

И * Л

Издательство
иностранно
литературы

*

MODERN MATHEMATICS FOR THE ENGINEER

ROYAL WELLER
SOLOMON LEFSCHETZ
RICHARD BELLMAN
JOHN W. GREEN
MAGNUS R. HESTENES
RICHARD COURANT
MENAHEM M. SCHIFFER
IVAN S. SOKOLNIKOFF
NORBERT WIENER
H. FREDERIC BOHNENBLUST
GEORGE W. BROWN
LOUIS A. PIPES
EDWIN F. BECKENBACH
CHARLES B. MORREY, JR.
GEORGE E. FORSYTHE
CHARLES B. TOMPKINS
DERRICK H. LEHMER

Edited by
EDWIN F. BECKENBACH
Professor of Mathematics
University of California
Los Angeles

1956
McGraw-Hill Book Company
NEW YORK TORONTO LONDON

СОВРЕМЕННАЯ МАТЕМАТИКА ДЛЯ ИНЖЕНЕРОВ

Под редакцией
Э. Ф. БЕККЕНБАХА

Перевод с английского
под общей редакцией
И. Н. ВЕКУА

**ИЗДАТЕЛЬСТВО
ИНОСТРАННОЙ ЛИТЕРАТУРЫ**
Москва, 1958

АННОТАЦИЯ

Книга „Современная математика для инженеров“ содержит ряд статей по различным вопросам математики, важным с точки зрения приложений к механике, физике, технике. В составлении книги участвовали многие крупные американские математики. Наряду с классическими разделами в ней рассмотрены также и новые области математики, представляющие живой интерес для инженерной практики. В частности, в книге значительное место уделено вероятностно-статистической и вычислительной проблематике (теория предсказания, теория игр, теория динамического планирования и др.).

Основная цель книги — расширить математический кругозор инженера. Книга будет полезна инженерам-теоретикам разных специальностей, студентам старших курсов и аспирантам технических вузов, преподавателям математики в технических вузах. Изложение ряда вопросов представляет интерес и для научных работников-математиков.

Редакция литературы по вопросам математических наук

ПРЕДИСЛОВИЕ

РЕДАКТОРА РУССКОГО ПЕРЕВОДА

Бурный рост техники, свидетелями которого мы являемся, требует от инженеров все более глубокой математической подготовки. Инженер в наши дни должен не только уметь практически решать задачи, укладывающиеся в традиционные рамки, но также формулировать и решать совершенно новые задачи, требующие применения новых математических методов.

По этим причинам математический кругозор инженера в современных условиях должен постоянно расширяться. Именно эту цель и преследует настоящая книга.

Написанная при участии многих крупных американских математиков, она выгодно отличается от известных книг подобного рода как общим замыслом, так и тем, что в ней наряду с классическими разделами рассмотрены также и новые области математики, представляющие живой интерес для инженерной практики. В частности, в книге значительное место уделено вероятностно-статистической и вычислительной проблематике (теория предсказания, теория игр, теория динамического планирования и др.).

Предлагаемую книгу ни в какой степени нельзя считать учебником. В ней дан обзор основных идей и методов тех направлений современной математики, которые теснее других связаны с прикладными областями. Конечно, в ней не могли быть охвачены все интересные направления, но все же многие отрасли математики освещены в достаточной мере широко. В подборе материала и характере изложения существенно отражены личные научные интересы и точки зрения авторов, и это обстоятельство придает книге дополнительный интерес.

Следует сказать, что не все главы книги написаны с одинаковой обстоятельностью. Авторам лучше удались классические разделы, которые составляют первую часть книги. Во второй и третьей частях, которые посвящены сравнительно новым областям математики, наряду, например, с главой о динамическом планировании, отличающейся большой методичностью и полнотой, имеются главы, написанные весьма бегло и схематично.

Для чтения книги нужно иметь подготовку по высшей математике по крайней мере в объеме программы втузов, однако не

все разделы читаются одинаково легко. Отдельные главы могут читаться независимо друг от друга.

В русское издание не включены две главы: „Функциональные преобразования в технических расчетах“, написанная Дж. Бэрнсом, и „Прикладная математика и исследование операций“, написанная Д. Кингом. При этом редакция учитывала, что первая из указанных глав в основном представляет собой обзор книги М. Гарднера и Дж. Бэрнса „Переходные процессы в линейных системах“ (Гостехиздат, 1949), известной советским читателям, а также, что по материалу этой главы имеется много доступной литературы на русском языке. Вторая же глава, на наш взгляд, написана столь бегло, что читателю трудно будет составить по ней впечатление о новой и важной отрасли науки, которой она посвящена. Интересующихся этой отраслью мы можем отослать к книге Ф. Морса и Дж. Е. Кимбела „Методы исследования операций“, недавно вышедшей в русском переводе (Советское радио, 1956).

При редактировании перевода существенную помощь оказали В. А. Диткин (прочитавший главы, написанные Л. Пайпсом и Ч. Томпкинсом), А. М. Яглом (по главе, написанной Н. Винером), и В. С. Владимиров (по главе, написанной Дж. Брауном). В редактировании всего текста книги участвовал Б. В. Боярский. Всем этим лицам я приношу искреннюю благодарность.

И. Н. Векуа.

К ЧИТАТЕЛЮ

Современная техника основывается как на научных экспериментальных и теоретических исследованиях, так и на опыте инженерной практики. Промежуток времени между научным открытием и техническим использованием в последние годы чрезвычайно сократился, поэтому инженеру больше чем когда-либо раньше необходимо быть в курсе значительных открытий в областях математики и физики.

Преследуя эту цель, технические отделения Калифорнийского университета совместно с научными отделениями организовали серию лекционных курсов по современной физике, математике и химии для сотрудников, студентов и аспирантов с тем, чтобы ознакомить их с научными достижениями. Каждый курс состоит из ряда лекций, составленных профессорами Калифорнийского и других университетов.

Нам доставляет удовольствие поделиться с читателем на страницах этой книги ценным опытом, собранным в серии лекций „Современная математика для инженеров“.

БОЛДУИН М. ВУДС,
Л. М. К. БЁЛТЕР,
МОРРОУ П. О'БРИЕН,

*Калифорнийский университет,
Лос-Анжелес, Беркли.*

ПРЕДИСЛОВИЕ

В то время как в Калифорнийском университете в Лос-Анжелесе проводился общедоступный курс лекций по современной физике для инженеров, имевший большой успех, была начата работа над соответствующим курсом лекций по математике под общим руководством Л. М. К. Бёлтера, декана Инженерного колледжа, и Клиффорда Белла, профессора математического факультета.

Организационный Комитет был составлен из сотрудников Калифорнийского инженерного колледжа и математического и физического факультетов в Лос-Анжелесе, Инженерного колледжа и математического факультета в Беркли, Лаборатории „Согопа“ Национального бюро стандартов, Испытательного центра морских ракет в Пойнт-Мугу, Калифорния, Испытательной станции военно-морского флота в Чайна-Лейк, Международной телеметрической корпорации и, наконец, авиационных компаний Валти, Дуглас, Хьюз и Локхид.

Комитет поставил своей целью организовать курс лекций, которые познакомили бы инженеров и специалистов, занятых исследованиями, проектированием и административной работой, с быстрым развитием прикладной математики — развитием, происшедшим в основном вследствие требований современного инженерного планирования и проектирования и сделавшимся возможным благодаря современным достижениям математики и статистики, а также благодаря развитию быстродействующих моделирующих устройств и цифровых вычислительных машин с чрезвычайно большими возможностями. Огромный успех курса в значительной степени обязан тщательности, с которой Комитет выбрал темы и лекторов.

Многие детали, связанные с этим курсом, были умело разрешены К. Беллом и Дж. К. Диллоном. Курс читался не только в Калифорнийском университете в Лос-Анжелесе, но также в Лаборатории „Согопа“ Национального бюро стандартов и под руководством Ч. Б. Моррея в Калифорнийском университете в Беркли. Кроме того, многие лекции были повторены перед многочисленными аудиториями в Сан-Диего и на Испытательной станции военно-морского флота в Чайна-Лейк. Общее число слушателей составило около 1000 человек.

Подобно курсу „Современная физика для инженеров“ (и будущему курсу „Современная химия для инженеров“), компания Мак Гроу-Хилл публикует указанные лекции в виде книги, с одной

стороны, как дополнительный материал, связанный с аналогичными курсами других институтов, с другой стороны — как книгу, имеющую самостоятельный интерес для специалиста-инженера в любой научной или технической области.

Как это обычно бывает в книгах по прикладной математике, каждая из глав затрагивает как математическую формулировку физических проблем (в общем случае рассмотренных, например, в § 5.2), так и конструктивные методы решения. Подобная система изложения принята и в настоящей книге, причем расположение глав соответствует тому, преобладает ли первый или второй из указанных аспектов. Классическая прикладная математика опирается в основном на дифференциальное и интегральное исчисления. В последнее же время (как указано во введении) все чаще встречаются понятия теории случайных процессов, применяемые как при математической постановке, так и при решении прикладных физических и социологических задач. В соответствии с этим центральная часть книги составлена из глав, в которых большое внимание уделяется теоретико-вероятностным задачам.

Книга, таким образом, разделяется на три части. В первой, озаглавленной „Математические модели“, последовательно рассматриваются физические проблемы, выраженные с помощью обыкновенных дифференциальных уравнений, интегральных уравнений и дифференциальных уравнений с частными производными. Во второй части („Теоретико-вероятностные задачи“) излагаются динамическое программирование, операционные задачи техники и использование вероятностных методов. Наконец, в третьей части („Методы вычислений“) основное внимание обращено на численные решения; как и в книге „Современная физика для инженеров“, в заключительных главах рассмотрены быстродействующие вычислительные устройства.

Эдвин Ф. Беккенбах.

ВВЕДЕНИЕ

РОЙЯЛ УЭЛЛЕР

Центр по испытанию управляемых снарядов ВМС

Пойнт-Мугу, Калифорния

Книга под названием „Современная математика для инженеров“ требует некоторых пояснений. Нижеследующие главы основываются на предпосылке, что современная инженерная практика все больше становится наукой и все меньше ремеслом и требует благодаря этому привлечения более совершенного математического аппарата.

Для большинства инженеров математика есть средство для получения результата. Этим результатом обычно является числовой ответ, хотя более тонкое и менее привычное применение математики можно найти при исследовании аналогий, при оценке применимости некоторых гипотез и т. п. Во многих случаях численные ответы, получаемые рядовым инженером, могут быть найдены с помощью простых арифметических формул. В случае успеха это удовлетворяет рядового инженера и не заставляет его искать более сложные математические методы. Однако, особенно в последние годы, все чаще возникают задачи, не поддающиеся решению с помощью столь простых формул. Чем это вызвано?

Один из ответов на этот вопрос характерен для авиационной промышленности, хотя и не ограничивается ею. Требование к эффективности самолетов приводит к проектированию по принципу „равной прочности“. Ни одна деталь не должна быть слишком тяжелой. Ни одна деталь не должна быть недостаточно прочной. Поэтому необходимо очень точно вычислять напряжения. Однако простые арифметические формулы сводят вычисление напряжений к грубым оценкам, кроме случаев очень простых геометрических форм. Становится необходимым расширить математическое исследование конструктивной схемы, чтобы охватить такие факторы, как сопротивление упругим деформациям, флаттер и концентрация напряжений. Необходимость эффективного проектирования заставила инженеров-самолетостроителей освоить более тонкие методы анализа и вычислений.

Второй ответ на поставленный вопрос мы находим в сложности современных технических систем. Примером может служить система автоматического управления огнем зенитной артиллерии. В нее входит

радарная установка, автоматически следящая за самолетом и дающая информацию о его курсе и скорости. Эти данные вводятся в счетно-решающее устройство, которое предсказывает будущее положение самолета и определяет азимут и угол возвышения батареи, необходимые для его поражения. Орудия наводятся и производят стрельбу с помощью сервосистемы, управляемой по данным счетно-решающего устройства.

Подобные системы не могут рассматриваться просто как совокупность элементарных компонент. Их надо рассматривать как серию функциональных устройств, каждое из которых производит некоторое преобразование над своей входной величиной. В результате этого преобразования должна получиться требуемая входная величина для следующего функционального устройства.

При исследовании подобных систем приходится сталкиваться с передаточными функциями, импедансами и устойчивостью по отношению к различным возмущающим функциям. Для проведения необходимых вычислений требуется знакомство с дифференциальными уравнениями, функциональными преобразованиями, матрицами и временными рядами.

Применение математики необходимо также при рассмотрении статистического характера физических явлений. Специалисты по естественным наукам с давних пор пользуются статистическими методами вследствие крайней сложности изучаемых ими систем, а также невозможности учесть действие многих переменных факторов. С другой стороны, в прошлом инженеры занимались главным образом детерминированными проблемами. Однако усложнение систем при развитии техники все чаще приводит к рассмотрению скорее стохастических, нежели детерминированных явлений. В примере упомянутой выше системы автоматического управления огнем счетно-решающее устройство выбирает *наиболее вероятное* будущее положение цели. Если обслуживающий инженер желает эксплуатировать подобную систему, он должен иметь в запасе те части, выход из строя которых *наиболее вероятен*. При определении боеспособности результатом является математическое ожидание уничтожения цели. Все это — характерные термины математической теории вероятностей. Вообще, можно считать, что при экспериментальном исследовании поведения системы необходимо учитывать наличие производственных допусков, различие в поведении материалов, различие мнений экспериментаторов и разнообразие ожидаемых условий работы. Подобные соображения сводят опытные испытания к рассмотрению вероятностей. Недостаточное понимание этого факта инженерами, возможно, объясняется тем, что литература по приложению математики к изучению случайных явлений изобилует примерами из сельского хозяйства и биологии, ничего не говорящими инженеру. Таков ответ на поставленный вопрос.

К сказанному добавим следующее:

После второй мировой войны выяснилось, что многие вопросы, часто решавшиеся интуитивно, а также на основании некоторых предвзятых мнений или „здорового смысла“ — на самом деле поддаются количественному исследованию. В частности, если удастся сопоставить числа возможным результатам различных способов выполнения операции, то иногда возможно выбрать аналитически наилучший способ.

Изучение указанного круга вопросов получило название „анализа операций“. Решение подобных задач требует привлечения различных математических методов, большинство из которых относится к теории вероятностей. В более развитых областях — например, в операционном программировании — могут оказаться необходимыми обширные познания основ математики.

Во многих случаях инженерной практики обнаруживается необходимость дать числовой ответ на задачу, не допускающую точной формулировки. Например, требуется указать напряжение вблизи отверстия произвольной формы, градиент электрического поля между электродами произвольной формы или скорость течения струи в трубе произвольного поперечного сечения. Трудность заключена в слове „произвольный“, которое в математическом смысле означает „неаналитический“. Простые решения для задач подобного рода существуют лишь в тех случаях, когда эти задачи могут быть выражены через разделяющиеся координаты. Например, контур зуба шестерни не выражается через разделяющиеся координаты и точное значение напряжения в нагруженном зубе поэтому не выражается в простой аналитической форме. Подобная трудность может быть преодолена несколькими способами. Можно, например, рассматривать зуб шестерни как консольную балку и пренебречь серьезной неточностью подобного допущения. С другой стороны, можно получить решение с помощью ряда последовательных приближений, которые приведут к ответу с любой наперед заданной точностью. В этой прикладной области используются так называемые „релаксационные“ и „итерационные“ методы и, вообще, замена дифференциальных уравнений конечно-разностными. Подобная методика применима к задачам с произвольными граничными условиями. Особый интерес к приближенным вычислениям связан с необычайным развитием цифровых счетно-решающих устройств в послевоенное время. Подобные устройства в основном представляют собой быстродействующие арифметические машины и обычно работают лишь при условии, что желаемый результат может быть получен путем сложения и вычитания чисел. Они хорошо приспособлены к численным приближениям, вычислению конечных разностей и подобным операциям и решают задачи, чрезмерно утомительные при использовании более медленных способов вычислений.

Просматривая оглавление, можно увидеть много разветвлений тематики, указанной в этом введении. Хорошо представлена, как

с теоретической стороны, так и в части приложений, математика случайных явлений. В ряде глав рассматриваются задачи преобразования и техника их решения. Значительное внимание обращено на численное решение различных типов дифференциальных уравнений.

В заключение необходимо отметить, что в настоящее время обстановка благоприятствует более тесной связи между техникой и математикой. Новые достижения техники требуют расширения математической подготовки; новые достижения математики оказываются хорошо соответствующими требованиям современной техники. Более тесному сближению этих двух мощных дисциплин и посвящена настоящая книга.

Ч А С Т Ь 1

Математические модели

Глава 1

ЛИНЕЙНЫЕ И НЕЛИНЕЙНЫЕ КОЛЕБАНИЯ

СОЛОМОН ЛЕФШЕЦ

Принстонский университет,
Национальный университет в Мексико

§ 1.1. Введение. В окружающем нас мире мы наблюдаем бесконечное разнообразие колебательных явлений: от маятника до приливов, от колебаний струны до бортовой и килевой качки корабля на море. Некоторые колебания оказываются для нас полезными, например, колебания маятника, струны, колебания в электронной лампе. Другие, напротив, являются вредными, например колебания корабля или самолета.

Для успешного применения полезных колебаний и контролирования вредных необходимо построить их теорию, что неизбежно приводит к математической модели. В некоторых случаях математическая теория может быть очень простой, в других — более сложной и глубокой.

Вначале мы рассмотрим некоторые из более простых явлений.

§ 1.2. Гармонические колебания.

Трудно представить себе более простой механизм, чем маятник (рис. 1.1). Если x есть угол отклонения (причем x остается малым) и если l есть длина, а m — масса груза, то основное соотношение:

$$\text{масса} \times \text{ускорение} = \text{сила}$$

приводит к дифференциальному уравнению

$$ml\ddot{x} = -mgx, \quad (1.1)$$

или

$$\ddot{x} + \omega^2 x = 0, \quad \omega^2 = \frac{g}{l}. \quad (1.2)$$

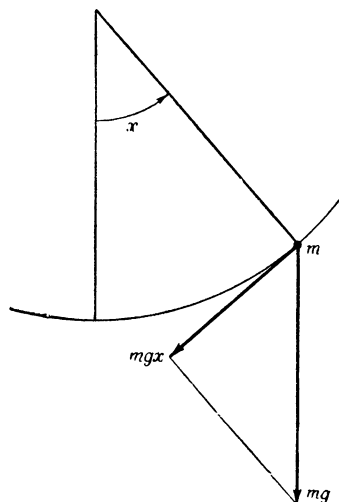


Рис. 1.1. Маятник.

Здесь $\dot{x}$, $\ddot{x}$, ... обозначают первую, вторую, ... производные по времени (скорость, ускорение, ...).

Крутильные колебания с малым углом отклонения будут выражаться таким же уравнением, но ω будет иметь другое значение.

Другим очень простым механизмом является пружина (рис. 1.2).

Пусть масса пружины пренебрежимо мала и пусть m — масса, приводимая в движение пружиной.

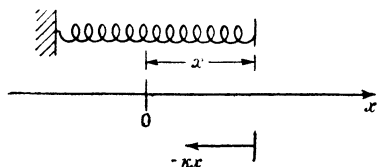


Рис. 1.2. Пружина.

Представим себе эту массу в виде материальной точки и обозначим через x ее (положительное или отрицательное) расстояние, отсчитанное от положения массы при нерастянутой пружине. Мы имеем здесь то же самое основное уравнение (1.1).

Простейшее допущение относительно силы состоит в том, что она пропорциональна $-x$; если масса m находится справа от точки 0, сила действует влево, и наоборот. Таким образом, уравнение (1.1) приводит к соотношению

$$m\ddot{x} = -kx. \quad (1.3)$$

Постоянная k есть так называемая *жесткость пружины*. Мы можем также написать

$$\ddot{x} + \omega^2 x = 0, \quad \omega^2 = \frac{k}{m}. \quad (1.4)$$

Таким образом, мы опять пришли к уравнению маятника.

Общее решение уравнения (1.2) хорошо известно читателю. Мы будем писать его в виде

$$x = a \cos(\omega t + \varphi).$$

Постоянная a есть *амплитуда*, постоянная φ — *фаза*. Обе они произвольны — это так называемые *постоянные интегрирования* — и могут быть выбраны так, чтобы удовлетворялись заданные начальные условия. В противоположность амплитуде и фазе период $T = 2\pi/\omega$ имеет вполне определенное (для данного механизма) значение, что относится и к частоте

$$f = \frac{1}{T} = \frac{\omega}{2\pi}.$$

Для маятника и для пружины имеем соответственно:

$$T = 2\pi \left(\frac{l}{g} \right)^{1/2}; \quad T = 2\pi \left(\frac{m}{k} \right)^{1/2}.$$

Приведем два способа графического представления колебаний, определяемых уравнением (1.4). Первый способ представляет собой построение графика зависимости x от t , несомненно знакомого читателю (см. рис. 1.3).

Второй способ не играет большой роли в данном случае, однако он имеет важное значение в более сложных вопросах, рассмотренных ниже. Сначала изменим масштаб времени так, чтобы ωt заменилось на t . В результате этого уравнение (1.2) заменится более простым уравнением

$$\ddot{x} + x = 0. \quad (1.5)$$

Примем в качестве новой переменной скорость $\dot{x}$ и обозначим ее через y , т. е. положим $\dot{x} = y$. Этим осуществляется замена уравнения (1.5) системой двух уравнений:

$$\dot{x} = y, \quad \dot{y} = -x.$$

Плоскость xu известна под названием *фазовой плоскости*. Рассматриваемый нами способ представления колебаний заключается в по-

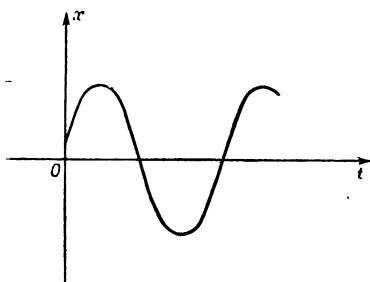


Рис. 1.3. Смещение как функция времени (для пружины).

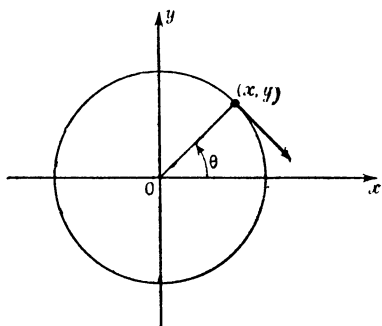


Рис. 1.4. Фазовая плоскость (для пружины).

строении графика зависимости между x и y . В явном виде общее решение уравнения (1.5) есть

$$x = a \cos(t + \varphi), \quad (1.6)$$

что дает

$$y = \dot{x} = -a \sin(t + \varphi). \quad (1.7)$$

При изменении t точка (x, y) , определяемая уравнениями (1.6) и (1.7) описывает кривую или *траекторию* на фазовой плоскости. Эта кривая и представляет собой график, который мы имеем в виду. Нетрудно найти эту кривую; возводя в квадрат уравнения (1.6) и (1.7) и складывая, получаем

$$x^2 + y^2 = a^2,$$

откуда следует, что искомая кривая является окружностью с центром в начале координат и радиусом, равным амплитуде колебаний a .

Если θ — полярный угол (рис. 1.4), то

$$\dot{\theta} = \frac{d}{dt} \arctg \frac{y}{x} = \frac{x\dot{y} - y\dot{x}}{x^2 + y^2} = -1.$$

Следовательно, кривая обходится по часовой стрелке с постоянной угловой скоростью, равной единице.

Уже на этой ранней стадии изложения целесообразно подчеркнуть два важных свойства этих кривых.

а) *Замкнутая кривая соответствует периодическому решению.* Действительно, если точка $[x(t), y(t)]$ возвращается в свое положение через время T , то

$$x(t+T) = x(t); \quad y(t+T) = y(t),$$

т. е. решение $x(t)$ имеет период T . Примеры таких движений будут показаны в дальнейшем. Если замкнутая кривая изолирована, она носит название *предельного цикла*. Таким образом, замкнутая кривая на рис. 1.4 не является предельным циклом, потому что она входит в семейство кривых, непрерывно зависящих от a и, следовательно, не является изолированной.

б) *Кривая, сводящаяся к одной точке, или особая точка — например, начало координат на рис. 1.4 — означает состояние равновесия*, так как для нее имеем

$$\frac{dx}{dt} = \frac{dy}{dt} = 0.$$

Рассматривавшиеся до сих пор колебания известны под названием *гармонических колебаний*, а системы, подверженные им, называются *гармоническими осцилляторами*.

§ 1.3. Затухающие колебания. Допустим теперь, что маятник или пружина испытывают сопротивление или некоторое диссипативное воздействие подобного же рода. Это внесет в правую часть уравнения (1.3) новый член — $\alpha\dot{x}$, так что мы будем иметь

$$m\ddot{x} = -\alpha\dot{x} - kx.$$

Перенося все члены в левую часть, мы приведем это уравнение к следующему общему виду:

$$A\ddot{x} + B\dot{x} + Cx = 0, \quad (1.8)$$

где A, B, C — положительные постоянные. Уравнение подобного вида получается также в связи с совершенно иным явлением — колебаниями в электрическом контуре. Рассмотрим простой контур (рис. 1.5), состоящий из батареи с постоянным напряжением E , катушки с самоиндукцией L , сопротивления R и емкости C . Если через x обозна-

чим силу тока, то будем иметь хорошо известное соотношение:

$$E = Rx + L\dot{x} + \frac{1}{C} \int x dt$$

или после дифференцирования

$$L\ddot{x} = R\dot{x} + \frac{1}{C}x = 0;$$

это уравнение имеет тот же вид, что и (1.8).

Общее решение уравнения (1.8) для малых B , т. е. если сопротивление воздуха или электрическое сопротивление мало, имеет хорошо известный вид

$$x = ae^{-\delta t} \cos(\omega t + \varphi),$$

где

$$\delta = -\frac{B}{2A}, \quad \omega = \frac{(4AC - B^2)^{1/2}}{2A}.$$

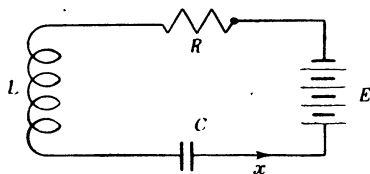


Рис. 1.5. Простой контур, состоящий из батареи с постоянным напряжением E , самоиндукцией L , сопротивлением R и емкостью C .

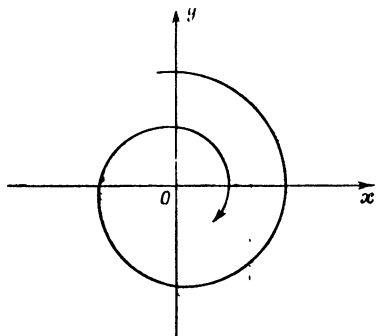


Рис. 1.6. Фазовая плоскость в случае затухающих колебаний.

Можно считать, что это решение описывает колебательное движение с *постоянным периодом* $2\pi/\omega$, но с *переменной амплитудой* $ae^{-\delta t}$.

Если изменить масштаб времени, как указано выше, и положить $\delta = \omega\Delta$, то решение примет вид

$$x = ae^{-\Delta t} \cos(t + \varphi). \quad (1.9)$$

Отсюда

$$y = \dot{x} = -ae^{-\Delta t} [\Delta \cos(t + \varphi) + \sin(t + \varphi)]. \quad (1.10)$$

Нетрудно доказать, что уравнения (1.9) и (1.10) в совокупности представляют кривую, спирально навивающуюся на начало координат. Однако общую картину явления можно получить и из рис. 1.4, если вспомнить, что мы имеем здесь колебания с убывающей амплитудой. Действительно, каждый полный оборот будет напоминать окружность, постепенно сжимающуюся в направлении начала координат, так что график представляет собой кривую, спирально подходящую к началу координат (рис. 1.6).

Аналогично вышеизложенному, нетрудно показать, что спираль обходит по часовой стрелке. Заметим, что если B сравнительно велико, т. е. если велико электрическое или воздушное сопротивление (например, если груз маятника представляет собой большой, но относительно легкий шар), то явление теряет колебательный характер и становится совершенно иным. Но так как мы ограничиваемся колебаниями, то этот вопрос рассматривать здесь не будем.

§ 1.4. Вынужденные колебания. Изучение вынужденных колебаний является более сложной задачей. Мы можем представить себе пружину с намагниченным грузом, на который действует соленоид, однако простейшим примером является электрический контур с приложенной переменной электродвижущей силой. Явление описывается

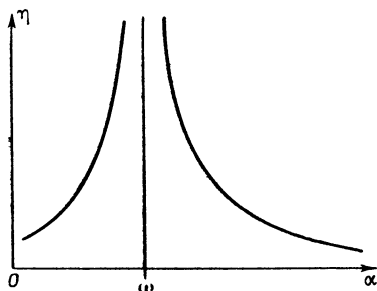


Рис. 1.7. Амплитудно-частотная характеристика при отсутствии затухания.

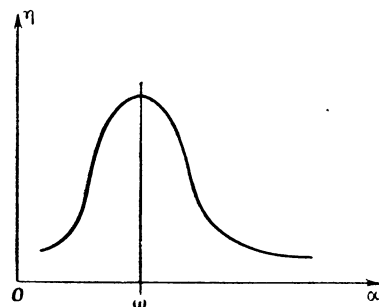


Рис. 1.8. Амплитудно-частотная характеристика при наличии затухания.

двумя типами уравнений; если затухание отсутствует (нет сопротивления), то

$$\ddot{x} + \omega^2 x = E \cos \alpha t, \quad (1.11)$$

а если имеется затухание (есть сопротивление), то

$$A\ddot{x} + B\dot{x} + Cx = E \cos \alpha t. \quad (1.12)$$

Обычно $\alpha \neq \omega$, тогда уравнение (1.11) имеет общее решение:

$$x = [a \cos(\omega t + \varphi)] + \left\{ \frac{E \cos \alpha t}{\omega^2 - \alpha^2} \right\}.$$

Первый член [...] представляет собой „свободные“ колебания, второй член {...} — „вынужденные“ колебания.

Второй член становится очень большим, когда α близко к ω , т. е. когда колебания возмущающей силы имеют период, весьма близкий к периоду свободных колебаний системы. Это — хорошо

известное явление *резонанса*, причинившее многочисленные бедствия. График зависимости амплитуды $|E/(\omega^2 - \alpha^2)|$ от α принимает вид, показанный на рис. 1.7, и наглядно объясняет характер явления.

Общее решение уравнения (1.12) имеет вид

$$x = [ae^{-\delta t} \cos(\omega t + \varphi)] + \left\{ \frac{D \cos(\alpha t + \lambda)}{(C - A\alpha^2)^2 + B^2\alpha^2} \right\},$$

где a и φ произвольны, а D и λ определены.

График, аналогичный рис. 1.7 (амплитудно-частотная характеристика), показан на рис. 1.8. Видно, что резонанс несколько смягчен и притом тем значительнее, чем выше диссипация в системе.

§ 1.5. Линейные и нелинейные системы. Рассмотренные до сих пор уравнения распадались на два типа:

а) *линейные однородные*: $A\ddot{x} + B\dot{x} + Cx = 0$;

б) *линейные неоднородные*: $A\ddot{x} + B\dot{x} + Cx = F(t)$.

Они называются линейными потому, что x , $\dot{x}$, $\ddot{x}$ входят только в первой степени. Соответствующие колебания также называются линейными.

Самым значительным свойством линейных однородных уравнений является то, что их решения подчиняются знаменитому *принципу суперпозиции*, а именно, если x и y — решения, то и $x + y$ является решением. Для типа (б) можно также сказать, что если x есть какое-либо решение неоднородного уравнения, а y *любое* решение соответствующего однородного уравнения (а), то $x + y$ есть решение неоднородного уравнения (б). Это, однако, свойства решений, а не колебаний. Колебания же, по крайней мере определяющиеся уравнением (а), обладают следующими двумя замечательными свойствами:

амплитуда каждого отдельного колебания произвольна;

частоты всех колебаний одинаковы.

Мы увидим, что нелинейные колебания могут обладать одним из этих двух свойств, но не обоими.

Рассмотренные до сих пор линейные системы давали возможность полного описания явлений, в основном, благодаря их крайней простоте. В частности, могли быть получены явные решения в замкнутой форме, что значительно облегчало задачу. Нелинейные системы — это те, в которых x , $\dot{x}$ встречаются в степени выше первой. Они охватывают все системы, кроме линейных, и потому дать исчерпывающую классификацию их представляет собой безнадежную задачу. Мы поэтому выберем для более детального изучения достаточно обширный класс таких систем.

Одной из причин, почему линейные системы встречаются столь часто в физике и технике, является стремление физика и инженера прилагать все усилия для упрощения явления до тех пор, пока рассматриваемая система не станет линейной. Это буквально верно,

например, при проектировании электрических машин. Лучшим свидетельством этому является господство различных векторных методов в указанной области.

Попытаемся, однако, избежать этих упрощений. Мы найдем при этом, что основное уравнение очень быстро теряет свою простоту. Например, неупрощенное уравнение маятника при конечных колебаниях, даже в случае вакуума, имеет вид

$$\ddot{x} + \frac{g}{l} \sin x = 0 \quad (1.13')$$

и сводится к уравнению (1.2) только при малых колебаниях, когда можно заменить $\sin x$ на x . Аналогично и для пружины: если материал неоднороден, поперечные сечения пружины неодинаковы и т. п.—

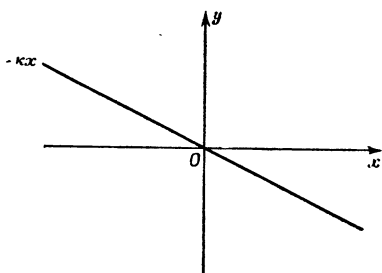


Рис. 1.9. Линейная упругая сила пружины.

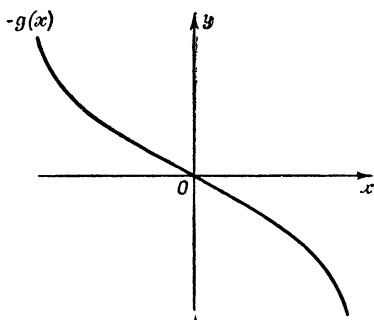


Рис. 1.10. Нелинейная упругая сила пружины.

упругая сила перестает выражаться через $-kx$ и превращается в более сложную функцию $-g(x)$, так что вместо уравнения (1.4) мы имеем более сложное уравнение:

$$\ddot{x} + g(x) = 0. \quad (1.13)$$

Уравнению (1.4) соответствует график, показанный на рис. 1.9, а уравнению (1.13)— более сложный график, например, такой, как на рис. 1.10. Если же мы хотим получить совершенно иное явление, достаточно ввести в рассмотренный выше электрический контур электронную лампу. В простейшем случае сопротивление выражается формулой $\beta x^2 - \alpha$, где α и β положительны, и уравнение новой системы будет иметь вид

$$L\ddot{x} + (\beta x^2 - \alpha)\dot{x} + \frac{x}{C} = 0. \quad (1.14)$$

Сопротивление отрицательно при достаточно малой силе тока x и становится положительным после того как x превзойдет некоторое

значение. Можно представить себе установившиеся колебания, при которых в течение некоторого времени система получает энергию и затем снова ее теряет. Это, в действительности, и происходит; имеется единственное установившееся колебание с определенной амплитудой и частотой. Мы видим здесь существенное отличие от случая линейных колебаний.

Если разделить уравнение (1.14) на L и изменить масштабы t и x , оно перейдет в знаменитое уравнение Ван дер Поля

$$\ddot{x} + k(x^2 - 1)\dot{x} + x = 0.$$

Все только что рассмотренные системы описываются следующим общим уравнением:

$$\ddot{x} + f(x, \dot{x})\dot{x} + g(x) = 0. \quad (1.15)$$

которое мы будем предполагать нелинейным. Нет сомнения, что соответствующие колебания значительно отличаются от возникающих в линейных системах.

§ 1.6. Некоторые нелинейные системы. В настоящем параграфе будет полезно рассмотреть системы, которые могут быть названы кусочно-линейными. В качестве особо простого случая рассмотрим пружину или часы с постоянной силой трения, направленной противоположно скорости. Эта система интересна тем, что ее можно полностью изучить с помощью простых графических построений и минимума вычислений. В то же время в ней проявляются основные особенности „нелинейных“ систем. Если сама пружина линейна, то основное уравнение будет:

$$\ddot{x} = -kx + h, \quad \text{если } \dot{x} < 0,$$

$$\ddot{x} = -kx - h, \quad \text{если } \dot{x} > 0,$$

где h есть постоянная, характеризующая силу трения. Заметим, что если $|\xi|$ обозначает абсолютную величину числа ξ , то сила трения может быть записана в виде $-h\dot{x}/|\dot{x}|$. Пользуясь таким обозначением, мы можем „двойное“ уравнение записать, используя только один знак равенства:

$$\ddot{x} + \frac{h\dot{x}}{|\dot{x}|} + kx = 0,$$

т. е. в форме уравнения (1.15) при $f = h/|\dot{x}|$. Удобнее, однако, записать это уравнение в виде

$$\ddot{x} + kx \pm h = 0,$$

где $\pm h$ имеет знак $\dot{x}$. Изменением масштаба времени можно привести это уравнение к виду

$$\ddot{x} + x \pm \alpha = 0,$$

где $\pm \alpha$ имеет знак $\dot{x}$. Как и раньше, заменяем это уравнение системой

$$\dot{x} = y, \quad \dot{y} = -(x \pm \alpha).$$

Мы можем также написать

$$(x \pm \alpha)' = y; \quad \dot{y} = -(x \pm \alpha),$$

где $\pm \alpha$ имеет знак y .

Предположим, что $y > 0$, и обозначим $x + \alpha = x^*$. Тогда будем иметь

$$\dot{x}^* = y, \quad \dot{y} = -x^* \quad (y > 0).$$

Это показывает, что над осью x траектории на фазовой плоскости представляют собой полуокружности с центром в точке A_1 с координатами

$$x^* = y = 0, \quad \text{т. е.} \quad x = -\alpha, \quad y = 0.$$

Допустим теперь, что $y < 0$ и обозначим $x - \alpha = x^{**}$. Тогда

$$\dot{x}^{**} = y, \quad \dot{y} = -x^{**} \quad (y < 0).$$

Следовательно, под осью x траектории являются полуокружностями с центром в точке A_2 , имеющей координаты $x^{**} = y = 0$, т. е. $x = \alpha$, $y = 0$.

Эта система пока еще не аналогична часам. Действительно, из графика (рис. 1.11) ясно видно, что, начав движение из любой точки, мы последовательно обходим полуокружности с центрами в A_1 , A_2 и останавливаемся при достижении отрезка $A_1 A_2$. Иначе говоря, мы описываем некоторую спираль, пока не достигнем A_1 , A_2 . Отсутствие аналогии с часами не удивительно, ибо система лишь поглощает энергию. Для системы, сходной с часами, необходим источник энергии. Пусть энергия подводится в виде постоянного импульса при прохождении x через нуль в положительном направлении. Этот случай изображен на рис. 1.12. Простой численный подсчет обнаруживает наличие единственной замкнутой траектории, к которой все другие (начинающиеся достаточно далеко) спиралевидно приближаются. Иными словами, система стремится иметь единственное и вполне определенное колебательное движение. Это явление весьма благоприятно для часового механизма и очень заметно отличается от характерных свойств линейных колебаний.

Рассмотренный случай является довольно общим для обширной группы нелинейных колебательных систем. Такие системы стремятся войти в режим единственного и вполне определенного колебательного

движения. Это происходит, например, для системы, подчиняющейся уравнению Ван дер Поля. В этом случае основная замкнутая кривая при малом коэффициенте k почти совпадает с окружностью радиуса 2 (рис. 1.13) и единственное колебательное движение представляет собой практически синусоидальное колебание $x = 2 \sin t$. С другой стороны, при большом k единственная замкнутая кривая принимает угловатую форму, показанную на рис. 1.14, и мы имеем почти прямоугольную волну, так как горизонтальные участки кри-

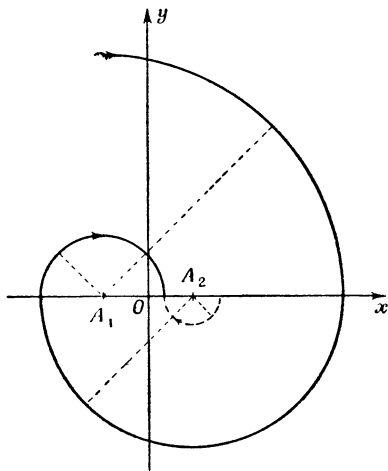


Рис. 1.11. Фазовая плоскость кусочно-линейной системы с поглощением, но без добавления энергии.

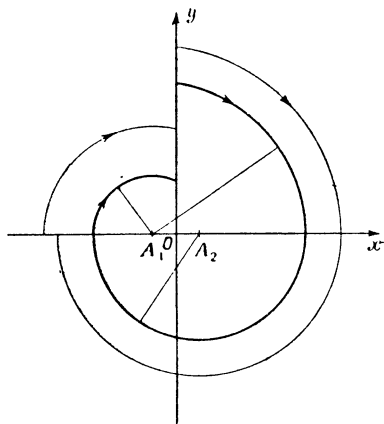


Рис. 1.12. Фазовая плоскость кусочно-линейной системы с поглощением и с добавлением энергии.

Внешняя дуга в первом квадранте является продолжением внешней дуги во втором.

вой обходятся с большой скоростью. Этот тип движений известен также под названием *релаксационных колебаний*.

Мы уже отмечали, что как в системе типа часов, так и в системе Ван дер Поля траектории стремятся к единственному колебательному движению. Иначе это выражается утверждением: колебания *устойчивы*. Если траектории стремятся отойти от колебательного движения, колебания *неустойчивы*. Явление устойчивости или неустойчивости колебаний обнаруживает еще одно отличие от случая линейных колебаний: в то время как линейные колебания всегда существенно устойчивы, нелинейные колебания могут оказаться неустойчивыми.

Устойчивость колебаний имеет большое практическое значение. Вообще говоря, можно утверждать, что только устойчивые колебания можно обнаружить и наблюдать. Неустойчивое колебание является как бы „отталкивающим“ фактором и не оставляет видимых следов.

Устойчивость колебаний не всегда является полезным свойством. В радио, телефоне и часах устойчивость может быть весьма желательным качеством, но во многих других системах ее можно расценить лишь как безусловную помеху, и при проектировании должны быть приняты меры к ее устранению. Вспомним, например, устройство Минорского для успокоения качки корабля с помощью двух резервуаров с водой, перетекающей из одного в другой. Эта схема подчиняется нелинейному уравнению и при определенных условиях

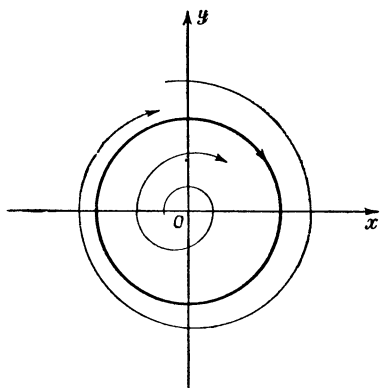


Рис. 1.13. Фазовая плоскость уравнения Ван дер Поля при малом k .

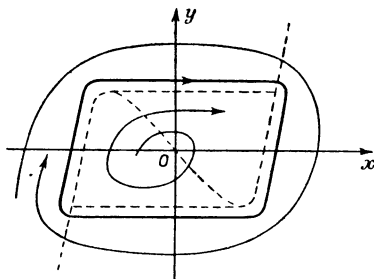


Рис. 1.14. Фазовая плоскость уравнения Ван дер Поля при большом k .

(соотношение емкостей резервуаров и соединяющего трубопровода) система стремится даже при тихой погоде войти в режим вполне определенной и устойчивой качки.

§ 1.7. Нелинейные колебания в консервативных системах. Существуют нелинейные системы, которые способны совершать бесконечное множество колебательных движений. Например, маятник с уравнением (1.13') является подобной системой. В общем случае подобная система подчиняется уравнению (1.13). Мы видим, что оно не имеет среднего диссипативного члена, чем и обеспечивается консервативность системы. Обозначим

$$G(x) = \int_0^x g(x) dx.$$

Умножая уравнение (1.13) на $\dot{x}$ и интегрируя, находим:

$$\frac{\dot{x}^2}{2} + G(x) = C. \quad (1.16)$$

Величина $G(x)$ определяет потенциальную энергию, а $\dot{x}^2/2$ — кинетическую энергию (для единичной массы), и соотношение (1.16) представляет собой обычный закон сохранения энергии для рассматриваемой системы. Отсюда получаем, что

$$\frac{dx}{dt} = [2[C - G(x)]]^{1/2},$$

$$\int_0^x \frac{dx}{\{2[C - G(x)]\}^{1/2}} = t + h. \quad (1.17)$$

Это соотношение недостаточно наглядно, поэтому предпочтительнее снова перейти к фазовой плоскости, полагая $\dot{x} = y$. Интегральные кривые таковы:

$$\frac{y^2}{2} + G(x) = C. \quad (1.18)$$

При довольно общих допущениях относительно потенциальной энергии $G(x)$ легко показать, что соотношение (1.18) представляет собой уравнение семейства concentric овалов (рис. 1.15), каждый из которых определяет колебательное движение системы. Амплитуда, очевидно, является переменной; *таким же свойством обладает и частота*; это можно показать, вычисляя интеграл (1.17)

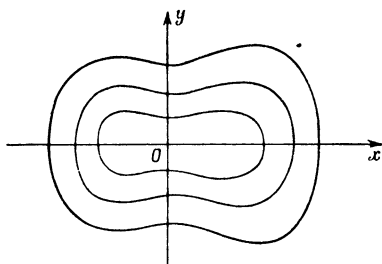


Рис. 1.15. Фазовая плоскость при нелинейных колебаниях в консервативной системе.

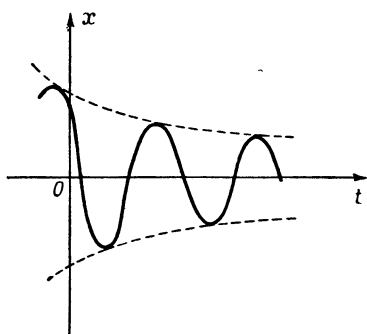


Рис. 1.16. Линейные колебания.

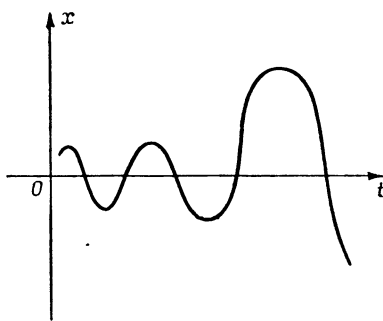


Рис. 1.17. Нелинейные колебания.

Переходный участок слева быстро сменяется установившимся режимом справа.

вдоль кривых семейства. Мы видим, что в данной нелинейной системе при изменении амплитуды изменяется также и частота, что

делает систему непригодной, например, для острой настройки. Отметим еще раз важное различие между линейными и нелинейными колебаниями. Линейные колебания происходят с постоянной частотой при любых начальных условиях.

Нелинейные колебания начинаются с переходного процесса и быстро переходят в установившийся режим. Это лучше всего видно на рис. 1.16 (случай линейных колебаний) и 1.17 (случай нелинейных колебаний).

§ 1.8. Вынужденные нелинейные колебания. Нелинейность может сочетаться с вынужденными колебаниями. Например, возможен контур с электронной лампой и переменной ЭДС (рис. 1.18). Уравнение такой системы имеет вид

$$L\ddot{x} + (\beta x^2 - \alpha)\dot{x} + \frac{x}{C} = E \cos \omega t,$$

или после упрощений, которые выше (стр. 25) привели к уравнению Ван дер Поля,

$$\ddot{x} + k(x^2 - 1)\dot{x} + x = e \cos \omega t.$$

Общее уравнение имеет вид

$$\ddot{x} + f(x, \dot{x})\dot{x} + g(x) = e(t),$$

где e — периодическая функция. Сведения о подобных общих системах немногочисленны. При некоторых условиях можно утверждать, что существует периодическое решение $x(t)$ с таким же периодом, как и у функции $e(t)$, и что это решение устойчиво.

Замечательное новое явление может возникнуть в нелинейных системах с периодическим возмущающим членом; если период возмущающего члена есть T , то могут возникнуть вынужденные колебания с периодами, кратными T : $2T$, $3T$, ... Это явление известно под названием *субгармонического резонанса*, причем его обнаружение

с помощью математических методов в лучшем случае затруднительно. Сравнительно хорошие результаты были недавно получены японским математиком Симидзу для систем, описываемых уравнением

$$\ddot{x} + c\dot{x} + g(x) = e \sin \omega t,$$

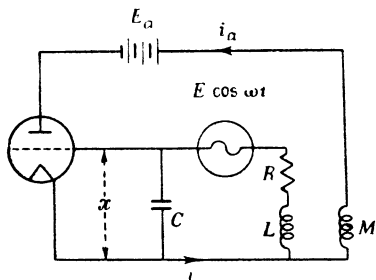


Рис. 1.18. Контур с электронной лампой и переменной ЭДС.

где $g(x)$ — многочлен нечетной степени с положительными коэффициентами, а коэффициенты s и e малы. Но рассмотрение таких систем завело бы нас слишком далеко.

Инженеру трудно представить себе субгармонический резонанс, однако он вполне реален. Л. Б. Такерман, бывший член Национального бюро стандартов, сообщил нам несколько лет назад о следующем случае: у одного из типов гражданских самолетов пропеллеры вызывали колебания крыльев с двукратным периодом, что в свою очередь приводило к колебаниям руля направления с четырехкратным периодом. Последние были столь интенсивными, что самолеты разрушались.

С точки зрения электротехники интенсивность низкочастотных колебаний хорошо объяснима, так как их импеданс значительно меньше импеданса тока *нормальной частоты*.

§ 1.9. Контур мультивибратора. Подобные контуры обладают по меньшей мере двумя особенностями, имеющими теоретический интерес. Первая состоит в том, что они порождают колебания, связанные не с одним дифференциальным уравнением второго порядка, а с *двумя* уравнениями первого порядка. Вторая состоит в том, что возбужденные колебания, вообще говоря, разрывны. В этом снова проявляется существенное отличие от линейных колебаний. Мы ограничимся предельно упрощенным изложением, достаточным лишь для обнаружения указанных двух особенностей.

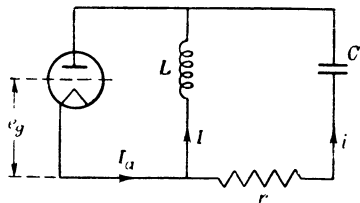


Рис. 1.19. Контур мультивибратора.

Интересующая нас часть контура схематически показана на рис. 1.19. Опущенная часть схемы включает в первую очередь электронную лампу — усилитель. Система подчиняется следующим двум основным уравнениям, которые вытекают из законов Кирхгофа:

$$I_a = \varphi(kri) = I + i, \quad L \frac{dI}{dt} = ri + \frac{1}{C} \int i dt.$$

Функция $I_a = \varphi(e_g)$ есть так называемая *характеристика лампы*, а $e_g = kri$ — сеточное напряжение усилительной лампы.

Обозначая $x = kri$ и $y = dI/dt$, приводим систему к виду

$$\varphi(x) = \int y dt + \frac{x}{kr}, \quad Ly = \frac{x}{k} + \frac{1}{Ckr} \int x dt,$$

или, после дифференцирования по t ,

$$\dot{x}\varphi'(x) = y + \frac{\dot{x}}{kr}, \quad \dot{y} = \frac{\dot{x}}{kL} + \frac{x}{krLC}.$$

Отсюда получаем

$$\dot{x} = \frac{y}{\varphi'(x) - 1/kr}, \quad \dot{y} = \frac{1}{kL} \frac{y}{\varphi'(x) - 1/kr} + \frac{x}{krLC}. \quad (1.19)$$

Примем простейшее возможное допущение, что $\varphi(x) = b^2 - x^2$. Вводя необходимые обозначения, приведем (1.19) к окончательному виду

$$\dot{x} = \frac{y}{a^2 - x^2}; \quad \dot{y} = \frac{by}{a^2 - x^2} + cx.$$

Эта система имеет разрывы непрерывности вдоль прямых $x = \pm a$, которые приводят к своего рода „внутренним“ толчкам. То, что

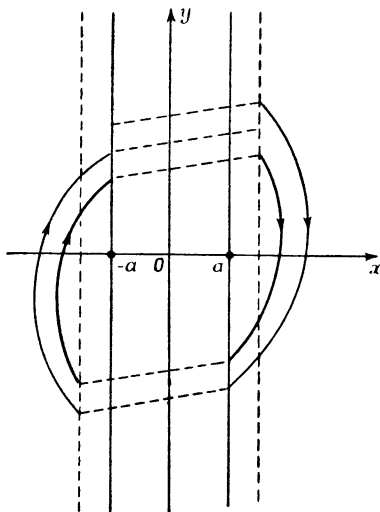


Рис. 1.20. Фазовая плоскость для контура мультивибратора.

происходит в фазовой плоскости x, y , схематически показано на рис. 1.20. По достижении полосы, заключенной между прямыми $x = \pm a$, происходит мгновенный скачок через полосу на прямые, лежащие вблизи полосы. Этот скачок подчиняется так называемому *закону Мандельштама*. Оказывается, что система ведет себя с удивительной закономерностью; существует единственная замкнутая траектория, которая весьма устойчива, причем все траектории спиралевидно приближаются к ней. Это свойство нетрудно обнаружить.

Только что намеченная схема принадлежит советским физикам Хайкину и Лошакову. Интересующихся деталями отсылаем к книгам Андронова и Хайкина [1] и Минорского [4, стр. 405].

§ 1.10. Математическое исследование нелинейных задач. Не будет преувеличением сказать, что методы математического изучения нелинейных задач на сегодня весьма несовершенны. Мы имеем, с одной стороны, качественные методы, в основном восходящие к Пуанкаре; с другой стороны, существуют методы аппроксимации, о которых можно сказать, что они используются за неимением лучшего, несмотря на их несовершенство.

Качественные методы имеют целью дать общую картину того, что происходит для всех возможных случаев. Возьмем в качестве отправной точки уравнение весьма общего вида

$$\ddot{x} + f(x)\dot{x} + g(x) = 0. \quad (1.20)$$

Введя обозначения

$$F(x) = \int_0^x f(x) dx; \quad y = \dot{x} + F(x),$$

заменяем уравнение (1.20) эквивалентной ему системой

$$\dot{x} = y - F(x), \quad \dot{y} = -g(x), \quad (1.21)$$

которая принадлежит к общему типу уравнений

$$\dot{x} = P(x, y), \quad \dot{y} = Q(x, y),$$

детально рассмотренному Пуанкаре. Остановимся на этом общем случае. Пуанкаре обнаружил, что для поведения подобной системы особенно важное значение имеют:

а) *особые точки*, или точки, в которых $\dot{x} = \dot{y} = 0$; они соответствуют состояниям равновесия;

б) *предельные циклы*, которые, как мы видели, соответствуют периодическим решениям¹⁾.

Особые точки являются пересечением кривых $P=0$, $Q=0$. Можно легко определить поведение решений вблизи особых точек.

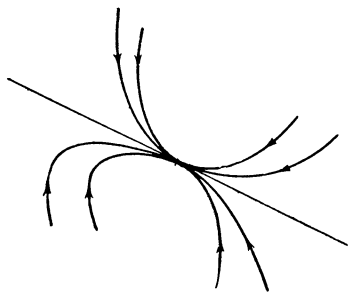


Рис. 1.21. Особая точка — узел.

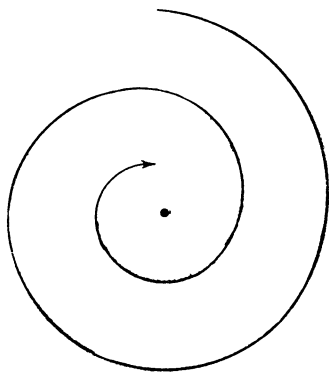


Рис. 1.22. Особая точка — фокус.

В сравнительно простых случаях особые точки распадаются на три типа, показанные на рис. 1.21—1.23. Изображенные на рис. 1.21 и 1.22 узел и фокус являются *устойчивыми*. Для получения *неустойчивых* узла и фокуса надо просто переменить направление стрелок.

Наиболее важным свойством обычного предельного цикла является то, что близлежащие решения спирально наматываются на него

¹⁾ Изолированным. — Прим. ред.

(устойчивый предельный цикл) или сматываются с него (неустойчивый предельный цикл). См. рис. 1.24.

Важное понятие *индекса*, введенное Пуанкаре, позволило ему обнаружить связь между особыми точками и предельными циклами. Пусть на рис. 1.25 V есть вектор с компонентами P, Q . Когда точка обходит замкнутую кривую γ , не проходящую через особые

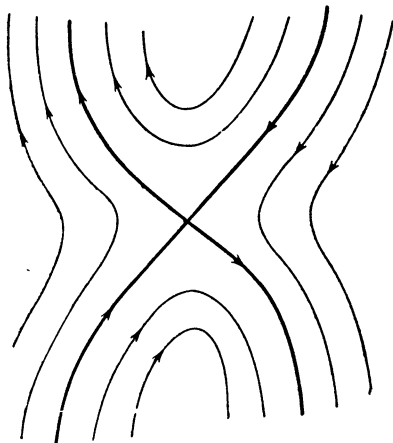


Рис. 1.23. Особая точка — седло.

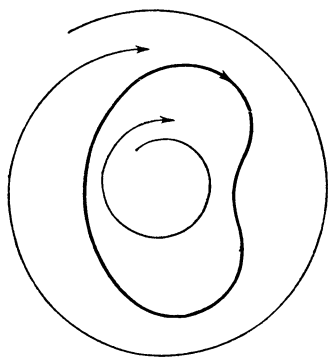


Рис. 1.24. Предельный цикл.

точки, то угол поворота вектора V оказывается равным произведению целого числа m на 2π ; m называется индексом кривой γ . Если кривая γ стягивается в точку A (см. рис. 1.26), то ее индекс остается

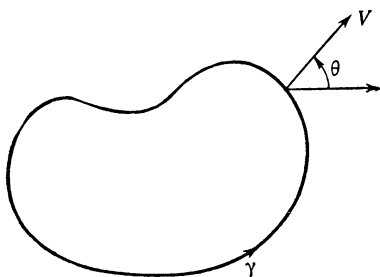


Рис. 1.25. К определению индекса кривой.

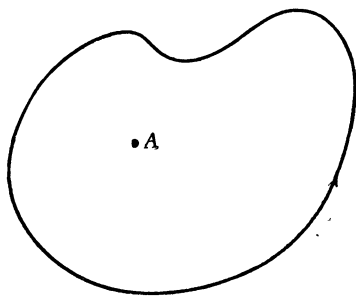


Рис. 1.26. К определению индекса точки.

постоянным и определяет индекс точки A . Основные свойства индекса таковы:

а) индекс неособой точки равен нулю;

б) индекс замкнутой кривой есть сумма индексов особых точек, лежащих внутри нее;

в) индекс предельного цикла равен единице;

г) индекс узла или фокуса равен 1, а седловой точки — 1.

Как следствие свойства (в) можно утверждать, что предельный цикл должен иметь внутри себя особую точку, причем он не может

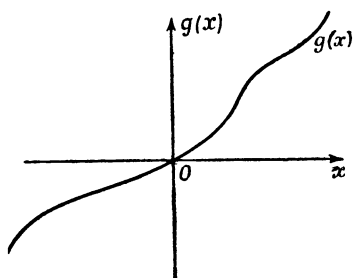


Рис. 1.27. График функции $g(x)$.

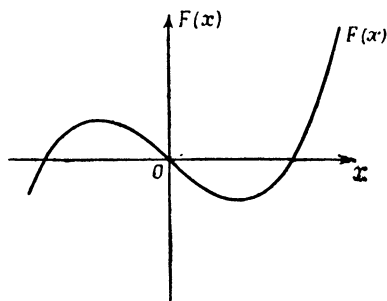


Рис. 1.28. Простейшая форма графика функции $F(x)$.

окружать седловые точки. На большую информацию трудно было и рассчитывать.

Применим изложенные соображения к уравнению (1.20) или, скорее, к эквивалентной ему системе (1.21). Сделаем упрощающие предположения, что $g(x)$ имеет знак x (см. рис. 1.27) и что

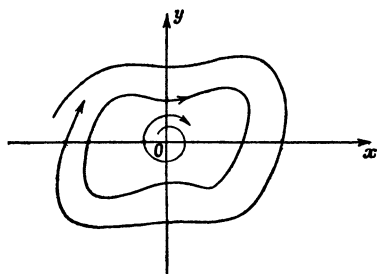


Рис. 1.29. Устойчивый предельный цикл и неустойчивый фокус.

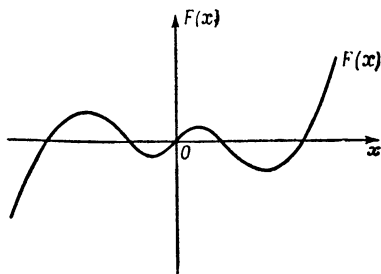


Рис. 1.30. Другая форма графика функции $F(x)$.

$F(x)$ — нечетная функция, положительная и превосходящая x при больших x . Простейшим допущением относительно $F(x)$ является то, что она ведет себя, как показано на рис. 1.28, т. е. похожа на кубический многочлен. При этих предположениях, как показал Льенар,

имеется в точности один предельный цикл (как на рис. 1.29), причем он устойчив. Единственной особой точкой является начало координат, причем это неустойчивый узел или фокус. Предельный цикл, очевидно, должен охватывать начало координат, так как последнее есть единственная особая точка. Таким образом, общая картина расположения траекторий такова, как показано на рис. 1.29. В частности, как мы видели, так располагаются интегральные кривые уравнения Ван дер Поля.

Здесь можно сделать интересное замечание. Как было отмечено самим Ван дер Полем, допущение, принятое при выводе уравнения,

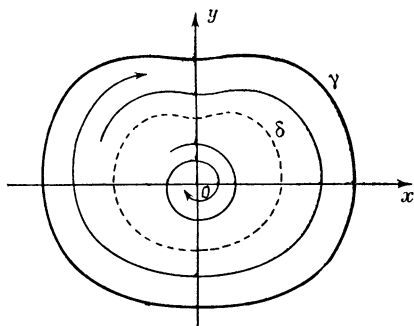


Рис. 1.31. Устойчивый фокус, окруженный неустойчивым предельным циклом δ и устойчивым предельным циклом γ .

не вполне соответствует наблюдаемым данным, и было бы точнее заменить выражение $\beta x^2 - \alpha$ выражением типа $\eta x^4 - \beta x^2 + \alpha$. Это означало бы, что график функции $F(x)$ на рис. 1.28 нужно заменить графиком, показанным на рис. 1.30. Вследствие этого начало координат становится устойчивым узлом или фокусом, окруженным двумя предельными циклами, причем внутренний δ неустойчив, а внешний γ устойчив (как показано на рис. 1.31). Для достижения режима установившихся колебаний, соответствующих циклу γ , необходимо начать движение *вне цикла* δ .

Если движение начнется в точке, лежащей внутри δ , то произойдет возвращение в начало координат, т. е. в положение равновесия. Иначе говоря, для возбуждения колебаний электронной лампы необходимо преодолеть инерцию этой лампы. Колебания подобного типа носят название *жестких*; рассмотренные ранее колебания, связанные с рис. 1.29, носят название *мягких*.

§ 1.11. Приближенные методы. Приближенные методы с наибольшим успехом применяются к системам, для которых уравнение движения имеет вид

$$\ddot{x} + \omega^2 x = kf(x, \dot{x}), \quad (1.22)$$

где k — малая положительная постоянная. Системы подобного рода близки к гармоническим осцилляторам, вследствие чего их часто называют *квазигармоническими*. Если коэффициент k в уравнении Ван дер Поля мал, то это уравнение принадлежит к рассматриваемому типу, так как его можно записать в виде

$$\ddot{x} + x = k(1 - x^2)\dot{x}.$$

Пользуясь методом, принадлежащим Ван дер Полю, в форме, предложенной советскими математиками Крыловым и Боголюбовым, попытаемся решить уравнение (1.22) с помощью функции

$$x = a \cos(\omega t + \varphi), \quad (1.23)$$

где a и φ выбираются так, что $\dot{x}$ имеет тот же вид, как если бы a и φ были постоянными. Можно ожидать, что a и φ изменяются медленно, так что выражение (1.23) для больших интервалов времени представляет гармонические колебания с периодом $2\pi/\omega$, хорошо аппроксимирующие истинное решение. Условие, наложенное на производную, дает

$$\dot{x} = -a\omega \sin(\omega t + \varphi)$$

и дополнительно

$$\dot{a} \cos(\omega t + \varphi) - a\dot{\varphi} \sin(\omega t + \varphi) = 0.$$

Удобно ввести новую переменную $\psi = \omega t + \varphi$. Предыдущее соотношение, выраженное через эту переменную, будет иметь вид

$$\dot{a} \cos \psi - a\dot{\varphi} \sin \psi = 0. \quad (1.24)$$

Составив выражение $\ddot{x} + \omega^2 x$, мы сможем переписать уравнение (1.22) в виде

$$-a\omega \sin \psi - a\omega\dot{\varphi} \cos \psi = kf(a \cos \psi, -a\omega \sin \psi). \quad (1.25)$$

Из уравнений (1.24) и (1.25) получаем

$$\begin{aligned} \dot{a} &= \frac{-k}{\omega} f(a \cos \psi, -a\omega \sin \psi) \sin \psi, \\ \dot{\varphi} &= \frac{-k}{\omega a} f(a \cos \psi, -a\omega \sin \psi) \cos \psi. \end{aligned} \quad (1.26)$$

Так как $\dot{a}$ и $\dot{\varphi}$ пропорциональны k , то a и φ действительно изменяются медленно. На первый взгляд система уравнений (1.26) более сложна для исследования, чем исходное уравнение (1.22). Однако благодаря медленному изменению a и φ правые части в системе (1.26) можно заменить их средними по времени за период, вычисленными так, как если бы a и φ были постоянными. Это приводит к новой и значительно упрощенной системе, так как $\sin \psi$ и $\cos \psi$ в результате интегрирования пропадают

$$\dot{a} = kG(a), \quad \dot{\varphi} = kH(a).$$

Решая уравнение $G=0$, получим стационарные амплитуды. Если $G(a_0)=H(a_0)=0$, то $x=a_0 \cos(\omega t + \varphi_0)$ является приближенным решением рассматриваемой системы при малых k . В применении

к уравнению Ван дер Поля мы находим, что $H(a) = 0$ и

$$\begin{aligned} G(a) &= \frac{-1}{2\pi} \int_0^{2\pi} (1 - a^2 \cos^2 \psi)(-a \sin^2 \psi) dt = \\ &= \frac{a}{2\pi} \int_0^{2\pi} (1 - a^2 \cos^2 \psi)(\sin^2 \psi) dt = \frac{a}{2\pi} \left(1 - \frac{a^2}{4}\right). \end{aligned}$$

Так как значение $a = 0$ исключается и a неотрицательно, то уравнение $G(a) = 0$ имеет решение $a = 2$. Следовательно, уравнение Ван дер Поля для малых k имеет приближенное решение $x = 2 \cos(t + \varphi)$, как мы уже установили выше.

§ 1.12. Метод Дуффинга. Примерно 35 лет назад Дуффинг исследовал уравнение частного вида

$$\ddot{x} + \alpha x + \beta x^3 = F \cos \omega t, \quad (1.27)$$

т. е. уравнение движения материальной точки, находящейся под действием упругой силы $-\alpha x - \beta x^3$ и возмущающей периодической силы $F \cos \omega t$. Он приложил к нему метод, тесно связанный с методом, применявшимся некоторыми астрономами конца девятнадцатого века. Грубо говоря, допуская, что система обладает периодическим решением, мы строим последовательные приближения таким образом, чтобы решение имело подходящий период на каждом шаге приближения.

Вернемся ненадолго к явлению резонанса. Для простоты рассмотрим уравнение

$$\ddot{x} + \omega^2 x = F \cos \omega t + G(t), \quad (1.28)$$

где $G(t)$ не обладает периодом $2\pi/\omega$. Частное решение уравнения (1.28) является суммой решения уравнения

$$\ddot{x} + \omega^2 x = G(t),$$

не представляющего интереса, и уравнения

$$\ddot{x} + \omega^2 x = F \cos \omega t.$$

Последнее имеет частное решение $x = (Ft/2) \sin \omega t$. Содержащий t член, появляющийся в решении, известен под названием *векового члена*. Он соответствует колебаниям со все время возрастающей амплитудой, или, иными словами, *резонансу*. Исключению векового члена астрономы посвятили значительные усилия; в этом же состоит и сущность метода Дуффинга. Излагаемый вариант метода мы заимствуем из книги Стокера [5, стр. 88].

Запишем уравнение (1.27) в форме

$$\ddot{x} + \omega^2 x = (\omega^2 - \alpha)x - \beta x^3 + \beta F_0 \cos \omega t, \quad (1.29)$$

имея в виду, что возмущающий член столь же мал, как и член βx^3 . Подставим в правую часть уравнения выражение $x_1 = A \cos \omega t$ в качестве первого приближения; для второго приближения получим уравнение

$$\begin{aligned} \ddot{x}_2 + \omega^2 x_2 &= (\omega^2 - \alpha) A \cos \omega t - \beta A^3 \cos^3 \omega t + \beta F_0 \cos \omega t = \\ &= \left[(\omega^2 - \alpha) A - \frac{3}{4} \beta A^3 + \beta F_0 \right] \cos \omega t - \frac{1}{4} \beta A^3 \cos 3\omega t. \end{aligned} \quad (1.30)$$

Это линейное уравнение относительно x_2 , которое легко решается.

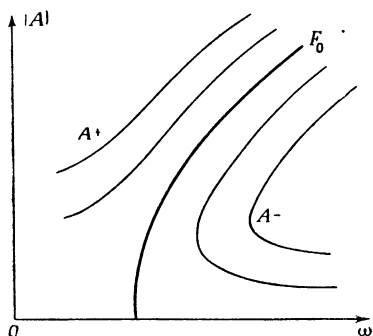


Рис. 1.32. График зависимости амплитуды первой гармоники от частоты.

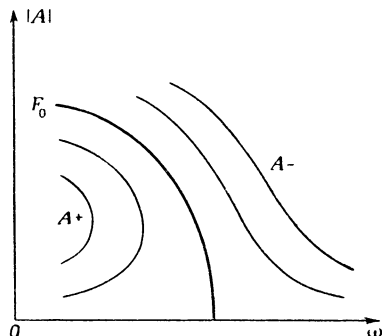


Рис. 1.33. Другой возможный вид графика зависимости амплитуды первой гармоники от частоты.

Для того чтобы решение не содержало векового члена, член с $\cos \omega t$ должен исключаться, что накладывает условие

$$(\omega^2 - \alpha) A - \frac{3}{4} \beta A^3 + \beta F_0 = 0,$$

откуда имеем

$$\omega^2 = \alpha + \frac{3}{4} \beta A^2 - \beta \frac{F_0}{A}. \quad (1.31)$$

Тогда решение уравнения (1.30) принимает вид

$$x_2 = A_1 \cos \omega t + (\dots) \cos 3\omega t,$$

где A_1 зависит от ω и A . Дуффинг полагает $A_1 = A$, что дает *поправленное* соотношение (1.31), и повторяет процесс, подставляя $x = x_2$ в правую часть уравнения (1.29) и определяя x_3 из уравнения

$$\ddot{x}_3 + \omega^2 x_3 = P_1 \cos \omega t + P_3 \cos 3\omega t + \dots$$

Полагаем $P_1 = 0$ и повторяем процесс. График зависимости амплитуды $|A|$ от ω (т. е. зависимости амплитуды первой гармоники от частоты), соответствующий выражению (1.31), имеет вид, показанный на рис. 1.32 и 1.33.

§ 1.13. Метод возмущений, предложенный Пуанкаре. Вернемся к уравнению (1.22) для квазигармонического осциллятора

$$\ddot{x} + \omega^2 x = kf(x, \dot{x})$$

и предположим для простоты, что f есть многочлен. Следствие из весьма общей теоремы, доказанной Пуанкаре, утверждает, что если существует периодическое решение, стремящееся, например, к решению $\lambda \sin \omega t$ при стремлении k к нулю, то оно может быть представлено рядом по степеням k

$$x = \lambda \sin \omega t + A_1(\lambda, t)k + A_2(\lambda, t)k^2 + \dots$$

Соответствующее значение $y = \dot{x}$ имеет вид

$$y = \lambda \omega \cos \omega t + \frac{dA_1}{dt}k + \frac{dA_2}{dt}k^2 + \dots$$

Мы должны выразить тот факт, что решение имеет период $2\pi/\omega + \tau(k)$, где

$$\tau(k) = \tau_0 + k\tau_1 + k^2\tau_2 + \dots,$$

т. е.

$$x\left(\frac{2\pi}{\omega} + \tau\right) = x(0), \quad y\left(\frac{2\pi}{\omega} + \tau\right) = y(0),$$

что дает два соотношения между τ , λ , k . Затем, подставляя x и y в уравнение (1.22) и приравнявая члены с одинаковыми степенями k , получаем последовательность дифференциальных уравнений для $A_1(\lambda, t)$, $A_2(\lambda, t)$, ..., которые при начальных условиях

$$A_1(\lambda, 0) = 0, \quad \left[\frac{dA_1}{dt}\right]_{t=0} = 0$$

и т. д., могут быть последовательно решены. При этом для x и y получаются ряды по степеням k , которые сходятся для малых k и обеспечивают приближение к решению со сколь угодно высокой точностью. Детальное изложение, выходящее за рамки рассмотренных общих замечаний, слишком сложно для воспроизведения в настоящей главе. Заметим лишь, что основой является представление решения в виде степенного ряда, и все факты группируются вокруг этого.

ЛИТЕРАТУРА

1. Андронов А. А. и Хайкин С. Э., Теория колебаний, Гостехиздат, М.—Л., 1937.
2. Ден-Гартог Дж. П., Теория колебаний, Гостехиздат, М.—Л., 1942.
3. Lefschetz S., Lectures on Differential Equations, Ann. Math. Stud., No 14, Princeton University Press, Princeton, N. J., 1946.
4. Minorsky N., Introduction to Non-linear Mechanics, Edwards, Ann Arbor, Mich., 1941.
5. Стокер Дж., Нелинейные колебания в механических и электрических системах, ИЛ, М., 1953.
6. Тимошенко С. П., Теория колебаний в инженерном деле, Гостехиздат, изд. 2, М.—Л., 1932.

Глава 2

ИССЛЕДОВАНИЕ РАВНОВЕСИЯ. ТЕОРИЯ УСТОЙЧИВОСТИ ПУАНКАРЕ И ЛЯПУНОВА

РИЧАРД БЕЛЛМАН

Корпорация Rand

§ 2.1. Введение. В предыдущей главе рассматривался вопрос о линейных и нелинейных колебаниях, имеющий огромное значение для теории и приложений дифференциальных уравнений. В этой главе мы попытаемся кратко наметить некоторые из наиболее важных черт другой области общей теории дифференциальных уравнений, а именно теории устойчивости состояний равновесия.

Основная проблема может быть сформулирована очень просто. Допустим, что мы имеем физическую систему, описываемую в любой момент времени t совокупностью переменных $[x_1(t), x_2(t), \dots, x_N(t)]$, определяющих ее состояние. Пусть эти переменные изменяются во времени согласно системе дифференциальных уравнений

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \dots, x_N), \quad x_i(0) = c_i, \quad (2.1)$$

где $i = 1, 2, \dots, N$.

Под *состоянием равновесия* или *стационарным состоянием* мы понимаем совокупность значений $(a_1, a_2, \dots, a_N)$, обладающих тем свойством, что

$$f_i(a_1, a_2, \dots, a_N) = 0, \quad i = 1, 2, \dots, N.$$

Эти значения определяют решение системы (2.1), являющееся точкой в пространстве $(x_1, x_2, \dots, x_N)$. При отсутствии внешней силы система останется в положении, определяемом равенствами $x_i = a_i$, если она находилась в нем в начальный момент.

Допустим теперь, что прикладывается некоторая внешняя сила (динамически определенная или случайная), вследствие чего точка $(a_1, a_2, \dots, a_N)$ смещается в точку $(a'_1, a'_2, \dots, a'_N)$. Вернется ли система в свое положение равновесия?

Если она возвращается при любых возможных возмущениях, сколь угодно больших, то положение равновесия называют *абсолютно устойчивым*. Если система возвращается в положение равновесия при достаточно малых возмущениях, то положение равновесия называют *устойчивым*.

Ясно, что абсолютная устойчивость представляет собой математическую фикцию и что в физических задачах величина возмущения играет существенную роль в определении последующего поведения системы.

В следующих параграфах мы остановимся на наиболее важном результате (теорема 2.1, см. ниже) теории устойчивости нелинейных дифференциальных уравнений, после чего рассмотрим в общих чертах некоторые родственные вопросы теории линейных дифференциальных уравнений, дифференциально-разностных уравнений и уравнений с частными производными параболического типа.

Дальнейшие результаты теории устойчивости и теории асимптотического поведения решений дифференциальных уравнений можно найти в книге автора [1], а также в более ранней его книге [2]. Рассмотрение результатов теории процессов с запаздыванием, приводящей к дифференциально-разностным уравнениям, можно найти в докладе автора [3], где имеются ссылки на более ранние результаты Е. М. Райта и автора. Ж. Проди недавно выполнил важное исследование по теории нелинейных параболических уравнений; обзор более ранних исследований по этому вопросу дан в статье автора [4].

§ 2.2. Теория устойчивости Пуанкаре и Ляпунова. Предположим, как это имеет место в большинстве физических задач, что каждая из функций $f_i(x_1, x_2, \dots, x_N)$ является степенным рядом относительно переменных $x_1, x_2, \dots, x_N$, и примем без ограничения общности, что все a_i равны нулю. Тогда основная система дифференциальных уравнений примет вид

$$\frac{dx_i}{dt} = \sum_{j=1}^N a_{ij}x_j + g_i(x_1, x_2, \dots, x_N), \quad i=1, 2, \dots, N, \quad (2.2)$$

где $A=(a_{ij})$ — постоянная матрица (см. гл. 13) и каждая функция $g_i(x)$ есть степенной ряд, не содержащий свободных членов и членов с первыми степенями переменных.

Зададим начальные условия

$$x_i(0) = c_i, \quad i=1, 2, \dots, N,$$

причем допустим, что c_i выбраны достаточно малыми, так что в полученной окрестности начала координат каждая из функций $g_i(x)$ мала по сравнению с линейными членами. Естественно предположить, что устойчивость положения равновесия $x_1=x_2=\dots=x_N=0$ для уравнения (2.2) зависит от того, будет ли это положение устойчивым для линейного приближения

$$\frac{dx_i}{dt} = \sum_{j=1}^N a_{ij}x_j, \quad i=1, 2, \dots, N. \quad (2.3)$$

То, что это действительно имеет место, и составляет сущность классического результата Пуанкаре и Ляпунова.

Прежде чем дать формулировку этой теоремы, рассмотрим асимптотическое поведение решений системы (2.3).

Хорошо известно [1], что все решения системы (2.3) можно получить с помощью частных решений вида $x_i = e^{\lambda t} c_i$ и их предельных образований. Подстановка их в уравнения (2.3) приводит к алгебраической системе

$$\lambda c_i = \sum_{j=1}^N a_{ij} c_j, \quad i = 1, 2, \dots, N.$$

Процесс исключения постоянных c_i , которые не могут быть одновременно равными нулю, приводит к равенству нулю детерминанта $|A - \lambda I|$, или к *характеристическому уравнению* для A . Корни этого уравнения называются *характеристическими корнями* матрицы A . Если встретятся кратные корни, мы можем получить решение вида

$$x_i = e^{\lambda t} p_i(t),$$

где $p_i(t)$ — многочлены относительно t .

Мы видим, что в любом случае асимптотическое поведение решений линейных дифференциальных уравнений определяется алгебраическим характером характеристических корней. Необходимое и достаточное условие для того, чтобы все решения системы (2.3) стремились к нулю при $t \rightarrow \infty$, состоит в том, чтобы все характеристические корни имели отрицательные вещественные части. Проверку выполнения этого условия без явного вычисления корней можно произвести с помощью критерия Гурвица [3], коль скоро характеристическое уравнение $|A - \lambda I| = 0$ написано в виде многочлена.

Перейдем теперь к формулировке основного результата Пуанкаре и Ляпунова.

ТЕОРЕМА 2.1¹⁾. *Рассмотрим систему дифференциальных уравнений (2.2).*

Допустим, что:

а) *все характеристические корни матрицы A имеют отрицательные вещественные части и*

б)

$$\frac{|g_i(x_1, x_2, \dots, x_N)|}{|x_1| + |x_2| + \dots + |x_N|} \rightarrow 0 \quad \text{при} \quad |x_1| + |x_2| + \dots + |x_N| \rightarrow 0.$$

Тогда любое решение уравнения (2.2), для которого $|c_1| + \dots + |c_N|$ достаточно мало, стремится к нулю при $t \rightarrow \infty$.

¹⁾ В доказанной Пуанкаре и Ляпуновым теореме предполагалось, что $g_i(x)$ являются степенными рядами, не содержащими свободных членов и членов первой степени. В приведенном виде теорема доказана Перроном.

Приведенная теорема подтверждает корректность использования линейного приближения для проверки устойчивости положения равновесия. Для нее было дано несколько доказательств, каждое из которых интересно по-своему.

§ 2.3. Теория устойчивости линейных уравнений. В проведенной выше математической формулировке поведения системы предполагалась стационарность, т. е. управляющие поведением системы законы принимались не зависящими от времени. Во многих прикладных задачах это не выполняется. Рассмотрим случай, когда система подчиняется линейным уравнениям с переменными коэффициентами:

$$\frac{dx_i}{dt} = \sum_{j=1}^N [a_{ij} + b_{ij}(t)] x_j, \quad i = 1, 2, \dots, N. \quad (2.4)$$

Естественно предположить, что если $b_{ij}(t) \rightarrow 0$ при $t \rightarrow \infty$, то асимптотическое поведение решений (2.4) будет повторять поведение решений аппроксимирующего уравнения

$$\frac{dx_i}{dt} = \sum_{j=1}^N a_{ij} x_j, \quad i = 1, 2, \dots, N. \quad (2.5)$$

К несчастью для инженеров и счастью для математиков, это не всегда верно. Приведем простой пример. Можно показать, что не все решения уравнения

$$\frac{d^2x}{dt^2} + \left(1 + \frac{\sin 2t}{t}\right) x = 0$$

ограничены, несмотря на то, что решения аппроксимирующего уравнения $d^2x/dt^2 + x = 0$ ограничены. Мы имеем здесь типичную задачу о сохранении свойств решений при малых изменениях формы уравнения.

Полезная теорема Хукухары (см. [1]) состоит в следующем:

ТЕОРЕМА 2.2. Если:

а) все решения уравнения (2.5) ограничены при $t \rightarrow \infty$ и

б) $\int_0^\infty |b_{ij}(t)| dt < \infty, \quad i, j = 1, 2, \dots, N,$

то все решения уравнения (2.4) ограничены.

Если относительно $b_{ij}(t)$ известно еще что-либо, то соответствующие дополнительные сведения могут быть получены и относительно решений. Особенно важным случаем является тот, когда каждая из функций $b_{ij}(t)$ имеет асимптотический ряд вида

$$b_{ij}(t) \sim b_{ij0} + \frac{b_{ij1}}{t} + \frac{b_{ij2}}{t^2} + \dots$$

В этом случае могут быть найдены асимптотические ряды для решений [1].

§ 2.4. Дифференциально-разностные уравнения. Во многих важных приложениях, относящихся к системам автоматического регулирования, для выполнения сигнала регулирующего механизма требуется известное время. В задачах этого типа мы получаем вместо дифференциального дифференциально-разностное уравнение, а иногда и функциональное уравнение более сложного типа.

Так, например, при описании движения маятника с демпфированием вместо обычного уравнения

$$\ddot{x}(t) + a\dot{x}(t) + bx(t) = 0 \quad (2.6)$$

мы получаем уравнение

$$\ddot{x}(t) + a\dot{x}(t - \tau) + bx(t) = 0, \quad (2.7)$$

если демпфирующий сигнал имеет запаздывание τ .

Соответствующее уравнению (2.6) характеристическое уравнение, полученное с помощью подстановки $x = e^{\lambda t}$, таково:

$$\lambda^2 + a\lambda + b = 0.$$

Если a и b — положительные числа, то легко видеть, что любое решение уравнения (2.6) стремится к нулю при $t \rightarrow \infty$. Уравнение, аналогичным способом получаемое из уравнения (2.7), таково:

$$\lambda^2 + ae^{-\lambda\tau} + b = 0.$$

В данном случае нахождение областей на плоскости ab , которым соответствуют уравнения, все корни которых имеют отрицательные вещественные части, уже не является тривиальной задачей. Важная работа, относящаяся к этой проблеме, выполнена Понтрягиным [8]. Ее изложение на английском языке, совместно с некоторыми приложениями содержащихся в ней методов, имеется в работе автора [3]. Другой подход к одному уравнению частного вида можно найти в статье Хэйса [6].

Теория устойчивости нелинейных дифференциально-разностных уравнений была предложена автором [3], а также Е. М. Райтом [11], весьма детально исследовавшим эти вопросы.

В заключение отметим, что теория периодических решений дифференциально-разностных уравнений была изложена Минорским [7] и Браунеллом [5].

§ 2.5. Уравнение теплопроводности. Задачи устойчивости вполне аналогичного типа возникают при рассмотрении нелинейного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + g(u), \quad (2.8)$$

где граничные и начальные условия имеют вид

$$u = 0 \text{ на } B, \text{ т. е. на границе области } R(x, y, z),$$

$$u = f(x, y, z) \text{ внутри } R \text{ при } t = 0.$$

Исходя из физических соображений, следует ожидать, что если R есть ограниченная достаточно правильная область, то условия

$$a) \max_R |f| \text{ достаточно мал,}$$

$$б) g(u) = o(u) \text{ при } u \rightarrow 0$$

гарантируют стремление решения уравнения (2.8) к нулю при $t \rightarrow \infty$.

Имеется много путей исследования этого вопроса. Библиография ранее опубликованных статей может быть найдена в работе автора [4], в которой приводится решение для того случая, когда R есть прямоугольный параллелепипед. Полное решение для областей общего вида было получено П. Лаксом (в неопубликованном сообщении), а в опубликованной литературе — Ж. Проди [9], который исследовал также другие интересные вопросы в этом направлении [10].

ЛИТЕРАТУРА

1. Беллман Р., Теория устойчивости решений дифференциальных уравнений, ИЛ, М., 1954.
2. Bellman R., On the Boundedness, Stability and Asymptotic Behavior of Solutions of Linear and Nonlinear Differential Equations, Office of Naval Research Report, Washington, D. C., 1949.
3. Bellman R., A Survey of the Mathematical Theory of Time-lag, Retarded Control and Hereditary Processes, Rand Corporation Report R-256 (with the collaboration of J. M. Danskin, Jr.), The Rand Corporation, Santa Monica, Calif., 1954.
4. Bellman R., On the Existence and Boundedness of Solutions of Nonlinear Partial Differential Equations of Parabolic Type, Trans. Amer. Math. Soc. **64**, 21—44 (1948).
5. Brownell F. H., Nonlinear Delay Differential Equations, Contributions to the Theory of Nonlinear Oscillations, Ann. Math. Stud., № 20, 89—148, Princeton University Press, Princeton, N. J., 1950.
6. Hayes N. D., Roots of the Transcendental Equation Associated with a Certain Differential-difference Equation, J. London Math. Soc., **25**, 226—232 (1950).
7. Minorsky N., Self-excited Oscillations in Dynamical Systems Possessing Retarded Action, J. Appl. Mech., **9**, 65—71 (1942).
8. Понтрягин Л. С., О нулях некоторых трансцендентных функций, Изв. АН СССР, сер. мат. **6**, 115—134 (1942).
9. Prodi G., Teoremi di esistenza per equazioni alle derivate parziali nonlineari di tipo parabolico, Rend. Inst. Lombardo Sci., **86**, 3—47 (1953).
10. Prodi G., Soluzione periodiche di equazioni alle derivate parziali di tipo parabolico e nonlineari, Riv. Mat. Univ. Parma, **3**, 265—290 (1952).
11. Wright E. M., The Stability of Solutions of Nonlinear Differential-difference Equations, Proc. Roy. Soc. Edinburgh, Sect. A, **63**, 18—26 (1950).

Глава 3

ВНЕШНЯЯ БАЛЛИСТИКА

ДЖОН В. ГРИН

Калифорнийский университет, Лос-Анжелес

§ 3.1. Введение. При изучении движения снаряда в атмосфере специалисты по внешней баллистике сталкиваются с весьма разнообразными математическими проблемами. Задачей настоящей главы, поскольку это допускает ее объем, является систематическое изложение этих проблем, а также математического аппарата, необходимого для их решения.

Обычно принимают, что движение снаряда описывается системой обыкновенных дифференциальных уравнений, связывающих его положение, скорость и ускорение с приложенными к нему силами. Вследствие этого оказывается необходимым решать довольно громоздкую систему дифференциальных уравнений. Однако еще до начала этого этапа работы приходится решать столь же важную, а иногда и более сложную задачу составления самих дифференциальных уравнений. Сюда относится выбор системы координат, наиболее подходящей для описания движения, и разработка теории, позволяющей установить, какого рода силы действуют на снаряд. После построения такой теории обычно оказывается возможным путем эксперимента определить действительные величины приложенных сил и отсюда получить вполне определенные уравнения движения. Например, естественно считать, что главная часть аэродинамической силы, действующей на снаряд, направлена противоположно скорости и зависит от ее величины; эту зависимость можно затем определить экспериментально для различных снарядов и различных скоростей.

После того как дифференциальные уравнения получены, обычно оказывается, что даже при использовании современных быстродействующих вычислительных средств решение этих уравнений для всех возможных сочетаний физических ситуаций и начальных условий представляет собой далеко не легкую задачу. Так, если результаты зависят от нескольких десятков параметров, их трудно даже просто табулировать и представить в удобном для использования виде. Поэтому специалисты по внешней баллистике постоянно заняты вопросом о том, какие упрощения являются допустимыми, какие силы можно отбросить и какие нет. Например, до самого последнего времени в задачах баллистики оказывалось возможным пренебрегать притяжением Солнца,

Луны и других светил, что в будущем, однако, может оказаться недопустимым. Поэтому возникают вопросы, чем именно, без вреда для результата, можно пренебречь вовсе, а чем — временно, имея в виду позднее каким-либо простым способом ввести необходимые поправки. Наиболее важных результатов по внешней баллистике обычно достигают ученые, хорошо владеющие искусством аппроксимации.

В остальных параграфах настоящей главы перечисленные здесь вопросы, а также и некоторые другие будут изложены с большей подробностью.

§ 3.2. Выбор системы координат. Выбор системы координат не всегда бывает столь тривиальной задачей, как это может показаться на первый взгляд. В элементарной механике мы обычно совмещаем начало координат с некоторой удобной точкой, например с начальной точкой траектории, направляем ось z вертикально и располагаем плоскость xu в плоскости, касательной к поверхности Земли. Затем мы исходим из второго закона Ньютона и пишем, что сила равна массе, умноженной на ускорение. При этом мы пренебрегаем тем обстоятельством, что наша координатная система участвует во вращении Земли и в ее движении вокруг Солнца, надеясь на то, что такое движение координатной системы для нас не существенно. Но это движение не является равномерным, вследствие чего второй закон Ньютона здесь неприменим.

Чтобы посмотреть, какие проблемы возникают в связи с этим, рассмотрим упрощенную задачу. Представим себе Землю в виде неподвижной плоскости и на ней поезд, движущийся вдоль оси x . Предположим, что поезд движется с постоянным ускорением a , что в момент $t=0$ он находился в начале координат и скорость его была равна нулю. Пусть в момент $t=0$ вперед по полу поезда, с начальной скоростью v_0 , без трения, начинает двигаться шарик. Скорость v_0 можно считать заданной как относительно поезда, так и относительно Земли, потому что в момент $t=0$ скорость поезда равна нулю.

Теперь сразу приходят на ум две системы координат: одна, связанная с Землей, другая — с поездом. Если O' означает ту точку поезда, из которой начал двигаться шарик, то O' имеет ускорение a и $OO' = (1/2)at^2$. Поэтому координаты шарика относительно Земли x и относительно поезда x' связаны уравнением

$$x = x' + \frac{at^2}{2}.$$

Так как в направлении оси x на шарик не действуют никакие силы, уравнение его движения имеет вид

$$m\ddot{x} = 0. \quad (3.1)$$

Подставив $\ddot{x} = \ddot{x}' + a$ в уравнение (3.1), получим

$$m\ddot{x}' = -ma. \quad (3.2)$$

Однако наблюдатель, находящийся в поезде, который не знает об его ускоренном движении или пренебрегает этим движением, будет считать, что истинное уравнение движения есть

$$m\ddot{x}' = 0,$$

вследствие чего он не сможет правильно предсказать поведение шарика. В частности, он будет чрезвычайно удивлен, когда шарик начнет двигаться обратно!

В большинстве случаев оказывается чрезвычайно неудобным пользоваться системами координат, движущимися без ускорения, в частности, очень неудобно и нежелательно искать выражения для аэродинамических сил, действующих на снаряд, в виде функций скорости, заданной относительно неподвижной системы координат, потому что в действительности эти силы зависят от скорости, отнесенной к системе координат, вращающейся вместе с Землей. Чаще всего пользуются удобными системами координат независимо от того, как они движутся, и изменяют законы движения таким образом, чтобы учесть ускорение системы координат. Заметим, что в предыдущем примере мы получим правильное уравнение движения, если введем фиктивную силу — ma и не будем обращать внимания на ускорение точки O' .

Этот прием является типичным; необходимая модификация законов движения может быть получена путем введения фиктивных сил, которые вычисляются из соотношений, связывающих неподвижную и движущуюся координатные системы. Для вращающихся координатных систем почти обязательным является применение векторных обозначений, и в связи с этой проблемой, а также в связи с проблемой описания движения снаряда как твердого тела исследователи в области баллистики должны уметь пользоваться векторной алгеброй.

Вероятно, каждый, изучавший курс векторного анализа, выводил уравнения движения частицы относительно системы координат, связанной с Землей; поэтому мы рассмотрим здесь именно такой пример. Будем считать Землю (пренебрегая ее движением вокруг Солнца) вращающейся вокруг неподвижной оси с постоянной угловой скоростью ω . (Векторы мы будем обозначать жирными буквами, например, ω . Под ω мы будем понимать модуль или длину вектора ω .) Вектор ω направлен от южного полюса к северному и имеет модуль, равный приблизительно $0,000073$ *радиан/сек*. Под вектором $x = x(t)$ мы будем понимать вектор, определяющий положение рассматриваемой материальной точки относительно системы координат, связанной с Землей. Начало этой системы мы расположим на поверхности Земли, ось x_3 направим вертикально вверх, ось x_1 — к востоку, ось x_2 — к северу. Если r есть вектор, определяющий положение начала координат относительно центра Земли, то $r + x$ определяет

положение рассматриваемой материальной точки относительно центра Земли.

Пусть F обозначает равнодействующую всех гравитационных аэродинамических и прочих сил, действующих на материальную точку. Тогда уравнение движения имеет вид

$$m\ddot{\mathbf{x}} = F + m \{ \omega^2 (\mathbf{r} + \mathbf{x}) - [\omega \cdot (\mathbf{r} + \mathbf{x})] \omega \} - 2m\omega \times \dot{\mathbf{x}}.$$

Таким образом, мы видим, что для того, чтобы Закон Ньютона „сила равна произведению массы на ускорение“ удовлетворялся, по крайней мере кажущимся образом, в движущейся системе координат необходимо ввести две фиктивные силы. Первая из них,

$$m \{ \omega^2 (\mathbf{r} + \mathbf{x}) - [\omega \cdot (\mathbf{r} + \mathbf{x})] \omega \},$$

зависит только от материальной точки, а не от ее скорости. Нетрудно видеть, что она направлена от оси Земли и имеет величину $m\omega^2 d$, где d — расстояние материальной точки от оси Земли. Это хорошо известная *центробежная сила*; для движений, происходящих в пределах малых участков земной поверхности, эта сила очень близка к постоянной, но не является пренебрежимо малой. Именно центробежная сила вместе с силой притяжения и определяют (при помощи отвеса) направление „вертикали“ и величину g , ускорение силы „тяжести“.

Вторая фиктивная сила, $-2m\omega \times \dot{\mathbf{x}}$, известна как *сила Кориолиса*; она появляется только в том случае, когда материальная точка движется относительно Земли. Эта сила обычно мала, но заметно изменяется вдоль каждой отдельной траектории. Для скоростей, не превосходящих 1000 *фут/сек*, производимое ею ускорение не превосходит 0,15 *фут/сек*². Ниже мы приближенно вычислим влияние Кориолисовой силы для простого случая. Мы сделаем это не столько потому, что нам очень часто придется уделять серьезное внимание эффекту Кориолиса, сколько для иллюстрации простого и удобного метода, которым следует пользоваться в подобных случаях.

Предположим, что мы бросаем материальную точку в пустоте с поверхности Земли вертикально вверх, придавая точке начальную скорость v_0 . Опыт учит, а грубая оценка, указанная выше, убеждает нас в том, что движение остается очень близким к оси x_3 и что $\dot{\mathbf{x}}$ имеет составляющие $(0, 0, \dot{x}_3)$. Тогда приближенные значения составляющих силы Кориолиса будут $(-2m\omega\dot{x}_3 \sin \lambda, 0, 0)$, где λ есть полярный угол точки O , т. е. угол, измеряемый дугой меридиана от северного полюса до точки O . Пренебрегая центробежной силой, мы получим уравнения движения в виде

$$\ddot{x}_1 = -2\omega\dot{x}_3 \sin \lambda, \quad \ddot{x}_2 = 0, \quad \ddot{x}_3 = -g. \quad (3.3)$$

После двукратной интеграции первого уравнения (3.3) получим

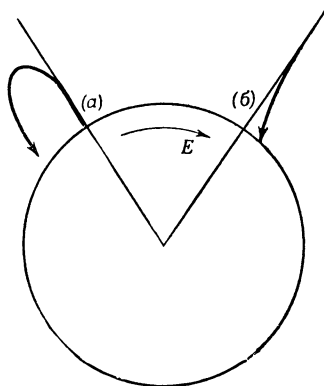
$$x_1 = -2\omega \sin \lambda \int_0^t x_3 dt. \quad (3.4)$$

Так как координата x_3 положительна между моментом выстрела и моментом падения на Землю, то из формулы (3.4) мы видим, что материальная точка упадет на поверхность Земли прямо к западу от места выстрела. Подставляя в (3.4) величину $x_3 = v_0 t - (1/2)gt^2$, полученную из третьего уравнения (3.3), мы найдем, что отклонение к западу в точке падения будет

$$\frac{4v_0^3 \omega \sin \lambda}{3g^2},$$

что составляет около 47 футов при $v_0 = 1000$ фум/сек и $\lambda = 30^\circ$. Если материальная точка, вместо того чтобы быть брошенной вертикально вверх, падает с некоторой высоты, она будет отклоняться к востоку, потому что координата x_3 вдоль траектории всюду отрицательна. На рис. 3.1 показаны отклонения в этих двух случаях. Для ясности эти отклонения увеличены.

Рис. 3.1. Эффект Кориолиса (увеличено) для тела, брошенного вертикально вверх (а) или падающего с некоторой высоты (б).



Простой и удобный метод решения подобных задач состоит в следующем. Пусть в дифференциальных уравнениях, подлежащих решению, содержатся силы в виде членов, которые затрудняют решение.

Пока уравнения не будут решены, точные значения этих сил остаются неизвестными, так как они зависят от неизвестной пока траектории. Однако поскольку силы малы, их влияние на траекторию незначительно; поэтому величину сил можно вычислить без большой ошибки при помощи траектории, полученной в результате пренебрежения малыми силами. Затем, используя вычисленные значения малых сил, мы снова можем определить траекторию и таким образом оценить, насколько изменяется траектория при учете малых сил. Такой процесс при многократном его повторении (что в вычислительном отношении, как правило, не является слишком простым) не очень сильно отличается по своей идее от решения дифференциальных уравнений методом последовательных приближений.

Точность оценок, получаемых подобным путем, например таких, как приведенные выше 47 футов, определить затруднительно. Было бы весьма желательно получить, например, математически строгое доказательство того, что в нашем примере отклонение материальной

точки по широте не превосходит некоторой определенной величины ϵ . Результаты такого рода иногда возможно получить, но обычно бывает так, что величина ϵ оказывается совершенно непригодной для использования. Например, мы, вероятно, смогли бы установить, что отклонение к западу не превосходит 500 футов, однако такой результат нам не особенно нужен. Оценки, получаемые указанным выше способом, часто оказываются лучше, чем те, к которым мы можем прийти путем строгого доказательства. Однако им нельзя слишком доверять без тщательной проверки при помощи эксперимента, безразлично, вычислительного или физического.

Грубые вычисления указанного выше типа и эксперименты показали, что для снарядов, рассчитанных на умеренные дальности, например в пределах ста миль, систему отсчета, связанную с Землей, можно рассматривать как инерциальную; при этом в худшем случае потребуются лишь малые поправки. Это уже не имеет места для снарядов будущего, обладающих, возможно, межконтинентальной дальностью. Еще более трудная ситуация возникает в связи со стрельбой с быстро движущегося самолета. Поскольку в этом случае скорости и ускорения самолета и снаряда имеют одинаковый порядок, пренебрежение силами инерции может привести к очень большим ошибкам. Действительно, нетрудно представить себе обстановку, в которой эти силы настолько велики, что их влияние уже не может быть учтено в виде простой поправки, как это было в приведенном выше примере.

§ 3.3. Аэродинамические силы, действующие на снаряд. Аэродинамические силы, действующие на тело, движущееся в воздухе, определяются распределением давления на поверхности тела. Распределение давления в свою очередь зависит от термодинамического состояния воздуха, а также от скорости и ориентации тела. Теоретически определить эти силы, исходя из формы тела и закона аэродинамики, очень трудно, если не невозможно, по крайней мере в настоящее время, поскольку сюда входит решение уравнений в частных производных с чрезвычайно сложными граничными условиями. Упрощающие предположения, которые иногда вводят, чтобы сделать задачу доступной математическому анализу, приводят к результатам, мало или совсем не совпадающим с опытом. Например, считая воздух идеальной несжимаемой жидкостью, приходят к выводу, что шар, движущийся в воздухе, не испытывает никакого сопротивления. Таким образом, появляется необходимость определять силы из опыта. Так как невозможно произвести измерения сил для всех желаемых физических условий, то необходимо построить обобщенную теорию того, как эти силы зависят от скорости, ориентации и т. д. изучаемого снаряда, с тем чтобы из измерения сил при каких-нибудь одних условиях получить их значения при всех прочих условиях.

Гипотеза, по своей простоте непосредственно следующая за пренебрежением аэродинамическими силами вообще, состоит, по-видимому, в том, чтобы, оставаясь в пределах динамики твердого тела, считать снаряд материальной точкой, к которой приложена единственная аэродинамическая сила, направленная по касательной к траектории; это означает, что за единственную аэродинамическую силу принимается лобовое сопротивление. Для некоторого определенного снаряда эта сила может быть измерена при различных скоростях в аэродинамической трубе или другим способом. При помощи интерполяции мы отсюда можем получить оценку силы сопротивления для других скоростей; таким образом, в пределах принятой гипотезы можно составить уравнения движения.

Этот способ особенно широко применяется в баллистике и приводит к существенным результатам. Большинство траекторий, на основе которых построены таблицы стрельбы и бомбометания, получены при помощи этого способа; мы к нему еще возвратимся в дальнейшем. Однако для многих снарядов подобная теория оказывается далеко недостаточной и при ее помощи нельзя объяснить многие явления. Например, было обнаружено, что некоторые бомбы в воздухе летят дальше, чем в пустоте. Это невозможно объяснить при помощи гипотезы о наличии только одного сопротивления, если не сделать невероятного предположения об отрицательном сопротивлении.

Другое соображение, говорящее за необходимость более полной теории, заключается в том, что нам нужно более подробно изучить движение снаряда с целью определить, будет ли снаряд послушным или он будет кувыркаться; это означает, что речь идет об устойчивости движения снаряда. Например, обычный артиллерийский снаряд, если он не вращается, будет кувыркаться и описывать неправильную траекторию. Если дать ему чрезмерное вращение, его ось не будет следить за траекторией, а останется приблизительно параллельной оси канала ствола; вследствие этого он будет лететь боком и, возможно, даже ударится о цель не носом, а тыльной частью. Вращательное движение снаряда должно быть подобрано таким образом, чтобы оно совместно с аэродинамическими силами заставляло ось снаряда оставаться вблизи касательной к траектории. Чтобы построить подобную теорию, необходимо изучить не только одну силу сопротивления, но и прочие аэродинамические силы.

Здесь мы сталкиваемся с той неприятностью, что снаряд не есть материальная точка и различные его части движутся по различным траекториям и с различными скоростями. Вследствие этого возникает вопрос о том, что именно следует понимать под скоростью снаряда вообще?

С целью показать, каков может быть менее наивный подход к решению задачи, рассмотрим снаряд, представляющий собой тело вращения относительно некоторой оси и движущийся таким образом, что его ось все время остается в некоторой плоскости. Приблизительно

так движется сброшенная бомба. Мы обойдем усложнения, вызываемые собственным вращением снаряда, если предположим, что это вращение отсутствует. Пусть O будет начало осей координат, связанных с снарядом, $\mathbf{v}$ — скорость точки O и ω — угловая скорость снаряда, направленная перпендикулярно к плоскости движения. Система аэродинамических сил, действующая на снаряд, в соответствии с основными законами механики может быть приведена к главному вектору $\mathbf{F}$, проходящему через точку O , и к главному моменту $\mathbf{M}$. Из соображений симметрии заключаем, что вектор $\mathbf{F}$ лежит в плоскости движения, а $\mathbf{M}$ — перпендикулярен к ней. Рис. 3.2 иллюстрирует эти обстоятельства.

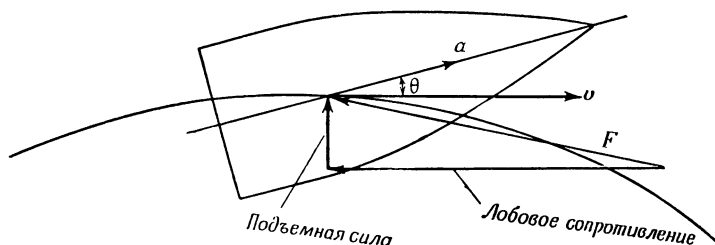


Рис. 3.2. Упрощенная система аэродинамических сил для плоского движения.

Составляющие вектора $\mathbf{F}$ вдоль $\mathbf{v}$ и перпендикулярно к ней называются *лобовым сопротивлением* и *подъемной силой* соответственно. Момент $\mathbf{M}$ будет восстанавливающим или опрокидывающим в зависимости от своего знака — опрокидывающим для большинства артиллерийских снарядов и восстанавливающим для бомб. Простейшую теорию получим, предполагая, что векторы $\mathbf{F}$ и $\mathbf{M}$ зависят только от скорости $\mathbf{v}$ и угла θ между $\mathbf{v}$ и осью снаряда, так называемого *угла атаки*. Если мы обозначим через $\mathbf{a}$ единичный вектор оси снаряда, это предположение означает, что

$$\mathbf{F} = \mathbf{F}(\mathbf{v}, \mathbf{a}), \quad \mathbf{M} = \mathbf{M}(\mathbf{v}, \mathbf{a}). \quad (3.5)$$

Таким образом, мы предположили, что система сил, действующих на снаряд, не зависит от его угловой скорости ω . На самом деле это не имеет места, и можно ожидать, что наша теория будет неполной; однако, как это ни удивительно, мы покажем, что она даже противоречива.

Предположим, что центр приведения O перемещен в O' , причем $OO' = \mathbf{r}$. Тогда скорость $\mathbf{v}'$ точки O' будет равна $\mathbf{v} + \mathbf{r} \times \omega$. Отсюда, следуя нашей гипотезе, получим, что система сил приведет к главному вектору $\mathbf{F}^*(\mathbf{v} + \mathbf{r} \times \omega, \mathbf{a})$, приложенному в O' , и к главному моменту $\mathbf{M}^*(\mathbf{v} + \mathbf{r} \times \omega, \mathbf{a})$. Здесь функции $\mathbf{F}^*$ и $\mathbf{M}^*$ не обязаны совпадать с функциями $\mathbf{F}$ и $\mathbf{M}$. Но главный вектор не зависит от

центра приведения, а главный момент изменяется на величину $\mathbf{r} \times \mathbf{F}$; поэтому

$$\begin{aligned} \mathbf{F}(\mathbf{v}, \mathbf{a}) &= \mathbf{F}^*(\mathbf{v} + \mathbf{r} \times \boldsymbol{\omega}, \mathbf{a}) \\ \mathbf{M}(\mathbf{v}, \mathbf{a}) &= \mathbf{M}^*(\mathbf{v} + \mathbf{r} \times \boldsymbol{\omega}, \mathbf{a}) + \mathbf{r} \times \mathbf{F}(\mathbf{v}, \mathbf{a}). \end{aligned} \quad (3.6)$$

Будем рассматривать $\mathbf{a}$ как неизменный вектор; тогда соотношения (3.6) должны быть тождествами относительно $\mathbf{v}$, $\mathbf{r}$ и $\boldsymbol{\omega}$.

Пусть $\mathbf{v}_1$ и $\mathbf{v}_2$ будут такие две скорости, что $\mathbf{v}_1 \cdot \mathbf{a} = \mathbf{v}_2 \cdot \mathbf{a}$. Тогда, очевидно, должно быть $\mathbf{v}_2 = \mathbf{v}_1 + \mathbf{n}$, где вектор $\mathbf{n}$ перпендикулярен к $\mathbf{a}$ и поэтому может быть представлен в форме $\mathbf{r} \times \boldsymbol{\omega}$, при соответствующем $\boldsymbol{\omega}$. Поэтому можно написать $\mathbf{v}_2 = \mathbf{v}_1 + \mathbf{r} \times \boldsymbol{\omega}$. Но из первого равенства (3.6) имеем:

$$\mathbf{F}(\mathbf{v}_1, \mathbf{a}) = \mathbf{F}^*(\mathbf{v}_1 + \mathbf{r} \times \boldsymbol{\omega}, \mathbf{a}) = \mathbf{F}^*(\mathbf{v}_2, \mathbf{a}) = \mathbf{F}(\mathbf{v}_2, \mathbf{a}).$$

Таким образом, мы видим, что $\mathbf{F}$ есть функция только $\mathbf{v} \cdot \mathbf{a}$. Из соображений симметрии нашего снаряда ясно, что составляющая $\mathbf{F}$, перпендикулярная к $\mathbf{a}$, изменяет свой знак вместе с углом атаки. Однако изменение знака угла атаки не влияет на $\mathbf{v} \cdot \mathbf{a}$; поэтому составляющая $\mathbf{F}$ по перпендикуляру к $\mathbf{a}$ просто отсутствует и $\mathbf{F}$ есть сила, направленная вдоль оси снаряда, или, как мы ее будем называть в дальнейшем, „осевое сопротивление“. Подобным же образом мы можем усмотреть, что $\mathbf{M}$ есть также функция только от $\mathbf{v} \cdot \mathbf{a}$; так как она должна изменять знак, если $\boldsymbol{\omega}$ и угол атаки меняют знак, то она должна быть равна нулю.

Мы видим, таким образом, что рассмотренная система сил невозможна, за исключением того случая, когда она состоит только из одной силы — осевого сопротивления; но и этот случай мы должны будем отбросить на основании рассуждения от противного, если посмотрим, что случится, когда угол атаки будет равен 90° . Следовательно, необходима новая гипотеза. Естественно предположить, что силы и моменты зависят не только от $\mathbf{v}$ и $\mathbf{a}$, но и от $\boldsymbol{\omega}$, т. е. что

$$\mathbf{F} = \mathbf{F}(\mathbf{v}, \boldsymbol{\omega}, \mathbf{a}), \quad \mathbf{M} = \mathbf{M}(\mathbf{v}, \boldsymbol{\omega}, \mathbf{a}).$$

Эти функции, конечно, должны удовлетворять соотношениям (3.6), чтобы система сил не зависела от выбора центра приведения. Нашей задачей является подобрать для этого наиболее простые функции, не впадая в противоречие с физическим смыслом. Последующую теорию развили Нилсен и Синг [5]; она основана на соображениях симметрии и на том принципе, что любая величина, которую мы имеем основания считать линейной, может и на самом деле рассматриваться как линейная, если эксперимент этому не противоречит.

Разложим $\mathbf{v}$ на две составляющие: $\mathbf{v}_1$, направленную по $\mathbf{a}$, и $\mathbf{v}_2$, перпендикулярную к $\mathbf{a}$; подобным же образом разложим $\mathbf{F}$ на $\mathbf{F}_1$ и $\mathbf{F}_2$. Так как $\mathbf{M}$ и $\boldsymbol{\omega}$ перпендикулярны к плоскости движения, то их отличные от нуля составляющие суть M_3 и ω_3 . Если снаряд хорошо спроектирован и выстрел произведен правильно, то можно ожидать,

что угол атаки будет оставаться малым вдоль всей траектории и его колебания относительно некоторого среднего значения будут только медленными. Поэтому целесообразно рассмотреть случай, когда угол атаки θ и вместе с ним составляющая v_2 малы ($\operatorname{tg} \theta = v_2/v_1$); подобным же образом мы считаем малой и ω_3 . Будем считать средние значения v_2 и ω_3 равными нулю и разложим составляющие силы и моменты в ряд Тэйлора по степеням v_2 и ω_3 ; сохранив в разложении только линейные члены, получим

$$\begin{aligned} F_1 &= F_1(v_1, v_2, \omega_3) = A_1 + B_1 v_2 + C_1 \omega_3, \\ F_2 &= F_2(v_1, v_2, \omega_3) = A_2 + B_2 v_2 + C_2 \omega_3, \\ M_3 &= M_3(v_1, v_2, \omega_3) = \alpha + \beta v_2 + \gamma \omega_3, \end{aligned}$$

причем коэффициенты этих разложений зависят от v_1 . Из соображений симметрии мы сразу видим, что B_1 , C_1 , A_2 и α равны нулю. Например, следует ожидать, что F_2 равно нулю, когда угол атаки и угловая скорость равны нулю; отсюда $A_2 = 0$. Таким образом, получаем возможные выражения для сил и момента, действующих на наш снаряд:

$$\begin{aligned} F_1 &= A_1, \\ F_2 &= B_2 v_2 + C_2 \omega_3, \\ M_3 &= \beta v_2 + \gamma \omega_3. \end{aligned} \tag{3.7}$$

Легко представить себе смысл отдельных членов в выражениях (3.7). Член $A_1 = A_1(v_1)$, который представляет собой всю силу F_1 , есть осевое сопротивление, т. е. при обычных условиях главная часть аэродинамической силы. Член $B_2 v_2$ представляет собой поперечную силу, вызванную поперечной скоростью, а $C_2 \omega_3$ — поперечную силу, вызванную поперечным вращением (мы пользуемся здесь термином „поперечное вращение“, потому что часто имеется вращение относительно продольной оси снаряда со значительно большей угловой скоростью). Член βv_2 есть поперечный момент, вызываемый поперечной скоростью; это — упомянутый выше восстанавливающий момент. Член $\gamma \omega_3$ есть поперечный момент, вызываемый поперечным вращением, или гасящий (демпфирующий) момент. Введенное выше лобовое сопротивление и подъемная сила здесь не участвуют вовсе просто потому, что мы разложили силу по оси снаряда и по перпендикуляру к ней вместо того, чтобы выполнить это разложение относительно траектории. Обычное лобовое сопротивление получается в качестве составляющей по направлению v и нетрудно видеть, что оно совпадает с осевым сопротивлением, т. е. с составляющей сопротивления вдоль продольной оси снаряда, с точностью до членов второго порядка относительно v_2 и ω_3 ; поэтому для определения траектории мы обычно принимаем лобовое сопротивление равным A_1 . Точно так же, с точностью до величины второго порядка малости, $v = v_1$, и поэтому в пределах линейного приближения мы можем принять лобовое сопротивление равным $A_1(v)$.

Описание системы сил с помощью соотношений (3.7) было получено путем отбрасывания в разложении Тэйлора членов порядка выше первого. Эти члены высших порядков в действительности могут присутствовать (и в некоторых случаях это в самом деле имеет место), вследствие чего равенства (3.7) можно рассматривать только как приближенное представление системы сил, являющееся хорошим только в случае малого и медленно изменяющегося угла атаки. Тем не менее нетрудно видеть, что такое описание системы сил согласуется с соотношениями (3.6); все что для этого необходимо сделать — это подчинить коэффициенты в (3.7) простым законам преобразования при переносе центра приведения. Таким образом, описание системы сил при помощи соотношений (3.7) во многих случаях может считаться вполне удовлетворительным.

В предыдущих рассуждениях мы исходили из предположения, что движение — плоское; это возможно только в случае невращающегося снаряда. Если применить тот же прием к вращающемуся снаряду, мы найдем пять дополнительных сил и моментов, зависящих от ω_1 . Эти силы и моменты известны под названием *эффекта Магнуса*; они в некотором смысле аналогичны силе, искривляющей полет вращающегося мяча при игре в гольф или бейзбол. Обычно эти силы и моменты малы; однако их введение необходимо для объяснения некоторых явлений, таких, как деривация вращающегося снаряда.

Мы рассматривали различные коэффициенты A_1 , B_2 и т. д. как зависящие только от v_1 , но, кроме того, они зависят от физического состояния атмосферы, через которую снаряд пролетает, от формы снаряда и от его калибра. Различные физические соображения и соображения, связанные с теорией размерности, для изложения которых здесь недостаточно места, приводят к появлению в формулах некоторых безразмерных аэродинамических коэффициентов, выражающих эту зависимость. Главный из них K_{DA} , *коэффициент осевого сопротивления*, определяется соотношением

$$F_1 = A_1 = \rho d^2 v_1^2 K_{DA},$$

где ρ есть плотность воздуха, d — диаметр снаряда. Так как A_1 есть сила, имеющая размерность (масса $\times$ длина)/время², а произведение $\rho d^2 v_1^2$ имеет ту же самую размерность, то K_{DA} есть величина безразмерная; это означает, что она будет иметь одно и то же значение в любой системе единиц¹⁾. K_{DA} может зависеть от многих безраз-

¹⁾ Баллистика всегда требует значительного количества утомительных вычислений. В результате попыток уменьшить объем вычислительных работ путем использования табулированных результатов опытов, проведенных другими учеными в другое время и в других местах, специалисты по внешней баллистике иногда применяют в пределах одной и той же задачи различные и очень странные системы единиц. Например, автор этой статьи видел расчет траектории, в котором дальность измерялась в футах, таблицы силы сопротивления были отнесены к метрам в секунду, диаметр снаряда измерялся в дюймах, а результат вычислений был приведен в ярдах.

мерных параметров, но оказалось, что этот коэффициент зависит главным образом от формы снаряда и отношения v_1/a , где a — местная скорость звука в воздухе, пропорциональная квадратному корню из абсолютной температуры. Отношение v_1/a называется *числом Маха*, соответствующим скорости v_1 . Поэтому для снаряда определенных размеров и формы мы можем написать $K_{DA} = K_{DA}(v_1/a)$. В действительности коэффициент K_{DA} лишь весьма слабо зависит от размеров снаряда; эту зависимость можно принять во внимание при помощи множителя d^2 . K_{DA} не очень сильно зависит даже от формы снаряда и поэтому грубое приближение для осевого сопротивления можно получить, используя любой из известных коэффициентов K_{DA} .

Как указывалось выше, мы можем приближенно принять, что осевое сопротивление равно полному лобовому сопротивлению, или $K_D = K_{DA}$, причем

$$\text{Лобовое сопротивление} = \rho d^2 v^2 K_D.$$

Безразмерное число K_D называется *коэффициентом лобового сопротивления*. Типичная зависимость K_D от числа Маха приведена на рис. 3.3.

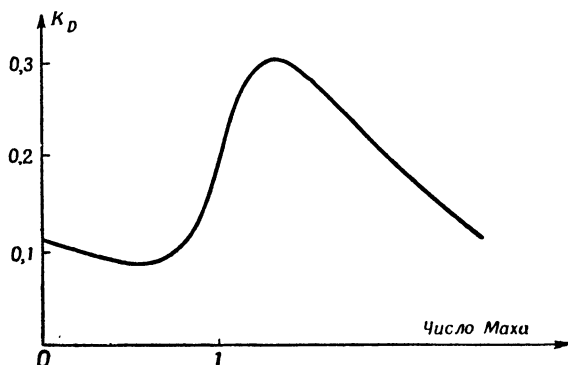


Рис. 3.3. Типичная зависимость коэффициента лобового сопротивления K от числа Маха.

Остальные аэродинамические силы также могут быть выражены при помощи ρ , d , v_1 и соответствующих безразмерных коэффициентов. Например, приведенная выше поперечная сила $B_2 v_2$ обычно записывается в виде

$$\rho d^2 v_1 K,$$

где $K = K(v_1/a)$ — безразмерная величина. Подъемная сила, или сила, нормальная к траектории, может быть получена из составляющих поперечной силы и осевого сопротивления по нормали к траектории.

В результате, положив $\cos \theta = 1$, $\sin \theta = v_2/v_1$, получим в качестве первого приближения

$$\begin{aligned}\text{Подъемная сила} &= \rho d^2 v_1^2 K_N - \rho d^2 v_1^2 K_{DA} \frac{v_2}{v_1} = \\ &= \rho d^2 v_1^2 \left(K_N - K_{DA} \cdot \frac{v_2}{v_1} \right).\end{aligned}$$

В этом приближении коэффициент подъемной силы K_L определяется выражением

$$K_L = K_N - K_{DA} \cdot \frac{v_2}{v_1}.$$

§ 3.4. Уравнения движения. Предположим, что нам известны введенные в предыдущем параграфе аэродинамические коэффициенты, а также плотность ρ , абсолютная температура T (и, следовательно, скорость звука) и скорость ветра в той части атмосферы, в которой, как мы предполагаем, будет пролетать снаряд. Тогда составление уравнений движения в осях, связанных с Землей, становится совсем простым делом. В наиболее общей форме мы получаем два векторных или шесть скалярных уравнений: три для движения центра инерции и три для вращательного движения. Нам не нужно выписывать их все, потому что, с одной стороны, они весьма сложны, а с другой стороны, их решение редко бывает необходимым. Однако, если для нас важно иметь траекторию снаряда, то ее, вообще говоря, всегда можно определить приближенно, с любой степенью точности, так как решение уравнений движения не лежит за пределами возможностей современных счетных машин.

Для составления таблицы стрельбы оказывается достаточным сильно упрощенный вариант. Он получается, если пренебречь всеми аэродинамическими силами, кроме лобового сопротивления, и предположить, что ветер отсутствует, а температура и плотность воздуха соответствуют нормальной атмосфере. Если начало координат расположено на обрезе ствола орудия и на уровне моря, причем ось x расположена горизонтально в направлении огня, а ось y вертикальна, то нормальная атмосфера определяется равенствами

$$\rho = \rho_0 e^{-hy}, \quad T = T_0 e^{-ay},$$

что соответствует опыту и является очевидным из теоретических соображений. Здесь ρ_0 , h , T_0 и a суть некоторые определенные постоянные, которые могут быть найдены в соответствующих трудах по внешней баллистике.

Предполагая теперь, что коэффициент K_D для рассматриваемого снаряда известен, мы получаем уравнения движения центра инерции в следующем виде:

$$m\ddot{x} = -\frac{D\dot{x}}{v}, \quad m\ddot{y} = -mg - \frac{D\dot{y}}{v}, \quad (3.8)$$

где $v^2 = \dot{x}^2 + \dot{y}^2$, $D = \rho v^2 d^2 K_D (v/a)$, и $a = (\text{const}) T^{1/2}$. Если начальные значения x и y заданы, то система (3.8) может быть решена (и была решена в действительности) более или менее точно при помощи самых разнообразных методов. Уравнения (3.8) часто называют *нормальными уравнениями движения*.

В прошлое время большая часть любого рассуждения о проблемах баллистики относилась к методам решения этих уравнений, потому что до появления современных быстродействующих счетных машин численное решение дифференциальных уравнений типа (3.8) было одной из основных, а возможно, и самой главной проблемой в баллистике; численное решение было медленным, утомительным и дорогим. В настоящее время такое решение, по-видимому, уже не является главной задачей. Методы численного решения представляют собой специальную ветвь математического анализа, применимую, кроме баллистики, к множеству различных отраслей науки. У нас нет особых оснований вдаваться здесь в эти вопросы, в частности потому, что их подробное изложение вывело бы нас далеко из рамок настоящей главы (см. часть третью этой книги).

Конечно, имеется много интересных и остроумных специально баллистических приближенных методов, таких, как методы Сначчи или метод Хитчкока — Кента и т. п., но объем настоящей главы не позволяет привести здесь их подробное изложение. Нечего и говорить, что человек, практически работающий в области баллистики, должен быть специалистом по численным решениям или по крайней мере иметь возможность пользоваться услугами такого специалиста.

В § 3.8 мы рассмотрим возможность решения уравнений движения в том случае, когда кроме силы сопротивления действуют другие аэродинамические силы, а также когда плотность воздуха отличается от плотности, соответствующей нормальной атмосфере, и т. п.

§ 3.5. Баллистические таблицы и таблицы стрельбы. Предполагая снова, что мы имеем снаряд, для которого K_D известно, мы можем найти решение уравнений (3.8) для большого числа различных начальных значений x и y , или, что то же самое, для различных значений начальной скорости снаряда и угла возвышения орудия. Результаты таких вычислений можно затем представить в виде удобных таблиц. Например, для артиллерийских снарядов особенно важна дальность, равная расстоянию между начальной точкой и точкой падения, и полное время полета снаряда. Эти величины табулируются по начальному углу возвышения, а если предполагается использовать переменный заряд, то и по начальной скорости. Полученные таблицы можно назвать таблицами стрельбы (в их простейшей форме); они представляют большую ценность для артиллерийского офицера, которому необходимо поразить определенный объект и правильно расположить для этого свои орудия. Если наша теория правильна, если

мы располагаем точным значением коэффициента лобового сопротивления K_D и если прочими аэродинамическими коэффициентами действительно можно пренебречь, то мы должны были бы получить хорошие таблицы стрельбы, по крайней мере для внешних условий, приближенно соответствующих нормальной атмосфере.

К сожалению, мы не можем особенно доверять нашей теории; мы редко знаем заранее коэффициент K_D , а остальными аэродинамическими силами, вероятно, нельзя пренебрегать. Поэтому изложенный выше упрощенный способ не может считаться удовлетворительным и должен быть заменен методом, принимающим во внимание результаты измерений при действительной стрельбе. В действительности поступают примерно так, как изложено ниже.

Примем коэффициент K_D в виде некоторой определенной функции ¹⁾, называемой функцией сопротивления. Перепишем уравнения (3.8) следующим образом:

$$\ddot{x} = -\rho \frac{vxK_D}{C}, \quad \ddot{y} = -g - \rho \frac{vyK_D}{C}. \quad (3.9)$$

Здесь мы ввели обозначение C для величины m/d^2 , называемой *баллистическим коэффициентом* снаряда. Теперь уравнения (3.9) можно решить для различных начальных условий и различных значений C , а результаты свести в таблицы. Полученные таким образом таблицы называются баллистическими таблицами, относящимися к тому конкретному значению K_D , которое было принято при расчетах.

Рассмотрим теперь некоторый снаряд, для которого коэффициент K_D неизвестен. Пусть мы стреляем этим снарядом при различных углах возвышения и различных начальных скоростях и измеряем дальность и полное время полета.

Если его коэффициент K'_D совпадает с принятым выше K_D , то наблюдаемые и вычисленные величины должны находиться в согласии. Предположим однако, что вследствие разницы в форме, размерах и т. п. K_D и K'_D не одинаковы, но $K'_D = kK_D$, где k — постоянная. В этом случае наблюдаемая дальность и полное время полета согласуются с баллистическими таблицами, соответствующими баллистическому коэффициенту m/kd^2 , а не m/d^2 . Другими словами, правильный результат получается и при неправильном законе сопротивления, если использовать фиктивный баллистический коэффициент. Потому мы называем величину m/kd^2 баллистическим коэффициентом снаряда, отнесенным к K_D . Баллистический коэффициент любого

¹⁾ В действительности большая часть баллистических расчетов до сих пор производится в предположении, что K_D определяется так называемой гаврской функцией сопротивления, которая была принята гаврской комиссией: во Фран. и в девятнадцатом веке. Она представляет собой нечто вроде смеси различных K_D , основанной на средних для различных снарядов.

снаряда, отнесенный к его собственному верному коэффициенту сопротивления, конечно, равен m/d^2 .

Теперь предположим, что мы не знаем коэффициента K'_D , но нам известно, что он имеет вид kK_D , где k — неизвестная постоянная. Все, что мы можем в таком случае сделать, это произвести выстрел при определенном угле возвышения и начальной скорости, после чего измерить дальность.

Имея эти три величины, мы можем обратиться к баллистическим таблицам и найти, если нужно, при помощи интерполяции величину C , которой они соответствуют. Это значение C есть баллистический коэффициент нашего снаряда, отнесенный к K_D . Используя найденное значение C , мы можем вычислить дальность для какого угодно числа углов возвышения и начальных скоростей и таким путем построить таблицы стрельбы.

Рассмотрим теперь такое положение дел, каким оно обычно бывает в действительности, а именно, когда нам совершенно ничего не известно относительно коэффициента сопротивления K'_D снаряда. Однако опыт показывает, что если снаряд не слишком сильно отличается от того, который положен в основу баллистических таблиц, соотношение $K'_D/K_D = \text{const}$ не слишком далеко от истины. Мы стреляем нашим снарядом (предполагая для простоты начальную скорость неизменной) под различными углами возвышения θ_i и измеряем дальности R_i и полные времена полета T_i . Входя в баллистические таблицы при помощи какой-нибудь пары θ_i, R_i , мы найдем, как и выше, соответствующую величину C_i . Неприятность состоит в том, что для разных значений i баллистический коэффициент C_i не один и тот же. Однако обычно оказывается, что C_i мало изменяется при переходе от одной пары начальных условий к другой. Поэтому мы можем проинтерполировать величины C_i при помощи плавной кривой; в результате мы получим C , например, как функцию дальности. Чтобы найти угол возвышения θ , под которым следует произвести выстрел, обеспечивающий достижение некоторой дальности R , мы берем $C(R)$ с графика и по нему в баллистических таблицах находим θ . Если вместо дальности использовано полное время полета, то можно ожидать больших различий в значениях C , но процесс решения задачи остается таким же.

Такой метод получения таблиц стрельбы из баллистических таблиц с неверным коэффициентом K_D в известной степени предохраняет нас от ошибки, обусловленной неполнотой теории. Если остальные аэродинамические силы существенно влияют на траекторию, то C будет изменяться вместе с начальными условиями даже в том случае, когда мы пользуемся верной величиной K_D . Однако чаще всего баллистический коэффициент изменяется лишь незначительно и изложенная выше схема интерполяции дает хороший результат. Но

иногда C будет изменяться столь быстро, что весь метод оказывается не заслуживающим доверия. Это относится, в частности, к крутым траекториям зенитных снарядов, когда у вершины можно ожидать большого угла атаки и к некоторым траекториям бомб, когда измеренная дальность указывает на отрицательный баллистический коэффициент, а время полета — на очень малый. В таких случаях может оказаться необходимым пересмотреть всю теорию и принять во внимание остальные аэродинамические силы, в частности подъемную силу.

Можно задать вопрос: если мы умеем строить таблицы стрельбы путем интерполяции между измеренными величинами — именно между указанными выше коэффициентами C , — то зачем нам вообще нужно развивать какую бы то ни было теорию? Почему просто не сделать достаточное число выстрелов, измерить дальности и т. п., затем проинтерполировать непосредственно эти результаты и таким путем получить таблицы стрельбы? Ответ заключается в том, что число различных условий, при которых в действительности может оказаться необходимым стрелять, настолько велико, что очень трудно или невозможно надлежащим образом перекрыть его экспериментами. Интервалы между начальными условиями оказались бы слишком большими; при этом изменения интересующих нас величин в зависимости от начальных условий оказываются не столь простыми и закономерными, и прямая интерполяция будет ненадежной. Вместе с тем описанный ранее метод предохраняет нас от неприятностей, вызываемых этими изменениями, а опыт показывает, что интерполяция по C дает надежные результаты. Более того, неизмеримо увеличивается также надежность экстраполяции.

Как крайний случай этого рода приведем следующий пример: по информации, полученной от одного из авторов „Внешней баллистики“ [4], в тридцатых годах, когда проводился режим экономии, он получил задание подготовить таблицы бомбометания для некоторой бомбы, только один экземпляр которой был сброшен в условиях, допускающих выполнение баллистических измерений. В подобном случае интерполяция невозможна, а экстраполяция без надежной теории рискованна.

§ 3.6. Поправки на малые влияния. В § 3.2 мы уже определили поправку на одно малое влияние, именно учли влияние вращения Земли при короткой траектории. Можно распространить использованный там простейший прием и на другие случаи, но при этом нужно быть достаточно осторожным. Возьмем, например, уравнения (3.9). Причина, делающая их трудными для решения, состоит в присутствии членов, зависящих от силы сопротивления; эти члены нелинейны и, конечно, неизвестны, пока задача не решена. Можно рассуждать следующим образом. Траектория есть парабола, модифицированная вследствие сопротивления воздуха. Если сопротивление воздуха

мало, траектория близка к параболе; поэтому, если мы воспользуемся решением для пустоты

$$x = v_0 \cos \theta t, \quad y = v_0 \sin \theta t - \frac{1}{2} g t^2, \quad (3.10)$$

ошибка будет небольшой. Если мы воспользуемся формулами (3.10), то нетрудно вычислить в уравнениях (3.9) члены, содержащие K_D , непосредственно как функции времени, после чего решение уравнений (3.9) находим простым интегрированием.

Однако условия, при которых описанный метод дает хорошие результаты, в действительности бывают редко. С одной стороны, обычно мы не знаем заранее, не возмущают ли силы сопротивления траекторию настолько сильно, что равенства (3.10) дают слишком большую ошибку; поэтому мы не знаем, в какой степени можно доверять полученным результатам. То же самое возражение можно сделать и по поводу большинства приближенных методов, но, как указывалось в § 3.2, все приближенные методы должны быть тщательно проверены опытом и более точными расчетами, несмотря на их большую трудоемкость. Опыт показал, что учет сопротивления при помощи указанного только что грубого способа чрезвычайно обманчив. Это, вероятно, происходит вследствие того, что сила сопротивления может оказаться значительно большей, чем это кажется на первый взгляд. (Пусть читатель вычислит силу сопротивления свинцовой пули калибра 0,22 дюйма длиной $\frac{1}{2}$ дюйма при $K_D = 0,1$ и скорости 2000 *фут/сек.* Результат окажется неожиданно большим.) Нетрудно указать такие условия, при которых подобный метод приведет к отрицательной дальности; тем не менее для тяжелых снарядов с небольшой скоростью этот метод достаточно надежен.

Для введения малых поправок разных типов был предложен довольно стройный метод, называемый методом *разностных поправок* или методом *линейных возмущений*. Идея этого метода очень проста и основана на предположении, что если некоторое малое изменение условий вызывает определенное изменение траектории, то n -кратное изменение условий вызовет n -кратное изменение траектории. Другими словами, это есть распространение на дифференциальные уравнения идеи замены Δu на du в случае обычных функций.

Рассмотрим для примера влияние на траекторию слабого постоянного горизонтального попутного ветра w . Координаты x и y снаряда теперь являются функциями t и w . Таким образом,

$$x = x(t, w), \quad y = y(t, w). \quad (3.11)$$

Это получается потому, что мы теперь должны в уравнения (3.9) подставить $\dot{x} - w$ вместо $\dot{x}$. Если мы разложим выражения (3.11)

для x и y в ряд Тэйлора по малой величине w и ограничимся только линейными членами, то получим линейное приближение

$$x = x(t, 0) + \left[\frac{\partial x(t, w)}{\partial w} \right]_{w=0} w, \quad y = y(t, 0) + \left[\frac{\partial y(t, w)}{\partial w} \right]_{w=0} w. \quad (3.12)$$

Члены $x(t, 0)$ и $y(t, 0)$ относятся к невозмущенной траектории, которую мы предположим уже вычисленной. Тогда для того, чтобы воспользоваться приближением (3.12), нужно вычислить член $[\partial x(t, w)/\partial w]_{w=0}$, а также второй аналогичный член. Как это делается¹⁾, лучше всего пояснить на простом примере.

Представим себе шар, катящийся без трения по полу с начальной скоростью v_0 и начинающий двигаться из начала координат при $t=0$. Дифференциальное уравнение движения имеет вид

$$\ddot{x} = 0,$$

а его решение есть

$$x = v_0 t.$$

Предположим теперь, что имеется сопротивление, пропорциональное квадрату скорости; дифференциальное уравнение движения примет вид

$$\ddot{x} = -k\dot{x}^2. \quad (3.13)$$

Теперь решение есть функция t и k , т. е. $x = x(t, k)$, и мы ищем приближенное решение в виде

$$x = x(t, 0) + \left[\frac{\partial x(t, k)}{\partial k} \right]_{k=0} k. \quad (3.14)$$

Дифференцируя уравнение (3.13) по k , получаем

$$\frac{\partial}{\partial k} \left[\frac{\partial^2}{\partial t^2} x(t, k) \right] = - \left[\frac{\partial}{\partial t} x(t, k) \right]^2 - 2k \left[\frac{\partial}{\partial t} x(t, k) \right] \left[\frac{\partial^2}{\partial k \partial t} x(t, k) \right]. \quad (3.15)$$

Переменим в уравнении (3.15) порядок дифференцирования и положим $k=0$. Если мы введем обозначение $\delta x = [\partial x(t, k)/\partial k]_{k=0}$, то получим

$$\delta \ddot{x} = - [\dot{x}(t, 0)]^2,$$

или, так как $\dot{x}(t, 0) = v_0$,

$$\delta \ddot{x} = -v_0^2. \quad (3.16)$$

При $t=0$, имеем $x(0, k) = \dot{x}(0, k) = 0$, так что при $t=0$ $\delta x = \delta \dot{x} = 0$. Поэтому решение уравнения (3.16) есть

$$\delta x = -\frac{1}{2} v_0^2 t^2,$$

и мы получаем приближенное решение (3.14) в виде

$$x = v_0 t - \frac{1}{2} v_0^2 t^2 k. \quad (3.17)$$

¹⁾ В гл. 14 содержится более подробное изложение подобных вычислений.

Для этого простого примера легко уяснить себе истинный смысл сделанной нами аппроксимации. Точное решение уравнения (3.13) есть

$$x = \frac{1}{k} \log(1 + kv_0 t) = v_0 t - \frac{1}{2} v_0^2 t^2 k + \dots,$$

где разложение получено для случая $|kv_0 t| < 1$ путем использования ряда Тэйлора для $\log(1+x)$. Таким образом, в этом частном случае наше приближение (3.17) определяется линейным и квадратичным членами при разложении x в ряд Тэйлора по t или постоянным и линейным членами при разложении x в ряд Тэйлора по k .

В баллистической задаче обычно не особенно интересуются вариациями координат x и y вдоль траектории, а интересуются лишь вариациями x и y в ее конце, чтобы можно было вычислить вариации дальности и полного времени полета. Вычислим в предыдущем примере вариацию времени полета T для некоторой заданной дальности R . Время T находим, решая уравнение

$$x(T, k) = R.$$

Для заданного R время T есть функция от k , $T = T(k)$; когда $k = 0$, мы имеем $T = T_0 = R/v_0$. Дифференцируя по k , получаем

$$\frac{\partial x}{\partial T} \frac{dT}{dk} + \frac{\partial x}{\partial k} = 0.$$

Если положить здесь k равным нулю и обозначить $(dT/dk)_{k=0}$ через δT , мы получим

$$\delta T = - \frac{\frac{\partial x}{\partial k}}{\frac{\partial x}{\partial T}} = - \frac{-\frac{1}{2} v_0^2 T_0^2}{v_0} = \frac{R^2}{2v_0}$$

и линейное приближение для T имеет вид

$$T = \frac{R}{v_0} + \frac{R^2}{2v_0} k.$$

Линейные поправки к баллистическим траекториям, вызываемые ветром и другими отклонениями от нормальных условий, могут быть найдены аналогичным способом. В этом случае дифференциальные уравнения, соответствующие уравнению (3.13), естественно, значительно более сложны, но они имеют и чрезвычайно важное преимущество: они линейны. Коэффициенты этих уравнений зависят от невозмущенной траектории, которая предполагается известной.

В некоторых случаях подобные поправки могут быть получены прямо из баллистических таблиц без решения каких бы то ни было дифференциальных уравнений. Поправка на постоянный ветер относится именно к таким случаям. Если пользоваться осями координат, движущимися вместе с воздухом, то дифференциальные уравнения траектории в них будут такими же, как в неподвижной системе координат при отсутствии ветра. Конечно, здесь имеет место изменение начальных условий; если x_0 и y_0 являются проекциями начальной

скорости снаряда в неподвижных осях, то в подвижных они будут $\dot{x}_0 - w$ и $\dot{y}_0$. Дальность R и время полета T могут быть получены из баллистических таблиц; однако это будет дальность относительно воздуха, истинная же дальность с учетом влияния ветра будет $R + wT$. Другой пример подобного типа есть случай отклонения плотности ρ от ее значения для нормальной атмосферы. Если ρ вместо того, чтобы быть равной $\rho_0 e^{-hy}$, отличается от этого значения постоянным множителем k , т. е. имеет место равенство $\rho = k\rho_0 e^{-hy}$, то влияние этого изменения на уравнения движения может быть точно компенсировано соответствующим изменением баллистического коэффициента. В самом деле, умножение плотности на k вызывает точно такой же эффект, как умножение баллистического коэффициента на $1/k$.

Пожалуй, следует остановиться на том, как можно учесть влияние ветра, изменяющегося по высоте. Рассмотрим случай, когда скорость ветра оказывается равной 10 *фут/сек* в первой тысяче футов высоты, 17 *фут/сек* — во второй тысяче, — 35 *фут/сек* — в третьей и т. д. Пользуясь изложенным выше методом поправок, можно найти влияние ветра, скорость которого равна 1 *фут/сек* в первой тысяче футов и нулю на всех остальных высотах. Обозначим это влияние через e_1 . Подобным же образом влияние ветра со скоростью, равной 1 *фут/сек* во второй тысяче футов и нулю на остальных высотах, обозначим через e_2 и т. д. Тогда полное влияние ветра будет

$$10e_1 + 17e_2 - 35e_3 + \dots$$

На практике для быстроты и упрощения вычислений данные такого рода, как числа e_i , являющиеся, очевидно, коэффициентами веса, с которыми ветер на разных высотах участвует в общем эффекте, обычно бывают представлены в графической форме, и получающаяся при этом кривая называется *кривой коэффициентов веса*. Использование этой кривой позволяет быстро вычислить постоянный ветер, который производит тот же самый эффект, что и заданный переменный ветер. Отметим, что в некоторых случаях ветры у поверхности земли и на больших высотах могут различаться больше чем на 100 миль в час.

Необходимо сделать замечание относительно одного из наиболее важных применений метода малых поправок — приведения результатов эксперимента к нормальным условиям. Возьмем, например, экспериментальные стрельбы для определения дальности. Если во время стрельбы был попутный ветер w , снаряд будет иметь дальность больше ожидаемой. Если на ветер w не будут внесены поправки, таблицы стрельбы, основанные на этом эксперименте, будут смещены, и снаряды при различных ветрах в среднем будут ложиться ближе, чем предписывается таблицами стрельбы. Таким образом, значительно более важно исключить влияние ветра из результатов,

полученных во время экспериментальных стрельб, чем учитывать его во время действительного применения.

