{aligned} \quad (11)$$

Это неравенство справедливо при любых $v \in E^n$, $t \in [t_0, T]$, если $(u(t), x(t))$, $t_0 \leq t \leq T$, — 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. Известное в классическом вариационном исчислении необходимое условие Вейерштрасса

$$E(t, x(t), u(t), v) \geq 0, \quad t_0 \leq t \leq T, \quad v \in E^n,$$

является следствием неравенства (11). Далее, из принципа максимума (см. теоремы 2.1—2.3 и замечание 1 к теореме 3.1) следует непрерывность функций $\psi(t)$ и $H(t) = \sup_{u \in E^n} H(x(t), u, t, \psi(t))$

на отрезке $[t_0, T]$. Поэтому с учетом соотношений (5), (7), (8) имеем

$$\begin{aligned} [f_u^0(x(t), u(t), t)]_t &= 0, \\ [\langle u(t), f_u^0(x(t), u(t), t) \rangle - f^0(x(t), u(t), t)]_t &= 0, \quad t_0 \leq t \leq T; \end{aligned} \quad (13)$$

здесь принято обозначение: $[z(t)]_t = z(t+0) - z(t-0)$. Поскольку равенства (13) выполнены при всех t , $t_0 \leq t \leq T$, 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

ления условия Эрдмана — Вейерштрасса в точках излома кривой $x(t)$, $t_0 \leq t \leq T$.

Перейдем к рассмотрению условий на правом конце оптимальной кривой $x(t)$, $t_0 \leq t \leq T$. Если конец $x(T)$ свободен, то в силу условия (2.24) $\psi(T) = 0$. Отсюда с учетом выражения (8) имеем

$$f_u^0(x(T), u(T), T) = 0. \quad (14)$$

Если правый конец $x(T)$ подвижен, точнее, $x(T) \in S_l(T) = \{x \in E^n: g_j(x, T) \equiv x^j - \varphi_j(T) \equiv 0, j = \overline{1, n}\}$, то согласно условиям (2.27), (2.43) существуют постоянные $a_1, a_2, \dots, a_n$ такие, что

$$\psi_i(T) = \sum_{j=1}^n a_j g_{jx_i}(x(T), T) = a_i,$$

$$\begin{aligned} H(x(T), u(T), T, \psi(T)) &= - \sum_{j=1}^n a_j g_{jt}(x(T), T) = \\ &= \sum_{j=1}^n a_j \dot{\varphi}_j(T) = \sum_{j=1}^n \psi_j(T) \dot{\varphi}_j(T) = \langle \psi(T), \dot{\varphi}(T) \rangle. \end{aligned}$$

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 [80, 136]. 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 ([14], стр. 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 $V \subseteq E^n$ 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 .

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 [11—13, 19, 36, 49, 56, 60, 71, 91, 152, 153, 162, 171, 205, 206, 208, 226, 231, 239, 241, 244—246].

§ 1. Схема Беллмана.

Проблема синтеза для дискретных систем

1. Рассмотрим следующую задачу оптимального управления:

$$J(t_0, x_0, u(\cdot)) = \int_{t_0}^T f^0(x(\tau), u(\tau), \tau) d\tau + \Phi(x(T)) \rightarrow \inf; \quad (1)$$

$$\dot{x}(\tau) = f(x(\tau), u(\tau), \tau), \quad t_0 \leq \tau \leq T, \quad x(t_0) = x_0, \quad (2)$$

$$x(\tau) \in G(\tau), \quad t_0 \leq \tau \leq T, \quad (3)$$

$$u = u(\tau) \in V(\tau), \quad t_0 \leq \tau \leq T; \quad u(\tau) - \text{кусочно непрерывна}; \quad (4)$$

моменты времени t_0, T будем считать заданными (описание обозначений см. в § 6.1).

Для приближенного решения этой задачи разобьем отрезок $t_0 \leq t \leq T$ на N частей точками $t_0 < t_1 < \dots < t_{N-1} < t_N = T$, и, приняв эти точки в качестве узловых, интеграл в (1) заменим квадратурной формулой прямоугольников, уравнения (2) — разностными уравнениями с помощью явной схемы Эйлера [2, 40, 129, 211]. В результате придем к следующей дискретной задаче

оптимального управления:

$$I_0(x; [u_i]_0) = \sum_{i=0}^{N-1} F_i^0(x_i, u_i) + \Phi(x_N) \rightarrow \inf, \quad (5)$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = \overline{0, N-1}; \quad x_0 = x \in G_0; \quad (6)$$

$$x_i \in G_i, \quad i = \overline{0, N}; \quad (7)$$

$$[u_i]_0 = (u_0, u_1, \dots, u_{N-1}): u_i \in V_i, \quad i = \overline{0, N-1}; \quad (8)$$

где $F_i^0(x, u) = f^0(x, u, t_i)(t_{i+1} - t_i)$, $F_i(x, u) = x + f(x, u, t_i)(t_{i+1} - t_i)$, $G_i = G(t_i)$, $V_i = V(t_i)$. 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 [47, 71, 196, 243].

Если задать какое-либо дискретное управление $[u_i]_0 = (u_0, u_1, \dots, u_{N-1})$ и начальное условие $x_0 = x \in G_0$, то система (6) однозначно определяет соответствующую дискретную траекторию $[x_i]_0 = [x_i(x; [u_i]_0)]_0 = (x_0, x_1, \dots, x_N)$. 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 $x_0 = x$, удовлетворяет 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 $\Delta_0(x) = \emptyset$ 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 $\Delta_0(x) \neq \emptyset$. Обозначим $X_0 = \{x: x \in G_0, \Delta_0(x) \neq \emptyset\}$. 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 $x, u_0, u_1, \dots, u_{N-1}$, и для ее решения в принципе могут быть использованы

методы, описанные в главах 1—3, 5. Однако в практических задачах 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:

$$I_k(x, [u_i]_k) = \sum_{i=k}^{N-1} F_i^0(x_i, u_i) + \Phi(x_N) \rightarrow \inf; \quad (9)$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = \overline{k, N-1}; \quad x_k = x; \quad (10)$$

$$x_i \in G_i, \quad i = \overline{k, N}; \quad (11)$$

$$[u_i]_k = (u_k, u_{k+1}, \dots, u_{N-1}); \quad u_i \in V_i, \quad i = \overline{k, N-1}, \quad (12)$$

где точка x и целое число k фиксированы, $x \in G_k$, $0 \leq k \leq N-1$. 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 $[x_i]_k = (x_k = x, x_{k+1}, \dots, x_N)$ 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 $I_k(x, [u_i^*]_k) = I_k^*(x) = \inf_{\Delta_k(x)} I_k(x, [u_i]_k)$; $[u_i^*]_k$ назовем оптимальным управлением, $[x_i^*]_k$ — оптимальной траекторией задачи (9)—(12).

Нетрудно видеть, что если $X_0 \neq \emptyset$, то $\Delta_k(x) \neq \emptyset$ хотя бы для одного $x \in G_k$. Введем функцию

$$B_k(x) = \inf_{\Delta_k(x)} I_k(x, [u_i]_k), \quad k = \overline{0, N-1},$$

называемую функцией Беллмана задачи (5)—(8).

Покажем, что функция Беллмана удовлетворяет следующим рекуррентным соотношениям, называемым уравнением Беллмана:

$$B_k(x) = \inf_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1}(F_k(x, u))], \quad k = \overline{0, N-1}, \quad (13)$$

$$B_N(x) = \Phi(x), \quad x \in G_N, \quad (14)$$

где $D_k(x)$ — множество всех тех $u \in V_k$, для которых существует хотя бы одно управление $[u_i]_k \in \Delta_k(x)$ с компонентой $u_k = u$. Отсюда 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 $\Delta_{k+1}(F_k(x, u)) \neq \emptyset$. Справедливость соотношения (13) при $k = N-1$ очевидным образом вытекает из условия $B_N(x) \equiv \Phi(x)$ и представления $I_{N-1}(x, [u_i]_{N-1}) \equiv F_{N-1}^0(x, u) + \Phi(F_{N-1}(x, u))$, верного для любого $u \in D_{N-1}(x) \equiv \Delta_{N-1}(x) \equiv \{u: u \in V_{N-1}, x_N = F_{N-1}(x, u) \in G_N, x \in G_{N-1}\}$. Докажем (13) при $k, 0 \leq k < N-1$. Для этого сначала убедимся в том, что

$$B_k(x) \leq \inf_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1}(F_k(x, u))]. \quad (15)$$

Возьмем произвольное $u \in D_k(x)$ (разумеется, предполагаем, что $D_k(x) \neq \emptyset$) 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 $\Delta_{k+1}(x_{k+1}) \neq \emptyset$. По определению $B_{k+1}(x_{k+1}) = \inf_{[u_i]_{k+1} \in \Delta_{k+1}(x_{k+1})} I_{k+1}(x_{k+1}, [u_i]_{k+1})$ для любого $\varepsilon > 0$ найдется управление $[u_i^e]_{k+1} \in \Delta_{k+1}(x_{k+1})$ такое, что $B_{k+1}(x_{k+1}) \leq I_{k+1}(x_{k+1}, [u_i^e]_{k+1}) \leq B_{k+1}(x_{k+1}) + \varepsilon$. Поскольку $[\bar{u}_i]_k = (u, u_{k+1}^e, \dots, u_{N-1}^e) \in \Delta_k(x)$, то $B_k(x) \leq I_k(x, [\bar{u}_i]_k) = F_k^0(x, u) + I_{k+1}(x_{k+1}, [u_i^e]_{k+1}) \leq F_k^0(x, u) + B_{k+1}(x_{k+1}) + \varepsilon \equiv F_k^0(x, u) + B_{k+1}(F_k(x, u)) + \varepsilon$. 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 $\inf_{[u_i]_k \in \Delta_k(x)} I_k(x, [u_i]_k) = B_k(x)$ для каждого $\varepsilon > 0$ найдется такое управление $[v_i^e]_k \in \Delta_k(x)$, что $B_k(x) \leq I_k(x, [v_i^e]_k) \leq B_k(x) + \varepsilon$. Но $[\bar{v}_i]_{k+1} = (v_{k+1}^e, \dots, v_{N-1}^e) \in \Delta_{k+1}(F_k(x, v_k^e))$, поэтому

$$\begin{aligned} F_k^0(x, v_k^e) + B_{k+1}(F_k(x, v_k^e)) &\leq F_k^0(x, v_k^e) + I_{k+1}(F_k(x, v_k^e), [\bar{v}_i]_{k+1}) = \\ &= I_k(x, [v_i^e]_k) \leq B_k(x) + \varepsilon. \end{aligned}$$

Так как $v_k^e \in D_k(x)$, то отсюда имеем: $\inf_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1}(F_k(x, u))] \leq B_k(x) + \varepsilon$, или в силу произвольности $\varepsilon > 0$:

$$\inf_{u \in D_k(x)} [F_k^0(x, u) + B_{k+1}(F_k(x, u))] \leq B_k(x).$$

Отсюда и из (15) немедленно следует равенство (13).

Пользуясь условиями (13), (14), можно последовательно определить функции $B_k(x)$ и их области определения X_k , $k = N, N-1, \dots, 1; 0$. 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} , $k \leq N-1$, то для определения $B_k(x)$ нужно решить задачу минимизации функции $\Phi(x, u) \equiv F_k^0(x, u) + B_{k+1}(F_k(x, u))$ переменных $u = (u^1, \dots, u^r)$ на известном множестве $D_k(x) = \{u: u \in V_k, F_k(x, u) \in X_{k+1}\}$. Здесь могут быть использованы методы глав 1—3, 5. Очевидно, функция $B_k(x)$ определена в точке x тогда и только тогда, когда $D_k(x) \neq \emptyset$. Таким образом, при определении значений функции $B_k(x)$ одновременно находится и область ее определения $X_k = \{x: x \in G_k, D_k(x) \neq \emptyset\} \equiv \{x: x \in G_k, \Delta_k(x) \neq \emptyset\}$. Так как $\Delta_k(x) \neq \emptyset$ хотя бы при одном $x \in G_k$, то $X_k \neq \emptyset$, $k = N, N-1, \dots, 1, 0$.

3. Предположим, что нам удалось найти функции $B_k(x)$ из условий (13), (14) и, кроме того, пусть также известны функции $u_k(x) \in D_k(x)$, $x \in X_k$, $k = 0, N-1$, 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 соответствующая траектория $[x_i^*]_0$ для задачи (5)—(8) определяются следующим образом: сначала из условия

$$\inf_{x \in X_0} B_0(x) = B_0(x_0^*) \quad (16)$$

находят $x_0^* \in X_0$, затем последовательно полагают

$$\begin{aligned} u_0^* &= u_0(x_0^*), & x_1^* &= F_0(x_0^*, u_0^*), & u_1^* &= u_1(x_1^*), \\ x_2^* &= F_1(x_1^*, u_1^*), & \dots, & & x_N^* &= F_{N-1}(x_{N-1}^*, u_{N-1}^*). \end{aligned} \quad (17)$$

Оптимальное управление $[u_i^*]_k$ и траектория $[x_i^*]_k$ для задачи (9)—(12) определяются аналогично:

$$\begin{aligned} x_k^* &= x, & u_k^* &= u_k(x_k^*), \\ x_{k+1}^* &= F_k(x_k^*, u_k^*), & \dots, & & x_N^* &= F_{N-1}(x_{N-1}^*, u_{N-1}^*). \end{aligned} \quad (18)$$

Для доказательства этих утверждений введем вспомогательные функции:

$$R_i(x, u) \equiv B_{i+1}(F_i(x, u)) - B_i(x) + F_i^0(x, u), \quad i = 0, N-1. \quad (19)$$

Очевидно, уравнения Беллмана (13) тогда можно переписать в эквивалентном виде:

$$\inf_{u \in D_k(x)} R_k(x, u) \equiv R_k(x, u_k(x)) = 0, \quad k = 0, N-1. \quad (20)$$

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

$$I_k(x, [u_i]_k) = \sum_{i=k}^{N-1} R_i(x_i, u_i) + B_k(x) \quad (21)$$

при всех $k=0, 1, \dots, N-1$. В самом деле, учитывая равенство $B_N(x) \equiv \Phi(x)$, из (10), (19) имеем $\sum_{i=k}^{N-1} R_i(x_i, u_i) =$

$$= \sum_{i=k}^{N-1} [B_{i+1}(x_{i+1}) - B_i(x_i) + F_i^0(x_i, u_i)] = B_N(x_N) - B_k(x) + \sum_{i=k}^{N-1} F_i^0(x_i, u_i) = I_k(x, [u_i]_k) - B_k(x), \text{ что равносильно (21).}$$

Теорема 1. Пусть найдены функции $B_k(x)$ из (13), (14) и их области определения X_k , а также функции $u = u_k(x)$, $x \in X_k$, $k = \overline{0, N-1}$, 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

$$R_i(x_i^*, u_i^*) \equiv R_i(x_i^*, u_i(x_i^*)) = \inf_{u \in D_i(x_i^*)} R_i(x_i^*, u) = 0, i = \overline{0, N-1}. \quad (22)$$

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

$$R_i(x_i, u_i) \geq \inf_{u \in D_i(x_i)} R_i(x_i, u) = R_i(x_i, u_i(x_i)) = 0, i = \overline{0, N-1}. \quad (23)$$

С помощью формулы (21) при $k=0$ с учетом соотношений (16), (22), (23) получаем $I_0(x, [u_i]_0) - I_0(x_0^*, [u_i^*]_0) = \sum_{i=0}^{N-1} [R_i(x_i, u_i) - R_i(x_i^*, u_i^*)] + B_0(x) - B_0(x_0^*) \geq 0$ для любых $x \in X_0$ и $[u_i]_0 \in \Delta_0(x)$, что и требовалось.

Теорема 2. 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 $i = \bar{k}, N-1$. Отсюда с помощью (21) получим

$$I_k(x, [u_i]_k) - I_k(x, [u_i^*]_k) = \sum_{i=k}^{N-1} [R_i(x_i, u_i) - R_i(x_i^*, u_i^*)] \geq 0,$$

что и требовалось.

Заметим, что $\inf_{[u_i]_k \in \Delta_k(x)} I_k(x, [u_i]_k) = I_k(x, [u_i^*]_k) = B_k(x)$, — это следует из формулы (21) с учетом равенств (22). Тем самым показано, что функции $B_k(x)$, определяемые из (13), (14), в самом деле являются функциями Беллмана для задачи (5)—(8).

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 2 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: $x_{i+1} = F_i(x_i, u_i(x_i))$, $i = \bar{k}, N-1$, $x_k = x$ (если $x \notin X_k$, то $\Delta_k(x) = \emptyset$, — движение с соблюдением условий (10)—(12) невозможно). Достаточные условия существования функции Беллмана и синтезирующей функции для задачи (5)—(8) даются в следующей теореме.

Теорема 3. Пусть множества G_k , $k = \bar{0}, N$, замкнуты, множества V_k , $k = \bar{0}, N-1$, замкнуты и ограничены, функция $F_k^0(x, u)$ полунепрерывна снизу, а функция $F_k(x, u)$ непрерывна по совокупности аргументов (x, u) при $x \in G_k$, $u \in V_k$, $k = \bar{0}, N-1$, $\Phi(x)$ полунепрерывна снизу на множестве G_N . Тогда: 1) множества X_k , $k = \bar{0}, N$, замкнуты, множества $D_k(x)$, $k = \bar{0}, N-1$, 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 , $k = \bar{0}, N$.

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 , $0 \leq k \leq N-1$. Докажем, что тогда X_k замкнуто и на X_k справедливы все утверждения

теоремы. Так как $D_k(x) = \{u: u \in V_k, F_k(x, u) \in X_{k+1}\} \subseteq V_k$ 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 $v_m \in D_k(x)$, $m = 1, 2, \dots$, $v_m \rightarrow v$ при $m \rightarrow \infty$. Это значит, что $v_m \in V_k$, $F_k(x, v_m) \in X_{k+1}$, $m = 1, 2, \dots$. Из замкнутости V_k , X_{k+1} и непрерывности $F_k(x, v)$ сразу имеем: $v \in V_k$, $\lim_{m \rightarrow \infty} F_k(x, v_m) = F_k(x, v) \in X_{k+1}$, т. е. $v \in D_k(x)$. Замкнутость $D_k(x)$ доказана.

Покажем замкнутость $X_k = \{x: x \in G_k, D_k(x) \neq \emptyset\}$. Пусть $y_m \in X_k$, $m = 1, 2, \dots$, $y_m \rightarrow y$ при $m \rightarrow \infty$. Из замкнутости G_k следует $y \in G_k$. Если мы еще покажем, что $D_k(y) \neq \emptyset$, то это будет означать, что $y \in X_k$, и замкнутость X_k будет доказана. Так как $D_k(y_m) \neq \emptyset$, то существует такое $v_m \in V_k$, что $F_k(y_m, v_m) \in X_{k+1}$, $m = 1, 2, \dots$. В силу компактности V_k из последовательности $\{v_m\}$ можно выбрать подпоследовательность $\{v_{m_p}\} \rightarrow v \in V_k$ при $p \rightarrow \infty$. Поскольку X_{k+1} замкнуто, $F_k(x, u)$ непрерывна, то $\lim_{p \rightarrow \infty} F_k(y_{m_p}, v_{m_p}) = F_k(y, v) \in X_{k+1}$, т. е. $v \in D_k(y)$. Таким образом, $D_k(y) \neq \emptyset$.

Далее, функция $\varphi(x, u) = F_k^0(x, u) + B_{k+1}(F_k(x, u))$ 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 $F_k^\lambda(x, u)$, $B_{k+1}(x)$. 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, $B_k(x) = \inf_{u \in D_k(x)} \varphi(x, u) = \varphi(x, u_k(x))$ в силу уравнения Беллмана (13).

Остается еще доказать полунепрерывность снизу $B_k(x)$ на X_k . Пусть $x, y_m \in X_k$, $y_m \rightarrow x$ при $m \rightarrow \infty$, $B_k(y_m) = \varphi(y_m, u_k(y_m))$. Так как $u_k(y_m) \in D_k(y_m) \subseteq V_k$, то в силу компактности V_k последовательность $\{u_k(y_m), m = 1, 2, \dots\}$ имеет хотя бы одну предельную точку $v \in V_k$. Можем считать, что сама последовательность $\{u_k(y_m)\} \rightarrow v$ при $m \rightarrow \infty$. 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 $\lim_{m \rightarrow \infty} F_k(y_m, v_k(y_m)) = F_k(y, v) \in X_{k+1}$. Это значит, что $v \in D_k(x)$. Тогда

$$\lim_{m \rightarrow \infty} B_k(y_m) = \lim_{m \rightarrow \infty} \varphi(y_m, u_k(y_m)) \geq \varphi(x, v) \geq \inf_{u \in D_k(x)} \varphi(x, u) = B_k(x).$$

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 управления $[u_i^*]_0$ 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 $B_i(x)$, $i = \overline{0, N}$. По аналогии с (19) введем функцию

$$S_i(x, u) \equiv K_{i+1}(F_i(x, u)) - K_i(x) + F_i^0(x, u), \quad i = \overline{0, N-1}, \quad (24)$$

и, кроме того, положим

$$s_N(x) = \Phi(x) - K_N(x), \quad x \in G_N. \quad (25)$$

Возьмем произвольную допустимую пару $[u_i]_0 = (u_0, u_1, \dots, u_{N-1})$, $[x_i]_0 = (x_0, x_1, \dots, x_{N-1})$, задачи (5)–(8). Тогда $x_0 \in X_0$, $[u_i]_0 \in \Delta_0(x_0)$, $u_i \in D_i(x_i)$, $i = \overline{0, N-1}$. Учитывая условие (6), из (24) имеем

$$S_i(x_i, u_i) = K_{i+1}(x_{i+1}) - K_i(x_i) + F_i^0(x_i, u_i), \quad i = \overline{0, N-1}.$$

Суммируя эти равенства по i от 0 до $N-1$, с помощью (25) получим формулу

$$I_0(x_0, [u_i]_0) = \sum_{i=0}^{N-1} S_i(x_i, u_i) + s_N(x_N) + K_0(x_0). \quad (26)$$

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. $\bar{x}_0 \in X_0$, $[\bar{u}_i]_0 \in \Delta_0(\bar{x}_0)$, $\bar{u}_i \in D_i(\bar{x}_i)$, $i = \overline{0, N-1}$. 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

$$\begin{aligned} I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0(x_0, [u_i]_0) &= \\ &= \sum_{i=0}^{N-1} [S_i(\bar{x}_i, \bar{u}_i) - S_i(x_i, u_i)] + s_N(\bar{x}_N) - s_N(x_N) + K_0(\bar{x}_0) - K_0(x_0) \end{aligned} \quad (27)$$

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

$$0 \leq I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^* \leq \sum_{i=0}^{N-1} [S_i(\bar{x}_i, \bar{u}_i) - \inf_{x \in X_i} \inf_{u \in D_i(x)} S_i(x, u)] + \\ + s_N(\bar{x}_N) - \inf_{x_N} s_N(x) + K_0(\bar{x}_0) - \inf_{x_0} K_0(x), \quad (28)$$

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 $K_i(x) = B_i(x)$, $i = \overline{0, N}$, то $S_i(x, u) = R_i(x, u)$, $s_N(x) = 0$. Кроме того, из (20) следует, что $\inf_{u \in D_i(x)} R_i(x, u) = 0$ для всех $x \in X_i$, так что $\inf_{x \in X_i} \inf_{u \in D_i(x)} R_i(x, u) = 0$. Поэтому при $K_i(x) = B_i(x)$ из (28) получим

$$0 \leq I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^* \leq \sum_{i=0}^{N-1} R_i(\bar{x}_i, \bar{u}_i) + B_0(\bar{x}_0) - \inf_{x_0} B_0(x). \quad (29)$$

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 $\inf_{u \in D_i(x)} R_i(x, u) = 0$, $x \in X_i$, а $B_0(\bar{x}_0)$ — поближе к $\inf_{x_0} B_0(x)$, и затем строить управление $[\bar{u}_i]$ и траекторию $[\bar{x}_i]$ следующим образом:

$\bar{u}_0 = \bar{u}_0(\bar{x}_0)$, $\bar{x}_1 = F_0(\bar{x}_0, \bar{u}_0)$, $\bar{u}_1 = \bar{u}_1(x_1)$, ..., $x_N = F_{N-1}(\bar{x}_{N-1}, \bar{u}_{N-1})$, то и величина $I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^*$ будет небольшой.

Замѣтим, 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

$$0 \leq I_0(\bar{x}_0, [\bar{u}_i]_0) - I_0^* \leq \sum_{i=0}^{N-1} [S_i(\bar{x}_i, \bar{u}_i) - \inf_{x \in G_i} \inf_{u \in V_i} S(x, u)] + \\ + s_N(\bar{x}_N) - \inf_{G_N} s_N(x) + K_0(\bar{x}_0) - \inf_{G_0} K_0(x). \quad (30)$$

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 4. 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 $K_i(x)$, $i = 0, N$, такая, что

$$S_i(\bar{x}_i, \bar{u}_i) = \inf_{x \in X_i} \inf_{u \in D_i(x)} S_i(x, u) = S_{i \min}, \quad i = \overline{0, N-1}; \quad (31)$$

$$s_N(\bar{x}_N) = \inf_{x_N} s_N(x) = s_{N \min}; \quad K_0(\bar{x}_0) = \inf_{x_0} K_0(x) = K_{0 \min}, \quad (32)$$

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 5. Пусть последовательность управлений $[u_{im}]_0$ и траекторий $[x_{im}]_0$, $m = 1, 2, \dots$, удовлетворяет условиям (6) — (8). Для того чтобы

$$\lim_{m \rightarrow \infty} I_0(x_{0m}, [u_{im}]_0) = I_0^*, \quad (33)$$

достаточно существования функции $K_i(x)$, $i = \overline{0, N}$, такой, что

$$\lim_{m \rightarrow \infty} S_i(\bar{x}_{im}, \bar{u}_{im}) = S_{i \min}, \quad i = \overline{0, N-1}, \quad (34)$$

$$\lim_{m \rightarrow \infty} s_N(x_{Nm}) = s_{N \min}, \quad \lim_{m \rightarrow \infty} K_0(x_{0m}) = K_{0 \min}, \quad (35)$$

где $S_{i \min}$, $s_{N \min}$, $K_{0 \min}$ определены в (31), (32).

Доказательство. В оценку (28) вместо $[\bar{u}_i]_0$, $[\bar{x}_i]_0$ подставим соответственно $[\bar{u}_{im}]_0$, $[\bar{x}_{im}]_0$ и перейдем к пределу при $m \rightarrow \infty$. С учетом условий (34), (35) получим равенство (33).

Утверждения, аналогичные теоремам 4, 5, доказаны в [91, 152, 153] для более общих задач, чем задача (5) — (8).

Всякую функцию $K_i(x)$, $i = \overline{0, N}$, удовлетворяющую условиям теоремы 4 [теоремы 5], назовем функцией Кротова задачи (5) — (8), соответствующей допустимой паре $[u_i]_0$, $[x_i]_0$ [или последовательности допустимых пар $[u_{im}]_0$, $[x_{im}]_0$, $m = 1, 2, \dots$].

Заметим, что если существует хотя бы одна функция Кротова $K_i(x)$ $i = \overline{0, N}$, то функция $K_i(x) + \alpha_i$, $i = \overline{0, N}$, при любых α_i также является функцией Кротова. Поэтому без ограничения

общности в теоремах 4, 5 можно принять $S_{i \min} = 0$, $i = \overline{0, N-1}$, $s_{N \min} = 0$, ибо в противном случае функцию $K_i(x)$ заменим новой функцией $K_i(x) + \alpha_i$, где

$$\alpha_i = S_{i \min} + S_{i+1, \min} + \dots + S_{N-1, \min}, \quad i = \overline{1, N-1}; \quad \alpha_N = s_{N \min}.$$

Таким образом, функция Кротова для допустимой пары $([\bar{u}_i]_0, [\bar{x}_i]_0)$ или последовательности $([u_{im}]_0, [x_{im}]_0)$, $m = 1, 2, \dots$, согласно теоремам 4, 5 удовлетворяет условиям

$$S_i(x, u) = K_{i+1}(F_i(x, u)) - K_i(x) + F_i^0(x, u) \geq 0, \quad u \in D_i(x), \quad x \in X_i, \quad i = \overline{0, N-1}, \quad (36)$$

$$s_N(x) = K_N(x) + \Phi(x) \geq 0, \quad x \in X_N = G_N;$$

$$K_0(x) \geq K_{0 \min}, \quad x \in X_0, \quad (37)$$

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 $m \rightarrow \infty$, $i = \overline{0, N}$.

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 удастся 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

$$I_0([u_i]_0) = I_0(u_0, u_1, \dots, u_{N-1}) = \sum_{i=0}^{N-1} f_i^0(u_i) \quad (38)$$

при условиях

$$u_i \in V_i, \quad i = \overline{0, N-1}; \quad (39)$$

$$\sum_{i=0}^{N-1} g_i^l(u_i) \leq b^l, \quad j = \overline{1, p}; \quad \sum_{i=0}^{N-1} g_i^l(u_i) = b^l, \quad j = \overline{p+1, n}, \quad (40)$$

где V_i — заданные множества из E^r ; $f_i^0(u)$, $g_i^l(u)$, $u \in V_i$ — заданные функции, b^l — заданные числа.

Задачу (38) — (40), оказывается, нетрудно записать в виде задачи (5) — (8). В самом деле, введем переменные x_i , $i = \overline{0, N}$,

как решение системы

$$x_{k+1} = x_k + g_k(u_k), \quad k = \overline{0, N-1}; \quad x_0 = 0, \quad (41)$$

где $g_k(u) = (g_k^1(u), \dots, g_k^n(u))$, $x_k = (x_k^1, \dots, x_k^n)$, $k = \overline{0, N}$. Так как из (41) следует, что $x_N = \sum_{k=0}^{N-1} g_k(u_k)$, то ясно, что ограничения (40) равносильны условию

$$x_N \in G_N = \{x: x \in E^n, x^j \leq b^j, j = \overline{1, p}; x^j = b^j, j = \overline{p+1, n}\}. \quad (42)$$

Таким образом, задача (38) — (40) эквивалентна задаче минимизации функции (38) при условиях (39), (41), (42) и является частным случаем задачи (5) — (8) при $F_i^0(x, u) = f_i^0(u)$, $F_i(x, u) = x + g_i(u)$, $\Phi(x) \equiv 0$, $G_i = E^n$, $i = \overline{1, N-1}$; $G_0 = \{0\}$, 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):

$$B_k(x) = \inf_{u \in D_k(x)} [f_k^0(u) + B_{k+1}(x + g_k(u))], \quad (43)$$

$$x \in X_k, \quad k = \overline{0, N-1}; \quad B_N(x) = 0.$$

В том случае, когда ограничения (40) и, следовательно, (42) отсутствуют, то $G_N = E^n$ и в (43) можно положить $D_k(x) = V_k$, $k = \overline{0, N-1}$. Уравнениями (43) можно пользоваться для решения задачи (38) — (40), как это показано выше в п. 3. Предлагаем читателю в качестве упражнения переформулировать теоремы 1—5 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 [71].

8. В заключение заметим, что 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 $\lim_{N \rightarrow \infty} \max_{0 \leq i \leq N-1} (t_{i+1} - t_i) = 0$ 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 удастся, 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).

Упражнения. 1. Найти функцию Беллмана для задачи

$$I_0([u]_0) = \sum_{i=0}^{N-1} [\langle a_i, x_i \rangle + b_i(u_i)] + \langle c, x_N \rangle \rightarrow \inf,$$

$$x_{i+1} = A_i x_i + B_i(u_i), \quad u_i \in V_i, \quad i = \overline{0, N-1}; \quad x_0 = a,$$

где A_i — матрица порядка $n \times n$; $B_i(u) = (B_i^1(u), \dots, B_i^n(u))$, $B_i^j(u)$, $b_i(u)$ — функции переменной $u \in V_i \subseteq E^r$, a_i , c , a — n -мерные векторы, $i = \overline{0, N-1}$.
Указание: функцию Беллмана искать в виде $B_k(x) = \langle \psi_k, x \rangle$, $k = \overline{0, N}$.

2. Найти функцию Беллмана для задачи: $I_0(u) = \Phi(x_1) \rightarrow \inf$; $x_1 = x_0 + u$, $x_i \in G_i$, $i = \overline{0, 1}$; $u \in V_0$, где $\Phi(x) = (1 + e^{-1/x})^{-1}$ при $x \neq 0$, $\Phi(0) = 1/2$; $G_0 = \{x \in E^1: |x| \leq 1/2\}$, $G_1 = E^1$; $V_0 = \{u \in E^1: |u| \leq 1\}$. 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 $B_k(x)$, $x \in X_k$, $k = \overline{0, N}$, удовлетворяет условиям (13), (14), а функции $u_{km}(x) \in D_k(x)$ и точки $x_{0m} \in X_0$, $m = 1, 2, \dots$, таковы, что $\lim_{m \rightarrow \infty} R_i(x, u_{im}(x)) = 0$, $\lim_{m \rightarrow \infty} B_0(x_{0m}) = \inf_{x_0} B_0(x)$. Пусть управление $[u_{im}]_0$ и траектория $[x_{im}]_0$ построены по правилу: $u_{0m} = u_{0m}(x_{0m})$, $x_{1m} = F_0(x_{0m}, u_{0m})$, $u_{1m} = u_{1m}(x_{1m})$, $\dots$, $x_{Nm} = F_{N-1}(x_{N-1,m}, u_{N-1,m})$. Тогда последовательность пар $(x_{0m}, [u_{im}]_0)$ — минимизирующая для задачи (5)–(8), т. е. $\lim_{m \rightarrow \infty} I_0(x_{0m}, [u_{im}]_0) = I_0^*$. Доказать.

4. Доказать, что последовательность функций $u_{im}(x)$, $i = \overline{0, N-1}$, $m = 1, 2, \dots$, 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 $x_{i+1,m} = F_i(x_{im}, u_{im}(x_{im}))$, $i = \overline{k, N-1}$; $x_{km} = x$, $m = 1, 2, \dots$, 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 1–3 для задачи:

$$I_0(x, [u]_0) = \sum_{i=0}^N F_i^0(x_i, u_i) + \Phi_0(x_0) + \Phi_1(x_N) \rightarrow \inf;$$

$$x_{i+1} = F_i(x_i, u_i), \quad i = \overline{0, N-1}; \quad x_i \in G_i, \quad u_i \in V_i, \quad i = \overline{0, N},$$

где $x_i = (x_i^1, \dots, x_i^n)$, $F_i = (F_i^1, \dots, F_i^n)$, $u_i = (u_i^1, \dots, u_i^r)$; множества $G_i \subseteq E^n$, $V_i \subseteq E^r$, $i = \overline{0, N}$, заданы; функции $F_i^j(x, u)$ определены при $x \in G_i$, $u \in V_i$, $i = \overline{0, N}$, $j = \overline{0, n}$; 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 4, 5 для задачи из упражнения 6.