§ 3.7. Бомбометание с самолетов. Баллистические задачи, с которыми приходится встречаться при бомбометании с самолетов, в основном сходны с баллистическими задачами при артиллерийской стрельбе, но могут отличаться в деталях и поэтому требуют иного подхода. Например, траектория бомбы, сброшенной при горизонтальном полете, по существу есть вторая половина траектории артиллерийского снаряда.

Интересно, пожалуй, отметить, что одно обстоятельство, которое, как кажется на первый взгляд, представляет значительные трудности, на самом деле может быть учтено точно и без всяких затруднений. Дело в том, что движение самолета и бомбы легче всего описать относительно воздуха, а движение цели — относительно земли.

Рассмотрим в качестве простого примера случай, в котором бомба сбрасывается с высоты H при скорости v_0 и при попутном ветре w на цель, находящуюся прямо впереди, как показано на рис. 3.4.

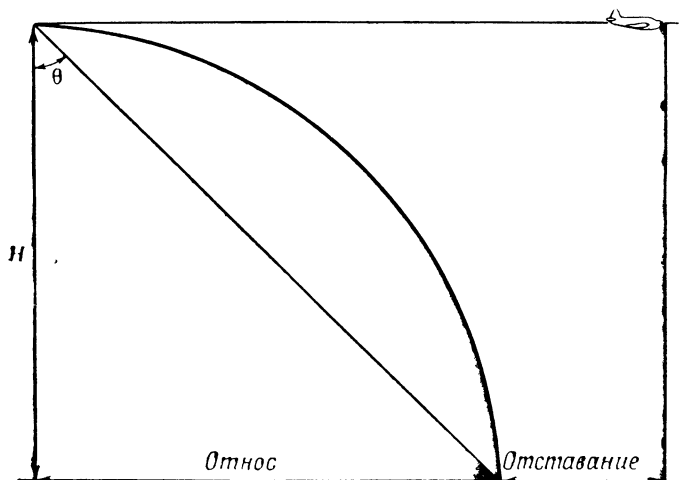


Рис. 3.4. Бомбометание при горизонтальном полете и при попутном ветре.

Приборы на самолете измеряют H и v_0 . Следовательно, мы знаем начальную скорость бомбы относительно воздуха и ее высоту; воздух мы предполагаем движущимся с постоянной скоростью w . Все это вместе определяет баллистическое поведение бомбы. Таблицы бомбометания для этой определенной бомбы содержат все необходимые элементы траектории: обычно в таблицах приводятся время полета (падения) T_f и отставание λ , равное $v_0 T_f$ минус горизонтальная

дальность относительно воздуха. *Отставание* есть расстояние, на котором бомба находится позади самолета в момент удара; очевидно, что оно не зависит от ветра.

Относ бомбы, или горизонтальное расстояние, пройденное бомбой относительно земли за время падения, зависит от w и выражается следующим образом:

$$\text{Относ} = \text{Дальность относительно воздуха} + wT_f = (w + v_0) T_f - \lambda.$$

Относ бомбы равен горизонтальному расстоянию от цели, на котором бомба должна быть сброшена; угол сбрасывания θ определяется из равенства

$$\operatorname{tg} \theta = \frac{(w + v_0) T_f - \lambda}{H}. \quad (3.18)$$

Бомбардир должен держать свой прицел направленным на цель и сбрасывать бомбу в тот момент, когда угол α между визирной линией и вертикалью становится равным θ . В формуле (3.18) все величины известны, за исключением ветра, или, что то же самое, путевой скорости $w + v_0$. Однако эта последняя может быть определена при помощи угловой скорости визирной линии, следящей за целью. Именно: путевая скорость $= d(H \operatorname{tg} \alpha)/dt$. Поэтому если перемещение визирной линии в прицеле осуществляется с помощью сервомотора, то угловую скорость последнего можно чрезвычайно просто использовать для получения путевой скорости, и тогда оказываются известными все данные, необходимые для вычисления $\operatorname{tg} \theta$. В действительности соответствующие вычисления производятся в самом прицеле. Нетрудно спроектировать прицел, в котором имеются шкалы для установки v_0 , H , T_f и λ и в котором все остальное, включая сбрасывание бомб, производится автоматически.

Учет ветра в том виде, как он только что описан, основывается на предположении, что ветер одинаков на всех высотах, — иногда это оказывается далеким от истины. В этом случае можно вводить поправки методом, изложенным в § 3.6.

Работы по подготовке таблиц бомбометания и стрельбы сходны между собой. Однако для получения данных из экспериментального бомбометания требуется более сложное оборудование. Обычно легко найти, где бомба упала, но не так просто установить, откуда она была сброшена.

§ 3.8. Влияние аэродинамических сил, отличных от силы лобового сопротивления. В § 3.3 мы провели длинное рассуждение о различных аэродинамических силах, действующих на снаряд, а в § 3.4 мы все эти силы просто отбросили, кроме силы лобового сопротивления. Однако существуют случаи, когда это безнаказанно сделать уже нельзя. Например, пуля из пулемета, стреляющего вбок с летящего самолета, будет иметь большой начальный угол атаки, возможно порядка 15° , вследствие чего поперечные силы и

моменты, несомненно, значительно повлияют на траекторию. При этих обстоятельствах существенным также является вопрос об устойчивости. Будут ли различные аэродинамические силы и моменты стремиться погасить этот угол атаки или, наоборот, увеличат его настолько, что произойдет опрокидывание?

Чтобы ответить на этот вопрос, нужно рассмотреть уравнения движения во всех подробностях с учетом различных аэродинамических сил. Уравнения становятся очень сложными и трудными для изучения вследствие их нелинейности. Нилсен и Синг [5], а также Макшейн, Келли и Рено [4] изучали эту проблему и получили интересные и полезные результаты, как качественные, так и количественные. Общий метод состоит в такой замене переменных, чтобы быстро меняющиеся величины были заменены медленно меняющимися, вследствие чего дифференциальные уравнения на значительной части траектории становятся близкими к линейным. Тогда угол атаки можно рассматривать как состоящий из двух частей, соответствующих решениям однородных и неоднородных дифференциальных уравнений. Первая часть зависит от начальных условий и имеет колебательный характер. Можно найти такие условия, при которых эта часть с течением времени затухает. Эти условия выражаются в виде некоторых неравенств, которым должны удовлетворять различные аэродинамические коэффициенты; они называются *условиями устойчивости*. Вторую часть можно назвать *остаточным углом атаки*; у устойчивого снаряда это есть тот угол атаки, который остается, когда переходный процесс затухнет. Для невращающихся снарядов, таких, как бомбы, этот остаточный угол таков, что ось бомбы отклонена от траектории немного вверх. В случае вращающегося снаряда его ось, в среднем, отклонена вправо и вверх (для снарядов правого вращения). Изложенное и является объяснением того, почему бомба имеет большую дальность, чем это следует из теории, учитывающей только лобовое сопротивление, и почему вращающийся снаряд отклоняется вправо. Очень точные оценки этих явлений могут быть получены при помощи метода поправок.

§ 3.9. Заключение и обзор литературы. Существует много отраслей баллистики, которые мы здесь совершенно не затронули. Во-первых, это относится к теории полета ракет, в которой появляются новые и запутывающие осложнения. Здесь у нас нет места даже для того, чтобы упомянуть о существовании вопроса. Во-вторых, мы совершенно опустили столь важный вопрос, как описание методов измерения различных аэродинамических коэффициентов. В этой области имеется широкое поле деятельности для инженеров, так как необходимо любым способом записать поведение снаряда в действительном или моделированном полете. Однако заметную роль играет здесь и математика — при анализе результатов измерений и

при попытках выделить из общих результатов влияние отдельных аэродинамических сил.

В связи с этими проблемами специалист по баллистике должен уметь пользоваться одной из отраслей математики, которая оказывается чрезвычайно полезной во всей его работе, — математической статистикой, изложенной во второй части настоящей книги. Ему всюду приходится иметь дело с неточно измеренными величинами или измерениями, произведенными при изменяющихся и не точно известных физических условиях; поэтому, чтобы проверить надежность и получить максимум сведений из некоторого фиксированного количества данных, он вынужден применять к ним статистические методы.

Имеется очень много источников, из которых можно черпать материалы по внешней баллистике. Большей частью это технические доклады правительственных лабораторий или поставщиков, в той или иной степени являющиеся секретными. Кроме того, имеются полезные общедоступные книги, более или менее охватывающие предмет. Одной из них является четырехтомный труд Кранца [2], охватывающий все вопросы баллистики. Этот труд основан на данных времен первой мировой войны и очень богат историческими материалами и изложением применявшихся тогда вычислительных методов. Небольшая книга Блисса [1], вышедшая около двенадцати лет тому назад, отличается очень хорошим выбором материала и доступностью изложения. В частности, в ней прекрасно изложен метод поправок и имеется очень легкое для понимания описание существующих механических интеграторов.

Большой курс „Внешней баллистики“, написанный Макшейном, Келли и Рено [4], является единственным по современности, совершенству, полноте и обстоятельности изложения вопроса. Автор чрезвычайно обязан этому труду, которым он пользовался при написании настоящей главы. Полет ракет трактуется подробно в современной книге Россера, Ньютона и Гросса [6]. Главы X и XII книги Хэйеса [3] „Основы артиллерии“ дают доступное введение во внешнюю баллистику.

ЛИТЕРАТУРА

1. Bliss G. A., *Mathematics for Exterior Ballistics*, John Wiley, New York, 1944.
2. Cranz K. J., *Lehrbuch der Ballistik*, 2. Aufl., Springer-Verlag, Berlin, 1927. Reprinted by J. W. Edwards, Ann Arbor, Mich.
3. Hayes T. J., *Elements of Ordnance*, John Wiley, New York, 1938.
4. McShane E. J., Kelley J. L., Reno F. V., *Exterior Ballistics*, University of Denver Press, Denver, 1953.
5. Nielsen K. L. and Synge J. L., *On the Motion of a Spinning Shell*, *Quart. Appl. Math.*, 4, 201—226 (1946).
6. Россер Д., Ньютон Р. и Гросс Г., *Математическая теория полета неуправляемых ракет*, ИЛ, М., 1950.

Глава 4

ЭЛЕМЕНТЫ ВАРИАЦИОННОГО ИСЧИСЛЕНИЯ

МАГНУС Р. ХЕСТЕНС

Калифорнийский университет, Лос-Анжелес

§ 4.1. Введение. Вариационное исчисление является одной из старейших математических дисциплин. Оно развивалось почти одновременно с анализом. Уже Ньютон и Лейбниц интересовались вариационными задачами. Однако обычно основателями вариационного исчисления считают братьев Бернулли, Эйлера и Лагранжа. В известном смысле вариационное исчисление является продолжением анализа. Оно в основном примыкает к теории максимумов и минимумов. При этом функции, у которых разыскивается экстремум, могут быть функциями точки, дуги, поверхности и т. п. Поэтому в вариационном исчислении мы имеем дело с теорией функций от функций, или, что то же, с теорией *функционалов*. Теория функционалов не отличается простотой, но она очень важна как для чистой, так и для прикладной математики. Вариационное исчисление представляет собой один из очень важных разделов этой теории.

Вариационное исчисление тесно связано с приложениями математики к механике, физике, к инженерному делу, численному анализу и ко многим другим областям. Существующие вариационные теории достаточно общи, однако они, конечно, не могут удовлетворить все возрастающего интереса к прикладным вопросам. Непрерывно возникают новые проблемы, в связи с чем вариационное исчисление все время продолжает быть объектом новых исследований.

Целью настоящей главы является описание некоторых типичных задач, которые могут быть изучены вариационными методами, и иллюстрация этой теории некоторыми простыми физическими приложениями. Это описание не претендует на полноту, многие интересные и важные вопросы вариационного исчисления при этом, естественно, будут опущены. Однако мы надеемся, что изложение будет достаточно полным для того, чтобы дать известное представление о красоте и пользе этого интересного предмета. Вариационное исчисление в приложениях имеет три особенности. Во-первых, существует много проблем, которые наиболее легко могут быть выражены при помощи понятий максимума и минимума и потому допускают непосредственную формулировку в виде вариационных принципов.

Во-вторых, существуют некоторые проблемы, которые хотя и формулируются сначала каким-то другим образом, но затем могут быть сведены к вариационным принципам, благодаря чему удается глубже проникнуть в сущность рассматриваемой проблемы. В-третьих, вариационные методы помогают не только при изучении свойств решений, но обычно дают и способ получения этих решений. Одним из преимуществ вариационных методов является их инвариантный характер и соответственно их независимость от применяемой системы координат.

Как было отмечено выше, в короткой главе невозможно изложить вариационное исчисление достаточно полно. Дальнейшие сведения по этому вопросу можно найти в книгах, список которых указан в конце этой главы. Этот список не полон, но он содержит большую часть основных книг по этому предмету, а дополнительные указания на литературу можно найти в этих книгах.

§ 4.2. Некоторые элементарные вариационные задачи. Исторически теория вариационного исчисления началась, по-видимому, с изучения задачи о брахистохроне, или линии кратчайшего времени падения. Эта задача изучалась Галилеем, Яковом и Иваном Бернулли, Ньютоном, Лейбницем и другими. Она состоит в отыскании кривой, двигаясь по которой под действием одной только силы тяжести,

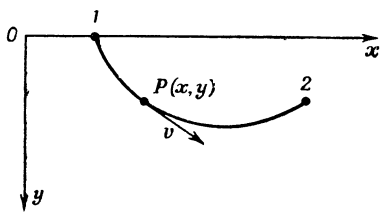


Рис. 4.1. Задача о брахистохроне.

материальная точка попадает из одной точки в другую в кратчайшее время. При математической формулировке задачи мы ограничимся рассмотрением плоского движения. Проведем ось x горизонтально через начальную точку 1 и направим ось y вниз, как указано на рис. 4.1. Из физики известно, что скорость v частицы P с массой m

в точке (x, y) кривой выражается формулой $v = (2mgy)^{1/2}$. Следовательно, время T , которое требуется для прохождения пути от точки 1 до точки 2 вдоль данной кривой, определяется формулой

$$T = \int_0^T dt = \int_0^l \frac{dt}{ds} ds = \int_0^l \frac{ds}{v} = (2mg)^{-1/2} \int_0^l y^{-1/2} ds,$$

где l обозначает длину кривой. Если ограничиться рассмотрением кривых вида

$$y = y(x), \quad x_1 \leq x \leq x_2, \quad (4.1)$$

то задача о брахистохроне сводится к задаче отыскания среди кривых (4.1), лежащих в полуплоскости $y > 0$ и соединяющих точки 1

и 2, кривой, для которой интеграл

$$I = \int_{x_1}^{x_2} y^{-1/2} (1 + y'^2)^{1/2} dx, \quad y' = \frac{dy}{dx},$$

принимает наименьшее значение.

В дальнейшем будет показано, что решением этой задачи является циклоида:

$$x = a - b(\varphi + \sin \varphi), \quad y = b(1 + \cos \varphi),$$

где a и b выбираются так, чтобы циклоида проходила через данные точки 1 и 2. Эта кривая описывается точкой P окружности радиуса b , катящейся по оси x .

Вторая задача, интересная как с исторической, так и с технической точки зрения, заключается в отыскании поверхности вращения наименьшей площади. Физически поверхность наименьшей площади может быть осуществлена посредством мыльной пленки. Если мы возьмем ось x в качестве оси вращения, как это указано на рис. 4.2, то задача будет заключаться в отыскании в классе кривых (4.1), соединяющих точки 1 и 2 и лежащих в верхней полуплоскости кривой, для которой интеграл

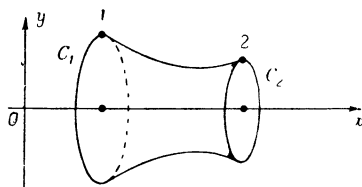


Рис. 4.2. Поверхность вращения наименьшей площади: катеноид.

$$I = 2\pi \int_0^l y ds = 2\pi \int_{x_1}^{x_2} y (1 + y'^2)^{1/2} dx$$

принимает наименьшее значение.

Эта задача интересна, в частности, еще тем, что является одной из простейших проблем, для которых минимизирующая дуга существует только в тех случаях, когда точки 1 и 2 расположены соответствующим образом. Если, например, окружности C_1 и C_2 на рис. 4.2 представить в виде двух проволочных петель, на которые, как указано, натянута мыльная пленка, то ясно, что если петлю C_2 удалять от C_1 , то мыльная пленка станет растягиваться и наконец лопнет, когда расстояние между двумя петлями станет слишком большим, чтобы пленка могла существовать. Минимизирующая кривая, если она существует, получается в виде цепной линии

$$y = b \operatorname{ch} \frac{x-a}{b},$$

где a и b выбираются так, чтобы кривая проходила через заданные точки 1 и 2.

Интересно отметить, что Иван Бернулли первый решил задачу о брахистохроне, представив последнюю как траекторию луча света, проходящего через среду с переменным коэффициентом преломления. Согласно *принципу Ферма*, траектория луча света есть линия, по которой свет проходит в кратчайшее время по сравнению с другими кривыми, имеющими те же концевые точки. Пусть в двумерном случае $v(x, y)$ есть скорость света в точке (x, y) изотропной среды и пусть c — его скорость в пустоте. Отношение $h(x, y) = c/v(x, y)$ называется *коэффициентом преломления*. Произведение $I = cT$, где T обозначает время, за которое свет проходит кривую длины l , определяется формулой

$$I = \int_0^l \frac{c ds}{v(x, y)} = \int_0^l h(x, y) ds = \int_{x_1}^{x_2} h(x, y) (1 + y'^2)^{1/2} dx, \quad (4.2)$$

причем последнее равенство справедливо, если рассматриваемая кривая имеет форму (4.1). Таким образом, согласно принципу Ферма, траектория луча света минимизирует интеграл (4.2), взятый по всевозможным кривым, соединяющим две данные точки. Задача о брахистохроне получается (формально) отсюда при $h = y^{-1/2}$. Поэтому нет ничего удивительного в том, что Ивану Бернулли удалось решить задачу о брахистохроне, используя ее аналогию с теорией распространения света (см. § 4.5).

По-видимому простейшей вариационной задачей является отыскание кривой наименьшей длины, соединяющей две данные точки. В этом случае минимизируемым интегралом является

$$I = \int ds.$$

Если мы рассмотрим трехмерный случай с кривыми, заданными в параметрической форме

$$x = x(t), \quad y = y(t), \quad z = z(t), \quad t_1 \leq t \leq t_2,$$

то указанный интеграл принимает вид

$$I = \int_{t_1}^{t_2} (\dot{x}^2 + \dot{y}^2 + \dot{z}^2)^{1/2} dt,$$

где точками обозначены производные по t . Решением в этом случае, очевидно, является прямая линия. Однако задача усложнится, если мы потребуем, чтобы сравниваемые кривые были расположены на данной поверхности. Например, мы можем потребовать, чтобы все кривые лежали на сфере

$$x^2 + y^2 + z^2 = a^2.$$

В этом случае можно легко показать, что решением являются дуги больших кругов. Эта задача может быть сведена к двумерной посредством введения сферических координат θ и φ по формулам

$$x = a \sin \theta \cos \varphi, \quad y = a \sin \theta \sin \varphi, \quad z = a \cos \theta.$$

Дуга кривой определяется тогда уравнениями

$$\theta = \theta(t), \quad \varphi = \varphi(t), \quad t_1 \leq t \leq t_2, \quad (4.3)$$

и соответствующая длина дуги выражается интегралом

$$I = a \int_{t_1}^{t_2} (\dot{\theta}^2 + \dot{\varphi}^2 \sin^2 \theta)^{1/2} dt.$$

Задача свелась, таким образом, к минимизированию этого интеграла, содержащего две независимые функции (4.3). В более общем случае, если потребовать, чтобы дуги лежали на поверхности

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v),$$

интеграл, выражающий длину кривой, принимает вид

$$I = \int_{t_1}^{t_2} ds = \int_{t_1}^{t_2} (E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)^{1/2} dt,$$

где

$$E = x_u^2 + y_u^2 + z_u^2, \quad G = x_v^2 + y_v^2 + z_v^2,$$

$$F = x_u x_v + y_u y_v + z_u z_v,$$

причем индексы внизу обозначают частные производные. Дуга кривой определяется уравнениями

$$u = u(t), \quad v = v(t), \quad t_1 \leq t \leq t_2,$$

в которых u и v опять представляют собой независимые функции.

Только что описанную задачу о геодезических (кратчайших) линиях математически удобнее формулировать, пользуясь системой обозначений и терминологией, обычно применяемыми в римановой геометрии. Здесь мы имеем n координат $x_1, \dots, x_n$, соответствующих переменным u и v , рассмотренным выше. Длина кривой

$$x_i = x_i(t), \quad t_1 \leq t \leq t_2, \quad i = 1, 2, \dots, n,$$

определяется формулой

$$I = \int_{t_1}^{t_2} [g_{ij}(x) \dot{x}_i \dot{x}_j]^{1/2} dt, \quad (4.4)$$

где по i и j происходит суммирование от 1 до n , а точкой, как и раньше, обозначается производная по t . Этот интеграл достигает

своего минимума на геодезической линии, т. е. на кривой кратчайшей длины. В случае сферы, рассмотренном выше, мы имеем

$$x_1 = 0, \quad x_2 = \varphi, \quad g_{11} = a^2, \quad g_{12} = g_{21} = 0, \quad g_{22} = a^2 \sin^2 \varphi.$$

Вернемся теперь к принципу Ферма и положим $g = h^2$. Если мы представим дугу кривой на плоскости (x, y) в параметрической форме

$$x = x(t), \quad y = y(t), \quad t_1 \leq t \leq t_2,$$

то

$$ds = (\dot{x}^2 + \dot{y}^2)^{1/2} dt$$

и интеграл (4.2) примет вид

$$I = \int_0^I h ds = \int_{t_1}^{t_2} (g\dot{x}^2 + g\dot{y}^2)^{1/2} dt.$$

Этот интеграл является частным случаем (4.4). Таким образом, существует тесная связь между теорией распространения света, основывающейся на принципе Ферма, и теорией геодезических линий. С точки зрения вариационного исчисления это по существу одна и та же математическая проблема, допускающая различные физические интерпретации.

Прежде чем закончить этот параграф, интересно отметить задачу типа задачи о брахистохроне, которую могли бы сформулировать братья Бернулли, если бы они жили в настоящее время. Мы имеем

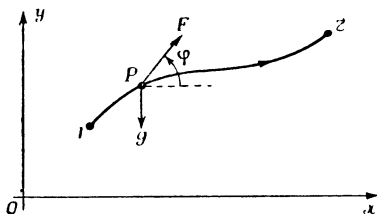


Рис. 4.3. Модификация задачи о брахистохроне: задача о ракете.

в виду следующее интересное видоизменение задачи о брахистохроне. Рассмотрим идеальную ракету, движущуюся в вертикальной плоскости. При этом мы будем считать ракету частицей, на которую действует сила тяжести g и реактивная сила постоянной величины F с переменным углом наклона φ , как это показано на рис. 4.3. Предположим, что на ракету не действуют никакие другие силы. Пилот ведет ракету, изменяя угол φ , т. е. изменяя направление реактивной силы. Задача состоит в отыскании пути наименьшего времени при соответствующих начальных и конечных условиях. Аналитически это сводится к отысканию среди всех функций

$$x = x(t), \quad y = y(t), \quad \varphi = \varphi(t), \quad 0 \leq t \leq T,$$

связанных соотношениями (при единичной массе)

$$\ddot{x} = F \cos \varphi, \quad \ddot{y} = F \sin \varphi - g$$

и с заданными значениями $x, y, \dot{x}, \dot{y}$ при $t=0$ и $t=T$, функции, минимизирующей время полета T . Можно показать, что путь наименьшего времени обладает следующим интересным свойством: существует определенным образом направленная в пространстве и определенным образом расположенная относительно ракеты прямая L , обладающая тем свойством, что если некоторая точка M будет двигаться вдоль этой прямой с постоянной скоростью, а реактивная сила, действующая на ракету, будет все время направлена в точку M , то ракета будет двигаться по пути наименьшего времени. Это задача другого типа по сравнению с рассмотренными выше. Она относится к более общему классу задач, называемых *задачами Больца*, которые будут описаны в § 4.12.

§ 4.3. Основные задачи. Необходимые условия минимума. Задачи, рассмотренные нами в § 4.2, были двух типов, — параметрического и непараметрического. Рассмотрим сначала непараметрический тип. В двумерном случае дуга в плоскости (x, y) может быть выражена в виде

$$y = y(x), \quad x_1 \leq x \leq x_2,$$

а минимизируемый интеграл будет иметь вид

$$I = \int_{x_1}^{x_2} f[x, y(x), y'(x)] dx. \quad (4.5)$$

В $(n+1)$ -мерном случае с координатами $(x, y_1, \dots, y_n)$ кривые можно задать в виде

$$y_i = y_i(x), \quad x_1 \leq x \leq x_2, \quad i = 1, \dots, n,$$

а для подинтегральной функции в минимизируемом интеграле можно воспользоваться сокращенным обозначением

$$f(x, y, y') = f(x, y_1, \dots, y_n, y'_1, \dots, y'_n),$$

так что интеграл сохранит прежний вид (4.5). Рассматриваемые кривые предполагаются непрерывными, с кусочно-непрерывно вращающейся касательной и расположенными в некоторой области R пространства (x, y, y') , в которой имеет смысл интеграл (4.5). Подинтегральная функция f предполагается непрерывной и имеющей непрерывные частные производные первого и второго порядков в области R . Часто представляет большой интерес и изучение разрывных подинтегральных функций, но в настоящий момент мы не будем касаться этого вопроса. Мы будем рассматривать задачи о минимизировании интеграла (4.5) в классе допустимых кривых, соединяющих две данные точки 1 и 2.

В n -мерном параметрическом случае кривые определяются уравнениями

$$x_i = x_i(t), \quad t_1 \leq t \leq t_2, \quad i = 1, \dots, n,$$

а интеграл принимает вид

$$I = \int_{t_1}^{t_2} f[x(t), \dot{x}(t)] dt,$$

где мы опять использовали сокращенное обозначение

$$f(x, \dot{x}) = f(x_1, \dots, x_n, \dot{x}_1, \dots, \dot{x}_n).$$

Для того чтобы обеспечить независимость интеграла I от параметризации кривых, подинтегральная функция должна обладать дополнительным свойством, состоящим в том, что соотношение

$$f(x, k\dot{x}) = kf(x, \dot{x}), \quad k > 0$$

выполняется для всех положительных значений k . Аналогично непараметрическому случаю и здесь накладываются ограничения на непрерывность и дифференцируемость как подинтегральной функции f , так и рассматриваемых кривых. Без особого ущерба для содержания мы ограничимся изучением непараметрического случая. Это — простейший случай, и большинство свойств параметрических задач может быть получено из соответствующих свойств непараметрических задач.

Вернемся снова к непараметрической задаче, в которой допустимые кривые соединяют две данные точки 1 и 2. Минимизирующая кривая в этом случае удовлетворяет четырем условиям. Мы сформулируем их сначала для двумерного случая, а затем укажем соответствующие их видоизменения для общего многомерного случая.

а) *Необходимые условия Эйлера.* Функции $f - y'f_{y'}$, $f_{y'}$ непрерывны вдоль минимизирующей кривой и между угловыми точками имеют непрерывные производные, удовлетворяющие дифференциальным уравнениям

$$\frac{d}{dx}(f - y'f_{y'}) = f_x, \quad \frac{d}{dx}f_{y'} = f_{y''}. \quad (4.6)$$

Здесь f_x , f_y , $f_{y'}$ обозначают частные производные функции f соответственно по x , y и y' . Дифференциальные уравнения (4.6) называются *уравнениями Эйлера* для интеграла I . Решения уравнений Эйлера называются *экстремальями* независимо от того, являются они минимизирующими, максимизирующими или седловыми кривыми. В многомерном случае уравнения Эйлера имеют вид

$$\frac{d}{dx}(f - y'_i f_{y'_i}) = f_{x_i}, \quad \frac{df_{y'_i}}{dx} = f_{y''_i}.$$

В первом из этих уравнений производится суммирование от 1 до n по повторяющемуся индексу i . Непосредственным дифференцированием можно убедиться в том, что первое из этих уравнений является следствием второго, при условии, что функции $y = y_i(x)$ обладают вторыми производными, как мы будем предполагать в дальнейшем. Необходимое условие Эйлера часто называют также *первым необходимым условием*.

б) *Необходимое условие Вейерштрасса*. Для каждого элемента (x, y, y') минимизирующей кривой выполняется неравенство

$$E(x, y, y', Y') \geq 0,$$

если только $(x, y, Y') \neq (x, y, y')$ и (x, y, Y') лежит в R и E есть функция Вейерштрасса

$$E = f(x, y, Y') - f(x, y, y') - (Y' - y')f_{y'}(x, y, y').$$

В общем случае функция E имеет вид

$$E = f(x, y, Y') - f(x, y, y') - (Y'_i - y'_i)f_{y'_i}(x, y, y'),$$

где по i производится суммирование от 1 до n .

в) *Необходимые условия Лежандра*. Вдоль минимизирующей кривой выполняется неравенство

$$f_{y'y'} \geq 0.$$

В общем случае вдоль минимизирующей кривой имеет место неравенство

$$f_{y'_i y'_k} \pi_i \pi_k \geq 0$$

для любой совокупности переменных $(\pi) \neq (0)$.

Условия (б) и (в) (если отбросить в них случай равенства) обычно выполняются для всех допустимых элементов (x, y, y') и (x, y, Y') в большинстве интересующих нас задач вариационного исчисления. Например, это имеет место для интеграла (4.2), если $h(x, y) > 0$. С помощью этих условий обычно устанавливают, что минимизирующая кривая не имеет угловых точек.

г) *Необходимое условие Якоби*. Это условие имеет весьма специальный характер и его частные случаи будут описаны в дальнейшем. Его изучение является одной из наиболее интересных и плодотворных глав вариационного исчисления. Если условие (в), за исключением случая равенства, выполняется, тогда условие Якоби также имеет место для достаточно короткой дуги кривой.

§ 4.4. Вывод уравнения Эйлера. В § 4.3 было указано, что на дуге

$$y = y(x), \quad x_1 \leq x \leq x_2, \quad (4.7)$$

расположенной в плоскости (x, y) и минимизирующей интеграл

$$I = \int_{x_1}^{x_2} f(x, y, y') dx$$

в некотором классе достаточно близких кривых, имеющих те же самые концевые точки, между угловыми точками удовлетворяется уравнение Эйлера

$$\frac{df_{y'}}{dx} = f_y \quad (4.8)$$

и, кроме того, производная $f_{y'}$ непрерывна вдоль кривой.

Для того чтобы дать некоторое представление о методах, применяемых в вариационном исчислении, докажем эту теорему. С этой целью рассмотрим вспомогательную функцию

$$y = \eta(x), \quad x_1 \leq x \leq x_2,$$

обращающуюся в нуль при $x = x_1$ и при $x = x_2$. Такая функция называется обычно *вариацией* и обозначается символом $\delta y(x)$. Это название возникло в связи с тем, что при каждом значении t кривая

$$y = y(x) + t\eta(x), \quad x_1 \leq x \leq x_2 \quad (4.9)$$

соединяет концевые точки кривой (4.7), а при $t = 0$ совпадает с этой кривой. Кроме того, производная от $y(x) + t\eta(x)$ по t равна самой вариации $\eta(x)$. Значение интеграла I вдоль кривой (4.9) дается функцией

$$I(t) = \int_{x_1}^{x_2} f[x, y(x) + t\eta(x), y'(x) + t\eta'(x)] dx. \quad (4.10)$$

Так как $I(0)$ является значением интеграла вдоль минимизирующей кривой (4.7), то для достаточно малых значений t получаем

$$I(t) \geq I(0).$$

Из элементарного анализа следует, что имеют место соотношения

$$I'(0) = 0, \quad I''(0) \geq 0, \quad (4.11)$$

где штрихи обозначают дифференцирование по t . Если продифференцировать интеграл (4.10) по t и положить $t = 0$, то получим

$$I'(0) = \int_{x_1}^{x_2} (f_y \eta + f_{y'} \eta') dx,$$

где производные функции f берутся вдоль минимизирующей кривой. Интеграл, стоящий справа, называется *первой вариацией* интеграла I вдоль данной кривой. Его обычно обозначают символом $I_1(\eta)$ или символом δI . Таким образом,

$$\delta I = I_1(\eta) = \int_{x_1}^{x_2} (f_y \eta + f_{y'} \eta') dx.$$

Пользуясь δ -обозначениями, обычно пишут

$$\delta I = \delta \int_{x_1}^{x_2} f dx = \int_{x_1}^{x_2} (f_y \delta y + f_{y'} \delta y') dx,$$

откуда виден дифференциальный характер первой вариации. Так как

$$I'(0) = I_1(\eta) = 0,$$

то мы имеем следующий результат:

Если кривая (4.7) осуществляет минимум интеграла I в классе достаточно близких к ней кривых, имеющих те же самые концевые точки, тогда вдоль нее выполняется условие

$$\delta I = \int_{x_1}^{x_2} (f_y \eta + f_{y'} \eta') dx = 0$$

при любых вариациях $\eta(x)$, обращающихся в нуль в точках x_1 и x_2 .

В общем случае первая вариация имеет вид

$$\delta I = \int_{x_1}^{x_2} (f_{y_i} \eta_i + f_{y'_i} \eta'_i) dx.$$

Из только что полученного результата мы заключаем, что производная $f_{y'}$ непрерывна и что уравнение (4.8) удовлетворяется вдоль минимизирующей кривой, чем и доказывается существенная часть условия (а) § 4.3. Этот факт вытекает из следующей основной леммы, если положить в ней $M = f_y$ и $N = f_{y'}$.

Основная лемма. Пусть $M(x)$ и $N(x)$ — кусочно непрерывные функции на отрезке $x_1 \leq x \leq x_2$. Предположим, что равенство

$$\int_{x_1}^{x_2} [M(x) \eta(x) + N(x) \eta'(x)] dx = 0 \quad (4.12)$$

имеет место для любой вариации $\eta(x)$, обращающейся в нуль в x_1 и в x_2 . Тогда функция $N(x)$ непрерывна, а $M(x)$ является ее производной, за исключением точек разрыва функции $M(x)$.

Эта лемма вместе с ее обобщениями является одной из самых фундаментальных теорем анализа.

Чтобы доказать эту лемму, отметим, что ее утверждение эквивалентно утверждению о существовании такой постоянной c , что

$$N(x) = \int_{x_1}^x M(s) ds + c$$

на $[x_1, x_2]$. Положим

$$P(x) = N(x) - \int_{x_1}^x M(s) ds - c$$

и выберем c так, чтобы

$$\int_{x_1}^{x_2} P(x) dx = 0. \quad (4.13)$$

Мы получим

$$\frac{d}{dx} [(N - P)\eta] = (N - P)\eta' + M\eta = M\eta + N\eta' - P\eta'.$$

Интегрируя от x_1 до x_2 и используя соотношение $\eta(x_1) = \eta(x_2) = 0$, мы видим, что

$$\int_{x_1}^{x_2} (M\eta + N\eta') dx = \int_{x_1}^{x_2} P\eta' dx.$$

Следовательно, согласно равенству (4.12),

$$\int_{x_1}^{x_2} P\eta' dx = 0. \quad (4.14)$$

Так как вариация

$$\eta(x) = \int_{x_1}^x P(t) dt,$$

согласно равенству (4.13), обращается в нуль в x_1 и в x_2 , то мы можем положить в равенстве (4.14) $\eta' = P$; получим

$$\int_{x_1}^{x_2} P^2 dx = 0.$$

Следовательно, $P(x) \equiv 0$, что и требовалось доказать.

Второе необходимое условие можно получить из неравенства $I''(0) \geq 0$ [см. (4.11)]. Если мы продифференцируем $I(t)$ два раза по t , положим $t = 0$ и обозначим результат через $I_2(\eta)$, то получим неравенство

$$I''(0) = I_2(\eta) = \int_{x_1}^{x_2} 2\omega(x, \eta, \eta') dx \geq 0,$$

где

$$2\omega(x, \eta, \eta') = f_{yy}\eta^2 + 2f_{yy'}\eta\eta' + f_{y'y'}\eta'^2.$$

Интеграл $I_2(\eta)$ называется *второй вариацией* интеграла I вдоль кривой (4.7). Мы получили, таким образом, следующий результат:

Если кривая осуществляет относительный минимум интеграла I в классе кривых, соединяющих две точки, то вдоль нее вторая вариация $I_2(\eta)$ интеграла I удовлетворяет условию $I_2(\eta) \geq 0$ при любых вариациях $\eta(x)$, обращающихся в нуль в x_1 и x_2 .

Это условие иногда рассматривают как условие, указанное в п. (г) § 4.3. Однако в таком виде это условие не вполне удовлетворительно, так как его очень трудно проверить. Правда, на основании полученных результатов ему можно придать более простой вид, но мы не будем здесь на этом останавливаться.

§ 4.5. Специальные случаи. В этом параграфе мы сначала установим следующий результат:

Уравнения Эйлера для интеграла

$$I = \int_{x_1}^{x_2} h(x, y)(1 + y'^2)^{1/2} dx \quad (4.15)$$

могут быть приведены к виду

$$\frac{d}{ds}(h \cos \theta) = h_x, \quad \frac{d}{ds}(h \sin \theta) = h_y, \quad (4.16)$$

где s обозначает длину дуги, а θ — угол наклона. В том случае, когда $h = y^r$, т. е. когда

$$I = \int_{x_1}^{x_2} y^r (1 + y'^2)^{1/2} dx, \quad y > 0, \quad (4.17)$$

экстремали (решения уравнения Эйлера) определяются уравнениями

$$x = a + \frac{b}{r} \int_0^\theta (\sec t)^{1/r} dt, \quad y = b (\sec \theta)^{1/r}, \quad (4.18)$$

где a и b — произвольные постоянные. Другая форма записи семейства кривых (4.18) имеет вид

$$x = a + \frac{b}{r} \int_0^\varphi (\operatorname{ch} u)^{(1-r)/r} du, \quad y = b (\operatorname{ch} \varphi)^{1/r}. \quad (4.19)$$

Для интеграла вида (4.15) мы имеем

$$f(x, y, y') = h(x, y)(1 + y'^2)^{1/2}.$$

Отсюда

$$f_x = h_x(1 + y'^2)^{1/2}, \quad f_y = h_y(1 + y'^2)^{1/2}, \\ f_{y'} = h y'(1 + y'^2)^{-1/2}, \quad f - y' f_{y'} = h(1 + y'^2)^{-1/2}.$$

Уравнения Эйлера (4.6) имеют соответственно вид

$$\frac{d}{dx} \left[\frac{h}{(1 + y'^2)^{1/2}} \right] = h_x(1 + y'^2)^{1/2}, \quad \frac{d}{dx} \left[\frac{h y'}{(1 + y'^2)^{1/2}} \right] = h_y(1 + y'^2)^{1/2}.$$

Уравнения (4.16) следуют из этих уравнений, если заметить, что

$$\frac{d}{ds} = \frac{dx}{ds} \cdot \frac{d}{dx}$$

и

$$\frac{dx}{ds} = \cos \theta = (1 + y'^2)^{-1/2}, \quad \sin \theta = y'(1 + y'^2)^{-1/2}.$$

В том случае, когда $h = y^r$, имеем $h_x = 0$; тогда первое из уравнений (4.16) показывает, что

$$y^r \cos \theta = b^r, \quad y = b(\sec \theta)^{1/r},$$

где b — постоянная. Это уравнение вместе с уравнением, получающимся в результате интегрирования равенства

$$dx = \frac{dx}{dy} dy = \operatorname{ctg} \theta dy = \frac{b}{r} (\sec \theta)^{1/r} d\theta,$$

и представляют собой уравнения (4.18). Уравнения (4.19) получаются из (4.18), если выбрать φ так, чтобы выполнялось первое из соотношений

$$\operatorname{sh} \varphi = \operatorname{tg} \theta, \quad \operatorname{ch} \varphi = \sec \theta, \quad d\varphi = \sec \theta d\theta.$$

Два других соотношения следуют из первого.

Характер экстремалей семейства (4.18) существенно зависит от знака r . Например, преобразовав уравнение Эйлера, получим

$$y y'' = r(1 + y'^2).$$

Следовательно, y'' имеет тот же знак, что и r , если только y положительно, что мы предполагали в наших рассуждениях. Сущность этой зависимости от r мы сформулируем в двух предложениях, первое из которых следующее:

Если $r < 0$, то экстремали (4.18) интеграла (4.17) направлены вогнутостью вниз, а прямая $x = a$ является осью симметрии. Через каждую пару точек 1 и 2 в верхней полуплоскости, не лежащих на одной вертикали, проходит одна и только одна экстремаль. Эта экстремаль минимизирует интеграл (4.17) в классе кривых, соединяющих точки 1 и 2 и лежащих в верхней полуплоскости. Если $r = -1$, то экстремаль (4.18) является дугой окружности

$$x = a - b \sin \theta, \quad y = b \cos \theta \quad (4.20)$$

радиуса b с центром $(a, 0)$. Если $r = -1/2$, то экстремалью (4.18) является дуга циклоиды

$$x = a - \frac{b}{2}(2\theta + \sin 2\theta), \quad y = \frac{b}{2}(1 + \cos 2\theta), \quad (4.21)$$

описываемой точкой P окружности с диаметром b , катящейся по оси x . Эта дуга циклоиды есть одновременно и линия кратчайшего времени падения, т. е. брахистохрона.

Уравнения (4.20) и (4.21) получаются из формул (4.18), если положить в них соответственно $r = -1$, $r = -1/2$ и проинтегрировать получившиеся выражения. Очевидно, что через каждую пару точек верхней полуплоскости проходит только одна окружность, имеющая центр на оси x . Следовательно, в случае $r = -1$ существует единственная экстремаль, соединяющая данную пару точек, расположенных в верхней полуплоскости и не лежащих на одной вертикальной прямой. Результат остается верным и в случае любого $r < 0$, однако в этом случае доказательство значительно труднее и мы не будем приводить его здесь. Если предположить существование минимизирующей кривой в случае $r < 0$, то ею обязательно будет экстремаль, поскольку существует единственная кривая с заданными концевыми точками, которая удовлетворяет первому необходимому условию.

Совсем иначе обстоит дело, когда $r > 0$. Основной результат в этом случае может быть сформулирован следующим образом:

Если $r > 0$, то экстремали (4.19) или, что то же, экстремали (4.18) направлены вогнутостью вверх и прямая $x = a$ является их осью симметрии. Для заданной пары точек 1 и 2, расположенных в верхней полуплоскости и не лежащих на одной и той же вертикальной прямой, существуют две, одна или ни одной соединяющей их экстремали. Если существуют две экстремали, то на одной из них достигается относительный минимум интеграла I , а на другой нет. Если существует только одна экстремаль, то она не минимизирует интеграла. Если не существует экстремали, соединяющей точки 1 и 2, то над осью x не существует минимизирующей кривой. На экстремали может достигаться относительный минимум интеграла (4.17) без одновременного достижения абсолютного минимума. Если $r = 1/2$, то экстремалью (4.19) будет дуга параболы

$$y = b \left[1 + \frac{1}{4} \left(\frac{x-a}{b} \right)^2 \right], \quad (4.22)$$

директрисой которой является ось x . Если $r = 1$, то экстремалью будет дуга цепной линии

$$y = b \operatorname{ch} \frac{x-a}{c},$$

имеющая ось x своей директрисой. Поверхность вращения наименьшей площади может быть получена посредством вращения соответственно выбранной дуги цепной линии вокруг ее директрисы.

Мы не в состоянии доказать здесь эти факты. Однако главные обстоятельства можно увидеть уже в случае $r = 1/2$. Поэтому мы проиллюстрируем некоторые из них на этом частном случае. С этой целью выберем в качестве точки 1 точку $(0, 1)$, затем положим

$$b = \cos^2 A, \quad a = 2 \sin A \cos A.$$

Формула (4.22) дает

$$y = 1 - x \operatorname{tg} A + \frac{1}{4} x^2 \sec^2 A. \quad (4.23)$$

Это семейство представляет собой всю совокупность экстремалей, проходящих через точку $(0, 1)$ и имеющих параболу

$$y = \frac{x^2}{4} \quad (4.24)$$

в качестве своей огибающей, как это указано на рис. 4.4. Для того чтобы убедиться в этом, следует продифференцировать уравнение (4.23) по A , тогда получим

$$0 = -x \sec^2 A + \frac{1}{2} x^2 \sec^2 A \operatorname{tg} A \quad \text{или} \quad \operatorname{tg} A = \frac{2}{x}.$$

Подставляя это в формулу (4.23), мы и получим огибающую (4.24). Если точка 2 расположена под огибающей, то не существует ни одной экстремали, соединяющей точки 1 и 2. Если она лежит на огибающей, то существует единственная экстремаль, соединяющая

точки 1 и 2. Если же точка 2 лежит над огибающей, то существуют две экстремали, соединяющие точки 1 и 2, как это показано на рис. 4.4. Та из них, которая не имеет точки касания с огибающей между точками 1 и 2, реализует относительный минимум интеграла I , в то время как другая такого минимума не дает.

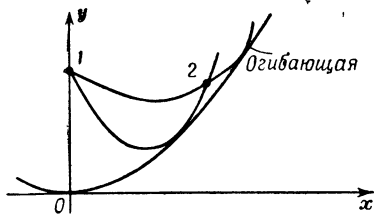


Рис. 4.4. Огибающая семейства кривых.

Вернемся теперь к более общему интегралу (4.15). В соответствии

с первым необходимым условием a , указанным в § 4.3, величины $h \cos \theta$ и $h \sin \theta$ (каждая из которых равна соответственно $f - y'f_y$ и $f_{y'}$) непрерывны вдоль минимизирующей кривой. Этот результат можно сформулировать в несколько другом виде:

Пусть C есть минимизирующая кривая интеграла (4.15). Если $\varphi(x, y)$ обозначает угол, который образует C с некоторым фиксированным направлением L в точке (x, y) кривой C , то величина $h \sin \varphi$ непрерывна вдоль C .

Это следует из того, что мы можем всегда так повернуть оси координат, что L будет параллельна одной из них. Вывод первого

необходимого условия показывает, что этот результат сохраняется и тогда, когда h имеет конечные разрывы вдоль одной или нескольких кривых, при условии, что минимизирующая дуга C не касается ни одной из этих кривых. Пусть D есть такая кривая разрыва и пусть 3 является точкой ее пересечения с C , как это указано на рис. 4.5. Обозначим через L нормаль к D в точке 3 . Тогда в соответствии с указанным свойством имеет место соотношение

$$h^- \sin \varphi^- = h^+ \sin \varphi^+,$$

где знак минус означает предельные значения слева, а знак плюс — предельные значения справа. В частности, если $h(x, y) = c/v(x, y)$, как в теории света, то

$$\frac{\sin \varphi^-}{v^-} = \frac{\sin \varphi^+}{v^+}.$$

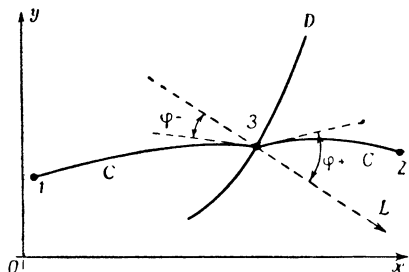


Рис. 4.5. Кривая разрывов. Оптический закон Снеллиуса.

Это оптический закон Снеллиуса. Именно это соотношение использовал Иван Бернулли при решении задачи о брахистохроне.

§ 4.6. Случай, когда подинтегральная функция имеет вид $f(x, y)$. В последнее время все большее значение начинает приобретать интеграл вида

$$I = \int_{x_1}^{x_2} f[x, y(x)] dx, \quad (4.25)$$

в котором на функцию $y(x)$ не накладывается требование непрерывности, а лишь предполагается ее кусочная непрерывность на отрезке $[x_1, x_2]$. Иногда приходится накладывать на $y(x)$ даже более слабые условия. Большая часть проблем этого типа, встречавшихся автору, имеет решения, которые частью лежат на границе рассматриваемой области. Этот вопрос слишком сложен, чтобы рассматривать его здесь. Однако при сравнительно слабых ограничениях верно следующее утверждение:

Если дуга

$$y = y(x), \quad x_1 \leq x \leq x_2$$

минимизирует интеграл (4.25) в области R плоскости (x, y) , то для каждого x имеет место неравенство

$$f(x, y) \geq f[x, y(x)], \quad (4.26)$$

если только величина y такова, что точка (x, y) находится в R .

Этот результат можно использовать в качестве руководящей идеи при изучении проблем этого типа. Из него следует, что во внутренних

точках условия

$$f_y = 0, \quad f_{yy} \geq 0$$

имеют место вдоль минимизирующей кривой. В случае большего числа измерений эти соотношения принимают вид

$$f_{y_i} = 0, \quad f_{y_i y_k} \pi_i \pi_k \geq 0, \quad (\pi) \neq (0).$$

Если в последнем неравенстве исключен знак равенства, то обычно минимизирующая кривая непрерывна.

Изучение только что описанной проблемы эквивалентно изучению интегралов типа

$$I = \int_{x_1}^{x_2} f[x, y'(x)] dx$$

при условии отсутствия ограничений на $y(x_2)$. Нужно только заменить $y'(x)$ на $y(x)$. Неравенство (4.26) в этом случае вытекает из условия Вейерштрасса, указанного в § 4.3. Таким образом, теория экстремалей этого интеграла может быть получена из теории экстремалей интеграла (4.5), если концевые точки могут перемещаться. Этот подход к изучению интеграла (4.25) не является, однако, удобным, гораздо лучше изучать интеграл (4.25) непосредственно.

§ 4.7. Принцип Гамильтона [10]. Один из важных принципов механики, обычно известный под названием *принципа Гамильтона*¹⁾, является вариационным принципом. При его описании мы будем пользоваться символикой, обычно применяемой в механике. Например, мы будем писать (q_1, q_2) вместо (x, y) , (q_1, q_2, q_3) вместо (x, y, z) и т. д. Временно мы будем предполагать, что q являются декартовыми координатами, хотя результаты, которые мы получим, справедливы и без этого ограничения. Кинетическая энергия частицы массы m определяется формулой

$$T = \frac{1}{2} m \dot{q}_i \dot{q}_i = \frac{1}{2} m (\dot{q}_1^2 + \dots + \dot{q}_n^2), \quad (4.27)$$

где точкой обозначены производные по времени, а повторяющийся индекс i означает суммирование относительно этого индекса.

Для консервативного поля сил существует потенциальная функция

$$U(q) = U(q_1, \dots, q_n),$$

такая, что уравнения движения частицы массы m имеют вид

$$m \ddot{q}_i = -U_{q_i}, \quad i = 1, 2, \dots, n. \quad (4.28)$$

Полагая

$$L(q, \dot{q}) = T - U,$$

¹⁾ В русской литературе этот принцип называют обычно принципом Гамильтона — Остроградского. — *Прим. ред.*

где T задана формулой (4.27), мы получаем

$$L_{\dot{q}_i} = m\dot{q}_i, \quad L_{q_i} = -U_{q_i},$$

и поэтому уравнения (4.28) принимают вид

$$\frac{dL_{\dot{q}_i}}{dt} = L_{q_i}, \quad i = 1, \dots, n. \quad (4.29)$$

Эти уравнения являются уравнениями Эйлера для интеграла

$$I = \int_{t_1}^{t_2} L(q, \dot{q}) dt. \quad (4.30)$$

Полученный результат подсказывает следующий принцип:

Принцип Гамильтона. В консервативном поле сил частица движется так, что минимизирует (за достаточно короткий интервал времени) интеграл действия (4.30).

Мы показали лишь, что траектория точки является экстремалью интеграла действия. Можно показать, что за достаточно короткий промежуток времени такие экстремали являются путями наименьшего действия. Для достаточно большого промежутка времени это может быть неверно, как мы вскоре увидим. Нижеследующее предложение, пожалуй, является более подходящей формулировкой принципа Гамильтона.

В консервативном поле сил траектории являются экстремальми интеграла действия (4.30).

Как было указано ранее, уравнение

$$\frac{d}{dt}(L - \dot{q}_i L_{\dot{q}_i}) = L_t$$

является следствием уравнений (4.29). Так как

$$q_i L_{q_i} - L = T + U, \quad L_t = 0,$$

то сумма $T + U = E = \text{const}$ Отсюда получаем:

Вдоль траектории в консервативном поле сил сумма кинетической и потенциальной энергии равна постоянной E , т. е.

$$T + U = E. \quad (4.31)$$

Учитывая этот результат, мы можем ограничить наше внимание рассмотрением путей, вдоль которых выполняется соотношение (4.31). Вдоль такого пути мы имеем

$$L = T - U = 2T - E.$$

Так как E является постоянной, то отсюда следует, что траектория минимизирует интеграл

$$\int_{t_1}^{t_2} 2T dt$$

за достаточно короткий промежуток времени. Введя

$$ds = (2T)^{1/2} dt$$

в качестве элемента длины дуги, мы получим, что этот интеграл может быть представлен в виде

$$I = \int_{t_1}^{t_2} (E - U)^{1/2} ds \quad (4.32)$$

с точностью до постоянного множителя. Таким образом (короткие) траектории минимизируют преобразованный *интеграл действия* (4.32). Этот результат обычно именуется *принципом наименьшего действия*.

В предыдущих рассмотрениях мы молчаливо предполагали, что T имеет вид (4.27) и что q являются декартовыми координатами. Однако наши результаты не зависят от применяемой координатной системы. Поэтому мы можем взять T в более общем виде

$$T = g_{ij}(q) \dot{q}_i \dot{q}_j, \quad i, j = 1, \dots, n,$$

предполагая, что эта квадратичная форма является положительно определенной. При допустимых преобразованиях системы координат выражение (4.27) всегда преобразуется в выражение этого же типа.

Сравнивая принцип наименьшего действия в механике с принципом Ферма в теории света и с теорией геодезических линий, мы видим, что все они приводят к одной и той же математической теории, и именно вариационное исчисление дает метод объединения этих различных теорий.

Выше отмечалось, что достаточно длинные траектории, вообще говоря, не минимизируют интеграл действия. Этот факт можно проиллюстрировать на простом примере. Рассмотрим поле тяготения в вертикальной плоскости (x, y) , в которой ось x направлена горизонтально, а ось y направлена по вертикали вверх. В качестве потенциальной функции для частицы единичной массы можно взять $U = gy$. Пусть из начала координат под положительным углом к оси x , отличным от прямого, вылетает пуля с начальной скоростью v_0 . Тогда $E = \frac{1}{2} v_0^2$ и интеграл действия (4.32) принимает вид

$$\int_{t_1}^{t_2} (E - gy)^{1/2} (\dot{x}^2 + \dot{y}^2)^{1/2} dt. \quad (4.33)$$

Этот интеграл похож на интеграл (4.17) при $r = 1/2$. Следовательно, траектории имеют вид

$$\frac{E}{g} - y = b \left[1 + \frac{1}{4} \left(\frac{x-a}{b} \right)^2 \right].$$

Частные значения

$$b = \frac{E}{g} \cos^2 A, \quad a = \frac{2E}{g} \sin A \cos A$$

дают семейство траекторий

$$y = x \operatorname{tg} A - \frac{g}{4E} x^2 \sec^2 A,$$

проходящих через начало координат с начальным углом наклона A . Это семейство имеет в качестве огибающей параболу

$$y = \frac{E}{g} - \frac{g}{4E} x^2.$$

Частица, вылетающая вверх из начала координат, не может пересечь эту огибающую. Как хорошо известно, через точку 2, находящуюся под огибающей, проходят два пути рассматриваемого семейства с углами наклона A_1 и A_2 , где, например, $A_1 > A_2$. Путь, определяемый углом A_2 , реализует относительный минимум интеграла действия (4.33), в то время как второй путь не дает такого минимума. В самом деле, если точка 2 достаточно близка к огибающей, то существуют пути под прямой линией $y = E/g$, на которых интеграл действия принимает меньшее значение, чем на каждой из этих двух траекторий. Этот результат показывает, что принцип наименьшего действия как минимальный принцип применим только к кривым, которые достаточно коротки. Тем не менее этот принцип является очень важным принципом механики.

Мы закончим этот параграф одной иллюстрацией принципа Гамильтона. Рассмотрим частицу, движущуюся под действием центральной силы, обратно пропорциональной квадрату расстояния от центра. Предполагая движение плоским и пользуясь полярными координатами с полюсом в центре сил, будем иметь

$$T = \frac{c}{r} (\dot{r}^2 + r^2 \dot{\theta}^2), \quad U = -\frac{c}{r}.$$

Полагая $L = T - U$, находим, что уравнения Эйлера получают вид

$$\frac{d}{dt} L_{\dot{r}} = L_r, \quad \frac{d}{dt} L_{\dot{\theta}} = L_{\theta}$$

и что

$$T + U = E = \text{const.}$$

Так как L не зависит от θ , то второе уравнение Эйлера дает

$$L_{\dot{\theta}} = m r^2 \dot{\theta} = \text{const.}, \quad \text{или} \quad r^2 \dot{\theta} = b.$$

Так как $\dot{r} = \theta dr/d\theta$, мы видим, что

$$T = \frac{m}{2} \left[\left(\frac{dr}{d\theta} \right)^2 + r^2 \right] \dot{\theta}^2 = \frac{m}{2} \left[\left(\frac{dr}{d\theta} \right)^2 + r^2 \right] \frac{b^2}{r^4}.$$

Предполагая, что $b \neq 0$, из соотношения $T + U = E$ получаем

$$a^2 \left[\left(\frac{dr}{d\theta} \right)^2 + r^2 \right] = Er^4 + cr^3, \quad a^2 = \frac{mb^2}{2},$$

откуда после интегрирования получаем решение в виде

$$r = \frac{k}{1 - e \cos(\theta - \gamma)},$$

где k , e и γ — постоянные. Траектория, таким образом, является одним из конических сечений.

§ 4.8. Гамильтонианы. В механике часто проще пользоваться *гамильтонианом*

$$H = \dot{q}_i L_{\dot{q}_i} - L = T + U \quad (4.34)$$

вместо лагранжиана L . В этом случае удобнее рассматривать *моменты*

$$p_i = L_{\dot{q}_i}(q, \dot{q}), \quad i = 1, 2, \dots, n \quad (4.35)$$

вместо переменных $\dot{q}_i$. Если мы обозначим решения уравнений (4.35) через

$$\dot{q}_i = Q_i(q, p) \quad (4.36)$$

и подставим результат в равенство (4.34), то получим выражение

$$H(q, p) = Q_i p_i - L(q, Q)$$

гамильтониана H через *канонические переменные* q_i, p_i . С помощью преобразования (4.36) нетрудно получить

$$H_{q_i} = -L_{q_i}, \quad H_{p_i} = Q_i = \dot{q}_i.$$

Применяя эти соотношения, находим, что уравнения Эйлера — Лагранжа (4.29) можно записать в канонической форме

$$\dot{q}_i = H_{p_i}, \quad \dot{p}_i = -H_{q_i}, \quad i = 1, \dots, n. \quad (4.37)$$

Эти уравнения являются основными в механике. Интересно отметить, что если мы будем рассматривать функции

$$q_i(t), p_i(t), t_1 \leq t \leq t_2, \quad i = 1, \dots, n \quad (4.38)$$

как произвольно заданные [а не подчиненные условию (4.35)], то уравнения (4.37) являются уравнениями Эйлера для интеграла

$$\int_{t_1}^{t_2} [\dot{q}_i p_i - H(q, p)] dt. \quad (4.39)$$

Таким образом, мы получаем другую форму принципа Гамильтона:

Траектории (4.38) в консервативном поле сил являются стационарными кривыми интеграла (4.39), в котором $H(q, p)$ есть соответствующий гамильтониан. Вдоль стационарной кривой H является постоянным.

Термин „стационарный“ употреблен вместо термина „экстремальный“, так как очевидно, что интеграл (4.39) не имеет ни минимизирующих, ни максимизирующих кривых, если p и q являются независимыми переменными.

Если T задается в виде (4.27), то

$$p_i = m \dot{q}_i, \quad T = \frac{p_i p_i}{2m}$$

и

$$H(q, p) = \frac{p_i p_i}{2m} + U(q).$$

В этом случае уравнения (4.37) принимают простой вид

$$\dot{q}_i = \frac{p_i}{m}, \quad \dot{p}_i = -U_{q_i}.$$

Интересно отметить, что, хотя рассматриваемый вариационный принцип не является принципом минимума или максимума, он может быть сформулирован как минимаксный принцип. Предположим, что мы рассматриваем фиксированный путь

$$q_i = q_i(t), \quad t_1 \leq t \leq t_2, \quad i = 1, \dots, n.$$

Вдоль этого пути интеграл (4.39) имеет вид

$$\int_{t_1}^{t_2} f[t, p(t)] dt,$$

где

$$f(t, p) = \dot{q}_i(t) p_i - H[q(t), p].$$

Если p — независимые переменные, то этот интеграл принимает максимальные значения в том случае, когда функция

$$p_i(t), \quad t_1 \leq t \leq t_2$$

выбрана так, что она удовлетворяет условиям

$$0 = f_{p_i} = \dot{q}_i - H_{p_i}. \quad (4.40)$$

Существование этого максимума обеспечивается тем обстоятельством, что f является полиномом второй степени относительно переменных p , у которого члены второго порядка образуют отрицательно определенную квадратичную форму. Принимая во внимание, что уравнения (4.40) и (4.35) эквивалентны, мы видим из уравнения (4.34), что

$$\max_{p_i} \int_{t_1}^{t_2} [\dot{q}_i p_i - H(q, p)] dt = \int_{t_1}^{t_2} L(q, \dot{q}) dt.$$

Поэтому интеграл действия (4.30) является максимальным значением интеграла (4.39) вдоль каждого фиксированного пути. Комбинируя этот результат со сформулированным в § 4.7 принципом Гамильтона, мы получаем следующий минимаксный принцип:

Для достаточно короткого промежутка времени траектории (4.38) в консервативном поле сил являются решениями минимаксной задачи

$$\min_{q_i} \max_{p_i} \int_{t_1}^{t_2} [\dot{q}_i p_i - H(q, p)] dt.$$

Значение этого результата, конечно, не ограничивается механикой. Важнейшие вариационные проблемы могут быть сформулированы аналогичным образом в виде подобных минимаксных принципов. При таком подходе теория полей оказывается более естественной, чем при прежнем способе изложения.

§ 4.9. Изопериметрическая задача. Одной из старейших задач вариационного исчисления является задача об отыскании вида простой замкнутой кривой данной длины, ограничивающей наибольшую площадь. Аналитически эта задача состоит в отыскании максимума интеграла

$$I(C) = \frac{1}{2} \int_{t_1}^{t_2} (x\dot{y} - y\dot{x}) dt$$

в классе простых замкнутых кривых

С: $x = x(t), \quad y = y(t), \quad t_1 \leq t \leq t_2,$

имеющих постоянную длину

$$J(C) = \int_{t_1}^{t_2} (\dot{x}^2 + \dot{y}^2)^{1/2} dt = \text{const.}$$

В этом виде задача не может иметь единственного решения, так как перенос решения снова дает решение. Однако это обстоятельство несущественно для наших целей. Мы покажем формальным путем, что решением является окружность. Пусть кривая C есть решение,

существование которого мы предполагаем. Так как J постоянно, то нас будут интересовать только вариации δx , δy , при которых первая вариация δJ интеграла J равна нулю вдоль C . Из того, что I принимает максимальное значение на C , следует, что первая вариация δI интеграла I равна нулю вдоль C при условии, что $\delta J = 0$. Вследствие линейности это возможно только в том случае, когда δI пропорционально δJ , т. е. когда

$$\delta I = \lambda \delta J \quad \text{или} \quad \delta(I - \lambda J) = 0.$$

Полагая

$$F = \frac{1}{2} (x\dot{y} - y\dot{x}) = \lambda (\dot{x}^2 + \dot{y}^2)^{1/2},$$

мы видим, что интеграл

$$I - \lambda J = \int_{t_1}^{t_2} F dt$$

обладает тем свойством, что его первая вариация $\delta(I - \lambda J)$ равна нулю вдоль C . Отсюда следует, что C должно удовлетворять уравнениям Эйлера

$$\frac{d}{dt} F_{\dot{x}} = F_x, \quad \frac{d}{dt} F_{\dot{y}} = F_y.$$

Следовательно,

$$-\frac{d}{dt} \left[\frac{y}{2} + \frac{\lambda \dot{x}}{(\dot{x}^2 + \dot{y}^2)^{1/2}} \right] = \frac{\dot{y}}{2}, \quad \frac{d}{dt} \left[\frac{x}{2} - \frac{\lambda \dot{y}}{(\dot{x}^2 + \dot{y}^2)^{1/2}} \right] = -\frac{\dot{x}}{2}. \quad (4.41)$$

Если выбрать длину дуги в качестве параметра, то

$$\dot{x}^2 + \dot{y}^2 = 1$$

и уравнения (4.41) принимают простой вид

$$-\lambda \ddot{x} = \dot{y}, \quad \lambda \ddot{y} = \dot{x}.$$

Умножив первое уравнение на $\dot{y}/\lambda$, второе на $\dot{x}/\lambda$ и сложив, получим соотношение

$$\dot{x}\ddot{y} - \ddot{x}\dot{y} = \frac{1}{\lambda} = \text{const.}$$

Выражение, стоящее в левой части, является кривизной кривой C . Поэтому решение нашей задачи должно иметь постоянную кривизну $1/\lambda$ и оно является окружностью радиуса λ , что и требовалось доказать.

Описанная здесь задача является частным случаем широкого класса задач, обычно называемых *изопериметрическими задачами*. Эти задачи состоят в отыскании максимума или минимума интеграла I в классе кривых, вдоль которых некоторые заданные интегралы

$J_1, \dots, J_p$ принимают фиксированные значения. Эти задачи могут формулироваться как в параметрическом, так и в непараметрическом виде.

По-видимому, простейшей изопериметрической задачей является минимизирование интеграла

$$I(C) = \int_{x_1}^{x_2} f(x, y, y') dx$$

в классе кривых вида

$$C: \quad y = y(x), \quad x_1 \leq x \leq x_2,$$

которые соединяют в плоскости (x, y) две фиксированные точки 1 и 2 и удовлетворяют единственному изопериметрическому условию

$$J(C) = \int_{x_1}^{x_2} g(x, y, y') dx = \text{const.}$$

Кроме некоторых особых случаев, минимизирующая кривая должна удовлетворять соответствующему уравнению Эйлера

$$\frac{d}{dx} F_{y'} = F_y, \quad (4.42)$$

где

$$F(x, y, y') = f(x, y, y') - \lambda g(x, y, y'),$$

а λ есть подходящим образом выбранная постоянная. Следует отметить, что если пытаться осуществить минимум или максимум $J(C)$, сохраняя постоянным $I(C)$, то уравнение (4.42) будет выполняться при $F = g - \mu f$. Соответствующее уравнение можно получить, деля уравнение (4.42), в том виде как оно написано, на λ и полагая $\mu = 1/\lambda$. Все это приводит к следующему принципу.

Принцип взаимности. Если C минимизирует $I(C)$ при условии $J(C) = \text{const}$, то обычно C максимизирует или минимизирует $J(C)$ при условии $I(C) = \text{const}$, смотря по тому, будет ли множитель λ , появляющийся в уравнении (4.42), положительным или отрицательным.

Нетрудно построить примеры, для которых принцип взаимности не имеет места.

Уравнение (4.42) представляет собой уравнение Эйлера для

$$I - \lambda J = \int_{x_1}^{x_2} F dx.$$

Из одного этого факта нельзя заключить, что кривая C минимизирует $I - \lambda J$. Это можно хорошо проиллюстрировать на следующем

примере. Как известно, подвешенная цепь принимает такое положение, при котором ее центр тяжести принимает самое низкое положение из всех возможных. Таким образом, абсолютно гибкую цепь постоянной плотности и заданной длины можно использовать для нахождения решения следующей задачи (в предположении, что ось y вертикальна): среди всех кривых C , соединяющих две фиксированные точки 1 и 2 в плоскости (x, y) (см. рис. 4.6) и имеющих данную длину

$$J(C) = \int_{x_1}^{x_2} (1 + y'^2)^{1/2} dx = l,$$

найти кривую, которая минимизирует интеграл

$$I(C) = \int_{x_1}^{x_2} y(1 + y'^2)^{1/2} dx.$$

Как отмечалось выше, решение C этой задачи должно удовлетворять уравнениям Эйлера для интеграла

$$I - \lambda J = \int_{x_1}^{x_2} (y - \lambda)(1 + y'^2)^{1/2} dx,$$

где λ есть соответствующим образом выбранная постоянная. Эта кривая является цепной линией с директрисой $y = \lambda$. Как было отмечено в § 4.5, для случая $\lambda = 0$ существуют две цепные линии, соединяющие точки 1 и 2 и имеющие прямую $y = \lambda$ в качестве своей директрисы. Более длинная из них не минимизирует $I - \lambda J$, но минимизирует I в классе кривых, соединяющих концевые точки и имеющих одну и ту же длину.

§ 4.10. Задачи с подвижными концами. До сих пор мы имели дело главным образом с задачей минимизирования интеграла

$$I(C) = \int_{x_1}^{x_2} f(x, y, y') dx$$

в классе кривых

$$C: \quad y = y(x), \quad x_1 \leq x \leq x_2,$$

соединяющих две фиксированные точки 1 и 2. Если мы теперь предположим, что 2 есть произвольная точка на некоторой кривой

$$D: \quad x = X(u), \quad y = Y(u),$$

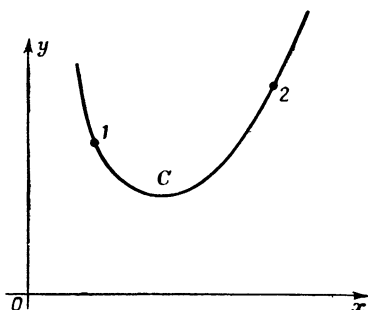


Рис. 4.6. Абсолютно гибкая тяжелая цепь с постоянной плотностью: цепная линия.

как это указано на рис. 4.7, то придем к двум новым понятиям, а именно к понятию *трансверсальности* и к понятию *фокальной точки*. Чтобы описать эти понятия, напомним сначала некоторые факты, касающиеся задачи отыскания кратчайшего прямолинейного отрезка, соединяющего фиксированную точку 1 с соответствующей

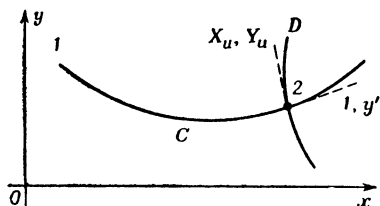


Рис. 4.7. Иллюстрация понятия трансверсальности (обобщение ортогональности).

точкой 2 на кривой D . Если C есть такой отрезок, то он должен обладать следующими двумя свойствами:

а) отрезок C перпендикулярен к кривой D в точке 2;

б) центр кривизны кривой D в точке 2 не может лежать на отрезке C между точками 1 и 2.

Чтобы убедиться в этом, достаточно нарисовать соответствующий чертеж.

В вариационном исчислении аналогом ортогональности и центра кривизны служат понятия трансверсальности и фокальных точек. Условие трансверсальности может быть сформулировано следующим образом.

Если дуга C минимизирует интеграл I в классе кривых, соединяющих точку 1 с точкой, лежащей на кривой D , то в точке 2 пересечения кривых C и D должно выполняться условие

$$(f - y'f_{y'})X_u + f_{y'}Y_u = 0, \quad (4.43)$$

где $(1, y')$ и (X_u, Y_u) обозначают векторы, касательные соответственно к C и D , как это указано на рис. 4.7.

Если $f = h(x, y)(1 + y'^2)^{1/2}$, то условие (4.43) эквивалентно условию

$$X_u + y'Y_u = 0,$$

если только $h \neq 0$. Следовательно, в этом случае трансверсальность превращается в ортогональность.

Чтобы получить уравнение (4.43), рассмотрим однопараметрическое семейство кривых

$$y = y(x, u), \quad x_1 \leq x \leq X(u),$$

соединяющих точку 1 с точкой $[X(u), Y(u)]$ на кривой D , и предположим, что при $u = u_0$ соответствующая кривая совпадает с C . Вдоль кривых этого семейства интеграл I принимает соответственно значения

$$I(u) = \int_{x_1}^{X(u)} f[x, y(x, u), y_x(x, u)] dx.$$

Так как C есть минимизирующая кривая, то мы будем иметь

$$I(u) \geq I(u_0) = \text{значение } I \text{ на } C.$$

Следовательно, $I'(u_0) = 0$, т. е.

$$I'(u_0) = fX_u|_{x_1}^{x_2} + \int_{x_1}^{x_2} (f_y y_u + f_{y'} y_{xu}) dx = 0,$$

где аргументы функции f , f_y и $f_{y'}$ берутся вдоль кривой C . Так как кривая C минимизирует интеграл I в классе кривых, соединяющих ее концевые точки, то вдоль C

$$\frac{d}{dx} f_{y'} = f_y.$$

Воспользовавшись этим, мы получим

$$I'(u_0) = fX_u|_{x_1}^{x_2} + f_{y'} y_u|_{x_1}^{x_2} = 0.$$

Но

$$y(x_1, u) = y_1, \quad y[X(u), u] = Y(u).$$

Дифференцируя по u , мы видим, что

$$y_u = 0 \text{ при } x = x_1,$$

$$y_x X_u + y_u = Y_u \text{ при } x = x_2 = X(u_0).$$

Следовательно, полученное выше выражение для $I'(u_0)$ приобретает вид

$$I'(u_0) = [(f - y'f_{y'}) X_u + f_{y'} Y_u]_{x=x_2} = 0,$$

что и требовалось доказать.

Остается описать обобщение понятия центра кривизны. Напомним, что геометрическое место центров кривизны является огибающей нормалей к D . Таким образом, центр кривизны кривой D на нормали C есть точка касания нормали C с этой огибающей. Это обстоятельство наводит на мысль заменить семейство нормалей к D семейством экстремалей, трансверсальных к D . Это семейство, вообще говоря, имеет огибающую E , как это указано на рис. 4.8. Рассмотрим экстремаль C этого семейства. Точка 3 касания кривых C и E называется *фокальной точкой* кривой D . Если дуга C_{12} кривой C является решением рассмотренной здесь задачи, то фокальная точка 3 не может лежать между точками 1 и 2. Доказательство этого свойства может быть получено как обобщение доказательства, приведенного для соответствующего результата в случае задачи о кратчайшей линии.

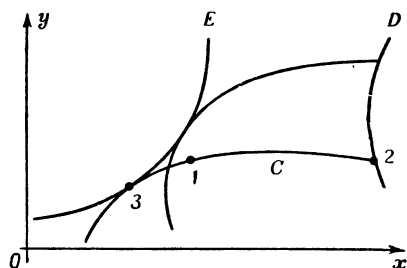


Рис. 4.8. Иллюстрация понятия фокальной точки (обобщение центра кривизны).

§ 4.11. Минимумы функций от интегралов. В предыдущем параграфе мы рассматривали минимумы отдельных интегралов. Полученные результаты легко распространяются на функции от интегралов. Чтобы проиллюстрировать этот метод, рассмотрим два интеграла

$$I(C) = \int_{x_1}^{x_2} f(x, y, y') dx, \quad J(C) = \int_{x_1}^{x_2} g(x, y, y') dx.$$

Попытаемся минимизировать отношение

$$\frac{I(C)}{J(C)}$$

в классе кривых

$$y = y(x), \quad x_1 \leq x \leq x_2,$$

соединяющих две фиксированные точки 1 и 2 в плоскости (x, y) . Пусть C_0 является минимизирующей дугой. Тогда первая вариация

$$\delta \frac{I}{J} = \frac{J \delta I - I \delta J}{J^2}$$

должна обращаться в нуль для всех вариаций δy , обращающихся в нуль в точках x_1 и x_2 . Полагая

$$\lambda = \frac{I(C_0)}{J(C_0)}, \quad (4.44)$$

мы видим, что

$$\delta I - \lambda \delta J = \delta \int_{x_1}^{x_2} (f - \lambda g) dx = 0$$

вдоль C_0 для всех указанных вариаций δy . Следовательно, C_0 должна удовлетворять уравнению Эйлера

$$\frac{d}{dx} (f_{y'} - \lambda g_{y'}) = f_y - \lambda g_y.$$

Это можно записать в другом виде

$$\frac{d}{dx} f_{y'} - f_y = \lambda \left(\frac{d}{dx} g_{y'} - g_y \right). \quad (4.45)$$

Чтобы проиллюстрировать этот результат, рассмотрим задачу о минимизировании отношения

$$\frac{\int_0^1 y'^2 dx}{\int_0^1 y^2 dx}$$

в классе кривых $y = y(x)$ ($0 \leq x \leq 1$), таких, что

$$y(0) = 0, \quad y(1) = 0. \quad (4.46)$$

Здесь $f = y'^2$, а $g = y^2$. Следовательно, уравнение (4.45) принимает вид

$$y'' + \lambda y = 0.$$

Решения этого уравнения, удовлетворяющие условиям (4.46), имеют вид

$$y = \sin \lambda^{1/2} x,$$

где λ принимает одно из значений $\pi^2, 4\pi^2, 9\pi^2, \dots, n^2\pi^2, \dots$. Кроме того,

$$\frac{\int_0^1 n^2 \pi^2 \cos^2 n\pi x \, dx}{\int_0^1 \sin^2 n\pi x \, dx} = n^2 \pi^2,$$

что согласуется с равенством (4.44). Это отношение имеет минимум при $n = 1$.

Можно легко поставить задачу этого типа, которая не будет иметь решения. Например, отношение

$$\frac{\int_{x_1}^{x_2} y (1 + y'^2)^{1/2} \, dx}{\int_{x_1}^{x_2} (1 + y'^2)^{1/2} \, dx}$$

представляет собой координату y центра тяжести дуги C , соединяющей точку (x_1, y_1) с точкой (x_2, y_2) . Очевидно, что это отношение не имеет минимума в классе всех кривых, соединяющих эти точки. Если взять уравнение Эйлера, соответствующее формуле (4.45), то решение его будет иметь вид

$$y = \lambda + b \operatorname{ch} \frac{x - a}{b}.$$

Прямая $y = \lambda$ является директрисой этой цепной линии, и потому λ не может быть координатой центра тяжести, как это должно было бы быть согласно равенству (4.44). Уравнения (4.44) и (4.45) для этой задачи не имеют совместного решения.

§ 4.12. Задача Больца. Автор на опыте убедился, что большое число задач, которые возникают в приложениях, не принадлежит к типам, описанным на предшествующих страницах. Многие из них, однако, являются частными случаями более общей задачи, называемой *задачей Больца*. Эта задача может быть сформулирована несколькими способами, которые на первый взгляд кажутся различными.

Пожалуй, одна из наиболее плодотворных формулировок этой задачи следующая:

Рассмотрим класс кривых, координаты $w_1, \dots, w_{r+n}$ точек которых определяются уравнениями

$$w_h = a_h, \quad w_{r+i} = y_i(x), \quad x_1 \leq x \leq x_2, \quad h = 1, \dots, r, \quad i = 1, \dots, n$$

(первые r координат — некоторые постоянные). На эти кривые накладывается ряд ограничений

$$\varphi_j(a, x, y, y') = 0, \quad j = 1, \dots, m < n,$$

$$I_k = g_k[a, x_1, y(x_1), x_2, y(x_2)] + \int_{x_1}^{x_2} f_k(a, x, y, y') dx = 0,$$

$$k = 1, \dots, p,$$

где $a \equiv (a_1, a_2, \dots, a_n)$. Задача состоит в отыскании кривой из этого класса, которая минимизирует функционал

$$I = g[a, x_1, y(x_1), x_2, y(x_2)] + \int_{x_1}^{x_2} f(a, x, y, y') dx.$$

Эту задачу мы будем называть *задачей А*.

Вторая формулировка проблемы Больца, используемая в ряде случаев, может быть выражена следующим образом:

Рассмотрим класс кривых, координаты точек которых $w_1, \dots, w_{r+n+m}$ определяются функциями

$$w_h = a_h, \quad w_{r+i} = y_i(x), \quad w_{r+n+j} = \theta_j(x) \\ x_1 \leq x \leq x_2, \quad h = 1, \dots, r, \quad i = 1, \dots, n, \quad j = 1, \dots, m,$$

удовлетворяющими условиям

$$y'_i = P_i(a, x, y, \theta),$$

$$x_s = X_s(a), \quad y_i(x_s) = Y_{is}(a), \quad s = 1, 2,$$

$$I_k = g_k(a) + \int_{x_1}^{x_2} f_k(a, x, y, \theta) dx = 0, \quad k = 1, 2, \dots, p.$$

Задача состоит в минимизировании функционала

$$I = g(a) + \int_{x_1}^{x_2} f(a, x, y, \theta) dx.$$

Эту задачу мы будем называть *задачей В*.

Мы не будем доказывать, что задачи *А* и *В* эквивалентны, хотя на самом деле это имеет место. Теорию этих задач можно найти в книге Г. А. Блисса по вариационному исчислению [1]. Задача,

рассмотренная Блоссом, представляет собой простое видоизменение задачи А. Полезно привести здесь формулировки первого необходимого условия минимума для задач А и В. Эти результаты имеют место при некоторых очень слабых допущениях, на которых мы не будем останавливаться.

Первое необходимое условие для задачи А. Для минимизирующей кривой C существуют функции

$$F = \lambda_0 f + \lambda_k f_k + \mu_j(x) \varphi_j, \quad G = \lambda_0 g + \lambda_k g_k,$$

такие, что

а) множители $\lambda_0 \geq 0$, λ_k — постоянные;

б) величины $F - y'_i F_{y'_i}$ и $F_{y'_i}$ непрерывны вдоль кривой C , а между угловыми точками этой кривой удовлетворяются уравнения Эйлера — Лагранжа

$$\frac{d}{dx} (F - y'_i F_{y'_i}) = F_x, \quad \frac{dF_{y'_i}}{dx} = F_{y_i}, \quad \varphi_j = 0;$$

в) концевые точки 1 и 2 кривой C таковы, что равенство

$$(F - y'_i F_{y'_i}) dx + F_{y'_i} dy_i \Big|_1^2 + dG + \int_{x_1}^{x_2} F_{a_h} da_h dx = 0$$

выполняется тождественно для dx_1 , dy_{i1} , dx_2 , dy_{i2} , da_h .

Первое необходимое условие для задачи В. Первое необходимое условие для задачи В можно сформулировать в несколько другой форме. В этом случае мы вводим функцию H , которая является аналогом гамильтониана; а именно для минимизирующей дуги C существуют функции

$$H = \lambda_0 f + \lambda_k f_k + z_i(x) P_i, \quad G = \lambda_0 g + \lambda_k g_k,$$

такие, что

а) множители $\lambda_0 \geq 0$, λ_k — постоянные;

б) функции H и z_i непрерывны вдоль кривой C , а между угловыми точками этой кривой удовлетворяются уравнения Эйлера — Лагранжа

$$\frac{dH}{dx} = H_x, \quad \frac{dy_i}{dx} = H_{z_i}, \quad \frac{dz_i}{dx} = -H_{y_i}, \quad H_{\theta_j} = 0;$$

в) концевые точки 1 и 2 кривой C таковы, что имеет место равенство

$$[HX_{sh} - z_i Y_{ish}]_{s=1}^{s=2} + G_h + \int_{x_1}^{x_2} H_h dx = 0,$$

где нижний индекс h обозначает частные производные по a_h .

В описанных только что условиях, за исключением особых случаев, всегда можно считать $\lambda_0 = 1$.

Задача, рассмотренная в конце § 4.2, является простым частным случаем задачи В. Если мы отождествим переменные $t, x, y, \dot{x}, \dot{y}$, употреблявшиеся при исследовании этой задачи, с переменными t, y_1, y_2, y_3, y_4 , то задача сведется к отысканию кривой, минимизирующей интеграл I в классе кривых

$$T: \quad y_i = y_i(t), \quad \varphi = \varphi(t), \quad t_1 \leq t \leq t_2, \quad i = 1, 2, 3, 4,$$

удовлетворяющих условиям

$$\begin{aligned} \dot{y}_1 &= y_3, & \dot{y}_2 &= y_4, & \dot{y}_3 &= F \cos \varphi, & \dot{y}_4 &= F \sin \varphi - g, \\ t_1 &= 0, & t_2 &= T, & y_i(t_s) &= Y_{is}, & s &= 1, 2. \end{aligned}$$

Здесь величины Y_{is} постоянны. В этом случае мы имеем

$$H = z_1 y_3 + z_2 y_4 + z_3 F \cos \varphi + z_4 (F \sin \varphi - g).$$

Из указанного выше условия (б) получаем соотношения

$$\begin{aligned} \frac{dH}{dt} &= 0, & \frac{dz_1}{dt} &= 0, & \frac{dz_2}{dt} &= 0, & \frac{dz_3}{dt} &= -z_1, & \frac{dz_4}{dt} &= -z_2 \\ H_\varphi &= -F \sin \varphi + F \cos \varphi = 0. \end{aligned}$$

Возвращаясь к первоначальным переменным и заменяя z_3, z_4 на u, v , мы видим из этих уравнений, что

$$u = at + b = p \cos \varphi, \quad v = ct + d = p \sin \varphi, \quad (4.47)$$

$$H = \dot{u}\dot{x} + \dot{v}\dot{y} + u\ddot{x} + v\ddot{y} = \frac{d}{dt}(u\dot{x} + v\dot{y}) = h, \quad (4.48)$$

где a, b, c, d, h суть постоянные, а p — множитель пропорциональности. Свойства траекторий, описанные в § 4.2, являются интерпретацией формулы (4.47). Кроме того, в силу уравнения (4.48) получаем соотношение

$$u\dot{x} + v\dot{y} = ht + k,$$

которое дает нам некоторые сведения относительно скорости ракеты. Теперь уже нетрудно найти и траекторию ракеты, так как она вычисляется с помощью квадратур; однако мы не будем здесь на этом останавливаться.

§ 4.13. Задачи с кратными интегралами. Мы закончим эту главу, сделав несколько замечаний о задачах с кратными интегралами. Математическая теория этих задач не так проста, как теория задач, рассмотренных раньше. Однако сравнительно нетрудно получить уравнения Эйлера, которым должно удовлетворять решение. Мы

проиллюстрируем это формальным образом на задаче для двойного интеграла в пространстве (x, y, z) . Попытаемся минимизировать интеграл

$$I = \iint_A f(x, y, z, p, q) dx dy, \quad p = \frac{\partial z}{\partial x}, \quad q = \frac{\partial z}{\partial y}$$

в классе поверхностей

$$z = z(x, y), \quad (x, y) \in A,$$

где функции z имеют одни и те же граничные значения. Поступая как и выше, мы видим, что первая вариация интеграла I вдоль минимизирующей поверхности при всех вариациях z , обращающихся в нуль на границе A , должна удовлетворять условию

$$\delta I = \iint_A (f_s \delta z + f_p \delta z_x + f_q \delta z_y) dx dy = 0.$$

Согласно формуле Грина

$$I = \iint_A \left(f_z - \frac{\partial}{\partial x} f_p - \frac{\partial}{\partial y} f_q \right) \delta z dx dy = 0,$$

это возможно лишь в том случае, когда

$$f_z - \frac{\partial}{\partial x} f_p - \frac{\partial}{\partial y} f_q = 0. \quad (4.49)$$

Таким образом, мы видим, что минимизирующая поверхность, имеющая непрерывные вторые производные, должна удовлетворять *уравнениям Эйлера* (4.49).

Пожалуй, наиболее известным кратным интегралом этого типа является интеграл Дирихле

$$\iint_A (p^2 + q^2) dx dy.$$

В этом случае уравнение Эйлера является уравнением Лапласа:

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0.$$

Решение этого уравнения называется *гармонической функцией*. Этот результат приводит к следующему вариационному принципу, обычно называемому *принципом Дирихле*:

Гармоническая функция минимизирует интеграл Дирихле в классе функций, имеющих одинаковые граничные значения.

Принцип Дирихле показывает, что для того, чтобы найти гармоническую функцию, принимающую на границе некоторые данные значения, мы должны минимизировать интеграл Дирихле в классе

функций, принимающих эти граничные значения¹⁾. Это дает численный метод для определения гармонической функции.

Изучение второй вариации для кратных интегралов приводит к изучению общих самосопряженных уравнений с частными производными второго порядка эллиптического типа. Это является очень интересной главой вариационного исчисления, и она лишь только недавно получила достаточное развитие.

ЛИТЕРАТУРА

1. Bliss G. A., Calculus of Variations, Mathematical Association of America, The Open Court Publishing Co., La Salle, Ill., 1925.
2. Блисс Г. А., Лекции по вариационному исчислению, ИЛ, М., 1950.
3. Bolza O., Lectures on the Calculus of Variations, Stechert-Hafner, New York, 1946.
4. Bolza O., Vorlesungen über Variationsrechnung, Teubner, Berlin, 1909, Reprinted by K. F. Koehlers, Leipzig, 1933.
5. Carathéodory C., Variationsrechnung und Partielle Differentialgleichungen Erster Ordnung, Teubner, Berlin, 1935.
6. Курант Р., Гильберт Д., Методы математической физики, т. I, изд. 3, и т. II, изд. 2, М., 1951.
7. Forsyth A. R., Calculus of Variations, Cambridge University Press, London, 1927.
8. Fox C., An Introduction to the Calculus of Variations, Oxford University Press, New York, 1950.
9. Hadamard J., Leçons sur le calcul des variations, Hermann, Paris, 1910.
10. Hestenes M. R., An Elementary Introduction to the Calculus of Variations, Math. Mag., 23, 249—257 (1950).
11. Hildebrand F. B., Methods of Applied Mathematics, Prentice-Hall, New York, 1952.
12. Morse M., The Calculus of Variations in the Large, Amer. Math. Soc. Colloquium Publ., 18, American Mathematical Society, New York, 1934.
13. Osgood W. F., Advanced Calculus, Macmillan, New York, 1933.
14. Weinstock R., Calculus of Variations with Applications to Physics and Engineering, McGraw-Hill, New York, 1952.
- 15.* Лаврентьев М. А., Люстерник Л. А., Курс вариационного исчисления, изд. 2, Гостехиздат, М. — Л., 1950.
- 16.* Михлин С. Г., Вариационные методы в математической физике, Гостехиздат, М., 1957.

¹⁾ Принцип Дирихле применим в том и только в том случае, когда заданные граничные значения допускают непрерывное продолжение внутри рассматриваемой области, обладающее конечным интегралом Дирихле. Как известно, не всякая непрерывная функция имеет это свойство (пример Адамара). — *Прим. ред.*

²⁾ Здесь и в дальнейшем звездочка обозначает литературу, добавленную переводчиками. — *Прим. ред.*

Глава 5

ГИПЕРБОЛИЧЕСКИЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ И ИХ ПРИЛОЖЕНИЯ

РИХАРД КУРАНТ

Нью-Йоркский университет

§ 5.1. Введение. Со времен Ньютона и Лапласа общие законы, управляющие физическим миром, формулируются в виде дифференциальных уравнений. Из множества решений таких уравнений с помощью конкретных начальных или граничных условий отбираются решения, соответствующие определенным явлениям. Явления, происходящие в сплошных средах, таких, как упругие тела, газы и жидкости, а также электромагнитные явления приводят к дифференциальным уравнениям с частными производными. В частности, процессы распространения „волн“ описываются классом дифференциальных уравнений с частными производными, называемых „гиперболическими“, которые по своим свойствам существенно отличаются от дифференциальных уравнений других классов, как, например, тех, которые описывают состояния равновесия (эллиптические уравнения) или диффузию и теплопередачу (параболические уравнения).

Типичными примерами таких уравнений являются классические уравнения математической физики с неизвестными функциями $u(x, y, \dots; t)$, зависящими от пространственных переменных $x, y, \dots$, и времени t . (Эти неизвестные функции могут обозначать давление, температуру, плотность, электромагнитный потенциал и другие подобные величины.)

Прототипами эллиптических уравнений являются *уравнение Лапласа*

$$\nabla^2 u \equiv u_{xx} + u_{yy} + u_{zz} = 0, \quad (5.1)$$

которое называется еще *гармоническим уравнением*, или *уравнением потенциала*, и неоднородное *уравнение Пуассона*

$$\nabla^2 u = g(x, y, z).$$

Переменное t не входит в эти уравнения явно, и в соответствии с этим их решения описывают состояния равновесия. Уравнение Лапласа соответствует „естественному“ равновесию, тогда как уравнение Пуассона соответствует равновесию под действием внешней силы с плотностью, пропорциональной g .

Процессы переноса тепла, так же как и явления диффузии, описываются *тепловым уравнением*

$$u_t = \nabla^2 u,$$

представляющим собой уравнение параболического типа.

Распространение различных видов волн — упругих, звуковых, электромагнитных и т. д., — а также колебательные явления описываются гиперболическим *волновым уравнением*

$$\frac{1}{c^2} u_{tt} = \nabla^2 u, \quad (5.2)$$

в котором постоянная c равна скорости распространения волны в данной среде. Это уравнение справедливо в предположении, что амплитуда волны мала и что возникающие взаимодействия могут рассматриваться как линейные. При больших амплитудах получаются нелинейные гиперболические уравнения. Такова, например, система нелинейных уравнений, которая описывает одномерное адиабатическое течение сжимаемой жидкости:

$$\rho_t + u\rho_x + \rho u_x = 0, \quad \rho u_t + \rho u u_x + p_x = 0$$

[ср. уравнения (5.20) и (5.21)].

Дифференциальным уравнениям эллиптического типа посвящена гл. 6; эти уравнения связаны с краевыми задачами. Параболические и гиперболические уравнения, напротив, приводят к задачам с начальными значениями. В случае параболических уравнений произвольно можно задавать начальные значения только самого решения u , а для уравнений гиперболических при $t=0$ нужно задавать и функцию u и ее производную по времени. В этой главе мы дальше будем заниматься главным образом гиперболическими уравнениями.

§ 5.2. Связь между дифференциальными уравнениями с частными производными и действительностью. Прежде чем изучать отдельные уравнения, обратим внимание на то, что дифференциальные уравнения являются математической идеализацией действительности и эта идеализация производится для того, чтобы упростить описание чрезмерно сложного физического мира. Подобного рода идеализацию представляют собой понятия материальной точки и непрерывного распределения материи; идеализация появляется, когда таким величинам, как производные, градиенты и плотности, полученным в результате предельного перехода, приписывается физический смысл. В статистической механике такая идеализация заключается в особенно характерном и остроумном выходе за пределы наблюдаемой физической действительности: средние статистические значения изучаемых величин по большому, но конечному числу частиц заменяются

непрерывными функциями, которые затем вольно рассматриваются как дифференцируемые.

Итак, поскольку точные физические законы являются всего лишь идеализацией и поскольку любую заданную физическую ситуацию можно математически идеализировать многими различными способами, то важно уметь выбирать приемлемые идеализации. Есть три критерия, которым должна удовлетворять всякая правильно поставленная математическая задача, имеющая своей целью описывать действительность. Во-первых, задача должна допускать решение. Во-вторых, решение должно быть единственным. И, в-третьих, решение должно быть устойчивым; это значит, что малые изменения любого из данных задачи должны вызывать только соответственно малое изменение решения. Критерии существования и единственности, оба вместе, выражают уверенность в детерминизме, без которой при повторении экспериментов нельзя ожидать совпадающих результатов. Критерий устойчивости нужен по двум причинам: в данных, особенно если они получены экспериментально, всегда содержится некоторая ошибка, и нужно избежать того, чтобы эта ошибка вызвала слишком большую неточность в решении; кроме того, задачи дифференциальных уравнений обычно решаются приближенными методами и поэтому должен быть какой-то способ убедиться в том, что полученные решения являются хорошими приближениями к „истинным“ решениям.

Надо, однако, отметить, что эти критерии относятся лишь к общему направлению мысли и не всегда должны применяться буквально. В действительности успешное решение ряда задач физики указывает на необходимость ограничения применимости таких критериев, что приводит к последствиям, важным как для математической физики, так и для практических приложений.

§ 5.3. Статистические процессы и дифференциальные уравнения с частными производными. Пример того, как в результате идеализации получается дифференциальное уравнение с частными производными, дает двумерное броуновское движение коллоидных частиц. Это движение можно рассматривать как „случайное блуждание“ по плоскости. Положение частицы в данный момент времени t не определяет ее положения в последующий момент $t + \varepsilon$ единственным образом, а допускает множество положений, в которых частица может оказаться в этот момент с определенной вероятностью. Пусть функция $K(x, y; \xi, \eta; t)$ равна плотности вероятности того, что частица, находившаяся в момент t в точке (ξ, η) , в момент $t + \varepsilon$ будет находиться в точке (x, y) , т. е. K равна пределу дроби, полученной от деления вероятности для частицы быть в малой области вокруг точки (x, y) на площадь этой области при стремлении величины области к нулю. Пусть функция $u(x, y; t)$ равна плотности массы коллоидного вещества в точке (x, y) в момент t . Тогда имеет место

соотношение

$$u(x, y; t + \varepsilon) = \int_P \int K(x, y; \xi, \eta; t) u(\xi, \eta) d\xi d\eta, \quad (5.3)$$

где область интегрирования P есть вся плоскость.

Предположим, что функция K симметрична:

$$K(x, y; \xi, \eta; t) = K(\xi, \eta; x, y; t)$$

и что, как это нам подсказывает вероятностное определение функции K ,

$$\int_P \int K dx dy = \int_P \int K d\xi d\eta = 1.$$

Соотношение (5.3) можно тогда записать в виде

$$\begin{aligned} u(x, y; t + \varepsilon) - u(x, y; t) = \\ = \int_P \int K(x, y; \xi, \eta; t) [u(\xi, \eta; t) - u(x, y; t)] d\xi d\eta. \end{aligned} \quad (5.4)$$

Наша идеализация состоит теперь в том, что мы предполагаем промежуток времени стремящимся к нулю. Так как частицы жидкости перемещаются с конечной скоростью, то ядро K при малых ε отлично от нуля только тогда, когда точка (ξ, η) лежит внутри малой окружности, описанной вокруг точки (x, y) , и, кроме того, радиус этой окружности пропорционален ε . Далее, деля обе части уравнения (5.4) на ε и приближая затем ε к нулю, мы можем получить, как показывает простой анализ, дифференциальное уравнение с частными производными

$$u_t = L[u], \quad (5.5)$$

в котором в выражение $L[u]$, стоящее в правой части, входят производные от u по пространственным переменным x и y .

Чтобы рассмотреть какой-либо конкретный случай, предположим, что первоначальное ядро K не зависит от направления от (ξ, η) к (x, y) — это соответствует случаю, когда наша физическая среда изотропна, — и что за время ε частицы перемещаются на один прямолинейный „скачок“ или „шаг“ фиксированной длины h . Тогда, если точка (x, y) не лежит на окружности радиуса h с центром в (ξ, η) , ядро K равно нулю, а на этой окружности величина K , или, точнее, соответствующая линейная плотность вероятности, равна $1/2\pi h$. Таким образом, уравнение (5.3) принимает вид

$$u(x, y; t + \varepsilon) = \frac{1}{2\pi h} \int_0^{2\pi} u(x + h \cos \theta, y + h \sin \theta; t) h d\theta. \quad (5.6)$$

Для того чтобы перейти к пределу, предположим, что в окрестности рассматриваемой точки функцию u можно разложить в ряд Тэйлора

по x и y . Подинтегральное выражение в правой части уравнения (5.6) можно тогда записать в виде

$$u(x + h \cos \theta, y + h \sin \theta; t) = u(x, y; t) + h \cos \theta u_x + h \sin \theta u_y + \\ + \frac{1}{2} h^2 (\cos^2 \theta u_{xx} + 2 \cos \theta \sin \theta u_{xy} + \sin^2 \theta u_{yy}) + h^2 r(h, \theta),$$

где все производные берутся в точке $(x, y; t)$, а $r(h, \theta)$ стремится к нулю равномерно по θ , когда $h \rightarrow 0$. Произведя интегрирование в явном виде, мы найдем, что уравнение (5.6) принимает вид

$$u(x, y; t + \epsilon) - u(x, y; t) = \frac{1}{4} h^2 (u_{xx} + u_{yy}) + h^2 R(h),$$

где $R(h)$ также стремится к нулю, когда $h \rightarrow 0$. Если теперь одновременно приближать h и ϵ к нулю так, чтобы отношение $h^2/4\epsilon$ стремилось к 1, то в пределе получится двумерное уравнение теплопроводности

$$u_t = u_{xx} + u_{yy}.$$

Нам надо определить еще некоторые условия, которые позволяя выделить отдельное решение. Это можно сделать, например, задавая начальное распределение плотности $u(x, y; 0)$. Таким образом, мы приходим к задаче Коши для параболического дифференциального уравнения (5.5).

Предыдущее рассмотрение показывает, как дифференциальное уравнение с частными производными может получиться в результате идеализации вероятностного процесса. С другой стороны, вероятностные процессы можно использовать для решения дифференциальных уравнений с частными производными; соответствующие методы получили название методов Монте-Карло. Более подробное описание этих методов см. в гл. 11.

§ 5.4. Классификация линейных дифференциальных уравнений с частными производными. Плоские волны. Классификацию дифференциальных уравнений с частными производными лучше всего пояснить на примере линейного уравнения второго порядка относительно функции $u(x_1, x_2, \dots, x_m)$:

$$\sum_{i, k=1}^m a_{ik} u_{x_i x_k} + \sum_{i=1}^m b_i u_{x_i} + cu + d = 0, \quad (5.7)$$

где коэффициенты $a_{ik} = a_{ki}$, b_i , c , d — заданные функции $x_1, \dots, x_m$. Это дифференциальное уравнение называется *гиперболическим*, если уравнение $\sum a_{ik} \xi_i \xi_k = 0$ в пространстве с координатами $\xi_1, \dots, \xi_m$ представляет действительный конус. (Если конус не имеет никаких действительных точек, кроме начала координат, то дифференциальное уравнение (5.7) называется *эллиптическим*.)

В дальнейшем мы будем предполагать, что $m=4$, и будем обозначать x_4 через t , а x_1, x_2, x_3 через x, y, z . Гиперболические дифференциальные уравнения отличаются тогда тем, что они допускают решения, которые представляют собой, по крайней мере приближенно, в малой области пространства „плоские волны“. Чтобы пояснить это, мы рассмотрим здесь линейные дифференциальные уравнения (второго или высшего порядка) с постоянными коэффициентами. Решение $u(x, y, z, t)$ такого уравнения называется *плоской волной*, если оно имеет вид

$$u(x, y, z, t) = f(\alpha x + \beta y + \gamma z - ct), \quad (5.8)$$

где α, β, γ удовлетворяют условию

$$\alpha^2 + \beta^2 + \gamma^2 = 1 \quad (5.9)$$

и могут рассматриваться как направляющие косинусы направления распространения волны, а c равно скорости этого распространения. Действительно, такое решение принимает постоянные значения на гиперплоскости

$$\alpha x + \beta y + \gamma z - ct = \text{const}$$

четырёхмерного пространственно-временного континуума, причем саму гиперплоскость можно рассматривать, как „траекторию“ двумерной плоскости, нормаль к которой имеет направляющие косинусы α, β, γ и которая движется в пространстве по направлению нормали со скоростью c . Такая движущаяся плоскость называется *плоскостью постоянной фазы* или *фронтом волны*.

В частности, мы можем отыскивать плоские волны вида

$$u = Ae^{j(\alpha x + \beta y + \gamma z - ct)}, \quad j^2 = -1, \quad (5.10)$$

где комплексная показательная функция применяется, конечно, только для удобства записи того, что ее действительная и мнимая части порознь являются решениями. Общую теорию таких экспоненциальных плоских волн мы проиллюстрируем на примере уравнений второго порядка.

Для уравнения потенциала

$$u_{xx} + u_{yy} + u_{zz} = 0 \quad (5.11)$$

не может существовать экспоненциальных плоских волн, потому что подстановка функции (5.10) в уравнение (5.11) приводит к условию $\alpha^2 + \beta^2 + \gamma^2 = 0$, которое несовместимо с условием (5.9). Этот результат есть следствие эллиптического характера уравнения (5.11).

С другой стороны, гиперболическому волновому уравнению

$$c^2(u_{xx} + u_{yy} + u_{zz}) - u_{tt} = 0$$

удовлетворяют плоские волны, которые перемещаются во всех возможных направлениях с постоянной скоростью c .

Можно показать [2], что линейное гиперболическое дифференциальное уравнение второго порядка с постоянными коэффициентами подходящей заменой переменных сводится к уравнению

$$u_{tt} = \nabla^2 u + ku. \quad (5.12)$$

Мы можем теперь спросить себя, какие типы плоских волн, не обязательно экспоненциальных, возможны для такого уравнения. Если мы подставим в уравнение (5.12) функцию (5.8) и примем во внимание условие (5.9), то получим, что скорость c и вид волны f необходимо должны удовлетворять условию

$$(1 - c^2)f'' + kf = 0. \quad (5.13)$$

Далее различаются два случая: если $k = 0$, то могут существовать волны произвольного вида f , но все они будут распространяться с одной и той же скоростью 1; если же $k \neq 0$, то условие (5.13) дает для вида волны f обыкновенное дифференциальное уравнение, которое имеет следующее решение:

$$A \exp \left[\pm j \left(\frac{k}{1 - c^2} \right)^{1/2} \right].$$

Таким образом, волны имеют здесь уже не произвольный, а экспоненциальный вид, и, кроме того, скорость этих экспоненциальных волн изменяется с изменением частоты их „колебаний“. В первом случае говорят, что отсутствует дисперсия, так как волна произвольного вида, перемещаясь в пространстве, сохраняет свою форму. Во втором случае, напротив, волна произвольного вида распадается на экспоненциальные составляющие, которые определяются разложением Фурье, и эти составляющие перемещаются с различными скоростями, поэтому если мы снова соединим их воедино через определенный промежуток времени, то увидим, что первоначальный вид волны исказился.

Плоские волны имеют практическое значение при решении линейных дифференциальных уравнений с частными производными, потому что общие решения таких уравнений нередко можно представить как суперпозицию плоских волн.

§ 5.5. Задача Коши для волнового уравнения. *Задача Коши* или *задача с начальными условиями* для волнового уравнения (5.2) состоит в отыскании при всех $t \geq 0$ решения u по значениям этого решения и его производной по t при $t = 0$. Сосредоточим сначала наше внимание на случае одного пространственного переменного и запишем волновое уравнение в виде

$$\frac{1}{c^2} u_{tt} = u_{xx} \quad (5.14)$$

с начальными условиями

$$u(x, 0) = f(x), \quad (5.15)$$

$$u_t(x, 0) = g(x). \quad (5.16)$$

Для того чтобы решить эту задачу, заметим, что самое общее решение уравнения (5.14) без всяких начальных условий можно записать в виде

$$u(x, t) = A(x + ct) + B(x - ct), \quad (5.17)$$

где A и B — произвольные (дважды дифференцируемые) функции. Действительно, легко проверить при помощи подстановки, что любая функция вида (5.17) удовлетворяет уравнению (5.14). Обратно, если $u(x, t)$ есть некоторое решение уравнения (5.14), то, введя вместо x и t новые переменные $\xi = x + ct$ и $\eta = x - ct$, можно из уравнения (5.14) вывести, что $u_{\xi\eta} = 0$, а значит решение u должно иметь вид $A(\xi) + B(\eta)$. Для того чтобы полностью найти решение задачи Коши, остается только определить функции A и B через f и g . Это можно сделать, подставив выражение (5.17) в уравнения (5.15) и (5.16), проинтегрировав второе из полученных уравнений и разрешив их относительно $A(x)$ и $B(x)$. В результате мы найдем, что

$$A(x) = \frac{1}{2} f(x) + \frac{1}{2c} \int_0^x g(\xi) d\xi + c_0,$$

$$B(x) = \frac{1}{2} f(x) - \frac{1}{2c} \int_0^x g(\xi) d\xi - c_0,$$

где c_0 — некоторая постоянная интегрирования. Таким образом, решение определяется формулой

$$u(x, t) = \frac{f(x + ct) + f(x - ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\xi) d\xi. \quad (5.18)$$

Заметим, что значение решения (5.18) в точке x в момент t зависит только от начальных данных в точках, лежащих между $x - ct$ и $x + ct$, и совсем не зависит от начальных данных вне этого интервала. Это можно интерпретировать следующим образом: процессы, которые подчиняются волновому уравнению, распространяются со скоростью, никогда не превышающей c . Отрезок $[x - ct, x + ct]$ называется *областью зависимости* точки (x, t) .

Мы видим также, что значения f и g в точке $(\xi, 0)$ начальной прямой могут влиять на значения u только внутри треугольной области $|x - \xi| \leq ct$. В частности, если начальные данные равны нулю всюду, за исключением очень маленькой окрестности точки

$(\xi, 0)$, то решение будет нулем на всей плоскости, за исключением внутренности этой треугольной области и некоторых полос, непосредственно прилежащих извне к границе. Волновые явления такого типа, в которых возмущение, сосредоточенное при $t=0$ в одной точке, растекается с течением времени, известны под названием явлений *излучения*; треугольная область, на которую распространяется такое возмущение, называется *областью влияния* точки $(\xi, 0)$.

Далее, решение задачи Коши для одномерного волнового уравнения определено единственным образом. В самом деле, это решение в явном виде дается формулой (5.18). Мы можем сделать даже более сильное заключение, что решение u определено в треугольной области $|x-\xi| \leq c(\tau-t)$, $0 \leq t \leq \tau$ единственным образом, коль скоро значения самой функции и ее производной по времени заданы на отрезке $|x-\xi| \leq c\tau$ начальной прямой $t=0$. Кроме того, это решение устойчиво в том смысле, что оно зависит непрерывным и даже дифференцируемым образом от значений функций f и g и от скорости распространения c .

В двумерном и трехмерном пространствах задачу Коши можно решить таким же способом: наложением решений типа плоских волн, аналогичных решениям (5.17), причем при наложении решений вместо суммирования здесь надо производить интегрирование. Мы не будем выполнять подробно соответствующие вычисления, однако мы дадим явную форму решений, которая позволит определить область зависимости и область влияния.

Решение двумерного волнового уравнения

$$\frac{1}{c^2} u_{tt} = u_{xx} + u_{yy}$$

с начальными условиями

$$u(x, y, 0) = f(x, y), \quad u_t(x, y, 0) = g(x, y)$$

имеет вид

$$u(x, y, t) = \frac{1}{2\pi c} \frac{\partial}{\partial t} \iint_{\rho \leq ct} \frac{f(\xi, \eta) d\xi d\eta}{(c^2 t^2 - \rho^2)^{1/2}} + \frac{1}{2\pi c} \iint_{\rho \leq ct} \frac{g(\xi, \eta) d\xi d\eta}{(c^2 t^2 - \rho^2)^{1/2}},$$

где

$$\rho = [(\xi - x)^2 + (\eta - y)^2]^{1/2}.$$

Решение трехмерного волнового уравнения

$$\frac{1}{c^2} u_{tt} = u_{xx} + u_{yy} + u_{zz}$$

с начальными условиями

$$u(x, y, z, 0) = f(x, y, z), \quad u_t(x, y, z, 0) = g(x, y, z)$$

имеет вид

$$u(x, y, z, t) = \frac{1}{4\pi c^2} \frac{\partial}{\partial t} \left(\frac{1}{t} \iint_{r=ct} f(\xi, \eta, \zeta) dS \right) + \frac{1}{4\pi c^2 t} \iint_{r=ct} g(\xi, \eta, \zeta) dS,$$

где $r = [(\xi - x)^2 + (\eta - y)^2 + (\zeta - z)^2]^{1/2}$, (ξ, η, ζ) — текущие координаты, dS — элемент поверхности сферы радиуса $r = ct$ с центром в точке (x, y, z) .

Непосредственно видно, что в двумерном случае область зависимости пространственной точки в момент t представляет собой всю *внутренность* (и границу) круга радиуса ct , описанного вокруг этой точки, тогда как в трехмерном случае соответствующая область зависимости состоит только из *поверхности* сферы радиуса ct . Соответственно и областью влияния точки в двумерном случае является внутренность одной полости конуса, т. е. траектория расширяющегося круглого диска, тогда как в случае трехмерного волнового уравнения областью влияния является траектория расширяющейся сферической поверхности, т. е. трехмерная поверхность четырехмерного конуса. Физически это означает, что двумерные волновые явления, так же как и соответствующие одномерные явления, распространяются с максимальной скоростью c , но могут распространяться и с меньшей скоростью, а трехмерные волновые явления всегда распространяются точно со скоростью c .

Говорят, что задача Коши в n -мерном пространстве подчиняется *принципу Гюйгенса*, если область зависимости каждой точки есть многообразие $n - 1$ или меньшего числа измерений. Благодаря тому что трехмерное волновое уравнение подчиняется принципу Гюйгенса, а соответствующее двумерное уравнение ему не подчиняется, мы в состоянии передавать четкие сигналы при помощи световых, звуковых или радиоволн, которые по своей природе существенно трехмерны, и не пользуемся для этого рябью на поверхности водоемов. Таким образом, в двумерном мире передавать сигналы с большой точностью было бы невозможно.

Математики в течение долгого времени старались установить, существуют ли какие-нибудь иные дифференциальные уравнения с частными производными, для решений которых справедлив принцип Гюйгенса. Оказалось, что таким свойством обладают только уравнения второго порядка, так как для уравнений высших порядков имеет место кратное преломление, в силу чего большая точность при передаче сигналов невозможна. Скоро было установлено, что волновые уравнения для пяти, семи, девяти... пространственных измерений обладают нужным свойством, тогда как у соответствующих уравнений в одномерном и во всех четномерных пространствах оно отсутствует. До сих пор, однако, не найдено уравнений, которые бы существенно отличались от волнового уравнения и удовлетворяли бы принципу Гюйгенса; более того, существует известная гипотеза Адамара, что других таких уравнений нет.

Оставляя в стороне гипотезу Адамара, можно поставить такой вопрос: чем наш трехмерный мир отличается, скажем, от допустимого семимерного мира, в котором сигналы распространяются согласно семимерному волновому уравнению

$$\frac{1}{c^2} u_{tt} = \sum_{i=1}^7 u_{x_i x_i}.$$

Оказывается, что трехмерное волновое уравнение имеет решающее отличительное свойство: только оно одно допускает в качестве решений сферические волны произвольной формы, которые сравнительно мало искажаются. *Сферической волной* мы называем решение u волнового уравнения, которое можно представить в виде

$$u(x_1, \dots, x_n; t) = h(r) f(r - ct), \quad (5.19)$$

где

$$r = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}.$$

Функция f называется *формой волны* или *профилем волны*, а функция h известна под названием *коэффициента затухания*. Для того чтобы проверить наше утверждение, подставим выражение (5.19) в n -мерное волновое уравнение

$$\frac{1}{c^2} u_{tt} = \sum_{i=1}^n u_{x_i x_i}$$

и заметим, что, поскольку функция f предполагалась произвольной, в окончательном выражении коэффициенты при f , f' и f'' должны быть нулями. Этим на функцию h накладываются условия, которые совместны друг с другом лишь в том случае, когда $n=3$, и в этом случае коэффициенты обращаются в нуль, когда $h(r) = 1/r$. Сигналы, которые обычно испускаются из источника как из центра, имеют вид сферических волн. Следовательно, такие сигналы при передаче могут оставаться неискаженными только в нашем трехмерном пространстве. Иными словами, только в трехмерном пространстве передача сигналов возможна с большой точностью.

§ 5.6. Нелинейные гиперболические уравнения. Легкость, с которой находятся явные решения волнового уравнения, и сравнительная простота этих решений своим происхождением обязаны главным образом линейному характеру этого уравнения, т. е. тому, что уравнение линейно относительно неизвестных функций и их производных. В случае нелинейных дифференциальных уравнений могут возникнуть серьезные математические трудности. Этими трудностями можно объяснить тот факт, что такие отделы физики, как гидродинамика, газовая динамика и теория волноводов, которые

в основном имеют дело с нелинейными явлениями, развивались сравнительно медленно.

Наиболее поразительное отличие линейных и нелинейных уравнений заключается в поведении допустимых разрывов решения. Все, что может произойти в линейном случае, ясно видно из выражения (5.18). Допустим, что начальные значения $f(x)$ решения $u(x, t)$ имеют разрыв в точке $x = \xi$. Тогда само решение будет разрывно во всех точках (x, t) , для которых $x + ct$ или $x - ct$ равно ξ , т. е. во всех точках, лежащих на двух прямых

$$x = \xi + ct, \quad x = \xi - ct,$$

которые имеют соответственно наклоны $\pm 1/c$ и выходят из точки разрыва $(\xi, 0)$. Эти прямые называются *характеристиками* волнового уравнения. Таким образом, можно сказать, что разрывы решения волнового уравнения распространяются вдоль характеристик с постоянной скоростью c , которая называется *скоростью звука*. Далее, из формулы (5.18) видно, что у решения не может возникнуть никакого разрыва, если соответствующего разрыва уже не было в начальных данных, и что разрыв начальных данных никогда не исчезнет, как бы велико ни стало t . Если мы продифференцируем выражение (5.18) один раз по x или по t , то увидим, что аналогичные утверждения справедливы для разрывов $u_x = f'(x)$ и $u_t = g(x)$ на начальной прямой.

По существу такие же результаты получаются для любого линейного гиперболического уравнения от двух независимых переменных. Если уравнение имеет переменные коэффициенты, то характеристики уже не будут прямыми линиями. Однако для любого гиперболического уравнения существуют два семейства характеристических кривых, которые обладают тем свойством, что через любую точку плоскости (x, t) проходит по одной кривой каждого из семейств. Разрывы решения $u(x, t)$ или разрывы его производных распространяются вдоль таких характеристик в точности так же, как и для волнового уравнения; в частности, они никогда не могут внезапно возникнуть или исчезнуть. В каждой точке скорость распространения разрывов равна обратной величине наклона соответствующей характеристики в этой точке; она называется *местной* скоростью звука.

При попытке распространить эти результаты на нелинейные уравнения возникают существенные трудности. Во-первых, чтобы определить характеристические кривые, надо заранее знать решение, а кроме того, здесь уже не верно утверждение, что разрывы решений распространяются вдоль характеристик. Более того, могут возникнуть случаи, когда имевшиеся в начале разрывы исчезают с течением времени, или, наоборот, случаи, когда непрерывное сначала состояние становится вдруг разрывным. Обе эти возможности можно показать на простой задаче гидродинамики, в которой рассматривается поток

газа в полубесконечной трубе, закрытой с одного конца подвижным поршнем.

Одномерный сжимаемый изэнтропический поток описывается дифференциальными уравнениями

$$\rho_t + u\rho_x + \rho u_x = 0, \quad (5.20)$$

$$\rho u_t + \rho u u_x + p_x = 0 \quad (5.21)$$

(см. стр. 90 книги Куранта и Фридрихса [1]). Это — нелинейная система двух дифференциальных уравнений первого порядка относительно двух неизвестных функций $u(x, t)$ и $p(x, t)$, которые представляют собой соответственно скорость и плотность газа в трубе в точке x и в момент t . Величина p , входящая в уравнение (5.21), равна давлению газа и предполагается известной функцией f от плотности:

$$p = f(\rho).$$

Систему (5.20), (5.21) можно назвать гиперболической, потому что, исключив из нее зависимое переменное p , мы получим для u одно гиперболическое уравнение второго порядка; удобнее, однако, не исключать p , а, наоборот, иметь дело с заданной системой.

В рассматриваемой задаче предполагается, что в начальный момент поршень находится в покое в точке $x=0$; газ также находится в покое и заполняет собой правую половину трубы $0 < x < \infty$, причем давление и плотность газа постоянны и равны p_0 и ρ_0 соответственно. Затем поршень приводится в движение и ставится вопрос, что при этом происходит со столбом газа, причем предполагается, что непосредственно вблизи поршня газ всегда движется с той же скоростью, что и поршень. Если описывать движение поршня, задавая его положение X как функцию времени $X(t)$, то условие, которое мы налагаем на движение газа, принимает вид граничного условия

$$u[X(t), t] = \dot{X}(t). \quad (5.22)$$

Таким образом, нашу задачу можно рассматривать как смешанную задачу с заданными начальными и граничными значениями. Начальные условия, очевидно, имеют вид

$$u(x, 0) = 0, \quad p(x, 0) = p_0.$$

Можно показать (см. гл. III упоминавшейся книги [1]), что одно из семейств характеристик этой задачи, а именно то, у которого больше величина dx/dt (равная обратной величине наклона характеристики), всегда состоит из прямых линий и что обратная величина наклона всякой такой прямой выражается через координаты (ξ, τ) точки пересечения этой прямой со следом поршня $x = X(t)$ следующим образом:

$$\frac{dx}{dt} = u(\xi, \tau) + c(\xi, \tau). \quad (5.23)$$

Функция $u(\xi, \tau)$ определяется из граничного условия (5.22), а c — местная скорость звука в точке (ξ, τ) вычисляется по формуле

$$c = \{f'[\rho(\xi, \tau)]\}^{1/2}, \quad \text{где } \rho(\xi, \tau)$$

определяется из условия

$$u(\xi, \tau) = \int_0^{\rho(\xi, \tau)} \frac{[f'(\rho)]^{1/2}}{\rho} d\rho. \quad (5.24)$$

Здесь возникают два совершенно различных явления в зависимости от того, выдвигается ли поршень из газа, или вдвигается в него. Рассмотрим первый из этих случаев, предполагая, что скорость выдвигаемого поршня равномерно возрастает. Тогда $u(\xi, \tau) < 0$ будет монотонно убывающей функцией τ и, как видно из уравнения (5.24), $\rho(\xi, \tau)$ тоже будет убывающей функцией τ . Все реальные газы подчиняются, однако, общему закону, согласно которому $dc/d\rho > 0$, поэтому $c(\xi, \tau)$ тоже убывает с ростом τ . Таким образом, характеристики, которые начинаются в точках на траектории поршня, лежащих левее, имеют меньшие значения обратных величин наклона; поэтому они перемещаются (вправо) более медленно. Следовательно, характеристики имеют тенденцию расходиться при возрастании t и, если предположить по аналогии с линейным случаем, что начальные возмущения распространяются вдоль этих характеристик подобно звуковым волнам, то окажется, что эти возмущения с течением времени растекаются все больше и больше и, следовательно, делаются слабее.

В предельном случае, когда возмущение вызвано мгновенным, конечным изменением скорости поршня (в отрицательном направлении), начальный разрыв сразу же сглаживается и движение делается непрерывным. Такое явление в линейных гиперболических уравнениях никогда не может возникнуть. При этом из точки на траектории поршня, в которой произошло внезапное изменение, выходит веерообразный пучок прямолинейных характеристик, соответствующих всем возможным скоростям поршня, заключенным между его начальной и конечной скоростями. В области, покрытой этими характеристиками, движение называется *центрированной волной разрежения*. При $t > 0$ в волне разрежения осуществляется непрерывный переход от состояния покоя, в котором находится часть газа, лежащая в области, куда еще не дошла первоначальная звуковая волна, к состоянию равномерного движения влево со скоростью, равной скорости движения поршня, и с более низкой, чем ρ_0 , плотностью. Это равномерное движение происходит в части газа, которую уже пересекла последняя звуковая волна (отвечающая окончательной скорости поршня и соответствующей плотности).

Если поршень вдвигается в газ, то непрерывное движение уже больше получиться не может, потому что на этот раз обратная вели-

чина наклона (5.23) характеристик, проходящих через (ξ, τ) , возрастает вместе с τ , в результате чего характеристики будут нагонять друг друга. При этом, как бы ни было гладко начальное движение поршня, в точках пересечения характеристик будут происходить разрывы. Следовательно, непрерывные начальные и граничные данные могут давать разрывы в последующее время. Такие разрывы называются ударами, они всегда движутся вперед со сверхзвуковой скоростью.

Простейшим случаем, в котором возникает удар, является случай, когда покоившийся поршень мгновенно ускоряется до конечной поступательной скорости u_F . В результате поток резко разделяется на две части поверхностью разрыва, которая называется *ударной волной* и перемещается вправо с постоянной сверхзвуковой скоростью U . Справа от этой ударной волны газ находится в покое и имеет начальное давление p_0 , а слева он имеет постоянное более высокое давление p_1 и движется с той же скоростью u_F , что и поршень. Кроме того, газ, через который прошла ударная волна, имеет большую, чем невозмущенный газ, энтропию. Следовательно, согласно второму закону термодинамики, удар является *необратимым* процессом, волна разрежения, напротив, не дает такого возрастания энтропии и поэтому она обратима.

За дальнейшими примерами ударов и других разрывов мы отсылаем читателя к цитированной выше книге Куранта и Фридрихса [1]. Для общего знакомства с теорией нелинейных гиперболических уравнений можно рекомендовать недавно вышедшую статью Куранта и Лакса [4].

В общем, даже в сравнительно простых видах нелинейного волнового движения уже проявляются существенные особенности, которые не встречаются в классической линейной теории. И хотя ударные разрывы подчиняются простым законам, которые так же, как и дифференциальные уравнения, выражают собой законы сохранения энергии массы и момента, положение этих ударов во времени и пространстве заранее неизвестно, а должно находиться как часть решения. В большинстве случаев аналитически этого сделать нельзя, и надежду на успешное решение задачи дают только численные методы.

§ 5.7 Конечно-разностный метод. Как мы уже отметили в § 5.2, дифференциальные уравнения с частными производными часто получаются при идеализации явлений, в которых участвует лишь конечное число величин. Кажется парадоксальным, что для численного решения таких дифференциальных уравнений нередко бывает необходимо частично пойти назад в процессе идеализации и заменить частные производные соответствующими разностными отношениями. При этом в пространстве независимых переменных выбирает фиксированную прямоугольную точечную сетку, которую берут настолько мелкой, насколько это возможно сделать, не загромождая чрезмерно

вычислений. В точках этой сетки ищут значения зависимой переменной, совместимые с уравнениями в частных разностях, полученными в результате упомянутой выше замены, а также с начальными или граничными условиями. Отыскание неизвестных значений состоит, вообще говоря, в решении системы совместных алгебраических или трансцендентных уравнений со многими неизвестными. Хотя в принципе это проще, чем решать само дифференциальное уравнение, реальные вычисления часто получаются очень громоздкими.

В некоторых случаях, однако, вид окончательной системы позволяет решать ее без особого труда. Такая система получится, например, из двумерного уравнения Лапласа (5.1), если мы выберем квадратную сетку с шагом h . Частные производные заменяем частными разностями по следующим формулам:

$$\begin{aligned} u_x(x, y) &\sim \frac{u(x+h, y) - u(x, y)}{h}, \\ u_{xx}(x, y) &\sim \frac{u_x(x, y) - u_x(x-h, y)}{h} \sim \\ &\sim \frac{u(x+h, y) + u(x-h, y) - 2u(x, y)}{h^2} \end{aligned}$$

и аналогично

$$u_{yy}(x, y) \sim \frac{u(x, y+h) + u(x, y-h) - 2u(x, y)}{h^2}.$$

Тогда уравнение Лапласа

$$u_{xx} + u_{yy} = 0 \quad (5.25)$$

заменится разностным уравнением

$$u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) = 4u(x, y), \quad (5.26)$$

которое показывает, что значение величины u в точке (x, y) сетки равно среднему арифметическому ее значений в четырех соседних точках. Это свойство решения аналогично хорошо известному „свойству среднего значения“ для гармонических функций.

Для гармонического уравнения обычно решается краевая задача, в которой требуется по заданным значениям функции u на границе найти ее значения внутри области. Соответствующую конечно-разностную задачу можно поставить следующим образом. Пусть дана связная область, покрытая точечной сеткой. Точка сетки называется *внутренней* точкой, если все ее четыре соседние точки принадлежат области, в противном случае точка называется *граничной*. В граничных точках задаются значения функции u . Задача состоит в отыскании значений u во всех внутренних точках. Это можно сделать любым из следующих трех методов. Первый из них основан на том, что так как система (5.26) является системой, в которой число линейных уравнений равно числу неизвестных, то ее можно решать при помощи

детерминантов и правила Крамера. Во втором методе применяются последовательные приближения: берут предположительные значения решения, а затем значения этого приближенного решения последовательно подправляют в каждой точке, пользуясь уравнением (5.26), причем на каждом шаге берут те значения u в четырех соседних для (x, y) точках, которые были получены в предыдущем приближении.

Третий метод аналогичен методу случайных блужданий, который мы применяли в § 5.3. Пусть дана точка (x, y) . Рассмотрим все ломаные траектории P_i , $i = 1, 2, \dots$, состоящие из горизонтальных и вертикальных отрезков длины h , которые начинаются в точке (x, y) , кончаются в одной из граничных точек p_i и больше ни через одну граничную точку не проходят. Пусть число отрезков, составляющих траекторию P_i , равно l_i . Легко показать, что из уравнения (5.26) вытекает соотношение

$$u(x, y) = \sum_{i=1}^{\infty} \frac{1}{4^{l_i}} u(p_i), \quad (5.27)$$

которое дает явное представление для решения разностного уравнения, если ряд (5.27) сходится.

Если в (5.27) сгруппировать вместе все траектории P_i , у которых конечные точки p_i совпадают с некоторой граничной точкой p , то это соотношение можно будет записать в виде

$$u(x, y) = \sum_{p \text{ на границе}} \left[\sum_{p_i=p} \frac{1}{4^{l_i}} u(p_i) \right] = \sum_{p \text{ на границе}} \left(\sum_{p_i=p} \frac{1}{4^{l_i}} \right) u(p). \quad (5.28)$$

Появившееся теперь в равенстве (5.28) выражение

$$G(x, y; p) = \sum_{p_i=p} \frac{1}{4^{l_i}}$$

называется *функцией Грина* разностного уравнения (5.26) и рассматриваемой сеточной области. Если эта функция известна, то уже можно решить любую краевую задачу для данного уравнения и данной области. Сама функция Грина тоже является решением краевой задачи, а именно задачи, в которой граничные значения равны единице в точке p и равны нулю во всех остальных точках. Интуитивно ясно, что функция Грина равна вероятности того, что частица, которая выходит из точки сетки (x, y) , впервые достигнет границы в точке p , если при этом она перемещается, последовательно переходя из точки сетки в одну из четырех ее соседних точек, причем направление перехода каждый раз оказывается случайным.

Если вместо эллиптического уравнения (5.25) рассматривается дифференциальное уравнение параболического или гиперболического типа, то соответствующие разностные уравнения можно решать прямым итерационным методом, в котором значения $u(x, t)$ последова-

тельно вычисляются по начальным значениям, причем значение $u(x, t)$ находится в конечном числе шагов. Такие итерационные процессы хорошо приспособлены для современных электронных вычислительных машин.

Для того чтобы разобрать подробно, какой вид может иметь этот итерационный процесс, рассмотрим одномерное (параболическое) уравнение теплопроводности

$$u_t = u_{xx},$$

которое можно заменить разностным уравнением, пользуясь прямоугольной сеткой (с шагом h в направлении x и с шагом k в направлении t):

$$u(x, t+k) = \frac{k}{h^2} [u(x+h, t) + u(x-h, t)] + \left(1 - \frac{2k}{h^2}\right) u(x, t). \quad (5.29)$$

Если для всех x заданы начальные значения $u(x, 0)$, то, пользуясь уравнением (5.29), мы можем найти сначала значения $u(x, k)$, потом значения $u(x, 2k)$ и т. д., пока не дойдем до любой нужной нам точки сетки.

Аналогичным образом волновое (гиперболическое) уравнение

$$u_{tt} = u_{xx},$$

в котором для простоты мы положили $c=1$, аппроксимируется конечно-разностным уравнением

$$u(x, t+k) = \frac{k^2}{h^2} [u(x+h, t) + u(x-h, t)] + 2\left(1 - \frac{k^2}{h^2}\right) u(x, t) - u(x, t-k). \quad (5.30)$$

В этом случае при $t=0$ задаются u и u_t ; по ним значение u при $t=k$ находится по приближенной формуле

$$u(x, k) = u(x, 0) + ku_t(x, 0).$$

Далее, пользуясь уравнением (5.30), мы можем по $u(x, k)$ и $u(x, 0)$ найти $u(x, 2k)$; затем по $u(x, 2k)$ и $u(x, k)$ найти $u(x, 3k)$ и т. д. Заметим, что по разностному уравнению легко установить число и вид начальных данных, которые необходимы для того, чтобы решение было определено.

Кроме того значения, которые конечно-разностные методы имеют для численного определения решений, они имеют еще и большую теоретическую ценность. В последнее время они часто применялись в качестве аппарата для доказательства теорем существования и единственности решений дифференциальных уравнений с частными производными. Как правило, техника таких доказательств состоит в следующем. Сначала для произвольной сети решают разностные уравнения, что теоретически не представляет трудности. Затем размеры сетки устремляют к нулю и показывают, что решения соответствующих

разностных уравнений стремятся к решению дифференциального уравнения, из которого эти разностные уравнения были получены. Обычно при этом необходимо наложить ограничения на вид сетки, т. е. на зависимость между h и k при их одновременном стремлении к нулю. Эти ограничения часто имеют не только теоретическое значение, так как даже при фиксированном размере сетки ими выражается необходимый критерий, который гарантирует, что решение разностных уравнений хорошо аппроксимирует точное решение дифференциального уравнения. Критерии этого рода называются *критериями устойчивости*. Для рассмотренных выше случаев они дают

$$\frac{k}{h^2} < \frac{1}{2} \quad \text{для уравнения теплопроводности,}$$

$$\frac{k}{h} < 1 \quad \text{для волнового уравнения.}$$

Предыдущие замечания относились к линейным задачам. Однако конечно-разностные методы естественным образом применяются и к нелинейным задачам. Большие практические результаты в этой области были достигнуты с помощью быстродействующих электронных вычислительных машин. В частности, конечно-разностные методы дают возможность преодолеть огромные трудности, с которыми мы встречаемся в случае ударов. В настоящей статье мы вынуждены ограничиться лишь краткой ссылкой на соответствующую литературу.

Для того чтобы решать задачи с ударами, предложены два существенно различных метода. В первом из них, принадлежащем Нейману и Рихтмайеру [7], в дифференциальные уравнения гидродинамики вводится добавочный малый член с вязкостью, который подбирается так, чтобы его влияние было заметно только в области быстрых изменений. Решение при этом получается непрерывным (хотя удары и аппроксимируются очень быстрыми переходами), приближенные решения ведут себя гладко, и можно показать, что результат аппроксимирует искоемое решение.

В другом методе, принадлежащем Лаксу [6], дифференциальные уравнения сначала записывают как законы сохранения и, исходя непосредственно из этого вида уравнений, переходят к подходящей конечно-разностной схеме. Вычисления при этом оказываются очень простыми; удары представляются областями, в которых разности зависимых функций получаются большими. Предварительные оценки показали, что этот метод приводит к удовлетворительным результатам¹⁾.

¹⁾ Математически строго метод конечных разностей для нелинейных уравнений гиперболического типа до сих пор не обоснован. Однако для некоторых нелинейных уравнений первого порядка такое обоснование метода Лакса дано Н. Д. Введенской. См. Олейник О. А., Разрывные решения нелинейных дифференциальных уравнений, УМН, 12, вып. 3 (75), (1957). — *Прим. ред.*

В общем теперь уже ясно, что современная вычислительная техника в соединении с глубоким анализом должны привести к решению многих, казавшихся ранее недоступными, задач газовой динамики и аэродинамики. Эти методы главным образом применимы к задачам распространения, т. е. к задачам, которые включают в себя параболические и прежде всего гиперболические дифференциальные уравнения.

ЛИТЕРАТУРА

1. Курант Р., Фридрихс К., Сверхзвуковое течение и ударные волны, ИЛ, М., 1950.
2. Курант Р., Гильберт Д., Методы математической физики, т. I, изд. 3, и т. II, изд. 2, Гостехиздат, М.—Л., 1951.
3. Courant R., Isaacson E., Rees M., On the Solution of Nonlinear Hyperbolic Differential Equations by Finite Differences, Comm. Pure Appl. Math., 5, 243—255 (1952).
4. Courant R., Lax P., On Nonlinear Partial Differential Equations with Two Independent Variables, Comm. Pure Appl. Math., 2, 255—273 (1949).
5. John F., On Integration of Parabolic Equations by Difference Methods, I. Linear and Quasi-linear Equations for the Infinite Interval, Comm. Pure Appl. Math., 5, 155—211 (1952).
6. Lax P., Weak Solutions of Nonlinear Hyperbolic Equations and Their Numerical Computation, Comm. Pure Appl. Math., 7, 159—194 (1954).
7. von Neumann J., Richtmyer R. D., A Method for the Numerical Calculation of Hydrodynamic Shocks, J. Appl. Phys., 21, 232—237 (1950).

Глава 6

КРАЕВЫЕ ЗАДАЧИ ДЛЯ ЭЛЛИПТИЧЕСКИХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

МЕНАХЕМ М. ШИФФЕР

Станфордский университет

§ 6.1. Что такое корректно поставленная задача для дифференциальных уравнений с частными производными? Дифференциальные уравнения с частными производными встречаются в большинстве разделов прикладной математики, теоретической физики и современной техники. Прототипом уравнений, которые мы будем здесь изучать, является классическое уравнение Лапласа

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0,$$

где ∇ — оператор

$$\nabla = i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z}.$$

Этим уравнением характеризуются гравитационный и электростатический потенциалы в точках свободного пространства, оно описывает потенциал скорости безвихревого потока несжимаемой жидкости и оно же справедливо для температуры изотропной среды при установившемся движении тепла. Уравнение такого же вида для двух переменных лежит в основе теории аналитических функций комплексного переменного.

В теории обыкновенных дифференциальных уравнений основная трудность заключается в отыскании общего решения уравнения, которое зависит от некоторого числа параметров. После того как общее решение найдено, эти параметры легко привести в соответствие с дополнительными условиями, заданными в рассматриваемой задаче, и тем самым получить нужное решение. В случае дифференциальных уравнений с частными производными мы оказываемся в совершенно ином положении.

Рассмотрим, например, уравнение Лапласа для функции двух переменных $u(x, y)$. Легко видеть (см. гл. 7), что всякое решение этого уравнения должно иметь вид

$$u(x, y) = f(x + jy) + g(x - jy) \quad f^2 = -1, \quad (6.1)$$

где f и g — две произвольные аналитические функции, и обратно, каждая функция вида (6.1) является решением нашего уравнения.

Таким образом, мы получаем общее решение задачи, но на этот раз оно зависит от двух произвольных функций f и g . Однако выбрать такое решение, которое соответствует данным задачи, вообще говоря, чрезвычайно трудно, а поэтому то, что мы знаем общее решение, в теории уравнения Лапласа не имеет большого значения.

Первое существенное продвижение в дифференциальных уравнениях с частными производными было достигнуто методом разложения в ряды. Предположим, что нам дано дифференциальное уравнение второго порядка

$$\frac{\partial^2 u}{\partial y^2} = F\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x \partial y}\right) \quad (6.2)$$

и мы ищем решение $u(x, y)$, для которого заданы $u(x, 0) = a(x)$ и $u_y(x, 0) = b(x)$. Тогда можно попытаться представить решение в виде ряда

$$u(x, y) = \sum_{v=0}^{\infty} \alpha_v(x) y^v \quad (6.3)$$

и найти $\alpha_v(x)$ из начальных условий и уравнения (6.2). Легко видеть, что это можно осуществить, если F , $a(x)$ и $b(x)$ суть аналитические функции своих аргументов, которые поэтому можно дифференцировать неограниченное число раз. Кроме того, можно показать, что ряд (6.3) действительно сходится при достаточно малых значениях y . Таким образом, в этом случае получается конструктивное доказательство существования решения уравнения (6.2) (Коши — Ковалевская). Аналогичный метод можно применить к дифференциальным уравнениям с частными производными любого порядка и для более чем двух независимых переменных.

Тот факт, что функции F , a , b должны быть аналитическими, сначала не тревожил большинство математиков, и тем более — физиков и инженеров. Практически все встречающиеся в приложениях функции F всегда аналитичны, а относительно начальных данных $a(x)$ и $b(x)$ приводили то соображение, что всякую непрерывную функцию можно в любом заданном интервале сколь угодно точно и равномерно аппроксимировать полиномами от x , которые, конечно, аналитичны. Таким образом, аналитичность функций казалась совсем несущественным ограничением. Ф. Клейн, ведущий математик начала этого века, написал работу „Точная и приближенная математика“, в которой он приводил аналогичные доводы, чтобы показать, что математик-прикладник может успешно ограничиться рассмотрением только аналитических функций.

Необоснованность таких соображений лучше всего видна на следующем примере, который принадлежит Адамару: Рассмотрим опять уравнение Лапласа от двух независимых переменных и будем искать решение, для которого

$$u(x, 0) = \frac{1}{n^2} \sin nx, \quad u_y(x, 0) = 0. \quad (6.4)$$

В силу теоремы Коши — Ковалевской решение существует, и притом его легко угадать; именно

$$u(x, y) = \frac{1}{n^\alpha} \sin nx \operatorname{ch} ny. \quad (6.5)$$

Если $\alpha > 0$ фиксировано, а $n \rightarrow \infty$, то начальные данные, очевидно, равномерно сходятся к нулевым значениям

$$u(x, 0) = 0, \quad u_y(x, 0) = 0,$$

которым отвечает решение $u(x, y) = 0$. Функции (6.5), однако, к нулю не сходятся, а становятся очень велики в любой окрестности оси x . Можно показать, выбрав α достаточно большим, что начальные значения (6.4) сходятся к нулю вместе с любым числом производных, но что это не влечет за собой сходимости решений. Следовательно, сколь угодно точная аппроксимация начальных данных поставленной задачи ни в коей мере не гарантирует соответствующей аппроксимации решений. Иными словами, решение не всегда устойчивым образом зависит от данных, которые ставятся в теории Коши — Ковалевской.

К счастью, никакие задачи естествознания или техники не приводятся к задаче с начальными значениями указанного типа для уравнения Лапласа.

Если для дифференциального уравнения задаются дополнительные условия, при помощи которых из всех возможных решений уравнения отбирается одна единственная функция, то такие условия называются *заданными корректно*, коль скоро указанное решение непрерывным образом зависит от дополнительных данных. Мы можем, таким образом, утверждать, что для уравнения Лапласа приведенные выше начальные значения $u(x, 0)$ и $u_y(x, 0)$ заданы некорректно.

§ 6.2. Теория теплопередачи; три основные краевые задачи.

Для того чтобы получить для уравнения Лапласа корректные условия, а также для того, чтобы дать интуитивное представление о тех понятиях, которые мы собираемся ввести, мы сейчас разберем довольно подробно задачу о распространении тепла в изотропном теле. Пусть B — тело в пространстве x, y, z , S — его граничная поверхность. Предположим, что в теле существует стационарное температурное поле $T(x, y, z)$ и постараемся вывести для функции T дифференциальное уравнение. Если теплопроводность тела равна k , то поток тепла в каждой точке тела описывается вектором

$$\mathbf{q} = -k \nabla T,$$

так что количество тепла, прошедшее за единицу времени через элемент поверхности $d\sigma$, расположенный в заданной точке и имеющий нормаль $\mathbf{n}$, равно

$$\mathbf{q} \cdot \mathbf{n} d\sigma = -k \frac{\partial T}{\partial n} d\sigma.$$

Пусть теперь объем b является частью тела B и поверхность c есть граница объема b . Если γ — удельная теплоемкость тела B , то количество тепла, заключенное в b , равно

$$Q = \gamma \int_b T dx dy dz,$$

причем эта величина остается постоянной, так как по предположению температура не меняется со временем. Поэтому общее количество тепла, прошедшего за единицу времени через границу c , равно нулю и, следовательно, для любой заключенной в B замкнутой поверхности c имеет место условие

$$\int_c \frac{\partial T}{\partial n} d\sigma = 0. \quad (6.6)$$

Мы покажем, что это условие эквивалентно некоторому дифференциальному уравнению для функции T .