8. Пусть в задаче (5)–(8) $F_i^0(x, u) \equiv 0$, $i = \overline{0, N-1}$. Показать, что тогда $B_k(x) = \Phi(x)$ для всех $k = \overline{0, N}$, 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

$$I_0(u_0, u_1, \dots, u_N) = \sum_{i=0}^{N-1} f_i(u_i, u_{i+1}) + f_N(u_N)$$

при условиях $u_i \in V_i$, $i = \overline{0, N}$. У к а з а н и е: 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 $(u_k = u, u_{k+1}, \dots, u_N)$, $u_i \in V_i$, $i = \overline{k, N}$ и показать, что $B_k(u) = \inf_{v \in V_{k+1}} [f_k(u, v) + B_{k+1}(v)]$, $k = \overline{0, N-1}$;

$$B_N(u) = f_N(u).$$

§ 2. Схема Моисеева

1. По-прежнему будем рассматривать задачу (1.1) — (1.4). Для приближенного решения этой задачи, как и раньше, разобьем отрезок $t_0 \leq t \leq T$ на N частей точками $t_0 < t_1 < \dots < t_{N-1} < t_N = T$. На множестве $G_i \equiv G(t_i)$ возьмем некоторую дискретную сетку точек $x_{ij} \in G_i$; следуя [12, 171] множество всех точек выбранной сетки будем называть шкалой состояний и обозначать через H_i , $i = \overline{0, 1, \dots, N}$. 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:

$$J_i(x, y, u(\cdot)) = \int_{t_i}^{t_{i+1}} f^0(x(t), u(t), t) dt \rightarrow \inf; \quad (1)$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_i \leq t \leq t_{i+1}; \quad x(t_i) = x, \quad x(t_{i+1}) = y, \quad (2)$$

$$x(t) \in G(t), \quad t_i \leq t \leq t_{i+1}, \quad (3)$$

$$u = u(\cdot) \text{ кусочно непрерывна и } u(t) \in V(t) \text{ при } t_i \leq t \leq t_{i+1}. \quad (4)$$

Следуя [12, 171], 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 $M_i(x, y) = \inf_{u \in \Delta_i(x, y)} J_i(x, y, u)$; если $\Delta_i(x, y) = \emptyset$, то $M_i(x, y) = +\infty$, по определению.

Пусть все точки всех соседних шкал попарно соединены элементарными операциями. Если $\inf_{\Delta_i(x, y)} J_i(x, y, u)$ достигается на некотором управлении $u_i(\cdot) \in \Delta_i(x, y)$ и соответствующей

траектории $x_i(\cdot)$, то величина

$$\sum_{i=0}^{N-1} M_i(x_{i|I_i}, x_{i+1, I_{i+1}}) + \Phi(x_{N|N}) \quad (5)$$

выражает собой значение исходной функции (1.1) на управляющей $u = u(t) = u_i(t)$, $t_i \leq t \leq t_{i+1}$, $i = \overline{0, N-1}$, и соответствующей траектории $x(t, u) = x_i(t)$, $t_i \leq t \leq t_{i+1}$, $i = \overline{0, N-1}$; $x(t_0) = x_{0|0}$, 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 $(x_{0|0}, x_{1|1}, \dots, x_{N|N})$, $x_{i|I_i} \in H_i$, $i = \overline{0, N}$.

Такая аппроксимация задачи (1.1)–(1.4) имеет смысл, конечно, и в том случае, когда $\inf_{\Delta_i(x, y)} J_i(x, y, u)$ 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 предоставляется 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 $\inf_{\Delta_i(x, y)} J_i(x, y, u)$, назовем дугой, соединяющей

точки x, y , а числа $M_i(x, y)$ — длиной этой дуги, $i = \overline{0, N-1}$. Дуги, последовательно соединяющие пары точек $x_{i|I_i}, x_{i+1, I_{i+1}}$ соседних шкал H_i, H_{i+1} , $i = \overline{0, N-1}$, назовем путем, соединяющим шкалы H_0 и H_N и проходящим через точки $x_{0|0}, x_{1|1}, \dots, x_{N|N}$, 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

$$C_k(x) = \inf \left\{ \sum_{i=k}^{N-1} M_i(x_{i|I_i}, x_{i+1, I_{i+1}}) + \Phi(x_{N|N}) \right\},$$

где нижняя грань берется по всем наборам точек $(x_{k|k} = x, x_{k+1, I_{k+1}}, \dots, x_{N|N})$, $x_{i|I_i} \in H_i$, $i = \overline{k, N}$. 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:

$$C_k(x) = \inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\}, \quad k = \overline{0, N-1}; \quad C_N(x) \equiv \Phi(x), \quad (6)$$

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 , $0 \leq k \leq N-1$. Для этого сначала убедимся в том, что

$$C_k(x) \leq \inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\}, \quad x \in H_k. \quad (7)$$

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 $C_k(x) \leq M_k(x, y) + C_{k+1}(y) + \varepsilon$. 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, $M_k(x, y_\varepsilon) + C_{k+1}(y_\varepsilon) \leq C_k(x) + \varepsilon$. Так как $y_\varepsilon \in H_{k+1}$, то отсюда имеем:

$$\inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\} \leq C_k(x) + \varepsilon, \quad \text{или в силу произвольности } \varepsilon: \inf_{y \in H_{k+1}} \{M_k(x, y) + C_{k+1}(y)\} \leq C_k(x).$$

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 , $i = \overline{0, N}$. 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 $\min_{y \in H_N} [M_{N-1}(x, y) + \Phi(y)] = C_{N-1}(x)$, 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 из 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 $\min_{y \in H_{k+1}} [M_k(x, y) + C_{k+1}(y)]$, 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 $k = N, N-1, \dots, 1, 0$. На вычисление всех $C_k(x)$ и $y = y_k(x)$, $x \in H_k$, $k = \overline{0, N-1}$, понадобится перебор $\sum_{i=0}^{N-1} p_i p_{i+1}$ величин. Наконец, находим $x_0^* \in H_0$ из условия $C_0(x_0^*) = \inf_{x \in H_0} C_0(x)$ — для этого нужно перебрать еще p_0 величин, и определяем последовательно точки $x_1^* = y_0(x_0^*)$, $x_2^* = y_1(x_1^*)$, $\dots$, $x_N^* = y_{N-1}(x_{N-1}^*)$. Путь, проходящий через найденные точки x_0^* , x_1^* , $\dots$, x_N^* , будет кратчайшим среди всех путей, соединяющих крайние шкалы H_0, H_N . В самом деле, по определению $y_k(x)$, $k = \overline{0, N-1}$, и $x_k^* \in H_k$, $k = \overline{0, N}$, имеем

$$C_k(x_k^*) = M_k(x_k^*, x_{k+1}^*) + C_{k+1}(x_{k+1}^*), \quad k = \overline{0, N-1}. \quad (8)$$

Возьмем произвольный путь, соединяющий шкалы H_0 и H_N и проходящий через точки $x_0, x_1, \dots, x_N$, $x_i \in H_i$, $i = \overline{0, N}$. Так как $x_{k+1} \in H_{k+1}$, то согласно (6) будем иметь: $C_k(x_k) \leq M_k(x_k, x_{k+1}) + C_{k+1}(x_{k+1})$. Отсюда и из (8) тогда следует: $M_k(x_k^*, x_{k+1}^*) + C_{k+1}(x_{k+1}^*) - C_k(x_k^*) = 0 \leq M_k(x_k, x_{k+1}) + C_{k+1}(x_{k+1}) - C_k(x_k)$, $k = \overline{0, N-1}$. Просуммируем это неравенство по k от нуля до $N-1$. Получим:

$$\begin{aligned} \sum_{k=0}^{N-1} M_k(x_k^*, x_{k+1}^*) + C_N(x_N^*) - C_0(x_0^*) &\leq \\ &\leq \sum_{k=0}^{N-1} M_k(x_k, x_{k+1}) + C_N(x_N) - C_0(x_0). \end{aligned}$$

Но $C_0(x_0^*) = \inf_{x \in H_0} C_0(x) \leq C_0(x_0)$, поэтому $\sum_{k=0}^{N-1} M_k(x_k^*, x_{k+1}^*) + C_N(x_N^*) \leq \sum_{k=0}^{N-1} M_k(x_k, x_{k+1}) + C_N(x_N)$ для любых путей, соединяющих шкалы H_0 и H_N . Таким образом, путь, проходящий через точки $x_0^*, x_1^*, \dots, x_N^*$, в самом деле кратчайший. Если $u_i^*(t)$ и $x_i^*(t)$, $t_i \leq t \leq t_{i+1}$, 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 ка-

честве приближенного решения исходной задачи (1.1) — (1.4) можно взять управление $u^*(t) = u_i^*(t)$ и траекторию $x^*(t) = x_i^*(t)$, $t_i \leq t \leq t_{i+1}$, $i = 0, N-1$.

Аналогично доказывается, что путь, проходящий через точки $x_k^* = x \in H_k$, $x_{k+1}^* = y_k(x_k^*) \in H_{k+1}$, ..., $x_N^* = y_{N-1}(x_{N-1}^*) \in H_N$, 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

$\sum_{i=0}^{N-1} p_i p_{i+1} + p_0$ величин, в то время как полный перебор всех путей, как нетрудно видеть, потребовал бы сравнения $p_0 p_1 \dots p_N$ 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 не кратчайшие 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 размерностях 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* [244]. 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 перестанет «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 $|\Delta x| = o(\Delta t)$ [171].

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* [12, 171, 244]. 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 $H_0, H_1, \dots, H_N$. Допустим, что шкалы $H_0, \dots, H_{i-1}$ уже просмотрены и получен путь $l_{1,i-1}$, соединяющий шкалу H_0 с 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_i и просмотр шкалы H_i на этом заканчиваем. Если же длины всех перебираемых путей оказались не меньше длины $l_{1,i-1}$, то полагаем $l_i = l_{1,i-1}$. После определения пути l_i переходим к просмотру следующей шкалы H_{i+1} и т. д. Такой перебор всех шкал $H_0, \dots, H_N$ 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 $H_0, \dots, H_N$ 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 более 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 удастся 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 удастся. 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 $ABEF$.

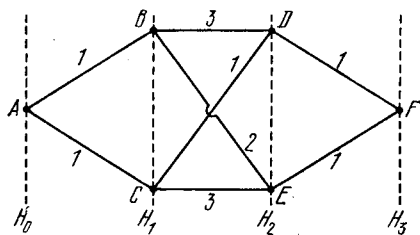


Рис. 7.1.

Различные модификации метода локальных вариаций, примеры задач, к которым применялся этот метод, см. в [12, 171, 244].

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 $t_i \leq t \leq t_{i+1}$ 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 $u(t) \equiv u(t, c_1, \dots, c_m)$, зависящих от параметров $c_1, c_2, \dots, c_m$ ($m \geq n$). Например, это могут быть алгебраические или тригонометрические многочлены с коэффициентами $c_1, \dots, c_m$ или кусочно постоянные функции со значениями $c_1, \dots, c_m$ и т. п. Значения этих параметров затем можно определить из следующей системы n уравнений с m неизвестными ($m \geq n$):

$$\int_{t_i}^{t_{i+1}} \dot{x}(t) dt = \int_{t_i}^{t_{i+1}} f(x(t), u(t, c_1, c_2, \dots, c_m)) dt = y - x, \quad (9)$$

$$u(t, c_1, \dots, c_m) \in V(t_i), \quad t_i \leq t \leq t_{i+1},$$

используя для этого различные методы [2, 39, 129]. Если $m > n$, то параметры $c_1, \dots, c_m$ 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:

$$\dot{x}(t) = f(x, u(t), t) \quad \text{или} \quad \dot{x}(t) = f\left(\frac{x+y}{2}, u(t), t\right)$$

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 [12, 171].

§ 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

$$J(t, x, u(\cdot)) = \int_t^T f^0(x(\tau), u(\tau), \tau) d\tau + \Phi(x(T)) \quad (1)$$

минимальное значение при условиях

$$\dot{x}(\tau) = f(x(\tau), u(\tau), \tau), \quad t \leq \tau \leq T; \quad x(t) = x; \quad (2)$$

$$x(\tau) \in G(\tau), \quad t \leq \tau \leq T, \quad (3)$$

$$u = u(\tau) \in V(\tau), \quad t \leq \tau \leq T; \quad u(\cdot) - \text{кусочно непрерывно}, \quad (4)$$

где x — произвольная точка множества $G(t)$, а t — произвольный фиксированный момент времени, $t_0 \leq t < T$. 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 $u(\tau)$, $t \leq \tau \leq T$, удовлетворяющих условиям (4) и $u(\tau) = u(\tau + 0) = \lim_{s \rightarrow \tau+0} u(s)$, $t \leq \tau < T$, и таких, что соответствующая траектория $x(\tau) = x(\tau, u)$, $t \leq \tau \leq T$, системы (2) определена на всем отрезке $[t, T]$ и удовлетворяет фазовому ограничению (3). Положим $X(t) = \{x: x \in G(t), \Delta(x, t) \neq \emptyset\}$, $t_0 \leq t < T$; $X(T) = G(T)$.

Пару $(u(\tau), x(\tau))$, $t \leq \tau \leq T$, 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 $(u(\tau), x(\tau))$, $t_0 \leq \tau \leq T$, 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 $(u_*(\tau), x_*(\tau))$, $t \leq \tau \leq T$, задачи (1)–(4) будем называть *решением* этой задачи, если $J(t, x, u_*(\cdot)) = \inf_{\Delta(x, t)} J(t, x, u(\cdot))$; 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) , $x \in G(t)$, $t_0 \leq t < T$, при которых эта задача имеет решение, 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) удастся 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):

$$\inf_{u \in D(x, t)} \{B_x(x, t), f(x, u, t)\} + B_t(x, t) + f^0(x, u, t) = 0, \quad (5)$$

$$x \in X(t), \quad t_0 \leq t < T,$$

$$B(x, T) = \Phi(x), \quad x \in X(T) = G(T), \quad (6)$$

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

$$B(x_0^*, t_0) = \inf_{x \in X(t_0)} B(x, t_0), \quad (14)$$

а пара $(u_*(\cdot), x_*(\cdot))$, где $x_*(\cdot)$ — решение задачи Коши

$$\dot{x}(\tau) = f(x(\tau), u(x(\tau), \tau), \tau), \quad t_0 \leq \tau \leq T, \quad x(t_0) = x_0^*,$$

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 $(u(\tau), x(\tau))$, $t_0 \leq \tau \leq T$, $x(t_0) = x_0 \in X(t_0)$ задачи (1.1)–(1.4). Из формулы (9) и неравенств (12) при $t = t_0$ с учетом условия (14) имеем

$$\begin{aligned} J(t_0, x_0, u(\cdot)) - J(t_0, x_0^*, u_*(\cdot)) &= \\ &= \int_{t_0}^T R(x(\tau), u(\tau), \tau) d\tau + B(x_0, t_0) - \inf_{x \in X(t_0)} B(x, t_0) \geq 0, \end{aligned}$$

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:

$$\begin{aligned} \inf_{u \in V(t)} [\langle B_x(x, t), f(x, u, t) \rangle + B_t(x, t) + f^0(x, u, t)] &= 0, \\ x \in G(t), \quad t_0 \leq t < T, \\ B(x, T) = \Phi(x), \quad x \in G(T), \end{aligned} \quad (15)$$

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 представляя 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

$$\inf_{u \in V(t)} [\langle B_x, f(x, u, t) \rangle + f^0(x, u, t)]$$

при фиксированных значениях параметров (x, t, B_x) . Подставив такое $u = u(x, t, B_x)$ в (15), приходим к следующей задаче Коши:

$$\begin{aligned} B_t + [\langle B_x, f(x, u, t) \rangle + f^0(x, u, t)]|_{u=u(x, t, B_x)} &= 0, \\ x \in G(t), \quad t_0 \leq t < T, \\ B(x, T) = \Phi(x), \quad x \in G(T), \end{aligned}$$

для нелинейного уравнения с частными производными первого порядка. Для численного решения этой задачи Коши можно воспользоваться известным арсеналом методов — разностными методами, методом характеристик, методом прямых, методом коллокации и т. п. [2, 40, 129, 211].

Иногда удается найти решение $B(x, t)$ задачи (15) в виде многочлена по переменным $x^1, x^2, \dots, x^n$ с неопределенными коэффициентами, зависящими от времени:

$$B(x, t) = \sum_{i_1=0}^{m_1} \sum_{i_2=0}^{m_2} \dots \sum_{i_n=0}^{m_n} \psi_{i_1 i_2 \dots i_n}(t) (x^1)^{i_1} (x^2)^{i_2} \dots (x^n)^{i_n}.$$

Если подставим это выражение для $B(x, t)$ в (15), то для определения коэффициентов $\psi_{i_1 i_2 \dots i_n}(t)$ получим дифференциальное уравнение следующего вида:

$$\begin{aligned} B_t(x, t) &= \sum_{i_1=0}^{m_1} \dots \sum_{i_n=0}^{m_n} \dot{\psi}_{i_1 \dots i_n}(t) (x^1)^{i_1} \dots (x^n)^{i_n} = \\ &= \inf_{u \in V(t)} F(\psi_0 \dots, (t), \dots, \psi_{m_1 \dots m_n}(t); x^1, \dots, x^n, u, t), \quad (16) \\ x &\in G(t), \quad t_0 \leq t \leq T, \end{aligned}$$

с начальным условием

$$\sum_{i_1=0}^{m_1} \dots \sum_{i_n=0}^{m_n} \psi_{i_1 \dots i_n}(T) (x^1)^{i_1} \dots (x^n)^{i_n} = \Phi(x), \quad x \in G(T). \quad (17)$$

Если $\Phi(x)$, $\inf_{V(t)} F$, в свою очередь, являются многочленами относительно

$x^1, \dots, x^n$, то, приравняв коэффициенты при одинаковых степенях в (16), (17), получим задачу Коши для системы обыкновенных дифференциальных уравнений относительно $\psi_{i_1 \dots i_n}(t)$, записанной в нормальной форме Коши.

Далее здесь можно использовать различные численные методы решения задачи Коши, такие, как методы Эйлера, Адамса, Рунге — Кутты и т. д. [2, 40, 129, 211].

Если же $\Phi(x)$ или $\inf_{V(t)} F$ не являются многочленами относительно $x^1, \dots, x^n$, то условия (16), (17) не могут быть, вообще говоря, удовлетворены во всей области $G(t)$, $t_0 \leq t \leq T$, ни при каком выборе $N = (m_1 + 1)(m_2 + 1) \dots (m_n + 1)$ коэффициентов $\psi_{i_1 \dots i_n}(t)$. В этом случае в [152] предлагается задать в области $G(t)$ N кривых $\xi_1(t), \dots, \xi_N(t)$ и рекомендуется определять $\psi_{i_1 \dots i_n}(t)$ из условия удовлетворения равенств (16), (17) не всюду в $G(t)$, а лишь на этих кривых. Этот подход переключается с известными методами коллокации и интегральных соотношений и приводит к задаче Коши для системы обыкновенных дифференциальных уравнений, не разрешенных относительно производной $\dot{\psi}_{i_1 \dots i_n}(t)$ (отметим, что эти производные в уравнения будут входить линейно). Кривые $\xi_1(t), \dots, \xi_N(t)$ 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 $G(t)$, $t_0 \leq t \leq T$.

Для иллюстрации вышесказанного приведем пример.

Пример 1. Пусть требуется минимизировать функцию

$$J(u) = \int_0^T u^2(t) dt + \lambda x^2(T), \quad \lambda = \text{const} > 0,$$

при условиях $\dot{x} = u(t)$, $x(0) = x_0$, $u = u(t)$ — кусочно непрерывная функция; числа T , x_0 заданы. Здесь $G(t) \equiv E^1$, $V(t) \equiv E^1$, $0 \leq t \leq T$.

Задача (15) в рассматриваемом случае имеет вид

$$\inf_{u \in E^1} [B_x(x, t)u + B_t(x, t) + u^2] = 0, \quad x \in E^1, \quad 0 \leq t \leq T; \quad (18)$$

$$B(x, T) = \lambda x^2, \quad x \in E^1. \quad (19)$$

Нижняя грань в (18) достигается при $u = -B_x/2$, поэтому уравнение (18) переписывается так:

$$B_t(x, t) - B_x^2(x, t)/4 = 0, \quad x \in E^1, \quad 0 \leq t \leq T. \quad (20)$$

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

$$\dot{\psi}_0 + \dot{\psi}_1 x + \dot{\psi}_2 x^2 - (\psi_1 + 2\psi_2 x)^2/4 = 0, \quad x \in E^1, \quad 0 \leq t \leq T;$$

$$\psi_0(T) + \psi_1(T)x + \psi_2(T)x^2 = \lambda x^2, \quad x \in E^1.$$

Приравнявая коэффициенты при одинаковых степенях x , придем к следующей задаче Коши:

$$\dot{\psi}_0 - \psi_1^2/4 = 0, \quad \dot{\psi}_1 - \psi_1\psi_2 = 0, \quad \dot{\psi}_2 - \psi_2^2 = 0, \quad 0 \leq t \leq T;$$

$$\psi_0(T) = 0, \quad \psi_1(T) = 0, \quad \psi_2(T) = \lambda.$$

Отсюда находим

$$\psi_0(t) \equiv \psi_1(t) \equiv 0, \quad \psi_2(t) = \frac{\lambda}{1 - \lambda(t - T)}.$$

Таким образом, функция Беллмана здесь имеет вид

$$B(x, t) = \frac{\lambda x^2}{1 - \lambda(t - T)};$$

синтезирующей является функция

$$u(x, t) = -B_x/2 = \frac{\lambda x}{1 - \lambda(t - T)}, \quad x \in E^1, \quad 0 \leq t \leq T.$$

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 $G(t)$, $t_0 \leq t \leq T$, 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

$$\begin{aligned}\dot{\bar{x}}(\tau) &= f(\bar{x}(\tau), u(\bar{x}(\tau), \tau), \tau), \quad t \leq \tau \leq T; \quad \bar{x}(t) = x, \\ \bar{x}(\tau) &\in G(\tau), \quad \bar{u}(\tau) = u(\bar{x}(\tau), \tau), \quad t \leq \tau \leq T.\end{aligned}\quad (21)$$

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 $\bar{u}(\tau) = u(\bar{x}(\tau), \tau)$, $t_0 \leq \tau \leq T$. 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 $x \in X(t)$, $t_0 \leq t \leq T$, обладает кусочно непрерывными производными K_x, K_t и такова, что для любой допустимой пары $(u(\cdot), x(\cdot))$ задачи (1)–(4) при всех $x \in X(t)$, $t_0 \leq t < T$, 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

$$\begin{aligned}S(x, u, t) &= \langle K_x(x, t), f(x, u, t) \rangle + K_t(x, t) + f^0(x, u, t), \\ x &\in X(t), \quad t_0 \leq t \leq T,\end{aligned}\quad (22)$$

и, кроме того, положим

$$s(x, T) = \Phi(x) - K(x, T), \quad x \in G(T). \quad (23)$$

Возьмем произвольную допустимую пару $(u(\cdot), x(\cdot))$ задачи (1)–(4). В силу уравнения (2) тогда имеем

$$\frac{dK(x(\tau), \tau)}{d\tau} = S(x(\tau), u(\tau), \tau) - f^0(x(\tau), u(\tau), \tau), \quad t_0 \leq \tau \leq T.$$

Учитывая непрерывность и кусочную гладкость функции $K(x(\tau), \tau)$, проинтегрируем это тождество по τ от t до T . Получим формулу

$$J(t, x, u(\cdot)) = \int_t^T S(x(\tau), u(\tau), \tau) d\tau + s(x(T), T) + K(x, t). \quad (24)$$

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 $(\bar{u}(\tau), \bar{x}(\tau))$, $t \leq \tau \leq T$, удовлетворяющую условиям (3), (4) и уравнению (2) с начальным условием $\bar{x}(t) = \bar{x} \in X(t)$. Согласно (24) тогда

$$\begin{aligned}J(t, \bar{x}, \bar{u}(\cdot)) - J(t, x, u(\cdot)) &= \int_t^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau) - S(x(\tau), u(\tau), \tau)] d\tau + \\ &+ [s(\bar{x}(T), T) - s(x(T), T)] + [K(\bar{x}, t) - K(x, t)]\end{aligned}\quad (25)$$

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

$$\begin{aligned}S_{\min}(\tau) &= \inf_{x \in X(\tau)} \inf_{u \in D(x, \tau)} S(x, u, \tau), \quad s_{\min} = \inf_{x \in X(T)} s(x, T), \\ K_{0 \min} &= \inf_{x \in X(t_0)} K(x, t_0).\end{aligned}\quad (26)$$

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 $x(t) = x \in X(t)$, $u(\cdot) \in \Delta(x, t)$, $x(\tau) \in X(\tau)$, $u(\tau) \in D(x(\tau), \tau)$, $t \leq \tau \leq T$, из (25), (26) получим требуемую оценку погрешности:

$$0 \leq J(t, x, \bar{u}(\cdot)) - \inf_{\Delta(x, t)} J(t, x, u(\cdot)) \leq \int_t^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau) - S_{\min}(\tau)] d\tau + [s(\bar{x}(T), T) - s_{\min}]. \quad (27)$$

Если же $(\bar{u}(\tau), \bar{x}(\tau))$, $t_0 \leq \tau \leq T$, — 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:

$$0 \leq J(t_0, \bar{x}_0, \bar{u}(\cdot)) - \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot)) \leq \int_{t_0}^T [S(\bar{x}(\tau), \bar{u}(\tau), \tau) - S_{\min}(\tau)] d\tau + [s(\bar{x}(T), T) - s_{\min}] + K(\bar{x}_0, t_0) - K_0 \min. \quad (28)$$

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 $\inf_{u \in D(x, t)} R(x, u, t) = 0$ при всех $x \in X(t)$, так что $\inf_{x \in X(t)} \inf_{u \in D(x, t)} R(x, u, t) = 0$. Поэтому при $K(x, t) = B(x, t)$ из (27), (28) соответственно получим

$$0 \leq J(t, x, \bar{u}(\cdot)) - \inf_{u(\cdot) \in \Delta(x, t)} J(t, x, u(\cdot)) \leq \int_t^T R(\bar{x}(\tau), \bar{u}(\tau), \tau) d\tau, \quad (29)$$

$$0 \leq J(t_0, \bar{x}_0, \bar{u}(\cdot)) - \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot)) \leq \int_{t_0}^T R(\bar{x}(\tau), \bar{u}(\tau), \tau) d\tau + B(\bar{x}_0, t_0) - \inf_{x \in X(t_0)} B(x, t_0). \quad (30)$$

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 $\inf_{\Delta(x, t)} R(x, u, t)$ [$B(\bar{x}_0, t_0)$ — поближе к $\inf_{x \in X(t_0)} B(x, t_0)$] и затем найти пару $(\bar{u}(\tau), \bar{x}(\tau))$ из условий

$$\begin{aligned} \dot{\bar{x}}(\tau) &= f(\bar{x}(\tau), \bar{u}(\bar{x}(\tau), \tau), \tau), \quad \bar{x}(\tau) \in G(\tau), \quad t \leq \tau \leq T; \\ \bar{x}(t) &= x; \quad \bar{u}(\tau) = \bar{u}(\bar{x}(\tau), \tau) \end{aligned}$$

[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 $J(x_0, u(\cdot)) = \int_0^T (x^2(t) + u^2(t)) dt$ 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 $\dot{x}(t) = u(t)$, $0 \leq t \leq 1$,

$x(0) = x_0$, $u(t) \in V = \{u \in E^1: 0 \leq u \leq 1\}$, $x(t) \in G(t) \equiv E^1$ при $0 \leq t \leq 1$.

Показать, что в этой задаче синтезирующих функций бесконечно много. Убедиться, например, что синтезирующими являются функции

$$u(x, t) = \begin{cases} 0, & x \geq 0, \\ 1, & x < 0, \end{cases} \quad \text{или} \quad u(x, t) = \begin{cases} 0, & x > t - 1, \\ 1, & x \leq t - 1. \end{cases}$$

3. Решить проблему синтеза для задачи минимизации функций

$$J(x_0, u(\cdot)) = x^2(T) \quad \text{или} \quad J(x_0, u(\cdot)) = \int_0^T x^2(t) dt$$

при условиях $\dot{x}(t) = u(t)$, $0 \leq t \leq T$; $x(0) = x_0$, $u(t) \in V(t) = \{u \in E^1: -1 \leq u \leq 1\}$, $x(t) \in G(t) \equiv E^1$, $0 \leq t \leq T$. Будет ли в этих задачах синтезирующая функция единственной?

4. Написать уравнения Беллмана (5), (6) для задачи быстрогодействия:

$$J(x_0, t_0, u(\cdot)) = T - t_0 \rightarrow \inf; \quad \dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T,$$

$x(t_0) = x_0$, $x(T) = x_1$, $u(\cdot)$ — кусочно непрерывна и принадлежит множеству $V \subseteq E^r$, $t \geq t_0$.

5. Показать, что функция Беллмана для задачи из примера 6.2.4 не является непрерывно дифференцируемой.

6. Найти функцию Беллмана для задачи

$$J(t_0, x_0, u(\cdot)) = \int_{t_0}^T [\langle a(t), x(t) \rangle + b(u(t), t)] dt + \langle c, x(T) \rangle \rightarrow \inf;$$

$$\dot{x}(t) = A(t)x(t) + C(t)u(t) + f(t), \quad t_0 \leq t \leq T; \quad x(t_0) = x_0, \quad (31)$$

$$u = u(t) \in V(t), \quad u(\cdot) \text{ — кусочно непрерывна,} \quad (32)$$

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 $V(t) \subseteq E^r$, $t_0 \leq t \leq T$. Указания: пользуясь уравнениями (5), (6), искать функцию $B(x, t)$ в виде $B(x, t) = \langle \psi(t), x \rangle$ — многочлена первой степени относительно переменных $x = (x^1, \dots, x^n)$.

7. Исследовать уравнения (5), (6) для задачи минимизации функции

$$J(t_0, x_0, u(\cdot)) = \alpha_1 \int_{t_0}^T x^2(t) dt + \alpha_2 x^2(T), \quad \alpha_1, \alpha_2 = \text{const} \geq 0,$$

при условиях (31), (32). Рассмотреть случаи $V(t) = E^r$, $V(t) = \{u \in E^r: |u| \leq 1\}$, $V(t) = \{u = (u^1, \dots, u^r): -1 \leq u^i \leq 1, i = \overline{1, r}\}$.

§ 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 [46, 71, 91, 152, 153, 163, 206]. 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 $(\bar{u}(t), \bar{x}(t))$, $t_0 \leq t \leq T$, $\bar{x}(t_0) = \bar{x}_0$, задачи (1.1)–(1.4) справедливы формула

$$J(t_0, \bar{x}_0, \bar{u}(\cdot)) = \int_{t_0}^T S(\bar{x}(t), \bar{u}(t), t) dt + s(x(T), T) + K(\bar{x}_0, t_0) \quad (1)$$

и оценка

$$0 \leq J(t_0, \bar{x}_0, \bar{u}(\cdot)) - J_* \leq \int_{t_0}^T [S(\bar{x}(t), \bar{u}(t), t) - S_{\min}(t)] dt + \\ + [s(\bar{x}(T), T) - s_{\min}] + [K(\bar{x}_0, t_0) - K_0 \min], \quad (2)$$

где

$$S(x, u, t) = \langle K_x(x, t), f(x, u, t) \rangle + K_t(x, t) + f^0(x, u, t), \quad (3)$$

$$s(x, T) = \Phi(x) - K(x, T), \quad (4)$$

$$S_{\min}(t) = \inf_{x \in X(t)} \inf_{u \in D(x, t)} S(x, u, t), \quad s_{\min} = \inf_{x \in X(T)} s(x, T), \quad (5)$$

$$K_0 \min = \inf_{x \in X(t_0)} K(x, t_0), \quad J_* = \inf_{x \in X(t_0)} \inf_{u(\cdot) \in \Delta(x, t_0)} J(t_0, x, u(\cdot));$$

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 непрерывными 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 $(\bar{u}(t), \bar{x}(t))$, $t_0 \leq t \leq T$, $\bar{x}(t_0) = \bar{x}_0$, 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

$$S(\bar{x}(t), \bar{u}(t), t) = S_{\min}(t), \quad t_0 \leq t \leq T, \quad (6)$$

$$s(x(T), T) = s_{\min}, \quad K(\bar{x}_0, t_0) = K_0 \min, \quad (7)$$

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 $(u_m(t), x_m(t))$, $t_0 \leq t \leq T$, $x_m(t_0) = x_{0m}$, $m = 1, 2, \dots$, 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

$$\lim_{m \rightarrow \infty} \int_{t_0}^T S(x_m(t), u_m(t), t) dt = \int_{t_0}^T S_{\min}(t) dt, \quad (8)$$

$$\lim_{m \rightarrow \infty} s(x_m(T), T) = s_{\min}, \quad \lim_{m \rightarrow \infty} K(x_{0m}, t_0) = K_{0 \min}. \quad (9)$$

Доказательство. В оценке (2) вместо $\bar{u}(t)$, $\bar{x}(t)$ подставим $u_m(t)$, $x_m(t)$ и перейдем к пределу при $m \rightarrow \infty$. С учетом условий (8), (9) получим $\lim_{m \rightarrow \infty} J(t_0, x_{0m}, u_m(\cdot)) = J_*$, 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 $(\bar{u}(t), \bar{x}(t))$, $t_0 \leq t \leq T$ [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 $\bar{K}(x, t) = K(x, t) + \alpha(t)$, где $\alpha(t)$ — произвольная непрерывная, кусочно гладкая (или абсолютно непрерывная) на $[t_0, T]$ функция. В частности, если $\alpha(t) = -\int_{t_0}^t S_{\min}(\tau) d\tau - s_{\min}$,

то функции $\bar{S}(x, u, t)$, $\bar{s}(x, T)$, построенные по формулам (3), (4) с заменой K на $\bar{K} = K + \alpha$, таковы, что $\bar{S}(x, u, t) = S(x, u, t) - S_{\min}(t)$, $\bar{s}(x, T) = s(x, T) - s_{\min}$ и, следовательно, $\inf_{x \in X(t)} \inf_{u \in D(x, t)} \bar{S}(x, u, t) \equiv 0$, $t_0 \leq t \leq T$, $\inf_{x \in X(T)} \bar{s}(x, T) = 0$, а $\inf_{x \in X(t_0)} \bar{K}(x, t_0) = K_{0 \min} + \alpha(t_0)$ отличается 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

$$S(x, u, t) = \langle K_x(x, t), f(x, u, t) \rangle + K_t(x, t) + f^0(x, u, t) \geq 0, \quad (10)$$

$$u \in D(x, t), \quad x \in X(t), \quad t_0 \leq t \leq T;$$

$$s(x, T) = \Phi(x) - K(x, T) \geq 0, \quad x \in X(T); \quad K(x, t_0) \geq K_{0 \min}, \quad x \in X(t_0), \quad (11)$$

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 $m \rightarrow \infty$.