Для этого мы воспользуемся формулой Грина, которая вообще имеет очень большое значение во всей теории линейных эллиптических дифференциальных уравнений:

$$\int_b \nabla U \cdot \nabla V dx dy dz = \int_c U \frac{\partial V}{\partial n} d\sigma - \int_b U \nabla^2 V dx dy dz. \quad (6.7)$$

Здесь через n обозначена внешняя нормаль, а U и V — две произвольные дважды непрерывно дифференцируемые в области b функции. Формула Грина выражает собой просто правило интегрирования по частям в пространстве. Применяя эту формулу к функциям $U = 1$ и $V = T$, мы найдем, что в силу уравнения (6.6)

$$\int_c \frac{\partial T}{\partial n} d\sigma = \int_b \nabla^2 T dx dy dz = 0.$$

Так как это равенство справедливо для произвольного объема b , то для его выполнения необходимо и достаточно условие

$$\nabla^2 T = 0. \quad (6.8)$$

Итак, мы знаем теперь, что стационарное температурное поле в изотропном однородном теле описывается гармонической функцией, т. е. функцией, которая удовлетворяет уравнению Лапласа. Поставим теперь такой вопрос: какие условия достаточно задать для того, чтобы температурное поле было определено единственным образом? Ясно, что температура T будет определена полностью, если известны ее значения на границе C тела B . Действительно, если бы существовало два распределения тепла T_1 и T_2 с одинаковыми граничными значениями на C , то разность $T = T_1 - T_2$ была бы гармонич-

ческой функцией, равной нулю на C . Положив $U=V=T$ в уравнении (6.7) и взяв $B=b$, $C=c$ мы получили бы тогда, что

$$\int_B (\nabla T)^2 dx dy dz = 0.$$

Это значит, что $\nabla T \equiv 0$ и $T \equiv 0$ в B , т. е. что $T_1 = T_2$.

Легко понять, что стационарно распределенная температура с нулевыми граничными условиями должна быть равна нулю всюду. В самом деле, такое состояние мы сможем получить, если поместим тело B в жидкость постоянной температуры иждемся того, что установится тепловое равновесие. Ясно, что тело будет остывать до тех пор, пока его температура не станет равна температуре термостата, т. е. до $T \equiv 0$.

Точно так же можно показать, что если в каждой точке границы тела известен поток тепла наружу $-k \partial T / \partial n$, то температурное поле T этого тела определено с точностью до постоянной. Действительно, для этого надо лишь показать, что решение уравнения Лапласа, у которого нормальная производная равна нулю на границе C , равно постоянной. Последнее непосредственно следует из формулы (6.7), если положить в ней $U=V=T$ и за b взять все тело B . Этому выводу можно дать следующее интуитивное объяснение. Изолируем тело таким образом, чтобы тепло не проходило через стенку C , так что на границе $\partial T / \partial n = 0$. Единственным стационарным распределением тепла в таком случае будет $T = \text{const}$ в B .

Мы можем, наконец, поместить тело в вакуум, так что никакого теплообмена с окружающей средой происходить не будет. Тогда тело станет излучать тепло, принесенное на его границу, согласно уравнению $-k \partial T / \partial n = \lambda T$ на границе C . Постоянный режим установится только тогда, когда все тепло излучится, т. е. когда $T = 0$. Более формальным путем это выводится из формулы (6.7). Положим $U=V=T$, тогда мы получим, что

$$\int_B (\nabla T)^2 dx dy dz = -\frac{\lambda}{k} \int_C T^2 d\sigma.$$

Так как левая часть неотрицательна, а правая — неположительна, то обе они должны быть равны нулю и, следовательно, $T \equiv 0$.

Наш вывод можно сформулировать в виде теоремы единственности. Температура $T(x, y, z)$ единственным образом определяется в теле B заданием на границе C комбинации $\partial T / \partial n + \mu T$, если при этом μ равно положительной постоянной. Легко видеть, что это утверждение остается справедливым, если μ — любая положительная функция на поверхности C .

Мы установили три типа граничных условий, которые единственным образом определяют решение уравнения Лапласа:

- а) можно задавать значения самого решения на границе C ;
- б) можно задавать нормальную производную решения на границе C и значение решения в одной точке в теле B или на границе C ;
- в) можно задавать комбинацию $\partial T / \partial n + \mu T$ на C .

Теперь возникает вопрос, существует ли решение при произвольном задании краевых условий трех перечисленных типов и зависит ли решение непрерывным образом от этих условий.

Заметим сразу, что в случае второй краевой задачи мы не совсем свободны в выборе значений для нормальной производной. Мы должны, очевидно, удовлетворить условию

$$\int_C \frac{\partial T}{\partial n} d\sigma = 0, \quad (6.9)$$

которое является частным случаем равенства (6.6). Это ограничение компенсируется правом наложить дополнительное требование на значение T в какой-либо точке B или C .

В § 6.3 мы покажем, что все три краевые задачи всегда имеют единственное решение, лишь бы было выполнено названное выше очевидное ограничение. Однако прежде чем приступить к доказательству этого утверждения, мы рассмотрим дифференциальное уравнение более общего вида, к которому можно применить тот же способ рассуждения.

Это более общее дифференциальное уравнение можно вывести, рассматривая стационарное распределение температуры в теле B , в котором происходит эндотермическая химическая реакция, скорость которой пропорциональна локальной температуре T . Количество тепла, которое при такой реакции поглощается частью тела b в единицу времени, равно

$$k \int_b p(x, y, z) T(x, y, z) dx dy dz.$$

Здесь $p(x, y, z)$ есть положительная функция, которая зависит от локальной плотности реагирующего вещества. Количество поглощенного тепла должно компенсироваться за счет притока тепла через границу c объема b .

Следовательно,

$$k \int_b p T dx dy dz = k \int_c \frac{\partial T}{\partial n} d\sigma = k \int_b \nabla^2 T dx dy dz.$$

Таким образом, мы получаем дифференциальное уравнение

$$\nabla^2 T = pT, \quad p > 0 \text{ в } B. \quad (6.10)$$

Аналогично тому, как это было сделано выше, можно установить, что граничные условия определяют решение единственным образом.

Теперь нам будет удобно представить формулу Грина в следующем виде:

$$\begin{aligned} \int_b (\nabla U - \nabla V + pUV) dx dy dz = \\ = \int_c U \frac{\partial V}{\partial n} d\sigma - \int_b U (\nabla^2 V - pV) dx dy dz, \end{aligned} \quad (6.11)$$

так как последний интеграл в правой части этой формулы обращается в нуль, если функция совпадает с решением уравнения (6.10). Положив $U = V = T$, мы найдем, что для каждого решения уравнения (6.10)

$$\int_B [(\nabla T)^2 + pT^2] dx dy dz = \int_c T \frac{\partial T}{\partial n} d\sigma. \quad (6.12)$$

Если мы предположим теперь, что $\partial T / \partial n = 0$ на C , то получим из уравнения (6.12), что T должно быть равно нулю, а не просто произвольной постоянной. Это ясно также из нашей физической интерпретации: если изолировать тело B , то эндотермический процесс будет продолжаться и будет понижать температуру до тех пор, пока она не достигнет значения $T \equiv 0$. В случае более общего уравнения (6.10) вторую краевую задачу можно сформулировать проще. На границе C можно задавать значения нормальной производной решения. В этом случае для значений нормальной производной уже нет ограничений типа (6.9). Последнее ограничение вообще характерно для уравнения Лапласа и его, как правило, нет для более широкого класса эллиптических дифференциальных уравнений с частными производными второго порядка. Поэтому в качестве представителя всего класса эллиптических дифференциальных уравнений часто предпочитают брать не классическое уравнение Лапласа, а уравнение (6.10). Для того чтобы обеспечить справедливость теорем единственности для рассмотренных краевых задач, при этом еще предполагают, что $p > 0$.

§ 6.3. Фундаментальные решения и функции Грина. Пусть P и Q суть точки области B с координатами (x, y, z) и (ξ, η, ζ) соответственно. Рассмотрим функцию

$$S(P, Q) = \frac{1}{4\pi r(P, Q)}, \quad r(P, Q) = [(x - \xi)^2 + (y - \eta)^2 + (z - \zeta)^2]^{1/2}$$

как функцию точки P при фиксированных значениях координат (ξ, η, ζ) . Ясно, что $S(P, Q)$ есть решение уравнения Лапласа (6.8)

и что это решение становится неограниченным, когда точка P приближается к точке Q . Легко видеть, что

$$-\int_c \frac{\partial}{\partial n} S(P, Q) d\sigma = 1, \quad (6.13)$$

где c есть произвольная замкнутая поверхность, заключенная в области B и окружающая точку Q .

Чтобы наши формулы были более простыми, мы предположим, что единицы измерения выбраны так, чтобы коэффициент теплопроводности k был равен 1. Тогда условие (6.13) означает, что внутри любой поверхности c , окружающей точку Q , за единицу времени образуется единица количества тепла, которое выделяется наружу через поверхность c . Таким образом, точка Q оказывается источником тепла мощности 1, а функция $S(P, Q)$ описывает распределение тепла, которое возникает в области B благодаря источнику в точке Q . Функция $S(P, Q)$ называется *фундаментальным решением* уравнения Лапласа с полюсом в точке Q .

Представим себе теперь, что тело B погружено в жидкость с постоянной нулевой температурой и что в фиксированной точке Q внутри B помещен тепловой источник единичной мощности. Очевидно, что при этом установится постоянный режим и создастся постоянное во времени температурное поле $G(P, Q)$, которое будет зависеть от положения источника тепла Q . Функция $G(P, Q)$ единственным образом определяется следующими тремя свойствами:

а) $G(P, Q)$, как функция точки P , есть решение уравнения Лапласа;

б) $G(P, Q)$ ведет себя вблизи точки Q так же, как функция $S(P, Q)$;

в) $G(P, Q)$ обращается в нуль на границе C тела B .

Функция $G(P, Q)$, в существовании которой мы убедились интуитивно на основании рассмотрения физической задачи, играет важную роль в теории первой краевой задачи. Она называется *функцией Грина первого рода* уравнения Лапласа для области B .

Чтобы изучить функцию Грина, нам понадобится новое интегральное тождество, которое тесно связано с формулой Грина (6.7). Поменяем в выражении (6.7) функции U и V местами; при этом левая часть останется без изменения, а правая изменится. Вычитая это новое тождество из первоначального, мы получим вторую формулу Грина

$$\int_c \left(U \frac{\partial V}{\partial n} - V \frac{\partial U}{\partial n} \right) d\sigma = \int_b (U \nabla^2 V - V \nabla^2 U) dx dy dz, \quad (6.14)$$

которая справедлива для любых функций U и V , дважды непрерывно дифференцируемых в замкнутой области b с границей c .

Пусть область b_ϵ получается из области B исключением из нее сферы радиуса ϵ с центром в точке Q ; граница c_ϵ области b_ϵ состоит тогда из границы C области B и из поверхности Σ_ϵ малой сферы.

Положим $U = G(P, Q)$ и $V = T(P)$, где $T(P)$ — произвольное решение уравнения Лапласа в области B . Заметим, что функция G регулярна всюду в b_ε , так как ее особая точка вырезана сферой Σ_ε . Следовательно, мы можем воспользоваться формулой (6.14), которая дает нам

$$-\int_C T(P) \frac{\partial G(P, Q)}{\partial n} d\sigma + \int_{\Sigma_\varepsilon} \left(G \frac{\partial T}{\partial n} - T \frac{\partial G}{\partial n} \right) d\sigma = 0.$$

Это равенство справедливо для любого ε . При $\varepsilon \rightarrow 0$ предел второго интеграла легко вычислить, так как функция G ведет себя на поверхности Σ_ε по существу так же, как функция $1/4\pi r(P, Q)$. Используя условие (6.13), мы найдем таким образом, что в пределе, при $\varepsilon = 0$,

$$T(Q) = - \int_C T(P) \frac{\partial G(P, Q)}{\partial n} d\sigma. \quad (6.15)$$

Эта формула имеет фундаментальное значение. В ней решение $T(Q)$ уравнения Лапласа выражается внутри области B через свои значения на границе C . Обратно, можно показать, что если в качестве $T(P)$ взять произвольную непрерывную на границе C функцию, то интеграл в правой части равенства (6.15) будет равен решению с крайними значениями $T(P)$. Таким образом, формула (6.15) дает решение первой краевой задачи в явном виде.

Для того чтобы понять, почему интеграл в формуле (6.15) всегда представляет собой решение уравнения Лапласа, рассмотрим область b^* , которая получается из области B удалением двух сфер радиуса ε с центрами в точках Q и Q' соответственно. Применим к области b^* тождество (6.14), положив в нем $U = G(P, Q)$ и $V = G(P, Q')$. Рассуждая так же, как раньше, мы получаем, что в пределе при $\varepsilon \rightarrow 0$,

$$G(Q, Q') = G(Q', Q). \quad (6.16)$$

Это значит, что функция $G(P, Q)$ симметрична относительно точки, играющей роль аргумента, и точки источника. В частности, $G(P, Q)$, как функция Q , также есть решение уравнения Лапласа. Этим фактом объясняется то, что функция (6.15) всегда является решением. В самом деле, для того чтобы показать, что оператор Лапласа, примененный к интегралу в правой части, дает нуль, достаточно изменить порядок интегрирования по P и дифференцирования по Q .

Закон симметрии (6.16) интерпретируется следующим образом: пусть B — теплопроводящее тело, помещенное в термостат; тогда температура, которая создается в точке P единичным источником тепла, помещенным в точку Q , равна температуре в точке Q , вызванной единичным источником в точке P . Это утверждение составляет одну из многих теорем взаимности, встречающихся в математике. Почти все они сводятся к симметричности некоторых функций Грина, и их можно доказывать аналогичным образом.

Из представления (6.15) решения через его граничные значения видно, что решение непрерывно зависит от своих граничных значений и это, следовательно, оно устойчиво относительно малых изменений этих значений. Это показывает, что первая краевая задача поставлена корректно.

Мы можем, далее, поместить единичный источник тепла в точку Q тела B , а единичный сток тепла — в точку Q' и так изолировать граничную поверхность C , чтобы через нее не было никакого потока тепла. Тогда мы получим стационарное распределение тепла $N(P; Q, Q')$, при котором имеет место постоянный поток тепла из точки Q в точку Q' . Из такой физической интерпретации функции N выводятся следующие три ее свойства:

- а) $N(P, Q, Q')$, как функция P , есть решение уравнения Лапласа;
- б) вблизи точки Q функция $N(P; Q, Q')$ ведет себя как функция $S(P, Q)$, а вблизи точки Q' — как $-S(P, Q')$;
- в) производная $\partial N / \partial n$ тождественно равна нулю на границе C тела B .

Легко видеть, что эти три свойства определяют функцию N с точностью до аддитивной постоянной; $N(P; Q, Q')$ называется *функцией Неймана* или *функцией Грина второго рода* уравнения Лапласа для области B .

Мы можем теперь применить вторую формулу Грина (6.14) к произвольному решению $V = T(P)$ и к функции Неймана $N(P; Q, Q')$. Для этого нам надо удалить из B две сферы радиуса ϵ с центрами в точках Q и Q' и затем устремить ϵ к 0. В результате мы получим, что

$$T(Q) - T(Q') = \int_C N(P; Q, Q') \frac{\partial T(P)}{\partial n} d\sigma. \quad (6.17)$$

Итак, решение $T(Q)$ всюду в области B выражается через значение нормальной производной этого решения на границе и через значение самого решения в некоторой фиксированной точке Q' . Это представляет формальное решение второй краевой задачи. То обстоятельство, что функция N определена только с точностью до постоянной, не имеет для нас значения, так как мы заранее знаем, что нормальная производная функция T должна удовлетворять условию (6.9) и что, следовательно, аддитивная постоянная в N не изменяет интеграла в выражении (6.17).

Функцию Неймана часто определяют несколько иначе. При этом от функции $N(P; Q)$ требуют, чтобы она удовлетворяла следующим свойствам:

- а) $N(P, Q)$, как функция P , есть решение уравнения Лапласа;
- б) $N(P, Q)$ ведет себя вблизи источника Q так же, как функция $S(P, Q)$;
- в) $\partial N(P, Q) / \partial n = \text{const}$ на границе C .

При таком определении функция N описывает температурное распределение в теле B , которое возникает благодаря единичному источнику в точке Q , когда поток тепла через все поверхностные элементы на границе C один и тот же. Так как эти три условия определяют функцию N лишь с точностью до постоянной, то к ним присоединяется еще условие нормировки

$$г) \int_C N(P, Q) d\sigma = 0.$$

Из формулы Грина (6.14) легко видеть, что эта функция Неймана симметрична:

$$N(P, Q) = N(Q, P).$$

Ясно, что

$$N(P; Q) - N(P; Q') = N(P; Q, Q').$$

Из этого уравнения видна простая связь между функциями Неймана двух родов.

Можно показать, что если производная $\partial T/\partial n$ задана на границе C как непрерывная функция, удовлетворяющая условию (6.9), то правая часть уравнения (6.17) представляет собой решение уравнения Лапласа, нормальная производная которого совпадает с $\partial T/\partial n$. Это показывает, что для уравнений с частными производными вторая краевая задача поставлена корректно.

Мы можем, наконец, рассмотреть тепловой источник единичной мощности, расположенный в точке Q тела B , которое обменивается теплом посредством излучения и находится при этом в равновесии с окружающей средой. Стационарное распределение тепла описывается тогда функцией $R_\mu(P, Q)$, обладающей следующими тремя свойствами:

- а) $R_\mu(P, Q)$, как функция P , есть решение уравнения Лапласа;
- б) $R_\mu(P, Q)$ вблизи точки Q ведет себя как $S(P, Q)$;
- в) $\partial R_\mu/\partial n + \mu R_\mu = 0$ на границе C .

Функция $R_\mu(P, Q)$ называется *функцией Робена* или *функцией Грина третьего рода* уравнения Лапласа для области B . Кроме того, она зависит от положительной постоянной μ .

Пользуясь формулой Грина (6.14), мы можем легко установить справедливость закона симметрии:

$$R_\mu(P, Q) = R_\mu(Q, P).$$

Это равенство показывает, что R_μ есть решение уравнения Лапласа так же, как и функция источника; отсюда вытекает закон взаимности в теплопроводности. Наконец, при помощи тождества (6.14) можно показать, что для произвольного решения $T(P)$ справедлива формула

$$T(Q) = \int_C \left(\frac{\partial T}{\partial n} + \mu T \right) R_\mu(P, Q) d\sigma.$$

Если на границе C мы произвольно зададим значение функции $\partial T/\partial n + \mu T$, потребовав от нее лишь непрерывность, то в правой части этой формулы мы получим решение, у которого названная комбинация принимает как раз нужные значения.

Таким образом, мы показали, что третья краевая задача поставлена корректно, а ее фактическое решение сведено к определению функции Робена.

Прежде чем заняться вопросом о том, какое значение имеют различные функции Грина при решении краевых задач, постараемся распространить эти понятия на случай более общих уравнений (6.10). Главная задача состоит здесь в том, чтобы получить представление о фундаментальных решениях, которые возможны для этих уравнений. Определим фундаментальное решение $S(P, Q)$ следующими условиями:

а) $S(P, Q)$, как функция P , является в области B решением уравнения (6.10);

б) функция $1/[4\pi r(P, Q)] - S(P, Q)$ остается ограниченной, когда $P \rightarrow Q$;

в) для сферы Σ_ϵ радиуса ϵ , описанной вокруг точки Q , справедливо уравнение

$$\lim_{\epsilon \rightarrow 0} - \int_{\Sigma_\epsilon} \frac{\partial S(P, Q)}{\partial n} d\sigma = 1.$$

Функцию $S(P, Q)$ можно опять интерпретировать как распределение температуры, возникающее благодаря источнику единичной мощности, помещенному в точке Q . Позже мы рассмотрим, как для данного уравнения (6.10) найти фундаментальное решение. Теперь же мы легко можем определить три рода функций Грина для уравнения (6.10) в области B :

а) функция Грина (первого рода) $G(P, Q)$ есть фундаментальное решение, которое обращается в нуль на границе C ;

б) функция Неймана $N(P, Q)$ есть фундаментальное решение, которое имеет на границе C нулевую нормальную производную;

в) функция Робена $R_\mu(P, Q)$ есть фундаментальное решение, которое удовлетворяет на границе C условию $\partial R_\mu/\partial n + \mu R_\mu = 0$.

Заметим, что для нового уравнения (6.10) функция Неймана определяется гораздо проще. В самом деле, мы можем поместить источник тепла в точку Q тела B , изолировать граничную поверхность C и получить окончательное стационарное распределение температуры. Как только тело B достаточно нагреется, тепло, создаваемое источником, будет целиком поглощаться эндотермической химической реакцией.

В случае уравнения Лапласа мы пользовались формулой Грина (6.14). Представим теперь эту формулу в виде

$$\int_C \left(U \frac{\partial V}{\partial n} - V \frac{\partial U}{\partial n} \right) d\sigma = \int_B [U (\nabla^2 V - pV) - V (\nabla^2 U - pU)] dx dy dz. \quad (6.18)$$

В таком виде его можно пользоваться и в случае уравнения (6.10). Таким же способом, как и раньше, выводятся законы симметрии

$$G(P, Q) = G(Q, P), \quad N(P, Q) = N(Q, P), \quad R_\mu(P, Q) = R_\mu(Q, P)$$

и получаются следующие тождества, которые справедливы для любого решения $T(P)$ уравнения (6.10):

$$\begin{aligned} T(Q) &= - \int_G \frac{\partial G(P, Q)}{\partial n} T(P) d\sigma = \int_G N(P, Q) \frac{\partial T(P)}{\partial n} d\sigma = \\ &= \int_G \left(\frac{\partial T}{\partial n} + \mu T \right) R_\mu(P, Q) d\sigma. \end{aligned} \quad (6.19)$$

Таким образом, пользуясь функцией Грина, можно опять решить все три краевые задачи.

На практике при решении краевых задач часто бывает трудно найти функцию Грина и во многих случаях задачу проще решить прямыми численными методами без применения общей теории. Важность различных функций Грина заключается не столько в их вычислительной ценности, сколько в том, что они выявляют связь между различными решениями. При помощи функций Грина можно получать тождества, неравенства и законы симметрии для всевозможных важных частных решений. Изучая функции Грина, можно установить законы зависимости решения от граничных значений, от области определения и от коэффициентов дифференциального уравнения.

В явном виде уравнения с частными производными решаются только в немногих частных случаях, когда симметричность задачи позволяет разделить переменные и свести дело к простому обыкновенному дифференциальному уравнению. Решение дифференциального уравнения с частными производными есть функция точки, но оно есть также и сложная функция граничных значений, граничной поверхности и коэффициентов уравнения. Нам кажется безнадежным и, может быть, даже бесполезным отыскивать явные решения в наиболее общем случае. С другой стороны, вычислительные процессы никогда не дадут ясного понимания взаимной зависимости решения и различных наложенных на него условий. Поэтому представляется желательным обладать качественной теорией, в которой изучаются общие свойства решений. А в такой теории функции Грина являются, по-видимому, необходимым аппаратом.

§ 6.4. Принцип максимума, ядро, интеграл Дирихле. Рассмотрим теперь некоторые общие свойства, которые присущи всем решениям дифференциального уравнения

$$\nabla^2 u = pu, \quad p > 0 \quad \text{в } B, \quad (6.20)$$

когда функции u непрерывны во всей рассматриваемой области B . Заметим прежде всего, что эти функции должны вполне определенным

образом вести себя во внутренних точках области B , в которых они имеют локальный максимум или минимум. Действительно, если в точке P_0 достигается, например, локальный минимум, то в ней выполняются необходимые условия минимума

$$\frac{\partial^2 u}{\partial x^2} \geq 0, \quad \frac{\partial^2 u}{\partial y^2} \geq 0, \quad \frac{\partial^2 u}{\partial z^2} \geq 0.$$

Следовательно, в этой точке $\nabla^2 u \geq 0$, а это, в силу уравнения (6.20), означает, что

$$u(P_0) \geq 0 \text{ в точках минимума внутри } B.$$

Аналогично можно показать, что

$$u(P_0) \leq 0 \text{ в точках максимума внутри } B.$$

Таким образом, произвольное непрерывное решение уравнения (6.20) может вести себя в области B тремя различными способами:

а) решение не имеет внутри области B ни максимума, ни минимума, а принимает свои экстремальные значения на границе C ;

б) решение имеет минимум внутри B ; тогда этот минимум неотрицателен и, следовательно, u неотрицательно в области B ;

в) решение имеет максимум внутри B и неположительно в B .

Из приведенного результата сразу вытекает, что если решение $u(P)$ неотрицательно на границе C области B , то оно должно быть неотрицательно и внутри B . В самом деле, если бы оно было отрицательно внутри B , то оно принимало бы свое минимальное значение внутри B и это минимальное значение тоже было бы отрицательным. Но внутренний минимум должен быть неотрицательным, согласно (б), что приводит к противоречию. Из (в) очевидно также, что такое неотрицательное решение внутри области всегда меньше своего наибольшего значения на границе.

На практике при изучении решений очень ценно то, что мы можем делать качественные заключения, касающиеся поведения решения внутри области, коль скоро нам известны некоторые его свойства на границе. Предположим, что мы интересуемся решением $u(P)$ с заданными краевыми значениями, а известно нам другое решение $u_1(P)$. Если при этом по крайней мере известно, что $u_1(P)$ всюду на границе C больше чем $u(P)$, то мы можем утверждать, что $u(P) \leq u_1(P)$ всюду в B , потому что $u_1(P) - u(P)$ также является решением, которое неотрицательно на C . Таким образом, при помощи только одного известного решения можно получить оценки для многих решений.

Рассмотрим, далее, функцию Грина $G(P, Q)$ для уравнения (6.20) и области B . Опишем вокруг точки источника Q сферу малого радиуса; так как $G(P, Q)$ ведет себя в окрестности точки Q как $1/4\pi r$, то этот радиус можно выбрать столь малым, чтобы на сфере функция G принимала большие положительные значения. На границе C по определению G равно нулю. Таким образом, мы видим, что на

границе области, ограниченной поверхностью S и выбранной сферой, G имеет неотрицательные значения. Следовательно, решение $G(P, Q)$ неотрицательно в этой области, а, значит, G неотрицательно и в B . Можно даже показать, что

$$G(P, Q) > 0 \text{ в } B. \quad (6.21)$$

Так как функция Грина положительна внутри области и обращается в нуль на границе, то ее производная по нормали должна быть отрицательна. Таким образом,

$$\frac{\partial}{\partial n} G(P, Q) < 0 \text{ на } S. \quad (6.22)$$

Мы вывели этот результат из свойств произвольного решения уравнения (6.20)¹⁾. Обратно, из неравенства (6.22) вытекает характер общего класса решений. Действительно, из первого из тождеств (6.19) и из неравенства (6.22) можно заключить, что если решение неотрицательно на границе, оно неотрицательно во всей области. Таким образом, в одной формуле (6.21) или (6.22) для функции Грина заключена теорема для целого класса решений.

Неравенства (6.21) и (6.22) мы можем использовать теперь следующим образом: пусть $U(P)$ есть дважды непрерывно дифференцируемая в области B функция и пусть

$$\nabla^2 U - \rho U = \rho(P) \leq 0 \text{ в } B. \quad (6.23)$$

Воспользуемся тождеством (6.18), положив $V = G(P, Q)$ и взяв за b область B , из которой вырезана сфера Σ_ε радиуса ε , описанная вокруг точки Q . Полагая $\varepsilon \rightarrow 0$, мы получим, что

$$U(Q) = - \int_C U(P) \frac{\partial G(P, Q)}{\partial n} d\sigma - \int_B \rho(P) G(P, Q) dx dy dz.$$

Первый интеграл в правой части представляет собой решение $u(Q)$, уравнения (6.20), краевые значения которого равны $U(P)$, а значение второго интеграла неотрицательно. Поэтому для искомого решения $u(Q)$ и вспомогательной функции $U(Q)$ имеет место следующее неравенство

$$U(Q) \geq u(Q). \quad (6.24)$$

Таким образом, для того чтобы оценить решение по его граничным значениям, не обязательно даже иметь для сравнения другое решение. Для этого годится просто функция $U(P)$, удовлетворяющая дифференциальному неравенству (6.23). Теперь уже ясно, что умелой

¹⁾ Из рассуждения автора вытекает более слабый результат: $\partial G / \partial n \leq 0$ (на S). На самом же деле имеет место действительно неравенство (6.22), которое строго доказано в работе О. А. Олейник [6].— *Прим. ред.*

комбинацией различных методов оценок можно получить достаточное количество сведений об искомом решении.

Положительный характер функции Грина становится неудивительным, если принять во внимание данную нами физическую интерпретацию. Действительно, представляется невероятным, чтобы совместное действие термостата с нулевой температурой, химического процесса, который выделяет тепло, если температура падает ниже нуля, и положительного источника тепла дало где-нибудь отрицательную температуру. Более того, мы предполагаем, конечно, что изолированное тело с положительным источником тепла всюду имеет неотрицательную температуру. Другими словами, мы можем ожидать для функции Неймана неравенство

$$N(P, Q) > 0 \text{ в } B. \quad (6.25)$$

Действительно, предположим, что при фиксированной точке Q существовали бы точки P в области B , в которых функция $N(P, Q)$ отрицательна. Такие точки заполнили бы подобласть b . Точка Q не принадлежит b , так как N делается большим и положительным, когда P приближается к Q . На границе области b либо $N = 0$ (если граничная точка b лежит внутри B), либо $\partial N / \partial n = 0$ (если граничная точка подобласти b является также и граничной точкой B). Так как функция N непрерывна в b , то мы можем применить формулу Грина (6.11), положив $U = V = N$. Тогда получится

$$\int_b \{ [\nabla N(P, Q)]^2 + p [N(P, Q)]^2 \} dx dy dz = 0.$$

Следовательно, $N(P, Q) = 0$ в b , а это противоречит нашему предположению, что в b есть точки, в которых функция N отрицательна. Итак, ясно, что N неотрицательна. Отсюда легко вывести более сильное неравенство (6.25).

Таким же способом можно показать, что функция Робина

$$R_\mu(P, Q) > 0 \text{ в } B.$$

и более того, что

$$R_\mu(P, Q) > R_{\mu'}(P, Q), \text{ если } \mu < \mu'.$$

Рассмотрим теперь функцию

$$K(P, Q) = N(P, Q) - G(P, Q). \quad (6.26)$$

Так как функции Неймана и Грина имеют в точке Q одинаковые особенности, их разность регулярна в Q , и представляет собой непрерывное решение уравнения (6.20) во всей области B . На границе S функция Грина обращается в нуль, тогда как функция Неймана неотрицательна. Поэтому функция $K(P, Q)$ неотрицательна на S и, следовательно,

$$K(P, Q) > 0 \text{ в } B.$$

Этот результат тоже можно было предсказать, пользуясь нашей моделью. Он утверждает, что $N > G$, или что температура в изолированном теле выше, чем в теле, которое охлаждается жидкостью постоянной температуры.

Функция $K(P, Q)$ называется *ядром* уравнения (6.20) для области B . Она имеет большое теоретическое значение и, кроме того, может использоваться для вычислительных целей. Заметим, что на границе C

$$K(P, Q) = N(P, Q), \quad \frac{\partial K(P, Q)}{\partial n} = -\frac{\partial G(P, Q)}{\partial n}, \quad P \in C,$$

так что, зная K , мы можем решать как первую, так и вторую краевую задачу. Действительно, согласно тождеству (6.19), всякое решение $u(P)$ уравнения (6.20) можно записать в следующих видах:

$$u(Q) = \int_C u(P) \frac{\partial K(P, Q)}{\partial n} d\sigma = \int_C \frac{\partial u(P)}{\partial n} K(P, Q) d\sigma. \quad (6.27)$$

Мы лучше всего поймем смысл ядра, если воспользуемся формулой Грина (6.11) и преобразуем два поверхностных интеграла в уравнении (6.27) в такие же интегралы по пространству:

$$u(Q) = \int_B [\nabla K(P, Q) \cdot \nabla u(P) + pK(P, Q)u(P)] dx dy dz. \quad (6.28)$$

Для дальнейшего будет полезно ввести следующее обозначение:

$$D[u, v] = \int_B (\nabla u \cdot \nabla v + puv) dx dy dz,$$

где $D[u, v]$ определено для любой пары непрерывно дифференцируемых в B функций u и v . Выражение $D[u, v]$ линейно относительно каждой функции, поэтому для произвольных значений постоянных α_1 и α_2 имеем

$$D[\alpha_1 u_1 + \alpha_2 u_2, v] = \alpha_1 D[u_1, v] + \alpha_2 D[u_2, v]. \quad (6.29)$$

Оператор $D[u, v]$ можно рассматривать как обобщение скалярного произведения двух векторов в пространстве функций; $D[u, v]$ называется *произведением Дирихле* функций u и v .

Пользуясь геометрическим языком, мы будем называть функции u, v *ортogonalными* между собой, если $D[u, v] = 0$. *Нормой* u назовем выражение

$$[D[u, u]]^{1/2} = \|u\| = \left\{ \int_B [(\nabla u)^2 + pu^2] dx dy dz \right\}^{1/2}.$$

Норма функции соответствует длине вектора, так как последняя определяется через скалярное произведение аналогичным образом.

Пусть теперь u и v — две произвольные функции, а t — постоянная. Из структуры интеграла Дирихле видно, что

$$D[u + tv, u + tv] \geq 0, \quad (6.30)$$

а из (6.29) легко следует, что

$$D[u + tv, u + tv] = D[u, u] + 2tD[u, v] + t^2D[v, v]. \quad (6.31)$$

Пусть теперь u и v фиксированы, а t принимает произвольные действительные значения. В силу неравенства (6.30) квадратичный по t полином, стоящий в правой части равенства (6.31), не может менять знак. Поэтому для его дискриминанта справедливо неравенство

$$\{D[u, v]\}^2 \leq D[u, u] D[v, v]. \quad (6.32)$$

Таким образом, мы вывели для произведения Дирихле неравенство Шварца; оно означает, что косинус угла между функциями u и v

$$\cos(u, v) = \frac{D[u, v]}{\{D[u, u] D[v, v]\}^{1/2}}$$

по абсолютной величине никогда не превосходит единицы. Это условие необходимо для того, чтобы геометрическая интерпретация имела место.

В наших новых обозначениях уравнение (6.28) можно записать в следующем изящном виде:

$$u(Q) = D[K(P, Q), u(P)]. \quad (6.33)$$

Так как функция $K(P, Q')$ сама является решением, то, в частности, мы найдем, что

$$K(Q, Q') = D[K(P, Q), K(P, Q')]. \quad (6.34)$$

Применив теперь неравенство Шварца (6.32), мы получим, что

$$u(Q)^2 \leq D[u, u] D[K(P, Q), K(P, Q)].$$

Отсюда, согласно уравнению (6.34), следует, что

$$u(Q)^2 \leq \|u\|^2 K(Q, Q).$$

Это важный теоретический результат. Он дает оценку для решения $u(Q)$ в каждой точке области через ядро и норму $\|u\|$. В большинстве приложений бывает интуитивно ясно, что величина $\|u\|$ имеет значение полной энергии поля или означает что-либо аналогичное. Мы получили наилучшую оценку. В частном случае, когда $u(P) = K(P, Q)$, неравенство превращается в равенство.

Нередко оказывается, что нам известно бесконечное множество решений дифференциального уравнения, но тем не менее краевую задачу решить трудно. Например, в случае уравнения Лапласа сферические функции дают нам такое бесконечное множество решений, но все же подойти к решению краевой задачи для области B произвольной

формы чрезвычайно трудно. В подобном случае ядро открывает нам путь к вычислению решения. Пусть $\{u_\nu(P)\}$ — множество решений некоторого уравнения (6.20). Пользуясь простым алгоритмом, предложенным Грамом и Шмидтом, можно так скомбинировать эти функции, что в результате будет

$$D[u_\nu, u_\mu] = \delta_{\nu\mu} = \begin{cases} 0, & \text{если } \nu \neq \mu, \\ 1, & \text{если } \nu = \mu. \end{cases} \quad (6.35)$$

Не ограничивая общности, можно предполагать, что функции $u_\nu(P)$ уже приведены к этому виду; назовем эту систему *ортogonalной системой* в метрике Дирихле. Предположим еще, что функции $u_\nu(P)$ образуют *полную систему*, т. е. что всякую функцию $u(P)$, которая является решением уравнения (6.20) в области B , можно разложить в ряд

$$u(P) = \sum_{\nu=1}^{\infty} c_\nu u_\nu(P).$$

Пользуясь уравнением (6.35), легко определить значение каждого c_ν :

$$c_\nu = D[u, u_\nu]. \quad (6.36)$$

Это равенство является обобщением хорошо известной формулы из теории рядов Фурье.

В частности, в такой ряд можно разложить ядро $K(P, Q)$, положив

$$K(P, Q) = \sum_{\nu=1}^{\infty} c_\nu(Q) u_\nu(P).$$

Если мы вычислим теперь коэффициенты $c_\nu(Q)$ по формуле (6.36) и воспользуемся уравнением (6.33), которое справедливо для всех решений, то получим, что

$$K(P, Q) = \sum_{\nu=1}^{\infty} u_\nu(P) u_\nu(Q).$$

Это очень изящная и короткая формула для ядра, которую можно положить в основу при численных расчетах.

Ядро можно определить и для уравнения Лапласа. Однако в этом случае необходимы некоторые видоизменения, потому что функция Неймана определяется здесь иначе. Заметим еще, что в соответствующей метрике Дирихле, которая основана на скалярном произведении

$$D[u, v] = \int_B \nabla u \cdot \nabla v \, dx \, dy \, dz,$$

существуют не равные тождественно нулю функции u с нулевой нормой. Эти функции равны постоянным, и их нужно исключить, наложив дополнительные ограничения. Если ограничиться рассмотрением

всех решений $u(P)$ уравнения Лапласа, которые удовлетворяют условию

$$\int_C u d\sigma = 0,$$

то в результате получится как раз ядро вида (6.26).

§ 6.5. Примеры из гидродинамики и электростатики. Теперь мы покажем на нескольких примерах, как, используя теорию функций Грина, можно единообразно подойти к изучению различных важных физических фактов. Рассмотрим опять тело B с граничной поверхностью C и рассмотрим область D , внешнюю относительно B . Область D содержит бесконечно удаленную точку, которую мы обозначим через ∞ . Рассмотрим класс всех гармонических в области D функций, т. е. всех функций, которые удовлетворяют в D уравнению Лапласа и которые убывают на бесконечности по крайней мере как $1/r$. Заметим, что при таком определении единственной гармонической постоянной будет постоянная, равная нулю. Все гармонические в D функции имеют конечный интеграл Дирихле

$$D[u, u] = \int_D (\nabla u)^2 dx dy dz,$$

и, следовательно, в силу неравенства Шварца, любые две гармонические функции u и v имеют конечное произведение Дирихле

$$D[u, v] = \int_D \nabla u \cdot \nabla v dx dy dz.$$

В этом классе единственной функцией с нулевой нормой является функция $u = 0$.

Мы видим, таким образом, что поведение гармонических функций в бесконечной области больше напоминает поведение решений уравнения (6.20), чем решений уравнения Лапласа в конечной области. Можно показать, что в области D существует гармоническая функция Неймана $N(P, Q)$, у которой нормальная производная на C равна нулю и которая положительна в D . Поэтому все результаты, которые мы получили в предыдущем параграфе, остаются справедливыми и в этом случае.

Для того чтобы интерпретировать функции Неймана, рассмотрим во внешней к телу B области D установившийся поток несжимаемой жидкости. Предположим, что поток безвихревой, т. е. что его поле скоростей $\mathbf{q}$ допускает потенциал скорости φ :

$$\mathbf{q} = -\nabla \varphi. \quad (6.37)$$

Из несжимаемости жидкости следует, что общее количество жидкости, которое втекает за единицу времени через границу внутрь области b ,

расположенной в D , должно быть равно нулю:

$$\int_C \mathbf{q} \cdot \mathbf{n} \, d\sigma = - \int_C \frac{\partial \varphi}{\partial n} \, d\sigma = 0.$$

В силу формулы Грина отсюда следует, что

$$\nabla^2 \varphi = 0.$$

Мы предполагаем, что граница C области течения непроницаема, так что нормальная составляющая скорости течения на C должна быть равна нулю. В силу уравнения (6.37) отсюда следует, что

$$\frac{\partial \varphi}{\partial n} = 0 \quad \text{на } C.$$

Так как потенциал φ определен только с точностью до аддитивной постоянной, то мы можем нормировать его так, чтобы он обращался в нуль на бесконечности и, следовательно, вел себя там как гармоническая функция.

Предположим, наконец, что поток порожден источником единичной мощности, расположенным в точке Q области D . Тогда φ ведет себя вблизи Q как $1/4\pi r(P, Q)$.

Собрав все данные, которыми характеризуется функция $\varphi(P)$, мы видим, что потенциал скоростей совпадает с функцией Неймана $N(P, Q)$ области D и тем самым придает последней простой гидродинамический смысл.

Считая Q постоянным, разложим $N(P, Q)$, как функцию P , в ряд в окрестности бесконечно удаленной точки. Мы получим разложение по сферическим функциям

$$N(P, Q) = \frac{a(Q)}{4\pi r} + \frac{\varphi_1(Q)x + \varphi_2(Q)y + \varphi_3(Q)z}{4\pi r^3} + \dots \quad (6.38)$$

Постараемся выяснить смысл коэффициентов этого разложения. Рассмотрим для этого область $D_{\epsilon, R}$, которая ограничена поверхностью C , малой сферой радиуса ϵ с центром в Q и большой сферой Σ_R радиуса R с центром в начале координат, от которого отсчитывается r . Мы предполагаем, что начало координат лежит внутри тела B . Применим к области $D_{\epsilon, R}$ формулу Грина (6.14), положив $U = N(P, Q)$ и $V = 1$. Замечая, что $\partial N / \partial n = 0$ на C , и переходя к пределу при $\epsilon \rightarrow 0$, мы находим, что

$$1 = \int_{\Sigma_R} \frac{\partial N}{\partial n} \, d\sigma = \int_{\Sigma_R} \frac{\partial N}{\partial r} \, d\sigma = a(Q) + O\left(\frac{1}{R}\right),$$

где $O(1/R)$ — остаточный член, который при неограниченном возрастании R стремится к нулю как $1/R$. При $R \rightarrow \infty$ мы найдем, что

$$a(Q) = 1.$$

Этот результат следовало ожидать; он состоит в утверждении, что функция Неймана с источником единичной мощности в точке Q должна иметь в бесконечности сток той же мощности 1.

Применим, далее, к области $D_{\varepsilon, R}$ вторую формулу Грина (6.14), положив $U = N(P, Q)$ и $V = x$. Если точка Q имеет координаты (ξ, η, ζ) , то в пределе, при $\varepsilon \rightarrow 0$, мы получим, что

$$-\xi + \int_C N(P, Q) \frac{\partial x}{\partial n} d\sigma + \int_{\Sigma_R} \left(N \frac{\partial x}{\partial r} - x \frac{\partial N}{\partial r} \right) d\sigma = 0. \quad (6.39)$$

Заметим, что по соображениям симметрии

$$\int_{\Sigma_R} x^2 d\sigma = \int_{\Sigma_R} y^2 d\sigma = \int_{\Sigma_R} z^2 d\sigma = \frac{4\pi}{3} R^4$$

и

$$\int_{\Sigma_R} xy d\sigma = \int_{\Sigma_R} xz d\sigma = \int_{\Sigma_R} yz d\sigma = 0.$$

Воспользуемся еще тем, что $\partial x / \partial r = x/R$ на Σ_R . Путем несложных выкладок из равенств (6.38) и (6.39) можно теперь получить, что

$$\varphi_1(Q) = \xi - \int_C N(P, Q) \frac{\partial x}{\partial n} d\sigma = \xi + \Phi_1(Q).$$

Ясно, что в каждой конечной точке области D $\varphi_1(Q)$ есть решение уравнения Лапласа. Так как функция $\varphi_1(Q)$ гармонична в D , а ее производная по нормали, очевидно, равна $-\partial \xi / \partial n$ на границе C , то производная по нормали функции $\varphi_1(Q)$ равна нулю на C . Поэтому $\varphi_1(Q)$ можно считать потенциалом скорости несжимаемого безвихревого стационарного потока, обтекающего тело B . Вблизи бесконечности $\varphi_1(Q) \sim \xi$; таким образом, $\varphi_1(Q)$ является потенциалом скорости обтекающего тело потока, направленного вдаль от тела в сторону отрицательных x . Это — потенциал относительной скорости, который возникает, когда тело B движется в покоящейся жидкости в направлении оси x .

Очевидно, что функции φ_2 и φ_3 играют роль, аналогичную роли φ_1 . Поток, обтекающий тело B и имеющий направление единичного вектора $(\alpha_1, \alpha_2, \alpha_3)$, можно задать его потенциалом скорости

$$\varphi(Q) = \alpha_1 \varphi_1(Q) + \alpha_2 \varphi_2(Q) + \alpha_3 \varphi_3(Q).$$

Каждую из функций $\varphi_v(P)$ вблизи бесконечности опять можно разложить по сферическим функциям

$$\varphi_v(P) = \frac{a_v}{4\pi r} + \frac{e_{v1}x + e_{v2}y + e_{v3}z}{4\pi r^3} + \dots,$$

а коэффициенты разложения можно найти точно таким же приемом. При этом мы получим, что $a_\nu = 0$ и

$$e_{\nu\mu} = - \int_C \varphi_\nu \frac{\partial x_\mu}{\partial n} d\sigma - \int_C x_\mu \frac{\partial \varphi_\nu}{\partial n} d\sigma. \quad (6.40)$$

(Здесь x, y, z мы для единообразия заменили на x_1, x_2, x_3 .) Воспользовавшись формулой Грина (6.7), преобразуем теперь первый интеграл в формуле (6.40) в объемный интеграл по области D , заметив при этом, что $\partial x_\mu / \partial n$ можно заменить на $-\partial \varphi_\mu / \partial n$. Второй поверхностный интеграл преобразуется в объемный интеграл по B при помощи той же формулы. Таким образом, мы найдем, что

$$e_{\nu\mu} = \int_D \nabla \varphi_\nu \cdot \nabla \varphi_\mu dx dy dz + \int_B \nabla x_\nu \cdot \nabla x_\mu dx dy dz.$$

Ясно, что последний интеграл равен нулю, когда $\nu \neq \mu$, и равен объему V тела B , если $\nu = \mu$. Таким образом, окончательно

$$e_{\nu\mu} = \int_D (\nabla \varphi_\nu \cdot \nabla \varphi_\mu) dx dy dz + \delta_{\nu\mu} V. \quad (6.41)$$

Эта формула показывает, что матрица коэффициентов $e_{\nu\mu}$ симметрична. Для того чтобы понять физический смысл коэффициентов $e_{\nu\mu}$, рассмотрим тело B , которое движется со скоростью c в направлении $(\alpha_1, \alpha_2, \alpha_3)$. При этом оно индуцирует в жидкости движение с потенциалом скорости

$$\varphi(P) = c(\alpha_1 \varphi_1 + \alpha_2 \varphi_2 + \alpha_3 \varphi_3)$$

и с кинетической энергией

$$T = \frac{c^2}{2} \int_D (\nabla \varphi)^2 dx dy dz = \frac{c^2}{2} \sum_{\mu, \nu=1}^3 \alpha_\mu \alpha_\nu \int_D (\nabla \varphi_\mu \cdot \nabla \varphi_\nu) dx dy dz.$$

(Мы считаем здесь, что плотность жидкости равна единице.) Таким образом, для кинетической энергии индуцированного движения жидкости мы получим, в силу уравнения (6.41), следующее выражение:

$$T = \frac{1}{2} c^2 \left(\sum_{\mu, \nu=1}^3 e_{\mu\nu} \alpha_\mu \alpha_\nu - V \right).$$

Суммарная кинетическая энергия, возникающая при движении тела B , равна $\frac{1}{2} mc^2 + T$, где m — масса тела. Выражение

$$W = \sum_{\mu, \nu=1}^3 e_{\mu\nu} \alpha_\mu \alpha_\nu - V \quad (6.42)$$

равно кажущемуся увеличению массы, которое возникает благодаря инерции жидкости; величина W и тензор $(e_{\mu\nu} - V\delta_{\mu\nu})$ называются соответственно *присоединенной массой* в направлении α и *тензором присоединенной массы*. Мы только что выяснили их простую связь с функцией Неймана области.

Такие же формальные операции можно повторить теперь в применении к функции Грина и получить соответствующие величины, которым на этот раз можно дать электростатическую интерпретацию. В самом деле, функцию Грина $G(P, Q)$ можно представлять себе как электростатический потенциал вне заземленного проводника B , который создается единичным зарядом, помещенным в точке Q . Если вблизи бесконечности разложить функцию Грина по сферическим функциям, то получится ряд

$$G(P, Q) = \frac{b(Q)}{4\pi r} - \frac{\Psi_1(Q)x + \Psi_2(Q)y + \Psi_3(Q)z}{4\pi r^3} + \dots$$

Коэффициенты этого разложения, как и раньше, можно определить, пользуясь для этого второй формулой Грина. Мы найдем, что

$$b(Q) = 1 + \int_C \frac{\partial G(P, Q)}{\partial n} d\sigma = 1 - V(Q).$$

Здесь $V(Q)$ — гармоническая в области D функция, которая принимает на границе C значение 1. Она называется *емкостным потенциалом* проводника B и описывает потенциал равновесного электрического заряда, который распределен по поверхности C проводника. Функция $b(Q)$ удовлетворяет уравнению Лапласа и обращается в нуль на границе C , но в бесконечности она не гармоническая. Ее можно рассматривать как функцию Грина области D с источником в бесконечности. Далее мы найдем, что

$$\Psi_v(Q) = \xi_v + \int_C x_v \frac{\partial G(P, Q)}{\partial n} d\sigma = \xi_v - \psi_v(Q).$$

Здесь функция $\psi_v(Q)$ гармоническая в области D , а на C она принимает граничные значения ξ_v . Следовательно, $\Psi_v(P)$ удовлетворяет уравнению Лапласа, обращается в нуль на границе C , а вблизи бесконечности ведет себя как x_v . Такое электростатическое поле можно получить, поместив заземленный проводник B в поле конденсатора, когда направление поля равно x_v . Это однородное поле создает в проводнике B индуцированный заряд и поляризует B таким образом, что возникает индуцированный потенциал $-\psi_v(P)$.

Разложение в ряд показывает, что

$$V(Q) = \frac{k}{4\pi r} + \dots,$$

где

$$K = \int_C \frac{\partial V}{\partial n} d\sigma$$

есть заряд, который надо сообщить проводнику B для того, чтобы создать на границе C потенциал, равный 1; величина K равна электростатической емкости тела B , она существенна во многих приложениях электростатики.

Заряд, который индуцируется на проводнике B внешним однородным полем направления x_v , определяется выражением

$$k_v = - \int_C \frac{\partial \Psi_v}{\partial n} d\sigma.$$

Если проводнику B сообщить такой заряд, то даже без внешнего поля он создаст в области D потенциал $(k_v/K)V(P)$. Остальное поле с потенциалом

$$-\chi_v(P) = \frac{k_v}{K}V(P) - \psi_v(P)$$

вызвано поляризацией зарядов на границе C при наличии электростатического поля. Это индуцированный потенциал, который возникнет, если во внешнее поле направления x_v внести изолированный проводник B . Ясно, что $\chi_v(P)$ принимает на C граничное значение $x_v - k_v/K$.

Мы можем взять разложение

$$\chi_v(P) = - \frac{f_{v1}x + f_{v2}y + f_{v3}z}{4\pi r^3} + \dots$$

и найти связь коэффициентов $f_{v\mu}$ с функцией $\chi_v(P)$ так же, как в предыдущих рассмотрениях мы нашли связь коэффициентов $e_{v\mu}$ с функцией $\varphi_v(P)$. При этом оказывается, что

$$f_{v\mu} = \int_D (\nabla \chi_v \cdot \nabla \chi_\mu) dx dy dz + \delta_{v\mu} V. \quad (6.43)$$

Интерпретация матрицы $(f_{v\mu})$ очень похожа на интерпретацию матрицы $(e_{v\mu})$. Предположим, что нам дано однородное электростатическое поле силы c , имеющее направление единичного вектора α , и пусть в это поле вносится изолированный проводник B . При этом возникнет индуцированное поле с потенциалом

$$\chi = -c(\alpha_1\chi_1 + \alpha_2\chi_2 + \alpha_3\chi_3),$$

энергия которого определяется выражением

$$E = \frac{1}{2} \int_D (\nabla \chi)^2 dx dy dz = \frac{c^2}{2} \left(\sum_{\mu, \nu=1}^3 f_{\nu\mu} \alpha_\nu \alpha_\mu - V \right).$$

Тензор $(f_{\nu\mu})$ называется *тензором поляризации* тела B . Ясно, что он весьма существенен для описания электростатического состояния проводника B .

Теперь нам уже очевидна формальная аналогия между тензорами $(e_{\nu\mu})$ и $(f_{\nu\mu})$. Но между ними существует также более глубокая связь. Заметим сначала, что мы можем вычислить произведение Дирихле

$$D[\varphi_\nu, \chi_\mu] = \int_D \nabla \varphi_\nu \cdot \nabla \chi_\mu \, dx \, dy \, dz = \int_C \chi_\mu \frac{\partial \varphi_\nu}{\partial n} \, d\sigma$$

в явном виде. Действительно, мы знаем, что

$$\chi_\mu = x_\mu - \frac{k_\mu}{K}, \quad \frac{\partial \varphi_\nu}{\partial n} = -\frac{\partial x_\nu}{\partial n} \quad \text{на } C.$$

Отсюда легко получить, что

$$D[\varphi_\nu, \chi_\mu] = - \int_C \left(x_\mu - \frac{k_\mu}{K} \right) \frac{\partial x_\nu}{\partial n} \, d\sigma = \delta_{\nu\mu} V.$$

Следует заметить, что и вообще произведение Дирихле функции, удовлетворяющей граничному условию первого рода, на функцию, удовлетворяющую граничному условию второго рода, можно вычислить в явном виде.

Рассмотрим произведение Дирихле

$$D \left[\sum_{\nu=1}^3 \alpha_\nu \varphi_\nu, \sum_{\mu=1}^3 \beta_\mu \chi_\mu \right] = \sum_{\nu=1}^3 \alpha_\nu \beta_\nu V.$$

Согласно неравенству Шварца (6.32), имеем

$$\left(\sum_{\nu=1}^3 \alpha_\nu \beta_\nu \right)^2 V^2 \leq D \left[\sum_{\nu=1}^3 \alpha_\nu \varphi_\nu, \sum_{\nu=1}^3 \alpha_\nu \varphi_\nu \right] D \left[\sum_{\mu=1}^3 \beta_\mu \chi_\mu, \sum_{\mu=1}^3 \beta_\mu \chi_\mu \right].$$

Воспользовавшись формулами (6.41) и (6.43), мы получим окончательно, что

$$\left(\sum_{\nu=1}^3 \alpha_\nu \beta_\nu \right)^2 V^2 \leq \left(\sum_{\mu, \nu=1}^3 e_{\nu\mu} \alpha_\nu \alpha_\mu - V \right) \left(\sum_{\mu, \nu=1}^3 f_{\nu\mu} \beta_\nu \beta_\mu - V \right).$$

Таким образом, зная присоединенную массу данного тела в данном направлении, мы получаем оценки для поляризации того же тела во всех остальных направлениях и наоборот. Зная электростатические свойства, мы получаем некоторые сведения о гидродинамических изменениях. Положив, в частности, $\alpha = \beta$, мы найдем, что

$$V^2 \leq WP, \quad (6.44)$$

где W — виртуальная масса, а P — поляризация в том же направлении. Так, почти совсем плоское тело B будет испытывать малое сопротивление при движении в жидкости, если это движение проис-

ходит в плоскости тела, но, согласно неравенству (6.44), оно будет сильно поляризоваться электростатическим полем того же направления. Это утверждение кажется интуитивным, неравенство же (6.44) дает его численную формулировку; оценка здесь тоже наилучшая, потому что для сферы она выполняется со знаком равенства.

§ 6.6. Изменение функций Грина с изменением области. Величины, которые мы встречали до сих пор, — функции Грина, различные потенциалы скорости и электростатические потенциалы, тензоры виртуальной массы и поляризации, емкость — все зависят от геометрической формы тела или, что то же самое, от граничной поверхности S . Теперь мы займемся выяснением того, как изменяются эти величины с изменением тела B или границы S , причем то обстоятельство, что все эти величины выражаются через функции Грина, подчеркивает важность этих последних.

Рассмотрим область D , внешнюю относительно тела B с граничной поверхностью S . Пусть B^* — меньшее тело, которое лежит в B , S^* — его граничная поверхность и D^* — область, внешняя относительно B^* . Ясно, что D является подобластью области D^* . Пусть $G(P, Q)$ и $G^*(P, Q)$ — функции Грина областей D и D^* соответственно. Функция

$$d(P, Q) = G^*(P, Q) - G(P, Q)$$

определена в области D и представляет собой гармоническую функцию в этой области. Применим к функции $d(P, Q)$ формулу Грина (6.7) и воспользуемся тем, что G обращается в нуль на границе S . Мы получим тогда, что для двух произвольных точек Q и Q' из области D

$$\int_D [\nabla d(P, Q) \cdot \nabla d(P, Q')] dx dy dz = \int_S G^*(P, Q) \frac{\partial d(P, Q')}{\partial n} d\sigma. \quad (6.45)$$

Применим теперь вторую формулу Грина (6.14), положив в ней $U = G^*$ и $V = d$. Так как G^* обращается в бесконечность при $P = Q$, то, прежде чем применять формулу Грина, эту точку надо вырезать сферой малого радиуса. Мы получим, что

$$\int_S G^*(P, Q) \frac{\partial d(P, Q')}{\partial n} d\sigma = \int_C d(P, Q') \frac{\partial G^*(P, Q)}{\partial n} d\sigma + d(Q, Q'). \quad (6.46)$$

Воспользовавшись опять тем, что функция G обращается в нуль на границе S , мы получим

$$\int_C d(P, Q') \frac{\partial G^*(P, Q)}{\partial n} d\sigma = \int_C G^*(P, Q') \frac{\partial G^*(P, Q)}{\partial n} d\sigma. \quad (6.47)$$

Применяя, наконец, формулу Грина (6.7) к функции G^* и области D^* — D , в которой G^* непрерывна, мы найдем (в силу того, что

внешняя нормаль к границе D есть внутренняя нормаль к границе $D^* - D$,

$$\int_C G^*(P, Q') \frac{\partial G^*(P, Q)}{\partial n} d\sigma = - \int_{D^* - D} \nabla G^*(P, Q') \cdot \nabla G^*(P, Q) dx dy dz. \quad (6.48)$$

Собрав выражения (6.45) — (6.48), мы получим в результате равенство

$$\begin{aligned} G^*(Q, Q') - G(Q, Q') = & \int_{D^* - D} [\nabla G^*(P, Q) \cdot \nabla G^*(P, Q')] dx dy dz + \\ & + \int_D [\nabla d(P, Q) \cdot \nabla d(P, Q')] dx dy dz. \end{aligned} \quad (6.49)$$

Предположим теперь, что поверхность C^* можно получить, сдвинув каждую точку P поверхности C вдоль внешней нормали на величину δn . Мы считаем, что поверхность C^* расположена близко к поверхности C , т. е. что $\varepsilon = \max(\delta n)$ есть малая величина. Функция $d(Q, Q') = G^*(Q, Q') - G(Q, Q')$ вместе со всеми своими производными стремится в области D к нулю, когда ε стремится к нулю. Заметим, что во второй интеграл в выражении (6.49) входят произведения производных от d и что поэтому он убывает быстрее, чем d , когда $\varepsilon \rightarrow 0$. Первый интеграл есть величина порядка ε , так как объем области $D^* - D$ имеет именно этот порядок стремления к нулю. Если в первом интеграле заменить G^* на G , то ошибка при этом будет малой высшего порядка по сравнению с ε . Таким образом, мы получим окончательно, что

$$G^*(Q, Q') - G(Q, Q') = \int_C [\nabla G(P, Q) \cdot \nabla G(P, Q')] \delta n d\sigma + o(\varepsilon),$$

где ошибка $o(\varepsilon)$ есть малая высшего порядка по сравнению с ε , т. е. $o(\varepsilon)/\varepsilon \rightarrow 0$, когда $\varepsilon \rightarrow 0$. Пользуясь обычными обозначениями вариационного исчисления, мы можем коротко записать, что

$$\delta G(Q, Q') = \int_C [\nabla G(P, Q) \cdot \nabla G(P, Q')] \delta n d\sigma. \quad (6.50)$$

Рассуждая точно так же, мы получаем соответствующую вариационную формулу для функции Неймана

$$\delta N(Q, Q') = - \int_C [\nabla N(P, Q) \cdot \nabla N(P, Q')] \delta n d\sigma. \quad (6.51)$$

Аналогичные формулы можно вывести и для функций Грина более общих дифференциальных уравнений.

Заметим, что градиент функции Грина имеет на границе C направление внешней нормали, так как C есть поверхность уровня функции G . Поэтому уравнение (6.50) можно записать в более простом виде

$$\delta G(Q, Q') = \int_C \frac{\partial G(P, Q)}{\partial n} \frac{\partial G(P, Q')}{\partial \tau} \delta n \, d\sigma. \quad (6.52)$$

Это *формула Адамара* для вариации функции Грина первого рода.

Вариационные формулы мы вывели здесь в предположении, что вариация δn положительна на C . Это было сделано лишь для простоты, результат остается справедливым для любой вариации δn границы C . В точности та же формула (6.52) справедлива также для функции Грина конечной области и для функции Грина любого дифференциального уравнения (6.20).

Рассмотрим для примера функцию Грина $G(P, Q)$ уравнения Лапласа для внутренности B замкнутой поверхности C . Ее можно рассматривать как электростатический потенциал внутри заземленного проводника C при наличии единичного заряда в точке Q . Положим

$$G(P, Q) = \frac{1}{4\pi r(P, Q)} - g(P, Q), \quad (6.53)$$

где $-g(P, Q)$ есть непрерывное решение, которое представляет собой потенциал, вызванный индуцированными зарядами. Так как для решений уравнения Лапласа справедлив принцип минимума и так как на границе C решение g принимает неотрицательные значения $1/4\pi r$, то мы можем утверждать, что решение g неотрицательно в B .

Кроме того, опять, в силу принципа минимума, функция G неотрицательна в B и равна нулю на C . Следовательно, в направлении внешней нормали она убывает и мы получаем

$$\frac{\partial G}{\partial n} < 0 \quad \text{на } C.$$

Из этого неравенства и из выражения (6.52) видно, что $\delta G > 0$, когда $\delta n > 0$, т. е. что функция Грина возрастает вместе с ростом области ее определения B .

Рассмотрим теперь положительную величину $g(Q, Q)$; она равна индуцированному потенциалу в той точке, где находится основной заряд, и является функционалом от B . Рассмотрим все области B объема V , в которые входит точка Q , и выясним, каково наименьшее значение $g(Q, Q)$. Довольно тонкое рассмотрение показывает, что такое минимизирующее тело существует. Для того чтобы охарактеризовать это тело по его экстремальному свойству, мы воспользуемся вариационной формулой (6.52).

Заметим, что вариация δn изменяет объем тела на величину

$$\delta V = \int_C \delta n \, d\sigma,$$

а из формул (6.53) и (6.52) немедленно следует, что

$$\delta g(Q, Q) = - \int_C \left[\frac{\partial G(P, Q)}{\partial n} \right]^2 \delta n \, d\sigma.$$

Свойство минимальности тела B означает, что $\delta g \geq 0$ при всех допустимых вариациях, т. е. при $\delta V = 0$. Отсюда, как показывают обычные в вариационном исчислении рассуждения (см. гл. 4), вытекает, что

$$\frac{\partial G(P, Q)}{\partial n} = \text{const} \quad \text{на } C. \quad (6.54)$$

Следующая задача состоит в том, чтобы найти поверхность C , на которой бы выполнялось условие (6.54). Другими словами, нам надо найти такую поверхность C , окружающую точку Q , на которой плотность индуцированных зарядов была бы постоянна.

Рассмотрим для этого наименьшую сферу Σ_R , описанную вокруг точки Q и заключающую в себе тело B . Пусть радиус Σ_R равен R . Сфера Σ_R касается границы C по крайней мере в одной точке P_R . Функция Грина для сферы Σ_R с источником в точке Q равна, очевидно,

$$G_R(P, Q) = \frac{1}{4\pi r(P, Q)} - \frac{1}{4\pi R}. \quad (6.55)$$

Так как тело B заключено в Σ_R , то в силу монотонной зависимости функции Грина от области

$$G_R(P, Q) \geq G(P, Q).$$

Но в точке P_R , представляющей собой общую точку сферы Σ_R и поверхности C , обе функции Грина равны нулю. Поэтому функция G_R должна убывать быстрее, чем G , когда мы приближаемся к точке P_R изнутри, и, следовательно,

$$\frac{\partial G_R(P, Q)}{\partial n} \leq \frac{\partial G(P, Q)}{\partial n} \quad \text{в } P_R.$$

Далее, в точке P_R нормаль к сфере Σ_R имеет радиальное направление и, согласно равенству (6.55),

$$\frac{\partial G_R}{\partial n} = - \frac{1}{4\pi R^2} \quad \text{в } P_R.$$

Итак, окончательно имеем

$$- \frac{1}{4\pi R^2} \leq \frac{\partial G(P, Q)}{\partial n} \quad \text{в } P_R.$$

Пусть теперь $\Sigma_{R'}$ есть наибольшая сфера, описанная вокруг Q и целиком лежащая в теле B , и пусть ее радиус равен R' . Сфера

$\Sigma_{R'}$ касается поверхности S по крайней мере в одной точке $P_{R'}$, и функция Грина этой сферы равна

$$G_{R'}(P, Q) = \frac{1}{4\pi r(P, Q)} - \frac{1}{4\pi R'}.$$

Далее мы можем провести те же рассуждения, что и раньше, меняя при этом во всех неравенствах знаки, так как сфера $\Sigma_{R'}$ лежит внутри B . При этом получим, что

$$\frac{\partial G(P, Q)}{\partial n} \leq -\frac{1}{4\pi R'^2} \quad \text{в } P_{R'}.$$

До сих пор мы не использовали еще свойства (6.54), характеризующего тело B . Из этого свойства вытекает неравенство

$$\frac{1}{R'^2} \leq \frac{1}{R^2}, \quad R \leq R'.$$

Очевидно, однако, что $\Sigma_{R'} \subset B \subset \Sigma_R$ и что, следовательно, $R' \leq R$. Отсюда мы получаем окончательно, что $R' = R$ и $\Sigma_{R'} = \Sigma_R = B$. Рассматриваемое тело должно быть сферой.

Таким образом, мы решили нашу задачу на экстремум и нашли, что

$$g(Q, Q) \geq (48\pi^2 V)^{-1/3},$$

где равенство имеет место только в случае сферы. Попутно мы доказали, что единственной поверхностью, для которой функция Грина имеет постоянную производную по нормали, является сфера. Мы привели эту задачу лишь для иллюстрации тех возможностей, которые открывает перед нами вариационная формула (6.52). Аналогичная задача для внешней функции Грина поверхности S встречается в теории равновесия гравитирующих жидких масс; эту задачу подобным же образом рассматривал Пуанкаре.

Осторожные математик-и могут подвергнуть известной критике вариационное рассмотрение экстремальных задач, потому что не было показано, что поверхность S , которая дает экстремум, имеет всюду нормаль, и потому неизвестно, можно ли применять к ней вариационную формулу. Однако современные исследования в вариационном исчислении подтвердили законность большого числа формальных операций, связанных с экстремальными задачами. Во всяком случае, вариационные формулы типа Адамара дают нам мощный эвристический аппарат для угадывания правильных решений экстремальных задач.

Так как мы уже вывели вариационные формулы для функций Грина, то можем теперь легко вычислить, как при разложении этих функций в ряды различные коэффициенты изменятся с изменением области. При этом формальные выводы § 6.5 позволят нам получить вариационные формулы для многих важных величин в газовой

динамике и в электростатике. Для краткости мы сосредоточим наше внимание только на одном таком примере.

Из вариационной формулы (6.51) для функции Неймана выводится следующее выражение для коэффициентов $e_{\nu\mu}$:

$$\delta e_{\nu\mu} = - \int_C (\nabla \varphi_\nu \cdot \nabla \varphi_\mu) \delta n \, d\sigma.$$

При помощи формулы (6.42) можно вычислить, чему равно изменение виртуальной массы в направлении α при вариации границы C тела B :

$$\delta W = - \int_C \left[\nabla \left(\sum_{\nu=1}^3 \alpha_\nu \varphi_\nu \right) \right]^2 \delta n \, d\sigma - \delta V. \quad (6.56)$$

Заметим, что в формуле (6.56) интеграл положителен, если δn отрицательно, т. е. если тело B возрастает (δn есть внешняя нормаль к области D , занятой потоком, и, следовательно, — внутренняя нормаль к границе области B). В то же время $\delta V > 0$, так что δW не имеет определенного знака. В самом деле, виртуальная масса W не обязательно возрастает вместе с телом B . Из равенства (6.56) можно, однако, заключить, что сумма

$$W + V = \sum_{\mu, \nu=1}^3 \alpha_\mu \alpha_\nu e_{\mu\nu}$$

монотонно растет вместе с телом B . Этот вывод полезен тогда, когда нужно бывает вывести грубые, но простые оценки для виртуальной массы тела B .

Из вариационных формул можно получить много аналогичных заключений о монотонном поведении физических величин при возрастании тела B .

Рассмотрим теперь следующую несколько искусственную экстремальную задачу: пусть нам дано тело B и требуется найти большее тело B_l заданного объема V_l , которое содержит в себе тело B и имеет наименьшую возможную виртуальную массу в направлении оси x . Чтобы задача имела решение, заданный объем V_l , очевидно, должен быть больше объема V тела B . Если теперь B_l — такое минимальное тело, C_l — его граничная поверхность, а $\varphi_{l\nu}(P)$ — соответствующие ему потенциалы скорости, то B_l характеризуется следующим свойством: вариация виртуальной массы в направлении оси x

$$\delta W_l = - \int_{C_l} (\nabla \varphi_{l1})^2 \delta n \, d\sigma - \delta V_l$$

должна быть положительной для всех вариаций δn границы C_l , которые сохраняют объем V_l и которые оставляют данное тело B внутри B_l . Ясно, что

$$\delta V_l = - \int_{C_l} \delta n \, d\sigma$$

и что сохранение объема достигается выполнением условия $\delta V_l = 0$. Чтобы гарантировать то, что тело B находится внутри тела B_l , нам надо разбить поверхность C_l на две части: (а) часть C_l , которая состоит из граничных точек тела B , (б) часть C_l , которая не пересекается с заданным телом B . Эту последнюю часть мы назовем *свободной границей* тела B_l . Свободную границу можно изменить произвольно, лишь бы размеры смещений были достаточно малы, чтобы при деформации границы C_l нам не дойти ни до одной точки тела B . Поэтому, если мы будем производить вариацию δn только свободной границы, то у нас останется условие $\delta V_l = 0$ и условие минимальности $\delta W_l \geq 0$. Это, очевидно, означает, что должна существовать такая постоянная λ , что

$$(\nabla \varphi_{li})^2 = \lambda \text{ на свободной границе.}$$

Наша экстремальная задача привела нас, таким образом, к следующему: имеется тело B_l , которое обтекается потоком с потенциалом скорости φ_{li} . В бесконечности скорость потока равна единице и направлена в сторону отрицательных x . Скорость потока постоянна на свободной границе тела B_l , и потому, согласно закону Бернулли, на свободной границе давление должно быть постоянным. Но это описание показывает, что поток, который описывается потенциалом φ_{li} , есть поток, который обтекает твердое тело B в направлении отрицательных x и оставляет вокруг этого тела след или мертвую зону $B_l - B$, т. е. область, заполненную жидкостью или газом, которые в потоке не участвуют. Описание мертвой зоны при помощи задачи на экстремум подобного вида было дано Рябушинским; оно послужило основанием для доказательств математических теорем существования этого важного явления. Его можно использовать для численного определения таких свободных границ.

Отметим еще задачу аналогичного рода. Пусть нам дано тело B и вне его кривая Γ , концы которой лежат на B . Нам надо провести поверхность Σ , стягивающую петлю, которая образована телом B и кривой Γ , таким образом, чтобы у нас получилось тело B_l без дырок, имеющее наименьшую возможную виртуальную массу в направлении x . Такими же вариационными методами эта задача сводится к задаче о потенциале скорости φ_{li} для минимального тела B_l , у которого величина $(\nabla \varphi_{li})^2$ принимает одинаковые значения в соответствующих точках на нижней и верхней сторонах поверхности Σ . Другими словами, давления на обеих сторонах поверхности Σ равны

друг другу, но векторы скоростей $\nabla \varphi_{1i}$ будут по обе стороны Σ , вообще говоря, различны. Таким образом, мы получаем поток, обтекающий тело B , у которого Σ есть поверхность разрыва поля скоростей. Эта поверхность разрыва находится, однако, в статическом равновесии, потому что давления по обе стороны уравновешивают друг друга. Такие поверхности разрыва встречаются в теории тонкого крыла, там они называются вихревыми листами. Мы получили их описание при помощи свойства экстремальности, пользуясь для этого вариационной формулой (6.56).

§ 6.7. Изменение функций Грина при изменении коэффициентов дифференциального уравнения. В предыдущем параграфе мы исследовали, как функции Грина зависят от области их определения. Покажем теперь, что можно построить аналогичную теорию для изучения зависимости функций Грина от коэффициентов уравнения. Центральную роль здесь опять будут играть формулы Грина.

Для простоты мы рассмотрим дифференциальное уравнение

$$\nabla^2 u = (p + tq)u, \quad p(x, y, z) > 0 \quad \text{в } B, \quad (6.57)$$

которое зависит от параметра t , а p и q суть непрерывные функции в рассматриваемой области B . Нам нужно выяснить, как функции Грина зависят от параметра t .

Обозначим через $G_1(P, Q)$ и $G_2(P, Q)$ функции Грина первого рода для уравнения (6.57) и для значений параметров t_1 и t_2 . Применим вторую формулу Грина (6.14) к функциям $U = G_1(P, Q)$, $V = G_2(P, Q')$ и к области B , из которой выброшены две бесконечно малые сферы, описанные вокруг точек Q и Q' . Так как обе функции Грина равны нулю на границе S тела B , то мы легко покажем, что

$$G_1(Q, Q') - G_2(Q, Q') = (t_2 - t_1) \int_B q(P) G_1(P, Q) G_2(P, Q') dx dy dz. \quad (6.58)$$

В точности такую же формулу можно получить и для других функций Грина дифференциального уравнения (6.57) в области B .

Разделив уравнение (6.58) на $t_1 - t_2$ и перейдя к пределу при $t_2 \rightarrow t_1$, мы получим

$$\frac{\partial G_t(Q, Q')}{\partial t} = - \int_B q(P) G_t(P, Q'), G_t(P, Q) dx dy dz, \quad (6.59)$$

где через G_t обозначена функция Грина уравнения (6.57), отвечающая значению параметра t . Предположим, в частности, что функция $q(P)$ положительна в B и что $t > 0$. Мы знаем, что в этом случае функции Грина тоже положительны, поэтому из уравнения (6.59) вытекает следующее:

Функции Грина убывают с возрастанием коэффициента уравнения (6.57).

На границе C области B функции Грина $G_t(P, Q)$ тождественно по t равны нулю. Внутри области они убывают при возрастании t , то же самое на границе C происходит и с производной по нормали, — $\partial G_t(P, Q)/\partial n$. Воспользовавшись теперь представлением (6.19), мы можем заключить, что:

Если $u(P)$ есть решение уравнения (6.57) с фиксированными граничными значениями на C , то $u(P)$ убывает с ростом коэффициента дифференциального уравнения.

Таким образом, если мы, например, умеем решать дифференциальное уравнение

$$\nabla^2 u = ku, \quad k = \text{const}, \quad (6.60)$$

то можем получить полезные оценки для решений более общего уравнения (6.57). В случае же уравнения (6.60) в нашем распоряжении находится по крайней мере полная система частных решений — функций Бесселя.

Предыдущая теорема, конечно, тесно связана с формулой сравнения (6.24) и может быть из нее выведена. Рассуждения, которые мы здесь провели, воспользовавшись сначала функцией Грина первого рода, можно повторить с другими функциями Грина и получить тот же результат.

Вернемся теперь к уравнению (6.58), которое связывает любые два решения уравнения (6.57) с различными значениями t . Пусть $G(P, Q)$ есть функция Грина уравнения (6.57) при $t=0$. Уравнение (6.58) можно представить в виде

$$G(Q, Q') = G_t(Q, Q') + t \int_B q(P) G(P, Q) G_t(P, Q') dx dy dz. \quad (6.61)$$

Будем считать функцию $G(P, Q)$ известной. Уравнение (6.61) можно тогда рассматривать как интегральное уравнение типа Фредгольма для функции $G_t(Q, Q')$. Мы имеем при фиксированном Q' уравнение вида

$$\varphi(Q) = f(Q) - t \int_B K(P, Q) \varphi(P) dx dy dz, \quad (6.62)$$

где

$$K(P, Q) = q(P) G(P, Q). \quad (6.63)$$

Функция $f(Q)$ известна, а неизвестна функция $\varphi(Q) = G_t(Q, Q')$.

Интегральное уравнение (6.62) лучше всего решать (если это возможно) методом последовательных приближений. Положим

$$\varphi_0(Q) = f(Q), \quad \varphi_n(Q) = f(Q) - t \int_B K(P, Q) \varphi_{n-1}(Q) dx dy dz.$$

Очевидно, мы можем последовательно вычислить все $\varphi_n(Q)$; таким образом мы приходим к разложению решения

$$\varphi(Q) = \lim_{n \rightarrow \infty} \varphi_n(Q)$$

в степенной ряд по параметру t . Этот ряд называется степенным рядом Неймана — Лиувилля; он сходится для достаточно малых значений t и для этих значений дает решение в наиболее удобном виде.

Для отыскания радиуса сходимости степенного ряда Неймана — Лиувилля общая теория интегральных уравнений предоставляет нам следующий критерий. Рассмотрим однородное интегральное уравнение

$$\rho(Q) + t \int_B K(P, Q) \rho(P) dx dy dz = 0.$$

Это уравнение всегда имеет решение $\rho(P) = 0$, и для достаточно малых значений $|t|$ это решение является единственно возможным. Однако могут существовать собственные значения t_* , для которых имеются не равные тождественно нулю решения $\varphi_*(P)$. Пусть $|t_1|$ равно наименьшей абсолютной величине всех собственных значений t_* . Тогда ряд Неймана — Лиувилля сходится при всех t , для которых $|t| < |t_1|$.

Пользуясь специальным видом (6.63) нашего ядра, можно легко показать, что у интегрального уравнения

$$\rho(Q) + t \int_B q(P) G(P, Q) \rho(P) dx dy dz \quad (6.64)$$

все собственные значения действительны. Опишем функцию $\rho(Q)$, пользуясь уравнением (6.64), которому она удовлетворяет.

Так как $G(P, Q)$ обращается в нуль, когда точка Q принадлежит границе C области B , то

$$\rho(Q) = 0 \text{ на } C.$$

Далее, $G(P, Q)$, как функция точки Q , есть решение уравнения

$$\nabla^2 G(P, Q) - p(Q) G(P, Q) = 0.$$

В точке $Q = P$, однако, эта функция имеет особенность; поэтому из уравнения (6.64) следует, что функция $\rho(Q)$ удовлетворяет обобщенному уравнению Лапласа — Пуассона

$$\nabla^2 \rho(Q) - p(Q) \rho(Q) = tq(Q) \rho(Q).$$

Предположим теперь еще, что

$$|q(Q)| < p(Q) \text{ в } B.$$

Функция $\rho(Q)$ удовлетворяет уравнению

$$\nabla^2 \rho = (p + tq) \rho \quad (6.65)$$

и обращается на границе C в нуль. До тех пор пока параметр t действителен и $|t| < 1$, коэффициент в уравнении (6.65) положителен в области B , и, в силу теоремы единственности первой краевой задачи для таких дифференциальных уравнений, мы получаем, что $\rho \equiv 0$ в B . Следовательно, первое собственное значение t может появиться только при $|t| > 1$, и, соответственно, ряд Неймана — Лиувилля, которым определяется решение неоднородного интегрального уравнения (6.61) или (6.62), сходится при $t = 1$.

Итак, если мы знаем функцию Грина дифференциального уравнения

$$\nabla^2 u = \rho u, \quad \rho > 0 \quad \text{в } B, \quad (6.66)$$

то мы можем получить функции Грина всех дифференциальных уравнений

$$\nabla^2 u = (p + q)u, \quad |q| < p \quad \text{в } B$$

в виде сходящихся степенных рядов. В частности, если мы знаем в некоторой области функции Грина уравнения (6.60), мы можем этим методом вычислить для той же области все функции Грина для уравнений (6.66), когда $\rho < k$.

В § 6.3 мы обсуждали значение фундаментальных решений уравнения (6.66). Теперь же мы вывели метод для получения таких решений, так как функция Грина представляет собой лишь частный случай фундаментального решения.

ЛИТЕРАТУРА

1. Bergman S., Schiffer M., Kernal Functions and Elliptic Differential Equations in Mathematical Physics, Academic Press, New York, 1953.
2. Курант Р. и Гильберт Л., Методы математической физики, т. I, изд. 3, и т. II, изд. 2, Гостехиздат, М. — Л., 1951.
3. Франк Ф. и Мизес Р., Дифференциальные и интегральные уравнения математической физики, Гостехиздат, М. — Л., 1937.
4. Kellogg O. D., Foundations of Potential Theory, Springer-Verlag, Berlin, 1929.
5. Морс Ф. и Фешбах Г., Методы теоретической физики, т. I, II, ИЛ, М., 1958.
- 6*. Олейник О. А., О свойствах решений краевых задач для уравнений эллиптического типа, Мат. сб., 30 (72); 3, 695—702 (1952).

Глава 7

КРАЕВЫЕ ЗАДАЧИ ТЕОРИИ УПРУГОСТИ

И. С. СОКОЛЬНИКОВ

Калифорнийский университет, Лос-Анжелес

Постановка задач

§ 7.1. Введение. В настоящей главе мы рассмотрим некоторый общий подход к одному широкому классу плоских краевых задач, часто возникающих при исследовании напряжений в упругих конструкциях. Излагаемый ниже метод имеет вполне современное происхождение, и, хотя он широко использовался последние 15 лет в работах советских авторов, он не нашел еще широкого применения на Западе.

Мы выбрали для иллюстрации применения метода теорию упругости, однако из математических формулировок рассматриваемых задач теории упругости будет ясно, что применение излагаемого метода целесообразно не только в этих задачах. Этот метод можно использовать и в других областях, он уже применялся для решения многих важных задач теории теплопроводности, электростатики и аэродинамики.

Поскольку не все читатели этой книги являются специалистами по теории упругости, мы сперва попытаемся описать интересующий нас класс задач теории упругости и затем введем аппарат, необходимый для математической формулировки этих задач. В заключение мы укажем некоторый метод решения этих задач, использующий лишь элементарные сведения из теории функций комплексного переменного. Имеются также и другие, более мощные методы, опирающиеся на решение некоторых функциональных и интегральных уравнений, которые естественно возникают из математической формулировки рассматриваемых задач, однако, эти методы требуют применения несколько более тонких методов анализа, нежели аппарат элементарной теории функций комплексного переменного.

§ 7.2. Два основных типа задач. Во всех краевых задачах теории упругости определяются напряжения и деформации упругого тела, подверженного действию внешних сил. Мы ограничимся в наших рассуждениях случаями, когда деформации тел не слишком велики, так что напряжения и деформации можно считать связанными некоторым линейным соотношением, известным под названием обобщенного закона Гука. Принятая нами гипотеза представляет собой обыч-

ную гипотезу классической линейной теории упругости, охватывающей большинство задач, привлекающих в настоящее время внимание специалистов по прочности. Имеются два основных типа этих задач. В задачах одного типа требуется определить поля напряжений и деформаций внутри тела, если заданы деформации на поверхности этого тела. Этот тип задач носит название *первой основной краевой задачи* теории упругости. *Вторая основная задача* представляет собой определение напряжений и смещений внутри некоторой конструкции, к поверхности которой приложено заданное распределение внешних сил¹⁾. На практике часто встречается комбинация этих двух основных типов задач, когда на некоторой части поверхности деформируемого тела заданы смещения, а на остальной части поверхности тела заданы внешние силы. В настоящей главе мы будем рассматривать задачи указанных здесь широких основных классов.

§ 7.3. Характеристики смещений; деформация.

Характеристики смещений упругого тела вводятся весьма просто. Рассмотрим некоторое тело τ (рис. 7.1) в его недеформированном состоянии и обозначим координаты произвольной точки этого тела P через (x_1, x_2, x_3) . После деформации тела материальная точка P смещается в некоторое новое положение P' с координатами (x'_1, x'_2, x'_3) , так что смещение точки P определяется вектором $PP' = \mathbf{u}$ с компонентами (u_1, u_2, u_3) . Мы будем предполагать, что функции u_i имеют непрерывные частные производные первого и второго порядков.

Вместо того чтобы рассматривать вектор смещения $\mathbf{u}(x_1, x_2, x_3)$, оказывается более удобным оперировать с системой шести скалярных функций e_{ij} , определяемых формулами

$$e_{ij} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right), \quad i, j = 1, 2, 3, \quad (7.1)$$

так что

$$e_{11} = \frac{\partial u_1}{\partial x_1}, \quad e_{12} = \frac{1}{2} \left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right) \text{ и т. д.}$$

Величины e_{ij} называются *компонентами тензора деформаций*; они имеют простое геометрическое истолкование как деформации растяжения (сжатия) и сдвига. Подобные истолкования, как бы интересны они ни были, мы здесь рассматривать не будем. Важно только отметить,

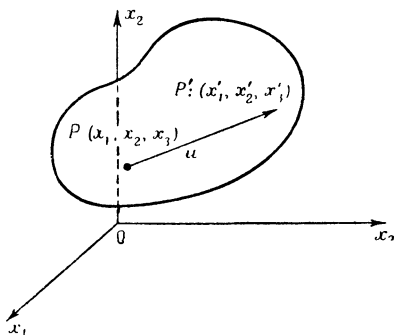


Рис. 7.1. Смещения в упругом теле.

1) Некоторые авторы называют „вторую“ задачу „первой“ и наоборот.

что если смещения u_i известны, то компоненты деформации можно вычислить по формулам (7.1). Наоборот, если в каждой точке тела известны компоненты деформации e_{ij} , то компоненты смещения легко получаются интегрированием, коль скоро e_{ij} представляют собой дифференцируемые функции с соответствующими свойствами. Так как система уравнений (7.1) представляет собой систему из шести уравнений с тремя неизвестными u_i , то функции e_{ij} нельзя выбирать произвольно — напротив, они должны удовлетворять условиям интегрируемости

$$\frac{\partial^2 e_{ij}}{\partial x_i \partial x_l} + \frac{\partial^2 e_{kl}}{\partial x_i \partial x_j} - \frac{\partial^2 e_{ik}}{\partial x_j \partial x_l} - \frac{\partial^2 e_{jl}}{\partial x_i \partial x_k} = 0, \quad i, j, k, l = 1, 2, 3, \quad (7.2)$$

известным под названием *условий совместности Сен-Венана*. Изложенные факты были получены еще во времена Коши и рассматриваются во всех книгах по теории упругости. Мы будем считать установленным, что деформированное состояние тела полностью определено, если для всех точек тела задана система шести функций e_{ij} , непрерывных, с непрерывными производными и удовлетворяющих

условиям (7.2). Основным преимуществом выбора величин e_{ij} в качестве характеристик деформации является то, что они более просто связаны с напряжениями (см. § 7.4), нежели компоненты вектора смещения u .

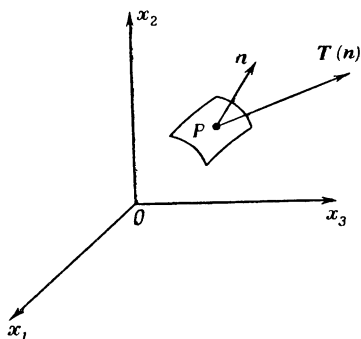


Рис. 7.2. Напряжения в упругом теле.

§ 7.4. Характеристики напряженного состояния. Характеристики напряженного состояния получаются, с другой стороны, несколько более сложными. Напряженное состояние в некоторой точке P непрерывной среды можно определить, задавая силу T , действующую на единицу площади каждого плоского элемента, проходящего через P

(рис. 7.2). Вектор напряжения T , вообще говоря, зависит от направления плоского элемента, проходящего через точку P , т. е. он зависит от направления единичной нормали n к этому плоскому элементу. Напряжение T в точке P не зависит от направления элемента поверхности только в простейшем случае, когда оно является гидростатическим давлением. Зависимость напряжения T от вектора n усложняет описание напряженного состояния, поскольку оказывается необходимым определить компоненты T_1 , T_2 и T_3 вектора T для всех возможных направлений плоских элементов, проходящих через заданную точку. Однако, как показал Коши, вместо задания векторной функции $T(n)$ для всех мыслимых направлений n достаточно задать систему девяти скалярных функций $\tau_{ij}(x_1, x_2, x_3)$, при помощи кото-

рых компоненты T_i вектора $\mathbf{T}(\mathbf{n})$ можно вычислить, пользуясь простой формулой

$$T_i = \tau_{i1}n_1 + \tau_{i2}n_2 + \tau_{i3}n_3, \quad i = 1, 2, 3, \quad (7.3)$$

где n_1, n_2, n_3 представляют собой компоненты вектора единичной нормали $\mathbf{n}$. Мы будем просить читателя принять на веру этот хорошо известный факт, доказываемый во всех учебниках теории упругости. Величины τ_{ij} образуют матрицу

$$\begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix} \quad (7.4)$$

и называются компонентами *тензора напряжений*.

Подобно компонентам деформации, компоненты напряжения τ_{ij} допускают простое физическое истолкование. Компоненты τ_{11}, τ_{22} и τ_{33} представляют собой *нормальные напряжения*, действующие на элементы, перпендикулярные к соответствующим координатным осям, а компоненты τ_{ij} при $i \neq j$ представляют собой *напряжения сдвига*. Важное соотношение (7.3) непосредственно получается из простых соображений равновесия.

§ 7.5. Условия равновесия. Если рассматриваемое тело находится в состоянии равновесия под действием внешних *поверхностных* сил, то легко показать, что матрица напряжений (7.4) симметрична и во всех внутренних точках тела функции τ_{ij} удовлетворяют системе трех линейных дифференциальных уравнений

$$\frac{\partial \tau_{i1}}{\partial x_1} + \frac{\partial \tau_{i2}}{\partial x_2} + \frac{\partial \tau_{i3}}{\partial x_3} = 0, \quad i = 1, 2, 3. \quad (7.5)$$

Описание деформации при помощи тензора деформаций e_{ij} и напряженного состояния при помощи тензора напряжений τ_{ij} возможно для любых непрерывных тел, независимо от их физических свойств. Если, однако, рассматриваемое тело является упругим и изотропным и если ограничиться рассмотрением малых деформаций, то компоненты e_{ij} тензора деформаций являются линейными однородными функциями компонент τ_{ij} тензора напряжений, так что мы можем написать следующую систему шести линейных соотношений:

$$e_{ij} = \frac{1+\sigma}{E} \tau_{ij} - \frac{\sigma}{E} \delta_{ij} \Theta, \quad i, j = 1, 2, 3, \quad (7.6)$$

где

$$\Theta = \tau_{11} + \tau_{22} + \tau_{33}, \quad \delta_{ij} = \begin{cases} 0 & \text{при } i \neq j, \\ 1 & \text{при } i = j. \end{cases}$$

Постоянная E называется *модулем Юнга*, а постоянная σ — *коэффициентом Пуассона* рассматриваемого упругого материала.

Формулы (7.6) представляют собой простое обобщение закона Гука; они устанавливают, что рассматриваемая среда является упругой и изотропной.

Подставив выражения (7.6) в условия совместности Сен-Венана (7.2), мы получим условия, которые следует наложить на τ_{ij} , чтобы поле τ_{ij} соответствовало некоторому напряженному состоянию изотропного упругого тела. Эти условия имеют вид

$$\nabla^2 \tau_{ij} + \frac{1}{1+\sigma} \frac{\partial^2 \theta}{\partial x_i \partial x_j} = 0, \quad \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2}. \quad (7.7)$$

Таким образом, напряженное состояние изотропного упругого тела, находящегося в состоянии равновесия под действием поверхностных сил, определяется шестью функциями τ_{ij} , удовлетворяющими уравнениям равновесия (7.5) и условиям совместности (7.7). Определив τ_{ij} , можно вычислить деформации e_{ij} из системы линейных соотношений (7.6), после чего можно определить смещения u_i , интегрируя уравнения (7.1). Разумеется, напряжения и смещения должны удовлетворять заданному распределению сил и смещений на поверхности тела.

§ 7.6. Плоские и пространственные задачи. Задача, сформулированная с той степенью общности, которая имеется в предыдущих рассуждениях, представляет собой проблему исключительной математической трудности, так что большинство так называемых общих решений основных краевых задач теории упругости, по крайней мере в пространственном случае, могут служить лишь для доказательства существования соответствующих решений. Инженер редко удовлетворяется заявлением математика, утверждающего, что он решил задачу, поскольку сделанное им предположение о несуществовании решения привело его к логическому противоречию. Инженер отказывается называть „решением“ какое-нибудь математическое построение, если оно оказывается неспособным привести к явному численному результату. В настоящее время для первой и второй пространственных краевых задач теории упругости не существует общих методов, приводящих к явному решению¹⁾.

Однако за последние 15 лет были развиты весьма эффективные общие методы решения плоских задач теории упругости. Для определения того, что мы понимаем под термином „плоские задачи“, рассмотрим физические явления, в которых эти задачи возникают.

¹⁾ Рассмотрено некоторое число частных пространственных задач, при решении которых используется либо симметрия геометрической конфигурации упругого тела, либо частный вид граничных условий. Однако полезные частные решения пространственных задач можно свободно пересчитать по пальцам.

§ 7.7. Цилиндр; приложенные силы постоянны по его длине.

Предположим, что некоторая деталь имеет форму длинного цилиндра произвольного поперечного сечения R (рис. 7.3) и что боковая поверхность цилиндра подвергается воздействию внешних сил, которые не меняются по длине цилиндра. Очевидно, что в этом случае все поперечные сечения цилиндра, нормальные к его оси, будут деформироваться одинаковым образом.

Если мы выберем ось цилиндра совпадающей с осью x_3 системы координат, то компоненты вектора смещения u не будут изменяться вдоль оси x_3 и будут зависеть, таким образом, только от координат x_1 и x_2 . Из уравнений (7.1) в этом случае следует, что компоненты деформации e_{ij} не зависят от координаты x_3 , а из соотношений между напряжениями и деформациями (7.6) в свою очередь следует, что $\tau_{13} = \tau_{23} = 0$ и что остальные функции τ_{ij} зависят только от координат x_1 и x_2 . Таким образом, первые два уравнения равновесия (7.5) приводятся к системе

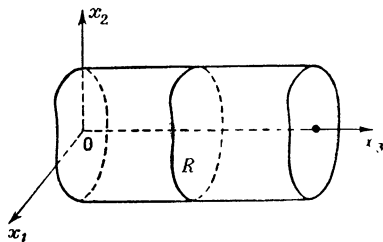


Рис. 7.3. Цилиндр с поперечным сечением R .

$$\left. \begin{aligned} \frac{\partial \tau_{11}}{\partial x_1} + \frac{\partial \tau_{12}}{\partial x_2} &= 0, \\ \frac{\partial \tau_{21}}{\partial x_1} + \frac{\partial \tau_{22}}{\partial x_2} &= 0. \end{aligned} \right\} \quad (7.8)$$

Простое вычисление, использующее уравнения (7.6), дает

$$\tau_{33} = \sigma(\tau_{11} + \tau_{22}). \quad (7.9)$$

Условия совместности (7.7) в рассматриваемом случае приводят к уравнению

$$\nabla^2(\tau_{11} + \tau_{22}) = 0, \quad \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}. \quad (7.10)$$

Таким образом, для решения задачи о деформации цилиндра силами, распределенными вдоль его боковой поверхности так, что они постоянны по длине цилиндра, следует определить из уравнений (7.8), (7.6) и (7.1) пять функций: $u_1(x_1, x_2)$, $u_2(x_1, x_2)$, $\tau_{11}(x_1, x_2)$, $\tau_{12}(x_1, x_2)$, $\tau_{22}(x_1, x_2)$. Функции $\tau_{\alpha 3}(x_1, x_2)$ должны, кроме того, как это следует из уравнений (7.3), удовлетворять следующим условиям на боковой поверхности:

$$\left. \begin{aligned} \tau_{11}n_1 + \tau_{12}n_2 &= T_1, \\ \tau_{21}n_1 + \tau_{22}n_2 &= T_2, \end{aligned} \right\} \quad (7.11)$$

где T_1 и T_2 представляют собой заданные компоненты вектора напряжений T на поверхности цилиндра.

Можно выписать систему граничных условий на поверхности цилиндра, соответствующих также случаю задания на этой границе смещений, а не напряжений. Однако оказывается, что получающуюся при этом краевую задачу можно рассматривать в точности таким же способом, так что по этим соображениям мы ограничимся рассмотрением второй основной краевой задачи.

Из уравнений равновесия (7.8) вытекает существование функции напряжений $U(x_1, x_2)$, такой, что

$$\tau_{11} = \frac{\partial^2 U}{\partial x_2^2}, \quad \tau_{12} = -\frac{\partial^2 U}{\partial x_1 \partial x_2}, \quad \tau_{22} = \frac{\partial^2 U}{\partial x_1^2}. \quad (7.12)$$

Легко проверить, что и обратно, если выбрать τ_{ij} таким образом, то уравнения (7.8) будут удовлетворяться тождественно, а из условия совместности (7.10) вытекает, что функция $U(x_1, x_2)$ удовлетворяет бигармоническому уравнению

$$\frac{\partial^4 U}{\partial x_1^4} + 2 \frac{\partial^4 U}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4 U}{\partial x_2^4} = 0.$$

Это уравнение можно записать в виде

$$\nabla^2 \nabla^2 U \equiv \nabla^4 U = 0,$$

где

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}.$$

Подставляя в формулы (7.11) соотношения (7.12) и замечая, что $n_1 = -dx_2/ds$, $n_2 = dx_1/ds$, мы можем записать граничные условия (7.11) в виде

$$\frac{\partial U}{\partial x_1} = \int_0^s T_2(s) ds \equiv f_1(s),$$

$$\frac{\partial U}{\partial x_2} = -\int_0^s T_1(s) ds \equiv f_2(s),$$

где $f_1(s)$ и $f_2(s)$ представляют собой некоторые известные функции длины дуги s . Функции $f_\alpha(s)$ полностью определяются заданными напряжениями $T_\alpha(s)$ на границе C .

Таким образом, решение задачи о деформациях цилиндра силами, распределенными по его боковой поверхности так, что они остаются постоянными вдоль образующих, приводится к определению бигармонической функции $U(x_1, x_2)$, производные которой принимают на границе C поперечного сечения R заданные значения $f_\alpha(s)$. Эта краевая задача, известная под названием *основной бигармонической краевой задачи*, была предметом многочисленных теоретических исследований. Частный случай ее впервые возник в связи с исследованием попе-

речных прогибов закрепленных упругих пластин; решение этой последней задачи было признано достойным специальной премии, учрежденной Французской Академией наук. Премия была разделена между несколькими первоклассными математиками, доказавшими существование решения задачи (для односвязной области) при сравнительно слабых ограничениях на функции, входящие в граничные условия.

За последние 15 лет доказательства существования были распространены на случай областей весьма общего характера. Кроме того, были развиты некоторые эффективные методы решения рассматриваемых задач.

§ 7.8. Цилиндр; приложенные силы распределены на торцах.

Из уравнения (7.9) следует, что торцы цилиндра, вообще говоря, не будут свободными от напряжений. Если же в исходной задаче торцы цилиндра являются свободными, то следует наложить на решение задачи о плоской деформации решение задачи о деформациях цилиндра со свободной от нагрузки боковой поверхностью, нагруженного силами, распределенными по торцам так, чтобы обратить в нуль распределение напряжений, определяемое формулой (7.9).

В настоящее время задача о деформациях цилиндра со свободной боковой поверхностью силами, приложенными на его торцах, является в некотором смысле одной из простейших задач теории упругости. Ее можно разбить на следующие четыре основные задачи:

- а) растяжение продольными силами;
- б) изгиб парами, приложенными на торцах цилиндра;
- в) кручение парами, момент которых направлен вдоль оси цилиндра;
- г) изгиб торцовыми нагрузками.

Все эти основные задачи были изучены за последние 90 лет. Известно, что технически решение задач (а) и (б) вполне элементарно, а решение задач кручения и изгиба упирается в решение одной из основных краевых задач теории потенциала, известной под названием задачи Дирихле.

Задача Дирихле представляет собой задачу об определении некоторой функции $V(x_1, x_2)$, удовлетворяющей во всей области R поперечного сечения цилиндра уравнению Лапласа

$$\nabla^2 V = 0,$$

а на границе C области R обращающейся в заданную непрерывную функцию¹⁾

$$V = f(s),$$

где s — длина дуги, отсчитываемая вдоль контура C .

¹⁾ Функция f в случае задач кручения представляется выражением вида $\frac{1}{2}(x_1^2 + x_2^2)$; она является несколько более сложной для задачи изгиба (см., например, гл. IV нашей книги [3], приведенной в списке литературы к этой главе).

Коль скоро функция $V(x_1, x_2)$ найдена, нетрудно уже определить напряжения и смещения в каждой точке цилиндра, деформированного нагрузками на торцах.

§ 7.9. Цилиндр; комбинированная задача. Таким образом, комбинируя решения соответствующих задач Дирихле с решением задачи плоской деформации, мы можем решить задачу о деформации цилиндра, торцы которого несут произвольную нагрузку, а нагрузка на боковую поверхность не меняется по длине цилиндра. Эта весьма общая задача охватывает большое число частных задач, возникающих при расчете конструкций, содержащих длинные распорки и балки.

Задача Дирихле занимает поистине исключительное место в математике. Без всякого сомнения, она изучалась более чем всякая другая краевая задача математической физики. К задаче Дирихле непосредственно приводится основная задача гидродинамики, заключающаяся в определении установившегося течения вокруг некоторого препятствия; она возникает также в теории теплопроводности и электростатике и, как мы только что выяснили, играет важную роль в исследовании деформации цилиндров. Бигармоническая краевая задача, как мы видели, также возникает при исследовании деформации тонких пластин и при исследовании деформации цилиндра под действием сил, распределенных по его боковой поверхности. Поскольку большинство инженерных сооружений состоит из деталей, являющихся либо длинными распорками, либо тонкими пластинами, трудно преувеличить практическую значимость эффективных методов решения подобных краевых задач. Обратимся теперь к рассмотрению одного из таких методов.

Метод решения

§ 7.10. Вычислительные соображения. Мы уже указывали в предыдущем параграфе, что некоторый широкий класс плоских задач теории упругости можно привести к решению задачи Дирихле

$$\left. \begin{aligned} \nabla^2 V &= 0 \text{ внутри } R, \\ V &= f(s) \text{ на границе } C \end{aligned} \right\} \quad (7.13)$$

и к решению бигармонической краевой задачи

$$\left. \begin{aligned} \nabla^4 U &= 0 \text{ внутри } R, \\ \frac{\partial U}{\partial x_1} &= f_1(s) \text{ на границе } C, \\ \frac{\partial U}{\partial x_2} &= f_2(s) \text{ на границе } C. \end{aligned} \right\} \quad (7.14)$$

Имеется огромное количество математической литературы, посвященной этим задачам; их исследование привело к открытиям первостепенного значения в теории дифференциальных и интегральных

уравнений, в вариационном исчислении, а также в некоторых других отделах анализа. Получившееся в результате развитие упомянутых разделов анализа далеко вышло за пределы ограниченной исходной задачи получения полезных явных решений; оно представляет собой прекрасный пример стимулирования физикой и техникой развития чистой математики. Однако потребности техники в явных решениях постоянно направляли усилия математиков на решение исходной задачи, так что не будет преувеличением сказать, что уравнение Лапласа и волновое уравнение послужили главным импульсом для наблюдающегося в последнее время прогресса методов численного анализа.

Граничные задачи (7.13) и (7.14) можно численно решать, хотя и с большими затратами труда, рассматривая, например, соответствующие разностные уравнения. Автор убежден, однако, что наиболее подходящим способом решения этих задач является применение аналитических методов, быть может в комбинации, на одном или двух промежуточных этапах, с численными расчетами. Для выяснения сущности одного простого аналитического метода решения этих задач мы предположим сначала, что область R является односвязной. Это означает, что пластина или поперечное сечение цилиндра R ограничены простым замкнутым контуром, подобным изображенному на рис. 7.4. Впоследствии мы укажем математический аппарат, необходимый для эффективного решения задач деформации пластин с отверстиями или для изучения составных и усиленных жесткими впамяками цилиндров.

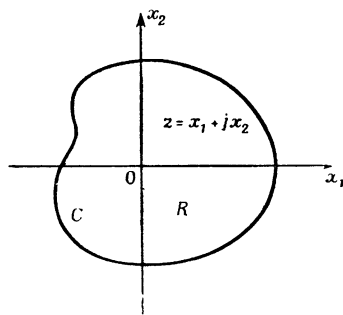


Рис 7.4. Поперечное сечение R цилиндра.

§ 7.11. Формулировка задачи в терминах теории функций комплексного переменного. Оказывается полезным несколько перефразировать задачи (7.13) и (7.14), произведя замену независимых переменных x_1 и x_2 , которая сводит отыскание гармонической функции V и бигармонической функции U к определению некоторых аналитических функций комплексного переменного. Основанием для этого служит то обстоятельство, что свойства аналитических функций комплексного переменного известны значительно лучше, нежели свойства действительных функций. Кроме того, вычисление интегралов в комплексной области осуществляется гораздо проще, нежели в соответствующей действительной области.

Предположим, что область, внутри которой ищется решение задач (7.13) и (7.14), расположена в плоскости комплексного переменного $z = x_1 + jx_2$. Предположим далее, что кривая C , ограничивающая область, представляет собой простой замкнутый контур,

и примем для определенности, что область R является внутренней по отношению к кривой C . Мы могли бы с тем же успехом считать, что область R является внешней по отношению к кривой C ; в этом случае решение будет соответствовать бесконечной пластине с отверстием, ограниченным кривой C , либо бесконечному телу с продольной цилиндрической полостью, ось которой перпендикулярна к плоскости x_1x_2 .

Обозначим сопряженное комплексное переменное через $\bar{z}$, так что

$$\bar{z} = x_1 - jx_2, \quad j^2 = -1;$$

тогда

$$x_1 = \frac{1}{2}(z + \bar{z}), \quad x_2 = \frac{1}{2j}(z - \bar{z}). \quad (7.15)$$

Заменяя в выражениях для $V(x_1, x_2)$ и $U(x_1, x_2)$ переменные x_1 и x_2 по формулам (7.15), мы получаем функции

$$V(x_1, x_2) = V\left(\frac{z + \bar{z}}{2}, \frac{z - \bar{z}}{2j}\right),$$

$$U(x_1, x_2) = U\left(\frac{z + \bar{z}}{2}, \frac{z - \bar{z}}{2j}\right),$$

которые можно рассматривать как функции независимых переменных z и $\bar{z}$.

Из соотношений (7.15) легко следует, что

$$\frac{\partial}{\partial z} = \frac{1}{2} \left(\frac{\partial}{\partial x_1} - j \frac{\partial}{\partial x_2} \right), \quad \frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left(\frac{\partial}{\partial x_1} + j \frac{\partial}{\partial x_2} \right),$$

откуда

$$4 \frac{\partial^2}{\partial z \partial \bar{z}} = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \equiv \nabla^2$$

и

$$16 \frac{\partial^4}{\partial z^2 \partial \bar{z}^2} = \frac{\partial^4}{\partial x_1^4} + 2 \frac{\partial^4}{\partial x_1^2 \partial x_2^2} + \frac{\partial^4}{\partial x_2^4} \equiv \nabla^4.$$

Таким образом, дифференциальные уравнения задач (7.13) и (7.14) принимают соответственно вид

$$\frac{\partial^2 V}{\partial z \partial \bar{z}} = 0 \quad (7.16)$$

и

$$\frac{\partial^4 U}{\partial z^2 \partial \bar{z}^2} = 0. \quad (7.17)$$

Уравнения (7.16) и (7.17) легко интегрируются в самой общей форме. В самом деле, интегрируя (7.16) по $\bar{z}$, получаем

$$\frac{\partial V}{\partial z} = \varphi_1(z),$$

где $\varphi_1(z)$ — произвольная функция $z = x_1 + jx_2$; интегрируя это равенство по z , получаем

$$V = \int^z \varphi_1(z) dz + \varphi_2(\bar{z}), \quad (7.18)$$

где $\varphi_2(\bar{z})$ представляет собой некоторую произвольную функцию $\bar{z}$. Однако V представляет собой действительную дифференцируемую функцию, откуда можно сделать заключение, что правая часть уравнения (7.18) должна представляться в виде

$$V = \frac{1}{2} [\varphi(z) + \overline{\varphi(z)}],$$

где $\varphi(z)$ — некоторая аналитическая функция z , а $\overline{\varphi(z)}$ означает функцию, сопряженную с $\varphi(z)$. Это соответствует известному факту, что каждая гармоническая функция V может быть представлена в виде действительной части некоторой аналитической функции $\varphi(z)$.

Интегрируя уравнение (7.17), мы получим в точности аналогичным способом, что общее решение бигармонического уравнения (7.14) представляется в виде

$$U = \frac{1}{2} [\varphi_1(z) + \overline{\varphi_1(z)} + z\overline{\varphi_2(z)} + \bar{z}\varphi_2(z)],$$

где функции $\varphi_1(z)$ и $\varphi_2(z)$ представляют собой аналитические функции z в области R .

Мы должны теперь выразить граничные условия в задачах (7.13) и (7.14) через функции φ_1 и φ_2 . В задаче Дирихле (7.13) функция V удовлетворяет условию обращения на границе C области R в заданную функцию $f(s)$, откуда имеем

$$\varphi(z) + \overline{\varphi(z)} = 2f(s) \quad \text{на } C. \quad (7.19)$$

Задача свелась, таким образом, к отысканию функции $\varphi(z)$, аналитической в области R и такой, что на границе C действительная часть этой функции обращается в заданную непрерывную действительную функцию $f(s)$. Хорошо известно, что решение этой задачи является единственным.

Аналогично можно выразить через функции φ_1 и φ_2 и граничные условия бигармонической задачи. Если заметить, что

$$\frac{\partial U}{\partial x_\alpha} = \frac{\partial U}{\partial z} \frac{\partial z}{\partial x_\alpha} + \frac{\partial U}{\partial \bar{z}} \frac{\partial \bar{z}}{\partial x_\alpha}, \quad \alpha = 1, 2,$$

то легко получаем

$$\frac{\partial U}{\partial x_1} + j \frac{\partial U}{\partial x_2} = \varphi_1'(z) + \varphi_2(z) + z\overline{\varphi_2'(z)}.$$

Поэтому граничные условия задачи (7.14) записываются в виде

$$\overline{\varphi_1'(z)} + \varphi_2(z) + \overline{z\varphi_2'(z)} = f_1(s) + jf_2(s), \quad (7.20)$$

где f_1 и f_2 — заданные действительные функции точки границы C области R . На этот раз задача сводится к определению пары аналитических функций φ_1 и φ_2 , удовлетворяющих на границе C условию (7.20).

Мы не собираемся здесь утомлять читателя подробным перечислением ограничений, которые математики находят желательным наложить на функции $f(s)$, $f_1(s)$ и $f_2(s)$ и на область R , чтобы гарантировать существование и единственность решения рассматриваемых нами задач теории упругости. Вопрос о существовании решения был очень подробно исследован для весьма общих областей при сравнительно слабых ограничениях на поведение упомянутых функций на контуре. Однако ввиду достаточной грубости средств, которые мы предполагаем использовать для решения этих задач, достаточно будет предположить, что функции $f(s)$ являются на контуре C дифференцируемыми функциями переменного s с ограниченными производными. Эти ограничения являются более жесткими, нежели необходимо, однако в обычно возникающих задачах теории упругости они выполняются.

§ 7.12. Задача Дирихле. Обратимся теперь к определению аналитической функции по граничному условию (7.19). Мы рассмотрим сперва эту задачу в предположении, что область R представляет собой круг, который, не уменьшая общности, можно считать имеющим радиус, равный в некотором масштабе единице. Наша задача сводится при этом к определению функции $\varphi(z)$, аналитической внутри круга $|z| \leq 1$ и удовлетворяющей условию

$$\varphi(z) + \overline{\varphi(z)} = 2f(s), \quad |z| = 1. \quad (7.21)$$

Полагая $z = re^{j\theta}$ и используя символ σ для обозначения значений z на границе, т. е. при $r = 1$, можно переписать граничное условие (7.19) в виде

$$\varphi(\sigma) + \overline{\varphi(\sigma)} = 2f(\theta), \quad (7.22)$$

поскольку длина дуги s на окружности единичного радиуса γ равна θ .

Умножая, далее, обе части уравнения (7.22) на $d\sigma/[2\pi j(\sigma - z)]$, где $|z| < 1$, и интегрируя по границе γ круга $|z| \leq 1$, мы получаем¹⁾

$$\frac{1}{2\pi j} \int_{\gamma} \frac{\varphi(\sigma)}{\sigma - z} d\sigma + \frac{1}{2\pi j} \int_{\gamma} \frac{\overline{\varphi(\sigma)}}{\overline{\sigma} - \overline{z}} d\sigma = \frac{1}{\pi j} \int_{\gamma} \frac{f(\theta)}{\sigma - z} d\sigma.$$

¹⁾ Имеется теорема, принадлежащая Гарнаку (см. стр. 160 нашей книги [3]) и устанавливающая, что из этого уравнения следует уравнение (7.22). (См. также книгу [2], стр. 276. — *Прим. ред.*)

Но первый из интегралов, в силу интегральной теоремы Коши, в точности равен $\varphi(z)$. Легко показать, что второй интеграл равен $\overline{\varphi(0)}$. В самом деле,

$$\varphi(\sigma) = \sum_{n=0}^{\infty} a_n \sigma^n \quad \text{и} \quad \overline{\varphi(\sigma)} = \sum_{n=0}^{\infty} \bar{a}_n \bar{\sigma}^n.$$

Но $\sigma = e^{j\theta}$, так что $\bar{\sigma} = e^{-j\theta} = 1/\sigma$, и мы получаем

$$\overline{\varphi(\sigma)} = \sum_{n=0}^{\infty} \bar{a}_n \bar{\sigma}^n = \sum_{n=0}^{\infty} \frac{\bar{a}_n}{\sigma^n}.$$

Далее,

$$\frac{1}{2\pi j} \int_{\gamma} \frac{d\sigma}{\sigma - z} = 1, \quad \frac{1}{2\pi j} \int_{\gamma} \frac{d\sigma}{\sigma^n (\sigma - z)} = 0, \quad n \neq 0,$$

откуда получаем

$$\frac{1}{2\pi j} \int_{\gamma} \sum_{n=0}^{\infty} \frac{\bar{a}_n}{\sigma^n (\sigma - z)} d\sigma = \bar{a}_0 = \overline{\varphi(0)}.$$

В соответствии с предыдущим имеем

$$\varphi(z) = \frac{1}{\pi j} \int_{\gamma} \frac{f(\theta)}{\sigma - z} d\sigma - \overline{\varphi(0)}. \quad (7.23)$$

Для определения постоянной $\overline{\varphi(0)}$ положим в формуле (7.23) $z = 0$, после чего получим

$$\varphi(0) = \frac{1}{\pi j} \int_{\gamma} \frac{f(\theta)}{\sigma} d\sigma - \overline{\varphi(0)},$$

так что

$$\varphi(0) + \overline{\varphi(0)} = \frac{1}{\pi j} \int_{\gamma} \frac{f(\theta)}{\sigma} d\sigma. \quad (7.24)$$

Таким образом, действительная часть $\overline{\varphi(0)}$ полностью определяется равенством (7.24). Но искомое решение — функция V — определяется действительной частью функции $\varphi(z)$; таким образом, формула (7.23) полностью определяет функцию V , поскольку все члены правой части равенства (7.23) известны, коль скоро задана функция $f(\theta)$. Задача приводится, таким образом, к вычислению одного простого интеграла. Этот интеграл можно вычислять разными способами; часто, например, простым способом его вычисления может служить метод вычетов.

В том случае, когда область R не представляет собой круга, задача легко сводится к соответствующей задаче для круговой области, если отобразить конформно область R на круг единичного

радиуса (см. гл. 13). Известно, что всегда можно указать аналитическую функцию

$$z = \omega(\zeta), \quad (7.25)$$

такую, что рассматриваемая односвязная область конформно отображается на круг единичного радиуса $|\zeta| \leq 1$ в комплексной плоскости ζ , так что граница C соответствует границе $|\zeta| = 1$.

Подставляя соотношение $z = \omega(\zeta)$ в граничное условие (7.21) и полагая $\varphi[\omega(\zeta)] = \varphi_1(\zeta)$, получаем

$$\varphi_1(\zeta) + \overline{\varphi_1(\zeta)} = 2f_1(\theta)$$

при $|\zeta| = 1$, где $f_1(\theta)$ представляет собой функцию, задающую граничные значения на окружности $|\zeta| = 1$. Эта функция легко находится по функции $f(s)$, так как значения отображающей функции $\omega(\zeta)$ на границе C области R известны. Так как решение задачи Дирихле для круговой области определяется формулой (7.23), то решение рассматриваемой задачи упирается в отыскание отображающей функции (7.25).

В последние годы значительное внимание (в частности, в Советском Союзе) уделялось задаче конформного отображения, так что в настоящее время мы располагаем эффективными методами построения отображающих функций для односвязных областей. Заметим здесь, что оказывается возможным отобразить, если не самую заданную область, то по крайней мере область, сколь угодно близкую к заданной, при помощи полиномов или рациональных отображающих функций. Следует отметить, что если функция f , заданная на границе C области R , представляет собой рациональную функцию переменных x_1 и x_2 и если отображающая функция $z = \omega(\zeta)$ также является рациональной функцией ζ , то интегралы типа (7.23) можно легко вычислить в замкнутой форме в элементарных функциях.

§ 7.13. Бигармоническая краевая задача. Вполне аналогичен предыдущему метод решения бигармонической краевой задачи (7.14). Если область R не является круговой, то следует сперва конформно отобразить ее на единичный круг $|\zeta| \leq 1$ в комплексной плоскости ζ при помощи отображающей функции (7.25). Вводя обозначения

$$\varphi'_1[\omega(\zeta)] = \Phi_1(\zeta),$$

$$\varphi_2[\omega(\zeta)] = \Phi_2(\zeta),$$

мы получаем путем непосредственного вычисления, используя при этом равенство (7.20), граничное условие в виде

$$\overline{\Phi_1(\sigma)} + \Phi_2(\sigma) + \frac{\omega(\sigma)}{\omega'(\sigma)} \overline{\Phi'_2(\sigma)} = F_1(\theta) + jF_2(\theta), \quad (7.26)$$

где $\sigma = e^{j\theta}$, а функции $F_1(\theta)$ и $F_2(\theta)$ полностью определяются функциями $f_1(s)$ и $f_2(s)$, так как $\omega(\zeta)$ известно.

Но функции $\Phi_1(\zeta)$ и $\Phi_2(\zeta)$ являются аналитическими внутри круга $|\zeta| \leq 1$, так что можно написать

$$\Phi_1(\sigma) = \sum_{n=0}^{\infty} a_n \sigma^n, \quad \Phi_2(\sigma) = \sum_{n=0}^{\infty} b_n \sigma^n,$$

где a_n, b_n — неизвестные коэффициенты. Известную функцию $\omega(\sigma)/\overline{\omega'(\sigma)}$ можно представить в виде ряда по степеням σ , а функцию $F_1(\theta) + jF_2(\theta)$ — в виде комплексного ряда Фурье

$$\sum_{-\infty}^{\infty} c_n e^{jn\theta} = \sum_{-\infty}^{\infty} c_n \sigma^n.$$

Приравнявая коэффициенты при одинаковых степенях σ в обеих частях результирующего уравнения, мы получаем систему уравнений для определения коэффициентов a_n и b_n . Этот элементарный метод всегда приводит к явным выражениям для a_n и b_n , когда отображающая функция $\omega(\zeta)$ является полиномом или рациональной функцией. В общем случае метод неопределенных коэффициентов является громоздким, и мы упоминаем его здесь главным образом потому, что он часто используется практиками, работающими в области прочности.

Часто оказывается проще поступать так, как мы делали в случае задачи Дирихле. Именно, умножая обе части уравнения (7.26) на $d\sigma/[2\pi j(\sigma - \zeta)]$ и интегрируя по единичной окружности γ , мы получаем интегродифференциальное уравнение для определения функции $\Phi_2(\zeta)$:

$$\Phi_2(\zeta) + \frac{1}{2\pi j} \int_{\gamma} \frac{\omega(\sigma) \overline{\Phi_2'(\sigma)}}{\omega'(\sigma)(\sigma - \zeta)} d\sigma + \overline{\Phi_1(0)} = \frac{1}{2\pi j} \int_{\gamma} \frac{F_1(\theta) + jF_2(\theta)}{\sigma - \zeta} d\sigma. \quad (7.27)$$

Это уравнение можно преобразовать в обычное интегральное уравнение Фредгольма, для которого существуют различные методы решения. Коль скоро функция $\Phi_2(\zeta)$ определена, функцию $\Phi_1(\zeta)$ можно получить из уравнения (7.26) посредством простого интегрирования; а именно, нужно написать уравнение, сопряженное с уравнением (7.26), умножить на $d\sigma/[2\pi j(\sigma - \zeta)]$ и проинтегрировать по γ . Можно показать, что если $\omega(\zeta)$ является рациональной функцией, то функции $\Phi_1(\zeta)$ и $\Phi_2(\zeta)$ можно определить элементарными средствами, не привлекая теории интегральных уравнений.

Намеченный здесь метод может показаться сложным, но следует иметь в виду, что мы рассматриваем здесь весьма общую задачу. В частных случаях, когда граничные условия определяются рациональными функциями и контур S не слишком сложен, функции $\Phi_1(\zeta)$ и $\Phi_2(\zeta)$ часто оказывается возможным получить из уравнения (7.27) без особого труда.

§ 7.14. Заключение. Метод решения, изложенный нами выше, использует только весьма простые соображения теории аналитических функций комплексного переменного и является непосредственно применимым только в том случае, когда область R односвязна. Этот метод можно обобщить на двусвязные области, вводя функцию $z = \omega(\zeta)$, отображающую заданную двусвязную область на круговое кольцо. Однако в общем случае многосвязных областей, который встречается, например, при исследовании деформации составных и усиленных вайками цилиндров или при исследовании пластин, имеющих отверстия, аналитический аппарат становится довольно сложным. Этот аппарат опирается на теорию интегральных уравнений Фредгольма и на теорию некоторых типов сингулярных интегральных уравнений.

За последние несколько лет было получено несколько полезных решений для ряда важных задач, таких, например, как задача о деформации упругого цилиндра, усиленного ядром эллиптического сечения, при различных типах нагрузки или задача о деформации круглой пластины с эллиптическим вырезом и некоторые другие задачи.

Существенное продвижение и обобщение этих методов позволило недавно применить их к исследованию плоских задач теории упругости для анизотропной среды. В этом случае фундаментальную роль играют аналитические функции двух комплексных переменных.

Систематическое использование аналитических функций в задачах теории упругости было начато в СССР Г. В. Колосовым и Н. И. Мусхелишвили, хотя некоторое число изолированных результатов было получено независимо в работах различных авторов в США и Англии. Подробное изложение методов и результатов применения комплексных переменных и сингулярных интегральных уравнений в краевых задачах математической физики содержится в двух обширных монографиях Н. И. Мусхелишвили [1, 2]; обе эти монографии недавно переведены на английский язык и доступны теперь для западного читателя.

ЛИТЕРАТУРА

1. Мусхелишвили Н. И., Сингулярные интегральные уравнения, Гостехиздат, М.—Л., 1946.
2. Мусхелишвили Н. И., Некоторые основные задачи математической теории упругости, изд. 3, Изд. АН СССР, М.—Л., 1954.
3. Sokolnikoff I. S., *Mathematical Theory of Elasticity*, 2d ed. McGraw-Hill, New York, 1956.

Ч А С Т Ь 2

**Теоретико-вероятностные
задачи**

Глава 8

ТЕОРИЯ ПРЕДСКАЗАНИЯ

НОРБЕРТ ВИНЕР

Массачусетский технологический институт

§ 8.1. Временные ряды и причинность. Центральная идея настоящей главы состоит в том, чтобы подчеркнуть статистический характер техники связи. Так как статистика является в сущности ветвью теории вероятностей, а теория вероятностей рядом исследователей, в том числе Колмогоровым, Хинчиным, Крамером, Дубом, Леви и автором в свое время была сведена к теории меры Лебега и интегрированию по Лебегу, то теория Лебега будет постоянно применяться в этой главе. Мы включили в список литературы несколько монографий [2—4, 6, 9], содержащих основные сведения из этой области.

В сжатом изложении, подобном настоящему, неизбежно приходится предполагать знакомство читателя с теорией Лебега. Кроме того, нам понадобятся далее элементарные сведения из теории рядов и интегралов Фурье, а также сведения по теории функций комплексного переменного в объеме начального курса этой теории.

§ 8.2. Эргодическая теорема. Так называемая *эргодическая теорема* по своей сложности ненамного превосходит те теоремы, которые мы предполагаем известными. Мы будем рассматривать сообщения, являющиеся фактически функциями от двух переменных. А именно: они образуют некоторую временную последовательность, и кроме того, мы должны еще учесть необходимость рассмотрения воздействия передающей системы не только на отдельные изолированные сообщения, но сразу на целый класс таких сообщений. Различные сообщения этого класса, следовательно, должны быть различимы одно от другого, т. е. должны отмечаться каким-нибудь индексом. В качестве такого индекса мы выберем число α со значениями из интервала $(0,1)$. Переменная α будет называться *фазовой переменной*. Для наших целей необходимо вывести некоторые соотношения, связывающие распределение во времени определенных типов сообщений и их распределение по фазе.

Продельвая это, мы будем идти по стопам замечательного американского ученого Уилларда Гиббса. В своей книге по статистической механике он существенно использовал связь между временными и

фазовыми средними. Однако его объяснение совпадения этих двух видов средних значений было совершенно ошибочно. С 1900 г., когда появилась работа Гиббса, вплоть до 1930 г. положение оставалось таким же неудовлетворительным. Появившаяся затем серия теорем дала правильное обоснование выводов работы Гиббса. Эти теоремы не могли быть доказаны самим Гиббсом, так как они опирались на работы Лебега, которых Гиббс никогда не изучал и которые были выполнены, по существу, одновременно с исследованиями Гиббса.

Простейшая форма эргодической теоремы, принадлежащая Купмену [5] и Нейману [7], может быть сформулирована следующим образом. Пусть T есть сохраняющее меру точечное преобразование интервала $(0,1)$ в себя, имеющее обратное преобразование T^{-1} . Определим целые степени T , положительные и отрицательные, как итерации соответственно операторов T и T^{-1} . Тогда, если функция $f(\alpha)$ принадлежит классу L^2 в интервале $(0,1)$, то существует предел в среднем

$$f^*(\alpha) = \text{l. i. m.}_{n \rightarrow \infty} \frac{f(\alpha) + \dots + f(T^n \alpha)}{n+1}.$$

Чтобы доказать этот результат, заметим, что если $f(\alpha)$ принадлежит L^2 , то, согласно неравенству Шварца,

$$\begin{aligned} \int_0^1 \left| \frac{f(T^n \alpha) + f(\alpha)}{2} \right|^2 d\alpha &\leq \frac{1}{4} \left\{ \int_0^1 |f(T^n \alpha)|^2 d\alpha + \int_0^1 |f(\alpha)|^2 d\alpha + \right. \\ &\quad \left. + 2 \left[\int_0^1 |f(\alpha)|^2 d\alpha \int_0^1 |f(T^n \alpha)|^2 d\alpha \right]^{1/2} \right\} = \int_0^1 |f(\alpha)|^2 d\alpha. \end{aligned}$$

Следовательно, если

$$f_{2^n}(\alpha) = \frac{f(\alpha) + \dots + f(T^{2^n-1} \alpha)}{2^n},$$

то последовательность

$$\left\{ \int_0^1 |f_{2^n}(\alpha)|^2 d\alpha \right\}$$

не возрастает и предел

$$A = \lim_{n \rightarrow \infty} \int_0^1 |f_{2^n}(\alpha)|^2 d\alpha$$

существует и неотрицателен. Таким образом, мы можем построить возрастающую последовательность $\{n_k\}$, для которой

$$\lim_{k \rightarrow \infty} \limsup_{j > k} \left[\int_0^1 |f_{2^{n_k}}(\alpha)|^2 d\alpha - \int_0^1 |f_{2^{n_j}}(\alpha)|^2 d\alpha \right] = 0.$$

Но если $j > k$, то мы имеем

$$f_{2^{n_j}}(\alpha) = \frac{f_{2^{n_k}}(\alpha) + f_{2^{n_k}}(T^{2^{n_k}}\alpha) + \dots + f_{2^{n_k}}(T^{2^{n_j-2^{n_k}}}\alpha)}{2^{n_j-n_k}},$$

или

$$\frac{1}{2^{n_j-n_k}} \sum_{\lambda=0}^{2^{n_j-n_k}-1} f_{2^{n_k}}(T^{\lambda 2^{n_k}}\alpha) - f_{2^{n_j}}(\alpha) = 0.$$

Таким образом, существует по крайней мере одно такое целое λ из интервала $(0, 2^{n_j-n_k} - 1)$, что

$$\int_0^1 |f_{2^{n_j}}(\alpha)|^2 d\alpha - \operatorname{Re} \int_0^1 f_{2^{n_k}}(T^{\lambda 2^{n_k}}\alpha) \overline{f_{2^{n_j}}(\alpha)} d\alpha \leq 0.$$

Следовательно,

$$\int_0^1 |f_{2^{n_k}}(\alpha)|^2 d\alpha - \int_0^1 |f_{2^{n_j}}(\alpha)|^2 d\alpha \geq \int_0^1 |f_{2^{n_k}}(T^{\lambda 2^{n_k}}\alpha) - f_{2^{n_j}}(\alpha)|^2 d\alpha$$

и при подходящем выборе n_j и n_k мы можем добиться выполнения неравенства

$$\int_0^1 |f_{2^{n_k}}(T^{\lambda 2^{n_k}}\alpha) - f_{2^{n_j}}(\alpha)|^2 d\alpha \leq 2^{-k}.$$

Таким образом, мы можем построить последовательность функций вида

$$f_{2^{n_k}}(T^{\lambda 2^{n_k}}\alpha),$$

которая в силу леммы Вейля¹⁾ будет сходиться в среднем

¹⁾ Лемма Вейля утверждает, что любая последовательность $\{f_m(x)\}$ функций из L^2 сходится к некоторому элементу L^2 в среднем, если только

$$\lim_{n \rightarrow \infty, m \rightarrow \infty} \int_0^1 |f_n(\alpha) - f_m(\alpha)|^2 d\alpha = 0. \text{ — Прим. ред.}$$

к некоторой функции $f^*(\alpha)$ при $k \rightarrow \infty$. Так как

$$\begin{aligned} \int_0^1 |f_{2^{n_k}}(T^{\lambda k} \alpha) - f_{2^{n_k}}(T^{\lambda k} \alpha)|^2 d\alpha = \\ = 2^{-n_k} \int_0^1 |f(T^{\lambda k} \alpha) - f(T^{\lambda k + 2^{n_k} - 1} \alpha)|^2 d\alpha \leq 4 (2^{-n_k}) \int_0^1 |f(\alpha)|^2 d\alpha, \end{aligned}$$

то отсюда следует, что

$$f^*(\alpha) = f^*(T\alpha) = f^*(T^{\pm 1} \alpha).$$

Следовательно,

$$f^*(\alpha) = \text{l. i. m.}_{k \rightarrow \infty} f_{2^{n_k}}(\alpha).$$

Пусть теперь целое положительное N имеет в двоичной системе счисления представление

$$N = 2^v + \sum_{\mu=0}^{v-1} a_\mu 2^\mu, \quad a_\mu = 0 \text{ или } 1.$$

Тогда

$$\int_0^1 |f^*(\alpha) - \frac{1}{N} \sum_{\lambda=0}^{N-1} f(T^\lambda \alpha)|^2 d\alpha = \left[\sum_{\mu=0}^v \frac{a_\mu 2^\mu}{N} \left(\int_0^1 |f^*(\alpha) - f_{2^\mu}(\alpha)|^2 d\alpha \right)^{1/2} \right]^2.$$

В силу рассмотренного выше случая отсюда вытекает следующий общий результат:

$$f^*(\alpha) = \text{l. i. m.}_{N \rightarrow \infty} f_N(\alpha).$$

Предположим теперь, что $f(\alpha)$ удовлетворяет всем предыдущим условиям, с тем лишь отличием, что сейчас $f(\alpha)$ принадлежит L , а не L^2 . Если

$$f_A(\alpha) = \begin{cases} f(\alpha) & \text{при } |f(\alpha)| \leq A, \\ A \frac{f(\alpha)}{|f(\alpha)|} & \text{при } |f(\alpha)| > A, \end{cases}$$

то предел $(f_A)^* \alpha$ существует и принадлежит L^2 . Кроме того, имеет место также и сходимости в смысле L :

$$(f_A)^*(\alpha) = \text{l. i. m.}_{N \rightarrow \infty} \frac{f_A(\alpha) + f_A(T\alpha) + \dots + f_A(T^{N-1}\alpha)}{N}.$$

Но так как

$$\begin{aligned} \int_0^1 \left| \frac{f_A(\alpha) - f_B(\alpha) + \dots + f_A(T^{N-1}\alpha) - f_B(T^{N-1}\alpha)}{N} \right| d\alpha \leq \\ \leq \int_0^1 |f_A(\alpha) - f_B(\alpha)| d\alpha, \end{aligned}$$

то, положив $N \rightarrow \infty$, получим

$$\int_0^1 |(f_A)^*(\alpha) - (f_B)^*(\alpha)| d\alpha \leq \int_0^1 |f_A(\alpha) - f_B(\alpha)| d\alpha.$$

Поскольку правая часть последнего выражения стремится к 0 при $A, B \rightarrow \infty$,

$$\lim_{A, B \rightarrow \infty} \int_0^1 |(f_A)^*(\alpha) - (f_B)^*(\alpha)| d\alpha = 0.$$

Отсюда с помощью L -формы леммы Вейля можно сделать вывод, что существует такая функция $f^*(\alpha)$, принадлежащая L , что

$$\lim_{A \rightarrow \infty} (f_A)^*(\alpha) = f^*(\alpha),$$

причем легко показать, что

$$\int_0^1 |f^*(\alpha) - (f_A)^*(\alpha)| d\alpha \leq \int_0^1 |f(\alpha) - f_A(\alpha)| d\alpha.$$

Отсюда мы заключаем, что в L -смысле

$$f^*(\alpha) = \text{l. i. m.}_{N \rightarrow \infty} \frac{f(\alpha) + \dots + f(T^{N-1}\alpha)}{N}.$$

Это — эргодическая теорема в смысле сходимости в среднем.

Перейдем теперь к более сильной эргодической теореме Биркгоффа [1]. Пусть

$$\varphi(\alpha) = \left| f^*(\alpha) - \frac{f(\alpha) + \dots + f(T^{N-1}\alpha)}{N} \right|$$

и число N выбрано таким образом, что

$$\int_0^1 \varphi(\alpha) d\alpha < \varepsilon.$$

Рассмотрим

$$\varphi(\nu, \alpha) = \frac{\varphi(\alpha) + \varphi(T\alpha) + \dots + \varphi(T^{\nu-1}\alpha)}{\nu}.$$

Существует теорема, согласно которой, если

$$\max_{0 < m \leq \nu} \varphi(m, \alpha) = A_\nu(\alpha),$$

то число значений ν из интервала $(0, n)$, таких, что

$$A_{n-\nu}(T^\nu) > c > 0,$$

не превосходит $n\varphi(n, \alpha)/c$. Интегрируя обе части этого неравенства по α , мы получаем

$$\begin{aligned} \int_0^1 [\text{число значений } \nu \text{ из } (0, n), \text{ для которых } A_{n-\nu}(T^\nu \alpha) > c] d\alpha &\leq \\ &\leq \frac{n}{c} \int_0^1 \varphi(n, \alpha) d\alpha = \frac{n}{c} \int_0^1 \varphi(\alpha) d\alpha, \end{aligned}$$

т. е.

$$\begin{aligned} \frac{1}{n} [\text{сумма мер множеств значений } \alpha, \text{ для которых } A_{n-\nu}(T^\nu \alpha) > c] &\leq \\ &\leq \frac{1}{c} \int_0^1 \varphi(\alpha) d\alpha \leq \frac{\varepsilon}{c}. \end{aligned}$$

Следовательно,

$$\frac{1}{n} \sum_{\nu=0}^{n-1} [\text{мера множества значений } \alpha, \text{ для которых } A_\nu(\alpha) > c] \leq \frac{\varepsilon}{c}.$$

Поскольку, однако, меры последних множеств возрастают (или по крайней мере не убывают) с ростом ν , то мера множества значений α , для которых $A_\nu(\alpha) > \varepsilon^{1/2}$, при соответствующем ν будет пределом последовательности возрастающих выражений, не превосходящих $\varepsilon^{1/2}$, и, следовательно, сама не может превосходить $\varepsilon^{1/2}$. Таким образом, за исключением множества значений α , мера которого не превосходит $\varepsilon^{1/2}$, мы будем иметь

$$\begin{aligned} \lim_{m \rightarrow \infty} \left| f^*(\alpha) - \frac{1}{(m+1)(n+1)} \{ f(\alpha) + 2f(T\alpha) + \dots + (n+1)f(T^n\alpha) + \right. \\ \left. + (n+1)[f(T^{n+1}\alpha) + \dots + f(T^{m-1}\alpha)] + (n+1)f(T^m\alpha) + \right. \\ \left. + nf(T^{m+1}\alpha) + \dots + 2f(T^{m+n-1}\alpha) + f(T^{m+n}\alpha) \} \right| = 0. \end{aligned}$$

Легко видеть, что первые и последние члены суммы, входящие с весом, отличным от $1/(m+1)$, не могут повлиять на значение этого предела и что

$$\lim_{m \rightarrow \infty} \left| f^*(\alpha) - \frac{1}{m+1} [f(\alpha) + \dots + f(T^m\alpha)] \right| = 0,$$

за исключением, быть может, множества значений α , мера которого не превосходит $2\varepsilon^{1/2}$. Так как ε сколь угодно мало, то, следовательно,

$$f^*(\alpha) = \lim_{m \rightarrow \infty} \frac{1}{m+1} [f(\alpha) + \dots + f(T^m\alpha)]$$

для почти всех значений α . Это и есть эргодическая теорема Биркгоффа.

Пусть теперь преобразование T оставляет измеримые множества инвариантными только тогда, когда их мера равна 0 или 1. Тогда мы можем доказать, что множество значений α , удовлетворяющих неравенству $a \leq f^*(\alpha) \leq b$, измеримо и что это неравенство удовлетворяется либо почти всюду, либо почти нигде. С помощью процесса последовательного вложения мы можем показать, что существует только одно такое значение A , что для почти всех значений α и для произвольного $\varepsilon > 0$

$$A - \varepsilon \leq f^*(\alpha) \leq A + \varepsilon.$$

Предположим, что такое A существует. Так как

$$\int_0^1 f(\alpha) d\alpha = \int_0^1 f^*(\alpha) d\alpha,$$

то мы имеем

$$A = \int_0^1 f(\alpha) d\alpha$$

и для почти всех α_0

$$\lim_{N \rightarrow \infty} \frac{f(\alpha_0) + \dots + f(T^N \alpha_0)}{N+1} = \int_0^1 f(\alpha) d\alpha.$$

Гипотеза о несуществовании инвариантных измеримых множеств, имеющих меру, отличную от 0 или 1, известна как „эргодическая гипотеза“. Она позволяет заменять (для почти всех значений α_0) средние по фазовой переменной α на средние по временной переменной N и, наоборот, — средние по N на средние по α .

Распространим эргодическую теорему на случай непрерывного времени. Пусть T^t — группа сохраняющих меру преобразований α в себя с обратными преобразованиями T^{-t} и пусть

$$T^\lambda T^n \alpha = T^{\lambda+n} \alpha.$$

Предположим, что функция $f(T^\lambda \alpha)$ интегрируема по α и измерима по λ при α , пробегающем значения из интервала $(0,1)$ и λ , меняющимся от $-\infty$ до $+\infty$. Тогда интеграл

$$\int_0^1 f(T^\lambda \alpha) d\lambda$$

является функцией α , принадлежащей L . Следовательно, существует (почти всюду) предел

$$\begin{aligned} f^*(\alpha) &= \lim_{N \rightarrow \infty} \frac{1}{N+1} \left[\int_0^1 f(T^t \alpha) dt + \int_0^1 f(T^{t+1} \alpha) dt + \dots \right. \\ &\quad \left. \dots + \int_0^1 f(T^{t+N} \alpha) dt \right] = \\ &= \lim_{N \rightarrow \infty} \frac{1}{N+1} \int_0^{N+1} f(T^t \alpha) dt. \end{aligned}$$

Сначала мы предположим, что функция f неотрицательна, а в качестве N выберем $[U]^1$; тогда

$$\frac{1}{U} \int_0^U f(T^t \alpha) dt \geq \frac{N}{U} \frac{1}{N} \int_0^N f(T^t \alpha) dt$$

и

$$\frac{1}{U} \int_0^U f(T^t \alpha) dt \leq \frac{N+1}{U} \frac{1}{N+1} \int_0^{N+1} f(T^t \alpha) dt.$$

Так как правые части обоих неравенств сходятся при $U \rightarrow \infty$ почти всюду к $f^*(\alpha)$, то, очевидно, для почти всех значений α

$$f^*(\alpha) = \lim_{U \rightarrow \infty} \frac{1}{U} \int_0^U f(T^t \alpha) dt.$$

Так как каждую интегрируемую функцию можно представить в виде разности двух положительных интегрируемых функций, то это соотношение можно распространить и на функции, не удовлетворяющие требованию положительности.

Эргодическая гипотеза в непрерывном случае несколько слабее, чем в дискретном. Она заключается в предположении о том, что не существует измеримого множества с мерой, отличной от 1 или 0, которое было бы инвариантно относительно *всех* преобразований T^t ; действительно, только такие множества могут быть множествами значений α , на которых $f^*(\alpha)$ остается постоянной. Если таких множеств с мерой, отличной от 0 и 1, не существует, то, как и прежде, мы имеем почти всюду (т. е. для почти всех α_0)

$$\lim_{A \rightarrow \infty} \frac{1}{A} \int_0^A f(T^t \alpha_0) dt = \int_0^1 f(\alpha) d\alpha.$$

¹⁾ Символ $[U]$ означает наибольшее целое число, не превосходящее U ; если же U — целое, то $[U] = U$.

§ 8.3. Временные ряды. Мы будем рассматривать теперь последовательности событий, распределенных во времени. Предположим сначала, что время представляет собой дискретную последовательность равноотстоящих „моментов наблюдения“, простирающуюся от бесконечного прошлого до бесконечного будущего. Мы можем считать, что такие моменты отвечают всем целым числам, отрицательным и положительным, а также нулю. Таким образом, временным рядом простейшего вида будет последовательность чисел A_n , где n пробегает все целые значения от $-\infty$ до $+\infty$.