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) показывает, 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 $D(x, t)$, $\bar{X}(t)$, то функцию Кротова можно попытаться определить из условий, получающихся из (10), (11) при замене $D(x, t)$, $\bar{X}(t)$ 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 $\dot{x}(t) = u(t)$, $x(0) = x(1) = 0$, $u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq 1$.

Здесь фазовые ограничения $G(t)$ такие: $G(0) = G(1) = \{0\}$, $G(t) = E^1$, $0 < t < 1$. 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

$$S(x, u, t) = x^2 \geq 0 = S(\bar{x}(t), \bar{u}(t), t) = S_{\min}(t);$$

$$s(x, 1) = -x = 0 = s(\bar{x}(1), 1) = \inf_{x \in G(1)} s(x, 1);$$

$$K(x, 0) = x = 0 = K(\bar{x}(0), 0) = \inf_{x \in G(0)} K(x, 0).$$

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 $\dot{x}(t) = u(t)$, $x(0) = 0$, $u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq 1$.

Здесь фазовые ограничения $G(t)$ такие: $G(0) = \{0\}$, $G(t) = E^1$, $0 < t \leq 1$. Возьмем последовательность пар функций $(u_m(\cdot), x_m(\cdot))$, где

$$u_m(t) = \begin{cases} +1 & \text{при } \frac{p}{m} < t \leq \frac{p}{m} + \frac{1}{2m}, \\ -1 & \text{при } \frac{p}{m} + \frac{1}{2m} < t \leq \frac{p}{m} + \frac{1}{m}, \end{cases}$$

$$x_m(t) = \begin{cases} t - \frac{p}{m} & \text{при } \frac{p}{m} < t \leq \frac{p}{m} + \frac{1}{2m}, \\ -t + \frac{p+1}{m} & \text{при } \frac{p}{m} + \frac{1}{2m} < t \leq \frac{p}{m} + \frac{1}{m}, \end{cases}$$

$$p = 0, m-1; \quad m = 1, 2, \dots$$

Нетрудно проверить, что пара $(u_m(\cdot), x_m(\cdot))$ допустима для рассматриваемой задачи при всех $m = 1, 2, \dots$. Покажем, что последовательность этих пар является минимизирующей. Для этого возьмем функцию $K(x, t) = t - 1$. Тогда $S(x, u, t) = x^2 + 1 - u^2 \geq 0 = \min_{x \in E^1} \min_{|u| \leq 1} S(x, u, t) = \lim_{m \rightarrow \infty} S(x_m(t),$

$u_m(t), t, s(x, 1) = -K(x, 1) \equiv 0 = \inf_{x \in E^1} K(x, 1) = \lim_{m \rightarrow \infty} s(x_m(1), 1),$
 $K(x, 0) \equiv -1 = \inf_{x \in G(0)} K(x, 0) = \lim_{m \rightarrow \infty} K(x_m(0), 0).$ Согласно теореме 2 по-
 следовательно $(u_m(\cdot), x_m(\cdot))$ будет минимизирующей: $\lim_{m \rightarrow \infty} J(u_m(\cdot)) =$
 $= -1 = J_*$.

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 $x^2(t) \neq 0$, то

$$J(u) > - \int_0^1 u^2(t) dt \geq -1. \text{ Таким образом, } J(u(\cdot)) > -1 \text{ для всех допусти-}$$

мых управлений и траекторий. В этой задаче мы имеем дело с так называемым скользящим режимом [54, 74, 91, 152, 153, 225]. Предлагаем читателю найти функцию Беллмана и приближенную синтезирующую функцию этой задачи.

2. Перейдем к рассмотрению следующей задачи оптимального управления с незакрепленным временем: минимизировать функцию

$$J(t_0, T, x_0, u(\cdot)) = \int_{t_0}^T f^0(x(t), u(t), t) dt + \Phi(x(T), T) \quad (12)$$

при условиях

$$\dot{x}(t) = f(x(t), u(t), t), \quad t_0 \leq t \leq T; \quad x(t_0) = x_0, \quad (13)$$

$$x(t) \in G(t), \quad t_0 \leq t \leq T; \quad x(T) \in S_I(T) \subseteq G(T); \quad (14)$$

$u = u(t) \in V(t), u(t)$ — кусочно непрерывна, $t_0 \leq t \leq T$,

$$u(t) = u(t+0) = \lim_{\tau \rightarrow t+0} u(\tau), \quad t_0 \leq t < T, \quad (15)$$

где моменты t_0, T , в отличие от задачи (1.1) — (1.4), неизвестны и таковы, что

$$t_0 \in \theta_0, \quad T \in \theta_1; \quad (16)$$

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 быстрогодействия. Всюду ниже будем предполагать, что $t_0 \leq T$.

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 $x(\tau) \in G(\tau), t \leq \tau \leq T, x(T) \in S_I(T)$. Положим $X(t, T) = \{x: x \in G(t), \Delta(x, t, T) \neq \emptyset\}$ при $t < T$; $X(T, T) \equiv S_I(T)$. 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 $(u(\tau), x(\tau)), t_0 \leq \tau \leq T$, 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 $t, t_0 \leq t < T$.

Моменты времени $t_0^* \in \Theta_0$, $T^* \in \Theta_1$ и допустимую пару $(u_*(t), x_*(t))$, определенную на отрезке $[t_0^*, T^*]$, назовем решением задачи (12) — (16), если

$$J(t_0^*, T^*, x_*(t_0), u_*(\cdot)) = \inf_{t_0 \in \Theta_0} \inf_{T \in \Theta_1} \inf_{x_0 \in X(t_0, T)} \inf_{u(\cdot) \in \Delta(x_0, t_0, T)} J(t_0, T, x_0, u(\cdot)) = J_*$$

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 $\lim_{m \rightarrow \infty} J(t_{0m}, T_m, x_m(t_{0m}), u_m(\cdot)) = J_*$.

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 $(u(t), x(t))$, $t_0 \leq t \leq T$, 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 $[t_0^*, T^*]$ и $[t_0, T]$ соответственно. Тогда из формулы (1) получим

$$J(t_0^*, T^*, x_*(t_0^*), u_*(\cdot)) - J(t_0, T, x(t_0), u(\cdot)) = \int_{t_0^*}^{T^*} S(x_*(t), u_*(t), t) dt - \int_{t_0}^T S(x(t), u(t), t) dt + [s(x_*(T^*), T^*) - s(x(T), T)] + [K(x_*(t_0^*), t_0^*) - K(x(t_0), t_0)]. \quad (17)$$

Обозначим

$$S_{\min} = \inf_{t_0 \in \Theta_0} \inf_{T \in \Theta_1} \int_{t_0}^T \inf_{x \in X(t, T)} \inf_{u \in D(x, t, T)} S(x, u, t) dt, \quad (18)$$

$$s_{\min} = \inf_{T \in \Theta_1} \inf_{x \in S_1(T)} s(x, T), \quad K_{0 \min} = \inf_{t_0 \in \Theta_0} \inf_{T \in \Theta_1} \inf_{x \in X(t_0, T)} K(x, t_0).$$

Учитывая, что для любой допустимой пары $(u(t), x(t))$, $t_0 \leq t \leq T$, задачи (12) — (16) имеют место включения $u(t) \in D(x(t), t, T)$, $x(t) \in X(t, T)$ при всех t , $t_0 \leq t < T$, 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:

$$0 \leq J(t_0^*, T^*, x_*(t_0^*), u_*(\cdot)) - J_* \leq \int_{t_0^*}^{T^*} S(x_*(t), u_*(t), t) dt - S_{\min} + [s(x_*(T^*), T^*) - s_{\min}] + [K(x_*(t_0^*), t_0^*) - K_{0 \min}]. \quad (19)$$

Неравенства (19) обобщают формулу (2) на случай задач оптимального управления с незакрепленным временем и представляют собой оценку погрешности

сти, которая будет допущена, если допустимую пару $(u_*(t), x_*(t))$, $t_0^* \leq t \leq T^*$, 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 $(u_*(t), x_*(t))$, $t_0^* \leq t \leq T^*$, 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

$$\int_{t_0^*}^{T^*} S(x_*(t), u_*(t), t) dt = S_{\min}, \quad (20)$$

$$s(x_*(T^*), T^*) = s_{\min}, \quad K(x_*(t_0^*), t_0^*) = K_0 \min. \quad (21)$$

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 $(u_m(t), x_m(t))$, $t_0 m \leq t \leq T_m$, $m = 1, 2, \dots$, 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

$$\lim_{m \rightarrow \infty} \int_{t_0 m}^{T_m} S(x_m(t), u_m(t), t) dt = S_{\min}, \quad (22)$$

$$\lim_{m \rightarrow \infty} s(x_m(T_m), T_m) = s_{\min}, \quad \lim_{m \rightarrow \infty} K(x_m(t_0 m), t_0 m) = K_0 \min. \quad (23)$$

Доказательство. В оценке (19) вместо $u_*(t)$, $x_*(t)$, t_0^* , T^* подставим соответственно $u_m(t)$, $x_m(t)$, $t_0 m$, T_m и перейдем к пределу при $m \rightarrow \infty$. 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 наиболее быстрым образом перевести точку $(x, y) \in E^2$ из начала координат $(0, 0)$ в точку $(1, 0)$, предполагая, что движение точки подчиняется условиям $\dot{x}(t) = -y^2(t) + u^2(t)$, $\dot{y}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq T$.

Здесь $G(0) = \{(0, 0)\}$, $G(t) = E^2$ при $0 < t \leq T$, $S_1(T) = \{(1, 0)\}$, $\theta_0 = \{0 = 0\}$, $\theta_1 = \{T: T \geq 0\}$, $j^0 = 1$, $\Phi = 0$, $J(T, u(t)) = T$.

Пусть T_m — корень уравнения $(T-1)^3 = 12(T-1) - m^{-2}$, расположенный в пределах $1 < T < 1 + m^{-2/3}$, $m = 1, 2, \dots$. Положим

$$u_m(t) = \begin{cases} 1 & \text{при } \frac{p}{m} < t \leq \frac{p}{m} + \frac{1}{2m} & \text{и } 1 \leq t \leq (T_m + 1)/2, \\ -1 & \text{при } \frac{p}{m} + \frac{1}{2m} < t \leq \frac{p}{m} + \frac{1}{m} & \text{и } (T_m + 1)/2 < t \leq T_m, \end{cases}$$

$p = 0, m-1, m=1, 2, \dots$ Через $(x_m(t), y_m(t))$, $0 \leq t \leq T_m$, обозначим траекторию точки, соответствующую управлению $u_m(\cdot)$ и начальным условиям $x_m(0) = y_m(0) = 0$. Нетрудно видеть, что $x_m(T_m) = 1, y_m(T_m) = 0$, $(1 - m^{-4/3})t \leq x_m(t) \leq t, 0 \leq y_m(t) \leq m^{-2/3}$ при $0 \leq t \leq T_m, m=1, 2, \dots$

Покажем, что $\lim_{m \rightarrow \infty} T_m = 1 = T^*$ — оптимальное время. Для этого возьмем функцию $K(x, y, t) = -x$. Тогда

$$S(x, y, u, t) = y^2 - u^2 + 1, \quad \inf_{(x, y) \in E^2} \inf_{|u| \leq 1} S(x, y, u, t) \equiv 0$$

при всех $t \geq 0, S_{\min} = 0, \int_0^{T_m} S(x_m(t), y_m(t), u_m(t), t) dt = \int_0^{T_m} y_m^2(t) dt - \int_0^{T_m} (u_m^2(t) - 1) dt = \int_0^{T_m} y_m^2(t) dt \rightarrow 0 = S_{\min}$ при $m \rightarrow \infty$; $s(x, y, T) = -K(x, y, T) = -x = -1$ при $(x, y) \in S_f(T), s(x_m(T_m), y_m(T_m), T_m) = -1 = s_{\min}, K(x, y, 0) = 0$ при $(x, y) \in G(0), K(x_m(0), y_m(0), 0) = 0 = K_{0 \min}, m=1, 2, \dots$

Кроме того, очевидно, формула (1) для данной задачи будет справедлива при всех допустимых $(x(\cdot), y(\cdot), u(\cdot))$. Таким образом, для последовательностей $\{T_m\}, \{(x_m(t), y_m(t), u_m(t)), 0 \leq t \leq T_m\}$ все условия теоремы 4 выполнены. Следовательно, $\lim_{m \rightarrow \infty} J(T_m, u_m(\cdot)) = \lim_{m \rightarrow \infty} T_m = 1$ — 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 быстройдействия.

Теорема 5. Пусть в задаче (12)–(16) $f^0 \equiv 1, \Phi \equiv 0; \theta_0 = \{t_0\}$, т.е. начальный момент t_0 закреплён; $\theta_1 \equiv \{T: T \geq t_0\}$. Пусть имеется некоторая последовательность $(u_m(t), x_m(t)), t_0 \leq t \leq T_m, m=1, 2, \dots$, допустимых пар рассматриваемой задачи быстройдействия, причём $\lim_{m \rightarrow \infty} T_m = T^*$. 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,

$$\lim_{m \rightarrow \infty} \int_{t_1}^{T_m} S(x_m(t), u_m(t), t) dt < \int_t^T S_{\min}(t, T) dt + T^* - T \quad (24)$$

для любого $T, t_0 \leq T < T^*$,

$$\lim_{m \rightarrow \infty} s(x_m(T_m), T_m) = s_{\min}, \quad \lim_{m \rightarrow \infty} K(x_m(t_0), t_0) = K_{0 \min}, \quad (25)$$

где

$$S_{\min}(t, T) = \inf_{x \in X(t, T)} \inf_{u \in D(x, t, T)} S(x, u, t),$$

$$s_{\min} = \inf_{0 \leq T \leq T^*} \inf_{x \in S_f(T)} s(x, T), \quad K_{0 \min} = \inf_{t_0 \leq T \leq T^*} \inf_{x \in X(t_0, T)} K(x, t_0).$$

Для получения формулировки достаточного условия оптимальности для фиксированной допустимой пары $(u_*(t), x_*(t)), t_0 \leq t \leq T^*$, в этой теореме надо принять $T_m = T^*, u_m(t) = u_*(t), x_m(t) = x_*(t), m=1, 2, \dots$ и в (24), (25) всюду опустить знак $\lim$.

Доказательство. Пусть вопреки утверждению теоремы T^* не является оптимальным временем. Тогда существуют момент T , $t_0 \leq T < T^*$, и допустимая пара $(u(t), x(t))$, $t_0 \leq t \leq T$. В формуле (17) вместо t_0^* , T^* , $(u_*(t), x_*(t))$, $t_0^* \leq t \leq T^*$, примем соответственно t_0 , T , $(u_m(t), x_m(t))$, $t_0 \leq t \leq T_m$, и перейдем к пределу при $m \rightarrow \infty$. С учетом условий (24), (25) будем иметь

$$\begin{aligned} \lim_{m \rightarrow \infty} J(t_0, T_m, x_m(t_0), u_m(\cdot)) - J(t_0, T, x(t_0), u(\cdot)) &= T^* - T \leq \\ &\leq \lim_{m \rightarrow \infty} \int_{t_0}^{T_m} S(x_m(t), u_m(t), t) dt - \int_{t_0}^T S(x(t), u(t), t) dt < T^* - T. \end{aligned}$$

Полученное противоречивое неравенство доказывает теорему.

Пример 4. Пусть требуется наиболее быстрым образом перевести точку $(x, y) \in E^2$ из положения $(1, 0)$ в начало координат $(0, 0)$, предполагая, что движение точки подчиняется условиям $\dot{x}(t) = y(t)$, $\dot{y}(t) = u(t)$, $u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}$, $0 \leq t \leq T$.

Здесь $G(0) = \{(1, 0)\}$, $G(t) \equiv E^2$ при $0 \leq t < T$, $S_f(T) = \{(0, 0)\}$, $\theta_0 = \{t_0 = 0\}$, $\theta_1 = \{T \geq 0\}$.