Представляется желательным рассматривать не индивидуальный временной ряд, а множество временных рядов с определенным распределением вероятностей. Любая форма введения вероятности эквивалентна введению меры в некотором пространстве вспомогательной переменной, имеющем конечную меру, удовлетворяющую обычной теории Лебега. Хорошо известно, однако, что теория Лебега не зависит от числа измерений пространства, так что нужная нам теория меры и интегрирования всегда будет эквивалентна теории меры и интегрирования для переменного α , принимающего значения из интервала $(0,1)$. Таким образом, мы приходим к временным рядам, рассматриваемым как последовательности функций $f_n(\alpha)$.

Не все такие последовательности нам одинаково интересны. В первую очередь наиболее интересны те из них, для которых в каком-то смысле безразличен выбор нулевого момента времени. Другими словами, $f_n(\alpha)$ должны иметь одно и то же распределение для всех значений n , хотя это и не означает, что после того, как мы изменили n , отдельные f остались без изменения. В сущности, мы хотим рассмотреть тот случай, для которого совместное распределение любого конечного набора $f_n(\alpha)$ остается неизменным при сдвиге всех индексов n на единицу вперед или назад. Для того чтобы яснее понять и сделать более четким наше предположение, рассмотрим функцию $f(\alpha)$ из класса L^2 и сохраняющее меру Лебега преобразование интервала $(0,1)$ в себя. Предположим также, что это преобразование T имеет обратное преобразование, которое мы обозначим через T^{-1} . Временной ряд $f(T^n\alpha)$, где n меняется от $-\infty$ до $+\infty$, мы назовем *стационарным временным рядом*. Специальная задача, которую мы хотим сначала рассмотреть, состоит в предсказании (иначе экстраполяции) будущего поведения таких временных рядов по известному их настоящему и прошлому.

Любопытно, что оказывается легче изучать временные ряды, для которых предсказание не может быть осуществлено с любой степенью точности. Под предсказанием значения $f(T^n\alpha)$ мы будем понимать проекцию функции $f(T^n\alpha)$ на подпространство гильбертова пространства L^2 , являющееся замыканием множества $f(\alpha), f(T^{-1}\alpha), \dots, f(T^{-n}\alpha), \dots$; отыскивая эту проекцию, мы будем предполагать, что она стремится в среднем к нулю при бесконечном возрастании

числа n (т. е. что точность прогноза неограниченно убывает при возрастании n).

Задача, которую мы пытаемся решить, может быть сформулирована и в другой форме. Именно требуется найти в пространстве, порожденном настоящим и прошлым функции f , такую функцию $g(\alpha)$, что интеграл

$$\int_0^1 |f(T^n \alpha) - g(\alpha)|^2 d\alpha$$

имеет наименьшее значение. При этом, естественно, предполагается, что функция $f(\alpha)$ не равна тождественно нулю.

Для того чтобы решить эту задачу, ортогонализируем множество функций $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, ..., $f(T^{-n}\alpha)$, ... Первая ортогонализованная функция

$$g_1(\alpha) = \frac{f(T^{-1}\alpha)}{\left[\int_0^1 |f(\beta)|^2 d\beta \right]^{1/2}}$$

принадлежит, очевидно, L^2 и не равна тождественно нулю (в смысле L^2). Строим далее

$$G_2(\alpha) = f(T^{-2}\alpha) - g_1(\alpha) \int_0^1 f(T^{-2}\beta) \overline{g_1(\beta)} d\beta.$$

Ясно, что эта функция ортогональна $g_1(\alpha)$ и также принадлежит L^2 . Она также не равна тождественно нулю, так как в противном случае мы могли бы выразить $f(T^{-1}\alpha)$ линейно через $f(T^{-2}\alpha)$ и, продолжая этот процесс далее, через $f(T^{-n}\alpha)$ при любом (сколь угодно большом) n , что противоречит сделанным предположениям. Если мы положим

$$g_2(\alpha) = \frac{G_2(\alpha)}{\left[\int_0^1 |G_2(\beta)|^2 d\beta \right]^{1/2}},$$

то пара функций $\{g_1(\alpha), g_2(\alpha)\}$ будет ортонормированной, и через эти функции можно будет выразить любую функцию, представимую посредством $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, и наоборот. Положим теперь

$$G_3(\alpha) = f(T^{-3}\alpha) - g_1(\alpha) \int_0^1 f(T^{-3}\beta) \overline{g_1(\beta)} d\beta - g_2(\alpha) \int_0^1 f(T^{-3}\beta) \overline{g_2(\beta)} d\beta.$$

Как и прежде, мы можем показать, что эта функция не равна тождественно нулю и может быть нормирована, после чего она даст нам функцию $g_3(\alpha)$. Продолжая этот процесс, мы получаем счетное множество $\{g_n(\alpha)\}$ ортонормированных функций, эквивалентное относительно задач разложения в ряды в пространстве L^2 множеству $\{f(T^{-n}\alpha)\}$, $n = 1, 2, \dots$.

Построим теперь функцию

$$H(\alpha) = f(\alpha) - \sum_{n=1}^{\infty} g_n(\alpha) \int_0^1 f(\beta) \overline{g_n(\beta)} d\beta,$$

которая, очевидно, будет линейно зависеть от $f(\alpha)$, $f(T^{-1}\alpha)$, ... и будет ортогональна к $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, ... Эта функция не может быть тождественно равна нулю, так как иначе функция $f(\alpha)$ принадлежала бы замыканию линейной оболочки функций $f(T^{-n}\alpha)$, $f(T^{-n-1}\alpha)$, ..., что не допускается нашими предположениями. Следовательно, существует такая функция $h(\alpha)$ из L^2 , что

$$h(\alpha) = \frac{H(\alpha)}{\left[\int_0^1 |H(\beta)|^2 d\beta \right]^{1/2}}.$$

Функция $h(\alpha)$ нормирована и ортогональна к $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, ... Поэтому она ортогональна и ко всем функциям $h(T^{-\nu}\alpha)$, $\nu > 0$, очевидно, линейно зависящим от $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, ... Так как преобразование T сохраняет меру и ортогональность, то множество $h(T^{\nu}\alpha)$, $-\infty < \nu < \infty$ является ортонормированным множеством функций. Следовательно, по теореме Рисса — Фишера, существует функция

$$F(\alpha) = f(\alpha) - \sum_{\nu=-\infty}^{\infty} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta,$$

где бесконечная сумма существует в смысле предела в L^2 . Так как

$$\int_0^1 h(\alpha) \overline{f(T^{-\nu}\alpha)} d\alpha = 0, \quad \nu = 1, 2, \dots,$$

то

$$\int_0^1 \overline{h(T^{\nu}\alpha)} f(\alpha) d\alpha = 0, \quad \nu = 1, 2, \dots,$$

и значит

$$F(\alpha) = f(\alpha) - \sum_{\nu=0}^{\infty} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta.$$

Покажем теперь, что $F(\alpha)$ равна нулю в L^2 . Ясно, что функция

$$f(\alpha) - \sum_{\nu=0}^{m-1} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta,$$

представляет собой разность между $f(\alpha)$ и ее проекцией на подпространство замкнутой оболочки функций $f(T^{-1}\alpha)$, $f(T^{-2}\alpha)$, ..., ортогональное к $f(T^{-m}\alpha)$, $f(T^{-m-1}\alpha)$, ... Эта разность является проекцией $f(\alpha)$ на замыкание линейной оболочки функций $f(T^{-m}\alpha)$, $f(T^{-m-1}\alpha)$, ... Отсюда

$$\begin{aligned} & \int_0^1 \left| f(\alpha) - \sum_{\nu=0}^{m-1} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta \right|^2 d\alpha = \\ &= \int_0^1 \left| \text{проекция } f(\alpha) \text{ на замыкание линейной оболочки} \right. \\ & \quad \left. \text{функций } f(T^{-m}\alpha), f(T^{-m-1}\alpha), \dots \right|^2 d\alpha. \end{aligned}$$

Поэтому

$$\lim_{m \rightarrow \infty} \int_0^1 \left| f(\alpha) - \sum_{\nu=0}^{m-1} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta \right|^2 d\alpha = 0.$$

Следовательно,

$$\int_0^1 |f(\alpha)|^2 d\alpha = \sum_{\nu=0}^{\infty} \left| \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta \right|^2$$

и

$$f(\alpha) = \sum_{\nu=0}^{\infty} h(T^{-\nu}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta$$

в пространстве L^2 , т. е. функция $F(\alpha)$, как элемент L^2 , равна нулю.

Очевидно, что

$$\sum_{\nu=0}^{\infty} \left| \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta \right|^2 < \infty.$$

Отсюда следует, что функция

$$\psi(z) = \sum_{\nu=0}^{\infty} z^{\nu} \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta$$

является аналитической внутри единичного круга и что при $r \rightarrow 1$ функция $\psi(re^{i\theta})$ сходится в среднем к функции

$$Q(\theta) = \sum_{\nu=0}^{\infty} e^{j\nu\theta} \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta, \quad j^2 = -1. \quad (8.1)$$

Докажем теперь, что функция $\psi(z)$ не имеет нулей внутри единичного круга. Заметим, что рассуждения, с помощью которых мы получили $h(\alpha)$, показывают, что существует такая функция $h_{\varepsilon}(\alpha)$ вида

$$h_{\varepsilon}(\alpha) = \sum_{n=0}^N a_{\varepsilon,n} f(T^{-n}\alpha),$$

что

$$\int_0^1 |h_{\varepsilon}(\alpha) - h(\alpha)|^2 d\alpha < \varepsilon.$$

Далее,

$$\begin{aligned} h_{\varepsilon}(\alpha) &= \sum_{n=0}^N a_{\varepsilon,n} \sum_{\nu=0}^{\infty} h(T^{-\nu-n}\alpha) \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta = \\ &= \sum_{m=0}^{\infty} h(T^{-m}\alpha) \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta, \end{aligned}$$

так что

$$\left| 1 - a_{\varepsilon,0} \int_0^1 f(\beta) \overline{h(\beta)} d\beta \right|^2 + \sum_{m=1}^{\infty} \left| \sum_{n=1}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta \right|^2 \leq \varepsilon.$$

Пусть теперь $\psi(r_0 e^{j\theta_0}) = 0$ для некоторого $r_0 < 1$; тогда

$$\sum_{m=0}^{\infty} r_0^m e^{jm\theta_0} \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta = 0.$$

С другой стороны,

$$1 + \sum_{m=0}^{\infty} \left| \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta \right|^2 - 2 \operatorname{Re} \left[a_{\varepsilon,0} \int_0^1 f(\beta) \overline{h(\beta)} d\beta \right] \leq \varepsilon.$$

Вспомним, что, согласно неравенству Шварца,

$$\varepsilon^{1/2} \geq \int_0^1 |h_\varepsilon(\alpha) - h(\alpha)| |h(\alpha)| d\alpha \geq \left| a_{\varepsilon,0} \int_0^1 f(\beta) \overline{h(\beta)} d\beta - 1 \right|.$$

Отсюда получаем

$$\left| 1 - \sum_{m=0}^{\infty} \left| \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta \right|^2 \right| \leq \varepsilon + 2\varepsilon^{1/2}.$$

Это приводит нас к соотношению

$$\begin{aligned} 0 &= \sum_{m=0}^{\infty} r_0^m e^{jm\theta_0} \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta = \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\varphi) \frac{1 - r_0^2}{1 - 2r_0 \cos(\varphi - \theta_0) + r_0^2} d\varphi, \end{aligned}$$

где

$$S(\varphi) = \sum_{m=0}^{\infty} e^{jm\varphi} \sum_{n=0}^m a_{\varepsilon,n} \int_0^1 f(\beta) \overline{h(T^{n-m}\beta)} d\beta.$$

Отсюда следует, что

$$\begin{aligned} 0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} [S(\varphi) - 1] \frac{1 - r_0^2}{1 - 2r_0 \cos(\varphi - \theta_0) + r_0^2} d\varphi + \\ &\quad + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1 - r_0^2}{1 - 2r_0 \cos(\varphi - \theta_0) + r_0^2} d\varphi, \end{aligned}$$

или

$$1 \leq (\varepsilon + 2\varepsilon^{1/2})^{1/2} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{(1 - r_0^2)^2}{[1 - 2r_0 \cos(\varphi - \theta_0) + r_0^2]^2} d\varphi \right\}^{1/2} \leq \frac{(\varepsilon + 2\varepsilon^{1/2})^{1/2}}{(1 - r_0^2)^{1/2}}.$$

Итак,

$$1 - r_0^2 \leq \varepsilon + 2\varepsilon^{1/2},$$

что приводит к противоречию, поскольку r_0 — фиксированная положительная величина, меньшая 1, а ε произвольно мало. Следовательно, $\psi(z)$ не имеет нулей внутри единичного круга.

Формально мы имеем

$$\begin{aligned} |Q(\theta)|^2 &= \left| \sum_{\nu=-\infty}^{\infty} e^{j\nu\theta} \int_0^1 f(\beta) \overline{h(T^{-\nu}\beta)} d\beta \right|^2 = \\ &= \sum_{\nu=-\infty}^{\infty} e^{j\nu\theta} \sum_{m=-\infty}^{\infty} \int_0^1 f(\beta) \overline{h(T^{-m}\beta)} d\beta \int_0^1 \overline{f(\beta)} h(T^{-m+\nu}\beta) d\beta = \\ &= \sum_{\nu=-\infty}^{\infty} e^{j\nu\theta} \int_0^1 f(T^\nu\beta) \overline{f(\beta)} d\beta. \end{aligned}$$

На самом деле, последний ряд может оказаться суммируемым только методом Чезаро, однако всегда будет справедливо равенство

$$\int_0^1 f(T^\nu\beta) \overline{f(\beta)} d\beta = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Q(\theta)|^2 e^{-j\nu\theta} d\theta.$$

Здесь $|Q(\theta)|^2$ принадлежит L . Следовательно, почти всюду

$$|Q(\theta)|^2 = \lim_{N \rightarrow \infty} \sum_{\nu=-N}^N \left(1 - \frac{|\nu|}{N}\right) e^{j\nu\theta} \int_0^1 f(T^\nu\beta) \overline{f(\beta)} d\beta$$

и

$$|\psi(z)|^2 = \lim_{N \rightarrow \infty} \sum_{\nu=-N}^N \left(1 - \frac{|\nu|}{N}\right) r^{2\nu} \int_0^1 f(T^\nu\beta) \overline{f(\beta)} d\beta.$$

Мы хотим получить новый метод нахождения $Q(\theta)$. Так как внутри единичного круга $\psi(z)$ — аналитическая функция, не имеющая нулей, то можно положить

$$\log \psi(z) = \log |\psi(z)| + j\psi_1(z),$$

где $z = re^{j\theta}$. Пусть

$$\log \psi(z) = \sum_{n=0}^{\infty} b_n r^n e^{jn\theta}.$$

Тогда

$$\log |\psi(z)| = \sum_{n=0}^{\infty} r^n \operatorname{Re} (b_n e^{jn\theta})$$

и

$$j\psi_1(z) = \sum_{n=0}^{\infty} r^n \operatorname{Im} (b_n e^{jn\theta}).$$

Для того чтобы получить $\psi(z)$, можно поступать следующим образом. Мы исходим из величины $|\psi(z)|^2$, из которой извлекаем

квадратный корень. Далее составляется логарифм получившегося выражения, который представляем в виде ряда Фурье по переменному θ . Производя затем так называемое преобразование Гильберта этого ряда, состоящее в замене коэффициентов при синусах на коэффициенты при косинусах, а коэффициентов при косинусах — на минус коэффициенты при синусах, мы получим ряд Фурье функции $\psi_1(z)$. Восстановив по $\log |\psi(z)|$ и $\psi_1(z)$ функцию $\log \psi(z)$, мы, естественно, получим

$$\psi(z) = e^{\log \psi(z)}.$$

При таком построении функции $\psi(z)$ используются только величины

$$\int_0^1 f(T^v \beta) \overline{f(\beta)} d\beta,$$

которые называются *автокорреляционными коэффициентами* функции f . Разлагая теперь функцию $\psi(z)$ в ряд Фурье, мы получаем величины

$$\int_0^1 f(\beta) \overline{h(T^{-v} \beta)} d\beta.$$

Можно показать, что, разлагая подобным же образом функцию $1/\psi(z)$, мы получим множество коэффициентов $\{c_n\}$, таких, что

$$h(\alpha) = \lim_{r \rightarrow 1} \sum_{n=0}^{\infty} r^n c_n f(T^{-n} \alpha).$$

Обычно этот метод построения $h(\alpha)$ оказывается более простым, чем приведенный выше метод, основанный на процессе ортогонализации.

Найдем теперь проекцию функции $f(T^v \alpha)$ на настоящее и прошлое последовательности $f(\alpha)$, которая будет вместе с тем проекцией $f(T^v \alpha)$ на настоящее и прошлое последовательности $h(\alpha)$. Соответственно этому, искомая проекция будет равна

$$\sum_{n=0}^{\infty} h(T^{-n} \alpha) \int_0^1 f(T^v \beta) \overline{h(T^{-n} \beta)} d\beta. \quad (8.2)$$

С другой стороны, $f(T^v \alpha)$ можно представить в виде

$$\sum_{n=0}^{\infty} h(T^{v-n} \alpha) \int_0^1 f(\beta) \overline{h(T^{-n} \beta)} d\beta = \sum_{n=-v}^{\infty} h(T^{-n} \alpha) \int_0^1 f(T^v \beta) \overline{h(T^{-n} \beta)} d\beta.$$

Разность этих двух выражений равна

$$\sum_{n=-v}^{-1} h(T^{-n} \alpha) \int_0^1 f(T^v \beta) \overline{h(T^{-n} \beta)} d\beta,$$

так что средний квадрат абсолютной ошибки наилучшего предсказания (т. е. минимум среднего квадрата абсолютной ошибки) будет равен

$$\sum_{n=-\nu}^{-1} \left| \int_0^1 f(T^n \beta) \overline{h(T^{-n} \beta)} d\beta \right|^2 = \sum_{n=0}^{\nu-1} \left| \int_0^1 f(\beta) \overline{h(T^{-n} \beta)} d\beta \right|^2. \quad (8.3)$$

Возникает еще вопрос о том, при каких условиях последовательность $\{A_\nu\}$ является последовательностью величин

$$\int_0^1 f(T^n \beta) \overline{f(\beta)} d\beta$$

для некоторой функции $f(\alpha)$, удовлетворяющей всем указанным выше условиям. Иначе говоря, при каких условиях $\{A_\nu\}$ есть последовательность коэффициентов Фурье

$$A_\nu = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Q(\theta)|^2 e^{-j\nu\theta} d\theta$$

некоторой положительной (или по крайней мере неотрицательной) функции $|Q(\theta)|^2$, принадлежащей L и такой, что $Q(\theta)$ является граничным значением в среднем функции $\psi(re^{j\theta})$, не имеющей нулей в единичном круге? При этом $\log \psi(re^{j\theta})$ будет аналитической функцией в единичном круге и мнимая часть этой функции будет отличаться только множителем j от преобразования Гильберта ее действительной части.

Предположим теперь, что $Q(\theta)$ есть предел в среднем функции $\psi(re^{j\theta})$ при $r \rightarrow 1$, где $\psi(z)$ — функция, не имеющая нулей и особых точек в единичном круге. Прежде всего так как интеграл

$$\int_{-\pi}^{\pi} |\psi(re^{j\theta})|^2 d\theta$$

ограничен, то легко показать, что и

$$\int_{-\pi}^{\pi} [V \log^+ \psi(re^{j\theta})]^2 d\theta$$

будет ограничен. Отсюда, принимая во внимание равенство

$$\int_{-\pi}^{\pi} [V \log \psi(re^{j\theta})]^2 d\theta = 0,$$

нетрудно показать, что интеграл

$$\int_{-\pi}^{\pi} |V \log \psi(re^{j\theta})|^2 d\theta$$

также ограничен. Следовательно, мы можем написать

$$V \log \psi(re^{j\theta}) = \sum_{n=0}^{\infty} a_n r^n e^{jn\theta},$$

где

$$\sum_{n=0}^{\infty} |a_n|^2 r^{2n} < \text{const}$$

для $r < 1$. Отсюда вытекает конечность интеграла

$$\int_{-\pi}^{\pi} |\log Q(\theta)| d\theta.$$

Позднее, с помощью теории броуновского движения, мы сможем показать, что если функция

$$\varphi(\theta) = |Q(\theta)|^2$$

принадлежит L в интервале $(-\pi, \pi)$ и

$$\int_{-\pi}^{\pi} |\log \varphi(\theta)| d\theta < \infty,$$

то на самом деле могут быть найдены такое сохраняющее меру преобразование T интервала $(0,1)$ в себя, а также такая функция $f(\alpha)$, принадлежащая L^2 в интервале $(0,1)$, не равная нулю почти всюду проекция которой на далекое прошлое стремится в среднем к нулю, что для этих T и $f(\alpha)$ будет иметь место равенство (8.1).

Те же методы оказываются пригодными для построения временных рядов $f_1(T_1^n \alpha)$ со всеми приведенными выше свойствами, для которых

$$\int_0^1 f_1(T_1^n \alpha) \overline{f_1(\alpha)} d\alpha = r^{|n|}, \quad r < 1.$$

Если теперь $f(T^n \beta)$ есть любой временной ряд, то совершенно ясно, что $f_1(T_1^n \alpha) f(T^n \beta)$ будет временным рядом, зависящим от пары переменных (α, β) , изменяющихся на интервалах $(0,1)$, $(0,1)$, допускающей

отображение на одну переменную, изменяющуюся в интервале $(0,1)$. Автокорреляционные коэффициенты этого нового временного ряда будут равны

$$r^n \int_0^1 f(T^n \beta) \overline{f(\beta)} d\beta.$$

Заметим, что проекция $f_1(\alpha)f(\beta)$ на $f_1(T_1^{-n}\alpha)f(T^{-n}\beta)$ будет меньше, чем постоянная, умноженная на r^n . Следовательно, проекция на все прошлое, удаленное на n и более шагов, не будет превосходить по величине $br^n/(1-r)$. Эта проекция стремится к нулю, когда $n \rightarrow \infty$ и, значит, мы имеем право пользоваться теорией предсказания, которую мы построили для $f_1(T_1^n\alpha), f(T^n\beta)$. Это даст нам приближение к этому выражению в виде

$$\sum_{v=0}^{\infty} A_v f_1(T_1^{-v}\alpha) f(T^{-v}\beta).$$

Умножая эту величину на $f(\alpha)$ и интегрируя, мы получаем в качестве приближения к $r^n f(T^n\beta)$ выражение

$$\sum_{v=0}^{\infty} A_v r^v f(T^{-v}\beta).$$

Предполагая теперь, что $r \rightarrow 1$, мы можем показать, что полученное приближение во всяком случае может быть сделано в среднем настолько близким к $f(T^n\beta)$, насколько близким может быть любое возможное предсказание.

Тем самым мы решили практическую задачу о предсказании даже и для того случая, когда проекция настоящего на далекое прошлое не стремится к нулю.

§ 8.4. Предсказание сообщений, заданных не полностью. Пусть $f(T^n\alpha)$ — временной ряд, удовлетворяющий условию, что проекция его настоящего на далекое прошлое стремится к нулю. Иногда бывает удобно рассматривать $f(\alpha)$ как сумму двух функций $f_1(\alpha)$ и $f_2(\alpha)$, где $f_1(\alpha)$ и $f_2(\alpha)$ играют роль соответственно истинного сообщения и искажения этого сообщения, которое мы будем называть *шумом*. Такое разделение на истинное сообщение и шум в какой-то степени произвольно; с точки зрения научного метода это разделение на две части имеет смысл тогда, когда мы обладаем теорией, дающей приближенное описание некоторого временного ряда, но при его рассмотрении должны учитывать еще возмущающее воздействие дополнительного фактора, добавляющее к нашему ряду «помехи», не поддающиеся теоретическому описанию. Эти «помехи» и представляют собой шум; при более полном анализе его, может быть, можно описать, исходя из некоторых других теоретических предположений,

и снова разложить на «точное сообщение» и «шум». Таким образом, понятие шума используется для описания той части временного ряда, которая не поддается точному измерению из-за неполноты имеющегося в нашем распоряжении анализа.

Чтобы дать наилучшее из возможных предсказаний величины $f_1(T^n\alpha)$ — сообщения в будущий момент времени — в терминах настоящего и прошлого функции $f(\alpha)$, мы сначала построим функцию $h(\alpha)$ и составим с ее помощью ортонормированную последовательность $h(T^n\alpha)$, такую, что настоящее и прошлые функции $h(\alpha)$ будут эквивалентны (в смысле линейной зависимости) настоящему и прошлому функции $f(\alpha)$. Положим теперь

$$\begin{aligned} f_1(T^n\alpha) &= \sum_{n=-\infty}^{\infty} h(T^{-n}\alpha) \int_0^1 f_1(T^n\beta) \overline{h(T^{-n}\beta)} d\beta = \\ &= \sum_{n=0}^{\infty} h(T^{-n}\alpha) \int_0^1 f_1(T^{n+n}\beta) \overline{h(\beta)} d\beta + \\ &+ \sum_{n=1}^{\infty} h(T^n\alpha) \int_0^1 f_1(T^{n-n}\beta) \overline{h(\beta)} d\beta. \end{aligned}$$

Первый член в правой части этой формулы представляет собой проекцию функции $f_1(T^n\alpha)$ на настоящее и прошлое функции $f(\alpha)$; поэтому второй член дает нам следующее выражение для среднего квадрата разности между $f_1(T^n\alpha)$ и этой ее проекцией:

$$\sum_{n=1}^{\infty} \left| \int_0^1 f_1(T^{n-n}\beta) \overline{h(\beta)} d\beta \right|^2.$$

Заметим, что, даже если n есть 0 или положительное число, задача предсказания при наличии «помехи» $f_2(\alpha)$, которую мы назовем *задачей фильтрации*, оказывается не простой. Причина этого заключается в том, что функция $f_1(T^n\alpha)$ не представима, вообще говоря, в терминах настоящего и прошлого функции f , а включает также и ее будущее.

Пусть теперь $g(\alpha)$ есть проекция функции $f_1(T^n\alpha)$ на настоящее и прошлое функции f , или в аналитическом представлении

$$g(\alpha) = \sum_{v=0}^{\infty} h(T^{-v}\alpha) \int_0^1 f_1(\beta) \overline{h(T^{-v-n}\beta)} d\beta.$$

Далее, если

$$G(z) = \sum_{v=0}^{\infty} z^v \int_0^1 f_1(\alpha) \overline{h(T^{-v-n}\alpha)} d\alpha,$$

то мы будем иметь

$$g(\alpha) = \sum_{\nu=0}^{\infty} h(T^{-\nu}\alpha) \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\theta}) e^{-j\nu\theta} d\theta.$$

Если мы обозначим

$$f(z) = \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) \overline{f(T^{-\nu-n}\alpha)} d\alpha,$$

то

$$\begin{aligned} f(z) &= \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) \overline{f(T^{-\nu-n}\alpha)} d\alpha = \\ &= \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) d\alpha \sum_{\mu=-\infty}^{\infty} \overline{h(T^{-\mu-n}\alpha)} \int_0^1 \overline{f(\beta)} h(T^{-\mu+\nu}\beta) d\beta = \\ &= \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) \overline{h(T^{-\nu-n}\alpha)} d\alpha \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 \overline{f(\beta)} h(T^{\nu}\beta) d\beta = \\ &= G(z) \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 \overline{f(\beta)} h(T^{\nu}\beta) d\beta. \end{aligned}$$

Следовательно,

$$G(z) = \frac{\sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) \overline{f(T^{-\nu-n}\alpha)} d\alpha}{\sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 \overline{f(\beta)} h(T^{\nu}\beta) d\beta},$$

причем на единичной окружности мы формально имеем

$$\left| \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f(\beta) \overline{h(T^{\nu}\beta)} d\beta \right|^2 = \sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f(\beta) \overline{f(T^{\nu}\beta)} d\beta.$$

Таким образом, определение $G(z)$, а следовательно, и определение $g(\alpha)$ сводится к изучению двух функций

$$\sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f_1(\alpha) \overline{f(T^{-\nu}\alpha)} d\alpha$$

и

$$\sum_{\nu=-\infty}^{\infty} z^{\nu} \int_0^1 f(\alpha) \overline{f(T^{\nu}\alpha)} d\alpha,$$

являющихся в некотором смысле преобразованиями Фурье взаимной корреляционной последовательности временных рядов f_1 , f и автокорреляционной последовательности ряда f . Эти корреляционные последовательности в свою очередь могут быть определены через спектры временных рядов $f_1(\alpha)$, $f(\alpha)$ и $f_2(\alpha)$, а если ряды $f_1(\alpha)$ и $f_2(\alpha)$ полностью некогерентны (т. е. некоррелированы), то и через спектры двух рядов $f_1(\alpha)$ и $f(\alpha)$. Последние соображения служат оправданием обычной интерпретации фильтрации в терминах гармонического анализа сигнала и шума.

§ 8.5. Предсказание в случае непрерывного времени. До сих пор мы рассматривали теорию предсказания, исходя из преобразования T переменной α . Однако ничто не препятствует тому, чтобы вместо $f(T^n\alpha)$ рассматривать выражение

$$T^n[f(\alpha)],$$

где $f(\alpha)$ принадлежит L^2 , а T — унитарное преобразование в L^2 . Мы используем это замечание для того, чтобы перейти от теории предсказания для дискретного времени к теории предсказания для непрерывного времени. С этой целью мы введем так называемые *функции Лагерра* $l_n(t)$, образующие ортонормированное семейство функций. При $n \geq 0$ эти функции обращаются в нуль в интервале $(-\infty, 0)$, а в интервале $(0, \infty)$ они имеют вид $L_n(t)e^{-t}$, где $L_n(t)$ — полином n -й степени, носящий название *полинома Лагерра*. При $n < 0$ мы положим

$$l_n(t) = l_{-n-1}(-t).$$

Все эти функции определяются формулой

$$l_n(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{(1-j\omega)^n}{\pi^{1/2}(1+j\omega)^{n+1}} e^{j\omega t} d\omega. \quad (8.4)$$

Пусть теперь S^λ — группа сохраняющих меру точечных преобразований отрезка $(0, 1)$ в себя, определенных для всех действительных λ и таких, что

$$S^\lambda S^\mu = S^{\lambda+\mu}.$$

Пусть T^{-1} есть такое унитарное преобразование, что оно переводит $f(\alpha)$ в $g(\alpha)$ тогда и только тогда, когда при всех n

$$\int_{-\infty}^{\infty} f(S^{-t}\alpha) l_n(t) dt = \int_{-\infty}^{\infty} g(S^{-t}\alpha) l_{n-1}(t) dt.$$

Чтобы показать, что это условие действительно определяет унитарное преобразование, мы запишем его в виде

$$g(S^{-t}\alpha) = \sum_n l_{n-1}(t) \int_{-\infty}^{\infty} l_n(\tau) f(S^\tau\alpha) d\tau.$$

С помощью равенства (8.4) последнее соотношение можно преобразовать к следующему виду:

$$g(\alpha) = -f(\alpha) + \int_0^{\infty} f(S^{-s}\alpha) e^{-s} ds,$$

представляющему собой явную форму преобразования одной функции в другую. Далее, мы имеем

$$\begin{aligned} \int_0^1 |g(\alpha)|^2 d\alpha &= \int_0^1 |f(\alpha)|^2 d\alpha - 4\operatorname{Re} \int_0^{\infty} e^{-s} ds \int_0^1 f(S^{-s}\alpha) \overline{f(\alpha)} d\alpha + \\ &+ 4 \int_0^{\infty} e^{-s} ds \int_0^{\infty} e^{-t} dt \int_0^1 f(S^{-s}\alpha) \overline{f(S^{-t}\alpha)} d\alpha. \end{aligned} \quad (8.5)$$

Положим

$$\int_0^1 f(S^{-s}\alpha) \overline{f(\alpha)} d\alpha = F(s).$$

Тогда соотношение (8.5) превратится в следующее:

$$\begin{aligned} \int_0^1 |g(\alpha)|^2 d\alpha - \int_0^1 |f(\alpha)|^2 d\alpha &= \\ &= -4\operatorname{Re} \int_0^{\infty} e^{-s} F(s) ds + 4 \int_0^{\infty} e^{-s} ds \int_0^{\infty} e^{-t} F(s-t) dt = \\ &= -4\operatorname{Re} \int_0^{\infty} e^{-s} F(s) ds + 4 \int_{-\infty}^{\infty} F(s) ds \int_{\substack{t \geq 0 \\ s+t \geq 0}} e^{-s-2t} dt. \end{aligned}$$

Далее,

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-s} F(s) ds \int_{\substack{t \geq 0 \\ s+t \geq 0}} e^{-2t} dt &= \\ &= \int_0^{\infty} e^{-s} F(s) ds \int_0^{\infty} e^{-2t} dt + \int_{-\infty}^0 e^{-s} F(s) ds \int_{-s}^{\infty} e^{-2t} dt = \\ &= \frac{1}{2} \int_0^{\infty} e^{-s} F(s) ds + \frac{1}{2} \int_0^{\infty} e^s F(-s) ds = \operatorname{Re} \int_0^{\infty} e^{-s} F(s) ds. \end{aligned}$$

Следовательно,

$$\int_0^1 |f(\alpha)|^2 d\alpha = \int_0^1 |g(\alpha)|^2 d\alpha.$$

Этим доказано, что наше преобразование унитарное.

Применим теперь теорию предсказания для случая дискретного времени к последовательности $\{T^n f(\alpha)\}$. Заметим, что настоящее и прошлое этой последовательности, т. е. замыкание множества элементов

$$\{T^n f(\alpha)\}, \quad n \geq 0$$

совпадает с замыканием множества

$$\{f(S^{\lambda}\alpha)\}, \quad \lambda > 0.$$

Пусть $h(\alpha)$ есть нормированная проекция $f(\alpha)$ на множество $\{T^{-n}f(\alpha)\}$, $n > 0$. Если проекция $f(\alpha)$ на далекое прошлое последовательности $\{T^n f(\alpha)\}$ стремится к нулю, то мы можем доказать, что проекция $f(\alpha)$ на прошлое $f(S^{\lambda}\alpha)$ также стремится к нулю, и обратно. Функцию $h(\alpha)$ можно выразить через автокорреляционные коэффициенты

$$A_n = \int_0^1 T^n f(\alpha) \overline{f(\alpha)} d\alpha.$$

Мы уже знаем, что

$$A_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} |Q(\omega)|^2 e^{-jn\omega} d\omega,$$

где

$$Q(\omega) = \sum_{n=0}^{\infty} e^{jn\omega} \int_0^1 T^n f(\alpha) \overline{h(\alpha)} d\alpha.$$

§ 8.6. Многомерное предсказание и причинность. Установим теперь теорему, играющую основную роль в теории линейной причинности и известную под названием теоремы о перемежающихся процессах. Пусть $\{\varphi_n(x)\}$ есть некоторое множество ортонормированных функций, вообще говоря, неполное, $\{\psi_n(x)\}$ — другое множество ортонормированных функций, также неполное и $\{\chi_n(x)\}$ — ортонормированное семейство функций, замкнутое на множестве всех функций из L^2 , таких, что

$$\begin{aligned} f(x) &\sim \sum_{n=0}^{\infty} \varphi_n(x) \int_0^1 f(\xi) \overline{\varphi_n(\xi)} d\xi \sim \\ &\sim \sum_{n=0}^{\infty} \psi_n(x) \int_0^1 f(\xi) \overline{\psi_n(\xi)} d\xi. \end{aligned}$$

Если

$$\begin{aligned} f_0(x) &= f(x), \\ f_{2v+1}(x) &\sim \sum_{n=0}^{\infty} \psi_n(x) \int_0^1 f_{2v}(\xi) \overline{\psi_n(\xi)} d\xi, \\ f_{2v}(x) &\sim \sum_{n=0}^{\infty} \varphi_n(x) \int_0^1 f_{2v-1}(\xi) \overline{\varphi_n(\xi)} d\xi, \end{aligned}$$

то предел $f^*(x) = \lim_{v \rightarrow \infty} f_v(x)$ существует, принадлежит L^2 и

$$\begin{aligned} f^*(x) &\sim \sum_{n=0}^{\infty} \chi_n(x) \int_0^1 f^*(\xi) \overline{\chi_n(\xi)} d\xi \sim \\ &\sim \sum_{n=0}^{\infty} \chi_n(x) \int_0^1 f(\xi) \overline{\chi_n(\xi)} d\xi. \end{aligned}$$

Для доказательства заметим, что существует линейное преобразование T в пространстве функций φ_n , переводящее каждую функцию $f_{2v}(x)$ в $f_{2v+2}(x)$ и обладающее свойствами

$$\begin{aligned} \int |Tf(x)|^2 dx &\leq \int |f(x)|^2 dx, \\ \int g(x) T\overline{h(x)} dx &= \int \overline{h(x)} Tg(x) dx. \end{aligned}$$

Последнее свойство вытекает из того, что

$$\begin{aligned} f_{2v+2}(x) &\sim \sum_{n=0}^{\infty} \varphi_n(x) \int \overline{\varphi_n(\xi)} \sum_m \psi_m(\xi) d\xi \int f_{2v}(\eta) \overline{\psi_m(\eta)} d\eta \sim \\ &\sim \int f_{2v}(\eta) \left[\sum_{m,n} \overline{\varphi_n(\eta)} \psi_m(x) \int \varphi_n(\xi) \overline{\psi_m(\xi)} d\xi \right] d\eta. \end{aligned}$$

Таким образом, T есть ограниченное (с нормой, меньшей или равной 1) и самосопряженное преобразование. Следовательно, по теореме Хеллингера и Вейля (см. [8]), коль скоро $F(x)$ принадлежит замыканию функций φ_n , то найдется такая связанная с $f(x)$ функция $\Lambda_f(u)$ с ограниченной полной вариацией в интервале $(-\infty, \infty)$, что

$$d\Lambda_{Tf}(u) = F(u) d\Lambda_f(u), \quad 0 \leq F(u) \leq 1,$$

$$\int_{-\infty}^{\infty} d\Lambda_f(u) = \int |f(x)|^2 dx.$$

Легко показать, что интеграл

$$\begin{aligned} \int_0^1 |T^{4n}f(x) - T^{4m}f(x)|^2 dx &= \\ &= \int_0^1 |T^{4n}f(x)|^2 dx + \int_0^1 |T^{4m}f(x)|^2 dx - 2 \int_0^1 |T^{2m+2n}f(x)|^2 dx = \\ &= \int_{-\infty}^{\infty} \{[F(u)]^{4n} + [F(u)]^{4m} - 2[F(u)]^{2m+2n}\} d\Lambda_f(u) \end{aligned}$$

стремится к нулю как при $F(u) = 1$, так и при $F(u) < 1$. Таким образом, по теореме Рисса — Фишера предел

$$f^*(x) = \lim_{v \rightarrow \infty} \text{i. m. } f_{4v}(x)$$

существует. Аналогично, существуют все пределы

$$\lim_{v \rightarrow \infty} \text{i. m. } f_{4v+1}(x), \quad \lim_{v \rightarrow \infty} \text{i. m. } f_{4v+2}(x), \quad \lim_{v \rightarrow \infty} \text{i. m. } f_{4v+3}(x),$$

а так как они представляют собой последовательные проекции одной и той же функции $f^*(x)$ и в свою очередь $f^*(x)$ есть проекция каждой из них, то все они эквивалентны. Тем самым мы доказали, что

$$f^*(x) = \lim_{v \rightarrow \infty} \text{i. m. } f_v(x).$$

Функция $f^*(x)$ принадлежит как замыканию функций φ_n , так и замыканию функций ψ_n . Следовательно, она принадлежит также замыканию функций χ_n и

$$f^*(x) = \sum_{n=0}^{\infty} \chi_n(x) \int_0^1 f^*(\xi) \overline{\chi_n(\xi)} d\xi.$$

Кроме того,

$$\int_0^1 f^*(\xi) \overline{\chi_n(\xi)} d\xi = \lim_{v \rightarrow \infty} \int_0^1 f_{4v}(\xi) \overline{\chi_n(\xi)} d\xi.$$

Однако легко показать, что

$$\int_0^1 f_{4v+4}(\xi) \overline{\chi_n(\xi)} d\xi = \int_0^1 f_{4v}(\xi) \overline{\chi_n(\xi)} d\xi,$$

откуда мы получаем

$$f^*(x) \sim \sum_{n=0}^{\infty} \chi_n(x) \int_0^1 f(\xi) \overline{\chi_n(\xi)} d\xi.$$

Рассмотрим теперь частный случай, представляющий интерес для теории причинности, а именно тот случай, когда ортонормированные

семейства $\{\varphi_n(x)\}$ и $\{\psi_n(x)\}$ суть семейства $h_1(T^{-n}\alpha)$ и $h_2(T^{-n}\alpha)$, $n > 0$, где h_1 и h_2 связаны с f_1 и f_2 так же, как в § 8.3 была связана функция h с f . Мы имеем

$$f_1(\alpha) \sim \sum_{n=0}^{\infty} h_1(T^{-n}\alpha) \int_0^1 \overline{h_1(T^{-n}\beta)} f_1(\beta) d\beta,$$

$$f_2(\alpha) \sim \sum_{n=0}^{\infty} h_2(T^{-n}\alpha) \int_0^1 \overline{h_2(T^{-n}\beta)} f_2(\beta) d\beta.$$

Нас интересует часть функции $h_1(\alpha)$, ортогональная как к $h_1(T^{-\nu}\alpha)$, так и к $h_2(T^{-\nu}\alpha)$, $\nu > 0$. Для того чтобы ее найти, надо сначала рассмотреть часть функции $h_1(\alpha)$, ортогональную к прошлому ряда $h_2(\alpha)$, т. е. величину

$$h_1(\alpha) - \sum_{m=1}^{\infty} h_2(T^{-m}\alpha) \int_0^1 h_1(\beta) \overline{h_2(T^{-m}\beta)} d\beta,$$

а затем уже перейти к части этого последнего выражения, ортогональной к прошлому ряда $h_1(\alpha)$, т. е. к величине

$$h_1(\alpha) - \sum_{m=1}^{\infty} h_2(T^{-m}\alpha) \int_0^1 h_1(\beta) \overline{h_2(T^{-m}\beta)} d\beta +$$

$$+ \sum_{m=1}^{\infty} h_1(T^{-m}\alpha) \sum_{n=1}^{\infty} \int_0^1 h_2(T^{-n}\beta) \overline{h_1(T^{-m}\beta)} d\beta \int_0^1 h_1(\beta) \overline{h_2(T^{-n}\beta)} d\beta.$$

Соответственно этому часть функции $h_1(\alpha)$, ортогональная к прошлому $h_1(\alpha)$ и к прошлому $h_2(\alpha)$, будет равна

$$k_1(\alpha) = \lim_{\nu \rightarrow \infty} \left[h_1(\alpha) - \sum_{m_1=1}^{\infty} h_2(T^{-m_1}\alpha) \int_0^1 h_1(\beta) \overline{h_2(T^{-m_1}\beta)} d\beta + \right.$$

$$+ \sum_{m_1=1}^{\infty} h_1(T^{-m_1}\alpha) \sum_{m_2=1}^{\infty} \int_0^1 h_2(T^{-m_2}\beta) \overline{h_1(T^{-m_1}\beta)} d\beta \int_0^1 h_1(\beta) \overline{h_2(T^{-m_2}\beta)} d\beta - \dots$$

$$\dots + (-1)^{\nu} \sum_{m_1=1}^{\infty} \sum_{m_2=1}^{\infty} \dots \sum_{m_{\nu}=1}^{\infty} h_{\frac{3}{2}-\frac{1}{2}}(T^{m_{\nu}}\alpha) \int_0^1 h_1(\beta) \overline{h_2(T^{-m_1}\beta)} d\beta \times$$

$$\times \int_0^1 h_2(T^{-m_1}\beta) \overline{h_1(T^{-m_2}\beta)} d\beta \times \dots$$

$$\left. \dots \times \int_0^1 h_{\frac{3}{2}-\frac{1}{2}}(T^{m_{\nu}}\beta) h_{\frac{3}{2}+\frac{1}{2}}(T^{m_{\nu}-1}\beta) d\beta \right].$$

Определим далее тем же способом $k_2(\alpha)$, отличающееся от $k_1(\alpha)$ заменой индексов 1 на 2 и 2 на 1. Мы найдем, что $k_1(\alpha)$ и $k_2(\alpha)$ ортогональны к $k_1(T^{-\nu}\alpha)$ и $k_2(T^{-\nu}\alpha)$, $\nu > 0$; следовательно,

$$\int_0^1 k_i(T^m\alpha) \overline{k_j(T^n\alpha)} d\alpha = 0 \quad \text{при } m \neq n.$$

Рассмотрим теперь матрицу $\|A_{ij}\|$ такую, что

$$\|A_{ij}\| \cdot \left\| \int_0^1 k_i(\alpha) \overline{k_j(\alpha)} d\alpha \right\| \cdot \|A_{ji}\| = I.$$

Такая матрица существует всегда, за исключением случая

$$\left| \int_0^1 k_i(\alpha) \overline{k_j(\alpha)} d\alpha \right| = 0,$$

когда имеется нетривиальное линейное соотношение между прошлым и будущим $f_1(\alpha)$ и прошлым и будущим $f_2(\alpha)$. Во всех других случаях, если мы положим

$$g_j(\alpha) = \sum_i A_{ij} k_i(\alpha),$$

то система функций $\{g_i(T^m\alpha)\}$ будет ортонормированной в том смысле, что

$$\int_0^1 g_i(T^m\alpha) \overline{g_j(T^n\alpha)} d\alpha = \delta_{ij} \delta_{mn},$$

где δ_{ij} , δ_{mn} — символы Кронекера.

Предсказание значения любой функции $F(\alpha)$ в терминах прошлого рядов $f_1(\alpha)$ и $f_2(\alpha)$ дается формулой¹⁾

$$\sum_{i=1,2} \sum_{n=0}^{\infty} g_i(T^{-n}\alpha) \int_0^1 F(\alpha) \overline{g_i(T^{-n}\alpha)} d\alpha,$$

¹⁾ Это — формальное разложение F по функциям g . Чтобы обосновать это разложение, необходимо указать условия, обеспечивающие принадлежность F к замыканию функций g . Тот случай, когда F совпадает с f_1 или f_2 , рассмотрен в моей и П. Масани совместной статье, которая должна появиться в журнале «The Annals of Mathematical Statistics». Эта статья содержит также исправление некоторых серьезных ошибок, имевшихся в моей более ранней работе. — Н. В. [Работа Винера и Масани на самом деле была опубликована в журнале «Acta Mathematica», 98 (1957), № 1—2 и 99 (1958), № 1—2. — Прим. ред.]

вполне аналогичной формуле (8.2). Среднеквадратичная ошибка этого предсказания будет равна

$$\sum_{i=1,2} \sum_{n=-\infty}^{-1} \left| \int_0^{\infty} F(\alpha) \overline{g_i(T^{-n}\alpha)} d\alpha \right|^2,$$

что аналогично выражению, которое дается уравнением (8.3). Исходя из этих формул, можно построить всю теорию двумерного предсказания в форме, аналогичной теории простого предсказания. Более того, можно распространить эту теорию также на случай n -мерного предсказания. В частности, при надлежащих условиях спектральную матрицу, задаваемую формальными соотношениями

$$Q_{ij}(\omega) = \sum_n e^{jn\omega} \int_0^1 f_i(T^n\alpha) \overline{f_j(\alpha)} d\alpha,$$

можно будет записать в виде

$$Q_{ij}(\omega) = \sum_k M_{ik}(\omega) \overline{M_{kj}(\omega)},$$

где $M_{ik}(\omega)$ есть предел в среднем аналитической функции $\Psi_{ik}(re^{j\omega})$, не имеющей особых точек при $r < 1$ и такой, что $|\Psi_{ij}(z)|$ не имеет нулей при $|z| < 1$. Подобное представление будет возможно тогда и только тогда, когда

$$\int_{-\pi}^{\pi} \log \begin{vmatrix} Q_{11}(\omega) & Q_{21}(\omega) \\ Q_{12}(\omega) & Q_{22}(\omega) \end{vmatrix} d\omega < \infty.$$

Все эти факты вместе с отображением единичного круга на полуплоскость, которое мы уже обсуждали, могут быть использованы для развития теории системы интегральных уравнений

$$\sum_{i=1,2} \int_0^1 f_i(x) K_{ij}(y-x) dx = g_j(y), \quad j=1, 2.$$

Возвращаясь к $h_1(\alpha)$ и $k_1(\alpha)$, отметим, что

$$\int_0^1 |k_1(\alpha)|^2 d\alpha$$

представляет собой ту нормированную часть функции $f_1(\alpha)$, которая остается непредсказуемой, если мы имеем всю информацию о прошлом как функции $f_2(\alpha)$, так и функции $f_1(\alpha)$. Мы имеем

$$\begin{aligned} \int_0^1 |k_1(\alpha)|^2 d\alpha &= \int_0^1 k_1(\alpha) \overline{h_1(\alpha)} d\alpha = 1 - \sum_{m=1}^{\infty} \left| \int_0^1 h_1(\beta) \overline{h_2(T^{-m}\beta)} d\beta \right|^2 - \\ &- \sum_{m=1}^{\infty} \left| \sum_{n=1}^{\infty} \int_0^1 h_1(\beta) \overline{h_2(T^{-n}\beta)} d\beta \int_0^1 h_2(T^{-n}\beta) \overline{h_1(T^{-m}\beta)} d\beta \right|^2 - \dots \end{aligned}$$

Таким образом, сумма

$$\sum_{m=1}^{\infty} \left| \int_0^1 h_1(\beta) \overline{h_2(T^{-m}\beta)} d\beta \right|^2 + \\ + \sum_{m=1}^{\infty} \left| \sum_{n=1}^{\infty} \int_0^1 h_1(\beta) \overline{h_2(T^{-n}\beta)} d\beta \int_0^1 h_2(T^{-n}\beta) \overline{h_1(T^{-m}\beta)} d\beta \right|^2 + \dots$$

есть неотрицательная величина, не превосходящая 1, которая измеряет дополнительную эффективность знания прошлого ряда $f_2(\alpha)$ при определении настоящего ряда $f_1(\alpha)$, и она может рассматриваться как мера C причинного воздействия $f_2(\alpha)$ на $f_1(\alpha)$.

В качестве приложения рассмотрим следующий случай. Пусть $f_1(\alpha)$ представляет собой температуру в 9 ч. утра в Бостоне, а $f_2(\alpha)$ — температуру в Олбани в то же самое время. Вообще говоря, мы предполагаем, что погода перемещается вместе с вращением Земли, с запада на восток; две величины, одна 1 — C и другая, получаемая из нее при помощи замены индексов 1 и 2 друг на друга, позволяют нам сформулировать точное утверждение, использующее указанное предположение, и затем проверить, верно ли это утверждение или нет. Другое приложение связано с изучением мозговых излучений; здесь мы можем получить электроэнцефалограммы, более или менее соответствующие электрической активности в различных частях мозга. В этом примере изучение коэффициентов причинности, соответствующих двум функциям $f(\alpha)$, взятым в том и в другом порядке, и их аналогов для более чем двух функций f может быть полезно при определении того, какая часть мозга обуславливает состояние нормальной активности в другой части мозга и в какой именно. Такое определение очень важно, так как указанный метод не вносит искусственных стимулов в различные части мозга. Опасность введения таких стимулов состоит в том, что благодаря своей интенсивности и резкости они могут проложить новые пути в мозгу, которые не используются при его нормальной активности и которые могут рассматриваться как в некотором смысле искусственные.

ЛИТЕРАТУРА

1. Birkhoff G. D., Proof of the Ergodic Theorem, Proc. Nat. Acad. Sci. USA, 17, 656—660 (1931).
2. Крамер Г., Математические методы статистики, ИЛ, М., 1948.
3. Дуб Дж. Л., Вероятностные процессы, ИЛ, М., 1956.
4. Колмогоров А. Н., Основные понятия теории вероятностей, Гостехиздат, М. — Л., 1936.
5. Koopman B. O., Hamiltonian Systems and Transformations in Hilbert Space, Proc. Nat. Acad. Sci. USA, 17, 315—318 (1931).

6. Lévy P., *Théorie de l'addition des variables aléatoires*, Gauthier-Villars, Paris, 1937.
7. Von Neumann J., *Proof of the Quasi-ergodic Hypothesis*, Proc. Nat. Acad. Sci. USA, **18**, 70—82 (1932).
8. Weyl H., *Über die Konvergenz von Reihen, die nach Orthogonalfunktionen fortschritten*, Math. Ann., **67**, 225—245 (1909).
9. Wiener N., *Generalized Harmonic Analysis*, Acta Math., **55**, 117—258 (1930).

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

- Колмогоров А. Н., Интерполирование и экстраполирование стационарных случайных последовательностей, Изв. АН СССР, сер. матем., **5**, № 1, 3—14 (1941).
- Яглом А. М., Введение в теорию стационарных случайных функций, Успехи мат. наук, **7**, № 5 (51), 3—168 (1952).
- Солодовников В. В., Введение в статистическую динамику систем автоматического управления, Гостехиздат, М., 1952.
- Колмогоров А. Н., Упрощенное доказательство эргодической теоремы Биркгоффа — Хинчина, Успехи мат. наук, вып. 5, 52—56 (1938).
- Гнеденко Б. В., Курс теории вероятностей, изд. 2, Гостехиздат, М., 1954, § 57.

Глава 9

ТЕОРИЯ ИГР

Г. ФРЕДЕРИК БОНЕНБЛАСТ

Калифорнийский технологический институт

§ 9.1. Введение. Хотя теория игр и применялась в разных областях, она является в основном теорией планирования и связана с анализом оптимальных направлений действия. Теория предполагает ясно очерченную общую картину игры. Точнее, предполагаются известными возможные направления действия, их последствия и цель каждого из участников игры. Отправляясь от этих общих предпосылок, теория пытается объективно анализировать положение, которое возникает из столкновения интересов участников. Она принимает во внимание тот факт, что в момент принятия игроком решения о направлении действия прошлые и будущие действия других участников могут быть ему известны только частично. Теория старается определить наилучшее направление действия для каждого участника отдельно и оценить размер выигрыша, на который он может надеяться. Пока еще не существует общей окончательной теории, которая могла бы ответить на эти вопросы в столь широкой форме. Возникающие здесь трудности носят как технический, так и принципиальный характер. Вполне понятно основное значение случая игр с двумя противниками, чьи интересы прямо противоположны. Более сложные случаи также изучались как непосредственно, так и путем сведения к случаю двух противников в предположении, что участники могут соединяться в группы.

Случаи, в которых может быть применена теория игр, встречаются в исследовании операций, стратегическом планировании и в тактических задачах. Теория игр весьма важна в эконометрике (математической экономике), в частности, она тесно связана с теорией линейного планирования. Наконец, как непосредственно по результатам, так и косвенно по методам теория игр связана с другими разделами математики, в наибольшей степени со статистикой. Такие игры, как шахматы, покер и бридж, представляют собой блестящие примеры для теории игр по крайней мере в идейном отношении. С технической стороны они слишком сложны, чтобы можно было полностью их анализировать. Правила игры перечисляют возможные ходы и обуславливают величину выигрыша или проигрыша каждого игрока в конце игры. Предполагается, естественно, что цель каждого игрока состоит в получении максимальной суммы выигрыша.

Теория вероятностей зародилась частично из попытки Паскаля анализировать игры, в которых игроки не имеют контроля над вы-

бором действий. Все ходы являются случайными событиями, вероятности осуществления которых известны. Анализ таких игр основывается на понятии математического ожидания. Если $p_1, p_2, \dots, p_n$ суть вероятности n событий, одно из которых обязательно происходит, и $a_1, a_2, \dots, a_n$ — соответствующие суммы выигрыша игрока в случае их осуществления, то сумма

$$a = p_1 a_1 + \dots + p_n a_n$$

называется *математическим ожиданием* игрока. Оно представляет собой предел среднего выигрыша игрока при бесконечном повторении той же ситуации; его можно также интерпретировать как наилучший размер ставки, которую нужно делать перед осуществлением события. Возможность случайных ходов теорией игр не исключается. Фактически они играют важную роль, и поэтому понятие математического ожидания имеет в теории игр основное значение. Однако игры, в которых имеют место только чисто случайные события, представляют для нас мало интереса. Игроки при этом не принимают решений, способных повлиять на ход игры, и потому возможный конфликт их интересов несущественен.

В начале 30-х годов Дж. фон Нейман заинтересовался теорией более общих случаев игр. В фундаментальной работе он заложил основы теории и получил главный результат. Позднее совместно с О. Morgenштерном он написал более исчерпывающую монографию [11], где, в частности, рассматривается вопрос об использовании теории игр в экономике. В течение последнего десятилетия в этой области появилась обширная литература. Результаты, полученные Нейманом, нашли свое применение в широком классе случаев. Теория оказалась приложимой ко многим специальным задачам, обнаружились интересные связи с другими областями математики. Наконец, были получены более эффективные методы решения задач теории игр. В последнее время появилось несколько книг, список которых приводится в конце настоящей главы [1, 6, 7, 8, 12].

§ 9.2. Предмет теории. Простейшая схема ситуаций, рассматриваемых в теории игр, обманчиво проста. В игре участвуют два человека, интересы которых строго противоположны. Представьте себе, что в каждом квадрате сетки прямоугольной системы координат или на каждом поле шахматной доски любых размеров написано число. Один игрок может выбрать строку, другой — столбец. Результат этих двух выборов определяется числом, находящимся в месте пересечения выбранных строки и столбца. Предполагается, что каждый игрок производит свой выбор, не зная решения противника. Задачей первого игрока является получение максимального числа, а задачей его противника — получение минимального числа. Каковы

те руководящие принципы, которые должны управлять решениями, и каков предполагаемый результат?

На этом примере видны важные отличительные черты теории игр. Она предполагает, что интерес каждого участника может быть количественно измерен. Фактически при этом еще предполагается, что количественное выражение интереса определяется математическим ожиданием. Точный смысл и причина этого предположения станут понятными позднее, когда будут рассматриваться оптимальные смешанные стратегии.

Для простейшей схемы характерны следующие два обстоятельства. Во-первых, каждый игрок принимает *одно* решение *совершенно независимо* от действия другого, и, во-вторых, число возможных выборов, т. е. число пересечений строк и столбцов, является *конечным*. Первое обстоятельство является только кажущимся упрощением задачи, поскольку более сложные ситуации могут быть сведены к этому случаю. Подробное описание этих случаев, однако, слишком громоздко, чтобы давать его здесь. Последний из приведенных примеров в § 9.10 был выбран, в частности, для иллюстрации методов, применяемых при таком сведении. По этому вопросу в том же § 9.10 приводятся дополнительные замечания. Второе обстоятельство более существенно. Когда число возможных действий является бесконечным, появляются дополнительные трудности, которые в некоторых случаях могут оказаться решающими. Непосредственное обобщение простейшей схемы дает простой пример бесконечного случая. Предположим, что дана функция $f(x, y)$ двух независимых переменных x и y . Для определенности будем считать, что x и y изменяются в единичном интервале: $0 \leq x \leq 1$, $0 \leq y \leq 1$. Первый игрок выбирает величину x , а второй — величину y . Первый добивается получения максимума функции $f(x, y)$, тогда как второй — минимума. Другие примеры легко получить, изменяя природу независимых переменных. Более сложные ситуации получаются, например, если x заменить точкой на плоскости, вынужденной, может быть, лежать в некоторой области плоскости, т. е. заменить переменную x двумя независимыми переменными x_1 и x_2 , связанными, возможно, некоторыми условиями. С другой стороны, если x и y считаются целыми числами со значениями соответственно из $1, 2, \dots, m$ и $1, 2, \dots, n$, то получается первоначальная простейшая схема. Методы и основные результаты теории будут изложены в терминах, относящихся к обобщенной простейшей схеме, которая считается полностью описанной, как только определены функция $f(x, y)$ и характер и пределы изменения независимых переменных x и y .

Более общий случай, пока еще полностью не разобранный, когда участвует больше двух игроков или когда интересы двух участников не являются прямо противоположными, может быть описан с помощью аналогичных простейших схем. Предполагается, что действия состоящих игроков определяются независимыми переменными

$x, y, \dots, z$. Принимается, что интересы участников измеряются соответственно известными функциями $f(x, y, \dots, z), g(x, y, \dots, z), \dots, h(x, y, \dots, z)$. Цель каждого — добиться максимального значения относящейся к нему функции. В частности, первоначальная простейшая схема отвечает специальному случаю двух игроков, в котором сумма соответствующих функций $f(x, y) + g(x, y)$ равна нулю для всех значений x и y . По этой причине этот случай известен как суммарно нулевая игра двух игроков¹⁾.

§ 9.3. Оптимальные чистые стратегии. Рассмотрим суммарно нулевую игру двух игроков, описываемую функцией $f(x, y)$, причем для простоты будем считать x и y целыми числами: $1 \leq x \leq m, 1 \leq y \leq n$. Для каждого выбранного x минимальное значение, принимаемое функцией $f(x, y)$ при изменении y от 1 до n , будет обозначаться через

$$\min_y f(x, y).$$

Другими словами, это значение является минимальным числом, которое появляется в x -ом ряду шахматной доски. Оно зависит только от x . Если теперь мы заставим x изменяться от 1 до m , то можно образовать величину

$$v_1 = \max_x [\min_y f(x, y)].$$

Пусть x_0 является значением x , при котором достигается этот максимум, т. е. пусть

$$v_1 = \min_y f(x_0, y),$$

тогда $v_1 \leq f(x_0, y)$ независимо от величины y . Другими словами, существует такая стратегия для первого игрока (а именно выбор $x = x_0$), что значение функции $f(x, y)$ будет не меньше v_1 независимо от действий его противника. Первый игрок может даже сообщить противнику о своем намерении выбрать x_0 и не заботиться, поверил ли тот ему или нет. Он знает, что при всех обстоятельствах значение $f(x, y)$ будет по крайней мере равно v_1 .

Если наши игроки поменяются ролями, то будет получена величина

$$v_2 = \min_y [\max_x f(x, y)].$$

В свою очередь для второго игрока существует такая стратегия, при которой значение $f(x, y)$ не меньше v_2 , что бы ни делал первый

¹⁾ См. примечание редактора в гл. 10, стр. 269. — *Прим. ред.*

игрок. Из этого определения чисел v_1 и v_2 следует, конечно, неравенство

$$v_1 \leq v_2.$$

Примеры показывают, что могут встретиться оба случая: $v_1 = v_2$ и $v_1 < v_2$. Различие между этими двумя случаями весьма существенно. В первом случае, когда $v_1 = v_2$, ситуация ясная и простая. Первый игрок придерживается стратегии, которая дает ему по крайней мере равные значения v_1 и v_2 ; кроме того, он знает, что его противник может воспрепятствовать достижению им более высокого значения функции $f(x, y)$. Аналогичное замечание может быть сделано относительно второго игрока. Значение $v_1 = v_2$ является компромиссным значением функции $f(x, y)$, стратегия x_0 является оптимальной стратегией для первого игрока и соответственно второй игрок имеет оптимальную стратегию y_0 . Следует обратить внимание на тот факт, что это значение получено не как результат сделки или какого-то другого соглашения между двумя игроками. Каждый действует индивидуально, не опасаясь, что его противник может получить перед ним преимущество, узнав его намерение.

Во втором случае, когда $v_1 < v_2$, остается некоторый промежуток, и достижение определенных компромиссных значений между v_1 и v_2 является главным моментом в суммарно нулевой игре. Решение этого вопроса не очевидно и потребует введения понятия так называемых *смешанных стратегий*. Во избежание путаницы мы будем называть стратегии, рассмотренные выше, в которых фиксировались определенные значения x (или y), *чистыми стратегиями*. Случай смешанных стратегий встретится в примере, рассматриваемом в следующем параграфе. Общий обзор теории дан в § 9.5.

Рассуждения этого параграфа проводились в предположении конечности области изменения x и y , т. е. для того случая, когда число фактически возможных выборов обоих игроков конечно. В случае бесконечных областей максимум и минимум рассматриваемых функций, вообще говоря, могут ими не достигаться, и вместо них приходится вводить понятия (точной) верхней (supremum) и (точной) нижней (infimum) границ. Это обстоятельство является основной причиной появления трудностей, часто встречающихся в этих случаях.

§ 9.4. Пример из теории планирования. В этом примере [4] функция $f(x, y)$ принимается равной сумме

$$f(x, y) = \sum_{i=1}^n k_i \max(x_i - y_i, 0).$$

Независимая переменная x здесь представляет собой совокупность n неотрицательных чисел $(x_1, x_2, \dots, x_n)$, сумма которых равна заданному числу A . Аналогично, независимая переменная y представляет

собой совокупность n неотрицательных чисел $(y_1, y_2, \dots, y_n)$ с суммой D . Постоянные $k_1, k_2, \dots, k_n$ являются заданными фиксированными положительными числами, которые, не уменьшая общности, можно считать невозрастающими: $k_1 \geq k_2 \geq k_3 \geq \dots \geq k_n$. Заданные числа A и D удовлетворяют неравенству $D \leq A$.

Этот пример имеет очевидные интерпретации в экономике или военно-стратегическом планировании. По существу, это задача атаки и обороны. Число A определяет все силы атакующего, а D — все силы обороны. Эти силы измеряются в любых подходящих единицах, выбираемых так, что одно соединение каждой из группировок нейтрализует одно соединение другой. Атакующие и обороняющиеся имеют n объектов, относительная важность или уязвимость которых измеряется постоянными $k_1, \dots, k_n$. Обе стороны должны решить вопрос о размещении своих сил по различным объектам. Число x_i представляет собой величину подразделения, атакующего i -й объект, а y_i — величину подразделения, на которое возложена задача обороны того же объекта. Величина $\max(x_i - y_i, 0)$, значение которой является большим из двух чисел $x_i - y_i$ и 0, представляет собой величину той части подразделения x_i , которая в состоянии проникнуть сквозь оборону к объекту. В этом примере считается, что величина $k_i \max(x_i - y_i, 0)$ количественно измеряет успех атаки на i -й объект. Вопрос о том, является ли нет данная частная оценка успеха атаки реалистичной, к теории игр уже не имеет отношения. В зависимости от характера обстановки может оказаться более подходящей какая-нибудь другая функция от x_i и y_i , что в результате приведет к другому примеру.

В качестве иллюстрации к рассуждениям предыдущего параграфа найдем значения v_1, v_2 для числового примера, полагая $n=3$, $A=16$, $D=12$, $k_1=6$, $k_2=2$, $k_3=1$. Начнем с v_2 . Прежде всего совершенно ясно, что против известного расположения сил обороны (y_1, y_2, y_3) атака будет наиболее успешной, если все силы будут направлены на одну из целей. Следовательно, для заданного расположения (y_1, y_2, y_3) величина максимума функции $f(x, y)$ по x будет равна наибольшему из трех чисел

$$6(A - y_1), \quad 2(A - y_2), \quad (A - y_3).$$

Для частного случая сил $y_1=11$, $y_2=1$, $y_3=0$ эта величина становится равной 30. Ни одна другая комбинация не уменьшит это число, так как из неравенств $6(16 - y_1) \leq 30$, $2(16 - y_2) \leq 30$ следует, что $y_1 \geq 11$, $y_2 \geq 1$, а поэтому $y_1=11$, $y_2=1$, так как $y_1 + y_2 + y_3 = 12$. Следовательно, величина v_2 равна 30, и чистая стратегия, которая разместила 11 соединений на защиту первой цели, одно соединение на защиту второй цели, оставив третью цель без защиты, гарантирует, что успех атаки, измеряемый величиной $f(x, y)$, никогда не превысит 30.

Обратимся теперь к величине v_1 . Лучшее, что может сделать оборона против известной атаки (x_1, x_2, x_3) , — это расположить силы, насколько позволяют возможности, на защиту различных целей в соответствии с их важностью или уязвимостью. Отсюда

$$\min_y f(x, y) = \begin{cases} 6(x_1 - 12) + 2x_2 + x_3, & \left\{ \begin{array}{l} x_1 \geq 12, \\ x_1 < 12 \text{ и } x_1 + x_2 \geq 12, \\ x_1 + x_2 < 12. \end{array} \right. \\ 2[x_2 - (12 - x_1)] + x_3, \\ x_3 - (12 - x_1 - x_2), \end{cases}$$

Нетрудно видеть, что любая из этих величин ≤ 24 , а число 24 достигается для стратегии $x_1 = 12, x_2 = x_3 = 0$. Другими словами, в то время как $v_2 = 30, v_1 = 24$.

Чтобы продолжить анализ дальше, рассмотрим смешанную стратегию. Такая стратегия представляет собой не точный выбор одной из чистых стратегий, а лишь набор вероятностей, с которыми они могут быть использованы. Заметим, что та смешанная стратегия, которая задает какую-нибудь чистую стратегию с вероятностью 1, а все остальные стратегии с вероятностью 0, конечно, предписывает пользоваться исключительно этой чистой стратегией. Это замечание показывает, что понятие смешанной стратегии является расширением понятия чистой стратегии. Иначе говоря, каждая чистая стратегия может быть представлена в виде смешанной стратегии. Для оценки эффекта, производимого смешанной стратегией, определяется математическое ожидание эффекта, даваемого каждой из чистых стратегий. В рассматриваемом примере особую важность имеет та смешанная стратегия атакующего, которая предписывает с вероятностью $1/4$ всеми силами атаковать первый объект, с вероятностью $3/4$ — второй объект и которая игнорирует все остальные направления атаки. Атака всеми силами на первый объект против расположения (y_1, y_2, y_3) обороны делает функцию $f(x, y)$ равной $6(A - y_1)$, а такая же атака второго объекта делает ее равной $2(A - y_2)$. Математическое ожидание функции $f(x, y)$ для смешанной стратегии равно $3/2(2A - y_1 - y_2)$, что в нашем случае всегда не меньше 30. Использование этой смешанной стратегии покрыло промежуток между v_1 и v_2 . Величина 30 представляет собой оценку успеха атаки. Оборона может использовать, как это случается, чистую стратегию, но атака должна пользоваться смешанной стратегией.

Излишне искать преимущества, получаемые обороной, от использования смешанных стратегий. Как только было установлено, что атака может обеспечить функции f значение, по крайней мере равное 30, ни одна смешанная стратегия обороны не даст лучших результатов, чем дает оптимальная чистая стратегия, которая сама по себе не позволяет атаке достигнуть значений, больших 30. Нетрудно дать формальное доказательство этого утверждения.

§ 9.5. Смешанные стратегии. Когда игрок принимает решение, руководствуясь чистой стратегией, то из всех вариантов, имеющих в его распоряжении, он выбирает один, который и будет использован. Если же он действует, исходя из смешанной стратегии, то вычисляет вероятность выбора каждого из возможных решений. Дальнейший выбор решений проводится случайно в соответствии с заданными вероятностями. Использование смешанных стратегий имеет два преимущества. Во-первых, достигается большая гибкость, так как каждая чистая стратегия может быть получена из смешанной, если соответствующую ей вероятность выбрать равной 1. Пример, рассмотренный в предыдущем параграфе, показывает, что этот выигрыш в свободе может оказаться существенным. Во-вторых, использование смешанных стратегий приводит к замене действительных значений функции $f(x, y)$ математическими ожиданиями. Игрок, пользующийся смешанной стратегией, больше не может быть уверен, что величина $f(x, y)$ будет равна определенному числу. Он уверен только в том, что *математическое ожидание* достигнет этого значения. Действительная величина может быть меньше. Использование смешанных стратегий или скорее подсчет математических ожиданий величины $f(x, y)$ для смешанных стратегий имеет смысл, если это математическое ожидание отражает действительные интересы участников. Теория игр предполагает это. Необходимо отметить, что во многих примерах подобная трудность не является результатом введения смешанных стратегий, а присуща самой задаче. В действительности часто случается, что значение функции $f(x, y)$ само является математическим ожиданием, усредняющим результаты некоторых случайных действий (см. § 9.6).

Математически введение смешанных стратегий приводит к изменению характера независимых переменных, входящих в задачу. Первоначальные независимые переменные x и y являлись чистыми стратегиями. Новые независимые переменные являются смешанными стратегиями. Обозначим их p и q . Если, например, чистая стратегия игрока определена целым числом, изменяющимся от 1 до m , то смешанная стратегия этого же игрока определяется последовательностью $(p_1, p_2, \dots, p_m)$ m неотрицательных чисел, сумма которых $\sum p_i$ равна 1. Переход к новым независимым переменным приводит, конечно, к изменению функции. Вместо функции $f(x, y)$ вводится новая функция $F(p, q)$. Ее значение для пары (p, q) является математическим ожиданием функции $f(x, y)$, соответствующей наборам вероятностей смешанных стратегий p и q . В частности, если оба игрока имеют конечное число стратегий, описываемых для первого игрока целыми числами x (от 1 до m), а для второго игрока целыми числами y (от 1 до n), то функция $F(p, q)$ является суммой mn слагаемых:

$$F(p, q) = f(1, 1)p_1q_1 + f(1, 2)p_1q_2 + \dots + f(m, n)p_mq_n.$$

Основной результат теории для суммарно нулевой игры в случае двух игроков состоит в том, что для обширного класса ситуаций эта новая функция удовлетворяет условию

$$\max_p [\min_q F(p, q)] = \min_q [\max_p F(p, q)].$$

К этому классу задач относится, в частности, тот случай, когда число стратегий обоих игроков является конечным. В терминах § 9.3 приведенная выше зависимость $\max\text{-}\min = \min\text{-}\max$ выражает тот факт, что для смешанных стратегий соотношение $v_1 = v_2$ сохраняется, и имеет, конечно, те же последствия, что и в случае чистых стратегий. Это общее значение является выражением того компромиссного решения, которое не могло быть определено в § 9.3, где использовались исключительно чистые стратегии. Каждый игрок имеет свою смешанную стратегию, обеспечивающую его интересы на уровне $v_1 = v_2$ независимо от действий его противника (такая стратегия может быть, в частности, чистой, как это было в случае обороны в примере предыдущего параграфа). Величина $v = v_1 = v_2$ называется *стоимостью* игры.

Простое следствие определения оптимальной смешанной стратегии приводит к общему принципу, имеющему важное значение при расчетах. Этот принцип будет использован в следующем параграфе при рассмотрении примера по тактическому выбору момента времени. Пусть p' и q' являются оптимальными стратегиями. По определению, величина $F(p', u)$ при всех u и величина $F(x, q')$ при любых x будет не меньше v . Эти неравенства показывают, что

$$F(p', u) = v$$

для каждого оптимального p' и для каждой чистой стратегии u , которая с положительной вероятностью задается какой-нибудь оптимальной стратегией q' . Как только были определены чистые стратегии, существенные для оптимальных смешанных стратегий, так вместо неравенств, с которыми трудно обращаться, получаются точные равенства.

Предыдущее рассуждение в основном относится к случаю, когда число чистых стратегий конечно. В случае бесконечного числа описание смешанных стратегий значительно сложнее. В частности, если бы чистые стратегии описывались действительными числами, скажем, в интервале $0 \leq x \leq 1$, то смешанные стратегии задавались бы функциями распределения $P(x)$. Они позволяют задавать как положительные вероятности отдельных чистых стратегий, так и плотность распределения вероятностей $dP = p(x)dx$ в случае непрерывного распределения чистых стратегий. В этом случае функция $F(p, q)$, выражающая ожидаемое значение функции $f(x, y)$ в терминах смешанных стратегий, является уже не суммой, а интегралом

$$\iint f(x, y) dP(x) dQ(y).$$

Плотности вероятностей распределения будут иметь существенное значение в следующем примере.

§ 9.6. Симметрия; пример тактического выбора момента времени. Задача является *симметричной*, когда оба участника действуют совершенно одинаковым способом. Так как оба они имеют одинаковые возможности, то их индивидуальные стратегии совпадают, а значит, будут совпадать также и соответствующие им смешанные стратегии. Симметричность задачи выражается соотношением $f(x, y) = -f(y, x)$, что приводит к соотношению $F(p, q) = -F(q, p)$ для смешанных стратегий. Когда решение существует, т. е. когда $\max \min$ этой последней функции равен ее $\min \max$, это общее значение v вследствие симметрии должно быть равно 0. Поэтому стратегия p оптимальна, если $F(p, y) \geq 0$ для каждого y . Заметим также, что если p оптимальна для одного участника, то та же смешанная стратегия автоматически оптимальна и для другого. Пример, рассматриваемый ниже, является симметричным.

Из двух соперников, борющихся за один и тот же объект, каждый должен независимо выбрать время для своей попытки захватить его. Значение известной функции $\varphi(t)$ выражает вероятность захвата этого объекта каждым из соперников, если он действует в момент времени t . Предполагается, что эта функция $\varphi(t)$ возрастает от нуля до максимального значения при $t=0$ и убывает после этого наиболее благоприятного момента. Чем раньше участник решит действовать, тем больше у него шансов опередить своего соперника, но тем меньше шансов успеха имеет его попытка. В настоящем варианте этого вида задачи каждый игрок имеет только одну возможность действовать, и он не знает, сделал ли его соперник свою попытку или нет.

Чистые стратегии соперников описываются моментами времени, в которые они решают действовать. Как обычно, для обозначения этих стратегий будут выбраны независимые переменные x и y . Они являются действительными числами, заключенными в пределах $x \leq 0$, $y \leq 0$, так как ни один из соперников не имеет побудительных причин действовать позже самого благоприятного момента времени $t=0$. Смешанные стратегии являются функциями распределения $P(x)$ и $Q(y)$, но оказывается, что в данном примере достаточно ограничить наши рассуждения случаем, когда смешанные стратегии задаются своими плотностями распределения вероятностей $p(x)$ и $q(y)$. Наконец, вследствие симметрии задачи достаточно рассматривать только $p(x)$. Сначала мы приведем решение, а потом покажем, как оно получается.

Оптимальная стратегия задается плотностью вероятностей

$$p(x) = \begin{cases} 0 & \text{для } x < a, \\ \frac{\varphi(0)\varphi'(x)}{2\varphi^3(x)[1+\varphi(0)]} & \text{для } a \leq x \leq 0, \end{cases}$$

где постоянная a является отрицательным числом, определяемым из уравнения

$$\varphi(a) = \frac{\varphi(0)}{1 + 2\varphi(0)}.$$

Это решение выражает в точной количественной форме интуитивные соображения об оптимальной схеме поведения соперников. Ни один не будет действовать до тех пор, пока шансы на успех ниже определенного уровня, который определяется уравнением для величины a .

Плотность вероятностей $p(x)$, которая управляет решением каждого соперника, велика там, где функция $\varphi(x)$ быстро возрастает.

Первым шагом в выводе окончательного результата является определение функции $f(x, y)$, которая математически описывает рассматриваемую ситуацию. Если первый игрок делает попытку раньше второго, т. е. если $x < y$, то шансы этого участника овладеть объектом определяются величиной $\varphi(x)$, а шансы проиграть своему сопернику — величиной $[1 - \varphi(x)]\varphi(y)$. Таким образом, вероятность его успеха при $x < y$ измеряется величиной $\varphi(x) - \varphi(y) + \varphi(x)\varphi(y)$. Если оба участника действуют одновременно, то эта вероятность равна нулю, т. е. каждый имеет одинаковые шансы на успех. Наконец, если $x > y$, то вероятность успеха равна $\varphi(x) - \varphi(y) - \varphi(x)\varphi(y)$. Следовательно, функция $f(x, y)$ определяется формулой

$$f(x, y) = \begin{cases} \varphi(x) - \varphi(y) + \varphi(x)\varphi(y) & \text{для } \begin{cases} x < y, \\ x = y, \end{cases} \\ 0 & \\ \varphi(x) - \varphi(y) - \varphi(x)\varphi(y) & \text{для } x > y. \end{cases}$$

Смешанная стратегия $p(x)$ является оптимальной, если для каждого y

$$F(p, y) = \int_{-\infty}^0 f(x, y) p(x) dx \geq 0.$$

Разбивая промежуток интегрирования на два участка, один от $-\infty$ до y , другой от y до 0 и заменяя $f(x, y)$ выражением, выведенным выше, приходим к соотношению

$$2 \int_{-\infty}^y \varphi(x) p(x) dx \geq 1 + [1 - \varphi^{-1}(y)]C,$$

где

$$C = \int_{-\infty}^0 \varphi(x) p(x) dx.$$

Общий принцип, упоминавшийся в конце предыдущего параграфа, будучи распространен на случай непрерывных переменных, показывает, что равенство в предыдущем соотношении должно выполняться при

каждом y , при котором $p(y)$ отлично от нуля. Дифференцирование по y получающегося равенства приводит нас к альтернативе: для каждого y

$$p(y) = 0 \quad \text{или} \quad p(y) = \frac{C}{2} \varphi'(y) \varphi^{-3}(y).$$

В дополнение к этому должны быть выполнены также следующие условия:

$$\int p(x) dx = 1 \quad \text{и} \quad \int \varphi(x) p(x) dx = C.$$

Теперь естественно предположить, что плотность $p(x)$ равна нулю для значений x , не превосходящих некоторой величины a , а при $x > a$ величина $p(x)$ определяется вторым из равенств приведенной выше альтернативы. Мы должны проверить это предположение для каждого случая. Второе дополнительное соотношение определяет величину a , при которой происходит разрыв, а значение постоянной C может быть определено из первого дополнительного условия. Наконец, мы должны проверить, что решение, полученное таким путем, действительно удовлетворяет условию, что величина $F(p, y)$ не меньше нуля при всех y . Эти подробности мы опускаем.

Этот пример является частным случаем более широкого класса рассматривавшихся задач, известных под названием „дуэлей“ [5]. Задача нахождения оптимальной смешанной стратегии сводится к решению линейных интегральных уравнений. Можно показать также, что решение является единственным, т. е. для каждого участника существует только одна смешанная стратегия, гарантирующая его интересы. Это положение характерно для всех „дуэлей“, а также и для многих других случаев. Однако оно не является справедливым всегда. Может случиться, что для одного участника окажется несколько оптимальных смешанных стратегий.

§ 9.7. Соотношение $\max\text{-}\min = \min\text{-}\max$. Предыдущие рассуждения показывают, что теория суммарно нулевой игры двух игроков основывается на соотношении $\max\text{-}\min = \min\text{-}\max$. Оптимальные чистые стратегии существуют в тех исключительных случаях, когда это соотношение справедливо для функции $f(x, y)$. Вообще говоря, оно справедливо для функции $F(p, q)$ и определяет смешанные стратегии для оптимального образа действий. Было принято без доказательства, что это соотношение справедливо для функции $F(p, q)$ по крайней мере во всех случаях, когда каждый участник имеет только конечное число чистых стратегий x, y . Доказательство этого последнего утверждения, которое одновременно показывает справедливость этого соотношения для более широкого класса функций, основывается на двух особых свойствах функции $F(p, q)$. Одно касается характера смешанных стратегий p, q , а другое относится к специальной форме функции $F(p, q)$.

По определению, смешанная стратегия p в случае конечного числа чистых стратегий является последовательностью $(p_1, p_2, \dots, p_m)$ неотрицательных чисел, сумма которых равна единице. Этот факт дает простую геометрическую интерпретацию смешанных стратегий. Положив $m=3$, выберем на плоскости треугольник. Каждая точка этого треугольника, внутри или на границе, может быть представлена как центр тяжести масс, расположенных в вершинах треугольника. Так как положение центра тяжести зависит только от отношений этих масс, то сумму их можно положить равной единице. Отождествление величин p_1, p_2, p_3 с этими массами приводит к отождествлению понятия смешанной стратегии p с понятием центра тяжести. Другими словами, смешанные стратегии могут быть представлены как точки внутри или на сторонах треугольника. Чистые стратегии при этом превращаются в вершины треугольника. Для других значений m конструкция видоизменяется очевидным образом. Для $m=2$ смешанные стратегии являются точками отрезка прямой, для $m=4$ они являются точками тетраэдра, для больших значений m смешанные стратегии суть точки геометрических фигур, соответствующих тетраэдру в пространствах большего числа измерений. Все эти геометрические фигуры являются *выпуклыми*, т. е. отрезок прямой, соединяющий любые две точки этой фигуры, полностью лежит внутри нее.

Второе свойство относится к форме функции $F(p, q)$. В случае конечного числа чистых стратегий функция $F(p, q)$ определяется суммой, приведенной в § 9.5. Заметим, что эта функция является билинейной формой относительно p и q , т. е. при заданных q функция $F(p, q)$ линейна относительно всех p_i и, наоборот, при заданных p она линейна относительно всех q_j . В частности, если p' и p'' — две любые смешанные стратегии, то

$$F\left(\frac{p' + p''}{2}, q\right) = \frac{F(p', q) + F(p'', q)}{2}.$$

Если в приведенном выше соотношении знак равенства заменить знаком $\leq$, то получается более широкий класс функций, которые называются *выпуклыми* по p . Аналогично, если знак равенства заменить знаком $\geq$, то получаются *вогнутые* функции.

Основной результат теории состоит в том, что если область изменения независимых переменных p и q является выпуклой фигурой в конечномерном пространстве и функция $F(p, q)$ вогнута по p при каждом q и выпукла по q при каждом p , то

$$\max_p \left[\min_q F(p, q) \right] = \min_q \left[\max_p F(p, q) \right].$$

(Предполагается, что выпуклые фигуры, представляющие области изменения p и q , включают граничные точки и заключены в конечную часть пространства, так как в противном случае максимумы и минимумы могут не существовать.) В действительности требуется еще

меньше; в некоторых случаях функция $F(p, q)$ представляется в виде частного с положительным знаменателем. Основное соотношение $\max\text{-}\min = \min\text{-}\max$ остается справедливым при некоторых условиях выпуклости и вогнутости для числителя и знаменателя: в частности, оно имеет место, если каждый из них билинеен. Эта более расширенная форма основного результата будет использована в примере, который мы рассмотрим в следующем параграфе.

Было получено распространение этой теоремы на случай бесконечномерных пространств. Это важно при изучении случаев с бесконечным числом чистых стратегий, но даже и конечномерная форма основной теоремы в этих случаях часто бывает полезна, что подтверждается примером атаки и обороны, рассмотренным в § 9.4. В этом примере чистые стратегии изменяются в пределах выпуклых фигур, а функция $f(x, y)$ выпукла по y для каждого x . Общий принцип теории утверждает, что в таких случаях второй участник имеет оптимальную чистую стратегию. Возможность немедленно установить этот факт, конечно, облегчает изучение и расчет оптимальной схемы поведения.

Другая возможность применения конечномерного варианта основной теоремы в случаях с бесконечным числом чистых стратегий появляется, когда чистые стратегии описываются действительными числами x, y , изменяющимися в интервале от 0 до 1, и когда функция $f(x, y)$ является многочленом

$$f(x, y) = \sum c_{i,k} x^i y^k.$$

Смешанные стратегии являются функциями распределения $P(x)$, $Q(y)$ и функция $F(p, q)$ принимает вид

$$F(p, q) = \sum c_{i,k} \int x^i dP \int y^k dQ.$$

Задача, по существу, сводится к конечномерному случаю, так как смешанные стратегии полностью могут быть представлены конечным числом соответствующих им так называемых моментов $\int x^i dP$ и $\int y^k dQ$. Эти рассуждения устанавливают интересную связь теории игр с теорией моментов.

§ 9.8. Приложение к экономике. Как методы, так и результаты теории игр часто используются при изучении моделей, представляющих различные экономические ситуации. В качестве примера рассмотрим модель устойчивого расширенного производства, связь которой с теорией игр была замечена фон Нейманом [10].

Это производство считается замкнутой системой, включающей n предметов и m операций (технологических процессов). Предметы будут обозначаться целыми числами j , $1 \leq j \leq n$, а операции — целыми

числами i , $1 \leq i \leq n$. При каждой операции используются некоторые из предметов и производятся другие. Выбираются единицы измерения количества предметов и интенсивности операций. Для каждой операции i числа

$$a_{i1}, a_{i2}, \dots, a_{in}$$

будут представлять количества различных предметов, которые используются в i -й операции, если эта операция выполняется с единичной интенсивностью. Числа

$$b_{i1}, b_{i2}, \dots, b_{in}$$

будут количествами предметов, производимых при этой операции. Относительные интенсивности различных операций, имеющихся в производстве, обозначаются через

$$x_1^{(0)}, x_2^{(0)}, \dots, x_m^{(0)},$$

а относительные цены различных предметов j через

$$y_1^{(0)}, y_2^{(0)}, \dots, y_n^{(0)};$$

как x , так и y являются неотрицательными числами, а их суммы равны единице. Предположение об устойчивости системы требует, чтобы эти числа были постоянными, а предположение о расширенном производстве означает, что в конце каждой единицы времени имеющиеся интенсивности операций умножаются на фиксированную величину $\xi > 1$.

Для каждого предмета j

$$\sum_i b_{ij} x_i^{(0)} \geq \xi \sum_i a_{ij} x_i^{(0)},$$

так как в противном случае спрос расширенного производства превысит реальное предложение. Более того, если для определенного предмета j сохраняется знак неравенства в предыдущем выражении, то получается избыточная стоимость, которая в устойчивом производстве заставляет относительную цену y_j этого предмета быть равной 0. Эти два условия могут быть заменены утверждением, что для любого возможного распределения цен y_j

$$\sum_{i,j} b_{ij} x_i^{(0)} y_j \geq \xi \sum_{i,j} a_{ij} x_i^{(0)} y_j$$

и что равенство имеет место для действительных относительных цен $y_j^{(0)}$.

Кроме того, согласно законам экономики конкуренция снижает прибыли. Если мы обозначим через η число $1 + p$, где p является степенью заинтересованности, то для каждой операции i должны быть справедливы следующие соотношения:

$$\sum_j b_{ij} y_j^{(0)} \leq \eta \sum_j a_{ij} y_j^{(0)}.$$

Более того, если для определенной операции i имеет место строгое неравенство, то $x_i^{(0)} = 0$. Эти два положения равносильны утверждению, что для любых возможных интенсивностей x_i

$$\sum_{i,j} b_{ij} y_j^{(0)} x_i \leq \eta \sum_{i,j} a_{ij} y_j^{(0)} x_i$$

и что неравенство сохраняется для имеющих в начальный момент интенсивностей $x_i^{(0)}$.

Теперь мы предположим, что каждое a_{ij} положительно, и рассмотрим функцию

$$f(x, y) = \frac{\sum b_{ij} x_i y_j}{\sum a_{ij} x_i y_j},$$

где суммы в числителе и знаменателе берутся по i и j . Условия, установленные выше, могут быть выражены с помощью этой функции в следующей форме:

$$\xi \leq \min_y f(x^{(0)}, y), \quad \xi = f(x^{(0)}, y^{(0)}),$$

$$\eta \geq \max_x f(x, y^{(0)}), \quad \eta = f(x^{(0)}, y^{(0)}),$$

откуда вытекает, что $\xi = \eta$ и что, кроме того,

$$\begin{aligned} \xi = \min_y f(x^{(0)}, y) &= \max_x \left[\min_y f(x, y) \right] = \\ &= \min_y \left[\max_x f(x, y) \right] = \max_x f(x, y^{(0)}) = \eta. \end{aligned}$$

Этот результат устанавливает связь между приведенной моделью и основной теоремой, рассмотренной в предыдущем параграфе. Заметим, что функция $f(x, y)$ является отношением двух билинейных форм, и поэтому теорема применима к этой функции. Она показывает, что для любых заданных a_{ij} и b_{ij} имеется по крайней мере одно устойчивое расширенное производство, соответствующее этим данным и обеспечивающее общее значение $\max\text{-}\min$ и $\min\text{-}\max$ функции f , большее единицы. Эта общая величина является значением коэффициента ξ , описывающего расширение производства, и равна также коэффициенту $\eta = 1 + p$, описывающему степень заинтересованности.

Предположение $a_{ij} > 0$ для любых i и j и было сделано для того, чтобы сумма $\sum a_{ij} x_i y_j$ была всегда положительна и чтобы можно было делить неравенства на это выражение. Сделанное предположение слишком стеснительно и может быть ослаблено, если рассматривать вместо $f(x, y)$ функцию

$$g(x, y) = \frac{\sum b_{ij} x_i y_j}{\sum (b_{ij} + a_{ij}) x_i y_j}$$

при новом предположении $a_{ij} + b_{ij} > 0$. Естественно предположить в силу характера самой задачи, что $a_{ij} \geq 0$ и $b_{ij} \geq 0$. К сожалению,

необходимо ввести дополнительное предположение, что сумма $a_{ij} + b_{ij}$ будет на самом деле положительна для любых i и j . Если полностью опустить это ограничение, равенство $\xi = \eta$ может уже не выполняться.

§ 9.9. Последовательные приближения. За исключением примера, рассмотренного в § 9.6, в этой главе было мало сказано о реальных расчетах оптимальных смешанных стратегий. Использовалось много различных методов, начиная от простых геометрических рассуждений и кончая последними результатами теории линейных операторов.

Если оба участника имеют одно и то же конечное число чистых стратегий и, кроме того, каждый имеет оптимальную смешанную стратегию, которая задает отличные от нуля вероятности отдельных чистых стратегий, то, как уже отмечалось в § 9.5, неравенства, определяющие оптимальные стратегии, могут быть заменены равенствами, а решения могут быть найдены как решения системы линейных уравнений. В преобразованном виде этот метод применим к общему случаю конечного числа чистых стратегий. Шахматная доска, описывающая положение, превращается в квадратную доску выбрасыванием соответствующих рядов и столбцов. Эта квадратная доска содержит существенную часть первоначальной, и предыдущий метод к ней применим. К сожалению, отыскание этой существенной части скоро становится очень громоздким, если число чистых стратегий возрастает, и метод теряет свою эффективность.

Браун [2] предложил метод последовательных приближений, который успешно применялся в тех случаях, когда сведение к существенной части не представлялось возможным. Метод состоит в построении для каждого участника последовательностей $p^{(1)}, \dots, p^{(r)}, \dots$ и $q^{(1)}, \dots, q^{(r)}, \dots$ смешанных стратегий, которые стремятся соответственно к оптимальным смешанным стратегиям p и q . Это очень удобно для числовых расчетов, так как легко может быть запрограммировано для счетных машин. Сходимость метода, доказанная Робинсоном [9], к сожалению, очень медленная. Последовательности строятся с помощью индуктивного процесса. Пусть задана смешанная стратегия q второго игрока. Обозначим через $\varphi(q)$ чистую стратегию первого игрока, оптимальную против заданной смешанной стратегии q . Аналогично, пусть $\psi(p)$ — чистая стратегия второго игрока, которая является оптимальной против стратегии p . Начиная теперь с произвольной стратегии $q^{(1)}$, положим $p^{(1)}$ равным $\varphi(q^{(1)})$. Последующие члены последовательности образуются парами. Если предположить, что $q^{(r)}$ и $p^{(r)}$ уже известны, то следующая пара получается из соотношений

$$q^{(r+1)} = \frac{r-1}{r} q^{(r)} + \frac{1}{r} \psi(p^{(r)}), \quad p^{(r+1)} = \frac{r-1}{r} p^{(r)} + \frac{1}{r} \varphi(q^{(r+1)}).$$

Очень близким к этой процедуре является метод, разработанный независимо Брауном и фон Нейманом, в котором оптимальные стратегии получаются как пределы решений системы дифференциальных уравнений при стремлении к бесконечности независимого переменного.

§ 9.10. Общее понятие стратегии; игра, подобная покеру. Все примеры, которые рассматривались нами до сих пор, выбирались в соответствии с видом простейшей схемы, описанной в § 9.2. Однако во многих случаях возникающие положения имеют более сложный характер. Например, двое участников принимают не одно, а несколько последовательных решений. Кроме этого, могут появиться случайные ходы. В тот момент, когда решение должно быть реализовано, участнику, контролирующему очередной ход, могут стать частично известны результаты предыдущих ходов, ответы противника или случайные ходы.

Основная трудность состоит в том, чтобы выделить случаи, охватываемые теорией, и установить точное определение понятия чистой стратегии. Мы сосредоточим наше внимание на том случае, когда общее число ходов и возможных путей изменения положения является конечным. Чистая стратегия отождествляется с совокупностью инструкций, охватывающих всевозможные непредвиденные обстоятельства и информирующих постороннего о тех действиях, которые следует предпринять. Общее число чистых стратегий конечно. Представим себе, что каждый участник имеет полный перечень своих чистых стратегий, собранных в книгу по одной на страницу. Игроки могут предоставить свои права посреднику. Каждый сообщает ему номер страницы чистой стратегии, которую он предполагает использовать. Следуя инструкции, данной ему этими двумя выбранными чистыми стратегиями, посредник может действовать без дальнейшей консультации с игроками. Заметим, что расширенный вариант сводится, таким образом, к простейшей схеме; каждый участник выбирает число, а именно номер страницы, полностью игнорируя соответствующие действия своего противника.

Пример поможет уяснить такое понимание чистой стратегии. Мы проанализируем суммарно нулевую игру с двумя участниками. Это упрощенный вариант покера, но он сохраняет все же важную характерную черту — возможность блефа. Правила игры таковы:

а) Каждый игрок делает вклад, равный по величине 2.

б) Каждый игрок получает независимо высшую или низшую карту (более точного различения вводиться не будет). Это является случайным ходом. Предполагается, что каждый игрок независимо имеет одинаковые шансы получить высшую или низшую карту. Каждый игрок видит свою карту и не видит карты своего противника.

в) Первый игрок может спасовать или увеличить свою ставку на 3 единицы.

г) Второй игрок может либо спасовать, либо остаться в игре, либо еще увеличить ставку (на 3 единицы).

д) Первый игрок может спасовать, остаться в игре, или еще увеличить ставку (на 3 единицы).

е) Второй игрок может либо спасовать, либо остаться в игре.

Игра заканчивается, как только игрок пасует или просто остается в игре без дальнейшего повышения ставки. В последнем случае игроки сравнивают карты. Выигрывает игрок с высшей картой. Если же их карты одинаковы, то каждый получает свою ставку обратно. В первом случае, когда игрок пасует, он проигрывает свою ставку партнеру. Это правило касается, в частности, п. (в).

Для ограничения общего числа стратегий вводится дополнительное соглашение. Игрок, имеющий высшую карту, будет автоматически повышать ставку до тех пор, пока не вынужден будет остановиться в соответствии с правилом (е). При этом условии чистая стратегия может давать указания лишь в том случае, когда игрок получил младшую карту. Первый игрок имеет четыре чистых стратегии:

	Π_1	Π_2	Π_3	Π_4
I_1	$-\frac{1}{2}$	$\frac{1}{4}$	1	$\frac{7}{4}$
I_2	$-\frac{1}{4}$	0	$-\frac{1}{2}$	$\frac{1}{4}$
I_3	-1	$-\frac{3}{4}$	0	$\frac{3}{4}$
I_4	$-\frac{7}{4}$	$-\frac{3}{2}$	$\frac{5}{4}$	0

I_1 : выйти из игры в положении (в);

I_2 : повысить ставку в положении (в), но выйти из игры в положении (д);

I_3 : повысить ставку в положении (в) и остаться в игре в положении (д);

I_4 : повысить ставку в положении (в) и повысить ее в положении (д).

Второй игрок оказывается тоже имеет четыре стратегии:

Π_1 : выйти из игры в положении (г);

Π_2 : остаться в игре в положении (г);

Π_3 : повысить ставку в положении (г), но спасовать в положении (е);

Π_4 : повысить ставку в положении (г) и остаться в игре в положении (е).

Рис. 9.1. Матрица выплат для игры, подобной покеру.

Шахматное поле, соответствующее этой игре, представлено на рис. 9.1. Такое поле носит название матрицы выплат. Для пояснения метода расчета чисел в различных квадратах этого поля рассмотрим число $-\frac{7}{4}$ в левом нижнем углу. Первый игрок использует свою четвертую стратегию, т. е. увеличивает ставки, как только представляется возможность. Второй игрок использует свою первую стратегию и выходит из игры в четвертом положении. Если оба игрока получают высшие карты, то величина их выигрыша равна нулю. Если первый игрок имеет высшую карту, а второй — низшую, то первый игрок выигрывает ставку своего противника, равную 2. Если же первый игрок имеет низшую карту, а второй — высшую, то первый теряет свою ставку и три повышения, всего — 11. Наконец, если оба игрока имеют низшие карты, то выигрыш первого игрока у второго равен 2. Эти четыре возможности встречаются с вероятностью $\frac{1}{4}$ каждая. Следо-

для первого игрока: использовать I_1 с вероятностью $\frac{1}{7}$;
 использовать I_2 с вероятностью $\frac{6}{7}$;
для второго игрока: использовать II_1 с вероятностью $\frac{6}{7}$;
 использовать III_3 с вероятностью $\frac{1}{7}$.

§ 9.11. Заключение. Рассуждения этой главы относились главным образом к случаю суммарно нулевой игры двух игроков. Большая работа проделана также для изучения других случаев игры. В общем случае возможность и последствия образования союзов между участниками были рассмотрены фон Нейманом. Интуитивно ясно, а примеры окончательно подтверждают это, что игроки, объединившись в группу, могут достигнуть в некоторых случаях большего, чем они могли бы получить, действуя индивидуально. Этот выигрыш является иногда результатом соглашения, по которому один из игроков должен предпринять действие, приносящее ему убыток, но которое даст выигрыш другим членам группы. Такое соглашение имеет смысл, если только общий выигрыш перераспределяется между участниками согласно некоему достигнутому соглашению. Это последнее должно быть в достаточной степени справедливо для всех участников, чтобы у них не было побуждений причинить вред группе и заключить соглашение с другими для приобретения более выгодного положения. Теория игр в настоящее время включает только обсуждение стабильности уже образовавшихся объединений, но не учитывает возможности текущего образования новых объединений. Уяснение этого момента имело бы фундаментальное значение для экономики. Подробности анализа, проведенного фон Нейманом, можно найти в книгах, список которых приведен ниже, в частности, в книге фон Неймана и Моргенштерна [11].

ЛИТЕРАТУРА

1. Блэкуэлл Д. и Гиршик М., Теория игр и статистических решений, ИЛ, М., 1958.
2. Brown G. W., Notes on the Solution of Linear Systems Involving Inequalities, Proc. Second Symposium on Large-scale Digital Calculating Machinery, Harvard University Press, Cambridge, Mass., 1951, pp. 137—140.
3. Dresher M., Karlin S., Shapley L. S., Polynomial Games, Contributions to the Theory of Games, Ann. Math. Stud., 24, (1950), pp. 161—180.

4. Gross O., Targets of Differing Vulnerability with Attack Stronger than Defense, Research Memorandum 359, The Rand Corporation, Santa Monica, Calif., 1950.
5. Index of Publications, The Rand Corporation, Santa Monica, Calif., February, 1954. См. стр. 14 о играх с выбором момента времени.
6. Kuhn H. W., Lectures on the Theory of Games, Princeton University Press, Princeton, N. J., 1952.
7. McDonald J., Strategy in Poker, Business and War, Norton, New York, 1950.
8. McKinsey J. C. C., Introduction to the Theory of Games, McGraw-Hill, New York, 1952.
9. Robinson Julia, An Iterative Method of Solving a Game, Ann. of Math. (2), 54, 296—301 (1951).
10. Von Neumann J., A Model of General Economic Equilibrium, Rev. Econ. Stud., 13, 1—9 (1945—1946). (Перевод на английский язык оригинальной статьи И. Неймана.)
1. Von Neumann J., Morgenstern O., Theory of Games and Economic Behavior, 2d ed., Princeton University Press, Princeton, N. J., 1947.
12. Williams J. D., The Compleat Strategyst, McGraw-Hill, New York, 1954.

Глава 10

ТЕОРИЯ ДИНАМИЧЕСКОГО ПЛАНИРОВАНИЯ

РИЧАРД БЕЛЛМАН

Корпорация RAND

§ 10.1. Введение. В этой главе дается обзор одного класса новых и трудных математических задач, возникших в связи с процессами многоэтапного выбора, и излагаются некоторые математические методы, разработанные для решения этих задач.

Задачи, о которых идет речь, носят специальное название задач по планированию. Чтобы подчеркнуть в них роль времени и наличие последовательности выборов, мы предположили термин „динамическое планирование“. При правильном толковании эта динамическая точка зрения способствует глубокому математическому проникновению в вопрос.

Многоэтапные процессы, которые мы будем изучать, состоят из последовательностей операций, в которых результат предыдущих операций можно использовать для управления ходом будущих операций. Сразу же можно различить два типа операций: операции, где результат полностью определен, и операции, где результат не определен, но может быть предсказан с помощью некоторого распределения вероятностей. Операции первого типа мы назовем *детерминистическими*, операции второго типа — *стохастическими*. Ниже мы приведем примеры операций обоих типов.

Сложные явления современной жизни выявили с большой отчетливостью эти процессы и связанные с ними задачи. Можно сказать с уверенностью, что любое реалистичное рассмотрение финансирования в экономике, промышленного производства, программ обучения и тренировки, научных методов выборочного контроля, если упоминать области первостепенной важности, и любое реалистичное рассмотрение игр, основанных на случае, и игр, основанных на искусстве, если говорить об областях, на первый взгляд весьма далеких от перечисленных выше, должны в большей или меньшей степени опираться на теорию динамического планирования.

Для ознакомления с новыми математическими методами, необходимыми для решения возникающих математических задач, трудных и своеобразных, мы рассмотрим несколько простых примеров, подобранных так, чтобы не загромождать картины техническими подробностями, которые поневоле сопровождают более сложные и более близкие к действительности рассуждения.

Мы начнем с задачи, которая возникает в теории закупок и потому принадлежит к общей теории снабжения. Речь идет о повторной закупке на протяжении ряда лет оборудования, которое по использованию продается в утиль. Эта задача находит непосредственную аналогию в теории капиталовложений.

Вторая задача относится к наиболее эффективному использованию чувствительного к повреждениям золотодобывающего оборудования при разработке нескольких приисков. Эта задача хотя и не имеет столь очевидного применения, как первая, но зато обнаруживает ряд весьма важных математических особенностей.

Третья задача остро стоит в области промышленного производства. Она заключается в использовании комплекса отраслей промышленности для производства одного продукта. Как таковая, она попадает в число так называемых „задач на узкие места“.

В заключение мы очень кратко остановимся на применении математических методов, использованных при решении этих задач, к теории многоэтапных игр и, в частности, к „играм на выживание“.

§ 10.2. Предварительные рассмотрения. Перед тем как приступить к математическому исследованию, мы обрисуем в весьма общих чертах отличительные свойства, присущие всем этим задачам, на первый взгляд очень далеким друг от друга.

а) В любой момент времени t состояние процесса описывается набором немногих параметров.

Так, в задаче о закупках нужна информация включает сумму наличных денег, количество оборудования, сдаваемого в утиль, количество оборудования, требуемого на следующий период, и текущие цены. В задаче о золотых приисках мы предположительно должны знать количество золота, оставшегося в приисках, количество уже добытого золота и состояние золотодобывающего оборудования.

В задачах, относящихся к производственным процессам, нам потребуются производственные мощности и запасы продукции в данный момент, а также время, оставшееся до конца процесса. В „игре на выживание“, например в покере для двоих, продолжаемом до разорения одного из игроков, состояние игры описывается капиталами игроков. Если сумма денег, находящихся в игре, не изменяется, то капитал одного из игроков полностью определяет состояние игры.

б) Операция выбора состоит в преобразовании этого набора параметров в такой же набор с другими численными значениями.

Так, в задаче „на узкое место“ операция выбора состоит в изменении текущих запасов сырья и производственных мощностей в разных отраслях промышленности путем выделения каждой отрасли большего или меньшего количества основных ресурсов, например стали.

в) *Прошлая история системы не имеет значения при определении будущих действий (марковское свойство).*

Обычно это свойство можно получить, вводя в описание процесса дополнительные параметры. Разумеется, оно будет полезно лишь при условии, что общее число требующихся параметров невелико.

При наличии запаздываний по времени может оказаться необходимым описывать систему не с помощью параметров, а с помощью функций.

Мы умышленно оставили описание этих отличительных свойств несколько туманным, так как, по нашему мнению, здесь важен дух, а не буква. При рассмотрении частных случаев может потребоваться некоторая изобретательность, чтобы поставить задачу математически. Но этого, конечно, и надо ожидать. Изобретательность никогда не может быть изгнана аксиоматикой.

Теперь введем некоторые основные термины. Во всех рассматриваемых процессах производится последовательность выборов. Такую последовательность выборов мы будем называть *стратегией*¹⁾. Стратегии, наиболее желательные с точки зрения какого-либо заранее заданного критерия, будут называться *оптимальными стратегиями*.

Для всей этой главы фундаментальным является следующий интуитивно очевидный принцип:

Принцип оптимальности. Оптимальная стратегия имеет то свойство, что, каково бы ни было начальное состояние и начальный выбор, остальные выборы должны составлять оптимальную стратегию относительно состояния, возникшего вследствие первого выбора.

Математическая запись этого принципа приводит к новому классу функциональных уравнений с интересными свойствами. С другой стороны, математическая методика, развитая при рассмотрении этих задач, позволяет по-новому подойти к некоторым классическим задачам вариационного исчисления, что может представить как теоретический, так и практический интерес.

Для тех, кто заинтересуется дальнейшими результатами в этой новой области, приводится библиография, содержащая обзор проделанной в этой области работы, включая ряд результатов, от обсуждения которых здесь пришлось отказаться из-за недостатка места. Подробный обзор теории вместе с приложениями и численным решением функциональных уравнений, получающихся в приложениях, можно найти в книге автора „Введение в теорию динамического планирования“, издаваемой Принстонским университетом²⁾.

¹⁾ Автор обозначает это понятие двумя терминами: „стратегия „и“ политика“, употребляя их равнозначно. При переводе выбран термин „стратегия“, как более отвечающий существу дела. — *Прим. ред.*

²⁾ По-видимому, речь идет о книге: Bellman R., Dynamic Programming, Princeton University Press, 1957. Перевод последней готовится в Издательстве иностранной литературы. — *Прим. ред.*

Оптимальное распределение ресурсов

§ 10.3. Введение. Как первый пример задачи, связанной с процессами многоэтапного планирования, мы рассмотрим задачу о закупках. Сначала мы подойдем к задаче с классической точки зрения и обсудим неудобства такого подхода. Затем мы опишем постановку задачи с точки зрения теории динамического планирования.

Этот второй подход приводит к новому важному классу функциональных уравнений, свойства которых мы кратко рассмотрим, сперва с аналитической точки зрения, а потом с расчетной стороны. Оба пути на самом деле тесно связаны.

В заключение мы рассмотрим соответствующую стохастическую ситуацию.

§ 10.4. Задача о закупках. В практике снабжения весьма часто встречается задача о закупках оборудования, предназначенного для выполнения определенной работы. Так как во многих случаях эта работа не является простой, а складывается из нескольких операций разного характера, то оборудование, идеально приспособленное к одной операции, может оказаться посредственным при другой операции и совершенно бесполезным при третьей. Во избежание этого обычно приходится закупать различные виды оборудования и комбинировать их для удовлетворительного выполнения работы.

Рассмотрим случай, достаточно реальный, чтобы быть интересным, и достаточно простой, чтобы допускать легкую аналитическую формулировку.

Задача. На протяжении периода в N лет необходимо в начале каждого года заказывать какое-то оборудование для выполнения определенных работ. Мы располагаем начальной суммой денег x , которая делится на две части: y и $x - y$. Первая сумма y используется для закупки оборудования типа A , остальная сумма $x - y$ — для закупки оборудования типа B . Затрата y долларов на оборудование A даст $g(y)$ часов полезной работы, а затрата $x - y$ долларов на оборудование B даст $h(x - y)$ часов работы. В конце года можно продать в утиль оборудование A , получив вместо него ay долларов на покупку нового оборудования, где $0 < a < 1$, и продать в утиль оборудование B , получив $b(x - y)$ долларов на покупку нового оборудования, где $0 < b < 1$. Этот процесс покупки и продажи повторяется каждый год, пока не кончится N -летний период.

Требуется найти такое распределение денег по годам, которое даст наибольшее общее число часов полезной работы оборудования за N -летний период.

§ 10.5. Классическая постановка задачи. Пусть $y_1, y_2, \dots, y_N$ — суммы денег, использованные для закупки оборудования A соответ-

ственно в начале первого, второго, ..., N -го годов. Общее число человеко-часов, полученных от оборудования обоих типов за N -летний период, будет равно

$$J(y_1, y_2, \dots, y_N) = g(y_1) + h(x_1 - y_1) + g(y_2) + h(x_2 - y_2) + \dots \\ \dots + g(y_N) + h(x_N - y_N),$$

где

$$0 \leq y_i \leq x_i \quad (10.1)$$

и

$$\left. \begin{aligned} x_1 &= x, \\ x_2 &= ay_1 + b(x_1 - y_1), \\ &\vdots \\ x_N &= ay_{N-1} + b(x_{N-1} - y_{N-1}). \end{aligned} \right\} \quad (10.2)$$

Задача, поставленная в § 10.4, равносильна задаче определения точки максимума функции $J(y_1, y_2, \dots, y_N)$ в N -мерной области, описываемой неравенствами (10.1). Хотя дифференциальное исчисление и пригодно для решения этой задачи, но его следует применять осторожно, поскольку некоторые (а в ряде случаев, как мы увидим, и все) максимумы могут лежать на границе области.

Решение найти нетрудно, если N мало. Однако при произвольно выбранном N , например при $N=10$, задача исследования области N -мерного пространства, определенной уравнениями (10.2), становится весьма трудоемкой.

Заметим, кроме того, что с практической точки зрения такой подход дает слишком много информации. При данных x и N нет необходимости знать $y_2, y_3, \dots, y_N$. В начале первого года требуется только y_1 . Следовательно, найдя y_1 как функцию от x_1 и N , мы полностью решим задачу.

Руководствуясь этим простым соображением, мы в следующем параграфе подойдем к задаче совершенно иначе и по-новому.

§ 10.6. Постановка задачи с точки зрения динамического планирования. Легко видеть, что при заданных функциях g и h , которые для простоты будут предполагаться независимыми от времени, общее число часов полезной работы, полученных за N -летний период благодаря оптимальной стратегии расходов, является функцией только от x (начальной суммы денег) и от N (числа этапов или лет в периоде). Для $N=1, 2, \dots$ определим $f_N(x)$ как число часов полезной работы, полученных за N -летний период при данной начальной сумме денег x благодаря оптимальной стратегии. Тогда

$$f_1(x) = \max_{0 \leq y \leq x} [g(y) + h(x - y)]. \quad (10.3)$$

Выведем основное рекуррентное соотношение между $f_{N+1}(x)$ и $f_N(x)$. Пусть на первом этапе $(N+1)$ -этапного процесса сумма x разбивается на y и $x-y$. Число человеко-часов, полученных на протяжении всех $N+1$ лет, будет равно $g(y) + h(x-y)$ (т. е. числу часов, полученных в первый год) плюс число человеко-часов, полученных за остальной N -летний период при начальной сумме $ay + b(x-y)$.

Ясно, что независимо от выбора начального y сумма $ay + b(x-y)$ должна использоваться оптимально в течение оставшихся N этапов. Поэтому число часов, полученных в течение последних N этапов, будет, по определению, равно $f_N[ay + b(x-y)]$. Следовательно, общая отдача при начальном разбиении на y и $x-y$ будет равна

$$R_{N+1}(x, y) = g(y) + h(x-y)f_N[ay + b(x-y)].$$

Чтобы сделать общую отдачу максимальной, надо найти y , при котором $R_{N+1}(x, y)$ будет максимальным. Это приводит к уравнению

$$f_{N+1}(x) = \max_{0 \leq y \leq x} R_{N+1}(x, y) = \max_{0 \leq y \leq x} \{g(y) + h(x-y) + f_N[ay + b(x-y)]\} \quad (10.4)$$

для $N = 1, 2, \dots$.

В следующем параграфе мы рассмотрим одну важную аппроксимацию этой системы нелинейных уравнений.

§ 10.7. Бесконечноэтапная аппроксимация. Очень полезным приемом в этой и аналогичных задачах, в которых отдаленное будущее играет незначительную роль, является предположение о бесконечности N . Эта математическая фикция приводит к существенно упрощению задачи, так как последовательность $\{f_N(x)\}$ заменяется одной функцией

$$f(x) = \lim_{N \rightarrow \infty} f_N(x).$$

Вместо уравнения (10.4) мы получим уравнение

$$f(x) = \max_{0 \leq y \leq x} \{g(y) + h(x-y) + f[ay + b(x-y)]\}. \quad (10.5)$$

Но хотя введение бесконечноэтапного процесса весьма упрощает исследование, взамен возникают новые трудности, которых вовсе не было при конечном процессе. Мы теперь должны рассматривать вопрос о существовании и единственности решения уравнения (10.5).

Так, например, ясно, что прибавление любой постоянной к решению уравнения (10.5) дает новое решение. Чтобы фиксировать функцию, определяемую исходной задачей, надо добавить условие $f(0) = 0$.

В следующем параграфе мы сформулируем теорему существования и единственности, охватывающую широкий класс задач этого типа.

§ 10.8. Теорема существования и единственности. Мы приведем теорему существования и единственности, относящуюся к уравнению (10.5), и ее доказательство, чтобы проиллюстрировать некоторые возможные методы.

ТЕОРЕМА 10.1. *Рассмотрим уравнение*

$$f(x) = \max_R \{a(x_1, x_2, \dots, x_N) + f[b(x_1, x_2, \dots, x_N)]\}, \quad (10.6)$$

где область $R = R(x)$ определена соотношениями

$$x_k \geq 0, \quad x = \sum_{k=1}^N x_k.$$

Если:

а) функция $a(x_1, x_2, \dots, x_N)$ непрерывна и неотрицательна на $R(x)$ для $0 \leq x \leq x_0$, причем $a(0, 0, \dots, 0) = 0$;

б) функция $b(x_1, x_2, \dots, x_N)$ непрерывна и неотрицательна на R , причем на $R(x_0)$

$$b(x_1, x_2, \dots, x_N) \leq c \sum_{k=1}^N x_k = cx, \quad 0 < c < 1;$$

$$\text{в) } \sum_{l=0}^{\infty} h(c^l x_0) < \infty,$$

где $h(x) = \max_{0 \leq y \leq x} [\max_{R(y)} a(x_1, x_2, \dots, x_N)]$, то для $0 \leq x \leq x_0$ существует непрерывное решение уравнения (10.6), для которого $f(0) = 0$.

Доказательство. Пусть $f_0(x)$ — значение, получаемое при выборе

$$x_1 = x, \quad x_2 = x_3 = \dots = x_n = 0,$$

где $n = 2, 3, \dots$. Тогда

$$f_0(x) = a(x) + a[b(x)] + \dots,$$

где $a(x) = a(x, 0, \dots, 0)$, $b(x) = b(x, 0, \dots, 0)$. Ряд справа мажорируется рядом

$$\sum_{l=0}^{\infty} h(c^l x_0)$$

и поэтому равномерно сходится.

Положим

$$f_{n+1}(x) = \max_{R(x)} \{a(x_1, x_2, \dots, x_N) + f_n[b(x_1, x_2, \dots, x_N)]\}. \quad (10.7)$$

Из определения f_0 вытекает, что $f_1 \geq 0$ и, следовательно, $f_{n+1} \geq f_n$. Пусть

$$M_n(x) = \max_{0 \leq y \leq x} f_n(y).$$

Тогда из уравнения (10.7) имеем

$$M_{n+1}(x) \leq h(x) + M_n(cx),$$

откуда

$$M_n(x) \leq \sum_{i=0}^{\infty} h(c^i x).$$

Поэтому $f_n(x)$ сходится к функции $f(x)$ для всех $x \in [0, x_0]$. Легко видеть, что $f(x)$ удовлетворяет уравнению

$$f(x) = \sup_R \{a(x_1, x_2, \dots, x_N) + f[b(x_1, x_2, \dots, x_N)]\}. \quad (10.8)$$

Чтобы установить существование непрерывного решения и заметить на этом основании верхнюю границу $\sup$ в уравнении (10.8) максимумом, мы должны действовать иначе. Пусть $(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)$ — точка в $R(x)$, в которой достигается максимум. Тогда

$$\begin{aligned} f_{n+1}(x) &= \{a(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N) + f_n[b(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)]\} \geq \\ &\geq \{a(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N) + f_n[b(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)]\}, \end{aligned} \quad (10.9)$$

где $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)$ — любая другая точка в $R(x)$. В частности, если в качестве $(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)$ взять точку, в которой достигает максимума соответствующее выражение для $n-1$, то мы получим дополнительное соотношение

$$\begin{aligned} f_n(x) &= \{a(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N) + f_{n-1}[b(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)]\} \geq \\ &\geq \{a(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N) + f_{n-1}[b(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)]\}. \end{aligned} \quad (10.10)$$

Из неравенств (10.9) и (10.10) вытекает важное неравенство

$$\begin{aligned} &|f_{n+1}(x) - f_n(x)| \leq \\ &\leq \max \left\{ \begin{aligned} &|f_n[b(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)] - f_{n-1}[b(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)]| \\ &|f_n[b(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)] - f_{n-1}[b(\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)]| \end{aligned} \right\}. \end{aligned} \quad (10.11)$$

Теперь положим

$$\begin{aligned} M_0(x) &= \max_{0 \leq y \leq x} f_0(y), \\ M_{n+1}(x) &= \max_{0 \leq y \leq x} |f_{n+1}(y) - f_n(y)|, \quad n \geq 0. \end{aligned}$$

Так как по предположению $b(x_1, x_2, \dots, x_N) \leq cx$ для всех $(x_1, x_2, \dots, x_N) \in R(x)$, то из неравенства (10.11) при $n > 0$ получим

$$M_{n+1}(x) \leq M_n(cx).$$

Следовательно, $M_n(x) \leq M_0(c^n x) = h(c^n x)$. Так как, по предположению, ряд

$$\sum_{n=0}^{\infty} h(c^n x_0)$$

сходится, то ряд

$$\sum_{n=0}^{\infty} [f_{n+1}(x) - f_n(x)]$$

равномерно сходится для $0 \leq x \leq x_0$. Отсюда вытекает, что

$$f(x) = \lim_{n \rightarrow \infty} f_n(x)$$

есть непрерывная функция на $[0, x_0]$. Следовательно, $f(x)$ есть решение уравнения (10.6).

Так как $M(x)$ непрерывно и $M(0) = 0$, то мы получим $M(x) \leq 0$, а это значит, что $M(x)$ тождественно равняется нулю. Тем самым доказательство закончено.

§ 10.9. Аналитические результаты. В предыдущем параграфе мы дали весьма общую теорему существования и единственности. Теперь мы дадим несколько примеров явных аналитических результатов, которые можно получить. Доказательства (см. [3, 6, 9]) здесь не приводятся.

Мы будем рассматривать только простое уравнение (10.5).

ТЕОРЕМА 10.2. Если $g(x)$ и $h(x)$ обе суть строго выпуклые функции от x , то оптимальная стратегия требует, чтобы или $y = 0$, или $y = x$.

Случай, когда g и h обе вогнуты, более сложен.

ТЕОРЕМА 10.3. Пусть

- а) $g(0) = h(0) = 0$;
- б) $g'(x) \geq 0$, $h'(x) \geq 0$ для $x \geq 0$;
- в) $g''(x) \leq 0$, $h''(x) \leq 0$ для $x \geq 0$.

Рассмотрим последовательность уравнений

$$f_1(x) = \max_{0 \leq y \leq x} [g(y) + h(x-y)],$$

$$f_{n+1}(x) = \max_{0 \leq y \leq x} \{g(y) + h(x-y) + f_n[ay + b(x-y)]\},$$

$$n = 1, 2, \dots$$

Для каждого n существует единственное $y_n = y_n(x)$, дающее максимум. Если $b < a$, то $y_1 \leq y_2 \leq y_3 \leq \dots$; если же $b > a$, то справедливы обратные неравенства. В частности, если $b < a$ и при этом $y_n(x) = x$ для некоторого n , то $y_m(x) = x$ для $m \geq n$.

Этот результат полезен для приближенных вычислений, так как y_1 , y_2 и даже y_3 находятся при ручных вычислениях довольно быстро.

Но если даже g и h выпуклы и мы знаем, что или $y=0$, или $y=x$, все же нелегко решить, какое значение y правильно. Поэтому для приближенных вычислений полезен следующий результат:

ТЕОРЕМА 10.4. Решение уравнения

$$F(x) = \max_{0 \leq y \leq x} [cx^a + F(ax), ex^f + F(bx)]$$

дается формулой

$$y = \begin{cases} x & \text{при } 0 \leq x \leq x_0, \\ 0 & \text{при } x_0 \leq x, \end{cases}$$

где

$$x_0 = \left[\frac{c/(1-a^a)}{e/(1-b^a)} \right]^{1/(f-a)}.$$

Решение получается легко также в том случае, когда g и h квадратичны относительно x .

§ 10.10. Методика вычислений. При построении математических моделей любых физических задач всегда имеется противоречие между требованием достаточно верного описания действительного физического явления и аналитической простотой соответствующей модели. Поэтому следует ожидать, что чем вернее передана действительность, тем меньше шансов разрешить задачу одними лишь аналитическими средствами.

Таким образом, ясно, что применение математики преследует две цели:

а) найти единообразную методику для решения задач малых размерностей;

б) дать достаточно широкую основу для приближенного решения задач больших размерностей и в том числе средства для повышения точности приближения за счет дополнительных затрат времени и труда.

Как пример из другой области рассмотрим задачу определения траектории снаряда (см. гл. 3). Мы знаем, что надо учитывать аэродинамические характеристики снаряда, сопротивление воздуха как функцию скорости, скорость ветра, кривизну Земли, изменение характеристик орудийного ствола как функцию числа предыдущих выстрелов и много других факторов. Тем не менее начинают с рассмотрения параболической траектории материальной точки в пустоте, где действует только сила тяжести.

Отсюда следует вывод, что единственной целью при рассмотрении простых моделей, обладающих определенными свойствами оригинала, являются поиски исходной точки для последующего исследования все более и более сложных моделей.

При изучении задач по динамическому планированию мы располагаем методом аппроксимации, которого мы лишены при изучении

большинства функциональных уравнений анализа. Речь идет о двойственности, существующей между решением функционального уравнения и оптимальной стратегией.

Обращаясь к уравнению (10.5), мы видим, что для данной функции $f(x)$ можно найти максимизирующее u и, более того, все максимизирующие u . Обратно, зная u как функцию от x , можно получить $f(x)$ простой итерацией. Таким образом, функция определяет все оптимальные стратегии, а оптимальная стратегия определяет функцию.

Отсюда следует, что при попытках найти оптимальную стратегию, которая составляет нашу главную цель, можно выбирать начальное приближение в *пространстве функций* и можно выбирать его в *пространстве стратегий*.

В большинстве случаев выгоднее выбирать начальную стратегию, так как практика большей частью дает прочную основу для пробного решения задачи. Вспомним игру в покер с тремя или более игроками. Хотя и нет математической теории, указывающей правильную игру, опыт многих лет существования игры дает эвристический набор правил хорошей игры, пренебрежение которыми обходится дорого. Заметим мимоходом, что покер с двумя участниками доступен математическому анализу [7,31].

Как раз при выборе начального приближения, первого в серии последовательных приближений, практический опыт и теоретическая подготовка, полученная при рассмотрении простых моделей, могут сочетаться наиболее плодотворно.

Кроме того, метод приближения с помощью стратегии обладает тем огромным преимуществом перед методом приближения с помощью функции, что последовательные приближения с помощью стратегии автоматически будут давать каждый раз улучшение результатов.

Подкрепим эти замечания рассмотрением уравнения (10.5). Если выбрать сначала $f_0 = f_g$, то дальнейшие приближения получатся из рекуррентного соотношения

$$f_{N+1}(x) = \max_{0 \leq y \leq x} \{g(y) + h(x-y) + f_N[ay + b(x-y)]\}, \quad N=0, 1, \dots$$

При тех же условиях, как в теореме 10.1, можно доказать, что f_N сходятся к решению, если только f_g — непрерывная функция в интервале $0 \leq x \leq x_0$ и $f_g(0) = 0$.

Важно отметить, что в соответствии с произвольным выбором f_g сходимость, вообще говоря, не обязательно будет монотонной. Если, однако, выбрать за начальное приближение $f_1(x)$, где f_1 — максимальная отдача одного этапа, определяемая уравнением (10.3), то сходимость наверняка будет монотонной, но зато обычно слишком медленной. Тем не менее этот метод довольно полезен, поскольку весьма часто функции $f_N(x)$ также представляют интерес.

Теперь рассмотрим подробнее методику приближения с помощью стратегии. Вот некоторые целесообразные аппроксимации:

а) Можно выбрать на каждом этапе такое y , которое делает максимальной одноэтапную отдачу $g(y) + h(x - y)$.

б) Можно аппроксимировать $g(x)$ и $h(x)$ выпуклыми, вогнутыми, квадратическими функциями или функциями вида ax^b и использовать в качестве приближенных стратегий те, которые были получены для этих простых случаев.

в) Можно принять за критерий себестоимость одного часа и выбрать такое y , чтобы соответствующие себестоимости одного часа сделались равными, т. е. чтобы

$$\frac{g(y)}{(1-a)y} = \frac{h(x-y)}{(1-b)(x-y)}.$$

г) Можно принять за критерий предельную себестоимость одного часа и выбрать такое y , чтобы

$$\frac{g'(y)}{1-a} = \frac{h'(x-y)}{1-b}.$$

Это последнее приближение и является наилучшим.

Во всех рассмотренных выше случаях мы располагаем систематическими методами, чтобы найти приближенное выражение для y как функции от x . Используя один из этих методов, можно вычислить приближенное решение рекуррентно. Обозначим полученную функцию через $F_g(x)$. Она удовлетворяет уравнению

$$F_g(x) = g(y) + h(x - y) + F_g[ay + b(x - y)], \quad y = y_g(x).$$

Затем рассмотрим второе приближение, определяемое формулой

$$f_2(x) = \max_{0 \leq y \leq x} \{g(y) + h(x - y) + F_g[ay + b(x - y)]\}.$$

Очевидно, что для всех $y \in [0, x]$

$$f_2(x) \geq \{g(y) + h(x - y) + F_g[ay + b(x - y)]\} \quad (10.12)$$

и, следовательно,

$$f_2(x) \geq F_g(x) = g(y) + h(x - y) + F_g[ay + b(x - y)], \quad y = y_g(x).$$

Так как

$$f_3(x) = \max_{0 \leq y \leq x} \{g(y) + h(x - y) + f_2[ay + b(x - y)]\}, \quad (10.13)$$

то, сравнивая уравнения (10.12) и (10.13), получим $f_3 \geq f_2$, поскольку $f_2 \geq F_g$. По индукции мы найдем в конце концов, что

$$f_{N+1} \geq f_N \geq \dots \geq F_g$$

со строгим неравенством для некоторых значений x , за исключением случая $F_g = f$.

§ 10.11. Неопределенность. В предыдущих параграфах мы предполагали полную определенность результата каждой затраты, составляющего твердое число человеко-часов и твердую выручку от продажи в утиль. Но в действительности можно предвидеть лишь среднее число человеко-часов, которые могут быть получены от затраты u долларов на оборудование A или B , и можно делать лишь предположения о выручке от продажи в утиль в конце года, исходя из развития экономики в сторону повышения или понижения цен, а также из предполагаемого состояния оборудования в конце года.

Сразу же возникает вопрос, как можно делать определенные утверждения о форме оптимальной стратегии перед лицом этой неопределенности в исходе любого действия?

Прежде всего ясно, что нельзя искать максимум *общего* числа человеко-часов, полученных за N -летний период, ибо эта величина является стохастической и зависит от различных факторов, многие из которых находятся совершенно вне нашего контроля. Следовательно, нужно условиться использовать какое-нибудь среднее значение этой величины. Как теоретически, так и практически проще всего взять обычное среднее. Математически это означает, что мы фиксируем наше внимание на математическом ожидании числа полезных рабочих часов.

Другой весьма важной величиной является вероятность получения не менее чем h часов полезной работы. Во многих случаях эта величина служит гораздо более важным критерием, чем более наглядное среднее значение.

Мы, однако, рассмотрим здесь только математическую постановку задач, связанных со средними значениями (математическими ожиданиями).

Чтобы не создавать аналитических трудностей, несущественных на этой стадии, предположим, что после затраты u долларов на покупку оборудования A существует только две возможности:

а) с вероятностью p_1 мы получим $g_1(u)$ часов работы и выручим $a_1 u$ долларов от продажи в утиль;

б) с вероятностью $p_2 = 1 - p_1$ мы получим $g_2(u)$ часов работы и выручим $a_2 u$ долларов от продажи в утиль.

Аналогично мы предположим, что при затрате $x - u$ долларов на оборудование B :

а) с вероятностью q_1 мы получим $h_1(x - u)$ часов работы и выручим $b_1(x - u)$ долларов от продажи в утиль;

б) с вероятностью q_2 мы получим $h_2(x - u)$ часов работы и выручим $b_2(x - u)$ долларов от продажи в утиль.

Указанные выше константы должны удовлетворять условию $0 < a_1, a_2, b_1, b_2 < 1$. Рассматривая для простоты только бесконечный процесс, определим $f(x)$ как математическое ожидание числа рабочих часов, полученных от начальной суммы x долларов благодаря оптимальной стратегии закупок. Тогда с помощью совершенно тех же

рассуждений, как и выше, мы увидим, что $f(x)$ удовлетворяет функциональному уравнению

$$f(x) = \max_{0 \leq y \leq x} \{p_1 q_1 \{g_1(y) + h_1(x-y) + f[a_1 y + b_1(x-y)]\} + \\ + p_1 q_2 \{g_1(y) + h_2(x-y) + f[a_1 y + b_2(x-y)]\} + \\ + p_2 q_1 \{g_2(y) + h_1(x-y) + f[a_2 y + b_1(x-y)]\} + \\ + p_2 q_2 \{g_2(y) + h_2(x-y) + f[a_2 y + b_2(x-y)]\}\}.$$

Это уравнение имеет такой же вид, как и в детерминистическом случае, и мы получим аналогичные результаты. Следует заметить, что после введения понятия о математическом ожидании мы уже для всех математических целей имеем дело с детерминистическим процессом.

Уравнение золотых приисков

§ 10.12. Введение. В первой части этой главы, названной „Оптимальное распределение ресурсов“, мы рассмотрели довольно подробно одну задачу детерминистического типа. Иными словами, исход наших действий был вполне определенным. Затем мы сказали о видоизменениях, необходимых для решения стохастического аналога этой задачи. В настоящей, второй части мы рассмотрим задачу чисто стохастического типа. Мы еще раз применим метод функциональных уравнений.

После постановки и математической формулировки задачи мы приведем общую теорему существования и единственности, относящуюся к функциональным уравнениям соответствующего класса. Затем мы найдем решение нашей первоначальной задачи и выясним его смысл. Рассмотрим задачу в том случае, когда функция полезности предполагается нелинейной, мы обратимся к непрерывному варианту задачи. Эта новая постановка, которую можно считать просто другим способом аппроксимации, допускает применение классических вариационных методов. Эти методы в соединении с той информацией о природе решения, которую мы получили из дискретного варианта, позволяют получить полное решение непрерывного варианта и в конечном счете много ценных сведений о первоначальном процессе.

§ 10.13. Задача о золотых приисках. Предположим, что мы владеем двумя золотыми приисками: прииском Анаконда с количеством золота x и прииском Бонанца с количеством золота y , а также одной золотодобывающей машиной, чувствительной к повреждениям конструкции.

При использовании машины на прииске Анаконда существует вероятность p_1 того, что $r_1 x$ золота будет добыто без повреждения машины и работа, таким образом, сможет продолжаться, и вероятность $1 - p_1$ того, что машина получит непоправимые повреждения и добыча прекратится. Аналогично, прииску Бонанца приписаны вероятности q_1 и $1 - q_1$ и количество $r_2 y$.

Требуется найти способ действия, обеспечивающий максимальное математическое ожидание количества золота, добытого до повреждения машины.

В следующем параграфе мы поставим задачу математически.

§ 10.14. Математическая постановка задачи. Так как мы имеем дело со стохастическим процессом, то, чтобы говорить о максимизации какой-либо величины, мы опять должны ввести подходящую метрику. Для начала предположим, что мы хотим сделать максимальным математическое ожидание количества золота, добытого до повреждения машины. Вследствие линейности этой метрики очевидно, что на каждом этапе совершенно безразлично, сколько золота уже добыто. Состояние процесса при этом полностью описывается в каждый момент времени величинами x и y , определенными выше. Пусть теперь $f(x, y)$ равняется математическому ожиданию количества золота, добытого до повреждения машины с помощью оптимальной стратегии при начальном количестве x в Анаконде и y в Бонанце.

При выборе для добычи Анаконды — обозначим эту операцию через A — мы с вероятностью p_1 получим золото в количестве r_1x и возможность продолжения работы и с вероятностью $1 - p_1$ не получим ничего ни в настоящем, ни в будущем. Так как оптимальная стратегия должна иметь оптимальное продолжение после любого начального хода, то математическое ожидание общей отдачи при стратегии, начинающейся с выбора A , будет

$$E_A = p_1 \{r_1x + f[(1 - r_1)x, y]\}.$$

Аналогично, математическое ожидание общей отдачи при стратегии, начинающейся с выбора B , будет

$$E_B = q_1 \{r_2y + f[x, (1 - r_2)y]\}.$$

Так как мы хотим сделать максимальным математическое ожидание отдачи, то

$$f(x, y) = \max(E_A, E_B),$$

что при полной записи дает функциональное уравнение

$$f(x, y) = \max \left(p_1 \{r_1x + f[(1 - r_1)x, y]\}, q_1 \{r_2y + f[x, (1 - r_2)y]\} \right). \quad (10.14)$$

§ 10.15. Существование и единственность. Хотя уравнение (10.14) такого же рода, как и уравнение, рассмотренное в § 10.7, оно не того же вида и поэтому требует отдельной теоремы существования и единственности.

ТЕОРЕМА 10.5. *Рассмотрим уравнение*

$$f(p) = \max_{1 \leq k \leq m} [g_k(p) + h_k(p) f(T_k p)] \quad (10.15)$$

и предположим, что:

а) точка p вынуждена оставаться в области R , обладающей тем свойством, что из условия $p \in R$ следует принадлежность преобразованной точки $p' = T_k p \in R$;

б) $|g_k(p)| \leq c_1$ для $p \in R$;

в) $|h_k(p)| \leq c_2 < 1$ для $p \in R$.

При этих условиях существует единственное ограниченное решение уравнения (10.15).

Доказательство этой теоремы во многом повторяет ход доказательства, данного в первой части главы. При этом используются последовательные приближения, определяемые соотношениями

$$f_{n+1}(p) = \max_k [g_k(p) + h_k(p) f_n(T_k p)], \quad n = 0, 1, 2, \dots,$$

где $f_0(p)$ — произвольная ограниченная функция в R . Легко показать, что ряд

$$\sum_{n=0}^{\infty} [f_{n+1}(p) - f_n(p)]$$

сходится равномерно для $p \in R$. Доказательство единственности проводится также по предыдущему образцу.

Таким же способом мы получаем следующий результат.

ТЕОРЕМА 10.6. Если функции $g_k(p)$ и $h_k(p)$ непрерывны в области R и если выполнены предыдущие условия, то и функция $f(p)$ непрерывна в R . Более того, если g_k и h_k — непрерывные функции параметров q , то и $f(p)$ будет непрерывной функцией от q .

§ 10.16. Решение. Мы найдем решение более общего функционального уравнения, чем (10.15), а именно уравнения

$$f(x, y) = \max \left\{ \begin{array}{l} A: \sum_{k=1}^N p_k [c_k x + f(c'_k x, y)] \\ B: \sum_{k=1}^N q_k [d_k y + f(x, d'_k y)] \end{array} \right\}, \quad (10.16)$$

где

$$а) \quad p_k \geq 0, \quad q_k \geq 0, \quad \sum_{k=1}^N p_k < 1, \quad \sum_{k=1}^N q_k < 1;$$

$$б) \quad 1 \geq c_k \geq 0, \quad 1 \geq d_k \geq 0, \quad c_k + c'_k = d_k + d'_k = 1;$$

$$в) \quad x \geq 0, \quad y \geq 0.$$

Уравнение (10.16) соответствует процессу, аналогичному описанному в § 10.13, но со многими возможными значениями отдачи при начальном выборе, допускающем продолжение.

ТЕОРЕМА 10.7. Решение уравнения (10.16) таково: если

$$\frac{\sum_{k=1}^N p_k c_k}{1 - \sum_{k=1}^N p_k} x > \frac{\sum_{k=1}^N q_k d_k}{1 - \sum_{k=1}^N q_k} y,$$

то надо выбрать A , а если справедливо обратное неравенство, то надо выбрать B . В случае равенства оба выбора одинаково удовлетворительны.

Геометрически решение можно выразить в очень изящной форме. Пусть прямая L (см. рис. 10.1) задана уравнением

$$\frac{\sum_{k=1}^N p_k c_k}{1 - \sum_{k=1}^N p_k} x = \frac{\sum_{k=1}^N q_k d_k}{1 - \sum_{k=1}^N q_k} y.$$

Если начальное состояние (x, y) есть точка под прямой L , то сначала применяем стратегию A , а если начальное состояние есть точка над L , то сначала применяем стратегию B . Если точка (x, y) лежит на самой прямой L , то безразлично, какую стратегию применять.

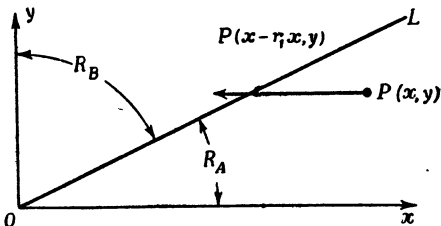


Рис. 10.1. Геометрическое представление решения уравнения золотых приисков.

Если наклон прямой L известен, то выбор между приисками A и B определяется относительными количествами x и y золота в приисках. Вектор из точки P может входить в соседнюю область.

§ 10.17. Интерпретация решения. Выясним смысл решения. Величина

$$\frac{\left(\sum_{k=1}^N p_k c_k \right) x}{1 - \sum_{k=1}^N p_k}$$

допускает весьма интересную интерпретацию. Числитель

$$\left(\sum_{k=1}^N p_k c_k \right) x$$

представляет собой математическое ожидание непосредственной прибыли от операции A , тогда как знаменатель

$$1 - \sum_{k=1}^N p_k$$

представляет собой математическое ожидание непосредственных потерь, т. е. вероятность повреждения машины. Следовательно, в целом

выражение равно отношению ожидания прибыли к ожиданию потерь; это отношение является весьма целесообразной и интуитивно оправданной мерой ценности отдельной операции.

К сожалению, в более сложных случаях это изящное решение теряет силу. Так, например, доказано, что в функциональном уравнении

$$f(x, y) = \max \begin{pmatrix} A: p_1 \{r_1 x + f[(1-r_1)x, y]\} \\ B: q_1 \{r_2 y + f[x, (1-r_2)y]\} \\ C: p_2 \{r_3 x + r_4 y + f[(1-r_3)x, (1-r_4)y]\} \end{pmatrix},$$

соответствующем фиктивному случаю, когда машину можно использовать в обоих приисках одновременно, существуют значения параметров, дающие решение такого вида, как на рис. 10.2, вместо предполагаемого решения такого вида, как на рис. 10.3.

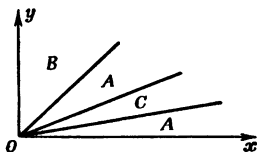


Рис. 10.2. Геометрическое представление обобщенного уравнения золотых приисков для некоторых значений параметров.

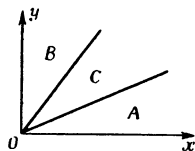


Рис. 10.3. Измененное решение обобщенного уравнения золотых приисков, обусловленное изменениями параметров.

Однако в качестве начального приближения критерий *отношения ожидания прибыли к ожиданию потерь*, несомненно, важен.

§ 10.18. Нелинейная полезность. В математической формулировке задачи выше мы предполагали, что требуется сделать максимальным математическое ожидание прибыли. Во многих случаях нас могут интересовать совсем другие критерии полезности. Например, мы можем пожелать сделать максимальной вероятность того, что общая добыча превышает заданное количество m . Эта вероятность является математическим ожиданием функции φ от общей добычи R , где

$$\varphi(R) = \begin{cases} 0 & \text{при } 0 \leq R < m, \\ 1 & \text{при } R \geq m. \end{cases}$$

Вообще говоря, если φ нелинейна, то для описания состояния процесса нам потребуются не только количества золота x и y , оставшиеся в приисках, но и уже добытое количество золота. Добытое количество золота мы обозначим через z .

Пусть теперь $f(x, y, z)$ равняется математическому ожиданию функции $\varphi(R)$ при условии, что уже добыто количество z , остались в приисках количества x и y и применяется оптимальная стратегия.

Функциональное уравнение для f таково:

$$f(x, y, z) = \max \left\{ \begin{array}{l} A: p_1 f[(1-r_1)x, y, z+r_1x] + (1-p_1)\varphi(z) \\ B: q_1 f[x, (1-r_2)y, z+r_2y] + (1-q_1)\varphi(z) \end{array} \right\}.$$

Начальным условием служит $f(0, 0, z) = \varphi(z)$, что очевидно из уравнения.

Решение уже не получается так же просто, как раньше, хотя и можно показать, что при некоторых подходящих допущениях оно имеет тот же самый вид.

Особенно интересен случай, когда $\varphi(R) = e^{bR}$. В этом случае (вследствие функционального уравнения для экспоненциальной функции) добытое количество золота не имеет значения при определении текущей стратегии. Если $f(x, y)$ — математическое ожидание от e^{bR} (т. е. $f(x, y) = E\{e^{bR}\}$) при использовании оптимальной стратегии, то функциональное уравнение для f имеет вид

$$f(x, y) = \max \left\{ \begin{array}{l} A: p_1 e^{br_1x} f[(1-r_1)x, y] + (1-p_1) \\ B: q_1 e^{br_2y} f[x, (1-r_2)y] + (1-q_1) \end{array} \right\}.$$

Его решение таково: если

$$\frac{p_1(e^{br_1x} - 1)}{1 - p_1} > \frac{q_1(e^{br_2y} - 1)}{1 - q_1},$$

то надо применить A ; если справедливо обратное неравенство, то надо применить B ; при равенстве же можно применить любую из этих операций.

Заметим, что предельная форма этого критерия при $b \rightarrow 0$ в точности совпадает с критерием, полученным для линейного случая в § 10.16.

§ 10.19. Непрерывный вариант. Так как дискретная формулировка приводит к ряду еще не решенных задач, то мы обратимся к непрерывному варианту задачи, где на помощь приходят мощные средства непрерывности.

Для того чтобы получить непрерывный аналог, мы предположим, что каждая операция имеет большую вероятность получения малого количества золота и отсутствия повреждений машины и малую вероятность повреждения машины и отсутствия добычи.

При малом $\delta > 0$ пусть $1 - q_1\delta$ равно вероятности получить $r_1x\delta$ и оставить машину неповрежденной, когда применяется A , и пусть $1 - q_2\delta$ равно вероятности получить $r_2y\delta$ и оставить машину неповрежденной, когда применяется B ; здесь $q_1, q_2 > 0$.

Как и прежде, перед нами уравнение

$$f(x, y) = \max \left\{ \begin{array}{l} A : (1 - q_1\delta) [r_1x\delta + f(x - r_1x\delta, y)] \\ B : (1 - q_2\delta) [r_2y\delta + f(x, y - r_2y\delta)] \end{array} \right\}.$$

Действуя формально, мы получим при $\delta \rightarrow 0$

$$f(x, y) = \max \left[\begin{array}{l} A : f(x, y) + r_1x\delta - q_1\delta f(x, y) - r_1x\delta \frac{\partial f}{\partial x} + O(\delta^2) \\ B : f(x, y) + r_2y\delta - q_2\delta f(x, y) - r_2y\delta \frac{\partial f}{\partial y} + O(\delta^2) \end{array} \right],$$

или

$$0 = \max \left[\begin{array}{l} A : r_1x - q_1f(x, y) - r_1x \frac{\partial f}{\partial x} \\ B : r_2y - q_2f(x, y) - r_2y \frac{\partial f}{\partial y} \end{array} \right]. \quad (10.17)$$

Но этот подход не представляется плодотворным ввиду трудностей, связанных с установлением теорем существования и единственности и с аналитическим исследованием уравнения (10.17).

Вместо дифференциального подхода, которым мы пользовались выше, применим теперь интегральный подход. Пусть мы подчинены стратегии S , состоящей из некоторой последовательности выборов операций A и операций B , и пусть мы выполнили n шагов этой стратегии (т. е. S_n). Соответствующее функциональное уравнение имеет вид

$$f(x, y) = \max_{S_n} \left[R_n(x, y) + \sum_k P_{nk}(x, y) f(x_{nk}, y_{nk}) \right], \quad (10.18)$$

где $R_n(x, y)$ — математическое ожидание отдачи за n этапов при стратегии S_n , а $P_{nk}(x, y)$ — вероятность сохранения машины и прихода в точку (x_{nk}, y_{nk}) при стратегии S_n .

Если выбрать такое $n\delta$, которое оставалось бы конечным и было бы равно t при $\delta \rightarrow 0$, $n \rightarrow \infty$, то аналогом уравнения (10.18) будет служить функциональное уравнение вида

$$f(x, y) = \max_{S(t)} \left[R(x, y, t) + \int_{r=0}^1 \int_{s=0}^1 f(xr, ys) dG_S(x, y, r, s) \right]. \quad (10.19)$$

Функциональные уравнения этого класса могут возникать в различных непрерывных вариантах задач по динамическому планированию. Мы не входим здесь в обсуждение этой постановки задач ввиду многих идейных и математических трудностей, связанных с понятием непрерывной стратегии для задач стохастического выбора. Вместо этого мы будем пользоваться третьим подходом, который находится в таком же отношении к уравнению (10.19), в каком

применение уравнения теплопроводности в теории диффузии находится к более строгому уравнению Чепмена—Колмогорова.

Прежде чем обсуждать подробности, заметим, что введение непрерывных процессов порождает трудность нового рода, которой не было в дискретном процессе. При $\delta \rightarrow 0$ пределом оптимальных процедур, описанных в § 10.16, является процедура, показанная на рис. 10.4. Она состоит в том, чтобы, будучи в $P(x, y)$, применять A , пока не достигнем прямой L , а затем идти по этой прямой.

Трудность заключается в том, что применение операции A или B , приводящее к горизонтальным или вертикальным перемещениям, никогда не даст перемещения вдоль прямой L .

Чтобы получить путь, следующий вдоль линии L , мы должны допустить бесконечно малое смещение стратегий A и B . Однако введение мгновенного (точечного) смещения приводит к ряду трудностей физического и математического характера. Математически мы сталкиваемся с той же трудностью, которая заставила нас отказаться от уравнения (10.19); физически мы не склонны считать стратегию этого типа применимой к первоначальной задаче, где требовался выбор между A и B в каждый момент времени.

Чтобы избежать этого понятия мгновенного смещения, мы воспользуемся приемом, который часто бывает полезен в задачах такого типа. Он основан на том, что для математических целей смещение в малых интервалах, содержащих рассматриваемый момент времени, эквивалентно смещению в самый момент t , при некоторых естественных допущениях о непрерывности [4, 31, 40, 41].

Предположим поэтому, что мы рассматриваем процесс, требующий выбора между A и B в моменты времени $0, \Delta, 2\Delta$ и т. д., и что в интервале $[k\Delta, (k+1)\Delta]$ мы применяем A для части $[k\Delta, k\Delta + \varphi_1\Delta]$ и B для части $[k\Delta + \varphi_1\Delta, (k+1)\Delta]$, где φ_1 зависит от k .

Предположив, что Δ мало и что членов первого порядка достаточно для описания процесса, мы выведем систему дифференциальных уравнений, определяющую процесс.

§ 10.20. Вывод дифференциальных уравнений. Пусть $x(t)$ равно количеству золота, оставшегося в A после того, как работа продолжалась до момента t , $y(t)$ — количеству золота, оставшегося в B после того, как работа продолжалась до момента t , $p(t)$ — вероятности сохранения машины до момента t , т. е. вероятности

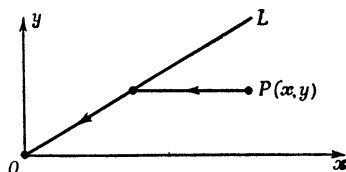


Рис. 10.4. Геометрическое представление решения непрерывного варианта уравнения золотых приисков. По достижении прямой L работа продолжается вдоль этой прямой.

того, что работа продолжалась до момента t , и $f(t)$ — математическому ожиданию количества золота, добытого до момента t . Здесь $t = n\Delta$; $n = 0, 1, 2, \dots$

Пренебрегая членами второго порядка относительно Δ , получим

$$\begin{aligned}x(t + \Delta) &= x(t) - r_1 \varphi_1(t) x(t) \Delta, \\y(t + \Delta) &= y(t) - r_2 \varphi_2(t) y(t) \Delta, \\p(t + \Delta) &= p(t) [1 - q_1 \varphi_1(t) \Delta - q_2 \varphi_2(t) \Delta], \\f(t + \Delta) &= f(t) + p(t) [\varphi_1(t) r_1 x(t) + \varphi_2(t) r_2 y(t)] \Delta,\end{aligned}$$

где $\varphi_2 = 1 - \varphi_1$. При $\Delta \rightarrow 0$ получим систему дифференциальных уравнений

$$\left. \begin{aligned}\frac{dx}{dt} &= -\varphi_1(t) r_1 x(t), & x(0) &= x_0, \\ \frac{dy}{dt} &= -\varphi_2(t) r_2 y(t), & y(0) &= y_0, \\ \frac{dp}{dt} &= -p(t) [\varphi_1(t) r_1 x(t) + \varphi_2(t) r_2 y(t)], & p(0) &= 1, \\ \frac{df}{dt} &= p(t) [\varphi_1(t) r_1 x(t) + \varphi_2(t) r_2 y(t)], & f(0) &= 0.\end{aligned}\right\} \quad (10.20)$$

Теперь задача сводится к определению такой функции $\varphi_1(t)$, подчиненной ограничению

$$0 \leq \varphi_1(t) \leq 1, \quad (10.21)$$

которая максимизировала бы $f(T)$ для данного промежутка времени длины T .

Нелинейность задачи важна при применении вариационных методов. Это свойство вместе с нашей априорной информацией о вероятной форме решения позволяет получить полное решение задачи. Однако важнее то обстоятельство, что многие задачи подобного типа, не поддающиеся решению в дискретной форме, могут быть полностью решены в этом варианте.

§ 10.21. Решение. Решение задачи максимизации $f(\infty)$ содержится в следующем результате.

ТЕОРЕМА 10.8. В случае уравнений (10.20) и ограничений (10.21) максимальное значение для $f(\infty)$ достигается применением стратегии

$$\varphi_1 = 1 \quad \text{при} \quad q_1 r_2 y < q_2 r_1 x,$$

$$\varphi_2 = 1 \quad \text{при} \quad q_1 r_2 y > q_2 r_1 x,$$

$$\varphi_1 = \frac{r_2}{r_1 + r_2} \quad \text{и} \quad \varphi_2 = \frac{r_1}{r_1 + r_2} \quad \text{при} \quad q_1 r_2 y = q_2 r_1 x.$$

Как показано в нашей с Леманом статье [41], где эти вопросы обсуждаются подробнее, решение, по существу, единственно. Там показано также, что если ввести любой критерий полезности $F(z)$, для которого $F'(z) > 0$ при $z \geq 0$, то задача максимизации ожидания прибыли будет иметь такое же решение. Это поднимает интересные теоретические вопросы, в которые, однако, мы не будем здесь входить.

Можно также получить решение для конечного T . Оно гораздо сложнее из-за наличия краткосрочных стратегий.

Задачи на узкие места

§ 10.22. Введение. В этой части главы мы рассмотрим класс задач, которые возникают при изучении производственных процессов, связанных с комплексом разных отраслей промышленности.

В частности, мы исследуем некоторые задачи, которые назовем „задачами на узкие места“, так как в них уровень активности регулируется ресурсами, подаваемыми в наименьшем количестве.

В задачах такого рода удачная постановка есть залог успеха. Все математические постановки задач по необходимости приближенны; поэтому первостепенную важность приобретает построение приближения, допускающего наиболее эффективное применение математического анализа. После рассмотрения постановки задачи, основанной на идее непрерывного приближения (см. параграф „Уравнение золотых приисков“ этой главы), мы изложим теоретический метод, обладающий широкой приложимостью, который основывается на идее применения функциональных уравнений — идее, также использованной в предыдущих частях. Этот метод, если его применять уверенно и решительно, часто приводит к очень быстрому решению.

Для доказательства того, что действительно получено решение, мы пользуемся линейностью задачи и строим двойственную ей задачу. Двойственная задача служит важным теоретическим инструментом и в то же время методом проверки.

Проиллюстрировав применение этих методов, мы вернемся на короткое время к обсуждению некоторых нелинейных задач и некоторых задач о собственных значениях. Применение функциональных уравнений сообщает совершенно новую постановку некоторым классическим задачам вариационного исчисления.

§ 10.23. Типичная задача. Мы рассмотрим простую модель трехотраслевого производственного процесса, где основными отраслями промышленности являются автомобильная, сталелитейная и инструментальная.

Эти отрасли промышленности используются для выпуска как можно большего числа автомобилей в течение промежутка времени T .

Для произвольного времени t , которое мы для начала считаем принимающим дискретные значения $0, 1, 2, \dots, T$, мы определяем:

$x_1(t)$ = количество автомобилей, выпущенных до момента t ;

$x_2(t)$ = производственная мощность автомобильных заводов;

$x_3(t)$ = запас стали;

$x_4(t)$ = производственная мощность сталелитейных заводов;

$x_5(t)$ = запас инструментов;

$x_6(t)$ = производственная мощность инструментальных заводов.

Интересно отметить, что в задачах этого типа непрерывные модели могут оказаться ближе к действительности, чем дискретные, которые мы рассмотрим в первую очередь, хотя и кратко.

Основные принципы таковы:

а) увеличение автомобильной, сталелитейной или инструментальной производственной мощности требует стали и инструментов;

б) производство автомобилей требует автомобильной производственной мощности и стали;

в) производство стали требует только сталелитейной производственной мощности;

г) производство инструментов требует инструментальной производственной мощности и стали.

Далее, мы предполагаем, что производство прямо пропорционально производственной мощности при обилии сырья и прямо пропорционально количеству выделенного сырья при обилии соответствующих производственных мощностей.

В начале каждого единичного интервала мы распределяем различные количества из запасов стали и инструментов для целей производства автомобилей, стали и инструментов и увеличения запасов автомобилей, стали и инструментов.

Пусть $z_i(t)$ равняется количеству стали, выделенному в момент t для увеличения $x_i(t)$, и пусть $w_i(t)$ равняется количеству инструментов, выделенному в момент t для увеличения $x_i(t)$.

В силу правил, сформулированных выше, легко видеть, что

$$z_3 = 0, \quad w_1 = w_3 = w_5 = 0.$$

Поэтому мы получаем следующие уравнения для x_i в момент $t+1$ в зависимости от x_i , z_i и w_i в момент t :

$$\left. \begin{aligned} x_1(t+1) &= x_1(t) + \min [\gamma_1 x_2(t), \alpha_1 z_1(t), \beta_1 w_1(t)], \\ x_2(t+1) &= x_2(t) + \min [\alpha_2 z_2(t), \beta_2 w_2(t)], \\ x_3(t+1) &= x_3(t) - z_1(t) - z_2(t) - z_4(t) - z_5(t) - z_6(t) + \gamma_3 x_4(t), \\ x_4(t+1) &= x_4(t) + \min [\alpha_4 z_4(t), \beta_4 w_4(t)], \\ x_5(t+1) &= x_5(t) - w_2(t) - w_4(t) - w_6(t) + \min [\gamma_5 x_6(t), \alpha_5 z_5(t)], \\ x_6(t+1) &= x_6(t) + \min [\alpha_6 z_6(t), \beta_6 w_6(t)], \end{aligned} \right\} (10.22)$$

где α_i , β_i и γ_i — постоянные.

Ограничения, которые должны быть наложены на z_i и w_i , очевидны:

$$z_i \geq 0, w_i \geq 0, z_1 + z_2 + z_4 + z_5 + z_6 \leq x_3, w_2 + w_4 + w_6 \leq x_5.$$

Здравый смысл подсказывает также ограничения:

$$\alpha_1 z_1 = \beta_1 w_1 \leq \gamma_1 x_2,$$

$$\alpha_2 z_2 = \beta_2 w_2,$$

$$\alpha_4 z_4 = \beta_4 w_4,$$

$$\alpha_5 z_5 = \gamma_5 x_6,$$

$$\alpha_6 z_6 = \beta_6 w_6.$$

Посредством этих дополнительных ограничений мы можем полностью исключить w_i и вместо (10.22) получить уравнения

$$\left. \begin{aligned} x_1(t+1) &= x_1(t) + \alpha_1 z_1(t), \\ x_2(t+1) &= x_2(t) + \alpha_2 z_2(t), \\ x_3(t+1) &= x_3(t) - z_1(t) - z_2(t) - z_4(t) - z_5(t) - z_6(t) + \gamma_3 x_4(t), \\ x_4(t+1) &= x_4(t) + \alpha_4 z_4(t), \\ x_5(t+1) &= x_5(t) - \varepsilon_2 z_2(t) - \varepsilon_4 z_4(t) - \varepsilon_6 z_6(t) + \alpha_5 z_5(t), \quad \varepsilon_i = \frac{\alpha_i}{\beta_i}, \\ x_6(t+1) &= x_6(t) + \alpha_6 z_6(t). \end{aligned} \right\} \quad (10.23)$$

Теперь для каждого t ограничения на $z_i(t)$ принимают вид:

$$z_i \geq 0,$$

$$z_1 + z_2 + z_4 + z_5 + z_6 \leq x_3, \quad (10.24)$$

$$\varepsilon_2 z_2 + \varepsilon_4 z_4 + \varepsilon_6 z_6 \leq x_5, \quad (10.25)$$

$$z_1 \leq \eta_1 x_2,$$

$$z_5 = \eta_5 x_6, \quad \eta_i = \frac{\gamma_i}{\alpha_i}.$$

Мы должны выбрать $z_i(t)$ для $t=0, 1, 2, \dots, T-1$ с учетом указанных ограничений и притом так, чтобы максимизировать $x_1(T)$.

В этой постановке задача целиком лежит в области линейного планирования и может быть решена при любой совокупности ограничений различными изящными итеративными процессами типа тех, которые разрабатывались Данцигом и другими.

Однако самый беглый подсчет числа переменных показывает, что даже эта простая трехотраслевая задача приводит к большому числу независимых переменных z_i , подчиненных еще большему числу ограничений. Если же мы пожелаем определить форму оптимальной стратегии (т. е. стратегии распределения, которая максимизирует $x_1(T)$),

исходя из многих различных начальных состояний и принимая различные совокупности коэффициентов пропорциональности (т. е. про- извести *анализ чувствительности*), то столкнемся с вычислитель- ной работой громадных размеров.

Встает вопрос, можем ли мы определить внутреннюю структуру оптимальной стратегии, не зависящую от численных значений?

На этот вопрос мы дадим частичный ответ в следующих пара- графах. Первый наш шаг будет состоять в важном видоизменении задачи.

§ 10.24. Непрерывный вариант. При исследовании уравнения золотых приисков мы уже отмечали, что непрерывные варианты задач легче решать, чем дискретные. В задачах же рассматриваемого сейчас типа непрерывные варианты нередко соответствуют реальной эконо- мической задаче ближе, чем дискретные варианты.

В непрерывном варианте задачи, поставленной в § 10.23, рас- пределения $z_i(t)$ для интервала времени $[t, t+1]$ заменяются распе- делениями $z_i(t)\Delta t$ для интервала времени $[t, t+\Delta t]$. Важные вели- чины $z_i(t)$ будут теперь плотностями распределения. Когда $\Delta t \rightarrow 0$, уравнения (10.23) переходят в следующие:

$$\dot{x}_1(t) = \alpha_1 z_1(t),$$

$$\dot{x}_2(t) = \alpha_2 z_2(t),$$

$$\dot{x}_3(t) = -z_1(t) - z_2(t) - z_4(t) - z_5(t) - z_6(t) + \gamma_3 x_4(t),$$

$$\dot{x}_4(t) = \alpha_4 z_4(t),$$

$$\dot{x}_5(t) = -\varepsilon_2 z_2(t) - \varepsilon_4 z_4(t) - \varepsilon_6 z_6(t) + \alpha_5 z_5(t),$$

$$\dot{x}_6(t) = \alpha_6 z_6(t).$$

Ограничения, накладываемые на z_i , теперь таковы:

$$z_i \geq 0$$

$$z_1 + z_2 + z_4 + z_5 + z_6 \leq \infty, \quad (10.26)$$

$$\varepsilon_2 z_2 + \varepsilon_4 z_4 + \varepsilon_6 z_6 \leq \infty \quad (10.27)$$

$$z_1 \leq \eta_1 x_2,$$

$$z_5 \leq \eta_5 x_6.$$

Теперь ясно, что ограничения, выражаемые неравенствами (10.26) и (10.27), отпадают и заменяются неравенствами

$$x_3 \geq 0, \quad x_5 \geq 0, \quad (10.28)$$

которые выполнялись в дискретном варианте тривиально.

Важность первого неравенства в (10.28) заключается в том, что при $x_3 = 0$ нам нужно дополнительное ограничение

$$z_1 + z_2 + z_4 + z_5 + z_6 \leq \gamma_3 x_4$$

и аналогично при $x_5 = 0$ ограничение

$$\varepsilon_2 z_2 + \varepsilon_4 z_4 + \varepsilon_6 z_6 \leq \alpha_2 z_5.$$

§ 10.25. Некоторые замечания об обозначениях. Прежде чем излагать наш общий теоретический подход, введем некоторые векторно-матричные обозначения. Пусть символы $\mathbf{x}(t)$ и $\mathbf{z}(t)$ обозначают n -мерные векторы-столбцы

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \quad \mathbf{z}(t) = \begin{pmatrix} z_1(t) \\ z_2(t) \\ \vdots \\ z_n(t) \end{pmatrix},$$

а символы A_i, B_j (для встречающихся далее значений i и j) обозначают $m \times n$ -матрицы.

Под неотрицательным вектором $\mathbf{x}(t)$ мы понимаем вектор, все компоненты которого неотрицательны, т. е. $x_i \geq 0$; мы пишем тогда $\mathbf{x} \geq 0$. Неравенство $\mathbf{x} \geq \mathbf{y}$ равносильно неравенству $\mathbf{x} - \mathbf{y} \geq 0$.

В этих обозначениях предыдущие уравнения и ограничения принимают вид

$$\frac{d\mathbf{x}}{dt} = A_1 \mathbf{x} + A_2 \mathbf{z}, \quad \mathbf{x}(0) = \mathbf{c},$$

$$\mathbf{z} \geq 0,$$

$$B_1 \mathbf{z} \leq B_2 \mathbf{x}.$$

Чтобы выразить нашу функцию-критерий в простом виде, введем скалярное произведение векторов

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n x_i y_i.$$

Тогда можно записать

$$x_1(T) = (\mathbf{x}(T), \boldsymbol{\alpha}),$$

где первая компонента вектора $\boldsymbol{\alpha}$ равна 1, а остальные компоненты равны нулю.

Наконец, будем писать

$$\sup_z (\mathbf{x}(T), \boldsymbol{\alpha}) = f(\mathbf{c}, T), \quad (10.29)$$

где f — скалярная функция вектора $\mathbf{c}$ и времени T , т. е.

$$f(\mathbf{c}, T) = f(c_1, c_2, \dots, c_n; T).$$

Заметим, что мы пользуемся в уравнении (10.29) верхней границей (sup), так как максимум может не достигаться. Ниже мы обсудим это подробнее.

§ 10.26. Основное функциональное уравнение. Записывая выше уравнение (10.29), мы молчаливо воспользовались тем крайне важным обстоятельством, что оптимальный выпуск $x_1(T)$ является функцией только начального состояния и продолжительности процесса, коль скоро заданы основные уравнения, управляющие процессом.

Заметим далее, что по самому характеру оптимального решения его продолжение в любом конечном подинтервале, например в $[S, T]$, должно быть оптимальной стратегией для процесса продолжительности $T - S$ с начальным состоянием $c(S)$. Это, конечно, лишь частный случай „принципа оптимальности“, высказанного в § 10.2.

Математическая формулировка этого принципа дает основное функциональное уравнение

$$f(c, S + T) = \max_{[0, S]} f[c(S), T], \quad (10.30)$$

где под $\max_{[0, S]}$ понимается максимизация по всем тем $z(t)$, $0 \leq t \leq S$, которые удовлетворяют ограничениям. Для простоты предположим пока, что максимум достигается.

§ 10.27. Аналогия в бесконечно малом. Чтобы полностью использовать уравнение (10.30), мы сделаем S бесконечно малым и получим нелинейное уравнение с частными производными, эквивалентное уравнению (10.30). Для этого надо сделать некоторые допущения. Допустим, что $c(S) = x(S)$ обладает кусочно непрерывными производными и что f обладает тем же свойством относительно компонент вектора c и T .

Приняв $S = 0$ за точку непрерывности производных, получим

$$\left. \begin{aligned} x(S) &= c + [A_1 c + A_2 z(0)] S + o(S), \\ f(c, S + T) &= f(c, T) + S \frac{\partial f}{\partial T} + o(S), \\ f[c(S), T] &= f\{c + [A_1 c + A_2 z(0)] S, T\} = \\ &= f(c, T) + S \left(A_1 c + A_2 z(0), \frac{\partial f}{\partial c} \right) + o(S), \end{aligned} \right\} \quad (10.31)$$

где

$$\frac{\partial f}{\partial c} = \begin{bmatrix} \frac{\partial f}{\partial c_1} \\ \vdots \\ \frac{\partial f}{\partial c_n} \end{bmatrix}.$$

Подставляя выражения (10.31) в уравнение (10.30) и заставляя S стремиться к 0, мы получим дифференциальную форму уравнения (10.30)

$$\frac{\partial f}{\partial T} = \max_{z(0)} \left[\left(A_1 c + A_2 z(0), \frac{\partial f}{\partial c} \right) \right],$$

где максимум берется по n -мерной области z , определяемой теми ограничениями, которые имеют место при $T=0$.

Аналогия в бесконечно малом важна тем, что она позволяет найти решение в интервале $[0, T + \Delta T]$, если решение уже найдено в интервале $[0, T]$ для *всех* начальных состояний. Так как в большинстве задач этого класса трудности легко преодолеваются для малого T , то тем самым мы получили возможный метод решения, обладающий и вычислительными и теоретическими преимуществами.

Так как мы действовали, основываясь на интуиции и предположениях, то нам крайне важно располагать методом, позволяющим проверить, что мы действительно получили решение.

§ 10.28. Двойственная задача. Возьмем наше основное уравнение в виде

$$\frac{dx}{dt} = Az, \quad x(0) = c$$

с ограничениями вида

$$z \geq 0, \quad Bz \leq x,$$

что всегда можно сделать за счет добавления некоторых переменных. Так как

$$x = c + \int_0^t Az dt,$$

то ограничение можно записать в виде

$$Bz + \int_0^t Cz dt \leq c \quad (C = -A).$$

Аналогично, задача максимизации $(x(T), \alpha)$ равносильна задаче максимизации выражения

$$\int_0^T (Az, \alpha) dt = \int_0^T (z, \alpha') dt.$$

Первоначальную задачу можно поставить тогда следующим образом: определить максимум выражения $\int_0^T (z, \alpha') dt$ для всех z , которые удо-

влетворяют ограничениям

$$z \geq 0, Bz + \int_0^t Cz dt \leq c. \quad (10.32)$$

Вообще говоря, максимума не существует. Это создает определенные трудности, требующие введения интегралов Стильтьеса или, что эквивалентно, дельта-функций. В настоящем элементарном изложении мы предпочитаем не входить в эти сложные подробности.

Пусть $w(t)$ — неотрицательный вектор той же размерности, что и c . Тогда, пользуясь вторым неравенством (10.32), мы получим

$$\int_0^T \left(w, Bz + \int_0^t Cz dt_1 \right) dt \leq \int_0^T (w, c) dt.$$

Пусть символы B' и C' соответственно обозначают транспонированные матрицы B и C . Тогда

$$(Bz, w) = (z, B'w),$$

$$\int_0^T \left(w, \int_0^t Cz dt_1 \right) dt = \int_0^T \left(\int_t^T C'w dt_1, z \right) dt,$$

второй результат получается интегрированием по частям.

Объединяя оба эти результата, найдем

$$\int_0^T \left(w, Bz + \int_0^t Cz dt_1 \right) dt = \int_0^T \left(B'w + \int_t^T C'w dt_1, z \right) dt.$$

Предположим, что можно найти неотрицательный вектор $w = w(t)$, удовлетворяющий неравенству

$$B'w + \int_t^T C'w dt_1 \geq \alpha'. \quad (10.33)$$

Тогда мы получим следующую цепочку равенств и неравенств:

$$\begin{aligned} \int_0^T (w, c) dt &\geq \int_0^T \left(w, Bz + \int_0^t Cz dt_1 \right) dt = \\ &= \int_0^T \left(B'w + \int_t^T C'w dt_1, z \right) dt \geq \int_0^T (\alpha', z) dt. \end{aligned} \quad (10.34)$$

Отсюда ясно, что

$$\min \int_0^T (w, c) dt \geq \max_z \int_0^T (z, \alpha') dt, \quad (10.35)$$

где минимум и максимум берутся по всем $\mathbf{w}$ и $\mathbf{z}$, которые удовлетворяют неравенствам (10.32) и (10.33). Заметим, что если даже существует максимум по $\mathbf{z}$, то это не гарантирует существования минимума по $\mathbf{w}$. Вопрос о том, когда действительно существуют максимумы или минимумы, тесно связан с экономическими предположениями.

Если оба экстремума в уравнении (10.35) равны, то, очевидно, справедливы следующие соотношения (разумеется, с точностью до множеств меры нуль):

$$\left. \begin{aligned} w_i = 0, \quad \text{если} \quad c_i &> \left(Bz + \int_0^t Cz \, dt \right)_i, \\ z_j = 0, \quad \text{если} \quad \alpha'_j &< \left(B'w + \int_t^T C'w \, dt \right)_j. \end{aligned} \right\} \quad (10.36)$$

Теперь установим то важное обстоятельство, что, обратно, любая пара неотрицательных $\mathbf{z}$ и $\mathbf{w}$, удовлетворяющих неравенствам (10.36) и первоначальным ограничениям, дает решение задачи на максимум и минимум.

Чтобы доказать это, заметим, что если справедливы неравенства (10.36), то все соотношения в (10.34) превращаются в равенства. Предположим, что какой-то другой вектор $\mathbf{z}$ удовлетворяет всем ограничениям и что для него

$$\int_0^T (\mathbf{z}, \boldsymbol{\alpha}') \, dt < \int_0^T (\bar{\mathbf{z}}, \boldsymbol{\alpha}') \, dt.$$

Тогда для $\mathbf{w}$, соответствующего вектору $\mathbf{z}$, мы получим противоречие

$$\begin{aligned} \int_0^T (\bar{\mathbf{z}}, \boldsymbol{\alpha}') \, dt &\leq \int_0^T \left(\bar{\mathbf{z}}, B'w + \int_t^T A'w \, dt_1 \right) dt = \\ &= \int_0^T \left(B\bar{\mathbf{z}} + \int_0^t A\bar{\mathbf{z}} \, dt_1, \mathbf{w} \right) dt \leq \int_0^T (\mathbf{c}, \mathbf{w}) \, dt = \int_0^T (\mathbf{z}, \boldsymbol{\alpha}') \, dt. \end{aligned}$$

Следовательно, мы нашли систематическую процедуру для проверки предполагаемого решения. Для данного $\mathbf{z}$ мы определяем $\mathbf{w}$ с помощью неравенств (10.36) и затем убеждаемся, удовлетворяет ли $\mathbf{w}$ или нет заданным ограничениям. Аналитические подробности для одной частной задачи можно найти в литературе [4, 16, 42].

§ 10.29. Более общие задачи. Весьма многие задачи математической физики и математической экономики сводятся к нахождению максимума интеграла вида

$$J(z) = \int_0^T F(x_1, x_2, \dots, x_n; z_1, z_2, \dots, z_m) dt,$$

где x_i связаны с z уравнениями вида

$$\frac{dx_i}{dt} = G_i(x, z), \quad i = 1, 2, \dots, n,$$

и ограничениями вида

$$R_k(x, z) \leq 0, \quad k = 1, 2, \dots, K.$$

К этим задачам можно также подойти [11] с помощью описанного выше метода функциональных уравнений. Так как определение собственных значений в задаче Штурма — Лиувилля

$$\begin{cases} u''(t) + \lambda \varphi(t) u(t) = 0, \\ u(0) = u(1) = 0 \end{cases}$$

равносильно при соответствующих предположениях относительно φ определению относительных минимумов интеграла $\int_0^1 (u'(t))^2 dt$ при

условии $\int_0^1 \varphi u^2(t) dt = 1$, то эта задача может решаться аналогичными методами.

Рассмотрение этого пути, приводящего к нелинейным уравнениям в частных производных для собственных значений, можно найти также в наших статьях [4, 17, 26].

§ 10.30. Задачи регулирования. Особый класс задач, имеющих важное технико-экономическое значение, составляют задачи, возникающие в связи с регулированием систем в данных интервалах времени. В некоторых случаях мы хотим удерживать динамическую систему физического или экономического происхождения в данном состоянии или вблизи его с минимальной стоимостью, причем стоимость состоит из двух частей: первая часть измеряется стоимостью отклонения системы от требуемого состояния, а вторая — стоимостью средств, использованных для регулирования.

Пусть состояние системы изображается вектором $x(t)$, удовлетворяющим линейному дифференциальному уравнению вида

$$\frac{dx}{dt} = Ax + f(t), \quad x(0) = c,$$

где A — постоянная, а $f(t)$ изображает регулирующий вектор.

Пусть $y(f)$ — требуемое состояние системы, а $F(x - y)$ — функционал, измеряющий стоимость отклонения состояния x от y . Аналогично, пусть $G(f)$ — функционал, измеряющий стоимость средств регулирования. Общая стоимость будет равна

$$J(f, c) = F(x - y) + G(f);$$

требуется выбрать f так, чтобы общая стоимость была минимальна.

В зависимости от классов используемых функционалов и классов функций, принятых за допустимые, мы получим вариационные задачи большей или меньшей трудности [5]. Подробное рассмотрение некоторых задач такого типа можно найти в наших с Гликсбергом и Гроссом статьях [33—38] и в издаваемой книге, на которую мы ссылались (см. стр. 239).

Игры на выживание

§ 10.31. Введение. В предыдущих частях мы рассматривали задачи, где человек выступал против природы вещей ¹⁾. Теперь мы покажем, что использованная нами методика применима также в тех случаях, когда имеется налицо конфликт между интересами двух противников.

Задачи, возникающие в этих случаях, побудили фон Неймана создать теорию игр — математическую теорию, проникающую в ряд прикладных областей, как, например, в математическую статистику и математическую экономику (см. гл. 9).

Для иллюстрации применения наших методов мы рассмотрим многоэтапную игру особого рода — „игру на выживание“ [3, 32]. Вслед за кратким обсуждением суммарно нулевого случая мы займемся аналогичным разбором суммарно ненулевой игры ²⁾.

¹⁾ Иногда эти задачи характеризуются выражением „упрямство неодолимых предметов“.

²⁾ Нейман и Моргенштерн (von Neumann J., Morgenstern O., *Theory of Games and Economic Behavior*, 2d ed., Princeton University Press, Princeton, 1947, p. 46—47) дают следующее определение суммарно нулевых и суммарно ненулевых игр:

„При классификации игр важно учитывать, равна ли всегда нулю или нет сумма всех платежей, полученных всеми игроками (в конце игры). Если эта сумма равна нулю, то можно сказать, что игроки платят только друг другу и что с этим не связано создание или разрушение благ. Все игры, в которые действительно играют для развлечения, принадлежат к этому типу. Но экономически важные деятельности большей частью не таковы. В них сумма всех платежей, совокупный общественный продукт, вообще говоря, не будет равна нулю и не будет даже постоянной. Иными словами, она будет зависеть от поведения игроков — участников общественной экономики... Игры первого типа мы назовем *суммарно нулевыми* играми (zero-sum games), игры второго типа — *суммарно ненулевыми* играми (non-zero-sum games)“. — *Прим. ред.*

§ 10.32. Игры на выживание. Рассмотрим суммарно нулевую игру, описываемую матрицей $A = (a_{ij})$, и предположим, что два игрока P и Q начинают ее с начальными суммами p и q соответственно. Предположим, что каждый игрок стремится разорить другого игрока. Так как игра суммарно нулевая, то сумма денег в игре остается постоянной, а именно равной $p + q$. Поэтому для того, чтобы характеризовать состояние игры, достаточно знать сумму денег, находящихся в распоряжении первого игрока P .

Не входя в подробности, далеко не тривиальные в отношении существования оптимальных стратегий, обозначим через $f(x)$ вероятность того, что P победит Q („выживет“) в том случае, когда P располагает суммой x и обе стороны применяют оптимальные стратегии, причем $0 < x < p + q$.

Так как игра суммарно нулевая, то лишь один игрок может разориться за одну партию. Положим

$$f(x) = \begin{cases} 0 & \text{при } x \leq 0, \\ 1 & \text{при } x \geq p + q. \end{cases}$$

Отсюда следует, что $f(x)$ удовлетворяет уравнению

$$f(x) = \min_{p_i} \max_{q_j} \sum_{i,j=1}^N f(x + a_{ij}) p_i q_j = \max_{q_j} \min_{p_i} \sum_{i,j=1}^N f(x + a_{ij}) p_i q_j \quad (10.37)$$

для $0 < x < p + q$.

Из уравнения (10.37) можно получить различные результаты относительно существования и единственности. Исследование здесь немного сложнее, чем в разобранных выше примерах функциональных уравнений [3]. Интересные результаты для сходных игр недавно были получены Шапли [55].

§ 10.33. Суммарно ненулевые игры на выживание. Рассмотрим теперь суммарно ненулевую игру для двух участников, в которой игроку P соответствует матрица $A = (a_{ij})$ и игроку Q — матрица $B = (b_{ij})$. Так как игра суммарно ненулевая, то для описания игры необходимо указать капитал каждого игрока.

Пусть при $x, y > 0$ функция $f(x, y)$ есть вероятность того, что P победит Q в том случае, когда P располагает суммой x , Q — суммой y и оба применяют оптимальные стратегии, и пусть

$$f(0, y) = 0 \quad \text{при } y > 0,$$

$$f(x, 0) = 1 \quad \text{при } x > 0,$$

а значение $f(0, 0)$ не определено. Всегда можно немного изменить игру, чтобы x и y никогда не были одновременно равны нулю.

Функциональное уравнение для $f(x, y)$ имеет вид

$$\begin{aligned} f(x, y) &= \min_{p_i} \max_{q_j} \sum_{i,j=1}^N f(x + a_{ij}, y + b_{ij}) p_i q_j = \\ &= \max_{q_j} \min_{p_i} \sum_{i,j=1}^N f(x + a_{ij}, y + b_{ij}) p_i q_j, \end{aligned}$$

где $x, y > 0$.

Как мы показали [24], это функциональное уравнение приводит к интересным результатам относительно стратегии в суммарно ненулевых играх.

ЛИТЕРАТУРА

1. Arrow K. J., Blackwell D., Girshick M. A., Bayes and Minimax Solutions of Sequential Decision Problems, *Econometrica*, 17, 214—244 (1949).
2. Arrow K. J., Harris T. E., Marschak J., Optimal Inventory Policy, Cowles Commission Paper, 1951.
3. Bellman R., An Introduction to the Theory of Dynamic Programming, Report R-245, The Rand Corporation, Santa Monica, Calif., 1953.
4. Bellman R., Dynamic Programming and Continuous Processes, Report R-271, The RAND Corporation, Santa Monica, Calif., 1954.
5. Bellman R., A Survey of the Mathematical Theory of Time Lag, Retarded Control and Hereditary Processes, Report R-256, The Rand Corporation, Santa Monica, Calif., 1954.
6. Bellman R., The Theory of Dynamic Programming, *Bull. Amer. Math. Soc.*, 60, 503—516 (1954).
7. Bellman R., On Games Involving Bluffing, *Rend. Circ. Mat. Palermo* (2), 1, 1—18 (1952).
8. Bellman R., On the Theory of Dynamic Programming, *Proc. Nat. Acad. Sci. USA*, 38, 716—719 (1952).
9. Bellman R., Some Problems in the Theory of Dynamic Programming, *Econometrica*, 22, 37—48 (1954).
10. Bellman R., On Bottleneck Problems and Dynamic Programming, *Proc. Nat. Acad. Sci. USA*, 39, 947—951 (1953).
11. Bellman R., On Computational Problems in the Theory of Dynamic Programming, „Symposium on Numerical Methods“, *Amer. Math. Soc. Symposia in Applied Mathematics*, 1956.
12. Bellman R., Some Functional Equations in the Theory of Dynamic Programming, *Proc. Nat. Acad. Sci. USA*, 39, 1077—1082 (1953).
13. Bellman R., Dynamic Programming and a New Formalism in the Calculus of Variations, *Proc. Nat. Acad. Sci. USA*, 40, 231—235 (1954).
14. Bellman R., Some Applications of the Theory of Dynamic Programming to Logistics, *Navy Quart. Logist.*, 1, 141—153 (1954).

15. Bellman R., Some Applications of the Theory of Dynamic Programming—
a Review, *Operations Res. Quart.*, June, 1954.
16. Bellman R., Bottleneck Problems, Functional Equations and Dynamic
Programming, *Econometrica*, 23, 73—87 (1955).
17. Bellman R., Monotone Convergence in Dynamic Programming and
the Calculus of Variations, *Proc. Nat. Acad. Sci. USA*, 40, 1073—1075
(1954).
18. Bellman R., Some Functional Equations in the Theory of Dynamic
Programming—I, Point Functions and Point Transformations, *Trans.
Amer. Math. Soc.* (в печати).
19. Bellman R., An Iterative Procedure for the Determination of the Perron
Root of a Positive Matrix, *Proc. Amer. Math. Soc.*, 6, 719—725 (1955)
20. Bellman R., On a Quasi-linear Equation, *Canad. J. Math.* (в печати).
21. Bellman R., Dynamic Programming and a New Formalism in the Theory
of Integral Equations, *Proc. Nat. Acad. Sci. USA*. (в печати).
22. Bellman R., A Problem in the Sequential Design of Experiments, *Sankhya*
(в печати).
23. Bellman R., Decision-making in the Face of Uncertainty—I, *Naval Res
Logist. Quart.*, 1, 230—232 (1954).
24. Bellman R., Decision-making in the Face of Uncertainty—II, *Naval
Res. Logist. Quart.*, 1, 327—332 (1955).
25. Bellman R., Dynamic Programming and Multi-stage Decision Processes
of Stochastic Type, „Symposium on Linear Programming“, Washington,
D. C., December, 1954.
26. Bellman R., Dynamic Programming and a New Formalism in the Calculus
of Variations, *Riv. Nat. Univ. Parma* (в печати).
27. Bellman R., Notes in the Theory of Dynamic Programming—I, Some
Recent Results on the Optimal Inventory Problem, Paper P-630, The Rand
Corporation, Santa Monica, Calif., 1955.
28. Bellman R., Notes in the Theory of Dynamic Programming—II, A Functional
Equation Arising in Allocation Theory, *J. Soc. Indust. Appl. Math.*
(в печати).
29. Bellman R., Notes in the Theory of Dynamic Programming—III, Equipment
Replacement Policy, *J. Soc. Indust. Appl. Math.*, 3, 133—136 (1955).
30. Bellman R., Notes in the Theory of Dynamic Programming—IV, A Variational
Problem with Constraints, *J. Soc. Indust. Appl. Math.* (в печати).
31. Bellman R. and Blackwell D., Some Two-person Games Involving
Bluffing, *Proc. Nat. Acad. Sci. USA*, 35, 600—605 (1949).
32. Bellman R. and Blackwell D., On a Particular Non-zero Sum Game,
рукопись The Rand Corporation, Santa Monica, Calif., 1949.
33. Bellman R., Glicksberg I. and Gross O., On Some Variational
Problems Occurring in the Theory of Dynamic Programming, *Proc. Nat.
Acad. Sci. USA*, 39, 298—301 (1953).
34. Bellman R., Glicksberg I. and Gross O., On Some Variational
Problems in the Theory of Dynamic Programming, *Rend. Circ. Mat.
Palermo* (в печати).

35. Bellman R., Glickberg I. and Gross O., The Theory of Dynamic Programming as Applied to a Smoothing Problem, *J. Soc. Indust. Appl. Math.*, **2**, 82—88 (1954).
36. Bellman R. and Gross O., Some Combinatorial Problems Arising in the Theory of Multi-stage Processes, *J. Soc. Indust. Appl. Math.* (в печати).
37. Bellman R. and Gross O., On Some Nonlinear Integral Equations Occurring in the Theory of Dynamic Programming, *Proc. Nat. Acad. Sci. USA*, **41**, 227—229 (1955).
38. Bellman R. and Gross O., On the Optimal Inventory Equation, *Management Sci.*, **2**, 83—104 (1955).
39. Bellman R., Harris T. E. and Shapiro H. N., Studies on Functional Equations Occurring in Decision Processes, Paper P-382, The Rand Corporation, Santa Monica, Calif., 1952.
40. Bellman R. and Lehman R. S., On the Continuous Gold-mining Equation, *Proc. Nat. Acad. Sci. USA*, **40**, 115—119 (1954).
41. Bellman R. and Lehman R. S., On a Functional Equation in the Theory of Dynamic Programming and its Generalizations, *Riv. Mat. Univ. Parma* (в печати).
42. Bellman R. and Lehman R. S., Studies on Bottleneck Problems in Production Processes, Paper P-492, The Rand Corporation, Santa Monica, Calif., 1954.
43. Bellman R., Fleming W. and Widder D. V., Variational Problems with Constraints, *Ann. Mat. Pura Appl.* (в печати).
44. Bush R. R. and Mosteller C. F., A Mathematical Model for Simple Learning, *Psychol. Rev.*, **58**, 313—325 (1951).
45. Dvoretzky A. J., Kiefer J. and Wolfowitz J., The Inventory Problem—I: Case of Known Distributions of Demand and The Inventory Problem—II: Case of Unknown Distributions of Demand, *Econometrica*, **20**, 187—222 (1952).
46. Dvoretzky A. J., Wald A. and Wolfowitz J., Elimination of Randomization in Certain Statistical Decision Procedures and Zero-sum Two-person Games, *Ann. Math. Statist.*, **22**, 1—21 (1951).
47. Estes W. K., Toward a Statistical Theory of Learning, *Psychol. Rev.*, **57**, 94—107 (1950).
48. Flood M. M., On Stochastic Learning Theory, Paper P-353, The Rand Corporation, Santa Monica, Calif., 1952.
49. Johnson S., Optimal Two- and Three-stage Production Schedules with Setup Times Included, Paper P-402, The Rand Corporation, Santa Monica, Calif., 1953.
50. Johnson S., Karlin S., On Optimal Sampling Procedure for a Problem of Two Populations—I, Paper P-328, The Rand Corporation, Santa Monica, Calif., 1952.
51. Karlin S., A Mathematical Treatment of Learning Models—I, Research Memorandum RM-921, The Rand Corporation, Santa Monica, Calif., 1952.

52. Karlin S. and Shapiro H. N., Decision Processes and Functional Equations, Research Memorandum RM-933, The Rand Corporation, Santa Monica, Calif., 1952.
53. Peisakoff M., More on Games of Survival, Research Memorandum RM-884, The Rand Corporation, Santa Monica, Calif., 1952.
54. Robbins H., Some Aspects of the Sequential Design of Experiments, Bull. Amer. Math. Soc., 58, 527—536 (1952).
55. Shapley L., Stochastic Games, Proc. Nat. Acad. Sci. ,USA, 39, 1095—1100 (1953).

Глава 11

МЕТОДЫ МОНТЕ-КАРЛО

ДЖ. В. БРАУН

*Международная телеметрическая корпорация и
Калифорнийский университет, Лос-Анжелес*

Введение

§ 11.1. Природа методов Монте-Карло. Отличительной чертой методов Монте-Карло является их экспериментальный характер. Подходит ли к данному экспериментальному методу название „метод Монте-Карло“, зависит в большой степени от личного вкуса. Не для всех задач известны методы Монте-Карло, и к конкретным задачам не обязательно существует единственный подход при помощи этих методов. Наоборот, для данной задачи могут существовать разные методы Монте-Карло, не связанные явно между собой.

Для определенности мы будем называть методом Монте-Карло любую процедуру, включающую использование приемов статистической выборки для приближенного решения той или иной математической или физической задачи. В качестве простой иллюстрации рассмотрим опытного игрока в покер, который на основании собственного опыта знает (по крайней мере подсознательно) приближенную вероятность прикупить „стрэт“. Требуемая вероятность может быть, конечно, непосредственно вычислена но, по-видимому, большинство игроков в покер обучались больше на опыте, а не на непосредственных расчетах. В таких условиях про игрока можно сказать, что он воспользовался подходом Монте-Карло.

Методы Монте-Карло имеют древнюю и почетную историю, примеры их применения зарегистрированы по меньшей мере несколько столетий назад. Первые исследователи в области теории вероятностей знали о связи некоторых функциональных уравнений с решениями вероятностных задач, возникающих при изучении азартных игр. В более позднее время физикам стали известны аналогии между вероятностными моделями поведения частиц и различными классическими функциональными уравнениями. Обычно при решении этих задач классическими методами анализа исходили из функционального уравнения, к которому приводит вероятностная модель. Термин „метод Монте-Карло“ стал популярным в последние годы главным образом в связи с задачами статистической механики, диффузии частиц и т. п., когда было замечено, что в случаях, не поддающихся классическому анализу, численный расчет приближенных решений можно иногда осуществить, исходя из вероятностной модели. Аналитическую модель обходят, прослеживая истории большого числа отдельных частиц,

каждая из которых ведет себя случайно в соответствии с вероятностной моделью, причем элементы случайности обеспечиваются каким-либо устройством или таблицами случайных чисел. История каждой частицы может рассматриваться как единичное испытание, а точность результата зависит от статистики, полученной из совокупности многих таких испытаний.

Таким образом, путь решения оказался противоположным классическому, и в первых применениях методов Монте-Карло расчеты проводились в полном соответствии с обычными статистическими моделями физических задач. С тех пор много изобретательности было направлено на изменение статистических моделей и изобретение новых моделей, не вполне соответствующих физической задаче, но позволяющих более экономично рассчитывать приближенные решения исходной задачи. Под влиянием этих приложений математики стали уделять большее внимание поискам вероятностных моделей, которые позволяют приближенно решать функциональные уравнения разных типов, не обращаясь к лежащим в их основе физическим системам. Существуют методы решения для некоторых классов дифференциальных уравнений в частных производных, интегральных уравнений, задач о собственных значениях и систем линейных алгебраических уравнений.

§ 11.2. Основные статистические понятия. В этом параграфе кратко рассматриваются некоторые общие статистические понятия, играющие основную роль в теории методов Монте-Карло. Полное изложение можно найти в обычных курсах, некоторые из которых перечислены в конце главы.

При этом рассмотрении понятие вероятности не определяется; свойства этого понятия описываются соответствующими аксиомами теории вероятностей. Постулированные свойства относятся, вообще говоря, к элементарным соотношениям между вероятностями и интуитивно достаточно очевидны, так что здесь они обсуждаться не будут. Важным определенным понятием является *независимость* событий, которая означает, что вероятность совместного наступления событий равна произведению индивидуальных вероятностей наступления этих событий.

Рассмотрим теперь последовательность повторяемых независимых испытаний, в каждом из которых вероятность наступления события равна p . Если в N испытаниях событие наступило n раз, то для оправдания частотной интерпретации вероятности естественно требовать, чтобы при возрастании N отношение n/N в каком-то смысле приближалось к p . Существуют много теорем относительно приближения n/N к p ; пожалуй, самая основная — это *закон больших чисел*:

$$\lim_{N \rightarrow \infty} \Pr \left\{ \left| \frac{n}{N} - p \right| > \varepsilon \right\} = 0 \quad \text{для любого } \varepsilon > 0.$$

Закон больших чисел утверждает, что для любого интервала, содержащего p , вероятность того, что n/N окажется вне этого интервала, стремится к нулю с ростом N . Именно эта теорема дает статистикам уверенность, что сосчитав n/N при достаточно большом числе испытаний, можно получить статистическую оценку неизвестной p .

События с двумя возможными исходами, например удачей и неудачей, с вероятностями p и $1-p$ соответственно легко обобщаются; при этом получаются случайные величины, принимающие численные значения, и соответствующие *распределения вероятностей*. Говорят, что *случайная величина X имеет дискретное распределение*, если существует конечная или бесконечная последовательность значений x_i , каждому из которых соответствует вероятность p_i , так что

$$\Pr \{X = x_i\} = p_i \geq 0, \quad \sum_i p_i = 1.$$

Несколько более тонко обобщение на непрерывные распределения вероятностей. Говорят, что случайная величина X имеет *плотность вероятностей $f(x)$* , если вероятность попадания X в любой заданный интервал равна интегралу от $f(x)$ по этому интервалу:

$$\Pr \{a \leq X \leq b\} = \int_a^b f(x) dx, \quad f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

Такие распределения вероятностей аналогичны распределениям массы или энергии, встречающимся в других вопросах. Можно представить себе смешанные распределения вероятностей: с дискретными и непрерывными частями (и даже более патологические случаи).

Обобщение понятия распределения вероятностей на многомерный случай не составляет трудностей. Можно рассматривать *случайный вектор* либо с дискретным распределением, либо с плотностью вероятностей, которая определяет вероятность попадания в данную область как интеграл от плотности по этой области.

Важным понятием является понятие *математического ожидания*. Если X — случайная величина с плотностью вероятностей $f(x)$ и если $g(X)$ — некоторая функция, то математическое ожидание $g(X)$, записываемое в виде $E\{g(X)\}$, определяется усреднением величин $g(X)$, взятых с весом, соответствующим распределению X , т. е.

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

если рассматриваемый интеграл существует. Если X имеет дискретное распределение $\Pr \{X = x_i\} = p_i$, то

$$E\{g(X)\} = \sum_i g(x_i) p_i$$

есть аналогичное взвешенное среднее значение величин $g(X)$.

Если в качестве $g(X)$ взять X , то математическое ожидание $E(X)$ называется *средним значением* распределения X в предположении, что интеграл существует. Среднее значение вполне аналогично центру тяжести распределения масс. Другой интересный частный случай получаем, полагая $g(X)$ равным $[X - E(X)]^2$. Величина $E\{[X - E(X)]^2\}$ часто обозначается σ_X^2 и называется *дисперсией* X . Дисперсия есть просто среднее квадратичное отклонение величины X от ее среднего значения, аналогичное моменту инерции распределения масс. Квадратный корень из дисперсии называется *стандартным отклонением* (σ_X) распределения X ; ясно, что в некотором смысле σ_X является мерой разброса распределения.

Необходимо отметить, что операция вычисления математического ожидания линейна, т. е.

$$E\{g(X) + h(X)\} = E\{g(X)\} + E\{h(X)\}$$

и

$$E\{cg(X)\} = cE\{g(X)\};$$

эти свойства следуют прямо из определения.

Так же, как мы представляем себе независимые испытания, в каждом из которых наступление события имеет вероятность p , можно рассматривать выборку из N независимых значений $X_1, X_2, \dots, X_N$ случайной величины с плотностью $f(x)$. В этом случае плотность вероятностей в точке $(x_1, x_2, \dots, x_N)$ будет равна произведению $f(x_1)f(x_2) \dots f(x_N)$. Здесь снова естественно ожидать, что существует связь между средним значением

$$\frac{1}{N} \sum_{\alpha=1}^N g(X_\alpha),$$

вычисленным по выборке, и величиной $E\{g(X)\}$, определенной на основе рассматриваемого распределения вероятностей. Из упомянутого выше свойства линейности ясно, что

$$E\left\{\frac{1}{N} \sum_{\alpha=1}^N g(X_\alpha)\right\} = E\{g(X)\}.$$

В частном случае независимых испытаний, в которых вероятность наступления события равна p , последнее выражение обращается в $E\{n/N\} = p$. Гарантию того, что среднее значение выборки приближается к $E\{g(X)\}$, дает прямое обобщение закона больших чисел: если $E\{g(X)\}$ существует, то

$$\lim_{N \rightarrow \infty} \Pr \left\{ \left| \frac{1}{N} \sum_{\alpha} g(X_\alpha) - E\{g(X)\} \right| > \varepsilon \right\} = 0 \quad \text{для любого } \varepsilon > 0.$$

Так же, как в законе больших чисел, сущность этого результата заключается в следующем: вероятность того, что среднее значение выборки лежит вне произвольного интервала, содержащего $E\{g(X)\}$, стремится к нулю с возрастанием N .

Мало того, что средние значения выборок сходятся к значению $E\{g(X)\}$, но можно даже определить характер этой сходимости. Известно, что при весьма широких условиях распределение величины

$$Y_N = N^{1/2} \left(\frac{1}{N} \sum_{\alpha} g(X_{\alpha}) - E\{g(X)\} \right)$$

с возрастанием N стремится к предельному распределению — гауссовскому или нормальному распределению с плотностью вероятностей вида

$$f(y) = \frac{1}{(2\pi)^{1/2} \sigma} e^{-y^2/2\sigma^2}, \quad \sigma^2 = [\sigma_g(X)]^2.$$

Знание предельного распределения позволяет делать весьма точные предсказания о близости средних значений выборки к неизвестному $E\{g(X)\}$ при больших N , даже если о распределении X известно сравнительно мало.

Оценка интегралов методом Монте-Карло

§ 11.3. Случайная выборка как метод вычисления. Рассмотрим задачу о вычислении интеграла вида

$$\int_{-\infty}^{\infty} g(x) f(x) dx,$$

где $f(x)$ — плотность вероятностей.

Откладывая пока обсуждение вопроса о том, как это сделать, предположим, что случайная выборка $X_1, X_2, \dots, X_N$ произведена в соответствии с плотностью распределения $f(x)$. Так как искомый интеграл равен $E\{g(X)\}$, то из предыдущего параграфа ясно, что величина

$$\frac{1}{N} \sum_{\alpha} g(X_{\alpha})$$

может служить некоторой оценкой интеграла, особенно если N велико. Иначе говоря, случайная выборка может рассматриваться как численный метод.

Существует несколько причин, чтобы рассматривать статистическую оценку интегралов вида

$$\int_{-\infty}^{\infty} g(x) f(x) dx.$$

Наиболее важно то, что многие полезные приложения методов Монте-Карло, по крайней мере неявно, приводят к интегралам такой формы, хотя на первый взгляд они не похожи на интегралы. Далее, такой подход может оказаться полезным для самих задач о вычислении интегралов. Нужно заметить, что присутствие под знаком интеграла функции плотности вероятностей не является ограничением, так как любой определенный интеграл может быть приведен к такому частному виду. Наконец, большая часть изложенного ниже применима одинаково хорошо и к многомерным интегралам, для которых метод Монте-Карло, как мы увидим, может оказаться весьма важным.

Прежде чем обсудить свойства статистической оценки интегралов, следует кратко рассмотреть процесс получения случайной выборки. Необходимо предполагать наличие какого-либо источника случайных событий — например монеты. Хотя существует много остроумных приемов выборки, все они отправляются от тех или иных исходных данных, которые предполагаются случайными и распределение вероятностей которых считается известным. Тогда можно вычислить такие функции от этих исходных данных, что их значения $\{X_a\}$ могут рассматриваться как значения случайной величины с плотностью $f(x)$, если только выполнены сделанные предположения относительно распределения исходных данных. Поэтому оценка

$$\frac{1}{N} \sum_a g(X_a)$$

в конце концов зависит от источника случайных величин; благодаря этой зависимости от экспериментального источника, окончательная оценка также носит экспериментальный характер.

Представление о том, что некоторую часть вычислений приходится сознательно оставлять на волю случая, на первый взгляд действует неприятно. Можно даже считать, что обращение к такому процессу доказывает, что не удалось найти лучшего метода решения задачи. В противоположность не совсем определенным статистическим оценкам результат обычного численного интегрирования выглядит определенным и однозначным. Критическим рассмотрением классических процессов численного интегрирования можно показать, что это отличие несущественно вследствие произвола, связанного с выбором интервала, формулы интегрирования, накопления ошибок округления и т. п. Опытные вычислители подтверждают, что обычные процессы счета, дающие „единственный“ результат, связаны с риском, возможно столь же серьезным, как и риск, присущий статистическим методам.

Тот факт, что экспериментальные методы внесли большой вклад в науку в других вопросах, может смягчить предвзятое отношение некоторых к методам Монте-Карло.

Можно показать, что статистическая оценка

$$\frac{1}{N} \sum_{a=1}^N g(X_a)$$

имеет дисперсию, равную σ_g^2/N , где

$$\begin{aligned} \sigma_g^2 &= \int_{-\infty}^{\infty} [g(x) - E\{g(X)\}]^2 f(x) dx = \\ &= \int_{-\infty}^{\infty} g^2(x) f(x) dx - [E\{g(x)\}]^2. \end{aligned}$$

Следовательно, если σ_g^2 конечно, то стандартное отклонение оценки, равное $\sigma_g/N^{1/2}$, показывает, что ошибка убывает, как $1/N^{1/2}$. Далее, при больших N закон распределения ошибок оказывается приближенно гауссовским, так что для получения вероятностей оценок можно использовать таблицы интеграла вероятностей. Например, с вероятностью приблизительно 0,95 погрешность не превосходит $1,96 \sigma_g/N^{1/2}$.

В некоторых случаях значение σ_g может быть заранее известно или может быть известна верхняя граница для σ_g ; тогда можно сделать соответствующие заключения о распределении ошибок. Если σ_g неизвестно, то, вообще говоря, можно оценить σ_g по тем же $\{X_a\}$, которые используются для оценки искомого интеграла, и получить статистические пределы ошибки. Итак, существуют вполне определенные процедуры для оценки надежности значения интеграла, полученного методом Монте-Карло, вследствие чего возможность получить неединственный результат при расчете не является серьезным неудобством.

§ 11.4. Частный случай. Более подробное изучение будет облегчено рассмотрением частного случая

$$I = \int_0^1 g(x) dx, \quad 0 \leq g(x) \leq 1,$$

в котором распределение $\{X_a\}$ является равномерным в интервале $(0,1)$, а приближенное значение интеграла I , полученное по методу Монте-Карло, равно

$$\frac{1}{N} \sum_a g(X_a).$$

Сравним теперь метод Монте-Карло с простой формулой численного интегрирования, согласно которой оценкой для I служит среднее арифметическое значений $g(x)$, вычисленных в точках $x = 1/2N$, $x = 3/2N$, $x = 5/2N$, и т. д. При таком методе интервал делится на N равных частей и каждый из интервалов длины $1/N$ представляется своей серединой. Важная особенность этого метода численного интегрирования состоит в том, что выборка точек из интервала $(0,1)$ производится систематически, а не случайно, как в методе Монте-Карло. В более тонких классических формулах интегрирования используются группы точек, улучшающих аппроксимацию функции $g(x)$ путем интерполирования вычисленных значений и поэтому позволяющих получить лучшие значения для интеграла при условии, что $g(x)$ удовлетворяет некоторым предположениям. Аналогичные варианты интерполирования могут быть при тех же предположениях применены и в методе Монте-Карло. Таким образом, наиболее существенное различие между классическими методами и методом Монте-Карло — это различие между систематической и случайной выборкой.

В этом месте можно заметить существенное различие в возможных оценках ошибок обоих методов. В обычном процессе численного интегрирования можно получить границы ошибки, основанные на некоторых предположениях о поведении $g(x)$ между систематически выбранными значениями, в то время как в статистическом методе вероятностные оценки ошибки, по существу, не зависят от поведения $g(x)$. При возрастании N систематический метод дает значения, которые обычно сходятся быстрее, чем оценки, полученные при случайной выборке, однако без дополнительных допущений нельзя исключить возможность того, что при каком-то N между сеткой интегрирования и экстремальными значениями $g(x)$ будет неблагоприятное соотношение. С другой стороны, статистический метод позволяет уменьшить возможность такой неприятности, так как можно, не увеличивая числа сосчитанных точек, брать больше случайных цифр для определения каждой точки. Поэтому случайная выборка охраняет процесс счета от нежелательных систематических ошибок, связанных с неизвестными причинами. В этом смысле обращение к случайной выборке означает использование неведения в целях защиты от неведения. Следует, конечно, заметить, что может быть получена также неудачная комбинация случайно выбранных точек, но вычислитель предпочитает риск, который он может оценить хотя бы в вероятностном смысле, риску, которого он не может оценить вообще; при этом есть уверенность, что для всего класса родственных задач такая неприятность почти невозможна. Возможности рандомизации хорошо известны в теории планирования экспериментов, развитой за несколько последних десятилетий; рандомизация устраняет возможные смещения, связанные с вполне систематической выборкой, включая в рассмотрение значительно большее число точек по сравнению с систематической выборкой того же размера. Забегая

вперед, можно отметить, что возможны компромиссные схемы, сочетающие достоинства как систематического, так и вполне случайного способов выборки.

Действительное преимущество случайной выборки становится очевидным, если рассмотреть задачу об оценке интеграла по k -мерному кубу от функции $g(x_1, x_2, \dots, x_k)$, где k очень велико. Если в этом случае для точного учета поведения $g(x_1, x_2, \dots, x_k)$ нужна сетка с размером ячеек $1/m$, то при обычном процессе интегрирования придется вычислить функцию в m^k точках. Если m равно только 20, а $k=6$, m^k будет равно 64 000 000. Нетрудно исчерпать любые счетные возможности. Строить многомерные сетки нелегко; гораздо легче все эти точки (или даже больше) получать единообразной выборкой. Таким образом, метод Монте-Карло позволяет использовать значительное число точек для получения несмещенных оценок требуемого интеграла с дисперсией, как и прежде, равной σ_g^2/N . Следует заметить, что σ_g^2 не связана непосредственно с размерностью k , а зависит от изменения $g(x_1, x_2, \dots, x_k)$ в области интегрирования. Методом Монте-Карло можно оценить многократный интеграл, пытаться вычислить который обычными классическими методами безнадёжно ¹⁾.

Возвращаясь к вычислению интеграла

$$I = \int_0^1 g(x) dx, \quad 0 \leq g(x) \leq 1,$$

можно получить его оценку по методу Монте-Карло, основанную на уравнении

$$I = \int_0^1 \int_0^1 h(x, y) dx dy, \quad \begin{cases} h(x, y) = 1, & \text{если } y \leq g(x), \\ h(x, y) = 0, & \text{если } y > g(x). \end{cases}$$

Интегрирование сначала по y дает предыдущее выражение для I . Для получения оценки по методу Монте-Карло, основанной на второй формулировке, нужно брать пары (X_α, Y_α) значений равномерно распределенной случайной величины и вычислять n/N , где n есть число тех α из общего числа N , для которых $Y_\alpha \leq g(X_\alpha)$. Можно показать,

¹⁾ Целесообразность применения методов Монте-Карло для оценки многомерных интегралов обусловлена не только единообразием счета. Гораздо более существенно то, что порядок погрешности методов Монте-Карло, $N^{-1/2}$, не зависит от размерности куба; если же брать равномерную сетку интегрирования, например разбивать куб на $m^k = N$ равных кубов и помещать узлы интегрирования в центры этих кубов, то погрешность интегрирования будет порядка $N^{-2/k}$. При больших k последняя оценка значительно хуже, чем $N^{-1/2}$. [Подробнее об этом см. статью И. М. Соболя, Многомерные интегралы и метод Монте-Карло, ДАН СССР, 114, № 4 (1957)]. — *Прим. перев.*

что дисперсия этой оценки будет $I(1 - I)/N$, а для статистической оценки

$$\frac{1}{N} \sum_{\alpha} g(X_{\alpha})$$

она будет

$$\frac{1}{N} \left[\int_0^1 g^2(x) dx - I^2 \right].$$

Далее, последнее выражение не превосходит первого и вообще будет меньше, если только не окажется $g(x) \equiv I$. Интуитивное объяснение заключается в том, что $g(X_{\alpha})$ заменяется случайной функцией, принимающей значения 0 или 1, которая с вероятностью $g(X_{\alpha})$ равна 1. Если значения $g(X_{\alpha})$ легко получить, то надо их использовать непосредственно; с другой стороны, возможны положения, в которых легче сравнить Y_{α} с $g(X_{\alpha})$, чем вычислить $g(X_{\alpha})$. Очевидный вывод из этого примера заключается в том, что обычно стоит рассмотреть, нельзя ли перед применением статистического метода снизить кратность интеграла, проинтегрировав его аналитически по некоторым из переменных.

Ясно, что дисперсия σ_g^2/N связана с изменением $g(x)$ в рассматриваемом интервале. Достоинство систематически расположенных точек заключается в том, что между смежными точками расстояние не становится слишком большим. Случайная выборка может сконцентрировать в одной области больше точек, чем нужно, за счет другой области. Метод выборки по группам обеспечивает компромисс между систематической и случайной выборками. Например, допустим, что произведены $N = km$ наблюдений, по m наблюдений в каждой из k групп, причем точки каждой группы выбирались случайно внутри интервала длины $1/k$. При этой процедуре каждая из выбранных частей интервала будет представлена равным числом точек. Статистическая оценка снова вычисляется как среднее значение по N точкам, но дисперсия определяется теперь в виде

$$\frac{1}{N} \left[\int_0^1 g^2(x) dx - k \sum_j I_j^2 \right],$$

где I_j суть входящие в оценку для I слагаемые из j -й группы. Эта дисперсия не превосходит дисперсии, связанной с вполне случайной выборкой, и будет меньше, если изменение $g(x)$ в среднем меньше на частях разбиения, чем во всем интервале $(0,1)$. Обычно за это улучшение приходится расплачиваться снижением точности статистической оценки дисперсии.

В более общем случае выборки по группам допустим, что произведены

$$N = n_1 + n_2 + \dots + n_k$$

наблюдений в интервалах длины $L_1, L_2, \dots, L_k$. Наблюдения j -й группы надо брать с весом L_j/n_j . Таким образом составленная оценка будет по-прежнему оценкой для I , но дисперсия будет зависеть от группировки и распределения наблюдений. Для реализации минимума дисперсии наблюдения должны распределяться по группам пропорционально $L_j\sigma_j$, где σ_j — стандартное отклонение $G(X)$ в j -й группе.

К методу выборки по группам близок прием, заключающийся в выборке X из распределения, соответствующего плотности, отличной от $f(x)$ ¹⁾. Допустим, что в нашем распоряжении имеются значения $\{X_\alpha\}$ случайной величины, распределенной с плотностью $f^*(x)$; тогда

$$I = \int_0^1 g(x)f(x)dx = \int_0^1 \frac{g(x)f(x)}{f^*(x)} f^*(x)dx$$

и, следовательно, I можем оценить при помощи величины

$$\frac{1}{N} \sum_{\alpha} \frac{g(X_{\alpha})f(X_{\alpha})}{f^*(X_{\alpha})}$$

при условии, что $f^*(x)$ не обращается в нуль там, где $f(x) \neq 0$. Если о функциях $f(x)$ и $g(x)$ известно достаточно много, то можно найти другие плотности распределения, для которых дисперсии оценки будут значительно меньше, чем для случая выборки из распределения, соответствующего $f(x)$.

Знание структуры задачи позволяет выбрать более эффективный план выборки. Часто могут быть использованы приемы последовательного анализа, позволяющие более эффективно планировать последующую выборку на основе информации, полученной при предшествующих наблюдениях.

§ 11.5. Тривиальный числовой пример. Допустим, что требуется оценить интеграл

$$I = \int_0^1 x dx$$

методами Монте-Карло и не известно, что $I=0,5$. После десяти случайных наблюдений в интервале $(0,1)$ путем выбора из таблицы случайных чисел двух десятичных цифр для каждого наблюдения получены следующие результаты:

$$X=0,69; 0,37; 0,39; 0,97; 0,66; 0,51; 0,60; 0,41; 0,76; 0,09.$$

¹⁾ Этот прием называется обычно существенной выборкой (importance sampling). Подробности о нем можно найти в статье Кана: Kahn, Use of different Monte Carlo sampling techniques, Symposium on Monte Carlo methods, New York, 1956, pp. 146—190. — *Прим. перев.*

Среднее значение

$$\frac{1}{10} \sum_{\alpha} X_{\alpha}$$

равно 0,545, что не является неприемлемым приближением к 0,5 при таком небольшом числе наблюдений. Отметим, что эти же самые случайные числа могли бы служить для обеспечения выборки из 10 наблюдений, состоящей из пяти наблюдений в интервале (0; 0,5) и пяти в интервале (0,5; 1). Это может быть достигнуто делением всех значений на 2 и добавлением 0,5 к каждому из пяти последних результатов деления. Получается следующая выборка:

$$X = 0,345; 0,185; 0,195; 0,485; 0,330; 0,775; 0,800; \\ 0,705; 0,880; 0,545.$$

Среднее значение этой выборки по группам есть 0,5225; ошибка в этом случае ровно в два раза меньше ошибки, полученной при простой выборке. Это, конечно, не случайно; в этом примере любые исходные случайные числа при таком же использовании дали бы половинную ошибку в случае выборки по группам. Далее, следует заметить, что если бы при выборке по группам использовалось одно наблюдение в каждой одной десятой части интервала, то ошибка была бы равна одной десятой ошибки в первоначальном случае. Такую выборку можно осуществить, если брать одну десятую каждого значения X_{α} и добавлять 0 к первому результату, 0,1 — ко второму, 0,2 — к третьему и т. д. Полученная выборка состоит из последовательности:

$$0,069; 0,137; 0,239; 0,397; 0,466; 0,551; 0,660; 0,741; 0,876; 0,909.$$

Среднее значение, как указывалось выше, есть 0,5045. Заметим, однако, что в этом последнем случае нельзя получить статистической оценки погрешности, так как в каждый интервал попадает всего одно наблюдение.

Уменьшение ошибки, связанное с выборкой по группам, в этом примере объясняется тем, что $g(X)$, равная в этом случае самому X , имеет в любом интервале изменение, пропорциональное длине интервала. Ограничивая случайное изменение малыми интервалами, мы уменьшаем соответственно изменение результата в этом частном примере.

§ 11.6. Приложения, приводящие к кратным интегралам. Очевидно, разные задачи прикладной математики могут приводить к кратным интегралам, которые должны быть вычислены тем или иным методом; менее очевидно, что во многих задачах появляются кратные интегралы от эмпирических функций, получаемых выборкой из физических совокупностей. Такие задачи возникают ежедневно в связи

с методами инспекции, контролем качества, анализом выборок, изучением почвы, полевыми экспериментами и т. д. до бесконечности.

Рассмотрим в качестве упражнения задачу, возникающую перед промышленным предприятием, которое купило большую партию химического сырья с разными, в известных пределах допустимыми примесями. Если представлять себе это сырье в виде большой кучи, лежащей во дворе, то задача сводится к вычислению нескольких кратных интегралов по объему, занимаемому кучей. Ясно, что для исследования можно брать пробы из разных частей кучи и оценивать процентные содержания примесей как средние от значений, полученных при выборках. Это выглядит весьма просто, но есть опасность, что самый легкий и очевидный способ выборки — взятие небольших проб с поверхности кучи — может не дать хороших оценок; если вещество неоднородно, то могут быть вполне определенные, но неизвестные соотношения между процентным содержанием некоторой примеси и расположением в куче. Процесс выборки из такой совокупности не прост, но в этом случае он может быть осуществлен с помощью случайных чисел и специальных проб. Такая хорошо спланированная выборка идентична в некотором смысле оценке интеграла методом Монте-Карло и, по существу, может привести ко всем специальным процедурам для снижения дисперсии.

Другой класс задач с квадратурами возникает непосредственно из вероятностных рассуждений. В задачах статистического происхождения обычно требуется найти функцию распределения вероятностей статистики, вычисленной как функция большого числа наблюдений, о которых предполагается, что они удовлетворяют некоторому постулированному закону распределения. Эта задача непосредственно сводится к задаче вычисления некоторых определенных многомерных интегралов; можно получать решения методом Монте-Карло, производя искусственные выборки и вычисляя соответствующую статистику, причем этот процесс можно повторять столько раз, сколько требуется для получения желаемой точности. Такие эксперименты часто проводились при развитии современной теории статистических выводов.

Задачи исследования операций дают много примеров, относящихся к последнему классу задач. При изучении военных ситуаций относительно легко составлять сложные описательные схемы, включающие вероятности, связанные с такими случайными элементами, как поражение целей, разрушения при попадании в цель и т. п. Способность аналитика строить гипотетические модели быстро обгоняет имеющиеся возможности математического анализа для изучения решений. Во многих таких случаях методы Монте-Карло оказались единственным практически приемлемым средством расчета. Методами Монте-Карло изучались также задачи, возникающие при исследовании производственных операций, и нужно ожидать, что число таких приложений возрастет.

Методы Монте-Карло, связанные с процессами диффузии

§ 11.7. Равномерное случайное блуждание на плоскости. Откладывая пока мотивировку изложения, представим себе частицу, которая вынуждена передвигаться по узлам сетки с целочисленными координатами (m, n) на плоскости. За каждый шаг частица будет передвигаться в один из соседних узлов вверх, вниз, вправо или влево от ее положения. Если предполагать, что вероятности всех четырех возможных переходов из (m, n) в $(m+1, n)$, $(m-1, n)$, $(m, n+1)$ или $(m, n-1)$ равны $1/4$ и каждый новый переход определяется случайно, независимо от положения или прошлого частицы, то частица будет описывать случайную траекторию, состоящую из последовательности таких единичных шагов. Допустим, что частица вышла из узла $(0, 0)$; тогда для каждого узла (m, n) и для каждого целого положительного числа s будет определена вероятность $P(m, n, s)$ того, что после s шагов частица окажется в узле (m, n) . Ясно, что вероятность $P(m, n, s)$ должна удовлетворять разностному уравнению

$$P(m, n, s+1) = \frac{1}{4} [P(m+1, n, s) + P(m-1, n, s) + P(m, n+1, s) + P(m, n-1, s)],$$

так как для того, чтобы после $(s+1)$ шага оказаться в узле (m, n) , частица должна была после s -го шага находиться в одном из четырех соседних узлов и иметь вероятность перехода в (m, n) , равную $1/4$ для каждого из этих положений.

Предыдущее разностное уравнение легко преобразуется в эквивалентное уравнение

$$\begin{aligned} P(m, n, s+1) - P(m, n, s) = \\ = \frac{1}{4} \{ [P(m+1, n, s) - P(m, n, s)] + [P(m-1, n, s) - P(m, n, s)] + \\ + [P(m, n+1, s) - P(m, n, s)] + [P(m, n-1, s) - P(m, n, s)] \}, \end{aligned}$$

которое связывает первую разность P относительно s со вторыми разностями P относительно m и n . Последняя форма разностного уравнения подобна уравнению теплопроводности

$$\frac{\partial P}{\partial t} = K \left\{ \frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2} \right\},$$

которое обычно сокращенно записывается в виде $P_t = K \nabla^2 P$. Мы не будем приводить здесь обоснование предельного перехода, связывающего разностное уравнение с дифференциальным, ограничившись указанием на чисто формальное сходство.

Как показано выше, распределение вероятностей частицы, подчиняющейся установленным простым правилам случайного поведения, должно как функция положения и пройденного числа шагов удовлетворять разностному уравнению, аналогичному известному дифферен-

циальному уравнению в частных производных. Если вместо того, чтобы начинать с частицы, выходящей из узла $(0, 0)$ при $s=0$, задавать распределение вероятностей $P(m, n, 0)$, которое определяет вероятность нахождения начальной частицы в (m, n) , то функция $P(m, n, s)$ может быть истолкована по-прежнему и по-прежнему будет удовлетворять тому же разностному уравнению. Начальное распределение $P(m, n, 0)$ соответствует произвольной начальной функции, порождающей частное решение разностного уравнения.

Случайная выборка, соответствующая постулированному здесь случайному блужданию, может быть осуществлена путем установления соответствия между интервалами изменения случайных чисел и конкретными решениями, которые нужно принимать при каждом шаге. Так, выбирая первое случайное число, можно определить начальную точку на сетке, выбирая следующую группу случайных чисел, можно определить первый переход и т. д. Последовательность таких выборов определит траекторию частицы, которую обычно называют „историей“ частицы. В принципе такая история может быть продолжена как угодно далеко. Выборка из таких историй может быть использована для оценки $P(m, n, s)$ в любой конкретной точке сетки (m, n) и при любом s , до которого истории были продолжены; для этого подсчитывается число историй, для которых соответствующие частицы попали в (m, n) после s -го шага, и делится на размер выборки. Такой статистический процесс является методом Монте-Карло, по существу, такой же структуры, как метод Монте-Карло для оценки интегралов. Ввиду дискретного характера рассматриваемых распределений вероятностей функция $P(m, n, s)$ представляется кратной суммой по многомерному пространству, а не кратным интегралом, но основания для этого те же.

Если обратить внимание на то, что разностные уравнения того типа, которому удовлетворяет $P(m, n, s)$, часто используются для приближенного численного решения соответствующего дифференциального уравнения, то легко видеть, что метод Монте-Карло позволяет построить численный метод решения дифференциального уравнения, если использовать сетку с надлежащим размером ячеек.

Таблица 11.1 (стр. 290) дает простой пример выборки 10 историй, начинающихся с $(0, 0)$ и продолженных до $s=10$. Источником случайных действий служила таблица случайных чисел, причем цифрам 0 и 4 соответствовал шаг вправо, 1 и 5 — шаг вверх, 2 и 6 — шаг влево, 3 и 7 — шаг вниз, а цифры 8 и 9 пропускались. Можно исследовать эти последовательности, чтобы посмотреть, какие из этих историй кажутся „необычными“ в том или другом отношении. Именно это неуловимое свойство случайных последовательностей трудно воспроизвести при попытках изобретать псевдослучайные последовательности.

§ 11.8. Уравнение Лапласа. Предположим теперь, что случайные блуждания описанного выше типа начинаются внутри области,

Таблица 11.1

ДЕСЯТЬ ИСТОРИЙ ДВУМЕРНОГО СЛУЧАЙНОГО БЛУЖДЕНИЯ С ЧИСЛОМ $s=10$ В КАЖДОМ СЛУЧАЕ

история s	1	2	3	4	5	6	7	8	9	10
0	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
1	(0, -1)	(-1, 0)	(0, -1)	(0, -1)	(0, 1)	(0, -1)	(-1, 0)	(1, 0)	(-1, 0)	(0, -1)
2	(1, -1)	(0, 0)	(1, -1)	(1, -1)	(-1, 1)	(-1, -1)	(-2, 0)	(0, 0)	(0, 0)	(0, -2)
3	(2, -1)	(0, 1)	(0, -1)	(2, -1)	(-1, 2)	(-2, -1)	(-2, 1)	(-1, 0)	(-1, 0)	(0, -1)
4	(3, -1)	(0, 2)	(-1, -1)	(2, 0)	(0, 2)	(-3, -1)	(-1, 1)	(0, 0)	(0, 0)	(0, -2)
5	(2, -1)	(-1, 2)	(0, -1)	(2, -1)	(1, 2)	(-3, -2)	(-1, 0)	(0, 1)	(1, 0)	(0, -3)
6	(1, -1)	(0, 2)	(1, -1)	(2, 0)	(1, 3)	(-4, -2)	(-2, 0)	(1, 1)	(2, 0)	(-1, -3)
7	(0, -1)	(0, 3)	(1, -2)	(3, 0)	(1, 2)	(-4, -3)	(-2, 1)	(0, 1)	(2, 1)	(-1, -4)
8	(-1, -1)	(0, 2)	(0, -2)	(2, 0)	(1, 1)	(-5, -3)	(-3, 1)	(1, 1)	(1, 1)	(0, -4)
9	(-1, -2)	(-1, 2)	(0, -3)	(3, 0)	(0, 1)	(-4, -3)	(-3, 0)	(0, 1)	(1, 0)	(0, -3)
10	(0, -2)	(-1, 3)	(1, -3)	(4, 0)	(1, 1)	(-3, -3)	(-3, -1)	(1, 1)	(2, 0)	(0, -2)

ограниченной определенной замкнутой границей, состоящей из точек сетки (m_i, n_i) . Здесь мы не будем пытаться уточнить понятие граничных точек на дискретной сетке, ограничиваясь интуитивным представлением о них. Для целей настоящего рассмотрения достаточно, чтобы граница отделяла внутреннюю часть от внешней; это обеспечивается тем, что траектория любого допустимого блуждания описанного типа между любыми двумя точками, из которых одна внутренняя, а другая внешняя, содержит граничную точку. Изменим теперь описанные выше случайные блуждания, а именно будем считать, что при первом же выходе на границу блуждание заканчивается, и будем отмечать соответствующую граничную точку как (m_i, n_i) . На языке физиков эта граница называется „поглощающей границей“ для частицы. Можно доказать, что блуждания в конце концов заканчиваются в том смысле, что вероятность оставаться всегда внутри области равна нулю. Поэтому удобно с любой начальной точкой (m, n) связать функцию $P(m_i, n_i, m, n)$, определяющую вероятность того, что частица, вышедшая из узла (m, n) , заканчивает свою историю в граничной точке (m_i, n_i) . Далее, если функция $U(m_i, n_i)$ определена сначала на границе, то оказывается интересным рассмотреть математическое ожидание величины U относительно распределения вероятностей конечных значений как функцию (m, n) . Обозначив эту функцию через $V(m, n)$, легко увидим, что

$$V(m, n) = \sum_i U(m_i, n_i) P(m_i, n_i, m, n).$$

Другими словами, если всем блужданиям, оканчивающимся в (m_i, n_i) , приписывается конечное значение $U(m_i, n_i)$, то $V(m, n)$ есть математическое ожидание конечного значения для блужданий, начинающихся в (m, n) .

Как и можно было ожидать, $V(m, n)$ удовлетворяет разностному уравнению

$$V(m, n) = \frac{1}{4} [V(m+1, n) + V(m-1, n) + V(m, n+1) + V(m, n-1)]$$

с граничным условием $V(m_i, n_i) = U(m_i, n_i)$. Это следует непосредственно из того, что четвертая часть случайных блужданий, начинающихся в (m, n) , становится новыми случайными блужданиями, начинающимися в каждом из четырех соседних узлов сетки.

Так же, как и в случае, рассмотренном в предыдущем параграфе, это разностное уравнение можно преобразовать к более привычному виду

$$[V(m+1, n) - 2V(m, n) + V(m-1, n)] + [V(m, n+1) - 2V(m, n) + V(m, n-1)] = 0,$$

показывающему, что сумма вторых разностей по m и n должна исчезать всюду в области. Соответствующее дифференциальное уравнение есть, конечно, уравнение потенциала (уравнение Лапласа)

$$\nabla^2 V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$$

с граничными условиями обычного типа, задающими значения V на границе.

Рассмотренное здесь простое случайное блуждание может лечь в основу статистической оценки $V(m, n)$ для внутренней точки (m, n) : частица начинает свою историю из точки (m, n) , и каждой истории приписывается значение $U(m_i, n_i)$, соответствующее граничной точке (m_i, n_i) , в которой блуждание заканчивается. Тот факт, что расчет дает приближенное значение решения в одной точке, не требуя при этом аппроксимации решения по всей области, резко отличается от обычного положения, встречающегося при большинстве стандартных методов решения уравнения Лапласа. Необходимо заметить, однако, что любое из блужданий, использованных при оценке $V(m, n)$, может быть использовано и при оценке решения в любой точке, через которую проходит это блуждание.

Одно из достоинств метода Монте-Карло заключается в том, что он обеспечивает одновременный учет всевозможных граничных условий, задаваемых на одной и той же границе. Так как блуждание не зависит от граничных условий, то можно использовать одни и те же конечные точки для получения приближенных решений с разными граничными условиями путем усреднения конечных значений, получающихся для каждого граничного условия. С другой точки зрения число блужданий, заканчивающихся в заданной точке границы (m_i, n_i) , пропорционально $P(m_i, n_i, m, n)$ — вероятности окончания в точке (m_i, n_i) истории, начавшейся в (m, n) . Так как

$$V(m, n) = \sum_i U(m_i, n_i) P(m_i, n_i, m, n),$$

то ясно, что приближенное распределение конечных точек является фундаментальным решением по отношению к общему решению с произвольными граничными условиями. Функция $P(m_i, n_i, m, n)$ эквивалентна функции Грина для рассматриваемой области (см. гл. 6).

Представляет некоторый интерес тот факт, что любое множество историй, заканчивающихся на заданной границе, может быть использовано для приближенного решения уравнения Лапласа с граничными условиями на границах, содержащихся внутри заданной границы, ибо каждая траектория этого множества может определить конечную точку на любой границе, содержащейся внутри первоначальной.

При решении уравнения Лапласа методом Монте-Карло можно многими путями применять принципы выборки по группам. Простей-

ший способ выборки по группам заключается в том, что начинают $N/4$ блужданий в каждом из четырех соседних с (m, n) узлов, вместо того, чтобы строить N блужданий из (m, n) . (Напомним, что вероятность каждого соседнего узла равна $1/4$.) Можно построить обобщение такого способа выборки по группам для любой границы, расположенной между точкой (m, n) и заданной границей, при условии, что общее решение для (m, n) и промежуточной границы уже известно. Это сводится к утверждению, что все блуждания, начинающиеся в точке (m, n) , должны пересечь промежуточную границу прежде чем достигнут внешней границы; если распределение вероятностей окончания в точках промежуточной границы уже известно, то блуждания можно начинать с промежуточной границы в количествах, пропорциональных известному распределению вероятностей. Если промежуточная граница близка к заданной границе, то блуждания будут в среднем значительно короче, и, кроме того, для достижения определенной точности потребуются меньше блужданий.

Выборку по промежуточным границам можно модифицировать, изменяя пропорции начальных блужданий и для компенсации приписывая результатам надлежащие веса. Искусным применением такой процедуры можно выделить при выборке более существенные части границы (где, например, $U(m_i, n_i)$ изменяется быстро) за счет менее существенных частей и, следовательно, улучшить точность результата. Процесс определения весов очень прост: если предполагаемая вероятность получить блуждание, соответствующее некоторой выборочной группе, равна p_α , а в действительности этой группе соответствует часть r_α всех блужданий, то результаты, связанные с этой группой, необходимо брать с весом p_α/r_α .

Возможны и другие модификации процесса выборки при условии, что делается надлежащая компенсация в счете. Предположим, например, что решено изменить вероятности перехода в некоторой точке. Это изменение может быть учтено, если приписать результатам веса так, как было описано выше: если p_α есть гипотетическая вероятность получения заданного исхода на любой стадии, а использованный процесс выборки реализует этот исход с вероятностью r_α , то любым результатам, связанным с этим исходом, нужно приписать вес p_α/r_α . Подобную процедуру „смещения“ можно применять на каждом этапе случайного процесса при условии, что вычисляются произведения соответствующих весов до конца каждой истории. Такие процедуры применяются с целью повышения эффективности выборки.

- Весьма досадной чертой немодифицированного случайного блуждания, на котором основывалось предыдущее рассмотрение, является медленная сходимость. Из самой природы случайного процесса следует, что требуется большое число шагов, чтобы прийти куда-либо. По этой причине обычно выгодно рассмотреть использование специальных методов выборки по группам. Результатом попыток развить эти идеи несколько дальше является новый (насколько автору известно)

метод Монте-Карло для приближенного решения рассматриваемого дифференциального уравнения без обращения к соответствующему разностному уравнению с сеткой. Для этого метода основным является тот факт, что значение решения уравнения Лапласа в любой внутренней точке равно среднему арифметическому его значений на любой окружности с центром в этой точке. Иначе говоря, предельные случайные блуждания, соответствующие бесконечно малой сетке, которые начинаются в некоторой точке, будут пересекать окружность с центром в этой точке с одинаковой вероятностью в любом направлении. Если исходная точка находится далеко от границы, то можно просуммировать результаты всех случайных движений до пересечения с произвольной окружностью с центром в начальной точке, выбирая на окружности точки под случайными равномерно распределенными углами по отношению к начальной точке. За один скачок траектория оказывается значительно ближе к границе, чем начальная точка, не требуя подробного изучения тягостных блужданий до пересечения с окружностью. Годится любая окружность, лежащая внутри границы с центром в начальной точке; ее следует выбирать по возможности больше. Повторение этого процесса порождает последовательность, сходящуюся к границе. Вопрос о том, как лучше закончить процесс вблизи границы, требует дальнейшего изучения, но похоже, что изложенная схема может служить основой для очень эффективного метода Монте-Карло ¹⁾.

§ 11.9. Обобщения. Простое случайное блуждание по прямоугольной сетке легко обобщается в нескольких направлениях. Все предшествующее рассмотрение допускает распространение на пространства с числом измерений, большим чем два, где потребности в практических методах расчета более значительны. Другие обобщения, которые будут рассмотрены ниже, также применимы в том случае, когда число измерений больше двух.

Видоизменение случайного блуждания с использованием разных вероятностей перехода, зависящих от положения частицы, получается при замене оператора Лапласа $\nabla^2 U$ более общими эллиптическими дифференциальными операторами вида

$$L(U) = \beta_{11} \frac{\partial^2 U}{\partial x^2} + 2\beta_{12} \frac{\partial^2 U}{\partial x \partial y} + \beta_{22} \frac{\partial^2 U}{\partial y^2} + 2\alpha_1 \frac{\partial U}{\partial x} + 2\alpha_2 \frac{\partial U}{\partial y},$$

где коэффициенты являются функциями от x и y . Для приближенного решения неоднородного уравнения $L(U) = g(x, y)$ видоизменение процесса счета состоит в том, что суммируются значения функции $g(x, y)$ по всем промежуточным точкам траектории и к конечному результату прибавляется величина, зависящая от конечного положения

¹⁾ Строгое обоснование и развитие этой идеи Брауна изложено в статье Маллера: Muller M. E., Some continuous Monte Carlo methods for the Dirichlet problem, Ann. of Math. Statistic, 27, No 3, 569 (1956). — *Прим. перев.*

точки. Параболические уравнения вида $\partial U/\partial t = L(U)$ получаются в тех случаях, когда процесс диффузии начинается с начальной плотностью и достигнутые при блуждании точки рассматриваются как функции числа шагов. Легко включить в такие процессы отражающие и поглощающие границы. Краевые задачи эллиптического типа соответствуют, по существу, асимптотическим или стационарным решениям параболических уравнений и, следовательно, зависят от конечных состояний, а не от числа шагов, нужных для достижения данных состояний.

В качестве обобщения в другом направлении рассмотрим вычисление интеграла

$$\int_0^a A(x) U(x) dx,$$

где функция $A(x)$ задана, а $U(x)$ удовлетворяет интегральному уравнению

$$U(x) = \int_0^a U(y) K(x, y) dy + B(x)$$

с заданными $K(x, y)$ и $B(x)$. Это интегральное уравнение может быть выведено из вероятностной модели, в которой предполагается, что частицы попадают в слой вещества $0 \leq x \leq a$. Допустим, что $K(x, y)$ есть условная плотность вероятностей для столкновения в точке x , если предыдущее столкновение было в точке y , $A(x)$ — вероятность вылета из слоя после столкновения в точке x и $B(x)$ — плотность вероятностей первого столкновения в точке x . Таким образом, предполагается, что история частицы состоит из последовательности столкновений, определяемой вероятностью $A(x)$ вылета из слоя после столкновения в точке x , условной плотностью вероятностей $K(x, y)$ для следующего столкновения в точке y и вероятностью

$$1 - A(x) - \int_0^a K(x, y) dy$$

исчезновения в точке x . Тогда функцию $U(x)$ можно интерпретировать как безусловную плотность вероятностей, связанную со столкновениями в точке x ; интегральное уравнение, которому удовлетворяет $U(x)$, выражает тот факт, что столкновения в точке x могут быть либо первичными столкновениями, представленными членом $B(x)$, либо вторичными столкновениями, происходящими с плотностью $K(x, y)$ в x после столкновений в каждой точке y . Тогда выражение

$$\int_0^a A(x) U(x) dx$$

дает вероятность того, что частица в конце концов вылетит из слоя, а не исчезнет при каком-либо столкновении.

Интенсивно изучались более сложные варианты рассмотренной выше задачи, которые составляют большую часть практических задач, решенных методами Монте-Карло. При изучении замедления нейтронов этот процесс соответствует обычно начальному выбору нейтронов согласно начальному распределению углов и энергии с последующим выбором точки первого столкновения; затем выбирается тип столкновения, согласно которому частица либо поглощается, либо испытывает упругое или неупругое рассеяние. Наконец с новым углом и новой энергией рассеянные частицы начинают новый цикл. Шаги повторяются до тех пор, пока нейтрон не поглощается, не отражается или не вылетает или пока его энергия не становится пренебрежимо малой. При помощи методов Монте-Карло были исследованы еще более сложные примеры процессов такого типа. Во многих таких задачах вероятность, соответствующая интересующим нас историям частиц, столь мала, что расчет методом Монте-Карло по естественной схеме сходной задачи практически невозможен. В этих случаях для повышения эффективности историй использовались многие остроумные способы выборки.

Кроме задач о замедлении, методы Монте-Карло применялись для определения собственных значений уравнений Шредингера; некоторые методы основывались на видоизменении процесса диффузии путем введения подходящих экспоненциальных весов, другие основывались на процедурах, не связанных явно с диффузией.

В качестве последнего примера рассмотрим случайное блуждание частицы, которая может находиться в одном из k положений, условно обозначенных $S_1, S_2, \dots, S_k$. Пусть для частицы, находящейся в S_j , определены вероятности $a_{ij} \geq 0$ перехода из S_j в S_i и вероятность $p_j \geq 0$ остановиться в S_j , так что

$$p_j + \sum_i a_{ij} = 1$$

при любом j . Обозначив через B_{ij} вероятность того, что частица, вышедшая из S_j , в конце концов остановится в S_i , легко видеть, что $\{B_{ij}\}$ должны удовлетворять уравнениям

$$B_{ij} = \sum_{r=1}^k a_{rj} B_{ir} + \delta_{ij} p_j, \quad i, j = 1, 2, \dots, k,$$

ввиду того, что частица, выходящая из S_j , либо перейдет в S_r с вероятностью a_{rj} и затем остановится в S_i с вероятностью B_{ir} , либо должна сразу остановиться с вероятностью p_j , которая учитывается в B_{ij} только в случае $i=j$. После замены $C_{ij} = B_{ij}/p_i$ уравнения можно переписать в матричной форме $C(I-a) = I$. Следовательно, описанное блуждание позволяет статистически обращаться матрицу $I-a$; для этого нужно строить блуждания частиц, используя слу-

чайные числа для определения переходов, и подсчитывать число остановок в каждом состоянии. Этот процесс можно применять не только для обращения матрицы, но и непосредственно для решения заданной системы уравнений; в этом случае при остановке в S_j будем записывать значение T_i и осреднять наблюдаемые T_i ; тогда математическое ожидание результата Y_j для историй, начинающихся из S_j , должно удовлетворять уравнениям

$$Y_j = \sum_{r=1}^k a_{rj} Y_r + p_j T_j, \quad j = 1, 2, \dots, k.$$

Однако эти же истории позволяют оценивать и обратную матрицу. Гибкость метода увеличивается другими модификациями, например включением весов, соответствующих состояниям, через которые проходит частица. Необходимо отметить, что разностные схемы, аппроксимирующие уравнения типа уравнения Лапласа, в действительности эквивалентны системе многих уравнений, и простое случайное блуждание, описанное в начале параграфа, является частным случаем только что изложенной ситуации.

Методы выборки из заданных распределений

§ 11.10. Источники случайных чисел. До сих пор очень мало было сказано о самом процессе выборки. Доказательство того, что данный источник случайных событий воспроизводит действительно случайные события с определенными распределениями является обычно весьма тонким и сложным. При использовании приспособлений, подобных монетам, игральным костям, рулеткам, по существу, невозможно заранее утверждать, что поведение получаемых случайных величин соответствует теоретической модели; можно лишь воспользоваться статистическими тестами для обнаружения специфических отклонений от теоретического поведения. Значительное число таких тестов стали более или менее стандартными и могут применяться для определения приемлемости воспроизводимых величин. Конечно, такие процедуры обходят основной вопрос, на который нет простого ответа.

Простейший способ заключается в том, чтобы использовать случайные числа, полученные из разных источников и опубликованные в таблицах случайных чисел. Некоторые из существующих таблиц были получены перестановками многозначных чисел из математических таблиц, другие же были составлены путем розыгрыша лотереи, перетасовки карт или с помощью специально спроектированного прибора. Все опубликованные таблицы были проверены при помощи стандартных статистических тестов и в результате считаются применимыми для использования при выборках. Существуют даже таблицы таких случайных чисел на перфокартах. Обычно в таблицах

даются случайные десятичные цифры в количестве от нескольких тысяч до миллиона. Для специальных целей было опубликовано несколько таблиц нормально распределенных (гауссовских) величин.

В последние годы в связи с увеличением применения быстроедействующих вычислительных машин были разработаны математические методы получения „псевдослучайных“ чисел на машине, позволяющие обходиться без большого запаса случайных чисел. Обоснованием таких методов служит главным образом то, что результаты проверки при помощи стандартных статистических тестов не противоречат гипотезе случайности. Один такой метод состоит в возведении в квадрат большого числа с последующим извлечением средних цифр квадрата; этот процесс повторяется неограниченно. Известно, что некоторые начальные числа дают быстро вырождающиеся последовательности, в то время как другие приводят к приемлемым результатам. Существуют другие методы, основанные на теоретико-числовых соображениях и поэтому кажущиеся не столь произвольными. В любом случае приемлемость определяется только статистическим испытанием результатов.

§ 11.11. Преобразования случайных чисел. Обычно для построения выборки требуется арифметическое преобразование исходных случайных чисел. В задачах, где требуется выборка из дискретного распределения, расчеты могут быть сравнительно простыми. Например, допустим, что требуются испытания события с вероятностью наступления 0,352. Если рассматривать группы по три случайных цифры, то комбинации 000—351 могут означать наступление, а комбинации 352—999 будут соответствовать ненаступлению события. Если нужны выборочные значения дискретного события с несколькими возможными исходами, то интервал 000—999 может быть разделен на части, содержащие столько чисел, сколько требуется. Если нужно учитывать вероятности с большим числом знаков, чем три, то следует использовать более длинные комбинации случайных цифр.

По аналогии с только что изложенным методом для дискретных распределений строится метод получения значений случайной величины X с плотностью распределения $f(x)$. Предположим, что Y есть случайное число из интервала $(0,1)$, взятое с требуемым числом случайных десятичных цифр после запятой. Решение X уравнения

$$\int_{-\infty}^X f(x) dx = Y$$

будет иметь требуемый закон распределения. Иногда этот общий метод трудно осуществить. Если $f(x)$ ограничена постоянной M и $0 < x < 1$, то можно использовать другой метод: берем два случайных числа X и Y из интервала $(0,1)$ и либо отбрасываем эту пару,

если $f(X) < MY$, либо выбираем X , если $f(X) \geq MY$. В частных случаях могут быть построены разные остроумные методы.

Получение выборки в многомерных случаях может оказаться весьма сложным делом, если законы распределения не просты. Один из общих приемов, упрощающих выборку, заключается в том, чтобы представить заданную плотность $f(x_1, x_2, \dots, x_k)$ в виде произведения

$$g_1(x_1) g_2(x_1, x_2) \dots g_k(x_1, x_2, \dots, x_k),$$

где $g_1(x_1)$ есть частная плотность распределения X_1 , полученная по формуле

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_k) dx_2 \dots dx_k,$$

а $g_2(x_1, x_2)$ есть условная плотность X_2 при условии, что $X_1 = x_1$ и т. д. Тогда случайный вектор $X_1, X_2, \dots, X_k$ может быть получен путем последовательного определения компонент, причем каждая получается в результате одномерной выборки. Различные частные случаи могут навести на мысль о разных специальных приемах. Допустим, что требуется найти случайную точку, равномерно распределенную по поверхности сферы в k -мерном пространстве, или, что то же, требуется найти случайный вектор с равномерно распределенным направлением. Если получить k независимых значений нормально распределенной величины $X_1, X_2, \dots, X_k$, то направление этого вектора будет требуемым. В большинстве случаев, когда требуется многомерная выборка, следует поискать какой-нибудь остроумный метод, при котором затраты вычислительной работы могут оказаться значительно ниже, чем при прямых способах ¹⁾.

ЛИТЕРАТУРА

1. Arley N., On the Theory of Stochastic Processes and Their Application to the Theory of Cosmic Radiation, John Wiley, New York, 1943.
2. Chandrasekhar S., Stochastic Problems in Physics and Astronomy, Rev. Mod. Phys., 15, 1—89 (1943).
3. Курант Р., Фридрихс К., Леви Г., Успехи матем. наук, вып. VIII, 125—160 (1940).
4. Крамер Г., Математические методы статистики, ИЛ, М., 1948.
5. Curtiss J. H., Sampling Methods Applied to Differential and Difference Equations, Proc. Seminar on Scientific Computation, November, 1949, International Business Machines Corporation, New York, 1950.

¹⁾ Более обстоятельное изложение этих вопросов дано, например, в статье Батлера: Butler I. W., Machine sampling from given probability distributions, Symposium on Monte Carlo methods, New York 1956, pp. 249—264. — *Прим. перев.*

6. Donsker M. and Kas M., The Monte Carlo Method and Its Applications, Proc. Computation Seminar, December, 1949, International Business Machines Corporation, New York, 1951.
7. Феллер В., Введение в теорию вероятностей и ее приложения (Дискретные распределения), ИЛ, М., 1952.
8. Forsythe G. E. and Leibler R. A., Matrix Inversion by a Monte Carlo Method, Math. Tables Aids Comput, 4, 127—129 (1950).
9. Fortet R., On the Estimation of an Eigenvalue by an Additive Functional of a Stochastic Process, with Special Reference to the Kac-Donsker Method, J. Res. Nat. Bur. Standards, 48, 68—75, 1952.
10. Hald A., Statistical Tables and Formulas, Table XIX, Random Sampling Numbers, John Wiley, New York, 1952.
11. Хальд А., Математическая статистика с техническими приложениями, ИЛ, М., 1956.
12. Kas M., Random Walk and the Theory of Brownian Motion, Amer. Math. Monthly, 54, 369—391 (1947).
13. Kas M. and Cohen M., A Statistical Method for Determining the Lowest Eigenvalue of Schrödinger's Equation, Nat. Bur. Standards Rept. NBS-1553, 1952.
14. Kahn H., Stochastic (Monte Carlo) Attenuation Analysis, Rand Corporation Report R-163, The Rand Corporation, Santa Monica, Calif., 1949.
15. Kahn H. and Marshall A. W., Methods of Reducing Sample Size in Monte Carlo Computations, J. Operations Res. Soc. Amer., 1, 263—278 (1953).
16. Kempthorne O., The Design and Analysis of Experiments, John Wiley, New York, 1952.
17. Kendall M. G. and Smith B. B., Random Sampling Numbers, 2d series „Tracts for Computers“, No. XXIV, Cambridge University Press, London, 1939.
18. Kendall M. G. and Smith B. B., Randomness and Random Sampling Numbers, J. Roy. Statist. Soc., 5, 147—166 (1938).
19. Kendall M. G. and Smith B. B., Second Paper on Random Sampling Numbers, J. Roy. Statist. Soc. Supp., 6, 51—61 (1939).
20. King G. W., Stochastic Methods in Quantum Mechanics, Proc. Seminar on Scientific Computation, November, 1949, International Business Machines Corporation, New York, 1950.
21. King G. W., The Monte Carlo Method as a Natural Mode of Expression in Operations Research, J. Operations Res. Soc. Amer., 1, 45—51 (1953).
22. King G. W., Monte Carlo Method for Solving Diffusion Problems, Indust. Eng. Chemistry, 43, 2475 (1951).
23. Lehmer D. H., Mathematical Methods in Large-scale Computing Units, Proc. Second Symposium on Large-scale Digital Calculating Machinery, 1949, p. 141—146; Harvard University Press, Cambridge, Mass., 1951.
24. Metropolis N. and Ulam S., The Monte Carlo Method, J. Amer. Statist. Assoc., 44, 335—341 (1949).
25. Monte Carlo Method, Nat. Bur. Standards Appl. Math. Ser., 12 (1951).

26. Mood A. M., Introduction to the Theory of Statistics, McGraw-Hill, New York, 1950.
27. Neyman J., First Course in Probability and Statistics, Henry Holt, New York, 1950.
28. Opler A., Monte Carlo Matrix Calculation with Punched Card Machines, Math. Tables Aids Comput., 5, 115—120 (1951).
29. Thomas L. H., A Comparison of Stochastic and Direct Methods for the Solution of Some Special Problems, J. Operations Res. Soc. Amer., 1, 181—186 (1953).
30. Thomas L. H., Numerical Solution of Partial Differential Equations of Parabolic Type, Proc. Seminar on Scientific Computation, November, 1949, International Business Machines Corporation, New York, 1950.
31. Tippett L. H. C., Random Sampling Numbers, 1st series, „Tracts for Computers“, No. XV, Cambridge University Press, London, 1927.
32. Wasow W., A Note on the Inversion of Matrices by Random Walks, Math. Tables Aids Comput., 6, 78—81 (1952).
33. World H., Random Normal Deviates, Cambridge University Press, London, 1948.
34. Yowell E. C., A Monte Carlo Method of Solving Laplace's Equation, Proc. Seminar on Scientific Computation, December, 1949, International Business Machines Corporation, New York, 1951.
35. Yule G. U., A Test of Tippett's Random Sampling Numbers, J. Roy. Statist. Soc., 5, 167—172 (1938).

Ч А С Т Ь 3

**Вычислительные
методы**

Глава 12

МАТРИЦЫ В ТЕХНИКЕ

ЛУИС А. ПАЙПС

Калифорнийский университет, Лос-Анжелес

§ 12.1. Введение. Матрицы впервые были введены в математику Кэли [6] в 1857 г. Матричные обозначения компактны, удобны и весьма полезны при выполнении линейных преобразований, и они позволяют развить стройный метод решения систем линейных дифференциальных уравнений. Интерес физиков к теории матриц особенно возрос после того, как в 1925 г. Гейзенберг, Борн и Иордан использовали ее в задачах квантовой механики [13]. После мемуара Кэли появилась обширная литература, посвященная приложениям матриц во многих областях математики, физики и техники. Обзор этой литературы (вплоть до 1936 г.), включающий 661 название, можно найти в книге Веддерберна [27]. Развитие современных цифровых вычислительных машин, легко осуществляющих основные матричные операции, сообщило дополнительный толчок широкому использованию матриц в различных областях знаний ¹⁾.

Матрицы находят вполне естественное применение во многих областях техники. Многие факты из теории упругих систем, электрических схем, распространения волн и механических колебаний могут быть сформулированы в компактной форме на языке матриц. Используя теоремы матричной алгебры, удается получить ряд практически важных численных результатов. В этой главе рассматриваются основные положения матричной алгебры, которые наряду с операциями дифференцирования и интегрирования матриц необходимы для понимания того, как используются матрицы в типичных задачах электротехники и механики. Кроме того, здесь будут приведены примеры приложения матричного исчисления к практическим задачам техники.

Основные положения матричной алгебры

§ 12.2. Определение матрицы. Матрицей называется совокупность *m* $\times$ *n* величин, размещенных в виде прямоугольной таблицы

¹⁾ Основные сведения по теории матриц и системам линейных уравнений читатель найдет в книге В. Н. Фаддеевой, *Вычислительные методы линейной алгебры*, Гостехиздат, М.—Л., 1950. — *Прим. ред.*

из m строк и n столбцов. Матрица обычно записывается в следующем виде:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = (a) = A.$$

Символ (a) используется в качестве краткого обозначения всей совокупности этих величин. Каждая величина a_{ij} из этой совокупности называется *элементом* матрицы. Прямоугольная матрица из m строк и n столбцов называется матрицей *порядка* (m, n) ; если $m = n$, то матрица называется *квадратной*.

§ 12.3. Основные типы матриц.

а) *Вектор-строка*. Совокупность n величин, расположенных в одну строку, образует матрицу порядка $(1, n)$. Такая матрица называется *матрицей-строкой*, *линейной матрицей* или *вектор-строкой*.

б) *Вектор-столбец*. Совокупность m величин, расположенных в один столбец, образует матрицу порядка $(m, 1)$ и называется *матрицей-столбцом* или *вектором-столбцом* (*вектором-колонкой*).

в) *Скаляр*. Матрица порядка $(1, 1)$, т. е. матрица, содержащая один элемент, отождествляется в некоторых случаях с соответствующим обычным числом, или *скаляром*.

г) *Диагональная матрица*. Квадратная матрица, элементы которой удовлетворяют условиям $a_{ij} = 0$ при $i \neq j$, так что все элементы ее, не стоящие на *главной диагонали*, равны нулю, называется *диагональной матрицей*.

д) *Единичная матрица*. *Единичной* называется диагональная матрица, все диагональные элементы которой равны единице. Единичная матрица порядка (n, n) будет обозначаться через U_n .

е) *Нулевая матрица*. Матрица порядка (m, n) , все элементы которой равны нулю, называется *нулевой матрицей* порядка (m, n) .

ж) *Транспонирование матрицы*. Транспонирование матрицы (a) состоит в переходе от нее к матрице $(a)_t$, строками которой являются столбцы матрицы (a) , а столбцами — ее строки. Так, например, если

$$(a) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix},$$

то

$$(a)_t = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}.$$

Отсюда следует, что транспонирование матрицы-столбца дает матрицу-строку, и наоборот.

з) *Симметрическая и кососимметрическая матрицы.* Симметрической называется квадратная матрица, симметричная относительно главной диагонали; следовательно, (a) — симметрическая тогда и только тогда, когда $(a)_t = a$. Если же $(a)_t = (-a)$, то (a) называется кососимметрической матрицей.

§ 12.4. Определения и основные алгебраические операции.

а) *Определитель матрицы.* Определитель, состоящий из элементов квадратной матрицы (a) , называется определителем матрицы (a) и обозначается через $|a|$.

б) *Особые и неособые матрицы.* Квадратная матрица (a) с детерминантом $|a|$, равным нулю, называется *особой (вырожденной)*. Если $|a| \neq 0$, то матрица (a) называется *неособой (невырожденной)*.

в) *Алгебраическое дополнение.* Алгебраическим дополнением элемента квадратной матрицы (a) , расположенного в i -й строке и j -м столбце, называется определитель матрицы, получаемой из (a) вычеркиванием i -й строки и j -го столбца, умноженный на $(-1)^{i+j}$.

г) *Присоединенная матрица.* Присоединенная матрица для заданной квадратной матрицы (a) получается заменой каждого ее элемента его алгебраическим дополнением и транспонированием. Присоединенная матрица обозначается через $\text{adj}(a)$.

д) *Равенство двух матриц.* Две матрицы (a) и (b) равны тогда и только тогда, когда они имеют один и тот же порядок и все их соответствующие элементы равны. Таким образом, равенство $(a) = (b)$ предполагает, что (a) и (b) одного порядка и $a_{ij} = b_{ij}$.

е) *Сложение и вычитание матриц.* Эти операции могут быть выполнены лишь при условии, что соответствующие матрицы одного и того же порядка. Если матрицы (a) и (b) одного и того же порядка, то суммой (a) и (b) называется такая матрица (c) , элементы которой $c_{ij} = a_{ij} + b_{ij}$. Другими словами, равенство

$$(a) + (b) = (c)$$

означает, что

$$c_{ij} = a_{ij} + b_{ij}.$$

Подобным же образом определяется *вычитание*: если

$$(a) - (b) = (d),$$

то элемент d_{ij} матрицы (d) равен

$$d_{ij} = a_{ij} - b_{ij}.$$

§ 12.5. Умножение матриц.

а) *Умножение на скаляр.* Умножение матрицы (a) на обычное число, или скаляр k (т. е. на величину, не являющуюся матрицей), определяется соотношением

$$k(a) = (a)k = (b),$$

в котором

$$b_{ij} = ka_{ij}.$$

Следовательно, матрица (a) оказывается умноженной на скаляр k , если каждый элемент ее умножен на этот скаляр. Так как $k(a) = (a)k$, то умножение на скаляр коммутативно.

б) *Умножение матриц*. Умножение матриц определяется таким образом, что оказывается возможным в компактном виде представлять результаты линейных преобразований [7]. Две матрицы могут быть перемножены лишь тогда, когда число столбцов первой из них равняется числу строк второй. Матрицы, удовлетворяющие этому условию, будем называть *соответственными* матрицами. Если две матрицы соответственные, то их произведение определяется соотношением

$$(a)(b) = (c),$$

в котором

$$c_{ij} = \sum_{k=1}^p a_{ik} b_{kj}, \quad (12.1)$$

причем порядки матриц (a) , (b) и (c) равны соответственно (m, p) , (p, n) и (m, n) . Для примера рассмотрим матрицы

$$(a) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad (b) = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

Матрица (a) имеет порядок $(3,3)$, а матрица (b) — порядок $(3,2)$, следовательно, эти матрицы — соответственные. Правило, выраженное равенством (12.1), дает следующее произведение этих матриц:

$$(c) = (a)(b) = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} \end{pmatrix}.$$

Матрица-произведение, как и должно быть, имеет порядок $(3,2)$. И в общем случае, если матрица порядка (m, p) умножается на матрицу порядка (p, n) , то произведение представляет собой матрицу порядка (m, n) . Символически это правило может быть выражено в следующем виде:

$$(m, p) \times (p, n) = (m, n). \quad (12.2)$$

в) *Произведение квадратных матриц*. Если (a) и (b) — две квадратные матрицы одного и того же порядка, то их произведение $(c) = (a)(b)$ является также квадратной матрицей того же порядка, что (a) и (b) . Из определения (12.1) следует, что, вообще говоря, $(a)(b) \neq (b)(a)$. Таким образом, вообще говоря, умножение матриц некоммутативно. Если две матрицы таковы, что $(a)(b) = (b)(a)$, то они называются *коммутирующими* или *перестановочными* матри-

цами. Единичная матрица U коммутирует с каждой квадратной матрицей одинакового с ней порядка

$$U(a) = (a)U = (a),$$

где (a) имеет тот же порядок, что и матрица U . Произведение $(a)(b)$ считают полученным *умножением* матрицы (b) *слева* на матрицу (a) или же *умножением* матрицы (a) *справа* на матрицу (b) .

г) *Умножение слева матрицы-столбца на матрицу-строку.* Рассмотрим следующее произведение:

$$(a_1 a_2 a_3) \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = (a_1 b_1 + a_2 b_2 + a_3 b_3).$$

Это есть произведение матрицы порядка $(1,3)$ на матрицу порядка $(3,1)$. Порядок матрицы-произведения равен $(1,3) \times (3,1) = (1,1)$, и она оказывается, таким образом, скаляром.

д) *Умножение справа матрицы-столбца на матрицу-строку.* Рассмотрим произведение

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} (b_1 b_2 b_3) = \begin{pmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 \end{pmatrix}.$$

Символическое правило (12.2) дает в этом случае, что $(3,1) \times (1,3) = (3,3)$, и потому произведение оказывается квадратной матрицей из трех строк и трех столбцов.

е) *Умножение квадратной матрицы на матрицу-столбец справа.* Рассмотрим следующее произведение матриц:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \quad (12.3)$$

Таким образом, произведение квадратной матрицы и матрицы-столбца оказывается матрицей-столбцом. В этом случае для порядков матриц имеет место соотношение $(3,3) \times (3,1) = (3,1)$

Систему линейных уравнений

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1,$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2,$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3$$

можно в силу (12.3) записать в следующей компактной матричной форме:

$$(a)(x) = (b).$$

Здесь (a) — квадратная матрица, элементами которой являются коэффициенты a_{ij} , а (x) и (b) — матрицы-столбцы, элементами которых являются величины x_i и b_i .

ж) *Сложные произведения матриц.* За исключением коммутативного закона, все обычные законы алгебры приложимы к матрицам. Особая роль принадлежит *ассоциативному* закону в сложных произведениях матриц. Если матрицы (a) и (b) , а также (b) и (c) соответственные, то может быть образовано следующее произведение:

$$[(a)(b)(c)] = [(a)(b)](c) = (a)[(b)(c)].$$

Это говорит о том, что прямые скобки могут быть опущены, и запись произведения $(a)(b)(c)$ может быть упрощена, так как двойное суммирование $\sum_k \sum_r a_{ik} b_{kr} c_{rj}$ может быть выполнено в любом из двух указанных порядков.

Следует, однако, отметить, что произведение ряда матриц будет иметь смысл только в том случае, если каждые две соседние матрицы будут *соответственными*. Если квадратная матрица умножается сама на себя n раз, то получающаяся в результате матрица обозначается через $(a)^n$. Таким образом, $(a)^n = (a)(a) \dots (a)$ (в правой части n сомножителей).

з) *Транспозиция произведения.* Пусть (a) — матрица порядка (p, n) и (b) — матрица порядка (m, p) . Тогда произведение $(c) = (b)(a)$ является матрицей порядка (m, n) . Транспонированные матрицы $(a)_t$ и $(b)_t$ имеют соответственно порядки (n, p) и (p, m) . Следовательно, они оказываются соответственными для составления произведения $(a)_t$ на $(b)_t$. Это произведение является матрицей порядка (n, m) и, как видно, получено транспонированием матрицы (c) . Следовательно, если транспонируется матрица-произведение, то порядок матриц-сомножителей, образующих это произведение, должен быть изменен на обратный, т. е.

$$[(a)(b)]_t = (b)_t(a)_t,$$

аналогично,

$$[(a)(b)(c)]_t = (c)_t(b)_t(a)_t$$

и т. д.

§ 12.6. Обратная матрица. Деление матриц. Если определитель $|a|$ квадратной матрицы (a) отличен от нуля, то матрица (a) называется неособой (невыврожденной) (см. „б“ в § 12.4). В этом случае она будет обладать *обратной* матрицей R , такой, что

$$(a)R = U = R(a),$$

где U есть единичная матрица того же порядка, что и (a) .

Можно доказать [11], что матрица R , обратная матрице (a) , определяется соотношением

$$R = \frac{\text{adj}(a)}{|a|} = (a)^{-1}.$$

Для матрицы R вводится удобное обозначение: $R = (a)^{-1}$. Так как обратная матрица $(a)^{-1}$ удовлетворяет соотношению

$$(a)(a)^{-1} = (a)^{-1}(a) = U,$$

то введенное обозначение оказывается оправданным. Если квадратная матрица невырожденная, то для нее существует обратная матрица, умножение на которую имеет много общего с обычным делением в алгебре. Матрица, обратная для произведения матриц, оказывается равной произведению матриц, обратных для каждого сомножителя, составленному в порядке, противоположном порядку в данном произведении, т. е. если $D = ABC$, то $D^{-1} = C^{-1}B^{-1}A^{-1}$. Если (a) — невырожденная матрица, то отрицательная степень (a) определяется путем возведения обратной матрицы $(a)^{-1}$ в положительную степень, т. е. $(a)^{-n} = [(a)^{-1}]^n$.

§ 12.7. Разбиение матрицы на подматрицы. В некоторых случаях оказывается удобным распространить основные законы матричной алгебры на матрицы, которые рассматриваются как составленные из миноров или алгебраических дополнений ее элементов. Рассмотрим, например, матрицу

$$(a) = \left(\begin{array}{ccc|ccc} a_{11} & a_{12} & a_{13} & & & \\ a_{21} & a_{22} & a_{23} & & & \\ \hline & & & & & \\ a_{31} & a_{32} & a_{33} & & & \end{array} \right) = \begin{pmatrix} P & Q \\ R & S \end{pmatrix},$$

где

$$P = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad Q = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}, \quad R = (a_{31} \ a_{32}), \quad S = (a_{33}).$$

В этом случае разбиение матрицы (a) произведено таким образом, что диагональные матрицы P и S оказываются квадратными и разбиение оказывается симметричным относительно диагонали. Пусть теперь в квадратной матрице (b) произведено аналогичное разбиение:

$$(b) = \left(\begin{array}{ccc|ccc} b_{11} & b_{12} & b_{13} & & & \\ b_{21} & b_{22} & b_{23} & & & \\ \hline & & & & & \\ & & & & & \\ b_{31} & b_{32} & b_{33} & & & \end{array} \right) = \begin{pmatrix} P_1 & Q_1 \\ R_1 & S_1 \end{pmatrix}.$$

Тогда, как легко проверить, при сложении и умножении этих матриц получим

$$(a) + (b) = \begin{pmatrix} P + P_1 & Q + Q_1 \\ R + R_1 & S + S_1 \end{pmatrix},$$

$$(a)(b) = \begin{pmatrix} PP_1 + QR_1 & PQ_1 + QS_1 \\ RP_1 + SR_1 & RQ_1 + SS_1 \end{pmatrix}.$$

Возможность разбиения громоздкой матрицы на миноры очень полезна при отыскании обратной матрицы. Например, квадратная

матрица M из $2p$ строк может быть разбита на четыре квадратных матрицы (a) , (b) , (c) и (d) , из p строк каждая, следующим образом:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Пусть R есть матрица, обратная для M , так что $MR = U$, где U — единичная матрица порядка $2p$. Пусть матрица R подразделена следующим образом:

$$R = \begin{pmatrix} A & C \\ B & D \end{pmatrix},$$

где A , B , C и D — квадратные матрицы порядка p . Так как $R = M^{-1}$, то

$$MR = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} U_p & 0 \\ 0 & U_p \end{pmatrix}, \quad (12.4)$$

где U_p — единичная матрица порядка p . Если выполненное в (12.4) умножение записать в развернутом виде, то получим четыре матричных уравнения, определяющих A , B , C и D :

$$\begin{aligned} (a)A + (b)B &= U_p, & (a)C + (b)D &= 0 \\ (c)A + (d)B &= 0, & (c)C + (d)D &= U_p. \end{aligned} \quad (12.5)$$

В них 0 поставлен вместо нулевой матрицы p -го порядка. Четыре уравнения (12.5) могут быть решены средствами элементарной алгебры, если только следить за тем, чтобы умножение соответствующих матриц осуществлялось в нужном порядке:

$$\begin{aligned} A &= [(a) - (b)(d)^{-1}(c)]^{-1}, & B &= -(d)^{-1}(c)A, \\ D &= [(d) - (c)(a)^{-1}(b)]^{-1}, & C &= -(a)^{-1}(b)D. \end{aligned}$$

Эти формулы требуют обращения четырех квадратных матриц порядка p . Возможны другие способы разбиения исходной матрицы, причем сокращение объема работ при отыскании обратной матрицы будет зависеть от формы этой матрицы. В книге Фрезера, Дункана и Коллара „Теория матриц“ [11] приведен еще один метод получения матрицы, обратной данной, с помощью предварительного разбиения исходной матрицы.

§ 12.8. Характеристическая матрица и характеристическое уравнение матрицы. Если μ — скалярный параметр и M — квадратная матрица порядка n , то матрица

$$K = \mu U - M,$$

где U — единичная матрица n -го порядка, называется характеристической матрицей для матрицы M . Определитель характеристической

матрицы, $\det K$, называется *характеристической функцией* матрицы M . Таким образом,

$$\Delta(\mu) = \det K \quad (12.6)$$

является характеристической функцией матрицы M .

Уравнение

$$\Delta(\mu) = \det K = 0 \quad (12.7)$$

называется *характеристическим уравнением* матрицы M . Таким образом, если M — квадратная матрица n -го порядка, то ее характеристическое уравнение является уравнением n -й степени относительно μ . n корней этого уравнения называются *характеристическими корнями*, или *собственными значениями*, матрицы M .

§ 12.9. Приведение матрицы к диагональной форме. Если матрица состоит только из одного столбца, то она называется *вектором-столбцом*. Если (x) есть вектор-столбец n -го порядка и M — квадратная матрица n -го порядка, то произведение

$$(y) = M(x)$$

выражает тот факт, что вектор (y) получается из вектора (x) путем *линейного преобразования*. Квадратная матрица M называется *матрицей преобразования*. Если потребовать, чтобы

$$(y) = \mu(x),$$

то последнее произведение можно записать в виде

$$\mu(x) = M(x), \quad (12.8)$$

и в этом случае действие матрицы M на вектор (x) сводится к умножению его на скалярный множитель μ . Уравнение (12.8) может быть записано в виде

$$(\mu U - M)(x) = 0 = K(x). \quad (12.9)$$

Это соотношение представляет собой систему линейных однородных уравнений с неизвестными $x_1, x_2, \dots, x_n$. Решение этой системы, кроме тривиального случая, когда все элементы (x) равны нулю, существует в том случае, когда определитель K обращается в нуль. Поэтому скаляр μ может принимать только значения $\mu_1, \mu_2, \dots, \mu_n$, которые являются корнями характеристического уравнения

$$\Delta(\mu) = \det(\mu U - M) = 0. \quad (12.10)$$

Таким образом, всевозможные значения μ , удовлетворяющие уравнению (12.10), оказываются собственными значениями матрицы M .

Для каждого собственного значения μ_i имеет место соответствующее уравнение (12.9), которое может быть записано в следующем виде:

$$\mu_i(x)_i = M(x)_i, \quad i = 1, 2, \dots, n, \quad (12.11)$$

где $(x)_i$ — ненулевой вектор-столбец. Эти n векторов называются *собственными векторами* матрицы M . Если уравнение (12.11) умножить слева на M , то получим

$$\mu_i M(x)_i = \mu_i^2 (x)_i = M^2(x)_i.$$

Повторными умножениями на M можем получить, что

$$M^r(x)_i = \mu_i^r (x)_i, \quad r = 1, 2, \dots \quad (12.12)$$

Обозначим i -й собственный вектор через

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{pmatrix}$$

и примем его за i -й столбец квадратной матрицы (x)

$$(x) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix}. \quad (12.13)$$

Если ввести в рассмотрение квадратную диагональную матрицу (d)

$$(d) = \begin{pmatrix} \mu_1 & 0 & \dots & 0 \\ 0 & \mu_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mu_n \end{pmatrix},$$

то уравнения (12.11), как легко видеть, могут быть записаны следующим образом:

$$(x)(d) = M(x). \quad (12.14)$$

Если n собственных значений матрицы M различны, то можно показать, что матрица (x) невырожденная и обладает обратной матрицей $(x)^{-1}$. Следовательно, уравнение (12.14) можно переписать и в таком виде:

$$M = (x)(d)(x)^{-1}.$$

Это говорит о том, что матрица (x) , приводящая матрицу M к диагональному виду, может быть составлена из собственных векторов матрицы M так, как это показано выше. Это оказывается возможным в том случае, когда все собственные значения M различны. Если два или более собственных значения матрицы M оказываются совпадающими, то приведение матрицы к диагональному виду не всегда оказывается возможным.

§ 12.10. След матрицы. Сумма всех диагональных элементов квадратной матрицы называется следом матрицы и обозначается следующим образом:

$$\text{Tr } M = \sum_{i=1}^n M_{ii}.$$

Можно показать, что

$$\text{Tr } AB = \text{Tr } BA$$

и что если $AB = C$, то $\text{Tr } A \text{Tr } B = \text{Tr } C$.

Можно показать также [3], что существует тесная связь между следами степеней матрицы и коэффициентами ее характеристического уравнения. Этот результат важен для получения характеристических уравнений матрицы и вычисления обратных матриц. Мы приводим его здесь без доказательства.

Характеристическое уравнение (12.7) квадратной матрицы M порядка n является, вообще говоря, уравнением n -й степени и может быть записано так:

$$\Delta(\mu) = \mu^n + c_1\mu^{n-1} + \dots + c_{n-1}\mu + c_n = 0. \quad (12.15)$$

Пусть

$$S_1 = \text{Tr } M, \quad S_2 = \text{Tr } M^2, \quad \dots, \quad S_n = \text{Tr } M^n.$$

Можно показать [3], что коэффициенты c_i характеристического уравнения (12.15) находятся из соотношений

$$\begin{aligned} c_1 &= -S_1, \\ c_2 &= -\frac{1}{2}(c_1 S_1 + S_2), \\ &\dots \dots \dots \\ c_n &= -\frac{1}{n}(c_{n-1} S_1 + c_{n-2} S_2 + \dots + c_1 S_{n-1} + S_n). \end{aligned} \quad (12.16)$$

Эти соотношения позволяют сократить объем работ при составлении характеристического уравнения матрицы.

§ 12.11. Теорема Кэли — Гамильтона. Одной из наиболее важных теорем теории матриц является *теорема Кэли — Гамильтона*. С помощью этой теоремы могут быть изучены многие свойства матриц, а также значительно упрощено рассмотрение функций от матриц. Теорема Кэли — Гамильтона часто формулируется в следующей форме: „Каждая квадратная матрица удовлетворяет своему собственному характеристическому уравнению, понимаемому в матричном смысле“. Это означает, что если задано характеристическое уравнение

$$\Delta(\mu) = \mu^n + c_1\mu^{n-1} + \dots + c_{n-1}\mu + c_n = 0$$

квадратной матрицы M порядка n , то имеет место матричное равенство

$$\Delta M = M^n + c_1 M^{n-1} + \dots + c_{n-1} M + c_n U = 0, \quad (12.17)$$

в котором U означает единичную матрицу n -го порядка, а 0 — нулевую матрицу того же порядка.

Доказательство этой теоремы можно получить, применяя результат (12.12). Если в выражении (12.15) характеристической функции $\Delta(\mu)$ матрицы M скаляр μ заменить квадратной матрицей M и коэффициент c_n умножить на единичную матрицу n -го порядка, то результат может быть записан в следующем виде:

$$\Delta M = M^n + c_1 M^{n-1} + \dots + c_n U. \quad (12.18)$$

Умножив обе части равенства (12.18) на собственный вектор $(x)_i$ справа, получим

$$\Delta M(x)_i = (M^n + c_1 M^{n-1} + \dots + c_n U)(x)_i.$$

На основании результата (12.12) это можно переписать в виде

$$\Delta M(x)_i = (\mu_i^n + c_1 \mu_i^{n-1} + \dots + c_n)(x)_i. \quad (12.19)$$

Однако, так как μ_i является одним из собственных значений M , первый множитель в правой части соотношения (12.19) обращается в нуль и оказывается, что

$$\Delta M(x)_i = 0, \quad i = 1, 2, \dots, n.$$

Эта система уравнений может быть записана короче в виде одного уравнения

$$\Delta M(x) = 0, \quad (12.20)$$

в котором (x) — квадратная матрица n -го порядка (12.13) и 0 — квадратная нулевая матрица того же порядка. Если M обладает n различными собственными значениями, то матрица (x) имеет обратную. Поэтому если (12.20) умножить справа на $(x)^{-1}$, то получится результат

$$\Delta M = 0,$$

который и доказывает теорему Кэли — Гамильтона. Можно показать [3], что теорема Кэли — Гамильтона справедлива также для вырожденных матриц и для матриц с совпадающими собственными значениями.

Если матрица M невырожденная, то матрица, ей обратная, может быть найдена с помощью теоремы Кэли — Гамильтона следующим способом: умножив уравнение (12.17) слева на M^{-1} , мы получим соотношение

$$M^{n-1} + c_1 M^{n-2} + \dots + c_n M^{-1} = 0,$$

которое может быть решено относительно M^{-1} :

$$M^{-1} = -\frac{1}{c_n}(M^{n-1} + c_1 M^{n-2} + \dots + c_{n-1} U). \quad (12.21)$$

Это соотношение выражает матрицу, обратную для M , через $n-1$ первых степеней матрицы M и через коэффициенты характеристического уравнения этой матрицы. Так как коэффициенты c_i даются равенствами (12.16) и могут быть, таким образом, получены путем умножения, то формула (12.21) дает практический метод вычисления матрицы, обратной данной [12]. Если требуется вычислить определитель M , то получить его можно из равенства

$$\det M = (-1)^n c_n.$$

Метод вычисления обратной матрицы, основанный на соотношении (12.21), очень прост, легко поддается проверке и хорошо приспособлен к умножению матриц с помощью перфорированных карт. Кроме того, при таком вычислении в качестве побочного результата оказываются найденными характеристическое уравнение и определитель матрицы.

§ 12.12. Решение линейных уравнений и обращение больших матриц. С развитием ряда новых мощных вычислительных средств, таких, как машины с перфорированными картами, электронные цифровые машины, машины непрерывного действия и т. д., значительно расширилась область применения математического анализа. Многочисленные проблемы в различных областях знания могут быть сведены к одной и той же математической задаче — к решению системы линейных уравнений типа

$$M(x) = (a), \quad (12.22)$$

где M — невырожденная квадратная матрица n -го порядка с известными элементами, (x) — вектор-столбец с n неизвестными элементами x_i и (a) — вектор-столбец из n заданных элементов a_i . Решение уравнения (12.22) может быть получено путем умножения обеих его частей слева на обратную матрицу M^{-1} . Это даст

$$(x) = M^{-1}(a).$$

Отсюда следует, что решение системы линейных уравнений (12.22) эквивалентно с математической точки зрения вычислению матрицы, обратной матрице M , составленной из коэффициентов системы.

Приводим несколько задач, сводящихся к этой общей математической задаче, т. е. к определению матрицы, обратной заданной (обращение матрицы):

а) распределение электрических токов в схемах с многими соединениями;

- б) распределение скоростей потоков воды в сложных гидравлических системах;
- в) приближенное решение задач теории потенциала;
- г) решение нормальных систем уравнений, получающихся при применении метода наименьших квадратов к решению различных физических задач;
- д) определение амплитуд при вынужденных колебаниях механических систем;
- е) приближенное решение некоторых проблем в вопросах электромагнитной радиации;
- ж) приложение статистического анализа к психологии, социологии и экономике.

Приведенный перечень проблем далеко не полон, но и он дает представление о разнообразии классов проблем, которые могут быть сведены к решению системы линейных уравнений.

На первый взгляд может показаться, что задача решения систем линейных уравнений стара и решение ее может быть получено с помощью хорошо известных средств, таких, например, как правило Крамера, изложенное в книгах по элементарной алгебре. Однако для решения задач перечисленных выше типов часто бывает необходимо рассмотреть от 150 до 200 совместных уравнений. Это обстоятельство создает на пути практического решения задачи столь значительные трудности, что они оказываются непреодолимыми, если пользоваться известными элементарными методами. В последние годы были разработаны мощные и не связанные явно друг с другом методы решения систем линейных уравнений. Весьма полный перечень подобных методов вместе с библиографией, включающей 131 наименование, можно найти в статье Форсайта [10]. В этой статье убедительно показано, насколько значительно уменьшается объем работ при решении большой системы линейных уравнений, если вместо элементарных методов пользоваться современными приемами вычислений. Решение системы n совместных уравнений с n неизвестными по правилу Крамера требует вычисления $n+1$ определителей. Если каждый определитель вычислять с помощью теоремы Лапласа, то нужно будет выполнить $n!$ умножений. Из этого следует, что отыскание решения методом определителей требует $(n+1)!$ умножений. При более эффективных методах для решения системы из n линейных уравнений потребуется осуществить только около $n^3/3$ умножений.

Подробная теория различных методов, созданных в последние годы, изложена в упомянутой статье Форсайта [10], к которой следует обращаться всем, заинтересованным в решении большого числа уравнений. Ниже приводится простой и эффективный метод, основанный на разбиении матрицы системы уравнений.

Пусть требуется определить матрицу, обратную для заданной квадратной матрицы n -го порядка. Тогда с помощью метода, изложенного в § 12.7, матрицу M можно разбить на четыре минора сле-

дующим образом:

$$M = \begin{pmatrix} M_{11}(r, r) & M_{12}(r, s) \\ M_{21}(s, r) & M_{22}(s, s) \end{pmatrix}.$$

Порядки различных матриц определяются буквами в скобках. Первая буква обозначает число строк, вторая — число столбцов. Так как исходная матрица имеет порядок n , то $r + s = n$. Допустим, что B — матрица, обратная матрице M . Подвергнем ее разбиению таким образом, чтобы

$$B = M^{-1} = \begin{pmatrix} B_{11}(r, r) & B_{12}(r, s) \\ B_{21}(s, r) & B_{22}(s, s) \end{pmatrix}.$$

В силу определения обратной матрицы

$$BM = U_n, \quad (12.23)$$

где U_n — единичная матрица n -го порядка. Если произвести указанное в (12.23) умножение, то получается следующий результат:

$$B_{11}M_{11} + B_{12}M_{21} = U_r,$$

$$B_{11}M_{12} + B_{12}M_{22} = 0,$$

$$B_{21}M_{11} + B_{22}M_{21} = 0,$$

$$B_{21}M_{12} + B_{22}M_{22} = U_s.$$

Эти уравнения можно, вообще говоря, полностью решить относительно матриц, составляющих матрицу B . Результат можно выразить в следующем виде. Пусть

$$X = M_{11}^{-1}M_{12}, \quad Y = M_{21}M_{11}^{-1},$$

$$\theta = M_{22} - YM_{12} = M_{22} - M_{21}X.$$

Тогда

$$\begin{aligned} B_{11} &= M_{11}^{-1} + X\theta^{-1}Y, & B_{21} &= -\theta^{-1}Y, \\ B_{12} &= -X\theta^{-1}, & B_{22} &= \theta^{-1}. \end{aligned} \quad (12.24)$$

Эти уравнения определяют $B = M^{-1}$ при условии существования обратных матриц M_{11}^{-1} и θ^{-1} . Если известна матрица M_{11}^{-1} , то матрицы X , Y и θ можно получить перемножением соответствующих матриц, а матрицу B можно определить с помощью уравнения (12.24). Выбор пути для разбиения исходной матрицы зависит от вида тех матриц, для которых нужно находить обратные матрицы. Из предыдущего видно, что приведенный метод допускает различные видоизменения. Уравнение (12.21) можно с успехом применить для определения M_{11}^{-1} и θ^{-1} . Благодаря этому задача отыскания матрицы, обратной для заданной большой матрицы, сводится к аналогичной задаче для нескольких матриц низшего порядка. Подробные численные примеры обращения больших матриц методом разбиения можно найти в „Теории матриц“ Фрезера, Дункана и Коллара [11].

Операции над матрицами

§ 12.13. Дифференцирование матриц. Если элементами матрицы (a) являются функции скалярного переменного t , то матрица $(a(t))$ называется *матричной функцией* t . Производная этой матрицы, если она существует, определяется соотношением

$$\frac{d}{dt}(a) = \left(\frac{d}{dt} a_{rs}(t) \right),$$

в котором $a_{rs}(t)$ есть элемент матрицы $(a(t))$, стоящей в r -й строке и s -м столбце. Следовательно, элементы $d(a)/dt$ являются производными соответствующих элементов матрицы (a) .

§ 12.14. Интегрирование матриц. Интеграл от матрицы $(a(t))$, элементы которой являются функциями скалярного переменного t , определяется следующим соотношением:

$$\int (\dot{a}(t)) dt = \left(\int a_{rs}(t) dt \right).$$

Интеграл от матрицы существует лишь тогда, когда существует интеграл от каждого ее элемента. Легко видеть, что интеграл от матрицы, элементы которой являются функциями скалярного переменного t , получается путем интегрирования каждого из этих элементов.

§ 12.15. Многочлены от матриц. Многочлен относительно квадратной матрицы (a) может быть записан в виде

$$P(a) = c_0(a)^k + c_1(a)^{k-1} + \dots + c_{k-1}(a) + c_n U,$$

где k есть целое положительное число (степень многочлена), коэффициенты $c_0, c_1, \dots, c_k$ — постоянные скаляры, а U — единичная матрица, порядок которой равен порядку матриц (a) .