В примере 6.2.4 с помощью принципа максимума была найдена допустимая пара $((x_*(t), y_*(t)), u_*(t))$, $0 \leq t \leq T^* = 2$, где

$$\begin{aligned} u_*(t) &= \begin{cases} -1, & 0 \leq t \leq 1, \\ +1, & 1 < t \leq 2, \end{cases} & x_*(t) &= \begin{cases} 1 - t^2/2, & 0 \leq t \leq 1, \\ (t-2)^2/2, & 1 \leq t \leq 2, \end{cases} \\ y_*(t) &= \begin{cases} -t, & 0 \leq t \leq 1, \\ t-2, & 1 \leq t \leq 2. \end{cases} \end{aligned}$$

Покажем, что эта пара является решением рассматриваемой задачи быстрого действия. Возьмем функцию $K(x, y, t) = x - (t-1)y$. Тогда $S(x, y, u, t) = K_{xy} + K_y u + K_t + 1 = -(t-1)u + 1$, $S_{\min}(t, T) = \inf_{(x, y) \in E^2} \inf_{|u| \leq 1} S(x, y, u, t) = -|t-1| + 1$ при всех T и $t \geq 0$, $S(x_*(t), y_*(t), u_*(t), t) = -|t-1| + 1$;

$\int_0^{T^*} S(x_*(t), y_*(t), u_*(t), t) dt - \int_0^T S_{\min}(t, T) dt = \int_0^2 (1 - |t-1|) dt < 2 - T = T^* - T$ для всех T , $0 \leq T < T^* = 2$; $s(x, y, T) = -K(x, y, T) = 0$ при $(x, y) \in S_f(T)$, $s(x_*(T), y_*(T), T^*) = 0 = s_{\min}$; $K(x, y, 0) = 1$ 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 [14, 46], 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 [6].

У п р а ж н е н и я. 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 $0 < T < \pi$. У к а з а н и е: 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 быстрогодействия: наиболее быстрым образом перевести точку $(x, y) \in E^2$ из заданного состояния (x_0, y_0) в начало координат $(0, 0)$, предполагая, что движение точки подчиняется одному из следующих условий:

а) $\dot{x}(t) = y(t), \dot{y}(t) = u(t), u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}, 0 \leq t \leq T$;

б) $\dot{x}(t) = y(t), \dot{y}(t) = -x(t) + u(t), u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}, 0 \leq t \leq T$;

в) $\dot{x}(t) = y(t) + u(t), \dot{y}(t) = -x(t) + v(t), (u(t), v(t)) \in V(t) = \{(u, v) \in E^2: |u| \leq 1, |v| \leq 1\}, 0 \leq t \leq T$.

У к а з а н и е: 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 $(a, 0, 0)$ быстрым образом, если $\dot{x}(t) = y(t), \dot{y}(t) = z(t), \dot{z}(t) = u(t), u(t) \in V(t) = \{u \in E^1: |u| \leq 1\}, 0 \leq t \leq T; a = \text{const}$. Показать, что оптимальное время $T^* = (32|a|)^{1/3}$. У к а з а н и е: 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.

паспорте. Это короткий отрезок времени, в течение которого температура двигателя не успевает достичь установившейся. После отключения двигатель длительно не работает и его температура постепенно снижается до температуры окружающей среды. В кратковременном режиме двигатели могут быть рассчитаны на стандартную продолжительность рабочего периода — 10, 30, 60 и 90 мин.

Повторно-кратковременным режимом работы двигателя называется такой, при котором кратковременные рабочие периоды номинальной нагрузки чередуются с паузами. Продолжительность рабочих периодов и пауз не настолько велика, чтобы перегревы отдельных частей двигателя при неизменной температуре окружающей среды могли достигнуть установившихся значений. Для этого режима установлены стандартные продолжительности включения (ПВ 15, 25, 40 и 60 %) с длительностью рабочего периода до 10 мин. При повторно-кратковременном режиме работы двигатель нагревается меньше, чем при непрерывной нагрузке, и поэтому его можно нагружать сильнее, чем при продолжительном номинальном режиме работы.

Перемежающимся номинальным режимом называют такой режим длительной работы электродвигателя, при котором кратковременные рабочие периоды чередуются с периодами холостого хода. При этом средняя температура достигает установившегося значения.

При эксплуатации электродвигателя недопустимы перегрузки сверх его максимального вращающего момента, так как это приводит к затормаживанию (опрокидыванию), которое при длительной перегрузке может вызвать выход двигателя из строя.

При выборе электродвигателя следует учитывать условия и его режимы работы, необходимую частоту вращения и т. д. Правильный выбор имеет большое значение. Если на машине установлен двигатель большей мощности, чем требуется, он работает с низким КПД. Двигатель же недостаточной мощности быстро перегревается и не обеспечивает нормальной работы установки.

В зависимости от условий окружающей среды применяют электродвигатели следующих исполнений: защищенные, закрытые обдуваемые, взрывозащищенные. В сельскохозяйственном производстве наиболее распространены электродвигатели первых двух групп.

В двигателях защищенного исполнения все вращающиеся и токоведущие части закрыты. Это предотвращает случайное прикосновение к токоведущим частям, попадание капель и посторонних предметов. Воздух, необходимый для охлаждения, поступает внутрь двигателя через вентиляционные окна.

Двигатели закрытого обдуваемого исполнения таких окон не имеют, поэтому их охлаждают с помощью вентилятора, установленного снаружи на валу двигателя и закрытого специальным ко-

жухом. Вентилятор засасывает воздух через отверстия в кожухе и направляет его вдоль ребер статора, предназначенных для улучшения охлаждения. Кроме того, предусмотрена вентиляция и внутри корпуса электродвигателя.

При выборе электродвигателя необходимо знать характер среды помещения, в котором он будет установлен.

При выборе электродвигателя по частоте вращения необходимо стремиться к тому, чтобы двигатель имел такую же частоту вращения, что и рабочая машина, так как это избавит от промежуточной передачи.

В общем случае задача выбора двигателя по частоте вращения — это задача экономическая. С увеличением частоты вращения двигателя его габаритные размеры уменьшаются, а значит, снижаются расход материалов и стоимость. Кроме того, высокоскоростные двигатели имеют большие значения коэффициентов полезного действия (η) и мощности ($\cos \varphi$) по сравнению с тихоходными. Однако большинство сельскохозяйственных машин тихоходные, поэтому между двигателем и рабочей машиной необходима механическая передача. С увеличением передаточного числа стоимость передачи растет. Таким образом, для каждой конкретной рабочей машины имеется своя оптимальная частота вращения.

Для правильного выбора мощности электродвигателя необходимо знать режим, в котором он будет работать. При продолжительном режиме работы правильно выбранный двигатель работает достаточно долго без перегрева сверх допустимой для данного класса изоляции температуры.

Различают работу двигателя при постоянной и переменной нагрузке. При постоянной нагрузке мощность двигателя $P_{\text{дв}}$ определяют в зависимости от мощности на валу машины $P_{\text{р.м}}$ по формуле

$$P_{\text{дв}} \geq P_{\text{р.м}} \quad (28.4)$$

Обычно мощность рабочей машины указывают в ее паспорте. Для некоторых машин мощность можно подсчитать по теоретическим или эмпирическим формулам, а также по нормативным показателям потребления электрической энергии на единицу продукции:

$$P_{\text{дв}} \geq W\Pi, \quad (28.5)$$

где W — потребление энергии на производство единицы продукции, кВт · ч/т; Π — производительность рабочей машины, т/ч.

Выбранный по каталогу двигатель проверяют на возможность пуска, так как рабочие машины обычно имеют повышенное сопротивление в момент трогания с места. В общем случае пуско-

ОСНОВНАЯ ЛИТЕРАТУРА

1. Алексеев В. М., Тихомиров В. М., Фомин С. В. Оптимальное управление. — М.: Наука, 1979.
2. Бахвалов Н. С. Численные методы, т. I. — М.: Наука, 1973.
3. Бублик Б. Н., Кириченко Н. Ф. Основы теории управления. — Киев: Вища школа, 1975.
- ✓ 4. Габасов Р., Кириллова Ф. М. Методы оптимизации — Минск: Изд-во Белорусского ун-та, 1975.
5. Гермейер Ю. Б. Введение в теорию исследования операций. — М.: Наука, 1971.
6. Зубов В. И. Лекции по теории управления. — М.: Наука, 1975.
7. Ильин В. А., Позняк Э. Г. Основы математического анализа. — М.: Наука, 1971, ч. I, 1973, ч. II.
8. Ильин В. А., Садовничий В. А., Сендов Бл. Х. Математический анализ. — М.: Наука, 1979.
- ✓ 9. Карманов В. Г. Математическое программирование. — М.: Наука, 1975.
10. Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — М.: Наука, 1976.
- ✓ 11. Ляшенко И. Н., Карагодова Е. А., Черникова Н. В., Шор Н. З. Линейное и нелинейное программирование. — Киев: Вища школа, 1975.
12. Моисеев Н. Н. Элементы теории оптимальных систем. — М.: Наука, 1975.
13. Моисеев Н. Н., Иванюков Ю. П., Столярова Е. М. Методы оптимизации. — М.: Наука, 1978.
14. Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф. Математическая теория оптимальных процессов. — М.: Наука, 1976.
15. Пшеничный Б. Н., Данилин Ю. М. Численные методы в экстремальных задачах. — М.: Наука, 1975.
- ✓ 16. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1979.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

17. Аграчев А. А., Гамкрелидзе Р. В. Принцип оптимальности второго порядка для задачи быстрого действия. — Матем. сборник, 1976, 100 (142), № 4, с. 610—643.
18. Адаменко Г. М., Габасов Р., Единолич А. А. Минимизация функций конечного числа переменных с помощью некоторых методов второго порядка. — Ж. вычислит. матем. и матем. физики, 1971, 11, № 5, с. 1139—1149.
19. Андреев Ю. Н. Управление конечномерными линейными объектами. — М.: Наука, 1976.

20. Анисимов И. В., Бодров В. И., Покровский В. Б. Математическое моделирование и оптимизация ректификационных установок. — М.: Химия, 1975.
21. Антипин А. С. Методы нелинейного программирования, основанные на прямой и двойственной модификации функции Лагранжа. — М.: Всесоюзный научно-исследовательский институт системных исследований, 1979. — Препринт.
22. Априон Р. Теория второй вариации и ее приложения в оптимальном управлении. — М.: Наука, 1979.
23. Аоки М. Введение в методы оптимизации. — М.: Наука, 1977.
24. Астафьев Н. Н. Двойственная регуляризация задачи линейного программирования, заданной последовательностью реализаций. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 5, с. 1129—1138.
25. Атанс М., Фалб П. Оптимальное управление. — М.: Машиностроение, 1968.
26. Афанасьев А. Ю., Новиков В. А. О поиске минимума функций с ограниченной третьей производной. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 4, с. 1031—1034.
27. Ахиев С. С. О необходимых условиях оптимальности для систем функционально-дифференциальных уравнений. — ДАН СССР, 1979, 247, № 1, с. 11—14.
28. Ахмедов К. Т., Ахиев С. С. Необходимое условие оптимальности для некоторых задач теории оптимального управления. — ДАН Азербайдж. ССР, 1972, 28, № 5, с. 12—15.
29. Ашманов С. А. Математические модели в экономике. Специальный курс лекций. — М.: Изд-во Московск. ун-та, 1976, ч. I; 1978, ч. II.
30. Батищев Д. И. Поисковые методы оптимального проектирования. — М.: Сов. радио, 1975.
31. Бегалишвили Г. М. Оптимальные процессы при слабодифференцируемых связях. — Дифференц. уравнения, 1976, 12, № 4, с. 626—636.
32. Бейко И. В., Бейко М. Ф. Выпуклая аппроксимация управляемого процесса и метод построения обобщенных оптимальных режимов. — Украинский матем. журн., 1973, 25, № 2, с. 343—346.
33. Бейко И. В., Зинько П. Н. Некоторые вероятностные оценки для алгоритмов стохастического программирования с постоянным шаговым множителем. — Кибернетика, 1978, № 2, с. 91—95.
34. Бейко И. В., Шпортюк З. М. О решении одного класса задач уклонения. — Украинский матем. журн., 1972, 24, № 5, с. 683—687.
35. Беленький В. З., Волконский В. А. и др. Итеративные методы в теории игр и программировании. — М.: Наука, 1974.
36. Беллман Р. Процессы регулирования с адаптацией. — М.: Наука, 1964.
37. Белолипецкий А. А. Численный метод решения линейной задачи оптимального быстрого действия сведением ее к задаче Коши. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 6, с. 1380—1386.
38. Белоусов Е. Г. Введение в выпуклый анализ и целочисленное программирование. — М.: Изд-во Московского ун-та, 1977.
39. Березин И. С., Жидков Н. П. Методы вычислений, т. I. — М.: Наука, 1966.
40. Березин И. С., Жидков Н. П. Методы вычислений, т. II. — М.: Физматгиз, 1962.
41. Береснев В. Л., Гимади Э. Х., Дементьев В. Т. Экстремальные задачи стандартизации. — Новосибирск: Наука, 1978.
42. Благодатских В. И. Линейная теория оптимального управления. — М.: Изд-во Московск. ун-та, 1978.
43. Благодатских В. И. Теория дифференциальных включений, ч. I. — М.: Изд-во Московск. ун-та, 1979.

44. Богомолов Н. А., Карманов В. Г. О методе вычисления стационарных точек общей задачи нелинейного программирования. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 1, с. 72—78.
45. Богомолов Н. А., Карманов В. Г. О сходимости метода случайного поиска для отыскания стационарных точек общей задачи нелинейного программирования. — Вестник Московск. ун-та. Сер. вычислит. матем. и кибернетика, 1977, № 1, с. 20—26.
46. Болтянский В. Г. Математические методы оптимального управления. — М.: Наука, 1969.
47. Болтянский В. Г. Оптимальное управление дискретными системами. — М.: Наука, 1973.
48. Болтянский В. Г. Метод шатров в теории экстремальных задач. — УМН, 1975, 30, вып. 3, с. 3—55.
49. Брайсон А., Хо Ю.-Ш.и. Прикладная теория оптимального управления. — М.: Мир, 1972.
50. Булавский В. А., Звягина Р. А., Яковлева М. А. Численные методы линейного программирования. — М.: Наука, 1977.
51. Буланый А. П., Данилин Ю. М. Квазиньютоновские алгоритмы минимизации, основанные на построении систем сопряженных векторов. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 4, с. 877—885.
52. Булатов В. П. Методы погружения в задачах оптимизации. — Новосибирск: Наука, 1977.
53. Вазан М. Стохастическая аппроксимация. — М.: Мир, 1972.
54. Варга Дж. Оптимальное управление дифференциальными и функциональными уравнениями. — М.: Наука, 1977.
55. Васильев О. В. Методы оптимизации в конечномерных пространствах. — Иркутск: Изд-во Иркутского ун-та, 1979.
56. Васильев Ф. П. Лекции по методам решения экстремальных задач. — М.: Изд-во Москов. ун-та, 1974.
57. Васильев Ф. П. О методе нагруженных функционалов. — Вестник Московск. ун-та. Сер. вычислит. матем. и киберн., 1978, № 3, с. 24—32.
58. Васильев Ф. П. О методе касательных. — Ж. вычислит. матем. и матем. физики, 1978, № 18, № 4, с. 1060—1061.
59. Васильев Ф. П. О регуляризации некорректных задач минимизации на множествах, заданных приближенно. — Ж. вычислит. матем. и матем. физики, 1980, 20, № 1, с. 38—50.
60. Вентцель Е. С. Элементы динамического программирования. М.: Наука, 1964.
61. Вилков А. В., Жидков Н. П., Щедрин Б. М. Метод отыскания глобального минимума функции одного переменного. — Ж. вычислит. матем. и матем. физики, 1975, 15, № 4, с. 1040—1042.
62. Владимиров А. А., Нестеров Ю. Е., Чеканов Ю. Н. О равномерно выпуклых функционалах. — Вестник Московск. ун-та. Сер. вычислит. матем. и кибернетика, 1978, № 3, с. 12—23.
63. Воеводин В. В. Линейная алгебра. — М.: Наука, 1974.
64. Воеводин В. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.
65. Вопросы теории и элементы программного обеспечения минимаксных задач: Сб. работ/под ред. Демьянова В. Ф., Малоземова В. Н. — Л.: Изд-во Ленинградск. ун-та, 1977.
66. Воробьев Л. М., Воробьева Т. М. Нелинейные преобразования в прикладных вариационных задачах. — М.: Энергия, 1972.
67. Воробьев Н. Н. Числа Фибоначчи. — М.: Наука, 1978.
68. Габасов Р., Кириллова Ф. Качественная теория оптимальных процессов. — М.: Наука, 1971.
69. Габасов Р., Кириллова Ф. М. Особые оптимальные управления. — М.: Наука, 1973.