§ 12.16. Функции матриц. Рассмотрим функцию $f(z)$ скалярного переменного z , например $e^z, \operatorname{sh} z, \sin z$ и т. д., которая определяется степенным рядом относительно z , так что

$$f(z) = \sum_{n=0}^{\infty} c_n z^n.$$

Определение такой функции можно распространить на случай матричного аргумента (a) с помощью ряда

$$f(a) = \sum_{n=0}^{\infty} c_n (a)^n \quad (12.25)$$

(в котором $(a)^0 = U$ означает единичную матрицу) при условии сходимости этого ряда. Так, например, экспоненциальная функция квадратной матрицы (a) определяется рядом

$$\exp(a) = e^{(a)} = U + (a) + \frac{(a)^2}{2!} + \frac{(a)^3}{3!} + \dots$$

Можно показать, что этот ряд сходится для всех квадратных матриц (a) (см. стр. 51 книги „Теория матриц“ [11]). Другим примером может служить функция $\sin(a)$, определяемая рядом

$$\sin(a) = U - \frac{(a)^3}{3!} + \frac{(a)^5}{5!} - \frac{(a)^7}{7!} + \dots$$

Этот ряд, как можно показать, сходится для всех квадратных матриц (a) . В результате изучения этого ряда можно выявить свойства матричной функции $\sin(a)$.

§ 12.17. Представление матричной функции с помощью многочлена. Если (a) есть квадратная матрица порядка n , то по теореме (12.17) Кэли—Гамильтона она удовлетворяет своему характеристическому уравнению. Запишем это уравнение в виде

$$(a)^n = k_1(a)^{n-1} + k_2(a)^{n-2} + \dots + k_{n-1}(a) + k_n U,$$

где величины k_i — скалярные коэффициенты. Это уравнение может быть использовано для сведения ряда (12.25) к многочлену вида

$$f(a) = b_{n-1}(a)^{n-1} + b_{n-2}(a)^{n-2} + \dots + b_1(a) + b_0 U \quad (12.26)$$

со скалярными коэффициентами b_i . Можно показать [1], что уравнение (12.26) может быть записано в следующем виде:

$$f(a) = \frac{D_{n-1}}{D}(a)^{n-1} + \frac{D_{n-2}}{D}(a)^{n-2} + \dots + \frac{D_1}{D}(a) + \frac{D_0}{D} U, \quad (12.27)$$

где D — определитель

$$D = \begin{vmatrix} 1 & 1 & \dots & 1 \\ \mu_1 & \mu_2 & \dots & \mu_n \\ \dots & \dots & \dots & \dots \\ \mu_1^{n-1} & \mu_2^{n-1} & \dots & \mu_n^{n-1} \end{vmatrix} \quad (12.28)$$

и $\mu_1, \mu_2, \dots, \mu_n$ — собственные значения матрицы (a) ; определитель D_{r-1} получается из определителя (12.28) путем замены элементов r -й строки на $f(\mu_1), f(\mu_2)$ и т. д. Можно заметить, что D есть произведение разностей собственных значений μ_i . В качестве примера уравнения (12.27) рассмотрим функцию $e^{(a)}$, где

$$(a) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Собственными значениями матрицы (a) оказываются $\mu_1 = 1$ и $\mu_2 = 2$. В этом случае имеем

$$D = \begin{vmatrix} 1 & 1 \\ \mu_1 & \mu_2 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix} = 1,$$

$$D_1 = \begin{vmatrix} 1 & 1 \\ e & e^2 \end{vmatrix} = e^2 - e; \quad D_0 = \begin{vmatrix} e & e^2 \\ 1 & 2 \end{vmatrix} = 2e - e^2.$$

Следовательно,

$$\exp(a) = e^{(a)} = \frac{D_1}{D}(a) + \frac{D_0}{D}U = (e^2 - e) \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + (2e - e^2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} e & 0 \\ 0 & e^2 \end{pmatrix}.$$

§ 12.18. Теорема Сильвестра. Можно доказать (см. [11], стр. 89), что если квадратная матрица (a) порядка n имеет различные собственные значения μ_i ($i = 1, 2, \dots, n$), то выражение (12.27) приводится к следующей двойственной форме:

$$f(a) = \sum_{r=1}^n f(\mu_r) Z_r, \quad (12.29)$$

где матрица Z_r определяется равенством

$$Z_r = \frac{\prod_s [\mu_s U - (a)]}{\prod_s (\mu_s - \mu_r)}, \quad r \neq s.$$

Эта форма записи функции матрицы (a) известна в математической литературе как *теорема Сильвестра*. Теорема Сильвестра оказывается ценной для представления матричных функций и весьма полезна для получения приближенных значений высоких степеней матриц. Так, если μ_m есть наибольшее по модулю собственное значение квадратной матрицы (a) и степень k достаточно велика, то можно заметить, что в выражении (12.29) член с μ_m^k будет доминировать над остальными членами. Это позволяет получить следующее приближенное значение матрицы $(a)^k$:

$$(a)^k \approx \mu_m^k \frac{\prod_s [\mu_s U - (a)]}{\prod_s (\mu_s - \mu_m)} \quad (s \neq m). \quad (12.30)$$

В качестве примера использования выражения (12.30) найдем приближенное значение матрицы

$$\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}^{256} = (a)^{256}.$$

В этом случае собственными значениями матрицы (a) оказываются $\mu_1 = 1$, $\mu_2 = 3 = \mu_m$, и, следовательно, из уравнения (12.30) может быть получен следующий результат:

$$(a)^{256} = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}^{256} \approx \frac{3^{256}}{2} [(a) - U] = 3^{256} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

§ 12.19. Решение алгебраических уравнений с помощью матриц. Теорема Сильвестра может служить основанием для эффективного метода отыскания корней алгебраических уравнений вида

$$F(z) = z^n + c_1 z^{n-1} + \dots + c_{n-1} z + c_n = 0. \quad (12.31)$$

Легко проверить, что это уравнение является характеристическим уравнением следующей квадратной матрицы:

$$(a) = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -c_n & -c_{n-1} & -c_{n-2} & \dots & -c_2 & -c_1 \end{pmatrix}.$$

Квадратная матрица (a) называется *присоединенной матрицей* уравнения (12.31). Если уравнение (12.31) имеет различные корни $r_1, r_2, \dots, r_n$, то, предположив, что r_m является наибольшим (по модулю) корнем, и применив теорему Сильвестра, можно получить из уравнения (12.30), что для достаточно больших k

$$(a)^{k+1} \approx r_m (a)^k. \quad (12.32)$$

Это означает, что для достаточно высоких степеней k матрицы (a) последнее умножение $(a)^k$ на (a) приводит к матрице $(a)^{k+1}$, элементы которой получаются из элементов матрицы $(a)^k$ умножением на один и тот же общий множитель r_m — доминирующий корень уравнения (12.31). Если наибольший из корней найден, то для отыскания наибольшего из оставшихся корней можно применить один из существующих методов. Практически это осуществляется следующим путем. Многочлен $F(z)$ делят на разность $z - r_m$. В частном получают многочлен $F_1(z)$, для которого записывают соответствующую ему присоединенную матрицу. Возведением в достаточно высокую степень новой присоединенной матрицы получают следующий по абсолютной величине корень уравнения (12.31). Повторяя этот прием, можно найти все корни уравнения. Этот метод может быть также распространен [18] на случай вычисления комплексных корней. Если достаточно высокие степени присоединенной матрицы (a) не подчиняются закону, выраженному уравнением (12.32), то это свидетельствует о наличии кратного или доминирующего комплексного

корня. Используя этот метод, работу можно организовать таким образом, что объем ее окажется несколько меньшим по сравнению с объемом работы при определении корней по известному методу Греффе и Данделена.

Интегрирование линейных дифференциальных уравнений

§ 12.20. Матрициант. Любую систему дифференциальных уравнений можно преобразовать в другую систему, которая будет содержать лишь уравнения первого порядка. Для этого достаточно вместо производных порядка выше первого от неизвестных функций ввести в рассмотрение новые неизвестные функции. С каждой новой вводимой неизвестной функцией к системе уравнений добавляется одно новое уравнение. Например, уравнение

$$\frac{d^2x}{dt^2} + x = 0 \quad (12.33)$$

может быть преобразовано в систему уравнений первого порядка введением новой зависимой переменной y с помощью уравнения (см. гл. 1)

$$\frac{dx}{dt} = y.$$

В результате уравнение (12.33) заменяется системой

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = -x. \quad (12.34)$$

Система (12.34) может быть записана в матричном виде

$$D \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}; \quad D = \frac{d}{dt}. \quad (12.35)$$

Это говорит о возможности записи в наиболее общем случае системы однородных линейных дифференциальных уравнений в виде

$$D(x) = (u(t))(x); \quad D = \frac{d}{dt}. \quad (12.36)$$

Здесь (x) — вектор-столбец из n неизвестных, а $(u(t))$ — квадратная матрица, элементы которой являются, вообще говоря, функциями переменной t . Если уравнения имеют постоянные коэффициенты, то и матрица (u) своими элементами имеет постоянные числа. Метод интегрирования системы (12.35) дифференциальных уравнений был впервые предложен в 1888 г. Пеано; Бейкеру [2] принадлежит дальнейшее усовершенствование этого метода. Пусть заданы начальные значения неизвестных при $t=0$. Будем их считать компонентами вектора $(x)_0$. Тогда система (12.36) может быть записана в виде интегрального уравнения

$$(x) = (x)_0 + Q(u(t))(x), \quad (12.37)$$

где

$$Q = \int_0^t () dt$$

есть интегральный оператор.

Метод Пеано — Бейкера состоит в решении интегрального уравнения (12.37) итерационным способом

$$(x) = (U + Q(u) + Q(u)Q(u) + Q(u)Q(u)Q(u) + \dots)(x)_0, \quad (12.38)$$

где U есть единичная матрица n -го порядка. Если элементы матрицы $(u(t))$ являются аналитическими во всей области изменения t , то можно доказать, что ряды в (12.38) сходятся абсолютно и равномерно. Матрица

$$H(u) = (U + Q(u) + Q(u)Q(u) + Q(u)Q(u)Q(u) + \dots) \quad (12.39)$$

называется *матрициантом* Бейкера матрицы (u) [2]. Решение системы (12.36) может быть, таким образом, записано в виде

$$(x) = H(u)(x)_0.$$

§ 12.21. Линейные уравнения с постоянными коэффициентами. Если уравнения системы (12.36) имеют постоянные коэффициенты, то и элементы матрицы $(u(t))$ оказываются *постоянными*. Можно показать, что в этом случае ряд (12.39) сводится к следующему:

$$H(u) = U + (u) \frac{t}{1!} + (u)^2 \frac{t^2}{2!} + (u)^3 \frac{t^3}{3!} + \dots = e^{(u)t},$$

а решение уравнения (12.36) оказывается в этом случае равным

$$(x) = e^{(u)t}(x)_0.$$

Функция $e^{(u)t}$ может быть определена на основании теоремы Сильвестра (12.29).

В специальном случае, когда система уравнений (12.36) с постоянными коэффициентами является системой двух совместных уравнений, имеем

$$(u) = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}.$$

Характеристическим уравнением для (u) будет уравнение

$$\mu^2 - (u_{11} + u_{22})\mu + (u_{11}u_{22} - u_{12}u_{21}) = 0.$$

Если предположить, что корни этого уравнения [собственные значения матрицы (u)] записаны в виде $\mu_1 = a + b$, $\mu_2 = a - b$,

то матрициант (u) может быть легко вычислен с помощью теоремы Сильвестра и представлен в виде

$$H(u) = e^{(u)t} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \quad (12.40)$$

где

$$H_{11} = \left[(u_{11} - a) \frac{\operatorname{sh} bt}{b} + \operatorname{ch} bt \right] e^{at},$$

$$H_{12} = u_{12} e^{at} \frac{\operatorname{sh} bt}{b},$$

$$H_{21} = u_{21} e^{at} \frac{\operatorname{sh} bt}{b},$$

$$H_{22} = \left[(u_{22} - a) \frac{\operatorname{sh} bt}{b} + \operatorname{ch} bt \right] e^{at}.$$

Если b — мнимое число, то гиперболические функции должны быть заменены соответствующими тригонометрическими функциями.

§ 12.22. Линейные уравнения с переменными коэффициентами; метод осредненных коэффициентов. Если коэффициенты уравнений системы (12.36) переменные, то определяемый уравнением (12.39) матрициант даст, вообще говоря, решение уравнений в виде степенного ряда. Если, однако, эти коэффициенты изменяются столь незначительно при изменении t , что их можно приближенно рассматривать как постоянные величины на конечном числе небольших интервалов изменения t , то может быть использован приближенный метод решения системы. Этот метод известен как *метод осредненных коэффициентов* (см. стр. 251—267 книги [11]). Этот метод состоит в том, что на каждом из указанных малых интервалов изменения t коэффициенты этих уравнений заменяются их осредненными значениями. Этим путем матрица дифференциального уравнения разбивается на несколько матриц, каждая из которых соответствует дифференциальным уравнениям с постоянными коэффициентами. Таким образом, уравнения при переходе от одного малого интервала к другому меняются, однако решения этих уравнений остаются связанными начальными условиями. Значения зависимых переменных на конце каждого из малых интервалов принимаются в качестве начальных значений для следующего интервала. Если $t_1, t_2, \dots, t_{n-1}$ означают любую последовательность из $n-1$ точек в интервале $(0, t)$, то искомое приближенное решение можно записать в следующем виде. Пусть $(u(p))$ означает матрицу, которой заменяется матрица $(u(t))$ [матрица $(u(p))$ составлена из осредненных значений коэффициентов $(u(t))$ на p -м интервале] и пусть

$$T_p = t_p - t_{p-1}.$$

Тогда из уравнения (12.40) получаем

$$H(u(p)) = \exp(u_p T_p); \quad (u_p) = (u(p)).$$

Полным приближенным решением уравнения (12.36) будет

$$(x) = [\exp(u_n T_n)] [\exp(u_{n-1} T_{n-1})] \dots [\exp(u_1 T_1)] (x)_0,$$

где $T_n = t - t_{n-1}$ и $(x)_0$ — вектор-столбец из начальных значений при $t = 0$. Такой метод оказывается наиболее подходящим при использовании современных вычислительных средств. Много примеров применения этого метода можно найти в „Теории матриц“ [11], где решения, получаемые методом аппроксимации, сравниваются с точными решениями. Установлено, что при надлежащем выборе интервалов T_p этот метод дает высокую точность. Метод осредненных коэффициентов дает возможность получить *точные* решения [22] некоторых уравнений, коэффициенты которых являются *периодическими* функциями t .

Приложение матриц к задачам об упругих системах

§ 12.23. Главные направления нагрузок в точке упругого тела. Рассмотрим точку P упругого тела (см. гл. 7) и систему декартовых координат с началом в точке P . Пусть в точке P приложена единичная нагрузка, направленная параллельно оси x . Приложенная в точке P единичная нагрузка связана законом Гука с вызываемыми ею статическими смещениями. Обозначим эти смещения в направлении осей x , y и z соответственно через f_{11} , f_{21} и f_{31} . Аналогично, обозначим смещения, вызываемые единичными нагрузками, приложенными в точке P и направленными параллельно осям y и z , соответственно через f_{12} , f_{22} , f_{32} и f_{13} , f_{23} , f_{33} . Тогда смещения x , y , z точки P в направлении координатных осей, вызванные приложением в этой точке силы, имеющей проекциями на оси координат X , Y и Z , выразятся следующим образом:

$$\begin{aligned} x &= f_{11}X + f_{12}Y + f_{13}Z, \\ y &= f_{21}X + f_{22}Y + f_{23}Z, \\ z &= f_{31}X + f_{32}Y + f_{33}Z. \end{aligned} \quad (12.41)$$

Коэффициенты f_{ij} называются *коэффициентами упругости*, а матрица (f) — *матрицей упругости*. Если предположить, что рассматриваемая система консервативна, т. е. не может служить источником энергии, то $f_{ij} = f_{ji}$, матрица (f) оказывается симметрической и, следовательно, удовлетворяет уравнению

$$(f) = (f)^t. \quad (12.42)$$

Легко показать, что если уравнение (12.42) не удовлетворяется, то при перемещении точки P по замкнутому контуру будет совершена

некоторая конечная отличная от нуля работа. Именно поэтому матрица упругости консервативной системы должна быть симметрической. Не все системы имеют симметрическую матрицу упругости. Так, например, системы, находящиеся под воздействием потоков, такие, как части самолета, оказываются неконсервативными и потому, вообще говоря, имеют несимметрическую матрицу упругости.

В общем случае уравнения (12.41) могут быть решены относительно всех трех компонент X , Y и Z силы и записаны в виде

$$\begin{aligned} X &= c_{11}x + c_{12}y + c_{13}z, \\ Y &= c_{21}x + c_{22}y + c_{23}z, \\ Z &= c_{31}x + c_{32}y + c_{33}z. \end{aligned} \quad (12.43)$$

Коэффициенты c_{ij} называются *коэффициентами жесткости*, а квадратная матрица $(c) = (c_{ij})$ — *матрицей жесткости*. Эта матрица удовлетворяет соотношению

$$(c) = (f)^{-1},$$

т. е. является обратной для матрицы упругости. Если система консервативна, то из уравнений (12.42) и (12.43) следует, что (c) — симметрическая матрица, $c_{ij} = c_{ji}$.

Главное направление статической нагрузки в данной точке упругого тела определяется как такое направление, в котором оказываются параллельными приложенная в этой точке сила и вызываемое ею смещение. Пусть m_1 , m_2 , m_3 есть *направляющие косинусы* главного направления, R — результирующая сила и r — результирующее смещение. Тогда в случае параллельности силы и смещения получаем

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = r \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} = r(m); \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = R \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} = R(m). \quad (12.44)$$

Если соотношения (12.44) подставить в систему (12.41), то результат может быть записан в матричной форме следующим образом:

$$\mu(m) = (f)(m); \quad \mu = \frac{r}{R}.$$

Возможные значения μ определяются уравнением

$$\det(\mu U - f) = 0,$$

где U — единичная матрица третьего порядка. Это уравнение аналогично уравнению (12.10). Из него следует, что возможные значения μ , например μ_1 , μ_2 и μ_3 , являются *собственными значениями* матрицы упругости (f) . Каждому собственному значению μ_i соответствует *собственный вектор* $(m)_i$. Каждый собственный вектор определяет три направляющих косинуса и, следовательно, три главных

направления. Легко убедиться, что три собственных вектора оказываются взаимно ортогональными, если соответствующие собственные значения различны.

Так как рассматриваемая упругая система предполагается консервативной, то накопленная в ней в данном состоянии энергия не зависит от способа, с помощью которого система была приведена в это состояние. Для накопленной в системе потенциальной энергии может быть получено в матричных обозначениях удобное выражение. Пусть

$$(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

есть вектор смещения, а

$$(X) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

— вектор силы. Потенциальная энергия V может быть представлена в следующих формах:

$$V = \frac{1}{2} (X)_t (x) = \frac{1}{2} (x)_t (X) = \frac{1}{2} (x)_t (c) (x) = \frac{1}{2} (X)_t (f) (X),$$

где (f) — матрица упругости и (c) — матрица жесткости. В силу симметричности матрицы упругости консервативной системы потенциальная энергия V может быть записана в виде

$$\frac{1}{2} f_{11} X^2 + \frac{1}{2} f_{22} Y^2 + \frac{1}{2} f_{33} Z^2 + f_{23} YZ + f_{31} ZX + f_{12} XY = V = \text{const.}$$

Это соотношение определяет в пространстве X, Y, Z центральную поверхность второго порядка. В аналитической геометрии доказывается, что в случае действительных коэффициентов f эта поверхность имеет три действительные взаимортогональные оси. Направления этих осей совпадают с главными направлениями нагрузки.

§ 12.24. Приложение к теории конструкций. Рассмотрим упругую конструкцию, схематически изображенную на рис. 12.1. Пусть F_i представляет систему из n вертикально направленных сил, действующих на конструкцию, и пусть y_i — соответствующие деформации, вызываемые силами системы в точках $i = 1, 2, \dots, n$. Если упругая конструкция консервативна и удовлетворяет обобщенному закону Гука, то эти деформации связаны с силами матричным уравнением

$$(y) = (a) (F). \quad (12.45)$$

В этом уравнении (y) означает вектор-столбец из n деформаций, (F) — вектор-столбец из n сил и (a) — квадратную матрицу из коэф-

эффициентов упругости (или влияния) a_{rs} . Так как рассматриваемая система консервативна, то $(a)_t = (a)$ и (a) есть симметрическая матрица. Соотношение (12.45) может быть переписано в форме

$$(F) = (a)^{-1} (y) = (k) (y); \quad (k) = (a)^{-1},$$

где матрица (k) является обратной для матрицы упругости (a) и называется матрицей жесткости. Матрица жесткости консервативной системы симметрическая.

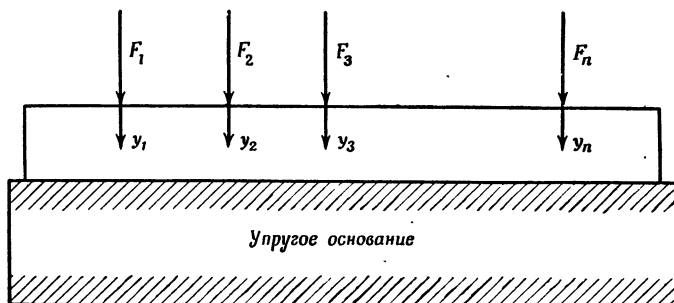


Рис. 12.1. Схема упругой конструкции; F_i — вертикальные силы, y_i — соответствующие деформации в точках $i = 1, 2, \dots, n$.

Если рассматриваемая упругая система консервативна, то накопленная в ней упругая энергия не зависит от способа приложения сил, ее деформирующих. Эта упругая энергия представляет собой потенциальную энергию V и определяется равенствами

$$V = \frac{1}{2} (F)_t (a) (F) = \frac{1}{2} (y)_t (k) (y).$$

Пусть (D) представляет следующий векторный оператор:

$$(D) = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{pmatrix}; \quad D_i = \frac{\partial}{\partial F_i}.$$

Тогда

$$(D)V = (D) \frac{1}{2} (F)_t (a) (F) = (a) (F).$$

Следовательно,

$$(y) = (D)V. \quad (12.46)$$

Соотношение (12.46) известно в теории конструкций под названием *теоремы Кастильяно*.

Тот факт, что обобщенный закон Гука [соотношение (12.45)] связывает силы и деформации линейных упругих конструкций с по-

мощью линейных преобразований, позволяет широко пользоваться в этой теории матричной алгеброй. Подобного рода приложения матричной алгебры можно найти в литературе, указанной в конце главы [11, 16].

Приложения матриц к задачам электротехники

§ 12.25. Задачи, относящиеся к сетям переменного тока. Матричные обозначения и матричная алгебра с большой пользой могут быть использованы для решения задач из теории сетей переменного тока. Рассмотрим общую сеть, состоящую из n контуров, находящихся в установившемся состоянии. Пусть соответствующие комплексные контурные токи имеют величины $(I_1, I_2, \dots, I_n)$, а комплексные напряжения, действующие в каждом из контуров, определяются величинами $(E_1, E_2, \dots, E_n)$. Напряжения и токи связаны системой линейных уравнений

$$\begin{aligned} Z_{11}I_1 + Z_{12}I_2 + \dots + Z_{1n}I_n &= E_1, \\ Z_{21}I_1 + Z_{22}I_2 + \dots + Z_{2n}I_n &= E_2, \\ \vdots & \\ Z_{n1}I_1 + Z_{n2}I_2 + \dots + Z_{nn}I_n &= E_n. \end{aligned} \quad (12.47)$$

Коэффициенты Z_{ij} являются, вообще говоря, комплексными числами. Коэффициенты Z_{rr} являются *собственными импедансами* различных контуров, а коэффициенты Z_{rs} при $r \neq s$ — *взаимными импедансами* системы. В сетях с установившимся током для взаимных импедансов выполняются условия

$$Z_{rs} = Z_{sr}.$$

Систему линейных уравнений (12.47) можно записать в компактной матричной форме

$$(Z)(I) = (E),$$

где (Z) — квадратная матрица n -го порядка, элементами которой являются комплексные числа Z_{rs} ; (I) — вектор-столбец из комплексных контурных токов и (E) — вектор-столбец из комплексных контурных напряжений.

В обычной задаче анализа электрических схем считаются заданными матрица (Z) и вектор напряжений (E) и требуется определить вектор тока (I). Эта задача имеет формальное решение

$$(I) = (Z)^{-1} (E) = (Y) (E), \quad (Y) = (Z)^{-1}. \quad (12.48)$$

Матрица (Y) обратна матрице импеданса и называется матрицей адмитанса. Уравнение (12.48) дает формальное решение задачи анализа схем. Для того чтобы избежать вычисления с комплексными величинами при отыскании матрицы, обратной матрице импеданса (Z),

удобно разбить эту матрицу и матрицу адмитанса (Y) на действительную и мнимую части с помощью следующих соотношений:

$$(Z) = (R) + j(X); \quad (\dot{Y}) = (G) + j(B); \quad j = \sqrt{-1}.$$

Матрица (R) называется матрицей сопротивлений, (X) — матрицей реактивных сопротивлений, (G) — матрицей проводимостей и (B) — матрицей реактивных проводимостей. Так как (Y) есть матрица, обратная (Z), то

$$\begin{aligned} (Z)(Y) &= (Z)(Z)^{-1} = U = (R + jX)(G + jB) = \\ &= (R)(G) - (X)(B) + j[(R)(B) + (X)(G)], \end{aligned} \quad (12.49)$$

где U — единичная матрица n -го порядка. Если приравнять действительные и мнимые коэффициенты в уравнении (12.49), то получим следующие выражения:

$$\begin{aligned} (R)(G) - (X)(B) &= U, \\ (X)(G) + (R)(B) &= 0, \end{aligned}$$

где 0 — нулевая матрица n -го порядка. Эти два уравнения могут быть решены относительно матриц (G) и (B), и результаты могут быть представлены в виде

$$\begin{aligned} (G) &= ((R) + (X)(R)^{-1}(X))^{-1}, \\ (B) &= -(R)^{-1}(X)(G); \end{aligned} \quad (12.50)$$

или

$$\begin{aligned} (B) &= -((X) + (R)(X)^{-1}(R))^{-1}, \\ (G) &= -(X)^{-1}(R)(B). \end{aligned} \quad (12.51)$$

Для вычисления матриц (G) и (B) может быть использована первая или вторая система уравнений. Выбор той или иной из этих двух систем уравнений определяется сложностью матриц (X) и (R). Определение матриц (G) и (B) по уравнениям (12.50) или (12.51) может быть осуществлено методом разбиения (см. § 12.7).

После определения матриц проводимости и реактивной проводимости можно найти вектор тока. Для этого удобно разложить векторы напряжения и силы тока на их действительные и мнимые части

$$\begin{aligned} (E) &= (E)_r + j(E)_i, \\ (I) &= (I)_r + j(I)_i. \end{aligned}$$

Здесь индексы r и i обозначают соответственно действительную и мнимую части. Тогда получаем

$$(I) = (Y)(E) = ((G) + j(B))((E)_r + j(E)_i).$$

Если приравнять соответственно действительную и мнимую части, получим

$$(I)_r = (G)(E)_r - (B)(E)_i,$$

$$(I)_i = (B)(E)_r + (G)(E)_i.$$

Эти соотношения определяют токи в различных контурах схемы.

§ 12.26. Задача на схемы трехфазного тока. Матричная алгебра находит практическое применение в электротехнике при изучении трехфазных схем [21, 26]. Для пояснения общего принципа рассмотрим прибор с четырьмя входами, изображенный на рис. 12.2.

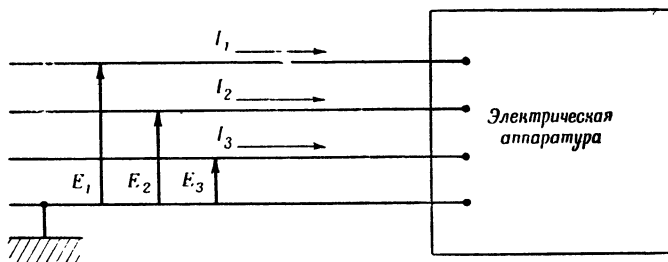


Рис. 12.2. Схема трехфазного электрического контура с четырьмя клеммами, комплексными токами I_1, I_2, I_3 и комплексными напряжениями E_1, E_2, E_3 относительно земли.

Пусть комплексные токи обозначены через I_1, I_2, I_3 , а комплексные напряжения E_1, E_2, E_3 (относительно земли) в линиях измеряются как указано на рисунке. Применяя закон Кирхгофа, можно составить следующие три уравнения, связывающие токи и напряжения:

$$\begin{aligned} E_1 &= Z_{11}I_1 + Z_{12}I_2 + Z_{13}I_3, \\ E_2 &= Z_{21}I_1 + Z_{22}I_2 + Z_{23}I_3, \\ E_3 &= Z_{31}I_1 + Z_{32}I_2 + Z_{33}I_3. \end{aligned} \quad (12.52)$$

Комплексные числа Z_{ij} являются собственными импедансами и взаимными импедансами этого прибора. Система уравнений (12.52) может быть записана в матричной форме

$$(E) = (Z)(I), \quad (12.53)$$

в которой (E) — вектор-столбец из напряжений, (Z) — квадратная матрица из импедансов и (I) — вектор-столбец сил тока. В общей трехфазной задаче требуется найти токи по заданной системе коэффициентов матрицы импедансов прибора и по вектору-столбцу напряжений. Это всегда может быть выполнено, так как (Z) , будучи невырож-

денной матрицей, имеет обратную. Таким образом, токи задаются равенством

$$(I) = (Z)^{-1} (E).$$

Практически было установлено, что матрица (Z) импедансов прибора, наиболее часто используемого в системах, работающих на трехфазном электрическом токе, обладает следующей симметрией:

$$(Z) = \begin{pmatrix} Z_0 & Z_1 & Z_2 \\ Z_2 & Z_0 & Z_1 \\ Z_1 & Z_2 & Z_0 \end{pmatrix}. \quad (12.54)$$

Матрица с подобного рода симметрией называется *циркулянтной*. Матрицы импедансов трехфазных синхронных и индукционных генераторов и моторов с равномерным воздушным зазором обладают указанной симметрией. Для облегчения изучения схем, матрица которых обладает циркулянтной симметрией, с помощью линейных преобразований

$$(I)_s = (S)^{-1} (I),$$

$$(E)_s = (S)^{-1} (E)$$

вводятся в рассмотрение новые токи и напряжения.

Эти равенства определяют два вектора-столбца $(I)_s$ и $(E)_s$, элементы которых называются *симметрическими составляющими* тока и потенциала. Матрица (S) преобразования имеет следующий вид:

$$(S) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & a^2 & a \\ 1 & a & a^2 \end{pmatrix}; \quad a = e^{2\pi j/3}, \quad j = \sqrt{-1}.$$

Число a — комплексное и является кубическим корнем из единицы, лежащим во втором квадранте. Матрица (S) преобразования — невырожденная и имеет своей обратной матрицу

$$(S)^{-1} = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{pmatrix}.$$

Вектор линейного тока и вектор напряжения даются уравнениями

$$(I) = (S) (I)_s; \quad (E) = (S) (E)_s.$$

Уравнение Кирхгофа в матричной форме (12.53) может быть переписано в виде

$$((S)^{-1} (Z) (S)) (I)_s = (E)_s = (Z)_s (I)_s.$$

Матрица

$$(Z)_s = (S)^{-1} (Z) (S) \quad (12.55)$$

называется *симметрической составляющей матрицы импедансов*. Если матрица (Z) импедансов схемы имеет циркулянтную симметрию (12.54), то легко можно показать, что симметрическая составляющая матрица (12.55) имеет следующий диагональный вид:

$$(Z)_s = \begin{pmatrix} Z_0 + Z_1 + Z_2 & 0 & 0 \\ 0 & Z_0 + a^2 Z_1 + a Z_2 & 0 \\ 0 & 0 & Z_0 + a Z_1 + a^2 Z_2 \end{pmatrix}.$$

Симметрические составляющие тока и потенциала оказываются, таким образом, несвязанными между собой. Это приводит к значительному упрощению исследования.

§ 12.27. Теория четырехполюсников. Одной из специальных форм электрической схемы, имеющей большое значение в технике является устройство с двумя входными и двумя выходными клеммами

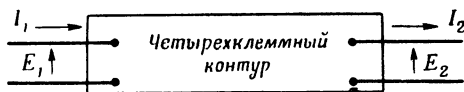


Рис. 12.3. Схема электрического контура с четырьмя клеммами или квадриполя с двумя входными и двумя выходными клеммами, с комплексными входными и выходными токами I_1 , I_2 и с комплексными потенциалами E_1 , E_2 , создаваемыми внешними источниками.

Обычно такая схема называется *четыреполюсником*, или *квадриполем*. Схематично четырехполюсник изображен на рис. 12.3 в виде прямоугольника с двумя парами клемм.

Основные направления для напряжения и тока примем такими как показано на рис. 12.3. Можно предположить, что в четырехполюсник входит электрическая схема с линейными параметрами, не имеющая источников энергии. Напряжения E_1 и E_2 — комплексные напряжения, создаваемые источниками, находящимися вне схемы. Токи I_1 и I_2 являются комплексными входными и выходными токами. Можно показать, что общим следствием закона Кирхгофа и предполагающейся линейности системы [5,20] являются следующие линейные соотношения:

$$E_1 = AE_2 + BI_2, \quad I_1 = CE_2 + DI_2 \quad (12.56)$$

между входными и выходными величинами четырехполюсника.

Коэффициенты A , B , C и D называются общими параметрами четырехполюсника. Они являются, вообще говоря, комплексными числами, которые зависят от внутренней конструкции схемы. Все эти

четыре параметра схемы не являются независимыми, а связаны соотношением

$$AD - BC = 1.$$

Если ввести в рассмотрение векторы-столбцы

$$(X)_1 = \begin{pmatrix} E_1 \\ I_1 \end{pmatrix}, \quad (X)_2 = \begin{pmatrix} E_2 \\ I_2 \end{pmatrix},$$

то уравнение (12.56) можно переписать в удобном виде

$$(X)_1 = (T)(X)_2, \quad (12.57)$$

где

$$(T) = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad \det(T) = AD - BC = 1.$$

Квадратная матрица (T) называется *передаточной матрицей* сети. Эта матрица невырожденная и имеет обратную

$$(T)^{-1} = \begin{pmatrix} D & -B \\ -C & A \end{pmatrix}.$$

Из уравнений (12.57) видно, что выходные величины могут быть выражены через входные с помощью уравнения

$$(X)_2 = (T)^{-1} (X)_1.$$

В случае симметричности четырехполюсника его работа может быть обращена без изменения связей между входными и выходными величинами. Для симметричной цепи выполняется условие $A = D$.

§ 12.28. Распространение волн по каскаду из симметричных четырехполюсников. Многие важные технические задачи в области техники связи требуют определения законов распределения токов и

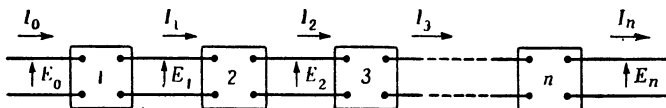


Рис. 12.4. Схема каскада, состоящего из одинаковых симметричных четырехполюсников.

напряжений в цепи, составленной из идентичных симметричных четырехполюсников. На рис. 12.4 изображена структура каскада из n идентичных четырехполюсников.

Пусть передаточная матрица одного из симметричных четырехполюсников указанного на рис. 12.4 каскада имеет вид

$$(T) = \begin{pmatrix} A & B \\ C & A \end{pmatrix}, \quad \det(T) = A^2 - BC = 1.$$

Введем вектор-столбец

$$(X)_i = \begin{pmatrix} E_i \\ I_i \end{pmatrix}.$$

Элементами этого вектора-столбца являются комплексный ток и потенциал на выходе i -го четырехполюсника каскада. Так как каждый из четырехполюсников каскада имеет своей передаточной матрицей одну и ту же матрицу, то напряжение и ток на выходе n -го четырехполюсника каскада удовлетворяют следующему соотношению:

$$(X)_0 = (T)^n (X)_n, \quad n = 0, 1, 2, \dots \quad (12.58)$$

Применяя теорему Сильвестра, матрице $(T)^n$ можно легко придать следующий вид:

$$(T)^n = \begin{pmatrix} A_n & B_n \\ C_n & A_n \end{pmatrix} = \begin{pmatrix} \operatorname{ch} an & Z_0 \operatorname{sh} an \\ \frac{\operatorname{sh} an}{Z_0} & \operatorname{ch} an \end{pmatrix}, \quad n = 0, 1, 2, \dots, \quad (12.59)$$

где

$$a = \operatorname{Ar} \operatorname{ch} A, \quad Z_0 = \left(\frac{B}{C} \right)^{1/2}. \quad (12.60)$$

Величина a называется *функцией распространения* четырехполюсника каскада, а Z_0 — *характеристическим импедансом* четырехполюсника.

Уравнение (12.58) мы можем переписать в виде

$$(X)_n = (T)^{-n} (X)_0,$$

и тем самым выразить выходные величины каскада через его входные величины. Можно показать, что уравнение (12.59) остается справедливым для целых отрицательных значений n и потому может быть использовано для вычисления матрицы $(T)^{-n}$. Многие характеристики периодических электрических систем, таких, например, как контуры фильтров, можно легко получить, если использовать уравнения (12.58). Можно показать, например, что если в конце каскада находится система с импедансом, равным Z_0 , так что

$$\frac{E_n}{I_n} = Z_0,$$

то как следствие уравнений (12.58) будет иметь место следующее распределение напряжений и токов в каскаде, показанном на рис. 12.4:

$$E_k = E_0 e^{-ak}, \quad I_k = \frac{E_0 e^{-ak}}{Z_0}, \quad k = 0, 1, 2, \dots, n.$$

Из этих уравнений видно, что в том случае, когда функция проводимости a есть величина чисто мнимая, прохождение тока и напряжения по каскаду четырехполюсников будет беспрепятственным (свободным), если же a — комплексное число, то токи и напряжения

оказываются ослабленными. В технической литературе по фильтрующим схемам принято говорить, что схема работает в *полосе пропускания*, если a — величина чисто мнимая; если же a — величина комплексная, то схема работает в *полосе затухания*. Так как на основании уравнения (12.60) мы имеем $\operatorname{ch} a = A$, то легко видеть, что

$$\text{для полосы пропускания } |A| < 1, \quad (12.61)$$

$$\text{для полосы затухания } |A| > 1. \quad (12.62)$$

Поведение каскада в отношении пропускания или затухания зависит от параметра A каскада. Так как A является функцией частоты, то можно заметить, что все частоты, удовлетворяющие неравенству (12.61), беспрепятственно пропускаются, тогда как на частоты, удовлетворяющие неравенству (12.62), этот каскад действует как фильтр.

С помощью уравнений, аналогичных уравнению (12.58), могут быть изучены свойства большого числа различных электрических систем, таких как трансформатор, трансмиссионные линии, линии задержки и др., если только оказываются известными их передаточные матрицы. В табл. 12.1 даны передаточные матрицы многих из часто встречающихся систем.

Приложения матриц к задачам о колебаниях

§ 12.29. Колебания консервативных систем. В фундаментальной работе [8] английских ученых Дункана и Коллара, опубликованной в 1934 г., теория матриц применялась для определения периодов свободных колебаний динамических систем, в которых демпфирующими силами можно пренебречь. Уравнения движения такой системы могут быть записаны в удобных матричных обозначениях:

$$(m)(\ddot{q}) + (k)(q) = 0. \quad (12.63)$$

Здесь (m) есть квадратная матрица, элементами которой являются коэффициенты m_{rs} инерции, (k) — квадратная матрица, элементы k_{rs} которой суть коэффициенты жесткости, и (q) — вектор-столбец из обобщенных координат q_i . Если система имеет n степеней свободы, то матрица (m) инерции и матрица (k) жесткости являются квадратными матрицами n -го порядка и координатный вектор (q) — матрицей-столбцом n -го порядка.

Кинетическая энергия T и потенциальная энергия V системы определяются выражениями

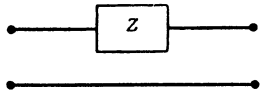
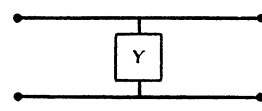
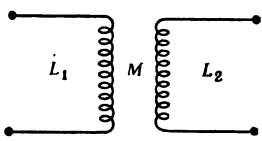
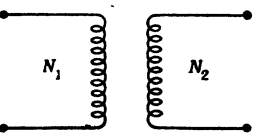
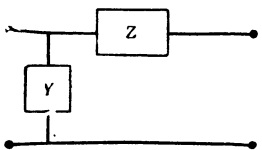
$$T = \frac{1}{2} (\dot{q})_t (m) (\dot{q}), \quad V = \frac{1}{2} (q)_t (k) (q).$$

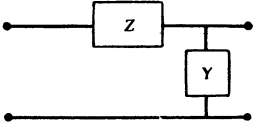
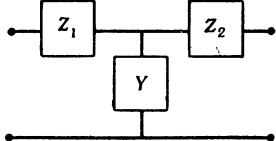
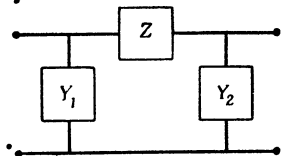
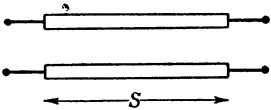
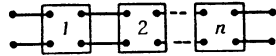
В целях определения собственных частот и типов колебаний системы допустим, что решение уравнения (12.63) имеет вид

$$(q) = (A) \sin(\omega t + \theta). \quad (12.64)$$

Таблица 12.1

ПЕРЕДАТОЧНЫЕ МАТРИЦЫ ОСНОВНЫХ ЧЕТЫРЕХПОЛЮСНИКОВ

№	Контур	Матри: $a = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = (T)$
1	 Последовательно включенное сопротивление	$\begin{pmatrix} 1 & Z \\ 0 & 1 \end{pmatrix}$
2	 Шунтированная проводимость	$\begin{pmatrix} 1 & 0 \\ Y & 1 \end{pmatrix}$
3	 Связанные контуры	$\begin{pmatrix} \frac{L_1}{M} & \frac{j\omega(L_1L_2 - M^2)}{M} \\ -\frac{j}{\omega M} & \frac{L_2}{M} \end{pmatrix}$
4	 Идеальный трансформатор $a = \frac{N_1}{N_2}$	$\begin{pmatrix} \frac{1}{a} & 0 \\ 0 & a \end{pmatrix}$
5		$\begin{pmatrix} 1 & Z \\ Y & 1 + YZ \end{pmatrix}$

№	Контур	Матрица = $\begin{pmatrix} A & B \\ C & D \end{pmatrix} = (T)$
6		$\begin{pmatrix} 1 + YZ & Z \\ Y & 1 \end{pmatrix}$
7	 Т-секция	$\begin{pmatrix} 1 + YZ_1 & Z_1 + Z_2 + YZ_1Z_2 \\ Y & 1 + YZ_2 \end{pmatrix}$
8	 П-секция	$\begin{pmatrix} 1 + ZY_2 & Z \\ Y_1 + Y_2 + ZY_1Y_2 & 1 + ZY_1 \end{pmatrix}$
9	 Линия передачи	$\begin{pmatrix} \operatorname{ch} aS & Z_0 \operatorname{sh} aS \\ \frac{\operatorname{sh} aS}{Z_0} & \operatorname{ch} aS \end{pmatrix}$ a — функция распространения Z_0 — характеристический импеданс S — длина линии
10	 Каскад из симметричных одинаковых контуров. Матрица каждого отдельного контура имеет вид $\begin{pmatrix} A & B \\ C & A \end{pmatrix}$	Общая матрица всех контуров $\begin{pmatrix} \operatorname{ch} an & Z_0 \operatorname{sh} an \\ \frac{\operatorname{sh} an}{Z_0} & \operatorname{ch} an \end{pmatrix}$ $a = \operatorname{Ar} \operatorname{ch} A$ — функция распространения $Z_0 = \sqrt{\frac{B}{C}}$ — характеристический импеданс

Это решение соответствует чисто гармоническим колебаниям системы. Здесь (A) есть вектор-столбец из амплитуд, ω — угловая частота колебаний и θ — фазовый угол. Если решение (12.64) подставить в уравнение (12.63) и результат разделить на скалярный множитель $\sin(\omega t - \theta)$, то можно получить следующее соотношение:

$$\begin{aligned} (k - \omega^2 m)(A) &= 0, \\ (U - \omega^2 D)(A) &= 0, \end{aligned} \quad (12.65)$$

где U — единичная матрица n -го порядка и (D) — динамическая матрица, определяемая соотношением

$$(D) = (k)^{-1} (m).$$

Если

$$\mu = \frac{1}{\omega^2},$$

то уравнение (12.65) можно записать в виде

$$(\mu U - D)(A) = 0. \quad (12.66)$$

Это уравнение представляет систему однородных линейных уравнений относительно элементов вектора (A) . Для существования нетривиальных решений уравнения (12.66) необходимо, чтобы

$$\det(\mu U - D) = 0. \quad (12.67)$$

Это есть *характеристическое уравнение* системы. Уравнение (12.67) является, как можно заметить, характеристическим уравнением динамической матрицы (D) . Возможные значения μ оказываются *собственными значениями* (D) и в общем случае при наличии n степеней свободы существуют, вообще говоря, n собственных значений $\mu_1, \mu_2, \dots, \mu_n$. Каждому собственному значению μ_i соответствует *собственная угловая частота* ω_i системы, определяемая соотношением

$$\omega_i = \left(\frac{1}{\mu_i} \right)^{1/2}. \quad (12.68)$$

Каждому собственному значению μ_i соответствует *собственный вектор* $(A)_i$. Такой собственный вектор удовлетворяет уравнению

$$(\mu_i U - D)(A)_i = 0, \quad i = 1, 2, \dots, n;$$

n собственных векторов $(A)_i$ называются *модальными столбцами* данной динамической системы. Общее решение уравнения (12.63) в том случае, когда все собственные значения различны, может быть записано через модальные столбцы и собственные частоты в виде

$$(q) = \sum_{r=1}^n C_r (A)_r \sin(\omega_r t + \theta_r). \quad (12.69)$$

Величины C_r и θ_r являются обычными константами и могут быть определены при известных начальных условиях.

Используя теорему Сильвестра, можно показать, что если μ_1 есть наибольшее собственное значение (D) , то для достаточно большого k получим

$$(D)^k(x) = c\mu_1^k(A)_1, \quad (12.70)$$

где (x) есть произвольный вектор-столбец n -го порядка, c — постоянная и $(A)_1$ — главный модальный столбец динамической системы. Применяя уравнение (12.70), можно определить собственное значение μ_1 и колонку, пропорциональную $(A)_1$. Если μ_1 известно, то из уравнения (12.68) можно найти нижнюю или основную угловую частоту ω_1 . После определения $(A)_1$ и ω_1 Дункан и Коллар [8] предлагают использовать итерационный метод для определения второго по величине собственного значения μ_2 и $(A)_2$. Итерационный метод может быть использован для определения всех последующих собственных значений и собственных векторов. Было проведено много исследований, целью которых являлось получение наиболее эффективного способа отделения собственных значений и соответствующих им собственных векторов динамической матрицы (D) . Перечень результатов, достигнутых за последнее время в этой области, можно найти в книге [17].

Если определены различные собственные векторы, то можно составить следующую квадратную матрицу:

$$(A) = ((A)_1 (A)_2 (A)_3 \dots (A)_n).$$

Различные столбцы этой квадратной матрицы (A) являются модальными столбцами. Можно показать, что если все n собственных значений динамической матрицы (D) различны, то матрица (A) невырожденная и имеет обратную $(A)^{-1}$. Если элементы $N_1, N_2, \dots, N_n$ вектора-столбца (N) связаны с обобщенными координатами $q_1, q_2, \dots, q_n$ при помощи линейного преобразования

$$(N) = (A)^{-1}(q),$$

то можно установить, используя уравнение (12.69), что координаты N_i изменяются по простому гармоническому закону:

$$N_i = C_i \sin(\omega_i t + \theta_i), \quad i = 1, 2, \dots, n,$$

где C_i — обычная константа. Таким образом, колебательные движения координат N_i оказываются несвязанными между собой. Эти координаты называются *нормальными координатами* системы.

Использование матриц оказалось весьма ценным при изучении технических проблем, связанных с колебаниями. Матрицы широко применяются при исследовании вибраций самолета и проблем флаттера [24], а также для характеристики электрических схем, имеющих наименьшее сопротивление [19].

§ 12.30. Колебания неконсервативных систем. Поведение динамических систем, колебания которых затухают вследствие наличия сил трения типа „вязкости“, также может быть изучено с помощью матриц. В статье [9], опубликованной в 1935 г., Дункан и Коллар обобщили свой метод анализа колеблющихся систем на случай наличия сил трения. Неконсервативная система, рассмотренная этими авторами, характеризовалась следующими функциями:

$$T = \frac{1}{2} (\dot{q})_t (m) (\dot{q}) = \text{кинетическая энергия,}$$

$$V = \frac{1}{2} (q)_t (k) (q) = \text{потенциальная энергия,}$$

$$F = \frac{1}{2} (\dot{q})_t (R) (\dot{q}) = \text{функция рассеяния.}$$

Квадратные матрицы (m) и (k) являются соответственно матрицами инерции и жесткости системы, (q) есть вектор обобщенных координат, R — квадратная матрица из коэффициентов демпфирования R_{ij} . Уравнение свободных колебаний системы имеет вид

$$(m)(\ddot{q}) + (R)(\dot{q}) + (k)(q) = 0. \quad (12.71)$$

Эта система уравнений может быть преобразована в систему первого порядка с помощью введения вектора-столбца

$$(Y) = \begin{pmatrix} (q) \\ (\dot{q}) \end{pmatrix}$$

и квадратной матрицы

$$(W) = \begin{pmatrix} (0) & (U) \\ -(m)^{-1}(k) & -(m)^{-1}(R) \end{pmatrix},$$

где (0) — нулевая, а U — единичная матрицы n -го порядка. Квадратная матрица (W) оказывается, таким образом, матрицей порядка $2n$ с указанным выше разбиением. Легко видеть, что с помощью этих обозначений уравнение (12.71) может быть записано в виде

$$D(Y) = (W)(Y); \quad D = \frac{d}{dt}.$$

Решение этого уравнения с помощью метода, изложенного в § 12.21, можно записать в компактной форме

$$(Y) = e^{(W)t} (Y)_0, \quad (12.72)$$

где $(Y)_0$ есть вектор-столбец, элементами которого являются начальные значения обобщенных координат и скоростей при $t=0$. Для того чтобы интерпретировать уравнение (12.72) с помощью формулы Сильвестра, необходимо определить собственные значения квадратной матрицы (W) . Эти собственные значения оказываются комплексными числами u_i и могут быть найдены с помощью приемов

матричного умножения [9, 11], примененных к матрице (W). Дункан и Коллар успешно использовали этот метод для изучения вибраций крыла.

§ 12.31. Заключение. В настоящей статье достаточно подробно изложены основные принципы, а также примеры использования матричной алгебры в различных областях техники. Эти примеры свидетельствуют о больших возможностях и практических удобствах, связанных с применением матричной алгебры к различным задачам техники. Естественно, что в такой небольшой статье невозможно подробно рассмотреть примеры из всех областей приложения матричной алгебры к важным техническим проблемам. Укажем дополнительно на следующие технические проблемы, в которых она находит применение:

- а) синтез микроволновых фильтров [15];
- б) схемы на кристаллических триодах [25];
- в) использование кристаллов в технике [4];
- г) многоканальные линии передач [23];
- д) многообмоточные трансформаторы [14, 26];
- е) схемы на многоэлектродных вакуумных лампах [14].

Этот список указывает на новые области техники, в которых находит применение матричная алгебра. Более подробные сведения можно найти в перечисленных ниже работах.

ЛИТЕРАТУРА

1. Baker H. F., On the Integration of Linear Differential Equations, Proc. London Math. Soc., 35, 333—378 (1903).
2. Baker H. F., Note on the Integration of Linear Differential Equations, Proc. London Math. Soc., 2, 293—296 (1905).
3. Бохер М., Введение в высшую алгебру, Гостехиздат, М. — Л., 1934.
4. Bond W. L., The Mathematics of the Physical Properties of Crystals, Bell System Tech. J., 22, 1—72 (1943).
5. Brillouin L., Wave Propagation in Periodic Structures, McGraw-Hill, New York, 1946, pp. 193—239.
6. Cayley A. A., Memoir on the Theory of Matrices, Philos. Trans. Roy. Soc. London, Ser. A, 148, 17—37 (1857).
7. Курант Р. и Гильберт Д., Методы математической физики, т. I, изд. 3, и т. II, изд. 2, Гостехиздат, М. — Л., 1951.
8. Duncan W. J. and Collar A. R., A Method for the Solution of Oscillation Problems by Matrices, Phil. Mag. (7), 17, 865—909 (1934).
9. Duncan W. J. and Collar A. R., Matrices Applied to the Motions of Damped Systems, Phil. Mag. (7), 19, 197—219 (1935).
10. Forsythe G. E., Solving Linear Equations Can Be Interesting, Bull. Amer. Math. Soc., 59, 299—329 (1953).
11. Фрезер Р., Дункан В. и Коллар А., Теория матриц и ее приложения к дифференциальным уравнениям и динамике, ИЛ, М., 1950.

12. Hotelling H., Some New Methods in Matrix Calculation, *Ann. Math. Statist.*, **14**, 1—34 (1943).
13. Kemble E. C., *The Fundamental Principles of Quantum Mechanics*, McGraw-Hill, New York, 1937, ch. 10.
14. Kron G., *Tensor Analysis of Networks*, John Wiley, New York, 1939.
15. Montgomery C. G., Dicke R. H. and Purcell E. M. (eds.), *Principles of Microwave Circuits*, pp. 405—411, Massachusetts Institute of Technology Radiation Laboratory Series, McGraw-Hill, New York, 1948.
16. Murnaghan F. D., *Finite Deformation of an Elastic Solid*, John Wiley, New York, 1951.
17. Paige L. J. and Taussky Olga (eds.), *Simultaneous Linear Equations and the Determination of Eigenvalues*, *Nat. Bur. Standards Appl. Math. Ser.*, No. 29, 1953.
18. Pipes L. A., Matrix Solution of Polynomial Equations, *J. Franklin Inst.*, **225**, 437—454 (1938).
19. Pipes L. A., The Matrix Theory of Oscillatory Networks, *J. Appl. Phys.*, **10**, 849—860 (1939).
20. Pipes L. A., The Matrix Theory of Four-terminal Networks, *Phil. Mag.* (7), **30**, 370—395 (1940).
21. Pipes L. A., Linear Transformations in Three-phase Circuits, *Trans. Amer. Inst. Electr. Engrs*, **30**, 351—356 (1941).
22. Pipes L. A., Matrix Solution of Equations of the Mathieu-Hill Type, *J. Appl. Phys.*, **24**, 902—910 (1953).
23. Rice S. O., Steady-state Solutions of Transmission Line Equations, *Bell System Techn. J.*, **20**, 131—178 (1941).
24. Scanlan R. H. and Rosenbaum R., *Aircraft Vibration and Flutter*, Macmillan, New York, 1951.
25. Shekel J., Matrix Representation of Transistor Circuits, *Proc. IRE*, **40**, 1493—1497 (1952).
26. Stigant S. A., *Modern Electrical Engineering Mathematics*, Hutchinson, London, 1946.
27. Wedderburn J. H. M., *Lectures on Matrices*, American Mathematical Society, J. W. Edwards, Ann Arbor, Mich., 1949.

Глава 13

МЕТОДЫ КОНФОРМНЫХ ОТОБРАЖЕНИЙ

ЭДВИН Ф. БЕККЕНБАХ

Калифорнийский университет, Лос-Анжелес

§ 13.1. Введение. Многие научные исследования основаны на *изоморфизме* (или структурном подобии) различных физических и математических объектов. В частности это относится к рассмотренным в настоящей книге вопросам и к математическим *моделям*, описывающим их иногда с большей, а иногда с меньшей степенью точности. Иногда эти модели могут значительно изменяться в результате *преобразований*, но вновь получаемые модели сохраняют свою значимость ввиду *инвариантности* некоторых их свойств относительно таких преобразований.

Важная роль конформных отображений в прикладной математике основана на следующих простых соображениях. Часто в математических моделях различных физических проблем фигурируют функции $x(u, v)$, которые являются решениями *уравнения Лапласа*

$$\nabla^2 x \equiv \frac{\partial^2 x}{\partial u^2} + \frac{\partial^2 x}{\partial v^2} = 0 \quad (13.1)$$

(например, бывает необходимо определить поведение таких функций внутри области, когда известны только их значения на границе). Решения таких задач известны или легко определимы для некоторых плоских областей канонического вида; заданная плоская область может быть конформно отображена на подходящим образом выбранную каноническую область; при этом равенство нулю *лапласиана* $\nabla^2 x \equiv \partial^2 x / \partial u^2 + \partial^2 x / \partial v^2$ инвариантно относительно конформных преобразований. Таким образом, решение уравнения Лапласа в заданной области определяется через решение в канонической области.

В этой главе мы разовьем отмеченные выше соображения, частично иллюстрируя их на примерах, и увидим, как различные стороны этого вопроса соответствуют друг другу.

§ 13.2. Картографические проекции. Мы введем понятие конформных отображений, следуя ходу исторического развития на основе рассмотрения задачи о построении плоских карт земного шара. По-видимому, все древние и первобытные народы в какой-то степени обладали умением составлять и понимать географические

карты. Изучение карт обогатило наш язык; например, слова „долгота“ и „широта“ берут свое начало от средневековых прямоугольных карт, а слово „ориентировка“ происходит от того, что на римских картах направление на восток изображалось как направление к верхнему краю листа.

Будем считать, что единичная сфера

$$S: x^2 + y^2 + z^2 = 1 \quad (13.2)$$

приблизненно подобна поверхности земного шара. Пусть θ и φ обозначают соответственно долготу и широту точек сферы и, следовательно,

$$x = \cos \theta \cos \varphi,$$

$$y = \sin \theta \cos \varphi,$$

$$z = \sin \varphi.$$

Давно известно, что никакая область на сфере не может быть отображена на плоскую область без искажения; иными словами — сфера не является *развертывающейся* поверхностью. Но существуют отображения, сохраняющие определенные желательные нам геометрические свойства, и сейчас мы рассмотрим некоторые из этих отображений.

Цилиндрическое отображение

$$u = \theta, \quad v = v(\varphi), \quad (13.3)$$

где $v(0) = 0$ и $v(\varphi) > 0$ при $\varphi > 0$, сферы S на плоскость uv можно получить, проектируя точки сферы S (эта проекция зависит от функции $v(\varphi)$) на прямой круговой цилиндр, касающийся сферы по экватору $\varphi = 0$. Затем цилиндрическая поверхность разрезается по вертикали (обычно там, где $\theta = -\pi$ и $\theta = \pi$) и разворачивается на плоскость (см. рис. 13.1).

Читатель может убедиться в том, что функция

$$v(\varphi) = \varphi$$

осуществляет равностороннее отображение, при котором клеткам на сфере, ограниченным дугами кругов долготы и широты, получающимся при возрастании долготы и широты на одну и ту же

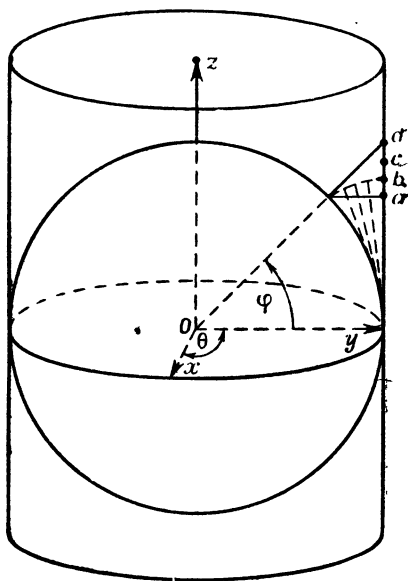


Рис. 13.1. Цилиндрические отображения единичной сферы для угла φ , равного $\pi/4$.

Высоты $a = \sin \varphi$, $b = \varphi$, $c = \log \operatorname{tg} (\pi/4 + \varphi/2)$, $d = \operatorname{tg} \varphi$ соответствуют проекциям равновеликой, равносторонней, конформной, или проекции Меркатора, и центральной.

величину, соответствуют на карте квадратные клетки, подобные клеткам на шахматной доске.

Функция $v(\varphi) = \operatorname{tg} \varphi$ осуществляет, очевидно, *центральную* цилиндрическую проекцию, т. е. проекцию на цилиндр из центра сферы.

В дифференциальной геометрии для поверхностей, заданных в виде

$$S: x = x(u, v), \quad y = y(u, v), \quad z = z(u, v), \quad (13.4)$$

фундаментальную роль играют величины

$$\begin{aligned} E &= \left(\frac{\partial x}{\partial u} \right)^2 + \left(\frac{\partial y}{\partial u} \right)^2 + \left(\frac{\partial z}{\partial u} \right)^2, \\ G &= \left(\frac{\partial x}{\partial v} \right)^2 + \left(\frac{\partial y}{\partial v} \right)^2 + \left(\frac{\partial z}{\partial v} \right)^2, \\ F &= \frac{\partial x}{\partial u} \frac{\partial x}{\partial v} + \frac{\partial y}{\partial u} \frac{\partial y}{\partial v} + \frac{\partial z}{\partial u} \frac{\partial z}{\partial v}. \end{aligned}$$

Действительно, как известно, элемент площади dA на поверхности S , соответствующий элементу da на плоскости uv , равен

$$dA = (EG - F^2)^{1/2} da, \quad (13.5)$$

а элемент длины dL на поверхности S определяется соотношением

$$dL^2 = E du^2 + 2F du dv + G dv^2. \quad (13.6)$$

Отсюда следует, что для приведенных выше цилиндрических отображений [уравнения (13.3)] сферы на плоскость

$$\begin{aligned} \frac{\partial x}{\partial u} &= \frac{\partial x}{\partial \theta} \frac{\partial \theta}{\partial u} + \frac{\partial x}{\partial \varphi} \frac{\partial \varphi}{\partial u} = \frac{\partial x}{\partial \theta} = -\sin \theta \cos \varphi, \\ \frac{\partial x}{\partial v} &= \frac{\partial x}{\partial \theta} \frac{\partial \theta}{\partial v} + \frac{\partial x}{\partial \varphi} \frac{\partial \varphi}{\partial v} = \frac{\partial x}{\partial \varphi} = -\cos \theta \sin \varphi \frac{\partial \varphi}{\partial v} \end{aligned}$$

и т. д., откуда

$$E = \cos^2 \varphi, \quad F = 0, \quad G = \left(\frac{d\varphi}{dv} \right)^2. \quad (13.7)$$

Из равенства $F = 0$ следует, что цилиндрические отображения являются *ортогональными*, т. е. ортогональной системе прямых $u = \text{const}$ и $v = \text{const}$ в плоскости uv соответствует на сфере *ортогональная сеть* — результат, который для данного случая легко было предвидеть из геометрических соображений.

Чтобы при отображении сферы на плоскость каждой области A на сфере S соответствовала равновеликая область a на плоскости uv , должно выполняться условие $dA = da$, или, в силу соотношения (13.5), равенство

$$EG - F^2 = 1.$$

В силу соотношений (13.7) при цилиндрических отображениях это имеет место тогда и только тогда, когда

$$\frac{dv}{d\varphi} = \cos \varphi, \text{ или } v = \sin \varphi = z.$$

Такое отображение получается проектированием точек сферы с помощью прямых, перпендикулярных к полярной оси. Это отображение является одновременно ортогональным и сохраняющим площади, но оно не свободно от искажений, так как все параллели, за исключением экватора, удлиняются, а все меридиональные большие круги укорачиваются.

Конформными отображениями называются отображения, сохраняющие углы. Это значит, что двум кривым, пересекающимся на S под некоторым углом, на плоскости соответствуют кривые, пересекающиеся под тем же углом. Конформные отображения можно эквивалентным образом охарактеризовать как „преобразования, подобные в малом“, или как преобразования, при которых *коэффициент линейного искажения* dL/dl (зависящий, вообще говоря, не только от точки (u, v) , но и от направления dv/du) является положительной функцией одной только точки (u, v) :

$$\frac{dL}{dl} = \lambda(u, v) > 0.$$

Из уравнения (13.6) следует, что условиями конформности являются

$$E = G > 0, \quad F = 0, \quad (13.8)$$

ибо тогда

$$dL^2 = E(u, v) dl^2. \quad (13.9)$$

В силу равенств (13.7), условия конформности (13.8) для цилиндрических отображений принимают вид

$$\frac{dv}{d\varphi} = \sec \varphi,$$

или

$$v = \log(\operatorname{tg} \varphi + \sec \varphi) = \log \operatorname{tg} \left(\frac{\pi}{4} + \frac{\varphi}{2} \right).$$

Эта функция вместе с функцией $u = \theta$ дает знаменитую *проекцию Меркатора*, введенную им в 1569 г.

Четыре цилиндрические проекции, которые мы рассмотрели выше, а именно равносторонняя, центральная, параллельная или равновеликая, и конформная, или проекция Меркатора, изображены на рис. 13.1 для значений углов $\theta = \pi/2$ и $\varphi = \pi/4$.

При *конических* отображениях описанный выше цилиндр заменяется конусом. Часто вместо них рассматриваются *поликонические* отображения, при которых главный меридиан отображается без искажений на прямолинейный отрезок, а „наилучший“, или каса-

тельный конус, используется для отображения части каждой параллели на дугу окружности, перпендикулярную к указанному выше прямолинейному отрезку (на плоскости дуги окружности не будут концентрическими).

Азимутальные отображения получаются при проектировании сферы S на касательную плоскость из точки, лежащей на диаметре, перпендикулярном к плоскости (см. рис. 13.2). Примером азимутальных отображений является *гномическая* проекция, у которой центр

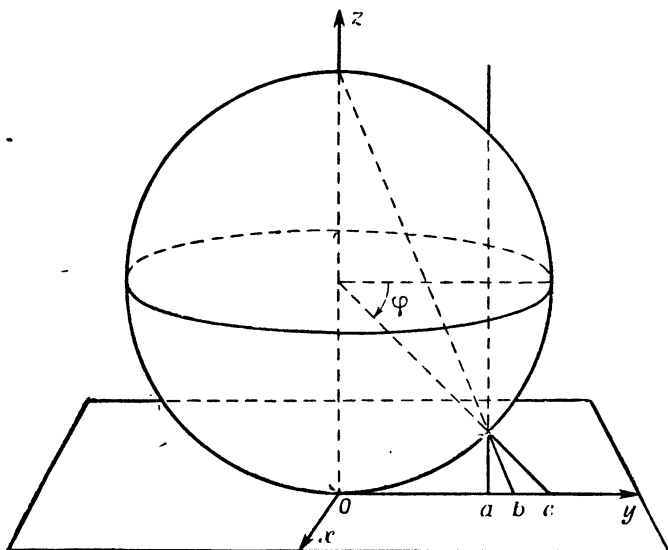


Рис. 13.2. Азимутальные отображения единичной сферы для угла φ , равного $-\pi/4$.

Расстояния $a = \cos \varphi$, $b = 2 \operatorname{tg} \varphi/2$, $c = \operatorname{tg} \varphi$ соответствуют ортографической, стереографической и гномической проекциям.

проекция совпадает с центром сферы. Построенные таким образом карты обладают тем ценным свойством, что всем большим кругам на сфере соответствуют прямые линии. Другим примером могут служить *ортографические* проекции, у которых проектирование производится из бесконечно удаленной точки. Такие карты удобны тем, что сохраняют внешнее сходство со сферой. К азимутальным отображениям относится также *стереографическая* проекция, у которой центром проекции является точка сферы, противоположная касательной плоскости. Это последнее отображение является, кроме того, и конформным. Все три азимутальные проекции изображены на рис. 13.2.

При стереографической проекции бесконечно удаленной точке плоскости соответствует единственная точка сферы. В силу этого свойства и в силу свойства конформности в теории функций ком-

плексного переменного пользуются стереографической проекцией для изображения комплексной плоскости на сфере, которая тогда называется *сферой Римана*.

Стереографическую проекцию сферы S можно рассматривать как часть конформного отображения T всего конечного трехмерного пространства самого на себя (стереографическая проекция может быть продолжена до такого отображения). Исключение составляет лишь центр проекции P , который можно считать соответствующим единственной бесконечно удаленной точке того же пространства. Это отображение T является *инверсией* относительно сферы, центр которой лежит в точке P и радиус равен диаметру сферы S .

Мы говорим, что две точки $P(x, y, z)$ и $P_1(x_1, y_1, z_1)$ *соответствуют друг другу* или *находятся в инверсии* относительно сферы S_0 радиуса r_0 и с центром в $P_0(x_0, y_0, z_0)$, если они лежат на одном и том же луче (полупрямой), исходящем из точки P_0 , и если расстояния $P_0P = r$ и $P_0P_1 = r_1$ удовлетворяют соотношению

$$rr_1 = r_0^2.$$

Таким образом, при инверсии точки сферы S_0 соответствуют самим себе, а точки, лежащие внутри сферы, — точкам, лежащим вне ее. В силу подобия треугольников точки P и P_1 находятся в инверсии относительно сферы S_0 тогда и только тогда, когда выполнены равенства

$$\frac{x - x_0}{x_1 - x_0} = \frac{y - y_0}{y_1 - y_0} = \frac{z - z_0}{z_1 - z_0} = \frac{r}{r_1} = \frac{r_0^2}{r_1^2},$$

в которых нулю в знаменателе одной из трех первых дробей соответствует нуль в числителе той же дроби. Из этих соотношений легко получить выражения для x, y, z через x_1, y_1, z_1 и известные величины x_0, y_0, z_0, r_0 :

$$\begin{aligned} x &= x_0 + \frac{r_0^2}{r_1^2}(x_1 - x_0), & y &= y_0 + \frac{r_0^2}{r_1^2}(y_1 - y_0), \\ z &= z_0 + \frac{r_0^2}{r_1^2}(z_1 - z_0), \end{aligned} \quad (13.10)$$

где

$$r_1^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2.$$

Из уравнений (13.10) можно получить соотношение

$$dx^2 + dy^2 + dz^2 = \frac{r_0^4}{r_1^4}(dx_1^2 + dy_1^2 + dz_1^2),$$

из которого [в силу равенств (13.8) и (13.9)] следует конформность инверсии относительно сферы.

Уравнение вида

$$A[(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2] + B(x-x_0) + C(y-y_0) + D(z-z_0) + E = 0, \quad (13.11)$$

в котором A, B, C, D не обращаются в нуль одновременно, изображает плоскость ($A=0$) или сферу ($A \neq 0$), хотя, конечно, при некоторых значениях коэффициентов мы получаем несуществующие, т. е. мнимые сферы. В формуле (13.11) взяты $(x-x_0), (y-y_0), (z-z_0)$ вместо x, y, z только для того, чтобы упростить подстановку в нее выражений (13.10). Сделав эту подстановку, мы получим уравнение

$$E[(x_1-x_0)^2 + (y_1-y_0)^2 + (z_1-z_0)^2] + Br_0^2(x_1-x_0) + Cr_0^2(y_1-y_0) + Dr_0^2(z_1-z_0) + Ar_0^4 = 0, \quad (13.12)$$

из которого следует, что сфера или плоскость (x, y, z) [уравнение (13.11)] отображается на сферу или плоскость (x_1, y_1, z_1) [уравнение (13.12)]. Действительно, уравнение (13.12) изображает плоскость ($E=0$) только в том случае, если точка P_0 лежит на многообразии (13.11) и, наоборот, если это многообразие не проходит через точку P_0 , то уравнение (13.12) изображает сферу ($E \neq 0$). Рассмотрим пример, показывающий, что стереографическая проекция является частным случаем инверсии относительно сферы S_0 . Пусть $(x_0, y_0, z_0) = (0, 0, 2)$ и $r_0 = 2$ (см. рис. 13.3). Точке $P_1(x_1, y_1, z_1)$, лежащей на сфере

$$S_1: x_1^2 + y_1^2 + (z_1 - 1)^2 = 1,$$

соответствует при инверсии точка P , координаты (x, y, z) которой даются формулами

$$\begin{aligned} x &= \frac{4x_1}{x_1^2 + y_1^2 + (z_1 - 2)^2} = \frac{2x_1}{2 - z_1}, \\ y &= \frac{4y_1}{x_1^2 + y_1^2 + (z_1 - 2)^2} = \frac{2y_1}{2 - z_1}, \\ z &= 2 + \frac{4(z_1 - 2)}{x_1^2 + y_1^2 + (z_1 - 2)^2} = 2 - 2 = 0. \end{aligned}$$

Таким образом, точка P лежит на плоскости $z=0$.

Аналогично стереографическую проекцию единичной сферы из ее *северного полюса* $(0, 0, 1)$ на экваториальную плоскость можно получить, рассмотрев инверсию относительно сферы, центр которой лежит в этом полюсе, а радиус равен $\sqrt{2}$.

Кроме инверсии относительно сферы, конформными отображениями трехмерных евклидовых областей друг на друга являются *только*

обычные отражения относительно плоскостей (которые являются предельным случаем инверсии относительно сферы), жесткие повороты, преобразования подобия и комбинации всех таких преобразований. Поэтому методы конформных преобразований применяются в основном либо в плоских задачах, либо в задачах, которые сводятся к таковым (см. §§ 7.6—7.9).

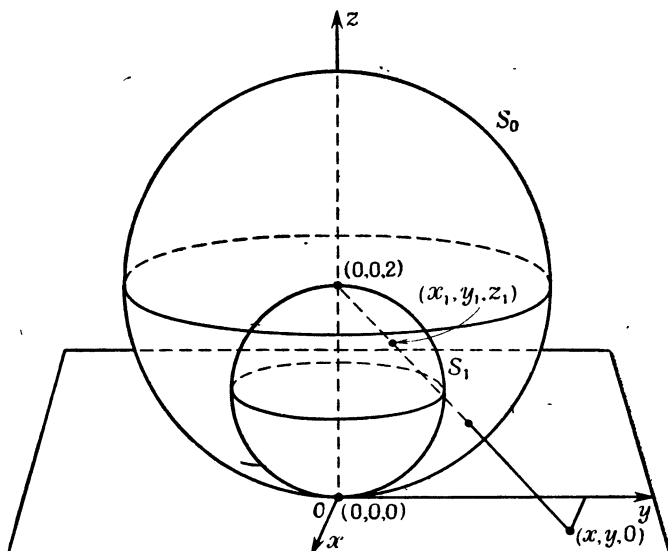


Рис. 13.3. Инверсия относительно сферы S_0 с центром в точке $(0, 0, 2)$ и радиусом, равным 2, приводит в соответствие точке (x_1, y_1, z_1) на сфере $S_1: x^2 + y^2 + (z-1)^2 = 1$ точку $(x, y, 0)$ на плоскости xy .

Это преобразование является стереографической проекцией сферы S_1 на плоскость.

§ 13.3. Конформные отображения и теория функций комплексного переменного. Для двух вещественных функций двух вещественных переменных

$$x = x(u, v), \quad y = y(u, v)$$

условия конформности (13.8) могут быть записаны в следующей форме:

$$E - G - 2jF = \left(\frac{\partial x}{\partial u} - j \frac{\partial x}{\partial v} \right)^2 + \left(\frac{\partial y}{\partial u} - j \frac{\partial y}{\partial v} \right)^2 = 0, \\ E > 0, \quad j^2 = -1,$$

откуда

$$\frac{\partial x}{\partial u} - j \frac{\partial x}{\partial v} = \pm j \left(\frac{\partial y}{\partial u} - j \frac{\partial y}{\partial v} \right),$$

или

$$\frac{\partial x}{\partial u} = \pm \frac{\partial y}{\partial v}, \quad \frac{\partial x}{\partial v} = \mp \frac{\partial y}{\partial u}. \quad (13.13)$$

Если в равенствах (13.13) взять верхние знаки, то мы получим хорошо известные уравнения Коши — Римана в частных производных первого порядка, характеризующие аналитические функции $x + jy$ комплексного аргумента $u + jv$ (см. ниже) или конформные отображения I рода областей в плоскости uv на области в плоскости xu . При нижних знаках уравнения (13.13) характеризуют аналитические функции $x + jy$ комплексного аргумента $u - jv$ или конформные отображения II рода (изменяющие направления вращения). Условие $E > 0$ выполнено во всех точках, исключая те, в которых производная комплексной функции равна нулю.

Наоборот, так как наши рассуждения можно провести в обратном порядке, то любая из пар (верхние или нижние знаки) уравнений (13.13) означает, что $x(u, v)$ и $y(u, v)$ удовлетворяют условиям $E = G$ и $F = 0$.

Тот факт, что аналитические функции комплексного переменного осуществляют конформные отображения, можно доказать следующим образом. Мы говорим, что комплексная функция $x + jy$ комплексного аргумента $u + jv$,

$$x + jy = f(w) = f(u + jv) = x(u, v) + jy(u, v), \quad (13.14)$$

определенная в открытой области D , аналитична в этой области, если она имеет производную

$$f'(w_0) = \lim_{w \rightarrow w_0} \frac{f(w) - f(w_0)}{w - w_0} \quad (13.15)$$

в каждой точке w_0 области D . Как хорошо известно, функция $f(w)$ допускает в этом случае разложение в ряд Тейлора, сходящийся в некоторой круговой окрестности точки w_0 :

$$f(w) = f(w_0) + f'(w_0)(w - w_0) + \dots + \frac{f^{(n)}(w_0)}{n!}(w - w_0)^n + \dots \quad (13.16)$$

Мы исключим из рассмотрений неинтересное для нас отображение области D на одну точку. Следовательно, производная $f'(w)$ может обращаться в нуль только в изолированных точках w_0 области D , и в каждой такой точке w_0 существует отличная от нуля производная $f^{(k)}(w)$ наименьшего порядка k .

Рассмотрим сначала обычный случай $f'(w_0) \neq 0$ и пусть

$$\begin{aligned} w - w_0 &= r(\cos \theta + j \sin \theta) = re^{j\theta}, \\ f(w) - f(w_0) &= \rho(\cos \varphi + j \sin \varphi) = \rho e^{j\varphi}, \\ f'(w_0) &= \alpha_0(\cos \beta_0 + j \sin \beta_0) = \alpha_0 e^{j\beta_0} \neq 0. \end{aligned}$$

В силу равенства (13.15)

$$\frac{f(w) - f(w_0)}{w - w_0} = \frac{re^{j\varphi}}{re^{j\theta}} = \frac{\rho}{r} e^{j(\varphi - \theta)} \xrightarrow{w \rightarrow w_0} f'(w_0) = \alpha_0 e^{j\beta},$$

откуда

$$\lim_{w \rightarrow w_0} \frac{\rho}{r} = \alpha_0 = |f'(w_0)| > 0$$

и

$$\lim_{w \rightarrow w_0} (\varphi - \theta) = \beta_0 = \arg f'(w_0) \pmod{2\pi}.$$

Таким образом, функция $f(w)$ осуществляет в этом случае „преобразование, подобное в малом“ (как мы его назвали в § 13.2), коэффициент линейного искажения которого равен $|f'(w)| > 0$. Если $w \rightarrow w_0$ вдоль кривой C_1 , угол наклона которой в точке w_0 равен θ_1 , то $f(w)$ стремится к $f(w_0)$ вдоль кривой γ_1 , угол наклона которой в точке $f(w_0)$ равен

$$\varphi_1 = \theta_1 + \beta_0 = \theta_1 + \arg f'(w_0) \pmod{2\pi},$$

так что $\arg f'(w_0)$ равен углу поворота при отображении. Если, наконец, мы рассмотрим другую кривую C_2 , угол наклона которой в точке w_0 равен θ_2 , так что угол между двумя кривыми равен $\theta_2 - \theta_1$, то угол между образами γ_1 и γ_2 этих кривых равен

$$\varphi_2 - \varphi_1 = (\theta_2 + \beta_0) - (\theta_1 + \beta_0) = \theta_2 - \theta_1. \quad (13.17)$$

Таким образом, углы сохраняются как по величине, так и по направлению.

Пусть теперь $f'(w_0) = 0$ и $f^{(k)}(w_0) \neq 0$, где $k \geq 2$ есть низший порядок производной, не равной нулю в точке w_0 . Положим

$$\frac{f^{(k)}(w_0)}{k!} = \alpha_0 (\cos \beta_0 + j \sin \beta_0) = \alpha_0 e^{j\beta_0} \neq 0.$$

В силу уравнения (13.16)

$$f(w) - f(w_0) = \frac{f^{(k)}(w_0)}{k!} (w - w_0)^k + \dots,$$

или

$$\rho e^{j\varphi} = \alpha_0 e^{j\beta_0} (r e^{j\theta})^k + \dots = \alpha_0 r^k e^{j(\beta_0 + k\theta)} + \dots \quad (13.18)$$

В данном случае неравенство

$$\lim_{w \rightarrow w_0} \frac{\rho}{r} > 0$$

уже не имеет места. Действительно, отношение радиусов-векторов соответствующих друг другу точек характеризуется в данном случае соотношением

$$\lim_{w \rightarrow w_0} \frac{\rho}{r^k} = \alpha_0 = \left| \frac{f^{(k)}(w_0)}{k!} \right| > 0.$$

Точно так же для углов мы имеем в этом случае соотношение

$$\lim_{w \rightarrow w_0} (\varphi - k\theta) = \beta_0 = \arg f^{(k)}(w_0) \pmod{2\pi}.$$

Таким образом, при отображении углы увеличиваются в k раз и поворачиваются дополнительно на угол $\arg f^{(k)}(w_0)$. Вместо равенства (13.17) мы имеем равенство

$$\varphi_2 - \varphi_1 = k(\theta_2 - \theta_1).$$

Оба вышеописанных случая $k=1$ и $k>1$ могут быть, конечно, описаны одновременно при $k \geq 1$.

Условие необращения в нуль коэффициента искажения (условие конформности) может быть записано при помощи якобиана

$$J\left(\frac{x, y}{u, v}\right) \equiv \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}.$$

Действительно, если $E=G$ и $F=0$, то из тождества

$$EG - F^2 \equiv \left\{ J\left(\frac{y, z}{u, v}\right) \right\}^2 + \left\{ J\left(\frac{z, x}{u, v}\right) \right\}^2 + \left\{ J\left(\frac{x, y}{u, v}\right) \right\}^2$$

следует, что неравенство $E=G>0$ имеет место в том и только в том случае, если три якобиана не обращаются в нуль одновременно. В то же время из равенства

$$f'(w) = \frac{\partial x}{\partial u} + j \frac{\partial y}{\partial u}$$

и уравнений (13.13) следует, что

$$|f'(w)|^2 \equiv J\left(\frac{x, y}{u, v}\right).$$

Поэтому условие $f'(w) \neq 0$ эквивалентно условию

$$J\left(\frac{x, y}{u, v}\right) \neq 0.$$

При любой паре знаков (как верхних, так и нижних) равенства (13.13) означают, что функции $x(u, v)$ и $y(u, v)$ являются *гармоническими*; иными словами, эти функции удовлетворяют уравнению Лапласа (13.1). Действительно, взяв равенства (13.13) с верхними знаками, продифференцировав первое из них по u , второе — по v и сложив между собой, мы получим уравнение (13.1) по отношению к функции $x(u, v)$.

Для трех и более действительных функций двух действительных переменных условие конформности (13.8) не влечет за собой гармоничность функций. Действительно, если поверхность S отображена конформно на плоскую область, то в силу теоремы Вейерштрасса

отображающие функции будут гармоничными только в том случае, когда поверхность S минимальная, т. е. когда средняя кривизна поверхности

$$K_m \equiv \frac{1}{\rho_1} + \frac{1}{\rho_2} = 0$$

тождественно равна нулю. Здесь ρ_1 и ρ_2 — главные радиусы кривизны поверхности. Естественным обобщением численного решения задачи о конформных отображениях в плоскости является, по-видимому, численное решение задачи о конформном отображении минимальной поверхности, заданной своей граничной кривой. Эта задача эквивалентна задаче о численном решении проблемы Плато для данной граничной кривой в пространстве. Последнее, очевидно, сложнее, чем численное решение [4] проблемы Плато в ее непараметрическом (и неконформном) виде, при котором ищется одна функция $z = z(x, y)$, принимающая заданные значения в каждой граничной точке и удовлетворяющая дифференциальному уравнению в частных производных

$$\left[1 + \left(\frac{\partial z}{\partial y}\right)^2\right] \frac{\partial^2 z}{\partial x^2} - 2 \frac{\partial z}{\partial x} \frac{\partial z}{\partial y} \frac{\partial^2 z}{\partial x \partial y} + \left[1 + \left(\frac{\partial z}{\partial x}\right)^2\right] \frac{\partial^2 z}{\partial y^2} = 0$$

в каждой внутренней точке.

В результате конформного преобразования (13.14) функция $\varphi(x, y)$ переходит в функцию от (u, v)

$$\varphi(x, y) = \varphi[x(u, v), y(u, v)] = \Phi(u, v).$$

Выполнив преобразования, мы получим равенство

$$\frac{\partial^2 \Phi}{\partial u^2} + \frac{\partial^2 \Phi}{\partial v^2} = \left(\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} \right) |f'(w)|^2,$$

из которого следует, что свойство гармоничности функции является инвариантным относительно конформных преобразований. То же самое имеет место и в отношении субгармонических функций, характеризующихся дифференциальным неравенством

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} \geq 0.$$

При более общем определении субгармоничности, использующем среднее значение функции,

$$\varphi(x_0, y_0) \leq \frac{1}{2\pi} \int_0^{2\pi} \varphi(x_0 + r \cos \theta, y_0 + r \sin \theta) d\theta,$$

также сохраняется свойство инвариантности.

Субгармонические функции являются мощным аппаратом в теории потенциала, теории функций комплексного переменного и дифференциальной геометрии, но мы не станем останавливаться на этом.

§ 13.4. Уравнение Лапласа в физике. Лапласиан $\nabla^2 x$ и уравнение Лапласа $\nabla^2 x = 0$ весьма часто встречаются и в математическом анализе (см. §§ 4.13 и 13.3) и в математической физике (см. §§ 2.5, 11.7, 11.8, 15.3 и 17.3 и гл. 5, 6 и 7). Соответствующие примеры приводятся в цитированных только что параграфах, и сейчас мы не будем останавливаться на иллюстрации этого положения. Рассмотрим только несколько случаев, при изучении которых используются гармонические функции двух действительных переменных.

Иногда гармонические функции или, что то же самое, решения уравнения Лапласа, встречающиеся в математической физике, являются функциями только двух переменных. Это имеет место, например, в таких задачах теории упругости, как задачи о кручении и изгибе. Иногда, однако, как, например, в аэродинамике, гармонические функции зависят от трех пространственных переменных, но и в этом случае часто оказывается возможным ограничиться изучением двумерных поперечных сечений. Это имеет место, например, при изучении аэродинамических свойств профиля крыла самолета. Серия прикладной математики Национального бюро стандартов содержит специальный том [1], посвященный построению и приложениям конформных отображений.

Для иллюстрации мы рассмотрим два простых классических примера применения гармонических функций в математической физике.

Пусть в фиксированной точке (x_0, y_0, z_0) помещено m единиц некоторой субстанции. Предположим, что векторное поле, индуцируемое этой субстанцией в переменной точке (x, y, z) , обратно пропорционально квадрату расстояния, и обозначим через $\mathbf{V}(x, y, z)$ силу, с которой это поле действует на одну единицу субстанции

$$|\mathbf{V}| = \frac{km}{r^2}, \quad r^2 = (x_0 - x)^2 + (y_0 - y)^2 + (z_0 - z)^2.$$

Таким образом, $\mathbf{V}(x, y, z)$ образует векторное поле с компонентами

$$V_x = \frac{km}{r^3} (x_0 - x), \quad V_y = \frac{km}{r^3} (y_0 - y), \quad V_z = \frac{km}{r^3} (z_0 - z).$$

Например, поле, образуемое вектором $\mathbf{V}(x, y, z)$, может быть гравитационным, отрицательным электростатическим или магнитным полем в пространстве. Легко проверить, что как вихрь, так и дивергенция вектора $\mathbf{V}$ тождественно обращаются в нуль:

$$\begin{aligned} \operatorname{rot} \mathbf{V} \equiv \nabla \times \mathbf{V} &\equiv \mathbf{i} \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z} \right) + \mathbf{j} \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x} \right) + \\ &+ \mathbf{k} \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y} \right) = 0 \end{aligned} \quad (13.19)$$

и

$$\operatorname{div} \mathbf{V} \equiv \nabla \cdot \mathbf{V} \equiv \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z} = 0. \quad (13.20)$$

Из равенства (13.19) следует, что существует исчезающая в бесконечно удаленной точке функция, называемая *скалярным потенциалом*,

$$P(x, y, z) = \int \mathbf{V} \cdot d\mathbf{r} = \int_{\infty}^{(x, y, z)} V_x dx + V_y dy + V_z dz = \frac{km}{r}, \quad (13.21)$$

частные производные которой равны соответствующим компонентам вектора $\mathbf{V}$:

$$\begin{aligned} \mathbf{V} = \nabla P &= \mathbf{i} \frac{\partial P}{\partial x} + \mathbf{j} \frac{\partial P}{\partial y} + \mathbf{k} \frac{\partial P}{\partial z} = \\ &= \mathbf{i} V_x + \mathbf{j} V_y + \mathbf{k} V_z. \end{aligned} \quad (13.22)$$

При более общем распределении субстанции скалярный потенциал имеет вид

$$P(x, y, z) = k \sum \frac{m_t}{r_t}$$

или

$$P(x, y, z) = k \iiint \frac{m}{r} dx_0 dy_0 dz_0 \quad (13.23)$$

и т. д. в зависимости от характера распределения. Нам остается заметить, что в силу уравнения (13.20) $P(x, y, z)$ является гармонической функцией ее трех переменных вне области, занятой субстанцией. [То же самое можно заключить, дифференцируя равенство (13.21) или одно из равенств (13.23) и используя равенство (13.22).]

Предположим теперь, что рассматриваемая субстанция равномерно распределена в направлении оси z от $-\infty$ до $+\infty$, например, вдоль самой оси z , и пусть плотность распределения равна m единицам на единицу длины. Пусть

$$\begin{aligned} R^2 &= x^2 + y^2, & r^2 &= x^2 + y^2 + z^2, & r &= R \sec \theta, \\ & & z &= R \operatorname{tg} \theta \end{aligned}$$

(см. рис. 13.4). Тогда вектор $\mathbf{V}$ не зависит от z , направлен к оси z и его величина определяется формулой

$$|\mathbf{V}| = \int_{-\infty}^{\infty} \frac{km \cos \theta}{r^2} dz = \frac{2km}{R} \int_0^{\pi/2} \cos \theta d\theta = \frac{2km}{R}.$$

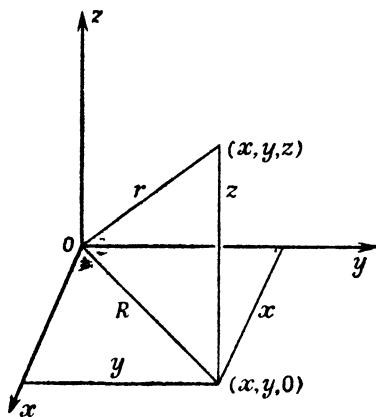


Рис. 13.4. Логарифмический потенциал.

Масса, индуцирующая гравитационное векторное поле, равномерно распределена вдоль оси z , и, следовательно, поле не зависит от z . Частица, на которую действует поле, помещена в плоскость xy .

Таким образом, вектор $\mathbf{V}$ можно рассматривать как образующий силовое поле в плоскости xu , в котором сила изменяется обратно пропорционально расстоянию. Компонентами вектора $\mathbf{V}$ являются

$$V_x = -\frac{2kmx}{R^2}, \quad V_y = -\frac{2kmy}{R^2}.$$

Как и выше, вихрь и дивергенция этого поля тождественно равны нулю, и поэтому его скалярный потенциал является гармонической функцией

$$P(x, y) = \int \mathbf{V} \cdot d\mathbf{r} = \int_{(0,1)}^{(x,y)} V_x dx + V_y dy = km \log \frac{1}{r},$$

называемой в этом случае *логарифмическим потенциалом*; частные производные этого потенциала равны соответствующим компонентам вектора $\mathbf{V}$:

$$\frac{\partial P}{\partial x} = V_x, \quad \frac{\partial P}{\partial y} = V_y.$$

В рассмотренном простейшем случае логарифмический потенциал обращается в нуль на единичной окружности $x^2 + y^2 = 1$. При более общем распределении того же типа мы положим [как и в равенствах (13.23)]

$$P(x, y) = k \sum m_t \log \frac{1}{r_t}$$

или

$$P(x, y) = k \iint m \log \frac{1}{r} dx_0 dy_0.$$

При этом мы отказываемся от требования обращения в нуль потенциала в некоторой фиксированной точке.

Конечно, рассмотренная двумерная идеализация не встречается на практике, но существует очень много случаев (длинный прямой проводник и т. д.), когда описанный метод дает прекрасное приближенное описание явления с весьма малой ошибкой.

В качестве второго примера рассмотрим плоское течение идеальной жидкости. Пусть $V_x(x, y)$ и $V_y(x, y)$ будут компонентами вектора скорости $\mathbf{V}$ течения соответственно вдоль осей x и y . Обозначим через

$$V_x(x, y) - jV_y(x, y), \quad j^2 = -1, \quad (13.24)$$

комплексную скорость течения. [Мы берем выражение (13.24) вместо *комплексно сопряженного* $V_x + jV_y$ по причинам, которые скоро станут понятными. Очевидно, оба выражения связаны инверсией относительно вещественной оси.] Плоскость комплексной скорости называется *плоскостью годографа*. Жидкость называется *несжимае-*

мой тогда и только тогда, когда дивергенция вектора скорости тождественно равна нулю [ср. равенство (13.20)]:

$$\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} = 0. \quad (13.25)$$

Течение называется *безвихревым* в том и только в том случае, если вихрь его вектора скорости тождественно равен нулю [ср. равенство (13.19)]:

$$\frac{\partial V_x}{\partial y} - \frac{\partial V_y}{\partial x} = 0. \quad (13.26)$$

Но равенства (13.25) и (13.26) являются в точности условиями Коши — Римана для функции (13.24). Таким образом, в рассматриваемом случае компоненты скорости $V_x(x, y)$ и $V_y(x, y)$ оказываются гармоническими функциями, а сама комплексная скорость — аналитической функцией комплексного переменного $x + jy$. Отсюда следует, что комплексная скорость допускает аналитическую *примитивную* функцию, называемую *комплексным потенциалом скоростей*

$$\int_{x_0 + jy_0}^{x + jy} (V_x - jV_y)(ds + jdt) = P(x, y) + jQ(x, y),$$

причем

$$\frac{\partial P}{\partial x} = \frac{\partial Q}{\partial y} = V_x(x, y), \quad \frac{\partial P}{\partial y} = -\frac{\partial Q}{\partial x} = V_y(x, y). \quad (13.27)$$

Из равенств (13.27) следует, что вектор скорости $\mathbf{V}$ является градиентом гармонической функции $P(x, y)$:

$$\frac{\partial P}{\partial x} = V_x(x, y), \quad \frac{\partial P}{\partial y} = V_y(x, y).$$

Таким образом, линии уровня $P = \text{const}$ являются *экипотенциальными линиями*, а ортогональные линии $Q = \text{const}$ — *линиями тока*. Гармоническая функция $Q(x, y)$ называется *функцией тока*. Из всего сказанного следует, что для изучения течения весьма полезно рассматривать поверхность

$$P + jQ = \text{потенциал скоростей}$$

в действительном пространстве xuP . Однако практически удобнее построить конформное отображение

$$P + jQ = F(x + jy),$$

переводящее экипотенциальные линии и линии тока в прямые, перпендикулярные в плоскости $P + jQ$ к действительной и мнимой осям соответственно.

§ 13.5. Теоремы существования. При построении конформных отображений необходимо: а) владеть основными фактами, относящимися к теоремам существования, которые позволяют понять, какие

типы отображений возможны; б) знать основные геометрические свойства конформных отображений, которые позволяют предвидеть результаты и избежать грубых ошибок; в) освоить отображения, осуществляемые элементарными функциями, чтобы иметь возможность находить достаточно точные первые приближения к искомым отображениям.

Сделаем краткий обзор всех трех аспектов этой теории по порядку.

В § 6.3 мы видели, что с помощью функции Грина $G(P, Q)$ можно решить первую граничную задачу (см. § 6.2), или задачу Дирихле, теории потенциала в трехмерном пространстве. Наоборот, решение задачи Дирихле при специальных граничных значениях $\frac{1}{r}$ непосредственно приводит к построению функции Грина. Таким образом, решение задачи Дирихле и построение функции Грина являются эквивалентными задачами. То же самое имеет место и в двумерном случае, с той разницей, что полюс решения в точке Q заменяется логарифмической особой точкой (ср. § 13.4). Функция Грина $G(P, Q)$ для плоской области D и логарифмической особенности Q , лежащей внутри D , однозначно определяется следующими свойствами:

а) $G(P, Q) = \log 1/r + h(P, Q)$, где r есть расстояние PQ , а $h(P, Q)$ как функция точки P является решением уравнения Лапласа.

б) $G(P, Q)$ обращается в нуль на границе C области D .

Если D односвязна, то построение функции Грина эквивалентно не только решению задачи Дирихле, но также и решению задачи о конформном отображении области D на внутренность единичного круга $|w|=1$, причем точке Q соответствует центр круга. (Функция Грина аналогичным образом связывается с задачей о конформном отображении многосвязных областей посредством понятия универсальной поверхности наложения, но мы на этом здесь не останавливаемся.) Действительно, пусть Q есть точка $\zeta_0 = x_0 + jy_0$. Функция, осуществляющая указанное отображение, будет иметь вид

$$w = f(\zeta) = (\zeta - \zeta_0) Q(\zeta), \quad |f(\zeta)| = 1 \text{ на } C,$$

причем функция $Q(\zeta)$ не обращается в нуль в области D . Поэтому функция

$$-\log |f(\zeta)| = \log \frac{1}{r} - \log |Q(\zeta)| = \log \frac{1}{r} + h(P, Q) = G(P, Q)$$

удовлетворяет написанным выше условиям (а) и (б). Наоборот, пусть задана функция Грина G и пусть H — сопряженная с ней функция, так что $G + jH$ является аналитической функцией комплексного аргумента ζ . Можно показать [6], что тогда функция

$$w = f(\zeta) = e^{-G - jH}$$

осуществляет искомое отображение. Сопряженная функция H опре-

деляется однозначно с точностью до аддитивной постоянной, которая в данном случае соответствует повороту единичного круга.

Основным результатом теории конформных отображений является *теорема Римана*:

Если D есть односвязная область, граница которой содержит более одной точки, то существует аналитическая функция $w = f(\zeta)$, осуществляющая взаимно однозначное и конформное отображение области D на единичный круг $|w| < 1$.

Из изложенного выше следует, что для области D указанного вида существуют функция Грина и решение задачи Дирихле при произвольных непрерывных значениях на границе.

Функция $w = f(\zeta)$, осуществляющая отображение Римана, может быть выбрана таким образом, чтобы произвольной внутренней точке ζ_0 области D соответствовала произвольная точка w круга $|w| < 1$ и, кроме того, произвольному направлению в точке ζ_0 соответствовало произвольное направление в точке w_0 . Этими условиями конформное отображение определяется однозначно. Обычно эти условия берут в виде

$$f(\zeta_0) = 0, \quad f'(\zeta_0) > 0. \quad (13.28)$$

Для областей, ограниченных простой жордановой кривой, теорема Римана может быть дополнена следующей *теоремой Осгуда — Каратеодори*:

Если область D ограничена жордановой кривой, то отображающая функция $w = f(\zeta)$ может быть непрерывно продолжена вплоть до границы области D . При этом она будет осуществлять взаимно однозначное и непрерывное отображение границы области на границу единичного круга.

Таким образом, конформное отображение области D , ограниченной жордановой кривой, на круг $|w| < 1$ автоматически приводит к топологическому отображению границы D на окружность $|w| = 1$. Этот результат остается в силе и для случая, когда граница области D содержит достижимую изнутри области D жорданову дугу C — в указанном случае устанавливается взаимно однозначное и непрерывное отображение дуги C на часть окружности $|w| = 1$. Рассмотрим, например, область, изображенную на рис. 13.5, граница которой содержит бесконечное число разрезов одинаковой длины, выходящих из точки F и сходящихся к отрезку FB . При отображении отрезку границы AB , так же как и обоим краям разреза EF ,

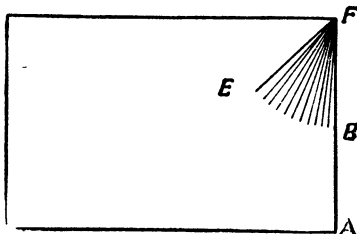


Рис. 13.5. Граничные элементы.

Бесконечное число разрезов, исходящих из точки F , сходятся к недостижимому отрезку границы BF . При отображении внутренней области на внутренность единичного круга отрезку границы AB , так же как обоим краям разреза EF , соответствуют дуги окружности, но недостижимому отрезку BF соответствует на окружности только одна точка.

соответствуют дуги окружности $|\omega| = 1$, но всему отрезку BF , точки которого недостижимы изнутри области, соответствует единственная точка на окружности.

Часть границы D , которой при отображении соответствует единственная точка на окружности, называется *простым концом*. Условия (13.28) единственности отображения часто заменяются условием соответствия трех точек A', B', C' на окружности $|\omega| = 1$ трем простым концам A, B, C на границе области D . Это так называемое *условие трех точек*. При отображениях I рода точки A', B', C' и A, B, C должны быть, очевидно, одинаково ориентированы, относительно своих областей.

Несколько сложнее решается вопрос об отображении многосвязных областей, т. е. областей с „отверстиями“. Например, кольцо $1 < |\zeta| < 2$ не может быть отображено на кольцо $1 < |\omega| < 3$. Чтобы такое отображение было возможно, отношения соответствующих радиусов должны быть равны. Но внутренность *любой* *двусвязной* области может быть отображена на *некоторое* кольцо $r_1 < |\omega| < r_2$, радиусы которого $r_1 < r_2$ подобраны соответствующим образом.

Для областей, более чем двусвязных, число соотношений, которые должны быть выполнены, увеличивается. При отображении многосвязных областей обычно задают только типы *канонических областей*, на которые осуществляется отображение, а указанные соотношения появляются автоматически.

Канонические типы областей при многосвязных отображениях возникают при обобщении и развитии тех экстремальных свойств, которые для случая односвязных областей приводят к отображениям на внутренность единичного круга. Достаточно вспомнить следующую теорему относительно отображения односвязных областей: пусть дана односвязная плоская область D с более чем одной граничной точкой, и пусть ζ_0 и ζ_1 — две ее различные внутренние точки; среди всех функций, аналитических в D и удовлетворяющих условиям

$$f(\zeta_0) = 0, \quad f'(\zeta_0) > 0, \quad |f(\zeta)| < 1,$$

существуют единственные функции, которые принимают значения

$$\sup_{f(\zeta)} [f'(\zeta_0)] \quad \text{или} \quad \sup_{f(\zeta)} [|f(\zeta_1)|]$$

или некоторые другие экстремальные значения; указанные функции осуществляют отображение области D на единичный круг. Аналогичные рассуждения некоторых экстремальных свойств приводят для случая многосвязных отображений к каноническим областям; примерами таких канонических областей служат плоские области, граница которых состоит из:

- параллельных надразов,
- круговых концентрических надразов,

радиальных надрезов, окружности и круговых концентрических надрезов, кольца и круговых концентрических надрезов и т. д.

Например, для произвольной многосвязной области D и двух ее внутренних точек ζ_0 и ζ_1 можно рассмотреть класс функций, имеющих вид

$$w = f(\zeta) = \frac{1}{\zeta - \zeta_1} + a_0 + a_1(\zeta - \zeta_1) + \dots$$

и удовлетворяющих условиям

$$f(\zeta_0) = 0, \quad f(\zeta_1) = \infty.$$

Среди этого класса функций существуют единственные функции, принимающие экстремальные значения

$$\sup_{f(\zeta)} [|f'(\zeta_0)|] \quad \text{или} \quad \inf_{f(\zeta)} [|f'(\zeta_0)|].$$

Эти функции осуществляют отображение области D на область, которая состоит из всей плоскости с некоторыми надрезами; в первом случае это надрезы по окружностям, центры которых лежат в начале координат, а во втором — надрезы по радиусам, исходящим из начала координат.

§ 13.6. Геометрические свойства конформных отображений.

Помимо сохранения систем ортогональных траекторий и вообще углов, конформные отображения обладают еще целым рядом геометрических свойств, знание которых помогает лучше уяснить характер этих отображений.

Как это следует из § 13.5, порядок связности области является *инвариантом* относительно конформных отображений. Это следует уже из того, что конформные отображения являются топологическими и, следовательно, все топологические свойства должны быть инвариантными относительно конформных отображений. Однако, хотя вся конечная плоскость может быть отображена топологически на внутренность единичного круга, из теоремы *Лиувилля* следует, что такое отображение не может быть *конформным*.

Наиболее известное и наиболее полезное геометрическое свойство конформных отображений выражается *леммой Шварца*:

Пусть функция $\zeta = g(w)$ осуществляет конформное отображение первого рода единичного круга $|w| < 1$ на область, лежащую внутри единичного круга, и пусть при этом начало координат переходит в начало координат, т. е.

$$|g(w)| \leq 1 \quad \text{и} \quad g(0) = 0.$$

Тогда любому кругу $|w| \leq r$ соответствует область, лежащая внутри круга $|\zeta| < r$, и коэффициент линейного искажения в центре

круга меньше или равен единице:

$$|g(w)| \leq |w| \quad \text{и} \quad |g'(0)| \leq 1.$$

Знаки равенства в двух последних соотношениях имеют место тогда и только тогда, когда отображение является простым поворотом единичного круга на угол α относительно начала координат

$$g(w) = e^{j\alpha} w.$$

В самом общем виде при отображении круга $|w| < 1$ на область D , лежащую в плоскости ζ , окружности $|w| = r_1$ соответствует, образно выражаясь, не более сложная кривая, чем кривая, соответствующая окружности $|w| = r_2$ ($0 < r_1 < r_2 \leq 1$). Если, например, область D — *выпуклая* (т. е. вместе с любыми двумя точками ζ_1 и ζ_2 содержит весь отрезок $\zeta_1\zeta_2$), то и любому кругу $|w| < r$, $r < 1$, соответствует выпуклая область. Точно так же, если область D — *звездная* относительно точки $\zeta_0 = g(0)$ (т. е. вместе с любой точкой ζ_1 содержит весь отрезок $\zeta_0\zeta_1$), то и любому кругу $|w| < r$, $r < 1$, соответствует область, *звездная* по отношению к точке ζ_0 .

Длина $l(r)$ образа дуги $|w| = r$ является выпуклой возрастающей функцией r . Более того, даже отношение $l(r)/2\pi r$ является выпуклой и неубывающей функцией. Точно так же, если $a(r)$ обозначает площадь образа круга $|w| < r$, то функция $a(r)/\pi r^2$ является выпуклой и не убывающей. В частности,

$$l(r) \geq 2\pi r |g'(0)| \quad \text{и} \quad a(r) \geq \pi r^2 |g'(0)|^2.$$

Знаки равенства имеют место только в том случае, если отображение $g(w)$ является преобразованием подобия

$$g(w) = \zeta_0 + |g'(0)| e^{j\alpha} w.$$

Если область D *однолистка*, т. е. не перекрывает саму себя (а именно с такими областями мы в основном сталкиваемся на практике), то круг с центром в точке $\zeta_0 = g(0)$ и радиусом $|g'(0)|/4$ принадлежит области D . (Поэтому $1/4$ называется иногда *постоянной Кёбе*.) Если область D *неоднолистка*, то она все же содержит круг, радиус которого не меньше $K|g'(0)|$, где постоянная $K > 0,5$ (*постоянная Блоха*), максимальное значение которой еще не известно.

Пусть $g(w)$ и $G(w)$ — две *однолистные* аналитические в круге $|w| < 1$ функции и $g(0) = G(0)$; тогда мы говорим, что функция $g(w)$ *подчинена* функции $G(w)$, если все значения, принимаемые $g(w)$ в круге $|w| < 1$, принимаются в том же круге и функцией $G(w)$; иными словами, если область, на которую $G(w)$ отображает единичный круг, покрывает аналогичную область для функции $g(w)$. В более общем случае, когда функции $g(w)$ и $G(w)$ *неоднолистные*, мы говорим, что $g(w)$ *подчинена* $G(w)$, если существует функция $\omega(w)$, удовлетворяющая условиям леммы Шварца, и такая, что

$$g(w) = G[\omega(w)].$$

Геометрически это означает, что *поверхность Римана* $\zeta = g(w)$ „лежит“ на поверхности Римана $\zeta = G(w)$, так что $g(0) = G(0)$ и каждой замкнутой кривой на первой поверхности соответствует такая же кривая на последней.

Из леммы Шварца мы можем получить следующий чрезвычайно полезный геометрический результат: если $g(w)$ подчинена $G(w)$ в круге $|w| < 1$, то $g(w)$ подчинена $G(w)$ в любом круге $|w| < r$.

В этом параграфе мы рассмотрели лишь немногие из большого числа изящных, поучительных и полезных геометрических свойств конформных отображений.

§ 13.7. Специальные функции. Для приложений очень большое значение имеет знакомство с теми отображениями, которые осуществляют некоторые специальные функции и их комбинации. Таким способом мы можем точно определить линии тока и т. д. (см. § 13.4) для многих плоских областей, встречающихся в физических приложениях, и для произвольных граничных значений. Большей частью, конечно, рассматриваемая область не имеется в существующих каталогах конформных отображений. В этом случае удачный выбор приближенной области дает хорошие начальные численные значения, значительно уменьшающие вычислительную или другую работу (см. § 15.9) при определении достаточно точного приближения.

Ввиду этого для Британского адмиралтейства в 1948—1952 гг. был составлен, а затем и издан сборник, или словарь, конформных отображений [7]. Этот сборник еще далек от желаемой полноты, но представляет собой первую и успешную попытку издания такого рода. С его помощью читатель может познакомиться с большим числом отображений, осуществляемых аналитическими и трансцендентными функциями. Поэтому мы здесь кратко познакомимся лишь с небольшим числом специальных отображений.

Чрезвычайно важную роль в теории конформных отображений первого рода играет *дробно-линейное* преобразование

$$\zeta = \frac{aw + b}{cw + d}, \quad ad - bc \neq 0. \quad (13.29)$$

Эти функции являются единственными функциями, осуществляющими взаимно однозначное отображение всей комплексной плоскости на себя. Они переводят окружности и прямые линии снова в окружности и прямые линии. Наоборот, если некоторая аналитическая функция осуществляет взаимно однозначное отображение внутренности круга или полуплоскости на внутренность круга или полуплоскость, то эта функция имеет вид (13.29).

Легко показать, что любое линейное преобразование (13.29) (не считая тождественного $\zeta = w$) имеет не более чем две *неподвижные точки*, являющиеся решениями уравнения

$$w = \frac{aw + b}{cw + d}.$$

Кроме того, линейное преобразование однозначно определяется заданием трех пар соответствующих друг другу точек $\zeta_1, \zeta_2, \zeta_3$ и w_1, w_2, w_3 . Это свойство объясняет условие „трех точек“ предыдущего параграфа. Соответствующее линейное преобразование можно получить, разрешая легко запоминающееся уравнение

$$\left(\frac{\zeta - \zeta_1}{\zeta - \zeta_2}\right) \left(\frac{\zeta_3 - \zeta_2}{\zeta_3 - \zeta_1}\right) = \left(\frac{w - w_1}{w - w_2}\right) \left(\frac{w_3 - w_2}{w_3 - w_1}\right) \quad (13.30)$$

(это уравнение выражает инвариантность *ангармонического отношения* четырех точек относительно линейных преобразований). В равенстве (13.30) должны быть использованы очевидные предельные соотношения, которые получаются, когда одна из точек ζ_t или w_t или две из них совпадают с бесконечно удаленной точкой. Так, например, линейное преобразование, для которого

$$(w_1, w_2, w_3) = (1, j, -1) \rightarrow (\zeta_1, \zeta_2, \zeta_3) = (0, 1, \infty),$$

дает отображение единичного круга плоскости w на верхнюю полуплоскость плоскости ζ . Подстановка в уравнение (13.30) дает

$$\left(\frac{\zeta - 0}{\zeta - 1}\right) \left(\frac{\infty - 1}{\infty - 0}\right) \sim \frac{\zeta}{\zeta - 1} = \left(\frac{w - 1}{w - j}\right) \left(\frac{-1 - j}{-1 - 1}\right),$$

откуда

$$\zeta = j \frac{1 - w}{1 + w}. \quad (13.31)$$

Любая аналитическая функция, взаимно однозначно отображающая круг $|w| = 1$ на себя и переводящая точку a (a — комплексное число, $|a| \neq 1$) в начало координат, имеет вид

$$\zeta = e^{j\alpha} \left(\frac{w - a}{1 - \bar{a}w} \right),$$

где $\bar{a}$ и a — комплексно сопряженные числа, а действительная постоянная α определяет угол поворота относительно начала координат. Этим объясняется условие (13.28), нормирующее отображение областей на единичный круг.

Преобразование

$$\xi = \zeta^2 \quad (13.32)$$

отображает первый квадрант плоскости ζ на верхнюю полуплоскость ξ , верхнюю полуплоскость ζ — на плоскость ξ , разрезанную вдоль положительной действительной полуоси, и т. д. (см. § 13.3). Комбинируя преобразования (13.31) и (13.32), мы найдем, что функция

$$\xi = - \left(\frac{1 - w}{1 + w} \right)^2$$

осуществляет отображение внутренности единичного круга в плоскости w на плоскость ξ , разрезанную вдоль положительной действ-

вительной полуоси. Поменяв ролями ξ и ζ , мы найдем функцию, отображающую внутренность единичного круга на первый квадрант. Для секторов, имеющих другой угол раствора, необходимо в уравнении (13.32) взять другие степени.

Преобразование Жуковского

$$\zeta = w + \frac{1}{w} \quad (13.33)$$

и его различные обобщения усиленно изучались ввиду их исключительной важности в теории обтекания профилей. Заметим, во-первых, что эта функция принимает одинаковые значения в точках w и $1/\bar{w}$ и, следовательно, отображает внутренность и внешность единичного круга на одну и ту же область.

Кроме того, при $|w|=1$, т. е. при

$$w = \cos \theta + j \sin \theta = e^{j\theta}$$

мы получаем

$$\zeta = w + \frac{1}{w} = 2 \cos \theta,$$

т. е. для $|w|=1$ переменная ζ принимает действительные значения. Наконец, действительным w соответствуют действительные ζ и при этом, если $w \rightarrow \pm \infty$, то и $\zeta \rightarrow \pm \infty$. Таким образом, функция

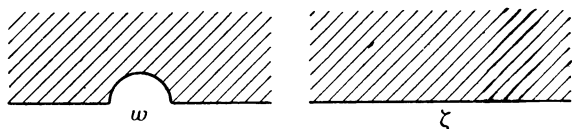


Рис. 13.6. Отображение верхней полуплоскости с исключенным из нее единичным полукругом на верхнюю полуплоскость ζ с помощью функции

$$\zeta = w + 1/w.$$

Жуковского отображает верхнюю полуплоскость w с исключенным из нее единичным полукругом на верхнюю полуплоскость ζ , как это изображено на рис. 13.6.

Производная этой функции имеет вид

$$\frac{d\zeta}{dw} = 1 - \frac{1}{w^2} = \frac{w^2 - 1}{w^2}$$

и, следовательно, отлична от нуля всюду, кроме точек $w = \pm 1$, где она имеет простой нуль. Отсюда следует, что рассмотренное отображение конформно всюду, за исключением указанных точек, в которых углы удваиваются. В частности, двум противоположным направлениям, исходящим из точки $w = -1$, соответствуют кривые, касающиеся друг друга в точке, соответствующей точке $w = -1$.

Если, таким образом, окружность γ проходит через точку $w = -1$ и содержит внутри себя точку $w = +1$, то при отображении внешность окружности γ переходит во внешность кривой Γ , называемой *профилем Жуковского*. В точке, соответствующей точке $w = -1$, эта кривая имеет острие. На рис. 13.7 изображен профиль Жуковского, соответствующий окружности γ :

$$|w - (0,12 + 0,08j)|^2 = (1,12)^2 + (0,08)^2.$$

Следует отметить, что во многих случаях *профили Жуковского* имеют поразительное сходство с профилями самолетных крыльев, применяемых на практике. Это сходство является чисто геометрическим. Если отвлечься от присущей конформным отображениям инвариантности

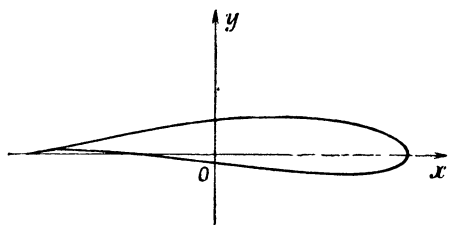


Рис. 13.7. Профиль Жуковского.

линии тока и т. д., то преобразование (13.33) и его обобщения сами по себе не имеют никакой связи с аэродинамическими принципами.

§ 13.8. Приложения. Конформные отображения находят применение во всех рассмотренных выше стационарных задачах математической физики ввиду того, что их изучение связано с уравнением Лапласа. Мы приведем здесь один простой пример обтекания профиля Жуковского однородным потоком воздуха.

Предположим, например, что направление потока параллельно оси x и рассмотрим течение, образуемое в плоскости xu . В этом случае

$$V_x = k, \quad V_y = 0$$

и, следовательно, потенциал комплексной скорости имеет вид

$$P(x, y) + jQ(x, y) = k(x + jy).$$

Функция тока имеет вид

$$Q(x, y) = ky,$$

а линиями тока $Q = \text{const}$ будут прямые $y = \text{const}$, параллельные оси x .

Введем в этот поток прямой круговой цилиндр $x^2 + y^2 = 1$ в качестве граничной поверхности и рассмотрим опять плоскость xu . Заметим прежде всего, что в силу симметрии ось x (за исключением отрезка $|x| \leq 1$) будет по-прежнему линией тока. Кроме того, линией тока станет окружность $x^2 + y^2 = 1$, как граничная поверхность. Наконец, мы предположим, что на достаточно большом расстоянии от цилиндра характер потока по существу не нарушается.

Требуется определить функцию скорости и линии тока изменившегося течения.

Чтобы установить связь между двумя разобранными случаями (при наличии цилиндра и без него), мы воспользуемся преобразованием

$$\zeta = x + jy = w + \frac{1}{w}, \quad w = u + jv,$$

которое, как мы знаем, отображает верхнюю полуплоскость с исключенным единичным полукругом на полуплоскость ζ (см. § 13.7).

Функция тока $Q = ky$ преобразуется в функцию тока

$$Q = ky = k \left(v - \frac{v}{u^2 + v^2} \right),$$

обладающую всеми необходимыми свойствами, указанными выше; линии тока для верхней полуплоскости отмечены на рис. 13.8.

Соответствующий потенциал скоростей имеет вид

$$P = kx = k \left(u + \frac{u}{u^2 + v^2} \right)$$

и, следовательно, компоненты скорости имеют вид

$$\begin{aligned} \frac{\partial P}{\partial u} &= V_u = k \left[1 + \frac{v^2 - u^2}{(u^2 + v^2)^2} \right], \\ \frac{\partial P}{\partial v} &= V_v = \frac{-2kuv}{(u^2 + v^2)^2}. \end{aligned}$$

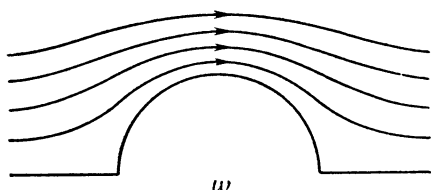


Рис. 13.8. Линии тока при обтекании прямого кругового цилиндра однородным потоком, перпендикулярным оси цилиндра.

Обе они обращаются в нуль в *критических точках* (1,0) и (−1,0).

Точно таким же образом мы можем изучить линии тока при обтекании профиля Жуковского, например при обтекании профиля, изображенного на рис. 13.7. Для этого достаточно отобразить его внешность на внешность соответствующего круга. Мы не будем останавливаться на этом подробнее, так как наша цель заключается лишь в пояснении метода; для более подробного ознакомления с этим вопросом читатель может обратиться к книге Милн-Томсона „Теоретическая аэродинамика“ [8].

§ 13.9. Построение конформных отображений. Эффективное определение функций, дающих требуемое конформное отображение, за исключением ряда известных частных случаев, представляет собой трудно выполнимую задачу. Чтобы убедиться в этом, достаточно рассмотреть формулу Шварца — Кристоффеля для конформного отображения полуплоскости на внутренность многоугольника. Определение входящих в нее постоянных представляет, вообще говоря (кроме некоторых простейших случаев), весьма значительные трудности.

Кроме того, затруднительным оказывается применение этой формулы для определения эквипотенциальных линий и т. д., даже если входящие в нее постоянные известны.

Тем не менее применение в технике метода конформных отображений непрерывно усиливается. В промышленных предприятиях, университетах и в ряде правительственных учреждений ведется большая работа по построению конформных отображений при помощи моделирующих устройств и современных электронных цифровых вычислительных машин. Такая работа [3] проводится, например, на вычислительных машинах в Кембридже, Принстоне, Вашингтоне и Лос-Анжелесе. Русские математики Канторович и Крылов посвящают конформным отображениям последние 300 страниц своей книги о приближенных методах высшего анализа [5]. Кроме того, имеется продолжение [12] книги, издаваемой Национальным бюро стандартов, к которой мы уже отсылали читателя. Авторы рассматривают не только новые методы, оценки ошибок и условия сходимости, но также и опыт использования вычислительной машины SEAC Национального бюро стандартов.

Мы закончим этот параграф замечаниями относительно некоторых методов построения отображающих функций.

При изучении сложных полей, связанных с электрическими машинами, электровакуумными приборами и т. д., инженеры часто используют первое упрощение, заключающееся в выпрямлении некоторых из граничных кривых с помощью соответствующего конформного отображения (см. § 13.7). В этом упрощенном поле от руки наносится криволинейная сетка эквипотенциальных линий и линий тока так, чтобы каждая ячейка сетки приблизительно имела форму квадрата. Иногда для этой цели используется электролитическая ванна или моделирующее устройство на постоянном токе. Применив к полученным таким образом результатам обратное отображение, мы найдем решение исходной задачи.

Применение электролитических ванн для изучения электрических полей описано Пирсом [10]. Использование этих ванн для построения линий уровня комплекснозначных функций, заданных явными выражениями, излагаются у Бердселла и Виннэри [2], которые иллюстрируют этот метод при изучении линий уровня комплексной функции $\omega J_1(\omega)/J_0(\omega)$, где J_i — функции Бесселя.

Другой, более наглядной, но несколько менее точной моделью является модель с резиновой пленкой. Рассмотрим горизонтальную, равномерно растянутую легкую резиновую пленку. Приложим к некоторым внутренним точкам или областям, или отрезкам прямых или кривых линий на этой пленке вертикально направленные силы, слегка перемещающие пленку. Тогда расстояния точек пленки от горизонтальной плоскости приблизительно будут удовлетворять уравнению Лапласа (см. § 17.3). Небольшой шарик (теоретически лучше рассматривать массу, скользящую без трения), положенный на пленку,

будет скатываться по направлению градиента (линия тока), ортогональному к линиям уровня (эквипотенциальные линии). Его скорость будет приближенно пропорциональна скорости электрона в аналогичном электрическом поле. Если шарик покрыт слоем порошка, то на поверхности пленки останутся следы.

Иногда составляются подробные каталоги конформных отображений, например отображений многоугольных областей с помощью формулы Шварца—Кристоффеля. Эти отображения можно использовать в качестве достаточно близких первых приближений к заданным областям произвольного вида. Однако, как показал Альфорс [12], уже для шестиугольников при сравнительно небольших изменениях параметров в разумных пределах потребуется таблица, содержащая до 10^{10} входов. Кроме того, все существующие каталоги определяют только границы и площади областей, в то время как нам необходимо знать также и линии уровня.

Таким образом, мы не можем рассчитывать на сколько-нибудь достаточную полноту каталогов конформных отображений, если не считать очень ограниченного числа исключительно важных случаев, таких, например, как профили Жуковского и их обобщения. Поэтому необходимо знать отображения, осуществляемые различными функциями, и при этом понимать геометрические значения параметров, входящих в эти функции.

Часто требуется составлять комбинации известных отображающих функций. Это легко осуществить на цифровых машинах, библиотеки подпрограмм которых содержат таблицы существующих функций. Так, например, Бердселл и Виннери (см. выше) отмечают, что для определения линий уровня функции $\omega J_1(\omega)/J_0(\omega)$ можно вместо электролитических ванн пользоваться с равным успехом численными методами.

Методы численного определения конформных отображений и решения задачи Дирихле основываются обычно на конструктивных доказательствах основных теорем существования. Поэтому они включают принцип Дирихле (см. § 4.13) и другие экстремальные методы (см. гл. 14 и 16), применимые к различным геометрическим вопросам конформных отображений (см. § 13.6), теоремы о среднем значении для гармонических функций (см. § 5.7, 11.8, 15.3 и 17.3, [11]), метод ядер (см. гл. 6) и т. д. В силу теорем Рунге и Уолша (см. § 7.12) конформные отображения односвязных областей можно приближенно с любой степенью точности осуществлять при помощи полиномов (в случае многосвязных областей при помощи рациональных функций). Поэтому при любом приближенном исследовании класс рассматриваемых функций может быть сильно сужен. Если граница области достаточно проста, то приближение равномерно близко во всей замкнутой области.

Альтернирующий метод Шварца, при котором отображаемая область покрывается конечным числом пересекающихся областей,

отображение которых на внутренность единичного круга известно, особенно легко применим к отображению полигональных областей, поскольку последние могут быть покрыты, например, частями кругов; пользуясь известными отображениями этих частей, мы получаем при помощи альтернирующего процесса отображение всей области.

Для построения конформных отображений на круг областей, близких к кругу, применяется метод последовательных сопряжений Теодорсена и Гаррика [1]. Этот метод применим также в качестве второго шага, когда известно отображение заданной области на область, близкую к кругу. Первый шаг часто осуществляется просто, если найти подходящее отображение в хорошем каталоге. Так, например, имея профиль, мало отличающийся от профиля Жуковского, мы находим в каталоге этих профилей и их обобщений профиль, наиболее близкий к заданному. Функция, отображающая внешность выбранного профиля на внешность единичного круга, отобразит соответствующим образом расположенный заданный профиль на область, близкую к окружности.

ЛИТЕРАТУРА

1. Beckenbach E. F. (ed.), Construction and Applications of Conformal Maps, Nat. Bur. Standards Appl. Math. Ser., 18 (1952).
2. Birdsall Ch. K. and Whinnery J. R., Waves in an Electron Stream with General Admittance Walls, J. Appl. Phys., 24, 315—323 (1953).
3. Birkhoff G., Young D. M., Zarantonello E. H., Numerical Methods in Conformal Mapping, Proc. Symposia in Applied Mathematics, IV: Fluid Dynamics, American Mathematical Society, McGraw-Hill, New York, 1953.
4. Douglas J., A Method of Numerical Solution of the Problem of Plateau, Ann. of Math., 29, 180—188 (1927—1928).
5. Канторович Л. В. и Крылов В. И., Приближенные методы высшего анализа, изд. 4, Гостехиздат, М., 1950.
6. Kellogg O. D., Foundations of Potential Theory, Springer-Verlag, Berlin, 1929.
7. Kober H., Dictionary of Conformal Representations, British Admiralty, Dover, New York, 1952.
8. Milne-Thomson L. M., Theoretical Aerodynamics, D. Van Nostrand, New York, 1947.
9. Nehari Z., Conformal Mapping, McGraw-Hill, New York, 1952.
10. Pierce J. R., Theory and Design of Electron Beams, D. Van Nostrand, New York, 1949.
11. Shortley G. H., Weller R., The Numerical Solution of Laplace's Equation, J. Appl. Phys., 9, 334—348 (1938).
12. Todd J. (ed.), Experiments in the Computation of Conformal Maps, Nat. Bur. Standards Appl. Math. Ser., 42 (1955).

Глава 14

НЕЛИНЕЙНЫЕ МЕТОДЫ

ЧАРЛЬЗ Б. МОРРЕЙ

Калифорнийский университет, Беркли

§ 14.1. Введение. Доклады, прочитанные в последнее время на заседаниях Американского математического общества (начиная с докладов фон Кармана в 1939 г. и Эрнеста Г. Келлера в 1943 г.), свидетельствуют о все возрастающем интересе к нелинейным проблемам. Нет сомнения в том, что развитие быстродействующих вычислительных машин способствовало повышению этого интереса, ибо с помощью этих машин практическое решение многих задач стало более доступным.

Работая в области чистой математики и лишь незначительно соприкасаясь с проблемами техники, автор не в состоянии перечислить интересные (для техники) типы линейных и нелинейных задач, не говоря уже об изложении методов их решения. Сведения подобного рода можно найти в упомянутых выше докладах и в периодической литературе по вопросам прикладной математики, выходящей все в большем и большем количестве. У многих из вас имеются свои собственные проблемы, которые можно было бы присоединить к этой коллекции.

В этой главе мы попытаемся изложить лишь общие идеи, которые в лучшем случае помогут некоторым из вас сосредоточить внимание на основных трудностях в ваших проблемах, а не оказаться в плену многочисленных деталей. Некоторые из рассматриваемых здесь идей и методов иллюстрируются специально подобранными примерами. Применяемые в этих примерах численные методы являются классическими, и реализация их возможна с помощью настольных вычислительных машин. Так как здесь приводятся только простейшие формулы, то упомянутые методы без сомнения покажутся слишком примитивными для тех, кому приходилось работать в области численного анализа. Объем литературы по этому вопросу непрерывно увеличивается; существует журнал «Mathematical Tables and Other Aids to Computation», который целиком посвящен численным методам анализа.

Наши примеры функциональных уравнений взяты главным образом из вариационного исчисления. Следовательно, в нашей статье не будут рассматриваться такие вопросы, как нелинейные колебания, сверхзвуковые потоки, распространение волн и дифференциальные

уравнения гиперболического типа. Некоторые из этих вопросов, а также многие специальные вычислительные методы, такие, как решение систем линейных уравнений, метод быстрого спуска (упомянутые здесь лишь вкратце), рассматриваются в других параграфах нашей книги.

В настоящей главе мы будем касаться только тех задач, которые не могут быть решены с помощью искусственных приемов или стандартных методов, описанных в учебных пособиях по дифференциальным уравнениям, математическому анализу и пр. Существует много подобных методов, и овладеть ими можно лишь решив большое число задач или, по меньшей мере, тщательно изучив соответствующие учебники. В связи с этим мы не будем останавливаться, например, на методе функций комплексного переменного, который весьма полезен при решении многих двумерных задач, связанных с уравнением Лапласа, и т. д. (см. гл. 7 и 13). Однако это несколько не уменьшает ценность подобных специальных методов, ибо если они и не позволяют найти решение заданной проблемы, то все же могут быть применены к отысканию точного решения другой задачи, которая может оказаться полезным приближением для заданной проблемы.

Каковы же те методы, которые могут быть использованы при решении задач упоминавшегося в предыдущем абзаце типа. Для систем алгебраических уравнений мы вкратце рассмотрим методы: возмущений, последовательных приближений (в их числе — метод Ньютона и два его видоизменения), непрерывности, бесконечных степенных рядов и быстрого спуска (применяется в вариационных задачах). Многие из этих методов имеют аналоги, предназначенные для решения функциональных уравнений. В то же время метод Рэлея — Ритца ¹⁾ и метод конечных разностей дают возможность перейти от функциональных уравнений к конечным системам алгебраических уравнений, которые решаются методами, изложенными ниже.

Системы алгебраических уравнений

§ 14.2. Метод Ньютона. Метод возмущений. Модификации этих методов. В этом параграфе мы остановимся на системах алгебраических уравнений с конечным числом неизвестных и посмотрим, каким образом различные упоминавшиеся выше методы могут быть применены для отыскания их решений. Все методы будут рассмотрены довольно подробно, ибо, как мы увидим далее, численное решение многих функциональных уравнений методом Рэлея — Ритца или методом конечных разностей сводится к решению систем алгебраических уравнений.

¹⁾ В советской литературе этот метод называется методом Ритца. — *Прим. ред.*

Метод возмущений для уравнения вида

$$f(x) = 0 \quad (14.1)$$

состоит в том, что начинают с приближенного решения, например x_0 , заменяют уравнение (14.1) линеаризованным уравнением

$$f(x_0) + f'(x_0)(x - x_0) = 0 \quad (14.2)$$

и решают последнее для получения лучшего приближения, скажем x_1 . Метод Ньютона состоит в повторении этого процесса и получении *последовательных приближений* $x_2, x_3, \dots$. Другими словами, метод Ньютона состоит в вычислении последовательных приближений путем решения уравнений

$$\begin{aligned} f(x_0) + f'(x_0)\xi_1 &= 0, & x_1 &= x_0 + \xi_1, \\ f(x_1) + f'(x_1)\xi_2 &= 0, & x_2 &= x_1 + \xi_2 = x_0 + \xi_1 + \xi_2, \\ & \dots & & \dots \\ f(x_n) + f'(x_n)\xi_{n+1} &= 0, & x_{n+1} &= x_n + \xi_{n+1} = x_0 + \sum_{i=1}^{n+1} \xi_i. \end{aligned} \quad (14.3)$$

Геометрический смысл метода возмущений состоит в замене графика функции $y = f(x)$ касательной, проведенной в точке $x = x_0$ этого графика. Метод Ньютона состоит в повторении этого процесса для получения каждого следующего приближения по предыдущему результату и продолжении его до достижения желаемой точности. Если $f'(x)$ и $f''(x)$ в интервале $a \leq x \leq b$ имеют один и тот же знак, а $f(a)$ и $f(b)$ — противоположные знаки, то легко заметить, представив все это геометрически, что процесс последовательных приближений сходится и довольно быстро, начиная с любого x_0 из этого интервала. Если функция $f'(x)$ непрерывна и $f'(x^*) \neq 0$, причем x^* является истинным решением, то последовательные приближения, находимые по методу Ньютона, будут сходиться к решению x^* при условии, что x_0 достаточно близко к x^* ; если $f''(x)$ меняет знак внутри интервала, то этот процесс может не сходиться, если только x_0 в действительности не окажется достаточно близким к x^* .

Мы остановимся на двух модификациях метода Ньютона. Первая заключается в получении последовательных приближений $x_1, x_2, \dots$ не из уравнений (14.3), а решением уравнений

$$\begin{aligned} f(x_0) + f'(x_0)\xi_1 &= 0, & x_1 &= x_0 + \xi_1, \\ f(x_1) + f'(x_0)\xi_2 &= 0, & x_2 &= x_1 + \xi_2 = x_0 + \xi_1 + \xi_2, \\ & \dots & & \dots \\ f(x_n) + f'(x_0)\xi_{n+1} &= 0, & x_{n+1} &= x_n + \xi_{n+1} = x_0 + \sum_{i=1}^{n+1} \xi_i, \end{aligned} \quad (14.4)$$

в которых на каждом шаге вместо x_n в $f'(x)$ подставляется x_0 . Геометрически это соответствует замене касательной в точке $\{x_n, f(x_n)\}$ прямой, проходящей через эту точку и параллельной касательной в точке x_0 . Этот метод имеет то преимущество, что величина $1/f'(x_0)$ может быть вычислена раз навсегда, а уравнения (14.4) можно переписать в виде

$$\xi_{n+1} = -\frac{1}{f'(x_0)} f(x_n), \quad n = 0, 1, 2, \dots$$

Преимущество этого приема по сравнению с методом Ньютона возрастает с увеличением числа уравнений, подлежащих решению. Однако, вообще говоря, этот процесс сходится медленнее, и для успешности его применения необходимо, чтобы x_0 было более близким к x^* , чем в методе Ньютона.

Вторая модификация состоит в замене уравнения (14.1) уравнением

$$f(x_0 + X) \equiv f(x_0) + f'(x_0) X + g(X) = 0, \quad (14.5)$$

в котором тождество определяет $g(X)$. Уравнение (14.5) может быть записано в виде

$$X = Y + G(X), \quad (14.6)$$

где

$$Y = -\frac{f(x_0)}{f'(x_0)}, \quad G(X) = -\frac{g(X)}{f'(x_0)}.$$

Затем мы определяем последовательные приближения $X_1, X_2, \dots$ с помощью следующих соотношений:

$$X_1 = Y, \quad X_{n+1} = Y + G(X_n), \quad n = 1, 2, \dots$$

В обеих модификациях первое приближение x_1 совпадает с первым приближением, полученным по методу возмущений.

Естественно, что до применения метода Ньютона и его упоминавшихся выше модификаций желательно тем или иным способом (например, построением графика) провести достаточно подробное исследование функции $f(x)$ с целью получения грубых приближений к корням. Если имеется таблица значений функции $f(x)$ или какой-либо части этой функции (как в случае уравнений, аналогичных уравнению $\lg x = 2x$), то ее корни могут быть быстро найдены интерполированием.

Идеи метода возмущений и метода Ньютона, а также его модификаций переносятся на системы N уравнений с N неизвестными. Для $N=2$, когда решаются уравнения

$$f(x, y) = 0, \quad g(x, y) = 0, \quad (14.7)$$

метод Ньютона состоит в том, что мы начинаем с некоторого приближения (x_0, y_0) и последовательно решаем линеаризованные уравнения

$$\begin{aligned} f(x_n, y_n) + f_x(x_n, y_n)\xi_{n+1} + f_y(x_n, y_n)\eta_{n+1} &= 0, & x_{n+1} &= x_n + \xi_{n+1}, \\ g(x_n, y_n) + g_x(x_n, y_n)\xi_{n+1} + g_y(x_n, y_n)\eta_{n+1} &= 0, & y_{n+1} &= y_n + \eta_{n+1}, \end{aligned} \quad (14.8)$$

аналогичные уравнениям (14.3) (здесь индексы при f_x, f_y и т. д. означают частное дифференцирование). Первая модификация состоит в решении уравнений (14.8) при условии замены (x_n, y_n) на (x_0, y_0) . Так как коэффициенты при ξ_{n+1} и η_{n+1} не зависят от n , то мы можем решить модифицированные уравнения (14.8) и получить

$$\xi_{n+1} = \alpha f(x_n, y_n) + \beta g(x_n, y_n), \quad \eta_{n+1} = \gamma f(x_n, y_n) + \delta g(x_n, y_n),$$

где α, β, γ и δ могут быть только приближенными значениями. Вторая модификация состоит в замене уравнений (14.7) уравнениями

$$\begin{aligned} f(x_0 + X, y_0 + Y) &\equiv f(x_0, y_0) + f_x(x_0, y_0)X + f_y(x_0, y_0)Y + h(X, Y) = 0, \\ g(x_0 + X, y_0 + Y) &\equiv g(x_0, y_0) + g_x(x_0, y_0)X + g_y(x_0, y_0)Y + k(X, Y) = 0, \end{aligned}$$

где тождества определяют h и k ; решая эти уравнения, получаем соотношения

$$X = Z + H(X, Y), \quad Y = W + K(X, Y),$$

аналогичные (14.6). Теперь из соотношений

$$X_1 = Z, \quad Y_1 = W, \quad X_{n+1} = Z + H(X_n, Y_n), \quad Y_{n+1} = W + K(X_n, Y_n)$$

могут быть получены последовательные приближения $(X_1, Y_1), (X_2, Y_2), \dots$

Как и в случае одной функции одной переменной, желательно иметь исходное приближенное решение (x_0, y_0) . В случае уравнений (14.7) это легко осуществляется путем графического построения линий, соответствующих этим уравнениям. Однако в случае N уравнений с N неизвестными для $N > 2$ графический метод практически неосуществим. В связи с этим мы предлагаем для $N = 2$ следующий прием, который может быть легко обобщен: составьте таблицу с двумя входами значений одной из этих функций, например $g(x, y)$, и воспользуйтесь ею для получения грубых решений $y = \varphi_1(x), \varphi_2(x)$ и т. д. (или $x = \varphi_1(y)$ и т. д.) уравнения $g(x, y) = 0$, считая y функцией x ; подставьте эти значения в другое уравнение, которое затем решите относительно x (или y), после чего найдите соответствующее значение y (или x). Такая программа действий осуществляется в примере (14.5) Приложения (§ 14.20) для пары уравнений

$$\begin{aligned} f(x, y) &\equiv 4 + x + y + x^2 + 2xy + 3y^2 = 0, \\ g(x, y) &\equiv 1 + 2x - 3y + x^2 + xy - 2y^2 = 0. \end{aligned} \quad (14.9)$$

За приближенное значение корня принимается $(3, 3; -3, 0)$; корень вычисляется с помощью первой модификации метода Ньютона с точностью до пяти десятичных знаков.

Для случая трех функций трех переменных, например x , y и z , надо начать с табулирования одной из функций, например по переменным y и z для каждого из нескольких значений x , использовать эти данные для отыскания решения z (зависящего от x , y) из уравнения, определяющего протабулированную функцию, а затем подставить эти решения во вторые два уравнения и закончить этот этап работы грубым определением корней этих уравнений. Указанный прием аналогичен обычному приему решения линейных уравнений методом исключения неизвестных. Конечно, может оказаться, что полученное таким путем грубое решение удается в достаточной степени уточнить „интерполяцией высшего порядка“.

§ 14.3. Замечания о системах линейных уравнений. Все рассмотренные выше методы связаны с решением систем линейных алгебраических уравнений. За исключением случая $N=2$ (два уравнения с двумя неизвестными) и некоторых других специальных случаев, решение с помощью определителей оказывается малопригодным. Вообще говоря, метод исключений почти столь же хорош, как и другие методы. Однако задача решения систем линейных алгебраических уравнений настолько важна, что за последнее время ей уделялось большое внимание, и было разработано много специальных методов решения подобных систем и вычисления матриц, обратных данным. В гл. 15 будут изложены многие из таких методов, поэтому здесь мы не будем на них останавливаться. В связи с изучением линейных систем в гл. 12 уже изучались матрицы и матричная алгебра. Это позволяет нам в настоящей главе свободно пользоваться терминологией, символикой и результатами теории матриц.

Для отыскания второго приближения в примере 14.8 Приложения (§ 14.20) линейная система из четырех уравнений с четырьмя неизвестными вида

$$\sum_{j=1}^4 a_{ij} \Delta \zeta_j = -k_i, \quad i = 1, \dots, 4,$$

в которой

$$\begin{aligned} \zeta_1 &= \eta_0, & \zeta_2 &= \eta_2, & \zeta_3 &= \eta_4, & \zeta_4 &= y_2, \\ k_1 &= f_0, & k_2 &= f_2, & k_3 &= f_4, & k_4 &= g, \end{aligned}$$

решается относительно $\Delta \zeta_i$ методом исключения. В результате $\Delta \zeta_i$ выражаются через k_i и мы получаем

$$\Delta \zeta_i = \sum_{j=1}^4 A_{ij} k_j,$$

где (A_{ij}) — матрица, обратная матрице (a_{ij}) , взятой со знаком минус.

Мы упомянем здесь только об одном методе получения обратной матрицы, а именно о классическом методе Хотеллинга [1], который может быть применен при наличии достаточно хорошего приближения A_0 к обратной матрице a^{-1} . Предположим, что I обозначает единичную матрицу, и напишем

$$aA_0 \equiv C = I - D, \quad B = C^{-1}.$$

Тогда, если элементы матрицы D достаточно малы, мы получим

$$B = C^{-1} = I + D + D^2 + D^3 + \dots \quad (14.10)$$

во всех тех случаях, когда этот ряд сходится. Этот ряд, конечно, можно применить для вычисления B . Когда матрица B получена, мы будем иметь

$$(aA_0)B = a(A_0B) = I,$$

так что $A_0B = a^{-1}$. Метод Хотеллинга значительно ускоряет сходимость ряда (14.10); он состоит в определении последовательности $B_1, B_2, \dots$ приближенных значений B :

$$B_1 = 2I - C, \quad B_2 = B_1(2I - CB_1), \quad \dots, \quad B_{n+1} = B_n(2I - CB_n), \quad \dots$$

Легко можно убедиться в том, что

$$B_1 = I + D, \quad B_2 = I + D + D^2 + D^3,$$

$$B_3 = I + D + D^2 + \dots + D^7,$$

$$B_4 = I + D + \dots + D^{15} \text{ и т. д.}$$

Конечно, никогда не возникает необходимости в фактической записи матрицы D .

§ 14.4. Замечания о решении обыкновенных дифференциальных уравнений. Два метода из тех, которые будут рассматриваться ниже, требуют, по меньшей мере, грубого числового решения системы обыкновенных дифференциальных уравнений. Техника „шагового“ интегрирования систем обыкновенных дифференциальных уравнений хорошо известна, и каждая вычислительная организация применяет, по-видимому, свои собственные приемы вычислений. Предположим, что

$$\frac{dy_i}{dx} = f_i(x, y_1, \dots, y_N), \quad i = 1, 2, \dots, N,$$

— заданная система дифференциальных уравнений. Для выполнения вычислений с помощью настольной вычислительной машины составляется бланк, в котором отводятся колонки для значений x , для каждой из переменных y'_i и y_i , по крайней мере одна колонка для разностей каждой из y'_i и y_i и дополнительная колонка, необходимая для вычисления f_i . Обычно задаются начальные значения x и y_i , которые

и используются для вычисления начальных значений y'_i . Эти значения в свою очередь используются для вычисления новых значений каждого из y_i при следующем значении x . Полученные значения y_i используются для отыскания y'_i , соответствующих этому новому значению x . Эти значения совместно с предыдущими значениями y'_i используются для вычисления новых значений y_i для очередного значения x с помощью какой-либо приспособленной для рассматриваемого случая формулы численного интегрирования. Весь этот процесс повторяется до тех пор, пока наблюдается изменение значений хотя бы одной из определяемых величин. На этом заканчивается первый этап вычислений. Следующие этапы вычислений аналогичны предыдущему. Интервалы изменения x должны быть выбраны достаточно малыми, чтобы свести к минимуму общую погрешность, возникающую вследствие округления и неточности формул численного интегрирования; с другой стороны, они не должны быть малы настолько, чтобы получение окончательного решения требовало слишком большого числа шагов. Обычно целесообразно начинать с нескольких коротких шагов, чтобы построить разности более высоких порядков.

Различные методы, реализуемые с помощью настоящих вычислительных машин, отличаются друг от друга способами вычисления и регистрации разностей, а также применяемыми формулами численного интегрирования. Приведем один хорошо зарекомендовавший себя метод. Регистрация разностей (см. пример 14.6 Приложения (§ 14.20)) каждой из величин y'_i или y_i производится в соответствующей колонке в правой ее части следующим образом: слева выписываются первые разности, правее — вторые разности и т. д. Каждая первая разность, отвечающая данному значению x , является именно той, которая получается вычитанием значения y_i (или y'_i), соответствующего предшествующему значению x , из значения, соответствующего заданному x . Каждая вторая разность, расположенная против данного значения x , равна значению первой разности, стоящей против того же x , минус разность, расположенная непосредственно выше ее; принимается, что все интервалы по x имеют равные длины (непосредственно в начале перехода к удлинненному интервалу изменения x новые разности должны быть вычислены с использованием только интервалов новой длины, для чего следует пропускать те из промежуточных значений, y_i или y'_i , соответствующих более коротким интервалам, индексы i которых не кратны длине нового интервала). Тогда формула для интегрирования $\varphi(x)$ от x_{n-1} до x_n имеет вид

$$\int_{x_{n-1}}^{x_n} \varphi(x) dx \sim \frac{h}{2} \left[\varphi(x_{n-1}) + \varphi(x_n) - \frac{1}{6} \Delta^2 \varphi(x_n) - \frac{1}{12} \Delta^3 \varphi(x_n) \right].$$

где $h = x_n - x_{n-1}$, а $\Delta^2 \varphi_n$ и $\Delta^3 \varphi_n$ являются второй и третьей разностями, соответствующими значению x_n (для первых нескольких шагов нужно использовать максимально возможное число разностей). Шаги должны быть настолько малы, чтобы для третьих разностей ошибка в округлении была наименьшей.

В примере 14.6 Приложения интегрируется очень простое уравнение $y' = y$ при начальном условии $y(0) = 1$. Точный результат есть e^x ; значение y для $x = 0,10$ с точностью до семи десятичных знаков оказывается равным 1,1051709. Интегрирование более сложной системы (пример 14.7 Приложения) осуществляется с меньшей точностью (при небольшом числе шагов интегрирования).

§ 14.5. Метод непрерывности. Метод бесконечных степенных рядов. В случае N уравнений с N неизвестными, когда N велико, определение приближенного решения может оказаться весьма трудоемким. В этом случае можно применить один из двух рассматриваемых в этом параграфе методов, причем оба они основаны на введении в уравнения дополнительной переменной, а именно: вместо системы уравнений

$$f_i(x_1, \dots, x_N) = 0, \quad i = 1, \dots, N,$$

можно рассмотреть систему

$$f_i(x_1, \dots, x_N, \lambda) = 0, \quad i = 1, \dots, N. \quad (14.11)$$

Функции $f_i(x_1, \dots, x_N, \lambda)$ следует выбрать с таким расчетом, чтобы зависимость их от λ была возможно более простой и чтобы при $\lambda = 1$ система (14.11) сводилась к исходной системе, а при $\lambda = 0$ сводилась к системе, решение которой легко найти. В каждом из двух методов нужно подобрать $x_1, \dots, x_N$ в виде некоторых функций от λ , удовлетворяющих измененным уравнениям (14.11) для всех значений λ от 0 до 1. При этом нужно начинать с решения, легко получаемого при $\lambda = 0$. В каждом из этих методов производные $x'_j(\lambda)$, $x''_j(\lambda)$ и т. д. для $j = 1, \dots, N$ должны удовлетворять соотношениям

$$\sum_{j=1}^N f_{ix_j}(x_1, \dots, x_N, \lambda) x'_j(\lambda) + f_{i\lambda}(x_1, \dots, x_N, \lambda) = 0 \quad (14.12)$$

и

$$\sum_{j=1}^N (f_{ix_j} x''_j + 2f_{ix_j \lambda} x'_j) + \sum_{j=1}^N \sum_{k=1}^N f_{ix_j x_k} x'_j x'_k + f_{i\lambda \lambda} = 0, \quad (14.13)$$

полученным путем дифференцирования измененных уравнений (14.11) по переменной λ при условии, что $x_1, \dots, x_N$ рассматриваются как функции λ .

Метод непрерывности состоит в решении системы (14.12) обыкновенных дифференциальных уравнений с начальными значениями x_i , равными соответствующим значениям этих переменных при $\lambda = 0$. Уравнения (14.12) линейны относительно x'_i . Если они могут быть решены при $\lambda = 0$ (и если f_{ix_j} , $f_{i\lambda}$ и производные второго порядка непрерывны), то из теоремы существования следует, что решение существует и единственно для всех значений λ , достаточно близких к 0, и может быть однозначно продолжено до тех пор, пока не встретится особая точка уравнений или решение не уйдет в бесконечность. Если решение может быть продолжено до $\lambda = 1$, то значения x_i для $\lambda = 1$ оказываются решениями исходных уравнений; действительно, если система (14.12) удовлетворяется для всех λ от 0 до 1 и уравнение (14.11) удовлетворяется для $\lambda = 0$, то уравнение (14.11) удовлетворяется для всех λ от 0 до 1.

Каждый шаг процесса численного интегрирования системы (14.12) требует решения N линейных уравнений и определения значений $x'_i(\lambda)$, и это, конечно, делает метод трудоемким. Если шаги не очень велики и решение рассматривается в области, не очень близкой к особой точке, то обратная матрица для предшествующих значений λ будет достаточно близка к обратной матрице, отвечающей заданному значению λ , чтобы классический метод Хотеллинга оказался эффективным. К сожалению, в общем случае не существует гарантии, что решение может быть продолжено для всех значений параметра λ от $\lambda = 0$ до $\lambda = 1$ или, если такое продолжение возможно на всем интервале, что оно приведет к желаемому решению, когда решений имеется несколько.

Этим методом решается система (14.9) (пример 14.7 Приложения (§ 14.20)). Здесь параметр λ введен в качестве коэффициента при квадратичных членах в f и g с таким расчетом, чтобы получившаяся после введения λ система (14.11) была линейной при $\lambda = 0$ и сводилась к исходной при $\lambda = 1$. Система (14.12) решена весьма грубо, причем результат для $\lambda = 1$ был улучшен применением первой модификации метода Ньютона. Решение, полученное этим методом, оказалось не совпадающим с решением, вычисленным ранее.

Метод степенных рядов состоит в применении разложения

$$x_i(\lambda) = x_i(0) + x'_i(0)\lambda + \frac{1}{2!}x''_i(0)\lambda^2 + \dots \quad (14.14)$$

Предполагается, что начальные значения $x_i(0)$ даны. Уравнения (14.12) при $\lambda = 0$ определяют значения $x'_i(0)$; уравнения (14.13) при $\lambda = 0$ определяют значения $x''_i(0)$ через $x_i(0)$ и $x'_i(0)$, а уравнения более высоких порядков, полученные из уравнений (14.13), определяют значения последующих производных от x_i при $\lambda = 0$ через вычисленные ранее величины. Преимущество этого метода состоит в том, что коэффициенты при производных высших порядков оказываются

всегда неизменными, так что необходимо вычислить только одну обратную матрицу. Однако формулы для производных высших порядков усложняются во всех случаях, кроме тех, в которых f_i представляют собой достаточно простые функции. Кроме того, точно не известно, будут ли члены высших порядков достаточно быстро стремиться к нулю для $\lambda=1$, что сделало бы этот метод практически ценным. Однако нет необходимости предполагать, что ряды (14.14) сходятся; их можно оборвать на любой степени λ^k , а весь остаток оценить по теореме Тейлора.

§ 14.6. Замечания, касающиеся метода быстрого спуска. К числу часто встречающихся задач относится задача об определении значений $(x_1^*, \dots, x_N^*)$, при которых функция $F(x_1, \dots, x_N)$ принимает минимальное или максимальное значение. В точке экстремума выполняются соотношения

$$F_{x_1} = F_{x_2} = \dots = F_{x_N} = 0, \quad (F_x = F_y = 0, \text{ если } N=2) \quad (14.15)$$

и, как известно, каждая определяемая уравнениями (14.15) точка называется *стационарной точкой* функции F . Уравнения (14.15) образуют систему частного вида, к которой могут быть приложены все ранее рассмотренные методы. Однако, если решение уравнений (14.15) определяет локальный минимум, имеется дополнительный метод, так называемый *метод быстрого спуска* (см. гл. 14). В своем чистом виде этот метод состоит в отыскании функций $x_i(\lambda)$, являющихся решениями уравнений

$$x'_i = -k F_{x_i}, \quad i = 1, \dots, N, \quad k > 0, \quad (14.16)$$

при некоторых начальных значениях, заданных для $\lambda=0$, и при некоторой подходящей функции (или постоянной) k . Если x_i являются решениями уравнений (14.16), то

$$\frac{d}{d\lambda} F[x_1(\lambda), \dots, x_N(\lambda)] = \sum_{j=1}^N F_{x_j} x'_j = -k \sum_{j=1}^N (F_{x_j})^2,$$

так что F убывает с возрастанием λ .

Если исходные значения достаточно близки к относительному (или абсолютному) минимуму, то каждая траектория при $k \equiv 1$ приводит к искомой точке. Однако по мере приближения траектории к искомой точке производные x'_i будут стремиться к нулю, и приближение к точке минимума будет иметь „асимптотический“ характер. Тем не менее, если значения x_i достаточно близки к искомой точке и F имеет в ней так называемый невырожденный минимум, то функ-

ция F достаточно точно аппроксимируется членами второго порядка соответствующего ряда Тэйлора:

$$F \sim F(x_1^*, \dots, x_N^*) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_{ij} (x_i - x_i^*) (x_j - x_j^*),$$

$$a_{ij} = F_{x_i x_j}(x_1^*, \dots, x_N^*) = a_{ji}.$$

Входящая в правую часть сумма оказывается положительной, если не все $x_i = x_i^*$. Поведение решений уравнений (14.16), например, при $k=1$, будет все более похожим на поведение решений линейной системы

$$x'_i = - \sum_{j=1}^N a_{ij} (x_j - x_j^*), \quad i = 1, \dots, N. \quad (14.17)$$

Вследствие того, что $a_{ij} = a_{ji}$, система (14.17) с помощью поворота осей в пространстве x может быть приведена к диагональной форме ($a_{ij} = 0$ при $i \neq j$). Следовательно, решения x_i^* системы (14.17) таковы, что каждая из разностей $x_i - x_i^*$ оказывается суммой экспоненциальных функций с отрицательными коэффициентами при λ в показателях. При достаточно больших λ одна из этих экспонент будет преобладать, остальными же можно будет пренебречь и, следовательно, начиная с достаточно большого λ , траектория будет стремиться к искомой точке по экспоненциальному закону. К сожалению, бывает и так, что траектория получает экспоненциальный характер лишь для очень больших значений λ .

Очевидно, что изложенный метод может быть легко модифицирован и обобщен на случай, когда $F(x_1, \dots, x_N)$ имеет локальный максимум. Заметим, что если искомая стационарная точка не является ни точкой максимума, ни точкой минимума, то этот метод не даст нужных результатов. Например, если $2F = x^2 - y^2$, то уравнения (14.16) при $k=1$ имеют вид

$$x' = -x, \quad y' = y,$$

что дает решения

$$x = x_0 e^{-\lambda}, \quad y = y_0 e^{\lambda}.$$

За исключением случая, когда $y_0 = 0$, траектория, определяемая этими решениями, будет с возрастанием λ уходить в бесконечность. В общем случае для системы (14.17) это, как правило, будет иметь место всегда, если только начальные значения не окажутся расположенными на некоторых кривых частного вида.

Рассмотренный метод до некоторой степени осложняется тем, что он связан с решением системы обыкновенных дифференциальных уравнений. Он, однако, обладает и определенными преимуществами, так как не требует вычисления всех производных $f_{ix_j} = F_{x_i x_j}$ или обра-

жения какой-либо матрицы. Кроме того, данная система уравнений может иметь вид (14.15) даже тогда, когда ее вывод не связан с какой-либо задачей на минимум. Эти преимущества способствовали широкому применению указанного метода и созданию его модификаций, облегчающих вычисления. Ясно, например, что нет необходимости решать с большой точностью уравнение (14.16).

Векторные обозначения

§ 14.7. Векторные обозначения. Операции над векторами. Длина. В этом параграфе мы введем векторные обозначения, которые подготовят путь к изложению в § 14.12 элементов теории нормированных линейных пространств.

Во всей этой части главы мы будем считать положительное целое число N фиксированным, но произвольным. Буквы, напечатанные жирным шрифтом, такие как $\mathbf{x}$, $\mathbf{f}$ и др., будут обозначать последовательности $(x_1, \dots, x_N)$, $(f_1, \dots, f_N)$ и т. д.; мы будем считать эти последовательности *векторами*. Функция (скалярная) $f(x_1, \dots, x_N)$ будет обозначаться через $f(\mathbf{x})$. Через $\mathbf{f}(\mathbf{x})$ будем обозначать векторную функцию $[f_1(\mathbf{x}), \dots, f_N(\mathbf{x})]$. Таким образом, соотношения

$$f_i(x_1, \dots, x_N) = 0, \quad i = 1, \dots, N,$$

могут быть записаны коротко в виде

$$\mathbf{f}(\mathbf{x}) = 0. \quad (14.18)$$

Если $\mathbf{x} = (x_1, \dots, x_N)$ и c есть *скаляр* (действительное число), то $c\mathbf{x}$ обозначает вектор $(cx_1, \dots, cx_N)$; если $\mathbf{y} = (y_1, \dots, y_N)$ есть другой вектор, то $\mathbf{x} \pm \mathbf{y}$ обозначает вектор $(x_1 \pm y_1, \dots, x_N \pm y_N)$. Таким образом, возможно определить линейную комбинацию $c_1\mathbf{x}_1 + \dots + c_k\mathbf{x}_k$, где $\mathbf{x}_1, \dots, \mathbf{x}_k$ являются векторами, а $c_1, \dots, c_k$ — скалярами. Определим длину $|\mathbf{x}|$ вектора $\mathbf{x}$ формулой

$$|\mathbf{x}| = \left[\sum_{i=1}^N x_i^2 \right]^{1/2}.$$

Легко показать, что

$$|c\mathbf{x}| = |c| |\mathbf{x}|, \quad |\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|.$$

Определим также *внутреннее* (скалярное) *произведение* двух векторов $\mathbf{x}$ и $\mathbf{y}$ равенством

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^N x_i y_i.$$

Из хорошо известного неравенства Шварца следует, что

$$|\mathbf{x} \cdot \mathbf{y}| \leq |\mathbf{x}| |\mathbf{y}|. \quad (14.19)$$

§ 14.8. Дифференцируемые функции и векторные функции. Предположим, что $f(x)$ есть скалярная функция, непрерывная вблизи некоторого x_0 вместе со своими частными производными первого порядка. Тогда существуют некоторое $h > 0$ и функция $\varepsilon_1(r)$, стремящаяся к нулю вместе с r , такие что

$$f(x + \xi) = f(x) + \sum_{j=1}^N f_{x_j}(x) \xi_j + g(x, \xi), \quad (14.20)$$

причем

$$|g(x, \xi)| \leq |\xi| [\varepsilon(|\xi|)]$$

для всех x и ξ , удовлетворяющих неравенствам

$$|x - x_0| \leq h \text{ и } |x + \xi - x_0| \leq h.$$

Перепишем уравнение (14.20) в виде

$$f(x + \xi) = f(x) + l(x, \xi) + g(x, \xi), \quad l(x, \xi) = \sum_{j=1}^N f_{x_j}(x) \xi_j,$$

где $l(x, \xi)$ представляет собой для каждого x линейную однородную функцию ξ , наиболее точно аппроксимирующую разность $f(x + \xi) - f(x)$ при малых значениях ξ . Назовем $l(x, \xi)$ полным дифференциалом функции f ; вектор ξ играет роль дифференциального вектора dx . Если f имеет также вторые производные, непрерывные вблизи x_0 , то существуют такое $h > 0$ и такая стремящаяся к нулю вместе с r функция $\varepsilon_2(r)$, что

$$f(x + \xi) = f(x) + l(x, \xi) + \frac{1}{2!} q(x, \xi) + g(x, \xi),$$

$$q(x, \xi) = \sum_{j=1}^N \sum_{k=1}^N f_{x_j x_k}(x) \xi_j \xi_k, \quad |g(x, \xi)| \leq |\xi|^2 [\varepsilon_2(|\xi|)] \quad (14.21)$$

для всех указанных выше x и ξ . Из соотношений (14.20) и (14.21) имеем

$$l(x, \xi) = \left[\frac{d}{d\lambda} f(x + \lambda \xi) \right]_{\lambda=0}, \quad q(x, \xi) = \left[\frac{d^2}{d\lambda^2} f(x + \lambda \xi) \right]_{\lambda=0}. \quad (14.22)$$

Соответствующие определения и обозначения могут быть введены также для векторной функции $f = (f_1, \dots, f_N)$. Мы будем иметь тогда

$$l(x, \xi) = [l_1(x, \xi), \dots, l_N(x, \xi)], \quad l_i(x, \xi) = \sum_{j=1}^N f_{i x_j}(x) \xi_j$$

и соответствующую формулу для $q(x, \xi)$. В уравнениях, соответствующих (14.20) и (14.21), $|g(x, \xi)|$ будет обозначать длину g . Соотношения (14.22) заменяются соответствующими векторными соотношениями; их можно рассматривать либо как систему соответствующих

уравнений относительно компонент, либо как настоящую векторную производную по λ . Другими словами, $df(x + \lambda \xi)/d\lambda$ означает такой вектор, для которого

$$\lim_{\Delta\lambda \rightarrow 0} \left| \frac{f(x + (\lambda + \Delta\lambda)\xi) - f(x + \lambda\xi)}{\Delta\lambda} - \frac{d}{d\lambda} f(x + \lambda\xi) \right| = 0.$$

Эта точка зрения будет полезна в дальнейшем.

Если $l(\xi)$ является линейной однородной векторной функцией

$$l = (l_1, \dots, l_N), \quad l_i(\xi) = \sum_{j=1}^N a_{ij} \xi_j,$$

то соотношение

$$l(\xi) = n, \quad n = (\eta_1, \dots, \eta_N) \quad (14.23)$$

представляет собой сокращенную форму записи системы линейных уравнений

$$\sum_{j=1}^N a_{ij} \xi_j = \eta_i, \quad i = 1, \dots, N. \quad (14.24)$$

Ясно, что l полностью определяется матрицей (a_{ij}) . Если определитель этой матрицы отличен от нуля, то система (14.24) имеет единственное решение ξ , которое выражается через n (при любых n) в следующем виде:

$$\xi_i = \sum_{j=1}^N A_{ij} \eta_j \quad \text{или} \quad \xi = L(n).$$

Здесь матрица (A_{ij}) является обратной для матрицы (a_{ij}) , и потому линейную векторную функцию L называют обратной для l и обозначают через l^{-1} . Если определитель матрицы (a_{ij}) равен нулю, то существует такой вектор $\xi \neq 0$, что $l(\xi) = 0$; система уравнений (14.24) для всех n не может быть решена и не имеет однозначного решения при некоторых n , т. е. не существует обратной матрицы $(a_{ij})^{-1}$ и функции l^{-1} . Мы будем говорить, что функция l *обратима*, если существует l^{-1} .

Если придерживаться таких обозначений, то метод Ньютона состоит в последовательном решении уравнений

$$f(x_n) + l(x_n, \xi_{n+1}) = 0, \quad x_{n+1} = x_n + \xi_{n+1}, \quad n = 0, 1, 2, \dots, \quad (14.25)$$

из которых по заданному значению x_0 находятся приближенные значения $x_1, x_2, \dots, x_n, \dots$. Его первая модификация состоит в замене уравнений (14.25) уравнениями

$$f(x_n) + l(\xi_{n+1}) = 0, \quad x_{n+1} = x_n + \xi_{n+1}, \quad l(\xi_{n+1}) = l(x_0, \xi_{n+1}). \quad (14.26)$$

Уравнения (14.26) могут быть преобразованы к виду

$$\xi_{n+1} = -L[f(x_n)],$$

в котором L равна I^{-1} . Вторая модификация метода Ньютона состоит в записи соотношений (14.18) в виде

$$f(x_0 + X) \approx f(x_0) + l(X) + g(X) = 0. \quad (14.27)$$

Здесь $l(X)$ есть дифференциал функции f при x_0 и $g(X)$ определяется левой частью тождества (14.27). Применяя к обеим частям (14.27) оператор I^{-1} , получаем

$$X = Y + G(X), \quad Y = -I^{-1}[f(x_0)], \quad G(X) = -I^{-1}[g(X)]. \quad (14.28)$$

Уравнения (14.28) решаем, находя последовательные приближения

$$X_1 = Y, \quad X_{n+1} = Y + G(X_n), \quad n = 1, 2, \dots \quad (14.29)$$

§ 14.9. Образец доказательства сходимости. Теорема об обратном отображении. В этом параграфе мы докажем, что последовательные приближения во второй модификации метода Ньютона сходятся к точному решению x^* , если $f(x_0)$ достаточно мало. Фактически мы приведем доказательство теоремы об обратном отображении, содержащее доказательство сходимости последовательных приближений. Аналогичным приемом можно доказать сходимость последовательных приближений, полученных с помощью первой модификации метода Ньютона. Если вторые производные $f(x)$ также непрерывны, то известно, что основной метод Ньютона дает более быструю сходимость, чем любая из двух его модификаций.

Прежде чем приступить к доказательству сходимости, отметим, во-первых, что

$$|x - y| \geq |x| - |y|, \quad (14.30)$$

так как

$$|x| = |y + (x - y)| \leq |y| + |x - y|$$

для любых двух векторов x и y . Далее, предположим, что функция $h(x)$ непрерывна вместе со своими первыми производными в области, содержащей отрезок от x_1 до x_2 . Полагаем для каждого i

$$\varphi_i(\lambda) = h_i[x_1 + \lambda(x_2 - x_1)].$$

Тогда

$$\begin{aligned} h_i(x_2) - h_i(x_1) &= \varphi_i(1) - \varphi_i(0) = \varphi_i'(\bar{\lambda}_i) = \\ &= \sum_{j=1}^N h_{ij} [x_1 + \bar{\lambda}_i(x_2 - x_1)] (x_{2j} - x_{1j}) = \\ &= Vh_i [x_1 + \bar{\lambda}_i(x_2 - x_1)] \cdot (x_2 - x_1) \end{aligned} \quad (14.31)$$

для некоторого $\bar{\lambda}_i$, заключенного между 0 и 1. Здесь Vh_i обозначает вектор $(h_{i x_1}, \dots, h_{i x_N})$. Из уравнения (14.31) следует, что

$$|h(x_2) - h(x_1)| \leq |x_2 - x_1| \left[\sum_{i=1}^N |Vh_i(\xi_i)|^2 \right]^{1/2} \quad (14.32)$$

для выбранных надлежащим образом на отрезке от x_1 до x_2 точек $\xi_1, \dots, \xi_N$.

Теорема об обратном отображении. Допустим теперь, что $f(x)$ есть функция, непрерывная вблизи x_0 вместе со своими частными производными первого порядка, и предположим, что ее дифференциал $l(x_0, \xi)$ разрешим относительно ξ . Тогда найдутся такие положительные числа h и k , что если y — некоторый вектор, для которого выполняется неравенство $|y - f(x_0)| \leq k$, то существует одно и только одно решение x уравнения

$$y = f(x), \quad (14.33)$$

для которого

$$|x - x_0| \leq h.$$

Мы придем к только что сформулированной теореме, показав, что последовательные приближения, полученные с помощью второй модификации метода Ньютона, сходятся к решению, если y достаточно близко к $f(x_0)$. Согласно этому методу запишем уравнение (14.33) в виде

$$f(x_0 + X) - f(x_0) \equiv l(X) + g(X) = y - f(x_0), \quad (14.34)$$

где l есть дифференциал f при $x = x_0$. Тогда g есть также непрерывно дифференцируемая функция вблизи $X = 0$, причем в этой точке все производные функции g равны нулю [см. уравнение (14.20)]. Далее решаем уравнение (14.34) относительно X , выражаем X через $g(X)$ и $y - f(x_0)$, которые рассматриваем как неизвестные, и получаем результат

$$X = Y + G(X); \quad Y = l^{-1}[y - f(x_0)], \quad G(X) = -l^{-1}[g(X)]. \quad (14.35)$$

Так как l^{-1} есть линейный оператор, то G и ее производные непрерывны вблизи $X = 0$, причем при $X = 0$ функция G и частные производные всех ее компонент равны нулю. В соответствии с этим существует такое достаточно малое $h > 0$, что

$$\left[\sum_{i=1}^N |v G_i(\xi_i)|^2 \right]^{1/2} \leq \frac{1}{2} \quad (14.36)$$

для всех возможных последовательностей $\xi_1, \dots, \xi_N$, для которых все $|\xi_i| \leq h$. Тогда если $|X_1| \leq h$ и $|X_2| \leq h$, то из уравнений (14.32) и (14.36) видно, что

$$|G(X_1) - G(X_2)| \leq \frac{1}{2} |X_1 - X_2|. \quad (14.37)$$

Предположим теперь, что X_1 и X_2 — две различные точки, для которых $|X_1| \leq h$ и $|X_2| \leq h$, и $Y_k = X_k - G(X_k)$. Тогда из уравнений (14.30) и (14.37) следует

$$\begin{aligned} |Y_1 - Y_2| &= |(X_1 - X_2) - [G(X_1) - G(X_2)]| \geq \\ &\geq |X_1 - X_2| - |G(X_1) - G(X_2)| \geq \frac{1}{2} |X_1 - X_2|. \end{aligned} \quad (14.38)$$

Таким образом, различным значениям X соответствуют различные значения Y . Теперь выберем $K = \frac{h}{2}$ и предположим, что $|Y| \leq K$. Пусть $X_1, X_2, \dots$ — последовательные приближения, определяемые уравнениями (14.29). Тогда

$$\begin{aligned} |X_1| &= |Y| \leq \frac{h}{2}, \quad |X_2 - X_1| = |G(Y)| \leq \frac{1}{2} |X_1| \leq \frac{h}{4} = \frac{h}{2^2}, \\ |X_{n+1} - X_n| &= |G(X_n) - G(X_{n-1})| \leq \frac{1}{2} |X_n - X_{n-1}| \leq \frac{h}{2^{n+1}}, \\ n &= 2, 3, \dots \end{aligned} \quad (14.39)$$

Пользуясь уравнением (14.39), убеждаемся последовательно для каждого n , что $|X_n| \leq h$. Далее, ряд $|X_1| + |X_2 - X_1| + \dots + |X_n - X_{n-1}| + \dots$ сходится, поэтому и ряд $X_1 + (X_2 - X_1) + \dots + (X_n - X_{n-1}) + \dots$ сходится к элементу X^* , $|X^*| \leq h$, который является пределом X_n . Так как функция G непрерывна, то мы можем положить в уравнении (14.29) $n \rightarrow \infty$, после чего убедимся, что X^* удовлетворяет уравнению (14.35). Соотношения (14.38) и последующие показывают, что X^* является единственным решением, удовлетворяющим условию $|X| \leq h$. Ясно, что существует такое $k > 0$, что $|Y| \leq K$, коль скоро $|y - f(x_0)| \leq k$. Этим теорема доказана.

Нормированные линейные пространства

§ 14.10. Абстрактные пространства. В этом параграфе мы вводим сначала понятие „абстрактного пространства“, а затем „нормированного линейного пространства“. К числу подобных пространств относится большая часть интересующих нас функциональных пространств.

Вообще говоря, абстрактное пространство представляет собой совокупность объектов, между которыми существуют некоторые соотношения или заранее определенные правила операций. Эти объекты или „элементы“ могут быть любыми объектами, представляющими математический интерес. В частности, они могут быть функциями или вектор-функциями одного или многих переменных, которые удовлетворяют определенным требованиям, связанным с непрерывностью и дифференцируемостью. В этом случае мы имеем так называемое функциональное пространство. Объекты, из которых состоит пространство, рассматриваются как точки или, в случае линейного пространства, как „векторы“.

§ 14.11. Линейные пространства. Действительное линейное пространство является абстрактным пространством, в котором определены операции сложения векторов и умножения вектора на скаляр, удовлетворяющие условиям 1—7 для любых элементов x, y и z пространства и для любых действительных чисел c и d (в случае ком-

плексного линейного пространства c и d могут быть комплексными числами):

1. $x + y$ определено и принадлежит пространству.

2. $x + y = y + x$ и $x + (y + z) = (x + y) + z$. (14.40)

3. Вычитание всегда осуществимо и притом однозначно (т. е. для заданных x и y существует одно и только одно z , для которого $x + z = y$).

Условия 1—3 показывают, что существует единственный нулевой вектор, обозначаемый через 0 , для которого $x + 0 = x$ для каждого x .

4. cx определено и принадлежит пространству.

5. $cx + dx = (c + d)x$ и $c(dx) = (cd)x$.

6. $1x = x$ и $0x = 0$.

7. $c(x + y) = cx + cy$.

Ясно, что пространство функций одного переменного, непрерывных в заданном интервале, с обычным определением сложения и умножения на постоянную величину, удовлетворяет всем перечисленным условиям. Ясно также, что пространство всех таких функций, обладающих непрерывными производными до некоторого фиксированного порядка k , при тех же самых операциях также удовлетворяет этим условиям. Далее, в пространстве всех векторных функций $x = [x_1(s_1, \dots, s_n), \dots, x_n(s_1, \dots, s_n)]$, определенных и непрерывных вместе с их частными производными до некоторого фиксированного порядка k в некоторой заданной области пространства переменных $s_1, \dots, s_n$, как легко убедиться, выполняются те же условия.

Очевидно, что в каждом линейном пространстве могут быть образованы линейные комбинации $c_1x_1 + \dots + c_kx_k$, в которых $x_1, \dots, x_k$ — некоторые элементы пространства, а $c_1, \dots, c_k$ — некоторые действительные числа. Говорят, что вектор x *линейно зависит* от векторов $x_1, \dots, x_k$ тогда и только тогда, когда соотношение

$$x = c_1x_1 + \dots + c_kx_k$$

выполняется по крайней мере для одной совокупности значений $c_1, \dots, c_k$. Совокупность векторов $x_1, \dots, x_k$ называется *линейно независимой*, если никакой из векторов x_i не является линейно зависимым от остальных; если вектор x линейно зависит от элементов такой совокупности, то постоянные $c_1, \dots, c_k$ определены однозначно.

Часть M пространства (которая может быть и всем пространством) называется *линейным многообразием*, если каждая конечная линейная комбинация элементов M есть также элемент M . Если каждый элемент такой части пространства M линейно зависит от некоторого фиксированного конечного числа элементов M , то говорят, что M имеет *конечное число измерений*; в противном случае говорят, что M имеет *бесконечное число измерений*. Все интересующие нас функциональные пространства имеют бесконечное число измерений.

§ 14.12. Нормированные линейные пространства. При изучении сходимости метода Ньютона и его вариантов и при доказательстве теоремы о существовании обратного отображения существенно использовалось понятие длины вектора. Хотя понятие „длины“ функции на первый взгляд не имеет смысла, некоторая мера величины функции часто оказывается желательной и возможной. Такая мера величины вектора x в линейном пространстве называется его нормой (и обозначается через $\|x\|$), если:

1. Она определена для каждого x из пространства.
2. $\|x\| \geq 0$ для всех x , причем она равна 0 только для $x = 0$. (14.41)
3. $\|x + y\| \leq \|x\| + \|y\|$.
4. $\|cx\| = |c| \|x\|$.

Линейное пространство, в котором определена норма [удовлетворяющая условиям (14.41)], называется *нормированным линейным пространством*. Каждая вводимая в пространстве норма является существенной частью определения этого пространства. Два нормированных линейных пространства, состоящие из одних и тех же элементов, являются различными пространствами, если они имеют различные нормы. Два таких пространства могут весьма существенно отличаться друг от друга [см. примеры 14.2, 14.4 и § 14.13].

Приведем несколько примеров нормированных линейных пространств, имеющих важное значение (в каждом случае сложение и умножение на скаляр определены обычным путем).

Пример 14.1. Действительное гильбертово пространство. Его элементами x являются бесконечные последовательности $(x_1, x_2, \dots, x_n, \dots)$ действительных чисел, для которых ряд $x_1^2 + x_2^2 + \dots + x_n^2 + \dots$ сходится. Норма определена соотношением

$$\|x\| = \left[\sum_{n=1}^{\infty} x_n^2 \right]^{1/2}.$$

Пример 14.2. Пространство C . Элементами x этого пространства являются функции или векторные функции $x(s_1, \dots, s_n)$ одного или многих (но конечного числа) переменных, непрерывных в некоторой фиксированной ограниченной замкнутой области D пространства $(s_1, \dots, s_n)$. При этом

$$\|x\| = \max |x(s_1, \dots, s_n)|.$$

Пример 14.3. Пространство $C^{(k)}$. Элементами этого пространства являются все те входящие в определенное выше пространство C элементы, которые имеют, кроме того, непрерывные (частные) производные до порядка k ; норма $\|x\|$ есть максимум из всех максимумов производных порядков $\leq k$ в области D (производной нулевого порядка является сама функция).

Пример 14.4. Пространство, элементами которого являются элементы пространства C , но норма определена следующим образом:

$$\|x\| = \left[\int \dots \int_D |x(s_1, \dots, s_v)|^2 ds_1 \dots ds_v \right]^{1/2}.$$

Естественно определять расстояние между точками (т. е. между концами двух векторов, исходящими из одной точки) x и y как норму $\|x - y\|$ разности этих векторов. Это в свою очередь позволяет определить *сходимость бесконечной последовательности векторов* x_n к предельному вектору x_0 ; сходимость имеет место тогда и только тогда, когда $\lim_{n \rightarrow \infty} \|x_n - x_0\| = 0$. А это уже позволяет определить *сходимость бесконечного ряда* $x_1 + x_2 + \dots + x_n + \dots$ к предельному вектору x_0 ; сходимость эта имеет место тогда и только тогда, когда последовательность $y_1, y_2, \dots, y_n, \dots$ частичных сумм ряда сходится к x_0 (здесь, как обычно, $y_n = x_1 + \dots + x_n$).

§ 14.13. Полнота. В предыдущих рассуждениях мы широко пользовались положением, что если $x_1, x_2, \dots, x_n$ есть бесконечная последовательность векторов (N фиксировано), для которой ряд $|x_1| + |x_2| + \dots + |x_n| + \dots$ сходится, то ряд $x_1 + x_2 + \dots + x_n + \dots$ сходится к некоторому вектору x_0 . Это положение имеет место вследствие того, что для каждого j от 1 до N j -е компоненты удовлетворяют неравенству $|x_{nj}| \leq |x_n|$ и образуют абсолютно сходящийся ряд чисел, который поэтому сходится в обычном смысле. Каждое нормированное линейное пространство, в котором сходимость рядов норм $\|x_1\| + \|x_2\| + \dots + \|x_n\| + \dots$ влечет сходимость рядов векторов $x_1 + x_2 + \dots + x_n + \dots$, называется *полным*. Известно, что пространства в примерах 14.1—14.3 (§ 14.12) являются полными, тогда как пространство в примере 14.4 — неполное.

В пространстве примера 14.2 ряд функций x_n сходится к x_0 (в смысле указанной нормы) тогда и только тогда, когда он сходится равномерно.

В пространстве примера 14.3 сходимость является равномерной сходимостью, имеющей место по отношению ко всем производным до порядка k . В примере 14.2, если x_n есть скалярная функция $x_n(s_1, \dots, s_v)$, то $|x_n(s_1, \dots, s_v)| \leq \|x_n\|$ для каждого $(s_1, \dots, s_v)$, так что сходимость ряда норм влечет за собой абсолютную сходимость этого ряда для каждого $(s_1, \dots, s_v)$ в области D к некоторой предельной функции $x_0(s_1, \dots, s_v)$. Более того, для каждого $(s_1, \dots, s_v)$

$$|x_n(s_1, \dots, s_v) - x_0(s_1, \dots, s_v)| \leq \sum_{j=n+1}^{\infty} \|x_j\| = \varepsilon_n, \quad \lim_{n \rightarrow \infty} \varepsilon_n = 0.$$

(14.42)

Так как ϵ_n не зависит от $(s_1, \dots, s_j)$, то сходимость является равномерной, так что x_0 есть непрерывная функция, также принадлежащая этому пространству, а неравенство (14.42) показывает, что $\|x_n - x_0\| \rightarrow 0$.

Чтобы убедиться в неполноте пространства примера 14.4, примем $v=1$, в качестве D возьмем интервал $0 \leq s \leq 1$ и построим последовательность $x_1(s), x_2(s), \dots, x_n(s), \dots$ непрерывных функций, такую, что ряд $\|x_1\| + \dots + \|x_n\| + \dots$ сходится, но ряд $x_1(s) + x_2(s) + \dots + x_n(s) + \dots$ не сходится ни к какой непрерывной функции.

Чтобы сделать это, для каждого n примем $x_n(s) = [X_n(s)]^{1/2}$, определив функцию $X_n(s)$ следующим образом: график $y = X_n(s)$, $0 \leq s \leq 1$ состоит из сторон каждого из $2^n - 1$ равнобедренных

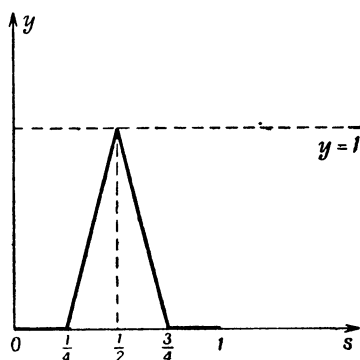


Рис. 14.1. График функции $y = X_1(s) = [x_1(s)]^2$.

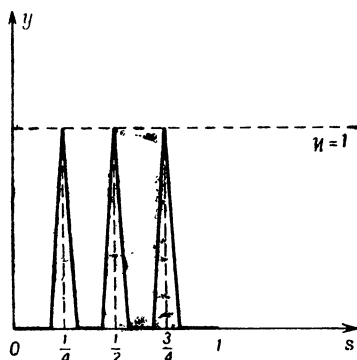


Рис. 14.2. График функции $y = X_2(s) = [x_2(s)]^2$.

треугольников (с основаниями, лежащими на оси s) и из тех частей отрезка $0 \leq s \leq 1$ оси s , которые не являются основаниями этих треугольников. Вершины этих треугольников лежат в точках $s = j/2^n$, $y = 1$, $j = 1, \dots, 2^n - 1$, и их основания настолько малы, что сумма их площадей составляет $\frac{1}{4^n}$. В этом случае

$$\|x_n\| = \left\{ \int_0^1 [x_n(s)]^2 ds \right\}^{1/2} = \left[\int_0^1 X_n(s) ds \right]^{1/2} = \frac{1}{2^n}.$$

Однако ряд $x_1(s) + x_2(s) + \dots + x_n(s) + \dots$ сходится к $+\infty$ для каждого s , равного $j/2^p$, $1 \leq j \leq 2^p - 1$, при любом p . Графики $y = X_n$ для $n=1$ и $n=2$ приведены на рис. 14.1 и 14.2.

Функциональные уравнения, вариационные задачи и нормированные линейные пространства

§ 14.14. Скалярные функции. Функционалы. Начнем с некоторых определений. Скалярная функция, определенная на нормированном линейном пространстве, называется обычно функционалом. Функционал $l(x)$ называется линейным тогда и только тогда, когда для всех $x_1, x_2, \dots, x_k$ и всех $c_1, c_2, \dots, c_k$

$$l(c_1 x_1 + c_2 x_2 + \dots + c_k x_k) = c_1 l(x_1) + \dots + c_k l(x_k)$$

и существует такое M , что для всех x выполняется неравенство

$$|l(x)| \leq M \|x\|.$$

Функционал $b(x, y)$ называется *билинейным*, если он линеен относительно x при каждом y и линеен относительно y при каждом x и если существует такое M , что $|b(x, y)| \leq M \|x\| \|y\|$ для всех x и y . Функционал $q(x)$ называется *квадратичным*, если существует такой билинейный функционал $b(x, y)$, что $q(x) = b(x, x)$; функционал $b(x, y)$ можно считать *симметричным*, т. е. удовлетворяющим условию $b(y, x) = b(x, y)$, так как в противном случае его можно заменить на $[b(x, y) + b(y, x)]/2$. Если $f(x)$ есть функционал, определенный вблизи x_0 , и $l(x_0, \xi)$ — линейный функционал относительно ξ , такой, что

$$f(x_0 + \xi) = f(x_0) + l(x_0, \xi) + g(x_0, \xi), \quad \lim_{\|\xi\| \rightarrow 0} \frac{g(x_0, \xi)}{\|\xi\|} = 0,$$

то f называют *дифференцируемым функционалом* при $x = x_0$. В этом случае

$$\lim_{\lambda \rightarrow 0} \left| \frac{f(x_0 + \lambda \xi) - f(x_0)}{\lambda} - l(x_0, \xi) \right| = 0$$

для каждого ξ , и $l(x_0, \xi)$ называют *дифференциалом* или *первой вариацией* функционала f при $x = x_0$. Если f — дифференцируемый функционал при каждом x , достаточно близком к x_0 , и если существуют такое $h > 0$ и такая стремящаяся к нулю вместе с r функция $\varepsilon_1(r)$, что

$$f(x + \xi) = f(x) + l(x, \xi) + g(x, \xi), \quad |g(x, \xi)| \leq \|\xi\| \varepsilon_1(\|\xi\|), \quad (14.43)$$

как только $\|x - x_0\| \leq h_1$, то f называется *непрерывно дифференцируемым функционалом* вблизи x_0 . Если функционал f непрерывно дифференцируем вблизи x_0 и если существуют функционал $q(x, \xi)$, квадратичный относительно ξ при каждом x , взятом вблизи x_0 , и такие $h > 0$ и функция $\varepsilon_2(r)$, стремящаяся к нулю вместе с r , что

$$f(x + \xi) = f(x) + l(x, \xi) + q(x, \xi) + g(x, \xi), \quad |g(x, \xi)| \leq \|\xi\|^2 \varepsilon_2(\|\xi\|),$$

как только $\|x - x_0\| \leq h$, то говорят, что f — функционал, дважды непрерывно дифференцируемый вблизи x_0 , а $q(x, \xi)$ часто называют второй вариацией функционала f в точке x . Если f — функционал, один или два раза непрерывно дифференцируемый вблизи x_0 , то получаем соответственно

$$l(x, \xi) = \left[\frac{d}{d\lambda} f(x + \lambda \xi) \right]_{\lambda=0}, \quad q(x, \xi) = \left[\frac{d^2}{d\lambda^2} f(x + \lambda \xi) \right]_{\lambda=0}. \quad (14.44)$$

§ 14.15. Простые вариационные задачи. Предположим, что $f(x, y, p)$ есть обычная скалярная функция трех переменных x, y и p класса C'' , т. е. непрерывная вместе со своими частными производными до второго порядка включительно для всех значений переменных. „Простейшая“ проблема вариационного исчисления (см. гл. 4) заключается в отыскании в классе C' такой функции $y = y(x)$, которая принимает заданные значения $y(a)$ и $y(b)$ в точках a и b и дает минимум (или максимум, или какое-либо иное стационарное значение) интегралу

$$I(y) = \int_a^b f[x, y(x), y'(x)] dx. \quad (14.45)$$

Очевидно, что интеграл $I(y)$ есть функционал в пространстве C' (на замкнутом интервале $[a, b]$). Если $n = \eta(x)$ — другая функция класса C' , то

$$I(y + \lambda n) = \int_a^b f[x, y(x) + \lambda \eta(x), y'(x) + \lambda \eta'(x)] dx. \quad (14.46)$$

Дифференцируя по λ при фиксированных y и n , получаем

$$\frac{d}{d\lambda} I(y + \lambda n) = \int_a^b \{f_y[*] \eta(x) + f_p[*] \eta'(x)\} dx, \quad (14.47)$$

$$\begin{aligned} \frac{d^2}{d\lambda^2} I(y + \lambda n) &= \\ &= \int_a^b \{f_{yy}[*] [\eta(x)]^2 + 2f_{yp}[*] \eta(x) \eta'(x) + f_{pp}[*] [\eta'(x)]^2\} dx, \end{aligned}$$

где аргументы $[*]$ такие же, что и в равенстве (14.46). Полагая в равенстве (14.47) $\lambda = 0$, мы видим из (14.44), что первая и вторая вариации $L(y, n)$ и $Q(y, n)$ функционала I имеют вид

$$L(y, n) = \int_a^b \{f_y[*] \eta(x) + f_p[*] \eta'(x)\} dx, \quad (14.48)$$

$$Q(y, n) = \int_a^b \{f_{yy}[*] [\eta(x)]^2 + 2f_{yp}[*] \eta(x) \eta'(x) + f_{pp}[*] [\eta'(x)]^2\} dx,$$

где $[\ast]$ означает те же аргументы, что и в равенстве (14.46) при $\lambda = 0$. Функция $\mathbf{n}$, входящая в равенства (14.48), часто обозначается через $\delta \mathbf{y}$.

О функции $\mathbf{y}$ говорят, что она дает интегралу $I(\mathbf{y})$ *стационарное значение* тогда и только тогда, когда $L(\mathbf{y}, \mathbf{n}) = 0$ для всех допустимых $\mathbf{n}$; если считаются допустимыми только те функции $\mathbf{y}$, которые имеют заданные значения в точках a и b , то все допустимые $\mathbf{n}$ должны равняться нулю в этих же точках для того, чтобы $\mathbf{y} + \lambda \mathbf{n}$ принимали на концах заданные значения. Конечно, понятие стационарного значения можно очевидным образом распространить на произвольный дифференцируемый функционал $f(\mathbf{x})$, потребовав, чтобы было $l(\mathbf{x}, \xi) = 0$ для всех ξ , и т. д. В этом можно убедиться на основании определения l для конечномерного случая, рассмотренного во второй и третьей частях настоящей главы, когда требование $l(\mathbf{x}, \xi) = 0$ соответствует хорошо известному условию, заключающемуся в исчезновении всех частных производных первого порядка.

Предположим теперь, что $\mathbf{y}$ дает стационарное значение интегралу $I(\mathbf{y})$ (14.45) и принадлежит также к классу C'' . Тогда член $f_p[\ast]$ в выражении $L(\mathbf{y}, \mathbf{n})$ будет дифференцируемым, так что член $f_p[\ast] \eta'(x)$ можно проинтегрировать по частям. При условии, что $\eta(a) = \eta(b) = 0$, это приводит к следующему выражению:

$$L(\mathbf{y}, \mathbf{n}) = \int_a^b \eta(x) \left\{ f_y[\ast] - \frac{d}{dx} f_p[\ast] \right\} dx. \quad (14.49)$$

Легко убедиться, что исчезновение выражения (14.49) для всех $\mathbf{n}$ из класса C' при условии, что $\eta(a) = \eta(b) = 0$, требует равенства нулю выражения, заключенного в скобки, т. е. приводит к *уравнению Эйлера*

$$\frac{d}{dx} f_p[\ast] = f_y[\ast]. \quad (14.50)$$

Выполнив дифференцирование в уравнении (14.50), получим

$$f_{pp}[\ast] y''(x) + f_{py}[\ast] y'(x) + f_{px}[\ast] = f_y[\ast]. \quad (14.51)$$

Если стационарное значение оказывается минимальным, то для каждого $\mathbf{n}$, удовлетворяющего условиям

$$\eta(a) = \eta(b) = 0,$$

должно быть

$$\frac{d^2 I(\mathbf{y} + \lambda \mathbf{n})}{d\lambda^2} \geq 0$$

при $\lambda = 0$ (исследование знаков вторых производных в элементарном анализе). Это как раз условие того, чтобы вторая вариация удовлетворяла неравенству $Q(\mathbf{y}, \mathbf{n}) \geq 0$ для всех указанных $\mathbf{n}$. Оно влечет

за собой неравенство $f_{pp}[*] \geq 0$ для всех $a \leq x \leq b$. Но, как видно из уравнения (14.51), обращение в нуль $f_{pp}[*]$ при некотором $x = x_0$ из отрезка $[a, b]$ приводит к некоторой сингулярности. Все это приводит к определению регулярной вариационной проблемы как проблемы, в которой $f_{pp}(x, y, p) > 0$ для всех значений (x, y, p) . Интересно отметить, что если f принадлежит к классу C'' и $f_{pp} > 0$, то каждая функция y из класса C' , которая дает стационарное значение интегралу I , принадлежит также к классу C'' и удовлетворяет уравнению Эйлера.

В качестве примера рассмотрим случай, когда

$$f(x, y, p) = e^y p^2, \quad I(y) = \int_a^b e^y(x) [y'(x)]^2 dx.$$

Ясно, что это регулярная вариационная проблема. Уравнение (14.51) принимает вид

$$2e^y y'' + 2e^y y'^2 = e^y y'^2$$

или

$$y'' = -\frac{1}{2} y'^2.$$

Отсюда получаем, что

$$y' = \frac{2}{x+c}, \quad y = 2 \ln(x+c) + d$$

(c, d — постоянные). Если положить $a=0$, $b=1$ и потребовать чтобы $y(0)=0$, $y(1)=2 \ln 2$, то решением будет функция

$$y = 2 \ln(1+x). \quad (14.52)$$

Подробное решение этой задачи приводится в примере (14.8) Приложения модифицированным методом Ритца, сущность которого изложена в следующем параграфе.

Естественно, что в вариационном исчислении часто встречаются более сложные интегралы. В качестве пространства, в котором отыскивается минимизирующая функция, может быть взято пространство $C^{(k)}$ векторов

$$z = [z_1(x_1, \dots, x_n), \dots, z_N(x_1, \dots, x_n)],$$

определенных в некоторой области пространства $(x_1, \dots, x_n)$. Функция f может зависеть от x, z и других переменных, являющихся частными производными от z_i до порядка k включительно. Если $k=1$, $v=2$, $N=1$ (и мы полагаем $x_1=x$, $x_2=y$), то $f=f(x, y, z, p, q)$ и

$$I(z) = \int_D \int f[x, y, z(x, y), z_x(x, y), z_y(x, y)] dx dy. \quad (14.53)$$

Уравнение Эйлера для функционала (14.53) имеет вид

$$\frac{\partial}{\partial x} f_p[*] + \frac{\partial}{\partial y} f_q[*] = f_z[*],$$

где $[*]$ обозначает те же аргументы, что и в уравнении (14.53).

Данное дифференциальное уравнение в частных производных или система таких уравнений могут оказаться уравнениями Эйлера для некоторой вариационной проблемы даже тогда, когда они получены на основании совсем других соображений. Например, уравнение Лапласа

$$z_{xx} + z_{yy} = 0$$

является уравнением Эйлера для вариационной задачи

$$I(z) = \iint_D (z_x^2 + z_y^2) dx dy, \quad f(x, y, z, p, q) = p^2 + q^2.$$

Кроме того, если $\rho(x, y, z)$ есть плотность, а $u(x, y, z)$, $v(x, y, z)$ и $w(x, y, z)$ — проекции скорости в установившемся безвихревом изэнтропическом сжимаемом потоке, то потенциал скорости V ($V_x = u$, $V_y = v$, $V_z = w$) удовлетворяет уравнению

$$\frac{\partial}{\partial x} [\Phi'(\omega) V_x] + \frac{\partial}{\partial y} [\Phi'(\omega) V_y] + \frac{\partial}{\partial z} [\Phi'(\omega) V_z] = 0,$$

где

$$\omega = V_x^2 + V_y^2 + V_z^2.$$

Это уравнение является уравнением Эйлера вариационной задачи, для которой

$$f = \frac{1}{2} \Phi(\omega),$$

где вид функции $\Phi(\omega)$ зависит от принятого соотношения между давлением p и плотностью ρ (энтропия считается постоянной). Если

$$p = A \rho^\gamma,$$

то

$$\Phi'(\omega) = \left[\rho_0^{\gamma-1} + \frac{\gamma-1}{2A\gamma} (\omega_0 - \omega) \right]^{1/(\gamma-1)},$$

где ρ_0 и ω_0 представляют собой значения величин ρ и ω в некоторой точке потока. Наконец, уравнение Эйлера для общего вида квадратической вариационной задачи (при $k=1$, $\nu=2$, $N=1$)

$$I(z) = \iint_D (a z_x^2 + 2b z_x z_y + c z_y^2 + 2d z z_x + 2e z z_y + f z^2 + \\ + 2g z_x + 2k z_y + 2l z) dx dy$$

оказывается линейным уравнением

$$\frac{\partial}{\partial x}(az_x + bz_y + dz + g) + \frac{\partial}{\partial y}(bz_x + cz_y + ez + k) + dz_x + ez_y = fz + l,$$

или

$$az_{xx} + 2bz_{xy} + cz_{yy} + (a_x + b_y)z_x + (b_x + c_y)z_y + (d_x + e_y - f)z = l - g_x - k_y. \quad (14.54)$$

Здесь коэффициенты могут быть функциями (x, y) и

$$f(x, y, z, p, q) =$$

$$= ap^2 + 2bpq + cq^2 + 2dpz + 2eqz + fz^2 + 2gp + 2kq + 2lz.$$

Уравнение (14.54) показывает, какие линейные дифференциальные уравнения второго порядка оказываются уравнениями Эйлера для вариационных задач. Кроме существующих методов интегрирования уравнений в частных производных, для решения таких уравнений применяются специальные методы вариационного исчисления. Для регулярных вариационных задач уравнения Эйлера оказываются всегда эллиптического типа.

§ 14.16. Векторные уравнения в нормированных линейных пространствах. Под *векторной функцией* или *оператором* $f(x)$ понимают закон, который каждому вектору x из определенной совокупности векторов некоторого нормированного линейного пространства ставит в соответствие единственный вектор $f(x)$ некоторого другого (возможно, того же самого) нормированного пространства. Оператор $I(x)$ является линейным тогда и только тогда, когда

$$I(c_1x_1 + \dots + c_kx_k) = c_1I(x_1) + \dots + c_kI(x_k)$$

и

$$\|I(x)\| \leq M\|x\|$$

для некоторого M и всех $x, x_1, \dots, x_k, c_1, \dots, c_k$ (нормы здесь те же, какие введены в обоих пространствах). Билинейные и квадратичные операторы могут быть определены аналогично тому, как были определены соответствующие функционалы в § 14.14. Ясно также, как перенести понятие непрерывной дифференцируемости на операторы. Любое функциональное уравнение (вместе с соответствующими граничными и начальными условиями) может быть интерпретировано как уравнение вида

$$f(x) = 0, \quad (14.55)$$

где f представляет собой подходящий оператор, и пространства векторов x и $f(x)$ выбраны надлежащим образом. Предположим, например, что нужно решить уравнение Лапласа (например, с двумя независимыми переменными) при заданных гладких граничных значениях

в некоторой области D с гладкой границей. Пространство, которое здесь будет использовано, является пространством функций $u(x, y)$ класса C'' в области D . Если мы рассматриваем уравнение (14.55) как систему уравнений $f_i(x) = 0$, то равенство $\Delta u \equiv u_{xx} + u_{yy} = 0$ можно считать одним из таких уравнений; остальные уравнения выражают тот факт, что u принимает заданные граничные значения. Это в некотором смысле неудобное обстоятельство можно обойти, если представить функцию u в виде суммы $u(x, y) = f(x, y) + U(x, y)$, в которой f есть некоторая функция, принимающая заданные граничные значения, а U — функция, исчезающая на границе и удовлетворяющая уравнению Пуассона $\Delta U + \Delta f = 0$. Часть C''_0 всего пространства C'' , состоящая из функций U , исчезающих на границе, является линейным многообразием в C'' и оказывается фактически полным нормированным линейным пространством. В этом пространстве $\mathbf{f}$ представляет собой дифференциальный оператор $\Delta U + \Delta f$. Если предположить, что U и f принадлежат классу C'' , то легко видеть, что вектор $\Delta U + \Delta f$ [$= \mathbf{f}(x)$] принадлежит лишь классу C . На этом примере видно, насколько важно рассматривать операторы, действующие между различными пространствами.

Рассмотрим теперь уравнения движения потока жидкости (без ударных волн)

$$\begin{aligned} \rho_t + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial y}(\rho v) + \frac{\partial}{\partial z}(\rho w) &= 0, \\ u_t + uu_x + vu_y + wu_z + \frac{p_x}{\rho} &= 0, \\ v_t + uv_x + vv_y + wv_z + \frac{p_y}{\rho} &= 0, \\ w_t + uw_x + vw_y + ww_z + \frac{p_z}{\rho} &= 0, \\ S_t &= 0, \quad p = f(\rho, S), \end{aligned} \tag{14.56}$$

где ρ — плотность, p — давление, S — энтропия, u , v и w — компоненты скорости и все эти величины являются функциями от (x, y, z, t) в некоторой области D пространства (x, y, z, t) . Здесь пространством векторов $\mathbf{x}$ является пространство векторных функций $[\rho(x, y, z, t), u(x, y, z, t), v(x, y, z, t), w(x, y, z, t), S(x, y, z, t)]$ класса C' в области D . „Значением“ оператора $\mathbf{f}(\mathbf{x})$ является вектор из класса C в области D , полученный в результате выполнения операций в левой части уравнений (14.56). Под нулевым вектором в уравнении (14.55) подразумевается нулевой вектор этого пространства. Конечно, если имеющиеся граничные условия неоднородны, то функции, удовлетворяющие им, нужно вычитать, а уравнения нужно изменить подобно тому, как показано в предыдущем примере.

Далее, в уравнении

$$\mathbf{f}(\mathbf{x} + \xi) = \mathbf{f}(\mathbf{x}) + \mathbf{l}(\mathbf{x}, \xi) + \mathbf{g}(\mathbf{x}, \xi), \tag{14.57}$$

представляющем собой обобщение уравнения (14.43) на случай операторов, имеем

$$I(x, \xi) = \left[\frac{d}{d\lambda} f(x + \lambda \xi) \right]_{\lambda=0} \quad (14.58)$$

или

$$\lim_{\lambda \rightarrow 0} \left\| \frac{f(x + \lambda \xi) - f(x)}{\lambda} - I(x, \xi) \right\| = 0.$$

Допустим, что x есть пространство векторов класса C' , как было предположено выше для уравнений движения потока, и пусть $f(x)$ — определенный там оператор. Требуется найти соответствующую $I(x, \xi)$. Допустим, что ξ — векторная функция $[\delta\rho, \delta u, \delta v, \delta w, \delta S]$ класса C' от (x, y, z, t) в области D . Тогда, согласно уравнению (14.58), мы найдем $I(x, \xi)$, заменив в выражении f вектор x на $x + \lambda \xi$, продифференцировав его по λ и положив $\lambda = 0$. В нашем случае заменяем ρ на $\rho + \lambda \delta\rho$, u на $u + \lambda \delta u$ и т. д., дифференцируем по λ , полагаем $\lambda = 0$ и получаем $I(x, \xi)$ как оператор в классе C , заданный своими пятью компонентами:

$$\begin{aligned} & \delta\rho_t + \frac{\partial}{\partial x} (\rho \delta u + u \delta\rho) + \frac{\partial}{\partial y} (\rho \delta v + v \delta\rho) + \frac{\partial}{\partial z} (\rho \delta w + w \delta\rho), \\ & \delta u_t + u \delta u_x + v \delta u_y + w \delta u_z + u_x \delta u + u_y \delta v + \\ & \quad + u_z \delta w + \rho^{-1} \frac{\partial}{\partial x} (f_p \delta\rho + f_S \delta S) - \frac{p_x \delta\rho}{\rho^2}, \\ & \delta v_t + u \delta v_x + v \delta v_y + w \delta v_z + v_x \delta u + v_y \delta v + \\ & \quad + v_z \delta w + \rho^{-1} \frac{\partial}{\partial y} (f_p \delta\rho + f_S \delta S) - \frac{p_y \delta\rho}{\rho^2}, \\ & \delta w_t + u \delta w_x + v \delta w_y + w \delta w_z + w_x \delta u + w_y \delta v + \\ & \quad + w_z \delta w + \rho^{-1} \frac{\partial}{\partial z} (f_p \delta\rho + f_S \delta S) - \frac{p_z \delta\rho}{\rho^2}, \\ & \delta S_t. \end{aligned}$$

Метод возмущений, метод Ньютона и его первая модификация, описанные в векторных обозначениях в § 14.7—14.9, могут быть без изменений перенесены на уравнения в полных нормированных линейных пространствах. Если пространство неполное, то рассматриваемый процесс может привести к последовательности $\xi_1, \dots, \xi_n, \dots$, для которой ряд $\|\xi_1\| + \dots + \|\xi_n\| + \dots$ оказывается сходящимся, а ряд $\xi_1 + \dots + \xi_n + \dots$ — расходящимся. Для ре-

шения уравнений потока методом возмущений нужно начинать с некоторого приближенного решения $(\rho_0, u_0, v_0, w_0, S_0)$ и находить

$$(\rho_1, u_1, v_1, w_1, S_1) = (\rho_0 + \delta\rho_1, u_0 + \delta u_1, \dots)$$

посредством решения уравнений

$$\begin{aligned} & \left[\rho_0 t + \frac{\partial}{\partial x} (\rho_0 u_0) + \frac{\partial}{\partial y} (\rho_0 v_0) + \frac{\partial}{\partial z} (\rho_0 w_0) \right] + \\ & + \left[\delta\rho_1 t + \frac{\partial}{\partial x} (\rho_0 \delta u_1 + u_0 \delta\rho_1) + \frac{\partial}{\partial y} (\rho_0 \delta v_1 + v_0 \delta\rho_1) + \right. \\ & \quad \left. + \frac{\partial}{\partial z} (\rho_0 \delta w_1 + w_0 \delta\rho_1) \right] = 0 \\ & \dots \end{aligned}$$

относительно $(\delta\rho_1, \delta u_1, \delta v_1, \delta w_1, \delta S_1)$; здесь точки поставлены вместо остальных уравнений, в каждом из которых левая часть является суммой двух прямых скобок: первая из них является значением левой части равенства (14.57) при $x_0 = (\rho_0, u_0, v_0, w_0, S_0)$, а вторая — значением левой части равенства (14.58) при x_0 и $\xi_1 = (\delta\rho_1, \delta u_1, \delta v_1, \delta w_1, \delta S_1)$. Метод Ньютона состоял бы в этом случае в повторении всего процесса, начиная с x_1 , в определении $x_2 = x_1 + \xi_2$ и т. д. При использовании же модифицированного метода Ньютона все осуществлялось бы аналогично предыдущему, за исключением того, что во второй из прямых скобок вместо $(\rho_n, u_n, v_n, w_n, S_n)$ всегда стояло бы $(\rho_0, u_0, v_0, w_0, S_0)$. В этом и состоит преимущество модифицированного метода Ньютона, так как уравнения для каждого ξ_{n+1} представляют собой систему линейных уравнений с не зависящими от n коэффициентами.

§ 14.17. Замечания о линейных уравнениях. Метод возмущений, метод Ньютона и его модификации сводят решение нелинейных систем к решению одной или нескольких линейных систем. Известно, однако, что системы линейных уравнений могут быть различной природы: обыкновенные дифференциальные уравнения, уравнения в частных производных с граничными (или начальными) условиями, интегральные или интегро-дифференциальные уравнения. В тех случаях, когда нельзя получить решение этих уравнений при помощи обычных методов, следует обратиться к численным методам; некоторые из таких методов рассмотрены в следующих параграфах настоящей главы.

Численные методы

§ 14.18. Метод Рэлея—Ритца и некоторые модификации вариационных задач. В § 14.2—14.6 изучались некоторые методы решения систем алгебраических уравнений; примеры, иллюстрирующие эти методы, приведены там же или в Приложении. Метод Рэлея—Ритца и его модификации позволяют получать приближенные решения функциональных уравнений путем сведения их к системам алгебраических уравнений.

Идея этих методов наиболее ясно раскрывается на примере вариационных задач. Возможность их применения основана на том факте, что почти во всех интересующих нас функциональных пространствах можно найти бесконечную последовательность конечномерных линейных многообразий $\mathcal{M}_i$, каждое из которых содержится в следующем многообразии, так что каждый элемент пространства либо входит в некоторое $\mathcal{M}_i$, либо является пределом последовательности x_k элементов, причем каждый x_k входит в $\mathcal{M}_k$. Например, каждая функция из $C^{(k)}$ в интервале $[a, b]$ может быть аппроксимирована с любой требуемой степенью точности (в соответствии с нормой, введенной в пространстве $C^{(k)}$) полиномами; очевидно, что все полиномы степени $\leq k$ образуют конечномерное линейное многообразие $\mathcal{M}_k$. Если $[a, b]$ есть интервал длины 2π , то каждая функция $f(x)$, определенная на $[a, b]$, принадлежащая классу $C^{(k)}$ и имеющая период 2π , может быть аппроксимирована на $[a, b]$ (по существу, на всей прямой) в классе $C^{(k)}$ конечными тригонометрическими полиномами, каждый из которых представляет собой линейную комбинацию функций

$$1, \cos nx, \sin nx, \quad n = 1, 2, \dots$$

Предположим теперь, что требуется найти стационарное значение некоторого функционала $f(x)$. Метод Рэля — Ритца состоит в выделении последовательности $\mathcal{M}_k$ конечномерных линейных многообразий пространства x , подчиняющихся указанным выше условиям, и в последующем нахождении стационарного значения $f(x)$ для всех x из $\mathcal{M}_k$ ($k = 1, 2, \dots$). Это приводит рассматриваемую задачу к задаче с конечным числом переменных. Кроме того, часто оказывается целесообразным на каждом k -м шаге заменять f приближенной функцией f_k .

Методы подобного типа оказываются весьма эффективными и часто применяются. В примере 14.8 Приложения этот метод применен для численного решения задачи об отыскании минимума функционала $I(y)$, где

$$I(y) = \int_0^1 e^y y'^2 dx, \quad y(0) = 0, \quad y(1) = 2 \ln 2;$$

точное решение этой задачи дается формулой (14.52) § 14.15. Здесь в качестве $\mathcal{M}_1$ мы выбираем класс всех многочленов третьей степени, обладающих тем свойством, что y и y' принимают заданные значения при $x = 0$ и $x = 1$. В качестве $\mathcal{M}_2$ берем множество функций класса C' на отрезке $[0, 1]$ с заданными значениями для y и y' при $x = 0, 1/2$ и 1 и кубических в каждом из двух интервалов и т. д. В качестве $\mathcal{M}_k$ берем множество всех функций класса C' на $[0, 1]$ с заданными значениями для y и y' при $x = j/2^{k-1}$, $j = 0, 1, \dots$, и кубических в каждом из 2^{k-1} малых интервалов. Для каждого k

величину $I(y)$ заменяем на $I_k(y)$, где $I_k(y)$ вычисляется в каждом интервале разбиения, соответствующего $\mathcal{M}_k$, по правилу Симпсона, причем необходимые для этого средние значения y и y' находятся по формулам

$$\bar{y} = \frac{1}{2} (y_1 + y_2) - \frac{h}{4} (y'_2 - y'_1),$$

$$\bar{y}' = \frac{3}{4h} (y_2 - y_1) - \frac{1}{4} (y'_2 - y'_1),$$

выражающим их через значения y_1 и y'_1 на левом и через значения y_2 и y'_2 на правом конце интервала длины h . Эти формулы являются точными для многочлена третьей степени. Так как значения y заданы при $x=0$ и $x=1$, то функции из $\mathcal{M}_1$ полностью определяются значениями их производных y' при $x=0$, $x=1$. Эти значения, умноженные на постоянные множители, были обозначены через η_0 и η_2 и приняты в качестве независимых переменных в первом приближении. Затем модифицированным методом Ньютона решались уравнения

$$\frac{\partial I_1}{\partial \eta_0} = 0, \quad \frac{\partial I_1}{\partial \eta_2} = 0.$$

Промежуточные значения y и y' , которые получались в процессе этих вычислений, принимались в качестве исходных данных для второго приближения с использованием I_2 . Так как значения y были даны при $x=0$ и $x=1$, то функции из $\mathcal{M}_2$ полностью определены значениями y при $x=1/2$ и значениями y' при $x=0$, $1/2$, 1 . Произведения этих величин на постоянные множители были приняты в качестве независимых переменных. Уравнения $\Delta I_2 = 0$ были решены сначала с помощью первой модификации метода Ньютона, а затем методом быстреего спуска. Первый из этих методов требует предварительной алгебраической работы большого объема и обращения одной матрицы, тогда как второй трудоемок в процессе вычислений, однако требует гораздо меньшей предварительной работы. Вычисления по второму методу в приложении не приведены.

§ 14.19. Замечания о методе Рэлея — Ритца и методах конечных разностей для решения функциональных уравнений. Метод Рэлея — Ритца часто применяется к функциональным уравнениям, которые не вытекают из вариационных задач. За некоторым исключением, в этих случаях необходимо заменить данное уравнение $\mathbf{f}(\mathbf{x}) = 0$ приближенным уравнением $\mathbf{f}_k(\mathbf{x}) = 0$, решения которого принадлежат многообразию $\mathcal{M}_k$.

В некотором смысле иным является метод конечных разностей. Этот метод широко и успешно используется для решения дифференциальных уравнений в частных производных гиперболического, параболического, а также эллиптического типов. В этом методе область, в которой решается уравнение, обычно покрывается сеткой из конечного

числа точек (узлов), а входящие в уравнение производные заменяются соответствующими разностями значений функций в узлах сетки. Так, например, если u_{ij} является значением функции $u(x, y)$ в узле сетки, находящемся в i -й строке и j -м столбце, то частные производные могут быть заменены разностями

$$u_{xx} \sim \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2}, \quad u_x \sim \frac{u_{i+1,j} - u_{i-1,j}}{2h},$$

$$u_{xy} \sim \frac{u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}}{4h^2}.$$

Соответствующие формулы имеют место в том же узле (i, j) для частных производных u_{yy} и u_y . В некоторых задачах для обеспечения лучшей сходимости были найдены другие формулы замены частных производных частными разностями.

§ 14.20. Приложение.

ПРИМЕР 14.5. Решение модифицированным методом Ньютона (см. § 14.2) уравнений (14.9)

$$f(x, y) \equiv 4 + x + y - x^2 + 2xy + 3y^2 = 0,$$

$$g(x, y) \equiv 1 + 2x - 3y + x^2 + xy - 2y^2 = 0.$$

Первый шаг. Отделение корней. Значения $g(x, y)$, соответствующие целым значениям x, y в подходящем интервале, содержащем искомые корни, даны в табл. 14.1.

Таблица 14.1

ЗНАЧЕНИЯ ФУНКЦИИ $g(x, y)$.

$y \backslash x$	-3	-2	-1	0	1	2	3	4
-4	-4							-11
-3	4	-2	-6	-8	-8	-6	-2	4
-2	8	3	0	-1	0	3	8	15
-1	8	4	2	2	4	8	14	22
0	4	1	0	1	4	9	16	25
1	-4	-6	-6	-4	0	6	14	24
2					-8	-1	8	19
3							-2	10
4								-3

Полагаем, что $\varphi_1(x)$ и $\varphi_2(x)$ представляют собой соответственно наименьшее и наибольшее из тех значений y , которые являются корнями уравнения $g(x, y) = 0$ и получены линейным интерполированием из значений, приведенных в табл. 14.1. Находим соответствующие значения $f[x, \varphi_1(x)]$ и $f[x, \varphi_2(x)]$; они приведены в табл. 14.2 и 14.3.

Таблица 14.2

ЗНАЧЕНИЯ ФУНКЦИИ

 $f[x, \varphi_1(x)]$

x	$\varphi_1(x)$	$f[x, \varphi_1(x)]$
1	-2,0	10,00
2	-2,3	6,37
3	-2,8	1,92
4	-3,3	-5,03

Таблица 14.3

ЗНАЧЕНИЯ ФУНКЦИИ

 $f[x, \varphi_2(x)]$

x	$\varphi_2(x)$	$f[x, \varphi_2(x)]$
-3	0,5	-9,75
-2	0,1	-2,27
-1	0	2,00
0	0,2	4,32

Табл. 14.2 показывает, что решения уравнений (14.9) локализуются примерно около значений $x = 3,3$, $y = \varphi_1(3,3) = -3,0$, тогда как табл. 14.1 указывает на локализацию корней вблизи значений $x = -1,5$, $y = 0$.

Второй шаг. Уточнение приближения. Для уточнения первого из предыдущих решений принимаем, что

$$x_0 = 3,3, \quad y_0 = -3,0.$$

Если пользоваться первой модификацией метода Ньютона, то нужно решить уравнения

$$\left. \begin{aligned} f(x_n, y_n) + f_x(x_0, y_0)\xi_{n+1} + f_y(x_0, y_0)\eta_{n+1} &= 0, \\ g(x_n, y_n) + g_x(x_0, y_0)\xi_{n+1} + g_y(x_0, y_0)\eta_{n+1} &= 0, \end{aligned} \right\} \quad (14.59)$$

$$\left. \begin{aligned} x_{n+1} &= x_n + \xi_{n+1}, \\ y_{n+1} &= y_n + \eta_{n+1}. \end{aligned} \right\}$$

Для (x_0, y_0) мы имеем

$$f_x = -11,6, \quad f_y = -10,4, \quad g_x = 5,6, \quad g_y = 12,3. \quad (14.60)$$

Подставляя (14.60) в (14.59) и решая, получаем приближенные значения

$$\left. \begin{aligned} \xi_{n+1} &= 0,14566 f_n + 0,12316 g_n, \\ \eta_{n+1} &= -0,06632 f_n - 0,13738 g_n. \end{aligned} \right\} \quad (14.61)$$

Вычисления последовательных значений x_n , y_n , ξ_{n+1} , η_{n+1} проведены в табл. 14.4. Необходимо отметить, что при этих вычислениях важно точно определять значения $f(x_n, y_n)$ и $g(x_n, y_n)$, тогда как для коэффициентов уравнения (14.61) не требуется высокой точности.

Таблица 14.4

ПОСЛЕДОВАТЕЛЬНЫЕ ПРИБЛИЖЕНИЯ К ПЕРВОМУ РЕШЕНИЮ УРАВНЕНИЙ (14.9)

n	x_n	y_n	x_n^2	$x_n y_n$	y_n^2	$f(x_n, y_n)$	$g(x_n, y_n)$	ξ_{n+1}	η_{n+1}
0	3,30	—3,00	10,89	—9,90	9,00	0,61	—0,41	0,0384	0,0159
1	3,338400	—2,984100	11,144915	—9,962119	8,904853	—0,000294	0,002190	0,000227	—0,000281
2	3,338627	—2,984381	11,146430	—9,963735	8,906530	—0,000064	0,000032	—0,000005	0,000000
3	3,338622	—2,984381	11,146391	—9,963720	8,906530	0	—0,000002	0	0

Из рассмотрения входов табл. 14.4 видно, что первое из двух решений уравнений (14.9), если ограничиться пятью десятичными знаками, будет таково:

$$x^* = 3,33862, \quad y^* = -2,98438.$$

Второе решение уточняется в примере 14.7 другим методом.

ПРИМЕР 14.6. Решение дифференциального уравнения

$$y' = y, \quad y(0) = 1 \quad (14.62)$$

методом численного интегрирования (см. § 14.4). Формула численного интегрирования, которую мы здесь применим, имеет вид

$$\int_{x_{n-1}}^{x_n} \varphi(x) dx = \frac{x_n - x_{n-1}}{2} \left[\varphi(x_{n-1}) + \varphi(x_n) - \frac{1}{6} \Delta^2 \varphi(x_n) - \frac{1}{12} \Delta^3 \varphi(x_n) + \dots \right].$$

Значения x и соответствующие значения y , y' , равно как и их последовательные разности, даны в табл. 14.5.

Точным решением является $y = e^x$. При $x = 0,10$

$$e^x = 1,1051709,$$

если ограничиться семью десятичными знаками.

ПРИМЕР 14.7. Решение уравнений (14.9) методом непрерывности (см. § 14.5).

Уравнения

$$\begin{aligned} f(x, y, \lambda) &\equiv (4 + x + y) + \lambda(-x^2 + 2xy + 3y^2) = 0, \\ g(x, y, \lambda) &\equiv (1 + 2x - 3y) + \lambda(x^2 + xy - 2y^2) = 0 \end{aligned} \quad (14.63)$$

Таблица 14.5

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ УРАВНЕНИЙ (14.62)

x	y'	y' (разности)	y	y (разности)
0	1,000000		1,000000	
0,0025	1,002503	2,503	1,0025031	25,031
0,0050	1,005012	2,509 6	1,0050125	25,094 63
0,0075	1,007528	2,516 7	1,0075282	25,157 63
0,0100	1,010050	2,522 6	1,0100502	25,220 63
		5,038 26		50,377 252
0,015	1,015113	5,063 25	1,0151131	50,629 252
0,020	1,020201	5,088 25	1,0202014	50,883 254 2
		10,151 101		101,512 1,010
0,03	1,030455	10,254 103 2	1,0304546	102,532 1,020 10
0,04	1,040811	10,356 102 —1	1,0408108	103,562 1,030 10
0,05	1,051271	10,460 104 2	1,0512711	104,603 1,041 11
0,06	1,061836	10,565 105 1	1,0618365	105,654 1,051 10
0,07	1,072508	10,672 107 2	1,0725081	106,716 1,062 11
0,08	1,083287	10,779 107 0	1,0832870	107,789 1,073 11
0,09	1,094174	10,887 108 1	1,0941742	108,872 1,083 10
0,10	1,105171	10,997 110 2	1,1051708	109,966 1,094 11

сводятся к уравнениям (14.9) при $\lambda = 1$; однако при $\lambda = 0$ уравнения (14.63) являются линейными и имеют единственное решение

$$x = -2,60, \quad y = -1,40.$$

Решая при помощи численного интегрирования (см. § 14.4) систему обыкновенных дифференциальных уравнений

$$f_x x' + f_y y' = -f_\lambda, \quad g_x x' + g_y y' = -g_\lambda, \quad (14.64)$$

находим решение $x = x(\lambda)$, $y = y(\lambda)$ уравнений (14.63) для $0 \leq \lambda \leq 1$. Из уравнений (14.64) имеем

$$x' = \frac{f_y g_\lambda - f_\lambda g_y}{f_x g_y - f_y g_x}, \quad y' = \frac{f_\lambda g_x - f_x g_\lambda}{f_x g_y - f_y g_x}, \quad (14.65)$$

где значения

$$\begin{aligned} f_x &= 1 - 2\lambda x + 2\lambda y, & g_x &= 2 + 2\lambda x + \lambda y, \\ f_y &= 1 + 2\lambda x + 6\lambda y, & g_y &= -3 + \lambda x - 4\lambda y, \\ f_\lambda &= -x^2 + 2xy + 3y^2, & g_\lambda &= x^2 + xy - 2y^2 \end{aligned}$$

получены из уравнений (14.63). Грубые результаты численного интегрирования приведены в табл. 14.6. Вспомогательные таблицы значений x^2 , xy , y^2 , f_x , f_y , f_λ , g_x , g_y , g_λ и $f_x g_y - f_y g_x$ не приводятся.

Таблица 14.6

ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ УРАВНЕНИЙ (14.65)

λ	x'	x' (разности)		x	x (разности)		y'	y' (разности)		x	y (разности)
0	-5,1			-2,60			-1,1			-1,40	
0,025	-5,4	-3		-2,73	-13		-0,9	2		-1,42	-2
0,050	-5,4	0	3	-2,87	-14	-1	-0,2	7	5	-1,44	-2 0
0,075	-4,5	9	9 6	-3,00	-13	1	1,0	12	5	-1,43	1 3
0,100	-2,7	18	9 0	-3,09	-9	4	2,4	14	2	-1,39	4 3
0,125	-0,6	21	3 -6	-3,13	-4	5	3,6	12	-2	-1,31	8 4
0,150	1,4	20	-1 -4	-3,12	1	5	4,3	7	-5	-1,21	10 2
		41	14 -16		-3	19 14		17	-9 -26		18 13 4
0,200	3,3	19	-22 -36	-2,99	13	16 -3	4,3	0	-19 -8	-0,99	22 4 -9
0,250	3,5	2	-17 5	-2,81	18	5 -11	3,6	-7	-7 12	-0,82	17 -5 -1
0,300	3,3	-2	-4 13	-2,64	17	-1 -6	2,8	-8	-1 6	-0,66	16 -1
0,350	2,9	-4	-2 2	-2,48	16	-1 0	2,3	-5	3 4	-0,53	13 -3
0,400	2,5	-4	0 2	-2,34	14	-2 -1	1,9	-4	-1 -2	-0,43	10 -3
0,45	2,2	-3	1 1	-2,22	12	-20	1,5	-4	0 -1	-0,35	8 -2
0,50	2,0	-2	1	-2,12	10	-2	1,3	-2	2	-0,28	7
		-5	3 11		22	-8 -3		-6	3 -3		15 -8 2
0,6	1,5	-5	0 -3	-1,94	18	-4 4	0,9	-4	2	-0,17	11 -4 4
0,7	1,2	-3	2 2	-1,81	13	-5 -1	0,7	-2	2	-0,09	8 -3
0,8	1,0	-2	1	-1,70	11		0,5	-2		-0,03	6
0,9	0,9	-1		-1,61	9		0,4	-1		0,01	4
1,0	0,7			-1,53	8		0,3	-1		0,04	4

Теперь с помощью первой модификации метода Ньютона может быть уточнено приближенное решение

$$x = -1,53, \quad y = 0,04$$

уравнений (14.9), соответствующее значению $\lambda = 1$. Уравнения, которые для этого должны быть решены, таковы:

$$f_x \Delta x + f_y \Delta y + f = 0, \quad g_x \Delta x + g_y \Delta y + g = 0.$$

При $\lambda = 1$ и коэффициентах, взятых с достаточной точностью, получаем:

$$4,14 \Delta x - 1,82 \Delta y = -f, \quad -1,02 \Delta x - 4,69 \Delta y = -g.$$

Отсюда

$$\Delta x = -0,2205 f + 0,0856 g, \quad \Delta y = 0,0480 f + 0,1946 g.$$

Вычисление последовательных приближений x_n, y_n показаны в табл. 14.7.

Таблица 14.7

ПОСЛЕДОВАТЕЛЬНЫЕ ПРИБЛИЖЕНИЯ КО ВТОРОМУ РЕШЕНИЮ УРАВНЕНИЙ (14.9)

n	x_n	y_n	x_n^2	$x_n y_n$	y_n^2	$f(x_n, y_n)$	$g(x_n, y_n)$	$\Delta x = \xi_{n-1}$	$\Delta y = \eta_{n-1}$
0	-1,53000	0,04000	2,34090	-0,06120	0,00160	0,05150	0,09650	-0,00310	0,02125
1	-1,53310	0,06125	2,35040	-0,09390	0,00375	0,00120	-0,00095	-0,00035	-0,00013
2	-1,53345	0,06112	2,35147	-0,09372	0,00374	-0,00002	0,00001	0,00001	0

Таким образом, решением уравнений (14.9) с пятью десятичными знаками будет

$$x^* = -1,53344, \quad y^* = 0,06112.$$

Это является вторым из решений, выделенных выше в примере 14.5.

Пример 14.8. Решение с помощью модифицированного метода Рунца (см. § 14.15) задачи

$$I(y) = \int_0^1 e^{y(x)} [y'(x)]^2 dx = \min, \quad y(0) = 0, \quad y(1) = 2 \ln 2.$$

Первое приближение. Полагаем $x_0 = 0, \quad x_1 = 0,5, \quad x_2 = 1,0, \quad h_1 = 0,5, \quad y_0 = 0, \quad y_2 = 2 \ln 2, \quad \eta_i = h_1 y_i, \quad i = 0, 1, 2$ и

$$I_1(y) = \frac{h_1}{3} [y_0'^2 + 4e^{y_1} y_1'^2 + 4y_2'^2] = \frac{1}{3h_1} [\eta_0^2 + 4e^{y_1} \eta_1^2 + 4\eta_2^2],$$

$$y_1 = \frac{1}{2}(y_0 + y_2) - \frac{1}{4}(\eta_2 - \eta_0) = 0,69315 + 0,25\eta_0 - 0,25\eta_2,$$

$$\eta_1 = \frac{3}{4}(y_2 - y_0) - \frac{1}{4}(\eta_2 + \eta_0) = 1,039725 - 0,25\eta_0 - 0,25\eta_2.$$

Принимаем η_0 и η_2 в качестве независимых переменных и решаем уравнения

$$f_0 = 3h_1 \frac{\partial I_1}{\partial \eta_0} = 2\eta_0 + e^{y_1} \eta_1^2 - 2e^{y_1} \eta_1 = 2\eta_0 - (2 - \eta_1) \eta_1 e^{y_1} = 0,$$

$$f_2 = 3h_1 \frac{\partial I_1}{\partial \eta_2} = 8\eta_2 - e^{y_1} \eta_1^2 - 2e^{y_1} \eta_1 = 8\eta_2 - (2 + \eta_1) \eta_1 e^{y_1} = 0$$

с помощью первой модификации метода Ньютона. Полагая $f_{00} = \partial f_0 / \partial \eta_0$ и т. д. и принимая за начальные значения $\eta_0 = \eta_2 = 0,693$, $y_1 = 0,693$, $e^{y_1} = 2$, получаем

$$\eta_1 = 0,693, \quad e^{y_1} \eta_1 = 1,386, \quad e^{y_1} \eta_1^2 = 0,960,$$

$$f_{00} = 2 + 0,50e^{y_1} - \eta_1 e^{y_1} + 0,25\eta_1^2 e^{y_1} = 1,854,$$

$$f_{02} = 0,50e^{y_1} - 0,25\eta_1^2 e^{y_1} = f_{20} = 0,760,$$

$$f_{22} = 8 + 0,50e^{y_1} + \eta_1 e^{y_1} + 0,25\eta_1^2 e^{y_1} = 10,626.$$

Мы должны решить уравнения

$$1,854 \Delta \eta_0 + 0,760 \Delta \eta_2 = -f_0,$$

$$0,760 \Delta \eta_0 + 10,626 \Delta \eta_2 = -f_2,$$

или

$$\Delta \eta_0 = -0,556f_0 + 0,040f_2,$$

$$\Delta \eta_2 = 0,040f_0 - 0,097f_2. \quad (14.66)$$

Решение этих уравнений выполнено в табл. 14.8.

Последовательные приближения к η_0, η_2 Таблица 14.8

n	η_0	η_2	y_1	η_1	e^{y_1}	$\eta_1 e^{y_1}$	f_0	f_2	$\Delta \eta_0$	$\Delta \eta_2$
0	0,693	0,693	0,693	0,693	2,000	1,386	-0,426	1,812	0,309	-0,193
1	1,002	0,500	0,819	0,664	2,268	1,506	-0,008	-0,012	0,004	0,001
2	1,006	0,501	0,819	0,663	2,268	1,504	0,001	0,003		

Таким образом, в качестве окончательного решения получаем

$$\eta_0 = 1,006, \quad \eta_1 = 0,663, \quad \eta_2 = 0,501, \quad y_1 = 0,819.$$

Значения, соответствующие истинному решению (см. § 14.15), следующие:

$$\eta_0 = 1,000, \quad \eta_1 = 0,667, \quad \eta_2 = 0,500, \quad y_1 = 0,811.$$

Второе приближение. Полагаем $h_2 = 0,25$, $x_i = 0,25i$, $i = 0, \dots, 4$, $y_0 = 0$, $y_4 = 2 \ln 2$, $\eta_i = h_2 y_i$, $i = 0, \dots, 4$ и

$$\begin{aligned} I_2(y) &= \frac{h_2}{3} [y_0'^2 + 4e^{y_1} y_1'^2 + 2e^{y_2} y_2'^2 + 4e^{y_3} y_3'^2 + 4y_4'^2] = \\ &= \frac{1}{3h_2} [\eta_0^2 + 4e^{y_1} \eta_1^2 + 2e^{y_2} \eta_2^2 + 4e^{y_3} \eta_3^2 + 4\eta_4^2]. \end{aligned}$$

Принимая в качестве независимых переменных η_0 , η_2 , η_4 и y_2 , получаем

$$y_1 = 0,50y_2 + 0,25\eta_0 - 0,25\eta_2,$$

$$y_3 = 0,69315 + 0,50y_2 + 0,25\eta_2 - 0,25\eta_4,$$

$$\eta_1 = 0,75y_2 - 0,25\eta_0 - 0,25\eta_2,$$

$$\eta_3 = 1,039725 - 0,75y_2 - 0,25\eta_2 - 0,25\eta_4$$

С помощью первой модификации метода Ньютона решаем уравнения

$$f_0 = 3h_2 \frac{\partial I_2}{\partial \eta_0} = 2\eta_0 - (2 - \eta_1) \eta_1 e^{y_1} = 0,$$

$$f_2 = 3h_2 \frac{\partial I_2}{\partial \eta_2} = 4\eta_2 e^{y_2} - (2 + \eta_1) \eta_1 e^{y_1} - (2 - \eta_3) \eta_3 e^{y_3} = 0,$$

$$f_4 = 3h_2 \frac{\partial I_2}{\partial \eta_4} = 8\eta_4 - (2 + \eta_3) \eta_3 e^{y_3} = 0,$$

$$g = 3h_2 \frac{\partial I_2}{\partial y_2} = 2e^{y_2} \eta_2^2 + (6 + 2\eta_1) \eta_1 e^{y_1} - (6 - 2\eta_3) \eta_3 e^{y_3} = 0.$$

Начальными значениями являются

$$\eta_0 = 0,50300, \quad \eta_2 = 0,33150, \quad \eta_4 = 0,25050, \quad y_2 = 0,81900,$$

$$y_1 = 0,45238, \quad y_3 = 1,12290, \quad \eta_1 = 0,40562, \quad \eta_3 = 0,27998,$$

$$e^{y_1} = 1,57205, \quad e^{y_2} = 2,26823, \quad 2e^{y_2} = 4,53646, \quad e^{y_3} = 3,07376,$$

$$\eta_1 e^{y_1} = 0,63765, \quad \eta_2 (2e^{y_2}) = 1,50384, \quad \eta_3 e^{y_3} = 0,86059,$$

$$\eta_1^2 e^{y_1} = 0,25864, \quad \eta_2^2 (2e^{y_2}) = 0,49852, \quad \eta_3^2 e^{y_3} = 0,24095.$$

Принимая f_{00} и т. д. такими, как и прежде, а $f_{i*} = \partial f_i / \partial y_2$ и т. д., получаем

$$f_{00} = 2 + 0,50e^{y_1} - \eta_1 e^{y_1} + 0,25\eta_1^2 e^{y_1} = 2,21304,$$

$$f_{02} = f_{20} = 0,50e^{y_1} - 0,25\eta_1^2 e^{y_1} = 0,72136,$$

$$f_{04} = f_{40} = 0,$$

$$f_{0*} = g_0 = -(1,50e^{y_1} - 0,50\eta_1 e^{y_1} - 0,50\eta_1^2 e^{y_1}) = -1,90993,$$

$$f_{22} = 4e^{y_2} + 0,50e^{y_1} + \eta_1 e^{y_1} + 0,25\eta_1^2 e^{y_1} + 0,50e^{y_3} - \eta_3 e^{y_3} + \\ + 0,25\eta_3^2 e^{y_3} = 11,29778,$$

$$f_{24} = f_{42} = 0,50e^{y_2} - 0,25\eta_3^2 e^{y_3} = 1,47664,$$

$$f_{2*} = g_2 = 4\eta_2 e^{y_2} + 1,50e^{y_2} - 2,50\eta_3 e^{y_2} + 0,50\eta_3 e^{y_3} - 1,50e^{y_1} - \\ - 2,50\eta_1 e^{y_1} - 0,50\eta_1^2 e^{y_1} = 1,50580,$$

$$f_{44} = 8 + 0,50e^{y_2} + \eta_3 e^{y_3} + 0,25\eta_3^2 e^{y_3} = 10,45771,$$

$$f_{4*} = g_4 = 1,50e^{y_2} + 0,50\eta_3 e^{y_3} - 0,50\eta_3^2 e^{y_3} = 4,92046,$$

$$g = 2\eta_2^2 e^{y_2} + 4,50e^{y_1} + 6\eta_1 e^{y_1} + \eta_1^2 e^{y_1} + 4,50e^{y_2} - 6\eta_3 e^{y_3} + \\ + \eta_3^2 e^{y_3} = 20,56662.$$

Таким образом, необходимо решить следующую систему уравнений:

$$2,21304 \Delta \eta_0 + 0,72136 \Delta \eta_2 - 1,90993 \Delta y_2 = -f_0 \quad (0,0928655),$$

$$0,72136 \Delta \eta_0 + 11,29778 \Delta \eta_2 + 1,47664 \Delta \eta_4 + 1,50580 \Delta y_2 = -f_2 \quad (-0,0732157),$$

$$1,47664 \Delta \eta_2 + 10,45771 \Delta \eta_4 + 4,92046 \Delta y_2 = -f_4 \quad (-0,2392450),$$

$$-1,90993 \Delta \eta_0 + 1,50580 \Delta \eta_2 + 4,92046 \Delta \eta_4 + 20,56662 \Delta y_2 = -g.$$

Исключая последовательно по одной переменной, получаем

$$20,56662 \Delta y_2 + 1,50580 \Delta \eta_2 + 4,92046 \Delta \eta_4 - 1,90993 \Delta \eta_0 = -g,$$

$$11,18753 \Delta \eta_2 + 1,11639 \Delta \eta_4 + 0,86120 \Delta \eta_0 = -f_2 + 0,07322g,$$

$$(-0,0997888) \quad 1,11639 \Delta \eta_2 + 9,28051 \Delta \eta_4 + 0,45694 \Delta \eta_0 = -f_4 + 0,23924g,$$

$$(-0,0769786) \quad 0,86120 \Delta \eta_2 + 0,45694 \Delta \eta_4 + 2,03567 \Delta \eta_0 = -f_0 - 0,09287g,$$

$$9,16911 \Delta \eta_4 + 0,37100 \Delta \eta_0 = 0,09979f_2 - f_4 + 0,23193g,$$

$$(-0,0404619) \quad 0,37100 \Delta \eta_4 + 1,96938 \Delta \eta_0 = -f_0 + 0,07698f_2 - 0,09851g,$$

$$1,95437 \Delta \eta_0 = -f_0 + 0,07294f_2 + 0,04046f_4 - 0,10789g.$$

Таким образом, получаем

$$\Delta \eta_0 = -0,51167f_0 + 0,03732f_2 + 0,02070f_4 - 0,05520g,$$

$$\Delta \eta_2 = 0,03732f_0 - 0,09319f_2 + 0,00937f_4 + 0,00805g,$$

$$\Delta \eta_4 = 0,02070f_0 + 0,00937f_2 - 0,10990f_4 + 0,02753g,$$

$$\Delta y_2 = 0,05520f_0 + 0,00805f_2 + 0,02753f_4 - 0,06092g.$$

Вычисление выполнено в табл. 14.9 (вторая и третья полосы являются продолжением первой и должны располагаться с ней в одну линию). В соответствии с этим получаем окончательные значения

$$\eta_0 = 0,50019, \quad \eta_2 = 0,33340, \quad \eta_4 = 0,25004, \quad y_2 = 0,81087,$$

$$y_1 = 0,44713, \quad y_3 = 1,11942, \quad \eta_1 = 0,39976, \quad \eta_3 = 0,28571.$$

Таблица 14.9

ПОСЛЕДОВАТЕЛЬНЫЕ ПРИБЛИЖЕНИЯ ДЛЯ $\eta_0, \eta_2, \eta_4, y_2$

n	η_1	η_2	η_4	y_2	y_1	y_3	η_1	η_3
0	0,50300	0,33150	0,25050	0,81900	0,45238	1,12290	0,40562	0,27998
1	0,50004	0,33339	0,25002	0,81094	0,44713	1,11946	0,39985	0,28567
2	0,50020	0,33341	0,25004	0,81087	0,44713	1,11943	0,39975	0,28571

n	$e y_1$	$2e y_2$	$e y_3$	$\eta_1 e y_1$	$2\eta_2 e y_2$	$\eta_3 e y_3$
0	1,57205	4,53646	3,07376	0,63765	1,50384	0,86059
1	1,56381	4,50004	3,06321	0,62529	1,50027	0,87507
2	1,56381	4,49972	3,06312	0,62513	1,50025	0,87516

n	f_0	f_2	f_4	g	$\Delta \eta_1$	$\Delta \eta_2$	$\Delta \eta_4$	Δy_2
0	-0,01066	-0,00650	0,04187	0,16007	-0,00276	0,00189	-0,00048	-0,00806
1	-0,00048	-0,00022	0,00004	0,00150	0,00016	0,00002	0,00002	-0,00007
2	0,00004	0,00007	-0,00004	-0,00011	-0,00001	-0,00001	0	0

Соответствующие точные значения (см. § 14.15) таковы:

$$\eta_0 = 0,50000, \quad \eta_2 = 0,33333, \quad \eta_4 = 0,25000, \quad y_2 = 0,81093, \\ y_1 = 0,44629, \quad y_3 = 1,11923, \quad \eta_1 = 0,40000, \quad \eta_3 = 0,28571.$$

ЛИТЕРАТУРА

1. Hotelling H., Some New Methods in Matrix Calculation, Ann. Math. Statist., 14, 1—34 и 440—441 (1943).
2. Keller E. G., Some Present Non-linear Problems of the Electrical and Aeronautical Industries, Quart. Appl. Math., 2, 72—86 (1944).
3. Von Kármán T., The Engineer Grapples with Non-linear Problems, Bull. Amer. Math. Soc., 46, 615—683 (1940).

Глава 15

ЧТО ПРЕДСТАВЛЯЮТ СОБОЙ РЕЛАКСАЦИОННЫЕ МЕТОДЫ?

ДЖОРДЖ Э. ФОРСАЙТ

Калифорнийский университет, Лос-Анжелес

§ 15.1. Введение. Вопрос, поставленный в заголовке, очень труден, и автору удалось ответить на него лишь в процессе подготовки этой статьи. Формулировка этого ответа будет приведена в § 15.7.

В качестве вступления в § 15.2 и 15.3 дается краткий обзор ряда этапов, в результате которых в течение последних 25 лет английские и американские инженеры воссоздали и реконструировали один из древних видов искусств, *искусство вычисления*. Изложение этого вопроса ведется путем разбора одного весьма простого примера, который позволяет проиллюстрировать характер примененных математических операций, минуя запутанность деталей обычных инженерных задач. В § 15.4 мы рассмотрим релаксационные методы в связи с различными способами решения систем линейных алгебраических уравнений. Материал § 15.5 иллюстрирует сущность методов численного анализа, т. е. вклад, вносимый математиками в дело развития и совершенствования вычислительных методов. В число этих усовершенствований входит особый прием, ускоряющий сходимость релаксационных методов. В § 15.6 изучается одна из задач о колебаниях, для решения которой могут быть использованы релаксационные методы. В конце главы приводится литература по этому вопросу.

§ 15.2. Решетки и аналогичные задачи. Простой пример может познакомить нас с релаксационными методами. Посмотрим на рис. 15.1. Мы видим нить с закрепленными концами, испытывающую постоянное натяжение T . Вдоль нити на равных расстояниях d располагаются 10 грузиков массы m каждый. Каково будет состояние равновесия грузиков под влиянием силы натяжения и силы тяжести? Допустим, что массой нити можно пренебречь и что благодаря большому натяжению нити смещение грузиков оказывается незначительным. (На рис. 15.1 провисание нити преувеличено.)

Пусть смещения грузиков, вызванные провисанием нити и отсчитываемые от горизонтальной пунктирной линии вниз (положительное направление), равны $u_1, \dots, u_{10}$. В силу симметрии мы можем пред-

положить, что $u_5 = u_6$ и т. п. Положим $u_0 = 0$. Для каждой симметрично расположенной последовательности точек u_i направленная вверх составляющая силы натяжения, действующей на i -й грузик, будет $-mg + (T/d)(-u_{i-1} + 2u_i - u_{i+1})$. Если мы выберем такую систему физических единиц, в которой $mgd/T = 1$, и обозначим

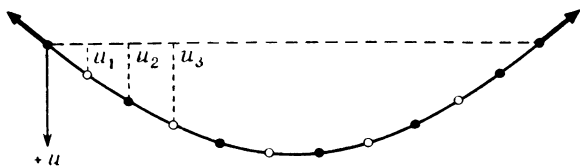


Рис. 15.1. Нить под постоянным натяжением T с концами, закрепленными на одной высоте, и с одинаковыми грузами, размещенными на равных расстояниях друг от друга.

направленные вверх составляющие силы натяжения нити через $r_1, \dots, r_5$, то будем иметь

$$\begin{aligned}
 -1 + 2u_1 - u_2 &= r_1, \\
 -1 - u_1 + 2u_2 - u_3 &= r_2, \\
 -1 - u_2 + 2u_3 - u_4 &= r_3, \\
 -1 - u_3 + 2u_4 - u_5 &= r_4, \\
 -1 - u_4 + u_5 &= r_5.
 \end{aligned} \tag{15.1}$$

Значения $u_1, \dots, u_5$, соответствующие положению равновесия нити, находятся путем решения уравнений (15.1) при условии, что $r_1 = \dots = r_5 = 0$. Прямой путь решения этой задачи заключался бы в решении системы (15.1) (при условии, что все $r_i = 0$) путем систематического исключения неизвестных приемами, известными из алгебры (см. ниже § 15.4). Система этих уравнений может быть решена также с помощью определителей. В то время как эти прямые методы довольно легко дают решение уравнений (15.1), они оказались бы весьма громоздкими для более сложной системы, подобной той, например, которая соответствует рис. 15.2.

Среди инженеров и вычислителей завоевал популярность иной метод решения уравнений (15.1). Сначала мы делаем первое предположение относительно значений u_i , скажем $u_1 = 4$, $u_2 = 8$, $u_3 = 12$, $u_4 = 14$, $u_5 = 15$. Из уравнений (15.1) находим, что $r_1 = r_2 = -1$, $r_3 = 1$, $r_4 = r_5 = 0$. Так как полученные *остатки* r_i не все равны нулю, мы будем улучшать „пробное“ решение с целью уменьшения остатков r_i . Выбираем одно из тех r_i , которые имеют наибольшие численные значения, скажем r_3 . Будем „приводить“ его к нулю путем соответствующего изменения Δu_3 значений только переменной u_3 . Из уравнений (15.1) ясно, что изменение u_3 на единицу ($\Delta u_3 = +1$) вызовет

изменения только у r_2 , r_3 и r_4 : $\Delta r_2 = -1$, $\Delta r_3 = +2$, $\Delta r_4 = -1$. Для обращения в нуль r_3 необходимо, чтобы $\Delta r_3 = -1$. Этого можно достигнуть, взяв $\Delta u_3 = -0,5$. При этом получаем $\Delta r_2 = \Delta r_4 = +0,5$. Окончательно находим, что поправленному решению соответствуют следующие остатки r_i : $r_1 = -1,0$; $r_2 = -0,5$; $r_3 = 0,0$; $r_4 = 0,5$; $r_5 = 0,0$. Среди них наибольшее численное значение имеет r_1 . Переходим к „ликвидации“ его, для чего выбираем $\Delta u_1 = +0,5$. Затем

выбираем $\Delta u_2 = +0,5$ и т. д. Восемь шагов этого процесса приведены в табл. 15.1. Опытный вычислитель выполняет расчет очень быстро, производя вычисления в уме и записывая остаток только в том случае, когда он изменяется.

Саусвелл [17] истолковывает остатки r_i как взятые со знаком минус величины внешних сил, реально приложенных к грузикам и обеспечивающих равновесие системы при происходящих смещениях. Каждый шаг в предыдущих вычислениях рассматривается тогда как релаксация (ослабление) одной из этих внешних сил. Отсюда и происходит предложенное Саусвеллом название *релаксационный метод*.

В нижней части табл. 15.1 указаны текущие значения различных u_i , например $u_1 = 4 + 0,5 + 0,3 = 4,8$. Повторные

Рис. 15.2. Сетка нитей, находящихся под постоянным натяжением T , с концами, закрепленными в горизонтальной плоскости, и с одинаковыми грузами, размещенными на равных расстояниях друг от друга.

вычисления остатков подтверждают правильность предшествующих им вычислений. В итоге восьми шагов вычислений, результаты которых указаны в таблице, максимальная погрешность в определении значений u_i снижается от 1,0 до 0,03. Дальнейшие вычисления опять повышают точность значений u_i , и для достижения обычной в инженерной практике точности не потребуется значительного времени.

В распоряжении вычислителей, пользующихся методом релаксации, имеется много остроумных приемов. Один из них (см. табл. 15.1) состоит в вычислении только одной значащей цифры, с тем чтобы не усложнять чисел поправками на излишнюю на данном шаге точность. Так, например, вместо $\Delta u_4 = -0,25$ принимают $\Delta u_4 = -0,3$. Таким образом, остатки ликвидируются только в наиболее важных

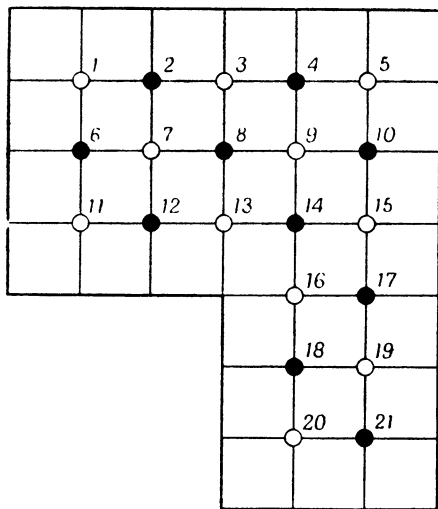


Таблица 15.1

РЕШЕНИЕ УРАВНЕНИЙ (15.1) РЕЛАКСАЦИОННЫМ МЕТОДОМ

		1	2	3	4	5	
Первое приближение u_i		4,0	8,0	12,0	14,0	15,0	
Остатки r_i		-1,0	-1,0	1,0	0,0	0,0	
i	Δu_i	Остатки					
3	-0,5	-1,0	-0,5	0,0	0,5	0,0	
1	+0,5	0,0	-1,0	0,0	0,5	0,0	
2	+0,5	-0,5	0,0	-0,5	0,5	0,0	
4	-0,3	-0,5	0,0	-0,2	-0,1	0,3	
1	+0,3	0,1	-0,3	-0,2	-0,1	0,3	
2	+0,2	-0,1	0,1	-0,4	-0,1	0,3	
3	+0,2	-0,1	-0,1	0,0	-0,3	0,3	
5	-0,3	-0,1	-0,1	0,0	0,0	0,0	
Текущее решение		4,8	8,7	11,7	13,7	14,7	Проверка
Остатки		-0,1	-0,1	0,0	0,0	0,0	
Истинное решение		5,0	9,0	12,0	14,0	15,0	

разрядах. Последующие шаги вычислений приводят автоматически к повышению точности.

Большая экономия времени в инженерной практике достигается путем использования (вместо табл. 15.1 или подобной ей) специальных вычислительных бланков. На этих бланках могут быть записаны значения Δu_i и r_i .

Для ускорения сходимости значений u_i к истинному решению могут быть также использованы и другие остроумные методы. С некоторыми из них можно познакомиться по указанной в приложении литературе.

Релаксация не представляет в действительности никакой трудности для вычислителя по следующим причинам: а) наличие промежуточных результатов вычислений, являющихся приближенными ответами, предопределяет цель каждого последующего шага вычислений и помогает снижению утомляемости в течение целого дня вычислительной работы; б) возможность ускорить сходимость последовательных приближений является постоянно стимулирующим фактором для проявления способностей вычислителя; в) никогда не приходится

терять много времени на производство ошибочных вычислений (когда вычисления продолжаются, хотя в результаты промежуточных вычислений уже вкралась ошибка), как это может иметь место при вычислениях по методу исключений.

Существует много видоизменений релаксационного метода. Все они связаны с решением систем уравнений, обычно линейных, и обладают, как нам кажется, следующими основными особенностями: а) для каждого пробного решения в каждом из уравнений существует мера погрешности; б) для каждого из уравнений, которому не удалось удовлетворить, существует своя собственная формула, позволяющая улучшать пробное решение; в) вычисления на каждом шаге производятся с тем из уравнений, в котором погрешность оказывается наибольшей.

Релаксационный метод первоначально предназначался для вычислений с помощью карандаша и бумаги без использования настольных вычислительных машин и идеально приспособлен для такой работы. Этот метод, естественно, может быть реализован и с помощью настольных вычислительных машин, однако в этом случае он теряет по сравнению с другими методами некоторые из своих преимуществ. О реализации этого метода на автоматических вычислительных машинах смотри ниже в § 15.5.

Релаксационный метод известен еще со времен Гаусса [22], который пользовался им и рекомендовал как сам метод, так и многие известные в настоящее время остроумные приемы. Зейдель [40] доказал сходимость этого метода для системы линейных уравнений с положительно определенной матрицей. Крсс [20] предложил для инженерных вычислений систему итеративных методов. В тридцатых годах Саусвелл переоткрыл метод Гаусса и назвал его релаксационным. Саусвелл и его школа развили этот метод и показали инженерам и научным работникам, сколь широка область его применения. Этот метод оказался особенно подходящим для анализа сложных статически неопределенных шарнирных ферм. Таким системам отвечают уравнения, подобные уравнениям (15.1), с большим числом отличных от нуля коэффициентов.

Следует отметить, что одномерные задачи, подобные той, которая показана на рис. 15.1, не следует на практике решать методом релаксации, так как они легко решаются рекуррентно, если исходить из некоторого пробного приближенного решения (даже при переменных массах). Так, например, последнее из уравнений (15.1) заменим на

$$-1 - u_4 + 2u_5 - u_6 = r_5.$$

Полагаем $r_1 = \dots = r_5 = 0$. В силу симметрии мы должны иметь $u_6 - u_5 = 0$. Испробуем значение $u_1^{(1)} = 3$. Из модифицированных уравнений (15.1) находим последовательно

$$u_2^{(1)} = 5, \quad u_3^{(1)} = 6, \quad u_4^{(1)} = 6, \quad u_5^{(1)} = 5, \quad u_6^{(1)} = 3,$$

откуда

$$u_6^{(1)} - u_5^{(1)} = -2.$$

Теперь используем значение $u_1^{(2)} = 6$. В этом случае мы последовательно находим $u_2^{(2)} = 11$, $u_3^{(2)} = 15$, $u_4^{(2)} = 18$, $u_5^{(2)} = 20$, $u_6^{(2)} = 21$, $u_6^{(2)} - u_5^{(2)} = +1$. Проводя линейную интерполяцию между $u_1^{(1)}$ и $u_1^{(2)}$, устанавливаем, что для выполнения соотношения $u_5 - u_6 = 0$ следует принять $u_1 = 5$, откуда и получаем рекуррентно истинное решение.

Таблица 15.1† показывает на простом примере технику реализации релаксационного метода. Практическое же использование этого метода начинается с двумерных задач, подобных задачам по расчету ферм. Предположим, что нужно исследовать L -образную сеть, расположенную в горизонтальной плоскости и состоящую из нитей с 21 грузиком. Каким образом можно рассчитать положение равновесия находящихся под действием силы тяжести грузиков при большой величине натяжения? При сделанных выше предположениях можно составить систему уравнений (21 уравнение), подобных уравнению

$$-1 - u_2 - u_6 + 4u_7 + u_8 - u_{12} = 0. \quad (15.2)$$

Использование метода исключений для решения полученной системы уравнений было бы связано с утомительными вычислениями, и никакая простая рекуррентная схема здесь не применима. Практический ответ дает метод релаксаций, который реализуется численно примерно так, как в табл. 15.1, если не считать более медленной сходимости. Здесь также можно применить упоминавшиеся выше остроумные приемы.

Характерной особенностью систем уравнений, подобных (15.2), является то, что большинство коэффициентов при неизвестных u_i равны нулю. Именно эта особенность и делает метод релаксации таким методом решения систем линейных уравнений, реализация которого возможна с помощью карандаша и бумаги.

§ 15.3. Другие источники получения систем уравнений, решаемых методом релаксации. Нагруженные фермы не являются основным источником систем линейных уравнений. Такие системы очень часто появляются в других разделах науки и техники. Группа важных задач возникает при упорядочении полученных данных. Так, например, в результате геодезических или астрономических измерений получают значения большого количества различных величин Q_i , связанных между собой известными или предполагаемыми соотношениями (так, например, углы плоского треугольника должны составлять в сумме 180°). Результаты этих измерений содержат обычно ошибки и оказываются между собой до некоторой степени несогласованными, а потому и не удовлетворяют в точности связывающим

их соотношениям. Задача состоит в вычислении по измеренным значениям наиболее вероятных значений Q_i .

Если упомянутые выше зависимости линейны относительно величин Q_i , то ответ находится с помощью решения некоторых *нормальных уравнений*, составляющих систему линейных уравнений. Подобного вида системы возникают также при подгонке результатов лабораторных измерений, подверженных погрешностям, путем вычерчивания кривых (подобных графикам полиномов пятой степени) с неизвестными и линейно входящими туда параметрами.

При определенных условиях, когда взаимодействия между различными из известных соотношений незначительны, нормальные уравнения могут напоминать систему (15.1) тем, что многие из коэффициентов в них равны нулю. Для подобных систем методы релаксации могут оказаться весьма полезными. В других задачах, как, например, в случае полиномов, возникающих при подгонке результатов измерения, взаимозависимостей оказывается много, соответствующие уравнения имеют мало нулевых коэффициентов, и метод релаксации оказывается мало подходящим.

Другими источниками получения систем с большим числом нулевых коэффициентов являются законы Кирхгоффа для простых электрических цепей, аналогичные им уравнения течения воды, разностные уравнения теории чисел и т. д.

Наиболее богатым источником получения систем уравнений, решаемых релаксационным методом, являются, по-видимому, задачи теории упругости непрерывных сред. Они приводят к обыкновенным дифференциальным уравнениям и уравнениям в частных производных, точное аналитическое решение которых оказывается обычно либо невозможным, либо неосуществимым практически из-за большого объема работы. В качестве примера предположим, что горизонтально расположенная L -образная наружная ферма (рис. 15.2) покрыта однородной тяжелой мембраной, испытывающей равномерное большое натяжение. Как будет провисать мембрана под действием силы тяжести? В определенной системе единиц вертикальное смещение $u(x, y)$ в каждой внутренней точке (x, y) оказывается решением дифференциального уравнения

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + 1 = 0 \quad (15.3)$$

в частных производных при граничном условии: на ферме $u(x, y) = 0$. К такому же уравнению приводит более важная для техники задача кручения однородных изотропных цилиндрических стержней, имеющих L -образное поперечное сечение.

Для указанной задачи не существует простого решения в замкнутом виде, и потому мы должны ограничиваться тем или иным приближением. Различные методы приближенного решения были описаны Коллатцом [3]. Один из широко применяемых методов, описан-

ный Курантом, состоит в замене непрерывной задачи на дискретную путем соответствующего разбиения непрерывно распределенной массы мембраны на изолированные части и концентрации натяжения на упругих нитях, соединяющих эти части. В литературе, посвященной релаксационным методам, рассматривается большое число разнообразных способов сведения непрерывной задачи к дискретной. Существуют глубокие соображения, которыми приходится пользоваться при рассмотрении особенностей в невыпуклых углах границы и при соответствующем задаче разбиении массы на отдельные части вблизи искривленных границ. См., например, книгу Милна [13] и статью Вудса [28].

Для рассмотренной выше L -образной мембраны квадратное распределение грузов при прямоугольном узоре нитей (рис. 15.2) оказалось бы наиболее удобным для работы. Пусть h есть длина стороны квадрата этой сетки. В каждом уравнении второго порядка, не содержащем члена $\partial^2 u / \partial x \partial y$, каждую из частных производных можно было бы заменить разностным отношением следующим образом:

$$\frac{\partial u(x, y)}{\partial x} \text{ заменяется на } \frac{u(x+h, y) - u(x-h, y)}{2h},$$

$$\frac{\partial^2 u(x, y)}{\partial x^2} \text{ заменяется на } \frac{u(x+h, y) - 2u(x, y) + u(x-h, y)}{h^2}.$$

Аналогично заменяются производные $\partial u / \partial y$ и $\partial^2 u / \partial y^2$. Таким образом, уравнение (15.3), отвечающее сплошной мембране, сводится при соответствующем выборе единиц к системе уравнений, подобных уравнению (15.2). Для решения таких систем с успехом можно применить, как это указывалось в § 15.2, релаксационный метод. То же самое оказывается справедливым и для задач, где сетки состоят из правильных треугольников, шестиугольников или даже из неправильных многоугольников.

Системы уравнений, связанных с квадратными или прямоугольными сетками (подобными сетке 15.2), обладают одним свойством, важность которого будет показана позднее. На рис. 15.2 грузы обозначены темными и светлыми кружочками. Нити соединяют только грузы противоположного цвета. В уравнение равновесия любого груза, например груза 7, входит член, соответствующий самому грузу ($4u_7$), члены, соответствующие грузам противоположного цвета ($-u_2, -u_6$ и т. д.), и постоянный член (-1). Таким образом, неизвестные могут быть разделены на две группы (на рис. 15.2 черные и белые кружочки) так, что не будет общих членов в уравнениях, связывающих неизвестные, относящиеся к одному цвету. Если расположить неизвестные и уравнения предшествующей системы так, чтобы вначале стояли все уравнения, соответствующие черным грузам, а затем — белым, то можно добиться того, чтобы матрица

Типичным прямым методом решения системы уравнений (15.4) является метод *исключения*, обычно характеризуемый в курсе алгебры как до некоторой степени неопределенный процесс исключения переменных из уравнений. Имеются различные способы систематизации этих исключений, каждый из которых и представляет собой одну из разновидностей метода исключения [4]. Различия между этими разновидностями состоят отчасти в порядке исключения неизвестных и отчасти в деталях записи промежуточных результатов вычислений. Приложение наиболее эффективного из этих методов к системе (15.4) требует проведения около $(1/3)n^3$ умножений с накоплением результата и примерно n^2 записей чисел. Никому не удалось найти метода решения общей системы уравнений (15.4), требующего меньшего числа умножений. Поэтому метод исключения считается обычно наиболее эффективным методом получения точного решения системы уравнений общего вида в том случае, когда наиболее важным фактором является число требуемых вычислительных операций. Это имеет место, например, при использовании для решения уравнений настольных вычислительных машин.

В курсе алгебры для колледжей уделяется значительное внимание решению систем линейных уравнений *методом Крамера* с использованием определителей. Этот метод удобен при $n=2$ или 3 и является полезным для теоретических исследований при больших значениях n . Однако на практике для больших n этот метод ставит вычислителя перед необходимостью вычисления $n+1$ определителей n -го порядка. Как известно, определитель n -го порядка является суммой $n!$ различных членов, для получения каждого из которых требуется произвести $n-1$ умножений. Если идти этим путем, то для отыскания значений $n+1$ таких определителей потребовалось бы произвести $(n-1)(n+1)!$ умножений. Для $n=10$ это составит 359 251 200 умножений, что потребует 114 лет работы одного вычислителя, выполняющего одно умножение за 10 сек. Даже машина SWAC¹⁾, которая выполняет 2600 умножений в секунду, смогла бы выполнить требуемое умножение лишь в течение 38 часов. Для $n=26$ машине SWAC понадобилось бы $3 \cdot 10^{18}$ лет. Выполненные недавно решения некоторых геодезических задач, связанных с системами (15.4) из $n=2500$ уравнений, были получены отнюдь не методом вычисления определителей! (На практике определители большей частью вычисляются методом исключения, требующим примерно $(1/3)n^3$ умножений.)

Один из известных вариантов метода исключения состоит в разделении n неизвестных из системы (15.4) на две или большее число

¹⁾ Автоматическая вычислительная машина Национального бюро стандартов — весьма быстродействующая электронная цифровая машина, предоставленная в пользование Исследовательской проектной организации в области численного анализа при Калифорнийском университете в Лос-Анджелесе. Подобные машины описаны Лемером [33].

групп и в соответствующем разделении уравнений системы (15.4) на то же самое число групп. После этого оказывается возможным составить схему исключения, которая сводит решение системы (15.4) к решению двух систем из $n/2$ уравнений с $n/2$ неизвестными каждая (если n — четное) при использовании некоторого дополнительного перемножения матриц. Этот метод называется методом *разбиения*. Применить метод разбиения легко может каждый, знакомый с матричными обозначениями (см. книгу Перлиса [15] или гл. 13 настоящей книги).

Итерационный процесс, приводящий к решению системы (15.4), начинается с некоторых первых предположительных значений неизвестных u_i , например $u_1^{(0)}, \dots, u_n^{(0)}$; совокупность этих значений можно кратко записать в виде вектора u^0 . Первый шаг алгоритма приводит к новому вектору $u^{(1)}$. После k шагов получаем $u^{(k)}$, и следующий типичный для данного метода шаг приводит к $u^{(k+1)}$. Один класс итерационных методов включает вычисление ошибки $\varepsilon(u) = r_1^2 + \dots + r_n^2$, соответствующей каждому пробному вектору u . На каждом шаге процесса пытаются уменьшить, насколько возможно, $\varepsilon(u)$ путем перемещения в n -мерном пространстве от точки $u^{(k)}$ в подходящем направлении. Такие методы рассматривались Томпкинсом [36], интересные модификации их предложены Хестенсом, Ланчосом и Штифелем.

Старый тип итерационного процесса, о котором упоминает Зейдель [40] и который обычно называется методом Зейделя или методом Гаусса — Зейделя, состоит в следующем: предполагается, что ни одно из a_{ii} не равно нулю. Решаем первое из уравнений (15.4) относительно u_1 , используя текущие значения $u_2, \dots, u_n$. Затем решаем второе из уравнений (15.4) относительно u_2 , используя последние из полученных значений $u_1, u_3, u_4, \dots, u_n$ и т. д. Таким образом, все уравнения системы (15.4) решаются в циклическом порядке относительно $u_1, \dots, u_n$, каждый раз с использованием самых последних значений остальных неизвестных. Чтобы описать это полнее, допустим, что известны $u_1^{(k)}, \dots, u_n^{(k)}$. После этого находим $u_1^{(k+1)}, \dots, u_n^{(k+1)}$ путем последовательного решения следующих n уравнений:

$$\begin{aligned} a_{11} u_1^{(k+1)} + a_{12} u_2^{(k)} + a_{13} u_3^{(k)} + \dots + a_{1n} u_n^{(k)} + b_1 &= 0, \\ a_{21} u_1^{(k+1)} + a_{22} u_2^{(k+1)} + a_{23} u_3^{(k)} + \dots + a_{2n} u_n^{(k)} + b_2 &= 0, \\ a_{31} u_1^{(k+1)} + a_{32} u_2^{(k+1)} + a_{33} u_3^{(k+1)} + \dots + a_{3n} u_n^{(k)} + b_3 &= 0, \\ \vdots & \\ a_{n1} u_1^{(k+1)} + a_{n2} u_2^{(k+1)} + a_{n3} u_3^{(k+1)} + \dots + a_{nn} u_n^{(k+1)} + b_n &= 0. \end{aligned} \quad (15.5)$$

Читатель может заметить, что метод Зейделя тесно связан с релаксационным методом, описанным в § 15.2. Единственное различие состоит в порядке решения уравнений (15.4). При методе релаксаций мы оперируем этими уравнениями в том порядке, который опреде-

ляется значением r_i . При методе Зейделя порядок фиксирован и является циклическим.

Существуют итерационные процессы, в которых n уравнений (15.4) интерпретируются как гиперплоскости в n -мерном пространстве. Начинают с произвольной точки P_0 n -мерного пространства, ортогонально проектируют ее на первую гиперплоскость и получают точку P_1 . Затем P_1 проектируют на вторую гиперплоскость и получают P_2 . Подобно методу Зейделя, этот процесс также циклический и родственен процессу релаксационного типа, в котором проектирование всегда производится на наиболее отдаленную гиперплоскость.

В заключение этого далеко не полного обзора скажем, что существует несколько методов Монте-Карло решения уравнений (15.4) путем случайных выборов.

§ 15.5. Математический анализ покомпонентной релаксации. В предыдущих параграфах мы рассмотрели некоторые стороны *искусства вычисления*, искусства использования технических средств для получения численного решения научных задач. С искусством вычисления непосредственно связан раздел науки, называемый *численным анализом*. Одним из видов деятельности научного работника, занимающегося численным анализом, является деятельность библиографического и исторического характера, имеющая своей целью составление обзоров известных численных методов, подобных краткому обзору, изложенному в § 15.4. Другой областью деятельности является критический анализ методов вычислений, разработка ответов на связанные с ним вопросы и приемов, улучшающих известные методы, а также создание новых вычислительных методов. В настоящем параграфе мы остановимся на разработке одного из таких улучшающих приемов и покажем, каким образом он обеспечивает более совершенную форму релаксации.

Рассматривая метод релаксации, изложенный в § 15.2, научный работник, занимающийся численным анализом, интересуется, конечно, вопросом о том, когда этот метод сходится к решению. (Если он не интересуется этим вопросом, то его спросит об этом вычислитель.) Даже для двух столь простых уравнений, как

$$x + 2y = 5, \quad -2x + y = 6,$$

метод релаксации сойдется только в том случае, если их предварительно видоизменить. Если метод сходится, то научный работник интересуется вопросом о том, как быстро он сходится. Если сходимость медленна, то он интересуется возможностью ее ускорения.

Для релаксационного метода Гаусса — Саусвелла весьма трудно дать удовлетворительные ответы на три поставленных выше вопроса. В результате серьезной работы различных математиков удалось получить только некоторые частичные ответы. Так, например:

Если матрица коэффициентов уравнений (15.4) симметрична и положительно определена¹⁾, то релаксационный метод будет сходиться к решению независимо от того, насколько неудачным было его начало. (Темпл [24]; тесно связанный с этим результат был получен Зейделем [40].)

Если матрица из коэффициентов уравнений (15.4) симметрична и имеет положительные диагональные элементы $a_{11}, \dots, a_{nn}$, но не является положительно определенной, то релаксационный метод должен расходиться при некоторых исходных данных, а практически, как правило, при любых исходных данных (Островский [38]; см. также Райх [39]).

В то время как трудности полного анализа релаксационного метода стимулируют математиков к их преодолению, появление быстродействующих цифровых вычислительных машин резко изменило всю обстановку. Ввиду не известного до конца характера этих машин более важным, чем прежде, оказывается проведение исчерпывающего анализа вычислительных процессов. Возрастает популярность итерационных методов, ибо на современных вычислительных машинах быстрее и дешевле работать с уравнениями в циклическом порядке, нежели искать наибольшие остатки.

Это обстоятельство оказывается благоприятным для численного анализа, так как циклические процессы гораздо легче анализировать ввиду их линейности. Дадим описание такого анализа.

Пусть матрица A из коэффициентов уравнений (15.4), предполагаемая невырожденной, записана в виде суммы трех матриц: $A = D + E + F$. Здесь D имеет ту же главную диагональ, что и матрица A , а остальные ее элементы равны нулю; E получается из матрицы A путем замены нулями всех элементов матрицы A , лежащих на главной диагонали этой матрицы и выше нее, и F — заменой нулями всех элементов матрицы A , лежащих на главной диагонали и ниже нее. Таким образом, схематически

$$A = \begin{bmatrix} & & & F \\ & & D & \\ E & & & \end{bmatrix}.$$

Теперь уравнения (15.5) могут быть записаны в следующей матрично-векторной форме:

$$(D + E) \mathbf{u}^{(k+1)} + F \mathbf{u}^{(k)} + \mathbf{b} = \mathbf{0}. \quad (15.6)$$

Если через $\mathbf{u}^{(\infty)}$ обозначить единственный вектор-решение системы (17.4), то получим

$$(D + E) \mathbf{u}^{(\infty)} + F \mathbf{u}^{(\infty)} + \mathbf{b} = \mathbf{0}. \quad (15.7)$$

¹⁾ Матрица системы (15.4) называется *положительно определенной*, если $\sum_{i,j=1}^n a_{ij} u_i u_j > 0$ для всех отличных от нуля значений $u_1, \dots, u_n$. К подобным матрицам приводят физические проблемы консервативного типа, подобные пассивным электрическим цепям.

Вычитая (15.7) из (15.6) и полагая *ошибку* в значении $\mathbf{u}^{(k)}$ равной $\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^{(\infty)}$, получаем

$$(D + E)\mathbf{e}^{(k+1)} + F\mathbf{e}^{(k)} = \mathbf{0}. \quad (15.8)$$

Далее, предполагая, как и ранее, что все $a_{ii} = 0$, видим, что матрица $D + E$ имеет обратную $(D + E)^{-1}$. Обозначая через H матрицу $-(D + E)^{-1}F$, находим из (15.8), что

$$\mathbf{e}^{(k+1)} = H\mathbf{e}^{(k)}, \quad (15.9)$$

откуда

$$\mathbf{e}^{(k)} = H^k \mathbf{e}^{(0)}. \quad (15.10)$$

Уравнение (15.9) характеризует линейный характер циклического итерационного процесса Зейделя, а уравнение (15.10) выражает ошибку $\mathbf{e}^{(k)}$ через исходную погрешность. Полный анализ погрешности основан на рассмотрении уравнения (15.10). Процесс Гаусса — Саусвелла теоретически более сложен именно вследствие того, что он не имеет простого аналога уравнения (15.10).

Из теории линейных преобразований (см., например, книгу Перлиса [15]) известно, когда и как $\mathbf{e}^{(k)}$ стремится к нулю. Уравнения

$$|H - \mu I| = 0$$

и

$$|(D + E)\mu + F| = 0 \quad (15.11)$$

имеют одни и те же n действительных или комплексных корней $\mu_1, \dots, \mu_n$. Если все $|\mu_i| < 1$, то в процессе Зейделя (17.5) $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^{(\infty)}$. Если какое-либо $|\mu_i| \geq 1$, то процесс Зейделя расходится. Этот факт является ответом (в отношении процесса Зейделя) на первый из вопросов о сходимости релаксационного метода, поставленных в начале этого параграфа.

Для большинства матриц A каждому корню μ_i уравнений (15.11) соответствует единственный вектор $\mathbf{y}^{(i)}$, такой, что

$$H\mathbf{y}^{(i)} = \mu_i \mathbf{y}^{(i)}.$$

Таким образом, преобразование H оставляет вектор $\mathbf{y}^{(i)}$ неизменным по направлению, но растягивает его (если $|\mu_i| > 1$) или сжимает ($|\mu_i| < 1$) в $|\mu_i|$ раз. Далее, все эти векторы $\mathbf{y}^{(i)}$ образуют координатный базис некоторой косоугольной координатной системы, в которой для исходного вектора ошибки $\mathbf{e}^{(0)}$ можно получить выражение

$$\mathbf{e}^{(0)} = \sum_{i=1}^n c_i \mathbf{y}^{(i)}.$$

Предположим, что $|\mu_1| > |\mu_2| \geq \dots \geq |\mu_n|$. Путем повторных умножений на H новый вектор $H^{(k)}\mathbf{e}^{(0)}$ перемещается примерно в напра-

влении, соответствующем корню μ_1 , имеющему наибольшую абсолютную величину. Следовательно, мы находим, что

$$\mathbf{e}^{(k)} = H^k \mathbf{e}^{(0)} = c_1 \mu_1^k \mathbf{y}^{(1)}. \quad (15.12)$$

Мы ответили на вопрос о том, как быстро $\mathbf{e}^{(k)} \rightarrow 0$. Если $|\mu_i| < 1$, то каждый шаг неизбежно уменьшает длину вектора $\mathbf{e}^{(k)}$ в $|\mu_1|$ раз.

Если $|\mu_1| < 1$, то $\mathbf{e}^{(k)} \rightarrow 0$ вдоль одного направления, т. е. направления $\mathbf{y}^{(1)}$. Следовательно, $\mathbf{u}^{(k)} \rightarrow \mathbf{u}^{(\infty)}$ вдоль направления $\mathbf{y}^{(1)}$. Зная это, мы можем без труда разработать метод для ускорения сходимости $\mathbf{u}^{(k)}$ к $\mathbf{u}^{(\infty)}$. Это является ответом на третий вопрос, поставленный в начале этого параграфа.

Случаи, когда доминирующими оказываются больше чем одно $|\mu_i|$, являются более сложными, однако поддаются исследованию.

В качестве примера, иллюстрирующего процесс Зейделя и характер его сходимости, решим систему (15.1), используя те же начальные данные, какие указаны в табл. 15.1. Мы получаем следующую итерацию:

$$\begin{aligned} u_1^{(k+1)} &= \frac{1}{2} (1 + u_2^{(k)}), \\ u_2^{(k+1)} &= \frac{1}{2} (1 + u_1^{(k+1)} + u_3^{(k)}), \\ u_3^{(k+1)} &= \frac{1}{2} (1 + u_2^{(k+1)} + u_4^{(k)}), \\ u_4^{(k+1)} &= \frac{1}{2} (1 + u_3^{(k+1)} + u_5^{(k)}), \\ u_5^{(k+1)} &= 1 + u_4^{(k+1)}. \end{aligned}$$

Несколько циклов таких итераций приведены в табл. 15.2. Остатки не указаны, и каждая строка таблицы соответствует полному циклу указанного выше алгоритма.

В первом столбце табл. 15.2 приведено отношение наибольших погрешностей в значениях u_i в предшествующей и последующей строках таблицы. Уравнение (15.2) показывает, что это отношение сходится к μ_1 . (Решение любого из уравнений (15.11) дает $\mu_1 = 0,9045$; $\mu_2 = 0,3455$; $\mu_3 = \mu_4 = 0$.) Сходимость $\mathbf{u}^{(k)}$ к $\mathbf{u}^{(\infty)}$ односторонняя и весьма регулярная. Требуется около 22 циклов, чтобы дать точность в 1 десятичный знак. При обычных вычислениях за столом было бы весьма легко сделать достаточно хорошие предположения о значении $\mathbf{u}^{(\infty)}$. Подобные „ускорения“ были бы также возможными при работе на электронной цифровой машине, однако требование иметь для этого отдельный код могло бы оказаться неудобным.

В своей диссертации Юнг [41] ставит вопрос о возможности такой модификации процесса Зейделя, при которой данный процесс сошелся бы гораздо быстрее и можно было бы обойтись без отдельных ускорений. В течение долгого времени вычислители, пользовавшиеся

Таблица 15.2

РЕШЕНИЕ УРАВНЕНИЙ (15.1) МЕТОДОМ ЗЕЙДЕЛЯ

Отношение наибольших ошибок	$u^{(\infty)}$	5	9	12	14	15
	$u^{(0)}$	4	8	12	14	15
0,50	$u^{(1)}$	4,50	8,75	11,88	13,94	14,94
0,24	$u^{(2)}$	4,88	8,88	11,88	13,91	14,91
0,75	$u^{(3)}$	4,94	8,91	11,91	13,91	14,91
0,96	$u^{(4)}$	4,955	8,932	11,921	13,916	14,916
0,92	$u^{(5)}$	4,966	8,944	11,930	13,923	14,923
0,909	$u^{(6)}$	4,972	8,951	11,937	13,930	14,930
0,906	$u^{(7)}$	4,9755	8,9562	11,9431	13,9366	14,9366

методом релаксации, наблюдали, что процесс Гаусса — Саусвелла часто протекает быстрее в том случае, если на каждом шаге производить небольшую „сверхрелаксацию“. Поэтому Юнг приступил к тщательному анализу сверхрелаксаций в связи с процессом Зейделя.

На каждом шаге процесса Зейделя (15.5) или (15.6) Юнг предлагает начать с вычисления зейделевой величины (он обозначает ее через $v_i^{(k)}$), а затем вычислять

$$u_i^{(k+1)} = u_i^{(k)} + \beta (v_i^{(k)} - u_i^{(k)}).$$

Это соответствует сверхрелаксации в $100(\beta - 1)\%$. Юнг спрашивает, выбор какого значения β из интервала $1 < \beta < 2$ является наилучшим.

Анализ, связанный с ответом на этот вопрос, во многом протекает аналогично процессу Зейделя, так как систематическая сверхрелаксация процесса является также линейной. Мы имеем

$$Eu^{(k+1)} + Dv^{(k)} + Fu^{(k)} + b = 0 \quad (15.13)$$

и

$$u^{(k+1)} = u^{(k)} + \beta (v^{(k)} - u^{(k)}) = (1 - \beta) u^{(k)} + \beta v^{(k)}.$$

Теперь исключаем $v^{(k)}$ из уравнения (15.13). Так как

$$Du^{(k+1)} = (1 - \beta) Du^{(k)} + \beta Dv^{(k)},$$

то получаем

$$Eu^{(k+1)} + \frac{1}{\beta} Du^{(k+1)} + \left(1 - \frac{1}{\beta}\right) Du^{(k)} + Fu^{(k)} + b = 0. \quad (15.14)$$

Уравнение (15.14) описывает систематическую сверхрелаксацию, подобно тому как уравнение (15.6) описывает процесс Зейделя. (Заметим, что уравнение (15.14) сводится к уравнению (15.6) при $\beta = 1$.) Быстрота сходимости процесса Юнга оценивается наибольшим из модулей корней σ_i следующего определяющего уравнения [оно подобно второму из уравнений (15.11)]:

$$\left| \sigma E + \frac{\sigma}{\beta} D + \left(1 - \frac{1}{\beta}\right) D + F \right| = 0.$$

Юнг ограничивает свое внимание положительно определенными симметрическими матрицами со свойством A (см. § 15.3). Для этих матриц и для определенного расположения соответствующих уравнений показано, что максимум для $|\sigma_i|$ окажется наименьшим, если выбрать $\beta = 2/[1 + (1 - \mu_1)^{1/2}]$, где μ_1 — наибольший из корней уравнений (15.11). Следовательно, это значение β и является оптимальным для сверхрелаксации.

Для иллюстрации этого метода в табл. 15.3 показаны результаты решения уравнений (15.1) методом систематической сверхрелаксации. В соответствии со значением $\mu_1 = 0,9045$ берем $\beta = 1,528$.

В таблице 15.3 приведено также отношение наибольших погрешностей, как и в табл. 15.2. Из табл. 15.3 трудно усмотреть закономерность в изменении этого отношения, однако в действительности имеет место сходимость его к значению $\beta - 1 = 0,528$. При этом значении для получения точности в один десятичный знак требуется всего лишь 3,6 цикла. Следовательно, значение $\beta = 1,528$ оказывается более подходящим, чем значение $\beta = 1$ в методе Зейделя.

Юнг и Лерч [42] показали, что изложенный выше метод действительно пригоден для электронных вычислительных машин. Наилучшее значение β , которое было бы трудно вычислить, на самом деле было легко оценено методом подбора и поправок.

На этом мы заканчиваем обзор одного из типичных примеров развития численного анализа. Существует ряд других методов, которые также могли бы быть избраны в качестве примера.

§ 15.6. Вычисление частот колебаний. Предположим, что необходимо исследовать динамику системы, изображенной на рис. 15.1, т. е. поведение различных грузов при движении. Пусть $u_1, \dots, u_{10}$ — положения статического равновесия, т. е. решение системы (15.1). Пусть $u_1 + x_1, \dots, u_{10} + x_{10}$ — положения грузов при движении, где $x_1, \dots, x_{10}$ зависят от времени t .

Уравнение движения i -го груза имеет вид

$$-m(\ddot{u}_i + \ddot{x}_i) = -mg + \frac{T}{d}(-u_{i-1} - x_{i-1} + 2u_i + 2x_i - u_{i+1} - x_{i+1}),$$

Таблица 15.3

РЕШЕНИЕ УРАВНЕНИЙ (15.1) МЕТОДОМ СИСТЕМАТИЧЕСКОЙ СВЕРХРЕЛАКСАЦИИ

Отношение наибольших ошибок	$u^{(\infty)}$	5	9	12	14	15
	$u^{(0)}$	4	8	12	14	15
0,34	$u^{(1)}$	4,76	9,34	12,26	14,20	15,31
0,51	$u^{(2)}$	5,39	9,32	12,26	14,33	15,34
0,54	$u^{(3)}$	5,039	9,060	12,161	14,209	15,140
0,76	$u^{(4)}$	5,025	9,110	12,159	14,119	15,108
0,74	$u^{(5)}$	5,071	9,118	12,097	14,093	15,085
0,52	$u^{(6)}$	5,053	9,052	12,059	14,061	15,048
0,59	$u^{(7)}$	5,012	9,027	12,036	14,032	15,024
0,67	$u^{(8)}$	5,014	9,024	12,024	14,020	15,018
0,58	$u^{(9)}$	5,011	9,014	12,013	14,013	15,010
0,56	$u^{(10)}$	5,0049	9,0063	12,0079	14,0068	15,0051

где $u_0 = x_0 = u_{11} = x_{11} = 0$. Далее, $\ddot{u}_i = 0$, и в соответствии с уравнениями статического равновесия

$$-mg + \frac{T}{d}(-u_{i-1} + 2u_i - u_{i+1}) = 0.$$

Следовательно,

$$-m\ddot{x}_i = \frac{T}{d}(-x_{i-1} + 2x_i - x_{i+1}). \quad (15.15)$$

Выберем систему единиц так, чтобы $T/md = 1$. Если следовать обычному приему разделения переменных, приняв $x_i = \tau y_i$, где τ зависит только от t , а y_i — только от i , то уравнение (15.15) примет вид

$$-y_i \frac{\ddot{\tau}}{\tau} = -y_{i-1} + 2y_i - y_{i+1}.$$

На основании этого можно показать, что $-\ddot{\tau}/\tau$ есть положительная постоянная, например λ . Это соответствует простой гармонической зависимости от времени $\tau = \sin \sqrt{\lambda}(t - \varphi)$, где φ — запаздывание по времени.

Для определения тех частот, с которыми грузы могут совместно осуществлять простые гармонические колебания, мы должны решить относительно λ систему

$$-y_{i-1} + 2y_i - y_{i+1} = \lambda y_i \quad (15.16)$$

или, в матричной форме,

$$Ay = \lambda y. \quad (15.17)$$

Это есть задача определения собственных значений, подобная рассмотренной в § 15.5.