70. Габасов Р., Кириллова Ф. М. Принцип максимума в теории оптимального управления. — Минск: Наука и техника, 1974.
71. Габасов Р., Кириллова Ф. М. Основы динамического программирования. — Минск: Изд-во Белорус. ун-та, 1975.
72. Габасов Р., Кириллова Ф. М. Методы линейного программирования. — Минск: Изд-во Белорусского ун-та, 1977, ч. 1; 1978, ч. 2; 1980, ч. 3.
73. Галиуллин А. С., Мухаметзянов И. А., Мухарлямов Р. Г., Фурасов В. Д. Построение систем программного движения. — М.: Наука, 1971.
74. Гамкрелидзе Р. В. Основы оптимального управления. — Тбилиси: Изд-во Тбилисского ун-та, 1977.
75. Гантмахер Ф. Р. Теория матриц. — М.: Наука, 1967.
76. Ганшин Г. С. Вычисление наибольшего значения функций. — Ж. вычислит. матем. и матем. физики, 1976, 16, № 1, с. 30—39.
77. Ганшин Г. С. Простейший последовательный алгоритм поиска наибольшего значения дважды дифференцируемой функции. — Матем. заметки, 1976, 19, № 6, с. 871—872.
78. Ганшин Г. С. Оптимальные пассивные алгоритмы вычисления наибольшего значения функций на отрезке. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 3, с. 562—571.
79. Герашенко Е. И., Герашенко С. М. Метод разделения движений и оптимизация нелинейных систем. — М.: Наука, 1975.
80. Гельфанд И. М., Фомин С. В. Вариационное исчисление. — М.: Физматгиз, 1961.
81. Гермейер Ю. Б. Игры с противоположными интересами. — М.: Наука, 1976.
82. Гермейер Ю. Б., Морозов В. В., Сухарев А. Г., Федоров В. В. Задачи по исследованию операций. — М.: Изд-во Московск. ун-та, 1979.
83. Гнеденко Б. В. Математика — народному хозяйству. — Новое в жизни, науке, технике. Сер. математика, кибернетика. М.: Знание, 1977, № 10.
84. Голиков А. И., Евтушенко Ю. Г. Об одном классе методов решения задач нелинейного программирования. — ДАН СССР, 1978, 239, № 3, с. 519—522.
85. Гольштейн Е. Г., Третьяков Н. В. Градиентный метод минимизации и алгоритмы выпуклого программирования, связанные с модифицированными функциями Лагранжа. Экономика и матем. методы, 1975, 9, № 4, 730—742.
86. Грачев И. И., Евтушенко Ю. Г. Пакет программ для решения задач оптимального управления. — М.: ВЦ АН СССР, 1978.
87. Григоренко Н. Л., Никольский М. С. Некоторые вычислительные аспекты метода убегания Л. С. Понтрягина и Е. Ф. Мищенко. — В сб.: Численные методы нелинейного программирования. Харьков, 1976, с. 143—147.
88. Гродзовский Г. Л., Иванов Ю. Н., Токарев В. В. Механика космического полета с малой тягой. — М.: Наука, 1966.
89. Гупал А. М. Стохастические методы решения негладких экстремальных задач. — Киев: Наукова думка, 1979.
90. Гурин Л. С., Дымарский Я. С., Меркулов А. Д. Задачи и методы оптимального распределения ресурсов. — М.: Советское радио, 1968.
91. Гурман В. И. Выврожденные задачи оптимального управления. — М.: Наука, 1977.
92. Гусятников П. Б., Никольский М. С. Об оптимальности времени преследования. — ДАН СССР, 1969, 184, № 3, с. 518—521.
93. Давыдов Э. Г. Методы и модели теории антагонистических игр. — М.: Изд-во Московск. ун-та, 1978.

94. Дамбраускас А. П. Симплексный поиск. — М.: Энергия, 1979.
95. Данилин Ю. М. Оценка эффективности одного алгоритма отыскания абсолютного минимума. — Ж. вычислит. матем. и матем. физики, 1971, 11, № 4, с. 1026—1031.
96. Даффин Р., Питерсон Э., Зенер К. Геометрическое программирование. — М.: Мир, 1972.
97. Демьянов В. Ф. Минимакс: дифференцируемость по направлениям. — Л.: Изд-во Ленинградск. ун-та, 1974.
98. Демьянов В. Ф. Многошаговый метод обобщенного градиентного спуска. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 5, с. 1112—1118.
99. Демьянов В. Ф., Малоземов В. Н. Введение в минимакс. — М.: Наука, 1972.
100. Денисов Д. В. Метод покоординатного спуска в задачах условной минимизации. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 4, с. 1034—1036.
101. Денисов Д. В. О методе случайного поиска в задачах условной минимизации. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 5, с. 1103—1111.
102. Дубовицкий А. Я., Милютин А. А. Задачи на экстремум при наличии ограничений. — Ж. вычислит. матем. и матем. физики, 1965, 5, № 3, с. 395—453.
103. Дубовицкий А. Я., Милютин А. А. Необходимые условия слабого экстремума в общей задаче оптимального управления. — М.: Наука, 1971.
104. Евтушенко Ю. Г. Численный метод поиска глобального экстремума функций (перебор на неравномерной сетке). — Ж. вычислит. матем. и матем. физики, 1971, 11, № 6, с. 1390—1373.
105. Евтушенко Ю. Г. Численные методы решения задач нелинейного программирования. — Ж. вычислит. матем. и матем. физики, 1976, 16, № 2, с. 307—324.
106. Еремин И. И. Стандартные итерационные процессы негладкой оптимизации для нестационарных задач выпуклого программирования. I, II. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 6, с. 1430—1442; 19, № 1, с. 112—120.
107. Еремин И. И., Астафьев Н. Н. Введение в теорию линейного и выпуклого программирования. — М.: Наука, 1976.
108. Еремин И. И., Мазуров В. Д. Нестационарные процессы математического программирования. — М.: Наука, 1979.
109. Ермольев Ю. М. Методы стохастического программирования. — М.: Наука, 1976.
110. Жданов В. А. О методе покоординатного спуска. — Математ. заметки, 1977, 22, вып. 1, с. 137—142.
111. Жилинскас А. Г. Одношаговый байесовский метод поиска экстремума функций одной переменной. — Кибернетика, 1975, № 1, с. 139—144.
112. Заботин Я. И., Кораблев А. И., Хабибуллин Р. Ф. О минимизации квазивыпуклых функционалов. — Известия вузов. Сер. матем., 1972, № 10 (125), с. 27—33.
113. Заботин Я. И., Кораблев А. И., Хабибуллин Р. Ф. Условия экстремума функционала при наличии ограничений. — Кибернетика, 1973, № 6, с. 65—70.
114. Заботин Я. И., Кораблев А. И. Псевдовыпуклые функционалы и их экстремальные свойства. — Известия вузов. Сер. матем., 1974, № 4(143), с. 27—31.
115. Заботин Я. И. Минимаксный метод решения задач математического программирования. — Известия вузов. Сер. матем., 1975, № 6(157), с. 36—43.

116. Заботин Я. И. Контрпримеры к алгоритмам Зойтендейка в методах возможных направлений. — Известия вузов. Сер. матем., 1976, № 11(174), с. 32—37.
117. Заботин Я. И., Крейнин М. И. К сходимости методов отыскания минимакса. — Известия вузов, сер. матем., 1977, № 10 (185), с. 56—64.
118. Зализняк Н. Ф., Лигун А. А. Об оптимальных стратегиях поиска глобального максимума функции. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 2, с. 314—321.
119. Зангвилл У. И. Нелинейное программирование. — М.: Советское радио, 1973.
120. Зеликин М. И., Тынянский Н. Т. Детерминированные дифференциальные игры. — Успехи матем. наук, 1965, 20, № 4(124), с. 151—157.
121. Зойтендейк Г. Методы возможных направлений. — М.: ИЛ, 1963.
122. Иванюлов Ю. П., Лотов А. В. Математические модели в экономике. — М.: Наука, 1979.
123. Иванов В. В. Об оптимальных алгоритмах минимизации функций некоторых классов. — Кибернетика, 1972, № 4, с. 81—94.
124. Иванов В. В., Березовский А. И., Задирака В. К., Здоренко Л. Д., Легеха Н. П. Методы алгоритмизации непрерывных производственных процессов. — М.: Наука, 1975.
125. Иванов В. В., Людвиченко В. А. Об одном методе последовательной безусловной минимизации решения задач математического программирования. — Кибернетика, 1977, № 2, с. 1—8.
126. Ивашкин В. В. Оптимизация комических маневров. — М.: Наука, 1975.
127. Ижуткин В. С. Мультиплексный метод для задачи выпуклого программирования. — Кибернетика, 1973, № 3, с. 123—127.
128. Ильин В. А., Позняк Э. Г. Линейная алгебра. «Наука», М, 1974.
129. Калиткин Н. Н. Численные методы. — М.: Наука, 1978.
130. Калихман И. Л. Сборник задач по математическому программированию. — М.: Высшая школа, 1975.
131. Канторович Л. В., Горстко А. Б. Оптимальные решения в экономике. — М.: Наука, 1972.
132. Карлин С. Математические методы в теории игр, программирования и экономике. — М.: Мир, 1964.
133. Карманов В. Г. О сходимости метода случайного поиска в выпуклых задачах минимизации. — Теория вероятностей и ее применения, 1974, 19, № 4, с. 817—824.
134. Карманов В. Г. Оценки сходимости итерационных методов минимизации. — Ж. вычислит. матем. и матем. физики, 1974, 14, № 1, с. 3—14.
135. Карманов В. Г. Об одном подходе к исследованию сходимости релаксационных процессов минимизации. — Ж. вычислит. матем. и матем. физики, 1974, 14, № 6, с. 1581—1585.
136. Карташев А. П., Рождественский Б. Л. Обыкновенные дифференциальные уравнения и основы вариационного исчисления. — М.: Наука, 1976.
137. Катковник В. Я. Линейные оценки и стохастические задачи оптимизации. — М.: Наука, 1976.
138. Кирич Н. Е. Методы последовательных оценок в задачах оптимизации управляемых систем. — Л.: Изд-во Ленинградск. ун-та, 1975.
139. Кириченко Н. Ф. Введение в теорию стабилизации движения. — Киев: Вища школа, 1978.
140. Ковач М. О регуляризации некорректных экстремальных задач с использованием метода барьерных функций. — В сб.: Численный анализ / Труды НИВЦ МГУ и БУВЦ. Будапешт, 1978, с. 62—78.
141. Ковач М. Непрерывный аналог итеративной регуляризации градиентного типа. — Вестник Московск. ун-та. Сер. вычислит. матем. и кибернетика, 1979, № 3, с. 36—42.

142. Конвей Р. В., Максвелл В. Л., Миллер Л. В. Теория расписаний. — М.: Наука, 1975.
143. Кононенко А. Ф. О равновесных позиционных стратегиях в неантагонистических дифференциальных играх. — ДАН СССР, 1976, 231, № 2.
144. Константинов М. С. Методы математического программирования в проектировании летательных аппаратов. — М.: Машиностроение, 1975.
145. Коротченко А. Г. Об одном алгоритме поиска наибольшего значения одномерных функций. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 2, с. 563—573.
146. Кошелев В. Н., Морозов С. Ф. Теоремы существования разрывных решений в пространственных вариационных задачах, I, II. — Известия вузов. Сер. матем., 1970, № 5(96), с. 47—52; 1977, № 2(177), с. 49—59.
147. Краснов М. Л., Макаренко Г. И., Киселев А. И. Вариационное исчисление. Задачи и упражнения. — М.: Наука, 1973.
148. Красовский Н. Н. Некоторые задачи теории устойчивости движения. — М.: Физматгиз, 1959.
149. Красовский Н. Н. Теория управления движением. — М.: Наука, 1968.
150. Красовский Н. Н. Игровые задачи о встрече движений. — М.: Наука, 1970.
151. Красовский Н. Н., Субботин А. И. Позиционные дифференциальные игры. — М.: Наука, 1974.
152. Кротов В. Ф., Букреев В. З., Гурман В. И. Новые методы вариационного исчисления в динамике полета. — М.: Машиностроение, 1969.
153. Кротов В. Ф., Гурман В. И. Методы и задачи оптимального управления. — М.: Наука, 1973.
154. Кузнецов Ю. Н., Кузубов В. И., Волощенко А. Б. Математическое программирование. — М.: Высшая школа, 1980.
155. Кукушкин Н. С., Морозов В. В. Теория неантагонистических игр. — М.: Изд-во Московск. ун-та, 1977.
156. Куржанский А. Б. Управление и наблюдение в условиях неопределенности. — М.: Наука, 1977.
157. Кюнц Г. П., Кресслер В. Нелинейное программирование. — М.: Советское радио, 1965.
158. Лагунов В. И. Нелинейная дифференциальная игра убегания. — ДАН СССР, 1972, 252, № 3, с. 522—525.
159. Лебедев А. А., Красильщиков М. Н., Малышев В. В. Оптимальное управление движением космических летательных аппаратов. — М.: Машиностроение, 1974.
160. Левитин Е. С., Милотин А. А., Осмоловский Н. П. Условия высших порядков локального минимума в задачах с ограничениями. — УМН, 1978, 28, № 6, с. 85—148.
161. Левитин Е. С., Поляк Б. Т. Методы минимизации при наличии ограничений. — Ж. вычислит. матем. и матем. физики, 1966, 6, № 5, с. 787—823.
162. Летов А. М. Динамика полета и управление. — М.: Наука, 1969.
163. Ли Э. Б., Маркус Л. Основы теории оптимального управления. — М.: Наука, 1972.
164. Лэсдон Л. С. Оптимизация больших систем. — М.: Наука, 1975.
165. Маергойз М. Д. Об одном классе вырожденных нелинейных экстремальных задач и методах их решения. — Кибернетика, 1977, № 4, с. 63—69.
166. Майстровский Г. Д. Градиентный метод для модифицированной функции Лагранжа. — Ж. вычислит. матем. и матем. физики, 1979, 19, № 1, с. 56—69.
167. Математика на службе инженера. Основы теории оптимального управления: Сб. работ. — М.: Знание, 1973.

168. Мезенцев А. В. Об одной дифференциальной игре. — Дифференц. уравнения, 1972, 8, № 10, с. 1767—1771.
169. Методы оптимизации с приложениями к механике космического полета: Сб. работ / Под ред. Дж. Лейтмана. — М.: Наука, 1965.
170. Мищенко Е. Ф. О некоторых игровых задачах преследования и уклонения от встречи. — Автоматика и телемеханика, 1972, № 9, с. 24—30.
171. Моисеев Н. Н. Численные методы в теории оптимальных систем. — М.: Наука, 1971.
172. Мухачева Э. А., Рубинштейн Г. Ш. Математическое программирование. — Новосибирск: Наука, 1977.
173. Небеснов В. И., Плотников В. И., Кузюшин А. Я. Оптимальное управление ВРШ на волнении. — М.: Пищевая промышленность, 1974.
174. Невельсон М. Б., Хасьминский Р. З. Стохастическая аппроксимация и рекуррентное оценивание. — М.: Наука, 1972.
175. Никайдо Х. Выпуклые структуры и математическая экономика. — М.: Мир, 1972.
176. Никольский М. С. Об идеально наблюдаемых системах. — Дифференц. уравнения, 1971, 7, № 4, с. 631—638.
177. Никольский М. С. Об одном способе убегающего. — ДАН СССР, 1974, 214, № 2, с. 287—290.
178. Никольский М. С. О квазилинейной задаче убегающего. — ДАН СССР, 1975, 221, № 3, с. 539—542.
179. Никольский С. М. Курс математического анализа. Т. I, II. — М.: Наука, 1973.
180. Нурминский Е. А. Численные методы решения детерминированных и стохастических минимаксных задач. — Киев: Наукова думка, 1979.
181. Оптимальное управление: Сб. работ. — М.: Знание, 1978.
182. Ортега Д., Рейнболдт В. Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.
183. Охрименко М. Г. Решение некоторых нелинейных задач оптимального управления. — Кибернетика, 1978, № 1, с. 117—118.
184. Петров Н. Н. О корректности задачи на быстроедействие. — Дифференц. уравнения, 1977, 13, № 11, с. 1974—1980.
185. Петров Ю. П. Вариационные методы теории оптимального управления. — Ленинград: Энергия, 1977.
186. Петросян Л. А. Дифференциальные игры преследования. — Л.: Изд-во Ленинградск. ун-та, 1977.
187. Пиявский С. А. Один алгоритм отыскания абсолютного экстремума функции. — Ж. вычислит. матем. и матем. физики, 1972, 12, № 4, с. 888—896.
188. Подиновский В. В., Гаврилов В. М. Оптимизация по последовательно применяемым критериям. — М.: Советское радио, 1975.
189. Полак Э. Численные методы оптимизации. Единый подход. — М.: Мир, 1974.
190. Половинкин Е. С. Об антагонистических играх двух лиц в динамических системах, I, II. — Вестник Московск. ун-та, 1975, № 1, с. 30—36; № 2, с. 15—21.
191. Поляк Б. Т. Методы минимизации при наличии ограничений. — В сб.: Матем. анализ. Сер. «Итоги науки и техники». — М.: ВИНТИ АН СССР, 1974, т. 12, с. 147—198.
192. Поляк Б. Т. Методы решения задач на условный экстремум при наличии случайных помех. — Ж. вычислит. матем. и матем. физики, 1979, 19, № 1, с. 70—78.
193. Понтрягин Л. С. Обыкновенные дифференциальные уравнения. — М.: Наука, 1965.