Часто требуется узнать только наименьшее значение λ из удовлетворяющих уравнению (15.16) или (15.17). Чтобы вычислить такое значение λ , мы сначала предположим, что в силу симметрии $y_6 = y_6$, $y_4 = y_7$ и т. д., что характеризует основной тип колебаний. После этого мы получим пять уравнений для определения $y_1, \dots, y_5$:

$$\begin{aligned} (2 - \lambda)y_1 - y_2 &= 0 \\ -y_1 + (2 - \lambda)y_2 - y_3 &= 0 \\ -y_2 + (2 - \lambda)y_3 - y_4 &= 0 \\ -y_3 + (2 - \lambda)y_4 - y_5 &= 0 \\ -y_4 + (1 - \lambda)y_5 &= 0 \end{aligned} \quad (15.18)$$

Эта система уравнений является уже билинейной, а не линейной. Однако она является линейной относительно y_i при фиксированном λ , и, следовательно, один путь для решения уравнений (15.18) состоит в том, что λ принимается равным некоторому значению, например 0,5. При этом фиксированном значении λ осуществите ряд релаксационных шагов с целью решения системы (15.18) относительно $y_1, \dots, y_5$. Остановитесь и произведите переоценку значения λ по формуле Рэлея

$$\lambda = \frac{\sum_{i=1}^5 (Ay)_i y_i}{\sum_{i=1}^5 y_i^2},$$

где $(Ay)_i$ является i -й компонентой Ay . Затем вновь проведите релаксационный процесс, подставив в уравнения (15.18) новое значение λ , и т. д. Процесс будет сходящимся.

Существует много модификаций этого процесса, включающих использование других методов для приближенного решения системы (15.18) при фиксированном λ . (См. статьи Айткена [43], Бодевига [44] и Семендяева [46].) Многообещающим является новый метод, предложенный Ланчосом [45].

Определение высших частот требует некоторых дополнительных приемов, и здесь релаксационный метод не всегда оказывается сходящимся в силу того, что матрица $A - \lambda I$ не обязана быть положительно определенной (см. § 15.5).

В этом параграфе речь шла лишь об одной нелинейной задаче, поддающейся методу релаксации благодаря тому, что она может быть сведена к последовательности линейных задач. Существует много других нелинейных задач, которые также могут быть разбиты на линейные подпроблемы и решены итеративно с помощью релаксационных методов или иным путем.

§ 15.7. Заключение. Слово „релаксация“ имеет для инженеров и математиков по крайней мере два значения. В узком смысле оно означает класс итерационных методов, предназначенных для решения систем линейных алгебраических уравнений. Как отмечалось в § 15.2, этот класс методов характеризуется последовательным сведением к нулю наихудшего из остатков.

В широком смысле слова релаксация означает приближенное описание широкого класса физических проблем с помощью систем линейных уравнений с последующим их решением. В этом смысле слова релаксация охватывает большую область искусства вычисления. Искусство-вычисления представляет собой использование всех видов вычислительных приемов, строго обоснованных или же не получивших до сих пор такого обоснования, с целью получения окончательных, т. е. доведенных до числа ответов на задачи численного характера.

Наряду с искусством вычисления существует наука о численном анализе. Работающие в этой области исследователи рассматривают и доказывают теоремы об известных вычислительных методах, разрабатывают новые приемы и испытывают их в сотрудничестве с вычислителями. Современный численный анализ часто ориентируется на применение новых электронных цифровых машин.

ЛИТЕРАТУРА

Избранные книги по релаксации, линейным задачам
и численному анализу

1. Allen D. N. de G., *Relaxation Methods*, McGraw-Hill, New York, 1954.
2. Collatz L., *Eigenwertaufgaben mit technischen Anwendungen*, Akademische Verlagsgesellschaft, Leipzig, 1949.
3. Collatz L., *Numerische Behandlung von Differentialgleichungen*, Springer-Verlag, Berlin, 1951.
4. Dwyer P. S., *Linear Computations*, John Wiley, New York, 1951.
5. Engineering Research Associates, *High-speed Computing Devices*, McGraw-Hill, New York, 1950.

6. Фадеева В. Н., Вычислительные методы линейной алгебры, Гостехиздат, М., 1950.
7. Фрезер Р., Дункан В. и Коллар А., Теория матриц и ее приложения к дифференциальным уравнениям и динамике, ИЛ, М., 1950.
8. Grinter L. E. (ed), Numerical Methods of Analysis in Engineering, Macmillan, New York, 1949.
9. Hartree D. R., Numerical Analysis, Clarendon Press, Oxford, 1952.
10. Хаусхолдер А. С., Основы численного анализа, ИЛ, М., 1956.
11. Канторович Л. В. и Крылов В. И., Приближенные методы высшего анализа, изд. 3, Гостехиздат, М. — Л., 1950.
12. Милн В. Э., Численный анализ, ИЛ, М., 1951.
13. Милн В. Э., Численное решение дифференциальных уравнений, ИЛ, М., 1955.
14. Paige L. J. and Taussky Olga (eds), Simultaneous Linear Equations and the Determination of Eigenvalues, Government Printing Office, Washington, D. C., 1953.
15. Perlis S., Theory of Matrices, Addison-Wesley, Cambridge, Mass., 1952.
16. Shaw F. S., An Introduction to Relaxation Methods, Dover Publications, New York, 1953.
17. Southwell R. V., Relaxation Methods in Engineering Science, Oxford Univ. Press, London, 1940.
18. Southwell R. V., Relaxation Methods in Theoretical Physics, Clarendon Press, Oxford, 1946.
19. Zurmühl R., Matrizen. Eine Darstellung für Ingenieure, Springer-Verlag, Berlin, 1950.

Избранные журнальные статьи

Содержание таких четырех периодических изданий, как: Proceedings of Royal Society of London, Series A; Philosophical Transactions of Royal Society of London, Series A; Quarterly Journal of Mechanics and Applied Mathematics; Mathematical Tables and Other Aids to Computation свидетельствует о наличии обширной и все возрастающей литературы по релаксационным методам и связанным с ними вопросам.

§ 15. 2.

20. Cross H., Analysis of Continuous Frames by Distributing Fixed-end Moments, Paper No 1793, Trans. ASCE, **96**, 1—10 (1932). (Воспроизведено в книге Grinter [8].)
21. Fox L., A Short Account of Relaxation Methods, Quart. J. Mech. Appl. Math., **1**, 253—280 (1948).
22. Gauss C. F., Brief an Gerling, Dec. 26, 1823, Werke, B. 9, S. 278—281. [Англ. перевод: Forsythe G. E., Math. Tables Aids Comput., **5**, 255—258 (1951)].
23. Николаева М. К., О релаксационном методе Саусвелла, Труды матем. института им. Стеклова, **23**, 160—182 (1949) (критический обзор).

24. Temple G., The General Theory of Relaxation Methods Applied to Linear Systems, Proc. Roy. Soc. London, Ser. A, 169, 476—500 (1939).
§ 15.3.
25. Курант Р., Фридрихс К. и Леви Х., О дифференциальных уравнениях математической физики в частных производных, Успехи матем. наук, VIII, 125—160 (1941).
26. Курант Р., гл. 5 этой книги.
27. Higgins Th. J., A Survey of the Approximate Solution of Two-dimensional Physical Problems by Variational Methods and Finite-difference Procedures, гл. 10 из книги Grinter [8].
28. Woods L. C., The Relaxation Treatment of Singular Points in Poisson's Equation, Quart. J. Mech. Appl. Math., 6, 163—185 (1953).
§ 15.4.
29. Bodewig E., Bericht über die verschiedene Methoden zur Lösung eines Systems linearer Gleichungen mit reellen Koeffizienten, I—V, Nederl. Akad. Wetensch. Proc., 50, 930—941, 1104—1116, 1285—1295 (1947) и 51, 53—64, 211—219 (1948). Некоторые статьи в Nederl. Akad. Wetensch. Indag. Math., 9, 441—452, 518—530, 611—621 (1947) и 10, 24—35, 82—90 (1948).
30. Forsythe G. E., Solving Linear Equations Can Be Interesting, Bull. Amer. Math. Soc., 59, 299—329 (1953).
31. Forsythe G. E., Tentative Classification of Methods and Bibliography on Solving Systems of Linear Equations, стр. 1—28 в книге Paige and Taussky [14].
32. Hotelling H., Some New Methods in Matrix Calculation, Ann. Math. Statist., 14, 1—34 и 440—441 (1943).
33. Лемер Д. Х., гл. 17 этой книги.
34. Пайпс Л., гл. 12 этой книги.
35. Taussky Olga and Todd J., Systems of Equations, Matrices and Determinants, Math. Mag., 26, 9—20 и 71—88 (1952).
36. Томпкинс К. Б., гл. 16 этой книги.
§ 15.5.
37. Frankel S. P., Convergence Rates of Iterative Treatments of Partial Differential Equations, Math. Tables Aids Comput., 4, 65—75 (1950).
38. Ostrowski A. M., On the Linear Iteration Procedures for Symmetric Matrices, Nat. Bur. Standards Rept., 1952.
39. Reich E., On the Convergence of the Classical Iterative Method of Solving Linear Simultaneous Equations, Ann. Math. Statist., 20, 448—451 (1949).
40. Seidel L., Ueber ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherungen aufzulösen, Abh. Bayer. Akad. Wiss. Math.-Phys. Kl., 11, 81—108 (1874).
41. Young D., Iterative Methods for Solving Partial Difference Equations of Elliptic Type, Trans. Amer. Math. Soc., 76, 92—111 (1954).

-
- 42. Young D. and Lerch F., The Numerical Solution of Laplace's Equation on ORDVAC, Ballistics Research Laboratories Memorandum Report 708, Aberdeen Proving Ground, Aberdeen, Md., 1953.
§ 15.6.
- 43. Aitken A. C., Studies in Practical Mathematics II. The Evaluation of the Latent Roots and Latent Vectors of a Matrix, Proc. Roy. Soc. Edinburgh, Sect. A, 57, 269—304 (1937).
- 44. Bodewig, E. Bericht über die Methoden zur numerischen Lösung von algebraischen Eigenwertproblemen I, II, Atti Sem. Mat. Fis. Univ. Modena, 4, 133—193 (1950) и 5, 3—39 (1951).
45. Lanczos C., An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators, J. Res. Nat. Bur. Standards, 45, 255—282 (1950).
- 46. Семендяев К. А., Определение связанных корней и инвариантных множеств матриц посредством итераций, Прикл. матем. и механика, 7, 293—322 (1943).

Глава 16

МЕТОДЫ БЫСТРОГО СПУСКА

ЧАРЛЬЗ Б. ТОМПКИНС

Калифорнийский университет, Лос-Анжелес

§ 16. 1. Введение. Главы этой книги распадаются, грубо говоря, на три группы. В одних главах рассматриваются общие теоремы, на которых могут основываться положения, относящиеся к общему поведению физических систем — природе некоторых колебаний или общей форме физически устойчивого состояния. Другие главы имеют дело с формулировкой задач, что облегчает их эффективное математическое исследование. Наконец, в последних главах, составляющих третью часть этой книги, рассматриваются вопросы, связанные с получением детальных численных решений конкретных задач (сущность которых могла быть рассмотрена в главах первой группы, а формулировка могла быть дана в главах второй группы).

Данная глава принадлежит к главам третьей группы. В ней некоторые сведения, известные каждому математику, но не сформулированные нигде в единообразном изложении, будут применены к ряду задач, важных для физики и техники. Для объяснения этих сведений требуется знание многомерной геометрии, т. е. геометрии четырех и более измерений. В конечном счете эта геометрия основана на алгебраических и аналитических понятиях и положениях, но используемые здесь идеи подсказаны аналогией с теоремами, известными и широко применяемыми в обычной евклидовой геометрии двух и трех измерений. Потребуется также знание дифференциального и интегрального исчисления; одна из идей здесь состоит в замене сложных функций более простыми (обычно линейными), которые настолько хорошо аппроксимируют сложные функции вблизи точки аппроксимации, что на изучении этих более простых функций могут быть основаны утверждения, справедливые и для сложных функций в небольшой окрестности точки аппроксимации.

В некоторых случаях потребуется знание отдельных формул. Те формулы из анализа, которые мы используем, известны большинству читателей, и, во всяком случае, они могут быть найдены в обычных учебниках анализа. Формулы же многомерной геометрии, на которые будут опираться некоторые положения, не так просто найти в учебниках, и многие из них будут здесь установлены, чтобы иметь необходимые основания для дальнейшего изложения.

Важно заметить, что многие из геометрических понятий, с помощью которых здесь предполагается вести исследование, являются очень глубокими, и читателя следовало бы предупредить, что предлагаемые здесь численные способы решения требуют больших знаний по математике, чем знание элементарной арифметики. Разработка эффективных численных методов является важной отраслью математики, которая глубоким и сложным образом связана со всеми другими математическими знаниями. Эта отрасль математики между прочим в последнее время изучалась с новой энергией и настойчивостью в связи с созданием больших быстродействующих счетных машин¹.

Методы, которые будут здесь рассматриваться, применяются для большого класса таких задач, которые можно решать при помощи минимизации некоторой функции многих переменных или минимизации интеграла. Принцип наименьшего действия и другие подобные минимальные принципы физики сводят многие физические и инженерные задачи к этому классу. Действительно, некоторые задачи могут быть сведены к таким задачам, в которых ищется минимум функции конечного числа переменных, если известно, что применим метод Рэлея—Ритца с конечным числом основных функций. В других случаях эти физические принципы обычно приводят к задачам вариационного исчисления, в которых должен быть минимизирован интеграл.

Наша цель заключается в том, чтобы изложить, не вдаваясь в подробности, один метод нахождения численных решений задач такого типа. Этот метод состоит в том, что во время решения задачи моделируется сама физическая система; однако в большинстве случаев имеют место некоторые абстракции. В физических задачах, которые удобнее всего рассматривать этим методом, процесс нахождения решения имитирует движение физической системы к положению равновесия от некоторого неравновесного состояния. Это движение происходит под влиянием сил, действующих в системе и дополняемых удобными демпфирующими силами (удобными скорее с точки зрения вычислителя, но не с точки зрения физической системы). Короче говоря, решение по этому методу напоминает движение шарика, скатывающегося с вершины возвышенности.

Сила этого метода заключается отчасти в его удобстве для вычислений. Без употребления каких-либо искусственных приемов вычислитель может получить некоторое приближенное решение своей задачи и он будет, вообще говоря, в состоянии в дальнейшем улучшить это приближенное решение.

Такой способ имеет, однако, преимущества более ценные, чем простое удобство. Настоящие математические преимущества этого

¹ См. работу Л. В. Канторовича «Функциональный анализ и прикладная математика», Успехи матем. наук, III, вып. 6 (1948). В гл. III изложен метод быстрого спуска. — *Прим. ред.*

способа заключаются в наличии более быстрых средств для улучшения приближенного решения задачи, чем приемы дробления сетки в конечно-разностных методах. Эти последовательные улучшения приближенного решения требуются во многих теоретических исследованиях труднейших вариационных задач, так как метод сеток может оказаться здесь непригодным. В таких задачах методы быстрого спуска являются необходимыми для любого конструктивного решения. В заключение можно отметить, что задачи, в которых исследование минимальных решений с помощью метода сеток невозможно, приобретают, по-видимому, все возрастающее практическое значение в технике. К ним относятся, например, задачи с двойными интегралами из классического вариационного исчисления.

Следовало бы отметить некоторые слабые места этого метода. Наиболее важным из них является некоторая неуверенность при его применении, вытекающая из того, что не все математические теоремы, на которых этот метод основан, в настоящее время доказаны. Однако это обстоятельство не является решающим, как это кажется с первого взгляда, потому что обычно имеются пути проверки решения, если таковое получено, и в практически важных случаях требуемые теоремы часто могут быть легко установлены.

§ 16.2. Многомерная аналитическая геометрия и метод Качмажа для линейных уравнений. Положения, которые здесь будут рассматриваться и использоваться, являются в конечном счете алгебраическими или аналитическими. Однако в прежнее время соображения, приводящие к этим положениям, были в основном геометрическими, и, во всяком случае, геометрическая терминология оказывается полезной для облегчения краткого описания используемых нами приемов.

Изложение будет здесь вестись в координатной форме, так как при этом упрощается рассмотрение задач на минимум. Однако нет оснований требовать, чтобы изложение многомерной геометрии велось только таким образом. Можно напомнить, что двух- и трехмерная евклидова геометрия обычно излагается в средней школе на основе аксиом, выдвинутых первоначально Евклидом. В высшей школе изложение геометрии повторяется, но уже в координатной форме, введенной Декартом.

Построение, полученное при координатном подходе, эквивалентно построению, создаваемому непосредственно на основании аксиом Евклида, но применять первое обычно намного проще. Однако аксиомы Евклида, несомненно, более привлекательны для большинства людей, так как они нагляднее при рассмотрении реальной геометрической ситуации, которую Евклид пытался описать абстрактно. Эквивалентность этих двух построений доказывается, например, в учебнике Л. П. Эйзенхарт [2]. Можно отметить между прочим, что математики уточняли и делали более ясной систему аксиом Евклида

в течение 2000 лет, прошедших со времен Эвклида; Эйзенхарт применяет систему, принадлежащую в основном немецкому математику Гильберту.

Здесь, вместо видоизменения аксиом, которые определяют размерность объектов, рассматриваемых в системе Гильберта, мы будем просто увеличивать число координат, используемых для описания точки в декартовой системе, и выведем формулы, вполне аналогичные формулам двух- и трехмерной геометрии. Одной из замечательных черт аналитической геометрии является подобие формул, встречающихся в обеих этих геометриях, если не принимать во внимание понятие наклона (которое существенно в случае двух измерений и не так существенно для трех измерений).

Это подобие лучше всего иллюстрируется при обозначении координат точки буквой с индексом, например x_i , вместо обычно употребляемых обозначений (x, y) и (x, y, z) для координат в двух и трех измерениях соответственно. Если обозначение x_i предназначено для представления точек в пространстве двух измерений, то индексу i следует давать два значения 1 и 2, если же это обозначение предназначено для представления точек в пространстве трех измерений, то индекс i должен принимать три значения 1, 2 и 3. Координаты соответствующих точек будут (x_1, x_2) и (x_1, x_2, x_3) . Принимая во внимание это условие, можно напомнить из начального курса аналитической геометрии следующие положения, относящиеся к точкам, прямым и плоскостям (при этом предполагается, что координаты являются ортогональными и что длины измеряются вдоль всех осей в одном и том же масштабе):

а) Направление и длина направленного отрезка, идущего из точки с координатами x_i к точке с координатами y_i , полностью определяются при помощи его проекций

$$u_i = y_i - x_i.$$

б) Длина любого отрезка с проекциями u_i равна L , где

$$L^2 = \sum_i u_i^2, \quad L \geq 0.$$

Здесь символ $\sum_i$ означает, что операцию справа от этого символа следует выполнить отдельно для каждого рассматриваемого значения i , а затем сложить результаты.

в) Два отрезка параллельны и одинаково направлены тогда и только тогда, когда их проекции пропорциональны и множитель пропорциональности положителен; они параллельны и противоположно направлены тогда и только тогда, когда их проекции пропорциональны и множитель пропорциональности отрицателен; они параллельны тогда и только тогда, когда их проекции пропорциональны,

причем множитель пропорциональности может быть произвольным. Таким образом, если u_i и v_i — проекции отрезков, то последние параллельны тогда и только тогда, когда

$$v_i = k u_i$$

для некоторого k , не зависящего от i ; они одинаково направлены, если k положительно, и противоположно направлены, если k отрицательно. Если k обращается в нуль (единственная оставшаяся возможность), то все числа v_i необходимо будут нулями и тогда два конца предполагаемого отрезка совпали бы — такое вырождение не допускается.

Приведенные выше положения говорят, помимо всего прочего, о том, как найти конечную точку отрезка, проведенного из заданной начальной точки при условии, что его длина и направление (также и в смысле знака) совпадают с длиной и направлением заданного отрезка. Следует напомнить, что прямая образуется из точек, которые являются концами отрезков, начинающихся в некоторой фиксированной точке и параллельных некоторому фиксированному направлению. В трехмерной геометрии плоскость есть совокупность точек, которая содержит каждую точку любой прямой, проведенной через две точки этой совокупности, не содержит по крайней мере одной точки пространства и является максимальной совокупностью точек, удовлетворяющих этим двум требованиям. В двумерной геометрии этим требованиям удовлетворяет прямая, и естественно, что некоторые свойства являются общими для прямых в двух измерениях и плоскостей в трех измерениях, в то время как другие свойства являются общими для прямых в двух измерениях и прямых в трех измерениях. Геометрический объект (точка в одном измерении, прямая в двух измерениях, плоскость в трех измерениях и, обобщая, плоское подпространство $n - 1$ измерений в пространстве n измерений), обладающий указанными тремя свойствами, мы будем называть *гиперплоскостью*. Это позволит избежать некоторого неудобства, связанного с ролью размерности пространства, в которое погружен объект. После этих предварительных замечаний можно закончить перечень результатов, напоминаемых из аналитической геометрии двух и трех измерений.

г) Координаты x_i точек прямой, проходящей через точку с координатами p_i и содержащей отрезок (или параллельной отрезку), проекции которого u_i , удовлетворяют уравнениям

$$\frac{x_i - p_i}{u_i} = \frac{x_j - p_j}{u_j}, \quad (*)$$

причем любая точка с координатами x_i , удовлетворяющая всем уравнениям (*), принадлежит этой прямой. Здесь i и j — два любых значения индексов у координат, а нулевое значение знаменателя истолковывается так, что соответствующий числитель также должен

обращаться в нуль. Числа u_i называются направляющими коэффициентами прямой (*).

д) Два отрезка, проекции которых u_i и v_i , перпендикулярны (т. е. они имеют перпендикулярные направления, хотя и не обязательно пересекаются) в том и только в том случае, если

$$\sum_i u_i v_i = 0.$$

е) Любая гиперплоскость может быть представлена уравнением вида

$$\sum_i a_i x_i = b,$$

где a_i и b являются действительными числами, причем по крайней мере одно из чисел a_i отлично от нуля. Это представление имеет тот смысл, что координаты любой точки на этой гиперплоскости удовлетворяют этому уравнению, а координаты точек, не лежащих на ней, не удовлетворяют ему.

ж) Отрезки, проекции которых равны a_i , перпендикулярны ко всем прямым, содержащимся в гиперплоскости, заданной уравнением

$$\sum_i a_i x_i = b,$$

а у любого отрезка, перпендикулярного всем прямым, лежащим в этой гиперплоскости, проекции пропорциональны числам a_i .

Замечательным здесь является то, что такое представление может быть дано в приведенной выше единой форме без указания числа измерений объемлющего пространства. Это сразу же наводит на мысль о том, что может представлять интерес обобщение таких представлений на большее число измерений. Формально это можно осуществить, отказавшись от ограничений на индекс i и разрешив этому индексу изменяться в случае n -мерной геометрии от 1 до n , что потребует некоторой аксиоматизации и распространения доказательств теорем на большее число измерений. Последнее (как было замечено выше) может быть сделано или путем распространения аксиом с сохранением употреблявшейся Евклидом терминологии или при помощи рассмотрения аксиом в терминах аналитической геометрии. Аксиомы в терминах аналитической геометрии были установлены О. Вебленом [12]. Разнообразные вопросы аналитической геометрии многих измерений рассмотрены в элементарном учебнике Эйзенхарта [2]. Жордан [5] излагает аналитическую геометрию многих измерений более полно; некоторые из соотношений между аналитической геометрией многих измерений и матричными операциями, широко применяемыми в технике, можно найти в книге О. Шрейера и Е. Шпернера [8].

Когда все это сделано, можно высказать следующий принципиальный результат: структура любых двух плоских евклидовых пространств

с одним и тем же числом измерений одинакова независимо от того, как они определены. Например, гиперплоскость в объемлющем пространстве n измерений есть $(n-1)$ -мерное пространство, эквивалентное $(n-1)$ -мерному объемлющему пространству с введенными в нем $n-1$ координатными осями. Кроме того, каждая совокупность k точек содержится в плоском евклидовом пространстве $k-1$ измерений; так, например, каждая пара точек лежит на прямой, каждая тройка точек лежит на обычной евклидовой плоскости (для которой справедливы все теоремы евклидовой планиметрии) и т. д.

С. Качмаж [6] использовал эти идеи для задачи отыскания приближенных решений систем, состоящих из большого числа линейных уравнений. Он заметил, что задача нахождения решения системы линейных уравнений эквивалентна задаче об определении координат точки пересечения гиперплоскостей, представляемых уравнениями этой же системы.

Качмаж свел эту алгебраическую задачу к геометрической следующим образом. Пусть даны алгебраические уравнения

$$\sum_i a_{hi} x_i = b_h, \quad (16.1)$$

где второй индекс h введен для того, чтобы различать уравнения. Для каждого значения h будем обозначать через H_h гиперплоскость, представляемую соответствующим уравнением. Теперь рассмотрим некоторую точку с координатами p_i . Методом, использующим приведенные выше положения, можно найти другую точку с координатами q_i , являющуюся основанием перпендикуляра, опущенного из p_i на любую гиперплоскость H_h . Чтобы найти q_i , надо выразить алгебраически следующие два факта: отрезок от p_i до q_i перпендикулярен к каждой прямой на H_h и точка q_i лежит на H_h . В соответствии с утверждениями (в), (ж) и (е) имеем

$$q_i - p_i = k a_{hi} \quad (16.2)$$

и

$$\sum_i a_{hi} q_i = b_h. \quad (16.3)$$

Здесь следует помнить, что рассматривается только одна гиперплоскость H_h и что значение h должно быть выбрано в соответствии с рассматриваемой гиперплоскостью.

Два последних уравнения легко могут быть разрешены относительно k . Заменим q_i в уравнениях (16.3), выразив их из уравнений (16.2) через p_i (которые предполагаются известными) и через k . В результате получим

$$k = \frac{b_h - \sum_i a_{hi} p_i}{\sum_i a_{hi}^2}.$$

После определения величины k координаты q_i могут быть сразу же получены из уравнений (16.2).

Если точка с координатами r_i есть решение задачи (т. е. если $x_i = r_i$ есть решение заданной системы уравнений), то треугольник с вершинами в точках p_i , q_i и r_i является прямоугольным; в соответствии с приведенными ранее положениями свойства этого треугольника в точности подобны свойствам прямоугольного треугольника, изучаемого в элементарной геометрии. В частности, гипотенуза $p_i r_i$ больше катета $q_i r_i$. Таким образом, независимо от того, как была первоначально взята точка p_i и какая из гиперплоскостей H_h была выбрана для проектирования, точка q_i или совпадает с p_i (если окажется, что p_i лежит на этой выбранной гиперплоскости) или лежит ближе, чем p_i , ко всякой точке пересечения всех гиперплоскостей.

Метод Качмажа состоит в следующем: выбирается произвольная точка p_i и проектируется на первую гиперплоскость, найденная проекция проектируется далее на вторую гиперплоскость, затем процесс проектирования на гиперплоскости продолжается в циклическом порядке, причем после каждого шага мы все ближе и ближе подходим к решению.

Метод Качмажа является одной из простейших иллюстраций метода быстрого спуска. Его можно рассматривать как метод минимизации некоторой функции; такой функцией здесь является расстояние от рассматриваемой точки до ближайшей точки пересечения всех гиперплоскостей. Тот факт, что эта функция не может быть указана явно до решения задачи, не имеет особого значения. Важным обстоятельством здесь является выполнение следующих требований:

Эта функция минимизируется искомым решением. Дается метод, посредством которого значение этой функции может быть уменьшено при движении в пространстве от одной точки к другой. Это уменьшение является достаточно быстрым, чтобы обеспечить стремление функции к ее минимуму, когда количество шагов увеличивается.

Выполнение третьего требования в методе Качмажа не установлено в данном здесь изложении. Однако, как показала Качмаж, легко установить, что и это требование выполняется.

Многие задачи могут быть приведены к форме, делающей их доступными для решения с помощью процесса, в котором удовлетворяются три сформулированных выше требования. Часто это можно сделать несколькими способами; ряд способов такого преобразования задачи решения системы линейных уравнений будет рассмотрен в дальнейшем. Вычислительная сторона задачи заключается при этом в построении подходящей функции и отыскании удобного метода быстрого спуска. Вопросы, которые должны быть здесь рассмотрены, будут, вероятно, более ясными, если излагать их, иллюстрируя примерами.

§ 16.3. Быстрый спуск и дифференциальные уравнения быстрого спуска для функций, определенных в многомерном евклидовом пространстве. В этом параграфе будет разобран один вопрос, относящийся к быстрейшему спуску в евклидовом пространстве. Мы будем рассматривать функцию $f(x)$; переменные будем обозначать через x_i , опуская индекс i в функциональном обозначении, за исключением тех случаев, когда он должен быть записан явно. Пусть индекс i меняется от 1 до n . Будем считать, что функция $f(x)$ задана в каждой точке n -мерного евклидова пространства; координаты точки являются аргументами x_i этой функции. Можно поставить вопрос о том, какое направление движения в некоторой точке p_i евклидова пространства дает наиболее быстрое уменьшение данной функции?

Введем здесь понятие кривой в евклидовом пространстве. В частности, кривую можно рассматривать как траекторию движущейся точки, координаты которой даются как функции времени. Кривая, начинающаяся в некоторой точке с координатами p_i в момент времени $t=0$, может быть описана путем задания ее координат в любой момент t при помощи уравнений

$$x_i = x_i(t), \text{ причем } x_i(0) = p_i, \quad t \geq 0.$$

В любой момент времени t эту кривую можно аппроксимировать прямой линией, проходящей через $x_i(t)$ и имеющей направляющие коэффициенты, равные значениям производных dx_i/dt , вычисленным в данный момент времени. Для точек, соответствующих моментам, близким к рассматриваемому, данная прямая аппроксимирует кривую более точно, чем любая другая. Эта прямая называется *касательной* к кривой, а производные от координат по времени дают приближенно среднюю скорость точки за малый промежуток времени вблизи рассматриваемого момента, и они называются *составляющими или компонентами мгновенной скорости* в этот момент времени.

Соотношение между длиной дуги и координатами точек кривой для двух и трех измерений может быть записано в общей форме, не зависящей от числа измерений. Эта формула легко распространяется и на многомерные пространства. Если параметр s представляет собой длину дуги, отсчитываемую от точки p_i до другой точки на кривой, соответствующей моменту времени t , то связь между s , t и функциями $x_i(t)$ можно записать так:

$$\frac{ds}{dt} = \left[\sum_i \left(\frac{dx_i}{dt} \right)^2 \right]^{1/2}, \quad s = 0 \text{ при } t = 0. \quad (16.4)$$

Длина дуги является пределом длин ломаных линий, вписанных в кривую. Вывод этой формулы представляет собой простое распространение вывода, приводимого в элементарных учебниках анализа для кривых в двух и трех измерениях.

Поставленный выше вопрос можно теперь выразить в аналитической форме. Рассмотрим кривую, начинающуюся в точке p_i . Другие точки можно обозначать или временным параметром t , или, если это более удобно, параметром s , представляющим собой длину дуги кривой, отсчитываемую от точки p_i . Выбор величины t или s определяет некоторую точку, и в каждой такой точке функция $f(x)$ имеет определенное значение. Следовательно, имеет смысл говорить о скорости изменения $f(x)$ в данной точке p_i в зависимости от s вдоль данной кривой.

Эта скорость может быть найдена по обычным формулам анализа. Предположим, что при $t=0$ не все производные dx_i/dt равны нулю (т. е. скорость движущейся точки не равна нулю в тот момент, когда она проходит через p_i). Тогда скорость изменения x_i по s можно записать так:

$$\frac{dx_i}{ds} = \frac{dx_i}{dt} \frac{dt}{ds} = \frac{dx_i/dt}{ds/dt}. \quad (16.5)$$

Скорость изменения $f(x)$ по s выразится следующим образом:

$$\frac{df}{ds} = \sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{ds}. \quad (16.6)$$

Таким образом, для определения скорости изменения $f(x)$ в точке p_i должны быть найдены в этой точке (т. е. при $s=0$ или $t=0$) все нужные производные.

Из формул (16.4) — (16.6) видно, что величина df/ds полностью определяется в некоторой точке величинами dx_i/dt ; это означает, очевидно, что скорость спуска в данный момент зависит только от направления спуска и не зависит от кривизны кривой спуска. Следует отметить, что df/ds не изменится, если все значения dx_i/dt умножить на одну и ту же постоянную величину k .

Будем величины dx_i/dt , определенные в точке p_i , обозначать для данного момента времени через u_i ; при этом возникает простая задача анализа об отыскании таких значений u_i , при которых величина df/ds максимальна. Эта задача может быть решена путем приравнивания нулю всех производных от df/ds по u_i :

$$\frac{\partial (df/ds)}{\partial u_i} = 0,$$

что приводит к написанному ниже несколько сложному уравнению, в котором обозначения u_i заменены на dx_i/dt :

$$\frac{\partial f}{\partial x_i} \sum_j \left(\frac{dx_j}{dt} \right)^2 = \frac{dx_i}{dt} \sum_j \frac{\partial f}{\partial x_j} \frac{dx_j}{dt}.$$

Хотя это уравнение может показаться сложным, решить его легко. Заметим, что первые множители в обеих частях уравнения,

df/dx_i и dx_i/dt соответственно, зависят от индекса i , а вторые множители (суммы по всем значениям индекса) не зависят от рассматриваемого значения i . Следовательно, эти уравнения можно интерпретировать как выражение того, что необходимое условие для быстрого подъема или быстрого спуска из точки p_i состоит в том, что производные dx_i/dt должны быть пропорциональны известным частным производным df/dx_i , определенным в той же точке p_i .

Более детальное исследование написанного выше выражения для df/ds проливает больше света на положение вещей. Однако выкладки при этом становятся несколько громоздкими, поэтому мы для пояснения рассматриваемого вопроса обратимся к численному примеру.

Поставленная первоначально задача является задачей минимизации функции $f(x)$ в n -мерном евклидовом пространстве. В нашем примере положим $n=4$, а функцию $f(x)$ возьмем в виде

$$f(x) = x_1^2 + 3x_2^2 + 5x_3^2 + 8x_4^2. \quad (16.7)$$

Очевидно, что эта функция достигает минимума, равного нулю, в точке $x_i=0$. Однако здесь мы будем игнорировать этот очевидный факт и начнем расчет от точки $x_1=2$, $x_2=3$, $x_3=4$, $x_4=5$. Эта точка в общем случае выше обозначалась как точка с координатами p_i . Частные производные df/dx_i соответственно будут $2x_1$, $6x_2$, $10x_3$ и $16x_4$. То обстоятельство, что частные производные по x_i зависят только от соответствующих переменных x_i и не зависят от других переменных, является следствием того, что нами была взята простая функция частного вида, — это положение не имеет места в общем случае. Благодаря указанному обстоятельству численное решение некоторых дифференциальных уравнений, которые будут рассматриваться далее, упрощается.

В выбранной нами точке p_i частные производные имеют соответственно значения 4, 18, 40 и 80. Уравнения (16.4)—(16.6), на которых основывается расчет df/ds , показывают, что df/ds обращается в нуль в точке p_i вдоль кривых, производные вдоль которых dx_i/dt удовлетворяют уравнению

$$4 \frac{dx_1}{dt} + 18 \frac{dx_2}{dt} + 40 \frac{dx_3}{dt} + 80 \frac{dx_4}{dt} = 0.$$

Гиперплоскость, проходящая через p_i , определяется уравнением с теми же самыми коэффициентами:

$$4x_1 + 18x_2 + 40x_3 + 80x_4 = 4 \times 2 + 18 \times 3 + 40 \times 4 + 80 \times 5.$$

Эта гиперплоскость делит все пространство на две части; вдоль лучей, проведенных из точки p_i (2, 3, 4, 5) в одну из этих частей пространства, величина f , по крайней мере вблизи p_i , увеличивается, а вдоль лучей, проведенных из p_i в другую часть пространства, величина f , по крайней мере вблизи p_i , уменьшается.

Производную можно рассматривать как приближенную меру скорости, с которой f в среднем изменяется на бесконечно малом расстоянии. Такое представление оказывается достаточно точным, чтобы сделанные выше утверждения были справедливы и для некоторого небольшого конечного расстояния, хотя величина этого расстояния, на котором их можно еще достаточно надежно применять, а priori неизвестна.

Общие утверждения относительно функции $f(x)$ и точки p_i , аналогичные уже приводившимся (и идущие несколько дальше), можно вывести из установленных выше формул (16.4)—(16.6), определяющих df/ds . Эти утверждения относятся к частным производным функции $\partial f/\partial x_i$, найденным в точке p_i , и к производным dx_i/dt , вычисленным при $t=0$, что также соответствует точке p_i . Утверждения эти таковы:

а) Быстрейший спуск получается, когда $dx_i/dt = k \partial f/\partial x_i$, где k — отрицательное число. Быстрейший подъем получается при положительном k .

б) Величины dx_i/dt , которые удовлетворяют уравнению

$$\sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} = 0,$$

определяют направления, вдоль которых мгновенная скорость изменения df/ds удовлетворяет уравнению $df/ds = 0$ (эти направления лежат в гиперплоскости, касательной к поверхности уровня f). Величина f при выходе из данной точки вдоль некоторого направления уменьшается, если

$$\sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} < 0,$$

и производные dx_i/dt , удовлетворяющие этому неравенству, дают направление спуска; величина f увеличивается вдоль направления, для которого

$$\sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} > 0.$$

Вектор, проекции которого суть $u_i = \partial f/\partial x_i$, определяет, таким образом, направление быстрейшего подъема. Этот вектор обычно называется *градиентом* функции f . *Отрицательный градиент* указывает направление быстрейшего спуска.

Легко составить дифференциальные уравнения, решением которых являются функции, определяющие в любой начальной точке траекторию быстрейшего спуска. Эти уравнения таковы:

$$\frac{dx_i}{dt} = - \frac{\partial f}{\partial x_i}. \quad (16.8)$$

Здесь предполагается, что функция $f(x)$ известна и что градиент может быть выражен как функция от переменных x_i . В соответствии с обычными теоремами, определяющими существование решений дифференциальных уравнений такого типа, должно существовать единственное решение, начинающееся в данной точке, если только градиент ведет себя достаточно хорошо. Уравнения (16.8) являются дифференциальными уравнениями быстрого спуска. Следует подчеркнуть, однако, что спуск является быстрым по отношению к метрике в евклидовом пространстве, которая была введена произвольно¹⁾. На практике из-за вычислительных трудностей направления быстрого спуска могут и не определяться по этим дифференциальным уравнениям — это положение будет пояснено в дальнейшем. Любое направление, идущее в сторону спуска от гиперплоскости, разделяющей подъем и спуск, является при решении задачи возможным направлением для спуска, если оно может быть легко определено и использовано. Открытым, однако, остается вопрос об эффективных путях использования любого такого найденного направления.

Теперь можно закончить пример, касающийся функции (16.7). Дифференциальные уравнения быстрого спуска будут иметь вид

$$\frac{dx_1}{dt} = -2x_1, \quad \frac{dx_2}{dt} = -6x_2, \quad \frac{dx_3}{dt} = -10x_3, \quad \frac{dx_4}{dt} = -16x_4.$$

Решения этой системы уравнений при начальной точке, имеющей координаты p_i , таковы:

$$x_1 = p_1 e^{-2t}, \quad x_2 = p_2 e^{-6t}, \quad x_3 = p_3 e^{-10t}, \quad x_4 = p_4 e^{-16t}.$$

Ясно, что эта кривая быстро стремится к точке минимума функции, когда t становится большим.

§ 16.4. Замечания о численном решении дифференциальных уравнений быстрого спуска. Желательность больших шагов. Метод сопряженных градиентов Хестенса — Штифеля. Как указывалось выше, известные теоремы существования утверждают, что уравнения (16.8) всегда определяют единственную траекторию быстрого спуска, проходящую через любую начальную точку. Метод, наиболее часто используемый для доказательства таких теорем существования, не всегда, однако, оказывается удобным для численного решения этих уравнений. Во всяком случае, применяется этот метод или нет, численное решение дифференциальных уравнений приводит в лучшем случае к приближенному решению. Если применяется цифровая вычислительная машина, такая как настольная или автоматическая электронная машина, то численное решение

¹⁾ Автор, по-видимому, имеет в виду возможность выбора произвольной римановой метрики. — *Прим. перев.*

получается при помощи интегрирования по шагам. Это означает, что дифференциальное уравнение заменяется некоторым аппроксимирующим разностным уравнением и решение является скорее решением по шагам величины t , имеющим положительную длину, чем решением при непрерывном изменении t . Когда такой способ применяется в методе спуска, это означает, что мы используем тот факт, что линейная аппроксимация, к которой прибегают в вычислениях, представляет собой приближение, справедливое в некотором положительном интервале, и в течение некоторого положительного времени мы движемся по направлению, близкому к направлению быстреего спуска.

Когда применяется такой подход, сразу же возникают два вопроса. Первый вопрос: где гарантия того, что принятые шаги не настолько велики, что они становятся шагами подъема? Второй вопрос таков: если шаги, примененные при численном решении дифференциальных уравнений (16.8), являются действительно шагами спуска, то почему бы не попытаться применить более крупные шаги в том же самом направлении (для определения которого были проделаны значительные вычисления), отказываясь при этом от желания спускаться наиболее быстро (относительно евклидовой метрики) в пользу желания спускаться настолько далеко, насколько это возможно при малом количестве вычислений?

Эти вопросы являются на вид противоречивыми, тем не менее каждый из них требует ответа. Второй вопрос обращает внимание на желательность эффективно использовать любое направление спуска и напоминает о сделанном выше замечании, что желательным является быстрый спуск при малом количестве вычислений.

М. Р. Хестенс и Эдуард Штифель [4] детально исследовали вопрос о том, насколько далеко вдоль найденного направления спуска целесообразно идти, отыскивая минимум функций, связанных с решением системы линейных уравнений. С этой точки зрения метод Качмажа можно рассматривать как метод, в котором направление быстрого, но не быстреего спуска выбирается после рассмотрения связанных с этим вычислений. Недавние исследования показывают, что может оказаться более выгодным продвигаться даже за гиперплоскость, используемую Качмажем, однако законченного исследования того, каков при этом может быть выигрыш, мы не имеем. Об этом кратко упоминалось в гл. 15, как о сверхрелаксации.

Хестенс и Штифель [4] рассмотрели решение системы уравнений (16.1), приведенной выше в связи с методом Качмажа. Они отметили, что эти уравнения могут быть решены путем минимизации квадратичных функций с нулевыми значениями минимумов. Одной из таких функций является функция, использованная Качмажем, — квадрат расстояния от рассматриваемой точки до ближайшего решения. Другой функцией, которая может быть использована и которая может

быть записана явно, является сумма квадратов остатков:

$$f(x) = \sum_h \left(\sum_i a_{hi} x_i - b_h \right)^2.$$

Важность работы Хестенса и Штифеля заключается, однако, не в частных функциях, выбранных для минимизации, а скорее в строгом исследовании величины расстояния, на которое можно продвигаться в одном направлении, прежде чем вычислять новое направление спуска, и в рассмотрении направлений, которые не являются направлениями быстрого спуска. Однако результаты Хестенса и Штифеля относятся только к функциям ограниченного класса, указанного выше.

Квадратичная функция, которая достигает минимума, равного нулю, в некоторой точке r_i , должна выражаться в виде

$$f(x) = \sum_{ij} a_{ij} (x_i - r_i) (x_j - r_j). \quad (16.9)$$

Теперь надо ответить на такой вопрос: может ли быть выбрана точка, а на направлении быстрого спуска из этой точки найдена вторая точка так, чтобы направление из второй точки в точку r_i было перпендикуляром к направлению быстрого спуска из первой точки? Конечно, это можно было бы просто сделать, если бы точка r_i была задана явно, но в практически важных случаях функция задается таким образом, что значение r_i неизвестно (так как это есть искомая точка, в которой функция имеет минимум).

Если на этот вопрос можно ответить утвердительно, то план действий будет таков: на направление спуска из второй точки накладывается связь, требующая, чтобы это направление лежало в гиперплоскости, проходящей через эту точку и перпендикулярной к направлению спуска из первой точки; затем вычисляется третья точка методом, использованным при вычислении второй точки, и процесс продолжается при все увеличивающемся числе связей. Если этот процесс можно таким путем продолжать до конца, то решение задачи (определение точки минимума квадратичной функции или решение системы линейных уравнений) может быть найдено с помощью такого числа шагов, которое не превосходит числа переменных.

Аналитический подход, посредством которого авторы провели первую часть своих расчетов, может быть изложен следующим образом. Производные функции (16.9), если коэффициенты a_{ij} симметричны, можно выразить в виде

$$\frac{\partial f}{\partial x_i} = 2 \sum_j a_{ij} (x_j - r_j). \quad (16.10)$$

Если начальная точка имеет координаты p_i , то любая точка с координатами q_i , лежащая на направлении быстрого спуска из p_i ,

должна иметь координаты

$$q_i = p_i - 2k \sum_j a_{ij} (p_j - r_j). \quad (16.11)$$

Если отрезок от q_i до r_i перпендикулярен к отрезку от p_i до q_i , то необходимо, чтобы выполнялось условие перпендикулярности

$$\sum_i (r_i - q_i) (q_i - p_i) = 0.$$

Предположим теперь, что числа r_i неизвестны, но пусть имеется способ вычислять в любой точке значение функции $f(x)$ и значения ее частных производных. При этом предположении условие перпендикулярности приводит непосредственно к определению множителя пропорциональности k , а следовательно, к координатам новой точки q_i . Значение k определяется равенством

$$k = \frac{2f(p)}{\sum_i \left(\frac{\partial f}{\partial x_i} \right)^2}, \quad (16.12)$$

где все производные также должны быть вычислены в точке p_i . Алгебраические выкладки, приводящие к формуле (16.12), опущены; они являются простыми преобразованиями, основанными на выражениях (16.9) — (16.11).

Этот результат представляет собой утвердительный ответ на первый вопрос Хестенса и Штифеля. Их метод можно использовать, если спуски вдоль направлений, ограниченных связями, могут быть осуществлены одним и тем же способом. Здесь мы сделаем еще только один шаг, так как цель настоящей статьи — отметить особенности методов быстрого спуска, а не решать линейные уравнения. Однако в цитированной работе авторы провели расчеты полностью.

Вопрос о быстром спуске при наличии связей будет рассмотрен в следующем параграфе этой главы. Чтобы завершить описание метода Хестенса—Штифеля, следует отметить, что квадратичная функция на плоскостях связи меньшего числа измерений (полученных после нескольких циклов процесса Хестенса—Штифеля) является по-прежнему квадратичной функцией в евклидовом пространстве в силу принципа, высказанного в § 16.2 и утверждающего, что все плоские подпространства равного числа измерений конгруэнтны. Поэтому соображение, высказанное выше, остается применимым к спускам, проводимым в этом пространстве.

Ясно, что при наложении связей направление спуска не будет направлением быстрого спуска в исходном объемлющем пространстве. С подобностями, не рассматриваемыми здесь, читатель может ознакомиться по оригинальной статье Хестенса и Штифеля.

§ 16.5. Спуск при наличии связей. В § 16.4 был описан метод Хестенса—Штифеля для решения систем линейных уравнений. Там приходилось иметь дело с точкой, в которой требуется проводить спуск при наличии связей. В этом параграфе мы применим один цикл метода Хестенса—Штифеля к численному примеру и рассчитаем спуск при условии связи. На этом простом примере мы проведем рассмотрение более серьезных вопросов, относящихся к быстрому спуску при наличии связи.

Будем решать методом Хестенса—Штифеля численную задачу из § 16.3. Здесь в качестве функции берется функция, определяемая равенством (16.7), а в качестве точки p_i —точка, применявшаяся в § 16.3, т. е. точка (2, 3, 4, 5). Значение функции в этой точке есть $f(p) = 311$, в то время как производные $\partial f / \partial x_i$ равны соответственно 4, 18, 40, 80. Сумма квадратов этих производных равна 8340, а множитель k , найденный по формуле (16.12), равен $k = 622/8340 \approx 0,07458$. Точка q_i , определяемая соотношениями

$$q_i = p_i - k \frac{\partial f}{\partial x_i},$$

после подстановки сюда соответствующих величин будет иметь приблизительно координаты (1,7017; 1,6576; 1,0168; —0,9664). Цель этого вычисления—обеспечить условие, чтобы прямая, идущая из точки q_i в точку минимума функции (в данном случае—начало координат, хотя этот факт и не был использован в расчете), была перпендикулярна к прямой, идущей из p_i в q_i .

В точке q_i частные производные $\partial f / \partial x_i$ имеют соответственно значения 3,4034; 9,9456; 10,1630; —15,4624. (Последний разряд здесь ненадежен, но он приведен, чтобы позволить читателю проверить вычисления.) Первой замечательной особенностью является то, что эти числа, если их интерпретировать как направляющие коэффициенты, не представляют направления, перпендикулярного к направлению первого спуска. Существо метода Хестенса—Штифеля состоит в том, что может быть найдено направление быстрого спуска, гарантирующее решение нашей задачи в результате четырех шагов (один из которых уже сделан), но в рассматриваемой точке q_i это направление не будет направлением быстрого спуска.

Направление, которое нужно использовать, должно быть направлением быстрого спуска, лежащим в гиперплоскости связей, перпендикулярно к направлению предыдущего спуска. Это направление может быть найдено при помощи той же техники проектирования, которая применялась Качмажем. Хотя предпосылки здесь несколько иные, чем прежде, но арифметические действия будут теми же самыми.

Геометрически мы имеем здесь направление быстрого спуска и имеем также набор направляющих коэффициентов, определяющих направление, перпендикулярное к гиперплоскости, в которой

мы должны двигаться. Быстрейший спуск вблизи рассматриваемой точки в этой гиперплоскости должен быть определен при помощи операций анализа, и мы знаем, что эти операции состоят в замене всех функций аппроксимирующими линейными функциями. Мы знаем также, что рассматриваемые направления — направление быстрейшего несвязанного спуска, направление быстрейшего связанного спуска и направление, перпендикулярное к гиперплоскости связи, — лежат в пространстве не более чем трех измерений. Тогда можно рассматривать эту геометрическую задачу в трех измерениях, и из геометрических соображений будет ясно, что направление быстрейшего связанного спуска должно быть ортогональной проекцией направления несвязанного спуска на гиперплоскость связи. Иначе говоря, направление быстрейшего связанного спуска из точки q_i определяется отрезком прямой, проведенным перпендикулярно к направлению, перпендикулярному ко всем прямым, лежащим в гиперплоскости связи. Оно должно получаться из направления быстрейшего несвязанного спуска на основании следующего соображения: направление быстрейшего связанного спуска есть направление отрезка прямой, начальной точкой которого является q_i , а конечная точка получается проектированием на гиперплоскость связи конечной точки другого отрезка прямой, начальная точка которого есть q_i , а конечная точка находится на направлении быстрейшего несвязанного спуска из точки q_i . Непосредственный подсчет приводит к точке с координатами $(-1,7017; -8,2880; -9,1512; 14,4960)$, как к соответствующей точке на направлении быстрейшего несвязанного спуска. Ее проектирование проводится непосредственно по формуле, полученной в связи с методом Качмажа.

В дальнейшие подробности метода Хестенса — Штифеля мы не будем входить. Численные упрощения могут быть найдены в их статье, либо интересующиеся читатели могут продолжить изучение при помощи сведений, почерпнутых из данной статьи, добавив свою собственную догадку. Рассмотренная задача способствовала пониманию путей решения задач связанного спуска. Она является примером того, что может быть достигнуто при использовании практически выгодных больших шагов и крутого спуска по удачно выбранным направлениям, которые могут и не быть направлениями быстрейшего спуска.

Задачи со связями являются обычными в физике и в технике. К числу их относится, например, задача о висящей тяжелой нити определенной длины (см. гл. 4). Здесь интеграл, выражающий потенциальную энергию системы, должен быть минимизирован при условии, что длина нити задана. Исходя из этой задачи, можно построить задачу с конечным числом переменных (как было упомянуто в гл. 15), рассмотрев конечное число грузиков, подвешенных на невесомой нити определенной длины. Все грузики будут стремиться занять положение с наименьшей общей потенциальной энергией, удовлетворяя при этом связям, наложенным длиной нити, на которой они подвешены.

Локально эти задачи напоминают задачу быстрого спуска со связью, которая возникает в методе Хестенса—Штифеля. Однако наложенные связи могут не быть линейными (как это имеет место в задаче Хестенса—Штифеля), и гиперплоскость связи в действительности будет только аппроксимирующей (касательной) гиперплоскостью к гиперповерхности связи.

Аналитически задачу желательно рассматривать при помощи функции $f(x)$, которая должна быть минимизирована при условии, что некоторая другая функция $g(x)$ должна сохранять фиксированное значение, скажем $g(x) = a$.

Если выбрана точка с координатами p_i и если она удовлетворяет связи, то допустимыми (бесконечно малыми) спусками являются те, при которых $f(x)$ убывает, а $g(x)$ остается постоянной. Рассуждение § 16.3 непосредственно применимо и в этом случае. Здесь желательно иметь совокупность таких направлений движения dx_i/dt , которые удовлетворяют соотношениям

$$\sum_i \frac{\partial g}{\partial x_i} \frac{dx_i}{dt} = 0, \quad \sum_i \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} < 0.$$

показывающим, что мгновенная скорость изменения $dg/ds = 0$ и что значение f в этом направлении убывает. Одним весьма простым способом нахождения такой совокупности значений dx_i/dt является проектирование отрицательного градиента f на касательную гиперплоскость к g . Такой путь дает возможность определить направление быстреего спуска при наложенных связях. Аналитически это приводит к системе обыкновенных дифференциальных уравнений, близких по форме к уравнениям (16.8). Любое решение этой системы, полученное на цифровой вычислительной машине, будет решением некоторых аппроксимирующих разностных уравнений, в которых переменная t меняется дискретно.

Последствия такого интегрирования по шагам системы дифференциальных уравнений быстреего связанного спуска представляются более серьезными, чем при интегрировании без связей, ибо можно ожидать, что приближенные решения заметно нарушат условия связи еще задолго до того, как они дадут движения, которые не ведут к быстрому спуску. Поэтому с первого взгляда на вопрос относительно желательности больших шагов можно, по-видимому, ответить так: шаги должны быть короткими для обеспечения условий связи. В общем случае это верно, но по крайней мере в ряде задач вопрос остается открытым, если спуск (по отношению к f) быстр по сравнению с отходом от условий связи и если при осуществлении большого шага или серии шагов условия связи восстанавливаются при помощи движения, предназначенного в первую очередь для ликвидации предыдущих отклонений от связи, а не для ускорения спуска.

§ 16.6. Обобщение метрики и расширение на случай пространств бесконечно большого числа измерений; проблемы вариационного исчисления. В этом параграфе будут сняты два ограничения, которые накладывались в предыдущих параграфах главы. Первое из них относится к метрике объемлющего пространства; нет никакого повода требовать евклидовой метрики в пространстве переменных, происхождение которого не было геометрическим с самого начала. Однако необходимо напомнить, что методы быстрого спуска требуют некоторой метрики для определения меры быстроты спуска. В дальнейшем будет показано, что, основываясь на анализе локального поведения функций, мы придем к более общему понятию быстрого спуска, для которого могут быть установлены удобные формулы и не потребуются никаких новых идей. Второе снимаемое ограничение — это ограничение конечности числа измерений пространства. Последнее соображение оказывается более важным.

Примером выгоды, которую можно получить, рассматривая задачу с неевклидовой метрикой, может служить трехмерная задача, в которой допустимые точки находятся на некоторой поверхности, например на поверхности сферы (см. гл. 4). Быстрый спуск в такой задаче возможен, однако в случае сохранения евклидовой метрики это достигается дорогой ценой. Так как никакая евклидова метрика двух измерений не подходит к сфере (хорошо известная теорема дифференциальной геометрии), то любое построение, основанное на евклидовых метриках, должно быть построением в трех измерениях, в котором связанные спуски лежат на поверхности связи — в данном случае на поверхности сферы. Однако если эту задачу трактовать при помощи параметров на сфере (например, долготы и широты), то она становится двумерной задачей без связей.

На самом деле имеются более серьезные возражения относительно применения евклидовой метрики, чем в только что приведенном примере. Некоторые задачи содержат свою собственную метрику, и для них не всегда возможно найти объемлющую евклидову метрику даже в том случае, когда допустимо включение большего числа измерений и связей. Разбор этого, однако, может завести нас в области дифференциальной геометрии, далекие от занимающих нас здесь вопросов.

Объединяющей теорией для проблем данного типа является тензорное исчисление. Читатели могут познакомиться с ним по книгам, подобным книге И. С. Сокольников [10]¹⁾, в которой естественное введение тензорных понятий, не основанных на евклидовых метриках, сопровождается некоторыми ссылками на физические явления. Теория тензоров опирается на некоторые дифференциальные концепции.

¹⁾ См. также: Кочин Н. Е., Векторное исчисление и начала тензорного исчисления, изд. АН СССР, М., 1951 и Рашевский П. К., Введение в риманову геометрию и тензорный анализ, Гостехиздат, М. — Л., 1936. — *Прим. ред.*

Она основана на серьезной аксиоматике, подобной той, которая приведена в монографии О. Веблена и Дж. Г. К. Уайтхеда [13], но может быть изложена также неформально, как теория пространств, локально аппроксимируемых координатными пространствами, подобными тем, в которых выше определялась евклидова метрика. В частности, важным классом тензорных пространств являлись римановы пространства, т. е. пространства с положительно определенной метрикой, задаваемой фундаментальным метрическим тензором g_{ik} . Важным обстоятельством является здесь то, что спуск в окрестности точки риманова пространства можно аппроксимировать спуском в окрестности этой точки в евклидовом пространстве, касательном к риманову пространству в данной точке, причем эти аппроксимации являются хорошими с точки зрения дифференциального исчисления. Перевод всех положений, высказанных в предыдущих параграфах, на язык римановых пространств есть простое упражнение по тензорному исчислению. Обычная формула направления быстрого спуска для функции $f(x)$ в римановом пространстве такова:

$$dx^i = g^{ik} \frac{\partial f}{\partial x^k} dt.$$

Здесь применяется соглашение Эйнштейна о суммировании; верхние индексы указывают контравариантные тензоры, а g^{ik} обозначает контравариантную форму фундаментального тензора. Подробности вычислений такого типа можно найти в цитированных выше работах, однако использование тензорных пространств в задачах быстрого спуска встречается редко. Замечания, которыми мы здесь ограничимся, состоят в том, что риманову метрику надо вводить тогда, когда она позволяет уменьшить размерность пространства или сократить число связей, и что все рассуждения предыдущих параграфов могут быть записаны в римановой метрике.

Второе обобщение, которое мы должны здесь обсудить, более серьезно. Оно касается распространения выведенных выше формул на пространства бесконечного числа измерений.

Такое обобщение можно начать без всякого затруднения. Напомним, что формулы этой главы получены на основании исследования сходимости математических формул, относящихся к геометриям двух и трех измерений. Формулы записывались при помощи индекса i и область изменения i увеличивалась для того, чтобы ввести понятия многомерной геометрии. Этот процесс можно распространить на случай бесконечномерной геометрии.

Однако при этом немедленно возникает одно небольшое затруднение. Если числа x_i , число которых бесконечно и которые представляют собой координаты точки, выбраны произвольно, то сумма их квадратов не обязательно ограничена и представляемая ими точка может быть бесконечно удалена от начала координат. Но это небольшое

затруднение устраняется без серьезных потерь путем отказа от использования такой совокупности координат, сумма квадратов которых бесконечна. Пространства с бесконечно большим числом измерений, представимые при помощи бесконечно большого числа координат, причем на координаты точек наложено ограничение, заключающееся в том, что сумма их квадратов сходится, называются *гильбертовыми пространствами* (см. гл. 14). Все теоремы, которые упоминались в связи с многомерной евклидовой геометрией, сохраняют смысл и верны и в геометрии гильбертова пространства. Однако имеются некоторые важные теоремы, справедливые в многомерном евклидовом пространстве и не справедливые в гильбертовом пространстве. Одной из таких теорем является утверждение, что любая последовательность (конечных) координат изображает некоторую точку; в гильбертовом пространстве необходимо еще добавить требование сходимости.

Более серьезная особенность гильбертова пространства наблюдается, когда рассматривается бесконечная последовательность точек в ограниченной части пространства, например бесконечная последовательность точек внутри сферы единичного радиуса с центром в начале координат. В евклидовом конечномерном пространстве любого числа измерений такие последовательности обязаны иметь предельные точки, но это не обязательно для гильбертова пространства. Действительно, если мы рассмотрим точки, все координаты которых нули, за исключением одной, равной 1, то в гильбертовом пространстве это будет бесконечная последовательность точек, каждая из которых отстоит на расстояние 1 от начала координат и на расстояние $\sqrt{2}$ от любой другой точки этой последовательности. Существование таких последовательностей налагает весьма значительное дополнительное ограничение на методы быстрого спуска; можно представить себе спуск, который не ведет к бесконечной точке, а проходит через точки, находящиеся на конечном расстоянии от начала координат, и тем не менее не сходится ни к какому решению. Быстрые спуски в евклидовом пространстве конечного числа измерений легко можно определить так, что это положение не будет иметь места, в случае же бесконечномерных пространств ситуация значительно осложняется.

При решении задач вариационного исчисления мы будем вводить гильбертовы пространства, несмотря даже на связанные с ними трудности. Вариационное исчисление можно рассматривать как изучение минимизации интегралов или, в общем случае, функционалов, зависящих от переменных, которые сами являются функциями одного или более переменных. Таким образом, базисное множество независимых переменных в задачах минимизации, рассмотренных выше, должно быть заменено одной или более функциями одного или более переменных, определенных в некоторой области. Эти функции обычно выбираются из достаточно большого класса функций, что делает

невозможным определение какой-либо функции из этого класса конечным набором чисел; этот класс часто включает в себя все функции с производными, непрерывными всюду, кроме конечного числа точек излома. Часто эту функцию можно задать, указав ее значения в некотором бесконечном числе точек.

Это позволяет строго обобщить понятие многомерных пространств на гильбертовы пространства таким способом, который будет иметь отношение к задачам вариационного исчисления. Мы дадим здесь такое обобщение, но оно никоим образом не является единственно приемлемым. Например, функция, представляющая расстояние между двумя точками, т. е. интеграл, представляющий расстояние между двумя функциями, может изменяться от одной задачи к другой в зависимости от соображений удобства. Такое изменение, поскольку это касается быстрых спусков в противоположность быстрейшим, ценно только с точки зрения удобства и не играет роли в связи с нынешним развитием теории.

Приводимые здесь рассуждения будут относиться к задаче вариационного исчисления с фиксированными граничными точками и с одной зависимой и одной независимой переменными. Они легко переносятся и на другие классические проблемы вариационного исчисления. Данная задача (как было описано в гл. 4) состоит в определении функции, принадлежащей допустимому множеству и минимизирующей интеграл относительно соседних функций этого же допустимого множества.

Допустимыми функциями являются функции $y(x)$, определенные в интервале $a \leq x \leq b$; они должны быть непрерывными, и обычно требуется, чтобы их первые производные были непрерывными на каждом конечном множестве замкнутых подинтервалов, покрывающих базисный интервал. Эти функции должны удовлетворять фиксированным граничным условиям

$$y(a) = \alpha \quad \text{и} \quad y(b) = \beta,$$

где α и β — некоторые заданные числа.

Допустимая функция будет рассматриваться как точка в гильбертовом пространстве. Функция, которая является допустимой, но не удовлетворяет граничным условиям, также будет рассматриваться как точка в гильбертовом пространстве.

Формулы и обозначения двух- и трехмерной аналитической геометрии, выведенные и применяемые в § 16.2, можно теперь прямо перенести на гильбертово пространство.

Проекция отрезка, соединяющего две точки $y(x)$ и $z(x)$ гильбертова пространства, выражаются через разность координат этих точек; $u(x) = z(x) - y(x)$ есть функция, представляющая направление и длину отрезка от $y(x)$ до $z(x)$, аналогично тому, что мы имели в п. (а) § 16.2. Интеграл заменяет знак суммы, поэтому

квадрат длины отрезка определяется выражением

$$L^2 = \int_a^b [u(x)]^2 dx, \quad L \geq 0.$$

Условия параллельности отрезков аналогичны условиям п. (в) § 16.2; можно по аналогии записать и уравнение прямой с заданными направляющими коэффициентами, проходящей через некоторую точку. Интеграл также появляется в определениях перпендикулярности и гиперплоскостей. Однако здесь уже невозможно рассматривать все плоские пространства низшего числа измерений в качестве пересечений конечного числа гиперплоскостей с той же легкостью, с какой это было сделано в § 16.2.

Наконец, метод Качмажа для решения линейных уравнений в гильбертовом пространстве оставляет желать лучшего; если имеется бесконечное число уравнений, то необходимо привлечь некоторые дополнительные средства, чтобы обеспечить сходимость бесконечного цикла проектирований. Однако при рассмотрении спусков со связями формулы Качмажа могут быть перенесены и на случай гильбертова пространства.

Все это кажется ясным, и детали здесь излагаться не будут.

В элементарных задачах вариационного исчисления, которые мы будем здесь рассматривать, функция, определенная во всех точках области гильбертова пространства, представляется в виде интеграла, и нужно найти минимум этого интеграла. Интеграл обычно зависит от функции $F(x, y, p)$, определенной для значений x на базисном интервале допустимых функций, причем область изменения y и p зависит от множества допустимых функций. Если $g(x)$ есть одна из допустимых функций, то интеграл выражается так:

$$J = \int_a^b F[x, g(x), g'(x)] dx, \quad (16.13)$$

где символ $g'(x)$ означает производную dg/dx . Задача состоит в определении допустимой функции $g(x)$, минимизирующей интеграл J по сравнению со всеми допустимыми функциями в окрестности $g(x)$.

Если ищется направление быстреего спуска и используется метод, развитый выше для функций конечного числа переменных, то никакой непосредственной трудности здесь не возникает. Действительно, применяемые при этом рассуждения достаточно точно воспроизводят вывод уравнений Эйлера из вариационного исчисления (см. гл. 4).

Однако, чтобы использовать до конца этот метод, необходимо быть уверенным в том, что метод Эйлера не выводит линию спуска

из гильбертова пространства, хотя возможность такого выхода вполне реальна (это будет подробно рассмотрено ниже). Трудность происходит в значительной степени от той решительности, с которой плохие точки ранее исключались из пространства. (Однако никакого реального выигрыша от восстановления слишком многих из этих исключенных точек мы не получим, ибо настоящая трудность более глубока.)

Если ищется направление быстреего спуска и если предполагается (это доказано для случая конечного числа измерений), что движение может быть ограничено движением по прямолинейному лучу, проведенному в этом направлении, то задача сводится к определению направляющих коэффициентов $\eta(x)$ быстреего спуска. Предположим, что $g(x)$ есть та точка, откуда необходимо начать спуск; тогда точки на луче, выходящем из точки $g(x)$ с направляющими коэффициентами $\eta(x)$, могут быть выражены в такой форме

$$y(x) = g(x) + t\eta(x).$$

Величину J и производную dJ/dt в любой точке $y(x)$ теперь можно легко вычислить. Производная в точке $t=0$, т. е. в точке $g(x)$ есть

$$J'(0) = \int_a^b (F_y \eta + F_p \eta') dx, \quad (16.14)$$

где индекс при F означает дифференцирование по соответствующему переменному. В этой формуле функция F должна быть выражена через функцию $g(x)$, поэтому, где бы функция F или ее производные ни встречались, аргументами у них будут x , $g(x)$ и $g'(x)$ [аналогично тому, что мы имели в формуле (16.13)].

Выражение в правой части формулы (16.14) может быть проинтегрировано по частям и приведено к одному из двух видов

$$J'(0) = \int_a^b \left(F_y - \frac{d}{dx} F_p \right) \eta dx \quad (16.15)$$

или

$$J'(0) = \int_a^b \left(F_p - \int_a^x F_y dx \right) \eta' dx. \quad (16.16)$$

Более удобным является выражение (16.15), так как в него направляющие коэффициенты луча входят в явном виде. Описанный выше способ определения направления быстреего спуска через направляющие коэффициенты может быть применен, если только условия в граничных точках все время выполняются. Если все это

имеет место, то направляющие коэффициенты направления быстрого спуска имеют вид

$$\eta(x) = \frac{d}{dx} F_p - F_y.$$

Вторая производная функции $g(x)$, выражающей начальную точку спуска, входит в эту формулу, и нет видимых оснований полагать, что луч находится в пространстве допустимых функций (ибо у него может оставаться слишком мало производных). Конечно, существуют функции, принадлежащие к пространству допустимых функций, которые определяют луч произвольно близко к желаемому лучу. Действительно, в ряде важных случаев кривая быстрого спуска, полученная интегрированием уравнений, аналогичных уравнениям (16.8), лежит в пространстве допустимых функций, но этот результат не элементарен и полный анализ методов быстрого спуска в вариационном исчислении требует более серьезного исследования, чем поверхностное геометрическое рассмотрение, достаточное для задач с конечным числом переменных.

И все же на этом пути был получен ряд плодотворных результатов. С этих позиций подходил Адамар к большой и знаменитой задаче о защемленной пластинке [3]. Современное развитие вариационного исчисления привело к почти исключительному применению идеализированного варианта метода быстрого спуска (метода минимизирующих последовательностей) в труднейших задачах, которые и были успешно решены. Такой подход применялся, в частности, к задачам, содержащим кратные интегралы и несколько независимых переменных интегрирования, ибо в этих задачах метод конечных разностей, применяемый для задач с однократными интегралами, неприложим. Курант [1] ясно изложил этот общий метод; список авторов, которые сделали вклад в эту область, был бы, по существу, и списком авторов работ по вариационному исчислению за последние 30 лет. Однако это развитие весьма редко шло по пути анализа поведения спуска по направлению градиента; минимизирующие последовательности обычно получались неконструктивными путями.

В случае однократных интегралов классическое развитие вариационного исчисления, основанное на лемме дю Буа-Реймона, дает удобные способы решения такой задачи, не выводящие за пределы гильбертова пространства допустимых функций. Это решение основано на формуле (16.16) и на замечании, что производная $J'(0)$ будет отрицательна для направляющих коэффициентов

$$\eta(x) = - \int_a^x \left(p - \int_a^b p \, dx \right) dx,$$

где

$$p(x) = F_p - \int_a^x F_y \, dx,$$

если она отрицательна для любого множества направляющих коэффициентов.

Поскольку в этой формуле интегрирование в противоположность дифференцированию является сглаживающей операцией, то в этом процессе производные не теряются, и направление η может быть использовано в качестве направления быстрого спуска.

Для задач, содержащих кратные интегралы, техника спасения производных значительно более сложна.

Во всяком случае, метод Рэлея—Ритца может рассматриваться как метод для аппроксимации направления быстреего спуска направлением в пространстве допустимых функций. Этот метод может применяться как минимизирующий метод быстрого спуска, если множество базисных функций полное и если каждая функция имеет производные любого порядка.

§ 16.7. Обзор математического исследования точек пространства, для которого $J \leq c$. Предыдущий параграф был посвящен описанию трудностей, которые необходимо преодолеть, когда в качестве пространства, в котором ищется быстрый спуск, вводится бесконечномерное пространство. Введение бесконечномерных пространств было вызвано потребностями вариационного исчисления. Настоящий параграф будет посвящен общему обзору тех свойств некоторых задач вариационного исчисления, которые позволяют применять метод быстрого спуска или использовать другие методы такого же рода.

Некоторые вычислительные стороны таких задач были рассмотрены Штейном [11], но надо еще проделать большую работу, прежде чем получится что-то подобное законченной теории. С другой стороны, многими авторами были проведены весьма глубокие исследования задач с кратными интегралами, а М. Морс и его сотрудники исследовали поведение минимизирующих и неминимизирующих критических точек в задачах с одно- и многократными интегралами. Работы Морса и его сотрудников (см. Морс [7], Зейферт и Трельфалл [9]) не могут быть непосредственно применены к вычислительным задачам, однако они проливают много света на эти задачи. В настоящем параграфе будут кратко приведены некоторые результаты, полученные этими авторами. Именно на этих результатах надо основывать все надежды на успех методов быстрого спуска в вычислительных задачах вариационного исчисления.

Если задача состоит в том, чтобы достигнуть минимума методами быстрого спуска, то очевидно, что решающим обстоятельством является такая постановка задачи, при которой минимум действительно существует. Это элементарное требование относится, конечно, как к задачам с конечным числом переменных, так и к задачам вариационного исчисления. В последнем случае задачи обычно ставятся так, чтобы величина интеграла J становилась бесконечно большой

по мере удаления от начала координат точки гильбертова пространства, в которой вычисляется этот интеграл. Такое положение обычно естественно возникает в любой интересной задаче вариационного исчисления. Например, не представляет настоящего интереса отыскивать путь наибольшей длины от одной точки до другой, ибо такого пути, очевидно, не существует. Задача же о кратчайшем пути интересна, и пути, представляемые точками в гильбертовом пространстве, далеко отстоящими от начала координат, оказываются достаточно длинными, что позволяет исключить их при отыскании кратчайшего пути между двумя данными точками.

Первое свойство, имеющее место во всех практически интересных (по крайней мере, с точки зрения физика) задачах из вариационного исчисления, состоит в том, что *минимизирующие функции можно искать только среди тех допустимых функций, которые лежат в части гильбертова пространства (допустимых функций) с точками, находящимися на ограниченном расстоянии от начала координат.*

Это свойство, однако, недостаточно для того, чтобы гарантировать существование минимума в пространстве допустимых функций. Даже для функции с конечным числом переменных это условие может выполняться, и в то же время функция может не достигать минимума. Тем не менее, если налагается дополнительное требование о непрерывности функции, которую необходимо минимизировать, то условие неограниченности в бесконечности уже гарантирует существование минимума для функции многих переменных.

В задачах вариационного исчисления необходим более осторожный подход. Достаточно хорошим примером этого является задача о линиях наименьшей длины. Линия, соединяющая две точки на плоскости, может быть составлена из равных отрезков, каждый из которых образует угол в 45° с прямой, соединяющей две конечные точки. Эти отрезки пересекают зигзагообразно прямую линию, соединяющую конечные точки. Число таких отрезков может беспрдельно увеличиваться, и по мере этого ломаная линия все ближе и ближе приближается к прямой. Однако ее длина всегда больше в $\sqrt{2}$ раз длины отрезка прямой линии, соединяющего наши точки.

Итак, функция, определяющая длину, которая столь обычна и распространена, не является непрерывной при таком геометрическом подходе. Некоторый успех в этом отношении может быть достигнут, если при вычислении длины рассматриваются как положение отрезков, так и их направление, но выигрыш здесь достигается лишь в результате больших усилий и его недостаточно для преодоления имеющих место трудностей.

Одна важная особенность задачи о длине состоит в том, что там имеет место полунепрерывность снизу. В терминах упомянутой выше задачи это означает, что можно найти множество линий, сходящихся к заданной линии так, что длины линий этого мно-

жества сходятся к некоторой величине, большей, нежели длина предельной линии, но невозможно найти множество линий, сходящихся к такой предельной линии, длина которой больше, чем предел длин аппроксимирующих линий. Это свойство само по себе гарантирует существование минимума в задаче с конечным числом переменных, где функция, которую необходимо минимизировать, стремится к бесконечности по мере того, как стремится к бесконечности расстояние между началом координат и точкой определения функции. Можно доказать, что свойство полунепрерывности снизу имеет место и для тех классов задач вариационного исчисления, которые интересны с точки зрения физики.

Свойство полунепрерывности снизу вместе со свойствами, гарантирующими отсутствие всякой конкуренции между функциями, достаточно удаленными от нулевой функции, является все еще недостаточным для того, чтобы гарантировать возможность решения задачи вариационного исчисления в множестве допустимых функций, ибо остается затруднение, связанное с бесконечномерностью объемлющего пространства. Однако для многих интересных задач решения все же были найдены, а физические модели указывают, что возможно решение и тех задач, которые не могут быть решены явно. Эти наблюдения заставили математиков обратиться к весьма серьезному изучению природы пространства точек, заполненного допустимыми функциями, для которого интеграл не принимает значений, больших некоторой постоянной c , т. е. пространства, для которого $J \leq c$. Такие исследования оказались очень плодотворными.

Простейший результат, который устанавливается для такого множества точек, состоит в том, что оно ведет себя как пространство конечного числа измерений, где каждая бесконечная последовательность точек этого множества должна иметь предельную точку в этом множестве. Это как раз и есть свойство ограниченных множеств в конечномерных пространствах, которое гарантирует, что полунепрерывная снизу функция, определенная в замкнутом ограниченном множестве, достигает в нем минимума. Рост в неограниченных точках позволяет направить наше внимание на ограниченные множества. Этот результат достаточен для доказательства существования минимума и уверенности в том, что этот минимум может быть достигнут методами быстрого спуска.

Упомянутая выше теория Морса развивалась дальше. Морса интересовали неминимизирующие экстремали-седловины и т. п., которые упоминались в гл. 4. Общая теория неминимизирующих критических точек для функций нескольких переменных, но не для функций с бесконечно большим числом переменных, создавалась Эйлером, Пуанкаре и др. Морс, а также и другие (в том числе Пуанкаре и Г. Д. Биркгофф) пытались распространить эту теорию на функции бесконечно большого числа переменных, встречающиеся в задачах вариационного исчисления. Морс достиг в этом успеха, доказав

в 1939 г., что минимальные поверхности согласуются с этой общей теорией (это положение, по крайней мере частично, было получено независимо от него некоторыми другими учеными). Проблема минимальной поверхности, которая включает в себя задачу о поведении тонкой мыльной пленки, оставалась математически не разрешенной в течение почти 100 лет, начиная со времени ее постановки бельгийским физиком Плато и до появления ряда блестящих работ различных авторов, написанных независимо друг от друга не более чем 25 лет назад. Для многих работающих над этой проблемой ее решение на основе теории Морса означало решение одной из наиболее трудных физических задач методами вариационного исчисления. Таким образом, в некотором смысле теория Морса может считаться законченной в ее приложении к задачам, которые берут свое начало в физике.

Важность изложенного здесь положения для задач вариационного исчисления с вычислительной точки зрения заключается в следующем: *теория Морса основана на ряде сложных свойств пространства точек, для которого $J \leq c$; эти вместе взятые свойства дают гарантию того, что это пространство ведет себя (в смысле, важном с точки зрения вариационного исчисления) как конечномерное пространство, в котором минимизируемая функция непрерывна.* Модификации теории Морса, которые необходимы для успешного проведения вычислений при использовании методов быстрого спуска, вовсе не являются невозможными. Хотя эти модификации еще не завершены полностью и хотя с точки зрения разработки эффективных методов вычислений остается сделать еще многое, результаты Морса и его сотрудников (а также авторов, которые занимались менее детальными исследованиями более общих проблем) и наличие быстродействующей цифровой вычислительной техники приводят к необходимости изучения методов быстрого спуска при решении физических задач вариационного исчисления.

§ 16.8. Заключение. В заключительной части этой главы мы должны по возможности суммировать все сказанное выше. В частности, нам следует добавить некоторые физически очевидные соображения к изложенным выше математическим доводам в пользу желательности применения методов быстрого спуска в задачах вариационного исчисления.

В гл. 4 профессор Хестенс указал (и это знают, конечно, все инженеры) на роль вариационного исчисления в статических задачах с консервативными полями. Примером служит висящая тяжелая нить. Задача здесь заключается в минимизации потенциальной энергии. Градиент потенциальной энергии есть сила, с которой консервативное поле действует на частицу, находящуюся в точке, где определяется градиент. Классическое решение этой задачи основано на решении системы дифференциальных уравнений (из которых существенно только одно, если точки подвеса нити находятся выше одной из координат).

натных осей, а сила направлена вниз) и на выборе значений параметров в общем решении для удовлетворения наложенных граничных условий и условия длины.

Изложенный здесь метод спуска более или менее имитирует движение тяжелой нити, когда она освобождена в начальном положении, не представляющем собой положение равновесия. Здесь направление быстрейшего спуска для каждой точки линии, как было вычислено с помощью вышеприведенных интегралов, есть направление результирующей силы, действующей в этой точке. Данная схема состоит в том, чтобы позволить тяжелой нити перемещаться в направлении результирующей силы на некоторое расстояние.

Если шаги этого движения относительно велики, то движение может считаться искусственно заторможенным; такое движение, будучи однажды начато, длится некоторое время, пока это позволяют изменения в системе. Это состояние действительно может быть достигнуто за счет демпфирующих сил. Для физика вполне очевидно, что затухание движения такой системы зависит от потери энергии тяжелой нитью. Если предположить, что движение прекратилось в конце одного из этих шагов, так что в моменты, соответствующие началу и концу одного из шагов, кинетическая энергия отсутствует, то потеря энергии будет иметь место только в том случае, когда потенциальная энергия в конце шага будет меньше, чем в начале его. Поэтому физически оказывается возможным продолжать спуск по любому пути без пересчета направления быстрейшего спуска на протяжении такого расстояния, на котором интеграл потенциальной энергии будет оставаться меньшим, чем он был в начале шага. Это можно предположить на основании общих положений предыдущего параграфа относительно природы пространства точек, для которого $J \leq c$.

О рекомендуемой величине шага можно сказать только то, что шаг должен быть достаточно длинным, но при этом величина интеграла, который надо минимизировать, должна быть меньше в конце шага, чем в начале. На больших вычислительных машинах удобно задавать произвольно величину множителя t при вычислении направляющих коэффициентов, определяющих быстрейший спуск, а затем делить эту величину на 2 для получения новой величины t , если значение интеграла в конце каждого шага не уменьшалось. Для этого правила нельзя дать особых обоснований, за исключением разве соображений удобства в случае использования машин, в которых применяется двоичная система.

В заключение необходимо отметить, что проблемы, возникающие в более сложных физических системах, не могут так просто анализироваться с точки зрения энергетики, результирующих сил или затухания движения, как задача о висящей тяжелой нити в консервативном поле. Например, задача о снаряде может быть рассмотрена вариационными методами (см. гл. 3 и 4), но физическое обоснование того,

что приближенная траектория преобразуется при быстром спуске в истинную траекторию, не так очевидно, как раньше. Здесь снова может быть поставлен вопрос о физическом смысле той энергии, которую необходимо добавлять для поддержания движения снаряда по принятой приближенной траектории. Эти соображения не являются особо убедительными с точки зрения физики, с математической же точки зрения они равнозначны проверке описанных выше свойств точек пространства, для которого $J \leq c$. Таким образом, в некоторых физических задачах метод быстрого спуска удивительно сходен с поведением самой физической системы. В других задачах этот метод является лишь абстракцией, которая (по крайней мере в некоторых случаях) может привести к формулировке другой эквивалентной физической задачи; в этой последней спуск уже является спуском, выполняемым самой системой. При этом спуск в задачах вариационного исчисления является своего рода абстрагированием движущейся физической системы, что обеспечивает ранее не достижимые вычислительные результаты.

ЛИТЕРАТУРА

1. Курант Р., Принцип Дирихле, конформные отображения и минимальные поверхности, ИЛ, М., 1953.
2. Eisenhart L. P., Coordinate Geometry, Ginn, Boston, 1939.
3. Hadamard J., Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées, Mem. Savants Etrangers Acad. Sci. Inst. France (2), 33, 1—128 (1908).
4. Hestenes M. R. and Stiefel E., Methods of Conjugate Gradients for Solving Linear Systems, J. Res. Nat. Bur. Standards, 49, 409—436 (1952).
5. Jordan C., Géométrie à n dimensions, Bull. Soc. Math. France, 3, 103—174 (1875).
6. Kaczmarz S., Angenäherte Auflösung von Systemen linearer Gleichungen, Bull. Internat. Acad. Polon. Sci., Cl. Sci. Math. Nat., Ser. A, 355—357 (1937).
7. Morse M., The Calculus of Variations in the Large, Amer. Math. Soc. Colloquium Publ., 18, American Mathematical Society, New York, 1934.
8. Schreier O. and Sperner E., Modern Algebra and Matrix Theory, Chelsea, New York, 1951.
9. Зейферт Г. и Трельфалл В., Вариационное исчисление в целом, ИЛ, М., 1947.
10. Sokolnikoff I. S., Tensor Analysis, John Wiley, New York, 1951.
11. Stein M. L., On Methods for Obtaining Solutions of Fixed End-point Problems in the Calculus of Variations, J. Res. Nat. Bur. Standards, 50, 277—297 (1953).
12. Veblen O., Certain Aspects of Modern Geometry I, Rice Inst. Pamphlet, 21, 209—222 (1932).
13. Veblen O. and Whitehead J. H. C., The Foundations of Differential Geometry, Cambridge Tract No 29, Cambridge University Press, London, 1932.

Глава 17

БЫСТРОДЕЙСТВУЮЩИЕ ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ И ИХ ПРИМЕНЕНИЯ

ДЕРРИК Г. ЛЕМЕР

Калифорнийский университет, Беркли

§ 17.1. Введение. Настоящая глава является последней в серии глав, посвященных тому вкладу, который на сегодняшний день вносят математики в развитие техники. Хотя иногда участие математиков ограничивается лишь формулировкой точек зрения или общих принципов, но чаще математикам удается разработать аппарат для решения задач, основанных на фактических измерениях. Этот аппарат может быть как математическим, так и физическим. Целью настоящей главы является описание и иллюстрация этих средств решения задач, и в особенности новых физических устройств. Во все математические и физические детали вопроса мы будем входить лишь постольку, поскольку это позволит объем главы. В известном смысле настоящая глава резюмирует все предыдущие, поскольку мы будем кратко заимствовать из их содержания иллюстративные задачи и общие точки зрения. Эта глава имеет также отношение и к технике, так как в ней рассматриваются некоторые детали конструкции вычислительных машин. Таким образом, это область, в которую внесли свой вклад как математики, так и инженеры. Мы надеемся, что читатель восполнит имеющиеся в настоящем изложении очевидные пробелы, обратившись к другим главам и к собственному техническому багажу. Мы надеемся также, что читатель, проявивший к какому-либо из рассматриваемых вопросов не только случайный интерес, продолжит изучение этого вопроса с помощью соответствующих работ, приведенных в списке литературы в конце каждой главы.

Каждому инженеру известно, что вычислительные машины бывают двух типов: машины-аналоги и цифровые машины (мы пользуемся здесь терминами, обычно применяемыми в этой области), и он хорошо знает отличительные черты тех и других. Беседуя с инженерами, автор пришел к заключению, что они предпочитают вычислять на машинах-аналогах, но конструировать предпочитают цифровые устройства. И это вполне естественно. Инженеру привычнее иметь дело с такими записывающими приборами, как вольтметр или осциллограф, чем со считающими механизмами. Кроме того, инженер приучен решать задачи

на моделях, изображающих изучаемый процесс в небольшом масштабе. Поэтому такие устройства или эквивалентные им электрические схемы кажутся инженеру значительно более „естественными“, чем действительно естественный метод построения математической модели и решения математической задачи. Между тем последнее является необходимым условием применения цифровой техники для решения инженерных задач.

Когда дело доходит до конструирования, цифровые машины требуют более широкого инженерного опыта. До недавнего времени основные усилия по улучшению машин-аналогов были направлены на разработку более точных и стабильных компонент — необходимая, но достаточно неинтересная работа. В настоящее время наблюдается тенденция использовать при конструировании машин-аналогов новые элементы для увеличения гибкости и области применения этих машин. Эта тенденция, если ей суждено будет развиваться, несомненно привлечет к машинам-аналогам значительно большее число инженеров, чем раньше.

Мы надеемся, что приводимое ниже описание цифровых машин подтолкнет некоторых из наших читателей-инженеров к более глубокому изучению возможностей этих машин. Решая задачу на таких сложных машинах, как современные цифровые вычислительные машины, мы в силу гибкости последних часто обнаруживаем, что наша задача допускает решение в значительно более общем виде, чем предполагалось первоначально. Более того, существуют интересующие инженера области, такие, как планирование, составление схем и задачи теории операций, которые по существу своему являются цифровыми задачами. Такие проблемы можно теперь снять с чертежных столов и решать с помощью цифровых вычислительных машин.

Заканчивая эти общие предварительные соображения, мы позволим себе сделать небольшое замечание в адрес математиков. Большинство из них, не считая тех, кто работает с вычислительными машинами, не знает, по существу, разницы между цифровыми машинами и машинами-аналогами. Как и инженеры, обучению которых математики помогают, большинство математиков приучено мыслить понятиями-аналогами, и поэтому машина-аналог является для них идеалом. Не только такие понятия, как линейные операторы и дифференциальные уравнения, но и более глубокие соображения, относящиеся к замкнутости и непрерывности, являются фундаментальными как в математическом анализе, так и для машин-аналогов. Таким образом, инженеры разделяют вместе с математиками некоторое недоверие к непривычным им цифровым вычислительным машинам.

§ 17.2. Машины-аналоги. Машинам-аналогам было дано следующее весьма общее определение: пусть существуют две физические системы, которые ведут себя аналогичным образом; тогда более привычная для нас система называется машиной-аналогом. Зарегистрированы примеры, когда эти две системы менялись ролями.

На рубеже нашего века наиболее совершенным типом машин-аналогов были гармонические анализаторы. Много выдающихся физиков и математиков работали над их усовершенствованием, и только инженеры уделяли им мало внимания. В 1920 г. появились две новые машины-аналоги: дифференциальный анализатор Буша и электронная машина Маллока для решения линейных уравнений. Первая из них была механическим устройством и оказала глубокое влияние на развитие техники. Логическая структура современных машин-аналогов все еще покоится на первоначальной идее Буша. Вместе с тем само выполнение операций перестало осуществляться механическими средствами и основано теперь на электрических устройствах.

Дифференциальный анализатор Буша в своем первоначальном виде основан на двух фундаментальных операциях: суммировании и интегрировании. Суммирование осуществляется с помощью дифференциальных коробок передач, а интегрирование — с помощью обычных колесно-дисковых механизмов. Умножение на постоянную величину выполняется с помощью зубчатой передачи, а умножение на переменную величину осуществляется на основании одной из двух следующих формул:

$$uv = \int u \, dv + \int v \, du,$$
$$uv = \left[\frac{1}{2} (u + v)^2 - \frac{1}{2} (u - v)^2 \right],$$

причем в последней квадраты вычисляются по формуле

$$x^2 = \int 2x \, dx.$$

Функции вводятся в машину либо с помощью их дифференциальных уравнений, либо с помощью графиков, которые опытный оператор строит от руки. Более поздние модели механических дифференциальных устройств стали оснащаться электронным и оптическим оборудованием, таким, как фотоэлектрические следящие системы и релейные усилители крутящего момента, и иногда, для сокращения времени установившихся машины, даже примитивными программными устройствами, оставляющими основные операции машины без изменения.

Появление интегрирующих усилителей заменило механические колеса и диски и, по существу, почти все движущиеся детали машины Буша. Но при этом основными операциями остались сложение и интегрирование. На таких машинах, как REAC, умножение на переменную величину выполняется с помощью потенциометра, снабженного сервоприводом. Но даже и эти движущиеся части постепенно заменяются электронными умножителями различных типов. В настоящее время функции образуются в машинах с помощью так называемых „функциональных датчиков“, т. е. соответствующим образом экранированных осциллографов, или с помощью линейных комбинаций основных

ступенчатых функций. Современный дифференциальный анализатор может быть собран из „черных ящиков“, т. е. основных элементарных устройств, указанных в каталогах.

Машины-аналоги предназначены в основном для решения дифференциальных уравнений, с помощью которых может быть описано большинство технических задач. Обычно говорят, что основное содержание математики группируется вокруг этой области. Особенно легко осуществить набор и решение на машине-аналоге обыкновенных дифференциальных уравнений и, в частности, линейных уравнений с заданными начальными условиями.

Хотя точность машин-аналогов не возросла вместе с их модернизацией, их все еще можно считать достаточно точными для изображения установившихся физических процессов, особенно если иметь в виду некоторую недоверенность математического описания процесса, с помощью которого было получено дифференциальное уравнение, и ненадежность начальных условий.

Много говорилось об исключительной скорости работы цифровых вычислительных машин, но надо сказать, что скорость работы машин-аналогов значительно выше и ограничивается исключительно переходным временем усилителей. Так, например, задачи, зависящие только от времени (одномерные задачи), являются более подходящими для машин-аналогов, и имеется ряд задач теории управления этого класса, для которых скорость работы цифровых машин все еще недостаточно высока.

Сделаем несколько замечаний о недостатках машин-аналогов и опишем, как в некоторых случаях преодолеваются эти трудности.

Вероятно, основным ограничением машин-аналогов является не их низкая точность, а отсутствие достаточной логической гибкости. Хорошо известно, что машины-аналоги могут обращаться только с одним независимым переменным — *временем*. Наиболее актуальные современные проблемы связаны с распределенными в пространстве системами. Такие системы не могут быть описаны с помощью обыкновенных уравнений и приводят к уравнениям в частных производных с двумя и более независимыми переменными. Правда, некоторые из этих проблем могут быть с помощью искусственных математических приемов „разложены“ на обыкновенные дифференциальные уравнения с краевыми условиями, приводящими к задачам о собственных значениях. Но очень часто такой метод оказывается неприменимым. Конечно, в распределенной системе иногда можно „сосредоточить“ величины и получить обыкновенное дифференциальное уравнение для некоторой эквивалентной задачи; однако для того, чтобы дать более чем качественное толкование слову „эквивалентная“, требуется глубокое понимание соответствующей задачи. Более поздние попытки решения дифференциальных уравнений в частных производных на машинах-аналогах основаны, по существу, на цифровой технике, т. е. на конечно-разностных операторах. Интуитивно ясно, что для этого

область, в которой ищется решение, разбивается на конечное число полос. Задача, содержащая зависящие от времени параметры, решается для короткого промежутка времени в каждой из этих полос, а затем соответствующим образом преобразуется и вновь решается для следующей полосы. При существующем оборудовании такие преобразования не могут выполняться автоматически. В отличие от цифровых машин, в которых эти действия выполняются автоматически, оператор принимает участие в работе машины-аналога даже в большей степени, чем это делает сама машина.

Приведем другой пример проявления ограничения, которое накладывает на машины-аналоги одномерность решаемых ими задач. Рассмотрим два известных интеграла

$$\int f(t) g(T-t) dt,$$

$$\int f(t) g(T+t) dt,$$

представляющих соответственно распределение суммы и разности двух случайных функций. Вычисление второго интеграла значительно проще, чем первого, так как переменная интегрирования может быть легко получена в этом случае простой задержкой времени на T секунд. В настоящее время это осуществляется следующим способом: функция $f(t)$ записывается на магнитную ленту или барабан и считывается T секундами позже. Вычисление первого интеграла предъявляет значительно большие требования к запоминающему устройству машин. Одна из функций должна быть полностью записана, а затем введена в машину в обратном направлении.

Для иллюстрации той простоты, с которой некоторые задачи могут быть поставлены на машине-аналоге, мы рассмотрим совершенно элементарный пример уравнения второго порядка

$$y'' + Ky' + Ly = f(t)$$

с начальными условиями

$$y' = a, \quad y = b \quad \text{при} \quad t = 0.$$

Схема на рис. 17.1 настолько проста, что практически понятна без объяснений. Инверторы меняют знак у вводимой величины. Интеграторы образуют интегралы вводимых величин с обратным знаком. Сумматор дает сумму входных величин с обратным знаком.

Мы молчаливо предполагали, что K и L являются положительными постоянными и что умножения на эти постоянные выполняются с помощью устанавливаемых вручную потенциометров. Если K и L являются функциями времени, то эти потенциометры будут устанавливаться посредством пальцев серво-привода с помощью соответствующих генераторов функций. Генерируемые функции могут либо

вводиться извне, либо вырабатываться в самой машине при помощи дополнительных интеграторов. Постановка таких задач на машинах-аналогах может оказаться значительно сложнее, чем в рассмотренном выше простом примере. Чем сложнее задача, тем сильнее проявляется здесь лавинообразный процесс нарастания сложности схемы. Большие машины содержат очень большое число усилителей; предполагается, что в одной из машин их будет 800. Специализированные машины-аналоги содержат значительно меньше оборудования и с успехом решают некоторые частные задачи.

Перед тем как перейти к чисто цифровым вычислительным машинам, мы обратим внимание читателя на цифровой дифференциальный анализатор, у которого логическая структура так же проста, как

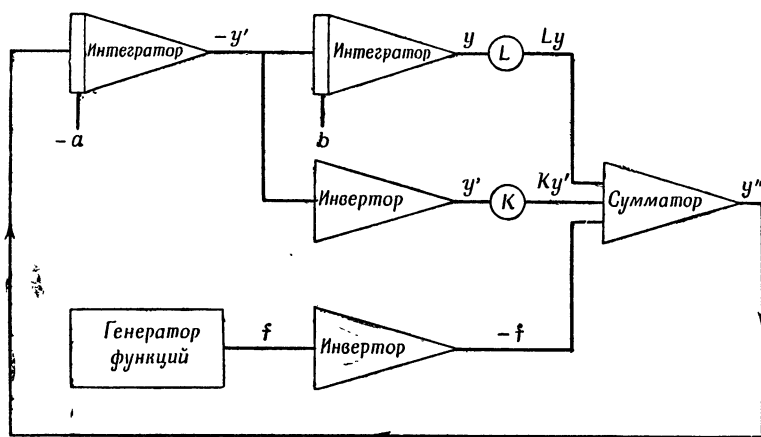


Рис. 17.1. — Решение на машине-аналоге дифференциального уравнения второго порядка $y'' + Ky' + Ly = f(t)$ при начальных условиях: $y' = a$, $y = b$ при $t = 0$.

и у машин-аналогов, но действия выполняются цифровым методом. Погрешности, содержащиеся даже в самых точных узлах машин-аналогов, заменяются здесь ошибками округления. Мы должны отослать читателя к специальной литературе по этим интересным машинам, так как недостаток места не позволяет остановиться на них подробнее ¹⁾.

§ 17.3. Цифровые вычислительные машины. Существенное свойство цифровых вычислительных машин гораздо лучше передается их менее употребительным названием „машины дискретного действия“. С точки зрения устройств дискретного действия объекты должны

¹⁾ См., например, вскоре выходящий в Издательстве иностр. литературы сборник «Цифровые дифференциальные анализаторы». — *Прим. ред.*

скорее считаться, чем измеряться; вселенная квантована, и в этом ее математика. Интегралы являются всего лишь суммами, а производные лишь разностными отношениями; функции везде разрывны; бесконечно малые и бесконечно большие пределы в действительности не существуют. Более того, деление не есть операция, обратная умножению, и даже сложение не всегда возможно (в силу переполнения разрядной сетки).

Таким образом, становясь на точку зрения дискретного переменного, мы как бы возвращаемся ко временам Пифагора. Это не должно пугать инженера, так как ему достаточно вспомнить, что, пользуясь каждый день настольными вычислительными машинами, он уже приобрел достаточно хороший опыт в технике дискретного переменного. Действительно, типичная цифровая вычислительная машина делает, по существу, то же самое, что обычная настольная машина и ее оператор, но только делает это быстро и автоматически.

В цифровых машинах время не является более независимой переменной. В действительности на машинах часто решаются задачи с двумя и более независимыми переменными; шаг этих переменных часто меняется по усмотрению машины. Однако почти все счетные устройства имеют только один арифметический узел, а это делает время важным фактором.

Основным действием, выполняемым машиной, является извлечение двух чисел из памяти, образование третьего числа в арифметическом узле и (обычно) возвращение этого числа в память. Систематическое шаг за шагом повторение подобных все время меняющихся действий составляет яркий контраст с *modus operandi* машины-аналога.

Легко понять, что с помощью цифровых устройств удобно решать такие задачи, которые по существу своему являются разрывными, как, например, задачи теории чисел. Однако в инженерной практике важную роль играют задачи другого типа, связанные с непрерывным распределением масс и с непрерывными функциями времени и пространства. В таких случаях вся задача разбивается на отдельные интервалы или ячейки, как мы это покажем ниже. При этом предполагается, что состояние, принимаемое в некоторой точке такой ячейки, принимается также и во всей ячейке. Это предположение ведет к ошибке, называемой *ошибкой метода*. Выяснение ее влияния на конечный результат является одной из задач численного анализа. Ошибками другого типа, имеющими место в цифровых машинах, являются так называемые *ошибки округления*. Они связаны с тем, что результаты таких действий, как деление, извлечение квадратного корня и даже умножение, округляются до фиксированного числа значащих цифр. Это вводит несистематические ошибки, аналогичные шумам в машинах-аналогах. При хорошо составленной программе вычислений ошибки округления не должны быстро накапливаться. Сохраняя надлежащее число значащих цифр (в должной мере превосходящее практическую точность вводимых в машину данных),

программист может предохранить значащую часть окончательного ответа от влияния ошибок округления, даже если миллионы отдельных мелких этапов счета содержат процесс округления. Ошибками дискретного анализа третьего типа являются *логические ошибки*. Им соответствуют логические ошибки при постановке задач на машинах-аналогах. Большая логическая гибкость цифровых машин налагает на программиста дополнительную ответственность. Маленькая ошибка в логической структуре программы очень часто оказывает громадное влияние на весь процесс вычислений, но при этом обнаружение самой ошибки бывает затруднено. Эти замечания будут иллюстрированы в рассматриваемых ниже вопросах программирования и кодирования задач.

Рассмотрим теперь дискретный метод решения уравнения

$$y'' + Ky' + Ly = f(t),$$

уже рассмотренного выше с точки зрения машин-аналогов. При решении системы уравнений на машинах-аналогах система должна быть заменена одним уравнением более высокого порядка. Дискретный метод решения требует обратного подхода: одно уравнение n -го порядка может быть заменено системой n уравнений первого порядка. В случае приведенного выше уравнения второго порядка мы можем ввести новую переменную p , положив

$$y' = p.$$

Тогда дифференциальное уравнение примет вид

$$p' = f(t) - Kp - Ly.$$

Два получившихся уравнения образуют систему двух уравнений первого порядка, вполне эквивалентную данному уравнению. Граничными условиями будут, естественно,

$$y(0) = b,$$

$$p(0) = a.$$

Мы можем также рассмотреть более общую систему двух уравнений

$$\begin{aligned} \frac{dy}{dt} &= \varphi_1(p, y, t), \\ \frac{dp}{dt} &= \varphi_2(p, y, t), \end{aligned} \quad (17.1)$$

где φ_1 и φ_2 — заданные функции, которые, конечно, не обязаны быть линейными.

Рассматривая эту задачу с точки зрения решения ее на цифровой машине, мы должны вычислить таблицу (а не кривую) многих тысяч

значений функции $y(t)$ и ее производной $p(t)$ для одинаковых интервалов Δt времени t , начиная с $t=0$ и начальных значений

$$y(0) = y_0, \quad p(0) = p_0.$$

Нас могут интересовать не все из этих значений. Практически, вероятно, около 90% вычисленных значений существуют в памяти машины всего несколько миллисекунд каждое. Заменим теперь производные в системе (17.1) соответствующими разностными отношениями

$$\frac{dy}{dt} \sim \frac{y_{n+1} - y_n}{\Delta t}, \quad \frac{dp}{dt} \sim \frac{p_{n+1} - p_n}{\Delta t}$$

и умножим последние на Δt . Тогда система (17.1) примет вид рекуррентных формул

$$\begin{aligned} y_{n+1} &= y_n + (\Delta t) \varphi_1(p_n, y_n, n \Delta t), \\ p_{n+1} &= p_n + (\Delta t) \varphi_2(p_n, y_n, n \Delta t) \end{aligned} \quad (17.2)$$

для последовательного вычисления y и p , исходя из начальных значений y_0 и p_0 . Этот метод решения является слишком простым и, как правило, ведет к нарастающей ошибке. При более точном применении метода конечных разностей значения, получаемые из уравнений (17.2), рассматривают в качестве предварительных. Обозначив

$$\begin{aligned} y_n^* &= y_n + (\Delta t) \varphi_1(p_n, y_n, n \Delta t), \\ p_n^* &= p_n + (\Delta t) \varphi_2(p_n, y_n, n \Delta t), \end{aligned}$$

определяют затем y_{n+1} и p_{n+1} по формулам

$$\begin{aligned} y_{n+1} &= y_n + \frac{1}{2} \Delta t [\varphi_1(p_n, y_n, n \Delta t) + \varphi_1(p_n^*, y_n^*, n \Delta t)], \\ p_{n+1} &= p_n + \frac{1}{2} \Delta t [\varphi_2(p_n, y_n, n \Delta t) + \varphi_2(p_n^*, y_n^*, n \Delta t)]. \end{aligned}$$

Результат, полученный по этим формулам, является более точным в силу меньшей ошибки метода. Для получения еще более точного результата можно пользоваться разностями более высокого порядка, но часто бывает выгоднее применять простые методы при сравнительно меньшем значении Δt .

Мы перейдем теперь к более сложному вопросу об уравнениях в частных производных и рассмотрим кратко примеры трех стандартных типов уравнений.

Одномерное волновое уравнение (см. гл. 5) имеет вид

$$\frac{\partial^2 \varphi}{\partial t^2} = c^2 \frac{\partial^2 \varphi}{\partial x^2}, \quad (17.3)$$

где коэффициент c есть действительная постоянная. Если заданы начальные значения $\varphi(x, 0)$ и $\partial \varphi / \partial t$ при $t=0$, то существует единст-

венное решение $\varphi = \varphi(x, t)$. Разобьем плоскость (x, t) на отдельные ячейки с помощью прямоугольной сетки, проходящей через точки

$$(x, t) = (n \Delta x, m \Delta t), \quad n = 0, 1, 2, \dots; \quad m = 0, 1, 2, \dots$$

Обозначим для простоты $\varphi(n \Delta x, m \Delta t) = \varphi_{n,m}$. Заменяя в уравнении (17.3) производные разностными отношениями и умножив на $(\Delta t)^2$, мы получим вместо уравнения (17.3) соответствующее конечно-разностное уравнение

$$\varphi_{n,m+1} - 2\varphi_{n,m} + \varphi_{n,m-1} = c^2 (\Delta t)^2 (\Delta x)^{-2} (\varphi_{n+1,m} - 2\varphi_{n,m} + \varphi_{n-1,m}).$$

Это уравнение служит рекуррентной формулой для последовательного получения

$$\varphi_{n,2}, \varphi_{n,3}, \varphi_{n,4}, \dots$$

причем начальные значения $\varphi_{n,0}$ и $\varphi_{n,1}$ определяются из граничных значений $\varphi(x, 0)$ и $\partial \varphi / \partial t$ при $t = 0$. Чтобы небольшие начальные ошибки не нарастали по мере увеличения аргумента, должно выполняться условие устойчивости метода

$$|c| \Delta t < \Delta x.$$

В качестве другого примера дифференциального уравнения в частных производных рассмотрим уравнение диффузии

$$\frac{\partial \varphi}{\partial t} = P(x) \frac{\partial^2 \varphi}{\partial x^2}.$$

Здесь $P(x)$ — заданная функция, значения которой в равноотстоящих точках x мы обозначим через

$$P_n = P(n \Delta x).$$

Кроме того, заданы начальные значения функции φ :

$$\varphi_{n,0} = \varphi(n \Delta x, 0).$$

Конечно-разностное уравнение, соответствующее уравнению диффузии, имеет вид

$$\varphi_{n,j+1} = \varphi_{n,j} + \frac{\Delta t}{(\Delta x)^2} P_n (\varphi_{n+1,j} - 2\varphi_{n,j} + \varphi_{n-1,j})$$

и также допускает в рассматриваемой области последовательное решение по «шагам». Это уравнение является рекуррентной относительно индекса j формулой, по которой, начиная с известных значений $\varphi_{n,0}$, мы можем последовательно определить

$$\varphi_{n,1}, \varphi_{n,2}, \dots$$

Для устойчивости метода мы имеем достаточное условие

$$\frac{\Delta t}{(\Delta x)^2} \max |P_n| \leq \frac{1}{2}.$$

Таким образом, желая сделать приближение

$$\frac{\varphi_{n+1,j} - 2\varphi_{n,j} + \varphi_{n-1,j}}{(\Delta x)^2}$$

возможно более близким ко второй производной $\partial^2 \varphi / \partial x^2$, нельзя слишком сильно уменьшать величину Δx , иначе будет нарушено условие устойчивости. Иными словами, должно быть обеспечено известное соотношение между сторонами прямоугольных ячеек, заменяющих двумерный континуум задачи диффузии. Эти рассуждения играют, очевидно, весьма важную роль при выборе экономичного способа решения практических задач.

Чтобы закончить это весьма краткое изучение уравнений в частных производных, мы рассмотрим в качестве примера уравнение, решение которого „по шагам“ невозможно. Это — знаменитое уравнение Лапласа

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0,$$

решение которого ищется в некоторой области R при заданных на границе значениях функции $\varphi(x, y)$. Пользуясь построенной выше прямоугольной сеткой, мы видим, что уравнению Лапласа может соответствовать одно из двух конечно-разностных уравнений

$$\varphi_{n+1,j} + \varphi_{n-1,j} + \varphi_{n,j+1} + \varphi_{n,j-1} = 4\varphi_{n,j},$$

$$\varphi_{n+1,j+1} + \varphi_{n-1,j+1} + \varphi_{n-1,j-1} + \varphi_{n+1,j-1} = 4\varphi_{n,j}.$$

Эти уравнения выражают закон Гаусса о среднеарифметическом значении: значение функции φ в любой точке равно среднеарифметическому ее значению в четырех соседних точках. Соседние точки в двух указанных уравнениях расположены соответственно так, как это указано на рис. 17.2.

Здесь нет начальных условий, исходя из которых можно было бы „по шагам“ построить все решение. Поэтому мы должны взять пробное решение и затем пытаться улучшить его так, чтобы оно



Рис. 17.2.

удовлетворило написанным выше разностным соотношениям. При этом мы должны поддерживать заданные значения на границе. Ясно, что в этой задаче возникают трудности, существенно отличные от возникающих при решении волнового уравнения, внешне похожего на уравнение Лапласа. В данном случае необходимо несколько раз пройти всю область, как-то сохраняя результаты предыдущего обхода.

Чтобы улучшить последовательные приближения к функции φ , машина должна выполнить некоторое количество оценок. Поэтому

экономичное решение этой задачи требует особенно тщательного планирования всего вычислительного процесса.

Три рассмотренных выше типа дифференциальных уравнений в частных производных: гиперболический, параболический и эллиптический — представляют лишь первые шаги на пути к реалистическому решению многих задач физики и техники. Существующие на сегодняшний день цифровые машины еще непригодны для решения уравнений, зависящих от трех пространственных переменных и времени. Этому препятствуют ограничения, налагаемые объемом памяти машины и временем выборки чисел из нее. Надо сказать, что и численный метод решения таких задач еще недостаточно развит.

§ 17.4. Матричные задачи. В главе 12 было ясно показано, какую важную роль играют матрицы и матричные задачи в инженерном деле. Мы приведем здесь одно сделанное там замечание:

„Развитие современных цифровых вычислительных машин, легко осуществляющих основные матричные операции, сообщило дополнительный толчок широкому использованию матриц в различных областях знания“.

К сказанному можно добавить, что развитие современных вычислительных машин и цифровых методов вызвало к жизни матричные задачи в таких областях, в которых они никогда не проявлялись раньше. Мы уже видели, что уравнения в частных производных непосредственно сводятся к матричным задачам. Точно так же достаточно сформулировать задачи о граничных значениях на языке интегральных уравнений, чтобы свести их к матричным задачам. Даже само понятие функции двух переменных становится матричным понятием.

Наиболее обычной матричной задачей является задача о решении линейных уравнений. В некоторой степени это связано с естественным стремлением математиков-прикладников упростить решение задачи с помощью ее линеаризации (см. гл. 1). Вместе с тем имеется очень много задач, линейных по своему существу. В общем виде задача ставится следующим образом: найти такие числа

для которых

$$x_1, x_2, \dots, x_n,$$

$$\sum_{j=1}^n a_{ij}x_j = b_i, \quad i = 1, 2, \dots, n,$$

где заданная матрица

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

предполагается невырожденной.

Геометрически это означает, что ищется единственная точка

$$x_1, x_2, \dots, x_n$$

пересечения n „плоскостей“

$$a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = b_i.$$

Однако не всегда эти плоскости пересекаются достаточно хорошо. Если система возникла из реальной физически устойчивой задачи, то решение такой системы не вызовет никаких осложнений, особенно для небольшого числа n , например для $n \leq 10$. Если же рассматривается поведение физической системы вблизи состояния неустойчивости, трудности численного решения возрастают. Каждый из большого числа существующих методов решения линейных уравнений оказывается неподходящим для какой-нибудь специальной матрицы A .

Полезно ознакомиться с методом решения линейных систем на машинах-аналогах. Для этого ищут минимум суммы квадратов невязок ε_i

$$\mu = \mu(x_1, x_2, \dots, x_n) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \sum_{j=1}^n (a_{ij}x_j - b_i)^2.$$

Так как время является независимым переменным, мы превращаем μ в функцию времени и стараемся сделать ее убывающей:

$$\frac{d\mu}{dt} < 0.$$

Производная

$$\frac{d\mu}{dt} = \sum_{j=1}^n \frac{\partial \mu}{\partial x_j} \frac{dx_j}{dt}$$

является скалярным произведением вектора градиента $\partial \mu / \partial x$ и вектора скорости dx/dt . Так как

$$\frac{\partial \mu}{\partial x_j} = 2 \sum_{i=1}^n \varepsilon_i a_{ij},$$

то в качестве отрицательной обратной связи можно принять

$$\frac{dx_j}{dt} = -\varepsilon_j,$$

откуда будет следовать, что

$$\frac{d\mu}{dt} = -2 \sum_{j=1}^n \sum_{i=1}^n a_{ij} \varepsilon_i \varepsilon_j < 0,$$

если только матрица A положительно определенная. Если указанные выше условия реализовать на машине-аналоге, то величина μ будет убывать экспоненциально вместе с временем, а x_i будут сходиться

к требуемому решению. Таким образом, обычная система линейных уравнений приводит к квадратичным формам и линейным дифференциальным уравнениям.

Численное решение системы уравнений может следовать по тому же пути или же осуществляться более прямыми методами. Таковым является *метод исключения*, при котором из уравнений исключается одно неизвестное за другим, пока не останется только одно уравнение, содержащее одно неизвестное. Из полученной таким образом треугольной системы последовательно определяются все неизвестные. Кроме того, для решения систем уравнений применяются различные итерационные методы, подробно изложенные в гл. 15.

Системы порядка $n = 200$ могут быть решены на цифровых машинах, оснащенных быстродействующей внешней памятью магнитного типа. Это требует выполнения около n^3 умножений.

Другой важной матричной задачей является задача об обращении матриц. Здесь требуется построить новую матрицу

$$A^{-1} = (a'_{ij}),$$

для которой

$$\sum_{j=1}^N a_{ij} a'_{jk} = \begin{cases} 1, & \text{если } i = k, \\ 0 & \text{в противном случае,} \end{cases}$$

или, символически,

$$AA^{-1} = I.$$

Из приведенных выше рассуждений следует, что при фиксированном k элементы a'_{jk} являются решением системы

$$\sum_{j=1}^n a_{ij} x_j = \delta_{ik},$$

где δ_{ik} равно 1 или 0 в зависимости от того, выполнено или нет равенство $i = k$. Таким образом, для обращения матрицы порядка n требуется решить n систем линейных уравнений с n неизвестными, причем все системы имеют одну и ту же матрицу коэффициентов. Иными словами, не нужно решать больше чем n систем с одной и той же матрицей коэффициентов A и различными правыми частями b_j . Выгоднее построить матрицу, обратную матрице коэффициентов, и затем подставлять различные правые части b_j в формулу

$$x_i = \sum_{j=1}^n a'_{ij} b_j.$$

Существует много других практически важных методов обращения матриц. Например, можно построить последовательность матриц

$$D_0, D_1, D_2, \dots,$$

определяемых из рекуррентного соотношения

$$D_{n+1} = D_n(2I - AD_n),$$

и исследовать сходимость такой последовательности. Если последовательность D_n сходится к некоторой предельной матрице, то последняя будет обратной по отношению к матрице A .

Другой метод основан на равенствах

$$A = I - B;$$

$$A^{-1} = (I - B)^{-1} = I + B + B^2 + \dots$$

Таким образом, если эта прогрессия сходится, то обращение матрицы A может быть получено вычислением степеней матрицы B . Для сходимости прогрессии необходимо и достаточно, чтобы все собственные значения матрицы B лежали внутри единичного круга в комплексной плоскости. Для эффективности метода эти собственные значения не должны располагаться вблизи границы круга.

Другой метод обращения матрицы основан на теореме Кэли — Гамильтона (см. гл. 12), согласно которой каждая матрица удовлетворяет своему характеристическому уравнению

$$A^n - a_1 A^{n-1} + \dots + (-1)^n a_n I = 0,$$

в котором a_i являются элементарными симметрическими функциями собственных значений матрицы A и, в частности, a_n является определителем матрицы A . Если записать это уравнение в форме

$$(-1)^{n-1} a_n A^{-1} = A^{n-1} - a_1 A^{n-2} + \dots + (-1)^{n-1} a_{n-1} I,$$

то мы получим представление обратной матрицы A^{-1} (умноженной на отличный от нуля определитель a_n) в виде простой линейной комбинации степеней матрицы A .

Все эти методы в очень большой степени опираются на умножение матриц и, следовательно, на скалярное произведение

$$\mathbf{a} \cdot \mathbf{b} = \sum_i a_i b_i$$

векторов $\mathbf{a}$ и $\mathbf{b}$. При составлении программы для машины следует добиваться того, чтобы эта сумма парных произведений вычислялась как можно быстрее. Если перемножаемые векторы не могут храниться внутри машины, то для увеличения скорости вычислений необходимо предусмотреть двойной ввод данных.

Отметим в заключение, что для обращения матриц можно применять метод случайных проб, использующий теорию игр (см. гл. 9).

§ 17.5. Задачи о собственных значениях. Третий тип задач, тесно связанный с двумя предыдущими, представляют собой задачи о собственных значениях, или характеристических числах, матриц.

Для некоторого заданного λ однородная система n уравнений

$$\sum_{j=1}^n a_{ij}x_j = \lambda x_i,$$

как правило, не имеет решений, отличных от тривиального

$$x_1 = 0, x_2 = 0, \dots, x_n = 0,$$

но для дискретного множества значений λ , зависящего от коэффициентов a_{ij} , такие решения существуют. Эти значения (не обязательно вещественные или различные) называются *собственными значениями*, или *характеристическими числами*, матрицы A ; эти числа, разумеется, являются корнями алгебраического уравнения

$$\det |A - \lambda I| = 0.$$

Много различных задач в большей или меньшей степени зависят от характеристических чисел. Иногда достаточно знать только наибольшее из этих чисел λ или знаки вещественных частей всех λ . В других задачах необходимо найти достаточно точное значение всех собственных значений λ и даже решения соответствующих однородных систем.

На машинах-аналогах эта задача решается подбором (вручную или автоматически) того значения λ , при котором появляется нетривиальное решение $(x_1, x_2, \dots, x_n)$. Если даже λ — комплексное число, этим методом можно определить его вещественную и мнимую части.

На цифровых машинах применяются обычно более сложные методы.

§ 17.6. Задачи, зависящие от дискретного переменного. Скажем несколько слов о последнем типе задач — о задачах, зависящих от дискретного переменного. Многие из них являются сравнительно новыми задачами. Своим появлением и своей важностью они обязаны развитию цифровых вычислительных машин, без которых их решение было бы невозможно.

Задачи, зависящие от дискретной переменной, возникают в таких областях, как анализ операций, логика, изучение кристаллического и молекулярного строения вещества и планирование экспериментов. Они возникают также при проектировании цифровых машин и программировании для них задач. Во всех этих задачах мы имеем конечное, но часто большое число различных значений переменных, меняющихся дискретно. Всякий выбор значений переменных приводит к некоторому состоянию рассматриваемой системы, которому мы можем приписать известную оценку. Задача состоит в том, чтобы выбрать ту систему значений переменных, при которой достигается наилучшее состояние. Обычная задача анализа операций допускает следующую геометрическую трактовку. Рассмотрим выпуклый полиэдр, заданный уравнениями своих граней. Требуется найти низшую вершину

этого полиэдра. Задачи другого типа могут быть высказаны в следующем виде. Колледж состоит из 16 отделений и в каждом отделении учреждены стипендии для первокурсников. Шестнадцати кандидатам предложены испытательные тесты, и пусть s_{ij} — число очков, набранных i -ым кандидатом при решении теста, предложенного j -ым отделением. Как должны быть присуждены стипендии?

Различные методы, которыми можно решать такие задачи, требуют для своей реализации высокой скорости и логической гибкости современных цифровых вычислительных машин. Эти методы, к которым относится также метод случайных проб, или метод Монте-Карло (см. гл. 11), представляют программисту довольно большую свободу. Программирование таких задач требует как большой изобретательности, так и мелочной педантичности в отдельных деталях — сочетание, которое редко встречается у математиков с обычной подготовкой. Поэтому не удивительно, что составление части этих программ поручается самой машине. Таким образом, первоначальная цель Бэббиджа, заключающаяся в механическом составлении таблиц аналитических функций, более чем достигнута.

§ 17.7. Замечания о численном анализе. В заключение мы сделаем несколько замечаний, касающихся весьма обширного круга вопросов. При подготовке и решении задач на больших цифровых вычислительных машинах можно выделить три различных этапа:

- а) формулировка и анализ задачи;
- б) планирование вычислений;
- в) программирование.

Самым важным, конечно, является этап (а). Удачным выбором соответствующих методов и использованием специфических преимуществ данной цифровой машины можно добиться десятикратного увеличения эффективности ее работы. На этом этапе учитываются значимость решаемой задачи, с одной стороны, и стоимость содержания машины и персонала — с другой. Математик, решающий задачу, должен определить шаг применяемой сетки, сколько членов разложения необходимо взять, какие ошибки допустимы в вычислениях и сколько времени займет у машины и у обслуживающего персонала решение данной задачи. Ясно, что на этом этапе (а) математик делает ошибки и научиться здесь можно в основном только на практике.

Под *планированием* вычислений понимается тщательное составление схем вычислений и обдумывание всей последовательности действий, выбор системы счисления, формы ввода и вывода, выбор контроля вычислений и возможных изменений последовательности вычислений для устранения возникающих у машины затруднений.

Под *программированием* мы понимаем перевод схемы вычислений на доступный машине язык, т. е. кодирование команд, или инструкций, а также монтирование из соответствующих подпрограмм единой программы, готовой для введения в память машины.

На практике разница между этапами (а) и (б) не всегда столь очевидна, как это может оказаться из приведенных выше формальных определений. Программист обычно составляет схему, состоящую из отдельных ветвей и направлений, показывающих связь всей задачи в одно целое. Если задача достаточно запутана, то составляется несколько таких схем, подробно описывающих отдельные узлы основной схемы. Некоторые наиболее опытные программисты начинают не с составления схемы, а прямо с первой команды программы. Позднее можно составить схему для проверки точности программы и для того, чтобы помочь оператору в выполнении программы.

Одна из обязанностей программиста заключается в том, чтобы предупредить излишнюю потерю информации во время вычислений. Числа, введенные в машину, часто претерпевают миллионы преобразований, прежде чем будет получен и отпечатан требуемый ответ. В простейшем случае программист может выделить под каждое число отдельное „слово“ или ячейку запоминающего устройства. Хотя количество информации, помещающейся в обычную ячейку, составляет от 36 до 40 двоичных единиц, этот способ не всегда является достаточным. Программист должен сохранить безопасный курс между находящейся слева Сциллой выхода чисел из разрядной сетки и находящейся справа Харибдой ошибок округления. Надо быть уверенным, что на протяжении миллионов операций сохраняется точность, превосходящая обычную техническую точность. Если достигнуть этого нельзя, то необходимо воспользоваться другой системой изображения чисел, как, например, изображением чисел с удвоенной точностью или по методу „плавающей запятой“. В первом случае под одно число отводятся две ячейки памяти, а во втором — числа представляются в виде

$$n = \pm 2^{\alpha} n_0 \text{ или } n = \pm 10^{\alpha} n_0,$$

где

$$1 \leq n_0 < 2 \text{ или } 1 \leq n_0 < 10$$

и α — целое число, лежащее между соответствующими пределами. Хотя оба эти метода допускают оперирование с числами значительно большего диапазона, однако они связаны с большой потерей скорости вычислений и особенно в случае метода плавающей запятой, когда время решения задачи нередко удлиняется в пять раз по сравнению с обычным временем, требующимся при применении стандартного представления чисел.

Задачи программирования многочисленны и разнообразны. Они меняются при переходе от одной задачи к другой и особенно при переходе от одной машины к другой. Применяются два различных вида команд, или инструкций, называемых соответственно *одноадресными* и *многоадресными*. При одноадресной системе каждая команда определяет, какое действие должно быть выполнено в данный момент над числом, хранящимся по указанному адресу (в указанной ячейке

памяти). Например, команда может передавать число из сорок первой ячейки в один из регистров арифметического узла. Для простой передачи числа из одной ячейки в другую требуются две такие команды, так как должны быть указаны два адреса. Одноадресная система является как бы попыткой разложить всю задачу на простейшие компоненты и в то же время упростить структуру машины.

При многоадресной системе каждая команда содержит три или четыре адреса. Например, при сложении два первых адреса указывают номера ячеек, в которых хранятся слагаемые, третий адрес указывает номер ячейки, в которую должна быть отослана сумма, и, наконец, четвертый адрес указывает номер ячейки, в которой хранится следующая команда. Управляемая такой командой цифровая машина выполняет одну из своих основных операций. Здесь в одной команде заключена операция, которая на одноадресной машине требует трех команд.

Из сказанного не следует делать заключения, что многоадресная система значительно эффективнее одноадресной. С точки зрения расхода времени разница между ними очень незначительна. Так как две одноадресные команды можно хранить в одной ячейке, то и загрузка запоминающего устройства при этих системах вполне сравнима. Обычно одноадресная машина имеет около 50 различных команд, в то время как при многоадресной системе достаточно около 15 команд. Большинство выпускаемых теперь машин — одноадресного типа. С другой стороны, конструкция некоторых из них такова, что программист может использовать обе эти системы. Большинство программистов, знакомых с обеими системами, предпочитают многоадресные коды.

Программирование является довольно медленным процессом, и раз составленная программа обладает некоторой экономической ценностью. Если она достаточно обща как по своей структуре, так и по назначению, то ее можно с успехом использовать в других программах. Такое включение одних программ в другие обычно выполняется самой машиной по так называемой „монтажной программе“. Эта последняя также состоит из последовательности команд, которые, однако, оперируют не столько с самими числами, сколько с некоторыми другими командами. В некоторых задачах такие монтажные программы бывают по меньшей мере трехступенчатыми.

Эти замечания не должны отпугивать инженеров от самостоятельного планирования вычислений и даже составления самой программы. Для постановки задачи на машине достаточно иметь только список выполняемых машиной операций и небольшое описание основных ее характеристик. Природная интуиция поможет инженеру найти большую часть приемов программирования, таких, как преобразования команд или вход в подпрограмму и выход из нее.

Содержание

Предисловие редактора русского перевода	5
К читателю	7
Предисловие	8
Введение	11

ЧАСТЬ 1. МАТЕМАТИЧЕСКИЕ МОДЕЛИ

<i>Глава 1. Линейные и нелинейные колебания. Соломон Лефшец. (Перевод В. И. Тумаркина)</i>	17

§ 1.1. Введение	17
§ 1.2. Гармонические колебания	17
§ 1.3. Затухающие колебания	20
§ 1.4. Вынужденные колебания	22
§ 1.5. Линейные и нелинейные системы	23
§ 1.6. Некоторые нелинейные системы	25
§ 1.7. Нелинейные колебания в консервативных системах	28
§ 1.8. Вынужденные нелинейные колебания	30
§ 1.9. Контур мультивибратора	31
§ 1.10. Математическое исследование нелинейных задач	32
§ 1.11. Приближенные методы	36
§ 1.12. Метод Дуффинга	38
§ 1.13. Метод возмущений, предложенный Пуанкаре	40

Литература	41

<i>Глава 2. Исследование равновесия. Теория устойчивости Пуанкаре и Ляпунова. Ричард Беллман. (Перевод В. И. Тумаркина)</i>	42

§ 2.1. Введение	42
§ 2.2. Теория устойчивости Пуанкаре и Ляпунова	43
§ 2.3. Теория устойчивости линейных уравнений	45
§ 2.4. Дифференциально-разностные уравнения	46
§ 2.5. Уравнение теплопроводности	46

Литература	47

Глава 3. Внешняя баллистика. Джон В. Грин. (Перевод Г. В. Коренева)	48
§ 3.1. Введение	48
§ 3.2. Выбор системы координат	49
§ 3.3. Аэродинамические силы, действующие на снаряд	53
§ 3.4. Уравнения движения	60
§ 3.5. Баллистические таблицы и таблицы стрельбы	61
§ 3.6. Поправки на малые влияния	64
§ 3.7. Бомбометание с самолетов	69
§ 3.8. Влияние аэродинамических сил, отличных от силы лобового сопротивления	70
§ 3.9. Заключение и обзор литературы	71
Литература	72
 Глава 4. Элементы вариационного исчисления. Магнус Р. Хестенс. (Перевод Л. Д. Кудрявцева)	 73
§ 4.1. Введение	73
§ 4.2. Некоторые элементарные вариационные задачи	74
§ 4.3. Основные задачи. Необходимые условия минимума	79
§ 4.4. Вывод уравнения Эйлера	82
§ 4.5. Специальные случаи	85
§ 4.6. Случай, когда подинтегральная функция имеет вид $f(x, y)$	89
§ 4.7. Принцип Гамильтона	90
§ 4.8. Гамильтонианы	94
§ 4.9. Изопериметрическая задача	96
§ 4.10. Задачи с подвижными концами	99
§ 4.11. Минимумы функций от интегралов	102
§ 4.12. Задача Больца	103
§ 4.13. Задачи с кратными интегралами	106
Литература	108
 Глава 5. Гиперболические дифференциальные уравнения с частными производными и их приложения. Рихард Курант. (Перевод Н. Д. Введенской)	 109
§ 5.1. Введение	109
§ 5.2. Связь между дифференциальными уравнениями с частными производными и действительностью	110
§ 5.3. Статистические процессы и дифференциальные уравнения с частными производными	111
§ 5.4. Классификация линейных дифференциальных уравнений с частными производными. Плоские волны	113
§ 5.5. Задача Коши для волнового уравнения	115
§ 5.6. Нелинейные гиперболические уравнения	119
§ 5.7. Конечно-разностный метод	123
Литература	128

Глава 6. Краевые задачи для эллиптических дифференциальных уравнений с частными производными. Менахем М. Шиффер. (Перевод Н. Д. Введенской)	129
§ 6.1. Что такое корректно поставленная задача для дифференциальных уравнений с частными производными?	129
§ 6.2. Теория теплопередачи; три основные краевые задачи	131
§ 6.3. Фундаментальные решения и функции Грина	135
§ 6.4. Принцип максимума, ядро, интеграл Дирихле	141
§ 6.5. Примеры из гидродинамики и электростатики	148
§ 6.6. Изменение функций Грина с изменением области	155
§ 6.7. Изменение функций Грина при изменении коэффициентов дифференциального уравнения	162
Литература	165

Глава 7. Краевые задачи теории упругости. И. С. Сокольников. (Перевод Г. И. Баренблатта)	166

Постановка задач

§ 7.1. Введение	166
§ 7.2. Два основных типа задач	166
§ 7.3. Характеристики смещений; деформация	167
§ 7.4. Характеристики напряженного состояния	168
§ 7.5. Условия равновесия	169
§ 7.6. Плоские и пространственные задачи	170
§ 7.7. Цилиндр; приложенные силы постоянны по его длине	171
§ 7.8. Цилиндр; приложенные силы распределены на торцах	173
§ 7.9. Цилиндр; комбинированная задача	174

Метод решения

§ 7.10. Вычислительные соображения	174
§ 7.11. Формулировка задачи в терминах теории функций комплексного переменного	175
§ 7.12. Задача Дирихле	178
§ 7.13. Бигармоническая краевая задача	180
§ 7.14. Заключение	182
Литература	183

ЧАСТЬ 2. ТЕОРЕТИКО-ВЕРОЯТНОСТНЫЕ ЗАДАЧИ

Глава 8. Теория предсказания. Норберт Винер. (Перевод В. М. Золотарева)	185
§ 8.1. Временные ряды и причинность	185
§ 8.2. Эргодическая теорема	185
§ 8.3. Временные ряды	193

§ 8.4. Предсказание сообщений, заданных не полностью	203
§ 8.5. Предсказание в случае непрерывного времени	206
§ 8.6. Многомерное предсказание и причинность	208
Литература	214
Дополнительная литература	215

Глава 9. Теория игр. Г. Фредерик Боненбласт. (Перевод В. М. Золотарева) 216

§ 9.1. Введение	216
§ 9.2. Предмет теории	217
§ 9.3. Оптимальные чистые стратегии	219
§ 9.4. Пример из теории планирования	220
§ 9.5. Смешанные стратегии	223
§ 9.6. Симметрия; пример тактического выбора момента времени	225
§ 9.7. Соотношение $\max\text{-}\min = \min\text{-}\max$	227
§ 9.8. Приложение к экономике	229
§ 9.9. Последовательные приближения	232
§ 9.10. Общее понятие стратегии; игра, подобная покеру	233
§ 9.11. Заключение	235
Литература	235

Глава 10. Теория динамического планирования. Ричард Беллман. (Перевод Г. Н. Поварова) 237

§ 10.1. Введение	237
§ 10.2. Предварительные рассмотрения	238

Оптимальное распределение ресурсов

§ 10.3. Введение	240
§ 10.4. Задача о закупках	240
§ 10.5. Классическая постановка задачи	240
§ 10.6. Постановка задачи с точки зрения динамического планирования	241
§ 10.7. Бесконечноэтапная аппроксимация	242
§ 10.8. Теорема существования и единственности	243
§ 10.9. Аналитические результаты	245
§ 10.10. Методика вычислений	246
§ 10.11. Неопределенность	249

Уравнение золотых приисков

§ 10.12. Введение	250
§ 10.13. Задача о золотых приисках	250
§ 10.14. Математическая постановка задачи	251
§ 10.15. Существование и единственность	251
§ 10.16. Решение	252
§ 10.17. Интерпретация решения	253

§ 10.18. Нелинейная полезность	254
§ 10.19. Непрерывный вариант	255
§ 10.20. Вывод дифференциальных уравнений	257
§ 10.21. Решение	258

Задачи на узкие места

§ 10.22. Введение	259
§ 10.23. Типичная задача	259
§ 10.24. Непрерывный вариант	262
§ 10.25. Некоторые замечания об обозначениях	263
§ 10.26. Основное функциональное уравнение	264
§ 10.27. Аналогия в бесконечно малом	264
§ 10.28. Двойственная задача	265
§ 10.29. Более общие задачи	268
§ 10.30. Задачи регулирования	268

Игры на выживание

§ 10.31. Введение	269
§ 10.32. Игры на выживание	270
§ 10.33. Суммарно ненулевые игры на выживание	270

Литература	271

Глава II. Методы Монте-Карло. Дж. В. Браун. (Перевод И. М. Соболя)	275

Введение

§ 11.1. Природа методов Монте-Карло	275
§ 11.2. Основные статистические понятия	276

Оценка интегралов методом Монте-Карло

§ 11.3. Случайная выборка как метод вычисления	279
§ 11.4. Частный случай	281
§ 11.5. Тривиальный числовой пример	285
§ 11.6. Приложения, приводящие к кратным интегралам	286

Методы Монте-Карло, связанные с процессами диффузии

§ 11.7. Равномерное случайное блуждание на плоскости	288
§ 11.8. Уравнение Лапласа	289
§ 11.9. Обобщения	294

Методы выборки из заданных распределений

§ 11.10. Источники случайных чисел	297
§ 11.11. Преобразования случайных чисел	298

Литература	299

ЧАСТЬ 3. ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Глава 12. Матрицы в технике. Луис А. Пайпс. (Перевод Л. Е. Садовского)	305
§ 12.1. Введение	305

Основные положения матричной алгебры

§ 12.2. Определение матрицы	305
§ 12.3. Основные типы матриц	306
§ 12.4. Определения и основные алгебраические операции	307
§ 12.5. Умножение матриц	307
§ 12.6. Обратная матрица. Деление матриц	310
§ 12.7. Разбиение матрицы на подматрицы	311
§ 12.8. Характеристическая матрица и характеристическое уравнение матрицы	312
§ 12.9. Приведение матрицы к диагональной форме	313
§ 12.10. След матрицы	315
§ 12.11. Теорема Кэли — Гамильтона	315
§ 12.12. Решение линейных уравнений и обращение больших матриц	317

Операции над матрицами

§ 12.13. Дифференцирование матриц	320
§ 12.14. Интегрирование матриц	320
§ 12.15. Многочлены от матриц	320
§ 12.16. Функции матриц	320
§ 12.17. Представление матричной функции с помощью многочлена	321
§ 12.18. Теорема Сильвестра	322
§ 12.19. Решение алгебраических уравнений с помощью матриц	323

Интегрирование линейных дифференциальных уравнений

§ 12.20. Матрициант	324
§ 12.21. Линейные уравнения с постоянными коэффициентами	325
§ 12.22. Линейные уравнения с переменными коэффициентами; метод осредненных коэффициентов	326

Приложение матриц к задачам об упругих системах

§ 12.23. Главные направления нагрузок в точке упругого тела	327
§ 12.24. Приложение к теории конструкций	329

Приложения матриц к задачам электротехники

§ 12.25. Задачи, относящиеся к сетям переменного тока	331
§ 12.26. Задачи на схемы трехфазного тока	333
§ 12.27. Теория четырехполюсников	335
§ 12.28. Распространение волн по каскаду из симметричных четырехполюсников	336

Приложения матриц к задачам о колебаниях

§ 12.29. Колебания консервативных систем	338
§ 12.30. Колебания неконсервативных систем	343
§ 12.31. Заключение	344
Литература	344

Глава 13. Методы конформных отображений. Эдвин Ф. Беккен-
бах. (Перевод М. Т. Тер-Микаэляна) 346

§ 13.1. Введение	346
§ 13.2. Картографические проекции	346
§ 13.3. Конформные отображения и теория функций комплексного переменного	353
§ 13.4. Уравнение Лапласа в физике	358
§ 13.5. Теоремы существования	361
§ 13.6. Геометрические свойства конформных отображений	365
§ 13.7. Специальные функции	367
§ 13.8. Приложения	370
§ 13.9. Построение конформных отображений	371
Литература	374

Глава 14. Нелинейные методы. Чарльз Б. Моррей. (Перевод
Л. Е. Садовского) 375

§ 14.1. Введение	375

Системы алгебраических уравнений

§ 14.2. Метод Ньютона. Метод возмущений. Модификации этих методов	376
§ 14.3. Замечания о системах линейных уравнений	380
§ 14.4. Замечания о решении обыкновенных дифференциальных уравнений	381
§ 14.5. Метод непрерывности. Метод бесконечных степенных рядов	383
§ 14.6. Замечания, касающиеся метода быстрейшего спуска	385

Векторные обозначения

§ 14.7. Векторные обозначения. Операции над векторами. Длина	387
§ 14.8. Дифференцируемые функции и векторные функции	388
§ 14.9. Образец доказательства сходимости. Теорема об обратном отображении	390

Нормированные линейные пространства

§ 14.10. Абстрактные пространства	392
§ 14.11. Линейные пространства	392
§ 14.12. Нормированные линейные пространства	394
§ 14.13. Полнота	395

**Функциональные уравнения, вариационные задачи и
нормированные линейные пространства**

§ 14.14. Скалярные функции. Функционалы	397
§ 14.15. Простые вариационные задачи	398
§ 14.16. Векторные уравнения в нормированных линейных про- странствах	402
§ 14.17. Замечания о линейных уравнениях	405

Численные методы

§ 14.18. Метод Рунге — Кутты и некоторые модификации вари- ационных задач	405
§ 14.19. Замечания о методе Рунге — Кутты и методах конечных разностей для решения функциональных уравнений	407
§ 14.20. Приложение	408
Литература	417

Глава 15. Что представляют собой релаксационные методы?

Джордж Э. Форсайт. (Перевод Л. Е. Садовского) . .	418
§ 15.1. Введение	418
§ 15.2. Решетки и аналогичные задачи	418
§ 15.3. Другие источники получения систем уравнений, решаемых методом релаксации	423
§ 15.4. Различные методы решения систем линейных уравнений . .	426
§ 15.5. Математический анализ покомпонентной релаксации . . .	429
§ 15.6. Вычисление частот колебаний	434
§ 15.7. Заключение	437
Литература	437

**Глава 16. Методы быстрого спуска. Чарльз Б. Томпкинс.
(Перевод О. М. Белоцерковского)**

§ 16.1. Введение	441
§ 16.2. Многомерная аналитическая геометрия и метод Качмажа для линейных уравнений	443
§ 16.3. Быстрый спуск и дифференциальные уравнения быстрей- шего спуска для функций, определенных в многомерном евклидовом пространстве	449
§ 16.4. Замечания о численном решении дифференциальных урав- нений быстрейшего спуска. Желательность больших шагов. Метод сопряженных градиентов Хестенса — Штифеля . . .	453
§ 16.5. Спуск при наличии связей	457
§ 16.6. Обобщение метрики и расширение на случай пространств бесконечно большого числа измерений; проблемы вари- ационного исчисления	460
§ 16.7. Обзор математического исследования точек пространства, для которого $J \leq c$	467
§ 16.8. Заключение	470
Литература	472

Глава 17. Быстродействующие вычислительные машины и их применения. Деррик Г. Лемер. (Перевод Т. М. Тер-Микаэляна)	473
§ 17.1. Введение	473
§ 17.2. Машины-аналоги	474
§ 17.3. Цифровые вычислительные машины	478
§ 17.4. Матричные задачи	484
§ 17.5. Задачи о собственных значениях	487
§ 17.6. Задачи, зависящие от дискретного переменного	488
§ 17.7. Замечания о численном анализе	489

СОВРЕМЕННАЯ МАТЕМАТИКА ДЛЯ ИНЖЕНЕРОВ

Редакторы А. В. ГЕРМОГЕНОВ и В. В. ЗАРУЦКАЯ

Технический редактор Н. И. Смирнова

Художник Б. Н. Гладков

Сдано в производство 22/V 1958 г. Подписано к печати 16/X 1958 г. Бумага 60×92¹/₁₆ = 15,6 бум. л.
31,2 печ. л. Уч.-изд. л. 30,2. Изд. № 1/4094. Цена 23 р. 15 к. Зак. 3190.

ИЗДАТЕЛЬСТВО ИНОСТРАННОЙ ЛИТЕРАТУРЫ

Москва, Ново-Алексеевская, 52.

Типография № 2 им. Евг. Соколовой УПП Ленсовнархоза,
Ленинград, Измайловский пр., 29.

ОПЕЧАТКИ

Стр.	Строка	Напечатано	Следует читать
57	20 св.	$M_3 = \beta v_2 = \gamma \omega_3$	$M_3 = \beta v_3 + \gamma \omega_3$
87	4 сн.	$\frac{x-a}{c}$	$\frac{x-a}{b}$
91	9 сн.	$q_i L_{q_i} - L = T + U$	$\dot{q}_i L_{\dot{q}_i} - L = T + U$
129	3 сн.	$g(x - jy) j^2 = -1$	$g(x - jy), j^2 = -1$
135	3 св.	$\nabla U - \nabla V$	$\nabla U \cdot \nabla V$
150	1, 18 сн.	φ	Φ
151	3—21 св.	φ	Φ
242	13 св.	$h(x - y) f_N [ay + b(x - y)]$	$h(x - y) + f_N [ay + b(x - y)]$
266	4 сн.	$\int_t^T$	$\int_0^T$
266	1 сн.	$\min$	$\min_z$