194. Понтрягин Л. С. О линейных дифференциальных играх, I, II. — ДАН СССР, 1967, 174, № 6, с. 1278—1280; 175, № 4, с. 764—766.
195. Понтрягин Л. С. Линейная дифференциальная игра убегания. — ДАН СССР, 1970, 191, № 2, с. 283—285.
196. Пропой А. И. Элементы теории оптимальных дискретных процессов. — М.: Наука, 1973.
197. Пшеничный Б. Н. Необходимые условия экстремума. — М.: Наука, 1969.
198. Пшеничный Б. Н. Структура дифференциальных игр. — ДАН СССР, 1969, 184, № 2, с. 285—289.
199. Пшеничный Б. Н. Метод минимизации функций без вычисления производных. — Кибернетика, 1973, № 4, с. 127—129.
200. Разумихин Б. С. Физические модели и методы теории равновесия в программировании и экономике. — М.: Наука, 1975.
201. Ракитский Ю. В., Устинов С. М., Черноуцкий И. Г. Численные методы решения жестких систем. — М.: Наука, 1979.
202. Растрингин Л. А. Системы экстремального управления. — М.: Наука, 1974.
203. Раушенбах Б. В., Токарь Е. Н. Управление ориентацией космических аппаратов. — М.: Наука, 1974.
204. Репников А. В. Колебания в оптимальных системах автоматического регулирования. — М.: Машиностроение, 1968.
205. Реттнев Н. С. О непрерывности и разрывности функции Беллмана. — Дифференц. уравнения, 1978, 14, № 12, с. 2178—2184.
206. Ройтенберг Я. Н. Автоматическое управление. — М.: Наука, 1978.
207. Рокафеллар Р. Выпуклый анализ. — М.: Мир, 1973.
208. Романовский И. В. Алгоритмы решения экстремальных задач. — М.: Наука, 1977.
209. Рудик А. П. Ядерные реакторы и принцип максимума Понтрягина. — М.: Атомиздат, 1971.
210. Рыбаков Л. М. Метод последовательного вычисления всех действительных корней уравнения. — В сб.: Математическое просвещение, вып. 6. — М.: Физматгиз, 1961, с. 262—263.
211. Самарский А. А. Теория разностных схем. — М.: Наука, 1977.
212. Сатимов Н. Ю. Об одном способе убегания в дифференциальных играх. — Матем. сборник, 1976, 99 (141), № 3, с. 380—393.
213. Саульев В. К., Самойлова И. И. Приближенные методы безусловной оптимизации функций многих переменных. — В сб. работ ВИНТИ: Матем. анализ. Сер. «Итоги науки и техники» — М.: ВИНТИ АН СССР, 1973, т. 11, с. 91—128.
214. Старосельский Л. А., Шелудько Г. А., Кантор Б. Я. Об одной реализации метода оврагов с адаптацией величины овражного шага по экспоненциальному закону. — Ж. вычислит. матем. и матем. физики, 1968, 8, № 5, с. 1161—1167.
215. Стронгин Р. Г. Численные методы в многоэкстремальных задачах. — М.: Наука, 1978.
216. Сухарев А. Г. Оптимальный поиск экстремума. — М.: Изд-во Московск. ун-та, 1975.
217. Табак Д., Куо Б. Оптимальное управление и математическое программирование. — М.: Наука, 1975.
218. Танаев В. С., Шкурба В. В. Введение в теорию расписаний. — М.: Наука, 1975.
219. Тарасова В. П. Оптимальные стратегии поиска области наибольших значений для некоторого класса функций. — Ж. вычислит. матем. и матем. физики, 1978, 18, № 4, с. 886—896.
220. Тер-Криков А. М. Оптимальное управление и математическая экономика. — М.: Наука, 1977.

221. Троицкий В. А. Оптимальные процессы колебаний механических систем. — Л.: Машиностроение, 1976.
222. Уайлд Д. Дж. Методы поиска экстремума. — М.: Наука, 1967.
223. Ульм С. Методы декомпозиции для решения задач оптимизации. — Таллин: Изд-во «Валгус», 1979.
224. Успенский А. Б., Федоров В. В. Вычислительные аспекты метода наименьших квадратов при анализе и планировании регрессионных экспериментов. — М.: МГУ, 1975.
225. Уткин В. И. Скользящие режимы и их применения в системах с переменной структурой. — М.: Наука, 1974.
226. Федоренко Р. П. Приближенное решение задач оптимального управления. — М.: 1978.
227. Федоров В. В. Принцип максимума для минимаксной задачи управления с фазовыми ограничениями. — Вестник Московск. ун-та. Сер. вычислит. матем. и кибернетика, 1978, № 4, с. 36—46.
228. Федоров В. В. Численные методы максимина. — М.: Наука, 1979.
229. Фиакко А., Мак-Кормик Г. Нелинейное программирование. Методы последовательной безусловной минимизации. — М.: Мир, 1972.
230. Филиппов А. Ф. О некоторых вопросах теории оптимального регулирования. — Вестник Московск. ун-та. Сер. матем., механика, 1959, № 2, с. 25—32.
231. Флеминг У., Ришел Р. Оптимальное управление детерминированными системами. — М.: Мир, 1978.
232. Формальский А. М. Управляемость и устойчивость систем с ограниченными ресурсами. — М.: Наука, 1974.
233. Фурасов В. Д. Устойчивость движения, оценки и стабилизация. — М.: Наука, 1977.
234. 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.
235. Харди Г. Г., Литтлвуд Дж. Е., Поля Г. Неравенства. — М.: ИЛ, 1948.
236. Химмельблау Д. Прикладное нелинейное программирование. М.: Мир, 1975.
237. Хоменюк В. В. Оптимальные системы управления. — М.: Наука, 1977.
238. Хромова Л. Н. Об одном методе минимизации с кубической скоростью сходимости. — Вестник Московск. ун-та. Сер. вычислит. матем. и кибернетика, 1980, № 3.
239. Хрусталева М. М. Необходимые и достаточные условия оптимальности в форме уравнения Беллмана. — ДАН СССР, 1978, 242, № 5, с. 1023—1026.
240. Ху Т. Целочисленное программирование и потоки в сетях. — М.: Мир, 1974.
241. Цирлин А. М., Балакирев В. С., Дудников Е. Г. Вариационные методы оптимизации управляемых объектов. — М.: Энергия, 1976.
242. Чан Хань. Некоторые методы минимизации при наличии ограничений в гильбертовом пространстве. — Ж. вычислит. матем. и матем. физики, 1977, 17, № 6, 1387—1395.
243. Черемных Ю. Н. Качественное исследование оптимальных траекторий динамических моделей экономики. — М.: МГУ, 1975.
244. Черноусько Ф. Л., Баничук Н. В. Вариационные задачи механики и управления. — М.: Наука, 1973.
245. Черноусько Ф. Л., Колмановский В. Б. Вычислительные и приближенные методы оптимального управления. — В сб. ВИНТИ: Математический анализ. М.; 1977, 14, с. 101—167.

246. Черноусько Ф. Л., Колмановский В. Б. Оптимальное управление при случайных возмущениях. — М.: Наука, 1978.
247. Черноусько Ф. Л., Меликян А. А. Игровые задачи управления и поиска. — М.: Наука, 1978.
248. Чикрий А. А. Об одном классе нелинейных игр убегания. — Кибернетика, 1976, № 3, с. 95—99.
249. Численные методы условной оптимизации: Сб. работ / Под ред. Гилла Ф., Мюррэй У. — М.: Мир, 1977.
250. Шепилов М. А. Непрерывные аналоги метода штрафов для задачи выпуклого программирования. — Экономика и матем. методы, 1975, 11, № 1, с. 130—140.
251. Шепилов М. А. О методе обобщенного градиента для экстремальных задач. — Ж. вычислит. матем. и матем. физики, 1976, 16, № 1, 242—247.
252. Шкадов Л. М., Буханова Р. С., Илларионов В. Ф., Плехин В. П. Механика оптимального пространственного движения летательных аппаратов в атмосфере. — М.: Машиностроение, 1972.
253. Шор Н. З. Методы минимизации недифференцируемых функций и их приложения. — Киев: Наукова думка, 1979.
254. Юдин Д. Б. Задачи и методы стохастического программирования. — М.: Советское радио, 1979.
255. Юдин Д. Б., Гольштейн Е. Г. Линейное программирование. — М.: Физматгиз, 1963.
256. 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.
257. Янг Л. Лекции по вариационному исчислению и теории оптимального управления. — М.: Мир, 1974.
258. Янушевский Р. Т. Управление объектами с запаздыванием. — М.: Наука, 1978.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Антициклон 133

Базис угловой точки 120

Базисные координаты 120

— переменные 120

Верхний предел последовательности 79

— — функции 86

Верхняя грань функции 14

Выпуклая комбинация точек 166

— оболочка 166

Гиперплоскость 158

— опорная 200

— отделяющая 198

Градиент 87

Двойственные переменные 232

Задача быстродействия 423

— двойственная 232

— классического вариационного исчисления 460

— Коши 414

— минимизации второго типа 11, 78

— — первого типа 11, 78

— на безусловный экстремум 90

— на условный экстремум 90

— невырожденная 92, 241

— оптимального управления 422

— — автономная 423

Замыкание множества 162

Заикливание 133

Золотое сечение отрезка 19

Игольчатая вариация 453

Квадратичная форма неотрицательная 174

— — отрицательно определенная 89

— — положительно определенная 89

Константа Липшица 37

— сильной выпуклости 186

Конус 204

— выпуклый, замкнутый, открытый 204

— двойственный (сопряженный) 205

Координата базисная 120

— отмеченная 151

— фазовая 414

Коэффициент барьерный 370

— штрафной 353

Краевая задача принципа максимума 430, 436, 439

Критерий выпуклости функции 46, 51, 52, 170, 172, 173

— оптимальности 49, 171, 197

— сильной выпуклости функции 188, 189

Лексикографически положительный вектор 144

Лексикографическое правило 144

Линейного программирования задача вырожденная 132

— — — каноническая 113

— — — невырожденная 132

— — — общая 108

— — — основная 113

Луч 160

Максимизирующая последовательность 14

Метод барьерных функций 369

— блуждающих трубок 483

— возможных направлений 296

— градиентный 260

— деления отрезка пополам 17, 58

— золотого сечения 19

— касательных 53

— классический 15, 86

— локальных вариаций 484

— ломаных 37

— множителей Лагранжа 343

— нагруженных функций 384

— Ньютона 316

— овражный 268

— оптимальный 23

- Метод парабол 66
 — пассивный 25, 43
 — — оптимальный 28
 — поиска глобального минимума 37, 61, 69, 335
 — по координатного спуска 329
 — последовательный 24, 44, 45
 — — оптимальный 30, 33
 — проекции градиента 273
 — — субградиента 281
 — равномерного перебора 43
 — симметричный 22
 — скорейшего спуска 261
 — с кубической скоростью сходимости 324
 — случайного поиска 399
 — — — без обучения 400
 — — — с обучением 401
 — сопряженных направлений 307
 — стохастической аппроксимации 74, 403
 — стрельбы 456
 — тяжелого шарика 272
 — условного градиента 288
 — Фибоначчи 30
 — штрафных функций 350
 Минимальный корень уравнения 387
 Минимизирующая последовательность 11
 Множество аффинное 159
 — выпуклое 100, 158
 — замкнутое 79
 — компактное 79
 — Лебега 81
 — ограниченное 79
 — регулярное 222
 Множитель Лагранжа 91, 219
 Модуль выпуклости 213
 — — точный 213
 Момент времени конечный 417
 — — — закрепленный 421
 — — — начальный 417
 — — — закрепленный 421
 Надграфик (эпиграф) функции 178
 Наибольшее (максимальное) значение функции 14
 Наименьшее (минимальное) значение функции 9
 Направление внутреннее 244
 — возможное 179
 — — убывания 296
 — касательное 245
 — рецессивное 184
 — убывания 244
 Неравенство Гельдера 360
 Неравенство Гронуолла 450
 — Коши — Буняковского 100
 Нижний предел последовательности 79
 — — функции 86
 Нижняя грань функции 10
 Нормальный вектор гиперплоскости 159
 Оболочка аффинная 160
 — выпуклая 166
 Ограничения интегральные 424
 — корректные 363
 — регулярные 222
 — согласованные 358
 — типа неравенств 94
 — — равенств 91
 — точечные 423
 — фазовые 420
 Опорный вектор 200
 Ортант неотрицательный 161
 Параллелепипед 161
 Полупространство замкнутое 159
 — открытое 159
 Поляра 207
 Принцип максимума 426
 Проблема синтеза 470, 486
 Программирование выпуклое 218
 — динамическое 464
 — линейное 109
 — стохастическое 404
 Проекция точки на множество 193
 Произведение множества на число 162
 Производная по направлению 179
 Прямая линия 160
 Прямое произведение множеств 203
 Размерность множества 160
 Разность множеств 162
 Разрешающий элемент 127
 Расстояние от точки до множества 11, 193
 Решение (траектория) задачи Коши 417
 Симплекс 167
 Симплекс-метод 121, 122
 Скользящий режим 499
 Сопряженная система 426
 Субградиент 207
 Субдифференциал 207
 Сумма множеств 161
 Схема Беллмана 464
 — Моисеева 479

- Сходимость последовательности к множеству 11
- Теорема Вейерштрасса 12, 78
 — Дубовицкого — Милютина 206
 — Куна — Таккера 218
 — Фаркаша 224
- Точка глобального (абсолютного) максимума 14
 — — минимума 9, 13
 — локального максимума 14
 — — минимума 12
 — множества внешняя 162
 — — внутренняя 162
 — — граничная 162
 — — изолированная 162
 — — относительно внутренняя 165
 — — угловая 118
 — — вырожденная 120
 — — невырожденная 120
 —, подозрительная на экстремум 15, 88
 — седловая 219
 — стационарная 88
 — строго локального максимума 14
 — — минимума 12
 — экстремума 15
- Точность метода гарантированная 24
 — — наилучшая 24
- Траектории левый конец 417
 — — закрепленный, подвижный, свободный 421
 — правый конец 417
 — — закрепленный, подвижный, свободный 321
- Траектория (решение) задачи Коши 417
 — оптимальная 422
- Управление 414
 — оптимальное 422
 — — особое 428
- Уравнение Беллмана 466, 487
 — Эйлера 461
- Условие Вейерштрасса 462
 — достаточное оптимальности (максимума, минимума) 16, 89, 93, 171, 180, 197, 220, 474, 496
 — Лежандра 462
 — необходимое оптимальности (максимума, минимума, экстремума) 15, 88, 171, 180, 197, 222, 228, 230, 241, 255, 433
 — Слейтера 222
 — трансверсальности 433, 463
 — Эрдмана — Вейерштрасса 463
- Фазовые координаты 414
 — ограничения 420
- Формула конечных приращений 99
- Функция барьерная 369
 — Беллмана 466, 487
 — Вейерштрасса 462
 — вогнутая 46, 169
 — выпуклая 46, 169
 — Гамильтона — Понтрягина 425
 — дважды дифференцируемая 88
 — — непрерывно дифференцируемая (дважды гладкая) 98
 — дифференцируемая 87
 — Кротова 474, 497
 — кусочно гладкая 414
 — кусочно непрерывная 414
 — Лагранжа 91, 219
 — — модифицированная 345
 — Ляпунова 272, 503
 — непрерывно дифференцируемая (гладкая) 98
 — овражная 267
 — ограниченная 13
 — — сверху 13
 — — снизу 10
 — полунепрерывная сверху 80
 — — снизу 80
 — равномерно выпуклая 212
 — сильно выпуклая 186
 — синтезирующая 470, 486
 — строго вогнутая 169
 — — выпуклая 169
 — — равномерная выпуклая 213
 — — унимодальная 13
 —, удовлетворяющая условию Липшица 37
 — унимодальная 13
 — штрафная 351, 353
- Шар 158
- Шкала состояний 479
- Элементарная операция 479

Федор Павлович Васильев

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ ЭКСТРЕМАЛЬНЫХ ЗАДАЧ

М., 1980 г., 520 стр. с илл.

Редакторы *Н. Л. Григоренко, Т. И. Кузнецова*

Техн. редактор *Н. В. Кошелева*

Корректор *Н. Д. Дорохова*

ИБ № 2289

Сдано в набор 20.04.80. Подписано к печати 25.09.80 Т-17808.
Бумага 60×90^{1/8}, тип. № 3. Литературная гарнитура. Высокая
печать. Условн. печ. л. 32,5. Уч.-изд. л. 38,48. Тираж 20 000 экз.
Заказ № 607. Цена книги 1 р. 50 к.

Издательство «Наука»

Главная редакция физико-математической литературы

117071, Москва, В-71, Ленинский проспект. 15

Ленинградская типография № 2 головное предприятие ордена
Трудового Красного Знамени Ленинградского объединения
«Техническая книга» им. Евгении Соколовой Союзполиграфпрома
при Государственном комитете СССР по делам издательств,
полиграфии и книжной торговли. 198052. г. Ленинград, Л-52.
Измайловский проспект, 29.

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
117071, Москва, В-71, Ленинский проспект, 15

*ГОТОВИТСЯ К ПЕЧАТИ
В 1981 ГОДУ*

Васильев Ф. П. Методы решения экстремальных задач: Учебное пособие. — М.: Наука. Главная редакция физико-математической литературы.

В книге излагаются методы минимизации функций (функционалов), заданных на множествах из функциональных пространств, рассматриваются приложения к задачам оптимального управления процессами, описываемыми системами обыкновенных дифференциальных уравнений и уравнениями с частными производными. Значительное место в книге занимают методы решения некорректных экстремальных задач, условия аппроксимации экстремальных задач.

Может рассматриваться как продолжение настоящей книги.

Книга предназначена студентам старших курсов и аспирантам университетов и других вузов по специальности «Прикладная математика».

Предварительные заказы на эту книгу принимаются без ограничения магазинами Книготорга и Академкниги.